An Insight-Based Methodology for Evaluating Bioinformatics Visualizations

Purvi Saraiya, Chris North, and Karen Duca

Abstract—High-throughput experiments, such as gene expression microarrays in the life sciences, result in very large data sets. In response, a wide variety of visualization tools have been created to facilitate data analysis. A primary purpose of these tools is to provide biologically relevant insight into the data. Typically, visualizations are evaluated in controlled studies that measure user performance on predetermined tasks or using heuristics and expert reviews. To evaluate and rank bioinformatics visualizations based on real-world data analysis scenarios, we developed a more relevant evaluation method that focuses on data insight. This paper presents several characteristics of insight that enabled us to recognize and quantify it in open-ended user tests. Using these characteristics, we evaluated five microarray visualization tools on the amount and types of insight they provide and the time it takes to acquire it. The results of the study guide biologists in selecting a visualization tool based on the type of their microarray data, visualization designers on the key role user interaction techniques, and evaluators on a new approach for evaluating the effectiveness of visualizations for providing insight. Though we used the method to analyze bioinformatics visualizations, it can be applied to other domains.

Index Terms—Evaluation/methodology, graphical user interfaces (GUI), information visualization, visualization systems and software, visualization techniques and methodologies.

1 Introduction

The advent of microarray experiments [1], [2] is causing a shift in the way biologists do research; a shift away from simple reductionist testing on a few variables toward systems-level exploratory analysis of thousands of variables simultaneously [3]. These experiments result in data sets that are very large. Biologists use these data to infer complex interactions between genes and proteins. Due to its magnitude, it is prohibitively difficult to analyze microarray data without the help of computational methods. Hence, the biologists use various data visualizations to derive domain-relevant insights. The main purpose in using these visualizations is to gain insight into the extremely complex and dynamic functioning of living cells.

In response to these needs, a large number of visualization tools targeted at this domain have been developed [4], [5], [6]. However, in collaborations with biologists, we received mixed feedback and reviews about these tools. With such a wide variety of available options, we need an evaluation method that allows biologists to choose the right tool for their needs. The method should address the open-ended and exploratory nature of the biologists' tasks, and allow us to determine if the tools provide insights valuable to their end users.

A primary purpose of visualization is to generate ***insight*** [7], [8]. The main consideration for any life science researcher is discovery. Arriving at an insight often sparks the critical breakthrough that leads to discovery: suddenly seeing something that previously passed unnoticed or seeing something familiar in a new light. The primary function of any visualization and analysis tool is to make it easier for an investigator to glean insight, whether from their own data or from external databanks. A measure of an effective visualization can also be its ability to generate unpredicted new insights, beyond predefined data analysis tasks. After all, visualization should not only enable biologists to find answers but to also find questions that identify new hypotheses.

We sought to evaluate a few popular microarray data visualization tools, such as Spotfire® [9]. Some research questions we addressed are: How successful are these tools in assisting the biologists in arriving at domain-relevant insights? How do various visualization techniques affect the users' perception of data? How does a user's background affect the tool usage? How do visualizations support hypothesis generation and suggest directions for future investigation? Most importantly, can insight be measured in a controlled experimental setting, uniformly across a group of participants? Our primary focus here is on insight.

Typically, visualization evaluations have previously focused on controlled measurements of user performance and accuracy on predetermined tasks [10], [11]. However, to answer these research questions requires an evaluation methodology that better addresses the needs of the bioinformatics data analysis scenario. Hence, we developed an evaluation protocol that focuses on recognition and quantification of insights gained from actual exploratory use of visualizations [12]. This paper presents a detailed explanation and discussion of the methodology, as well as detailed results of applying the method to bioinformatics visualizations.

- *P. Saraiya and C. North are with the Department of Computer Science, Virginia Tech, Blacksburg, VA 24061. E-mail: {psuraiya, north}@vt.edu.*
- *K. Duca is with the Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA 24061. E-mail: kduca@vbi.vt.edu.*

Manuscript received 18 Sept. 2004; revised 22 Dec. 2004; accepted 30 Dec. 2004; published online 10 May 2005.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org, and reference IEEECS Log Number TVCGSI-0111-0904.

Published by the IEEE Computer Society

2 Related Work

A variety of evaluation methodologies have been used to measure effectiveness of visualizations.

2.1 Controlled Experiments

Many studies have evaluated visualization effectiveness through rigorously controlled experiments [10], [11] for summative testing 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 efficiency measures. Accuracy measures include precision, error rates, number of correct and incorrect responses, whereas efficiency includes measures of time to complete predefined benchmark tasks. Such studies compare the effectiveness of two or more tools (e.g., [13] compares three different visualization systems) or examine human visual perception (e.g., [14] compares mappings of information to graphical design).

2.2 Formative Usability Testing

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 [15]. Refer to [16] for an example of a professional formative usability study of a visualization.

2.3 Metrics, Heuristics, and Models

Analytic evaluations include inspections of user interfaces by experts, such as with heuristics [17]. Examples of specific metrics for visualizations include expressiveness and effectiveness criteria [18], data density and data/ink [19], criteria for representation and interaction [20], high-level design guidelines [21], principles based on preattentive processing and perceptual independence [22], and rules for effectiveness of various visual properties [23]. Cognitive models, such as CAEVA [24], can be used to simulate visualization usage and thereby examine the low-level effects of various visualization techniques.

2.4 Longitudinal, Case, and Field Studies in Realistic Settings

A longitudinal study of information visualization adoption by data analysts [25] suggests advantages when visualizations are used as complementary products rather than standalone products. Rieman [26] examines users' long-term exploratory learning of new user interfaces, with "eureka reports" to record learning events. These studies come the closest to examining open-ended usage.

Thus, a range of evaluation methods has been used to measure effectiveness of visualizations. In the literature, controlled experiments using predefined tasks are the most prevalent for identifying and validating more effective visualizations. Unfortunately, these studies provide results for only the set of predefined tested tasks. These predefined tasks are often a poor representation of actual visualization usage because they must be overly simplistic and search-like to enable definitive scoring.

In an attempt to promote evaluation, benchmark data sets and tasks were created for the IEEE InfoVis 2003 contests [27] with the hope of focusing the submitters' attention on insights. However, it was difficult to judge the visualizations based on just the tasks and data sets [28]. To better measure the benefits of open-ended discoveries using visualizations, we need a new evaluation method that focuses primarily on the visualizations ability to generate insight.

3 Pilot Study

The main challenge we faced was precisely defining insight and how to measure it. The word "insight" in ordinary usage is vague and can mean different things to different people. However, for the purpose of our study, we needed this term to be quantifiable and reproducible. To examine this, we undertook an initial pilot study to observe how users recognized and categorized information obtained from microarray data using visualization tools with limited training. We used both GeneSpring® [29] and Spotfire® [9] to ascertain that these commercial tools were not too difficult to learn and could be used by novice as well as expert users.

As the pilot experiment was exploratory in nature, we presented no strict protocol as to how users ought to proceed. We recruited five subjects at our institute to participate. As our recruits had no prior experience using these particular tools, we reduced their initial learning time by offering a brief introduction to the tool they would use along with a summary of the different visualization techniques provided by the tool. Users were encouraged to think aloud and report any findings they had about the data set. Pilot participants were supplied two data sets to work with, a table containing fake data that contained information about just 10 genes and the Lupus data set used in the final experiment (Section 4.1). We selected the smaller data set to help users become familiar with the visualization techniques. Once comfortable with using the visualization tool, users were instructed to move onto the Lupus data.

Due to the volume and rapidity of observations reported, we concluded that we needed to record any future sessions on videotape. We also discovered that the users grew weary analyzing the practice data set, despite being told that it was just a learning aid. They tended to spend too much time on it and, by the time they began looking at actual data, they were already fatigued. We found that our test subjects could learn a visualization technique just as quickly from real data, hence, we decided to use only the real data for final experiments. From the users' comments, we recognized various quantifiable characteristics of "insight."

3.1 Insight Characteristics

To measure insights gained from visualization, a rigorous definition and coding scheme is required. We recognized in the pilot that we could capture and characterize specific individual insights as they occurred in the participants' open-ended data analysis process. This provided more detailed information about the insight capabilities of the tools than subjective measures from postexperiment surveys.

