Pan-genome structural analysis and visualisation

Paulina Dziadkiewicz, Jakub Tyrek, Norbert Dojer
pedziadkiewicz@gmail.com, jakubtyrek@gmail.com, dojer@mimuw.edu.pl

UNIVERSITY OF WARSAW

Introduction-OK

Multiple sequence alignment is an information-rich object. Our aim is to analyze its component sequences by building a tree which consists of consensuses extracted from the alignment.

A powerful way to achieve this result is to use graph representation of multiple alignment. It can be examined visually and be processed efficiently.

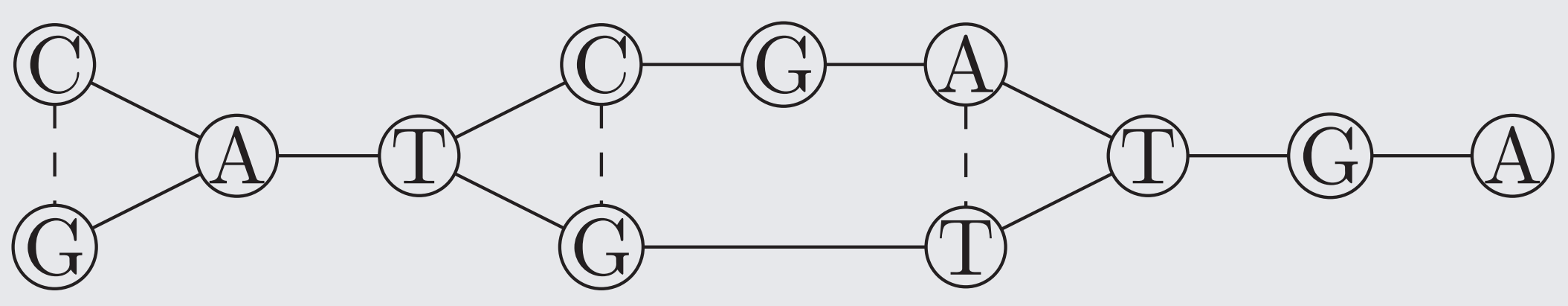
Graph Idea-OK

Graph representation of multiple alignment is based on partial order alignment graph[1]. It reflects the multiple alignment structure in a more concise and intuitive way 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 the POA graph and effectively finds possible consensuses.

Future Work-OK

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.

Algorithms: Development of fast algorithm for handling cycles in genome graphs.

References-OK

[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 et al. *The UCSC Ebola Genome Portal.*, PLOS Currents Outbreaks. 2014 Nov 7 . Edition 1. doi: 10.1371/currents.outbreaks.386ab0964ab4d6c8cb550bfb6071d822.

[4] *Mycoplasma genomes phylogeny* https://www.patricbrc.org/view/Taxonomy/2093#view_tab=phylogeny, Accessed: April 2018

Acknowledgements-OK

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

How to get the tree of consensuses?-TODO

Firstly, convert a multiple sequence alignment (eg. MAF file) into POA graph **G**. The consensuses tree is being built in a breadth first manner. In the following iterations, on the subsequent subgraphs **SG** of the original graph **G**, the procedure of consensus generation and sequences assignment is carried out:

1. Run consensus generation algorithm on the **SG** to get a consensus **C**
2. Choose the most compatible to **C** sequences from **SG** and generate a consensus **BestC** for them only.
3. Set group **S** of the most compatible to **BestC** sequences from the **SG**.
4. Add **BestC** with sequences **S** to the consensuses tree and assign the minimum compatibility among **S** to **BestC**.
5. Remove from **SG** sequences **S** and go to 2 if any sequences are left.

