Ebola virus multialignment Analysis and visualization

Paulina Hyzy, Jakub Tyrek, Norbert Dojer
paulinahyzy@gmail.com, jakubtyrek@gmail.com, dojer@mimuw.edu.pl

UNIWERSYTET WARSZAWSKI

Introduction

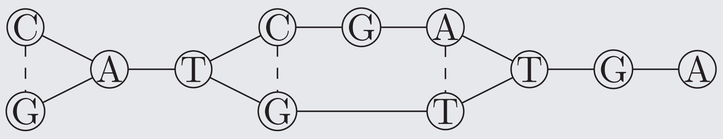
Multiple sequence alignment is an information-rich object. Our aim is to partition its component sequences into clusters and extract consensuses.
A powerful way to achieve this result is to use graph representation of multiple alignment which is not only computationally efficient but also, allows visual analysis of the alignment structure.

Graph Idea

Graph representation of multiple alignment is based on partial order alignment graph[1], which is concise and intuitive. It reflects the multiple alignment structure better than typical approaches like MAF files or alignment browsers.

The basic idea is to merge aligned nucleotides which are the same into single nodes, create directed edges between subsequent nodes and undirected edges between aligned but different nucleotides.

```
CATCGATGA
GATG-TTGA
```



Consensuses can be find in such a graph by using heaviest bundle algorithm implemented in software called *poa*[2]. It reads multiple alignment in POA graph format and effectively finds possible consensuses.

Future Work

Visualisation: Development of the graph visualisation, especially to make it more interactive.

Import/Export: More formats of input and output files will be introduced in order to assure broad software compatibility.

Data: The approach will be tested with other species whose genome graphs contain cycles.

References

[1] Lee C., Grasso C., Sharlow M.F. *Multiple sequence alignment using partial order graphs*, Bioinformatics (2002) 18 (3): 452-464.

[2] Lee C. *Generating consensus sequences from partial order multiple sequence alignment graphs*, Bioinformatics. (2003) 22;19(8):999-1008.

[3] Haeussler M, Karolchik D, Clawson H, Raney BJ, Rosenbloom KR, Fujita PA, Hinrichs AS, Speir ML, Eisenhart C, Zweig AS, Haussler D, Kent WJ. *The UCSC Ebola Genome Portal.*, PLOS Currents Outbreaks. 2014 Nov 7 . Edition 1. doi: 10.1371/currents.outbreaks.386ab0964ab4d6c8cb550bfb6071d822.

Acknowledgements

This work was supported by the National Science Centre, Poland, under grant number 2016/21/B/ST6/01471.

Methods and Data

Data used in this research come from USCB Ebola Portal[3]. There is a multiple alignment (in MAF format) available, generated from 158 Ebola and 2 Marburg viruses genomes coming from all over the world, which were sequenced at different times. The alignment has no cycles. According to the authors', the genomes can be split into 7 groups:

- Ebola 2014 (101)
- DRC 2007 (20)
- Zaire(DRC) 1967-7 (8)
- Bundibugyo 2007 (8)
- Reston 1989-90 (9)
- Sudan 1976 (11)
- Marburg 1987 (2)

In order to analyze this multiple alignment: generate consensuses, partition data into subsets and compare the division with the one discribed above, the available data must have been converted into POA graph and used as *poa*'s input. The output has been visualized and summarized in a readible manner.

As a consequence of taking this approach, a tool for performing the above operations was developed. It is also possible to acquaint with some online results:

Single Alignment Block Analysis

An example block of the Ebola virus multiple alignment (ca. 3008 - 3200bp) can clearly demonstrate the multiple alignment visulisation and consensuses generation:

