Featured IMAS multi-alignment
For exploration of the evolutions of a virus strains
To link mutations and disease characteristics

Mahshid Z. Baraghoush (mzeinaly@sfu.ca)

Domain, Task and Data

Domain

My dataset and tasks is part of the Vast Challenge 2010, mini-challenge 3. The VAST Challenge is a part of the IEEE Symposium (VisWeek) that invites visual analytics researchers and practitioners around the world to solve a suite of problems using their Visual Analytics Applications.

One of the purposes of this challenge is to speed the transfer of VA technology from research labs to commercial products, and to provide opportunities for developers to evaluate techniques, pushing the forefront of visual analytics tools using benchmark data sets and establishing a forum to advance visual analytics evaluation methods.

VAST Challenge 2010 consisted of three mini- challenges, plus a grand challenge that knits the 3 mini-challenges together.

Data Set

The VAST mini Challenge 3 dataset consist of three files.

CurrentOutbreakSequences.txt

NativeSequences.txt

DiseaseCharacteristics.txt

File1

“This dataset contains sequences collected during the current viral outbreak. There are a number of entries in this text file. For each entry, the first line is a label identifying the new viral strain (mutant). The next line contains the viral strain’s genetic sequence. ”

>118

ATGTCACCGCCCTGCGCAGTTCATAGGGCCTCTCTTCGCCGGAACACGGGTCCTTTCTGGATGG

TGAGGGTTGTGGGAAAGACTTGTAGCCATAACGCATATCC … (1400)

>770

ATGTCACCGCCCTGCGCAGTTCATAGGGCCTCTCTTCGCCGGAACACGGGTCCTTTCTGGATGG

TGAGGGTTGTGGGAAAGACTTGTAGCCATAACGCATATCC …

… (56 Outbreak sequences)

File 2

“This dataset contains sequences of native viral strains. There are 10 entries in this text file. For each entry, the first line is a label identifying the origin (the name of the country or region in Africa) of the native viral strain. The next line contains the viral strain’s genetic sequence. “

>West_And_Central_Africa

ATGTCTCCGCCCTGCGCAGTTCATAGGGCCTCTCTTCGCCGGAACACGGGTCCTTTTTGGATGGT

GAGGGTTTTGGGAAAGACTTGTAGCCATAACGCATACCC… (1400)

>Nigeria_B

ATGTCACCGCCCTGCGCAGTTCATAGGGCCTCTCTTCGCCGGAACACGGGTCCTTTCTGGATGG

TGAGGGTTGTGGGAAAGACTTGTAGCCATAACGCATATCC…

… (10 country sequences)

File 3

There are some Characteristics for each of the Outbreak sequences.

And an explanation about the table:

"The third dataset is a tab-delimited table. This table categorizes virulence and drug resistance characteristics we selected related to the viral strains collected during the current outbreak. "

Example:

Table 1 Sequence Characteristics Table

Sequence ID	Symptoms	Mortality	Complications	Drug Resistance	At-Risk Vulnerability
32	Mild	High	Minor	Resistant	High
256	Severe	Medium	Major	Susceptible	Medium
19	Severe	Low	Major	Intermediate	Low
4	Moderate	High	Minor	Resistant	High
200	Mild	High	Major	Resistant	Low

… 56 sequences

Definitions

Symptoms – what a patient experiences (e.g., pain, sore throat, vomiting, swelling, tremors)

Mortality – number of deaths as a result of disease

Complications – unfavourable evolution of illness (e.g. deafness, spontaneous abortion)

Drug Resistance – mutant vulnerability to anti viral drugs

At Risk Vulnerability – disproportional effect on certain risk groups (e.g. children, elderly)"

Tasks

I have chosen the following two tasks from the mini challenge. Here I summarized the tasks description:

Task 1

> Identify mutations that lead to an increase in symptom severity (a disease characteristic)

Task 2

Identify mutations that lead to the most dangerous viral strains.

The mutations involve one or more base substitutions.

