FilooT, a Visualization Tool for Exploring Genomic Data

Mahshid Zeinaly[a], Mina Soltangheis[a], and Chris Shaw[a]

[a] School of Interactive Arts and Technology, Simon Fraser University, Surrey, BC, CANADA

ABSTRACT

In order to enhance analysis of synthetic health data of the IEEE VAST Challenge 2010, we introduce an interactive Visual Analytics tool called FilooT designed as a part of the Interactive Multi-genomic Analysis System (IMAS) project. In this paper, we described different interactive views of FilooT: Tabular View for exploring and comparing genetic sequences, Matrix View for sorting sequences according to the values of different characteristics, P-value View for finding the most important mutations across a family of sequences, Graph View for finding related sequences and Group View to group them for further investigation. We followed the Nested Process Model framework throughout the design process and the evaluation. To understand the tools design capabilities for target domain analysts, we conducted a User Experience scenario-based study followed by an informal interview. The findings indicated how analysts employ each of the visualization and interaction designs in their Bioinformatics task-analysis process. The critical analysis of the results inspired design informing suggestions.

Keywords: Human Computer Interaction, Information Visualization, Visual Analytics, Visual Encoding and Interaction Design, User Experience.

1. INTRODUCTION

Prior to the existence of computers, advanced hand-drawn pictures in scientific publications shows that utilizing the human vision system was grounded in biology many years ago.[1] As the biological data-sets scales are increasing rapidly, custom software combined with manual intervention is replacing manual data analysis in biological sciences.[2] These computer-based visualization tools have enhanced our ability to communicate with the large amount of genomic data. Although classical visualization techniques[1] are used in the field of biology, researchers define new and creative ways to meet the target domain visualization needs.[3,4] Usually these tools are designed for a specific data-set/task-set in the domain. Advantages of these custom tools are twofold. First, they solve target analysts problems, which are part of the domain problems. Second, by analyzing the successful tools, researchers can eventually extract the target domains design guidelines and patterns. In this paper we introduce a custom-designed visualization tool to facilitate exploration of a virus gene family of sequences. This tool can assist health investigators in finding the relation of the gene substitutions to disease characteristics. This biology-specific tool also helps analysts to understand how characteristics of the disease relate to virus strain mutations. To design such systems, we applied the principles of Information Visualization as acknowledged in the literature.[1,2,4] To the best of our knowledge Interactive Tabular View slider-bar (filtering), as well as defining two modes for the system (Row mode and Column mode) is completely novel. The domain user participants showed great interest in the Tabular View slider-bar. They mentioned that filtering option does not exist in available tools and it could enhance their work greatly.

2. LITERATURE REVIEW

Once a virus infects a host, it makes copies of itself and spreads to other people. During this replication process, typically some substitutions appear in genetic sequence.[5] One way for characterizing DNA is to compare their sequences with each other.[6] In bioinformatics, Multiple Sequence Alignment helps to compare more than a pair of sequences and to find the similar regions between them.[6] For representing Multiple Sequence Alignment, there are two types of visualizations: the Sequence Logo[7] and Multiple Sequence Alignment viewers.[8,9] Our focus is on Multiple Sequence Alignment view which is a table. Each tables row corresponds to a sequence and each column is a position in all the sequences. Each cell represents a DNA letter in each sequence.[1] The goal of this view is to show the variations in the sequences to the analyst. There are different tools available for sequence analysis. IMAS[9] is a visual analysis tool for rapid analyses of DNA sequences. This tool visualizes the output of common bioinformatics tools such as BLAST and Clustal-W in a unified framework with semantic zooming navigation.[9, 10] Jalview is one of the most commonly installed tools. Its Multiple Alignment view is capable of hiding and grouping multiple sequences (rows).[8] It also allows the user to sort the sequences with different criteria. Historically, the table view provides interactive features to allow the users to gain insight about the data.[8, 9] Sequnce-Juxtaposer[11] is an example of applying Focus+Context in bioinformatic sequences alignment explorations. The multiple alignment views usually accompanied by another metadata matrix view which is a kind of Table Lens where each column contains information about one metadata and each row represents the value of that metadata for each strain.[12–16] Freire et al.[14] used horizontal bars for encoding each cells data, as well as vertical bars on top of each column to show overall column distributions. Others, such as Sopan et al.,[15] used colour saturations in different cells to encode their values. The colour saturation usually is a better choice for visualizing the information in a cell as it has higher accuracy for encoding ordered data.[16] For representing a hierarchical structure in data tree is a common visualization technique.[17] In our tool we also utilized prior knowledge produced by solutions that has been proposed for solving VAST challenges 2010. The Noblis Team[17] used sunburst layout to represent the evolutionary tree of the current outbreak sequences and utilized colour to represent the degree of the overall danger level of the sequences. Freire et al.[14] used the basic Node-Link layout for the evolutionary tree information. And we adopted the same representation in FilooT. For other types of relationships between data items, the Network representation is used.[3] ManyNets[14] is a network visualization tool with tabular interface in which its tabular view was a kind of Table Lens that the disease characteristics were shown in columns. This tool enables users to create a new column with the existing characteristics and sort all the rows according to the values of the particular characteristics associated to a column. GeneTracer[18] provided three views: Gene Sequence view, Disease Characteristic view, and Graph view. Disease Characteristic view used a Table Lens where each column had a different colour and each cell had different saturations of that colour. The Graph View visualized the relations among the sequences via a Minimum Spanning Tee representation where the weight of an edge is the Hamming distance between the two sequences.[6] Wood et al.[19] utilized a heat map for Table Lens representation. It had two levels of sorting of the rows according to one characteristic, and sorting them again within each category of the first sort, according to the second characteristic.

3. DESIGN

The primary interface for FilooT is shown in Figure 1.

3.1 Interactive Tabular View

The Tabular view (see figure 2) is an interactive visualization for exploring genetic sequences. The first row represents the genetic information about the original sequence. The second row shows position numbers and numbers start from one and end with the length of the sequences. Each of the subsequent rows indicates one sequence. Each cell contains the result of the comparison of each sequence with the original sequence appeared in the first row. The purple colour is used to represent those cells that did not change in comparison with the original sequence and the yellow colour highlights cells with a change in a particular row and column. The letter indicates a change in the information of the specific cell in comparison to the corresponding column in the original sequence. This view supports the following user interactions:

Navigation: The horizontal and vertical scroll bars at the bottom and on the right are so that the user can explore more of the sequences and the positions data.

