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 datasets. 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. To evaluate and rank bioinformatics visualizations based on real world data analysis scenarios, we need a more relevant evaluation method that focuses on data insight. This paper presents several characteristics of insight that enable us to recognize and quantify it. Ideally, users want maximum insight from the data in the least possible time. Based on this, we evaluate five popular microarray visualization tools on the amount and types of insight they provide and the time it takes to reach the insight. Though we use this technique to analyze bioinformatics visualizations, it can be applied in other domains.

Index Terms— H.5.2 User Interfaces - Evaluation/methodology, Graphical user interfaces (GUI), I.6.9 Visualization - Information visualization, Visualization systems and software, Visualization techniques and methodologies

1 Introduction

Biolgists use high-throughput experiments such as microarray to answer complex biological research questions. Experiments, such as gene-expression microarrays [1], [2] result in datasets that are very large. The datasets usually contain information about several thousand genes. Scientists use these datasets to infer complex interactions between genes and proteins. But due to its magnitude, microarray data is prohibitively difficult to analyze without the help of computational methods.

The advent of high-throughput experiments is causing a shift in the way biologists do research, a shift away from simple reductionist testing on a few variables towards systems-level exploratory analysis of 1000s of variables simultaneously [3]. Hence, they use various data visualizations to derive biologically 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. First, with a wide variety of available tools, there is significant confusion among the biologists about which tool to use. Second, because of the open-ended and exploratory nature of the tasks, it is unclear how and if these tools meet their needs in providing insight.

The ultimate goal of the research reported in this paper is to evaluate some of the popular microarray data visualization tools, such as Spotfire® [7]. The key research questions are: How successful are these tools in assisting the biologists in arriving at domain-relevant insights? How do various visualization techniques affect users' perception of data? How does user's background affect the tool usage?

The immediate goal is to devise an evaluation methodology that better reflects the needs of the bioinformatics data analysis scenario. Typically, visualization evaluations have focused on controlled measurements of user performance and accuracy on predetermined tasks [8], [9]. However, to answer these research questions requires an evaluation method that more closely matches the exploratory nature of the biologists' goals. The main consideration for any 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. Ultimately the real 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. Thus, our primary focus is on ***insight***. We devise and deploy a novel insight-based approach to visualization evaluation that can be generally applied in other data visualization domains.

2 Related Work

A large number of studies have been conducted to measure effectiveness of visualizations using different evaluation methods.

Controlled experiments: Many studies have evaluated visualizations through rigorous controlled experiments [8], [9]. 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. E.g., [10] compares three different visualization systems on different tasks in terms of solution time and accuracy.

Usability testing: Usability tests typically evaluate visualizations to identify and solve user interface problems. Methods involve observing participants as they perform designated tasks using a 'think aloud' protocol, noting the usability incidents that may suggest incorrect use of the interface, and comparing results against a predefined usability specification [11]. Refer to [12] for an example of a professional usability study of a visualization.

Metrics, Heuristics, and Models: Different from empirical evaluations are 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 pre-attentive 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 and thereby examine the low-level effects of various visualization techniques.

Longitudinal and Field Studies: A longitudinal study of information visualization adoption by data analysts is presented in [21]. Their work suggests advantages when visualizations are used as complementary products rather than stand alone products. [22] examines users' long-term exploratory learning of new user interfaces, with 'eureka reports' to record learning events.

Thus, a range of evaluation methods has been used to measure effectiveness of visualizations [23]. In the literature, controlled experiments are the most prevalent for identifying and validating more effective visualizations. Unfortunately, these studies evaluate visualizations based only on a set of predefined tasks. Test subjects are instructed to use the visualizations to find answers to specific questions that are given by the test administrators. While this approach is useful, it is too narrow to evaluate the benefits of open-ended discovery as needed by biologists.

A primary purpose of visualization is to generate *insight* [24]. A measure of an effective visualization is its ability to generate unpredictable new insights that might not be the result of a preplanned benchmark task. Visualization can enable biologists to not only find answers but to also find questions, to identify new hypotheses. To evaluate this capability, visualizations could be measured in terms of insight generation. Hence, we developed an evaluation protocol that focuses on recognition and quantification of insights gained from actual exploratory use of visualizations [38]. This paper presents a detailed explanation and discussion of the methodology, as well as detailed results of applying the method to bioinformatics visualizations.

3 Pilot Study

The main challenge we faced in designing the experiment 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 meaningfully reproducible. To do this we undertook an initial pilot study to observe how users' recognized and categorized information obtained from microarray data using visualization tools. We used both GeneSpring® [25] and Spotfire® [7] 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 dataset. Pilot participants were supplied two datasets to work with, a table containing fake data that contained information about just ten genes, and the Lupus Dataset used in the final experiment (Section 4.1). We selected the smaller dataset for training as we believed this would facilitate users becoming familiar with the visualization techniques faster. 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 dataset, despite us telling them 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 just the real data for final experiments. From the users' comments we recognized the following 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 participants' visual data analysis process. This provided more detailed information about the insight capabilities of the tools than subjective measures from post-experiment surveys.

