VIP: Visualization of Integrated Proteomics Data

Eugenia G. Giannopoulou, George Lepouras, and Elias S. Manolakos, *Senior Member, IEEE*

Abstract—**The post-genomic era is characterized by the rapid data accumulation leading to unwieldy and large volumes of biological data. The proteomics results (large sets of identified proteins or peptides) that originate from several workflow steps, play an important role in the analysis of a proteomics experiment. As a result, the area of high-throughput proteomics created new visualization challenges in interpreting large-scale datasets. We present VIP, an interactive visualization tool for proteomics data, which integrates protein or peptide features emanating at every step of the proteomics analysis and combines them visually in synthetic maps. Our novel tool offers the flexibility to choose any desired features according to the analysis objectives, examine simultaneously more than one maps and interact with the visualization by querying and filtering the results. The synthetic maps visualization aims not only at summarizing proteomics experiments in a unified manner for both 2DE-MS and LC-MS based analyses, but also at providing a quick and combined overview of protein/peptide features, thus facilitating the data analysis and interpretation.**

I. Introduction

PROTEOMICS was initially defined as the large-scale study of the functions of all expressed proteins within an organism. Nowadays, it also involves the set of all protein isoforms and modifications as well as the interactions between them [1], [2]. In other words, proteomics expanded to the point that incorporates information that could be characterized as "post-genomic". Importantly, differential proteomics stands for the comparison of proteomes in two or more biological states and aims at finding proteins that can be reliable biomarkers for these different states [3].

A proteomics analysis consists of several sequential steps, including the separation, differential expression analysis, protein identification, meta-data analysis and possibly even more. All these steps create a plethora of protein (or peptide) -related data (Figure 1), coming from different software tools, that play an important role in the analysis of the proteomics results. We call these data *proteomic features,* and we will use this notion throughout the paper extensively.

Typically, the proteomics workflow starts with the application of a separation technique (e.g., 2-D Gel Electrophoresis (2DGE) or Liquid Chromatography (LC)). During this step, the protein spots in a 2DGE-MS experiment are associated with features such as the *isoelectric point* (pI), the *molecular weight* (MW) (i.e., the two dimensions of a 2D gel), the spot *volume* and so on, all coming from the image analysis software used. At the differential expression analysis step several statistical and quantitative methods are applied (e.g., Student's t-test and fold factor criterion) in order to find the spots that discriminate the biological conditions. Hence, protein spots obtain features such as the *p-value* of a statistical test and the *volume fold factor*, to name a few. In the step of the mass spectrometry (MS)-based identification, each protein is associated with a *mass-to-charge* (m/z) ratio, a *score* value, as well as other features resulting from the search engines used to achieve protein identification. Similarly, in the case of the LC-MS separation method, the corresponding analysis steps accumulate proteomic features for each peptide or protein. Finally, the step of meta-data analysis (e.g., application of data mining and statistical methods [4], [5], Gene Ontology annotation [6], [7], pathway information retrieval [8]) follows the protein identification and attaches additional features to proteins and peptides. This step is of great importance since it focuses on inferring protein-protein interactions, understanding how proteins are organized into biological networks and, generally speaking, exploring proteomics data from a systems biology perspective.

The researchers' focus on the promising high-throughput proteomics analysis, leads to large volumes of data [9], which are definitely altering the field of life sciences and create the need for new areas of expertise, not only in terms of data storage and handling, but also in terms of information visualization. Working on extended lists of identified and possibly differentially expressed proteins, 2D gel images and database results (i.e., data usually heterogeneous and distributed in multiple computers) is an ineffective and time-consuming task for a biologist who attempts to discover biologically relevant relations suggested by her/his results. Therefore, the proteomics visualization has to deal with new challenges [10] which involve disengaging the proteomics data analysis from non-friendly tables and hand-annotated images, and providing new ways to combine data stemming from multiple proteomic steps in a way that meaningful biological assumptions can be derived.

In this work we present *VIP*, a visualization tool which

This paper is part of the 03ED306 research project, implemented within the framework of the "Reinforcement Programme of Human Research Manpower" (PENED) and co-financed by National and Community Funds.

E. G. Giannopoulou is with the Department of Computer Science and Technology, University of Peloponnese, Tripolis, 22100 Greece (e-mail: egian@uop.gr).

G. Lepouras, is with the Department of Computer Science and Technology, University of Peloponnese, Tripolis, 22100 Greece (e-mail: G.Lepouras@uop.gr)

E. S. Manolakos is with the Department of Informatics and Telecommunications, University of Athens, Athens, 15784 Greece (phone: 0030-210-7275312; fax: 0030-210-7275214; e-mail: eliasm@di.uoa.gr).

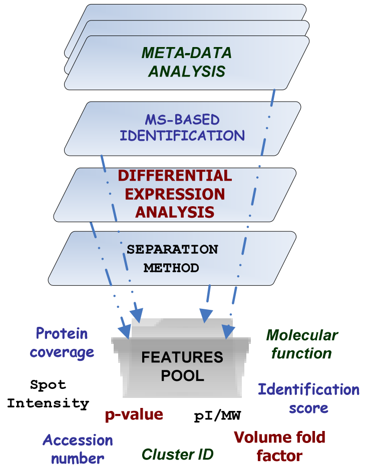


Fig. 1. The accumulation of features per protein/peptide produced by the proteomics analysis steps.

integrates data (i.e., protein/peptide features) from all steps of the proteomics workflow and combines them visually in easy to interpret synthetic maps. The visualizations provided by VIP assist in summarizing a 2DGE-MS or LC-MS experiment based on user-selected proteomic features and in creating scenarios that complement the interpretation of the proteomics analysis results. The novel aspect of VIP lies on the visual exploration of the integrated proteomics information, which eventually leads in revealing patterns, or trends, that could not be discovered by examining the proteomics data from each step separately. To the best of our knowledge, VIP is the first tool which incorporates the notion of integrated proteomic features and provides their joint visualization on synthetic maps.