Personal Expertise

I participated in vast challenge 2010 as part of a mini challenge 3 group and the to SIAT's Grand Challenge entry to that contest. Although our team did not win a prize for mini-challenge 3, our proposed solution for that mini challenge, contributed to the grand challenge and our team has won the Excellent Student Team Analysis award. In the mini challenge 3, we solved the problems with 78 percentage of accuracy, which was a good result, but the measure of being winners was more about the process and how much the tool could be helpful on that. From the reviewers and others solution, I realized that IMAS mainly has lack of interactions with the user and the user has to do a lot of parts by integrating the other tools along with IMAS (example: photo shop to extract the desired part from the images)

From that experience, I learnt how to improve the tool more specifically IMAS multi-alignment panel. Now it has been one year that I am thinking about the solutions, preparing IMAS documentations to enable to implement the solution and reading related papers. This work will be linked to masters my thesis work.

Proposed Solution

The image below shows IMAS multi-alignment panel. A multi-alignment view is created by choosing sequences that you want to multi-align together.

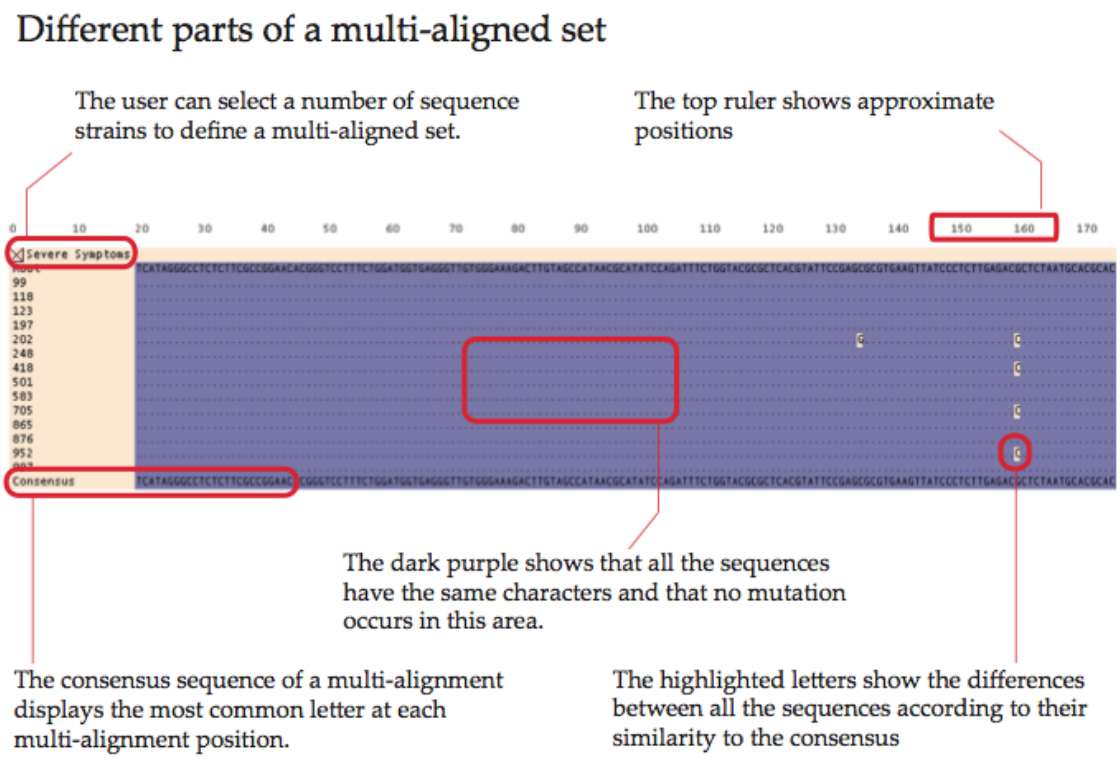

Figure 1 Different areas of the IMAS multialignment view

There are several features that could possibly make IMAS more usable for that dataset-task. I have categorize them into four based on the problems that we currently have:

