GeneShelf: A Web-based Visual Interface for Large Gene Expression Time-Series Data Repositories

Bohyoung Kim, Bongshin Lee, Susan Knoblach, Eric Hoffman, and Jinwook Seo

Abstract—A widespread use of high-throughput gene expression analysis techniques enabled the biomedical research community to share a huge body of gene expression datasets in many public databases on the web. However, current gene expression data repositories provide static representations of the data and support limited interactions. This hinders biologists from effectively exploring shared gene expression datasets. Responding to the growing need for better interfaces to improve the utility of the public datasets, we have designed and developed a new web-based visual interface entitled GeneShelf (http://bioinformatics.cnmcresearch.org/GeneShelf). It builds upon a zoomable grid display to represent two categorical dimensions. It also incorporates an augmented timeline with expandable time points that better shows multiple data values for the focused time point by embedding bar charts. We applied GeneShelf to one of the largest microarray datasets generated to study the progression and recovery process of injuries at the spinal cord of mice and rats. We present a case study and a preliminary qualitative user study with biologists to show the utility and usability of GeneShelf.

Index Terms—bioinformatics visualization, augmented timeline, animation, zoomable grid, gene expression profiling.

1 Introduction

The biomedical research community has witnessed tremendous advances in genetic probing technologies in recent years. Researchers can now measure the activity of tens of thousands of genetic markers in a single high-throughput gene chip (aka microarray). As the microarray technology becomes a commodity in the biomedical research community, there have been collaborative efforts to share the valuable microarray datasets with other researchers of similar interest in the world. For example, NIH has been running the GEO (Gene Expression Omnibus) repository since 2002, where biomedical researchers submit their microarray datasets for public access. The total number of public gene chips (or samples) that GEO hosts reaches about 318 thousand as of 2009. The number of microarray datasets in public repositories is expected to grow steadily as most biomedical journals require researchers to make their datasets publically available to publish their articles.

As public repositories with a large body of gene chip datasets become popular, there is a growing need for efficient interfaces to help users explore the datasets in the repositories. However, current gene expression data repositories provide static representations of the data and support limited interactions. Researchers either search or browse the repositories using simple Boolean text searches to find samples or datasets that they are interested in. A major problem is that it may not be possible for researchers to check the appropriateness of samples or datasets acquired from search or browse using the meta-data and static representations available at the repositories, which is often an arduous process.

To improve the utility of current public microarray data repositories, we designed and developed GeneShelf, a novel web-based visual interface (Fig. 1). Our primary goal was to build a new light-weight visual interface that serves as a front-end web user interface for public microarray data repositories. This makes it possible for GeneShelf to be complementary to the current existing static (and text-based) interface for the repositories. Our secondary goal was to support a majority of the microarray projects that may have different study designs. To that end, we first categorized microarray projects into three main classes according to their study designs: cross-sectional, time-series, and hybrid. Cross-sectional studies compare samples from different groups (e.g. normal vs. diseased). Time-series studies focus on change of gene expression patterns across samples over time after intervention or onset. The hybrid ones are a combination of cross-sectional and time-series. For cross-sectional study designs, GeneShelf assigns a separate grid block for each condition in a grid display, for example, one for normal and the other for diseased. For time-series study designs GeneShelf uses augmented timelines. To support hybrid study designs, GeneShelf is based on a zoomable grid display where timeline views can be embedded in each block.

One of the most popular research trends in microarray research is to integrate genetic pathways into the analysis process. A genetic pathway is a network to show interactions occurring between a set of genes depending on each other's individual functions. Pathway-based analysis allows researchers to view their data in much broader biological context, increasing in part the sensitivity limit of general microarray experiments.. We support this research trend by using a pathway as a unit of visualization in GeneShelf.

To show the utility and usability of GeneShelf, we applied GeneShelf to the spinal cord injury (SCI) project, one of the world's largest microarray projects. As the name implies, the SCI project deals with spinal cord injuries in mice and rats from the PEPR GeneChip data warehouse, a public microarray data repository [5]. We performed a case study and a preliminary qualitative user study with biologists at a large microarray research center. All the participants gave us strong positive feedback on GeneShelf as well as several constructive suggestions about desirable functions.

We organize this paper as follows. We first present the background about microarray data analysis, visualizations for gene expression data, and tabular visualization techniques. After we present GeneShelf along with design rationales and description of user interactions, we summarize our evaluation results as a validation of the usability and efficacy of GeneShelf. Finally, we conclude this paper with future work.

- *Bohyoung Kim is with Seoul National University, Email:bhkim@cse.snu.ac.kr*
- *Bongshin Lee is with Microsoft Research, Email:bongshin@microsoft.com*
- *Susan Knoblach and Eric Hoffman are with Children's National medical Center, Email: {sknoblach, ehoffman}@cnmcresearch.org*
- *Jinwook Seo is the corresponding author with Seoul National University, Email:jwseo@cse.snu.ac.kr*

Manuscript received 31 March 2009; accepted 27 July 2009; posted online 11 October 2009; mailed on 5 October 2009.
For information on obtaining reprints of this article, please send email to: tvcg@computer.org .

Fig. 1. GeneShelf (available at http://bioinformatics.cnmcresearch.org/GeneShelf) showing a relevant genetic pathway at the injury site (T9), 7 days after a mild spinal cord injury at a vertebral level T9. The gene "*sc4mol*" is highlighted in red in all views. "*sc4mol*" is the most active gene for moderate injury at T9, but it is defeated by "*hmgcs1*" in all other conditions. GeneShelf consists of four visualization components: top 10 pathway list view (A), pathway view (B), gene list view (C), and *nTimeLines* grid view (D). Top 10 pathway list view (A) shows the top 10 pathway names. The current pathway is shown in the pathway view (B) at the top center and the gene names in the pathway are shown in the gene list view (C) at the top right. The horizontal axis of the *nTimeLines* grid view (D) is for the severity of spinal cord injuries induced (sham, mild, moderate, and severe). The vertical axis is for the sampling location relative to the injury site (T8: above, T9: at, and T10: below). The brown vertical line in the grid block for mild injury at T9 indicates the condition under which the top 10 pathways are selected.

