



**HAL**  
open science

# Concepts of Classification and Taxonomy. Phylogenetic Classification

Didier Fraix-Burnet

► **To cite this version:**

Didier Fraix-Burnet. Concepts of Classification and Taxonomy. Phylogenetic Classification. D. Fraix-Burnet & S. Girard. Statistics for Astrophysics: Clustering and Classification , 77, EDP Sciences, pp.221-257, 2016, EAS Publications Series, 978-2-7598-9001-9. 10.1051/eas/1677010 . hal-01324988

**HAL Id: hal-01324988**

**<https://hal.science/hal-01324988>**

Submitted on 3 Jun 2016

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

## CONCEPTS OF CLASSIFICATION AND TAXONOMY PHYLOGENETIC CLASSIFICATION

Didier Fraix-Burnet<sup>1, 2</sup>

**Abstract.** Phylogenetic approaches to classification have been heavily developed in biology by bioinformaticians. But these techniques have applications in other fields, in particular in linguistics. Their main characteristics is to search for relationships between the objects or species in study, instead of grouping them by similarity. They are thus rather well suited for any kind of evolutionary objects. For nearly fifteen years, astrocladistics has explored the use of Maximum Parsimony (or cladistics) for astronomical objects like galaxies or globular clusters. In this lesson we will learn how it works.

### 1 Why phylogenetic tools in astrophysics?

#### 1.1 History of classification

The need for classifying living organisms is very ancient, and the first classification system can be dated back to the Greeks. The goal was very practical since it was intended to distinguish between eatable and toxic aliments, or kind and dangerous animals. Simple resemblance was used and has been used for centuries. Basically, until the XVIIIth century, every naturalist chose his own criterion to build a classification. At the end, hundreds of classifications were available, most often incompatible to each other. The criteria for this traditional way of classifying is the subjective appearance of the living organisms.

During the XVIIIth a revolution occurred. Scientists like Adanson and Linné devised new ways of classifying the objects and naming the classes. Adanson realised that all the observable traits should be used, giving birth to the multivariate clustering and classification activity (Adanson, 1763). Linné based his binomial nomenclature on neutral names unrelated whatsoever to any property of the classes. We can realise the success of these two ideas more than two centuries and a half later!

---

<sup>1</sup> Univ. Grenoble Alpes, IPAG, F-38000 Grenoble, France

<sup>2</sup> CNRS, IPAG, F-38000 Grenoble, France

The hierarchical organization of the living organisms was already established when Linné devised his nomenclature, but its origin became understood in the mid-XIXth thanks to Darwin. Evolution was the key to understand the hierarchy and interpret the scheme as depicting the evolutionary relationships between the species. Then, biologists have since devoted themselves to establish the phylogenetic tree of all living organisms, the Tree of Life.

Partitioning or hierarchical clustering techniques were used in this purpose by comparing species by their global similarity. However, William Hennig in 1950 translated the transmission with modification idea of Darwin into a new concept of classification: species are not compared anymore on the resemblance, but on their ascendance. We are not looking anymore for who is like who, but who is cousin of whom. Cladistics was born (Hennig, 1965).

Indeed, it seems that linguists already used a similar approach to compare the languages and their diversification, but never formalised it like Hennig. Harsh debates occurred in the evolutionary biology community, but this approach, better called now Maximum Parsimony, was definitively adopted in the 1980s.

Despite that other techniques were developed and are very often used nowadays, the philosophy of cladistics is still behind all these approaches. Cladistics is conceptually simple and very general, yet it is also very demanding computationally.

Table 1 summarizes the evolution of the concepts of classification. It is important to keep this in mind. In astrophysics, we are still essentially using the traditional way of classifying objects.