Sorting the sequences in the multi-alignment view on a given characteristics:

In the current IMAS, multi-alignment sequences are placed based on the order of the selection of them at the first place. In these tasks each sequence has a different disease characteristics. We need to separate those that are most dangerous from those that are least dangerous. In the current version you have two choices:

- When you are selecting the sequences, select them in order that you want to be displayed. This is an overwhelming process as there are 56 strains in the dataset and you have to make an order somewhere else, and then try to find them one by one and then select them from the upper panel. And once you created the view, you cannot change your mind to change the order. In order to change the order,

you could make a new multi – alignment that would be placed below this old multi-alignment view.

- You may split the sequences into different groups of high, medium or low dangerous groups and multi align each group separately. The result in this example would be three multi-alignments each separated by a white space. This is similar to the above solution just deals with less number of sequences each time, but still has all the cons of the other solution.

Sorted sequences means the most severe ones go next to each other. If there is limited number of dynamic range then you could redundantly code each region based on each scale.

Selecting the sorting variable

The criteria here would be directed disease characterization or a derived variable constructed by one or some of those characterization. The tool could provide a way for the user to create this derived variable by selecting some characteristics and use some operator in between of them.

Also for selecting between derived or directed characteristics that which one is good for the sort, or which one is good to choose to make derived variables with them, it would be beneficial to be able to explore those characteristics, sort sequences based on each of them and be able to see the correlations between each of the characteristics and also the multi-alignment result based on each sort.

Filtering the positions

The analyzer could look at the entire 1400 positions at a time using the zoomed view. But this does not give the details about each position. Also there are some positions that do not give her any extra information. For example those positions that have not any pair substitutions happen in them.

Solution1: The ability to Hide / Unhide selected positions or a region of positions in the alignment. If the analyzer hides something, she should be able to unhide it by clicking on

some graphics showing that the region is hidden. Also there could be suggested built-in filters in the system that for example filters all of the positions that do not have any substitutions happening in all the sequences. (or have just one, or have just two…)

Seems that this filtering is a multi-level action that filters some positions in each level of analysis. Then if the user wants to go back, it unhide the previous levels not all the levels

Solution2: Sequence Juxtapose

Finding the interesting positions and being able to rank them

Those positions of interest in these particular tasks are those that they have more SNP on top of a sorted multi-alignment (high is more dangerous low is less). That by going down the rate of them decreases.

Finding the related positions to report mutations

The mutations involve one or more base substitutions (one or more positions). The question is when to combine two positions, and when to report them separately. Answering to this question is dealing with finding those positions that are correlated or complement each other. And combine them and report them as one mutation.

A Scenario of Use

For task 2: Identify mutations that lead to the most dangerous viral strains:

Lets assume that the multi-alignment view is ready for the user. The user would turn on the table lens view which looks like the picture below:

Figure 2 Early sketch of the table lens view

This view would give the user to select the sorting variable. The user also could define a new derived variable and see how the sorted multi-alignment would look like. So lets say the user see none of these characteristics mean the overall danger level, but the sum of them would make sense. So she asks to create a new column. To do that, she select all of these columns and put "+" as her operator in between of them. The system created this new column and shows its different values for each sequence. The user also has access to the range of the variables in each column and change the automatic assigned numbers to that. For example, symptoms is an ordinal variable, so the system automatically assigned 0-3 to its range, however the user might want to change it to be (1,2,4).

Sometimes in order to find this new derived variable, she might want to explore the correlations between different variables, that is why this table lens exists.

If the user satisfies with this sorting, then she closes that table lens view. The user then could filter the positions and just look at those that have at least one base pair substitutions occurring on them. The next step is to find the positions that we want from these filtered positions, which are about 57 positions. Now the p value visualization would help us find those with the least p values, which are about 21 positions. (The user hides all the other positions). Now the user should come up with 3 mutations each of them consists of some positions. The system will allow the user to find those related positions so that the user can come up with the groupings.