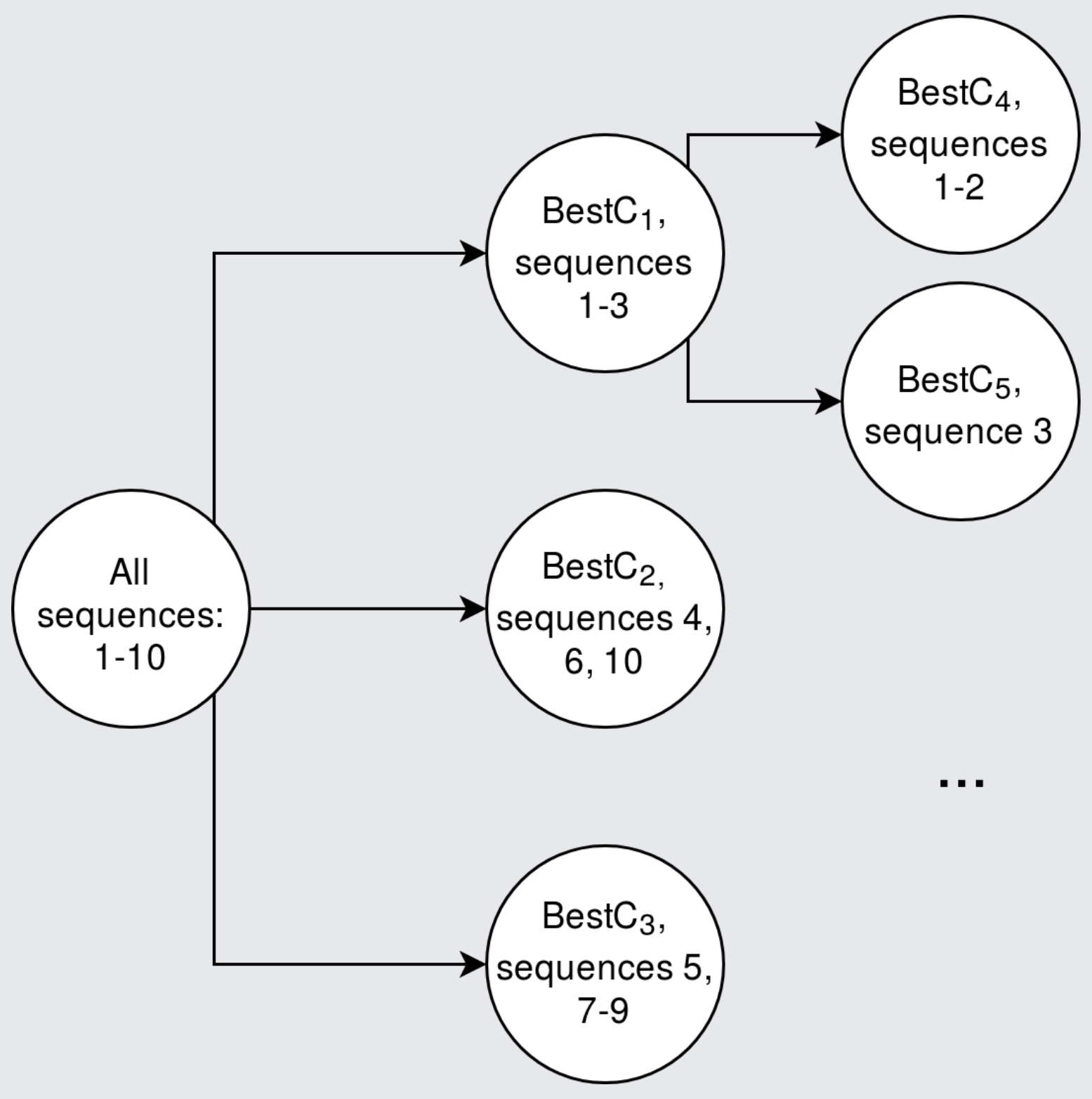


Figure 1: Tree of consensuses

Use the above schema to split group of sequences and assign a best consensus for them, until required level of fragmentation is reached.

