An Evaluation of Microarray Visualization Tools for Biological Insight

Purvi Saraiya[1], Chris North[2]
Department of Computer Science
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061 USA

Karen Duca[3]
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061 USA

ABSTRACT

High-throughput experiments such as gene expression microarrays in the life sciences result in large datasets. In response, a wide variety of visualization tools have been created to facilitate data analysis. Biologists often face a dilemma in choosing the best tool for their situation. The tool that works best for one biologist may not work well for another due to differences in the type of insight they seek from their data. A primary purpose of a visualization tool is to provide domain-relevant insight into the data. Ideally, any user wants maximum information in the least possible time. In this paper we identify several distinct characteristics of insight that enable us to recognize and quantify it. Based on this, we empirically evaluate five popular microarray visualization tools. Our conclusions can guide biologists in selecting the best tool for their data, and computer scientists in developing and evaluating visualizations.

CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation/Methodology, I.6.9 [Visualization] – Information Visualization, Visualization Systems and software, Visualization techniques and Methodologies

Keywords: Data visualization, empirical evaluation, insight, high throughput experiments, microarray data, bioinformatics

1 INTRODUCTION

Biologists use high-throughput experiments to answer complex biological research questions. Experiments, such as gene-expression microarrays [8], result in datasets that are very large. 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. Hence, they use various data visualizations to derive biological inferences. 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 [2], [19] and [26].

However, in collaborations with biologists, we received mixed feedback and reviews about these tools. First, with so many tools to choose from, 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 are meeting their needs in providing insight.

The main goal of the research reported in this paper is to evaluate some of the most popular visualization tools for microarray data analysis, such as Spotfire® [30]. The key research questions are: How successful are these tools in assisting the biologists in arriving at domain-relevant insights? How do the various visualization techniques affect users' perception of data? How does the user's background affect the tool usage?

Typically, visualization evaluations have focused on controlled measurements of user performance and accuracy on predetermined tasks. However, to answer these research questions requires an evaluation method that more closely matches the exploratory nature of the biologists' goals. We devise and deploy an insight-based approach to visualization evaluation that we believe can be generally applied in other data 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 [4], [5]. 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., [18] 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 [14]. Refer to [24] for a professional example.

Metrics, Heuristics, and Models: Different from empirical evaluations are inspections of user interfaces by experts, such as with heuristics [21]. Examples of specific metrics for visualizations include expressiveness and effectiveness criteria [20], data density and data/ink [31], a variety of other criteria for representation and interaction [10], as well as high-level design principles [28]. Cognitive models, such as CAEVA [17], can be used to simulate visualization usage and thereby study 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 [13]. Their work suggests advantages when visualizations are used as complementary products rather than stand alone products. [25] examines users' long-term exploratory learning of new user interfaces, with 'eureka reports' to record learning events.

Thus, a range of evaluation methods have been used to measure effectiveness of visualizations [22]. 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.

[1] email: psaraiya@vt.edu
[2] email: north@cs.vt.edu
[3] email: kduca@vbi.vt.edu

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* [29]. 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. It 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 insights. Hence, we developed an evaluation protocol that focuses on recognition and quantification of insights gained from actual exploratory use of visualizations.

3 Experiment Design

The main aim of this 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

3.1 Microarray Datasets

To examine a range of data scenarios, we used data from three common types of gene-expression microarray experiments. The 3 datasets are summarized in Table 1. 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 [6] and [23]) 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.

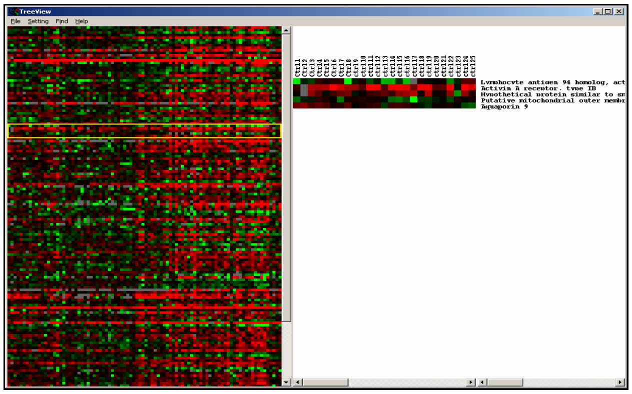
Table 1: Microarray datasets in the experiment

