Project Proposal

Team: Junhao Wang, Weiren Lai

Describe what you intend to do

Visually show how the Covid-19 spike protein sequence will change during the pandemic.

1. At different times and locations, look at the specific mutations in the spike protein sequence and visualize them.

2. To understand the changes in the sequence of the spike protein by looking at the number of different changes in the specific mutation on the spike protein in different time frames.

3. To understand the changes of the spike protein sequence by looking at the mutation sequence that is different from the known dominant sequence in a fixed time period.

Describe the methods you plan to use or develop

The data visualization method is used to show the change process of the spike protein. The main steps are:

1. Data collection. To have accurate analysis results, you must have high-quality data. The method of data collection and the quality of the data largely determine the final effect of data visualization. The official data is used here, which can be obtained by downloading or crawling.

2. Data processing. Data processing and data transformation are prerequisites for data visualization, and data needs to be cleaned and denoised. By preprocessing the data collected in the previous period to filter out the useless fields, large error data and invalid data, several key data indicators needed for this project are obtained. The methods used here are generally dimensionality reduction, clustering, segmentation, sampling, decision trees, etc.

3. Visual mapping. This is the core of the entire data visualization process, the process of mapping processed data information into visualization elements.

The visual element is composed of 3 parts: visual space, mark, and visual channel.

- The display space of data visualization is usually two-dimensional, but it can also be formed into three-dimensional through graphics rendering technology.
- Markers are the mapping of data attributes to visual geometric graphic elements and are used to represent the classification of data attributes.
- The visual channel, the mapping of the value of the data attribute to the visual presentation parameters of the mark, is usually used to display the quantitative information of the data attribute.

By combining the "marker" and "visual channel", the data information can be completely visualized.

4. Human-computer interaction

The purpose of visualization is to reflect the value, characteristics and patterns of the data, and to present the information hidden behind the data in an intuitive way. So the common interaction methods are:

- Scrolling and zooming: When the data cannot be fully displayed on the device of the current resolution, scrolling and zooming are a very effective way of interaction, such as the information details of maps and line graphs.
- Color mapping control: users can configure the color of the visual graphics according to their preferences.

- Control of data mapping methods: A data set generally has multiple sets of features. Users can choose dimensions according to their preferences or choose specific dimensions to explore the hidden information behind the data according to the target requirements.
- Data level of detail control: you can hide the details of the data, and it will only be displayed by clicking the trigger.

Describe the data you will use and discuss how you plan to obtain it

For this project you will be provided with a set of Covid-19 spike protein data collected between January 1st and August 7th of 2020. The data include the following attributes:accession ID,which is a code assigned to each published sequence by the popluar data repository Genbank, RNA sequence, collection date, and for some sequences collection location.These data all come from all the sequences in the data set we collected in the United States.

The method of obtaining data can be directly from the official website to download the official data set or through the "web crawler" to obtain the data set.

Discuss relevant background work

The covid-19 pandemic that broke out at the beginning of this year has had a serious impact on our lives and caused a global economic setback. In order to better control the epidemic, many scientists have done their best to understand the structure and composition of the virus in order to better formulate effective plans for it. One of the key proteins required by a virus to infect its host is called a spike protein. Its job is to bind to certain receptors in the host cell to access its machinery and achieve infection and reproduction. These spike proteins are considered to be one of the key goals of vaccine and antiviral drug development, so understanding their structure can provide valuable insights for fighting viruses.

Discuss your tentative plan

The visual method is used to display the changes of the covid-19 spike protein sequence. First explain the source of the data and how to obtain it. Secondly, it is necessary to explain in detail the data cleaning process and the useful data finally filtered out. Third, the data needs to be visualized, and the changes in the spike protein over time are displayed through commonly used line graphs, histograms, pie charts, etc. Finally, analyze their internal connections and the information behind them based on the visualized data. Make detailed explanations based on data information and present them in the form of reports.