Zoom: The "+" and "-" buttons allow user to zoom in and out in order to control the level of detail.[10]

Re-ordering: To compare different columns with each other, placing the columns close to each other frees up the cognitive load of the users to focus on their desired task.[20] One way of putting columns close to each other is to allow the user to drag and drop the columns next to each other. However, the natural order of nucleotides in a sequence is meaningful Therefore the "reset" button returns the columns to their original sequence from one to the length.

Filter: Tabular view provides two filtering capabilities; a) **Basic Filtering:** the user can separate out a group of columns (or one column). The transition between hidden/ unhidden state is animated so that the view does not jump to a new state. b) **Augmented Filtering:** While having basic filtering seems useful for exploring the data, finding relevant columns still requires manual work (exploring all the columns to find relevant ones).

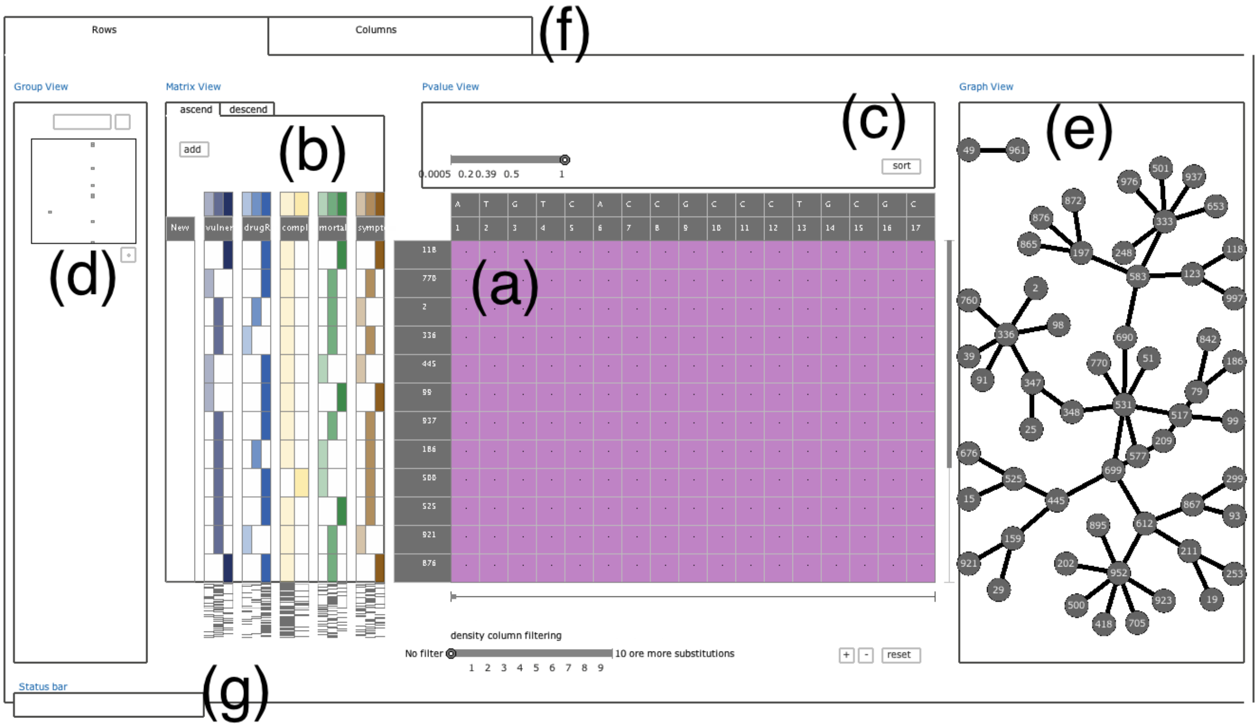
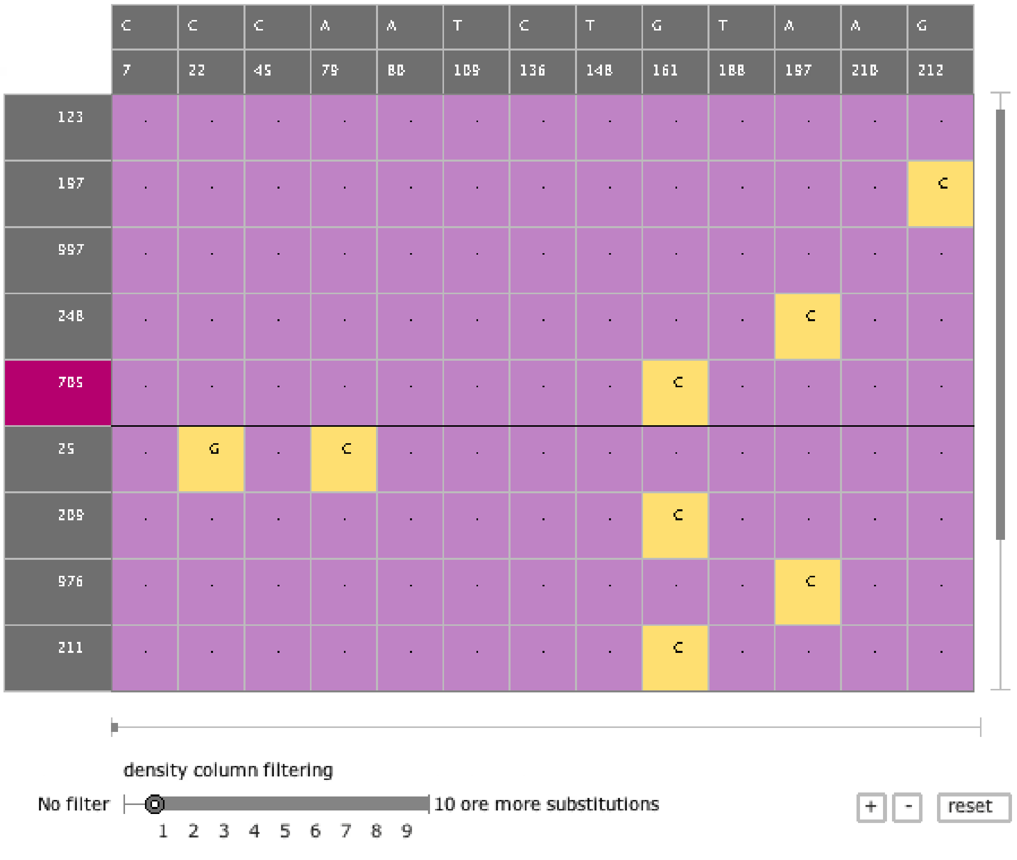


Figure 1: FilooT overall picture. FilooT visualization system consists of (a) an interactive visualization table to represent the genetic sequence information (b) a matrix visualization for interacting with the disease characteristics data (c) The P-Value bars to show a metric (reverse of P-value in Mann- Whitney U test) about each column (d) The Group View containing the user created groups along with an overview of each group (e) a graph visualization for representing row (or column) relationships depending on the system mode (Row based or Column based) (f) two buttons enable the user to choose between the Column and Row mode (g) the Status bar is being updated after each action that the user makes.

