Comparing Benchmark Task and Insight Evaluation Methods on Timeseries Graph Visualizations

Purvi Saraiya[1], Chris North[1], Karen Duca[2]
Department of Computer Science[1]
Virginia Bioinformatics Institute[2]
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061, USA
north@vt.edu; http://infovis.cs.vt.edu/

ABSTRACT
A study to compare two different empirical research methods for evaluating visualization tools is described: the traditional benchmark-task method and the insight method. The methods were compared using different criteria such as: the conclusions about the visualization tools provided by each method, the time participants spent during the study, the time and effort required to analyze the resulting empirical data, and the effect of individual differences between participants on the results. The studies used three graph visualization alternatives to associate bioinformatics microarray timeseries data to pathway graph vertices, based on popular approaches used in existing bioinformatics software.

Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces – Evaluation/Methodology.

Keywords: Empirical evaluation, graph visualization, timeseries data analysis.

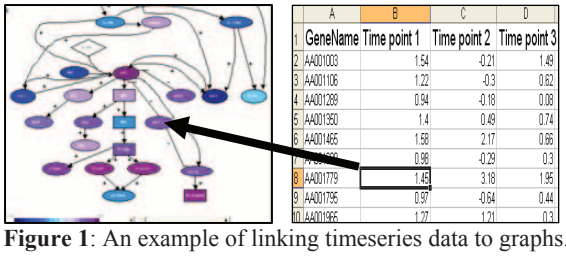
1. INTRODUCTION
Visualization tools are often evaluated in controlled studies that use benchmark tasks [1, 2]. Participants are usually given a variety of predefined tasks to perform on pre-selected data during the course of the study. The performance time and accuracy of the participants' responses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

An attempt to capture the real world exploratory data analysis scenario in a short-term study, using an insight-based method, is reported in [4]. The method used an unguided protocol requiring the participants to think aloud the insights they glean from the data. The visualization tools were then analyzed based on the quantifiable characteristics of the insights that can be measured uniformly across participants. Thus, in contrast to the controlled studies that use benchmark tasks, the insight method does not use predefined tasks and instead treats tasks as a dependent variable in the experiment.

While the insight method appeared useful, there are open questions about how the method compares to the traditional benchmark task method, and whether the method should be used instead of the benchmark task method to provide meaningful statistical analyses between visualizations or as a complementary approach. Thus, the goal of this paper is to compare both methods: the task-based and insight-based methods. Such studies to compare empirical research methods are more common to the field of usability engineering, but less frequent in the information visualization domain. Thus, a broader research goal is to investigate if such a comparison between methods can be done for information visualization.

A secondary goal is to examine visualization of graphs with associated timeseries data for the bioinformatics domain. In bioinformatics, graphs or 'pathways' with a node-link representation are typically used to represent interactions between bio-molecules (genes, proteins, etc). Multidimensional timeseries data from high throughput experiments such as gene expression microarrays [5] are often analyzed in context of these biological pathway graphs. The graphs provide important biological context to otherwise raw timeseries numerical data analysis [4].

Figure 1 shows the overlay of timeseries data in context of a graph. Each vertex in the graph corresponds to a row in the timeseries dataset, and each experiment treatment or timepoint is a column. Three common visualization methods used by current bioinformatics software tools to overlay multidimensional data on graphs are summarized in [6] and include: (a) Overlaying data on graph vertices for one timepoint at a time (Figure 2) by manipulating a visual property (e.g. color) of the vertex, and using sliders or similar interaction to animate the graph to other timepoints; (b) Data from all the timepoints can be overlaid simultaneously by using complex node glyphs (Figure 3); Or, (c) small multiples can be used to simultaneously display a miniature graph for each timepoint (Figure 4). A detailed description of the design space for graph visualization with associated timeseries is presented in [7]. Though many user studies have been conducted to evaluate graph visualization, few studies have evaluated alternatives for graphs with associated multidimensional data.



	A	B	C	D
1	GeneName	Time point 1	Time point 2	Time point 3
2	AA001003	1.54	-0.21	1.49
3	AA001106	1.22	-0.3	0.62
4	AA001289	0.94	-0.18	0.08
5	AA001350	1.4	0.49	0.74
6	AA001465	1.58	2.17	0.66
7	AA001...	0.98	-0.29	0.3
8	AA001779	1.45	3.18	1.95
9	AA001795	0.97	-0.64	0.44
10	AA001965	1.27	1.21	0.3

Figure 1: An example of linking timeseries data to graphs.

2. LITERATURE SURVEY
The literature for visualization of graphs with multidimensional data is summarized in [7]. While the study in [7] examined the use of multiple views including parallel-coordinates plots, this paper focuses on the primary graph representation itself.

BELIV'10, April 10–11, 2010, Atlanta, GA, USA.

2.1 Comparison of Information Visualization Studies

Different types of studies have been used to evaluate visualization tools, as summarized in [1, 4]. The shortcomings of these studies and the need to develop new evaluation methods for visualization tools that better represent real world data analysis scenarios and also provide better feedback about the usability of the data representation method have been suggested [1, 4, 8].

The literature for comparisons of empirical research methods used to evaluate information visualization tools is sparse, and mostly anecdotal. General guidelines for better tasks and methods to evaluate visualizations can be found in [9]. Recommendations for more consistent and comparable user studies based on a meta analysis is presented in [10]. Authors' comments about user studies for information visualization, and the lessons learned from these studies and how these were used to design more effective visualization tools and evaluation studies is presented in [11]. A panel discussion summarizing research for visualization evaluation using human subjects, and suggestions and guidelines for conducting such studies by several visualization experts based on their experiences is presented in [12]. Expert reviews as an alternative in certain contexts where designing and conducting user studies can be difficult is suggested in [13].

