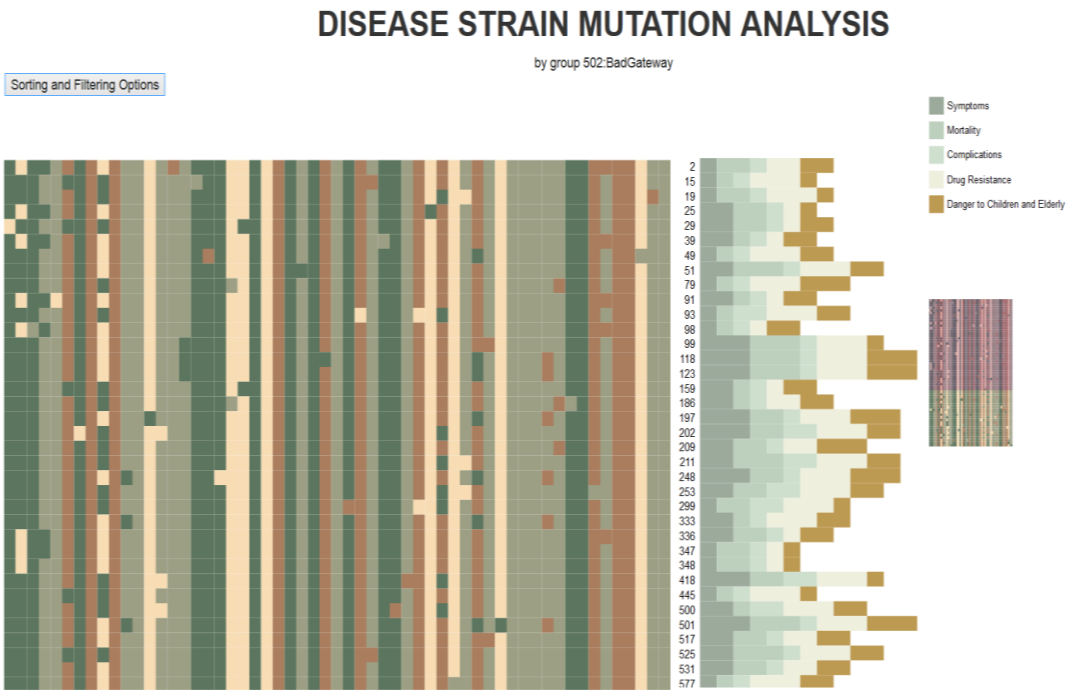
Project Design: Phase II

Group name: 502 bad gateway



Members name:

Li Jiaqi,

Timothy Elbert,

HE Junfeng,

ZHUANG Yufan;

Date: 6/12/2015

What Problem Are We trying to Solve?

The raw data given contains genetic sequences of 58 viral mutations of a Drafa virus which has broken out in Africa. Each strain has 1404 bases and 5 different attributes that: Symptom severity, tendency of mortality, severity of complications, strength of drug resistance, and the danger it poses to the weak in society, primarily children and the elderly. That comes out to roughly 82 thousand pieces of information, which is really difficult to look at, much less spot patterns and draw conclusions from without data mining. Our goal was to make a visualization of this data in a simple and intuitive way to allow people to explore this mass of data and be able to spot patterns at a glance that might give insight that would allow them to make predictions in the future. The problem we were ultimately trying to solve was the difficulty of analyzing the data in its raw form.

Our final design is radically different from our original one. At first we wanted to display the mutations in a hierarchical structure to visualize the order in which the mutations may have happened. However, upon running our program we realized that all the mutations came from a single original strain and none of them could be said to be the predecessor of any other.

