Università
Ca'Foscari
Venezia

Master's Degree programme – Second Cycle
(*D.M. 270/2004*)
in Computer Science

Final Thesis

Ca' Foscari
Dorsoduro 3246
30123 Venezia

A new similarity measure for phylogenetic trees

Supervisor
Prof. Nicoletta Cocco

Graduand
Antonio Panzetta
834125

Academic Year
2015 / 2016

Acknowledgements

Firstly, I would like to express my sincere gratitude to my supervisor Prof. Cocco Nicoletta for the continuous support, useful comments, remarks and for her patience. Her guidance helped me in all the time of writing of this thesis. My sincere thanks also goes to my fellow classmates Gianluca, Alberto, Stefano and Enrico for the sleepless nights passed to finish homeworks together before deadlines, and for all the fun we have had in this last five years.

I am also thankful to my parents, my brothers and my sister, who have supported me throughout entire university studies, both economically and psychologically.

A heartfelt thanks goes out to all people for all the support and patience especially during my bad times.

Abstract

Comparing phylogenetic trees is a crucial task in computational biology since the various inference techniques may produce different trees for the same set of organisms. The more natural way to do this comparison is by using a distance or a similarity measure. In this thesis the most used measures in the literature are discussed, highlighting their advantages and their weaknesses. Then a new similarity measure between two phylogenetic trees for the same set of taxa is proposed. It is defined as the weighted sum of three indexes: *SSL*, that computes the ratio of taxa that have the same speciation level in both the phylogenetic trees; *NCE*, that singles out the *non-trivial common evolutionary histories* for each taxa in the two phylogenetic trees and *MCB*, that considers *non-trivial minimum common ancestors* for each taxa in the two phylogenetic trees. A parametric version of the measure is also given, that returns individual information of a specific taxon or on a set of taxa. A prototype tool for computing the similarity measure and a general discussion on the results is also given.

Contents

Chapter 1

Introduction

Comparing phylogenetic trees is a crucial task in computational biology since the various inference techniques may produce different trees for the same set of organisms. The more natural way to do this comparison is by using a distance or a similarity measure. In literature we found many similarity and distance methods that take into account different phylogenetic tree features: the speciation path and time for each taxon; kinship between a pair of taxa and kinship among a group of taxa (clusters). The Robinson-Foulds and cousin pair distance are methods that define global measure on kinship, while the MC distance and the multilevel comparison of dendrograms are methods that try to maximize the most similar bipartitions in the two trees and the speciation level with the strongest similarity wrt. relations among groups of taxa.

In this thesis we define and compute a similarity measure, called COMETH, that allows us to state the global similarity of two rooted phylogenetic trees that share the same set of taxa by considering the similarities in the evolutionary history of each single taxa x. Such measure can be also used wrt. a specific taxon to extract and display individual information.

In Chapter 2 a general overview on phylogenesis background is given. Some basic notions and definitions are introduced, which are used in the subsequent chapters.

Next, Chapter 3 describes some measures for the comparison of phylogenetic trees found in the literature, discussing their advantages and their weaknesses. The main features from the phylogenetic point of view captured by such comparison measures are highlighted.

In Chapter 4 we propose COMETH as a similarity measure between two rooted phylogenetic trees. It is based on non-trivial common evolutionary histories and non-trivial minimum common ancestors for each taxa x in the two

phylogenetic trees T_1 and T_2. A prototype tool for COMETH similarity measure has been implemented in JAVA (complete code is available in Appendix **??**) and its main components are described. A brief computational complexity analysis of the main methods is also given.

Chapter 5 reports and discusses some computational experimentations in which the COMETH similarity measure is used in contrast with other measures in the literature.

Chapter 6 contains some conclusive remarks and further future developments of the proposal.

Chapter 2

Basic notations

In this chapter we introduce some basic notions and definitions with the main aim of helping the reader to better understand the subsequent chapters. Notions and definitions are taken from [15, 18, 21, 22, 28]. It is fundamental to point out our attention on the background that we will go to analyze: **phylogeny**.

Phylogeny is the science that studies the evolution and the evolutionary relations among the organisms. Initially it was based on the comparison of physical features (*phenotypes*) and more recently on the analysis and comparison of molecular sequences, with the goal to determine the relationships between known ancestral species. The result obtained from a phylogenetic analysis is a **phylogenetic tree**, a particular structure representing relations between species (ancestral/descendant relationship). A simple general definition of phylogenetic tree is given below.

Definition 2.1 *A **phylogenetic tree** is a particular tree structure, in general a binary tree, in which we identify:*

- *nodes, called **taxonomic units** (TUs);*
- *leaves as extant species, called **operational taxonomic units** (OTUs) or taxa;*
- *internal nodes as extint species, each one represents the common **ancestor** of all the species (called in this case **descendants**) in the corresponding subtree;*
- *edges, representing **ancestor-descendant relationship**.*

Phylogenetic trees can be represented with or without root (**rooted** and **unrooted**). The former category denotes all the phylogenetic tree in which we

have a particular element, the root, that represents the farthest common ancestor of all the nodes in the evolution process. In rooted tree, the branches are time oriented, highlighting the distance in time between nodes. An unrooted phylogenetic tree, instead, describes the relations between the operational taxonomic units, without providing complete information of the evolutionary process (we are not able to distinguish if an internal node is older than another one). It shows only the topological relationship among nodes.

However, it is possible to transform a rooted phylogenetic tree in an unrooted one (**unrooting**) and viceversa (**rooting**). The process to transform an unrooted phylogenetic tree in a rooted one is called *rooting*: an element called *outgroup*, that has nothing in common or is distantly related with the analyzed species (*ingroups*), is added to the taxa and the root is placed in the edge that connects the outgroup element to the ingroups. On the other hand, the unrooting process takes out from the rooted tree the more distant ancestor. Figure A.1 shows a rooted phylogenetic tree and an unrooted phylogenetic

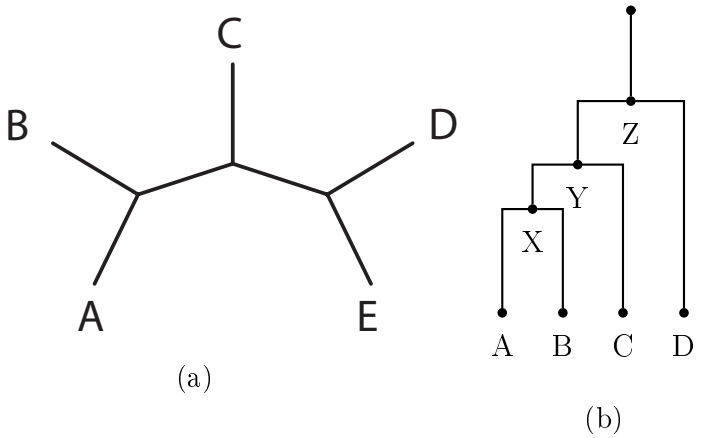


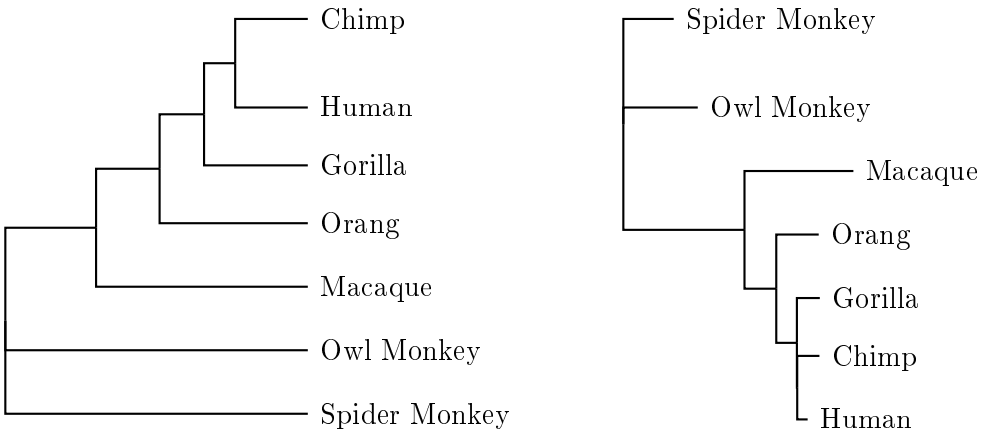
Figure 2.1: Example of phylogenetic trees: (a) unrooted phylogenetic tree (b) rooted phylogenetic tree.

tree. In the former we see that X is the common ancestor of the extant species A and B, Y is the common ancestor of A,B and C, while Z is the common ancestor of all the extant species A,B,C and D.

Generally, as described in Definition 2.1, phylogenetic trees are binary trees: each node has at least three branches, one directed to the ancestor node and the other two directed to descendants. We denote these trees as **fully resolved trees**. Neverthless, we can have a node that has more than tree branches (named **polytomy**), meaning that the tree doesn't show a fully bifurcated

topology. In that cases, the tree is considered **not resolved**. A polytomy can be always resolvable by adding more information to the taxa in the tree.

In phylogenetic trees, different weights associated to edges induce different tree classifications. A **cladogram** is a particular representation in which different branch lengths have no particular significance and do not represent time. It is used when we are interested only in inferring ancestor-descendant relationships among nodes. On the contrary, if the branch lengths is proportional to the evolutive distance among nodes, the tree is denoted as **phylogram**. In phylogram, the evolutive distance within nodes reflects their genetic divergence. An example of cladogram and phylogram is given in Figure 2.2.



(a) Cladogram (b) Phylogram

Figure 2.2: Example of cladogram (a) and phylogram (b) that describes the phylogenetic relationship among seven sequences of primates $\psi\eta$-globin. Both are not resolved trees and taken from [28].

The first goal in phylogenetic analysis is to determine the tree topology, among all possible ones, that describes the phylogenetic relationship between species that we are considering. Another important goal is to compute the branches length if we are interested in that measure. The number of all possible trees is exponentially proportional to the taxonomic units. Let us consider n as the number of leaves. The number of possible trees with or without root, respectively N_R and N_U, is defined as follows:

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \qquad N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

Note that the number of all possible unrooted trees considering n OTUs is equal to the number of all possible rooted tree computed considering $(n-1)$ OTUs.

2.1 Phylogenetic trees: building methods

Reconstructing a phylogeny starting from a set of molecular sequences is not an easy task and finding the most accurate representation of the species' relationships is not straightforward. Many methods are available in literature, even if none of them are guaranteed to reconstruct the "true" phylogenetic tree and then they are considered *estimation methods*. Based on the nature of data analyzed and the typology of the employed algorithm, methods are classified as shown in Table 2.1.

Type of algorithm	**Type of data**	
	Distances	**Nucleotide sites**
Clustering algorithms	UPGMA Neighbor joining Fitch Margoliash	
Optimization algorithms	Minimal Evolution	Maximum Parsimony Maximum Likelihood

Table 2.1: Phylogenetic trees reconstruction methods classification

In clustering algorithms, phylogenetic tree topologies are built by combining most related OTUs into clusters and computing branch lengths, while in the optimization algorithms the reconstruction is done by determining a relationship among OTUs which maximizes a certain tree feature (optimality criterion). In the following sections we will briefly introduce the most relevant phylogenetic trees reconstruction methods, trying to point out the essential features of each method.

2.1.1 Distance-based methods

Among all the methods based on clustering, **UPGMA** (Unweighted Pair Group Method with Aritmetic means) is probably the simplest method based on distances between sequences. Developed by Sokal and Michener in 1958,

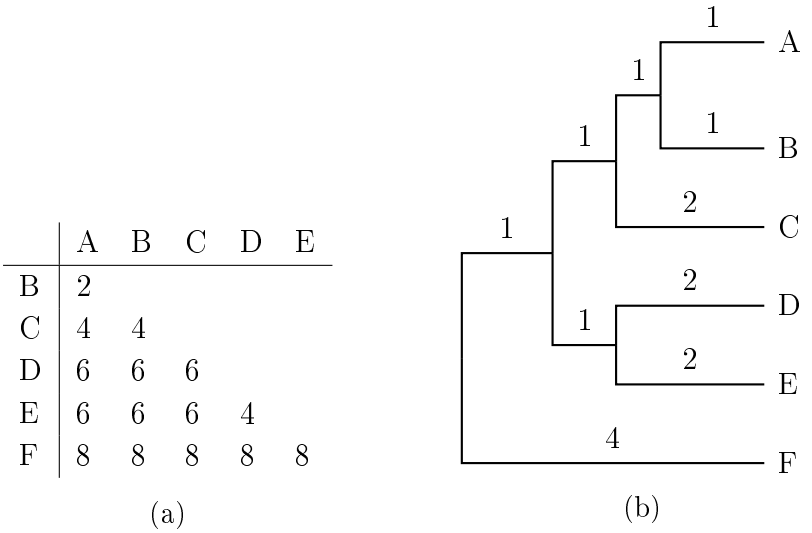


Figure 2.3: Example of phylogenetic tree reconstruction using UPGMA method. (a) denotes the initial genetic distances matrix and (b) the resulting unrooted phylogenetic tree.

this method assumes the validity of the *molecular clock*[1] that allows to evolutionary biologists to deduce how species evolve on the evolutionary timeline and to establish a date when two species diverged. It is an iterative clustering algorithm, meaning that at each step, it groups the two sequences (clusters) that are most similar to each other: in other words, based on a genetic distances matrix, the algorithm locates the pair of OTUs characterized by the minimal distance. Such pair forms a cluster that it will be treated as a single OTU in the subsequent algorithm steps. The genetic distances matrix is then recomputed, a new cluster is choosen, considering again the pair for which the similarity is the highest, and so on, until we reach the stage in which we have only two clusters that will be linked by an edge. In that stage, we define the midpoint that highlights the common ancestor of all OTUs (root). It produces an unweighted tree since all the groups are treated equally, namely UPGMA produces an *ultrametric* tree, meaning that the leaves are all equidistant from the root (branches have the same length). An example of ultrametric tree obtained through UPGMA is shown in Figure 2.3.

UPGMA is strongly related to the molecular clock hypotesis: if this is not verified, the method can not be used. When the molecular clock hypothesis, corresponding to an ultrametric tree, is not assumed, phylogenetic tree recon-

[1]The molecular clock hypothesis is a strong assumption first attributed to Zuckerkandl and Pauling in 1962 that asserts that genetic mutations, although random, occur at a relatively constant rate.

struction can be entrusted to the **Neighbor-joining (NJ)** method, developed by Saitou and Nei in 1987. It reconstructs the phylogenetic tree using minimum evolution criterion, i.e. the best estimation is defined by the tree that minimizes the lengths of all the branches in the tree. Starting from a not-fully resolved tree, NJ defines branches between closed OTUs (neighbors) and the remaining OTUs through subsequent iterative steps. The algorithm evolution is represented in Figure 2.4. The method produces an unrooted additive tree, a tree in which the branch lengths are proportional to the evolutionary change.

2.1.2 Nucleotides-based methods

The main distance-based methods' limitation is that we loose information when we synthesize phylogenetic information from a pair of aligned sequences to a distance measure. Unlike distance-based methods, in nucleotides-based methods alignment information are used as pillars of phylogenetic reconstruction. Compared to the previous described methods, these methods are slower than the distance-based ones, but they are advantageous in terms of precision and quality of the resulting tree. The most commonly used nucleotides-based methods are **Maximum Parsimony** and **Maximum Likelihood** method.

