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

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.

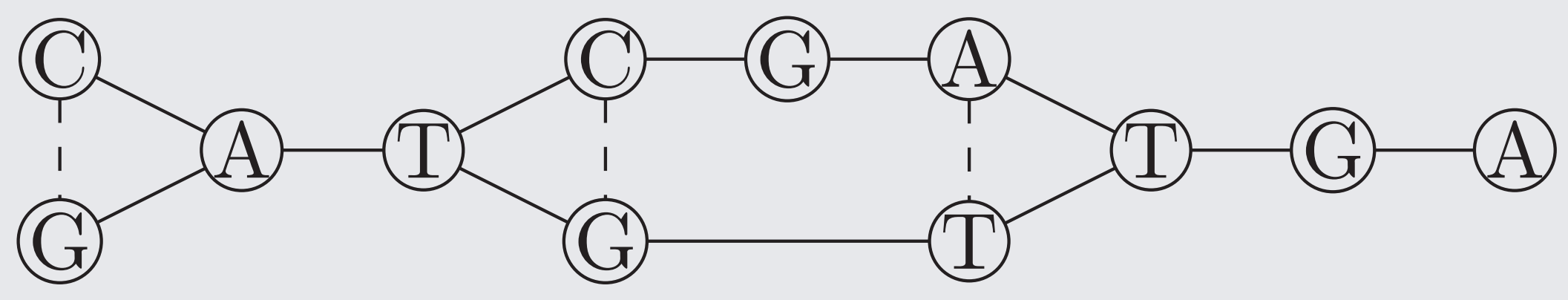
POA graph

Graph representation of multiple alignment is based on partial order alignment graph (POA 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
```

↓



Consensus in a POA graph is a path representing some sequences present in the multi-alignment. By using heaviest bundle algorithm implemented in software called *poa*[2] many consensuses can be defined.

Conclusion

This work represents an approach to analyse **a pan-genome**. An insight into complex multi-alignments is given by **a tree of consensuses**. This is not only an attempt to reconstruct phylogenetic tree but also identification of sequences patterns shared by individuals and analysis of the alignment structure.
Enhancements planned to be undertaken are: a fast algorithm for handling cycles in genome graphs and visualization development.

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.

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

Acknowledgements

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?

The tree of consensuses is being built in the top-down manner from the POA graph **G** representing the multiple sequence alignment. Given a **G**-subgraph **SG** reflecting sequences assigned to the current node, the following procedeure creates this node's children:

1. Run consensus generation algorithm on the **SG** to get a consensus **C**
2. Choose the most compatible with **C** sequences from **SG** and generate a consensus **BestC** for them only.
3. Set group **S** of the most compatible with **BestC SG**-sequences.
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 1 if any sequences are left.