2 Background

2.1 Microarray Data Analysis

To promote sharing of biological projects, several microarray data repositories have been developed in the public domain. The NIH is hosting the Gene Expression Omnibus (GEO) repository [1] since 2002, where biomedical researchers submit their microarray datasets for public access. European Bioinformatics Institute (EBI) is also hosting a public data repository called ArryExpress [16]. Children's National Medical Center is running a repository called PEPR (Public Expression Profiling Resource) [5]. To help users find relevant microarray projects, they all support Boolean queries. They also support a so-called single gene query [5] with which users can see the expression pattern of a single gene by project.

When using these public repositories, most users are interested in finding microarray experiments that are closely related to their own projects. Boolean queries and the single gene query can assist users with the search of related projects. However, the search results often turn out to be irrelevant or of poor-quality after users download the data and explore them using local analysis tools. When showing each individual experiment, GEO supports a visual interface that shows hierarchical clustering results of each project in a dendrogram with a heatmap [9]. We believe a lightweight visual web interface to show each project can significantly increase the utility of current public microarray data repositories.

One of the most important factors in microarray data analysis is the signal algorithm selection. Signal algorithms draw numerical gene expression values from gene chip scans. They are a critical part of microarray experiments because different signal algorithms can make completely different gene expression values out of the same gene chip with the same sample, which even affects the conclusion of the experiment [23]. However, the signal algorithm is not seriously taken into account in the interfaces for the repositories, if not completely ignored. We incorporated signal algorithms in our visual query interface in addition to preprocessing the microarray project data in a repository, which will be described later.

The microarray technology has been useful for sifting a limited set of genes that show significant changes across experimental conditions. In these days, researchers want to go one step forward by examining genetic pathways in which the selected genes take part. There are many public or commercial pathway databases ([7, 12, 17], ww.ingenuity.com, www.pathwaycommons.org) that host thousands of pathways of many kinds. Users can examine how the genes of their interest interact with biological processes represented in pathways. These pathways give researchers better biological insights into what is really happening in the biological system. In spite of potential benefits from combining pathway information with microarray datasets, most current microarray data repositories are not integrating the pathway resources. GeneShelf supports this by using a pathway as a unit of visualization as well as a unit of analysis.

2.2 Microarray Data Visualization

Cytoscape [25] is a popular pathway visualization tool in the bioinformatics field. It is a stand-alone, open source bioinformatics software platform for visualizing biological pathways and networks. It allows users to connect to external data sources to integrate pathway data and other related annotations.

Information visualization community has been contributing to this field by applying various techniques such as graph visualization methods to biological pathway visualization [22]. The pathway visualization tool in GeneSpring (by Agilent Technologies) implements brushing and linking between pathway view and time-series charts. Craig et al. designed an interactive visualization tool called "Time-series Explorer," where they used animated scatterplot views through a close interaction between a scatterplot and two parallel coordinate views [6]. Users can select a region of interest in a scatterplot to highlight the genes in the parallel coordinate views or specify a time interval of interest in the parallel coordinate views to see the animated scatterplot of the selected genes within the specified time interval. Eichler et al. implemented the Gene Expression Dynamics Inspector (GEDI), where they show a visual summary of each sample using the self-organizing map technique [8]. Hence, users can examine the global activity patterns of genes by examining the color mosaics of samples of a microarray project.

Although there have been many microarray data visualization tools [11, 21, 24], most of them are heavyweight stand-alone tools. Biomedical researchers often have to download several projects to a local machine and check them using these heavyweight analysis tools to find the most relevant one. This problem degrades major microarray repositories' utility. Thus, biomedical researchers are in need of lightweight visualization tools that can serve as a front-end interface to help them interactively find the most relevant projects for ever-growing public microarray data repositories. Techniques and frameworks for incorporating public genetic pathway data repositories within the lightweight visualization will be necessary as well. In designing GeneShelf, we sought to construct a model interface that can help researchers determine the most relevant projects in the public microarray data repositories through efficient visual encodings and smooth animation.

2.3 Tabular Visualization Techniques

TableLens is a seminal work on visualizing a large multivariate dataset on a tabular display [19]. It can handle quantitative and categorical data types. It employs a focus+context technique with which users can zoom in to columns and rows. Each table cell shows an aspect of a single item when zoomed in. Items are presented as single-pixel-width bars or as colored swatches. Users can sort columns in descending or ascending order. GeneShelf is also based upon the focus+context technique, and it shows up to dozens of items in a single cell (or a grid block) since genes in a pathway are the unit of analysis and visualization. Each item (or gene) is presented as a line graph by default or as a bar on a bar chart when a time point is expanded (or zoomed in).

Line Graph Explorer is an interactive visualization tool for large line graph collections [20]. It encodes the *y*-dimension of individual line graphs with color, which enables users to see global patterns of the data. Users can zoom in to a row to see the detail pattern of an individual item using a standard line graph. Like TableLens, it shows a single item in each row (cell), but unlike TableLens, it employs only a single column. GeneShelf shows more than one item with standard line graphs in a single cell (or grid block). It also supports multiple columns because it has to show multiple combinations of experimental conditions in a compact grid view.

Most similar to our work is LiveRAC. It is a visualization tool for computer systems and network performance monitoring data over a long period of time [14]. It also supports the focus+context technique (accordion drawing [15]), which is the most scalable in terms of the number of data items that it can handle. Together with the focus+context technique, LiveRAC incorporates the semantic zooming technique [3] to show different visualizations at different zoom levels. LiveRAC reduces clutter in each cell by showing only one line graph or two for a device. Since GeneShelf deals with a pathway consisting of up to dozens of genes as a unit of analysis and visualization, it has to show a group of related line graphs in a single cell. GeneShelf allows for two-level zooming, especially for zooming in to a specific time point to examine the details of the time point using a different visual representation, or bar chart.

