

IMAS and Genetic Sequences: Tracing the Mutations of a Disease

VAST 2010 Mini Challenge 3

Mahshid Z. Baraghoush,

Chris D. Shaw

School of Interactive Arts and Technology, Simon Fraser University

ABSTRACT

We provide a description of the IMAS (The Interactive Multi-genomic Analysis System) tool (Shaw et al. 2007) and our analytical experience from the contest entry submitted to the VAST 2010 Mini Challenge 3 dealing with a synthetic dataset.

There were 58 mutant strains of a disease with a length of 1400 nucleotides and 10 country strains with the same length. Each sequence has some disease characteristics such as Symptom, Mortality and so on.

This paper also highlights IMAS strengths and weaknesses in dealing with each task. We used IMAS to find the root sequence of all strains, explore nucleotide substitutions, define different groups of sequences and multi-align all the sequences in each group.

KEYWORDS: Design, Bioinformatics, Visual Analytics, VAST Contest

INDEX TERMS: J.3 [Life and Medical Sciences] Biology and Genetics, I.3.3 [Computer Graphics]: Picture/Image Generation - Viewing Algorithms; I.3.6 [Computer Graphics]: Methodology and Techniques - Interaction Techniques

1 INTRODUCTION

The VAST Challenge is a part of the IEEE VAST 2010 Symposium (VisWeek 2010) that invites visual analytics researchers and practitioners around the world to solve a suite of problems using their Visual Analytics Applications.

This year's VAST Challenge consisted of three mini-challenges, plus a grand challenge that knits the 3 mini-challenges together. This paper is a summary of our Mini Challenge 3 entry. We also contributed to SIAT's Grand Challenge entry and our team has won the Excellent Student Team Analysis award.

The vast challenge 2010 provides opportunities for us as developers of IMAS to evaluate our techniques. We decided to use the benchmark data set for our future testing and establish a more advanced visual analytics bioinformatics tool.

2 TOOL

IMAS is a Visual Analytics system for the discovery of knowledge in genomic information. This software is available on SourceForge at imas.sourceforge.net under the GPL 3 license.

IMAS enables the user to load various FASTA format files. One or more sequences can then be selected for analysis. This tool visualizes the output of common bioinformatics tools such as BLAST and ClustalW in a unified framework. In this challenge, BLAST is used for pair-wise nucleotide sequence alignment, and

ClustalW is used to perform NT sequence multi-alignment. Pair-wise alignment visualizes the character-by-character similarities and highlights the differences between sequences with color. IMAS also enables users to select BLAST hits and their corresponding sequences for multi-alignment. Multi-alignment results are displayed such that identical NT letters are given a background color, and NT letters different from the consensus are highlighted.

IMAS provides the user with a horizontal zooming of the sequences. This interaction assists the user in discovering patterns in the sequence by controlling its level of detail. Those patterns emphasize non-conserved regions at a glance. The user can then zooms into a region of interest and examine it in more detail.

3 FINDING PAIR-WISE RELATIONSHIPS

When determining the original country of all the virus strains, we assumed that the country strain which displays the most similarities to each of the 58 outbreak strains is the ancestor of all the mutant strains. We defined the similarity between two strains as the number of different bases.

Figure 1 shows one BLAST run result for strain 118 against all the countries. The light blue areas show the similar regions and the green rectangles highlight the differences, which indicates that at least one substitution takes places at that area.



Figure 1. A window of the results of the pair-wise alignment of sequence 118 against each of the countries

By looking at the image it is clear that the first sequence (Nigeria B) has the least number of differences with the strain 118 in comparison with others.

4 FINDING RELATIONSHIPS AMONG MULTIPLE SEQUENCES

IMAS allows the user to define different sets consisting of a selection of the sequence strains in the dataset. The user can multi-align all the sequences in each set together. This way,

mutations that occur within a set are highlighted by a different color. In order to solve one of the tasks of the contest, we decided to split up the sequences into distinct subsets based on the disease characteristics of each strain which were provided in a table.

Figure 2 shows four different sets of strains where the sequences have been aligned together within each set. The sets are sorted by their overall danger level from the most dangerous set on the top to the least dangerous one on the bottom. Having the different sets sorted is useful to analyze their behavior; however their order of placement in the window cannot be changed after the first creation. Our future goal is to provide the option for the users to be able to rearrange the sets after the first definition.

The highlighted letters show the differences between all the sequences according to their similarity to the consensus. The dark purple shows that all the sequences have the same characters and that no mutation occurs in that area.