6. Re-assign sequences to **BestC** consensuses.

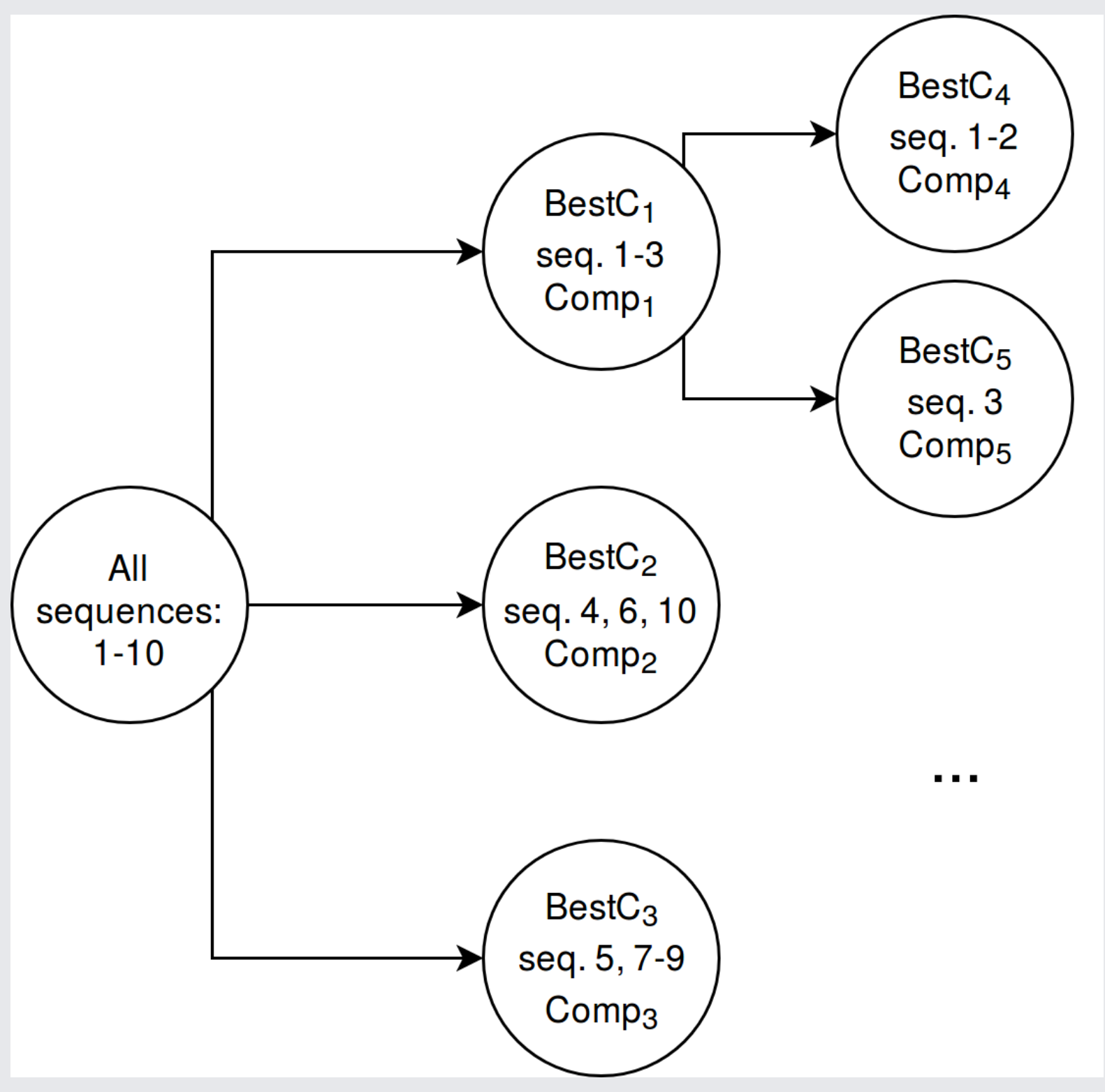
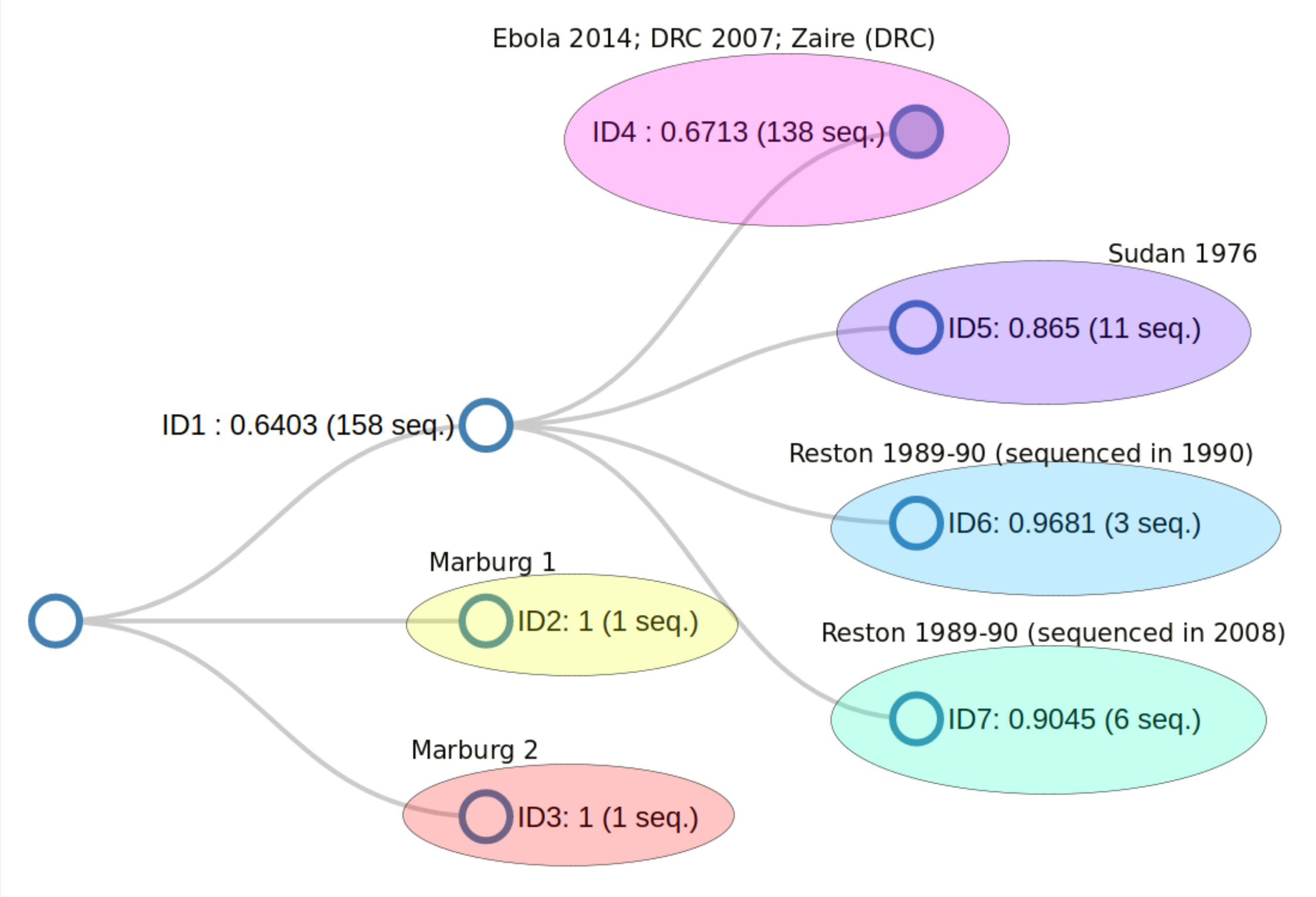

All sequences: 1-10
$BestC_1$ seq. 1-3 $Comp_1$
$BestC_4$ seq. 1-2 $Comp_4$
$BestC_5$ seq. 3 $Comp_5$
$BestC_2$ seq. 4, 6, 10 $Comp_2$
...