Moreover, a small number of substitutions in a column may occur randomly and do not reveal any valuable information to the analysts. Therefore, an augmented filtering excludes the columns that have fewer yellow cells than the filter number. These interactions affect the other views linked to Tabular View.

3.2 Matrix View

Matrix View (see figure 3) enables the user to sort the rows according to the values of different characteristics (for example a disease characteristic such as severity). Design of this view is inspired by the Table Lens.[14] In table lens, the levels are shown by the length of horizontal bars or colour saturation per cell.[15] However, we utilize position and redundantly colour saturations to encode the same property of the data. Each column is divided by the number of its characteristics levels. The coloured label on top of each column shows the different levels in that particular column. The darker the colour is, the higher the level of the characteristics. The coloured labels are placed from right to left respective to color saturation level. We exploit position channel for representing discrete ordered data-type because it is the most powerful visual property for encoding all kinds of data.[16] In addition, the colour saturation is a better alternative for the length channel for encoding this ordered information.[16] We also used hue to separate different characteristics that are nominal data and the hue channel is appropriate for separating different categories.[16] The user can perform the following list of interactions in the Matrix View:
Sort: The rows can be sorted ascending or descending according to the values of a selected column header.
Aggregation: The add button enables the user to make a new column by combining the existing ones with a simple mathematic function in between them.
Zoom: The user can zoom in and out to the view using + and - buttons from Tabular View.



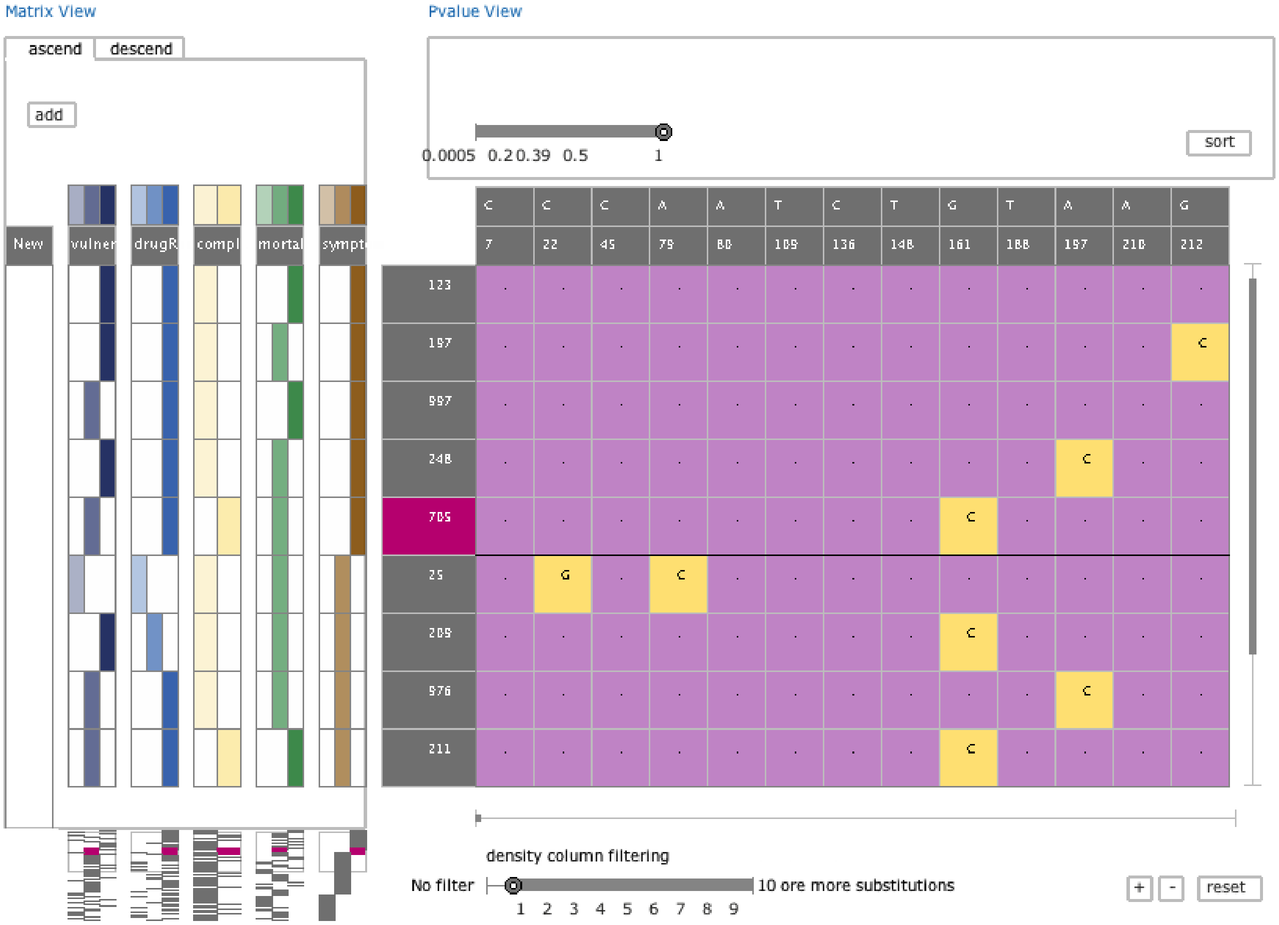
	C	C	C	A	A	T	C	T	G	T	A	A	G
	7	22	45	79	88	109	136	148	161	188	197	210	212
123	.	.	.	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	.	.	.	.	.	.	.	.	C
997	.	.	.	.	.	.	.	.	.	.	.	.	.
248	.	.	.	.	.	.	.	.	.	.	C	.	.
705	.	.	.	.	.	.	.	.	C	.	.	.	.
25	.	G	.	C	.	.	.	.	.	.	.	.	.
209	.	.	.	.	.	.	.	.	C	.	.	.	.
976	.	.	.	.	.	.	.	.	.	.	C	.	.
211	.	.	.	.	.	.	.	.	C	.	.	.	.

Figure 2: Interactive Tabular View. The features including scrollbars, filtering, zooming buttons, reset buttons, and mouse highlighting.