Dataset	Description
Time Series	Data for 1060 genes over 5 time points of a viral infection cycle in human embryonic kidney cells [7]. (1060 rows, 5 columns)
Viral Conditions	Data for 861 genes for 3 related viral infections at 8 hrs post infection of human lung epithelial cells [11]. (861 rows, 3 columns)
Lupus vs. Control	Data for 170 genes from 42 control (healthy) people and 48 people suffering from systemic lupus erythematosus (SLE), an autoimmune disease [1]. (170 rows, 90 columns)

3.2 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) [9], TimeSearcher [16], and Hierarchical Clustering Explorer (HCE) [27] are free tools, while Spotfire® [30] and GeneSpring® [12] are commercial tools. Table 2 and Figures 1-5 summarize these tools.

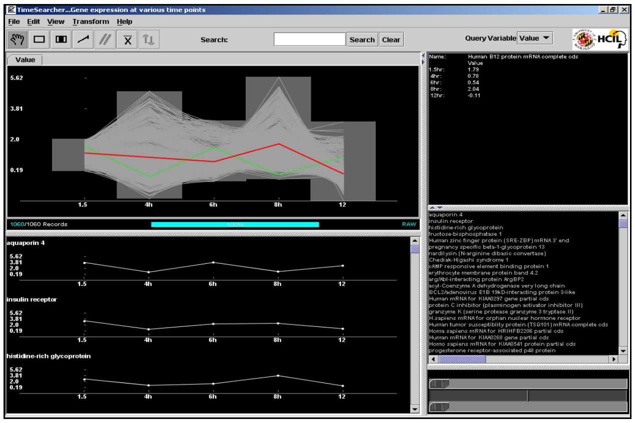
Clusterview (Figure 1) uses a heat-map visualization for both data overview and details. A compressed heat-map provides an overview of all values in the dataset, in row-column format. Users can select a part of the overview to study in more detail. It is standard practice in bioinformatics to visually 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.



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

Figure 1: Cluster/Treeview (Clusterview) [9]

TimeSearcher (Figure2) uses a parallel-coordinate visualization for data overview. Line graphs and detailed information are also provided for each individual data entity. TimeSearcher provides dynamic query widgets directly in the parallel-coordinate overview to support interactive filtering based on user specified time-series patterns.



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
Homa 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

Figure 2: TimeSearcher [16]

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 places similar data items near each other in the tree display. 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.

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.

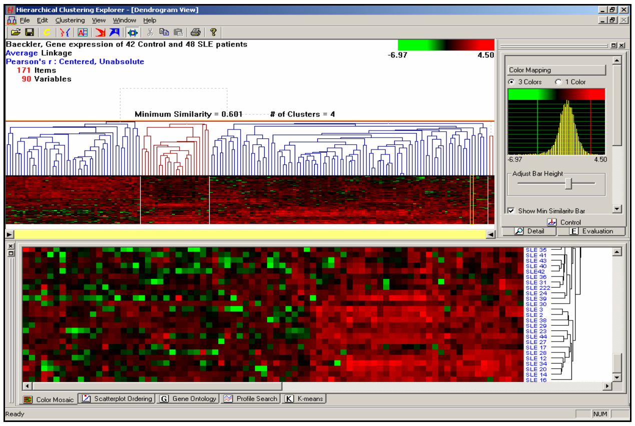


Figure 3: Hierarchical Clustering Explorer (HCE) [27]

