Discovering mutation paths in sets of genetic sequences and determining critical mutations

Georgios Petkos* CERTH / ITI
Konstantinos Moustakas† CERTH / ITI
Dimitrios Tzovaras‡ CERTH / ITI

ABSTRACT

This short summary presents the approach used by the authors to gain insights in the genetic sequence data of the third mini challenge of VAST 2010. We employ multidimensional scaling and a minimum spanning tree to determine relationships between genetic sequences and discover likely mutation paths.

Index Terms: K.6.1 [Management of Computing and Information Systems]: Project and People Management—Life Cycle; K.7.m [The Computing Profession]: Miscellaneous—Ethics

1 GENETIC SEQUENCE ANALYSIS

The third mini challenge of VAST 2010 presents a set of genetic sequences of viruses along with some characteristics of the diseases that they cause. The questions asked are about identifying likely sequences of mutation and about identifying mutations that worsen disease characteristics.

One of the main issues with the challenge is the discovery of the most likely sequences of mutations. In order to deal with this problem, we use two alternative but complementary methods. In both, the distances between all pairs of genetic sequences are computed. That is, a symmetric matrix D is computed, where the element D_{ij} holds the number of bases that are different between sequences i and j.

In the first method, this distance matrix is used to compute an allocation of genetic sequences in 2 dimensional space using multidimensional scaling (MDS). Similar genetic sequences, i.e. sequences with small distance D_{ij} will be placed close, whereas dissimilar genetic sequences will be placed far apart.

In the second method, the dissimilarity matrix D is treated as the weighted adjacency matrix of a fully connected graph. Subsequently, the minimum spanning tree of this fully connected graph is computed. This minimum spanning tree contains the most likely mutation path between each pair of sequences.

This analysis assumes that mutations are a rather rare event and that they usually occur in a very small number of bases during the replication process of a sequence.

The tool described has been built from scratch using Processing. The only external dependency is the MDSJ library for computing multidimensional scaling.

2 VISUALIZATION AND INTERACTION MECHANISMS

The two analysis methods that were mentioned in the previous section, provide two alternative visualization methods that the user can switch between. In both, the characteristics of each sequence, along with an aggregate severity measure (computed as a simple sum of individual symptoms' numerical representations of characteristics) are displayed with color coding and are allocated according either to the tree based layout (where nodes are placed on a regular manner) or the MDS based layout. In addition, the user can mark sequences in any of the two representations and the set of bases of the sequences appear below the central visualization, in an auxillary visualization, as displayed if Fig. 1. It is important to mention that in order to make the displayed information more clear, only the bases where at least one of the selected sequences differs are shown.

The user can perform any of the following actions:

- Choose if the set of native, outbreak or both native and outbreak sequences is displayed.
- Choose between the tree based and the multidimensional scaling based methods.
- Combine the two methods by allocating nodes in 2D space using MDS but also displaying the links of the tree. This differs from the default tree representation in that it does not have a regular allocation of the nodes. It can provide useful insights to the relationships between the sequences.
- Select the set of disease characteristics to be displayed in the visualization (only when displaying the outbreak sequences, since there are no disease characteristics available for the native sequences).
- Select the set of disease characteristics to be used for computation of the multidimensional scaling based representation (only when displaying the outbreak sequences).
- Mark sequences in the main visualization (in order to compare them in the auxillary sequence visualization).
- Perform linked highlighting of marked sequences between the main and auxillary visualizations.
- Display mutation details when a mutation link is highlighted in the tree based visualization. It is important to mention that the tool also highlights any other mutation link that corresponds to a mutation in the same position of the sequence.
- Select the root node of the mutation tree so that alternative layouts are obtained.
- Order the markeed sequences by gene or by characteristic.

3 HYPOTHESIS GENERATION AND ANALYSIS USING THE TOOL

We display the utility of the developed visualizations and interaction mechanisms by attempting to answer the questions of the challenge.

Question 1

The first question is about identifying the most similar sequence within a group of sequences (the native sequences) to another group of sequences (the outbreak sequences). We opt to display simultaneously both the outbreak and native sequences using either the tree or the MDS based representation (native sequences are represented by elements with different colors). In the MDS representation, it is