2.2 Comparisons of Studies in Usability Engineering

Several studies have been conducted to analyze and compare methods typically used for evaluating user interfaces. A comparison of usage based evaluation techniques and inspection method for groupware systems is provided in [14]. A study to compare the effectiveness of local vs. remote usability studies is reported in [15]. Two methods for children's computer games are compared in [16]. Usability testing methods with multiple participants is compared to heuristic evaluation in [17]. A list of criteria that can be used to compare usability evaluation methods is presented in [18]. Detailed case study of 6 usability methods that evaluates each method's usability error predictive power to actual user tests is reported in [19]. A comparison of different usability testing methods for information retrieval tasks is provided in [20].

Though studies have been conducted to evaluate usability methods that analyze user interfaces with respect to each other, studies to evaluate empirical research methods for evaluating visualization tools are rare. Most of the usability methods are compared based on the number of usability errors found, severity of these errors and participants' and facilitators' experience in the study. Since the dependent variables for the usability methods are usually the same (usability errors) such direct comparisons between the evaluation methods are possible. However, the dependent variables for the benchmark task based method (performance time, accuracy), and the insight method (data insights) are different. Also, the evaluation for visualization tools investigates a wider range of options (e.g., data representation method, interaction mechanisms used, etc) as compared to the user interface evaluation. Hence higher level measures such as the conclusions about the visualization tools, time spent by the participants in the study, effort spent to analyze the resulting empirical data, etc. need to be used for meaningful comparisons between these two evaluation methods.

3. EXPERIMENT DESIGN

The aim of this study is to analyze and compare two empirical evaluation methods, using three different visualization alternatives that support analysis of timeseries data in context of graphs. Conceptually, a 2X3 between subjects design examines the following two independent variables:

1 Two empirical evaluation methods: benchmark tasks method, and the insight method.
2 Three graph visualization alternatives.

3.1 Data

The biologists we were collaborating with conducted a gene expression microarray experiment to analyze impacts of tobacco smoking on flu infection immune response. The actual data was 45,001 rows (genes) X 72 columns (timepoints and conditions). The biological significance of the data and the actual analysis process for this data by bioinformaticians are presented in [21].

A directed graph, having 46 vertices (or genes) and 36 edges (representing gene interactions) representing an actual immune response pathway, was linked to a timeseries dataset representing gene expression for 12 timepoints (Table 1). Thus, the participants in the experiment were working with a small subset of the actual data. However, the graph size was based on the typical size of the pathways used by the biologists, corroborated in general by [22].

Table 1: Data used for the study.

Data Type	Description
Graph	A directed graph having 46 vertices and 36 edges. Each node had an out degree of 0 to 3.
Timeseries data	Gene expression values for 12 timepoints for each vertex. Of these, 6 timepoints measured expression values for flu infection for non-smokers, and the remaining 6 timepoints corresponded to flu infection for smokers.

3.2 Graph Visualization Alternatives

Three graph visualization alternatives were used in the study. The visual encoding of the data was based on the common color scheme used in bioinformatics, i.e. the color scale from yellow to green was used to display negative data values, and yellow to red was used to display positive data values [6,7].

1 ***Single Timepoint (1 Tpt):*** This visualization overlays values for one timepoint on a node at a time (Figure 2). A slider lets users iterate over all the timepoints in the data. Mousing over nodes displays the corresponding numerical data value.
2 ***Multiple Timepoint (M Tpts):*** This visualization overlays data from all the timepoints on a node using a heat map (Figure 3). Mousing over the heatmap cells displays the corresponding numerical value and the timepoint.
3 ***Multiple Graphs (M Graphs):*** This visualization displays a miniature graph for all the timepoints in the data (Figure 4). Mousing over a node displays its numerical value, the name of the node (because nodes are too small to clearly show name labels), and also the time point corresponding to it.

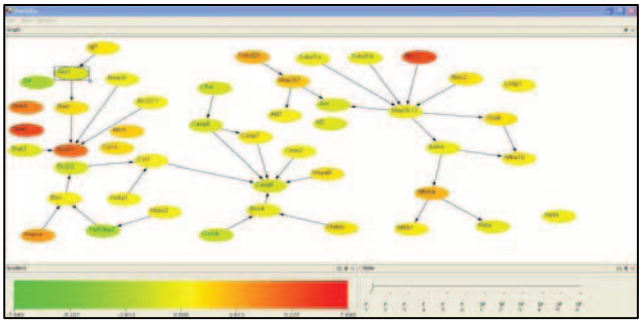


Figure 2: Overlay a single timepoint on graph vertices. A slider is used to navigate between different timepoints.

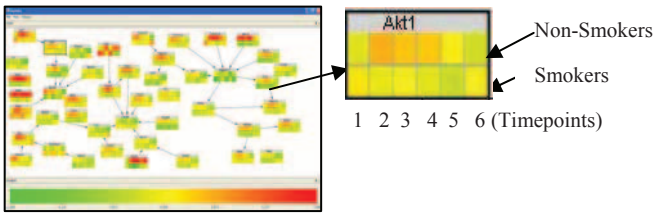


Figure 3: Overlay all the timepoints on graph vertices.