For example lets say position A has the least P value and position B has the next low p value. The user should report A then B? Or she can report A+B as the first mutation, then maybe C as the second mutation.

Implementation Approach

IMAS is written in C++ and use OpenGl for its graphics. In the time of this course, I would like to finalize my ideas about the design of these features in the context of IMAS and after this course implement those ideas or a selected of them in IMAS. I would like to propose that I work on my ideas and give feedbacks from others or by reading more literature and backgrounds before implementation. In order to be able to communicate with others, I would create some interfaces that show my designs with Protovis or Processing. In the final report I would have a series of interfaces from the early idea to the final design.

Milestones and Schedule

November 16: Prototype for p-value design complete

November 21: previous work of the other's solution in vast + refine my ideas

November 23: previous work of the literature and papers complete+ refines ideas

November 28: prototypes complete

November 30: discussed my ideas with at least 3 VA profs + 3 bioinformatics students

December 5: compile all the ideas + update prototypes + write down the justifications

December 12: Finish writing the final report

December 14: Edit the final report

There would be 5 designs for this work. Another safe iteration could be picking one of them at a time and do all the above step for that specific problem. As I do some part, I will figure out which works best.

Previous Work

Previous work has two parts: 1.Previous solutions for this challenge 2. Related papers

In this proposal I am going to discuss one of the successful solutions for the Vast mini-challenge 3, translations of what they did in IMAS, and lessons that I could use in my work:

Noblis Team:

1. Sort strains on a characteristics:

They sort strains by color-coding each sequence based on each characteristic. The picture below shows one example of the sorting by color on severity. Their types of graphics is different from us that we use a rectangular multi-alignment panel and they have a sunburst plot. This plot does not deal with positions and it just shows each sequences in a tree diagram with phylo-genetic information that we did not deal with in our solution. The main point is they use color to sort sequences. I would like to change the positions of the strains in a multi-alignment for sorting, and possibly using color to redundantly show this sorting. (It would create three regions)