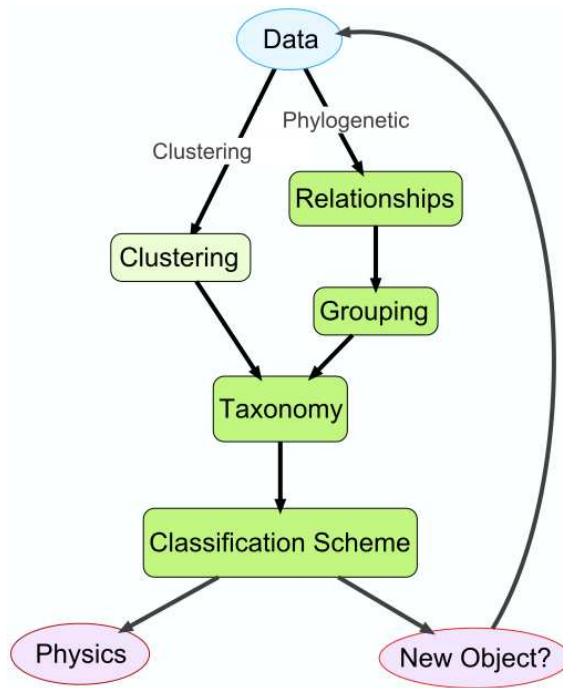
|        |                 |                       |                               |
|--------|-----------------|-----------------------|-------------------------------|
| ~ -300 | <b>Aristote</b> | First classification  | <i>Appearance, usefulness</i> |
| ~ 1730 | <b>Linné</b>    | Binomial nomenclature | <i>Objective taxonomy</i>     |
| 1763   | <b>Adanson</b>  | Use of all parameters | <i>Global similarity</i>      |
| 1850   | <b>Darwin</b>   | Evolution             | <i>Inheritance, mutations</i> |
| 1950   | <b>Hennig</b>   | Cladistics            | <i>Common history</i>         |

**Table 1.** A summary of the evolution of the concepts of classification.

Since astronomical objects are evolving, it is natural to consider that phylogenetic tools might be adapted. The transmission with modification notion is not restricted to the living organisms. Two spiral galaxies that merge create a new object, possibly a galaxy of elliptical shape, which contains the material from its progenitors (transmission) but with modified properties like the kinematics or new generations of stars (modification). It can also be purely virtual. For instance, if you want to transform a red triangle to a red circle, you have to transmit the color but modify the shape.

## 1.2 Two routes for classification

Two thousands years of classification in biology have taught us that there are two paths toward devising a classification scheme (Fig. 1).



**Fig. 1.** The two routes to build a classification scheme. The clustering path gathers objects together, any relationships between the classes are established afterwards using physical or biological arguments. Conversely, the phylogenetic approach looks for the relationships from which groups are defined.

The first path is to gather objects according to appearance or global similarity. This is clustering. The groups are then characterized, described and defined, and are given names respecting rigorous taxonomic rules. All this provides the classification scheme that will later be used for (supervised) classification.

Having grouped objects together is the end of the statistical investigation. Now we wish to understand why the objects are grouped in this way, and what are the relationships between the groups. In other words, what is the origin of the groups?

In biology, the key is evolution and Darwin showed that the transmission with modification mechanism creates the hierarchical organization of the living organisms, the diversity of which can thus be well represented on a tree-like diagram.

In astrophysics, Hubble made a clustering of the galaxies he discovered, defining categories like elliptical or spiral, and later hypothesized that the dissipation would

flatten elliptical galaxies into disk-like ones like spiral galaxies. This is the famous Hubble Tuning Fork diagram that established the relationships between the classes after the clustering process.

Since we are interested in the relationships between species, why not try to gather the living organisms according to their relatedness? This is the idea of the phylogenetic path toward a classification. It establishes the relationships first, and then the grouping of the objects is deduced from the tree.

The phylogenetic approach is the subject of this lecture. In Sect. 2 we present a few techniques based on the pairwise distances between the objects in study. In Sect. 3 we present the techniques that use the parameters themselves, and detail the Maximum Parsimony (or cladistics).

### 1.3 Taxonomy

Taxonomy is the science of naming, defining, describing and classifying that have been developed for the living organisms.

It is not simple to cluster objects into classes, and it is not obvious to classify new objects into known classes. This depends very much on the description of the classes. In particular, the names of the classes can have a strong influence on this process.

This difficulty has been recognised by Linné in the XVIIIth century. The enormous success of the nomenclature he proposed is due to the fact that the name of a class is not related to any of its specific properties. Why? Let us take an example: what do you think the class of “elliptical galaxies” is made of? Star-forming galaxies? Radio galaxies? Massive galaxies? No, of course, all galaxies with an elliptical shape belong to this class: the name of the class inevitably plays also the role of its definition and its description, whatever is the diversity in the other properties.

