An Insight-Based Longitudinal Study of Visual Analytics

Purvi Saraiya, Chris North, Vy Lam, and Karen A. Duca

Abstract—Visualization tools are typically evaluated in controlled studies that observe the short-term usage of these tools by participants on preselected data sets and benchmark tasks. Though such studies provide useful suggestions, they miss the long-term usage of the tools. A longitudinal study of a bioinformatics data set analysis is reported here. The main focus of this work is to capture the entire analysis process that an analyst goes through from a raw data set to the insights sought from the data. The study provides interesting observations about the use of visual representations and interaction mechanisms provided by the tools, and also about the process of insight generation in general. This deepens our understanding of visual analytics, guides visualization developers in creating more effective visualization tools in terms of user requirements, and guides evaluators in designing future studies that are more representative of insights sought by users from their data sets.

Index Terms—Evaluation/methodology, Graphical User Interface (GUI), information visualization, visualization systems and software, visualization and methodologies.

1 Introduction

Visualization tools are often evaluated in controlled studies [1], [2]. Participants are usually given a range of predefined tasks to perform on preselected data during the course of the study. The performance time and accuracy of the participants' responses for the selected tasks are recorded and later analyzed to evaluate the visualization tools [3]. However, such studies often fail to represent the real-world data analysis scenario, which is usually less guided and much more in-depth.

An initial attempt to capture the real-world exploratory data analysis scenario in a short-term controlled study using an insight-based methodology is reported in [4]. Though the study provided interesting observations about the visualization tools, it had limitations. It measured the insight process for an initial one to two hours of data analysis by the participants and, thus, failed at capturing the long-term insight gained by users who spend more time analyzing the data. Most of the participants in the study were unfamiliar with the tool they were working with. This was to measure the time a participant takes in getting familiar with a particular tool. However, the amount, time, and type of insight generated may change as one becomes more familiar with the visualization tool as compared to using it for the first time.

Most importantly, the participants in the insight study were unfamiliar with the experimental context of the data used in the study. Hence, the data did not mean as much to them because, simply put, it was not their data. Since the participants were not self-motivated to perform data analysis, they had to be prompted during the study to report insights. Thus, the study failed to address the most important factor—***motivation***—that drives a data analyst to spend days and often months analyzing a particular data set. Also, the study did not capture the ability of a data analyst to judge the ***significance*** of reported insights, which is usually based on users' domain knowledge and familiarity with the data background and the experimental context.

To address these issues, we performed a longitudinal study by working closely with bioinformaticians who were ready to start analyzing data from a microarray experiment [5], [6] using visualization tools. The goal of the study was to gain basic understanding into the visual analytic process. The primary research questions addressed by the longitudinal study were: How are different visualization tools used to gain insight into the data? How much effort and time are required to derive the most interesting insights (e.g. hypothesis generation [4])? What process is followed by users to get needed insights? How is insight synthesized over time? Is it by constantly discovering the unexpected trends in the data, or is it a gradual process that builds newer and deeper insights in the context of previously generated ones?

A primary use of the visualization tool is to gain insight into the data [7], [8]. For this, a visualization tool not only provides data representations, but also supports interaction mechanisms. We were also interested to learn: Which visualization techniques and interaction mechanism combinations were most effective in providing insights? And, more importantly, how do users overcome the shortcomings of a visualization tool that does not provide an interaction feature or visual representation that is needed?

One of the main purposes of conducting the studies reported in [1], [2], [3], [4] was to evaluate the visualization

- *P. Saraiya and C. North are with the Department of Computer Science, Virginia Tech, 660 McBryde Hall, Blacksburg, VA 24061-0106. E-mail: {psaraiya, north}@vt.edu.*
- *V. Lam and K.A. Duca are with Virginia Tech/Virginia Bioinformatics Institute, 1 Washington Street, Blacksburg, VA 24061. E-mail: {vlam, kduca}@vbi.vt.edu.*

Manuscript received 28 Nov. 2005; revised 27 Feb. 2006; accepted 28 Mar. 2006; published online 8 Sept. 2006.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCG-0174-1105.

1077-2626/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

tools under investigation with respect to one another. The aim of the longitudinal study reported here was not to evaluate or compare bioinformatics visualization tools with respect to each other. Instead, the primary focus was to capture the entire data analysis process that began from a raw data set and was continued until the desired insights were obtained from the data, and to examine the visualization tools' capabilities in supporting the analysis.

The data analysis scenario reported here is from the bioinformatics domain. However, the process of interpreting visual representations and interacting with them to analyze multidimensional data is not limited to the bioinformatics field. Bioinformatics involves deep analysis of large, noisy, and incomplete data to derive high-level models of reality, and hence is well representative of visual analytics in general. The study reported here provides interesting observations as to how visualization tools are actually used for data analysis. Lessons learned here should be applicable for visual analytics in other domains.

2 Literature Survey

2.1 Methods to Evaluate Visualizations

A variety of evaluation methodologies have been used to measure the effectiveness of visualizations. Many studies have evaluated visualization effectiveness through rigorously controlled experiments [2], [9] for summative or scientific hypothesis testing. In these studies, typical independent variables control aspects of the tools, tasks, data, and participant classes. Dependent variables include accuracy and performance measures. Accuracy measures include precision, error rates, number of correct and incorrect responses, whereas performance includes measures of time to complete predefined benchmark tasks. Such studies compare effectiveness of two or more tools (e.g., [3] compares three different visualization systems), or examine human visual perception (e.g., [10] compares graphical mappings of information).

Formative usability tests typically evaluate visualizations to identify and solve user interface problems. A typical method for usability studies involves observing participants as they perform designated tasks, using a "think aloud" protocol. Evaluators note the usability incidents that may suggest incorrect use of the interface, and compare results against a predefined usability specification [11]. Refer to [12] for an example of a professional formative usability study of a visualization.

Analytic evaluations include inspections of user interfaces by experts, such as with heuristics [13]. Examples of specific metrics for visualizations include expressiveness and effectiveness criteria [14], data density and data/ink [15], criteria for representation and interaction [16], high-level design guidelines [17], principles based on preattentive processing and perceptual independence [18], and rules for effectiveness of various visual properties [19]. Cognitive models, such as CAEVA [20], can be used to simulate visualization usage for analyzing the low-level effects of various visualization techniques.

A longitudinal study of information visualization adoption by data analysts [21] suggests advantages when visualizations are used as complementary products rather than stand alone products. Rieman [22] examines users' long-term exploratory learning of new user interfaces, with "eureka reports" to record learning events. An insight-based study to evaluate microarray data visualization using more realistic exploratory data analysis is reported in [4]. Three case studies, and a user survey to evaluate effectiveness of Hierarchical Clustering Explorer (HCE), a visualization tool, are reported in [23]. The authors also compare both the evaluation methods used to measure tool effectiveness based on results they provided about the tool usage.

2.2 User Studies for Bioinformatics

Biologists use microarray experiments [5], [6] to answer complex biological research questions. As these experiments result in very large data sets, biologists need computational methods to derive domain-relevant insights. A detailed description of the microarray data analysis process is in [24], [25]. Since this process is very complicated, considerable research is currently being conducted to search for new and improved methods [26], [27]. Extensive evaluations for raw data normalization and statistical algorithms for data analysis have been conducted. For example, different normalization methods based on data variance and bias are compared in [28], and [29] lists a review of statistical methods to discover differentially expressed genes. Case studies describing data analysis procedures using clustering algorithms and suggestions for new and improved methods have been published [30]. A comprehensive list of publications for this area can be obtained from [31].

