Phylogenetic inference with Multi-Objective algorithms
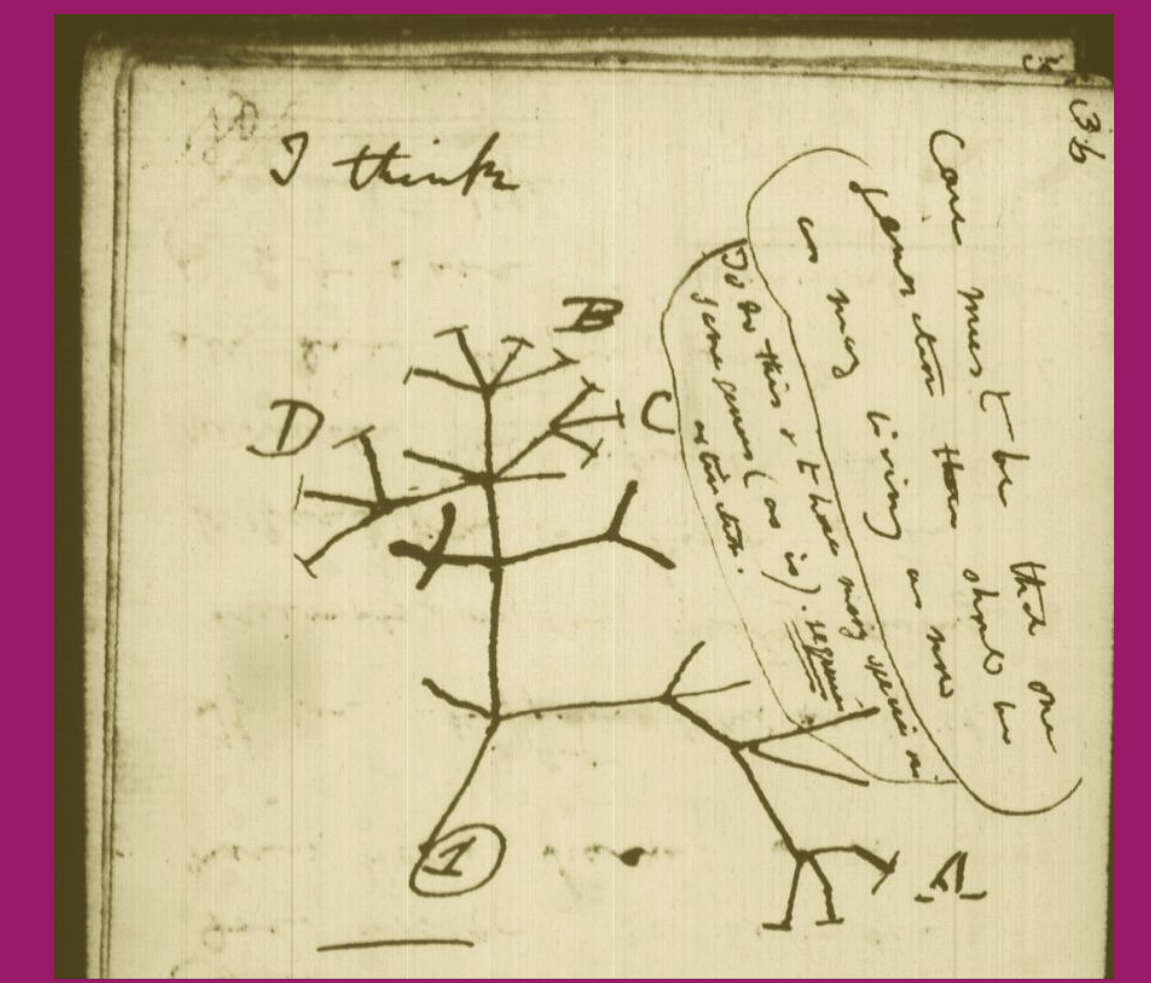
E.T.S.I. Informática

Alicia Fernández Rovira and Álvaro Gómez Jáuregui,
Ingeniería de la Salud, mención en Bioinformática

Introduction

Phylogenetic inference consist in the search of an evolutionary tree to explain the best way possible genealogical relationships of a set of species. Therefore the search of a phylogenetic tree can be formulated as a multi-objective optimization problem, which aims to find trees which satisfy simultaneously (and as much as possible) both criteria of **parsimony** and **likelihood**.

What do we need?