This rule is indeed quite general. I give below a few examples of some classes of objects with their name and (short) definition. Try to deduce the kind of objects with the name only:

- Bird: class *Aves*, a group of endothermic vertebrates, characterised by feathers, a beak with no teeth, the laying of hard-shelled eggs, a high metabolic rate, a four-chambered heart, and a lightweight but strong skeleton.
- Mammals: class *Mammalia*, any members of a clade of endothermic amniotes distinguished from reptiles and birds by the possession of hair, three middle ear bones, mammary glands, and a neocortex (a region of the brain).
- Particle: a minute fragment or quantity of matter. Classified as electron, proton, neutron, Z...
- Star: a luminous sphere of plasma held together by its own gravity. Classified as O A B..., red giant, dwarf...

- Galaxy: a gravitationally bound system of stars, stellar remnants, interstellar gas and dust, and dark matter. Classified as spiral, elliptical, ultra-luminous infrared, blue, giant....

In biology, many species have a common name which is not the scientific name. Note that the name of the class Mammalia strongly suggests the presence of mammary glands... Interestingly, in particle physics, it is impossible to guess the nature of the particles from their names. The quarks have poetic names that has nothing to do with their properties (charm, beauty...).

In astrophysics, this is not so rigorous. The first and most used classification of stars respects the fundamental taxonomic rule, but other classes have been added to this picture following apparent physical or chemical properties, and mainly describing evolutionary stages.

For galaxies, this is far worse, there are hundreds of classifications according to the many observables we managed to obtain. I cannot refrain from seeing here a parallel with the situation of biological classification at the end of the XVIIIth century that motivated the works by people like Linné and Adanson.

So the requirement of multivariate classifications in astrophysics appears not only natural, but somewhat urgent.

#### 1.4 Stars, galaxies: multivariate objects in evolution

Evolution is ubiquitous in our expanding Universe, and a multivariate description of astrophysical objects is necessary to distinguish different species, to relate together the evolutionary stages or the progenitors, and to build and constrain our physical models.

When two spiral (disky) galaxies merge, they generally yield a galaxy with an elliptical shape. When a huge cloud of gas collapses, it can also yield an elliptical galaxies. These two scenarios, validated by numerical simulations, implies extremely different evolutionary histories for the evolution and formation of galaxies. It has also great consequences for our knowledge of the evolution of our Universe: in the first case a rather long time is necessary, while in the second case it can happen rather fast and early. Yet, the final product looks the same when considering only the morphology.

It is quite instructive to recognise these two different populations of elliptical galaxies. For this purpose, other properties like the kinematics and the gas content should be of good help.

This little example is a convincing case for multivariate and evolutionary studies of the diversity of astrophysical entities. This is precisely the goal of phylogenetic analyses (e.g. see a review of multivariate approaches for galaxy classification in Thuillard & Fraix-Burnet, 2015).

## 2 Distance-based approaches

### 2.1 Minimum Spanning Tree (MST)

A spanning tree is an acyclic, connected graph  $G = (V, E)$ , where  $V$  are vertices (nodes) and  $E$  edges (branches) plus a weight function:

$$w : E \rightarrow \mathbb{R} \quad (2.1)$$

$$e \rightarrow w(e) \quad (2.2)$$

The Minimum Spanning Tree is the tree  $T$  minimizing:

$$w(T) = \sum_{e \in T} w(e) \quad (2.3)$$

If the weights  $w(e)$  are distinct, then the solution is unique.

To perform this computation, we need a matrix with the objects, that will be the nodes, and the weights between any two of them. These weights can be pairwise distances computed from the parameters describing the objects, or dissimilarities, or any other “cost” that is pertinent for the study.

How can we make a clustering from a MST? The branches depict some relationships between the connected points according to the minimisation. By cutting the longest branches, the ones with  $w(e) > w_{\text{threshold}}$ , we can separate bunches of branches, effectively creating groups. This is sometimes called the separation of the tree. This process is similar to the simple linkage criterion in the hierarchical clustering method, and the friends-of-friends algorithm or Nearest Neighbor algorithm known to astronomers (Gower & Ross, 1969; Feigelson & Babu, 2012).