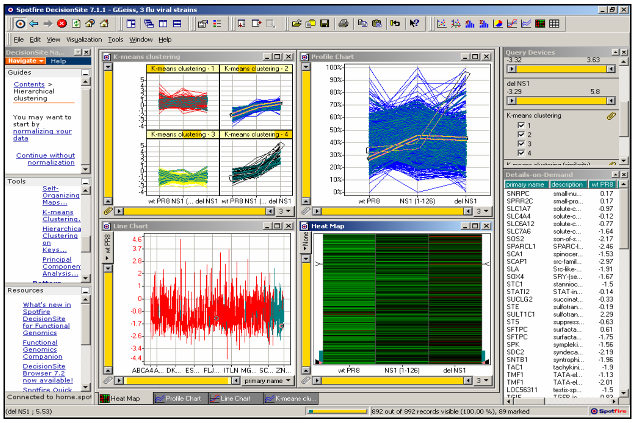
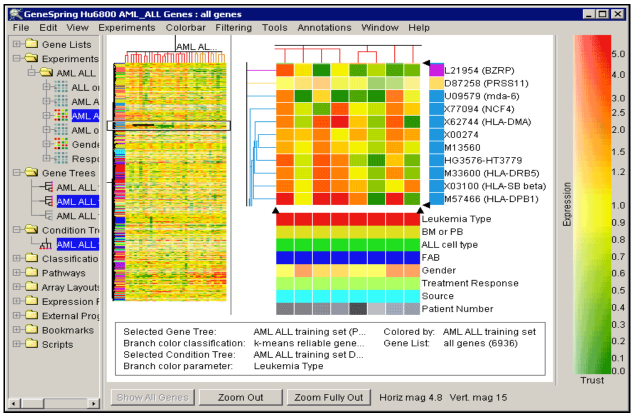


Figure 4: Spotfire® [30]



Selected Gene Tree: AML ALL training set (P...
Branch color classification: k-means reliable gene...
Selected Condition Tree: AML ALL training set D...
Branch color parameter: Leukemia Type
Colored by: AML ALL training set
Gene List: all genes (6936)

Figure 5: GeneSpring® [12]

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.

Table 2: 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, cluster dendrogram	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, clustering	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, clustering	Brushing, Zooming

3.3 Participants

30 test participants volunteered from the university community. We allotted six participants per tool, with two per dataset per tool. We required all participants 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 participants a tool that they had never worked with before. Based on their profiles, the participants fit into one of three categories summarized in Table 3.

Table 3: 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

3.4 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 participant was assigned one dataset and one tool. Before starting their analysis, participants were given a background description about the dataset. To reduce initial learning time, the participants were given a 15-minute tutorial about the visualization and interaction techniques of the tool. Participants 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. Participants 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, participants were asked to comment on their observations, inferences and conclusions. Approximately every 15 minutes, participants 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, participants 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 4 summarizes the dependent variables.

Table 4: 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

3.5 Insight Definition and Characteristics

To measure insights gained from visualization, a more rigorous definition and coding scheme is required. While the subjective 'amount learned' metric provides a measure of overall insight level, a mechanism is needed to capture more specific individual insight occurrences. Through a pilot study with 5 participants, we found that it is possible to recognize and characterize insights as they occur. 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. Although we present them here in the context of biological and microarray data, this can be applied to other data domains as well.

Characteristics of an insight:

- **Fact:** The actual finding about the data. We counted distinct facts for each participant.
- **Time:** The amount of time required 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 was coded on a scale of 1 to 5 by a biology expert familiar with the results of all 3 datasets. In general, trivial observations earned 1-2 points, insights about a particular process earned an intermediate value of 3, and insights that confirmed, denied, or created a hypothesis earned 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 'closing the loop' of the experimental process, in which data analysis feeds back into design of the next experimental iteration [15].
- **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 was 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.
- **Correctness:** Some insights are incorrect observations that result from misinterpreting the visualization. This was coded by an expert biologist and visualization expert together.
- **Category:** Insights were 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 categories were not predefined, but were identified from the experiment results after all insights were collected.

The result of this coding is a single table containing all the insight occurrences and their characteristics, for each participant. Note that insights are distinct within a given participant.

4 Results

Results are presented in terms of participants' data questions, insights, visualization usage, and participant background.

4.1 Initial Questions

At the start of each session, participants were requested to formulate questions about the data that they expected the visualization to answer. Almost all the participants 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 participants working with time series data had questions that focused more on time related changes in gene expression. Most expert participants 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 participants wanted to start by taking averages of both the groups to see what genes changed the most from one group to another. One expert participant said it would be interesting to see how patient characteristics such as their age, gender, and ethnic race affect and cause variance in the data.