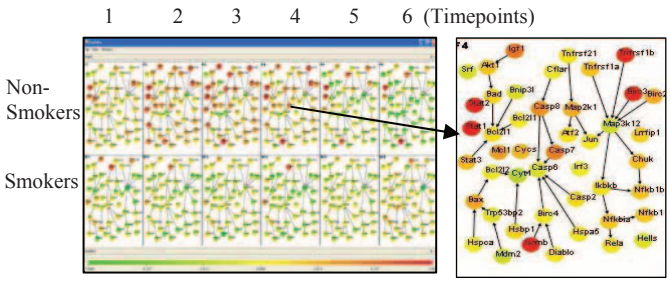


Figure 4: Multiple small graphs to display all data timepoints.

3.3 Participants

60 participants, 10 for each visualization alternative for each evaluation method, participated in the study. Since the data had a biological background, all the participants in the study were sophomore or junior biology students. Hence, the participants were all familiar with the basic concepts of the data, although were not familiar with this specific data set.

3.4 Experiment Protocol

Before beginning, the participants were given a brief introduction to the visualization alternative that they were assigned and the data background used in the study. Then, the protocols were different depending on the assigned evaluation method:

3.4.1 Benchmark Task Method Protocol

Participants were required to perform 7 tasks listed in Table 2. All the tasks were multiple choice questions, with five possible choices. The tasks were based on the observed analysis tasks of the bioinformaticians who designed the biology experiment and analyzed the actual data [21]. Time and correctness were measured for each task (Table 3).

Table 2: Benchmark tasks for the Task-based method.

No.	Task
T1	Which of the following nodes shows a positive value for all Flu timepoints but negative value for all Smoking+Flu timepoints?
T2	What is the overall expression pattern for Flu timepoints vs. Smoking+Flu timepoints?
T3	Which of the following nodes is negative for all 12 timepoints?
T4	Which of the following timepoints has the maximum number of positive nodes?
T5	Which of the following timepoints has the maximum number of negative nodes?
T6	At which of the following timepoints, for both conditions, do most nodes change their expression values from previous timepoints?
T7	How many nodes are between Map3k12 and Rela?

Table 3: Dependent variables for the Task-based method.

Dependent variables	• Time to answer each question • Correctness of answers • Overall time spent in the study • Feedback about the visualization alternative

3.4.2 Insight Method Protocol

Participants were asked to analyze the data in a think-aloud fashion, reporting any insights, until they felt that they had learned all they could from the data. The experimenter sat next to the participants during the study, silently observing the participants' data analysis process and also recording (on a laptop) the data insights and the times at which these were made since beginning the study (Table 4).

Table 4: Dependent variables for the Insight method.

Dependent variables	• Data insights reported • Time at which each insight was reported • Overall time spent in the study • Feedback about the visualization alternative

4. RESULTS

4.1 Benchmark Task Method Results

4.1.1 Overall Performance

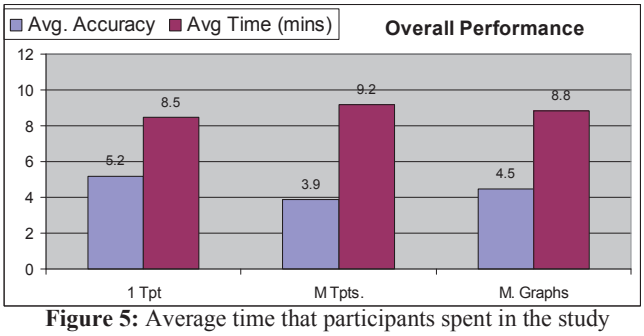


Figure 5: Average time that participants spent in the study (minutes), and the average number of correct responses (out of 7 tasks), for all three visualization types.

On performing ANOVA analysis, we found that there were no significant differences between the participants on the total time spent in the study or overall differences on the accuracy for the tasks, for all the three visualization alternatives. However, the participants using single timepoint visualization were somewhat ($p=0.09$) more accurate than multi timepoint visualization.

4.1.2 Performance for Individual Tasks

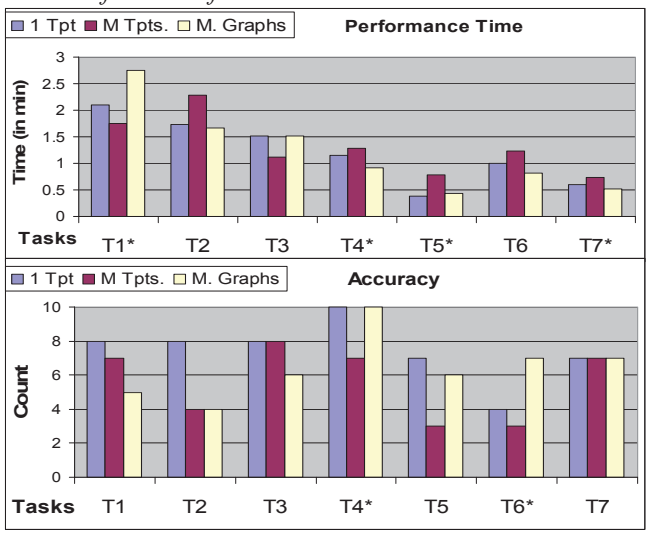


Figure 6: Average time (minutes), and total count of correct responses (out of 10 participants) for each task, for all three visualization types. * indicates significant differences.

Significant results from paired t-tests on tasks between the three visualization alternatives for time and accuracy are summarized in Table 5. Though Tasks 4 and 5 were equivalent, Task 5 required more careful analysis as compared to Task 4, as the timepoint at which most nodes were positive was more obvious as compared to the timepoint at which there were most negative nodes.

Table 5: Summary of results for individual tasks.