We define an *insight* as an individual observation about the data by the participant, a unit of discovery. It is straightforward to recognize insight occurrences in a think-aloud protocol as any data observation that the user mentions. The following quantifiable characteristics of each insight can then be encoded for analysis. We applied this scheme in the main experiment. Although we present them here in the context of biological and microarray data, we believe that this can be applied to other data domains as well. The characteristics of each insight are:

- **Observation:** The actual finding about the data. We counted distinct data observations by each participant.
- **Time:** The amount of time taken to reach the insight. Initial training time is not included.
- **Domain Value:** The value, importance, or significance of the insight. Simple observations such as "Gene A is high in experiment B" are fairly trivial, whereas more global observations of a biological pattern such as "deletion of the viral NS1 gene causes a major change in genes relating to cytokine expression" are more valuable. The domain value is coded on a scale of 1 to 5 by a biology expert familiar with the results of the data. In general, trivial observations earn 1-2 points, insights about a particular process earn an intermediate value of 3, and insights that confirm, deny, or create a hypothesis earn 4 or 5 points.
- **Hypotheses:** Some insights lead users to identify a new biologically relevant hypothesis and direction of research. These are most critical because they suggest an in-depth data understanding, relationship to biology, and inference. They lead biologists toward "continuing the feedback loop" of the experimental process in which data analysis feeds back into design of the next experimental iteration [30].
- **Directed versus Unexpected:** Directed insights answer specific questions that users want to answer. Unexpected insights are additional exploratory or serendipitous discoveries that were not specifically being searched for. This distinction is recognized by asking participants to identify specific questions they want to explore about the data set at the beginning of the trial.
- **Correctness:** Some insights are incorrect observations that result from misinterpreting the visualization. This is coded by an expert biologist and visualization expert together.
- **Breadth versus Depth:** Breadth insights present an overview of biological processes, but not much detail, e.g., "there is a general trend of increasing variation in the gene expression patterns." Depth insights are more focused and detailed, e.g., "gene A mirrors the up-down pattern of gene B, but is shifted in time." This also is coded by a domain expert.
- **Category:** Insights are grouped into four main categories: overview (overall distributions of gene expression), patterns (identification or comparison across data attributes), groups (identification or comparison of groups of genes), and details (focused information about specific genes). These common categories were identified from the pilot experiment results after insights were collected.

4 Experiment Design

The aim of the main study is to evaluate five popular bioinformatics visualization tools in terms of the *insight* that they provide to the users. A 3×5 between-subjects design examines these two independent variables:

1. Microarray data sets, three treatments:
   - Timeseries data set—five time-points
   - Virus data set (Categorical)—three viral strains
   - Lupus data set (Multicategorical)—42 healthy, 48 patients
2. Microarray visualization tool, five treatments:
   - Clusterview
   - TimeSearcher
   - HCE
   - Spotfire
   - GeneSpring

4.1 Microarray Data Sets

To examine a range of data scenarios, we used data from three common types of microarray experiments. The data sets are all quantitative, multidimensional data. Values represent a gene's measured activity level (or *gene expression*) with respect to a control condition. Hence, higher (lower) values indicate an increased (decreased) gene activity level. Since our study is focused on the interactive visualization portion of data analysis, the data sets were preprocessed, normalized, prefiltered, and converted to the required formats (as discussed in [31] and [32]) in advance. In general, the biologists' goal is to identify and understand the complex interactions among the genes and conditions, essentially to reverse engineer the genetic code. The following three data sets were used.

4.1.1 Time-Series Data Set

Users were given an unpublished data set from Karen Duca's lab [33]. HEK293 cells, a human embryonic kidney cell line, were infected with the A/WSN/33 strain of influenza virus in vitro at an MOI of 5. At defined time points across the entire viral replication cycle in vitro, mRNA was extracted from infected and mock-infected cultures. The values in the columns of Table 1 were the $\log_2$ of the normalized ratios of experimental signal to control signal. The data set used for analysis had 1,060 rows (genes) over five time points. Two additional columns represent the gene name and standard ID.

4.1.2 Viral Data Set

Part of a published data set from Michael Katze's lab [34] was given to users. A549 cells, a human lung epithelial cell line, were infected with one of three influenza viruses in vitro (wild type A/PR/8/34, recombinant strain of PR8 with the NS1 partially deleted, called NS1 (1-126), recombinant strain derived from PR8 with the NS1 gene completely deleted, called delNS). Other than in the NS1 gene, all three viruses are identical. At 8 hours postinfection, mRNA was extracted from infected and mock-infected cultures. The data set used for analysis (shown in Table 2) had three columns (representing the three viral conditions) and 861 rows (genes). Two additional columns represent the gene name and standard ID.

4.1.3 Lupus Data Set

Participants were presented a subset of published data from Timothy Behren's lab [35]. In this study, after blood draw, peripheral blood mononuclear cells (PBMCs), comprised of monocytes/macrophages, B and T lymphocytes, and NK

TABLE 1
Time-Series Data Set Used in the Experiment

GeneName	GenBankId	1.5 Hr	4 hr	6 Hr	8 Hr	12 Hr
aquaporin 4	AA001003	1.54	-0.21	1.49	-0.12	0.96
…	…	…	…	…	…	…

TABLE 2
Viral Data Set Used in the Experiment

Name	Description	wt PR8	NS1 (1-126)	delNS1
ADCY9	adenylate-cyclase-9	0.54	0.91	5.8
…	…	…	…	…

TABLE 3
Lupus Data Set Used in the Experiment

Accession #	Gene	Ctrl 1	…	Ctrl 42	SLE 1	…	SLE 48
AB008775	Aquaporin 9	-63.7	…	100.1	4418.	…	3433.2
…	…	…	…	…	…	…	…

cells, were isolated from control and Systemic Lupus Erythematosus (SLE) samples. mRNA was harvested for expression profiling using Affymetrix technology [36]. The column values in Table 3 represented expression values (average difference or AD) for each gene. Scaling was performed to allow comparison between chips. The data set had 90 columns (consisting of gene expression from 48 SLE samples and 42 healthy control samples) and 170 rows (genes). Two additional columns represent the gene name and standard ID.

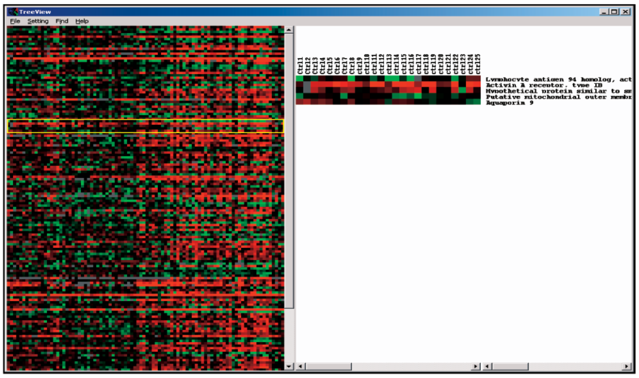
4.2 Microarray Visualization Tools

For practical reasons, we limited this study to five microarray visualization tools. We chose the tools based on their popularity and availability. We attempted to select a set of tools that would span a broad range of analytical and visual capabilities and techniques. Cluster/Treeview (Clusterview) [37], TimeSearcher [38], and Hierarchical Clustering Explorer (HCE) [39], [40] are free tools, while Spotfire [9] and GeneSpring [29] are commercial. Table 4 summarizes the visualization and interaction techniques supported by each tool.

TABLE 4
Visualization and Interaction Techniques

Tool	Visual Representations	Interactions
Cluster/ Treeview	Heat-map, Clustered heat-map	O+D
Time-Searcher	Parallel coordinates, line graph	Brushing, O+D, DQ
HCE	Cluster dendrogram, parallel coordinates, heat-map, scatterplot, histogram	Brushing, Zooming, O+D, DQ
Spotfire® 7.2 Functional Genomics	Parallel coordinates, heat-map, scatterplots (2D/3D), histogram, bar/pie chart, tree view, spreadsheet view, Clustered parallel coordinates	Brushing, Zooming, O+D, DQ
GeneSpring ® 5.0	Parallel coordinate, heat-map, scatterplots (2D/3D), histogram, bar chart, block view, physical position view, array layout view, pathway view, spreadsheet view, compare gene to gene, Clusterested parallel coordinates	Brushing, Zooming

This table summarizes the visualization and interaction techniques supported by each visualization tool O+D = overview+details; DQ = dynamic queries.



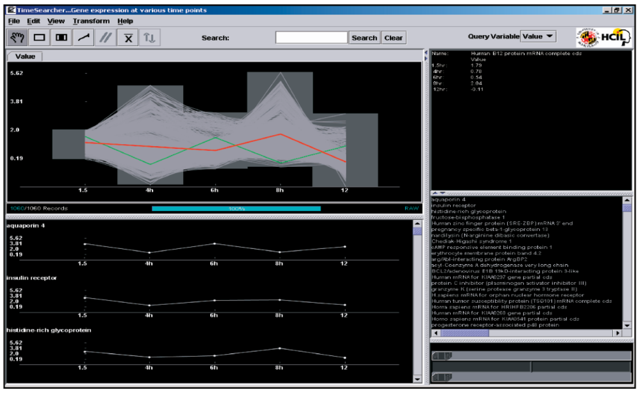
Lymphocyte antigen 94 homolog, act
Activin A receptor, type IB
Hypothetical protein similar to sm
Putative mitochondrial outer membr
Aquaporin 9

Fig. 1. Clusterview [37] visualization of the Lupus data set.