A large number of information visualization tools targeting this domain have been developed [31], [32], and a number of user studies have also been conducted. A case study using GeneSifter [33] to analyze microarray data is reported in [34]. A survey of biologists' tasks for a general query system is reported in [35]. O'Day et al. [36] report observations from user studies with molecular biologists to identify information needs unmet by the current tools. End user participatory design process is used in [37] to create prototype electronic laboratory notebooks. A combination of end user interviews, heuristic evaluations, and surveys was used to elicit the end user requirements for pathway visualization software [38].

Thus, though there has been significant emphasis placed on improving data analysis techniques for bioinformatics, very few studies have actually been conducted to investigate the analytic process and the use of visualization tools from the end user's perspective.

2.3 Visual Analytics

The research agenda in [39] provides a comprehensive list of key aspects that influence visual analytics, the process by which users gain insight into complex data. For discussion here, the most relevant aspects are: science of analytical reasoning and visual representations and interaction techniques.

Visual analytics deals with the capabilities of visualization tools that help users make judgments about the data. It is important to create visualization tools that maximize human capabilities to perceive and understand complex and dynamic data. Though visual representations provide

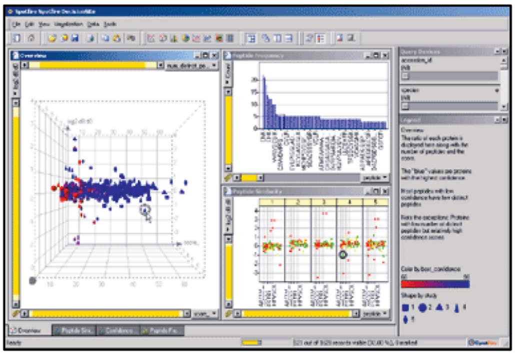


Fig. 1. Spotfire®.

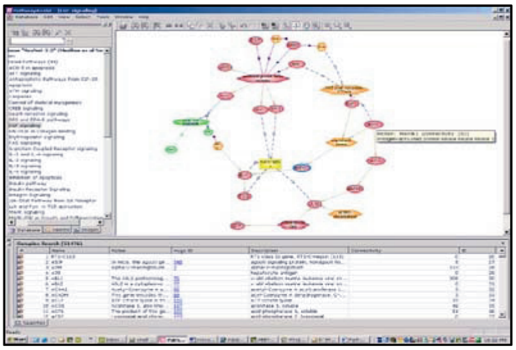


Fig. 2. PathwayAssist®.

an initial direction to data analysis, they will not be effective without interaction mechanisms that let users explore data. It is through a combination of visual representations and interaction mechanisms that a user achieves insight into the data. A detailed description about each aspect of visual analytics, along with an extensive literature survey, and suggestions for future research work is presented in [39].

Summarizing the current literature, though several studies have been conducted for visualization tools, very few have observed their long-term usage for data analysis. A longitudinal study can provide interesting observations about the analysis process involving use of visualization tools by users to gain insight into data.

3 Experiment Design

In this longitudinal study, we observed bioinformaticians over a long period of time as they analyzed their data from a microarray experiment.

3.1 User Background

Two bioinformaticians worked closely together to analyze the data and interpret the results. A postdoctorate was mainly in charge of performing the bioinformatics data analysis using software visualization tools. A bioinformatics faculty member supervised the overall analysis. Though not new to microarray technology, the bioinformaticians had little previous experience with the specific software tools used for this data analysis. Later in their analysis, they collaborated with a larger group of biologists to examine broader impacts.

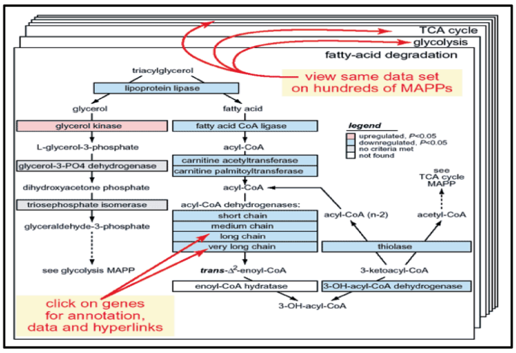
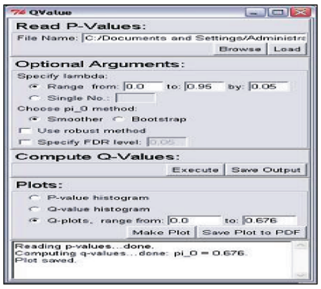


Fig. 3. GenMapp.



Read P-Values:

File Name: C:/Documents and Settings/Administra

Optional Arguments:

Specify lambda:

● Range from: 0.0 to: 0.95 by: 0.05

○ Single No.:

Choose pi_0 method:

● Smoother ○ Bootstrap

□ Use robust method

□ Specify FDR level: 0.05

Compute Q-Values:

Plots:

○ P-value histogram

○ Q-value histogram

● Q-plots, range from: 0.0 to: 0.676

```
Reading p-values...done.
Computing q-values...done: pi_0 = 0.676.
Plot saved.
```

Fig. 4. Q-Value analysis.

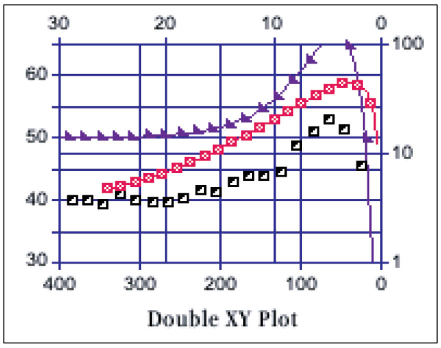


Fig. 5. KaleidaGraph.

3.2 Visualization Tools

The following visualization tools were chosen by the bioinformaticians for data analysis and reporting. Microsoft Excel was also used extensively for data formatting.

- Spotfire® [40] (Fig. 1).
- PathwayAssist® 3.0 [41] (Fig. 2).
- GenMapp [42] (Fig. 3).
- Q-Value Software [43] (Fig. 4).
- KaleidaGraph [44] (Fig. 5).

The bioinformaticians started with Spotfire and PathwayAssist 3.0 because software licenses for them were already purchased by their lab. They also tried to use other tools like Affymetrix GCOS [45], and R [46], and different versions of PathwayAssist (2.0, 3.0, and 4.0). They found that they preferred PathwayAssist 3.0. However, they did not search for other software tools rigorously, as on performing some data analysis they felt that both Spotfire and PathwayAssist supported their tasks very well. GenMapp was used because the bioinformaticians liked the mouse signaling pathways

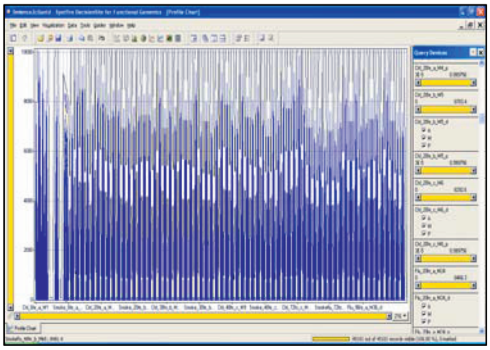


Fig. 6. The entire data set (45,001 rows × 72 columns) visualized using the profile chart in Spotfire®.

provided by that tool. They also used a Q Value software package to minimize false discoveries and, finally, KaleidaGraph to create readable static graphs to present their results.