Tasks	1 Tpt	M Tpts	M Graphs
T1	-	Fastest (p=0.038)	-
T2	-	-	-
T3	-	-	-
T4	-	Least accurate (p=0.033)	Faster than M Tpts (p=0.01)
T5	Weakly more accurate than M Tpts (p=0.08)	Slowest (p=0.044)	-
T6	-	Weak slowest (p=0.098)	More accurate than M Tpts (p=0.01)
T7	-	-	Faster than M Tpts (p=0.033)

4.1.3 Task-based Conclusions of Visualization Alternatives

The discussion about individual task performances and summary in Table 5 leads to the conclusions about the visualization alternatives summarized in Table 6. Most of the conclusions about the single timepoint and multiple timepoints visualization alternatives are similar to the ones from the study reported in [7]. However, in the previous study it was also found that the multiple timepoints were faster and more accurate to search for the outlier nodes, i.e. nodes that display different behavior than most other nodes. Since we did not have a task to represent this information, it was not possible to make this conclusion about the visualization from the task based method.

Table 6: Conclusions about visualization alternatives from the task-based study.

1 Tpt	M Tpts	M Graphs
+ Consistent performance for all tasks. + Accurate for timepoints	+ Faster performance for single node analysis. – Slower and less accurate for overall graph expression at a particular timepoint.	+ More accurate and faster for finding interesting timepoints. + Faster than M Tpts for graph topology tasks.

4.2 Insight Method Results

4.2.1 Overall Performance

On performing ANOVA analysis we found that the participants using single timepoint visualization spent significantly more time in the study as compared to other participants ($p = 0.02$). The participants using single timepoint visualization had a greater number of distinct data insights than both the multiple timepoint ($p= 0.07$) and multiple graphs ($p=0.04$) users. These results are summarized in Figures 7. Note that the data insights are distinct for a participant. However, the data insights may be repeated across participants when more than one participant reported the same data insights.

4.2.2 Performance based on Insight Category

On analyzing the participants' data insights, we found that all of these could be grouped into 7 distinct categories. Each data insight belongs to only one of these categories. Figure 8 summarizes the participants' performance based on these insight categories for all three visualization alternatives.

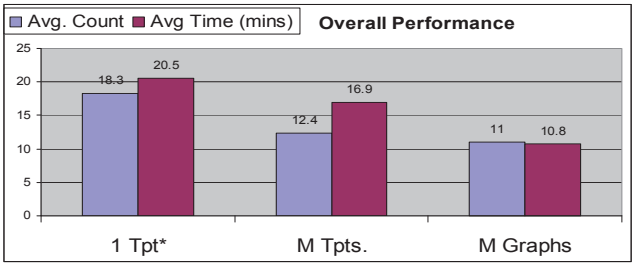


Figure 7: Average amount of time (minutes) participants spent in the study, and average count of data insights reported, for all three visualization types. * indicates significant differences.

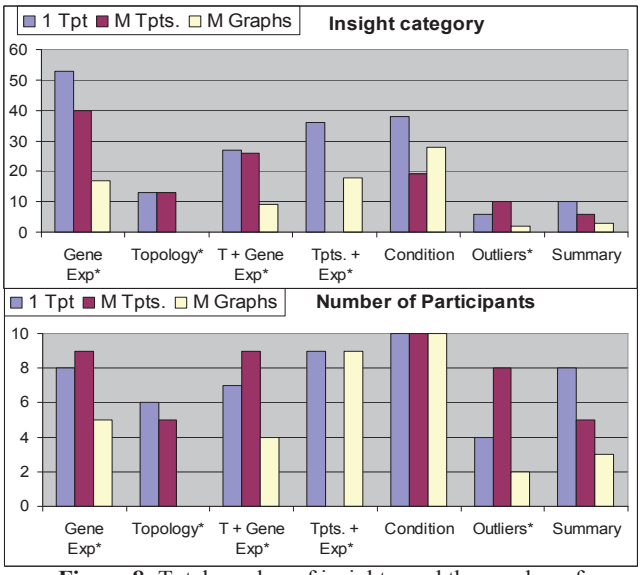


Figure 8: Total number of insights, and the number of participants (out of 10) who reported these, for each insight category. * indicates significant differences.

Gene expression: Most frequent data insights reported expression pattern for just one gene. E.g.: "Gene Gzmb displays positive values for all the non-smaking timepoints except the first timepoint, but is negative for all the smoking timepoints".

Topology: Some of the insights reported involved only the graph topology. This did not include any information about the associated timeseries data. E.g.: "The map3k12, casp6, and bcl2ll genes seem to be major focal points in the graph as they have a lot of arrows pointing towards them". None of the participants using M Graphs reported such insights.

Topology + expression: Some of the insights reported by the participants investigated gene expression based on graph topology or effects of genes on each other connected directly or indirectly through other genes. E.g.: "All the genes towards the outside, i.e. Trnf2, birc3, etc. are more positive for almost all the timepoints as compared to the inside ones that they are supposed to affect".

Timepoint analysis: Some participants reported insights that investigated overall graph expression at a particular timepoint. E.g.: "A lot of genes are negatively expressed at timepoint 5 for smokers as compared to all other timepoints". None of the participants using Multiple Timepoints reported such insights.

Experiment conditions: All participants in the study evaluated the differences in the gene expression between smokers and non-smokers, which appropriately reflects the goals of biology experiment from which the dataset originated. E.g.: "Overall, non-smokers have more positively expressed genes than smokers".

Outliers: Some participants identified a few genes that displayed different expression values than other genes in the graph. E.g.: "Stat1 gene is different than other genes, as it up-regulates with time for non-smokers, whereas most other genes down-regulate".
Summary: Some participants tried to summarize their findings about the data or suggested future research based on their data analysis. These insights are most similar to the hypothesis insight characteristic that was ranked very important in [4]. E.g.: "Smokers don't have many highly expressed genes, and a lot of them may reduce the gene expression of the subsequent genes. This may eventually lead to less expression for the overall immune system against the flu for smokers".