The remainder of the paper is organized as follows: Related studies are presented and discussed in Section II. The VIP system overview is thoroughly explained in section III, followed by the architectural and implementation details in section IV. In Section V we provide some examples of synthetic maps produced by VIP and finally, summarize our work and discuss future research directions in Section VI.

II. Related Work

As far as 2DGE-MS is concerned, there are several software suites for handling and visualizing proteomics data. Genedata Expressionist™ [11], and Bruker's Proteinscape [12] are indicative examples. Using the Expressionist suite, users can link the actual 2D gel images with statistical results (e.g., bar charts) to present the same information in various formats for comparison and verification. Proteinscape links the spots in a 2D gel image with their identification data and annotates them with a colored cross according to the level of identification. Yet, these platforms not only lack in the flexibility of combining any desired features throughout a 2DGE-MS analysis, but also do not offer the vital capability of visualizing them jointly.

Among the numerous 2DGE image analysis tools, Delta2D [13] stands out for its impressive differential display. In particular, by enabling the "ratio mode", the spots on a gel image get color-coded according to the intensity ratio between two different gels, thus assisting in locating quickly the "interesting" spots. However, Delta2D does not consider the visual combination of 2D-gel spots with MS-related information, apart from labeling the spots with the respective identified protein names.

Regarding the LC-MS technique, existing efforts focus on producing visualization tools that cope with the large datasets produced by this technique. Pep3D, described in [14], [15] is a tool which summarizes an LC-ESI-MS experiment by placing the peaks (i.e., peptides) in a 2D gel-like image (also called "density plot"), using the RT (retention time) feature, from LC, and the m/z feature, from MS, for the two dimensions. In a Pep3D image two more features can be displayed by colored boxes: the score values of peptide identifications and the precursor ions selected for fragmentation. However, the main drawbacks of Pep3D are: (a) the visualization of a single experiment at a time, and (b) the difficulty to distinguish the color difference of the identification score due to the small size of the boxes used.

In the tool presented by Linsen et al. [16], which displays LC-MS data in a 3D space, color is used to depict the intensity values of each peak, the condition under which the maximum peak intensity has been observed and the up/down regulated peaks. In the visualization presented by Turner et al. [10], color is also used to display the ratio of differential expression levels of identical peptides in two different data sets, in a 2D plot (RT vs. m/z). However, these efforts are limited on visualizing specific features using color only.

Despite the various attempts to visualize proteomics data, either from 2DGE-MS or LC-MS, the existing tools fail to allow for simultaneous visualization of multiple features from different stages of the proteomics analysis workflow.

Our tool moves a step beyond these efforts, by integrating protein (or peptide) features that originate from different proteomic steps and supporting their joint visualization on synthetic maps. These maps contain glyphs (i.e., geometric shapes) that represent proteins or peptides. Every proteomic feature is assigned to a glyph's visualization attribute, such as size, color, label and so on. VIP offers the user the remarkable flexibility to select any desired features and create multiple maps for her/his dataset, thus allowing her/him to explore an experiment from multiple perspectives. We will show that the visualization of integrated proteomic features on synthetic maps, serves as am effective mechanism for information extraction and interpretation of proteomics results.

III. System Requirements

Our work aims at designing and implementing a novel and interactive tool that provides multidimensional visualization of user-defined proteomic features on, easy to grasp and interpret, synthetic maps. These maps assist the biologist in the important task of data analysis of the output of high-throughput proteomics experiments. Our tool also offers data integration of protein (or peptide) features coming from multiple steps of a proteomics workflow, in order to create an abstraction layer, for the features, on which visualization will be applied. We have exploited the knowledge we gained by participating in proteomics studies [17], in order to partition our system into components that meet the end-user needs during a proteomics data analysis. These components can be implemented independently using tools from different scientific fields. Based on the gathered functional requirements, in this section, we describe the main components of VIP: (A) *Data Import and Unified Representation*, (B) *Features Management*, (C) *Visualization Engine* and (D) *Graphical User Interface*.

A. Data Import and Unified Representation

The plethora of tools used in proteomics, result in a great variety of, usually incompatible, data formats. Hence, there is a need to collect all proteomics data and create a common interface which will facilitate the independent modular design of the components that follow (i.e., Features Management and Visualization Engine). To overcome the diversity of existing formats used in proteomics, the first component of our tool offers data aggregation, and ensures the capability of importing any type of features (e.g., numerical, categorical) deriving from any software or database. Moreover, this component facilitates the efficient manipulation and visualization of features by supporting their storage in well-designed data structures (see section IV.A.2). This way, we also offer a familiar representation to the user, since the used data structures imitate the real proteomics data hierarchy (e.g., project → experiment → category → gel → spot).

B. Features Management

This component captures the user's need for intervention over the proteomic features by offering: (a) preprocessing and (b) graphical encoding capabilities.

As far as the features preprocessing is concerned, this component must allow the user to transform her/his data according to her/his needs (see section IV.B.1).

Graphical encoding (i.e., mapping features to visualization attributes) is a very important task for, if it is well designed, it can ensure rapid awareness of the different features of a visualized object, while alleviating the need for excessive cognitive effort from the user [18]. This module gives the user the capability to select the desired features, perform their mapping to the visualization attributes offered for a concrete glyph, and finally forward them to the visualization engine. Therefore, if the proteins are displayed as circles, s/he can assign the *abundance ratio* feature to the circle's color and the *identification score* to the radius, providing an indication of the proteins differential expression and identification confidence respectively.