We define an *insight* as an individual observation about the data by the participant, a unit of discovery. These can be recognized in a think-aloud protocol. 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:

Fact: The actual finding about the data. We counted only distinct facts for 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 [26].

Breadth vs. 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.

Directed vs. Unexpected: Directed insights are those that answer a specific question that the user was searching for. 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 dataset 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.

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 3x5 between-subjects design examines these two independent variables:

1. Microarray dataset, 3 treatments
2. Microarray visualization tool, 5 treatments

4.1 Microarray Datasets

To examine a range of data scenarios, we used data from three common types of microarray experiments. The datasets are all quantitative, multi-dimensional 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 datasets were preprocessed, normalized, pre-filtered, and converted to the required formats (as discussed in [27] and [28]) 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 datasets were used.

1) Time-series dataset: Users were given an unpublished dataset from Karen Duca's lab [29]. 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 were the $\log_2$ of the normalized ratios of experimental signal to control signal. The dataset used for analysis had 1060 rows (genes) over 5 time points. Two additional columns represent the gene name and standard ID.

Table 1: Time-series dataset 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
…	…	…	…	…	…	…

2) Viral dataset: Part of a published dataset from Michael Katze's lab [30] 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 post infection, mRNA was extracted from infected and mock-infected cultures. The dataset used for analysis had 3 columns (representing the 3 viral conditions) and 861 rows (genes). Two additional columns represent the gene name and standard ID.

Table 2: Viral dataset used in the experiment

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

3) Lupus dataset: Participants were presented a subset of published data from Timothy Behren's lab [31]. In this study, after blood draw, peripheral blood mononuclear cells (PBMCs), comprising monocytes/macrophages, B and T lymphocytes, and NK cells, were isolated from control and Systemic Lupus Erythematosus (SLE) samples. mRNA was harvested for expression profiling using Affymetrix technology [32]. The column values represented expression values (average difference or AD) for each gene. Scaling was performed to allow comparison between chips. The dataset 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.

Table 3: Lupus dataset used in the experiment

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

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) [33], TimeSearcher [34], and Hierarchical Clustering Explorer (HCE) [35] are free tools, while Spotfire® [7] and GeneSpring® [25] are commercial tools.

Clusterview (Figure 1) uses a heat-map visualization for both data overview and details. A compressed heat-map provides an overview of all values in the dataset, 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.

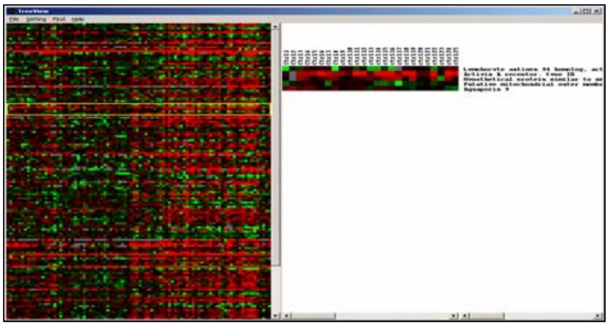

Figure 1: Cluster/Treeview (Clusterview) [33]

TimeSearcher (Figure2) uses a parallel-coordinate visualization for data overview. Line graphs and detailed information are also provided for each individual data entity. The views are tightly coupled using the concept of 'brushing and linking', selecting a gene in one view highlights it in all views. TimeSearcher provides dynamic query widgets directly in the parallel-coordinate overview to support interactive filtering based on user specified time-series patterns.

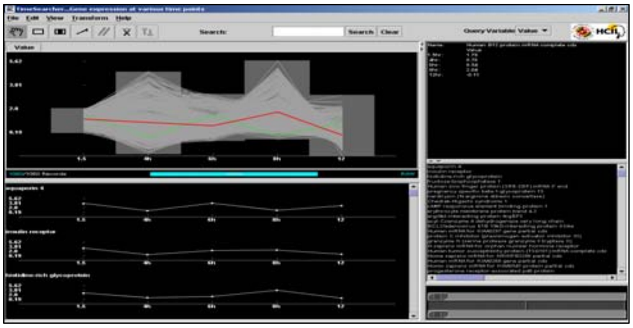

Figure 2: TimeSearcher [34]

HCE (Figure 3) provides several different visualizations: scatter plots, histograms, heat maps, and parallel-coordinate displays for data. 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 dataset, the number of scatterplots possible is very large. HCE introduces a new concept of 'rank by feature' [36] to allow users to quickly find interesting Histograms and Scatterplots. The visualizations are tightly coupled using the interactive concept of brushing and linking. Users can manipulate various properties of the visualizations and also zoom into areas of interest.