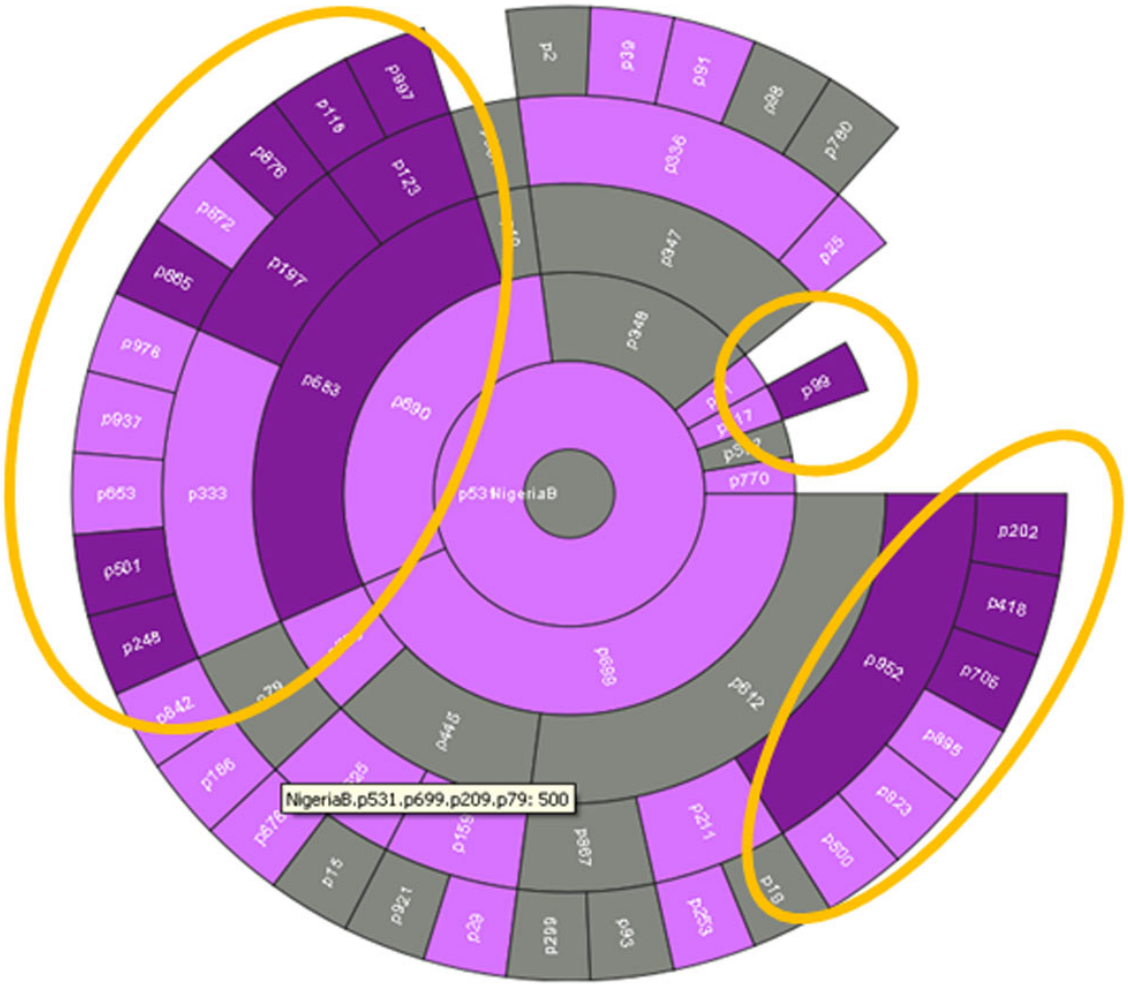

Figure 3: sunburst plot highlighting severity of symptoms. Noblis team used directed variable (severity) to color different sequences.

They also have another plot called polar plot to display different strains as well as positions in them.

Again their polar plot they did not sort the sequences, instead they color each sequence's SNPs based on that directed attribute (severity). So those SNPs along with each diameter have the same color that comes from the range of the severity in the characteristics table.