Maximum Parsimony (MP) is the most used method. It is based on the identification of the phylogenetic tree that requires the smallest number of substitutions (or evolutionary changes) to explain the observed differences within the sequences that we are considering (parsimony criterion). The most parsimonious tree is chosen among all trees which describe the phylogenetic relationships. However, searching the tree that minimizes the number of evolutionary changes is not straighforward, as well as determining the length of the tree. Indeed, the method doesn't estimate accurately the effective genetic distance, because it doesn't consider multiple or converging substitutions. Another relevant limitation of the method is that often the solution found is not unique: many trees equally parsimonious are determined, then a tree that synthesize the common features of these trees can be built (consensus technique: we'll discuss it in the following section). The resulting tree is a cladogram since we have no information on the edge lengths. An example is shown in Figure 2.5.

Another important sequence-based method is the **Maximum Likelihood method (ML)**. This is the best approach to determine the more consistent tree by considering molecular sequences. It is an optimization method w.r.t. a model evaluation and it returns a tree with edge lengths. Given a specific set of data D, given by a multiple alignement, and an hypotesis H corresponding

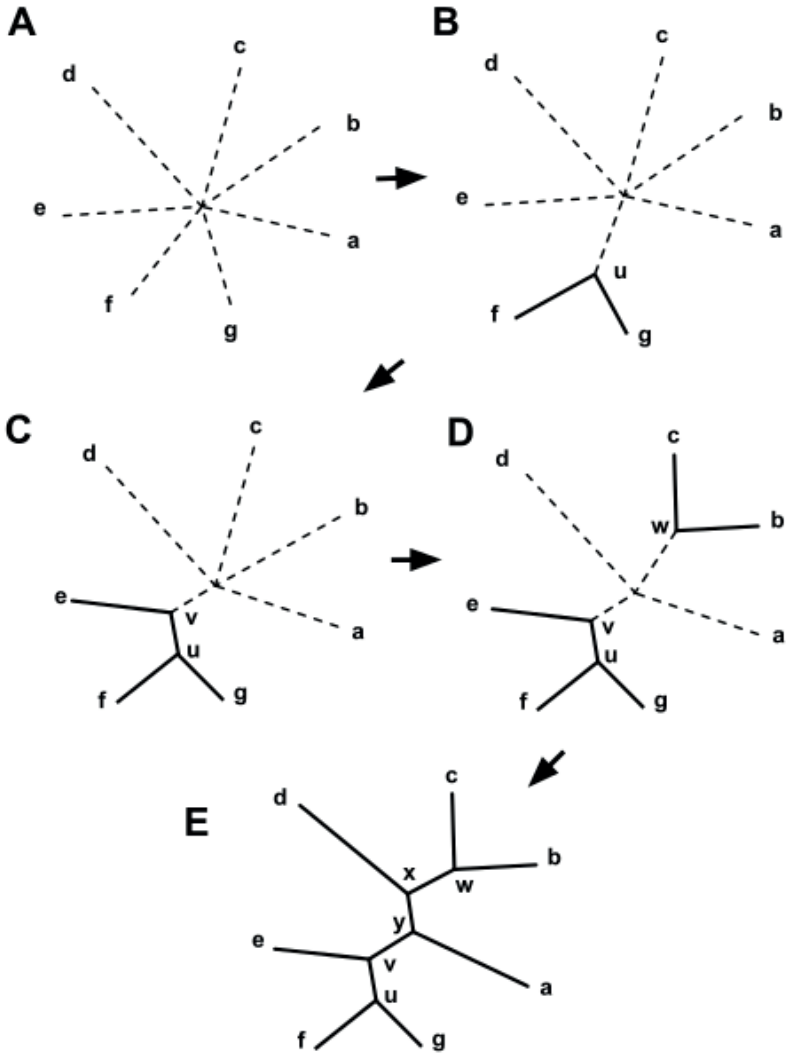


Figure 2.4: Example of NJ algorithm (A) Starting from a not-fully resolved tree (star topology) choose the closest OTUs, f and g in this case, from a distance matrix, and join them creating a new node u (B). Now, computes new distances between u and the remaining nodes and choose the lowest distance: e, shows in (C). After the (D) iteration, the algorithm ends and (E) represents the obtained fully resolved tree. Example taken from `https://en.wikipedia.org/wiki/Neighbor_joining#The_algorithm`

to a specific phylogenetic tree that depicts relations among the sequences, the probability L of observing the data, stated H, is given by $L = P(D|H)$. Between all the generated trees, the one that has the highest probability value represents the maximum likelihood estimate of the OTUs phylogeny. Despite it is considered the best method in terms of quality of tree reconstruction, the

Taxa	Sequence position (sites) and character								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G

Table 2.2: Four aligned sequences used to find the correct unrooted tree by the maximum parsimony method taken from [21].

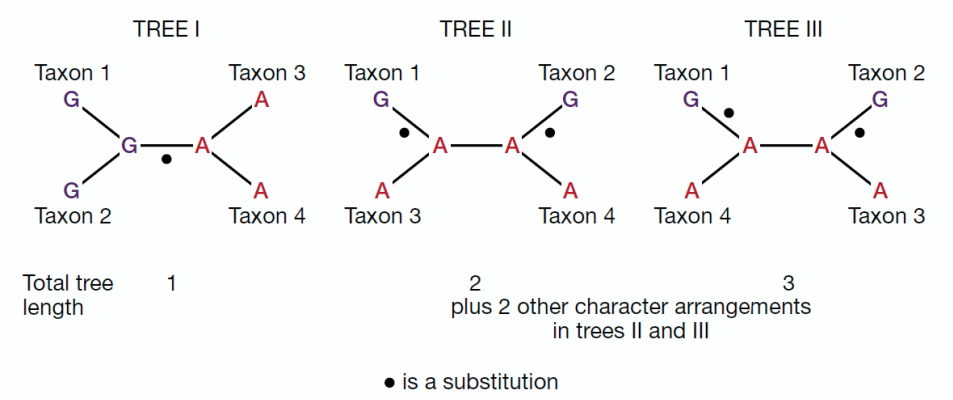


Figure 2.5: Example of phylogenetic analysis based on sequence position 5 from Table 2.2 using the maximum parsimony taken from [21] (redrawn from [19]).

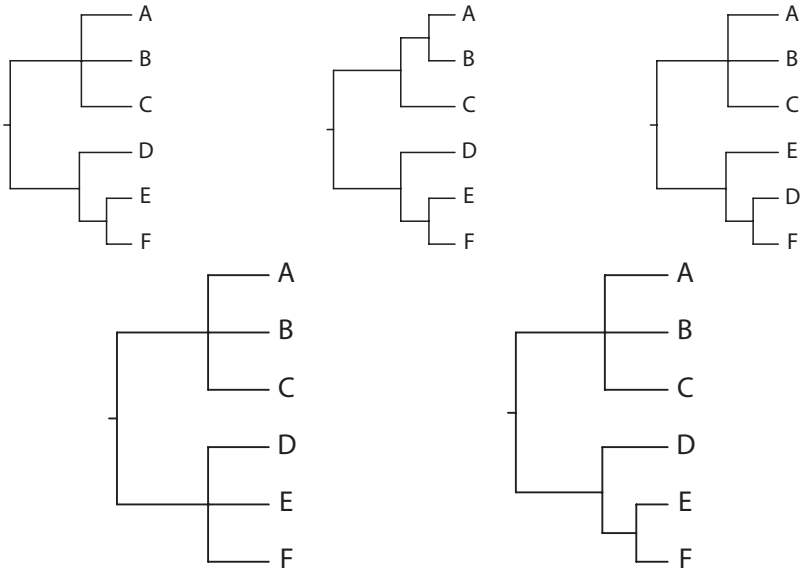
ML method is computationally very expensive, hence the use of heuristics is preferable.

2.1.3 Consensus tree

Given the same group of taxa, different reconstruction methods could produce different phylogenetic trees. Moreover, the majority of the methods that have been introduced previously generate more than one reconstructed tree and it is not easy to choose only one tree among them. Usually, an algorithm to establish a *consensus* among the reconstructed trees is used. A *consensus tree* is a tree that represents what is in common among two or more trees obtained from different phylogenetic reconstructions for the same group of taxa. There are different ways to get a consensus tree: the simplest ones are the **strict consensus** and the **majority rule consensus**.

The former method generates a consensus tree by considering only the

clades[2] (or cluster) belonging to all the trees that are considered, while the latter method produces a consensus tree by using a more relaxed criterion. In fact, the majority rule consensus considers all the groups that are present in at least the 50% of all the trees. An example of application of these methods is shown in Figure 2.6.



(a) Strict consensus (b) Majority rule consensus

Figure 2.6: Given the three rooted trees on the top, on the left side it is shown a strict consensus tree and on the right side a majority rule consensus tree.

2.2 Tools review

Some tools are available for phylogenesis reconstruction. These tools help us to display, edit and manipulate phylogenetic trees by using evolutionary reconstruction methods and analyses. Table 2.3 shows a summary of relevant packages and programs: a large list is avaible on `https://en.wikipedia.org/wiki/List_of_phylogenetics_software`.

2.3 Phylogenesis and phylogenomics

As stated in the beginning of this chapter, phylogenetics nowadays is the science that studies evolutionary relationships among species by inferring them

[2]A clade is a subset of organisms that share a common ancestor.

Name	Description
PHYLIP	(*PHY*logeny *I*nference *P*ackage) Phylogenetic tree reconstruction tool using Maximum Parsimony, Maximum Likelihood, UPGMA, Neighbor-Join and Fitch-Margoliash methods. Tool available at `evolution.genetics.washington.edu/phylip.html`
PAUP*	Originally based on the Maximum Parsimony method, the tool at present permits also phylogenetic tree reconstruction by using Maximum Likelihood and distances-based methods. Tool available at `paup.csit.fsu.edu/`
PAML	Package of programs for phylogenetic analyses of DNA or protein sequences using Maximum Likelihood. Tool available at `abacus.gene.ucl.ac.uk/software/paml.html`
Tree-Puzzle	Computer program used to construct phylogenetic trees from sequences based on Maximum Likelihood method. Tool available at `www.tree-puzzle.de/`
MEGA	*(Molecular Evolutionary Genetics Analysis)* software for phylogenetic analysis including methods like Maximum Parsimony, Maximum Likelihood and genetic distances-based methods. Tool available at `www.megasoftware.net/`
T-REX	Webserver that offers tree inference and visualization by using methods like Neighbor Joining, Maximum Parsimony and Maximum Likelihood. Tool available at `www.trex.uqam.ca/`

Table 2.3: Main packages and software for phylogenetic analyses.

from comparing their genetic sequences. Generally, a single gene is take into account for reconstruction of relationships between species. Due to the rapid progress in genome sequencing, it is now feasible to compare entire genomes instead of single genes, providing the possibility to infer not only evolutive relationships among species, but also evolutive relationships among genomic data. This intersection between genomics and phylogenetics engenders to a new discipline called **phylogenomics** (blend of the words phylogenetics and genomics indeed [25]). It is defined as the study of evolutionary relationships like phylogenesis, but it is based on comparative analysis of full genomes instead of a

single gene. The usage of whole-genomes allows us to expand considerably the number of features involved in the phylogenetic analysis.

Phylogemomics reconstruction methods are based on distances, parsimony and likelihood and they can be grouped in *sequence-based* methods and *whole-genome features-based* methods, explained in the following sections. The quality of the reconstructed phylogenetic tree derives from the quality of the features that are considered and from the accuracy of the reconstruction methods [12].

2.3.1 Sequence-based methods

Sequence-based methods in phylogenomics start generally from a multiple alignment. Two approaches can be used: **supermatrix** and **supertree** approach, both summarized in Figure 2.7.

The former approach considers a concatenation of selected genes from the whole genome into a data structure called **supermatrix**. The phylogenetic tree is built by using a tree reconstruction method (likelihood method is preferable).

The **supertree** approach, instead, analyses each gene separately. Each analysis provides a phylogenetic tree as a result. Such trees is assembled into a supertree for all the species, often built through a parsimony method.

2.3.2 Whole-genome features-based methods

Whole-genome features-based methods have recently been developed. Differently from the sequence-based methods, they are not based on a multiple-sequence alignment. These methods infer phylogenetic trees from the comparison of some genome features, such as **gene order** and **gene content**.

Gene-order methods provide a phylogenetic tree reconstruction by minimizing *breakpoints*[3] within genomes, using parsimony and distance methods. This minimization corresponds to define the number of rearrangements that are needed to transform a genome into another.

Gene-content methods may reconstruct phylogenetic trees with two different techniques. The first one uses distance-based methods on shared *orthol-*

[3]A breakpoint in two genomes signals an adjacent pair of genes which is present in one genome but not in the other.

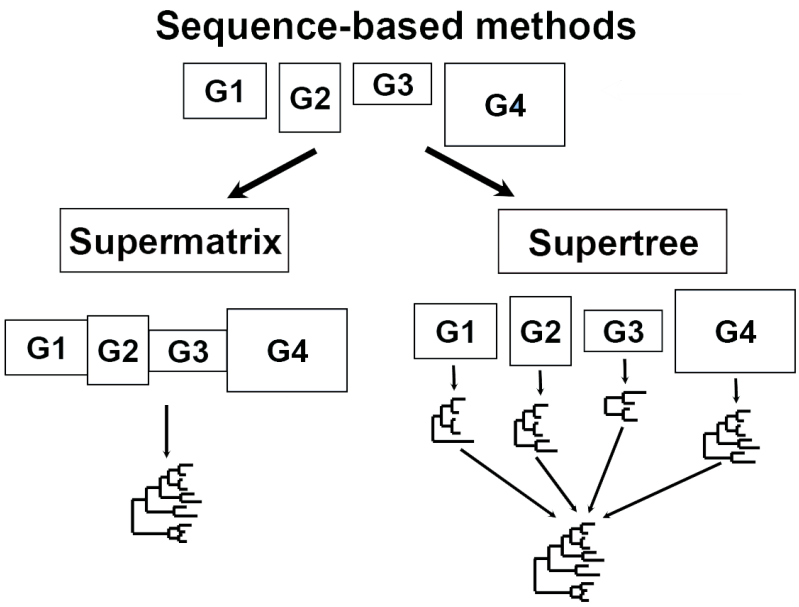


Figure 2.7: An example of (a) supermatrix tree and (b) supertree strategy taken from [12]

ogous genes[4] between genomes. Such distances are computed with one of the distance-based algorithms described in Section 2.1.1. The second technique uses matrices that represent the presence or absence of homologues/orthologues genes in a pair of genomes with a score (0 denotes the absence while 1 denotes the presence). The resulting matrices are then compared using maximum parsimony method.

