Pan-genome structural analysis and visualisation

UNIVERSITY OF WARSAW

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

Introduction

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.

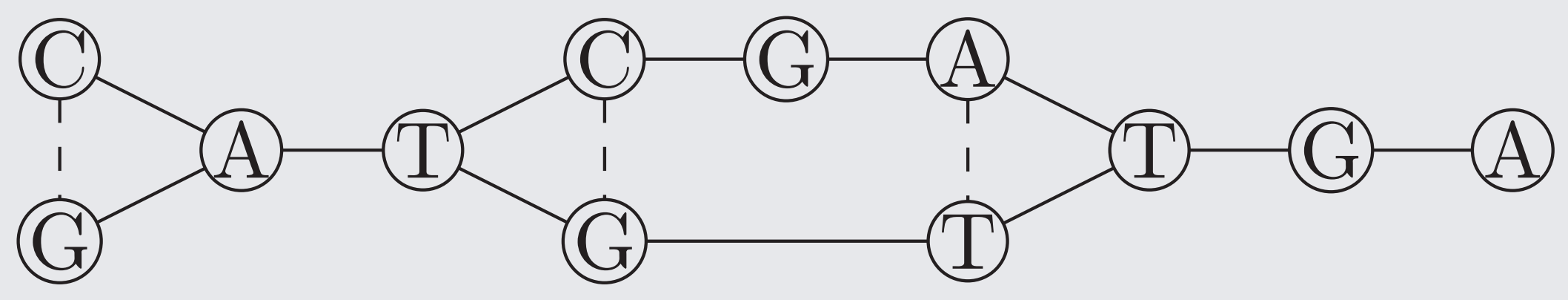
Graph idea

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

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

[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?

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 to it the minimum compatibility **Comp** to **BestC** among **S**.
5. Remove from **SG** sequences **S** and go to 2 if any sequences are left.