3.3 Data Set

The data set measured mRNA expression levels from mouse lung tissue under four different conditions (control group, flu infected, tobacco smoke exposed, and both flu infected and tobacco smoke exposed). The measurements were taken for six timepoints post infection (6, 20, 30, 48, 72, and 96 hours) with three replicates for each timepoint, resulting in: four conditions × six Timepoints × three Replicates = 72 data conditions for 45,001 probe sets (genes). Thus, the data set was 45,001 rows × 72 columns.

In general, these bioinformaticians' scientific goal is to understand the pathogenesis of flu infection and the impact of tobacco smoking on that process. Their analysis is exploratory, and is not limited to simply verifying a specific hypothesis.

3.4 Protocol

To keep the experiment as close to real-world data analysis as possible, we did not require the bioinformaticians to follow an unusual protocol. They were requested to keep a diary of the process they undertook, the insights gained from the data, the visualization and interaction techniques that led to the insights, and the successes and frustrations they experienced with the software tools. We also met regularly, once every two to three weeks over a three month period, to discuss the data insights and their experience with the tools. The bioinformaticians did not perform data analysis every day, but rather based on how it fit with their normal job activities. However, when analyzing the data they usually spent about three to four hours at a time. To judge the significance of insights, at the end of data analysis, we requested the bioinformaticians to rank the data insights on a scale of 1-5, with 5 being the most significant.

An important requirement of the study was that we did not impact their normal data analysis process in any way, except for the diary keeping and debriefing meetings. We did not provide any help with the software tools or guide their data analysis in any way. The analytic process, the selection of tools, and the data were all determined by their own normal procedures that they had planned regardless of our observation.

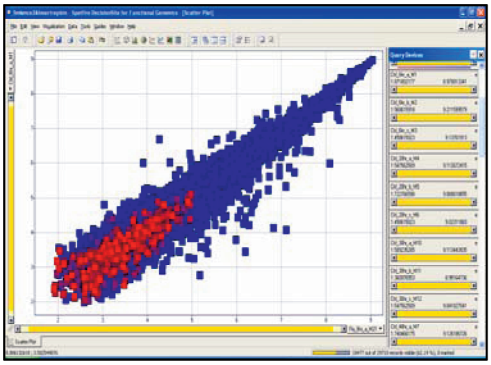


Fig. 7. Visualization of control six hours versus flu six hours using the scatter plot in Spotfire®. Each dot in the figure corresponds to a probe (or a row) in the data set. The color of each dot corresponds to expression value in smoke exposed six hours.

4 Data Analysis Procedure and Insights

The bioinformaticians started from a raw Affymetrix microarray data set. They used Microsoft Excel to convert the data into the format they needed for further analysis. Description about different file formats used by Affymetrix and their meaning and significance can be obtained from [45]. This process was nontrivial and required about 15 hours of extensive data manipulation.

Once the data was in the required format, they loaded it into Spotfire® to get an initial overview. Fig. 6 displays the visualization of the entire data set (45,001 × 72) using the profile chart (similar to parallel coordinate visualization) provided by Spotfire®. To make data analysis more manageable, they decided to filter some genes. They first removed the genes that had absent or null values in the data, by using sorting and column reordering features of Excel. For further filtering, they decided to remove genes that did not show much change from one condition to another, using dynamic queries provided by Spotfire®. The final data set had 30,000 rows.

They began the data analysis by using the scatterplot visualization in Spotfire® to plot expression (data values) for each control timepoint with respect to timepoints of the other conditions (Fig. 7). Each point in Fig. 7 corresponds to a probe value (or a row) in the data set. This was an extremely time consuming process due to combinatorial explosion. They initially wanted to use the profile chart to get an overview, however, due to the sheer volume of data they found it confusing due to visual clutter (Fig. 8 shows the profile chart for control versus flu timepoints). Thus, they had to manually check individual time points to make data size manageable.

One of their data analysis aims was to search for probes, from the entire data set, that displayed different expression values for selected conditions. Hence, they used scatter plots, since that view made it easier for them to identify outliers that displayed distinct behavior for the selected time points. They also tried to increase the dimensions visualized by coloring the plot using a third dimension. For example, in Fig. 7, though the plot visualizes control six hours versus flu six hours, the color for each dot is based on its expression values for smoke exposed at six hours.

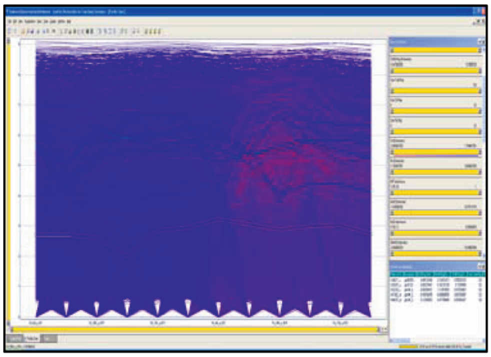


Fig. 8. Visualization of all the six timepoints for control and flu conditions in profile chart in Spotfire® for 30,000 probes.

However, they found this confusing and focused more on the layout without taking color into account.

They tried to use 3D scatterplot visualizations (Fig. 9) to have an overview of more timepoints simultaneously and save some data analysis time. However, they immediately gave up the idea as they had difficulty interpreting the visualization and found it actually took longer for them to think through the meaning this way.

They also tried K-means and SOMS clustering algorithms, and the treatment comparison feature provided by Spotfire® to get an overview of the common gene expression trends in the data. Fig. 10 shows the visualization resulting from grouping the data by 3 × 3 SOMS clustering. They also checked the clusters to verify if various familiar genes displayed the behavior they expected, and if biologically functionally related genes were appropriately grouped together. Table 1 lists the insights they obtained using these views.

The clustering algorithms group genes based on the similarity in their expression profiles. The bioinformaticians were worried to discover that genes with distinct time profiles were also grouped together. Also, the algorithms do not take into account the biological functionality of the genes. However, they liked the dynamic query interaction method provided by Spotfire® as a way to quickly explore many criteria. Hence, they decided to focus on the profile chart and scatter plot visualizations for more detailed analysis. For example, Fig. 11 displays a profile chart visualization for all genes that are up regulated for flu as compared to control conditions. Table 2 lists the insights obtained by this process. The bioinformaticians also made a list of interesting genes from the queries and saved them for further investigation.

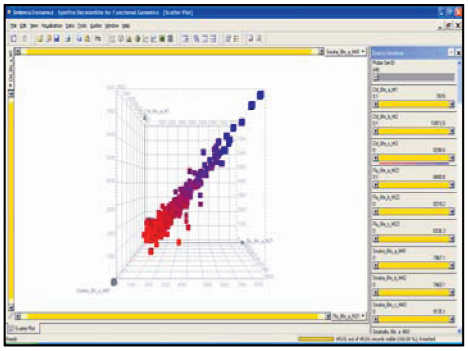


Fig. 9. Visualization of control six hours versus flu six hours versus smoke exposed six hours using the 3D scatter plot in Spotfire®. The dots are colored based on the expression value in smoke exposed + flu six hours.