$BestC_3$ seq. 5, 7-9 $Comp_3$

Figure 1: Tree of consensuses

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

Results - Ebola

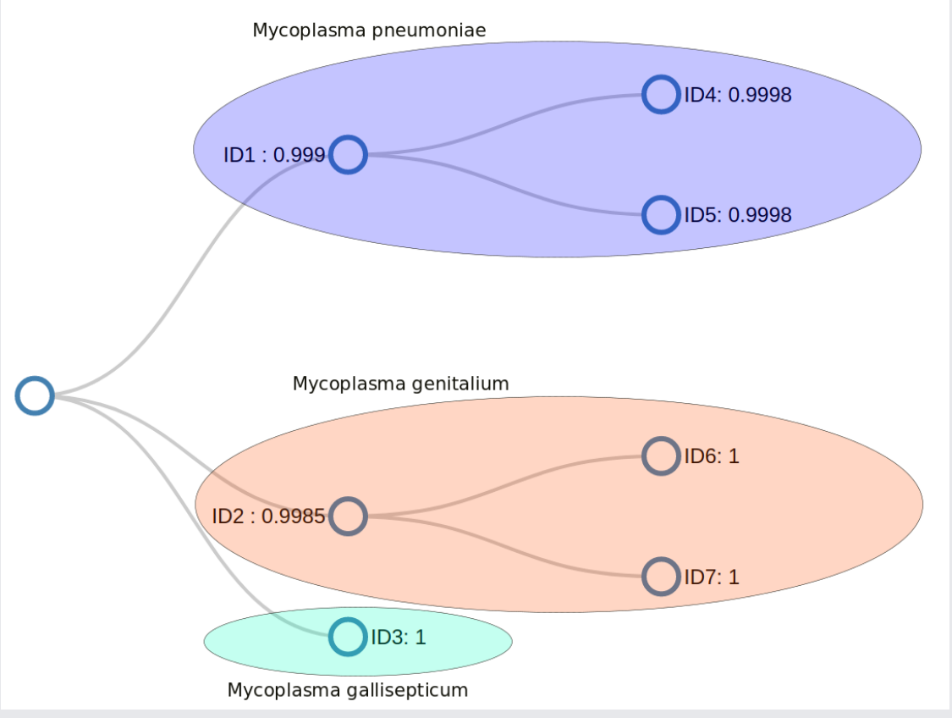
One dataset used in this research comes from USCB Ebola Portal [3]. There is a multiple alignment created for 158 Ebola and 2 Marburg viruses coming from all over the world, sequenced at different times. The received consensus tree is compatible with the biologically substantiated sequences division.



Results - Mycoplasma