Here, we should note that if it had not been for the abstraction layer provided by the data aggregation mentioned earlier, it would be far more complicated to design this component due to the variety of proteomics formats, that would dictate creating multiple graphical encoding and preprocessing schemes.

C. Visualization Engine

The Visualization Engine satisfies the visualization demands by creating the synthetic maps, using the previous component's output (i.e., the user selected and mapped features), and offering several capabilities from the human-computer interaction perspective [19], to facilitate the maps exploration. The synthetic maps visualization resembles the 2D gel images, thus assisting the user to perceive her/his data using a familiar representation. This component offers basic navigation techniques for the maps (e.g., zooming, rotation, translation), as well as advanced distortion techniques [20] that aim to deal with the problem of displaying a large information space through a relatively small window (e.g., fisheye view, perspective wall).

Importantly, the component should also support the concurrent visualization of multiple synthetic maps, as well as interaction between them, in order to facilitate the differential expression analysis. Visual queries [21] that allow the user to filter dynamically the proteins/peptides on a synthetic map should also be supported in order to facilitate trial-and-error efforts in searching the large volume of the proteomics results.

D. Graphical User Interface

The Graphical User Interface is an essential component for it ensures the interoperability among all previous components and offers the user effective and quick ways to interact with the features and the derived maps. An informal user requirement analysis revealed that it must allow the user to import, aggregate and store the proteomic features, select the visualization parameters, create synthetic maps, perform queries and interact with the visualization. Finally, it should incorporate multiple controls and interaction procedures so that the overall functionality is efficiently supported and provide an intuitive, user-friendly and easy-to-use interface.

IV. Architecture and Implementation

This section describes the architectural details, basic characteristics and benefits of every VIP module (Figure 2), stemming from the corresponding component of the previous section. In particular, we give an inside view of every module and justify our design choices. Additionally, this section serves as a summary for all best practices we distilled from the design and realization of every module. Specific implementation details are also provided at the end.

A. Backend

This module is composed of two sub-modules which have been designed based on the requirements of component A.

1) Data Aggregator: To support the proteomics data

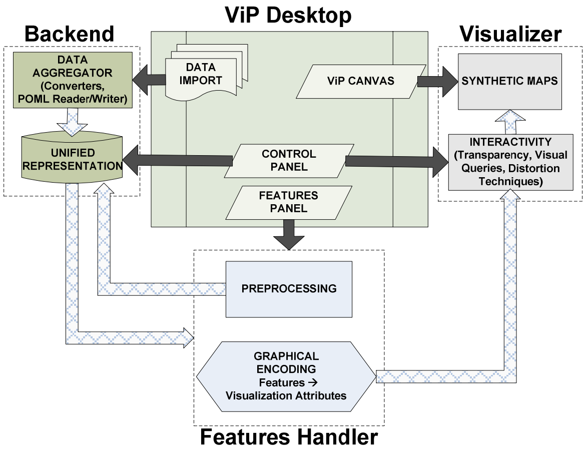
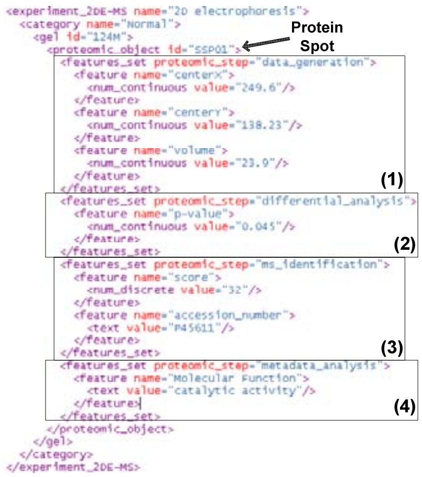

ViP Desktop
Backend
Visualizer
DATA AGGREGATOR (Converters, POML Reader/Writer)
DATA IMPORT
VIP CANVAS
SYNTHETIC MAPS
UNIFIED REPRESENTATION
CONTROL PANEL
INTERACTIVITY (Transparency, Visual Queries, Distortion Techniques)
FEATURES PANEL
PREPROCESSING
GRAPHICAL ENCODING Features → Visualization Attributes
Features Handler

Fig. 2. Schematic overview of the VIP architecture. VIP Desktop ensures the connectivity among Backend, Features Handler and Visualizer.

aggregation we have established *POML*, ***P**roteomic **O**bjects **M**arkup **L**anguage*, a unified XML schema for representing the features that originate from multiple proteomic steps and concern a large number of proteins or peptides. Among the numerous advantages of XML (extensibility, system independency, lossless exchange of data between systems using different formats), we exploit its capability to: (1) preserve the existing hierarchy among proteomics concepts and more importantly (2) add a semantics perspective to raw data, by using XML schema validators (e.g., XSD). POML is designed to represent all available protein/peptide-related features (numerical or categorical) deriving from *any* proteomic step, along with the corresponding experimental information. The design and full description of POML is beyond the scope of this paper. In Figure 3 we only present and discuss a fragment of POML, to complete the module description. Currently, VIP relies on conversion tools in order to transform native data files, from different software tools used in proteomics, into POML data files. Although this need is mandatory at the time, since POML is under development, in the future we hope this to change with the adoption of POML by standard proteomics tools.

2) Unified Representation: The internal representation of proteins/peptides and their features, is based on associative data structures, which support encapsulation and mimic the actual hierarchy of proteomics data. More specifically, VIP is based on associative data containers, collections of keys and values, where each key is connected to one and only value. The most essential operation offered by an associative container, is finding the value that is coupled to a key (also called lookup or indexing). For example, to obtain a spot from the *Gel* container is as simple as to search for the *spot name* (i.e., key) and retrieve the *Spot* container (i.e., value). Similarly, the Spot container includes feature names as keys and feature objects as values.