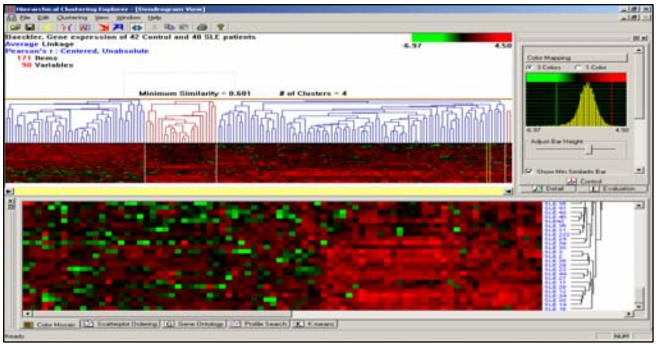
Hierarchical Clustering Explorer - [Dendrogram View]
File Edit Clustering View Window Help
Baechler, Gene expression of 42 Control and 48 SLE patients
Average Linkage
Pearson's r : Centered, Unabsolute
171 Items
90 Variables
-6.97
4.50
Minimum Similarity = 0.601
# of Clusters = 4
Color Mapping
Detail
Control
Evaluation
Color Mosaic
Scatterplot Ordering
Gene Ontology
Profile Search
K-means
Ready

Figure 3: Hierarchical Clustering Explorer (HCE) [35]

Spotfire® (Figure 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.

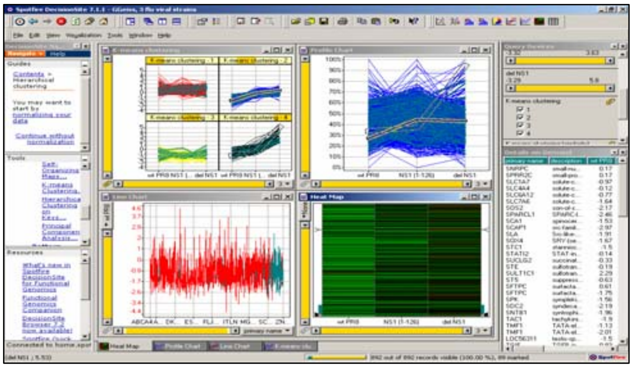

Figure 4: Spotfire® [7]

GeneSpring® (Figure 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. We could not use some of the visualizations, such as physical position and array layout views, for this experiment due to lack of sufficient data. 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.

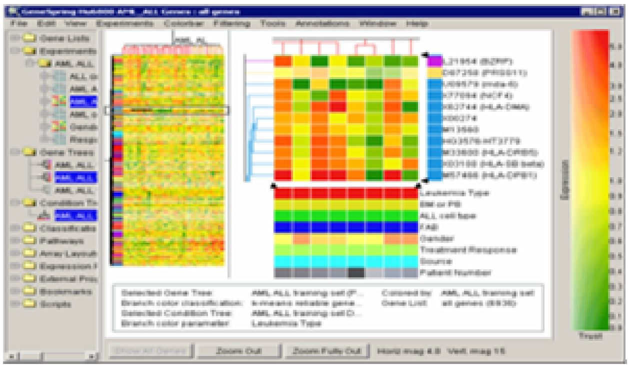

Figure 5: GeneSpring® [25]

Table 4: Summarizes the visualization and interaction techniques supported by each tool (O+D = Overview+Detail; DQ = Dynamic Queries).

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

4.3 Participants

30 test subjects volunteered from the university community. We allotted six users per tool, with two per dataset 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 also to measure learning time, we assigned users to a tool that they had never used before. Based on their profiles, the users fit into one of three categories summarized in Table 5.

Table 5: Participant background and number for each category

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

4.4 Protocol and Measures

To evaluate these 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 overall amount of learning while participants analyze data in an open-ended think-aloud format. 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 dataset and one tool. Before starting their analysis, users were given a background description about the dataset. To reduce initial learning time, the users were given a 15-minute tutorial about the visualization and interaction techniques of the tool. Users then listed some analysis questions they would typically ask about such a dataset. 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%. 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, as described in the next subsection. Table 6 summarizes the dependent variables.

Table 6: Dependent measures

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

5 Results

Results are presented in terms of 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 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 location of genes of interest on chromosomes and pathways. They said that it would be valuable to know what pathways show correlations.

The users working with time series data had questions that focused more on time related changes in gene expression. Most expert users were interested in finding a set of genes that responded earlier to a treatment and was later followed by other genes. Rather than analyzing information for individual patients, the Lupus dataset users were more interested in comparing the overall expression between control and lupus groups. Most novice users wanted to start by taking averages of both the groups to see what genes changed the most from one group to another.

Table 7: Lists all the data questions asked by the participants. The number of participants who asked a particular question is also listed.

	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 expressions to their position on chromosomes	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 having high or low expression for each viral strain	10/10
14	Different patterns of gene expression	10/10
15	Pathway information for genes showing a particular expression pattern	3/10
16	Correlations between different pathways	3/10
17	Chromosomal location of genes that show similar change	3/10
18	Functional information of selected genes	1/10
19	Statistical significance in overall 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 2 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 list all 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

There were collectively 31 distinct questions for all the datasets. It was not possible to answer some of the questions during the experiment, due to insufficient data. GeneSpring® (31/31) and Spotfire® (27/31) can potentially address most of the questions posed by the participants. Clusterview (11/31), TimeSearcher (14/31), and HCE (15/31) answer more specific questions.