Overview: At the bottom of each column in Matrix View, an overview of that specific column is provided so that the user can see the pattern of the change for all the row values for that specific disease characteristics column, without the need to zoom. When the Row mode is activated, and a sequence header is highlighted to show the mouse position, it also highlights a row in the overview of Matrix View.

3.3 P-Value View

There is a pattern within some of the columns that makes them interesting candidates to form new hypothesis.[17] This pattern suggests a relationship between substitutions in a particular column and one of the characteristic of the rows. As humans do not complete pattern-detection tasks very well,[21] we cannot rely on them to find this pattern in columns. Commonly biologists use metrics to detect interesting patterns. Mann-Whitney U tests p-value is one of the metrics used for finding relevant positions.[17] Using the Mann-Whitney U test, the severe rows can be separated from others by splitting all the rows into two groups based on the existence substitutions in them. The negative of the logarithm of the P-Value suggests likeliness of the significant difference between the two groups. This value is shown by the bar lengths in P-value View (see figure 4) to help users find relevant columns. We used length to represent the p-value metric because the length channel is the second most powerful channel for encoding the ordinal values.[16] The P-value view also provides the filtering feature. This feature enables the user to filter out any column where the length of the bar is smaller than the filter number. This view also lets the user sort the positions based on the bar length. The columns will be sorted from high to low and placed from right to left. In general, assuming that sorting all the rows are sorted according to one of the characteristics from top to bottom, a significantly larger proportion of substitutions appear at the top rather than the bottom. As the user might want to focus on those columns with the higher bar length, merely hide/unhide all the other columns is not efficient. Instead, it would be more productive to sort the columns



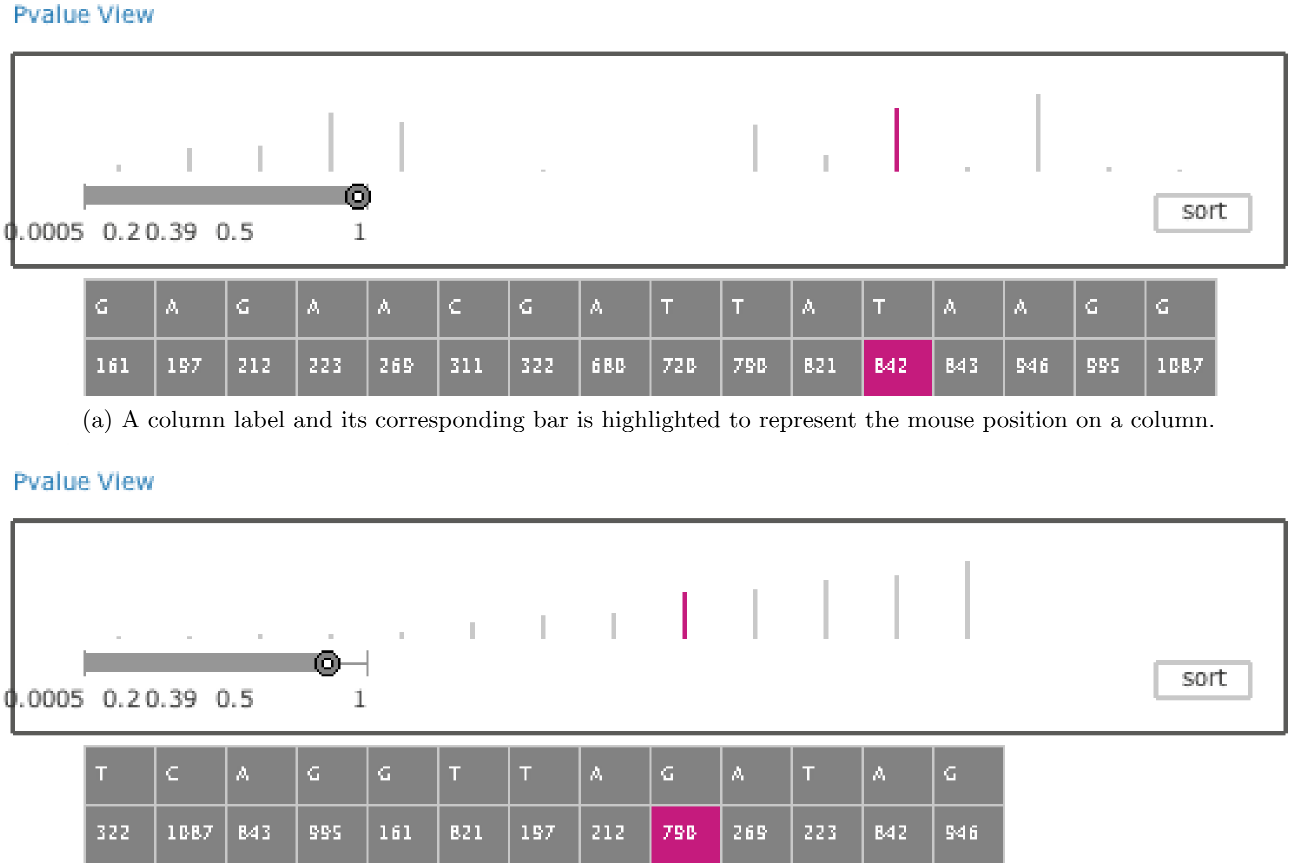
	C	C	C	A	A	T	C	T	G	T	A	A	G
	7	22	45	79	88	109	136	148	161	188	197	210	212
123	.	.	.	.	.	.	.	.	.	.	.	.	.
197	.	.	.	.	.	.	.	.	.	.	.	.	C
997	.	.	.	.	.	.	.	.	.	.	.	.	.
248	.	.	.	.	.	.	.	.	.	.	C	.	.
705	.	.	.	.	.	.	.	.	C	.	.	.	.
25	.	G	.	C	.	.	.	.	.	.	.	.	.
209	.	.	.	.	.	.	.	.	C	.	.	.	.
976	.	.	.	.	.	.	.	.	.	.	C	.	.
211	.	.	.	.	.	.	.	.	C	.	.	.	.

Figure 3: Left: The Matrix View. Matrix View and Tabular View are linked together by shared row labels. Consequently, when the rows positions are changed in one view, for example if the user sorts the rows, their vertical positions will be changed in the other view accordingly.

based on the reverse of the p-value (The length of the bars). Keeping the bars on top of the columns in Tabular View allow the user to go over the bars while observing the columns pattern. The Tabular view and the P-Value view are linked so that if the user reorders the positions in one view, the corresponding columns order will be changed in the other. Also they can use the reset button to go back to the original domain ordering.