[4]Two genes are orthologous when they diverge after a speciation event

Chapter 3

Comparison methods in the literature for phylogenetic trees

In this chapter we present some comparison methods found in the literature for phylogenetic trees. This is a crucial task in computational biology for different reasons. First, different methods produce different results and it is necessary some technique to understand which is the best one or to find common properties. Second, considering a reference phylogeny, it is relevant to understand how much different reconstruction methods are able to approximate it. To do these comparisons, metrics and measures are needed and in the literature we find many proposals.

The **Maximum Agreement SubTree** (MAST) is a particular technique that, given a set of trees, allows us to remove the smallest set of leaves from a tree in order to produce the same tree for each tree in the set. The issue of finding the MAST is typically NP-hard as described in [1]. Due to overcommitment to find the exact agreement among trees, many variations have been developed like **maximum information subtree (MIST)** [7] and **maximum information subtree consensus (MISC)** [24]. They work well in presence of a small number of "rogue" taxa[1], while they achieve poorly results when rogue taxa are numerous. They are used also in finding the consensus.

A measure that allows us to compare two trees is the *edit distance* on trees. Given a specific collection of well-defined transformation operations on trees, the edit distance between two trees T_1 and T_2 is defined by the minimal number of operations in the set that will transform T_1 into T_2. Many operations (well discussed in [8]) have been considered to convert a tree into

[1] A rogue taxa is a particular taxa for which the placement in the tree is variable and not clear.

another one: **NNI** (Nearest Neighbor Interchange) that allows us to exchange subtrees across a defined edge; **SPR** (Subtree Prune and Regraft) that selects and removes a subtree from the main tree and puts it elsewhere creating a new node; **TBR** (Tree Bisection and Reconnection) that detaches a subtree from the main tree and performs all possible combinations between branches of the two trees newly created. Computing the edit distance is NP-hard. In this thesis we will not consider the previous approaches.

Since we are interested in which aspects the comparison methods capture from the phylogenetic point of view, we try to list the features that a phylogenetic tree represents:

F1 speciation path and time for each taxon.

F2 kinship between a pair of taxa;

F3 kinship among a group of taxa (clusters);

We describe now four proposals of comparison methods for phylogenetic trees found in the literature, focusing on the main ideas on which the methods are based. For each proposal we will underline the main features, the tools that are available, what kind of phylogenetic trees are considered and which features are captured from the phylogenetic point of view. We will focus on methods which apply to binary phylogenetic trees, considering both rooted and unrooted ones. In some methods, clustering is examined too.

3.1 Robinson-Foulds Distance

The **Robinson-Foulds (RF)** distance [26], introduced in 1981, is widely used in biology to compute dissimilarity between phylogenetic trees. It considers unrooted binary trees, even if rooted ones can be evaluated in some generalizations of the distance, like in [4]. It is based on comparing non-trivial bipartitions. In fact, given an unrooted binary phylogenetic tree, taking out an edge produces a bipartition on the tree, creating two disjoint subsets of taxa. The tree T is uniquely represented by its bipartitions $\Gamma(T) = \{\pi_e | e \in E(T)\}$, where $E(T)$ is the set of all internal edges in T and e denotes the edge that produces the bipartition.

An internal edge always produces a non-trivial bipartition[2]. The tree T in Figure 3.1 can be described by two non-trivial bipartitions $\{AB|CDE, ABC|DE\}$

[2]Let us consider a tree T and a bipartition $\pi_e = P_1/_e P_2$ on T produced by the edge e. The bipartition π_e is non-trivial if the both cardinalities of P_1 and P_2 are greater than 1.

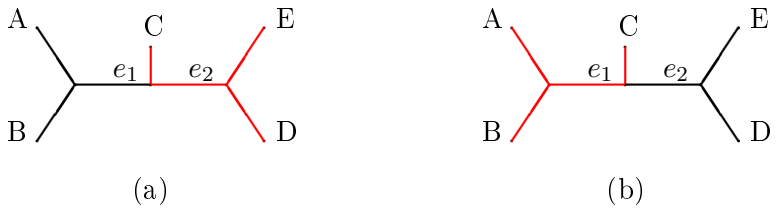


Figure 3.1: The unrooted tree is represented by its non-trivial bipartitions (a) $AB|CDE$ by considering the edge e_1, (b) $ABC|DE$ by considering the edge e_2

corresponding to the internal edge e_1 and e_2 . Let us consider two unrooted tree T_1 and T_2 that share the same set of taxa. The RF distance is defined as the number of bipartitions provided by one tree and not by the other one, as expressed by the following formula:

$$d_{RF}(T_1, T_2) = \frac{1}{2}(|\Gamma(T_1) - \Gamma(T_2)| + |\Gamma(T_2) - \Gamma(T_1)|). \tag{3.1}$$

A binary tree with n leaves, $n \geq 4$, is represented by $n-3$ non-trivial bipartitions. Hence, the maximal possible distance using the RF distance is exactly $n-3$. Let us consider the example in Figure 3.1: the unrooted labeled tree is a 5-leaves tree. The largest possible distance computable with RF is 2, that is $n - 3 = 5 - 3 = 2$.

The Robinson-Foulds distance can be computed in linear time w.r.t. number of nodes, considering the algorithm proposed by William H.E. Day in 1985 and well-described in [9], based on labeled trees. Despite the linear time computation complexity, when faced with the need to compare a large number of large trees, even a linear time becomes difficult to manage. Pattengale et al. in [23] describe a scheme based on randomized hash tables that yields an approximation of the RF distance in sublinear time [20]. The main shortcoming of this distance is that it is excessively sensitive to small changes in the tree.

The RF distance is computable in PHYLIP[3], a package of programs for inferring and analysing phylogenies. `treedist`[4] is the software in PHYLIP that computes the symmetric difference of Robinson-Foulds among two or more trees. When given as input six trees in Newick format, as shown in 3.2, `treedist` returns as output the file listed in Figure 3.2, in which it is highlighted simply a count of how many partitions there are, between two trees, that are on one tree and not on the other.

[3] `http://evolution.genetics.washington.edu/phylip.html`
[4] `http://evolution.genetics.washington.edu/phylip/doc/treedist.html`

$$
\begin{aligned}
&(A,(B,(H,(D,(J,(((G,E),(F,I)),C))))));\\
&(A,(B,(D,((J,H),(((G,E),(F,I)),C)))));\\
&(A,(B,(D,(H,(J,(((G,E),(F,I)),C))))));\\
&(A,(B,(E,(G,((F,I),((J,(H,D)),C))))));\\
&(A,(B,(E,(G,((F,I),(((J,H),D),C))))));\\
&(A,(B,(E,((F,I),(G,((J,(H,D)),C))))));
\end{aligned}
\tag{3.2}
$$

```
1
2  Tree distance program, version 3.7a
3
4  Symmetric differences between adjacent pairs of trees:
5
6  Trees 1 and 2:  4
7  Trees 3 and 4:  10
8  Trees 5 and 6:  4
```

Figure 3.2: Output file generated from the computation of Symmetric difference of Robinson-Foulds distance using trees defined in Equation 3.2

We could consider RF distance as a comparison technique which represents, with a global measure, the differences between two phylogenetic trees w.r.t. relations among group of taxa (F3).

3.2 Cousin pairs Distance

Shasha et al. in [27] propose a technique that aims to find frequent occurring patterns in a structured data set and more precisely in a tree. The cornerstone of this technique for finding patterns is the *cousin pairs* idea: a cousin pair is a pair of nodes in the tree that share either the same parent, or the same grandparent, or the same great-grandparent and so on. In phylogenetic trees, it represents the evolutionary relationship between two species that have an ancestor in common. The cousin pair distance between two nodes is related to the concept of *kinship* between nodes, i.e a cousin pair distance equal to 0 means that the two nodes are siblings, 0.5 depicts the aunt-niece relationship and 1 denotes the sharing of the same grandparent. It is applicable both to rooted and unrooted unordered labeled tree, where the order of siblings is not

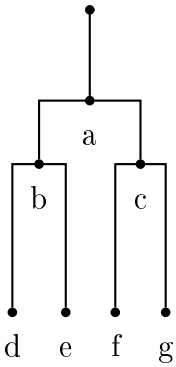


Figure 3.3: Cousin pairs distance

relevant.

Definition 3.1 *Given two labeled nodes u,v of tree T and the least common ancestor (lca) w,* ***the cousin pairs distance*** $d_{CP}(u,v)$ *is defined as follow*

$$d_{CP}(u,v) = \begin{cases} height(u,w) - 1 & \text{if } height(u,w) = height(v,w) \\ max\{height(u,w), height(v,w)\} - 1.5 & \text{if } |height(u,w) - height(v,w)| = 1 \end{cases}$$

Considering the rooted tree shown in Figure 3.3, nodes (b,c), (d,e), and (f,g) have distance 0 (siblings), (c,d), (c,e), (b,f) and (b,g) have distance 0.5 (aunt-niece), while (d,g), (d,f), (e,f) and (e,g) have distance 1 (share the same grandparent).

To find all frequent cousin pairs in a set of trees $\{T_1, \ldots, T_k\}$, it is mandatory to find first all *cousin pair items* in each tree that satisfy a specific cousin pair distance and to count the number of occurrences of a specific cousin pair in all the trees.

Definition 3.2 *Let us consider a tree* $T \in \{T_1, \ldots, T_k\}$. *A* ***cousin pair item*** *is a quadruple* $(L(u), L(v), d_{CP}(u,v), occur(u,v))$, *where u and v are cousin in the tree T,* $L(u)$ *and* $L(v)$ *denote labels for u and v respectively and* $occur(u,v) \geq 0$ *is the number of occurrences of the cousin pair* (u,v) *in T with the specified* d_{CP} *distance, defined in 3.1.*

Shasha et al. develop an algorithm[5] and a software[6] that find all cousin pairs in $O(|T|^2)$, called `Single_Tree_Mining`, which is extended in the `Multiple_Tree_Mining`

[5]Source code at `http://cs.nyu.edu/cs/faculty/shasha/papers/cousins.k`

[6]Available at `http://www.cs.nyu.edu/shasha/papers/cousins.html`

algorithm when a set of trees $\{T_1, \ldots, T_k\}$ is considered, with complexity $O(k\, m^2)$, where $m = max\{|T_1|, \ldots, |T_k|\}$ and k is the cardinality of the set of trees.

When considering the comparison between two phylogenetic trees, Sasha et al. define a measure for finding *kernel trees* from a group of phylogenies.

Definition 3.3 *Let T_1 and T_2 be two phylogenetic trees and let $cpi(T_1)$ and $cpi(T_2)$ be the sets containing all the cousin pair items of T_1 and T_2 respectively. The tree distance $t\_dist(T_1, T_2)$ is defined as follows*

$$t\_dist(T_1, T_2) = \frac{|cpi(T_1) \cap cpi(T_2)|}{|cpi(T_1) \cup cpi(T_2)|} \tag{3.3}$$

This measure can be applied in four different ways to cousin pairs items.

- $t\_dist_{null}(T_1, T_2)$ does consider neither the cousin distance nor the occurrence number in each cousin pair item;
- $t\_dist_{occ\_d_{CP}}(T_1, T_2)$ considers the cousin distance and the occurrence number in each cousin pair item;
- $t\_dist_{d_{CP}}(T_1, T_2)$ considers only the cousin distance in each cousin pair item;
- $t\_dist_{occ}(T_1, T_2)$ considers only the occurence number in each cousin pair item.

In phylogenetic trees we have no duplicated node and we are interested only in the taxa (leaves) relations, hence $t\_dist_{null}(T_1, T_2)$ and $t\_dist_{d_{CP}}(T_1, T_2)$ are the only interesting measures. Referring to Equation 3.3, we can notice that the higher the value of $|cpi(T_1) \cap cpi(T_2)|$, the higher is the similarity between the trees T_1 and T_2.

In phylogenesis, the cousin pair distance captures a global measure of the relationships between pairs of taxa (F2). It can be used to compare phylogenies or to evaluate the goodness of a consensus tree (from multiple phylogenies, it can identify how many times a specific pair of taxa occurs).

3.3 Matching Cluster distance

The Matching Cluster (MC) distance can be considered as an example of refinement of the RF distance. It is a distance between rooted trees proposed

by Damian Bogdanowicz and Krzysztof Giaro in 2013 [6] based on *clusters*. Clusters of taxa (leaves) in a rooted phylogenetic tree correspond to bipartitions in unrooted ones. The MC distance is an extension of the Matching Split (MS) distance metric presented in [5], namely the method based on *splits* for creating matching in unrooted trees, where a split represents a bipartition on a set of taxa induced by a single edge. Two such splits can be compared using a metric h, that defines the level of dissimilarity between them. This approach is then extended to whole phylogenetic trees to define a measure of distance between trees.

The MC distance is based on a *metric* [7] that represents a measure of dissimilarity between two clusters, A and B. It is defined as $h_C(A, B) = |A \oplus B| = |(A \setminus B) \cup (B \setminus A)|$, where $h_C : 2^L \times 2^L \to \mathbb{Z}_{\geq 0}, L = (A \cup B)$, it returns the number of elements that are present in one cluster but not in the other. This approach is extended to the phylogenetic trees T_1 and T_2 to define the least expensive cluster matching between all possible matchings as their distance.

The MC distance is computed by using an algorithm for finding the *minimum-weight perfect matching* in a *complete bipartite graph* [8]. A *minimum-weight perfect matching* is a perfect matching[9] in a graph where the sum of the edge's weights has the minimum value.

By $\sigma(T)$ and $\sigma^*(T)$ we define the set of all clusters and the set of *non-trivial clusters* on T respectively. Each non-trivial cluster corresponds to an internal node n (except the root) in the tree and then it corresponds to a bipartition

[7] A **metric** on a set X is a function

$$d : X \times X \to R$$

where R is the set of real numbers, and for all x, y, z in X, the following conditions are satisfied:

- $d(x, y) \geq 0$ (non-negativity, or separation axiom)
- $d(x, y) = 0$ if and only if $x = y$ (coincidence axiom)
- $d(x, y) = d(y, x)$ (symmetry)
- $d(x, z) \leq d(x, y) + d(y, z)$ (subadditivity / triangle inequality).

[8] A *complete bipartite graph* is a bipartite graph $G = (U, V, E)$ where U and V are two disjoint sets of vertices, E is the set of all the edges in G and every vertex of U is linked to each vertex of V.

[9] A perfect matching of a graph is a matching in which every vertex of the graph is incident to exactly one edge of the matching.

given from considering the leaves descending from n and all the other leaves.

Given the metric h_C, it is possible to define the distance $d_{h_C}(A, B)$ equal to the minimum-weight perfect matching in a complete bipartite graph, where $d_{h_C} : 2^D \times 2^D \to \mathbb{R}_{\geq 0}$ is a metric on a given finite set D and $A, B \in 2^D$. In phylogenetic trees, D is the set of taxon.

Let us consider two arbitrary trees T_1 and T_2. The MC distance is defined as follows

$$\begin{aligned} d_{MC}(T_1, T_2) &= d_{h_C}(\sigma(T_1), \sigma(T_2)) = \\ &= d_{h_C}(\sigma^*(T_1), \sigma^*(T_2)) \end{aligned} \tag{3.4}$$

Each non-trivial cluster identifies a vertex in the complete bipartite graph. If the bipartite graph is not complete, it is possible to add an element $O = \emptyset$ different from all the non-trivial clusters found. The weight associated to an edge between a cluster A in T_1 and a cluster B in T_2 is given by $h_C(A, B)$.

