GeneTracer: Gene Sequence Analysis of Disease Mutations
VAST 2010 Mini Challenge 3 Award: Excellent Process Explanation

Hanseung Lee* Jaegul Choo† Carsten Görg† Jaeeun Shim* Jaeyeon Kihm* Zhicheng Liu†
Haesun Park† John Stasko†

Georgia Institute of Technology

ABSTRACT

Our visual analytics tool GeneTracer, developed for the VAST 2010 genetic sequence mini challenge, visualizes gene sequences of current outbreaks and native sequences along with disease characteristics. We successfully used GeneTracer in combination with data mining techniques to solve the challenge.

Keywords: VAST challenge, visual analytics, gene sequence analysis

Index Terms: H.1.2 [Information Systems]: User/Machine Systems—Human information processing; H.2.8 [Database Management]: Database Applications—Data mining

1 PROBLEM OVERVIEW

The task of the VAST 2010 Mini Challenge 3 (Genetic Sequences) is to find the country of origin for a disease outbreak, which patient contracted the disease from the initial carrier, and the gene bases that led to particular characteristics and mutations. The provided data included the gene sequences from the current outbreak and from native people in several regions/locations, as well as the disease characteristics. The current outbreak and the native sequences each have two fields: a sequence identifier (or country name in the native sequences) and the actual gene sequence. A gene sequence consists of a list of gene bases coded as A, T, C, or G. The disease characteristic data has an identifier field and five characteristic fields: symptoms, mortality, complications, drug resistance, and risk vulnerability. Each characteristic field's value is a categorical value taking on one of either two or three severity levels.

2 GENETRACER

The GeneTracer system provides three views, the Gene Sequence View, the Disease Characteristic View, and the Graph View.

The Gene Sequence View (right view in Figure 1) presents the current outbreak virus sequences and the native sequences as horizontal rows in a grid. The base for each gene in a row is color-coded: A (red), T (green), C (purple), G (blue). A special heatmap row vector (top row of the grid) represents how varied the gene bases for that position (column) are across regions that have different characteristic values. Similarly, a special heatmap column vector (leftmost column of the grid) represents how much each sequence (row) is different from the selected row. The user can interactively reorder and remove rows and columns from the grid.

The Disease Characteristic View (left view in Figure 1) presents each sequence's characteristics using shaded color cells. The different sequences are in the rows and the five characteristics are in the columns. The color mapping is Symptoms (red), Mortality (blue), Complications (green), Resistance (purple), and Vulnerability (orange). Darker values (cells) indicate more severe characteristics for a particular gene sequence. Additionally, a total characteristic weight is calculated and shown to the right by adding up the individual characteristics. The user can sort the sequences (rows) by a specific characteristic or total weight.

These two views can be linked: when a sequence is selected in one view, it is also selected in the other. Additionally, the row order of the sequences can be synchronized across the two views.

Finally, the Graph View (not shown) visualizes the relations among the sequences via a minimum spanning tree representation where the weight of an edge between two sequences is the Hamming distance between the two. Selecting a node in the Graph View will also select that node (sequence) in the other views.

3 ANALYTICAL PROCESS

Initially, we viewed all of the current outbreak sequences and the native (region) sequences in the Gene Sequence View. It showed 68 different sequences (rows) with 1404 gene bases (columns). We used GeneTracer's functionality to remove all gene bases that were identical across all the sequences, thus resulting in a much more manageable number of gene bases. With this reduced data, we directed GeneTracer to construct the graph and calculate the minimum spanning tree (MST) that shows distances between the sequences. From the visualization of this MST in the Graph View, we observed that Nigeria-B was the nearest native sequence to the current outbreak sequences. Therefore, we suspected Nigeria-B as the country of origin of the current outbreak. To check this hypothesis, we analyzed the data in the Gene Sequence View and interactively changed the order of the rows by dragging them upwards closer to the outbreak sequences, making the comparison easier. In addition, we filtered out some of the sequences that were clearly dissimilar. By interactively exploring the data in this manner we found that Nigeria-B was by far the most similar sequence to the current outbreak, which matched with the result of the Graph View.