Table 7 lists the results from ANOVA analysis between participants, on number of distinct insights for each category reported by each participant using each visualization alternative.

Table 7: Summary of number of insights for insight categories.

Category	1 Tpt	M Tpts	M Graphs
Gene Expression	-	-	Weak least (p=0.069)
Topology	-	-	Least (p=0.04)
Topology + Expression	-	-	Weak least (p=0.09)
Timepoint Analysis	Most (p=0.03)	Least (p=0.00002)	-
Condition	-	-	-
Outliers	-	Weak most (p=0.06)	-
Summary	More than M Graph (p=0.015)	-	-

4.2.3 Insight Conclusions about Visualization Alternatives
Participants' performance on the insight categories and the summary of data analysis results in Table 7 lead to the conclusions about the visualization alternatives listed in Table 8.

Table 8: Conclusions about the visualization alternatives from the insight-based study.

1 Tpt	M Tpts	M Graphs
+ Somewhat better to summarize findings. + Best for single timepoint analysis. + More consistent performance for all insight categories. + Encouraged users to work longer and find more total insights.	+ Best for identifying outlier nodes. – Difficult to analyze a single timepoint. + Fast rate of insights relating expression to topology.	– Difficult to focus on expression values for a single node. – Difficult to analyze graph topology. + Fast rate of insights comparing experiment conditions.

5. COMPARISON BETWEEN METHODS

5.1 Total Time Spent

On performing ANOVA analysis, overall participants in the insight method spent significantly more total time in the study as compared to the task-based method (Figure 9). Participants using Single Timepoint and Multiple Timepoint visualizations spent significantly more time ($p<0.01$) in the insight method as compared to the task method. Thus, normalizing the results to compute *insights per minute* would produce more advantage for the M Tpts and M Graphs visualizations. It would also be possible to conduct further analysis in the insight method of the sequence of insights gained by participants (as in [4]), which might provide further information about the salience of the different categories of insights in each visualization.

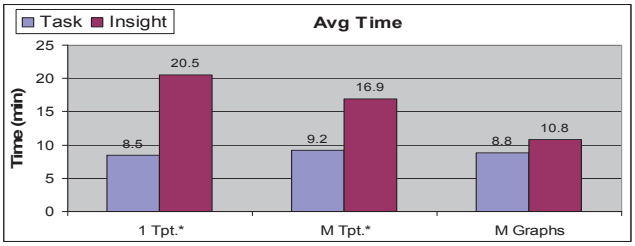


Figure 9: Average time participants spent in the study for each visualization type, for Task and Insight methods. * indicates significant performance differences.

5.2 Comparison on Pre-selected Task Results

Number of participants who reported equivalent insights: Some participants in the insight method reported insights very similar to the tasks that were used in the task-based method. E.g.: Similar to task 4 in the task method, the participants in the insight method reported that timepoint 4 for non-smokers has maximum number of positively expressed genes. Figure 10 summarizes the number of participants that made insights comparable to a particular task. Since task 7 was very specific, we omitted it from the analysis. This chart is an interesting hybrid of the insight and task methods, because it shows how well each visualization enables a specific predefined set of insights.

None of the participants using Multiple Timepoints reported insights involving analysis for a single timepoint (tasks 4, 5, and 6). The participants using Multiple Timepoints were significantly slower or less accurate on these tasks in the task-based method. The effect was confirmed and found even more significant in the insight method when analyzing the insight category: timepoint analysis. This indicates that providing a set of predefined tasks potentially forces participants to perform a type of analysis that they would not otherwise perform and that the visualization would not otherwise encourage. This brings into question the ecological validity of prescribed task-based studies.

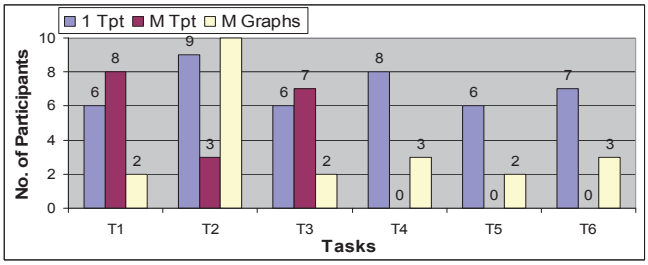


Figure 10: Number of participants for each visualization who reported insights equivalent to the tasks in the task-based method.

Tasks vs. insight categories: Table 9, lists the tasks and the corresponding insight categories. Conclusions about the visualization alternatives based on the method type are also listed. For most of the tasks, each of the two evaluation methods confirmed or partially confirmed the other's results. Tasks 1 and 3 required the participants to analyze expression values for a single node. The task-based method found that M Tpts was faster at one of these tasks. Similarly, the insight method found that M Graphs produced the least of that type of insight. For task 2, which examined differences between conditions, neither method produced significant differences.

The most confirmatory results are for Tasks T4 – T6. These tasks required analysis of the graphs at a single timepoint. Both evaluation methods found significant disadvantages of the M Tpts visualization and advantages of the 1 Tpt visualization in these

types of tasks. The task-based method also found advantages of M Graphs for some of these tasks. The conclusions for Task T5 are the most confirming between the two methods, suggesting identical visualization design recommendations for that case. This suggests a link between how efficiently and accurately a visualization supports a particular task and how much of that type of insight a user gains from using the visualization.