3.4 Group View

Group View enables the users to create different groups of columns or rows, to separately load each group into the views to investigate the group information, and to focus on the relationships between the columns/or rows. It is more likely that the users make these groups from the relevant columns/rows however they have to find the related columns/rows manually (Basic Grouping, as oppose to Augmented Grouping that users are guided to find related columns/rows more effectively). Basic grouping feature is defined for both columns and rows. Therefore row and column mode is defined to separate the behaviour of the system.
Column Grouping: This feature would let the user create and add different groups from a combination of different columns, see an overview of the group and its general pattern.
Row Grouping: This feature is built so that in row mode the user can select different rows to make a group of them and all the features are similar to column grouping. When the user selects a group among the previously created groups from the Group View, the chosen groups data will be uploaded into the system. Therefore, the



G	A	G	A	A	C	G	A	T	T	A	T	A	A	G	G
161	197	212	223	269	311	322	680	720	790	821	842	843	946	995	1087

(a) A column label and its corresponding bar is highlighted to represent the mouse position on a column.

T	C	A	G	G	T	T	A	G	A	T	A	G
322	1087	843	995	161	821	197	212	790	269	223	842	946

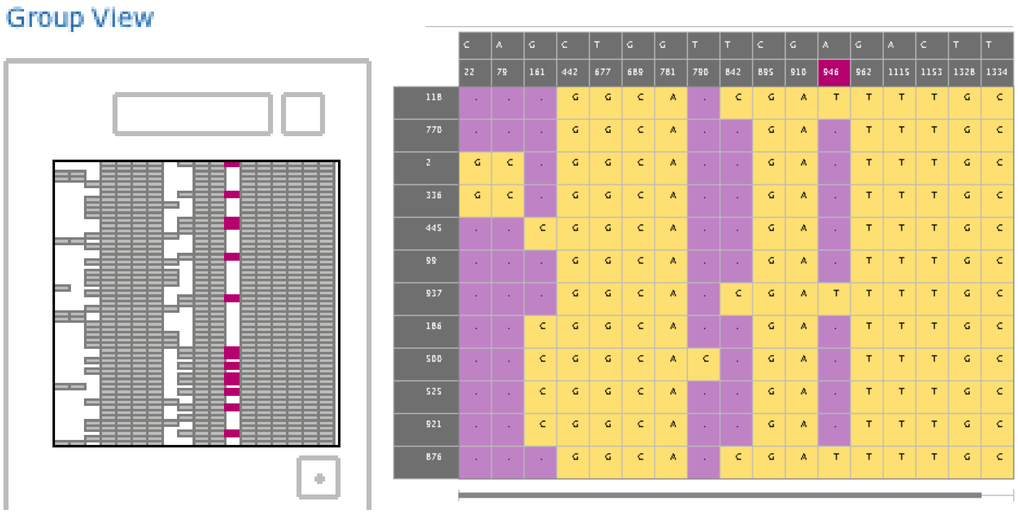
(b) The columns are sorted from left to right. The user also filtered some columns with smaller bars.

Figure 4: P-value View

data in the Tabular View matches the data in the selected group. There is a predefined group in each Row and Column mode that contains the entire data set for the user to be able to go back to the original step.
The Overview feature consists of a larger window than Tabular View (prior to zooming) that allow analysts to see the big picture, and identify clusters, trends and outliers that may be candidates for detailed inspection.[15]

3.5 Graph View

Graph View is a node-link representation that visualizes the relationship between the columns (or rows) and helps users to find related columns (or rows) and group them together to focus on fewer rows (less data dimensions) for future analysis. the Graph view is motivated by the augmented grouping in which the users can more effectively detect the columns (or rows) of the same group because the system highlights the relationships between columns (or rows) to guide group creation; In contrast to the Basic Grouping where groups are created by the users based on their prior knowledge or observations. Grouping feature is defined in both row and column mode:
Column Relations: Between any pairs of columns, two kinds of relationship are supported: Complementary patterns and Correlation. The correlation pattern between a pair of column means that the substitutions in both columns appear in the same rows. On the other hand, the complementary pattern between two columns means that the substitutions of the two columns appear in the opposite rows and wherever there is a substitution in one column, there is no substitution in the other one. There are several alternatives representations for column relations in this view; One alternative representation for relationship between a pair of columns is the matrix visualization.[3] One benefit of using this matrix is that, by re-arranging the rows and columns, some interesting patterns would be revealed. However, it requires a large screen space and we cannot eliminate the cells with 0 correlation from the space. The second option is a node-link graph, where there is a link between a pair of



	C	A	G	C	T	G	G	T	T	C	G	A	G	A	C	T	T
	22	79	161	442	677	689	781	790	842	895	910	946	962	1115	1153	1328	1334
118	.	.	.	G	G	C	A	.	C	G	A	T	T	T	T	G	C
770	.	.	.	G	G	C	A	.	.	G	A	.	T	T	T	G	C
2	G	C	.	G	G	C	A	.	.	G	A	.	T	T	T	G	C
336	G	C	.	G	G	C	A	.	.	G	A	.	T	T	T	G	C
445	.	.	C	G	G	C	A	.	.	G	A	.	T	T	T	G	C
99	.	.	.	G	G	C	A	.	.	G	A	.	T	T	T	G	C
937	.	.	.	G	G	C	A	.	C	G	A	T	T	T	T	G	C
186	.	.	C	G	G	C	A	.	.	G	A	.	T	T	T	G	C
500	.	.	C	G	G	C	A	C	.	G	A	.	T	T	T	G	C
525	.	.	C	G	G	C	A	.	.	G	A	.	T	T	T	G	C
921	.	.	C	G	G	C	A	.	.	G	A	.	T	T	T	G	C
876	.	.	.	G	G	C	A	.	C	G	A	T	T	T	T	G	C

(a) An overview of a pattern in a column in Tabular View is shown in Group View.

(b) Tabular View

Figure 5: A group of columns in Group View helps the user to analyze patterns within his/her selected columns.

columns only if their correlation is non-zero. The link is coloured blue for correlation (numbers greater than 0) and red for complementary (numbers less than 0). Colour saturations and line weights are also redundantly used to encode the same information. As there are a considerable number of columns with zero correlations, this option conserves the space better than the table representation. Given that the substitutions in data-set are represented by values 0 and 1, one may define the similarity between two columns with measures such as Pearsons correlation calculation for any pair of column. This is however not optimal because many zeros (no substitution) in the columns result in a correlation close to 1 indicating they are highly correlated however it is not the case. To alleviate this problem, we propose to use a new measure which ignores the common zeros between columns in equation 1. Assuming that two columns, X and Y, each have n members, X = x1, ...,xn and Y = y1, . . . , yn. The measure is defined as follows:

$$M(X,Y) = \sum_{i=1}^{n} x_i y_i, - \sum_{i=1}^{n} x_i \oplus y_i \tag{1}$$

Where $\oplus$ is the logical XOR operation. This measure ignores entries with no substitution in columns, increases when entries with substitutions occur together and decreases when substitution complements each other. Given that, both positive and negative values are expected. This view is also linked to Tabular View; if the user deletes a row from theTabular View, the matching node in the Graph View will be deleted as well. Moreover, when the user filters the columns with different number of substitutions or by the p-value measure, the corresponding nodes will be filtered. There is also another filter mechanism built into the view that removes the columns based on the strengths of their connection (see figure 6).