Clusterview (Fig. 1) uses a heat-map visualization for both data overview and details. A compressed heat-map provides an overview of all values in the data set, in row-column format. Users can select a part of the overview to study in more detail. It is standard practice in bioinformatics to visually encode increased gene-expression values with a red brightness scale, decreased gene-expression values with a green brightness scale, and no-change as black. As a slight variation, some tools use a continuous red-yellow-green scale with yellow in the no-change region.

TimeSearcher (Fig. 2) introduces a new concept of time-boxes [38] to query a set of entities with temporal attributes. The visualization used for data overview is a time series display of all the data attributes. Line graphs and detailed information are also provided for each individual data entity. The views are tightly coupled using the concept of interactive "brushing and linking," selecting a gene in one view highlights it in all views.

HCE (Fig. 3) provides several different visualizations: scatter plots, histograms, heat maps, and parallel-coordinates. HCE's primary display uses dendrogram visualizations to present hierarchical clustering results. This clusters similar data items near each other in the tree display. HCE also provides histograms and scatter plots for data analysis. In a multidimensional data set, the number of scatterplots possible is large. HCE introduces a new concept of "rank by



Name: Human B12 protein mRNA complete cds
Value
1.5hr: 1.79
4hr: 0.78
6hr: 0.54
8hr: 2.04
12hr: -0.11

aquaporin 4
insulin receptor
histidine-rich glycoprotein
fructose-bisphosphatase 1
Human zinc finger protein (SRE-ZBP) mRNA 3' end
pregnancy specific beta-1-glycoprotein 13
nardilysin (N-arginine dibasic convertase)
Chediak-Higashi syndrome 1
cAMP responsive element binding protein 1
erythrocyte membrane protein band 4.2
arg/Abl-interacting protein ArgBP2
acyl-Coenzyme A dehydrogenase very long chain
BCL2/adenovirus E1B 19kD-interacting protein 3-like
Human mRNA for KIAA0297 gene partial cds
protein C inhibitor (plasminogen activator inhibitor III)
granzyme K (serine protease granzyme 3 tryptase II)
H.sapiens mRNA for orphan nuclear hormone receptor
Human tumor susceptiblity protein (TSG101) mRNA complete cds
Homo sapiens mRNA for HRIHFB2206 partial cds
Human mRNA for KIAA0268 gene partial cds
Homo sapiens mRNA for KIAA0541 protein partial cds
progesterone receptor-associated p48 protein

Fig. 2. TimeSearcher [38] visualization of Time-series data set.

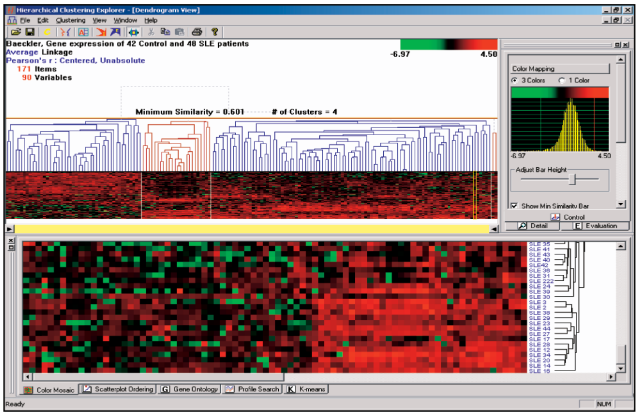


Fig. 3. HCE [39] visualization of the Lupus data set.

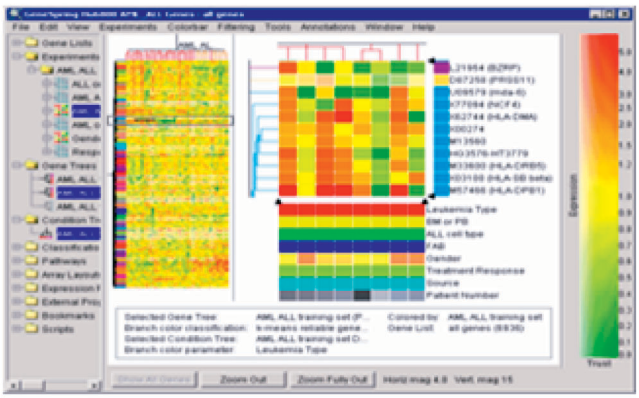


Fig. 5. GeneSpring [29] visualization.

feature" [40] to allow users to quickly find interesting histograms and scatterplots, although this feature was not available for this study. The visualizations are tightly coupled for interactive brushing. Users can manipulate various properties of the visualizations and also zoom into areas of interest.

Spotfire (Fig. 4) offers a wide range of visualizations: scatter plots, bar graphs, histograms, line charts, pie charts, parallel coordinates, heat maps, and spreadsheet views. Spotfire presents clustering results in multiple views, placing each cluster in a separate parallel coordinate view. The visualizations are linked for brushing. Selecting data items in any view shows feedback in a common detail window. Users can zoom, pan, define data ranges, and customize visualizations. The fundamental interaction technique in Spotfire is the dynamic query sliders, which interactively filter data in all views.

GeneSpring (Fig. 5) provides the largest variety of visualizations for microarray data analysis: parallel coordinates, heat-maps, scatter plots, histograms, bar charts, block views, physical position on genomes, array layouts, pathways, ontologies, spreadsheet views, and gene-to-gene comparison. As we did not have information such as position of genes on chromosome and organization of gene clones on microarray chip for all the experiments, we could not use some of the visualizations, such as physical position and array layout views, provided by GeneSpring. The visualizations are linked for brushing. Users can manipulate the visualizations in several ways, e.g., zooming, customizing visualizations by changing the color, range, etc. GeneSpring also includes data clustering capabilities.

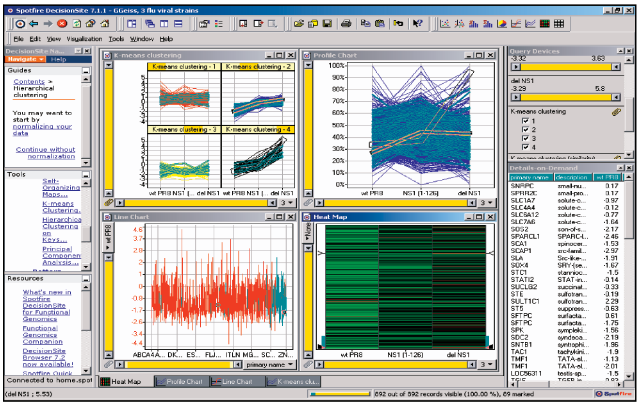


Fig. 4. Spotfire [9] visualization of the Viral data set.

4.3 Participants

Thirty test subjects volunteered from the university community. We allotted six users per tool, with two per data set per tool. We required all users to have earned at least a Bachelor's degree in a biological field and be familiar with microarray concepts. To prevent undue advantage and to measure learning time, we assigned users to a tool that they had never used before. Users were randomized within this constraint. Based on their profiles, the users fit into one of three categories summarized in Table 5.

4.4 Protocol and Measures

To evaluate the visualization tools in terms of their ability to generate insight, a new protocol and set of measures is used that combines elements of the controlled experiment and usability testing methodologies. This approach seeks to identify individual insight occurrences as well as the overall amount of learning while participants analyze data in an open-ended think-aloud format. No benchmark tasks were assigned. Also, we decided to focus on new users of the tools with only minimal tool training. We have found that success in the initial usage period of a tool is critical for tool adoption by biologists.

Each user was assigned one data set and one tool. Before starting their analysis, users were given a background description about the data set. To reduce initial learning

TABLE 5
Participant Background

Category	Participant Background	N
Domain Expert	Senior researchers with extensive experience in microarray experiments and microarray data analysis. Possess a Ph.D. in a biological field.	10
Domain Novice	Lab technicians or graduate student research assistants, having an M.S. or B.S. in a biological field. Some experience with microarray data analysis.	11
Software Developers	Professionals who implement microarray software tools. Have an M.S. in a biological field and also M.S. in computer science.	9

This table summarizes the number of participants (N) and their backgrounds.

TABLE 6
Dependent Variables

1	User's initial questions about the dataset
2	Total time spent with the tool
3	Amount learned (as a percentage), periodic and final
4	List of insights and characteristics
5	Visualization techniques used
6	Usability issues
7	Participant demographics

time, the users were given a brief 15-minute tutorial about the primary visualization and interaction techniques of the tool. Users then listed some analysis questions they would typically ask about such a data set. Then, they were instructed to continue to examine the data with the tool until they felt that they would not gain any additional insight. The entire session was videotaped for later analysis. Users were allowed to ask the administrator about using the tool if they could not understand a feature. The training in this protocol was intended to simulate how biologists often learn to use new tools from their colleagues.