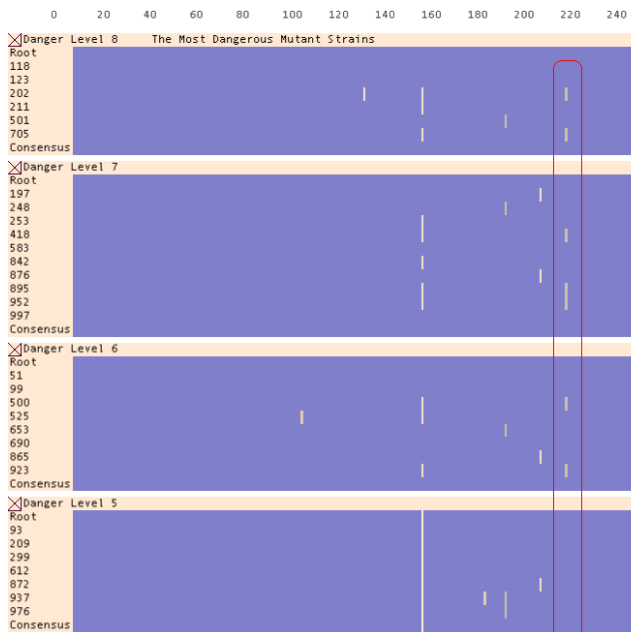


Figure 2. An example window of four multi-aligned sets. The red rectangle shows an interesting position in which there is an increase in substitutions from low danger to high danger groups

4.1 INTERACTION

The ruler on top of the snapshot shows approximate positions in the sequences. The example shown in figure 2 contains 240 Nucleotides, from position 0 to 240. Scrolling left to right will allow the user to see more of the alignment. Zooming out enables the user to see the entire alignment at a glance; by looking at the fully zoomed-out picture, the user might see some patterns and get a primary idea of the different positions. By zooming in, the analyst can go to the positions of interest and get a more detailed view. There is also a vertical ruler that shows the exact positions of each nucleotide.

The next step to solve the challenge’s task was to compare the positions of interest and make a conclusion about their impact on disease characteristics. To do this, we copied our selected snapshots into an image editing software and made a new image of all them together. IMAS is currently unable to display non-contiguous areas of interest. We plan to add hide-unhide feature to the system for this purpose.

Figure 3 shows four of the positions that seemed interesting to us. The reason is positions 946 and 842 are highly correlated and

their substitutions occur more commonly in the most dangerous groups. We will name this pair of substitutions as *mutation 1*. However, in dangerous groups where mutation 1 does not occur, there is another pair of substitutions that occur in tandem: positions 161 and 790. We call this substitution pair *mutation 2*.

Each sequence in danger group 8 and 7 has either mutation 1 or mutation 2, but not both. We called this a complementary pattern and we plan to will visualize that in future work.

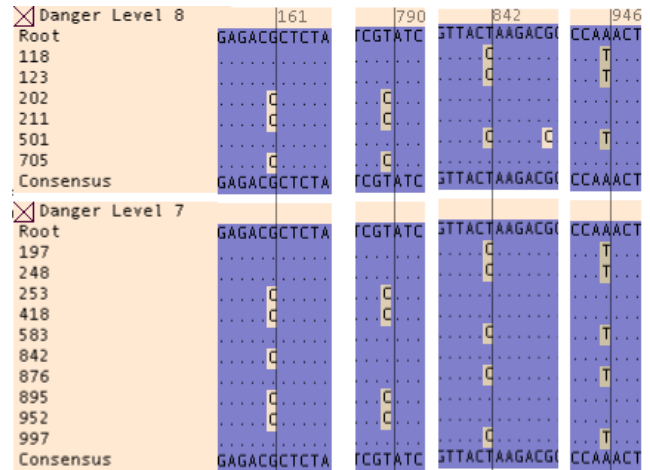


Figure 3. Interesting positions that we cut their image and put them together in a separate picture.

5 CONCLUSION AND FUTURE WORK

We believe our tool has a number of strengths. The application provides the users with an overview of the entire dataset. IMAS also enables the user to rapidly select and run BLAST and Clustal-W analyses on selected sequences. Data management and ingest from these external tools is automatic.

Starting from a complete picture, the zooming feature facilitates exploratory data analysis. The ability to define different group of strains coupled with multi-align visualization allow analyst to gain insight into the data about different positions with various levels of detail. That helped us discover the complementary visualization pattern between some of the positions.

However we believe our tool should interact more with the user. Right now the user has to define separate groups for multi-alignment purpose manually; we believe that an external sorting mechanism on different features would be beneficial. At present we have not provided any capability of editing or changing the place of each row, column, or base in the multi-alignment view, we are looking forward to add functionality to provide users with a more general set of tools for exploring genomic sequences.

ACKNOWLEDGMENTS

We would like to thank Ji-Dong Yim¹ for his valuable contributions to the initial idea about complementary positions.

REFERENCES

- [1] C. Shaw, G. Dasch, and M. Ereemeeva, “IMAS: The Interactive Multigenomic Analysis System”, Proceedings of IEEE Visual Analytics Science & Technology 2007, IEEE, Sacramento, CA, Oct 30-Nov 1, 2007, pp. 59-66.

¹ e- mail: jdyim at sfu.ca