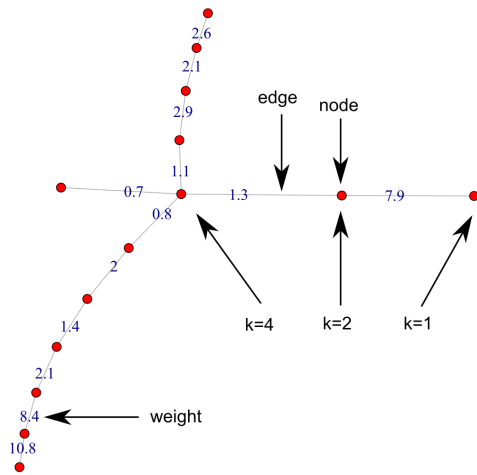
The MST technique has been long used by cosmologists to map the cosmic web and spatially clustering galaxies (Barrow et al., 1985). To get a better idea of the distinct filaments and their respective orientations, it is possible to prune the tree, i.e. to remove some branches, in particular the ones which are leading to a single terminal node (degree  $k = 1$ ) and originate from a node of degree  $k=3$  (three branches connects to this node).

### 2.2 Neighbor Joining Tree Estimation (NJ)

Neighbor-Joining (NJ, Saitou & Nei, 1987; Gascuel & Steel, 2006) is the most popular distance-based approach to construct a phylogenetic tree. This is a bottom-up hierarchical clustering methods starting from a star tree (unresolved tree). The procedure runs as follows (see Fig. 3):

1. from the distance (or dissimilarity) matrix for the  $n$  objects giving the distances  $d(i, j)$  between objects  $i$  and  $j$ , compute the “corrected” distance  $Q(i, j)$  which represents a kind of evolutionary distance measured on the tree:

$$Q(i, j) = (n - 2)d(i, j) - \sum_{k=1}^n d(i, k) - \sum_{k=1}^n d(j, k) \quad (2.4)$$



**Fig. 2.** An example of a Minimum Spanning Tree with some definitions (see text). The index  $k$  is the degree of a node, i.e. the number of branches connected to it.

2. find the lowest  $Q$  and relate the two corresponding objects with a new node  $u$ . This new node is called an internal node.

3. replace  $i$  and  $j$  by  $u$ , rebuild the new distance matrix with the  $n - 1$  objects using the following definition:

$$d(u, k) = \frac{1}{2} [d(i, k) - d(i, u)] + \frac{1}{2} [d(j, k) - d(j, u)] \quad (2.5)$$

4. iterate from 1 until all the objects have been considered.

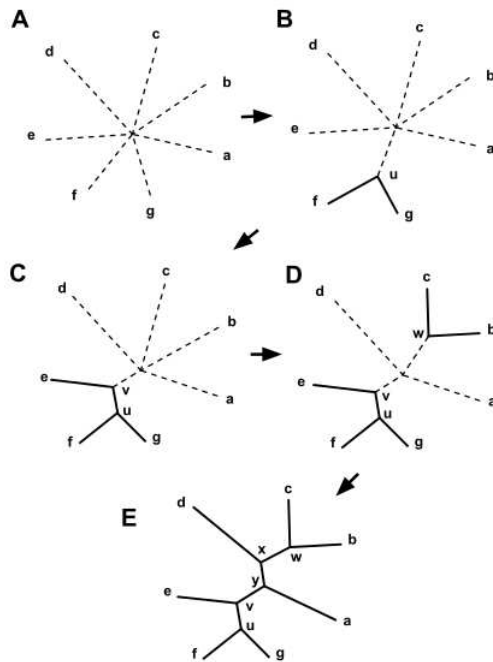
Neighbor-Joining minimizes a tree length, according to a criteria that can be viewed as a Balanced Minimum Evolution (Gascuel & Steel, 2006). It furnishes a simple algorithm to reconstruct a tree from the distance matrix.

### 2.3 Difference between a hierarchical tree and a phylogenetic tree

The hierarchical clustering trees, the MST and the NJ trees are all based on distances computed from the parameters and are used to define groups. Are there really different and does it matter? Let us take a simple and concrete example (Fig. 4).