While they were working, users were asked to comment on their observations, inferences, and conclusions. Approximately every 10-15 minutes, users were asked to estimate how much of the total potential insight they felt they had obtained so far about the data, on a scale of 0-100 percent. When they felt they were finished, users were asked to assess their overall experience with the tool, including any difficulties or benefits.

Later, we analyzed the videotapes to identify and codify all individual occurrences of insights. Table 6 summarizes the dependent variables.

5 Results

Results are presented in terms of the users' data questions, insights, visualization usage, and user background.

5.1 Initial Questions

At the start of each session, users were requested to formulate questions about the data that they expected the visualization tool to answer (Table 7). Almost all the users wanted to know how the gene expression changed and its statistical significance with each experimental condition, different expression patterns, and obtain pathway information and known literature for the genes of interest. More biologically specific questions focused on the location of genes of interest on chromosomes and pathways. They said that it would be valuable to know what pathways show correlations.

There were, collectively, 31 distinct questions for all the data sets. It was not possible to answer some of the questions during the experimen, due to insufficient data, e.g., the Lupus data set did not have information about disease severity or patient demographics, as would be required for questions 23 and 26 in Table 7. Nor did the data sets include pathway information for questions 4, 7, 15, 18, and 30 listed in Table 7. However, GeneSpring (31/31) and Spotfire (27/31) can potentially address most of the questions posed by the participants, if adequate data were provided. Clusterview (11/31), TimeSearcher (14/31), and HCE (15/31) can answer more specific subsets of the questions.

TABLE 7
List of Data Questions Asked by the Participants

	Information participants wanted from the data	Num.
Questions for Time series dataset		
1	Change in overall expression with time	10/10
2	Different patterns of expression	10/10
3	Genes that responded early to a treatment and were later followed by other genes	5/10
4	Functional details of genes showing high change	2/10
5	Genes showing similar expression pattern to a specific gene of interest	1/10
6	Relate change in gene expressions to physiological changes in the cells	1/10
7	Pathway information for genes having similar expression patterns	2/10
8	Relate gene expression to chromosome position	1/10
9	Retrieve known information for selected genes	10/10
Questions for Viral dataset		
10	Difference in overall expression for three viruses	10/10
11	Genes that show similar/different behavior to the experimental hypothesis	3/10
12	Expression patterns different from the hypothesis	3/10
13	Genes with high or low expression for each virus	10/10
14	Different patterns of gene expression	10/10
15	Pathway information for genes of interest	3/10
16	Correlations between different pathways	3/10
17	Chromosom location of genes with similar change	3/10
18	Functional information of selected genes	1/10
19	Statistical significance in changes between different viral strains	1/10
Questions for Lupus dataset		
20	Difference in expression between 2 groups	10/10
21	Statistical significance of difference between groups	3/10
22	Different patterns of gene expression	10/10
23	Relate expressions to severity of disease	1/10
24	The range of gene expression for each group	1/10
25	Statistical significance of variability of expression for genes in each group	4/10
26	In case of variability, if this is based on patients' age, sex, race, etc.	1/10
27	Analyses such as genes that show more than 50% increase from control to lupus patients	1/10
28	A list of housekeeping genes to evaluate experiment results	1/10
29	Patient characteristics such as those who used some drug vs. those who did not use any drug, males vs. females etc.	1/10
30	Behavior of Immune pathway genes	2/10
31	Calculate average expression for each group	6/10

5.2 Evaluation on Insight Characteristics

Listed here are the measured results for each insight characteristic described earlier, aggregated by visualization tool. Since this evaluation method is more qualitative and subjective than quantitative and the number of participants is limited, a general comparison of tendencies in the results is most appropriate (Fig. 6 and Table 8). However, we include some statistical analysis that provides useful indicators.

Insights. We counted the total number of insights, i.e., distinct observations about the data by each participant. Participants who analyzed the same data set with a particular tool reported very similar insights about the data. Thus, the reported insights were repetitive across participants. As shown in Fig. 6, the count of insights was highest for Spotfire and lowest for HCE.

Total Domain Value. The sum of the domain value of all the insight occurrences. Insight value was highest for Spotfire. Participants using Spotfire gained significantly

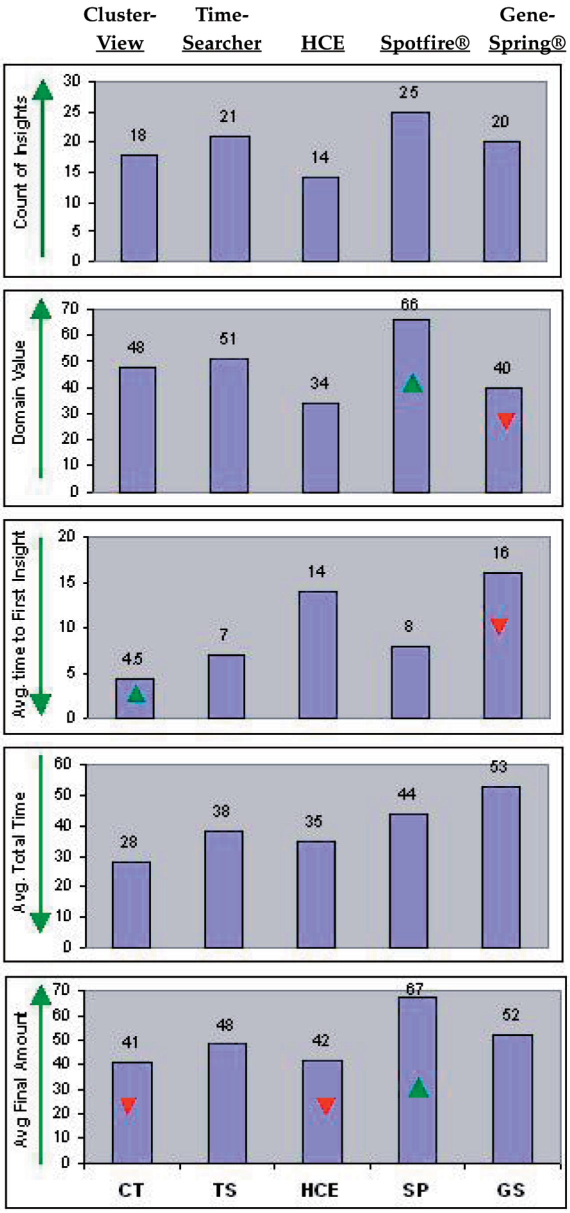


Fig. 6. Count of insights, total insight domain value, average time to first insight, average total time, and average final amount learned for each tool. △/▽ indicates significantly better/worse performance differences. Y-axis arrows indicate direction of better performance.

more insight value than with GeneSpring ($p < 0.05$). Though the numeric value was lowest for HCE, there were no significant differences between Spotfire or other tools and HCE due to high variance in the performance of HCE users, as explained in Section 5.4.

Time. The following two temporal characteristics (average time to first insight and average total time) summarize the time to acquire insights:

Average Time to First Insight. The average time into the session, in minutes, of the first insight occurrence of each participant. Lower times suggest that users are able to get immersed in the data more quickly and, thus, may indicate a faster tool learning time. The participants using Clusterview took a very short time to reach first insight. TimeSearcher and Spotfire were also fairly quick to first insight, while HCE and GeneSpring took twice as long on average. Clusterview users took significantly less time ($p < 0.01$) to reach the first insight than the other users, while GeneSpring took significantly longer ($p < 0.01$).

TABLE 8
Insight Characteristics

	Cluster-view	Time-Searcher	HCE	Spotfire®	Gene-Spring®
Hypotheses	2	1	1	3	0
Unexpected Insights	3	3	5	2	0
Incorrect Insights	0	0	2	0	0

This table summarizes the total number of hypotheses generated, unexpected insights, and incorrect insights for each tool.

Average Total Time. The average total time users spent using the tool until they felt they could gain no more insight. Lower times indicate a more efficient tool or, possibly, that users gave up on the tool due to lack of further insight. In general, Clusterview users finished quickly, while GeneSpring users took twice as long.

Average Final Amount Learned. The average of the users' final stated estimate of their amount learned. The amount learned is a percentage of the total potential insight, as perceived by users. In contrast to other insight characteristics reported, this metric gauges the users' belief about insight gained and about how much the tool is or is not enabling them to discover. Spotfire users were most confident in their perceived insight. The similarity between this metric and total domain value might indicate that the users are fairly accurate in their assessment.