However, Task T7 shows opposite results between the two evaluation methods. For Task 7 we found that the participants performed this task faster using M Graphs. However, in the insight method participants using M Graphs produced significantly less Topology insight (no insight, in fact) than the other tools. This result breaks the previously mentioned link, and highlights the difference between imposed tasks and self-discovered tasks. It is possible for a visualization to support a given task more efficiently than other visualizations, but *not* to promote or encourage the use of that task to gain insight about it as much as the other visualizations. Some possible explanations for this phenomenon are: (1) the visualization steered the user towards other types of insights that were made more perceptually obvious; (2) the visualization made insights of that type appear as Uninteresting or irrelevant to the problem; or (3) the methods are measuring effects and different levels – the task-based method is more perceptually oriented, capturing perceptual efficiency, whereas the insight method is more cognitively oriented, probing at the users thought process. Interestingly there were no clear cases of the opposite effect, where a visualization performed particularly poorly in the task-based method but well in the insight method.

Furthermore, the insight method found additional important task categories that had not been considered for the task-based method, including the Topology+Expression and Outlier categories. Hence, the insight method revealed further differences between the visualization alternatives for those categories.

Table 9: Compares participants' performance on the selected tasks for both the methods. * indicates opposite conclusions about the visualizations between the two methods.

Tasks	Insight Category	Task-based study result	Insight study result
T1	Gene expression	M Tpts fastest.	M Graphs least insights.
T2	Condition	No differences	No differences
T3	Gene expression	No differences	M Graphs least.
T4	Timepoint analysis	M Tpts least accurate. M Graphs faster.	M Tpts least. 1 Tpt most.
T5	Timepoint analysis	M Tpts slowest. 1 Tpt more accurate.	M Tpts least. 1 Tpt most.
T6	Timepoint analysis	M Tpts slowest. M Graphs more accurate.	M Tpts least. 1 Tpt most.
T7*	Topology	M Graphs faster.	M Graphs least.

5.3 Empirical Data Analysis

The data analysis process for the task-based method was more straightforward in comparison to the insight method. It required the use of standard statistical analysis methods like ANOVA and paired t-tests. It took about 6-7 hours to finish the entire process, as the investigators had previous experience analyzing such data.

The data analysis process for the insight method is more complex. The amount of empirical data collected for the insight method supports richer analysis options. The participants' insights were analyzed first to find suitable categories to group the insights. The choice of categories can be dependent on the investigators' preferences and data understanding. A discussion was required between the investigators and domain biology experts to finally agree to a list. With meetings involved it took about 3-4 days to finish the data analysis. Thus in contrast to the task based method, data analysis for insight method is more complicated and subjective. It is possible that other analysts may have grouped the insights differently. For future work, a more generalized insight categorization such as [23] can be attempted. Also, a more in-depth analysis of insights is possible such as quantifying domain value of individual insights (as in [4]), but in this instance the domain experts only found the Summary category of insights to be more valuable, probably due to the novice experience level of the subject pool.

This effort is partly offset during the design phase of the task-based method by the need to design the benchmark task set to test. This requires time and subjectivity by the investigator to interview domain experts and decide on the most import task set.

5.4 Feedback about the Visualization Interface

Much more valuable user feedback was collected from the insight method, even though we did not require it.

Usability issues: Though both methods were conducted to evaluate visualization alternatives, the insight method required more interaction with the participants. The experiment protocol for the insight method required a closer observation of the participants' data analysis procedure and one-to-one interaction. This made it easier to notice if the participants were having any difficulties with the user interface. Also while performing data analysis in the insight method, participants commented about the visualization interfaces such as "the choice of color is weird", "the timepoint labels are difficult to understand", etc. Such valuable information was missed in the task-based method. For example, we also noticed in the insight method that participants using Single Timepoint visualization enjoyed the study because the visualization was more interactive in comparison to the other visualization alternatives. This may have prompted these participants to spend more time in the study as compared to the other participants.

Visual representations: Participants in the insight method provided more feedback about the visual representations of the graph. While analyzing the data, the participants would comment on the difficulties and suggest other data representation methods that they thought would support some of their data analysis tasks in a better way. E.g., the participants using Multiple Graph visualization commented that it was difficult for them to focus on a single gene only. The participants using Multiple Timepoints visualization commented that they were having trouble focusing on a single timepoint. They said that somehow the visualization was prompting them to focus on the overall node expressions, and that various interactions could be used to drill down.

5.5 Effect of Individual Differences

The task-based method produced less overall variance within conditions, and thus more statistically significant differences (e.g. Table 5 vs Table 7). The task-based method provides all the participants with an equivalent set of tasks. The list of tasks provides very specific direction to the participants throughout the study. This prevents the participants from getting confused about what to do next. Also, it makes the experience similar for most of the participants.

The insight method is open-ended. It is important for the study that the participants think aloud. It is possible that some participants are more communicative than others, and may report more insights as compared to other participants who may have actually had similar data insights but choose not to verbalize

them. Sometimes participants, depending on the type of visualization alternative they were using, felt that some insights were so noticeable that they may be too trivial and not worth reporting. Thus, findings from the insight study are more likely to be affected by the individual differences between the participants, causing higher variance in the results.

The participants in the insight method were suspicious of our intentions, and some asked if the data insights they were reporting made sense, or if they can be provided with more idea as to what they should be reporting so that they can be more helpful. When the participants in the insight method became confused, sometimes they needed to be encouraged to report insights. We would just say "yes, that makes sense". Some users required more prompting than the others. It may be helpful in the future to decide if the participants should be provided with such encouragement to make the study more uniform. A few participants reported that the entire study felt as if there was some catch involved to it. They thought there was either something that they were supposed to definitely notice, or that we wanted them to completely miss. At the end of the study, when participants were ready to leave, they wanted to know if they behaved as we expected them to or what was the point of the entire study.