There were collectively 31 distinct questions. 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.

4.2 Insight Gained

Figure 6 summarizes several aggregate measures of insight gained and usage time, combining all 6 participants and all 3 datasets for each visualization tool. It shows the following measures:

- **Count (of insights):** the total number of insights acquired.

- **Total Domain Value:** the sum of the domain value of all the insight occurrences. Together, higher total value and count indicate a more effective tool for providing useful insight.
- **Average Final Amount Learned:** the average of the participants' final stated amount learned. The amount learned is a percentage of total potential insight, as perceived by participants. In contrast to total value and count, this metric gauges users' belief about insight gained, and about how much the tool is *not* enabling them to discover.
- **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 participants are able to get immersed in the data more quickly, and thus may indicate a faster tool learning time.
- **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 participants gave up on the tool due to lack of further insight.

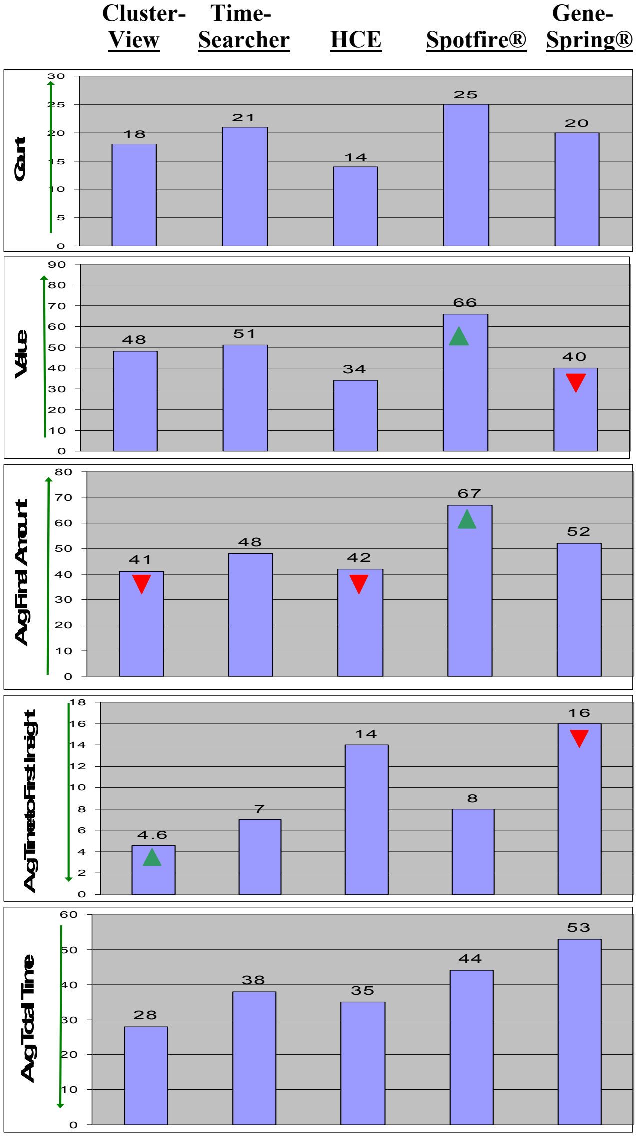


Figure 6: Count of insights, total insight value, average final amount learned, average time to first insight, and average total time for each tool. ▲/▼ indicate significantly better/worse differences. Y-axis arrows ↑ indicate direction of better performance.

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. However, we include some statistical analysis that provides useful indicators. Insight value was highest for Spotfire®. Participants using Spotfire® gained significantly more insight value than with GeneSpring® ($p<0.05$). Similarly, the count of insights was highest for Spotfire® and lowest for HCE. Participants felt they learned the most from the data using Spotfire®. Spotfire® users felt they learned significantly more from the data than participants using HCE ($p<0.05$) and Clusterview ($p<0.05$). 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 participants took significantly less time ($p<0.01$) to reach the first insight than the other users, while GeneSpring® took significantly longer ($p<0.01$). In general, Clusterview users finished quickly while GeneSpring® users took twice as long.