Single Alignment Block Analysis-TODO

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

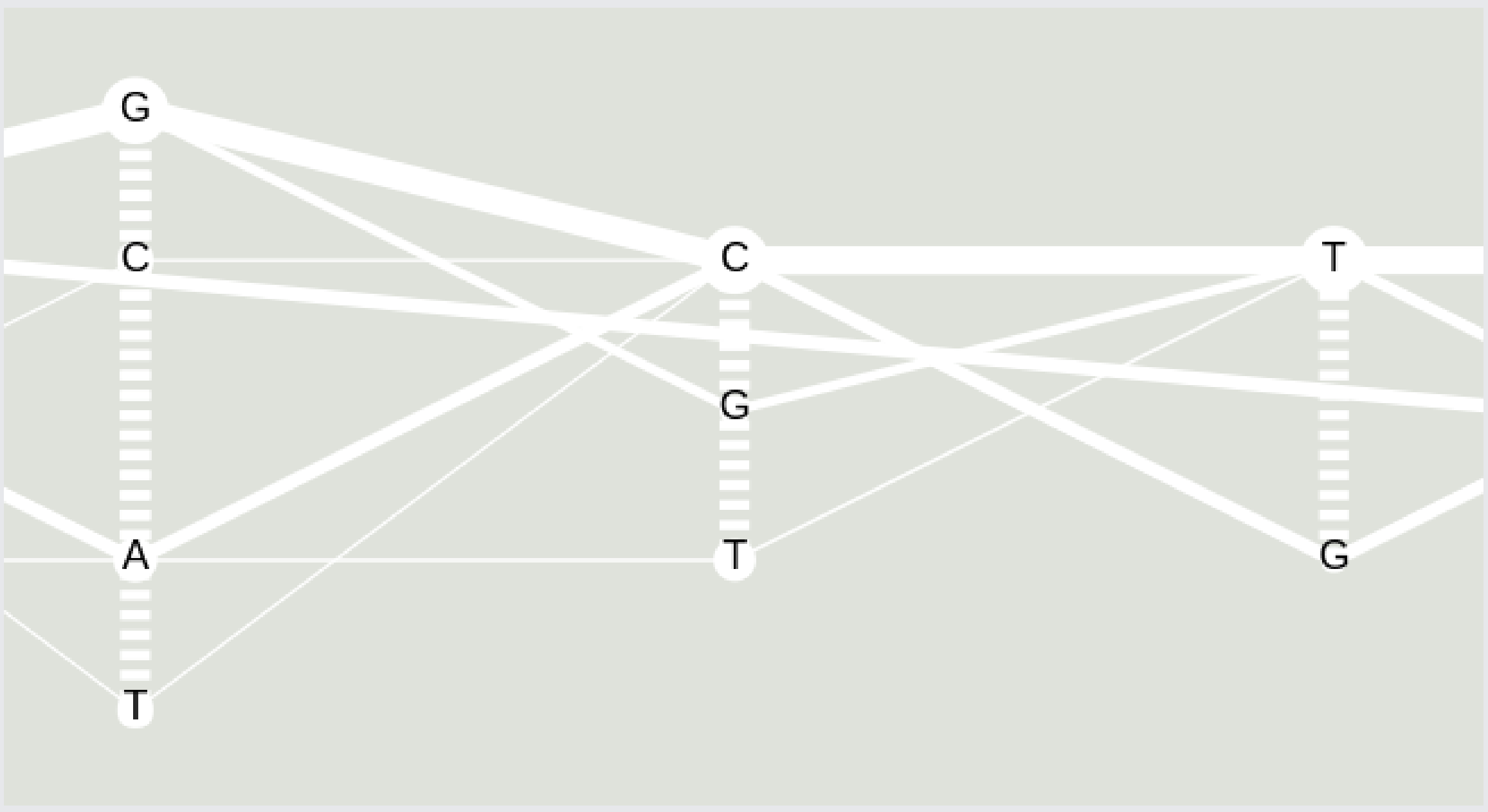
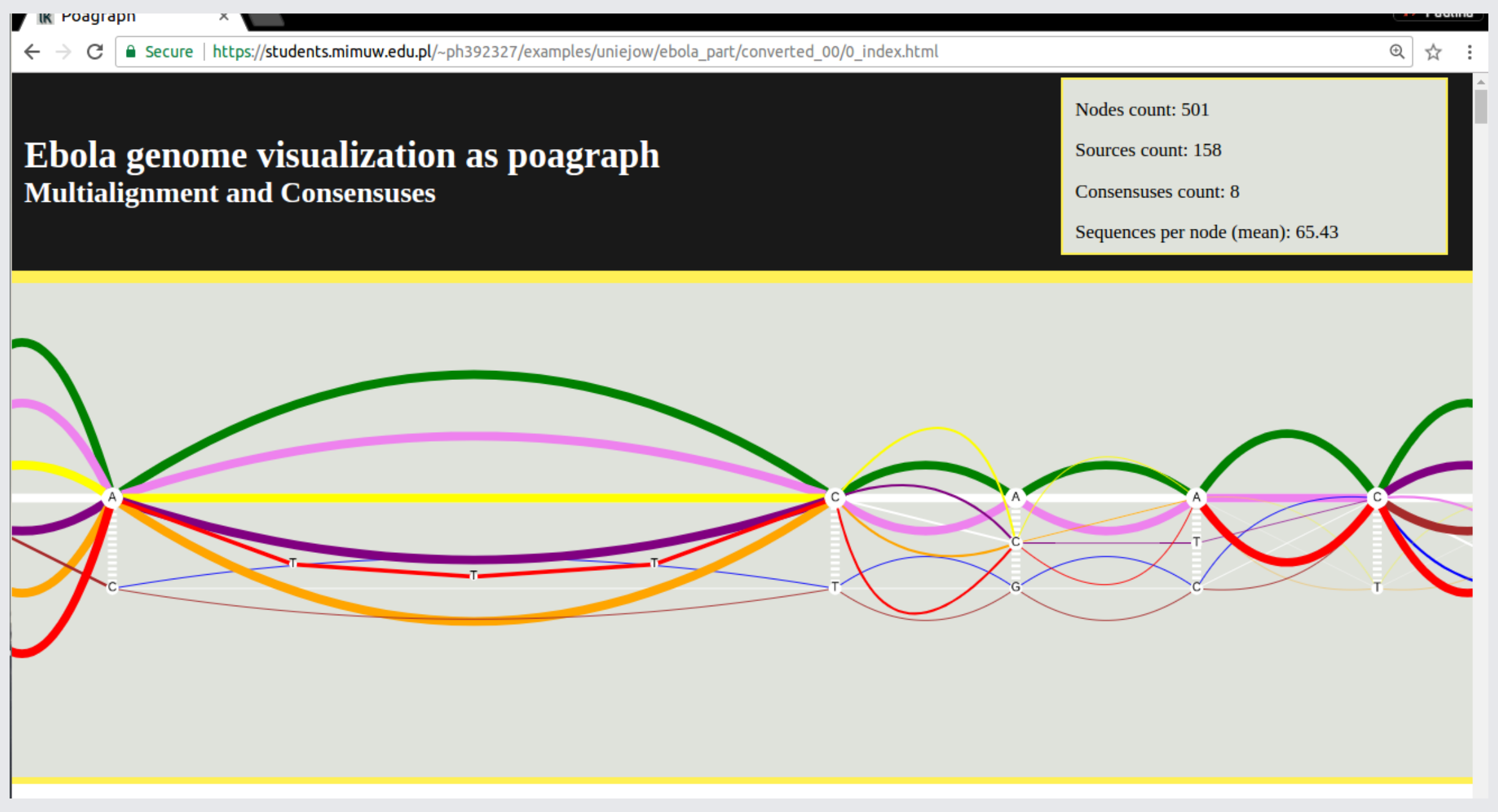