5.6 Participant Motivation for Data Analysis

Unmotivated subjects were easier to recognize in the insight method. All the participants in the study were undergraduate biology students. To encourage participation in the study, they received some course credit. It is likely that some participants came only for the credit and were not motivated to perform data analysis. For the task-based method, it could be either lack of motivation or usability of the visual representation that can affect participants' performance (especially accuracy). For the insight method it was easier to notice such unmotivated participants because there was more communication with the investigator. The participants would often comment that they were tired or say "I just came from class, my mind is blank, please give me a minute to rest". We also noticed that participants who came during the weekend were more relaxed and interactive in the study, whereas participants who came during the weekdays were less inclined to spend as much time in the study. Potentially, since such unmotivated subjects can be recognized in the insight method, they could be filtered from the study so as to focus on a more realistic scenario. Motivational rewards could also be offered.

5.7 Conclusions about Visualization Alternatives

Table 10 summarizes conclusions for the visualization alternatives using both methods. Since the dependent variables for both methods are different, they provided different conclusions about the visualizations. The task-based method provides feedback in terms of accuracy and performance time. The insight method provides feedback based on the types of data insights the visualization generated. Since the tasks are pre-selected, they provide a more reliable feedback for the visualization in terms of the tasks. This allows designers to judge accurately if a visualization design *supports* a particular task or not. An unguided method provides feedback at a higher level, suggesting what kinds of data analysis a particular visualization method *motivates*. The fact that users may not perform certain tasks with it may not mean that the task is not supported, but that the visualization encourages the users to focus on other data analysis aspects.

Table 10: Comparison of the conclusions about the visualization alternatives from both evaluation methods.

Vis.	Task-based method	Insight-based method
1 Tpt	+ Consistent + Timepoint analysis	+ Consistent + Summary insights + Timepoint analysis + More time and insights
M Tpts	– Timepoint analysis + Single node analysis	– Timepoint analysis + Outlier nodes + Topology+expression
M Graphs	+ Timepoint analysis + Graph topology	+ Experiment conditions – Graph topology – Single node analysis

6. DISCUSSION

The insight method presented in [4] recognized several characteristics of an *insight* such as hypothesis generation, breadth vs. depth, directed vs. undirected, and domain value. For the data analysis discussed here, we decided to focus just on the category of data insights. Grouping insights by categories provided us with sufficient basis to compare the studies for the present discussion.

Also, the data and tools used here were more simplistic to reduce the learning time and allow users to complete the analysis in limited time. For real world data analysis scenarios, a data analyst spends much more time analyzing the data. The original data set from which the data for this study was selected was 45,001 rows X 72 timepoints and required about 3 months of data analysis by the bioinformaticians. The most important subgraph found after a few months of data analysis, and the associated time series dataset which was just 46 rows X 12 timepoints, was used in this study. Thus, though the short term studies provide important feedback and enable rigorous comparison of alternatives, they miss the amount of feedback provided by a longitudinal study [21, 24] for visualization tool usage. However, an advantage of the insight method is that it can be applied in a longitudinal study as reported in [21].

Most of the participants in this study were undergraduate biology students. For the insight method, at the end of the data analysis some participants were confident about the data analysis and could summarize the data or make hypothesis about the biological phenomenon suggested by the data. Such comments were ranked very high in the earlier insight study [4]. However, though the participants had biological background, they did not have enough familiarity with the specific immunity phenomena examined by this data set. Any such hypotheses were just speculations. They would not be able to judge the actual value of such findings. The data insights from the visualization tools ranked and evaluated by the actual data analysts [21] will be different than those by the actual user. An attempt at ranking the insights in this study by the experts resulted in most insights being rated at a similar value, so was not a useful measure in this case.

7. CONCLUSION

The study reported here was conducted to compare two empirical research methods for evaluating visualization alternatives. Since the dependent variables for both the methods are different, the studies were compared based on their results and on higher level criteria most relevant to evaluating visualization tools. A fundamental difference between the insight method and the task-based method is that the task-based method is more uniform across the participants both in terms of the user experience and the data collected from the experiment. The insight method, on the other hand, is more qualitative and thus involves some subjectivity to produce quantitative results. It is possible that

given a dataset two participants may analyze it in very different ways and report different insights. Hence, a higher level analysis such as grouping insights into categories or assigning domain value is needed, making the data analysis partly subjective. Through this, though, the insight-based method provides a way to capture a real world data analysis scenario and a wider range of comparison factors for the visualization.

Table 11: Comparison of the benchmark task and insight evaluation methods.

Comparison Factor	Task-based Method	Insight-based Method
Design Prep	• Prepare benchmark tasks • Better with simple data, tools, tasks	• Prepare problem scenario • Better with complex data and tools
Experiment Design	• Benchmark task protocol • Form based • Time & accuracy • Can be multiplexed • Short term study only	• Open-ended protocol • Think aloud • Capture insights • Interaction with user • Can be longitudinal [21]
User Tasks	• Determined by experimenter	• Determined by user
Participants	• Any users • Many users	• Expert, motivated users • Motivation is detectable • Training without biasing
Empirical Data Analysis	• Quantitative statistical analysis	• Coding rich insight and usability data • Very high variance • Longer analysis time
Primary Outputs	• Identify tasks supported by a visualization • Task efficiency; Perceptual, mechanical • Feedback on selected tasks only, ensures coverage of those tasks • Statistical differences	• Identify tasks promoted by a visualization • Learning efficiency; Cognitive, interactive • Detects new tasks, ignores unneeded tasks • User hypotheses, Summary task
Subjectivity	• Choice of tasks (ecological validity)	• Coding of insights and categories (repeatability)

There are several key findings between the evaluation methods in the comparison of the visualization alternatives:

Insight confirms task method: Many of the findings in the task method were confirmed, or even amplified, in the insight method. For example, both methods showed that 1 Tpt is the most successful and M Tpts is the least successful at timepoint analysis. This may provide some validation of the insight method to detect effects found by the task method.