5.2 Evaluation on Insight Characteristics

We list here measures for each characteristic of the insight described earlier. Since this evaluation method is more qualitative and subjective than quantitative, and the number of participants is limited, general comparison of tendencies in the results is most appropriate (Figure 6 and Table 6). However, we include some statistical analysis that provides useful indicators.

Facts: We counted the total number of facts i.e. distinct insights about the data for each participant. As shown in figure 6, the count of insights was highest for Spotfire® and lowest for HCE.

Time: The following two temporal characteristics 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 ($p<0.01$).

Average Total Time: The average total time each user 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.

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 more insight value than with GeneSpring® ($p<0.05$). Though, 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, explained later.

Hypotheses: Only a few insights led users to new biological hypotheses. 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 viral dataset 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 these genes belong to and relate all this biology to the data. Spotfire® resulted in one hypothesis for each dataset, thus a total of three. Clusterview also led users to a hypothesis for the Viral and Lupus datasets.

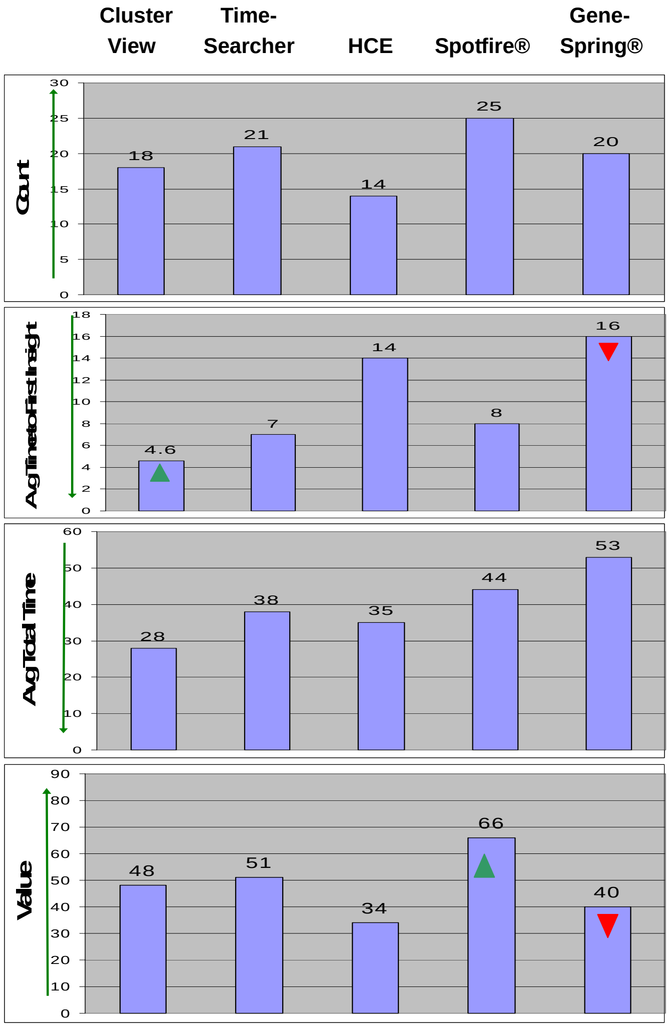

Figure 6: Count of insights, average time to first insight, average total time for each tool, and total insight value, ▲/▼ indicates significantly better/worse performance differences. Y-axis arrows indicate direction of better performance.

Breadth vs. 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'.

Directed vs. Unexpected Insights: The participants using HCE with the Viral dataset 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 dataset. TimeSearcher provided unexpected insights about the time series data, whereas Spotfire® had one each for time series and Lupus

Incorrect Insights (Correctness): HCE proved very helpful to users working with the viral dataset. However, users working with the time series or Lupus datasets did not gain much insight from the data. When prompted to report their data findings, they stated some observations about the data that were incorrect. None of the other tools resulted in incorrect findings.

Table 6: shows the total number of unexpected insights, hypotheses generated, and incorrect insights from the insight occurrences for each tool

Visualization Tool	Unexpected Insights	Hypotheses Generated	Incorrect Insights
Clusterview	3	2	0
TimeSearcher	3	1	0
HCE	5	1	2
Spotfire®	2	3	0
GeneSpring®	0	0	0

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 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. Using 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 Dataset

Now we compare the tools within each dataset.

Time series data: In general, Spotfire® and TimeSearcher performed the best of the 5 tools in this dataset. 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%) compared to TimeSearcher (53%). 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 dataset. 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 5 unexpected insights that were different than the initial information users were searching for in this dataset.

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 vs. Datasets

This section examines individual tools across the three datasets. TimeSearcher and HCE had interesting differences among the datasets (Figure 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 datasets. With time series data, they had over double the value and number of insights than the participants using Viral and Lupus datasets.