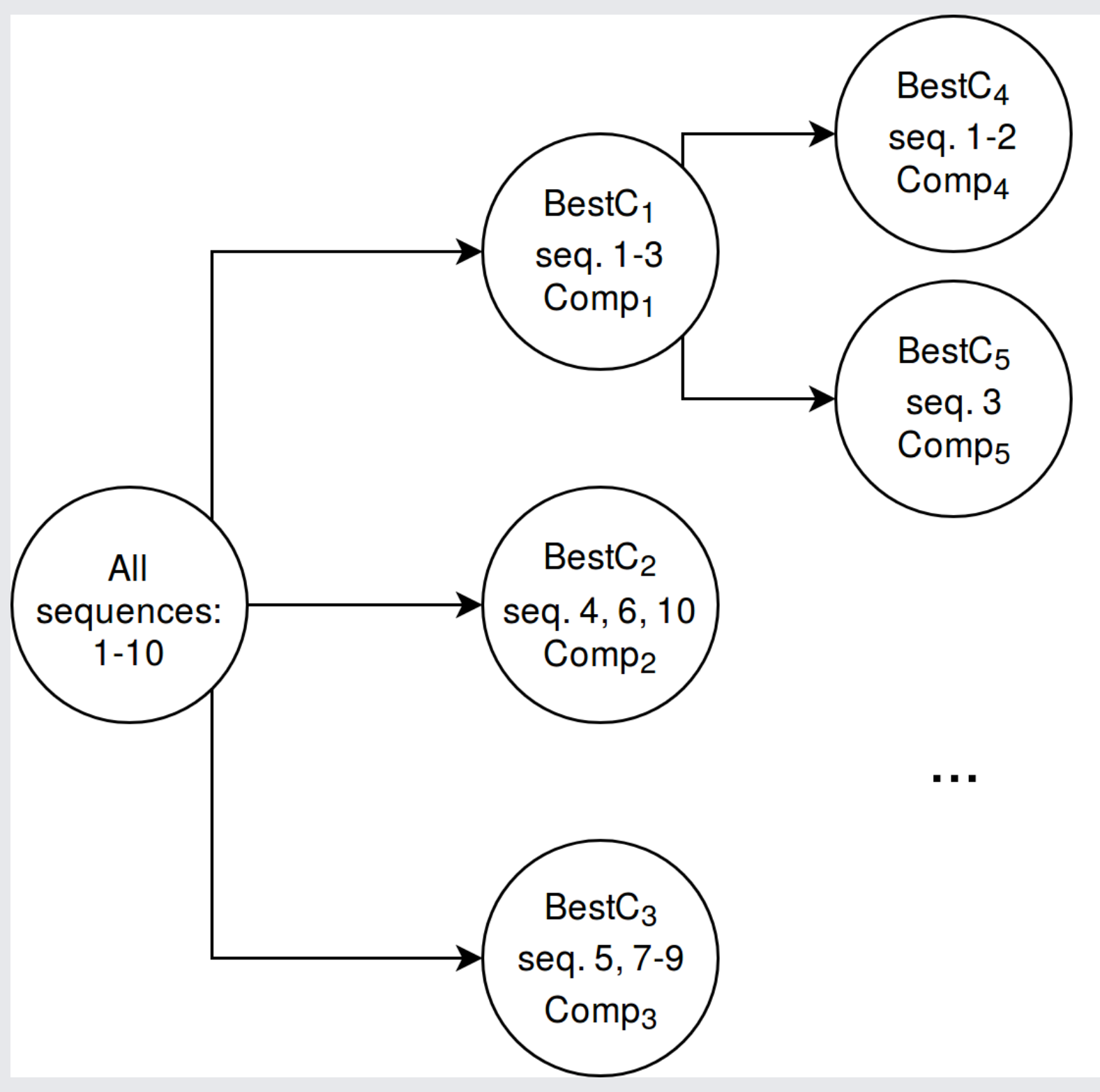

All sequences: 1-10
BestC$_1$ seq. 1-3 Comp$_1$
BestC$_4$ seq. 1-2 Comp$_4$
BestC$_5$ seq. 3 Comp$_5$
BestC$_2$ seq. 4, 6, 10 Comp$_2$
...
BestC$_3$ seq. 5, 7-9 Comp$_3$

Figure 1: Tree of consensuses

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

Data and results

The above solution was tested on two multiple alignments. One contains 159 genomes of the Ebola virus and the other one 7 genomes of the bacteria Mycoplasma (only a part of the genomes was used due to efficiency issues). The results were confronted with existing taxonomic databases [3][4].