The selection of data structures for the unified representation of proteomic features is a critical task, since the data structures are used by all modules of VIP and have an impact on its overall efficiency. Our choice to use associative data structures achieves the required efficiency, since we can succeed a very rapid lookup that is independent of the data volume, in contrast to linear data structures for example, where both the look up and the insertion costs increase proportionally to the size of the data.



```
<experiment_2DE-MS name="2D electrophoresis">
  <category name="Normal">
    <gel id="124M">
      <proteomic_object id="SSP01">
        <features_set proteomic_step="data_generation">
          <feature name="centerX">
            <num_continuous value="249.6"/>
          </feature>
          <feature name="centerY">
            <num_continuous value="138.23"/>
          </feature>
          <feature name="volume">
            <num_continuous value="23.9"/>
          </feature>
        </features_set>
        <features_set proteomic_step="differential_analysis">
          <feature name="p-value">
            <num_continuous value="0.045"/>
          </feature>
        </features_set>
        <features_set proteomic_step="ms_identification">
          <feature name="score">
            <num_discrete value="32"/>
          </feature>
          <feature name="accession_number">
            <text value="P45611"/>
          </feature>
        </features_set>
        <features_set proteomic_step="metadata_analysis">
          <feature name="Molecular Function">
            <text value="catalytic activity"/>
          </feature>
        </features_set>
      </proteomic_object>
    </gel>
  </category>
</experiment_2DE-MS>
```

Fig. 3. Indicative example of a POML file: A gel contains protein spots (i.e., proteomic objects), the gels belong to a category (describing the biological state) and the categories belong to an experiment (in this case a 2DGE based experiment). Each protein spot has sets of features which emanate at a specific step of the proteomics workflow (e.g., spot "SSP01" gets the *volume*="23.9" feature from the separation step (1), the *p-value*= "0.045" from differential analysis step (2), the *score*="32" from the mass spectrometry identification step (3), the *molecular function*="catalytic activity" from the metadata analysis step (4)).

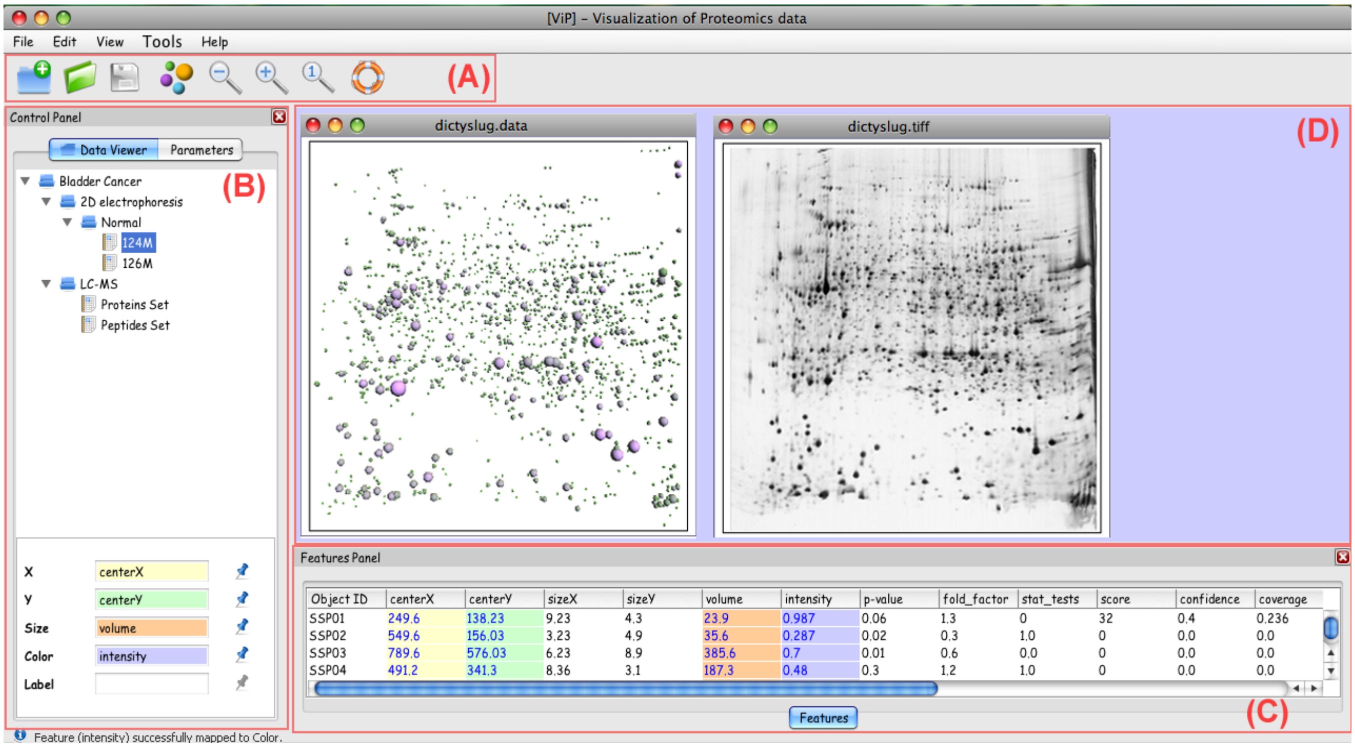
B. *Features Handler*

The Features Handler module is responsible for the features preprocessing and graphical encoding.

We chose to show the imported proteomic features as columns of a table (Figure 4-C), whose rows represent the proteins/peptides. By selecting a column (i.e., a feature), the user can either choose to: (1) perform preprocessing to the values of the feature, or (2) map the feature to a specific visualization attribute of a glyph.