Row Relations: The relations between rows are hierarchical. The existed Graph View is used to make a Tree for the representation of this relationship. Some of the submissions of VAST 2010[18,22] used the Minimum Spanning Tree for constructing the evolutionary tree. The weight of the edges was the Hamming distance between the two nodes.[6] The Hamming distance was the number of positions that differed in any two rows that implied the required number of changes to transform one sequence to the other. The Minimum Spanning Tree is a tree in a graph that connects all nodes (rows) and its total edge weight is the minimum of total edge weights of all the possible trees.

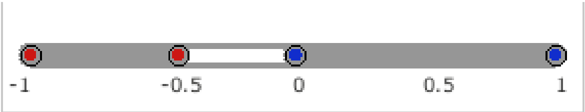


Figure 6: Graph View has two filters to delete the columns in Column mode from the column graph.

4. EVALUATION

To evaluate our tool we followed Nested Model Process (NMP) which was also used for Visualization Design of the tool. Our work falls into data/operation abstraction, and encoding/interaction technique design layers of this framework.[23] For evaluation at the visual encoding/interaction layer, NMP provides two recommendations: a) The design needs to follow perceptual and cognitive principles.[23] We used heuristics for information visualization to meet this requirement.[24] b) The design should be able to communicate with the analyzer and be useful towards solving their problem[23] .Therefore our general research question was set to: How the design of FilooT could help the domain users solve the tasks problems. Considering the formative nature of our research question, we took User Experience scenario (UE) to understand "what do our target users think of the visualization?"[25] However, our evaluation methodology is user-based in which problems are found through the observation of and interaction with users while they use or comment on the interfaces.[25, 26]

4.1 Method

We conducted a qualitative, User Experience study for system evaluations[25] where users played with the system to answer some tasks. The participants were asked to talk freely about anything comes to their mind about the tool during and after the process. The study follows by an informal interview with open-ended questions to understand users opinion about the tool. The followings are the interview guide questions. a) How did each feature help you to find the answers? b) Did you have any difficulties understanding any of the presented information in any of the views? c) Please explain your steps of finding the answers. d) Give us your feedbacks to make each of the views better. Your suggestions may include your ideas of new interactions or refinement of some parts of the design. e) Are there limitations of the current system that would hinder its adoption?

4.2 Participants

We had four participants over the age of 19 who are undergraduate/graduate students or postdoctoral researchers in the lower mainland. They are required to be familiar with DNA multiple sequence alignment concepts and they should be interested in this study. Therefore our participants had an academic background in either Bioinformatics, Computational Biology or Biological Physics.

4.3 Procedures

The experiment took less than 1.5 hours. The studies took place in the lab environment on a laptop computer. The participants read and signed the consent form and filled out a pre-study about their familiarity with the domain and their experience with similar tools. Then they were asked to read the Task/Data Description and they were trained for 10 minutes to learn to use the basic features of the software. After that, they used the system for 30 minutes and they were encouraged to write down their findings, think aloud and express their thoughts, concerns and their questions at any time. The experimenter used and paper to take notes from observations of participants use of the tool. After the 30 minutes, in a semi-structured interview participants were asked about their experience with the tool. At the end, they were thanked for the participation and they received the compensation by signing the compensation form.

4.4 Data

We use a synthetic data-set for the VAST Challenge 2010, Mini Challenge 3.[27] It is about an arms dealing scenario in which one of the dealers, Nicolai, died in a hospital with symptoms consistent with Drafa Fever. To help Public health organizations to develop pandemic response plans, analysts seek for more information about the disease. The data-set consists of 56 strains of a particular original virus spread over time to different infected people. Each strain has a gene sequence of 1400 nucleotides with one or more nucleotide changes from the original viruss sequence. There is also information about some characteristics for each of the evolved viral strains.

4.5 Tasks

In this health data-set, signs and symptoms of the Drafa virus are varied and humans react differently to infection. Some mutant strains from the current outbreak have been reported as being worse than others for the patients that come in contact with them. Our three selected tasks for understanding how the domain users interact with the system and use the visualizations to sole their domain problems are as following:

- Task 1: Identify mutations that lead to an increase in symptom severity (a disease characteristic). For example, C → G, 456 shows C changed to G at position 456.
- Task 2: Identify mutations that lead to the most dangerous viral strains.
- Task 3: Nicolai has a strain identified by sequence 583. One patient has a strain identified by sequence 123 and the other has a strain identified by sequence 51. Which patient contracted the illness from Nicolai?

Note that we select the three above tasks among a benchmark data-set/task-set[27] and we assume that the tasks are already validated and they reflect the target domain users work.

4.6 Limitations

The given time can interfere with participants solving task. However we explained the users that the problem solving process is more important than reaching to answers and their accuracy.

5. RESULTS

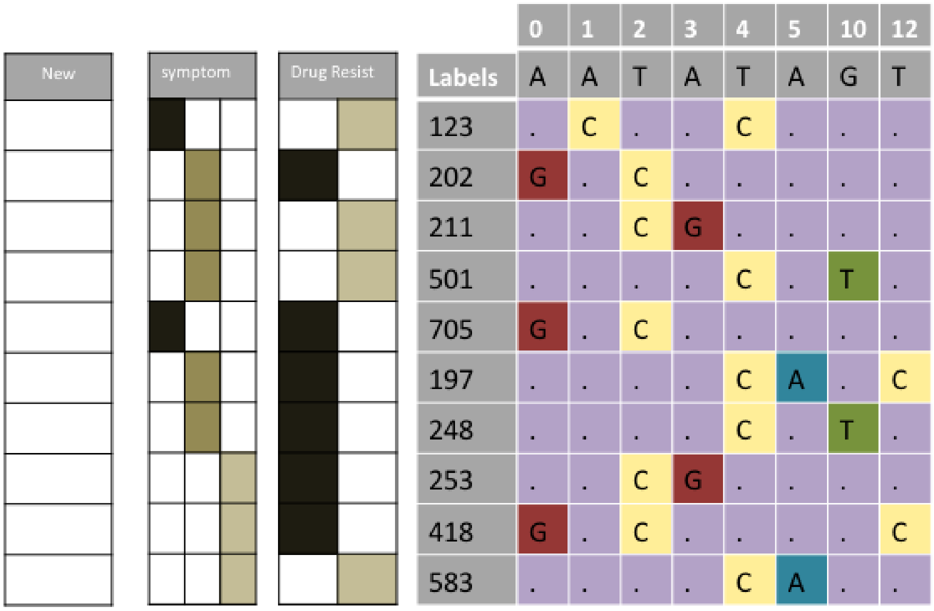
The results of the user study led us to create a list of requirements which informed design of the tool as well as develop a guideline for future work. In this section, we provide a summary of the comments and observations for each view. More details can be found in.[28]