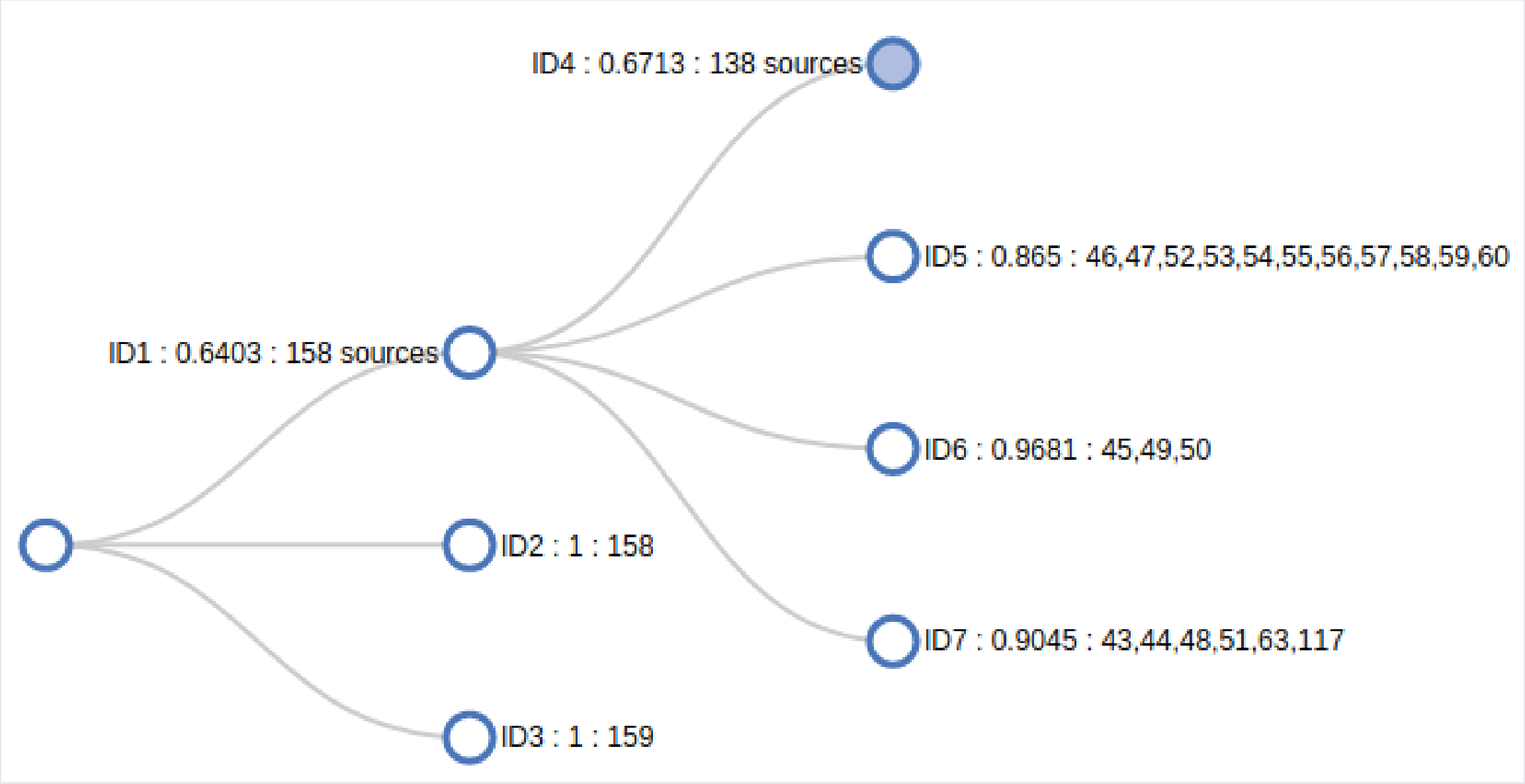


Figure 2: Tree of consensuses generated for Ebola

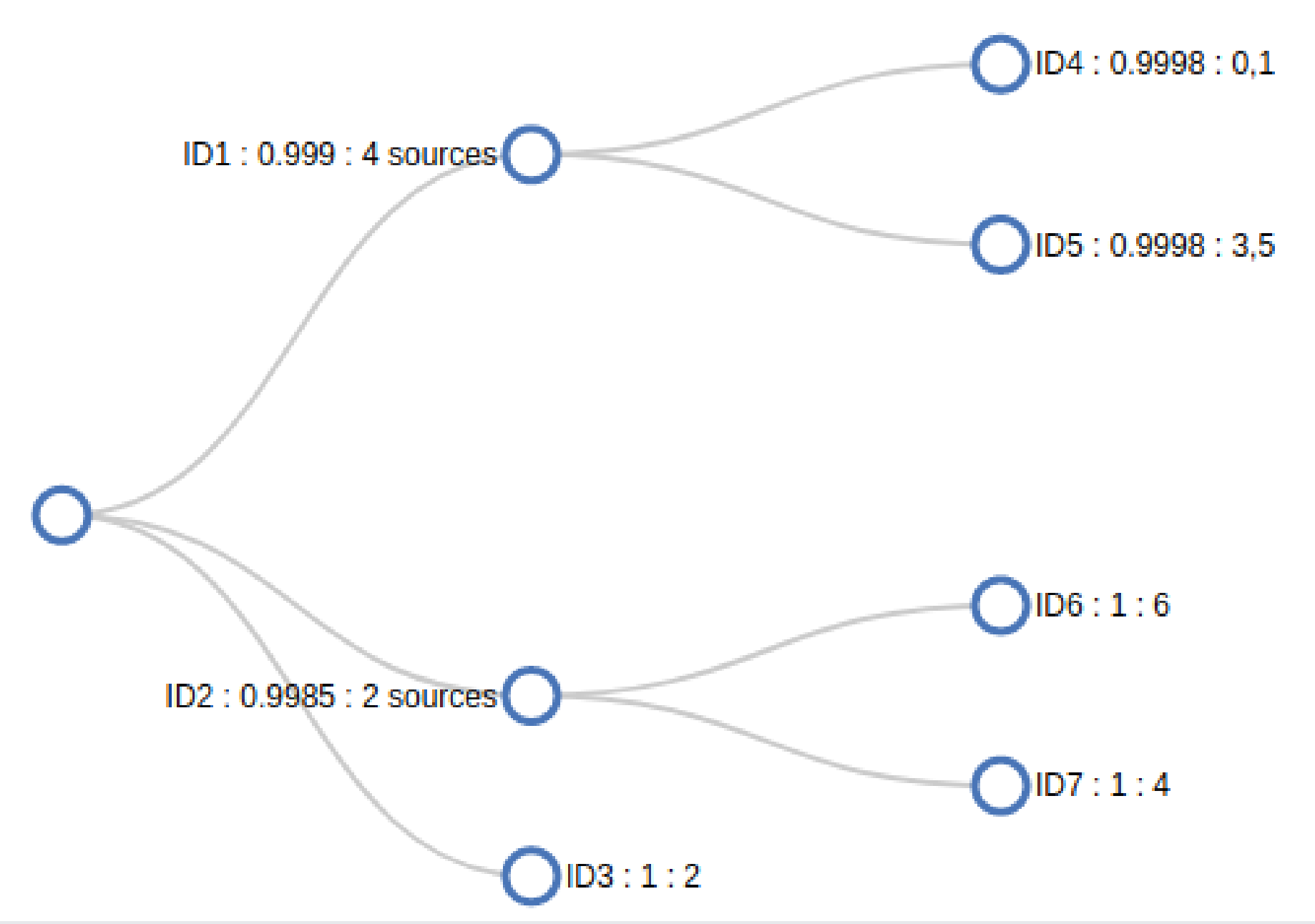


Figure 3: Tree of consensuses generated for Mycoplasma

The received sequences division for both alignments is compatible with the biological knowledge. The results are enriched with visualization and available online:

Visual representation

The high-level purpose of the tool under development is pan-genome analysis and visualisation. Currently the visualization consists of:

- POA graph built from the multiple alignment which shows its structure on a nucleotides level
- interactive tree of consensuses generated with the algorithm described above
- tabular summary of sequences present in the alignment