1) Preprocessing: The preprocessing module includes: (a) appropriate forms that guide the user to select and perform a transformation on the selected feature, (b) controls that aim at minimizing user errors (e.g., prevent the user from performing normalization on text values) and (c) sub-modules that offer standard transformation (e.g., scaling numerical values so as to give a better visualization result, log transformation to account for a large dynamic range, converting categorical data into numerical).

2) Feature-to-Attribute Mapping: Through this module, VIP allows the user to select: (a) which features will move ahead to the visualization engine and (b) the appropriate graphical encoding of the feature. In particular, the user chooses the visualization attribute that best suits her/his needs, from a list of available attributes, such as color, size, label, texture, to name a few. The attributes that could be used are many and is part of further study to find the ones that are most suitable for a specific proteomic feature, based on the feature type, range of values and significance. Importantly, this module keeps a record of the user's previous selections and of her/his feature-to-attribute



X	centerX
Y	centerY
Size	volume
Color	intensity
Label	

Feature (intensity) successfully mapped to Color.

Object ID	centerX	centerY	sizeX	sizeY	volume	intensity	p-value	fold_factor	stat_tests	score	confidence	coverage
SSP01	249.6	138.23	9.23	4.3	23.9	0.987	0.06	1.3	0	32	0.4	0.236
SSP02	549.6	156.03	3.23	4.9	35.6	0.287	0.02	0.3	1.0	0	0.0	0.0
SSP03	789.6	576.03	6.23	8.9	385.6	0.7	0.01	0.6	0.0	0	0.0	0.0
SSP04	491.2	341.3	8.36	3.1	187.3	0.48	0.3	1.2	1.0	0	0.0	0.0

Fig. 4. The VIP Desktop. (A) The *Toolbar* contains the basic command buttons. (B) The *Control Panel* holds the *Data Viewer* and the *Parameters* panes. (C) The *Features Panel* shows the proteins/peptides along with their features and allows their handling. (D) The *VIP Canvas* displays the synthetic maps. Here, *VIP Canvas* shows a protein-based map and its corresponding 2D gel image.

mapping history to facilitate or enhance the decision-making process.

C. *Visualizer*

The Visualizer constructs synthetic maps according to user-selected features and provides interaction mechanisms.

The synthetic maps are 2D panels that reproduce the 3D space by exploiting 3D glyphs (e.g., spheres) and interaction methods. More specifically, the Visualizer offers 3D rotation, moving and zooming capabilities for the maps, using the mouse. Currently, we support two types of synthetic maps: (a) the protein-based and (b) the peptide-based. Protein-based maps summarize visually spots/proteins from a 2DGE-MS experiment and the proteins summary of an LC-MS experiment, whereas peptide-based maps visualize the peptides output of LC-MS experiments only.

Through this module, VIP supports visual queries, allowing the user to filter the glyphs of her/his map and delve deeper into her/his dataset. In particular, when the user performs a visual query (e.g., proteins with *score*>40 AND *abundance ratio*<1), the proteomic objects glyphs, whose features meet the given criteria, are detached from the level of the synthetic map and are raised up (Figure 5).

Furthermore, to deal with the common problem of 2DGE overlapping spots [22], we added to the Visualizer the gradient transparency mode functionality, as well as control over the glyph size, by defining its minimum and maximum values (Figure 6).

Last but not least, the Visualizer incorporates distortion techniques (e.g., fisheye view, perspective wall, multifocal display) that aim to improve the exploration and legibility of "dense" synthetic maps (Figure 8) (i.e., usually peptide-based maps, which contain thousands of glyphs displayed very close to each other).

D. *VIP Desktop*

The VIP desktop serves as the "glue-component" of all previous modules. Figure 4 shows the VIP desktop as well as its partial components. The *Toolbar* contains shortcut buttons to provide easy access to basic commands, as it is obvious in Figure 4-A. The *Control Panel* (Figure 4-B) contains the proteomics *Data Viewer*, a tree structure which holds the loaded data and allows the user to access quickly the desired protein/peptides container (gel, peptides set), while preserving the data hierarchy. Moreover, Control Panel contains the *Parameters*, a pane which includes the visualization parameters (e.g., transparency mode, coloring scheme, size control). In its lower part it also contains the feature-to-attribute mapping to remind the user the selected features, while exploring the maps. The *Features Panel* (Figure 4-C) holds the table of loaded features and offers for each column preprocessing and graphical encoding. Note that each selected column/feature for visualization is highlighted uniquely to stand out from the rest of the features. This panel is initially absent from the VIP desktop and is shown when the user selects a gel or a peptides set from the data viewer. Finally, the *VIP canvas* (Figure 4-D)

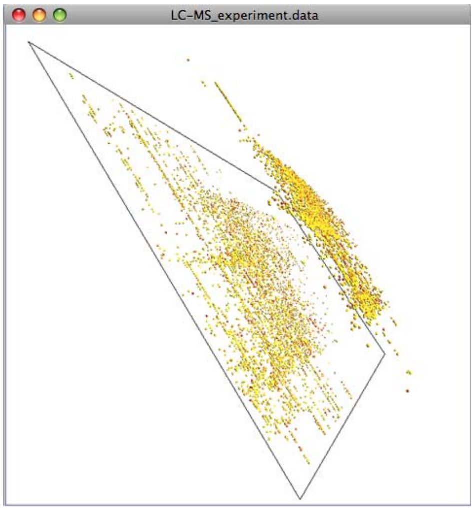


Fig. 5. A rotated peptide-based map after performing a visual query. The peptides that meet the criterion *identification score > 90* are detached from the map's level to facilitate the user's searching.