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.

Hypotheses: Only a few insights led participants 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 participant 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. Spotfire® resulted in one hypothesis for each dataset, thus a total of three. Clusterview also led participants to a hypothesis for the Viral and Lupus datasets.

Incorrect Insights: HCE proved very helpful to participants working with the viral dataset. However, participants working with the Time series or Lupus datasets could not get 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.

Table 6: Unexpected, hypotheses, and incorrect insights

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

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®, participants could infer the overall behavior of the data and the patterns of gene expressions. However because the participants found the tool complicated to use, most of them were overly consumed with learning the tool rather than analyzing the data, and were frustrated. They had difficulty getting beyond simple insights. HCE's strengths will become clear in the next two sections.

4.3 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 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 from the initial information participants were searching for in the dataset.

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

4.4 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.

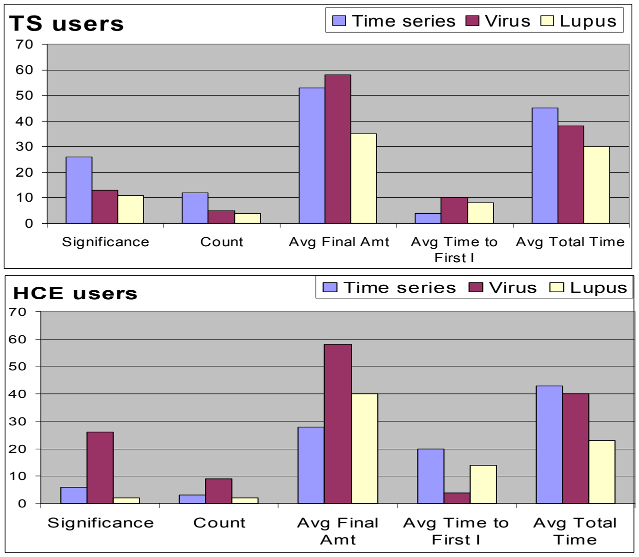


Figure 7: TimeSearcher and HCE specialize in the Time series and Viral datasets respectively.

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 data. On this dataset, they had a significant better performance 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.

4.5 Insight Categories

Though a wide variety of insights were made, most could be categorized into a few basic groups through a clustering process. 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 participant 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".

Expression Patterns: Most participants 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 participants 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, e.g. 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. Participants found this functionality very helpful. Also most of the participants were very pleased to learn that they could link the biological information 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

4.6 Visual Representations and Interaction

Spotfire® participants 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.

We noticed that for the Lupus dataset Spotfire® and Clusterview users liked 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 this dataset using these visualizations. This finding is strengthened by the fact that both TimeSearcher and GeneSpring® users showed average performance on this dataset. Users of these tools used parallel coordinate visualizations to analyze the data.

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 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.

4.7 Participants' Background

One might conjecture that participants 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 participant 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 participants' behavior during the experiment. Novice participants needed more prompting to make comments about the datasets. They were less confident to report their findings.

5 Discussion

This study attempts to measure insight. We accomplish this by defining insight, identifying several measurable characteristics of insight, and establishing methods to recognize insight. These measures are based on our observations of scientists doing data analysis. This measurement process also enables recognition of qualitative aspects of user behavior. Clearly, true insight has a much broader meaning. However, although our definition is not comprehensive, it does provide an approximation of participants' insight. This, in turn, has enabled us as evaluators to gain insight into the effectiveness of these visualization tools.

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. We hypothesized that someone more expert 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 insights generated. 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.

Choice of visualization tool can have major impacts. Both Spotfire® and Clusterview participants 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.

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 an occluding visual highlight that made it very difficult for users to determine which genes were selected. 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.

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®.

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 those who had designed the 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.

6 Conclusions

Our study suggests the following major conclusions for life scientists, visualization designers, and evaluators.

Biologists: A visualization tool clearly influences the insight gained. Hence, it is imperative that the appropriate tool for the dataset be chosen. 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, 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.