The hierarchical tree depicts the relative distances between the objects. For instance, it tells you that Les Houches and Chamonix are much closer to each other than they are from Geneva. But this does not tell you whether, when you will leave the conference centre in Les Houches to go the Geneva airport, you should go through Chamonix. A hierarchical clustering tree does not provide you with the relationships between the cities.





**Fig. 3.** A diagram to show the successive steps of a Neighbor-Joining Tree Estimation (Source: [https://en.wikipedia.org/wiki/Neighbor\\_joining](https://en.wikipedia.org/wiki/Neighbor_joining)).

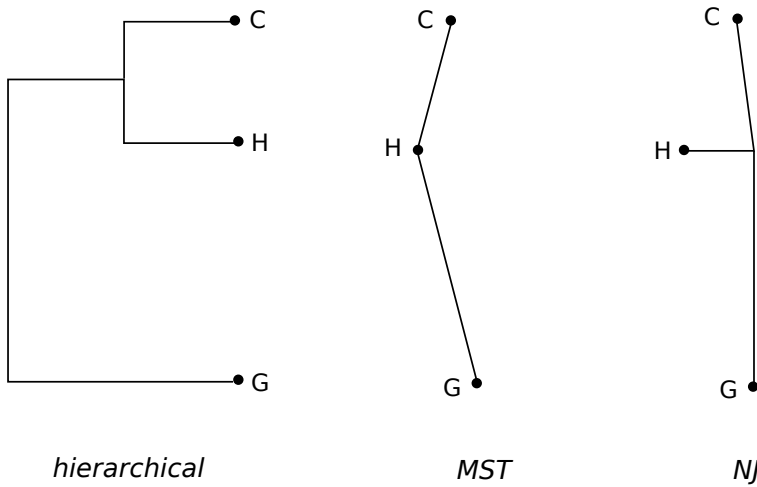
For this you need a phylogenetic tree. The MST tells you that from Les Houches you should go directly to Geneva without visiting Chamonix. The relationships situates Les Houches in between Chamonix and Geneva.

But this is not entirely correct, since the road between Chamonix and Geneva does not go through the village of Les Houches. There is crossroad just outside Les Houches. The reason is that whatever your departure point to go to Geneva, you will take a shorter path. This crossroad is an “internal node” introduced in the NJ method, as well as the cladistics we are going to present.

### 3 Character-based approaches

Parameters that, after discretization, can be given ancestral and derived states are called characters. This means that these parameters describing the objects to classify have kept a trace of the historical evolution of the different species.

A big advantage of character-based approaches is their ability to take uncertainties or unknowns into account. It is of course not possible to compute a distance



**Fig. 4.** Three different kinds of trees representing the respective situations of the cities of Chamonix (C), Les Houches (H) and Geneva (G). Left: a hierarchical clustering tree. Middle: a MST tree. Right: A NJ tree showing the addition of an internal node. See text for details.

when a data is missing. Here, one can simply evaluate the different possibilities allowed by a proposed range of values, and then select the best ones according to the optimization criterion. In the case of unknown parameters, the phylogenetic tree thus provides a prediction for the unknown values.

### 3.1 Maximum Parsimony (cladistics) and Maximum Likelihood

The Maximum Parsimony (cladistics) algorithm compares objects according to their shared common histories using the characters. It selects the phylogenetic tree that minimizes the total number of state changes, which depicts the simplest evolution scenario given the data set (Felsenstein, 1984).

Maximum Parsimony is a powerful approach to find tree-like arrangements of objects. Its main drawback is that the analysis must consider all possible trees before selecting the most parsimonious one. The computation complexity depends on the number of objects and character states, so that too large samples (say more than a few thousands) cannot be analyzed.

Another class of character-based techniques relies on an a priori models of character evolution (probabilities of state changes). For instance, the Maximum Likelihood algorithm selects the most probable phylogenetic tree given the proposed evolutionary model (Williams & Moret, 2003). This kind of approaches are mainly developed for genetic data (evolution of nucleotide sites, mutation and substitution rates of genes...) and they can be applied to big data sets.

In this chapter we present cladistics, or Maximum Parsimony, in great details since it is conceptually the simplest and also the most general phylogenetic method.