HCE: In contrast, participants using HCE did best on the Viral dataset. On Viral dataset, they had a significant 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 datasets. They also felt they learned much more from the data. Participants using 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.

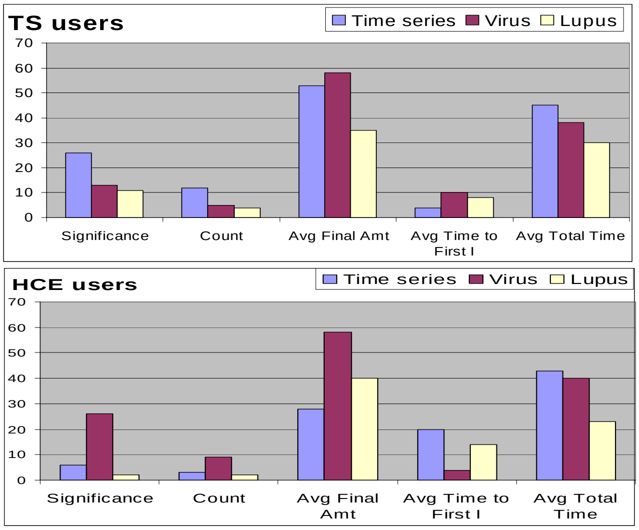

Figure 7: TimeSearcher and HCE specialize in the Time series and Viral datasets respectively.

5.5 Insight Categories

Though a wide variety of insights were made, most could be categorized into a few basic groups. Table 7 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 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 dataset 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 dataset 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 Lupus group as compared to 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 different 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% 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.

Table 7: Total number of insights in each category

Tool	Overview	Patterns	Groups	Detail
Clusterview	9	10	0	2
TimeSearcher	10	8	0	3
HCE	6	5	0	1
Spotfire®	13	10	1	1
GeneSpring®	5	8	4	1

5.6 Learning Curves

During the course of the experiment, users were asked every 10-15 minutes after they began data analysis about how much they felt they learned about the data using the tool. The amount learned is a percentage of total potential insight, as perceived by users. In contrast to other parameters reported earlier, this metric gauges users' belief about insight gained, and about how much the tool is or is not enabling them to discover. Figure 8 presents the average learning curves for all the three datasets for each tool.

The findings reported earlier are further strengthened by the graphs in Figure 8. As shown, participants using the Lupus dataset felt they learned more using Clusterview as compared to the other participants, though they spent almost the same amount of time in the study. Participants using the timeseries dataset felt they learned more as compared to the others with TimeSearcher, participants using the Viral dataset felt they learned the most using HCE. Also, these users spent more time in the study analyzing data as compared to the participants who worked with other datasets. Participants using Lupus dataset spent less time on average in the study for both HCE and TimeSearcher. On average, participants did not report much learning difference across the datasets for Spotfire®. Though, participants analyzing timeseries dataset spent more time in the study as compared to the others. Participants using Virus dataset felt they learned most using GeneSpring®, whereas participants using Lupus dataset felt they learned the least. Also, the participants analyzing time series data spent the least amount of time in the study.