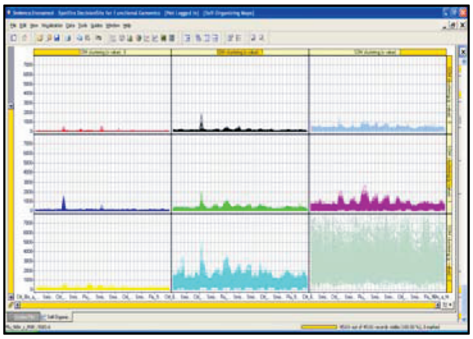


Fig. 10. A visualization of 3×3 SOMS clustering of the data.

Now that the bioinformaticians were more familiar with the data, they needed different visualizations to get more biologically relevant insights. They decided to use PathwayAssist for further data analysis involving biological pathways. Pathways are network-based models of complex biological processes [38]. They had already made lists of genes they needed to investigate further. They wanted to build pathways involving these genes, using search capabilities provided by PathwayAssist. This would show other genes that have a direct influence on these genes of interest. PathwayAssist uses NLP algorithms to extract information about relationships between genes from various search engines such as PubMed. Fig. 12 shows an initial pathway created for genes they selected. Since the visualization had more information than they could handle, they abandoned the idea of depending on pathways created

TABLE 1
Insights Gained at the Start of Data Analysis

Date	Visualization	Insight	Value
8/12	Scatter Plots	• Noticed very up-regulated genes in flu 96 vs. ctrl 96 on scatter plot. Same effect is seen in the time series.	1
8/12	SOMS Clustering	• Self-organizing-maps at 3x3 grid show some interesting profiles e.g., one where genes are only up-regulated in flu 20 hr and others only up regulated at smoke exposed+flu 30 hr. If these are same genes, then smoking delays flu induction. • Certain matrix metalloproteases are up regulated along with interferon activated genes.	4
8/12	K-Means Clustering	• K mean clustering appear to be much better than SOM in sorting out different dynamics of gene regulation, especially on IFN genes.	3
8/12	Treatment comparisons	• An important proteolytic enzyme of relevance to the group appears to be down regulated in flu, smoke exposed, and smoke exposed +flu. • Several immune system activating genes are all up regulated by smoke exposure	3

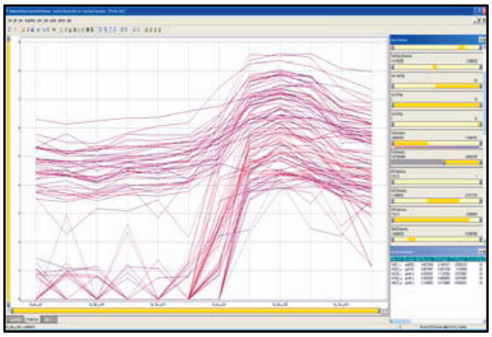


Fig. 11. Visualizes control versus flu for all six timepoints for one replicate. The display is manipulated to show genes that were upregulated for flu as compared to control condition.

TABLE 2
Insights by Using Scatter Plot and Profile Chart Visualizations along with Dynamic Queries

Date	Visualization	Insight	Value
8/15	Scatter plots	• Heat shock proteins, Retnla, cathepsins, serum amyloid A3, interferon induced proteins, certain matrix based proteins are up-regulated in flu infected mice. Also slower in up-regulated items include MHC molecules by 10 hours (in smoking).	3
8/29	Profile Chart	• Noticed that some heat shock proteins are up-regulated only at 30 hours in control mice. • Cathepsins are upregulated by flu.	4

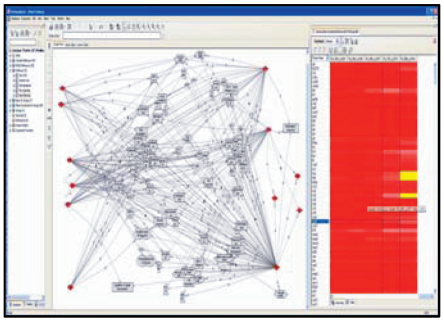
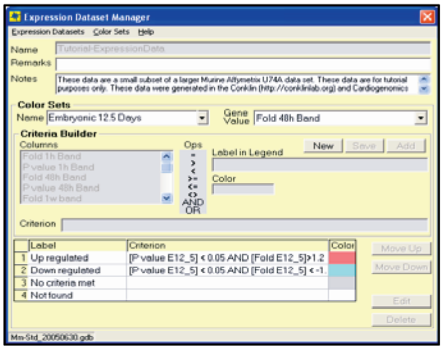


Fig. 12. Automatic pathway created by PathwayAssist for a selected list of genes.

automatically. Also, they cannot completely trust the automatic pathways created by the tool. They would have to manually curate the pathways, because the NLP algorithms usually provide some level of information that is irrelevant to their data analysis or is incorrect.

The bioinformaticians decided to focus on the apoptosis signaling pathway because their research group is most interested in that topic. They knew that GenMapp provides prebuilt pathways. Although they preferred the pathway provided by GenMapp, they had problems overlaying the expression data onto it. The expression data manager in GenMapp (Fig. 13) required them to define color scales for each individual column. Since they had 72 columns, they thought this would be time consuming. They decided to transport the GenMapp pathway to PathwayAssist and then link it to the microarray data. This involved importing genes and reconstructing the pathway.



Name Tutorial-ExpressionData

Remarks

Notes These data are a small subset of a larger Murine Affymetrix U74A data set. These data are for tutorial purposes only. These data were generated in the Conklin (http://conklinlab.org) and Cardiogenomics

Name Embryonic 12.5 Days

Gene Value Fold 48h Band

Label in Legend

Color

Criterion

	Label	Criterion	Color
1	Up regulated	[P-value E12_5] < 0.05 AND [Fold E12_5]>1.2	🟥
2	Down regulated	[P-value E12_5] < 0.05 AND [Fold E12_5] < -1.	🟦
3	No criteria met		▨
4	Not found		

Mm-Std_20050630.gdb

Fig. 13. Expression data set manager in GenMapp.

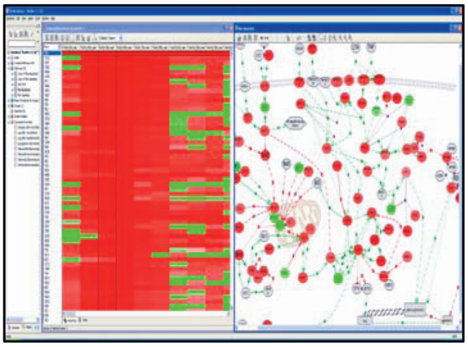


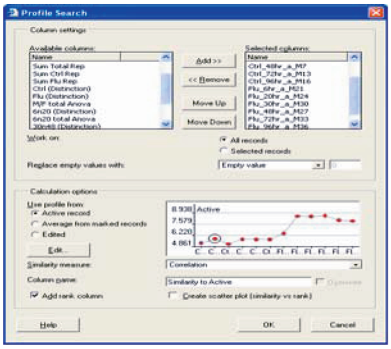
Fig. 14. Data from flu 20 hours is overlaid on cell apoptosis pathway and linked to heatmap visualization in PathwayAssist. The color is used to denote expression value, red denotes upregulated genes, whereas green implies down regulation.

