**1. Introduction**

Applying the principles of Information Visualization to design biology-specific systems

is highly acknowledged in the literature [18, 25, 26, 30, 32]. Advanced hand-drawn pictures

in scientific publications prior to existing computers shows that utilizing the human vision

system was grounded in biology many years ago [32]. As the biological data-sets scales are increasing rapidly, custom software combined with manual intervention is replacing manual data analysis in biological sciences [18].

These computer-based visualization tools have enhanced our ability to communicate with the large amount of scientific data. 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 domain’s design guidelines and patterns. A special issue of Nature Methods gives biologists an overview of current computational methods and tools used for visualizing biological data [32]. Although classical visualization techniques are used in the field of

biology, researchers define new and creative ways to meet the target domain visualization

needs [15]. One such example is the work of Nielsen et al. in creating a novel graph

representation for visualizing genome sequence assembly structures [30].

**2. Literature review**

Once a virus infects a host, it makes copies of itself, growing the population of virus within the same host and eventually spreading to other people. During the viral replication process, its gene sequence has to copy and transmit the exact same sequence (between 7000 to 500,000 nucleotide bases) to its child cells. During this process, typically some mistakes can be made and as a result, some changes appear in genetic sequence [31].

One way for characterizing DNA is to compare their sequences with each other [19].

In bioinformatics, aligning the sequences in rows helps finding the similar regions between them. In the case of having Pairwise Alignment, analyzers compare the sequence of one gene to the sequence of another; that, in many cases, is useful for their tasks. In the case of having a set of DNA sequences however, comparing each sequence against all the others in

the set is not only time consuming, but also leads to inaccurate results.

health investigators are interested

in finding the relation of the gene substitutions to disease characteristics in a virus gene family of sequences.

**2. design**  
Figure 1 FilooT visualization system. (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 pvalue in  
Mann- Whitney U test 3.2.4.3 ) 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 Statusbar is being updated after each action that the user makes.

1. **2.1. Interactive Tabular view**

The tabular view 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 color is used to represent those cells that did not change in comparison with the original sequence and the yellow color 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. These interactions affects the other views linked to Main View.

:

**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.  
**Reset:** In order to compare different columns with each other, placing the columns close to each other frees up the cognitive load of the users and enables them to use their memory to focus on their desired task [40]. 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, enabling this feature admits that the user can change the natural order of nucleotides in a sequence. One must realize that the natural order is meaningful in the original domain. In order to keep the natural position orders, the “reset” button returns the columns to their original sequence from one to the length. This feature is used whenever the user previously changed the column positions, and wants to reset the position numbers.  
**Filter:** 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.

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). Moreover, a small number of substitutions in a column may occur randomly and do not reveal any valuable information to the analyzers.

Therefore, an augmented filtering excludes the columns that have fewer yellow cells than the filter number.

1. **2.2. Matrix View**

The matrix view 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 [11]. In table lens, the levels are shown by the length of horizontal bars or colour saturation per cell [43]. 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. On top of each column, there is a coloured label that shows the different levels in that particular column. The darker the colour, 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 [23]. In addition, the colour saturation is a better alternative for the length channel for encoding this ordered information [23].  
We also used hue to separate different characteristics that are nominal data and the hue channel is appropriate for separating different categories [23].  
  
The user can perform the following list of interactions in Matrix View:  
• **Sort:** If the user selects a column header, the rows will be sorted according to the values of that column. Besides, the user can choose between ascending and descending sorting.  
• **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 Main View  
• **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.  
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.

1. **2.3. The P-Value View**

# 

There is a pattern [4] within some of the columns that makes them interesting candidates to form new hypothesis.

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 [28], we cannot rely on them to find this pattern in columns.

Commonly biologists use metrics detect interesting patterns. Mann-Whitney U test’s p-value is one of the metrics used for finding relevant positions [4].

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 of a 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 to help users find relevant columns.