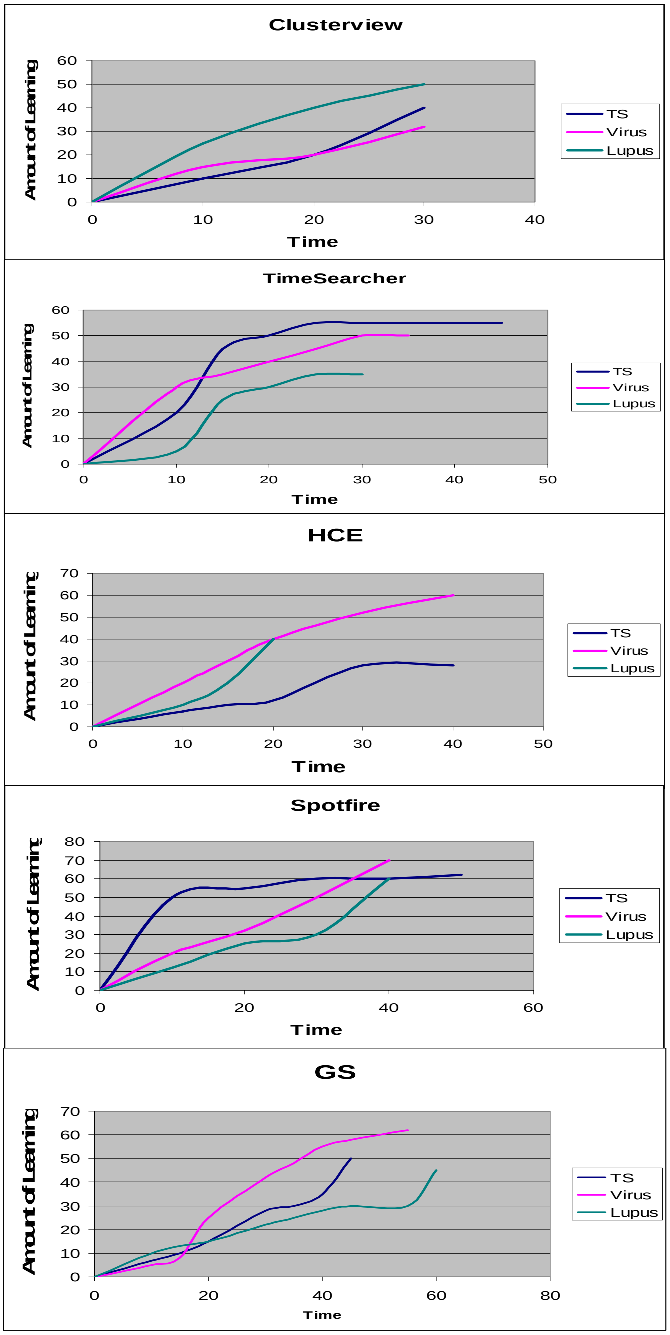

Figure 8: The average learning curves for each dataset for the tools, showing users' estimated insight percentage over time.

Average Final Amount Learned: Figure 9 shows the average of the participants' final stated amount learned for all the datasets for each tool.

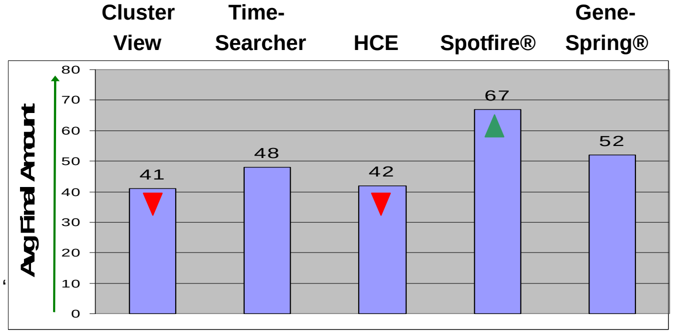

Figure 9: Average final amount learned for each tool. ▲/▼ indicate significantly better/worse differences. Y-axis arrow↑ indicate direction of better performance

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 the datasets, we noticed that for the Lupus dataset 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 datasets 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 datasets.

We noticed that even though tools like Spotfire® and GeneSpring® provides 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 re-ordered based on the results of hierarchical clustering algorithms. Columns and samples with the most similar expression values are placed together. Thus, for both the Time series and Lupus datasets, where a particular column arrangement is useful to recognize changes across the experimental conditions, HCE showed poorer performance. Users focused primarily on the clustering, and apparently did not consider the potential benefits of turning off that feature.

5.8 Participant Comments on Visualization Tools

At the end of each experiment, users were requested to summarize their experience with the tool they used. The following sections summarize 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 timeseries data commented that 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 datasets. The users found the parallel-coordinate visualization provided by TimeSearcher easy to understand. Users working with the timeseries data found the tool very helpful. They were able to easily identify trends and patterns in the data. Users working with Lupus dataset said that it was very difficult for them to see all 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 'timeboxes' as shown in Figure 10, which can prove tedious in a large dataset.

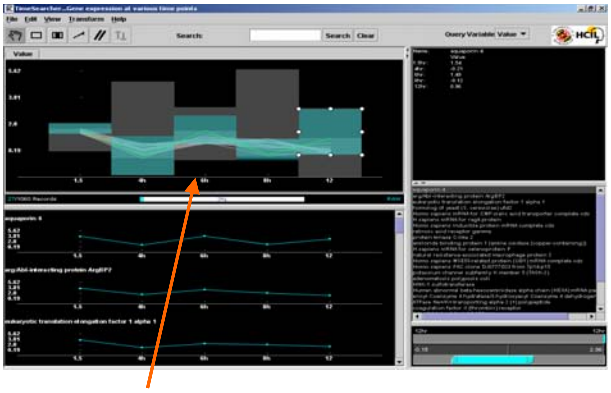

Figure 10: In TimeSearcher, users search for individual patterns of expression using time-boxes.

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 viral dataset. The users working with Lupus dataset 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 visualization were easy to understand. Most users preferred the heat-map visualization of the Spotfire over its parallel coordinate or Profile chart display (Figure 11). Though, the user 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.

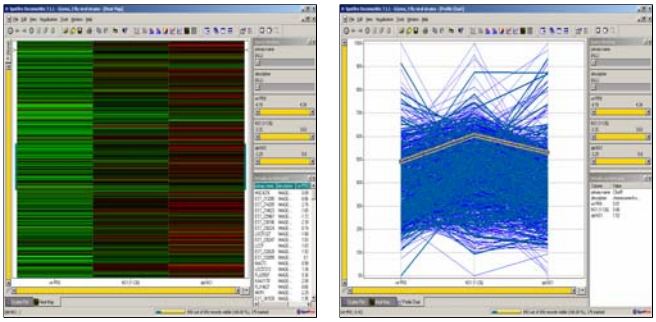

Figure 11: The Heat-map and Parallel Coordinate visualizations in Spotfire [7].