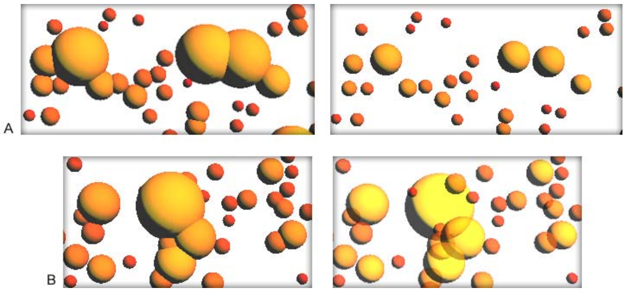


Fig. 6. The problem of overlapping spots is tackled either by controlling the size of a sphere (A) or by enabling the transparency mode (B).

displays the synthetic maps and is initially blank.

E. Implementation

VIP is implemented using the Java platform (Sun JDK 1.6) as well as external libraries and released under the *GNU General Public License* (GPL). Our reference implementation is available upon request since it is still under development (we plan to make it publicly available soon).

Backend, the module responsible for aggregating proteomic features using POML, our proposed markup language, uses the *Java API for XML Processing* (JAXP) an open source API for efficient XML validation and parsing [23]. More specifically, we used the *Document Object Model* (DOM) parsing interface, which constructs a complete in-memory representation of an XML document in order to keep the Backend module compact (in terms of structure complexity and lines of code).

As already mentioned, the unified representation of proteomics data is achieved by utilizing associative data structures. We used the standard Java `HashMap` library included in `java.util` package, which supports efficient insertion and lookup in O(1) time [24].

The Visualizer module is based on the Java 3D API [25]. Using Java 3D as a visualization engine has the advantage of rapid application development, since it incorporates a high-level scene-graph model, allowing developers to focus on the objects and the scene composition. Since Java 3D runs on top of either OpenGL or Direct3D technologies, it takes advantage of hardware acceleration and releases the CPU of the system from the burden of drawing complex 3D scenes.

The VIP Desktop has been implemented using the Swing widget toolkit included in `javax.swing` package. It provides native and pluggable look and feels for achieving consistency with different operating system user interfaces.

The source code of VIP is organized in classes and packages that are portable, reusable and extensible for "plugging-in" new data sources and models. VIP has been tested under Microsoft Windows XP and Vista, Mac OS X and GNU/Linux.

V. Examples

In order to demonstrate some of the VIP functionalities, as well as benefits and advantages over similar works, we present two indicative examples of synthetic maps, one for a 2DGE-MS and one for an LC-MS experiment. Both maps depict protein/peptides as spheres and exploit two basic visualization attributes, the sphere size and color. Though these examples, we show that we have embraced the functional requirements of VIP as they were analyzed in section III and that our suggested visualization can be applicable for both types of proteomics experiments. The datasets used have been provided by collaborators in the Biotechnology Division of the Biomedical Research Foundation, Academy of Athens, Greece [26] and are part of the studies presented in [27] and [28] respectively.

A. Proteins Map

The work presented in [27] is a typical experimental study involving 2D electrophoresis and mass spectrometry, in order to compare protein expression profiles in different biological states (in this case between neonates under intrauterine growth restriction and neonates appropriate for gestational age) and possibly discover the importance of specific proteins in a pathophysiology.

During this study, the biologists' interest focuses on: (a) detecting the up and down- regulated protein spots and (b) possibly conducting further analysis on some of the positively identified spots. To perform these analysis tasks, the routine method is to annotate the original gel image with the spot labels and to consult a table, which summarizes all the feature values (e.g., identification score, abundance ratio, p-value, protein name) per spot, using these labels. To offer an effortless and integrated analysis of the proteomics data, that does not require searching in lengthy tables, we exploit the advantages of multidimensional visualization and we propose the synthetic maps visualization using VIP.

For the particular example case, we have constructed a synthetic map using spherical glyphs to represent protein spots and the sphere radius and color to encode the protein *identification score* and *abundance ratio* respectively. The

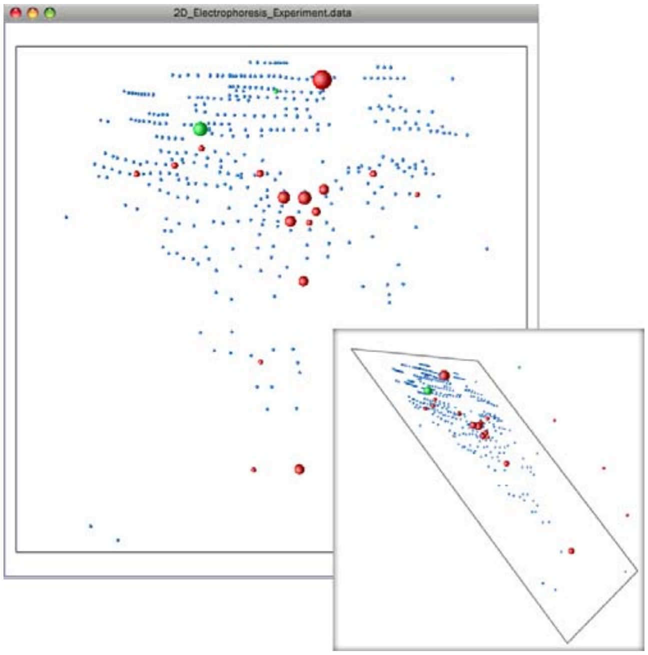


Fig. 7. Protein-based synthetic map which depicts the up/down-regulated protein spots (green/red spheres) along with an indication about their identification score (spheres size). The same map is also shown, with the four non-positively identified spots detached from the map's level.

resulting map is shown in Figure 7 and corresponds to the annotated gel image in [27]. We chose to visualize the *abundance ratio* with color because color-coded expression level is familiar to the user since the first microarrays studies (i.e., up/down regulation depicted with red/green color). All spots that were detected by the image analysis software (e.g., PDQuest) are displayed on the synthetic map as spheres, located in the exact positions of the corresponding detected spots on the 2D gel in order to main a visual reference to the original gel image (Figure 4-D). Those with red or green color are the ones that were found to be differentially expressed[1] and were selected for mass spectrometry. From the twenty spots that were subjected to mass spectrometry, only sixteen were found to be positively identified, based on the searching criteria. The rest four protein spots that did not obtain a good identification score and possibly need further analysis (very small-size spheres, 1 green and 3 red), can be seen in Figure 7 as the spheres that are raised up from the map's level.