TABLE 3
Insights Resulting from Using the Heatmap and Pathway Visualizations

Date	Visualization	Insight	Value
9/01	Heatmap	• A list of pathway genes that are suppressed by smoking but up-regulated by flu.	4
9/12 9/13 9/21	Pathway visualization	• The up-regulation of Mx by flu is suppressed by smoking even though smoking itself did not have an effect on basal Mx activity. • Genes involved in apoptosis are regulated, particularly DAXX which is up-regulated in flu infections. • Flipping through time points on PA, noticed that CHUK and IRAK1 of the NFKB signaling is only up-regulated in flu vs. control.	3

They utilized the heatmap visualization provided by PathwayAssist to investigate time-dependent regulation of the pathway. They found it easier to click on the column name to display data related to a particular condition on pathways in PathwayAssist (Fig. 14). Using this, they found genes that were suppressed by smoke exposure but up regulated in flu. Table 3 lists insights resulting from pathway visualization.

Along with the data analysis, the bioinformaticians also became more familiar with additional features and func-



Work on: ● All records ○ Selected records

Replace empty values with: Empty value

Use profile from: ● Active record ○ Average from marked records ○ Edited

Edit...

Active

Similarity measure: Correlation

Column name: Similarity to Active □ Overwrite

☑ Add rank column □ Create scatter plot (similarity vs rank)

Fig. 15. Profile search feature in Spotfire®.

tionalities of the visualization tools by reading the help documentation and calling technical support. They used profile search (Fig. 15) for genes that display expressions similar to a specified pattern, and statistical analysis methods such as t-tests and Anovas. Table 4 lists insights from this process.

Table 5 lists the process toward the end of their data analysis. The bioinformaticians were trying new methods to get more insights from the data to ensure that they did not unintentionally neglect any unexpected results. The complexity of the procedure indicates more familiarity with detailed features of the tools. They were also refining their findings to ensure the most accurate insights.

Toward the end of the study, the bioinformaticians were evaluating the best statistical tests to apply to the data. In addition to using t-tests and p-values to minimize the number of false-positive tests, they were also using q-value analysis to minimize the number of false discoveries [47]. The q-value offers a less conservative approach to measuring the statistical significance of genomic data than the traditional Bonferoni-corrected p-value. Although Microsoft Excel supports this analysis, they felt that the Q-value software was more suited for bioinformatics data analysis.

TABLE 4
Insights Obtained from Profile Search and Statistical Analysis

Date	Visualization	Insight	Value
9/28	Profile Search	• Used profile search to find genes that are regulated (over all conditions) similar to Mx. • Smoke exposure REALLY suppresses some heat shock accessory proteins. The NKκB system is responding similarly. Maybe through TLRs?	3
9/28	Data-Pattern-Distinction	• Few distinctive genes that may be indicative of smoking. There are several candidates. • Influenza infection is typified by the up-regulation of certain genes that are activated very early by interferon. • Retnla has a very interesting profile, up-regulated in flu and EARLY in smoking recovery! VERY interesting.	5
9/30	Anovas, t-tests	• T-test/ANOVA shows that three genes are the optimal indicators of smoking, including flu infected individuals.	4

TABLE 5
The Later Data Analysis Procedures

Date	Visualization	Data Analysis Procedure
10/31	Data formatting + Pattern Distinction + Biology Database search	• Removed absent call data points and used the discovered binary sorting to count replicate present calls. Using distinction factor (correlated to t-test) to find flu indicators in Spotfire and then export to Pathway Assist to find biological significance.
11/01	Profile search	• Examining profile search and treatment distinctions in Spotfire. Trying to find the best way to differentiate different time profiles of expression. Particularly, should absent calls be considered 0 or null?

5 Insight Presentation

The bioinformaticians work in close collaboration with another large international biological research group. They recently presented their data analysis results to the other group. Most of their presentation was related to immune system genes and used Microsoft PowerPoint slides. Since the international group is less conversant with microarray data analysis, the bioinformaticians shared their time-series data analysis experiences including data filtering and normalization methods.

The international group is primarily interested in chronic respiratory diseases, not flu infections per se. Hence, they have a different set of genes of interest than the bioinformaticians. However, the bioinformaticians were able to easily provide information about the other genes during their presentation and later meetings, by using the data filtering capabilities of Spotfire®. Spotfire® allowed them to easily narrow down the genes of interest using text filters. For a given text string, Spotfire® can list all the genes containing that text. Many genes having similar functionalities have similar names like Casp1, Casp2, etc. The search capabilities worked well to find such groups. Spotfire® also let them analyze the time profiles for selected genes (Fig. 16). The audience found the dynamic query mechanism provided by Spotfire® to be helpful because it allowed them to search for genes based on the expression values. They also performed t-testing on the fly to check significance of the results.

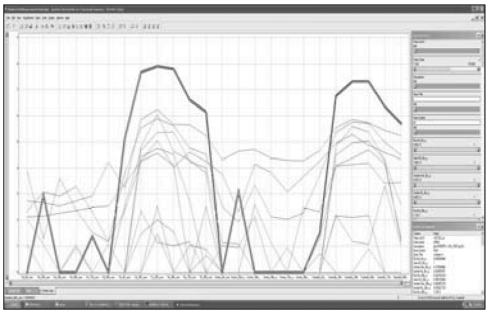


Fig. 16. Time profiles for a group of selected genes. The profile indicates suppressed values for smoking + flu condition as compared to the flu condition.

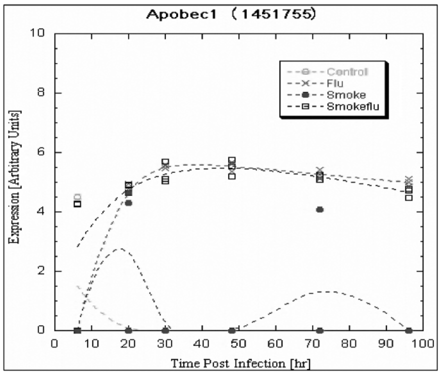


Fig. 17. Time profiles for a selected gene in KaleidaGraph. The profile is divided into four different colored lines to represent the four conditions in the data.

Based on this interactive collaboration, they discovered that smoking suppresses expression of Slfn genes. This finding was considered very exciting, since not much data is currently available for genes belonging to that family. They also concluded that smoking affects genes involved in DNA repair and some that facilitate cell cycle advance (insight value = 5). It should be noted that the domain expertise provided by the international group added considerable value to the analysis done by the two investigators.

The biologists are currently working toward publishing their data analysis results. Their main conclusions are that smoke exposure suppresses overall gene expression under conditions of flu infection. They will also report a list of genes that were found to be significantly affected, the biological functions of these genes and the overall significance of these effects on biological processes. There was one major new insight involving the DNA repair mechanisms that will be explored in future collaborative work. It should be emphasized that, despite use of software tools, a significant amount of manual exploration and the input of several biological domain experts was necessary to derive useful biological understanding from the experiment data.

They will use KaleidaGraph to graph the results and gene expression time profiles. Though other software tools allow them to easily transfer screenshots to Microsoft Word, they are accustomed to KaleidaGraph and find it better suited for simple static data presentation. They also prefer the quality of images in terms of print resolution. KaleidaGraph also provides better capabilities to manipulate graph display details, such as labeling, for presenting information. An example of a graph presenting expression profiles for a selected gene using KaleidaGraph is shown in Fig. 17. The profiles are color coded based on the four main conditions in the experimental paradigm.