Figure 3.4 shows an example of the MC distance taken from [6]. T_1 has as non-trivial clusters $\{a, b\}$ and $\{c, d\}$, while T_2 has $\{a, b, c\}$ and O. Applying h_C to the two trees, we obtain that $h_C(\{a, b\}, \emptyset) = 2$, $h_C(\{a, b\}, \{a, b, c\}) = 1$, $h_C(\{c, d\}, \emptyset) = 2$ and $h_C(\{c, d\}, \{a, b, c\}) = 3$. The mimimum-weight perfect matching is given by $h_C(\{c, d\}, \emptyset) = 2$ and $h_C(\{a, b\}, \{a, b, c\}) = 1$, that is equal to 3.

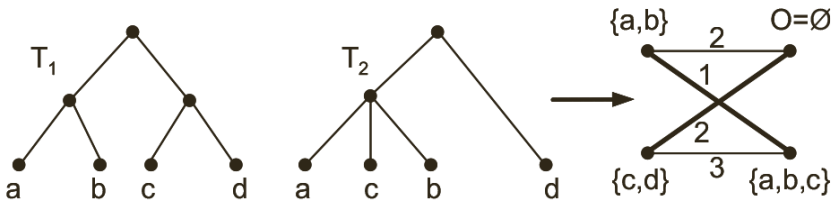


Figure 3.4: d_{MC} computed between two rooted trees T_1 (on the left side) and T_2 (in the middle). The complete bipartite graph (on the right side) highlights the perfect matching among clusters with minimum value, that is 3. [6]

The MC distance is computable in $O(|L|^{2.5} log|L|)$ [14], where L denotes the set of leaves. When compared with RF, even if MC has a worse complexity bound, it can still be used in practical applications.

From the phylogenetic point of view, the MC distance compares the bipartitions in two phylogenetic trees and returns a measure that characterizes the distance between the most similar bipartitions in the trees. We could consider

the MC distance as a comparison technique which represents the differences between two phylogenetic trees wrt. relations among groups of taxa (F3) wrt. the clades which are more similar in the two trees.

3.4 Multilevel comparison of dendrograms

Podani et al. in [17] propose an exhaustive search procedure that compares partitions between two *dendrograms*[10]. The comparison detects at which hierarchical level two dendrograms show the maximum agreement.

A dendrogram G may be viewed as a nested system of k partitions P_1, P_2, $\ldots$, P_k, $k \leq n-1$, where m denotes the number of the items in the dendrogram (the leaf of the corresponding tree). A partition in G is obtained by cutting G at a specific hierarchical level ($h_1 < h_2 < \cdots < h_k$). A partition at a certain level h_k is a union of the partitions at level h_{k-1}. Each partition P_k can be described also by an incidence matrix $\mathbf{X_k}$ in which $x_{ij} = 1$ if objects i and j belong to partition P_k, otherwise $x_{ij} = 0$.

The procedure proposed in [17] compares two dendrograms G_1 and G_2 by comparing all their non-trivial partitions in all possible ways w.r.t a distance measure. The maximum cluster agreement is reached at the minimum distance. Five different dissimilarity measures are considered for comparing two partitions $\mathcal{C}$ and $\mathcal{C}'$: Rand and Jaccard indexes [29], Euclidean distance and Sørensen index [2], and the Adjusted Rand index, that is an adjusted version of the Rand index proposed by Hubert and Arabie in [16].

Let us consider the two partitions of n items $\mathcal{C} = \{C_1, \ldots, C_k\}$ and $\mathcal{C}' = \{C'_1, \ldots, C'_l\}$ and the following sets:

- $S_{11} = \{$ pairs of items that are in the same cluster in both $\mathcal{C}$ and $\mathcal{C}'\}$
- $S_{00} = \{$ pairs of items that are in different clusters in both $\mathcal{C}$ and $\mathcal{C}'\}$
- $S_{01} = \{$ pairs of items that are in the same cluster in $\mathcal{C}$ but in different ones in $\mathcal{C}'\}$
- $S_{10} = \{$ pairs of items that are in the same cluster in $\mathcal{C}'$ but in different ones in $\mathcal{C}\}$

Let $n_{ab} = |S_{ab}|, a, b \in \{0, 1\}$, denote the respective sizes. Then

$$n_{11} + n_{00} + n_{10} + n_{01} = \binom{n}{2}$$

[10]A dendrogram is a treelike diagram depicting evolutionary changes from ancestor to descendant forms, based on shared characteristics, without considering time.

The five indexes are computed as follows.

Rand index

$$\mathcal{R}(\mathcal{C},\mathcal{C}') = \frac{2(n_{11}+n_{00})}{n(n-1)}$$

Jaccard index

$$\mathcal{J}(\mathcal{C},\mathcal{C}') = \frac{n_{11}}{n_{11}+n_{10}+n_{01}}$$

Euclidean index

$$\mathcal{E}(\mathcal{C},\mathcal{C}') = n_{01}+n_{10}$$

Sørensen index

$$\mathcal{S}(\mathcal{C},\mathcal{C}') = \frac{n_{11}}{2n_{11}+n_{10}+n_{01}}$$

Adjusted Rand index

$$\mathcal{R}_{adj}(\mathcal{C},\mathcal{C}') = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1+t_2) - t3}$$

$$\text{where} \quad t_1 = \sum_{i=1}^{k}\binom{|C_i|}{2}, t_2 = \sum_{j=1}^{l}\binom{|C'_l|}{2}, t_3 = \frac{2t_1t_2}{n(n-1)}$$

and m_{ij} is the ij-th entry of the confusion matrix $M = (m_{ij})$ of the pair $\mathcal{C},\mathcal{C}'$ that identifies the number of elements in the intersection of the clusters C_i and C'_j.

DENCOMPAR[11] is a tool written by Jànos Podani in 2009 that takes in input two dendograms in a specific format and returns a file with the five computed coefficients. Let us consider the two dendrograms G_1 and G_2 in Figure 3.5, described in `test1.dat` and `test2.dat` input files. Figure 3.6 reports a screenshot of the results: the maximum agreement is obtained at level 3 of both the dendrograms for all the five indexes considered. The output file that shows the dissimilarity values is `matrices.dat` which is shown in Table 3.1.

Since it compares the *level* of the dendrograms, the multilevel comparison applied to phylogenetic trees can represent information both on speciation time

[11] `http://ramet.elte.hu/ podani/subindex.html`

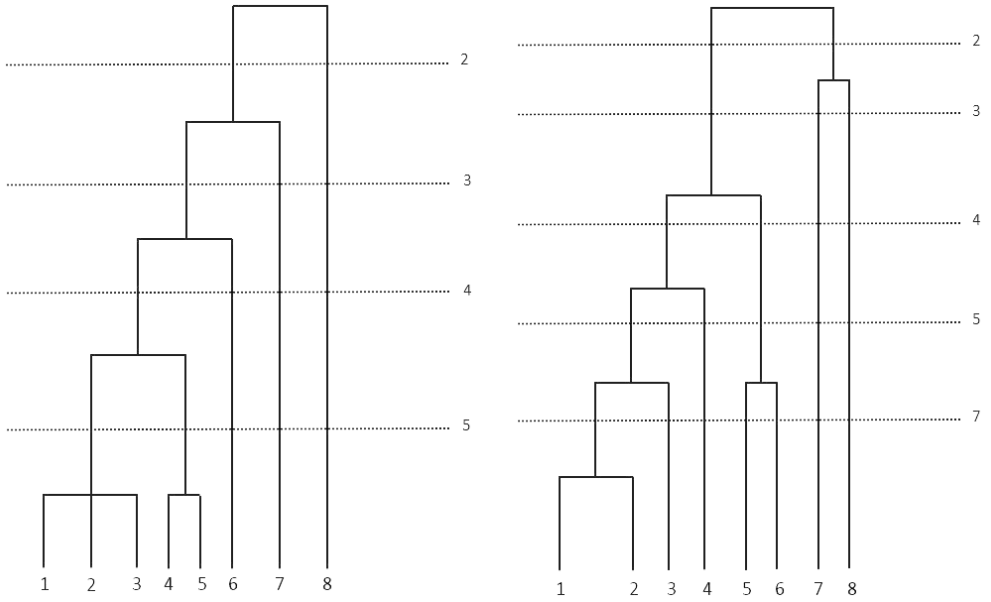


Figure 3.5: Two dendrograms (a) G_1 and (b) G_2 (taken from [17])

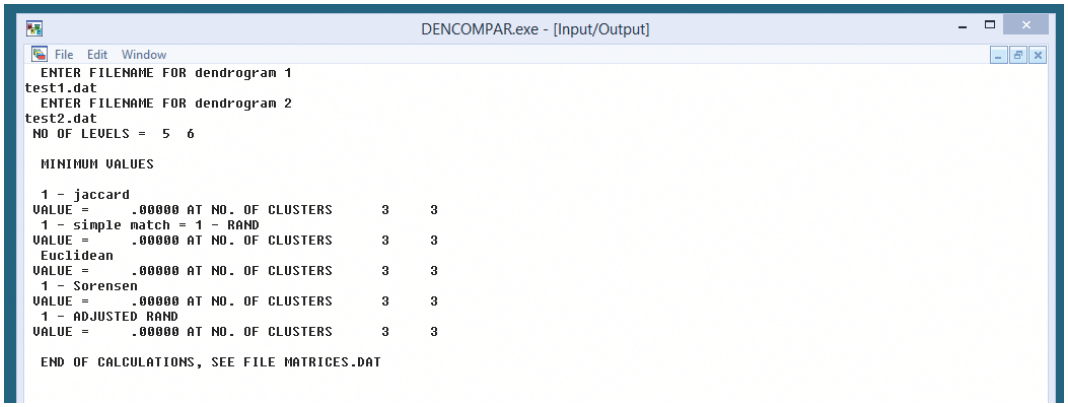


Figure 3.6: Screenshot taken from DECOMPAR tool. It shows the minimum value taken from `matrices.dat` file (reported below) and the hierarchical levels that show the maximum agreement, that is 3.

(F1) and on relations among groups of taxa (F3).
In particular it characterizes the speciation level with the strongest similarity wrt. relations among groups of taxa in the two trees.

0	7	5	4	3	2
			1 - jaccard		
5	.750000	.857143	.625000	.733333	.750000
4	.900000	.727273	.454545	.333333	.375000
3	.933333	.733333	.533333	.000000	6.25000E-02
2	.952381	.809524	.666667	.285714	.318182
			1 - simple match = 1-RAND		
5	.107143	.214286	.178571	.392857	.428571
4	.321429	.285714	.178571	.178571	.214286
3	.500000	.392857	.285714	.000000	3.571429E-02
2	.714286	.607143	.500000	.214286	.250000
			Euclidean		
5	1.73205	2.44949	2.23607	3.31662	3.46410
4	3.00000	2.82843	2.23607	2.23607	2.44949
3	3.74166	3.31662	2.82843	.000000	1.00000
2	4.47214	4.12311	3.74166	2.44949	2.64575
			1 - Sorensen		
5	.600000	.750000	.454545	.578947	.600000
4	.818182	.571429	.294118	.200000	.230769
3	.875000	.578947	.363636	.000000	3.225806E-02
2	.909091	.680000	.500000	.166667	.189189
			1 - ADJUSTED RAND		
5	.636364	.875000	.555556	.747573	.777778
4	.875000	.717949	.416667	.350000	.411765
3	.937799	.747573	.551724	.000000	7.216495E-02
2	.975610	.894737	.800000	.444444	.538462

Table 3.1: Dissimilarity values computed by DECOMPAR tool with the five different indexes. The first row identifies the number of clusters in the second dendrogram and the first column the number of clusters in the first dendrogram. For each dissimilarity measure, the minimum value (.000000) is detected when both the dendrograms have 3 clusters.

Chapter 4

A new similarity measure for phylogenetic trees: COMETH

In this chapter we propose a new similarity measure for comparing phylogenetic trees. It is based on the identification of common evolutionary histories which are computed by using a suffix function on the lists of ancestors in the two phylogenetic trees for each taxa. First, a short intuitive description of the method is given by using a small example. Then, we give the formalization of the measure, we analyse its complexity and give the corresponding pseudocode.

4.1 Phylogenetic tree computable features

Before introducing the COMETH similarity measure, let us define some features of a rooted phylogenetic tree. These features will be used in order to compute the COMETH similarity measure between two rooted phylogenetic trees that share the same set of taxa.

Let us consider a rooted phylogenetic tree T for a set of taxa $\mathcal{D}$. We can single out a structural feature: the *number of levels* of the phylogenetic tree T, k^T. It can be used to compute the depth of a taxon in a tree. In Figure 4.1 we show a simple example T_1, having the set $\mathcal{D} = (A, B, C, D, E)$ as set of taxa. In the example, $k^{T_1} = 3$. Each internal node has an associated level. Let r be the root node, then $l^T(r) = k^T$. Then, for each taxa $x \in \mathcal{D}$, we can define its *speciation level* $l^T(x)$ as the level in which the taxa diverges in T namely the level of its closest ancestor. Levels are numbered bottom-up starting from the leaves, which are at level 0, as shown in Figure 4.1. Referring to the example, the taxa A, B, C and D have speciation level equal to 1, while E has speciation level equal to 2.

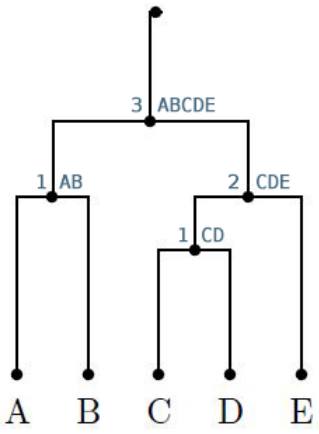


Figure 4.1: The rooted phylogenetic tree T_1 with the set of taxa $\mathcal{D} = (A, B, C, D, E)$

Each internal node y in T may be labelled by using an *unique identifier* $I(y)$. It is the concatenation in lexicographic order of each taxa belonging to the subtree rooted in y. It is unique for each internal node since the set $\mathcal{D}$ contains distinct elements. In Figure 4.1 the identifier for the internal nodes of T_1 are shown. We can determine, for each taxa x, in T the identifier of its *closest ancestor* $g^T(x)$.