To solve the second problem, we examined the strains identified by sequences 583, 123, and 51. We dragged these three sequences to the top of the Gene Sequence View. By observing the heatmap column vector, we found that 583 was more similar (lighter color) to 123 than to 51. We then filtered out the columns exhibiting the same gene bases among the three sequences. From this analysis, we found that sequence 123 has only one different gene base (column index 269) whereas sequence 51 has three different gene bases (column indices 494, 842, and 946) compared to gene sequence 583. Therefore, we concluded that the patient identified by sequence 123 is likely to have contracted the disease from Nicolai (sequence 583).

To identify the top three mutations that lead to an increase in symptom severity, we used the Disease Characteristic View and the Gene Sequence View together with interactions. At first, we sorted and reordered the sequences based on the symptom severity in the Disease Characteristic View. We applied this order to the Gene Sequence View, which then showed the boundary of different levels of a certain characteristic with a blue thick line. Next, we reorganized the Gene Sequence View by moving the columns (gene bases) with

*e-mail: {hanseung.lee, jaeeun.shim, jkihm3}@gatech.edu
†e-mail: {joyfull, goerg, zcliu, hpark, stasko}@cc.gatech.edu

IEEE Symposium on Visual Analytics Science and Technology
October 24 - 29, Salt Lake City, Utah, USA
978-1-4244-9486-6/10/$26.00 ©2010 IEEE

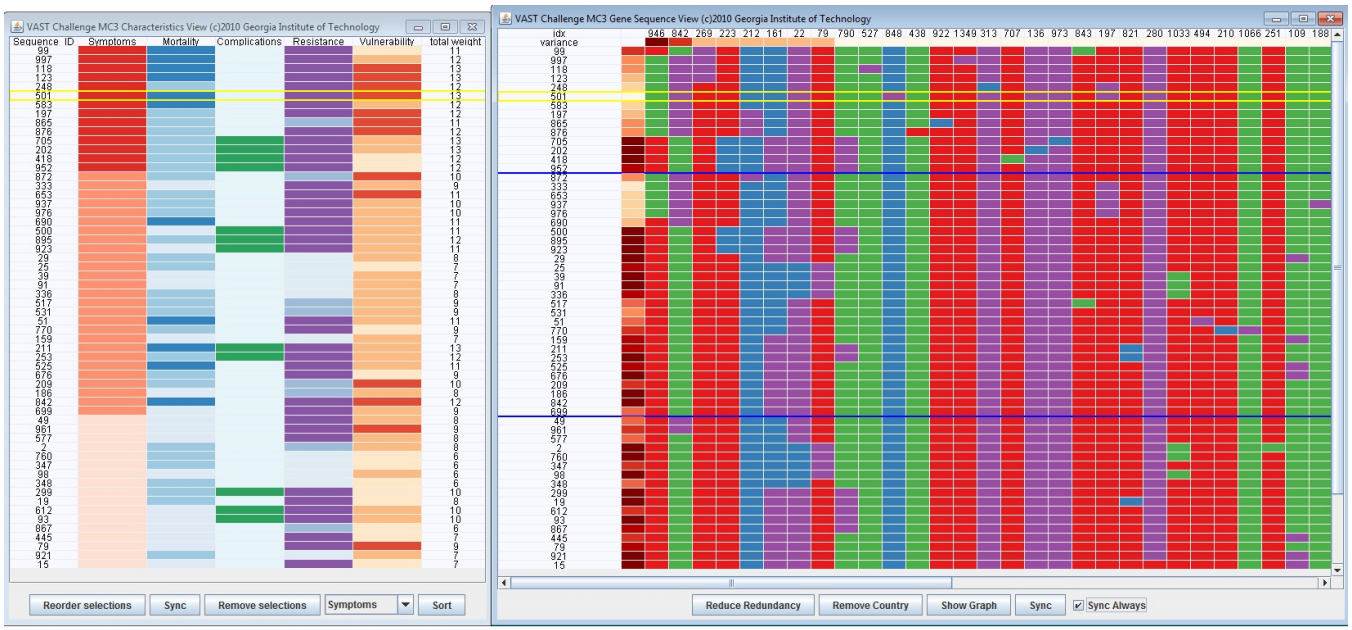

Figure 1: The Disease Characteristic View (left) and the Gene Sequence View (right) are linked and synchronized after sorting the rows in the Disease Characteristic View by symptom severity. Gene sequence 501 is selected.