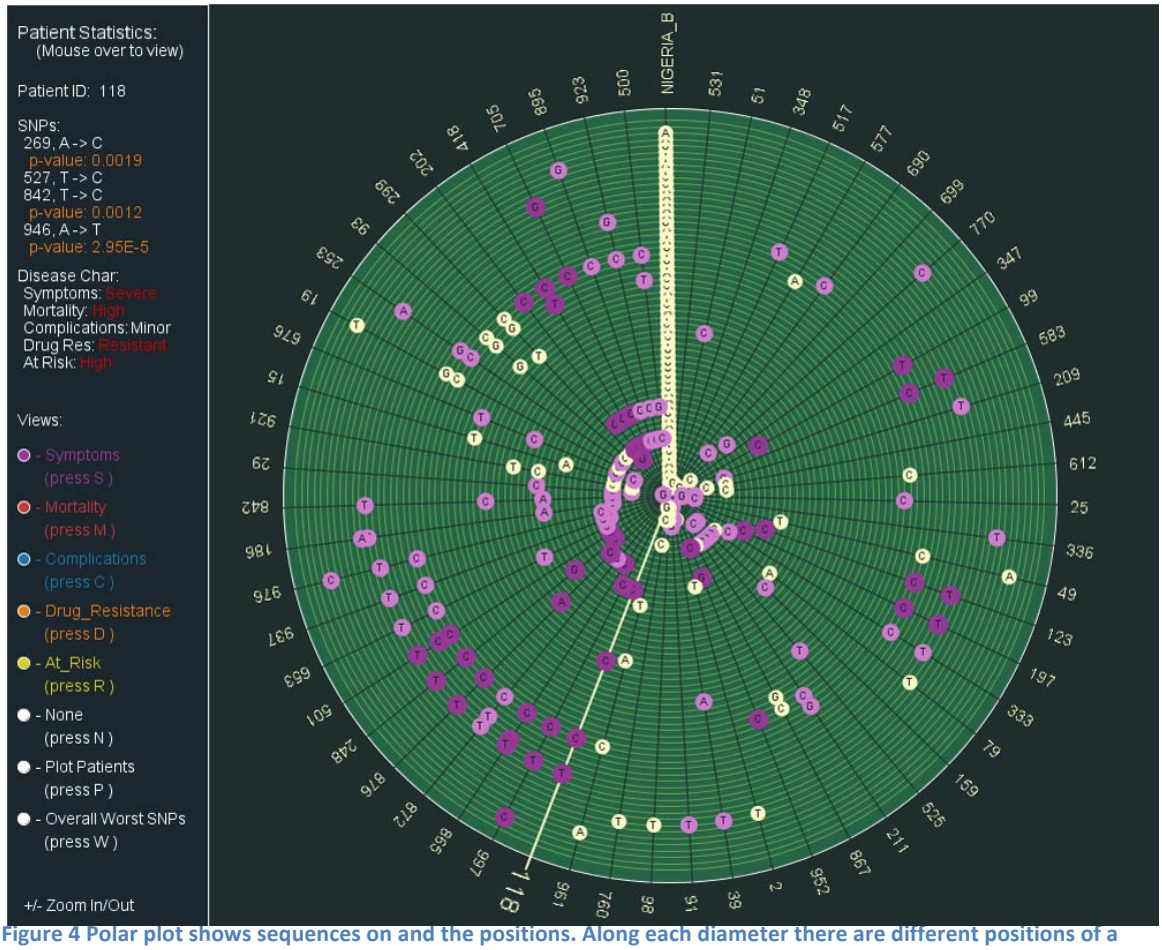

Figure 4 Polar plot shows sequences on and the positions. Along each diameter there are different positions of a partcular sequence, labled on the outer round.

2. How to change sort:

As I mentioned above, they sort sequences but not by changing their positions on the plot, instead they use the color for each variable. On the left panel they listed the different characteristics so that by selecting each of those the color would be different based on the dynamic range of that variable.

3. The ability of directing the user to find interesting positions:

One bioinformatician in their team spent two days examining the association between SNPs and symptom severity using the Mann-Whitney U test (in R).

They used the same graphics but with different approach of color-coding. Now they color each SNP along a diameter based on its own specific P value. The most significant p values are colored in red and the least significant are colored in yellow. So now the

color along a circle is the same because the p value for SNPs in each position is the same. They also use their previous sorting method by coloring each line in a diameter for sorting each strain. So now each sequence has a color based on a derived charatcristics (overall severity).