- MULTIFASTA with gene's sequences of the organism we want to infer.
- Compute a MSA (Multiple Sequence Alignement) over the target dataset.
- Convert the output of the MSA process to Newick Hamsphire tree format.
- Run the MO-Phylogenetics tool.

http://onlinelibrary.wiley.com/doi/10.1111/2041-210X.12529/abstract

Algortihm workflow

1. Choose multi-objective evolutionary algorithm.
2. Configure Genetic Operators.
3. Configure Phylogenetic Optimization Strategies.
4. Read input MSA.
5. Read Partitions.
6. Read or Generate initial Population.
7. Execute MOEA (MO-Evolutionary algorithm).
8. Generate best MO results: phylogenetic trees in newick format and Pareto Fronts approximations.

Results

The results of the algorithm workflow are two output files. One containing the **Pareto front's approximation** (Fig.1) which is the representation of the Pareto optimal set in the **objective space** and the other containing the **optimized phylogenetic trees** in newick format.

rbcL_55 data set (55 sequences of the rbcL chloroplast gene of 1314 nucleotides per sequence from species of green plants)
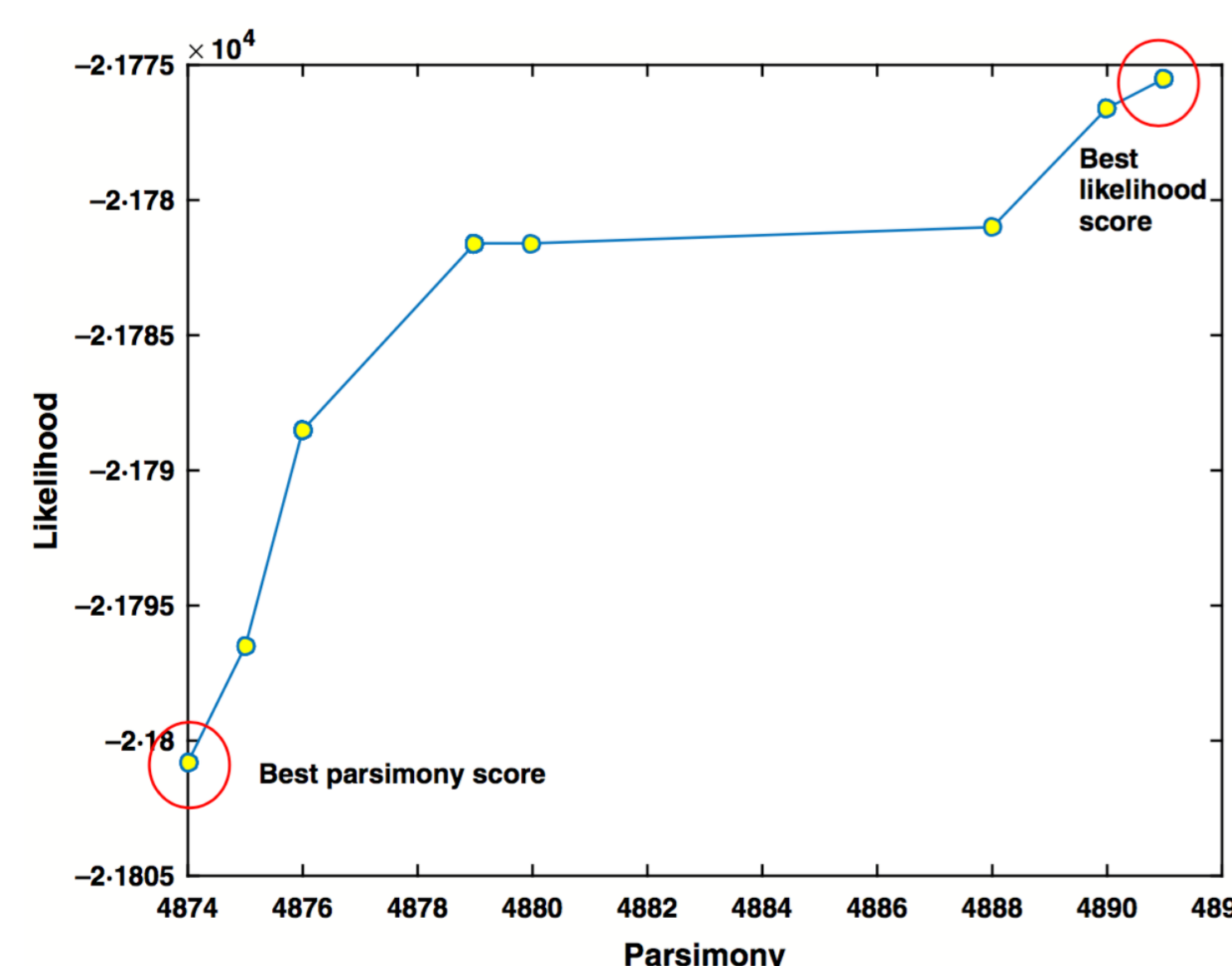