the largest variance to the left side. We also changed the order of sequences (rows) to place similar gene bases together within the same category of characteristics. Through these interactive steps, we could see the patterns of gene sequences and find some potential mutations that were critical for each characteristic.

Mutation $A \rightarrow C$ at position 269 only occurred in severe symptoms (at sequence 99, 118, 123 and 997), so it was strongly related with symptom severity. The most common mutation in the severe symptoms is $A \rightarrow T$ at 946 and $T \rightarrow C$ at 842. It occurred frequently in the severe symptom sequences (9 times out of 14 sequences) and also occurred five times in the moderate symptoms. If $T \rightarrow C$ at 842 occurs, but $A \rightarrow T$ at 946 does not occur, then this mutation results in mild symptoms (e.g., in sequences 49 and 961). Therefore, if these two mutations occur at the same time, it increases the symptom's severity.

For the third mutation, we obtained three candidates, $A \rightarrow G$ at 223, $A \rightarrow C$ at 197, and $G \rightarrow C$ at 212. We considered them as candidates since these mutations occurred in severe and moderate symptoms more than twice. We finally selected $A \rightarrow G$ at 223 as our third mutation, since the sequences of the other two candidates overlapped with the other two mutations we first found.

For considering all the characteristics, our approach consisted of three steps. First, we found some candidate mutations for each characteristic using the previously explained strategy. For each characteristic, we sorted the sequences and analyzed them using hints from the heatmap and filtering interactions. As a result, we found two to five candidates of critical mutations for each characteristic. In a second step we applied the same process as in the first step, except that the gene sequence was sorted by total weight. We already assigned a weight severity from one to three to each characteristic value, and the total weight was determined by aggregating all the characteristic weights. We reorganized the resulting Gene Sequence View based on the total weights. We then focused on critical mutation candidates from the first strategy and analyzed again by moving, removing, or filtering columns and rows.

In a third step we applied several data mining techniques, such as decision tree and regression, to choose candidates of genes that carry significant information. Regression and decision trees were helpful to provide initial clues. From those, we chose the column indices with the highest coefficient values and also examined a few nodes near the root of the decision tree. This led to an improved classification result. We used first-order and second-order linear regression to explore initial column indices (gene base positions) that could be potential genes carrying significant information. We selected the top few coefficients with the corresponding column indices and started exploring the data with the views in the GeneTracer tool. Even though these data mining techniques did not directly provide the answers we expected, they still suggested some column indices for closer examination in the visualizations. Finally, we were able to verify that the hypotheses we formed from the qualitative decision making process were correct, based on the results from the quantitative analysis.

4 Conclusion and Future Development

GeneTracer successfully facilitated the process of identifying the country of origin for the current outbreak, similarity of a pair of genes, and the top few mutations that led to an increase in a specified disease characteristic. For future development, implementing algorithms that perform automatic reordering of matrices [1, 2] will improve the effectiveness of GeneTracer. Since each position of the gene sequence has at most two kinds of gene bases, we can model it as a binary matrix and can easily solve the matrix reordering algorithm.

Acknowledgements

The work of these authors was supported in part by the National Science Foundation under awards CCF-0728812, CCF-0808863, and IIS-0915788, as well as the VACCINE Center, a Department of Homeland Security Center of Excellence in Command, Control and Interoperability.

References

[1] I. Liiv. Seriation and matrix reordering methods: An historical overview. *Statistical Analysis and Data Mining*, 3(2):70–91, 2010.

[2] L. Wilkinson and M. Friendly. The history of the cluster heat map. *The American Statistician*, 63(2):179–184, 2009.