Each taxa x in T can be associated to a pair $S_i{}^T(x) = (l^T(x), g^T(x))$ called its *speciation level pair*. Such pair shows, for each taxa, "when" it speciated and the corresponding taxa bipartition, that is the group of "its relatives" after speciation. Table 4.1 summarizes the features described above for T_1, where x is a taxa in $\mathcal{D}$.

x	$l^{T_1}(x)$	$g^{T_1}(x)$	$S_i{}^{T_1}(x)$
A	1	AB	(1,AB)
B	1	AB	(1,AB)
C	1	CD	(1,CD)
D	1	CD	(1,CD)
E	2	CDE	(2,CDE)

Table 4.1: Speciation level, identifier and speciation level pair for each taxa x in T_1

The evolutionary history of each taxa x in T is well-represented by the *list of its ancestors* $anc^T(x)$, which is the ordered list of pairs *(level, identifier)* for all x ancestors. For example, let us consider the taxon A. The list of its ancestors is $anc^{T_1}(A) = [(1, AB), (3, ABCDE)]$. It underlines that A

belongs to the subtrees with leaves AB, $ABCDE$ and that it diverges at level 1. Table 4.2 reports the list of the ancestors for each taxa x in T_1. Note that for each taxa x, each element in the list of its ancestors identifies a *bi-partition* at the specified associated level in the rooted phylogenetic tree T wrt. x. For instance, let us consider the pair $(2, CDE)$ for the taxon C in $anc^{T_1}(C) = [(1, CD), (2, CDE), (3, ABCDE)]$. It denotes a bipartition defined by the cluster $\{CDE\}$ and the remaining taxa $\mathcal{D} \setminus \{CDE\}$ at level 2.

Let us consider two taxa X and Y. The lists of ancestors $anc^T(X)$ and $anc^T(Y)$ permit to determine also the *kinships* between the pair of taxa (X, Y). It is sufficient to single out the leftmost common element between $anc^T(X)$ and $anc^T(Y)$. More precisely, X and Y are siblings if they share the same closest ancestor, otherwise they are cousins of some degree, determined by the level of their first common element minus 1. Clearly, taxa are all related in T and in fact they share, in the list of their ancestors, at least the root. In Table 4.2 we can see that A and B are siblings, since they share the first element $(1, AB)$, C and D are also siblings (they share $(1, CD)$), (E, C) and (E, D) are cousins of degree 1 since they share $(2, CDE)$.

$anc^{T_1}(A)$	[(1,AB),(3,ABCDE)]
$anc^{T_1}(B)$	[(1,AB),(3,ABCDE)]
$anc^{T_1}(C)$	[(1,CD),(2,CDE),(3,ABCDE)]
$anc^{T_1}(D)$	[(1,CD),(2,CDE),(3,ABCDE)]
$anc^{T_1}(E)$	[(2,CDE),(3,ABCDE)]

Table 4.2: List of the ancestors for each taxa x in T_1

Hence, for each rooted phylogenetic tree T, the following features can be singled out:

- the number of its levels k^T;
- for each taxa x in T, its speciation level $l^T(x)$;
- for each internal node y in T, its unique identifier $I(y)$;
- for each taxa x in T, the identifier of its closest ancestor $g^T(x)$ and then its speciation level pair $S_i{}^T(x) = (l^T(x), g^T(x))$;
- for each taxa x in T, the list of its ancestors $anc^T(x)$;
- for each pair of taxa (x, y) in T, the kinship relationships between (x, y).

4.2 Comparison between phylogenetic trees

If we consider now another rooted phylogenetic tree T_2 that share the same set of taxa $\mathcal{D}$ with T_1, we can compute all the features reported in the previous section also for T_2. In Figure 4.2 we show a tree T_2 and in Table 4.3 all its features are listed.

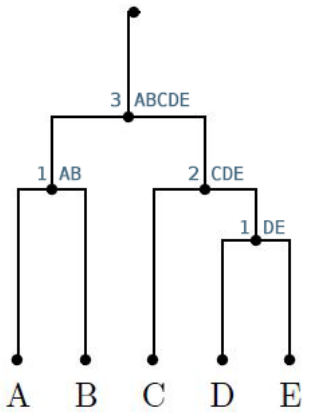


Figure 4.2: The rooted phylogenetic tree T_2 with the set of taxa $\mathcal{D} = (A, B, C, D, E)$

x	$l^{T_2}(x)$	$g^{T_2}(x)$	$S_i{}^{T_2}(x)$	$anc^{T_2}(x)$
A	1	AB	(1,AB)	[(1,AB),(3,ABCDE)]
B	1	AB	(1,AB)	[(1,AB),(3,ABCDE)]
C	2	CDE	(2,CDE)	[(2,CDE),(3,ABCDE)]
D	1	DE	(1,DE)	[(1,DE),(2,CDE),(3,ABCDE)]
E	1	DE	(1,DE)	[(1,DE),(2,CDE),(3,ABCDE)]

Table 4.3: Features computed for each taxa x in T_2

When we want to compare the two rooted phylogenetic trees, we can compare their features. We can define the set of taxa, called $Sim(T_1, T_2)$, containing all the taxa with the same *speciation level*.

In our example, we can easily compute $Sim(T_1, T_2) = (A, B, D)$. $Sim(T_1, T_2)$ can be normalised wrt. $|\mathcal{D}|$ and used as an index, that is the ratio of how many taxa maintain the same speciation level in T_1 and T_2.

We can also compare, for each taxa $x \in \mathcal{D}$, the list of its ancestors $anc^{T_1}(x)$ and $anc^{T_2}(x)$ in the two rooted phylogenetic trees T_1 and T_2. This comparison may allow us to infer something on the common evolution of x in

T_1 and T_2. In our example, let us consider first $A \in \mathcal{D}$. We know that $anc^{T_1}(A) = [(1, AB), (3, ABCDE)]$ and $anc^{T_2}(A) = [(1, AB), (3, ABCDE)]$. Since $anc^{T_1}(A) = anc^{T_2}(A)$, we can assert that the taxon A has the same evolutionary history in both the rooted phylogenetic trees. Also the taxon B has the same evolutionary history in both rooted phylogenetic trees, since $anc^{T_1}(B) = anc^{T_2}(B)$.

The property of having the same evolutionary history in two different trees for a taxa x is a very strong property. It means that the taxon x has the same ancestors at the same levels in both the rooted phylogenetic trees. Since this is extremely rare in practice, we can consider weaker properties. First of all, we can ignore the level information in the list of the ancestors, since the structure and the number of levels in the two trees, in general, can be very different. Then we define the *common evolution history* of the taxa $x \in \mathcal{D}$ in T_1 and T_2.

Definition 4.1 *Let us consider two lists L_1 and L_2. L_1 is a suffix of L_2, denoted $L_1 \sqsupset L_2$, iff $L_2 = L \cdot L_1$ for some list L.*

Definition 4.2 *The **common evolution** for a taxa x in two rooted phylogenetic trees T_1 and T_2, denoted by $CE_{T_1,T_2}(x)$, is the longest common suffix computed on the lists of its ancestors $anc^{T_1}(x)$ and $anc^{T_2}(x)$, when considering only the second components in the pairs, that are the identifiers.*

We are interested only in *non-trivial common evolutions*, called $CEvo_{T_1,T_2}$, that are suffixes of length greater than one, unless the length of the list of ancestors itself is 1 in both T_1 and T_2.

Let us consider the two rooted phylogenetic trees T_3 and T_4 shown in Figure 4.3 and the taxon D. We see that the non-trivial common evolution for the taxon D is $CEvo_{T_3,T_4}(D) = [ABCD]$, that is the root identifier. It is consider non-trivial since the list of the ancestors for D has length one in both the rooted phylogenetic trees.

Let us consider the taxon C in our running example and the list of its ancestors in T_1 and T_2:

$$anc^{T_1}(C) = [(1, CD), (2, CDE), (3, ABCDE)]$$
$$anc^{T_2}(C) = [(2, CDE), (3, ABCDE)]$$

Starting from the second ancestor, C maintains the same evolution in both the rooted phylogenetic trees. Also D and E have the same evolution starting from the second ancestor, since

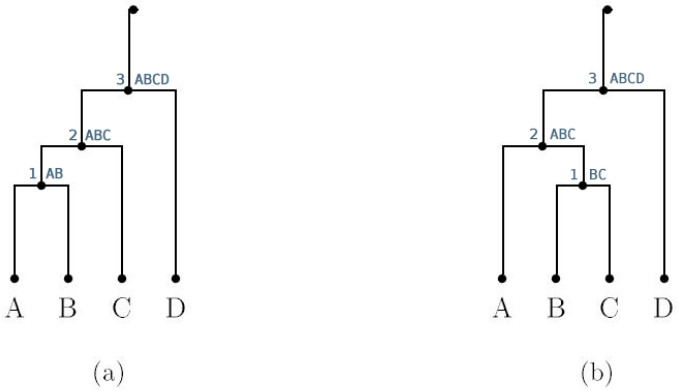


Figure 4.3: Two rooted phylogenetic trees (a) T_3 and (b) T_4 with the same set of taxa: A, B, C, D.

$$anc^{T_1}(D) = [(1, CD), (2, CDE), (3, ABCDE)]$$
$$anc^{T_2}(D) = [(1, DE), (2, CDE), (3, ABCDE)]$$

$$anc^{T_1}(E) = [(2, CDE), (3, ABCDE)]$$
$$anc^{T_2}(E) = [(1, DE), (2, CDE), (3, ABCDE)]$$

In this example the levels are also the same, but this is not generally the case, as shown in the example of Figure 4.4. We report in Table 4.4 the non-trivial common evolution for each taxa $x \in \mathcal{D}$.

$CEvo_{T_1,T_2}(A) = [AB, ABCDE]$
$CEvo_{T_1,T_2}(B) = [AB, ABCDE]$
$CEvo_{T_1,T_2}(C) = [CDE, ABCDE]$
$CEvo_{T_1,T_2}(D) = [CDE, ABCDE]$
$CEvo_{T_1,T_2}(E) = [CDE, ABCDE]$

Table 4.4: Non-trivial common evolutions for each taxa $x \in \mathcal{D}$.

A non-trivial common evolution is rather rare in practice. Hence, when considering the lists of the ancestors of a specific taxon $x \in \mathcal{D}$, we can identify, if it exists, the *non-trivial minimum common ancestor* in both the rooted phylogenetic trees T_1 and T_2: $MCA_{T_1,T_2}(x)$. It is the minimum common ancestor of the taxon x in T_1 and T_2, different from the root, unless the root is the only ancestor of x in T_1 and T_2. The non-trivial minimum common ancestor of a taxa x corresponds to the non-trivial minimum common bipartition wrt. x in

the two rooted phylogenetic trees, when it exists.

Let us consider our example, where the taxon D has the following lists of ancestors in T_1 and T_2:

$$anc^{T_1}(D) = [(1, CD), (2, CDE), (3, ABCDE)]$$
$$anc^{T_2}(D) = [(1, DE), (2, CDE), (3, ABCDE)]$$

The non-trivial minimum common ancestor is $MCA_{T_1,T_2}(D) = [CDE]$ since it is the first common ancestor of x in both the rooted phylogenetic trees. The levels associated to the first common ancestor are available, but they are not considered since, in general, they can be different.

Let us consider a different example, given by the two rooted phylogenetic tree T_5 and T_6 sharing the same set of taxa, shown in Figure 4.4, and the taxon F that has the following lists of ancestors:

$$anc^{T_5}(F) = [(1, FG), (2, DEFG), (3, ABCDEFG)]$$
$$anc^{T_6}(F) = [(2, DEF), (3, DEFG), (4, ABCDEFG)]$$

The non-trivial minimum common ancestor is $MCA_{T_5,T_6}(F) = [DEFG]$, with different levels in the two rooted phylogenetic trees.

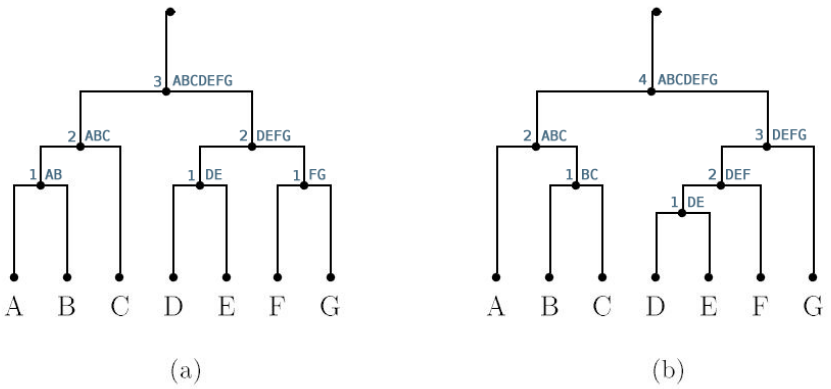


Figure 4.4: Two rooted phylogenetic trees (a) T_5 and (b) T_6 with the same set of taxa: A, B, C, D, E, F, G.

Table 4.5 reports the non-trivial minimum common ancestors for each taxa x in T_1 and T_2, and Table 4.6 for each taxa in T_5 and T_6.

Let us summarize the main features we consider in comparing two rooted phylogenetic trees T_1 and T_2 with the same set of taxa $\mathcal{D}$:

$MCA_{T_1,T_2}(A) = [AB]$
$MCA_{T_1,T_2}(B) = [AB]$
$MCA_{T_1,T_2}(C) = [CDE]$
$MCA_{T_1,T_2}(D) = [CDE]$
$MCA_{T_1,T_2}(E) = [CDE]$

Table 4.5: Non-trivial minimum common ancestors for each taxa $x \in \mathcal{D}$ in T_1 and T_2.

$MCA_{T_5,T_6}(A) = [ABC]$
$MCA_{T_5,T_6}(B) = [ABC]$
$MCA_{T_5,T_6}(C) = [ABC]$
$MCA_{T_5,T_6}(D) = [DE]$
$MCA_{T_5,T_6}(E) = [DE]$
$MCA_{T_5,T_6}(F) = [DEFG]$
$MCA_{T_5,T_6}(G) = [DEFG]$

Table 4.6: Non-trivial minimum common ancestors for each taxa $x \in \mathcal{D}$ in T_5 and T_6.

- the set of taxa with the same speciation level $Sim(T_1, T_2)$;
- for each taxa $x \in \mathcal{D}$, the non-trivial common evolution $CEvo_{T_1,T_2}(x)$;
- for each taxa $x \in \mathcal{D}$, the non-trivial minimum common ancestor $MCA_{T_1,T_2}(x)$.

4.3 Similarity indexes for comparison

In order to do the comparison between two rooted phylogenetic trees T_1 and T_2, we define some similarity indexes based on the previous features. We define three indexes which may be weighted and combined to define a similarity measure.

The SSL index computes the percentage of *taxa that have the same speciation level* in both the rooted phylogenetic trees T_1 and T_2. It is defined as

$$\mathrm{SSL}_{\mathrm{T_1,T_2}} = \frac{|Sim(T_1, T_2)|}{n} \tag{4.1}$$

where $n = |\mathcal{D}|$. The SSL index can assume values in the interval $[0, 1]$.

$\mathrm{SSL}_{\mathrm{T}_1,\mathrm{T}_2} = 0$ when $Sim(T_1, T_2) = \emptyset$. $\mathrm{SSL}_{\mathrm{T}_1,\mathrm{T}_2} = 1$ when all the taxa in $\mathcal{D}$ have the same speciation level in both the rooted phylogenetic trees.

In our example $Sim(T_1, T_2) = (A, B, D)$ and $n = 5$. Hence

$$\mathrm{SSL}_{\mathrm{T}_1,\mathrm{T}_2} = \frac{|Sim(T_1, T_2)|}{n} = \frac{3}{5} = 0,6$$

This means that 60% of the taxa in $\mathcal{D}$ maintains the same speciation level in both the rooted phylogenetic trees T_1 and T_2.