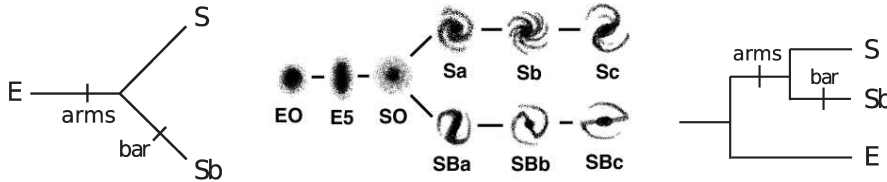
### 3.2 Cladistics: constructing a Tree

Let us begin with a very simple case. Consider three families of objects, E, S, and Sb, characterized by two parameters: Arm and Bar. These parameters can be present (code 1) or absent (code 0). The Table 2 gives our data.

|    | Arm | Bar |
|----|-----|-----|
| E  | 0   | 0   |
| S  | 1   | 0   |
| Sb | 1   | 1   |

**Table 2.** Three families E, S and Sb, of objects described with two parameters Arm and Bar which can be present (1) or absent (0).

There is only one possible tree to link the three objects with a tree. It is given in the graph below (left).



**Fig. 5.** Left: the unrooted tree that can be built with three objects (see Table 2). Middle: the famous Hubble Tuning Fork diagram. Right: the rooted tree corresponding to the left figure.

If we follow the evolution from E to S, we have to change the Arm parameter from 0 to 1. This is indicated by the tick mark on the branch starting from E. To go from E to Sb, we have to add a tick mark before Sb to indicate a change in the parameter Bar from 0 to 1. Going from Sb to S means changing the Bar parameter from 1 to 0 as already indicated by the tick mark on the Sb branch. With such a diagram and two tick marks, we have depicted entirely the table above with the possible (but still hypothetical at this stage) evolutionary scenario.

Obviously this sample describes the simplest classification of galaxies, the one using morphology: Elliptical, Spiral and Barred Spiral galaxies. It is striking to note that the cladogram on the left is so much similar to the famous Hubble Tuning

Fork diagram (middle) that Hubble himself drew to depict is thought on galaxy evolution in the 1930's!

With a cladogram, we can go one step further by adding an arrow of evolution (biologists prefer to say diversification). In this purpose, if we impose or know that the state “0” is more primitive (ancestral) than the state “1” (derived), in other words spiral structures and bars appeared in the course of the evolution of the Universe, we obtain the so-called rooted diagram depicted on the right. Note that the leftmost branch and node are not attributed to any objects. This is a rule, since it is always possible that a new object will be discovered and will have to be inserted somewhere in the tree. Hence, the common ancestor of a group is always assumed to be unknown.

Consider the case with four objects described by only one parameter as given in Table 3.

|   |   |
|---|---|
| O | 0 |
| A | 1 |
| B | 1 |
| C | 0 |

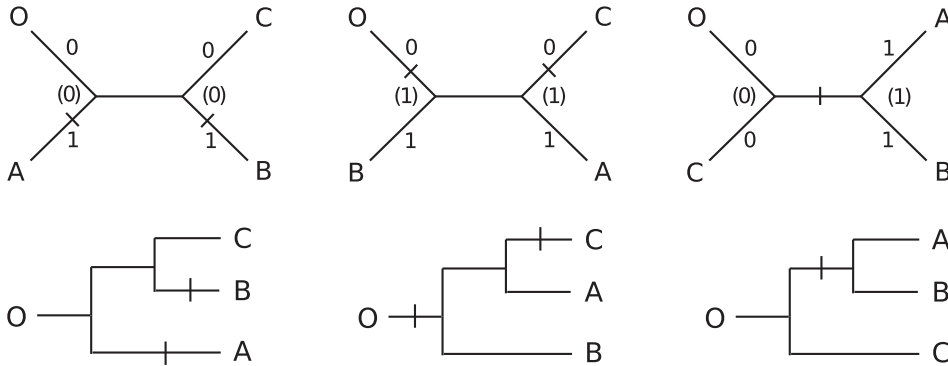
**Table 3.** Four objects described by one parameter with two states, 0 and 1.