Figure 2: Example POA graph fragment



https://students.mimuw.edu.pl/~ph392327/examples/uniejow/ebola_part/converted_00/0_index.html

Ebola genome visualization as poagraph
Multialignment and Consensuses

Nodes count: 501
Sources count: 158
Consensuses count: 8
Sequences per node (mean): 65.43

Figure 3: 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.

All Alignment Blocks Analysis-TODO

The above described approach applied to the whole Ebola virus multiple alignment resulted in 17 consensuses:

		Consensuses ID																
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Ebola groups	Ebola 2014	101	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	DRC 2007	0	0	9	10	0	0	0	0	0	0	0	0	0	0	0	0	1
	Zaire (DRC)	0	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Bundibugyo	0	0	0	0	0	0	0	4	0	0	0	0	2	2	0	0	0
	Reston 1989-90	0	0	0	0	0	2	2	0	0	2	0	3	0	0	0	0	0
	Sudan	0	0	0	0	4	0	0	0	3	0	3	0	0	0	0	0	1
	Marburg	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0

As one can see, the sequences where quite correctly distinguished. However, some groups (e.g. ”DRC 2007”, ”Bundibugyo”) were split into smaller groups. This uncovers heterogeneity inside the original groups. Here is the example ”DRC 2007” and the dates of origin of its sequences which explains the received partition:

Genomes from 2007		Genomes from ca. 1995		Genome from 2002	
Name	Cons. ID	Name	Cons. ID	Name	Cons. ID
Luebo43_2007	2	1Mbie_Gabon_1996	3	Ilembe_2002	16
Luebo5_2007	2	Gabon_1994	3		
Luebo4_2007	2	1Eko_1996	3		
Luebo9_2007	2	1Oba_Gabon_1996	3		
Luebo23_2007	2	Zaire_1995	3		
Luebo1_2007	2	13709Kikwit_1995	3		
Luebo0_2007	2	1Ikot_Gabon_1996	3		
034-KS_2008	2	13625Kikwit_1995	3		