Overall severity = Sum of all the charactristics

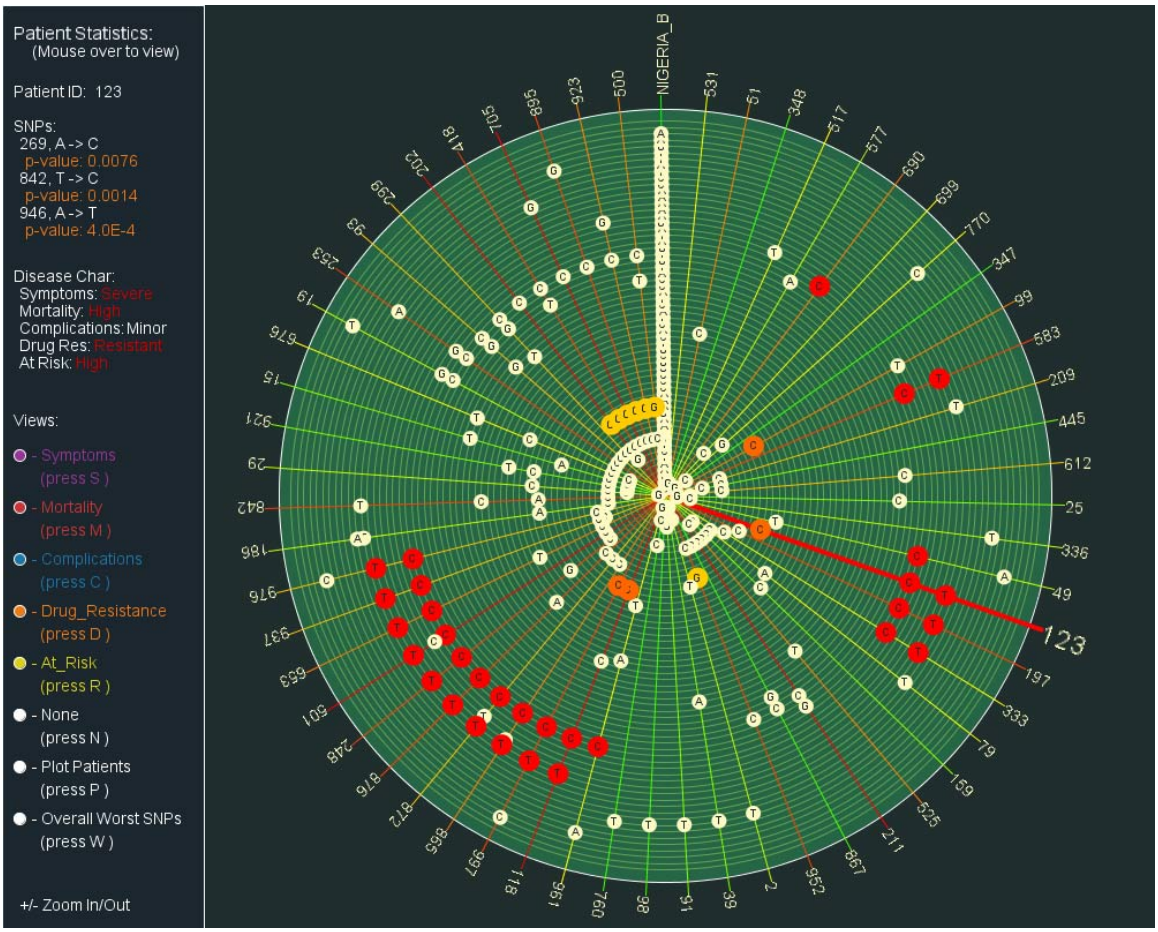

Figure 5 Polar plot color-coded for adifferent perpose

Their work could be translated in IMAS by approach in IMAS by coloring each column's SNP's. Then also color each line, or a boarder for each sorted regions base on the derived variable overall severity. However I do not think color is the most effective channel to encode this statistics number for each position. Although I would keep coloring for the sorted regions as it shows different levels of a characteristics.

They did not talk about the details of their statistical analysis. This is the process that seems that they did:

We want to see if there is any association between SNPs and characteristics. Lets pick one of the characteristics for example the overall danger. For each position a certain SNP happens at some of the sequences or does not happen. Lets split the sequences into two groups for a specific position: Those who have this SNP, and those who do not.

They suggest we use Mann-Whitney U test, I think it is because the severity is an ordinal variable (8 being the most dangerous strain and 1 is the least dangerous) .
The null hypothesis would be there is no significant difference between the mean of the two groups. The P value is the probability of Ho being correct.
For example for the position 842:

SNP) 8,8,8,7,7,7,7,7,6,6,6,5,5,5,4,4,3

Not SNP) 8,8,8, 7,7,7,7,7,6,6,6,6,6,5,5,5,5,4,4,4,4,4,4,3,3,3,3,3,3,2,2,2,2,2,2,2,1,1,1,1,1

Mann-Whitney U test:

P-value for the test is 0.0015. So with alpha 0.5 this P-value is very small, and the chance of Ho being correct is so low. So we reject the null hypothesis and it seems there is a difference in overall severity between those who that SNP appears on them in that position and who do not. Then we say the less P-value means the more responsible this position for being responsible for that disease overall severity in those sequences that they have it!