3 GeneShelf

GeneShelf is designed as a web-based visual interface for large gene expression time-series data repositories. In this paper, we applied it to a large microarray project (SCI project) of a hybrid experiment design from the PEPR repository, the largest contributor to the NIH GEO. The SCI project consists of about 1,000 gene chips and it has three dimensions: severity of injury, sampling location, and the time after injury; a group of researchers traumatized the spinal cord at a specific location of the spinal cord and took samples below, above, and at the injury site at various time points after the injury.

3.1 Experimental Design for Microarray Projects

To cover most microarray studies, GeneShelf is designed considering all the three most commonly used study designs: cross-sectional, time-series, and hybrid (combination of cross-sectional and time-series).

Cross-sectional studies compare different conditions, often for multiple factors. For example, one factor can be disease status (normal vs. diseased) and another can be treatment (drug A vs. drug B). In order to visualize all combinations of conditions, GeneShelf uses a grid display, where each factor is assigned to each grid axis, for example, disease status to the *x*-axis and treatment to the *y*-axis. Then each grid block displays the expression values of genes for the corresponding condition.

In addition to the cross-sectional design, biologists often need a time-series design to investigate the progress after a certain intervention. If there are just a few time points (e.g., two: pre- and post-treatment), this could be treated as a cross-sectional study even though it involves time variable. For the cases where there are many time points involved, GeneShelf uses the timeline display representing the progress of gene expression values as line graphs, each of which represents an individual gene's expression pattern. We will describe a drawback of the traditional timeline display and how GeneShelf tackles it in the next section.

The number of hybrid designs combining cross-sectional and time-series designs is increasing since longitudinal studies tend to control confounding factors. For the hybrid designs, GeneShelf combines the grid display with the timeline display by embedding a timeline display within each block of the grid display, which implements a small multiple visualization [26]. In other words, each gird block visualizes gene expression patterns for the corresponding experimental condition.

GeneShelf is currently designed to support only two factors in addition to time. One way to extend GeneShelf to support more factors is to recursively subdivide grid axes. However, we suspect this makes it difficult to compare gene expression patterns by an individual factor. Another possible, simple but powerful solution is to enable users to assign a factor to each axis. It is important to note that microarray studies with such a complex design are rare in public microarray repositories.

3.2 *nTimeLines* (Enhanced TimeLines) Control

3.2.1 Expandable Time Points

The small multiple visualization in a grid display helps users compare the overall time-varying gene expression patterns across all grid blocks covering all combinations of two experimental conditions. Tight coupling among the small multiples further helps users interactively compare an individual gene's expression pattern

across all grid blocks. However, it is not easy for users to compare gene expression values at a specific time point across all grid blocks. In some cases where many lines overlap with each other, it is difficult to compare values or to see the overall gene expression patterns even at a single time point.

To address this problem, we augmented the traditional timeline display. Since a bar chart is very effective for the comparison of multiple values, we combined the traditional timeline of line graphs and the bar chart to take the full advantage of their own benefits. The enhanced timeline display entitled "*nTimeLines*" allows users to expand a time point to show the gene expression values at the selected time point as a bar chart within the timeline display. Each bar in a bar chart represents a gene.

In GeneShelf, a time point is represented by a vertical line and a label right below the vertical line. A mouse-over on a time point (vertical line or label for the time point) changes the cursor icon to a hand, indicating that this time point can be expanded. When users click on a time point, GeneShelf duplicates the vertical line for the time point and slides open them to have a rectangle representing the time point (Fig. 2). A bar chart is embedded within the rectangle. To preserve the connectivity of line graphs, horizontal lines will connect the two duplicated vertical lines. The height of each bar for a gene will exactly reach the parallel horizontal line that represents the same gene. This alignment of the bar chart within the expanded time point makes apparent the relationship between bars and the corresponding line graphs. In addition, a gray bounding box for the rectangle surrounding a bar chart is displayed to indicate that the enclosed area represents a single time point.

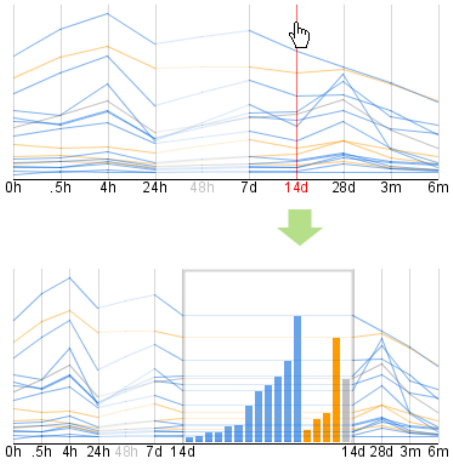

Fig. 2. Expanding a time point in an *nTimeLines* control. When users click on the vertical line for "14d", the time point expands to show a bar chart for the gene expression values at the time point. In the expanded view, it is much easier to see the overall distribution and compare individual values.

To help users follow this transformation, GeneShelf uses a two-step animation; it first opens a rectangle and then bars are smoothly growing downward from their corresponding horizontal lines (Fig. 4). Users can close an expanded time point either by clicking on the vertical line or label for the expanded time point or by clicking on the background of the bar chart. If users click on a time point other than an already expanded time point, the expanded time point first closes and then a new selected time point expands.

3.2.2 Support for Missing Values

Depending on the actual experiment, there could be no gene expression data at a certain time point. Regardless of the data availability, line graphs should be connected to show the data trends over time. However, this could mislead users as if the data exists there. To tackle this issue, GeneShelf not only hides the vertical line for the time point with the missing value but also attenuates the label and the line segments for the missing value. Furthermore, GeneShelf does not display line segments at any leading and trailing time points with missing values. For example, Fig. 4 shows that there is no gene expression data available at 0h, 48h, 28d, 3m, and 6m. For leading time point (0h) and trailing time points (28d, 3m, and 6m), GeneShelf hides both the vertical lines for those time points and the line segments from their adjacent time points. For the time point 48h, GeneShelf attenuates the line segments connecting the two nearest neighbors with values (24h and 7d) to indicate that they are not real values.