Hypotheses. Only a few insights led users to new biological hypotheses (Table 8). These insights are most vital because they suggest future areas of research and result in real scientific contributions. For example, one user commented that parts of the time series data showed a regular cyclic behavior. He searched for genes that showed similar behavior at earlier time points, but could not find any. He offered several alternative explanations for this behavior related to immune system regulation and said that it would compel him to perform follow-up experiments to attempt to isolate this interesting periodicity in the data. For the viral data set, two users commented that there were two patterns of gene expression that showed negative correlation. They inquired whether this means that the transcription factors of these genes have inhibitory or stimulatory effects on each other. They said that they wanted more information about the functions and pathways to which these genes belong to better relate the data to biological meaning. Spotfire resulted in one hypothesis for each data set, thus a total of three. Clusterview also led users to a hypothesis for the Viral and Lupus data sets.

Directed versus Unexpected Insights. The participants using HCE with the Viral data set noticed several facts about the data that were completely unrelated to their initial list of questions. Clusterview provided a few unexpected insights from the Lupus data set and TimeSearcher provided unexpected insights about the time series data. Spotfire had one each for time series and Lupus.

Incorrect Insights (Correctness). HCE proved helpful to users working with the viral data set. However, users working with the time series or Lupus data sets did not gain

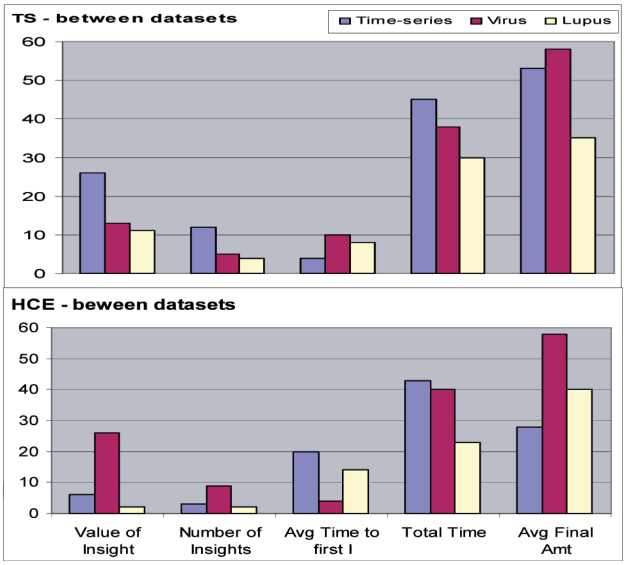


Fig. 7. Timesearcher and HCE specialize in the time-series and viral data, respectively.

much insight from the data. When prompted to report their data findings, they stated some observations about the data that were incorrect. The two users that reported incorrect insights were in the domain expert and software developer categories. The errors may have been due to inferring the color scale backward or due to misinterpreting the way that HCE reorders the rows and columns of the heat map by hierarchical clustering. None of the other tools resulted in incorrect findings.

Breadth versus Depth. Though we had initially thought this to be an interesting criterion, on data analysis we found that most user comments were of the type "breadth." For this experiment, all the users worked with a visualization tool they were not familiar with. It will be difficult for first time users to learn all the features of both Spotfire and GeneSpring within the time span of the experiment. Also, many users were not familiar with the specific genes in the data sets used for the study. We discovered that to get deeper insights into the data, the participants need to be more familiar with the data background. Hence, for the purpose of this study, we did not pursue this characteristic in detail.

Together, higher total value and count indicate a more effective tool for providing useful insight. Lower time to first insight indicates a faster learning curve for a tool. Ideally, a visualization tool should provide the maximum amount of information in shortest possible time.

Overall, Spotfire resulted in the best general performance, with higher insight levels and rapid insight pace. Clusterview and TimeSearcher appear to specialize in rapid insight generation, but to a limit. With GeneSpring, users could infer the overall behavior of the data and the patterns of gene expressions. However, because the users found the tool complicated to use, most of them were overly consumed with learning the tool rather than analyzing the data. They had difficulty getting beyond simple insights. HCE's strengths will become clear in the next two sections.

5.3 Insight per Data Set

This section compares the tools within each data set.

TABLE 9
Insight Categories

Category	Cluster-view	Time-Searcher	HCE	Spotfire®	Gene-Spring®
Overview	9	10	6	13	5
Patterns	10	8	5	10	8
Groups	0	0	0	1	4
Details	2	3	1	1	1

Time series data. In general, Spotfire and TimeSearcher performed the best of the five tools in this data set. Participants using Spotfire and TimeSearcher felt they learned significantly more ($p < 0.05$) from time series data than the other tools. Participants using Spotfire felt they learned more from the data (73 percent) compared to TimeSearcher (53 percent). Both Spotfire and TimeSearcher had nearly equivalent performance in terms of value and number of insights. Time to first insight was slightly lower for TimeSearcher (4 min) as compared to Spotfire (6 min). At the bottom, participants using HCE took significantly longer ($p < 0.01$) to reach the first insight than the other tools. Participants using GeneSpring took significantly longer ($p < 0.05$) than TimeSearcher and Clusterview.

Virus data. HCE proved to be the best tool for this data set. Participants using HCE had better performance in terms of insight value as compared to other users. However, there were no significant differences between the other users. HCE provided five unexpected insights that were different than the initial information users were searching for in this data set.

Lupus data. Participants using Clusterview and Spotfire had more insight value as compared to the other tools ($p < 0.05$) in this data.

5.4 Tools versus Data Sets

This section examines individual tools across the three data sets. TimeSearcher and HCE had interesting differences among the data sets (Fig. 7), while the other tools were well rounded.

TimeSearcher. Participants using TimeSearcher performed comparatively best with the time series data as compared to the other two data sets. With time series data, they had over double the value and number of insights than with the Viral and Lupus data sets.

HCE. In contrast, participants using HCE did best on the Viral data set. On the Viral data set, they had a significantly better performance advantage on insight value ($p < 0.01$), number of insights ($p < 0.05$), and time to first insight ($p < 0.05$) as compared to the other data sets. They also felt they learned much more from the data. Participants using the Lupus data spent significantly less overall time with the tool ($p < 0.05$) as they felt they could not learn much from the data using HCE.

5.5 Insight Categories

Though a wide variety of insights were made, most could be categorized into a few basic groups. Table 9 summarizes the number of each type of insight by tool.

Overall Gene Expression. These described and compared overall expression distributions for a particular experimental condition. For example, a user analyzing time

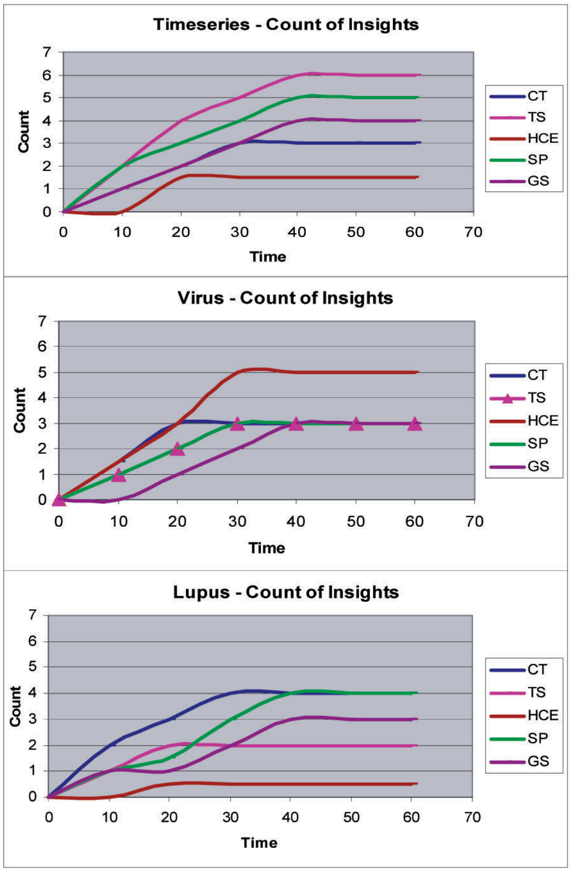


Fig. 8. Average number of insights, over time, for each data set and tool.