5.1 Tabular View Visualization Comments

According to users suggestions and our observations, the slider bar was the most usable feature. Also it was suggested to use different hues to show different nucleotides. However depending on the number of columns in the Matrix View, this could interfere with the hue used in that view and using more than eight hue colours on screen is not recommended.[29] Instead, we suggest using only one hue per column for Matrix View and distinguishing between columns by adding extra space between them (Figure 6 shows this idea). More specifically we suggest using Proximity to show the organization of inter and intra columns and prevent hue overloading. In addition, users in case of working with authentic data-sets need to know if a column is related to its neighbours and what the nature of that relationship is. Examples of these relationships include codon and motif information in data-sets. However, because the VAST Challenge stated that the DNA is non-coding, codon analysis and AA sequence analysis cannot be considered. All of the users were familiar with at least one Visual Analytics tool similar to the Tabular View. Although some of the interactions that they used to see in similar tools were not relevant to the study tasks, the users desired to see all of the familiar features in Tabular View.

5.2 Matrix View Comments

Matrix View is one of the most frequently used views. In this view, clicking on the coloured label at the top of each column and will sort all the rows according to the values of that columns characteristics which users enjoyed as a cool feature. However, when the user clicks again, the rows order might be changed because the sorting algorithm allows for sequences that are different but still correct. Therefore, although the rows will be sorted each time a user clicks on a coloured label, the rows order within a level might be changed. The user did not like this and preferred consistency and predictability in the column sorting behaviour. The labels could be made more useful if the different levels were clickable separately so that the system would jump to a state in which the Tabular View and the Matrix View contained the rows with the selected level. Moreover, the Matrix View could have a built-in option to keep track, and sort the next column based on previous selections; specially when the analyst had certain priorities of columns and want to see a level of disease characteristic in that priority list.



	0	1	2	3	4	5	10	12
Labels	A	A	T	A	T	A	G	T
123	.	C	.	.	C	.	.	.
202	G	.	C	.	.	.	.	.
211	.	.	C	G	.	.	.	.
501	.	.	.	.	C	.	T	.
705	G	.	C	.	.	.	.	.
197	.	.	.	.	C	A	.	C
248	.	.	.	.	C	.	T	.
253	.	.	C	G	.	.	.	.
418	G	.	C	.	.	.	.	C
583	.	.	.	.	C	A	.	.

Figure 7: Increasing the number of hue in Tabular View led to using limited number of hue in Matrix View. Instead of Hue, Proximity is used to differentiate columns

5.3 P-Value View Comments

Users tend to use P-value slider bar instead of the Tabular Views Slider bar in order to filter out the columns that have mutations in all the rows. In addition, users used the slider bar in Tabular View in conjunction with the P-Value sider to understand the reliability of the data. Columns that contain a lot of mutations tend not to be informative, because it is ambiguous what the original nucleotide should be. Alternatively, columns with high mutation rates could be the result of a data error, so users would use the Tabular View table to observe columns containing a lot of mutations (yellow cells), and would use the P-Value slider to filter these out.

5.4 Graph View Comments

This view was useful for tasks 1 and 2, but not as much as the P-value View. Users commented that the tree structure could show the row clusters better if the node position changed so that each node correlated to its matching row in Tabular View. Having such layout would enable the user to sort the rows based on clusters that appear in the tree structure. Moreover, it can be helpful to select multiple nodes in Graph View by holding down the ctrl key and highlighting the matching information in all the other views. Also the users expected that selecting a node would scroll the Tabular View to contain the matching row (or column) and deleting a Tabular View's column would update the corresponding tree structure. In addition, the metrics used for calculating the column relationships in the Graph view were based on the information available in previous VAST Challenge submissions, however, our domain users preferred existing Bioinformatics metrics.

5.5 Group View Comments

Users could detect the pattern using the overview feature. However to get an overview they would have preferred simply to zoom out within the Tabular View since all the sequences are already depicted there. A contributing design factor to the low use of the overview feature was that the button to activate it was hard to find. User created a new group of rows as an alternative for filtering the unwanted rows. We think this is because the new group of rows could be further filtered and interacted with independently of any previous groups that may have been created. This is an example of what is called the "sandbox".[30] The users wanted to separate their analytical ideas into independent subsets that could be independently controlled and filtered, and the group view provided this.

5.6 Row/Column Mode Comments

Although the users liked the idea of having two viewing modes, they wanted to be able to go across rows and columns and use checkbox to select both Row and Column mode at the same time instead of a tabbed pane (or buttons). However, the user could create a new group only by selecting rows or columns, and then they could load the group and try the other mode to go across both rows and columns.

5.7 General Comments on FilooT Design

The users also suggested adding scripting languages for direct data access as well as a help system, especially for P-value and Graph View, to explain the graph relations, weights and bar meanings.

6. CONCLUSION AND FUTURE WORK

We highlight some of the observations and users comments and suggestions which can be considered as an agenda for further research. We had two metrics for the tool: One for making P-value View bars and one for making Graph View edge weight. The first one seems acceptable to the participants, but the second still needs to be checked by domain experts. Although the colours of Tabular View are checked with VisCheck website, it seems that the system still needs colour checking for highlighting caused by mouse overs. Overviews need to show the entire sequence rather than just a larger window. To guarantee the visibility of all rows without overlaps, there should be at least one vertical overview pixel per row. If there are more rows than available pixels, some aggregation is needed. To present more than one distribution in one vertical pixel, we can calculate the minimum, average or maximum of the intensity values for that pixel.[15] One of the critical contributions of this work is several filtering options for users to exclude irrelevant data to solve the study tasks. The filtering option allows the user to work with a subset of data set. Also to the best of our knowledge having two modes for the system (row and column mode) as well as the Tabular View slider bar is novel. Based on the users comments the most useful feature is the design of the slider bar in the Tabular View that filters those columns with less information it can be either columns that are not informative because of significantly high (or low) number of mutations.