Insight refutes task method: However, some findings were counter, indicating that users behave differently when not in the forced direction of a task-based method. Overall, the task method tended to favor the Multi Graphs visualization, while the insight method emphasized advantages of the Single Timepoints visualization. As specific example, even though participants performed the graph topology task fastest using the Multi Graph visualization in the task-based method, they gained the least insight about topology with Multi Graphs in the insight method. In fact, none of the Multi Graph users made any topology insights. Thus, because of its unguided protocol, the insight method may allow participants to miss certain type of tasks. The fact that participants did not gain topology insight does not mean that the task is not supported by the visualization, but indicates that the visualization does not provoke the participants to look for it.

Insight expands task method: Though the task-based method is more uniform, it provides feedback only on the selected tasks. Designing proper benchmark tasks is non trivial [13] and requires deep domain knowledge. For example, the insight study found that the Multi Timepoints visualization performed well at finding outlier nodes. We did not get this information from the task based method because we did not have benchmark tasks to reflect that potential insight category. The insight method offers the opportunity to discover important new task types from users.

8. ACKNOWLEDGEMENTS

We thank Peter Lee for developing the visualization tools, Vy Lam for providing data, and Dr. Joe Cowles and Dr. Carla Finkielstein for encouraging student participation in the study.

9. REFERENCES

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

[2] C. Chen and M. Czerwinski, "Empirical evaluation of information visualizations: an introduction", *IJHCS*, vol 53, pp 631-635, 2000.

[3] C. North, "Toward Measuring Visualization Insight", *IEEE Computer Graphics & Applications*, 26(3): 6-9, May/June 2006.

[4] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualization", *IEEE Transactions on Visualization and Computer Graphics*, Vol 11, No.4, 2005.

[5] Duggan, D., Bittner, B., Chen, Y., Meltzer, P., Trent, J. "Expression profiling using cDNA microarrays", *Nat. Gen.,* vol 21, 11-19, 1999.

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

[7] P. Saraiya, P. Lee, C. North, "Visualization of Graphs with Associated Timeseries data", *Proceedings of Infovis* 2005.

[8] Shneiderman, B. and Plaisant, C. "Strategies for evaluating information visualization tools: multi-dimensional in-depth long-term case studies." *Proc. of AVI BELIV 2006*, pg. 1-7.

[9] G. Golovchinsky, and N.J Belkin, "Innovation and Evaluation of Information Exploration Interfaces: A CHI98 Workshop." *SIGCHI Bulletin* 31(1):22-25, 1999.

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

[11] R. Kosara, C. G. Healey, V. Interrante, D. Laidlaw, C. Ware, "Thoughts on User Studies: Why, How, and When", *IEEE CG&A*, vol. 23, no. 4, pp. 20-25, July/August 2003.

[12] D. House, V. Interrante, D. Laidlaw, R. Taylor, C. Ware, "Design and Evaluation in Visualization Research", Panel, *IEEE Vis*, 2005.

[13] M. Tory and T. Möller, "Evaluating Visualizations: Do Expert Reviews Work?", *IEEE CG&A*,.25(5), pp.8-11, Sept./Oct. 2005.

[14] Steves, M.P., Morse, E., Gutwin, C., Greenberg, S., "A Comparison of Usage Evaluation and Inspection Methods for Assessing Groupware Usability", 2001 *ACM Conf on Supporting Group Work*.

[15] A.J. Brush, M. Ames, J. Davis. "A Comparison of Synchronous Remote and Local Usability Studies for an Expert Interface". Extended Abstracts, *CHI 2004*.

[16] Ester Baauw and Mathilde M. Bekker, "A comparison of two analytical evaluation methods for children's computer games" *Interact 2005 Workshop on Child Computer Interaction*.

[17] Jeffries, R., Desurvire, H. "Usability Testing vs. Heuristic Evaluation: Was there a Contest?", *SIGCHI Bulletin*, 24(4):39-41, 1992.

[18] H R Hartson, and T S Andre, "Criteria for evaluating usability methods", *IJHCI*, 2001, Vol. 13, No. 4, Pages 373-410.

[19] John, B., Marks, S. "Tracking the effectiveness of usability evaluation methods", *BIT*, 16, 4/5, 188-202, 1997.

[20] A. Doubleday, M. Ryan, M. Springett, A. Sutcliffe, "A Comparison of Usability Techniques for Evaluating Design", *DIS 1997*, 101-110.

[21] P.Saraiya, C.North, V.Lam, K.Duca, "An Insight-based Longitudinal Study of Visual Analytics", *IEEE TVCG*, 12(6): 1511-1522, 2006.

[22] STKE database, www.stke.org

[23] R. Amar, J. Eagan, and J. Stasko, "Low-Level Components of Analytic Activity in Information Visualization", *Infovis 2005*.

[24] J. Seo, B. Shneiderman, "Knowledge Discovery in High Dimensional Data: Case Studies and a User Survey for the Rank-by-Feature Framework", *IEEE TVCG*, Vol. 12, No. 3, 2006.