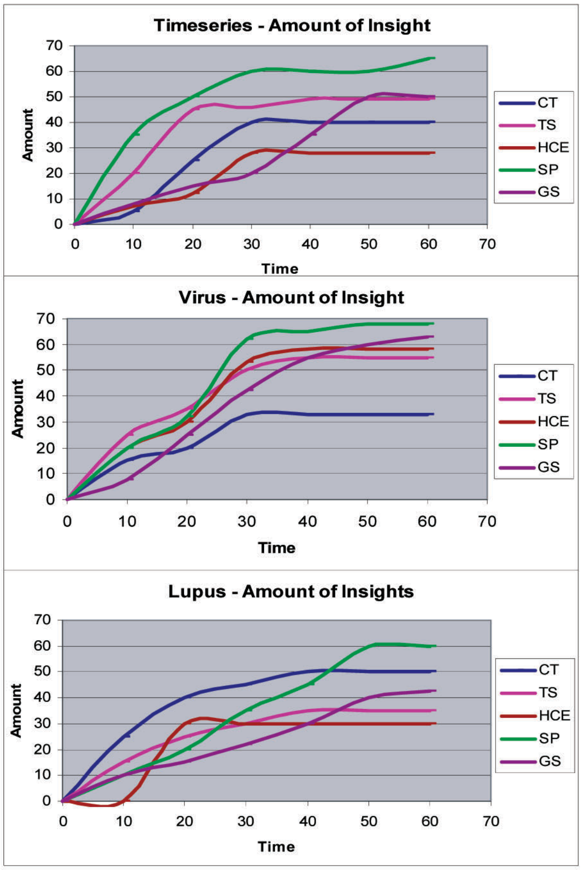


Fig. 9. Average percentage of total insight gained as periodically estimated by participants, over time, for each data set and tool.

series data reported that "at time points 4 and 8 a lot of genes are up regulated, but at time point 6 a lot are down regulated." Several users analyzing the virus data set commented that more genes showed a higher expression level for delNS1 virus as compared to wt virus and the gene expression seems to be increasing with the deletion. Most users working with the Lupus data set reported that gene expression for SLE patients appeared higher than the control group.

Expression Patterns. Most users considered the ability to search for patterns of gene expressions very valuable. Most started by using different clustering algorithms (e.g., K-Means, SOMS, Hierarchical Clustering) provided by the tools to extract the primary patterns of expression. They compared genes showing different patterns. For example, some users noted that, while most genes showed higher expression value for the Lupus group as compared to the Control group, there were other genes that were less expressed for the Lupus group. They thought it would be interesting to obtain more information about these genes in terms of their functions and the pathways they belong to.

Grouping. Some users, mainly those working with Spotfire and GeneSpring, grouped genes based on some criteria. For example, a user working with Spotfire wanted to know all genes expressed similarly to the gene HSP70. Users working with GeneSpring used gene ontology categories to group genes. GeneSpring provides several ways in which users can group their data. They found this functionality very helpful. Also, most of the users were very pleased to learn that they could link the biological information, such as gene functions, with the groups.

Detail Information. A few users wanted detailed information about particular genes that were familiar to them. For time series data, a user noticed about 5 percent of genes high at 1.5 hr were also high at 12 hr and followed a regular cycle. He looked up the annotations for a few of these genes and tried to obtain more information about them to see if they could be responsible for the cyclic nature of the data.

5.6 Insight Curves

This approach to measuring insight also enables the examination of how insight accumulates over time. This section shows the insight curves for actual insight counts as well as users' perceived insight amount. These graphs show the rate of insight generation for the tools.

Fig. 8 represents the average accumulation of insight occurrences over time for each tool and data set. Fig. 9 shows the users' average estimated percentage of total insight acquired over time. During the course of the experiment, users were asked every 10-15 minutes to report how much they felt they had learned about the data as a percentage of total potential insight.

Some of the tools stand out on certain data sets as providing a faster or slower rate of insight and strengthen findings reported earlier. TimeSearcher and Clusterview provide an early jump in insight on the time series and Lupus data sets, respectively. While Spotfire eventually catches up, other tools plateau sooner. HCE rises above other tools in the viral data set in actual insight count. However, in the other data sets, HCE shows a step-like curve, perhaps indicating an initial period of learning the tool, followed by a small number of insights, followed by a plateau and termination by the users.

There is some similarity between Figs. 8 and 9 for the time series and lupus data sets, in terms of the general shape of curves and order of the tools. This could indicate some relative accuracy of participants' insight estimates. An interesting difference is that, for Spotfire and GeneSpring, the users' estimated insight curves continue to rise even after their corresponding curves in actual insight counts plateau. That is, even after they make no new insights, they still felt they were gaining more insight. This may be due to the fact that, after continuing to explore the data in the many different visual representations within these tools, participants became more confident in their findings and felt that they had not missed much after all.

5.7 Visual Representations and Interaction

Spotfire users preferred the heat-map visual representation, whereas GeneSpring users preferred the parallel coordinate view. This is despite the fact that both of these tools offer both representations. Most of these users performed the same analyses, but using different views.

Though there were no particular preferences of visualizations for particular data sets, we noticed that, for the Lupus data set, Spotfire and Clusterview users preferred the heat-map visualization. The heat-map allowed them to group Control and Lupus data neatly into two distinct groups and they could easily infer patterns within and across both groups. Participants using these tools showed a higher performance on these data sets using these visualizations. This finding is strengthened by the fact that both TimeSearcher and GeneSpring users showed average performance on this data set. Users of these tools used parallel coordinate visualizations to analyze the data sets.

We noticed that, even though tools like Spotfire and GeneSpring provide a wide range of visualizations to users, only a few of these were used significantly during the study. Most users preferred visualizations showing outputs of clustering algorithms, such as provided by Clusterview, Spotfire, and GeneSpring. These enabled the users to easily see different patterns in the data. However, many said that it would be more helpful to them if the interaction capabilities of this representation were increased, e.g., to better enable comparison of the groups, subdividing, etc.

HCE's primary overview presents the data in a dendogram heat-map that is reordered based on the results of hierarchical clustering algorithms. Columns and samples with the most similar expression values are placed near each other. Thus, for both the time series and Lupus data sets, where a particular column arrangement is useful to recognize changes across the experimental conditions, HCE showed poorer performance. Users were not aware of the fact that they could turn off that feature (such customization capabilities of views were not demonstrated in the initial short training session). Also, none of the four users who would have benefited the most from turning off this feature considered the possibility of turning it off and they did not inquire about it. This turned out to be a critical feature that should be made more prominent in the tool or, in hindsight, should be included in the training.

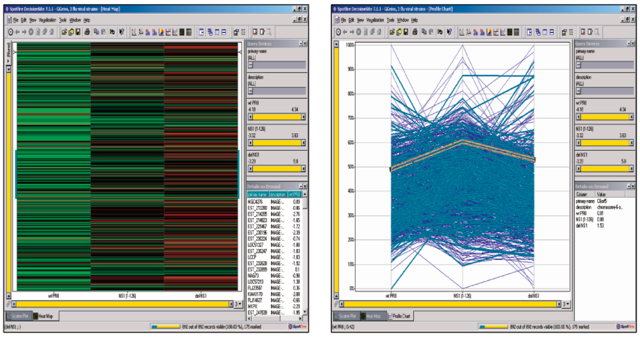


Fig. 10. The heat map and parallel coordinate views in Spotfire.

5.8 Participant Comments on Visualization Tools

At the end of each experiment, users were requested to comment on their experience with the tool they used. The following sections summarize the users' comments.

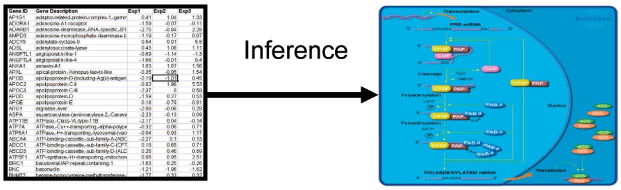
Clusterview. Users felt that the tool was extremely simple to use. Some users (3/6) required a brief explanation of the heat-map view of the data. The users felt that the information provided by Clusterview is very basic and they will need to perform additional analysis with other methods to get further information from the data. The users who worked with time series data commented that the heat map was not a very efficient way to represent data and they preferred visualizations similar to parallel-coordinates.

TimeSearcher. Feedback on TimeSearcher varied for different data sets. The users found the parallel-coordinate overview provided by TimeSearcher was easy to understand. Users working with the time series data found the tool very helpful. They were able to easily identify trends and patterns in the data. Users working with Lupus data set said that it was very difficult for them to see all of the 90 data points clearly. Some participants found a few features of TimeSearcher, such as "Angular Queries" and "Variable Time-Boxes," difficult to interpret. As TimeSearcher does not provide any clustering capabilities, users have to manually search for every pattern in the data using "time boxes," which can prove tedious in a large data set.

HCE. Most users were impressed with HCE. The tool provides a wide variety of features for data analysis. HCE was more helpful to participants working with the viral data set. Users working with the Lupus data set gave up data analysis within 20 minutes, complaining that it was very difficult for them to analyze data using HCE.

Spotfire. Users working with Spotfire were impressed with it. They did not require any special assistance to understand the tool. They said that most of the visualizations were easy to understand. Most users preferred the heat-map visualization of the Spotfire over its parallel coordinate or Profile chart display (Fig. 10). Though the users found the visualization displaying different clusters in the data helpful, they said that it should be easier to interact with. They found it annoying that they could not select and focus on a particular cluster of interest.