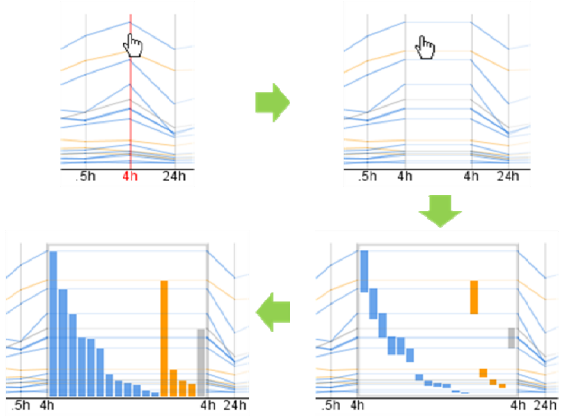

Fig. 3. Animation on expanding a time point in an *nTimeLines* control. When a time point expands to show a bar chart, each bar of the bar chart smoothly grows from top to bottom staring from the horizontal line for the corresponding line graph.

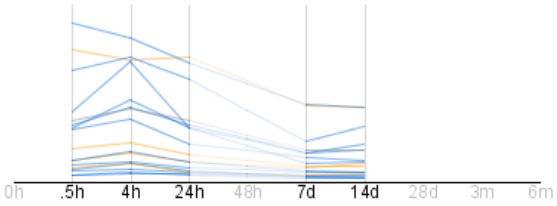

Fig. 4. Missing value representation. Labels for the time points with missing values are shown in an attenuated color. Line segments for a missing time point (48h) in the middle are interpolated, but the line segments for leading or trailing time points with missing values disappear. Line segments passing the interpolated values are also rendered using an attenuated color.

3.3 Interactions

3.3.1 Pathway Retrieval

To use a genetic pathway as a unit of analysis, GeneShelf allows users to query the pathway database by specifying the experimental condition (or the **reference condition**) including animal model, sampling location, injury severity, and time point in the "Fetch top 10 pathways" dialog box (Fig. 5), under which users filter significant genes. Furthermore, users can specify how to filter significant genes (comparing the selected time point to the time 0 or to the previous time point) using the "Normalized to" combo box in the dialog. Then these significant genes are used to determine the top 10 most relevant pathways from Wikipathways.org [17], an open public pathway database using Fisher's exact test [10]. Users can also specify the signal algorithm to generate gene expression values from gene chips using the "Signal Algorithm" ratio buttons. Once users click on the "Fetch" button on the dialog box, GeneShelf downloads the top 10 pathways with their gene expression data and lists them in the top 10 pathways list view. GeneShelf keeps the history of the reference conditions to fetch the pathways so that users can revisit those reference conditions by clicking the "Prev list" and "Next list" near top left corner of GeneShelf (Fig. 1). When users mouse-over these

buttons, GeneShelf shows a reference condition that will be reloaded when the button is pressed.

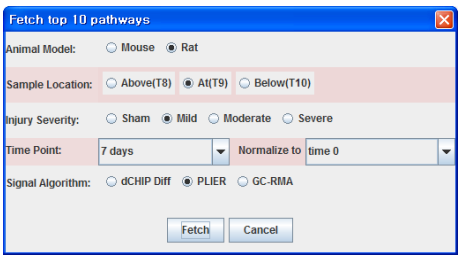

Fig. 5. "Fetch top 10 pathways" dialog box. Users can learn about the study design of the given experiment from this dialog box. Users select values for experimental design parameters and select a signal algorithm.

To examine the gene expression data of a pathway, users can select a pathway item from the top 10 pathway list view (Fig. 1A). Previously examined pathways are shown in purple to indicate that they are already visited.

3.3.2 Multiple Coordinated Views

As genetic pathways are becoming a very important part of microarray data analysis, we incorporated a lightweight pathway visualization tool, or the pathway view (Fig. 1B), as a component of GeneShelf. In the pathway view, each gene is represented as a rectangle. GeneShelf uses different background colors according to gene's neuron cell type information suggested in [4]: orange for astrocyte, blue for neuron, red for oligodendrocyte, and gray for the undefined. The same color-coding scheme is used throughout all other views including the gene list view and the *nTimeLines* grid view. For the genes that are in the current pathway but are not present in the current gene expression dataset, GeneShelf uses lighter background color; for example, light pink for oligodendrocyte. Users can zoom in and out the pathway view using the vertical slider on the right side of the view and pan the view by dragging the mouse with the left button depressed. Users can also see the pathway at the WikiPathways.org site by clicking on the button at the top right corner of the pathway view. The gene list view (Fig. 1C) shows the list of genes in the current pathway. In addition to their names and other annotations, GeneShelf shows the name of cell type in the third column of the gene list view with different background colors for each cell type using the same color-coding scheme.

The gene expression patterns of the current pathway are displayed in the *nTimeLines* grid view (Fig. 1D), which is built upon the small multiples and the focus+context technique. This grid display shows the two dimensions of the cross-sectional aspect of the study design: horizontal axis for the severity of injury and vertical axis for the sampling location relative to the injury site. The severity of spinal cord injury increases from left to right (sham, mild, moderate, and severe). For each severity column, there are three sampling locations from top to bottom (T8: above the injury site, T9: the injury site, and T10: below the injury site). For example, the second grid block of the top row shows the time-varying gene expression patterns for mild injury samples at T8 (above the injury site). The row and column titles highlighted in bold brown represent the reference condition specified in the "Fetch top 10 pathways" dialog box. The reference time point in the grid block for the reference condition is also highlighted in bold brown. All the *nTimeLines* controls in the grid display are synchronized to help users compare gene expression values at a specific time point across them. For example, clicking on a time point in one *nTimeLines* control will trigger expansions of the same time point in all other grid blocks.

When users mouse over an item representing a gene in any view including a line graph or a bar in a bar chart of an expanded time point, it is highlighted in red. The corresponding genes in all other views are also highlighted in red. Furthermore, users can right-click on a gene to see a popup menu, where they can select the gene name to visit the detail information page for the selected gene at the NCBI website.