Figure 1.
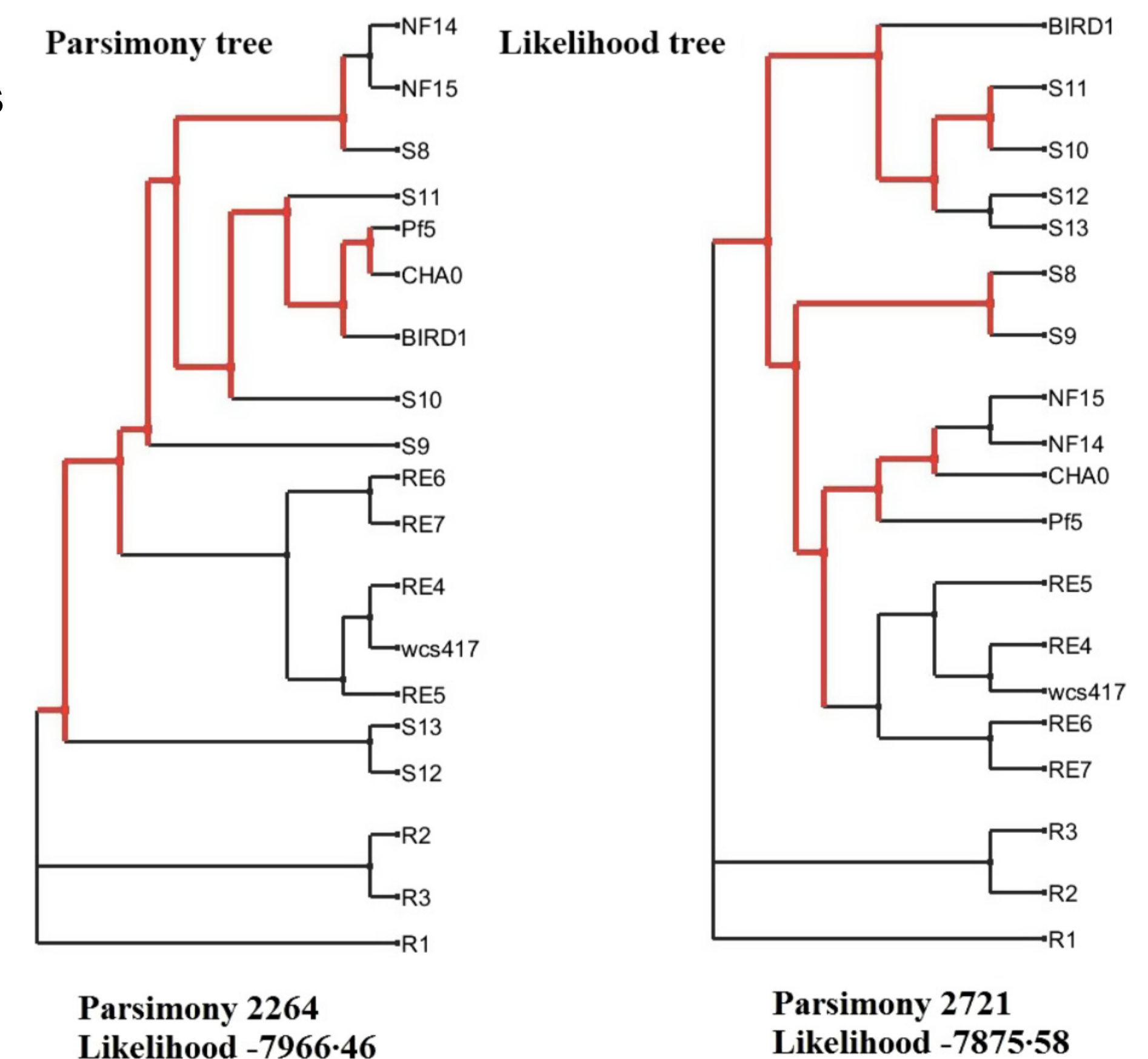Pareto Front Approximation using rbcL_55 data set.

Figure 2.
Best parsimony tree

Figure 3.
Best likelihood tree

Conclusions

Thanks to the solution proposed by this type of methodology based on evolutionary algorithms, we obtain a set of trees result. This provides great versatility compared to other approaches due to the fact that we can choose between the clustering tree which best fits our problem.

Due to the high algorithmic complexity and specificity of the problem we are trying to solve, collaboration between the fields of engineering and biology is needed.