GeneSpring®: Users felt that they will 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 will take them too long to comprehend what the tool does. A few users commented that it will 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 dataset should not be so difficult. None of the users could change different properties of visualization such as color, scale, 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.

5.9 Participants' Background

One might conjecture that users with more domain experience or software development experience would gain more insight from the data. Yet, we found that the insight 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. However, 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 datasets. They were less confident to report their findings.

6 Discussion of Results

Commercial vs. Free: Both Spotfire® and Clusterview users resulted in equivalent insight from the Lupus dataset. 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 dataset can have performance comparable to more comprehensive software containing many different visualizations and features.

Free research software like TimeSearcher and HCE tend to address a smaller set of closely related tasks. Hence, they provide excellent insight on certain datasets. 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 did 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 beginner. 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 (Figure 12). Visualization must support this level of inference.

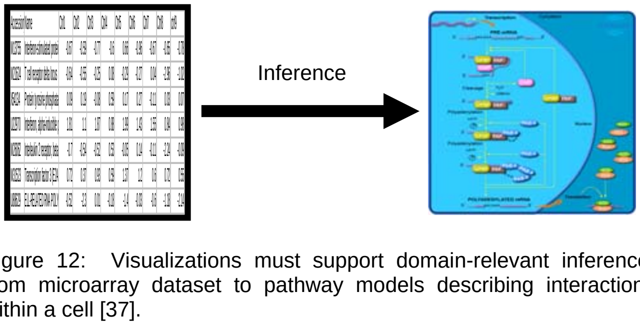

Figure 12: Visualizations must support domain-relevant inference, from microarray dataset to pathway models describing interactions within a cell [37].

Interaction Design: The design of interaction mechanisms in visualization is critically important. Usability 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. Selected lines in its parallel coordinates 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.

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 tools. 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.

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 higher quantity of clicks required to accomplish tasks) reduced their overall insight performance.

User Preferences: Certain visualizations, such as the clustering vsualizations for both Spotfire® and GeneSpring® were the most widely used in the study. The users commented that it would be very helpful if the interaction techniques for these were improved, so that they were better integrated into the overall interaction model.

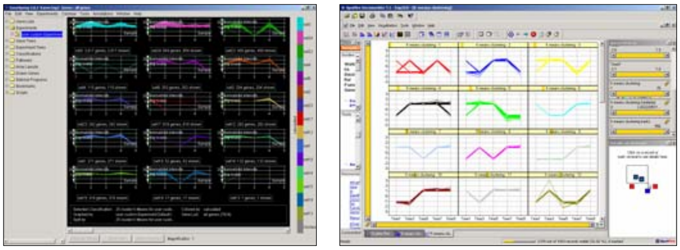

Figure 13: Clustering visualizations were the most widely used in the study. GeneSpring® left, Spotfire® right.

Clustering was a very useful feature throughout, but care should be taken to provide non-clustered 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 non-clustered 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 sizeable number of breadth insights in spite of low motivation levels.

7 Discussion of Methodology

This study takes as its major premise the belief that insight can be measured. We defined insight based on our observations of scientists doing visual data analysis. We recognized, in their comments and actions, characteristics that revealed insight. This measurement process also enables recognition of qualitative aspects of user behavior.

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, which is still useful for detailed testing of specific targeted tasks.

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

- Labor intensive. It is time consuming for the experimenters to capture and code insights.
- 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 more realistic visualization usage.

The study reported here measures insight from short term usage, typically under 2 hours per user. In real world scenarios, users spend days, weeks and even months analyzing data. Moreover, the participants in the study were unfamiliar with the data. The only background knowledge they had was what we provided during the course of study. It is very difficult to appreciate the biological relevance of the microarray data they were analyzing. In this case, the hypotheses they reported were more speculative. Yet, they were not trivial. This suggests that the visualizations are provoking the users to think deeply about the data and also to apply the insight in their domain. Once a user is familiarized with a visualization, the method in which it is used may change. Furthermore, the long-term insight may be very different than short term insight. Long term insight could be broader understanding that guides biologists through multiple cycles of microarray experiments.

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. [21] and [22] present such longitudinal studies that included frequent user interviews, diary studies and 'Eureka' reports. Such studies can help to identify the broader information needs, and develop more meaningful tools that leverage users' domain knowledge and expertise.

8 Conclusions

This study suggests the following major conclusions for life scientists, 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 dataset. We sought to answer the question of which is the best tool to use. Some tools work more effectively with certain types of data. Both TimeSearcher and HCE performed better with the Time series and viral datasets respectively, for others they provided below average results. Thus, dataset dictates which tool is best to use. Additionally, larger software packages like Spotfire® and GeneSpring® work consistently across different datasets. If a researcher needs to work with multiple kinds of data, software like Spotfire® and GeneSpring® 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. 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 are 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 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. This methodology can prove helpful for future studies for analyzing the effectiveness visualizations in many domains.