The length channel is the second most powerful channel for encoding the ordinal values [reference]. Therefore we used length to represent the p-value metric.

The P-value view also provides the filtering feature.

The filtering 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, sorting all the rows 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 based on the reverse of the p-value (length of bars. keep the bars on top of the columns in Main View, so that the user could go over the bars while observing the columns’ pattern.

The Tabular view and the P–Value view are linked so that if the user re-orders the positions in one view, the corresponding column’s order will be changed in the other. Also they can use the reset button to go back to the original domain ordering.

1. **2.4. Group View**

The group view helps Users to find related columns (or rows) and group them together to focus on fewer rows (less data dimensions) for future analysis.

grouping feature is defined for both columns and rows. the user is allowed to click on rows and add them to a newly created group. When the user is in

row mode, the user can select different rows to make a group of them.

The user also can separately load each group into the views in order to investigate the group information and to focus on the relationships between the columns. It is more likely that they will make these groups from the relevant columns. Column Grouping The idea is to let the user make different groups from a combination

of different columns. The user can see an overview of the group and its general pattern.

This view supports two categories of interactions with groups: basic grouping and augmented Grouping.

In the Basic Grouping, groups are created by user based on their prior knowledge or observations.

In contrast to the Basic Grouping, in the Augmented Grouping, 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.

The augmented grouping motivates the graph view.

An overview of the distributions should allow analysts to see the big picture, and identify clusters, trends and outliers that may be candidates for detailed inspection [43]. Therefore, an overview of a group consists of a larger window than Main View information (prior to zooming).

If the user clicks on the overview the content of that specific group is loaded in all other Views. There is a predefined group that contains the entire data-set for the user to be able to go back to the original data (with latest changes).

In order to guide the users to find related columns or related row, an augmented grouping feature is designed. This feature is different for rows and column because they could have different kind of relationships. The visualization of augmented feature is supported in Group View.

There is linkage between Main View and the selected group. When the user selects a group among the previously created groups from Group View, the chosen group’s data will be uploaded into the system. Therefore, the data in all Views matches the data in the selected group. when the user’s mouse hovers a column, the corresponding column is highlighted in the overview of the currently selected group.

1. **2.5. Graph View**

A node-link representation visualizes the relationship between the columns (or rows).

Column relationship:

Two kinds of relationships between any pair of columns are supported: Complementary patterns and Correlation.

Given that the substitutions in this data-set are represented by values 0 and 1, one may define the similarity between two columns with measures such as Pearson’s 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 is not true.

To alleviate this problem, we propose to use a new measure which ignores the common zeros between columns. 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) = n

i=1

xiyi −

n

i=1

xi ⊕ yi, (3.2)

where ⊕ is the logical XOR operation and results in 1 when one of the side equals to 1 the other side equals to 0. This measure, ignores entries with no substitution in both columns, increases when entries with substitutions occurs together and decreases when substitution complements each other. Given that, both positive and negative values are expected.

Row Relation:

The relations between rows are hierarchical. The already designed

Graph View tis used to make a Tree for the representation of this relationship.

One alternative representation for relationship between a pair of columns is the matrix visualization [15]. One benefit of using this matrix is, by re-arranging the rows and columns, some interesting patterns would be revealed. However, this option requires a large screen space. One drawback is that we cannot eliminate the cells with 0 correlation from the space.

The second option is using a node-link graph, where there is a link between a pair of 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.

Assuming we know the degree of the relationship between a pair of columns with numbers from 0 to 1 for correlated ones and -1 to 0 for complementary ones, my initial suggestion to encode this information was using a graph or a table representation.

Interaction:

There is also another filter mechanism built into the view that removes the different levels of correlation links.

the user’s mouse position highlights the corresponding row label in Main View (in

as well as the equivalent node label in Graph View the user deletes a row from the matching node in Graph View will be deleted.

Two filters are placed to enable the user to sort out the columns based on the strengths of their connection.

**Evaluation**

**Discussion**