6 Discussion

6.1 Data Analysis Procedure

Microarray experiments result in very large data sets that require extensive preprocessing before they can be analyzed for insight. The bioinformaticians spent about 15 hours formatting the data in Microsoft Excel. Excel was used because it provides an extensive and efficient functionality for data manipulation. Another reason for selecting Excel was the bioinformaticians' familiarity with it. They commented that even though a visualization tool may provide ways to manipulate data, they preferred Excel to save time learning new software and that Excel enabled better data sharing with other colleagues, since everyone can read the files.

Once they had formatted data they needed help to load the data into Spotfire®. They had separate files for information relate to the genes and their expression values. They called Spotfire® technical support to figure out a way to efficiently combine both the data sets into one so that they could proceed with data analysis. Although Spotfire® allowed them to combine more than one column by taking averages, they did not find a mechanism for row arithmetic. To get around this limitation, the bioinformaticians had to format the data in Excel and reimport it to Spotfire®.

The first step in data analysis described here was getting familiar with the data. The bioinformaticians used multiple visualization representations in Spotfire® for this. They started initially by using scatter plots, profile charts, and cluster visualizations provided by Spotfire®, and eventually used features such as statistical algorithms and data pattern distinction for more complex analysis. They seemed to follow the general HCI approach of "overview, zoom, and filter" in their process. For biologically relevant insights they found Spotfire® alone inadequate and needed to rely on other tools, as well as domain experts. They used the list of genes selected in Spotfire® to examine their biological functionalities and relationships with other genes in PathwayAssist. Toward the end of their process, they were trying different statistical analysis methods to ensure more accurate and statistically sound results. Moreover, they needed to ascertain that their results were robust to various choices that are commonly made in the community. There is no single accepted method as yet for microarray data analysis and their process reflected their professional judgment.

Once, the data analysis is completed, it is equally important to have an efficient mechanism to present information. Though the bioinformaticians worked with Spotfire® and PathwayAssist for analyzing the data, they needed to use yet another tool, KaleidaGraph for creating readable graphs to present their results. They found both Spotfire® and Microsoft Excel inadequate to create all data representations needed for publication, although they will use exported plots from both SpotFire and Pathway Assist in their published research report.

Thus, the bioinformaticians used a combination of different software tools during the course of their data analysis process. They picked out key features from different tools for different purposes. The use of multiple visualization tools required the bioinformaticians to export and import information, several times, from one tool to another. This required additional data formatting that was time consuming. Hence, it is important to provide better interfaces for the software tools to facilitate data exchange between them.

6.2 Effect of Interaction Mechanisms

The bioinformaticians used multiple visual representations in Spotfire® to get an initial overview of the data. One of the main reasons that they spent time exploring the data in Spotfire®, using scatter plots and profile charts, is the dynamic queries provided by the tool. They said that this might have even motivated them to spend more time with the visualization then they wanted initially. Dynamic queries provided an efficient way for them to manage a large amount of data. Rather than worrying about 45,000 probe-set values they could easily focus on the genes of their interest.

For pathway analysis, the bioinformaticians preferred PathwayAssist because the tool allowed them to easily overlay data values for a selected condition on the pathway of interest using color coding. The tool also automatically filtered out all the genes that did not belong to the pathways. Though seemingly trivial, this was one of the main reasons that encouraged the analysts to continue working with these tools. Filtering is critical for making the data set tractable for human exploration, yet they worried that they may have been missing important information in this process. They did not rigorously search for other software visualization tools, as they felt both these tools supported their tasks well.

Spotfire® served well for providing dynamic queries. Even during their short term presentations to the international collaborators, the tool provided them with an efficient way to highlight genes of interest for other researchers. Using text search mechanisms, the bioinformaticians could easily create lists of genes of interest to the other researchers and also display their time profiles.

It is important to maintain a history of user actions, and provide replication capability. The bioinformaticians spent a lot of time rearranging the data columns in Spotfire® to visualize timepoints of interest next to each other. However, each time they restarted Spotfire®, these rearrangements were lost and they had to redo them again. They had the same experience with zooming on areas of interest within visualizations. For example, when they changed the data columns in scatter plots, the zoom position was lost. Thus, these seemingly minor usability problems that developers might not have considered important had a major effect on these bioinformaticians when they had to repeat arduous operations 72 times.

Thus, methods to efficiently interact with and selectively filter the data to focus on points of interest were considered equally or even more important than the visual representations. In fact, they tended to prefer the more simple visual representations. The bioinformaticians had no trouble restarting data analysis with the selected tools even after a gap of a few days. An efficient interaction method can make the entire experience with the visualization tool and, thus, the insight generation process, more rich and enjoyable.

6.3 The Process of Insight Generation

In the reported scenario, the subjects started the data analysis process by searching for potential insights. They did not have a prior list of specific hypotheses to validate, although they used the past 20 years of interferon research as a benchmark for validating their experimental results. They wanted to find as many interesting facts as possible in the data for more detailed later exploration. Though they wanted to use profile charts, they found the scatterplot visualization more informative due to the data size (Figs. 7 and 8). However, gaining an overview of the data by examining only two columns at a time was time consuming.

They used clustering algorithms and treatment comparison features to get an initial idea about various patterns in the data. Most of these visualization features were not considered difficult to learn. But, there were many steps to execute and many combinations to explore. It is also important to interpret results from each combination in terms of biological domain knowledge to ensure that the results make sense. In fact, the most novel insights were not revealed directly by tools, but by experienced investigators who connected the patterns of changes in two particular pathways to their prior knowledge about the underlying biological processes.

An important process is not just analyzing data using different combinations, but also interpreting which combination is best suited to analyze a particular data set. For instance, Spotfire® provides several different clustering algorithms including SOMS, k-means, and hierarchical clustering. Interpreting how each method groups the genes, and resolving conflicting results from these methods can take time. Similarly, different normalization methods yield different results from the data. Hence, selecting the appropriate method depends on understanding how each method affects the data in terms of experimental context. This clearly suggests the influence of domain knowledge on data analysis.

The bioinformaticians decided not to rely on clustering algorithms for data analysis because the algorithms grouped genes with nondistinct time profiles in similar groups, and the algorithms did not take biological functions of the genes into account.

Later, the bioinformaticians used profile chart visualizations to explore the data in more details. The visual representation along with the dynamic query interaction mechanism provided a valuable combination to explore the data. They found they could easily combine many different queries to filter data, resulting in a high user satisfaction. They used this technique to find a list of interesting genes specific to a particular biological function to focus on. They were especially interested in finding genes that were differentially expressed in smoke exposure + flu condition as compared to the flu condition. This would indicate infection-related genes that were affected by smoking. They spent about one to two weeks exploring the data in profile charts. They needed the experience of exploring many possible combinations to simply observe all facets of the data. This gave them confidence in their coverage, and resulted in some serendipitous findings as well.

For domain specific information, they needed more biologically relevant visualizations. Though Spotfire® ontology gave them some clues about patterns of expression for functionally related genes, it was not sufficient. The bioinformaticians needed to see the interactions of the genes that they selected with respect to other genes that have a direct influence on them. They decided to use pathway visualizations in PathwayAssist for this. They initially decided to use the gene list to create pathways automatically. However, since the queries resulted in too much information that was difficult to comprehend and interact with in the visualization (Fig. 12), they decided to manually curate pathways. The process of pathway analysis was more complex and required about two to three weeks of data analysis and interaction with the tools. In general, it seemed a constant struggle for the bioinformaticians to continually reduce the complexity of the data to a comprehendible amount. Even with the use of visualization tools, they were forced to focus on smaller pieces so that they could wrap their minds around the observed biological behaviors.