The other dataset was a multiple alignment built from 7 genomes of the bacteria Mycoplasma (M. pneumoniae, M. genitalium, M. gallisepticum). Only the part of the genomes could be used that satisfies POA graph requirement for being cycles-free.

The results were successfully confronted with existing taxonomic databases [4].



Visual representation

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

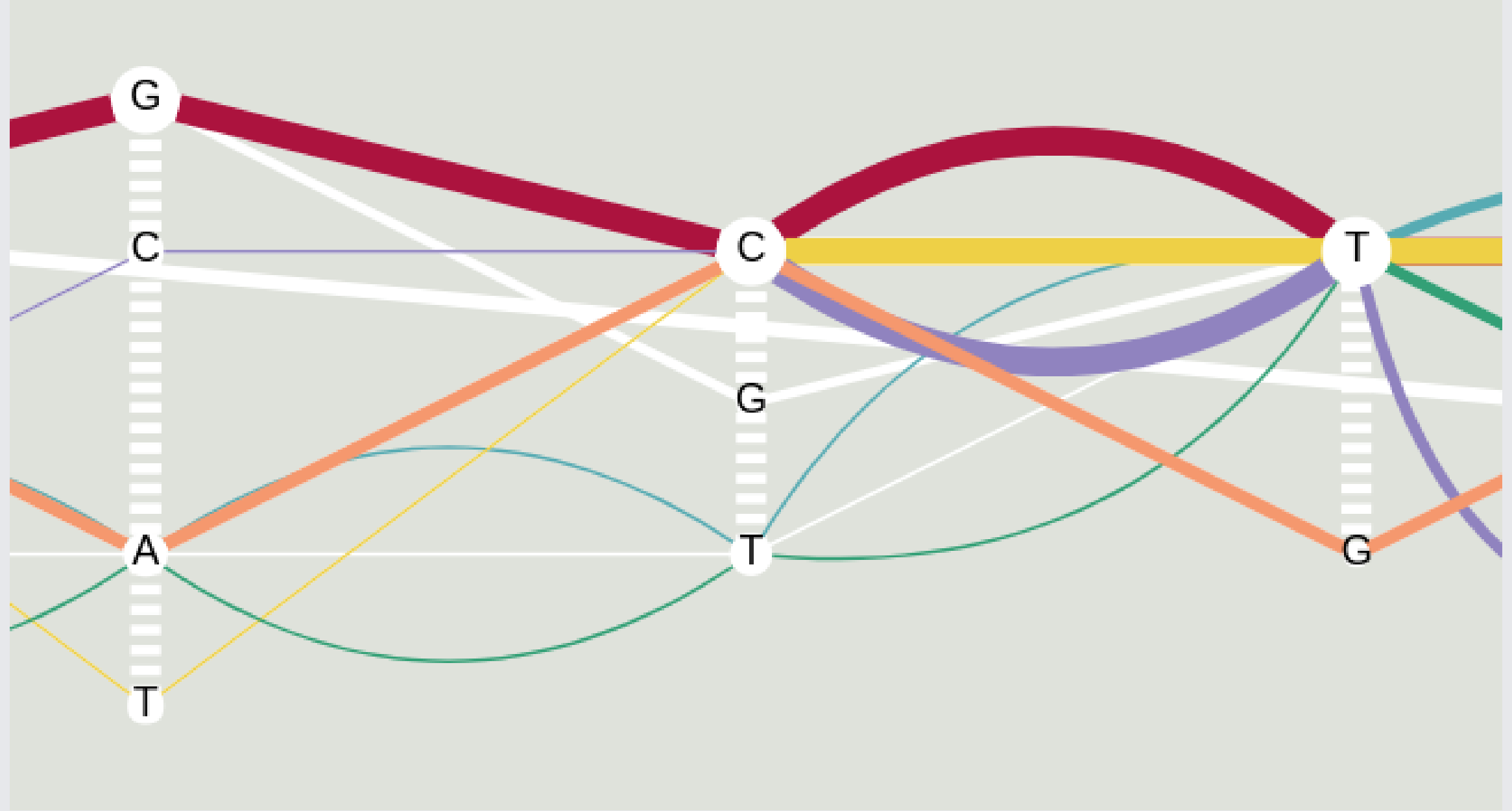
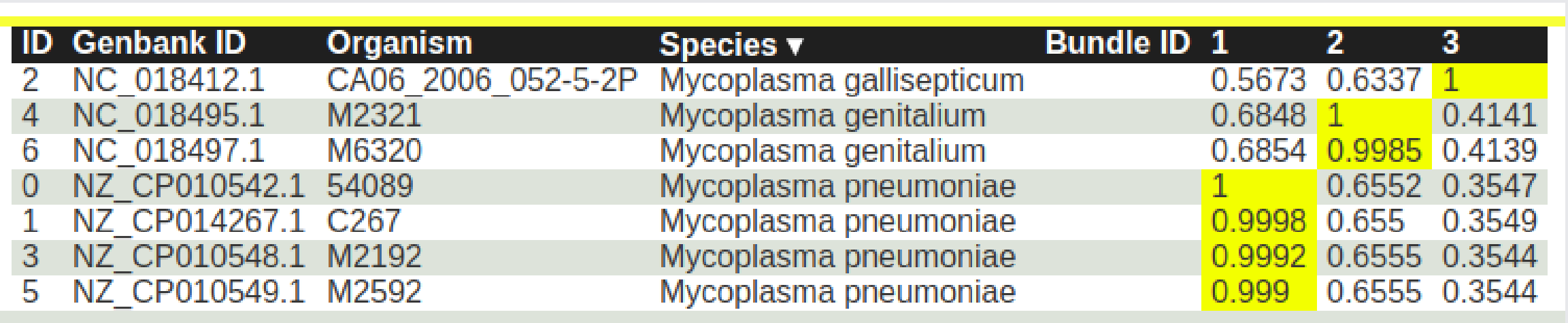


Figure 2: An excerpt of a POA graph.



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 3: Example sequences summary for Mycoplasma.