By using our map, the user can easily inspect the green/red spheres that correspond to the up/down- regulated spots and also obtain a quick estimation of the regulation trend seamlessly (e.g., here we observe a reduction rather that an over-expression of the proteins).

The interactivity provided by VIP allows the user to control the spheres size so as to minimize or prevent overlapping, enable the transparency mode to make hidden spots visible and link a protein spot on the synthetic map with all of its features. By clicking on a sphere, the corresponding row in the Features Table is highlighted enabling the user to explore all the features coupled to the specific protein spot. As it has become obvious, the synthetic maps visualization as well as the interactivity provided by VIP, allow the overall inspection of the dataset according to different needs and criteria, by simply selecting the appropriate features for visualization.

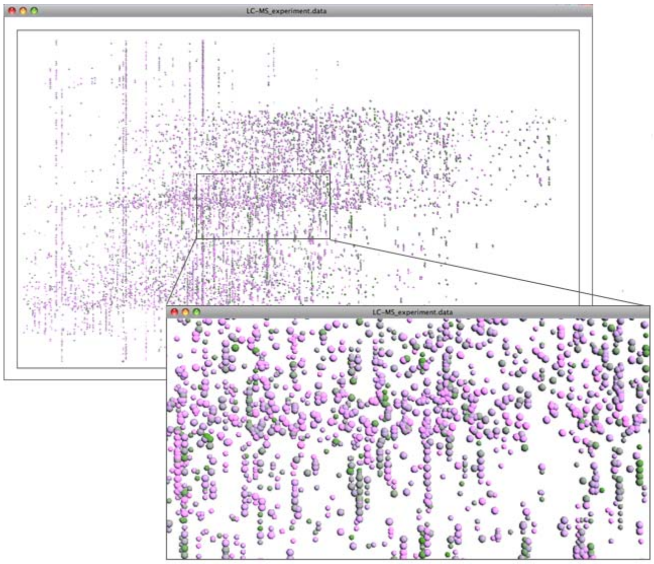


Fig. 8. Peptide-based map which summarizes an LC-MS experiment. The map is crowded and therefore zooming in specific regions is needed to comprehend the color and size differences.

B. Peptides Map

As already mentioned in the Related Work section, existing tools try to visualize the high dimensional LC-MS data in 2D or 3D space. These representations help the user gain a better understanding of a large dataset in its entirety, locate patterns that can be used to identify chemical contaminants or post-translationally modified peptides, to reveal quantitative regulation trends and to diagnose the LC-MS system performance (e.g., low peptide concentration, insufficient LC separation) [14], [15], [16].

However, our suggested visualization of an LC-MS dataset offers in addition:

(a) The flexibility to create alternative views of the dataset, by selecting the appropriate features.

(b) The capability to delve deeper into the large dataset by performing visual queries that allow the user to search for peptides that meet certain criteria.

(c) The effective representation of combinations of proteomic features by using the visualization attributes of a glyph. Thus, comparing the values of a feature is as easy as perceiving size and color differences.

The synthetic map shown in Figure 8 is created using the dataset described in [28], an LC-MS/MS study which compares normal and cancerous prostate tissue extracts in order to determine statistically significant differences

[1] Ratio of the mean percentage of the Optical Density of the protein spot in one condition, to the mean percentage of the protein spot in the other condition. Ratio > 1 indicates up-regulation, whereas ratio < 1 indicates down-regulation.

between them and discover novel biomarkers for the prostate cancer. In this map, the 8284 identified peptides of the experiment are located on their [retention time, m/z] position and depicted with spheres, whose size and color represent the *identification confidence* and *abundance ratio*, two features emanating from the identification and quantitation steps of the analysis respectively. Exploring this map with the use of the VIP navigation capabilities as well as the visualization attributes, the user can obtain an estimation of the expression ratio for her/his experiment and locate the highly identified peptides. Additionally, s/he can effortlessly discover interesting peptide patterns (e.g., vertical streaks), and make observations that might lead in the optimization of the experiment's methodological conditions.

Due to the high dimensionality of the LC-MS experiment, it would have been difficult for the user to perceive mainly the size differences, if VIP did not offer zooming at specific regions of the synthetic map (Figure 8). To deal with the large number of peptides, the user can also perform visual queries on the map and isolate the peptides that meet the given criteria by detaching them from the map's level (Figure 5). This way, s/he can search into the dataset for specific peptides of interest (e.g., highly up-regulated peptides).

VI. Conclusion

In this paper we have presented VIP, a novel interactive tool that provides visualization for integrated proteomics data. We have described the main components of VIP, focusing on how they satisfy the gathered functional requirements and justified our design choices using an in-depth architectural description. Using two indicative examples, one for each type of differential proteomics experiments, we have shown that by exploiting the joint visualization of proteomic features, the proposed synthetic maps can work as an effective mechanism for the difficult task of multidimensional feature space exploration and data interpretation.