3.3.3 Two-level Animated Zooming

In addition to supporting the zooming for the expansion of the time point described above, GeneShelf enables users to zoom in and out a grid block. Users click on anywhere in the background of a grid block to smoothly enlarge the grid block and shrink other grid blocks. If users click again on the enlarged gird block, all the grid blocks return to their regular size, in which all the blocks have an equal size. If users click on a grid block other than the already expanded grid block, the expanded grid block first closes and a new selected grid block expands.

Animation is an effective technique to keep users alerted to the zooming events since the human visual system is preattentively sensitive to motion across the whole visual field [18]. It also helps users keep track of changes to visualization components without increasing cognitive overhead especially in users' peripheral view [13]. Thus, in GeneShelf we smoothly animate the two levels of zooming to help users remain focused during the transformation.

GeneShelf currently allows users to expand only one time point at a time, but it can be easily extended to expand two time points when there is enough space (e.g., when they zoom in to a grid block). Then, users can compare the change in value between two time points in a selected grid block.

3.3.4 Sorting Bar Charts in the *nTimeLines* Grid View

The *nTimeLines* control enables users to sort the genes represented in the embedded bar chart in 5 different ways: increasing/decreasing order of gene expression values, increasing/decreasing order of gene expression values for each cell type, and the same order as in the gene list view. Users can right-click on the background of a bar chart in an *nTimeLines* control to select a way to sort the bar chart. For example, a right-click on the bar chart in the grid block for moderate injury at injury site (T9) at the 4h time point will show a popup menu (Fig. 6). Once users select the menu item, "sort in increasing order for each cell type," GeneShelf will sort the selected bar chart highlighted with a green bounding box (Fig. 1). All corresponding bar charts in other grid bocks are arranged in the same order as in the selected bar chart (Fig. 1).

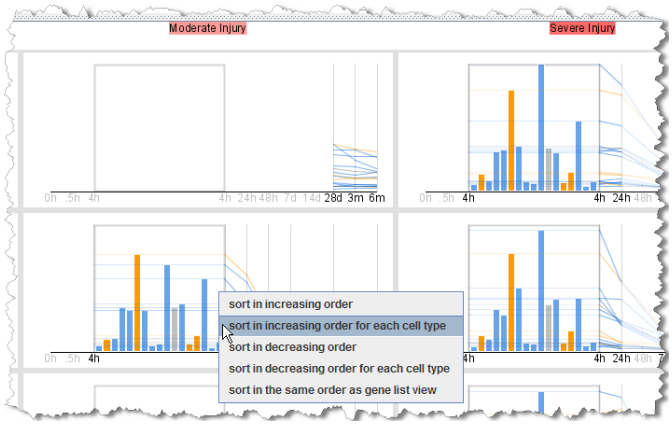

Fig. 6. Sorting bar charts in the *nTimeLines* grid view.

The resulting view (shown in Fig. 1) reveals that the gene expression pattern of the neuronal cell type genes (in blue) is unique for the selected condition compared to all other conditions. There is a peak in the middle of the blue bars for all other conditions especially for the sham and mild conditions, but no peak for the selected condition. This sorting function can support the correlation analyses done with TableLens.

3.4 Implementation Notes

GeneShelf is implemented with Java SE 6 and available at http://bioinformatics.cnmcresearch.org/GeneShelf as a Java Applet. GeneShelf communicates with a public microarray data repository (PEPR) using Java Servlet technology, and it downloads pathways data in GPML format [17] from wikipathways.org (Fig. 7).

To reduce the data loading time, we preprocessed the data. Gene expression values were calculated in advance using three well-known signal algorithms because it would take too long to run the algorithms on the fly. We also pre-calculated a significant gene list and corresponding top 10 pathway list for each of all possible experimental conditions. We save the pre-calculated gene expression data and the ranked pathway information in two separate database tables. As a trade-off, these extra database tables require some additional space on the data repositories. It usually takes about 10~15 seconds to download a pathway data from our microarray data repository. We do not cover the details of data preparation stage in this paper because of space limitations and low relevance of the topic to the information visualization community. GeneShelf's website has some more details on data preprocessing.

To save time to fetch the same data from the server again when users revisit the same pathway, gene expression data of every selected pathway are stored in the local machine's memory.

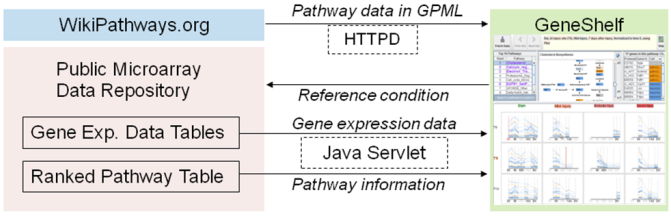

Fig. 7. Data transfer diagram.

To implement smooth animation of zooming in GeneShelf, we used Piccolo toolkit [2] developed at the human-computer interaction laboratory of the University of Maryland. Its infrastructure for structured graphical applications development enabled us to implement our pathway visualization tool extremely efficiently in terms of time and quality.

Biological pathways have become an essential part of contemporary microarray data analysis, which is significantly different from the traditional analysis where biologists just wanted to filter a set of significant genes through microarray studies. Major microarray analysis tools such as GeneSpring (http://www.agilent.com/chem/genespringGeneSpring) and IPA (http://www.ingenuity.com/products/pathways_analysis.html) treat pathways as units of analysis. GeneShelf uses pathways as a unit of visualization, showing a selected pathway at a time. Since most pathways in pathway databases contain only dozens of genes, there is no problem showing the gene expression changes across all conditions in GeneShelf. However, we have to come up with a different interface design if we have to show multiple pathways at a time. The current version of GeneShelf utilizes only the WikiPathways.org pathways. It can be extended to make use of other public pathway databases such as KEGG, BioCarta, and Pathway Commons.

4 Evaluation

We performed a case study and a preliminary qualitative user study at a microarray research lab to show the utility and usability of GeneShelf.