10 Acknowledgements

We would like to thank all participants without whom the study would not have been possible. We 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. We also extend our 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.genechips.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.http://ihome.cuhk.edu.hk/%7Eb400 559/, 2004.

[6] A. Robinson, “Bioinformatics Visualization”, http://industry.ebi.ac.uk/~alan/, 2002.

[7] SPOTFIRE® Decisionsite™ for functional Genomics, http://www.spotfire.com.

[8] C. Chen and M. Czerwinski, “Empirical evaluation of information visualizations: an introduction”, Int. J. Human- Computer Studies, vol 53, pp 631-635, 2000

[9] C. Chen and Y. Yu, “Empirical studies of information visualization: a meta-analysis”, Int. J. Human- Computer Studies, vol 53, PP 851-866, 2000

[10] A. Kobsa, “An Empirical Comparison of Three Commercial Information Visualization Systems”, Proceedings of InfoVis 2001, pp: 123-130, 2001.

[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” Proceedings of ACM CHI’92, pp 373-380, 1992.

[14] J. D. Mackinlay, “Automating the design of graphical presentations of relational information”, ACM Transactions on Graphics, vol 5, pp 110 – 141, 1986.

[15] E. Tufte, “The Visual Display of Quantitative Information.”, 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 – AVI’02, poster, pp. 373-374, 2002.

[17] B. Shneiderman, “The eyes have it: a task by data type taxonomy”, Proc. IEEE Symp. on 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,* Wadsworth Advanced Books and Software, Monterey, California, 93940, 1980

[20] O. Juarez, “CAEVA: Cognitive Architecture to Evaluate Visualization Applications, Proc. Intl. Conf on Information Visualization-- IV’03, pp: 589-595, 2003.

[21] V. Gonzales and A. Kobsa, “A workplace study of the adoption of information visualization systems”, Proceedings of I-KNOW'03: 3rd International Conference on KnowledgeManagement, Graz, Austria, pp 92-102, 2003.

[22] J. Rieman, “A field study of exploratory learning Strategies, ACM Transactions on Computer-Human Interaction, vol 3, 189-218, 1996.

[23] C. Plaisant, “The Challenge of Information Visualization Evaluation, Proc. of Advanced Visual Interfaces --AVI’04”, 2004.

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

[25] GENESPRING®, Cutting-edge tools for expression analysis, www.silicongenetics.com.

[26] L. Heath and N. Ramakrishnan, “The Emerging Landscape of Bioinformatics Software Systems”, IEEE Computer 35(7), 41-45, 2002

[27] G. Churchill, “Fundamentals of experimental design for cDNA microarrays”, Nature Genetics, vol 32, pp 490-495, 2002

[28] J. Quackenbush, “Microarray data normalization and Transformation”, Nature Genetics, vol 32, 496-501, 2002

[29] K. A. Duca, H. Goto, Y. Kawaoka, and J. Yin, “Time-Resolved mRNA Profiling During Influenza Infection: Extracting Information from a Challenging Experimental System”. American Society for Virology, 20th Annual Meeting, Madison, WI. Data Website: http://infovis.cs.vt.edu/cs5764/ fall2003/ideas/influenza.doc., 2001

[30] 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, Issue 16, 10736–41, 2002.

[31] E. Baechler, F Batliwala, G. Karvpis, P. Gaffney, W. Ortmann, K. Espe, K. Shark, W. Grande, K Hughes, K Kapur, P. Gregersen, and Behrens T., “Interferon-inducible gene expression signature in peripheral blood cells of patients with severe SLE” PNAS vol 100, Issue 5, 2610-5. 2003.

[32] Affymetrix Technologiyes, http://www.affymetrix.com

[33] M. Eisen, P Shellman, P Brown, and D Bostein, “Cluster analysis and display of genome-wide expression” patterns, PNAS vol 95, Issue 25, 14963-68, 1998.

[34] H. Hochheriser, E. H. Baehrecke, S. M. Mount, and B. Shneiderman, “Dynamic Querying for Pattern Identification in Microarray and Genomic Data, Proc. of IEEE International Conference on Multimedia and Expo. 2003.

[35] J. Seo and B. Shneiderman, “Interactively Exploring Hierarchical Clustering Results”, IEEE Computer, vol 35, 80-86, 2002.

[36] J. Seo and B. Shneiderman, “A Rank-by-Feature Framework for Unsupervised Multidimensional Data Exploration Using Low Dimensional Projections”, Proceedings of InfoVis 2004.

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

[38] P. Saraiya, C. North, K. Duca, “An Evaluation of Microarray Visualization Tools for Biological Insight”, Proceedings of InfoVis 2004.

Purvi Saraiya received the B.E. degree in Computer Engineering from L.D. College of Engineering, Ahmedabad, India in 2000. She received the MS degree in Computer Science from Virginia Tech in 2002. She is a Ph.D. candidate in Computer Science at Virginia Tech. Her research interests are Design and evaluation of Information Visualization software, Human computer Interaction, and User Interface Software.

Chris North

Karen Duca