The bioinformaticians found the most exciting insights after almost 1.5 months of data analysis and several months of "learning" time with the software. However, from the values of insights reported earlier, it is clear that later analysis is influenced by findings from the earlier analysis. Also, the bioinformaticians used more complex queries and features in the tools to reach them. This suggests more familiarity and confidence with the tools. Moreover, they feel that despite a state-of-the-art analysis, there is much untapped information waiting for mining by different domain experts.

Once they were done with pathway analysis, they then used other visualizations in Spotfire® to ensure that they did not unintentionally miss any unexpected insight from the data. The later data analysis process dealt with analyzing their insights and to ensure correct statistical interpretation. They also tried another data formatting method to check if this resulted in any other insights or conflicts with earlier observations. Their most recent data analysis involves capabilities of more than one visualization tool simultaneously, requiring a lot of back-and-forth processing.

From the discussion, it is clear that the choice of visualization methods used to analyze the data is based on the subjects' domain knowledge. Discovering an appropriate visual representation and procedure to interpret the data could be considered procedural insight. This is usually a nontrivial task, and requires trial-and-error attempts with many combinations. The subjects reported that in the future they will be able to analyze a similar data set in a relatively shorter time. Such use of learned domain knowledge is very difficult to reproduce in short-term experiments.

6.4 Longitudinal Methodology

We performed a longitudinal study to analyze how visualizations are used to gain insight into the data. To do this, we worked with bioinformaticians who were doing data analysis. Since they had an undeniable motivation in performing the data analysis, it was possible for us to observe the process for an extended period of time. It would have been impossible to perform this study if the subjects were not intrinsically interested in the data. Hence, they were able to provide us with more meaningful feedback about insights and their utility than [4].

To keep the data analysis as natural as possible, we worked primarily through a research diary maintained by the subjects. This saved us from having to continuously observe the user. It also indicates the viability of a self-reporting approach to longitudinal insight studies. The bioinformaticians did not have to do anything difficult beyond maintaining the research diary, in which they noted insights and captured screen shots. Most of these notes are things they would want to capture anyway. The data analysis process proceeded according to their normal job activity. Thus, the longitudinal study requirements were very light in extra effort for subjects and straightforwardly manageable for the evaluators.

Since the study lasted an extended period of time, it was possible for us to study the long-term insight generation process. We were also able to observe the use of different visual representations and interaction techniques over a long period of time. Due to their familiarity, the subjects could provide more relevant feedback about their insights and about the visualization tools and their limitations, including long-term usability problems. Thus, the longitudinal study enabled observations that would not have been possible in a short term study.

The goal of this methodology was not to make statistically valid comparative claims, as in a controlled experiment. However, it is possible that a comparison of tools could be made according to how the tools' use evolves. In this study, we see clear distinctions between Spotfire® and PathwayAssist, and between different visual representations within Spotfire®, in terms of how users eventually used them and what types of insights were gained. In any case, this phenomenological approach presumably best represents actual tool usage.

Though the study reported here provides interesting observations about the visualization usage, it may have been more helpful to maintain a complete detailed record of the data analysis process. The subjects could have recorded an hourly diary of their thought processes that led to specific insights. Similarly, a complete digital record of actions performed in the visualizations could be captured by instrumenting the software or recording the screen. Correlating these records could allow a much deeper investigation of users' thought processes, leading to a more detailed insight generation pattern. However, the greater the amount of instrumentation, the less natural it will be for subjects (the Uncertainty Principle). In this study, we balanced the trade-off toward minimizing efforts on the subjects part, while still getting a good amount of detailed feedback on the visualization usage.

7 Conclusions

The longitudinal study reported here serves three important functions:

First, it gives us a deeper understanding about the visual analytics process and practices of actual data analysts, in this case bioinformatics scientists, about the nature of the insight that analytics produces, and about the way that visualizations generate this insight. We can see how the analysts proceed through several phases of analysis, and how they gather and build up insight over long periods of time through the exploratory use of multiple visualization tools and techniques, and by deeply connecting the data to extensive domain knowledge and expertise. They begin at a high level to familiarize themselves with the data, and then focus in on many combinations of details, simplifying by filtering out data they deem irrelevant or unhelpful, and carefully refining their findings over time. They can also then re-expand their foci by interactively collaborating with other experts to broaden the picture.

Second, it guides visualization designers in constructing tools that better match this deep analytic process. For example, designers must recognize the way in which analysts exhaustively explore many alternatives in combinatorial fashion, how a significant amount of manual manipulation is needed and many tedious actions must be repeated, how multiple tools must be combined at the individual feature level, and how specific visual representations and interaction techniques are perceived by users and lead to insight.

Third, it guides visualization evaluators in designing studies that better examine the long-term effect of visualization tools on visual analytics. This case study indicates the viability and importance of a longitudinal, motivated, domain embedded, self-reporting approach to evaluating visualizations, identifying their learnability and insight generation capability, and examining their role in visual analytics.

Clearly, this longitudinal study is just the beginning of this line of work, and there is much more research to be done. More studies need to be conducted with more subjects and tools in diverse domains, in order to extract broader abstractions and patterns of the visual analytics process. Further analysis can lead to a more rigorous diary keeping framework that enables better coding of insight and the analytic process. Eventually, this could lead to improved insight representation methods, and a better ability to compare visualization tools within the deeper analytic context.

ACKNOWLEDGMENTS

The authors gratefully thank all the domain experts who helped make this study possible, and Dr. Rosa Gualano and Dr. Gary Anderson in the Department of Pharmacology at the University of Melbourne for contributions and helpful discussions of results.

REFERENCES

[1] C. Plaisant, "The Challenge of Information Visualization Evaluation," *Proc. Conf. Advanced Visual Interfaces (AVI '04),* 2004.

[2] C. Chen and M. Czerwinski, "Empirical Evaluation of Information Visualizations: An Introduction," *Int'l J. Human-Computer Studies,* vol. 53, pp 631-635, 2000.

[3] A. Kobsa, "An Empirical Comparison of Three Commercial Information Visualization Systems," *Proc. IEEE InfoVis Conf.,* pp. 123-130, 2001.

[4] P. Saraiya, C. North, and K. Duca, "An Insight-Based Methodology for Evaluating Bioinformatics Visualization," *IEEE Trans. Visualization and Computer Graphics,* vol. 11, no. 4, July/Aug. 2005.

[5] D. Duggan, B. Bittner, Y. Chen, P. Meltzer, and J. Trent, "Expression Profiling Using cDNA Microarrays," *Nature Genetics,* vol. 21, pp. 11-19, Jan. 1999.

[6] L. Shi, "DNA Microarray—Genome Chip," http://www.genechips.com/GeneChips.html#What, 2002.

[7] R. Spence, *Information Visualization.* Addison-Wesley, 2001.

[8] S. Card, J.D. Mackinlay, and B. Shneiderman, *Readings in Information Visualization—Using Visualization to Think.* San Francisco: Morgan Kaufmann, 1999.

[9] C. Chen and Y. Yu, "Empirical Studies of Information Visualization: A Meta-Analysis," *Int'l J. Human-Computer Studies,* vol. 53, pp. 851-866, 2000.