Visualization Designers: Interaction techniques play a key role in determining visualization effectiveness. It is imperative that users are able to access and link biological information to their data. Designers should emphasize consistent usable interaction design models with clear visual feedback. The grouping interaction and clustering is a must. Identify which visualization technique in a given software is used the most and improve it.

Evaluators: The main purpose of visualization is to provide insight. This can be difficult to measure. Our insight definition allowed us to quantify insight using different characteristics. This can prove helpful for future studies for analyzing visualizations for effectiveness.

7 Future Work

In the real world, researchers spend days, weeks and often months analyzing their data. While this short-term study was useful for gauging early insight generation, current problems with this approach include lack of sufficient training and high coding effort. 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 visualization tools influence knowledge acquisition. These studies should be conducted with researchers analyzing their own experimental results for the first time. [25] and [18] present such longitudinal studies that included frequent user interviews, diary studies and 'Eureka' reports. Such studies could identify the broader information needs, and help to develop more meaningful tools that leverage biological knowledge and users' domain expertise. Moreover, the analysis of high-throughput data is in its infancy and improved analysis frameworks within the life sciences are needed.

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

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

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

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

[5] Chen, C., and Yu, Y. 2000, Empirical studies of information visualization: a meta-analysis, *IJHCS* vol 53, 851-866.

[6] Churchill, G. 2002, Fundamentals of experimental design for cDNA microarrays, *Nature Genetics*, vol 32, 490-495.

[7] Duca, K., A., Goto, H., Kawaoka, Y. and Yin, J. 2001, 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.xls.

[8] Duggan, D., Bittner, B., Chen, Y., Meltzer, P., and Trent, J. 1999, Expression profiling using cDNA microarrays, *Nature Genetics* vol 21, 11-19.

[9] Eisen, M., Spellman, P., Brown, P., and Botstein, D. 1998, Cluster analysis and display of genome-wide expression patterns, *PNAS* vol 95, Issue 25, 14963-68.

[10] Freitas, C., Luzzardi, P., Cava, R., Pimenta, M. Winckler, A., and Nedel, L. 2002, Evaluating Usability of Information Visualization Techniques. *Proc. Advanced Visual Interfaces – AVI'02*, poster, p. 373-374.

[11] Geiss, G., Salvatore, M., Tumpey, T., Carter, V., Wang, X., Basler, C., Taubenberger, J., Bumgarner, R., Palese, P., Katze, M., and Garcia-Sastre, A. 2002, 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.

[12] GeneSpring®, Cutting-edge tools for expression analysis, www.silicongenetics.com.

[13] González, V., and Kobsa, A. 2003, A workplace study of the adoption of information visualization systems, *Proceedings of I-KNOW'03: 3rd International Conference on Knowledge Management*, Graz, Austria, 92-102.

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

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

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

[17] Juarez, O. 2003, CAEVA: Cognitive Architecture to Evaluate Visualization Applications, *Proc. Intl. Conf on Information Visualization-- IV'03*, 589-595.

[18] Kobsa, A. 2001, An Empirical Comparison of Three Commercial Information Visualization Systems, *Proceedings of InfoVis 2001*, pg. 123-130.

[19] Lueng, Y. F. 2004, Functional Genomics, http://genomicshome.com.

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

[21] Nielsen, J. 1992, Finding usability problems through heuristic evaluation. *Proceedings ACM CHI'92*, 373-380.

[22] Plaisant, C., 2004, The Challenge of Information Visualization Evaluation, *Proc. of Advanced Visual Interfaces --AVI'04.*

[23] Quackenbush, J. 2002, Microarray data normalization and Transformation, *Nature Genetics*, vol 32, 496-501.

[24] Rao, G., and Mingay, D. 2001, Report on usability testing of census bureau's dynamaps CD-ROM product, http://infovis.cs.vt.edu/cs5764/papers/dynamapsUsability.pdf.

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

[26] Robinson, A. 2002, Bioinformatics visualizations, http://industry.ebi.ac.uk/~alan/.

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

[28] Shneiderman, B. 1996, The eyes have it: a task by data type taxonomy, *Proc. IEEE Symp. on Visual Languages '96*, pg 336-343.

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

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

[31] Tufte, E. 1983, *The Visual Display of Quantitative Information*.