4.1 A Case Study

To validate and improve our interface design, an author of this paper who is a neuroscientist conducted a case study by trying GeneShelf herself. Since GeneShelf was designed as a front-end interface for a large public microarray data repository, the most appropriate usage scenario is that users visit the repository and try to find microarray experiments relevant to their own experiments. Once they find interesting experiments, they need to understand experiments from the context of the original experiment design. We assumed a situation where a neuroscientist studying spinal cord injury came across a potentially interesting experiment (SCI project) at a repository and tried to make sense of it using GeneShelf before she actually download it for further analysis using heavy-weight standalone tools. The goal of this case study was two-fold. First, we wanted to see if GeneShelf's features could help us find any interesting patterns in the dataset. Second, we were also interested in measuring GeneShelf's potential as a front-end interface for microarray experiments in large data repositories. We summarize her experience and discovery with GeneShelf.

While exploring top ten pathways for rat model, she identified ***oxidative stress*** as one of the major biological pathways involved after a spinal cord injury. The oxidative stress pathway was ranked number 2 for the reference condition of T9 (at the injury site), moderate injury, and 4 hours after injury. It was the number 3 for the same reference condition but 7 days after injury. This result corroborated that oxidative stress is a key component at the early stages of injury. Moreover, GeneShelf enabled her to visualize simultaneously many genes involved in *oxidative stress*, across sampling location, injury severity and time point (Fig. 8 and Fig. 9).

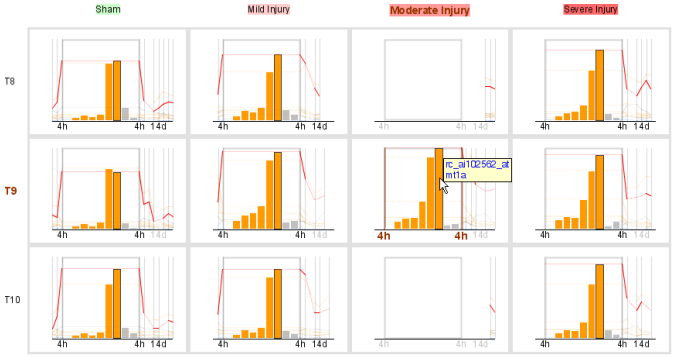

Fig. 8. *nTimeLines* grid view showing the oxidative stress pathway with the 4h time point expanded as a bar chart. Note the marked increase in ***mt1a*** 4 hours after injury.

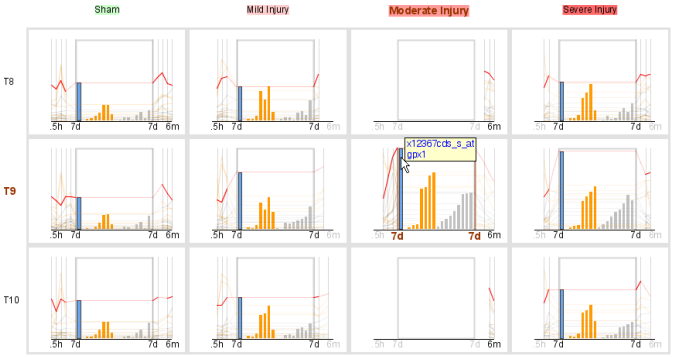

Fig. 9. *nTimeLines* grid view showing the oxidative stress pathway with the 7d time point expanded as a bar chart. Note the marked increase in ***gpx1*** 7 days after injury at the injury site (T9).

She particularly liked the *nTimeLines* grid view that simultaneously shows four different injury severities at three different sampling locations and the time point expansion in a bar chart on the gird view. She also enjoyed utilizing the multiple view coordination. These interactive features enabled her to observe increased expression of genes that are of a specific cell type. For example, ***mt1a*** (highlighted in Fig. 8), a gene associated with the astrocytes cell type (in orange), showed an elevated expression level

4 hours after all types of injuries, whereas ***gpx1*** (highlighted in Fig. 9), a gene associated with neurons (in blue), increased after 7 days at the injury site (T9) (Fig. 9). These data suggest the hypothesis that specific antioxidant genes such as ***mtla*** and ***gpx1*** may differ not only in time course, but also by cell type. In addition, the bar charts embedded in an expanded time point yielded another observation regarding the change of gene expression values across the injury severity: she could notice that gene expression level of ***gpx1*** at the injury site (T9) increases as the injury becomes severer (note the height of blue bars in the second row in Fig. 9).

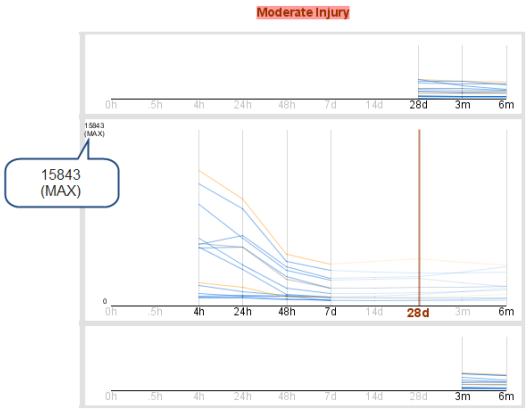

Fig. 10. *nTimeLines* grid view showing the cholesterol biosynthesis pathway for rat model with the *nTimeLines* control for moderate injury at the injury site (T9) enlarged. Note that most line graphs progress downward and the maximum expression value displayed in the top left corner of the enlarged view is 15843.

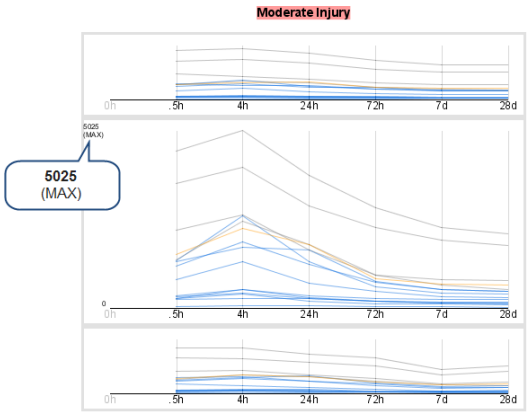

Fig. 11. *nTimeLines* grid view showing the cholesterol biosynthesis pathway for mouse model with the *nTimeLines* control for the same injury severity and location as in Fig. 10 enlarged. Expressions of many genes are also reduced, though to a lesser degree than in the rat. Note that the maximum expression is 5025, indicating much lesser degree of decrease in the gene expression value.