[10] P. Irani and C. Ware, "Diagramming Information Structures Using 3D Perceptual Primitives," *ACM Trans. Computer-Human Interaction,* vol. 10, no. 1, pp. 1-19, 2003.

[11] H. Hartson and D. Hix, *Developing User Interfaces: Ensuring Usability through Product and Process.* John Wiley, 1993.

[12] G. Rao and D. Mingay, "Report on Usability Testing of Census Bureau's Dynamaps CD-ROM Product," http://infovis.cs.vt.edu/cs5764/papers/dynamapsUsability.pdf, 2001.

[13] J. Nielsen, "Finding Usability Problems through Heuristic Evaluation," *Proc. ACM Int'l Conf. Human-Computer Interaction (CHI '92),* pp 373-380, 1992.

[14] J.D. Mackinlay, "Automating the Design of Graphical Presentations of Relational Information," *ACM Trans. Graphics,* vol. 5, pp. 110-141, 1986.

[15] E. Tufte, *The Visual Display of Quantitative Information.* Graphics Press, 1983.

[16] C. Freitas, P. Luzzardi, R. Cava, M. Pimenta, A. Winckler, and L. Nedel, "Evaluating Usability of Information Visualization Techniques," *Proc. Advanced Visual Interfaces Conf. (AVI '02),* pp. 373-374, 2002.

[17] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy," *Proc. IEEE Symp. Visual Languages '96,* pp. 336-343, 1996.

[18] C. Ware, *Information Visualization: Perception for Design.* Morgan Kaufmann, 2004.

[19] W.S. Cleveland, *The Elements of Graphing Data.* Hobart Press, 1994.

[20] O. Juarez, "CAEVA: Cognitive Architecture to Evaluate Visualization Applications," *Proc. Int'l Conf. Information Visualization (IV '03),* pp. 589-595, 2003.

[21] V. Gonzales and A. Kobsa, "A Workplace Study of the Adoption of Information Visualization Systems," *Proc. I-KNOW '03: Third Int'l Conf. Knowledge Management,* pp 92-102, 2003.

[22] J. Rieman, "A Field Study of Exploratory Learning Strategies," *ACM Trans. Computer-Human Interaction,* vol. 3, pp. 189-218, 1996.

[23] J. Seo and B. Shneiderman, "Knowledge Discovery in High Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework," *IEEE Trans. Visualization and Computer Graphics,* vol. 12, no. 3, May/June 2006.

[24] H. Causton, J. Quackenbush, and A. Brazma, *A Beginners Guide to Microarray Data Analysis.* Blackwell Publishing, 2003.

[25] D. Berrar, W. Dubitzky, and M Granzow, *A Practical Approach to Microarray Data Analysis.* Kluwer Academic Publishers, 2004.

[26] Cambridge Healthtech Inst. Data Visualization and Interpretation, Making Breakthroughs Possible in the Omics Research, http://www.healthtech.com/2002/dvs/, 2002.

[27] Cambridge Healthtech Inst. Data Visualization and Interpretation, Deciphering the Data Deluge, http://www.healthtech.com/2003/mde/index.asp, 2003.

[28] B.M. Bolstad, R.A. Irizarry, M. Astrand, and T.P. Speed, "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance," *Bioinformatics,* vol. 19, pp. 185-193, 2003.

[29] W. Pan, "A Comparative Review of Statistical Methods for Discovering Differentially Expressed Genes in Replicated Microarray Experiments," *Bioinformatics,* vol. 18, no. 4, pp. 546-554, 2002.

[30] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Science US,* vol. 95, pp. 14863-14868, 1998.

[31] Y. Leung, "Functional Genomics," http://genomicshome.com, 2004.

[32] N. Bolshakova, *Microarray Software Catalogue,* http://www.cs.tcd.ie/Nadia.Bolshakova/softwaretotal.html, 2004.

[33] GeneSifter, http://www.genesifter.net/web/, 2006.

[34] N.E. Olson, "Identification of Cell Cycle Genes Regulated during Erythroid Differentiation," http://www.microarraysuccess.org/web/, 2005.

[35] R. Stevens, C. Goble, P. Baker, and A. Brass, "A Classification of Tasks in Bioinformatics," *Bioinformatics,* vol. 17, no. 2, pp 180-188, 2001.

[36] V.L. O'Day, A. Adler, A. Kuchinsky, A. Bouch, "When Worlds Collide: Molecular Biology as Interdisciplinary Collaboration," *Proc. European Conf. Computer Supported Cooperative Work,* 2001.

[37] W.E. Mackay, G. Pothier, C. Letondal, K. Bøegh, H.E. Sørensen, "The Missing Link: Augmenting Biology Laboratory Notebooks," *Proc. Symp. User Interface Software and Technology,* pp. 41-50, 2002.

[38] P. Saraiya, C. North, and K. Duca, "Visualization of Biological Pathways: Requirements Analysis, Systems Evaluation, and Research Agenda," *Information Visualization,* vol. 4, no. 3, 2005.

[39] J. Thomas and K. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytics,* IEEE Press, http://nvac.pnl.gov/agenda.stm, 2006.

[40] Spotfire® Decisionsite for Functional Genomics, http://www.Spotfire.com, 2006.

[41] PathwayAssist® Ariadne Genomics, 2006, http://www.ariadnegenomics.com/products/pathway.html.

[42] GenMapp, Gene Map Annotator and Pathway Profiler, 2006, www.genmapp.org.

[43] Q Value Software, http://faculty.washington.edu/jstorey/qvalue/, 2006.

[44] KalaidaGraph, http://www.synergy.com/, 2006.

[45] Affymetrix, http://www.affymetrix.com/index.affx, 2006.

[46] The R Project for Statistical Computing, http://www.r-project.org/, 2006.

[47] J. Storey and R. Tibshirani, "Statistical Significance for Genome-wide Studies," *Proc. Nat'l Academy of Sciences,* vol. 100, no. 16, pp. 9440-9445, 2003.

Purvi Saraiya received the BE (Gujarat University, India) and MS (Virginia Tech) degrees in computer engineering and computer science. She is a PhD candidate in the Department of Computer Science at Virginia Tech. She is a member of the Information Visualization Group and Center for Human Computer Interaction at Virginia Tech. Her research interests include design and evaluation of information visualization software and user interface software.

Chris North received the PhD degree from the University of Maryland, College Park. He is an associate professor of computer science at Virginia Polytechnic Institute and State University, is head of the Laboratory for Information Visualization and Evaluation, and a member of the Center for Human-Computer Interaction. His current research interests are information visualization, high-resolution displays, and evaluation methods.

Vy Lam received the BS (Northwestern University) and the PhD (University of Wisconsin—Madison) degrees in chemical and biological engineering. He is currently a postdoctoral fellow at the Virginia Bioinformatics Institute at Virginia Tech. His interests are in virology, viral immunology, bioinformatics, the systems biology of virus-host interactions, and the development of innovative techniques to modulate these interactions for biomedical applications.

Karen A. Duca received the BS (University of Massachusetts at Boston) and MS (Northeastern University) degrees in chemistry and the PhD degree in biophysics and structural biology (Brandeis University). She is a research assistant professor at the Virginia Bioinformatics Institute and an adjunct assistant professor of biology at Virginia Tech. Her interests are in the development of linked experimental and computational methods for biotechnology/biomedicine, the systems biology of host-virus interactions, and quantitative imaging methods in virology and viral immunology.