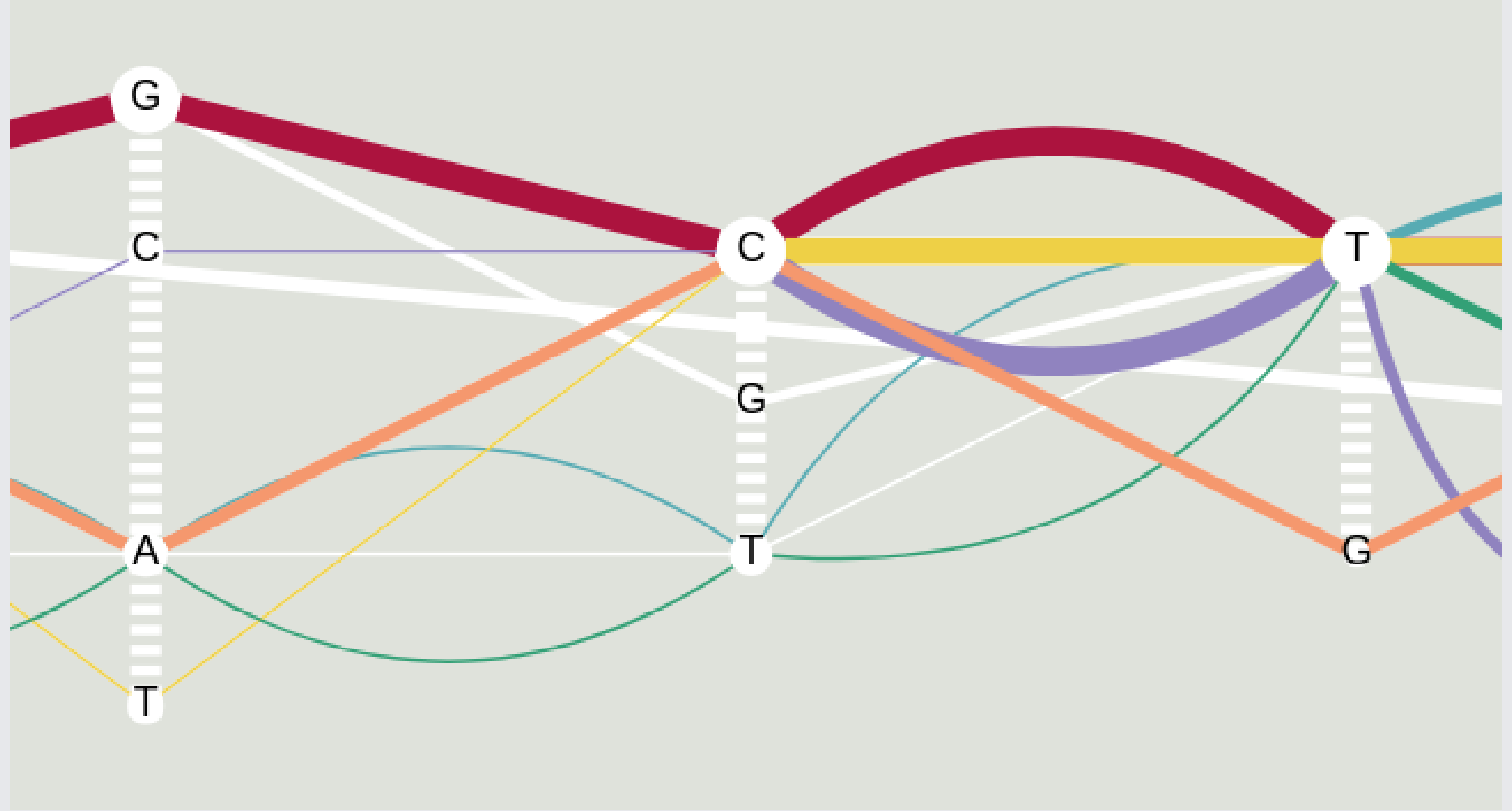
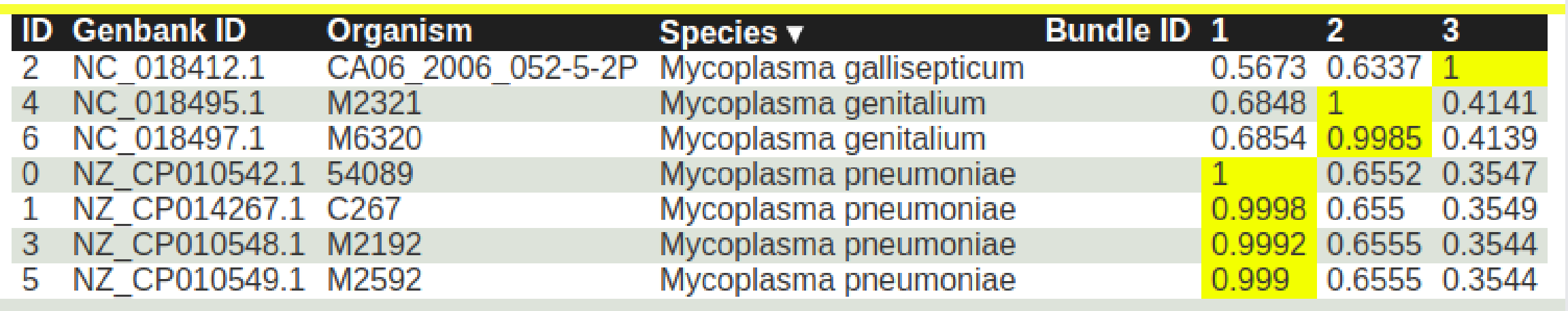


Figure 4: An excerpt of a POA graph.



ID	Genbank ID	Organism	Species ▾	Bundle ID	1	2	3
2	NC_018412.1	CA06_2006_052-5-2P	Mycoplasma gallisepticum		0.5673	0.6337	1
4	NC_018495.1	M2321	Mycoplasma genitalium		0.6848	1	0.4141
6	NC_018497.1	M6320	Mycoplasma genitalium		0.6854	0.9985	0.4139
0	NZ_CP010542.1	54089	Mycoplasma pneumoniae		1	0.6552	0.3547
1	NZ_CP014267.1	C267	Mycoplasma pneumoniae		0.9998	0.655	0.3549
3	NZ_CP010548.1	M2192	Mycoplasma pneumoniae		0.9992	0.6555	0.3544
5	NZ_CP010549.1	M2592	Mycoplasma pneumoniae		0.999	0.6555	0.3544

Figure 5: Example sequences summary for Mycoplasma.