GeneSpring. Users felt that they would have to spend a long time learning GeneSpring. A few users (2/6), spent an initial 45 minutes just trying to get familiar with GeneSpring, after which they gave up the data analysis, saying that it would take them too long to comprehend what the tool does. A few users commented that it would be great to have some sort of automation that would show them which visualization to begin the data analysis and how to change the visualization properties. One user said that the basic things should be easy and visualizing an already normalized data set should not be so difficult. None of the users could change different properties of visualization, such as color, scale, or amount of data to be visualized, without help. Users were pleased to know that GeneSpring provided features to make lists of genes based on different criteria. The users commented that such features could prove to be very helpful. Also, features that allow users to add pathway information to gene lists were considered very useful.



Gene ID	Gene Description	Exp1	Exp2	Exp3
AP1G1	adaptor-related-protein-complex-1,-gamm	0.41	1.04	1.33
ADORA1	adenosine-A1-receptor	-1.59	-0.07	-0.11
ADARB1	adenosine-deaminase,-RNA-specific,-B1-	-2.75	-0.84	2.28
AMPD3	adenosine-monophosphate-deaminase-(i	-1.19	-0.17	0.07
ADCY9	adenylate-cyclase-9	0.54	0.91	5.8
ADSL	adenylosuccinate-lyase	0.43	1.08	1.11
ANGPTL1	angiopoietin-like-1	-0.69	-1.14	-1.3
ANGPTL4	angiopoietin-like-4	-1.66	-0.01	0.4
ANXA1	annexin-A1	1.03	1.87	1.56
APXL	apical-protein,-Xenopus-laevis-like	-0.35	-0.95	1.54
APOB	apolipoprotein-B-(including-Ag(x)-antigen	-2.19	-1.01	0.45
APOC2	apolipoprotein-C-II	-0.82	1.96	5.53
APOC3	apolipoprotein-C-III	-2.37	0	0.59
APOD	apolipoprotein-D	-1.59	0.27	0.53
APOE	apolipoprotein-E	0.18	-0.79	-0.81
ARG1	arginase,-liver	-2.08	-0.06	0.28
ASPA	aspartoacylase-(aminoacylase-2,-Canava	-2.23	-0.13	0.09
ATP11B	ATPase,-Class-VI,-type-11B	-2.17	0.04	-0.14
ATP7A	ATPase,-Cu++-transporting,-alpha-polype	-0.32	0.06	0.71
ATP6A1	ATPase,-H+-transporting,-lysosomal-(vac	-0.04	0.93	1.17
ABCA4	ATP-binding-cassette,-sub-family-A-(ABC	-2.27	0.1	0.13
ABCC1	ATP-binding-cassette,-sub-family-C-(CFT	0.18	0.65	0.71
ABCD3	ATP-binding-cassette,-sub-family-D-(ALD	0.26	0.46	0.89
ATP5F1	ATP-synthase,-H+-transporting,-mitochon	0.06	0.95	2.51
BIRC1	baculoviral-IAP-repeat-containing-1	-1.63	0.25	-0.26
BNC	basonuclin	-1.21	-1.96	-1.62
BHMT2	betaine-homocysteine-methyltransferase-	-1.77	0.33	0.32

Fig. 11. Visualizations must support domain-relevant inference, from microarray data set to pathway models describing interactions within a cell [41].

5.9 Participants' Backgrounds

One might conjecture that users with more domain experience or software development experience would gain more insight from the data visualizations. Yet, we found that the insight domain value and total number of insights did not appear to depend on participant background. Averages were similar and no significant difference between user categories was detected. Due to the limited number of subjects, full factorial analysis within tool or data set groups is not feasible. Trends within user categories followed the same general trends for tools and data sets identified previously. We did not find any differences in the number of insights, value of insight, and hypothesis generation based on the participants' background. Rather, we found that these factors were more dependent on the user motivation.

Software developers on average felt that they learned less from the data as compared to others, whereas domain novices felt they learned more from the data. Novices also spent comparatively more time in the study as compared to others. A noticeable difference was in the users' behavior during the experiment. Novice users needed more prompting to make comments about the data sets. They were less confident to report their findings. Software developers almost always made the first insight faster than the novice users.

6 Discussion of Results

Commercial versus Free. Both Spotfire and Clusterview users resulted in equivalent insight from the Lupus data set. However, participants using Spotfire felt they learned much more from the data as compared to Clusterview. Analyzing data in multiple visual representations gave Spotfire users more confidence that they did not miss any information, whereas Clusterview users were more skeptical about their progress, believing that they must be missing something. A simple visualization tool used on an appropriate data set can have performance comparable to more comprehensive software containing many different visualizations and features.

Free research software like TimeSearcher and HCE tends to address a smaller set of closely related tasks. Hence, they provide excellent insight on certain data sets. Also, since they are focused on specific tasks, they have simpler user interfaces that emphasize a certain interaction model. This reduces the learning time and enables users to generate insights quickly. Spotfire, despite having a large feature set, has a learning time almost equivalent to the simple tools, which is commendable. This is likely due to Spotfire's unified interaction model. The brushing and dynamic query concepts were quickly learned by users, and resulted in early rapid insight generation.

Domain Relevance. A serious shortcoming of the tools is that they do not adequately link the data to biological meaning. The fact that domain experts performed on par with domain novices and the small numbers of hypotheses generated indicates that the tools did not leverage the domain expertise well. Before we conducted the study, we believed that users with more expertise in biology would gain more from visualizations than a novice. We were also curious about whether software development experience would lead to better usage of the tools. However, these background differences did not reveal themselves in the actual insights generated. The difference was only in the users' believed insight, in which novices were overconfident and developers were skeptical.

If the tools could provide a more information-rich environment, such as linking data directly to public gene databases or literature sources, expert biologists could better exploit their domain knowledge to construct higher level, biologically relevant hypotheses. In this experiment, the tools helped users identify patterns in the data, but did not enable them to connect these numerical patterns to the underlying biological phenomena. A critical need is for highly integrated visualization environments that excel at domain relevance and inference. In this case, understanding gene expression patterns must lead to inference of underlying pathways that model the interactions of the genes (Fig. 11). Visualization must support this level of inference.

Interaction Design. The design of interaction mechanisms in visualization is critically important. Usability of interactions can outweigh the choice of visual representation. Spotfire users mainly focused on the heat-map representation, while GeneSpring users focused on the parallel coordinates, even though both tools support both representations. The primary reason for this, based on comments from users, was that users preferred parallel coordinates, but Spotfire's parallel coordinates view employs a poorly designed selection mechanism. Selecting lines in its parallel coordinates view results in unusual and occluding visual highlight feedback that made it very difficult for users to determine which genes were selected and what other genes were nearby (Fig. 10).

The ability to select and group genes was the most common interaction that users performed. The grouping of genes into semantic groups is a fundamental need in bioinformatics visualization. GeneSpring provided useful grouping features that enabled more insights in the "groups" category. More tools need better support for grouping items, based on interactive selections as well as computational clustering, and managing groups.

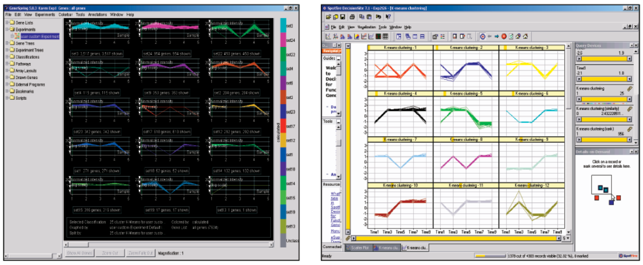


Fig. 12. Clustered views were the most commonly used in the study. GeneSpring (left) and Spotfire (right).

GeneSpring is the most feature-rich tool of the five and, therefore, perhaps the most difficult to learn. However, even though users tended to focus on a small number of basic visualization features, usability issues (such as the large quantity of clicks required to accomplish tasks) reduced their overall insight performance.

Clustering. Certain visualizations, such as the clustering visualizations for both Spotfire and GeneSpring, were the most popular in the study. Users commented that it would be very helpful if the interaction techniques for these clustered views were improved so that they were better integrated into the overall interaction model.

Clustering (Fig. 12) was a very useful feature throughout, but care should be taken to provide nonclustered overviews first. As in HCE, clustering can potentially bias users into a particular line of thought too quickly. In comparing Spotfire and Clusterview, users were also more confident when they could confirm their findings between clustered and nonclustered views of Spotfire.

User Motivation. We noticed that an important factor in gaining insight is user motivation. Clearly, participants in our study did not analyze the data with as much care as they would if the data were from their own experiments. They mainly focused on discovering the overall effects in the data, but were not sufficiently motivated to extreme details. Most of the insights generated were classified as breadth rather than depth. However, the visualizations were able to provide a sizeable number of breadth insights in spite of low motivation levels.