She found it useful that GeneShelf can rapidly visualize data from two different species (mouse and rat). It is known that the lesion that forms at the injury site differs markedly between rat and mouse. In rats, an open cavity develops at the injury region, but in the mouse, fibrous tissue fills the injury region and no cavity is present. Since cholesterol is a major component of neuronal membranes, she looked at the changes in *cholesterol biosynthesis* in both species, with the hypothesis that there may be less cholesterol biosynthesis in the rat after injury, due to the cavitation lesion. In the rat, *cholesterol biosynthesis* was the number 2 ranked pathway for the reference condition of T9 (at the injury site), moderate injury, 28 days after injury. GeneShelf showed a marked decrease in many of neuron-associated genes (in blue) in the pathway at the reference condition (Fig. 10). In the mouse, however, the cholesterol biosynthesis pathway was much lower ranked at 7 for the same reference condition. The decreases of expression values in many of the genes in the pathway were less marked than in the rat (Fig. 11). She could clearly notice this difference by enlarging the corresponding *nTimeLines* controls in both species. It was also noted that the maximum expression value (5025) for the mouse, displayed in the top left corner of the enlarged *nTimeLines* control, is only about one third of that for the rat (15843), indicating much lesser degree of decrease in the gene expression value.

In summary, we found that GeneShelf supported users in both learning experimental design parameters and performing visual exploration. The two level animated zooming and coordinated highlight of genes in multiple views helped us stay focused. Even though further evaluations with other datasets of different study designs are necessary to gauge the general utility of GeneShelf, this case study hinted that GeneShelf has the potential for serving as a lightweight front-end interface for public microarray data repositories. We also identified a task that was not efficiently supported in the current version of GeneShelf: side-by-side comparison of mice and rats. A potential solution would be to enable users to popup a window where an *nTimeLines* control visualizes gene expression patterns for the other species.

4.2 Qualitative User Study

To understand the usability of GeneShelf, we performed a preliminary qualitative user study. We sent a solicitation email to the researchers at a microarray research center and a small group of neuroscientists who are interested in the SCI research at Washington DC region. We recruited six researchers including one neuroscientist. In addition to the URLs for GeneShelf Java Applet and the user manual, we sent a questionnaire consisting of 10 questions where participants had to rate their response on 1-to-7 Likert scale (1 for "strongly disagree" and 7 for "strongly agree"). We asked them to fill it out after they tried GeneShelf and they sent us their questionnaires and additional comments via email. One participant gave us general comments but not a questionnaire. Therefore, average numbers reported here is with five participants.

Since we could not meet with the participants face to face just because they were far from us, we used an online user manual as a medium to teach them how to use GeneShelf. We provided many images and videos to explain important features in case participants did not read the text carefully.

All the participants strongly agreed that GeneShelf was easy to learn (avg. score 7). Three participants specifically mentioned in their comments that the images and videos in the user manual were very helpful and they did not have any problems learning GeneShelf. Participants also strongly agreed that GeneShelf was easy to use; (avg. 6.8). We also received very positive responses on the appropriateness of color, highlights, labels and messages (avg. 6.2).

The missing value representation received mixed responses (avg. 5.5). Three participants said that they could easily identify missing values in GeneShelf (two 6's and one 7). However, one participant gave the score 3 and another participant indicated that she did not see any missing values in the SCI data even though there were many missing values in there. We suspect that the participant thought that she had to do an action to find missing values instead of understanding the representation of the missing values.

Most participants agreed that the animation speed (0.7 second for animating blocks and 0.8 second for expanding a time point) was neither too fast (avg. 2.6) nor too slow (avg. 2) to follow. Even though the pathway view was given a relatively small portion of screen space, it seems to be useful (avg. 5.8). Two participants wanted to have a scroll bar in the pathway view even though we explained the panning function in the user manual using a video clip. We suspect they are not used to the panning operation itself.

We learned that all participants strongly agreed that the *nTimeLines* grid view was useful (avg. 6.8). They liked the fact that they could see and compare all conditions in one view using multiple

view coordination upon highlighting a gene. Two participants mentioned that the favorite feature of GeneShelf was that it enabled them to click on a time point and see the detail of the time point in a bar chart without losing a sense of overall gene expression patterns. We also learned that participants strongly agreed that they would like to use GeneShelf again for their research (avg. 6.6).

5 Conclusion

In this paper, we presented a web based visual interface, called GeneShelf, for understanding and exploring large time-series microarray datasets. We based the design of GeneShelf on our categorization of microarray projects available at public microarray data repositories. By demonstrating a model example with one of the most complex and largest microarray projects, we showed our interface templates' potential to be applied to most experiments in public microarray datasets. An enhanced timeline called *nTimelines*, one of the main components of GeneShelf, allows users to expand a time point to show the gene expression values at the selected time point as a bar chart within the timeline display. In addition to smooth animations and multiple view coordination with interactive highlighting, the *nTimeLines* grid view helped users examine and compare a selected time point across all conditions. Finally, we described a case study and a preliminary qualitative user study with biologists at a microarray research center. All the participants gave us positive feedback on GeneShelf as well as several constructive suggestions about desirable functions.

As for future work, we are interested in generalizing the GeneShelf visualization design for other microarray projects. As more and more public biomedical data is expected to accumulate in public repositories, we expect GeneShelf to serve as a good model example of exerting a decade-long achievement of information visualization research on biomedical problem solving.

Acknowledgments

This work was supported by NIH NINDS-01 (NS-1-2339), NIH NCMRR/NINDS 5R24 HD 050846 (Integrated Molecular Core for Rehabilitation Medicine), the Engineering Research Center of Excellence Program of Korea MEST/KOSEF (R11-2008-007-01002-0), and the Brain Korea 21 Project. The ICT at Seoul National University provided research facilities for this study.