More interesting are the indexes that are based on the non-trivial common evolutions and on the non-trivial minimum common ancestor, called respectively NCE and MCB indexes.

The NCE index computes the percentage of *non-trivial common evolutions* for each taxa $x \in \mathcal{D}$ in both the rooted phylogenetic trees. It is defined as

$$\mathrm{NCE}_{\mathrm{T}_1,\mathrm{T}_2} = \frac{\sum_{x \in \mathcal{D}} min\left(\frac{|CEvo_{T_1,T_2}(x)|}{|anc^{T_1}(x)|}, \frac{|CEvo_{T_1,T_2}(x)|}{|anc^{T_2}(x)|}\right)}{n} \tag{4.2}$$

We recall that non-trivial means that for each taxa x $|CEvo_{T_1,T_2}(x)| > 1$ or ($|anc^{T_1}(x)| = 1$ and $|anc^{T_2}(x)| = 1$). NCE ranges in the interval $[0, 1]$. The index is equal to 0 when, for any taxa x in the two rooted phylogenetic trees, there is no non-trivial common evolution. On the other hand, the index is equal to 1 when all the taxa x have the same non-trivial evolution in both the rooted phylogenetic trees.

We compute the index in our running example. Let us start from the first taxon A. We know that

$$anc^{T_1}(A) = [(1, AB), (3, ABCDE)]$$
$$anc^{T_2}(A) = [(1, AB), (3, ABCDE)]$$

$$CEvo_{T_1,T_2}(A) = [AB, ABCDE]$$

In this case $|anc^{T_1}(A)| = 2$, $|anc^{T_2}(A)| = 2$ and $|CEvo_{T_1,T_2}(A)| = 2$. Then,

$$min\left(\frac{|CEvo_{T_1,T_2}(A)|}{|anc^{T_1}(A)|}, \frac{|CEvo_{T_1,T_2}(A)|}{|anc^{T_2}(A)|}\right) = min\left(\frac{2}{2}, \frac{2}{2}\right) = 1$$

For taxon B, we have the same result of A that is 1, since they are siblings. Let us consider now the taxon C that have the lists of ancestors and the common evolution as follows:

$$anc^{T_1}(C) = [(1, CD), (2, CDE), (3, ABCDE)]$$
$$anc^{T_2}(C) = [(2, CDE), (3, ABCDE)]$$

$$CEvo_{T_1,T_2}(C) = [CDE, ABCDE]$$

Since $|anc^{T_1}(C)| = 3$, $|anc^{T_2}(C)| = 2$ and $|CEvo_{T_1,T_2}(C)| = 2$, we have that

$$min\left(\frac{|CEvo_{T_1,T_2}(C)|}{|anc^{T_1}(C)|}, \frac{|CEvo_{T_1,T_2}(C)|}{|anc^{T_2}(C)|}\right) = min\left(\frac{2}{3}, \frac{2}{2}\right) = min(0.67, 1) = 0.67$$

Taxa D have the same result of C that is $0,67$ since they are siblings. For the taxon E, we have

$$anc^{T_1}(E) = [(2, CDE), (3, ABCDE)]$$
$$anc^{T_2}(E) = [(1, DE), (2, CDE), (3, ABCDE)]$$

$$CEvo_{T_1,T_2}(E) = [CDE, ABCDE]$$

Since $|anc^{T_1}(E)| = 2$, $|anc^{T_2}(E)| = 3$ and $|CEvo_{T_1,T_2}(E)| = 2$, we have that

$$min\left(\frac{|CEvo_{T_1,T_2}(E)|}{|anc^{T_1}(E)|}, \frac{|CEvo_{T_1,T_2}(E)|}{|anc^{T_2}(E)|}\right) = min\left(\frac{2}{2}, \frac{2}{3}\right) = min(1, 0.67) = 0.67$$

So,

$$\begin{aligned}\mathrm{NCE}_{\mathrm{T_1,T_2}} &= \frac{\sum_{x\in\mathcal{D}} min\left(\frac{|CEvo_{T_1,T_2}(x)|}{|anc^{T_1}(x)|}, \frac{|CEvo_{T_1,T_2}(x)|}{|anc^{T_2}(x)|}\right)}{n} \\ &= \frac{(1 + 1 + 0,67 + 0,67 + 0,67)}{5} \\ &= \frac{4,01}{5} = 0,802\end{aligned}$$

In our example, the NCE index returns a value equal to 0,802, that denotes that the 80% of the evolutions are in common between the two rooted

phylogenetic trees.

The MCB index defines the percentage of *the non-trivial minimum common bipartitions* for each taxa x in both the rooted phylogenetic trees. It is defined as

$$\mathrm{MCB}_{\mathrm{T_1,T_2}} = \frac{\left|\{MCA_{T_1,T_2}(x)\}_{x\in\mathcal{D}}\right|}{n} \tag{4.3}$$

We recall that non-trivial means that $MCA_{T_1,T_2}(x)$ doesn't exist when the minimal common ancestor of x in T_1, T_2 is $I(r)$, where r is the root, unless ($|anc^{T_1}(x)| = 1$ and $|anc^{T_2}(x)| = 1$). MCB can assume values in the interval $[0, 1]$. If the index is equal to 0, it means that each taxa x has no non-trivial minimum common ancestor in the rooted phylogenetic trees T_1 and T_2, while it is equal to 1 when all the taxa x have a non-trivial minimum common ancestor in T_1 and T_2.
In our example, we have that $\left|\{MCA_{T_1,T_2}(x)\}_{x\in\mathcal{D}}\right| = 5$ and $n = 5$. So, $\mathrm{MCB}_{\mathrm{T_1,T_2}} = \frac{5}{5} = 1$

4.3.1 COMETH similarity measure

We can define a global similarity measure for comparing two rooted phylogenetic trees T_1 and T_2, called COMETH, as the weighted sum of the previous indexes. It is defined as

$$\mathrm{COMETH}_{T_1,T_2} = w_1\mathrm{SSL}_{T_1,T_2} + w_2\mathrm{NCE}_{T_1,T_2} + w_3\mathrm{MCB}_{T_1,T_2} \tag{4.4}$$

where w_1,w_2 and w_3 are weights that we associate to each index and $w_1 + w_2 + w_3 = 1$. The value of each weight w_i is determined by the importance for the comparison of the information captured by the index. In fact, the weight w_1 associated to SSL should be the smallest since it represents only how many taxa in the two phylogenetic trees maintain the same speciation level. By default, we set $w_1 = 0,2$. The weight w_2 associated to NCE should be the highest since it singles out how many taxa in the two phylogenetic trees have the same non-trivial common evolution, rarely satisfied as the number of taxa increases. By default, we set $w_2 = 0,5$. Since $w_1 + w_2 + w_3 = 1$, we set $w_3 = 1 - (w_1 + w_2) = 0,3$ for the last index MCB, that highlights how many taxa have the same non-trivial minimum common bipartition in the two phylogenetic trees. The weights associated to the indexes should be modifiable, depending on what index we want to highlight for our purposes.

COMETH ranges in the interval $[0, 1]$. When $T_1 = T_2$, it assumes value equal to 1.

In COMETH, we can assert these properties:

1. $\text{NCE} = 1 \implies T_1 = T_2$;
2. $\text{NCE} = 1 \implies \text{SSL} = 1$;
3. $\text{NCE} = 1 \implies \text{MCB} = 1$;

To demonstrate property (1), we have, for each internal node z_1 in T_1, find an internal node z_2 in T_2 such that $anc^{T_1}(z_1) = anc^{T_2}(z_2)$ and they have the same structure. The proof can be given by induction on the depth of the tree rooted in z_1.

Since the COMETH similarity measure depends on the taxa in $\mathcal{D}$, it can be also parametrically computed wrt. the set of taxa. It means that the COMETH similarity measure can be computed for any subset of $\mathcal{D}$ or even for each single taxa x and, in this case, it may give us information in the comparison for one taxon at time. This can be used to individually display the non-trivial common evolutions and the non-trivial minimum common bipartition for a specific taxon. We define COMETH similarity measure, wrt. a simple taxa x, as

$$\text{COMETH}_{T_1,T_2}(x) = w_1\text{SSL}_{T_1,T_2}(x) + w_2\text{NCE}_{T_1,T_2}(x) + w_3\text{MCB}_{T_1,T_2}(x) \quad (4.5)$$

Also this measure assumes values in the interval $[0, 1]$.

The COMETH similarity measure considers all the features listed in Chapter 3 and denoted as F1, F2 and F3. More precisely, COMETH consider the speciation time feature through the speciation level $l^T(x)$, while the kinship relations are highlighted in the list of the ancestors $anc^T(x)$. COMETH similarity measure is easy to compute since it is the weighted sum of three simple indexes. The weights associated to indexes give to the user the possibility to determine which index has more importance on the global similarity measure.

COMETH similarity measure considers features that have not been taken into account from the measures described in Chapter 3, like the non-trivial common evolution, the non-trivial minimum common ancestor and the number of taxa that have the same speciation level in the two rooted phylogenetic trees. The latter feature *SSL* allows us to give a general structural information about the taxa in the two rooted phylogenetic trees, while the non-trivial common evolutionary history and the non-trivial minimum common ancestors give us more information on the evolutions of the two rooted phylogenetic trees. For

this reason, the three indexes are weighted. These can be used to graphically display the common evolutionary history and the minimum common ancestor both for the whole rooted phylogenetic trees and for a taxon suitably specified. The graphical visualization has not been implemented in this thesis, but it is an interesting further development.

COMETH similarity measure is defined for rooted phylogenetic trees only. For the unrooted ones, it is not possible to specify the evolutionary history for each taxa in the two trees, that is at the base of the COMETH measure.

Although we could compute information on kinships among taxa, this feature is not considered in the COMETH measure, since starting from the list of ancestors of each taxa it is feasible but costly to infer kinship relations and common substructures in the two rooted phylogenetic trees.

4.4 COMETH tool

A prototype tool for comparing two rooted phylogenetic trees by using the COMETH similarity measure has been implemented in JAVA. It is defined in the `cometh` package composed by different functions that compute the features and the indexes defined in the previous sections.

The `main` function accepts as input two rooted phylogenetic trees T_1 and T_2 in Newick format and allows the user to perform either the global similarity measure or its parametric version wrt. a group of taxa or single taxon. The tool alerts if the input trees are not in the Newick format by displaying an error. When we choose to perform the parametric measure wrt. single taxa, it is mandatory to select one taxon from the list of available ones. In the case in which the user selects a taxon that is not in the list of the available taxa, the tool displays an error. It is possible to define the weights associated to the SSL, NCE and MCB indexes. The sum of the weights must be equal to 1, otherwise the tool signals an error through a message. Default values are also available.

To compute the COMETH similarity measure, it is necessary to:

- Select the typology of the measure to be computed (global, single taxon, group of taxa);
- Supply the two rooted phylogenetic trees in the Newick format;
- Specify new weights for the indexes or proceed with the default ones;

- Specify the single taxon or a group of taxa when the parametric measure is performed.

Let us consider as input to the tool the Newick format of the two rooted phylogenetic trees T_1 and T_2 shown in Figure 4.1 and Figure 4.2 and reported below

$$T_1 = \texttt{((A,B),((C,D),E));}$$
$$T_2 = \texttt{((A,B),(C,(D,E)));}$$

Figure 4.5 shows the results of the `cometh` tool performing the global measure for the two rooted phylogenetic trees T_1 and T_2 by considering the default weights.

Figure 4.6 shows the results of the `cometh` tool performing the parametric similarity measure for the specific taxon C in the two rooted phylogenetic trees T_1 and T_2 by considering the default weights.

4.4.1 Main functions in the COMETH tool

Let us briefly discuss the main methods in the `cometh` package, specifing the functions that they perform and highlighting their complexity.

```
1 private static List<String> newick(String str,List<String> lst_taxa)
```

This `newick` method allows the user to detect first if the string `str` is in the Newick format. If the tree doesn't respect the Newick format, an error is generated and a message is displayed. Then, it computes the list of ancestors of each taxa x, extracted from the `lst_taxa` list, for the current tree `str`. It is the first called function since all the features and indexes to be computed derive from information that are available in the list of the ancestors of each taxa x. In this prototype, extracting the information is computationally expensive (in the worst case $O(n^2)$, where $n = |\mathcal{D}|$). In fact for each taxon x the unique identifier $I(y)$ is computed by the function `get_unique_identifier()`, which traverses the subtree with root x with linear worst case complexity $O(n)$. Then, the list of the ancestors, computed by the function `update_taxa_features()`, that has $O(k^T)$ as complexity, is updated. By using different data structures to store and update the list of the ancestors for each taxa might reduce the complexity. On the other hand, in general applications, the number of taxa n is not so large, it could be tens of hundreds of taxa.

```
1 private static List<String> find_Sim(List<String> T1,List<String> T2)
```

```
-------- COMETH tool --------
Panzetta Antonio  834125
Version 1.0
-----------------------------

1) COMETH global measure
2) COMETH parametric measure (single taxa)
3) COMETH parametric measure (subset of taxa)

0) Exit

Select an option: 1

Insert T1 in Newick format: ((A,B),((C,D),E));
Insert T2 in Newick format: ((A,B),(C,(D,E)));
Do you want to specify weight for COMETH? (y,n) n
Default weights selected


-------- Computing COMETH global measure --------
Weights:
        w1 = 0.2
        w2 = 0.5
        w3 = 0.3

T1: ((A,B),((C,D),E));
        A:(1,AB)(3,ABCDE)
        B:(1,AB)(3,ABCDE)
        C:(1,CD)(2,CDE)(3,ABCDE)
        D:(1,CD)(2,CDE)(3,ABCDE)
        E:(2,CDE)(3,ABCDE)

T2: ((A,B),(C,(D,E)));
        A:(1,AB)(3,ABCDE)
        B:(1,AB)(3,ABCDE)
        C:(2,CDE)(3,ABCDE)
        D:(1,DE)(2,CDE)(3,ABCDE)
        E:(1,DE)(2,CDE)(3,ABCDE)

Sim(T1,T2) set: [A, B, D]

Common evolutionary histories:
        A:[(AB),(ABCDE)]
        B:[(AB),(ABCDE)]
        C:[(CDE),(ABCDE)]
        D:[(CDE),(ABCDE)]
        E:[(CDE),(ABCDE)]

Minimum common non-trivial bipartition:
        A: [AB]
        B: [AB]
        C: [CDE]
        D: [CDE]
        E: [CDE]

Indexes:
        SSL index: 60,00%
        NCE index: 80,00%
        MCB index: 100,00%

        COMETH global measure: 0.2*0.6 + 0.5*0,80 + 0.3*1.0 = 0,82 = 82,00%

##### END COMPUTATION ###
```

Figure 4.5: Output of the global similarity measure computed by COMETH tool with $T_1 =$`((A,B),((C,D),E));` and $T_2 =$`((A,B),(C,(D,E)));` as input.

```
-------- COMETH tool --------
Panzetta Antonio  834125
Version 1.0
-----------------------------

1) COMETH global measure
2) COMETH parametric measure (single taxa)
3) COMETH parametric measure (subset of taxa)

0) Exit

Select an option: 2

Insert T1 in Newick format: ((A,B),((C,D),E));
Insert T2 in Newick format: ((A,B),(C,(D,E)));
Do you want to specify weight for COMETH? (y,n) n
Default weights selected

Choose a taxon in ("A" "B" "C" "D" "E" ): C

Weights:
        w1 = 0.2
        w2 = 0.5
        w3 = 0.3

T1: ((A,B),((C,D),E));
        A:(1,AB)(3,ABCDE)
        B:(1,AB)(3,ABCDE)
        C:(1,CD)(2,CDE)(3,ABCDE)
        D:(1,CD)(2,CDE)(3,ABCDE)
        E:(2,CDE)(3,ABCDE)

T2: ((A,B),(C,(D,E)));
        A:(1,AB)(3,ABCDE)
        B:(1,AB)(3,ABCDE)
        C:(2,CDE)(3,ABCDE)
        D:(1,DE)(2,CDE)(3,ABCDE)
        E:(1,DE)(2,CDE)(3,ABCDE)

- Taxon "C" is NOT present in Sim(T1,T2)

- Non-trivial common evolutionary history for "C"
        C:[(CDE),(ABCDE)]

- Non-trivial minimum common bipartition for "C"
        C: [CDE]

Indexes:

        COMETH global measure: 0.2*0.0 + 0.5*0,67 + 0.3*1.0 = 0,63 = 63,33%

##### END COMPUTATION ###
```

Figure 4.6: Output of the similarity measure computed by COMETH tool for the taxon C with T_1 =`((A,B),((C,D),E));` and T_2 =`((A,B),(C,(D,E)));` as input.