REFERENCES

[1] Procter, J. B., Thompson, J., Letunic, I., Creevey, C., Jossinet, F., and Barton, G. J., "Visualization of multiple alignments, phylogenies and gene family evolution," *Nature methods* **7**, S16–S25 (2010).

[2] Joachimiak, M. P., Weisman, J. L., and May, B. C., "Jcolorgrid: software for the visualization of biological measurements," *BMC bioinformatics* **7**(1), 225 (2006).

[3] Heer, J., Bostock, M., and Ogievetsky, V., "A tour through the visualization zoo.," *Commun. ACM* **53**(6), 59–67 (2010).

[4] Nielsen, C. B., Jackman, S. D., Birol, I., and Jones, S. J., "Abyss-explorer: visualizing genome sequence assemblies," *Visualization and Computer Graphics, IEEE Transactions on* **15**(6), 881–888 (2009).

[5] Pray, L., "Dna replication and causes of mutation," *Nature Education* **1**(1) (2008).

[6] Jones, N. C. and Pevzner, P., [*An introduction to bioinformatics algorithms*], MIT press (2004).

[7] Schneider, T. D. and Stephens, R. M., "Sequence logos: a new way to display consensus sequences," *Nucleic acids research* **18**(20), 6097–6100 (1990).

[8] Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., and Barton, G. J., "Jalview version 2a multiple sequence alignment editor and analysis workbench," *Bioinformatics* **25**(9), 1189–1191 (2009).

[9] Shaw, C. D., Dasch, G. A., and Eremeeva, M. E., "Imas: the interactive multigenomic analysis system," in [*Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*], 59–66, IEEE (2007).

[10] Dunsmuir, D., Baraghoush, M. Z., Chen, V., Joorabchi, M. E., Alimadadi, S., Lee, E., Dill, J., Qian, C., Shaw, C., and Woodbury, R., "Czsaw, imas & tableau: Collaboration among teams: Vast 2010 grand challenge award: Excellent student team analysis," in [*Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*], 267–268, IEEE (2010).

[11] Slack, J., Hildebrand, K., Munzner, T., and John, K. S., "Sequencejuxtaposer: Fluid navigation for large-scale sequence comparison in context.," in [*German Conference on Bioinformatics*], **53**, Citeseer (2004).

[12] Rao, R. and Card, S. K., "The table lens: merging graphical and symbolic representations in an interactive focus+ context visualization for tabular information," in [*Proceedings of the SIGCHI conference on Human factors in computing systems*], 318–322, ACM (1994).

[13] Vehlow, C., Heinrich, J., Battke, F., Weiskopf, D., and Nieselt, K., "ihat: Interactive hierarchical aggregation table," in [*Biological Data Visualization (BioVis), 2011 IEEE Symposium on*], 63–69, IEEE (2011).

[14] Freire, M., Plaisant, C., Shneiderman, B., and Golbeck, J., "Manynets: an interface for multiple network analysis and visualization," in [*Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*], 213–222, ACM (2010).

[15] Sopan, A., Freier, M., Taieb-Maimon, M., Plaisant, C., Golbeck, J., and Shneiderman, B., "Exploring data distributions: Visual design and evaluation," *International Journal of Human-Computer Interaction* **29**(2), 77–95 (2013).

[16] Mackinlay, J., "Automating the design of graphical presentations of relational information," *ACM Transactions on Graphics (TOG)* **5**(2), 110–141 (1986).

[17] Campbell, C., Blanchard, S., Chin, S., Henderson, C., Holland, M., Jennings, K., Kuehl, P., Lucey, D., McCoy, M., McCracken, J., et al., "Multi-viz data fusion vast 2010 grand challenge award: Outstanding debrief," in [*Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*], 265–266, IEEE (2010).

[18] Lee, H., Choo, J., Görg, C., Shim, J., Kihm, J., Liu, Z., Park, H., and Stasko, J., "Genetracer: Gene sequence analysis of disease mutations," 291–292 (2010).

[19] Wood, J., Slingsby, A., and Dykes, J., "Designing visual analytics systems for disease spread and evolution," 285–286 (2010).

[20] Shirley, P., Ashikhmin, M., Gleicher, M., Marschner, S., Reinhard, E., Sung, K., Thompson, W., and Willemsen, P., [*Fundamentals of computer graphics*], AK Peters Natick (2002).

[21] Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L., and Zhou, Y., "Treejuxtaposer: scalable tree comparison using focus+ context with guaranteed visibility," *ACM Transactions on Graphics (TOG)* **22**(3), 453–462 (2003).

[22] Freire, M. and Sopan, A., "Gene similarity uncovers mutation path vast 2010 mini challenge 3 award: Innovative tool adaptation," in [*Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*], 287–288, IEEE (2010).

[23] Munzner, T., "A nested model for visualization design and validation," *Visualization and Computer Graphics, IEEE Transactions on* **15**(6), 921–928 (2009).

[24] Zuk, T., Schlesier, L., Neumann, P., Hancock, M. S., and Carpendale, S., "Heuristics for information visualization evaluation," in [*Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*], 1–6, ACM (2006).

[25] Lam, H., Bertini, E., Isenberg, P., Plaisant, C., and Carpendale, S., "Empirical studies in information visualization: Seven scenarios," *Visualization and Computer Graphics, IEEE Transactions on* **18**(9), 1520–1536 (2012).

[26] Shneiderman, B., [*Designing The User Interface: Strategies for Effective Human-Computer Interaction, 4/e (New Edition)*], Pearson Education India (2003).

[27] Grinstein, G., Konecni, S., Scholtz, J., Whiting, M., and Plaisant, C., "Vast 2010 challenge: arms dealings and pandemics," in [*Visual Analytics Science and Technology (VAST), 2010 IEEE Symposium on*], 263–264, IEEE (2010).

[28] Zeinaly Baraghoush, M., *Visualizing mutations of a virus sequence*, Master's thesis, Communication, Art & Technology: School of Interactive Arts and Technology (2012).

[29] Stolte, C., Tang, D., and Hanrahan, P., "Polaris: A system for query, analysis, and visualization of multidimensional relational databases," *Visualization and Computer Graphics, IEEE Transactions on* **8**(1), 52–65 (2002).

[30] Pirolli, P. and Card, S., "Information foraging in information access environments," in [*Proceedings of the SIGCHI conference on Human factors in computing systems*], 51–58, ACM Press/Addison-Wesley Publishing Co. (1995).