Future work on VIP includes the incorporation of several distortion techniques to deal with the dense peptides-based maps, as well as the support of connections and interactions between maps to facilitate the differential expression analysis. We also plan to exploit additional glyphs and visualization attributes, as well as investigate the optimal number of features that can be visualized simultaneously per glyph, as far as the user comprehension is concerned.

References

[1] M. Tyers and M. Mann, "From Genomics to Proteomics", *Nature*, Vol. 422, March 2003, pp. 193-197.

[2] D. C. Liebler, *Introduction to Proteomics.* Humana Press, 2002, pp. 55-76.

[3] L. Monteoliva and J. P. Albar, "Differential proteomics: An overview of gel and non-gel based approaches", *Briefings in Functional Genomics and Proteomics*, Vol. 3, 2004, pp. 220-239.

[4] M. Hilario, A. Kalousis, C. Pellegrini and M. Muller, "Processing and Classification of Protein Mass Spectra", *Mass Spectrometry Reviews*, Vol. 25, 2006, pp. 409– 449.

[5] I. Beer, E. Barnea, T. Ziv, and A. Admon, "Improving large-scale proteomics by clustering of mass spectrometry data", *Proteomics*, Vol. 4, 2004, pp. 950– 960.

[6] The Gene Ontology url: http://www.geneontology.org/

[7] B. D. Halligan, S. P. Mirza, M. C. Pellitteri-Hahn, M. Olivier and A. S. Greene, "Visualizing Quantitative Proteomics Datasets using Treemaps", in *Proc. 11th Int. Conf. Information Visualization*, 2007, pp. 527-534.

[8] L. Krishnamurthy, J. H. Nadeau, G. Ozsoyoglu, Z. M. Ozsoyoglu, G. Schaeffer, et al., "Pathways Database System: An Integrated System for Biological Pathways", *Bioinformatics*, Vol. 19(8), 2003, pp. 930-937.

[9] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics", Nature, Vol. 422, 2003, pp. 198-207.

[10] R. J. Turner, K. Chaturvedi, N. J. Edwards, D. Fasulo, A. L. Halpern, et al., "Visualization challenges for a new cyberpharmaceutical computing paradigm", in *Proc. IEEE 2001 Symp. on Parallel and Large-data Visualization and Graphics*, 2001, pp. 7-18.

[11] GeneData Expressionist®
url: http://www.genedata.com/products/expressionist/

[12] bdal.de | home | Bruker Daltonics url: http://www.bdal.de/

[13] S. luhn, M. Berth, M. Hecker and J. Bernhardt, "Using standard positions and image fusion to create proteome maps from collections of two-dimensional gel electrophoresis images", *Proteomics*, Vol. 3, 2003, pp. 1117-1127.

[14] SPC Proteomics Tools: Pep3D
url: http://tools.proteomecenter.org/Pep3D.php

[15] X. Li, P. G. A. Pedrioli, J. Eng, D. Martin, E. C. Yi, H. Lee and R. Aebersold, "A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry", *Anal. Chem.*, Vol. 76, 2004, pp. 3856-3860.

[16] L. Linsen, J. Locherbach, M. Berth, J. Bernhardt and D. Becher, "Differential protein expression analysis via Liquid-Chromatography/Mass-Spectrometry data visualization", in *Proc. IEEE Visualization 2005*, pp. 447-454.

[17] S. D. Garbis, S. I. Tyritzis, T. Roumeliotis, P. Zerefos, E. G. Giannopoulou, et al, "Search for potential markers for prostate cancer diagnosis, prognosis and treatment in clinical tissue specimens using amine-specific isobaric tagging (iTRAQ) with two-dimensional liquid chromatography and tandem mass spectrometry", *J. of Proteome Res.* 2008, In Press.

[18] R. Spence, *Information Visualization: Design for Interaction,* Pearson Education, 2006, pp. 20.

[19] A. Dix, J. Finley, G. D. Abowd, R. Beale, *Human-Computer Interaction*, 3rd Edition Prentice-Hall, 2004.

[20] Y. K. Leung, and M. D. Apperley, "A review and taxonomy of distortion-oriented presentation techniques", *ACM Transactions on Human-Computer Interaction,* Vol. 1, 1994, pp. 126-160.

[21] B. Shneiderman, "Dynamic queries for visual information seeking", *IEEE Software*, Vol. 11, 1994, pp. 70-77.

[22] A. W. Dowsey, M. J. Dunn, G. Yang, "The role of bioinformatics in two-dimensional gel electrophoresis", Proteomics, Vol.3, 2003, pp. 1567-1596.

[23] jaxp: JAXP Reference Implementation url: http://jaxp.dev.java.net

[24] HashMap (Java platform SE 6)
url: http://java.sun.com/javase/6/docs/api/java/util/HashMap.html

[25] java3d: Java 3D Parent Project url: http://java3d.dev.java.net

[26] Biomedical Research Foundation url: http://www.bioacademy.gr

[27] P. M. Karamessinis, A. Malamitsi-Puchner, T. Boutsikou, M. Makridakis, K. Vougas et al., "Marked defects in the expression and glycosylation of α-2-HS glycoprotein/fetuin-A in plasma from neonates with intrauterine growth restriction: Proteomic screening and potential clinical implications", *Mol Cell Proteomics,* Vol. 7, 2008, pp. 591-599

[28] S. D. Garbis, P. Zerefos, A. Papadopoulou, A. Vlahou, C. Tamvakopoulos, et al., "Quantitative proteomic approaches for the determination of carcinogenesis biomarkers in prostate tissue", in *Proc. 54th ASMS conference on Mass Spectrometry,* 2006, Seattle, Washington.