7 Discussion of Methodology

The main purpose of visualization is to provide insight. This can be difficult to measure. Although our definition of insight is not comprehensive, it does provide an approximation of users' learning. This, in turn, enabled us, as evaluators, to gain insight into the effectiveness of these visualization tools. The definition of insight and the methodology presented are domain independent and can be applied for similar data analysis scenarios in other domains. The technique evaluates users' findings from the data. More, valuable, faster, and deeper data findings correspond to more effective visualizations as it suggests users can gain more insight from the data.

The methodology succeeded in measuring open-ended insight generation by not restricting users to a set of preplanned benchmark tasks. This approach closely matches the purpose of visualization—to discover unforeseen insights, rather than to perform routine tasks. This provided a good analysis of the insight capabilities of these visualization tools. However, this method does not replace the need for controlled experimentation on benchmark tasks, which is still useful for detailed testing of specific targeted tasks.

This new approach has shown promise, but some difficulties remain to be overcome:

- Labor intensive. It is time consuming for the experimenters to capture and code insights. Self-reporting by subjects could be a solution.
- Requires domain expert. The available population of capable experts in the bioinformatics domain for coding the value of insights is not large. This coder must also be removed from the subject pool.
- Requires motivated subjects. Since benchmark tasks are not given, subjects must self motivate to accomplish anything.
- Training and trial time. Longer time periods would better reflect realistic visualization usage.

Three major limitations of this study need to be overcome. First, the study reported here measures insight from short term usage, typically less than 2 hours per user. In real-world scenarios, biologists spend days, weeks, and even months analyzing data. Long term insight may be very different than short term insight. Long term insight can provide broader understanding that guides biologists through multiple cycles of microarray experiments. Second, the participants in the study were unfamiliar with the data and not personally invested in its creation. The only background knowledge they had was what we provided during the course of study. It was very difficult to appreciate the biological relevance of the microarray data they were analyzing. Hence, the hypotheses they reported were more speculative. Yet, the insights were not trivial, which suggests that the visualizations are provoking users to think deeply about the data and to apply the insight in their domain. Third, each participant was unfamiliar with the visualization tool that they used. Gaining expertise with a visualization tool may change the method in which it is used and the insight it provides.

We now recognize that it would be very valuable to conduct a longitudinal study that records each and every finding of the users over a longer period of time to see how the visualization tools influence and adapt to their knowledge acquisition. These studies should be conducted with researchers analyzing their own experimental results for the first time and preferably through multiple experimental cycles. This could be done using long-term ethnographic methods or subjects' self-reporting. Gonzales and Kobsa [25] and Rieman [26] present such longitudinal studies that included frequent user interviews, diary studies, and "Eureka" reports. Such studies can help to identify the broader information needs of biologists and to develop more meaningful tools that leverage their domain knowledge and expertise.

8 Conclusions

This study suggests the following major conclusions for biologists, visualization designers, and evaluators.

Biologists. A visualization tool clearly influences the interpretation of the data and insight gained. Hence, it is imperative that the appropriate tool be chosen for a given data set. We sought to answer the question of which is the best tool to use. Some tools work more effectively with certain types of data. Clusterview, TimeSearcher, and HCE performed better with the Lupus, time series, and viral data sets, respectively. For other data, they provided below average results. Thus, the data set dictates which tool is best to use. Additionally, larger software packages, like Spotfire and GeneSpring, work consistently across different data sets. If a researcher needs to work with multiple kinds of data, those packages would be better. But, if a researcher needs to work with just one kind of data, more focused tools can provide better results in a much faster time frame. Spotfire proved to be an excellent tool all around for rapid insight generation.

Visualization Designers. Interaction techniques play a key role in determining visualization effectiveness. Designers should emphasize consistent usable interaction design models with clear visual feedback. Grouping and clustering is a must. Multiple representations can help provide user confidence. It would be helpful to identify which visualization technique in a given software package is used the most by users and improve it. It is imperative that users be able to access and link biological information to their data. Visualizations should strive to support higher-level domain relevant inference.

Evaluators. The main purpose of visualization is to provide insight. This can be difficult to measure with controlled experiments on benchmark tasks or other methods. Our insight definition allowed us to quantify insight generation using a variety of insight characteristics, which enabled us to gauge the open-ended insight capability of bioinformatics visualization tools. Simultaneously, the use of the think-aloud protocol provides deeper qualitative explanations for quantitative results. Further work to overcome difficulties (such as labor intensiveness) and limitations (such as user motivation) would produce a more practical and effective evaluation method. This methodology can prove helpful in future studies for analyzing the effectiveness of visualizations in many domains.

Acknowledgments

The authors would like to thank all the experiment participants, without whom the study would not have been possible. They also thank Hideo Goto, Yoshi Kawaoka, and John Yin for assistance related to mRNA extraction, as well as Jeff Furlong and Emily Baechler for assistance with data sets. They also extend their thanks to Lee LeCuyer and Edward Fulton for their help with the study.

References

[1] 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.

[2] L. Shi, "DNA Microarray—Genome Chip," http://www.gene-chips.com/GeneChips. html#What, 2002.

[3] D. Bassett, M. Eisen, and M. Boguski, "Gene Expression Informatics—Its All in Your Mine," *Nature Genetics Supplement,* vol. 21, Jan. 1999.

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

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

[6] A. Robinson, "Bioinformatics Visualization," http://industry.ebi.ac.uk/alan/, 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] SPOTFIRE® Decision site for functional Genomics, http://www.Spotfire.com, 2005.

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

[11] 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.

[12] P. Saraiya, C. North, and K. Duca, "An Evaluation of Microarray Visualization Tools for Biological Insight," *Proc. IEEE InfoVis,* 2004.

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

[14] 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.

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

[16] 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.

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

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

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

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

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

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

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

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

[25] 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.

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

[27] J.-D. Fekete and C. Plaisant Infovis 2003 Contest, http://www.cs.umd.edu/hcil/iv03contest/, 2003.

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

[29] GENESPRING®, Cutting-Edge Tools for Expression Analysis, www.silicongenetics.com, 2005.

[30] L. Heath and N. Ramakrishnan, "The Emerging Landscape of Bioinformatics Software Systems," *Computer,* vol. 35, no. 7, pp. 41-45, July 2002.

[31] G. Churchill, "Fundamentals of Experimental Design for cDNA Microarrays," *Nature Genetics,* vol. 32, pp. 490-495, 2002.

[32] J. Quackenbush, "Microarray Data Normalization and Transformation," *Nature Genetics,* vol. 32, pp. 496-501, 2002.

[33] K.A. Duca, H. Goto, Y. Kawaoka, and J. Yin, "Time-Resolved mRNA Profiling during Influenza Infection: Extracting Information from a Challenging Experimental System," Am. Soc. Virology, 20th Ann. Meeting, http://infovis.cs.vt.edu/cs5764/fall2003/ideas/influenza.doc, 2001.

[34] G. Geiss, M. Salvatore, T. Tumpey, V. Carter, X. Wang, C. Basler, J. Taubenberger, R. Bumbarner, P. Palese, M. Katze, and A. Garcia-Sastre, "Cellular Transcriptional Profiling in Influenza A Virus-Infected Lung Epithelial Cells: The Role of the Nonstructural NS1 Protein in the Evasion of the Host Innate Defense and Its Potential Contribution to Pandemic Influenza," *PNAS,* vol. 99, no. 16, pp. 10736-10741, 2002.

[35] E. Baechler, F Batliwala, G. Karvpis, P. Gaffney, W. Ortmann, K. Espe, K. Shark, W. Grande, K. Hughes, K. Kapur, P. Gregersen, and T. Behrens, "Interferon-Inducible Gene Expression Signature in Peripheral Blood Cells of Patients with Severe SLE," *Proc. Nat'l Academy of Science,* vol. 100, no. 5, pp. 2610-2615, 2003.

[36] Affymetrix Technologies, http://www.affymetrix.com/, 2005.

[37] M. Eisen, P Shellman, P Brown, and D Bostein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Science,* vol. 95, no. 25, pp. 14963-14968, 1998.

[38] H. Hochheiser, E.H. Baehrecke, S.M. Mount, and B. Shneiderman, "Dynamic Querying for Pattern Identification in Microarray and Genomic Data," *Proc. IEEE Int'l Conf. Multimedia and Expo,* 2003.

[39] J. Seo and B. Shneiderman, "Interactively Exploring Hierarchical Clustering Results," *Computer,* vol. 35, pp. 80-86, 2002.

[40] J. Seo and B. Shneiderman, "A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections," *Proc. IEEE InfoVis,* 2004.

[41] Biocarta™, Charting Pathways of Life, http://www.biocarta.com/genes/index.asp, 2005.

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

Chris North received the PhD degree from the University of Maryland, College Park. He is an assistant 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 in the HCI aspects of data visualization, including interaction techniques for large high-resolution displays, evaluation methods, and multiple-view strategies. In applied work, he collaborates with faculty in bioinformatics, network security, and construction engineering.

Karen 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.