References

[1] T. Barrett, T. O. Suzek, D. B. Troup, S. E. Wilhite, W.-C. Ngau, P. Ledoux, D. Rudnev, A. E. Lash, W. Fujibuchi, and R. Edgar, "NCBI GEO: mining millions of expression profiles--database and tools," *Nucleic Acids Research,* vol. 33, pp. D562-D566, 2005.

[2] B. B. Bederson, J. Grosjean, and J. Meyer, "Toolkit design for interactive structured graphics," *IEEE Trans. Software Eng.,* vol. 30, pp. 535-546, 2004.

[3] B. B. Bederson and J. D. Hollan, "Pad++: a zooming graphical interface for exploring alternate interface physics," in *Proceedings of the 7th annual ACM symposium on User interface software and technology*, Marina del Rey, California, United States, 1994, pp. 17-26.

[4] J. D. Cahoy, B. Emery, A. Kaushal, L. C. Foo, J. L. Zamanian, K. S. Christopherson, Y. Xing, J. L. Lubischer, P. A. Krieg, S. A. Krupenko, W. J. Thompson, and B. A. Barres, "A transcriptome database for astrocytes, neurons, and oligodendrocytes: a new resource for understanding brain development and function," *The Journal of Neuroscience,* vol. 28, pp. 264-278, 2008.

[5] J. Chen, P. Zhao, D. Massaro, L. B. Clerch, R. R. Almon, D. C. DuBois, W. J. Jusko, and E. P. Hoffman, "The PEPR GeneChip data warehouse, and implementation of a dynamic time series query tool (SGQT) with graphical interface," *Nucl. Acids Res.,* vol. 32, pp. D578-D581, 2004.

[6] P. Craig, J. Kennedy, and A. Cumming, "Animated interval scatter-plot views for the exploratory analysis of large-scale microarray time-course data," *Information Visualization,* vol. 4, pp. 149-163, 2005.

[7] K. D. Dahlquist, N. Salomonis, K. Vranizan, S. C. Lawlor, and B. R. Conklin, "GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways," *Natutre Genetics,* vol. 31, pp. 19-20, 2002.

[8] G. S. Eichler, S. Huang, and D. E. Ingber, "Gene Expression Dynamics Inspector (GEDI): for integrative analysis of expression profiles," *Bioinformatics,* vol. 19, pp. 2321-2322, 2003.

[9] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proc. Natl. Acad. Sci. U. S. A.,* vol. 95, pp. 14863-14868, 1998.

[10] R. A. Fisher, *Statistical methods for research workers*, 14th ed. Darien, Conn.: Hafner Pub. Co., 1970.

[11] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: timebox widgets for interactive exploration," *Information Visualization,* vol. 3, pp. 1-18, 2004.

[12] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, "KEGG for linking genomes to life and the environment," *Nucleic Acids Research,* vol. 36, pp. D480-D484, 2008.

[13] B. Lyn, "Perceptual and interpretative properties of motion for information visualization," in *Proceedings of the 1997 workshop on New paradigms in information visualization and manipulation*, Las Vegas, Nevada, United States, 1997, pp. 3-7.

[14] P. McLachlan, T. Munzner, E. Koutsofios, and S. North, "LiveRAC: interactive visual exploration of system management time-series data," in *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, Florence, Italy, 2008, pp. 1483-1492.

[15] T. Munzner, F. Guimbretiere, S. Tasiran, L. Zhang, and Y. Zhou, "TreeJuxtaposer: scalable tree comparison using Focus+Context with guaranteed visibility," *ACM Trans. Graphics,* vol. 22, pp. 453-462, 2003.

[16] H. Parkinson, M. Kapushesky, M. Shojatalab, N. Abeygunawardena, R. Coulson, A. Farne, E. Holloway, N. Kolesnykov, P. Lilja, M. Lukk, R. Mani, T. Rayner, A. Sharma, E. William, U. Sarkans, and A. Brazma, "ArrayExpress--a public database of microarray experiments and gene expression profiles," *Nucleic Acids Research,* vol. 35, pp. D747-D750, 2007.

[17] A. R. Pico, T. Kelder, M. P. van Iersel, K. Hanspers, B. R. Conklin, and C. Evelo, "WikiPathways: pathway editing for the people," *PLoS Biol,* vol. 6, p. e184, 2008.

[18] Z. Pylyshyn, J. Burkell, B. Fisher, C. Sears, W. Schmidt, and L. Trick, "Multiple parallel access in visual attention," *Canadian Journal of Experimental Psychology,* vol. 48, pp. 260-283, 1994.

[19] R. Ramana and K. C. Stuart, "The table lens: merging graphical and symbolic representations in an interactive focus + context visualization for tabular information," in *Proceedings of the SIGCHI conference on Human factors in computing systems: celebrating interdependence*, Boston, Massachusetts, United States, 1994, pp. 318-322.

[20] K. Robert and L. Heidi, "Line graph explorer: scalable display of line graphs using Focus+Context," in *Proceedings of the working conference on Advanced visual interfaces*, Venezia, Italy, 2006, pp. 404-411.

[21] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *IEEE Trans. Visual Comput. Graphics,* vol. 11, pp. 443-456, 2005.

[22] P. Saraiya, C. North, and K. Duca, "Visualizing biological pathways: requirements analysis, systems evaluation and research agenda," *Information Visualization,* vol. 4, pp. 191-205, 2005.

[23] J. Seo, M. Bakay, Y.-W. Chen, S. Hilmer, B. Shneiderman, and E. P. Hoffman, "Interactively optimizing signal-to-noise ratios in expression profiling: project-specific algorithm selection and detection p-value weighting in Affymetrix microarrays," *Bioinformatics,* vol. 20, pp. 2534-2544, 2004.

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

[25] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research,* vol. 13, pp. 2498-2504, 2003.

[26] E. R. Tufte, *The visual display of quantitative information*, 2nd ed. Cheshire, Conn.: Graphics Press, 2001.