TME2 - PHYG

Witold PODLEJSKI & Jérémie PERRIN

October 28, 2019

1 Exercise 1 : Genome evolution simulation

1.1 Question 1

Since the evolution of genomes is a stochastic process, it is a good idea to simulate it to compare simulated results to what we observe. In order to simulate accurately such a process, we need a theoretical model of evolution. The statistical accordance of the simulated data with the data observed is one way we can validate a model of evolution, or at least have an idea of the most important effects one needs to take into account.
In order to fully describe the evolution, a model needs :

- A representation of the genome
- A mutation rate (how often does a rearrangement happens)
- A list of rearrangements events and their relative occurence

1.2 Question 2

The choice of parameters is important because the model needs to reflect the evolutionnary process. Or at least should describe a process comparable to that of evolution so as to be able to compare the results of the simulation to the observed data.
If the model fails to describe the evolutionnary process then there would be no sense comparing its result to what we observe. Of course we cannot be sure that the model will indeed represent (even partially) the evolutionnary process. But if we are careful not too have a too complicated model (which might indeed have results corresponding to observation), and still observe concordance with our data than we can safely assume that the model describes some part of what really happens.

1.3 Question 3

We cannot use the same parameters in between the different clades. That is so because the genomes might be very different and in some cases more prone to

rearrangements. One example of such a distinction is noted in the thesis report where thay have to consider different number of events occuring depending on wether they study vertebrates or yeast.

2 Exercise 2 : Implementation of simple events

Cf. folder "code".

3 Exercise 3 : Small dataset simulation

3.1 Question 1

Cf. folder "code/results/[1-10]".

3.2 Question 2

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Inversion	9	12	13	6	8	9	10	11	9	13
Translocation	5	6	3	9	3	4	4	4	7	4
Duplication	0	0	1	2	0	1	8	0	0	0
Deletions	0	1	1	0	3	0	0	1	0	0
Fusions	0	0	0	0	0	0	0	0	0	0
Fissions	0	0	0	0	0	0	0	0	0	0
WGD	0	0	0	0	0	0	0	0	0	0

Inversions are most frequents, out of the events we coded ourselves fission is the least present.

3.3 Question 3

Fissions did not appear in any of the simulations, since it only occurs once out of a thousand event on average. And since we only simulated around 150 events throughout the simulations it is understandable that not once did it occur.

3.4 Question 4-5

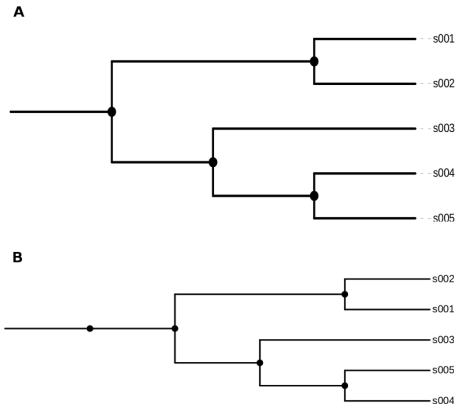

Figure 1: Simulation 3 : (A) Real tree (B) Gene Order Inferred

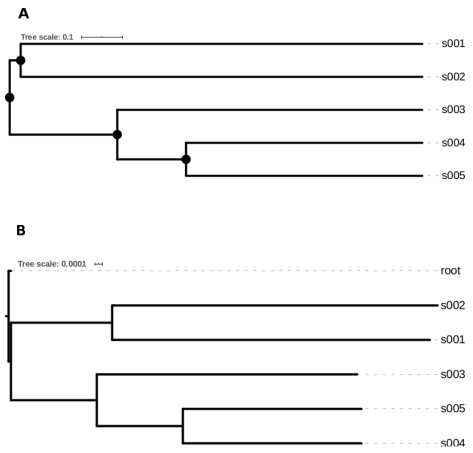

Figure 2: Simulation 7 : (A) Real tree (B) Gene Order Inferred

When we compare the two trees for both simulations, we notice that Gene Order algorithm does a convincing job in finding the tree underlying the simulated evolutions. Although the branch lengths are not exactly accurate, especially concerning the positionning of the root, the topology and speciation events are correctly inferred from the sequences.

4 Exercise 4 : Large dataset simulation

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
Inversion	4656	5415	5086	5617	5315	5103	5056	4928	4847	5002
Translocation	2357	2639	2532	2897	2599	2591	2580	2407	2340	2465
Duplication	384	397	412	484	447	402	465	430	408	440
Deletions	397	464	417	445	444	421	427	433	411	415
Fusions	10	4	5	10	8	4	7	6	6	6
Fissions	7	11	13	4	10	3	12	6	10	12
WGD	1	0	3	1	1	1	2	0	0	1

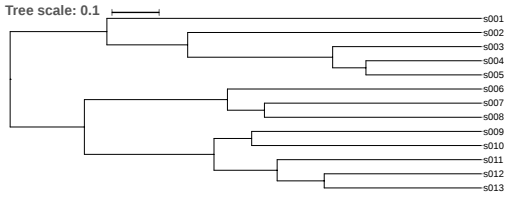

Figure 3: Real tree for the large dataset simulation

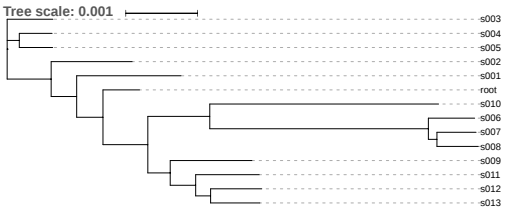

Figure 4: The resulting tree of Maximum of likelihood algorithm for the large dataset simulation

As we can see the Maximum Likelihood Gene Order algorithm did not achieve to construct the right tree topology. This is due to the large amount of recombinaison event, in that case the simplest model is not necessarily the good one.