Jason McGough

Michael Wilder

Waid Wrege

Thomas Wynn

Introduction to Bioinformatics

Project 2: Understanding the Ebola Virus 2014 Outbreak

Project Report

**Abstract**

The Ebola virus is posing a huge threat to the world. It is not very well understood, and is constantly evolving. In order to better understand how the virus works, and how it may change in the future, pairwise alignment is a very powerful tool we can use. We can use it to compare sections of the Ebola virus to find regions of similarity that

may be a consequence of functional, structural, or evolutionary relationships between the sequences.

We have designed a program to perform pair wise alignment for us, and output any gaps, mismatches or insertions that have taken place between two different strands. These differences can then be compared to data we already have in order to find out what outcomes these differences may cause (synonymous or nonsynonymous mutations). This project we’ve completed has shown us that the Zaire Ebola virus is much more similar to the Ebola Viruses we are seeing now in 2014.

**Introduction**

The length of the Ebola Virus genome is 18959 (made of RNA). It is made up of seven proteins (see ref. 3). These seven proteins make up the seven genes named GP, L, NP, VP24, VP30, VP35 and VP40 (see ref. 6). Eukaryotic protein-coding genes, such as Ebola, use a Poly(A) signal when making copies of itself (see ref. 5).

A quick literature survey will reveal that many people all over the world are concerned about this Ebola epidemic. The United States Government was so concerned, it declared an all out war on it on September 16, 2014. The Liberian government has set up a command center in Monrovia with The United States Military (see ref. 2). In the past, Ebola started in remote regions of central Africa, and didn’t spread further. This outbreak is significantly larger than any other in the past, and started in Guinea, West Africa. This time, it has spread to numerous other countries, and is causing many more causalities (see ref. 1). That is why so many people are worried about the damage this virus has caused, and frightened as to how much more damage I may end up causing. In order to aid with spreading information about the virus, and to help prevent the virus from infecting more people as fast, the CDC has been updating their website with the latest information on the virus and what steps to take to help with the situation (see ref. 4).

**Materials and Methods**

First, a script was written to download sequences. It then reads in data from the files. The sequences are split by description comments, so parsing is needed to get the useful information. The algorithm used to do this has a complexity of O(n2). The information comes from http://www.ncbi.nlm.nih.gov. To avoid having to read from the website every time, after the information is read in the first time, the files are saved to a repository. Once this information has been obtained, the sequence alignment is run on it. The program goes through the file and parses out each individual gene to compare them.

**Implementation**

Our implementation only needs to consider two rows at a time, unlike using dynamic programming methods that would require calculating the entire table. The implementation works in the following way:

We save the last row of scores from comparisons, and then add each column of scores with the last two rows. Then find the max score of all those additions. At the index of this max score, we split the second string. At this point, take the first half of the first string, and compare it to the second string. Then reverse the second half of the first string, and compare it to the second string reversed. This function is recursively called until one of the strings is one or zero character long. At this point, we start aligning.

The base case is one of the strings being one or zero characters long. If one of the strings is zero characters long, the entire other string is aligned with nothing but gaps. If one of the strings is one character long, then we run the Needleman–Wunsch algorithm to get the alignment.

**Results and Discussion**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Zaire  Ebola Virus | Insertions | Deletions | Synonymous Mutations | Nonsynonymous Mutations | Matches | Mismatches |
| 2014  GP Gene | 0 | 119448 | 15484 | 316 | 69994 | 632 |
| 2014  L Gene | 0 | 11297 | 167309 | 6570 | 511184 | 13297 |
| 2014  NP Gene | 79 | 59329 | 51032 | 1898 | 171029 | 4272 |
| 2014  VP24 Gene | 0 | 69362 | 13193 | 158 | 59011 | 713 |
| 2014  VP30 Gene | 79 | 48294 | 20619 | 474 | 67624 | 790 |
| 2014  VP35 Gene | 0 | 27887 | 23778 | 296 | 79947 | 870 |
| 2014  VP40 Gene | 0 | 41396 | 23699 | 949 | 76466 | 1033 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sudan  Ebola Virus | Insertions | Deletions | Synonymous Mutations | Nonsynonymous Mutations | Matches | Mismatches |
| 2014  GP Gene | 474 | 18249 | 11218 | 7347 | 52298 | 17854 |
| 2014  L Gene | 8453 | 7979 | 108846 | 53577 | 369859 | 146169 |
| 2014  NP Gene | 3950 | 3713 | 33652 | 18883 | 123712 | 47718 |
| 2014  VP24 Gene | 790 | 790 | 12798 | 6004 | 42579 | 16355 |
| 2014  VP30 Gene | 1501 | 1501 | 12798 | 7742 | 46373 | 20619 |
| 2014  VP35 Gene | 3871 | 1264 | 15879 | 7821 | 55458 | 21488 |
| 2014  VP40 Gene | 1185 | 1185 | 15883 | 8291 | 55825 | 20462 |

git add filename

git commit -m "I worked on blah blah blah"

git push origin master

literature survey: what people are saying about it

**References:**

1. http://www.sciencemag.org/content/345/6202/1369.full.pdf

2. http://news.sciencemag.org/africa/2014/09/u-s-declares-war-ebola-epidemic

3. http://viralzone.expasy.org/all\_by\_species/207.html

4. http://www.cdc.gov/vhf/ebola/

5. http://genesdev.cshlp.org/content/25/17/1770.full

6. http://en.wikipedia.org/wiki/Ebola\_virus#Genome