This function computes the set of all the taxa that maintain the same speciation level in the two phylogenetic trees. It is called by the tool without the interaction of the user. It accepts two lists, in which each element is defined as `taxa:<ancestors>`, where each element in `<ancestors>` is a pair *(level, identifier)*. This data structure will be called *list_taxa_ancestors* in all the functions where it is used. The `find_Sim` function complexity is linear in the number of taxa ($O(n)$).

```
private static List<String> find_CEvo(List<String> T1,List<String> T2)
```

It computes is the non-trivial common evolution $CEvo_{T_1,T_2}$. It is an internal function that accepts as parameters two *list_ taxa_ ancestors* lists. It computes the longest common suffix between the lists of the ancestors for any taxa x in the two trees and returns the list of all the non-trivial common evolutions. Regarding complexity, the longest common suffix between the lists of the ancestors can be, in the worst case, the whole list (at most its length is equal to k^T). Since it is computed for each taxa x, the complexity is $O(nk^T)$.

```
private static List<String> find_mca(List<String> T1,List<String> T2)
```

`find_mca` function computes the non-trivial minimum common ancestor for each taxa x in T_1 and T_2. It is an internal function that accepts as parameters two *list_ taxa_ ancestors* lists. It returns the list of all non-trivial minimum common ancestors for each taxa. It has the same complexity $O(nk^T)$ of `find_Sim` function, since, in the worst case, the non-trivial minimum common ancestor is reached at the end of the list of the ancestors, that has at most length equal to k^T.

```
private static double compute_COMETH( double SSL, double NCE, double MCB, double
    w1, double w2, double w3, boolean perc )
```

The `compute_COMETH` function computes the global similarity measure. To compute it, it is mandatory to compute the three indexes SSL, NCE and MCB, computed by `compute_SSL`, `compute_NCE` and `compute_MCB` function respectively. While the complexity of `compute_SSL` is linear in n, `compute_MCB` and `compute_NCE` functions have in the worst case complexity $O(k^T n)$.

`compute_COMETH` accepts as input parameters the three indexes `SSL`, `NCE` and `MCB` and the corresponding weights `w1`, `w2` and `w3`, and it returns the weighted sum of the indexes as global measure.

The COMETH similarity measure wrt. a single taxa is computed by using the function `compute_COMETH` where the input indexes SSL, NCE and MCB are filtered by a specific taxon given as input parameter.

Chapter 5

Experimenting COMETH

In this chapter we discuss some experiments by using the COMETH measure, the Robinson-Foulds distance and the cousin pair distance. While COMETH and cousin pairs are measures of similarity, the RF is a dissimilarity measure. We use the COMETH tool and some other tools available for the other distances, namely `treedist` from the PHYLIP package and `cousin.k` respectively.

5.1 First experiment: comparing with a consensus

As first experiment we consider 9 rooted phylogenetic trees and their consensus, computed by using the `consense` program in the PHYLIP package. It accepts as input a list of phylogenetic trees in the Newick format and return their consensus tree. The consensus tree is obtained by using the majority rule. The list of the rooted phylogenetic trees and their consensus is reported below, while their graphical representation is given in Appendix A.

```
 Tree num.    Newick tree format
--------------------------------------------------------------
   1       (A,(B,(H,(D,(J,(((G,E),(F,I)),C))))));
   2       (A,(B,(D,((J,H),(((G,E),(F,I)),C))));
   3       (A,(B,(D,(H,(J,(((G,E),(F,I)),C))))));
   4       (A,(B,(E,(G,((F,I),((J,(H,D)),C))))));
   5       (A,(B,(E,(G,((F,I),(((J,H),D),C))))));
   6       (A,(B,(E,((F,I),(G,((J,(H,D)),C))))));
   7       (A,(B,(E,((F,I),(G,(((J,H),D),C))))));
   8       (A,(B,(E,((G,(F,I)),((J,(H,D)),C))));
   9       (A,(B,(E,((G,(F,I)),(((J,H),D),C))));
--------------------------------------------------------------
Consensus tree: ((((((C,((H,D),J)),(F,I)),G),E),B),A);
```

Each phylogenetic tree is compared with the consensus tree by using the three distances COMETH, RF and cousin pairs. In COMETH the default weights

are used, while the cousin pairs computes kinships up to the second level. We list the result in Table 5.1.

Tree num.	COMETH	`cousin.k`	`treedist`
1	0.6598	0.175	10
2	0.6496	0.146067	10
3	0.6589	0.119047	10
4	1	1	0
5	0.9413	0.75	2
6	0.8486	0.441176	2
7	0.8050	0.484849	4
8	0.7886	0.294117	2
9	0.7486	0.447368	4

Table 5.1: Tabular results for the first experiment on 9 rooted phylogenetic trees compared with their consensus.

The `treedist` program determines RF distance, namely how many bipartitions there are, between the two trees, that are on one tree and not on the other. Generally, the less the number of partitions, the higher is the similarity between the two rooted phylogenetic trees. Let us consider the comparison of the rooted phylogenetic tree 4: `(A,(B,(E,(G,((F,I),((J,(H,D)),C))))));` and its consensus. Since the trees are equal, the similarity computed by COMETH and `cousins.k` is 1, while the number of bipartitions computed by `treedist` is 0. This means that we have no different bipartitions in the two trees, hence the two trees are equal.

The three measures capture well the most similar trees (tree 4 and 5) and the most dissimilar ones (tree 1,2 and 3) from the consensus tree. Table 5.1 highlights also an irregular behaviour for the tree 8. In fact, since for COMETH and `treedist` the tree is quite similar to the consensus tree, `cousin.k` returns a very low value of similarity. This happens because the taxon G has different speciation levels in the two trees and this difference in the subtree `CDFGHIJ` causes a dizzy fall of the entire measure up to 20% of similarity.

For what concern the range of values of each measure, we see clearly that COMETH decrease linearly from 1 to 0.65, while the `cousin.k` has a scattered range of values (from 1 to 0.12). In `treedist`, the range of value is from 0 to 10, but it is not the general range. It depends on the structure of the two trees that we compare.

The COMETH similarity measure allows us to infer also the common evolutionary history for a specified set of taxa and to compute the similarity value only for this subset. In our experimentation, the subset (A, B, E) has an high value of similarity in all the comparisons since each taxa in the set has an high value of non-trivial common evolution in all the rooted phylogenetic trees.

5.2 Second experiment

In this experiment we consider two rooted phylogenetic trees taken from [3] that describe two different representation, reported in Figure 5.1, of four eucaryotes - *Homo Sapiens* (HSA) *Rattus norvegicus* (RNO) *C. elegans* (CEL) *Drosophila melanogaster* (DME) - and a bacterium - *E. coli* (ECO). The first tree matches the standard NCBI taxonomy, while the second is generated by considering three metabolic pathways, *glycolysis*, *private metabolism* and *purine metabolism*.

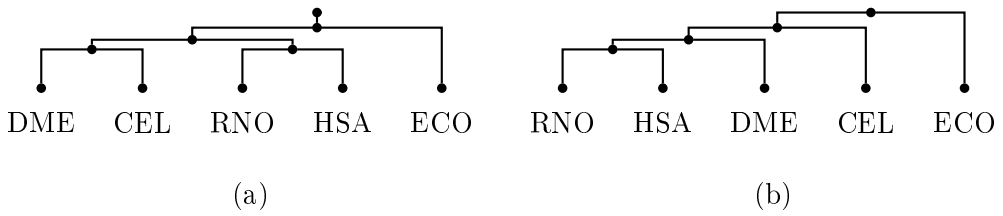


Figure 5.1: Rooted phylogenetic trees for the experiment 2.

We report in Table 5.2 the results computed by using the COMETH tool with the default weights, the `cousins.k` program that computes kinship relations up to the second level, and the `treedist` software to compute RF dissimilarity measure.

COMETH	`cousin.k`	`treedist`
0.7133	0.333333	2

Table 5.2: Tabular results for the second experiment on the comparison of the two rooted phylogenetic trees shown in Figure 5.1.

In this experiment, the `cousins.k` output value signals that the similarity between the two rooted phylogenetic tree is small since the only taxa that have the same kinships are `RNO`,`HSA` and `ECO` in the two trees. For what concern the COMETH result and `treedist` value, they suggest that the two trees are quite similar, except for two taxa that have different structure in the two trees.

COMETH returns an high value since it weights more the non-trivial common evolutions index in the two trees, at the expense of the two other indexes.

Since the only difference in the two rooted phylogenetic trees are the two taxa `CEL` and `DME`, it could be interesting compute the COMETH similarity measure only for these two and discuss the output. For both the taxa, COMETH returns a value of similarity equal to 0.6333. While in the `cousins.k` the two taxa have been not considered since they haven't the same kinship relations in the two trees, in COMETH the two taxa have an high value of similarity since they share the same non-trivial common evolution history in both the trees. The complete output for `CEL` and `DME` is reported in Appendix A.

Chapter 6

Conclusions

The aim of this thesis is to study the comparison techniques between two rooted phylogenetic trees that share the same set of taxa. This can be useful in various contexts, for evaluating the quality of an inferred tree wrt. a reference one or to compare the results of two different phylogenetic inferences. After discussing the most used techniques proposed in the literature, a new measure of similarity is proposed by considering the similarities in the evolutionary history of each single taxa. Such measure can be also used wrt. a subset of taxa or a specific taxon to extract and display individual information.

Three similarity indexes have been defined: *SSL*, that computes the ratio of the taxa that have the same speciation level in both the rooted phylogenetic trees; *NCE* that singles out the non-trivial common evolution for each taxa in the two rooted phylogenetic trees and *MCB*, that considers the non-trivial minimum common ancestor, if it exists, for each taxa in the two rooted phylogenetic trees. The new similarity measure called COMETH is defined as the weighted sum of these indexes.

A prototype tool for comparing two rooted phylogenetic trees by using the COMETH similarity measure has been implemented in JAVA. The prototype tool has been used in different experiments to assert the quality of the proposal. COMETH measure has been compared with some existing methods like the Robinson-Foulds distance and the cousin pairs distance, pointing out the advantages and the weakness of the proposal.

Further developments can be considered. First of all, the tool can be extended with a graphical plugin to highlight the common evolution of a specific taxon or of a subset of taxa. This would exploit the specific characteristic of COMETH that allows one to analyse the comparison wrt. a simple taxon

or a specific group of taxa. In the present version COMETH doesn't consider time on branches in the rooted phylogenetic trees. It would be interesting to develop a timed version of the measure. This would give a stronger value to the comparison based on common evolution.
A further direction for extending the present proposal is to exploit also kinship relationships. This can be done in two ways. One is to explore the integration of COMETH with some measure able to capture kinships among taxa such as *cousin* measure [27]. The other possible direction is to define a further similarity index based on kinship relations to be integrated into COMETH measure.

Appendix A

Experiments

A.1 First experiment: rooted phylogenetic trees representation

Here the representation of the rooted phylogenetic trees used in the first experiment.