*e-mail: gpetkos@iti.gr
†e-mail:moustak@iti.gr
‡e-mail:Dimitrios.Tzovaras@iti.gr

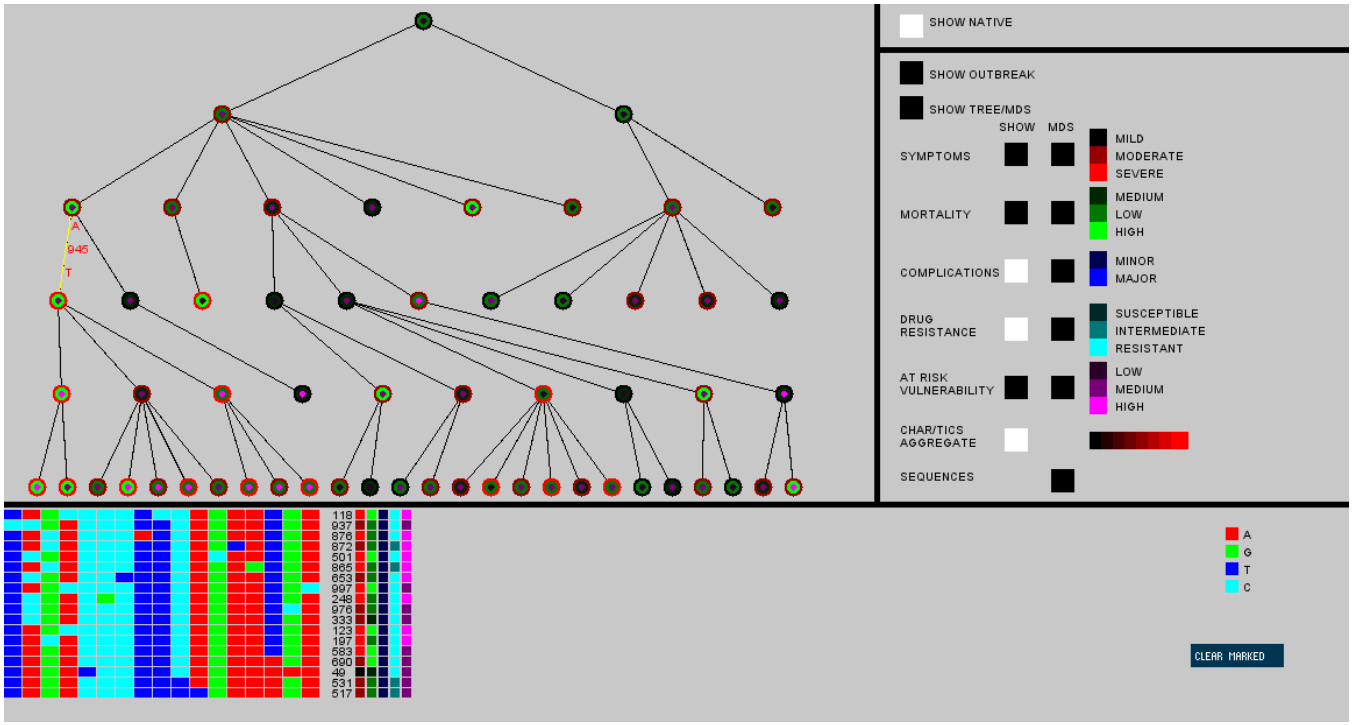

Figure 1: Overall view of the tool

seen that the outbreak sequences form a tight cluster, whereas the native sequences lie farther apart. The closest native sequence is Nigeria_b, therefore Nigeria is identified as the most likely origin of the outbreak. This can be verified by switching to the tree based representation: the outbreak sequences form a completely separate subtree and the native sequence that is connected to this subtree is Nigeria_b, therefore the same conclusion can be reached.

Question 2

The second question is again about sequence matching and can be answered in a similar manner. This time only the outbreak sequences are displayed and using the tree representation it can be seen that the patient with the strain identified by sequence 123 most likely contracted the virus from Nicolai (sequence 583), since only one mutation link exists between sequences 583 and 123, whereas three mutation links separate sequences 583 and 51. A similar conclusion can be reached using the MDS representation. Therefore, it is safe to assume that it is quite more likely that the patient who carries the sequence with id 123 contracted the virus from Nicolai rather than the patient who carries the sequence with id 51.

Question 3

The third question asks for the three mutations that cause the largest increase in symptom severity. To answer this question, we choose to display only the relevant disease characteristic and use the tree based visualization. Subsequently, we look for three subtrees that carry sequences that cause increased symptom severity compared to their neighbours. The ability that the tool provides to find mutations that occur in different parts of the tree has been helpful at this point, as it made it quite clear that the mutation in position 268 from A to C always results in severe disease symptoms. Three other subtrees are easily identified as candidates and the two with the largest ratio of sequences that cause severe symptoms are selected (please see answer form or video).

Question 4

The final question is similar to the third and asks for the three mutations that cause the more severe worsening of all disease characteristics. Analysis can be carried out either using the tree based or the MDS based visualizations. Instead of displaying symptom severity, we show the aggregate characteristic, which, as it was mentioned, is the simple sum of a number encoding of the individual characteristics. Again, the way to achieve this is to look for the three subtrees with the most intense coloring. Moreover, the ability to mark sequences and order them according to their symptom severity has been very helpful for this task.

4 Comments

The described tool performs analysis of a set of genetic sequences by means of multidimensional scaling and minimum spanning tree computation and offers alternative visualization methods. Moreover, it provides a rich set of interaction mechanisms that have allowed us to answer the questions of the challenge.

Although only marginally useful for answering the questions of the challenge, interesting patterns could be obtained by selecting different sets of characteristics for inclusion in the computation of the MDS. For instance, attempting to answer the last question, groups of sequences that had high values in the aggregate characteristic could be observed when selecting all features for inclusion in the computation of the MDS.

Finally, alternative tree drawing mechanisms could significantly improve the visual clarity of the tree based representation (e.g. a tree with the root lying in the center and the branches expanding outwards).