Our Final Design

Our final visualization is a table of the gene sequences of the viral strains. The table intuitively plots each base (A,T,G, or C) based on the strain that it belongs to and its position in the sequence. The base type is encoded with color. To the right of the main display of the gene sequences, we have horizontal stacked bars to reflect the attributes of each strain. There are five different colored bars in each stack, each color representing one of the 5 attributes, explained in the legend to the right of the stacked bars. The length of the bar represents the severity or danger of the corresponding attribute for that strain. This makes for an easy comparison of the overall danger presented by the strain, as the longest bars have the more severe attributes. In between the stacked bars are the strain identification numbers, lined up properly with the stacked bars and the DNA table to allow an easy connection between the two. This covers the data display.

What makes this visualization worth anything is the utility provided to explore the data. Beneath the legend is a miniature version of the main DNA display with a translucent dragable window. Since the display is rather large, it does not entirely fit on the page. Having a miniature version of it on the side allows the user to see it whole and potentially notice something of interest, such as an anomalous block of DNA. Then the user may use the dragable window to reposition the main display to that point.

Above the main display is a collapsible sorting and filtering interface. It allows the user to sort the strains based on the attributes or the overall danger. It also allows the display to be filtered by the severity of attributes and even by sections of DNA the user suspects may be relevant.

Implementation

To generate this visualization we used D3.js. To use this, we needed to convert the data into proper files. We generated JSON files using java. The data was read from the generated JSON and the provided TSV and then combined into a single workable array. It was then used to generate a modified heat-map and stacked bars using the update function, which filters and sorts the data before generating anything. We wrote all our own sorting and filtering functions as well.

One of the most difficult parts to create proved to be the miniature display on the right most side. At first we tried to encode it as a view of the main display, but this was disastrous because if the main display was moved, so was the miniature window. Around this time, we realized that display all 82,000 bits of information was infeasible. The browser would stop responding and updates from filtering and sorting would take 10 seconds at a time. Chrome refused to work at all. We realized that it was completely unnecessary to display all the entire DNA sequence. So we modified the JSON to only contain the parts of the strains that were not the same across all 58. Once we did that, all the lag issues dissipated and it became practical to simply generate another heat-map in miniature, which would stay stationary and not lose its perspective.

Getting rid of all the identical columns in the table proved to not only remove all lag from the display, but also to make sense with what we were trying to accomplish. Our goal was to help find patterns in the DNA and to possible identify sections of the DNA responsible for certain attributes. If a section of DNA is the same across all strains, it holds no relevant information for these tasks.

Findings

Although we were not able to find a single section wholly responsible for an entire attribute, which was an unreasonable expectation but would have been the best case scenario, the visualization proved very effective at generating questions and then easily answering them with a little exploration. These are the findings 1 of us got after only 20 minutes of exploring the visualization:

The base at position 108 is T across most strains, but there is a significant minority of strains where it is actually C. We found that if a strain has base C as opposed to T in that position then one can confidently say that:

- Only minor complications may be expected form
- Symptoms will not be severe
- The strain will either be susceptible to drugs or, unfortunately, be highly resistant to them, no middle ground.
- The good news is that this strain does not pose a high danger to the weak of a population

If a strain contains G at position 21, as opposed to C, then you are in luck because it will be pretty harmless with no extreme attributes, mostly mild, and only a few moderate. This is all the more important because the majority of strains are resistant to drugs, but not ones with this mutation.

If a strain presents a high danger to the weak, then the good news is, the complications expected from this strain will be minor. The reverse cannot be said, however.

Symptoms being sever means more bad news because in addition to the symptoms, the strain will not be susceptible to drugs and mortality will be medium to high.

If the base at position 945 is T instead of A, then it will not be easily susceptible to drugs and will probably be very resistant to them, and will pose a non-negligible danger to the weak. However, it seems to have almost no effect on the overall danger (the combined severities of the 5 attributes).

Position 160 seems to be one of the most evenly split positions with about half being G and half being C. However, it seems to be completely irrelevant which base is at that position as it does not affect the attributes.

Playing around with the display is really fun and patterns are pretty easy to find. Unfortunately, as we are not biologists, we will clearly miss certain implications of the genetic code and differences, but we can spot correlations just the same. It might be interesting to see if an expert's opinion on the Drafa strain matches the results we find in the visualization.