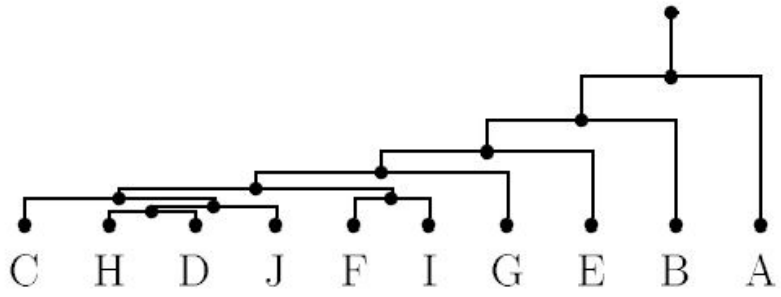


Figure A.1: Consenus: `((((((C,((H,D),J)),(F,I)),G),E),B),A);`

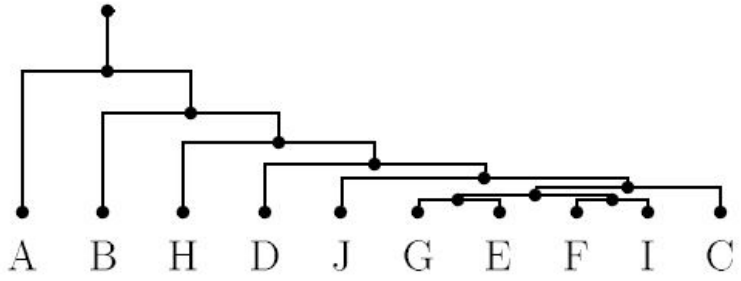


Figure A.2: Tree 1: `(A,(B,(H,(D,(J,(((G,E),(F,I)),C))))));`

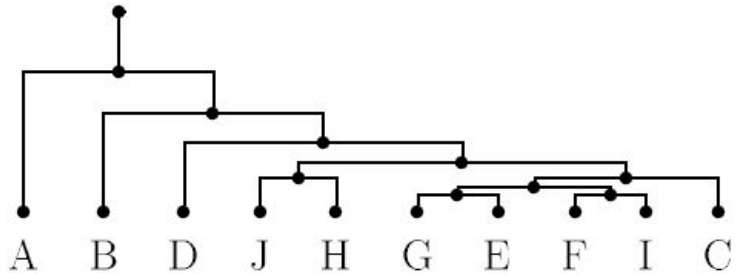


Figure A.3: Tree 2: `(A,(B,(D,((J,H),(((G,E),(F,I)),C))));`

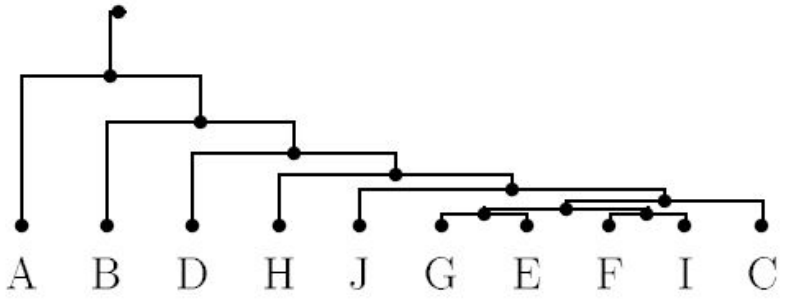


Figure A.4: Tree 3: `(A,(B,(D,(H,(J,(((G,E),(F,I)),C))))));`

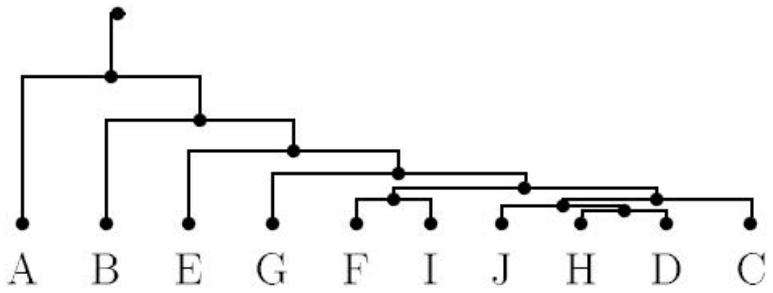


Figure A.5: Tree 4: `(A,(B,(E,(G,((F,I),((J,(H,D)),C))))));`

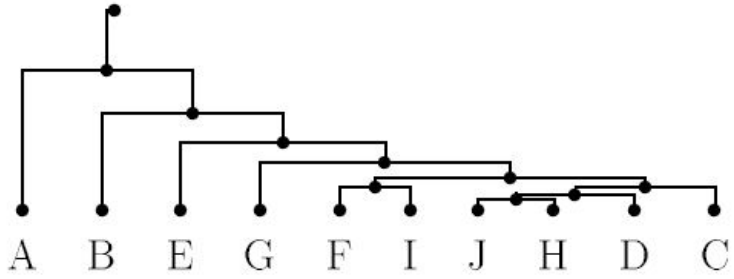


Figure A.6: Tree 5: `(A,(B,(E,(G,((F,I),(((J,H),D),C))))));`

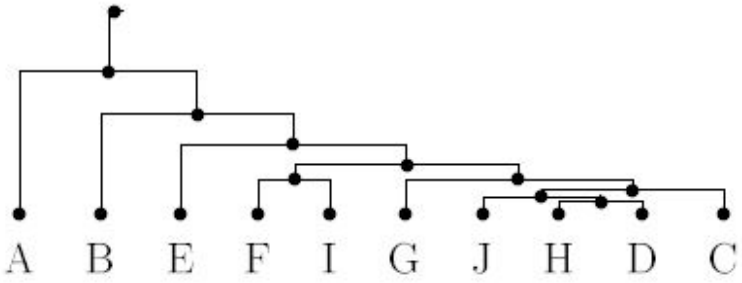


Figure A.7: Tree 6: `(A,(B,(E,((F,I),(G,((J,(H,D)),C))))));`

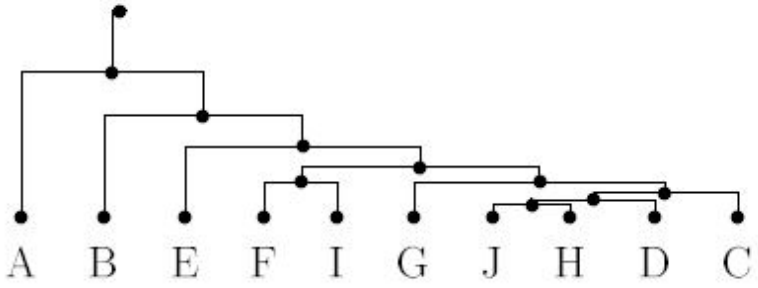


Figure A.8: Tree 7: `(A,(B,(E,((F,I),(G,(((J,H),D),C))))));`

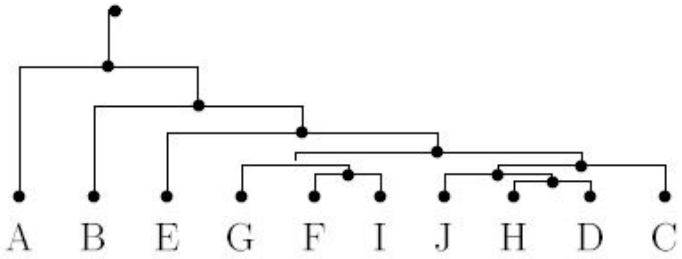


Figure A.9: Tree 8: `(A,(B,(E,((G,(F,I)),((J,(H,D)),C)))));`

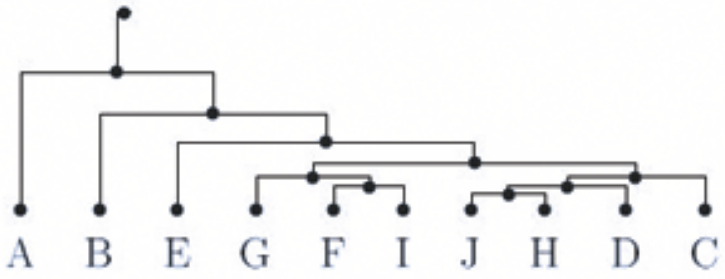


Figure A.10: Tree 9: `(A,(B,(E,((G,(F,I)),(((J,H),D),C)))));`

A.2 Second experiment: COMETH output for CEL and DME

Listing A.1: COMETH output for CEL

```
run:
-------- COMETH tool --------
Panzetta Antonio 834125
Version 1.0
----------------------------

1) COMETH global measure
2) COMETH parametric measure (single taxa)
3) COMETH parametric measure (subset of taxa)

0) Exit

Select an option: 2

Insert T1 in Newick format: ((((RNO,HSA),DME),CEL),ECO);
Insert T2 in Newick format: (((DME,CEL),(RNO,HSA)),ECO);
Do you want to specify weight for COMETH? (y,n) n
Default weights selected

Choose a taxon in ("CEL" "DME" "ECO" "HSA" "RNO" ): CEL

-------- Computing COMETH for "CEL" --------

Weights:
   w1 = 0.2
   w2 = 0.5
   w3 = 0.3

T1: ((((RNO,HSA),DME),CEL),ECO);
   CEL:(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
   DME:(2,DMEHSARNO)(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
   ECO:(4,CELDMEECOHSARNO)
   HSA:(1,HSARNO)(2,DMEHSARNO)(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
   RNO:(1,HSARNO)(2,DMEHSARNO)(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)

T2: (((DME,CEL),(RNO,HSA)),ECO);
   CEL:(1,CELDME)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
   DME:(1,CELDME)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
   ECO:(3,CELDMEECOHSARNO)
   HSA:(1,HSARNO)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
   RNO:(1,HSARNO)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)


- Taxon "CEL" is NOT present in Sim(T1,T2)

- Non-trivial common evolutionary history for "CEL"
   CEL:[(CELDMEHSARNO),(CELDMEECOHSARNO)]

- Non-trivial minimum common bipartition for "CEL"
   CEL: [CELDMEHSARNO]

Indexes:
```

```
53
54    COMETH global measure: 0.2*0.0 + 0.5*0,67 + 0.3*1.0 = 0,63 = 63,33%
55
56  ##### END COMPUTATION ###
57  BUILD SUCCESSFUL (total time: 1 minute 27 seconds)
```

Listing A.2: COMETH output for DME

```
1   run:
2   -------- COMETH tool --------
3   Panzetta Antonio 834125
4   Version 1.0
5   ----------------------------
6
7   1) COMETH global measure
8   2) COMETH parametric measure (single taxa)
9   3) COMETH parametric measure (subset of taxa)
10
11  0) Exit
12
13  Select an option: 2
14
15  Insert T1 in Newick format: ((((RNO,HSA),DME),CEL),ECO);
16  Insert T2 in Newick format: (((DME,CEL),(RNO,HSA)),ECO);
17  Do you want to specify weight for COMETH? (y,n) n
18  Default weights selected
19
20  Choose a taxon in ("CEL" "DME" "ECO" "HSA" "RNO" ): DME
21
22  -------- Computing COMETH for "DME" --------
23
24  Weights:
25     w1 = 0.2
26     w2 = 0.5
27     w3 = 0.3
28
29  T1: ((((RNO,HSA),DME),CEL),ECO);
30     CEL:(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
31     DME:(2,DMEHSARNO)(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
32     ECO:(4,CELDMEECOHSARNO)
33     HSA:(1,HSARNO)(2,DMEHSARNO)(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
34     RNO:(1,HSARNO)(2,DMEHSARNO)(3,CELDMEHSARNO)(4,CELDMEECOHSARNO)
35
36  T2: (((DME,CEL),(RNO,HSA)),ECO);
37     CEL:(1,CELDME)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
38     DME:(1,CELDME)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
39     ECO:(3,CELDMEECOHSARNO)
40     HSA:(1,HSARNO)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
41     RNO:(1,HSARNO)(2,CELDMEHSARNO)(3,CELDMEECOHSARNO)
42
43
44  - Taxon "DME" is NOT present in Sim(T1,T2)
45
46  - Non-trivial common evolutionary history for "DME"
47     DME:[(CELDMEHSARNO),(CELDMEECOHSARNO)]
48
49  - Non-trivial minimum common bipartition for "DME"
50     DME: [CELDMEHSARNO]
```

```
51
52 Indexes:
53
54    COMETH global measure: 0.2*0.0 + 0.5*0,67 + 0.3*1.0 = 0,63 = 63,33%
55
56 ##### END COMPUTATION ###
57 BUILD SUCCESSFUL (total time: 26 seconds)
```

Bibliography

[1] Amihood Amir and Dmitry Keselman. Maximum agreement subtree in a set of evolutionary trees: Metrics and efficient algorithms. *SIAM J. Comput.*, 26(6):1656–1669, 1997.

[2] P. Arabie and S. A. Boorman. Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology*, 10:148–203, 1973.

[3] Paolo Baldan, Nicoletta Cocco, and Marta Simeoni. Comparison of metabolic pathways by considering potential fluxes. In *BioPPN2012 - 3rd International Workshop on Biological Processes and Petri Nets, satellite event Proc. BioPPN 2013, a satellite event of PETRI NETS 2013 Metabolic Pathways through Potential Fluxes 15 of Petri Nets 2012*, 2012.

[4] Sebastian Böcker, Stefan Canzar, and Gunnar W. Klau. The Generalized Robinson-Foulds metric. In Aaron Darling and Jens Stoye, editors, *Algorithms in Bioinformatics*, volume 8126 of *Lecture Notes in Computer Science*, pages 156–169. 2013.

[5] Damian Bogdanowicz and Krzysztof Giaro. Matching split distance for unrooted binary phylogenetic trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(1):150–160, 2012.

[6] Damian Bogdanowicz and Krzysztof Giaro. On a matching distance between rooted phylogenetic trees. *Applied Mathematics and Computer Science*, 23(3):669–684, 2013.

[7] David Bryant. *Building trees, hunting for trees and comparing trees.* PhD thesis, Department of Mathematics, University of Canterbury, 1997.

[8] B. DasGupta, X. He, T. Jiang, M. Li, J. Tromp, and L. Zhang. On distances between phylogenetic trees. In *Proceedings 8th ACM/SIAM Symposium Discrete Algorithms*, pages 427–436. (SODA), 1997.

[9] William Day. Optimal algorithms for comparing trees with labeled leaves. *Journal of Classification*, 2(1):7–28, 1985.

[10] Alan de Queiroz and John Gatesy. The supermatrix approach to systematics. *Trends in Ecology & Evolution*, 22(1):34–41, 2007.

[11] Frederic Delsuc, Henner Brinkmann, Nicolas Lartillot, and Herve Philippe. Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics, Annual Reviews*, 36:541–562, 2005.

[12] Frederic Delsuc, Henner Brinkmann, and Herve Philippe. Phylogenomics and the reconstruction of the tree of life. *Nature Review Genetics*, 6:361–375, 2005.

[13] Jonathan A. Eisen. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis, genome res. In *Genome Research*, pages 163–167. Cold Spring Harbor Laboratory Press, 1998.

[14] Harold N. Gabow and Robert Endre Tarjan. Faster scaling algorithms for network problems. *SIAM J. Comput.*, 18(5):1013–1036, 1989.

[15] Matthew E. Hayes. Phylogeny evaluation in biology. In *Simulating Malware Evolution for Evaluating Program Phylogenies*, pages 5–15. University of Louisiana at Lafayette, 2008.

[16] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[17] Podani Jànos, Engloner Attila, and Major Agnes. Multilevel comparison of dendrograms: A new method with an application for genetic classifications. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–14, 2009.

[18] Philippe Lemey, Marco Salemi, and Anne-Mieke Vandamme. *The Phylogenetic Handbook. A Practical Approach to Phylogenetic Analysis and Hypothesis Testing.* Cambridge University Press, 2nd edition, 2009.

[19] Wen-Hsiung Li and Dan Graur. *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, Massachusetts, 1991.

[20] Yu Lin, Vaibhav Rajan, and Bernard M. E. Moret. A metric for phylogenetic trees based on matching. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(4):1014–1022, 2012.

[21] David W. Mount. Phylogenetic prediction. In *Bioinformatics. Sequence and Genome Analysis*, pages 237–279. Cold Spring Harbor Laboratory Press, 2001.

[22] Roderick Page and Eward Holmes. *Molecular Evolution: a phylogenetic approach*. Blackwell Science Ltd, 1998.

[23] Nicholas D. Pattengale, Eric J. Gottlieb, and Bernard M. E. Moret. Efficiently computing the Robinson-Foulds metric. *Journal of Computational Biology*, 14(6):724–735, 2007.

[24] Nicholas D. Pattengale, Krister M. Swenson, and Bernard M. E. Moret. Uncovering hidden phylogenetic consensus. In Mark Borodovsky, Johann Peter Gogarten, Teresa M. Przytycka, and Sanguthevar Rajasekaran, editors, *ISBRA*, volume 6053 of *Lecture Notes in Computer Science*, pages 128–139. Springer, 2010.

[25] Herve Philippe and Mathieu Blanchette. Overview of the first phylogenomics conference. *BMC Evolutionary Biology*, 7(Suppl 1), 2007.

[26] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.

[27] Dennis Shasha, Jason Tsong-Li Wang, and Sen Zhang. Unordered tree mining with applications to phylogeny. In *ICDE*, pages 708–719. IEEE Computer Society, 2004.

[28] Giorgio Valle, Manuela Helmer Citterich, Marcella Attimonelli, and Graziano Pesole. Evoluzione molecolare. In *Introduzione alla bioinformatica*, pages 121–129. Zanichelli, 2003.

[29] Silke Wagner and Dorothea Wagner. Comparing Clusterings – An Overview. Technical Report 2006-04, Universität Karlsruhe (TH), 2007.