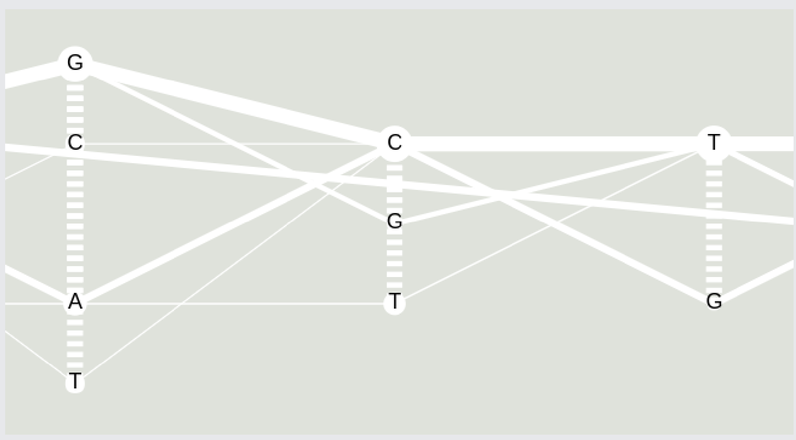


Figure 1: Example POA graph fragment

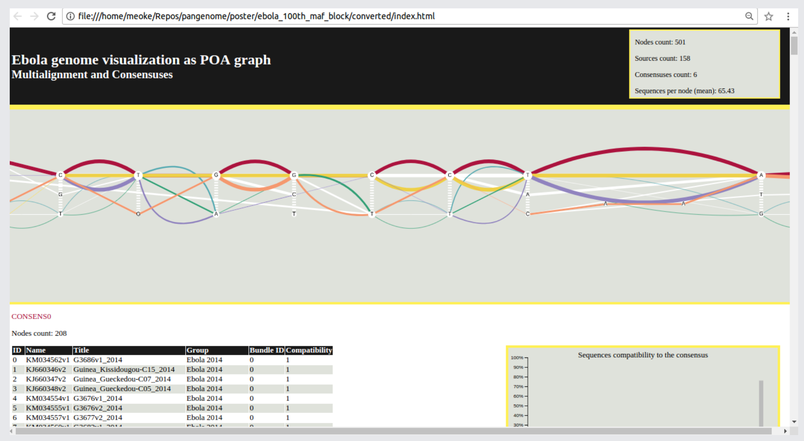


Figure 2: Example with generated consensuses

The tool output is not only the POA graph visualization (Fig.1) but also source sequences and generated consensuses summary (Fig.2). It is possible to assess, whether the original Ebola virus genomes clustering is common with the one being this method's result.

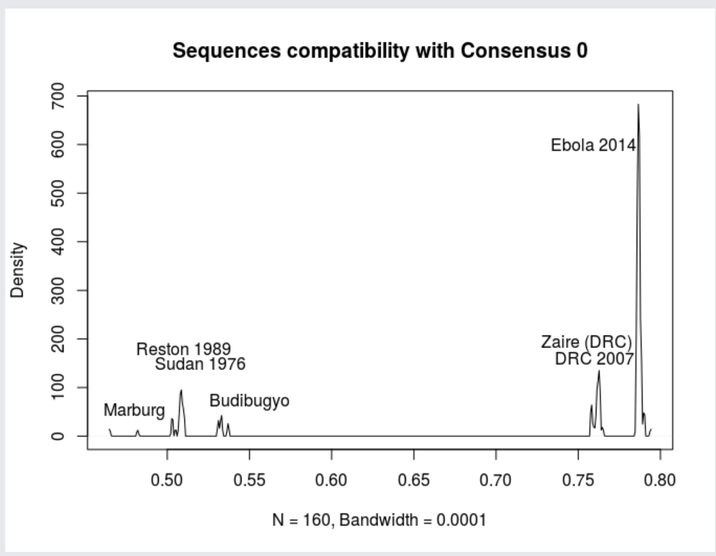
Whole Multiple Alignment Analysis

The described before approach applied to the whole Ebola virus multiple alignment resulted in 6 consensuses, which is quite close to the original 7 clusters. A threshold of 0.75 must have been exceeded in order to classify a source sequence as mapped to the particular consensus ID.

As one can see, the sequences where quite correctlly distinguished. However, some of them (Ebola 2014, DRC 2007, Zaire (DRC)) are so similar to each other that they were mapped to the same consensus. There is another powerful analysis possible - a measure of compatibility of every single source sequence with given (e.g. 0th) consensus.

		Consensuses ID						
		0	1	2	3	4	5	-1
Ebola groups	Ebola 2014	101	0	0	0	0	0	0
	DRC 2007	20	0	0	0	0	0	0
	Zaire (DRC)	8	0	0	0	0	0	0
	Bundibugyo	0	0	4	2	2	0	0
	Reston 1989-90	0	0	9	0	0	0	0
	Sudan	0	11	0	0	0	0	0
	Marburg	0	0	0	0	0	1	1

Table 1: Confusion matrix



By looking at the density plot of compatibility values, it is easy to notice that even if some sequences group were mapped to the same consensus ID, the distance between them and the consensus is different. The local maxima correspond, more or less, to the original genomes distribution. Combination of different analysis can result in accurate genomes partitioning and finding adequate consensuses.