You can make the exercise to find all the possible arrangements of the objects on a tree (unrooted). The result is that there are now three possible solutions. This number depends only on the number of objects, and increases incredibly fast with it. It becomes very difficult (and boring) to draw them all by hand for more than 7 or 8 objects. Even modern computers take a huge amount of time to explore the exhaustive tree space for more than a hundred trees.

The three possible arrangements with four objects are shown in Fig. 6. For each tree, we have to indicate the parameter changes to evolve from any one of the objects to any other ones, choosing the simplest solution to put as few tick marks as possible. Each tick mark is called a step, and the total number of steps found on a tree indicates the cost of evolution. This is a measure of the complexity of the evolutionary scenario depicted by the tree. We notice that the tree on the right has only one step, it is the simplest of all the possible arrangements for this four objects, i.e. the most parsimonious tree that is chosen as the most probable evolutionary scenario.

On the unrooted trees, we have added in parentheses the value of the parameter at the node. This is for clarity, and it is not necessary in practice to count the number of steps.

In Fig. 6, the rooting of the trees has assumed that the value “0” is the ancestral state. As an exercise you can draw the rooted trees assuming the alternative that the value “1” is the ancestral value, and count the number of required changes.



**Fig. 6.** The three possible arrangements of four objects on a tree (top: unrooted, bottom: rooted). The two trees to the left have a total score of two steps, and the right tree is the simplest scenario with only one step.

You will find that the total score of a tree does not depend on the choice of the root.

Consider the sample data given in Table 4. We have the same four objects as previously, so that, the possible arrangements on a tree depending only on the number of objects, the possible trees with four objects are identical to the ones depicted in Fig. 6. But we have more information on these objects since they are described by five parameters. The total number of steps depends strongly on the number of parameters.

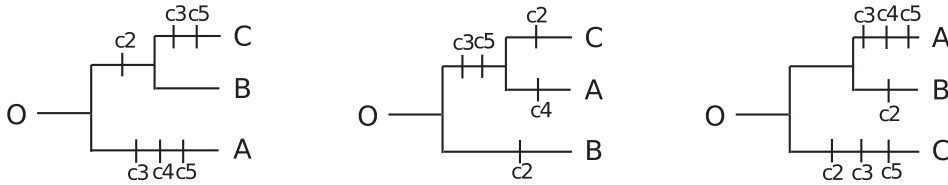
|   | c1 | c2 | c3 | c4 | c5 |
|---|----|----|----|----|----|
| O | 0  | 0  | 0  | 0  | 0  |
| A | 0  | 0  | 1  | 1  | 1  |
| B | 0  | 1  | 0  | 0  | 0  |
| C | 0  | 1  | 1  | 0  | 1  |

**Table 4.** Four objects described by five binary parameters.

Note that the parameter  $c_0$  is constant, it is non-informative and can be dropped. The unrooted trees are the same as previously, but the rooting depends on the parameter states and the choice of the ancestral states. Assuming “0” as the ancestral state for all the parameters, object O is obviously the most ancestral object and we show only the rooted trees.

As shown in Fig. 7, the number of tick marks is necessarily higher than previously (Fig. 6), and the simplest tree is now the one in the middle. This illustrates

the important point that a classification, or an evolutionary scenario changes depending on the data availability.



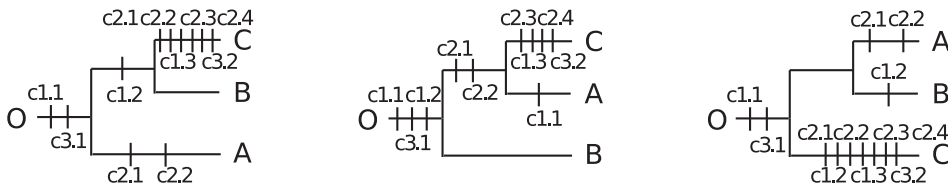
**Fig. 7.** Trees obtained with the sample data given in Table 4. From left to right, the scores of the trees are 6, 5 and 7 steps.

Up to now, we have considered parameters with only two states, 0 or 1. The Table 5 contains four objects as before, and three parameters with different number of states (up to 5 for parameter c2).

|   | c1 | c2 | c3 |
|---|----|----|----|
| O | 0  | 0  | 0  |
| A | 1  | 2  | 1  |
| B | 2  | 0  | 1  |
| C | 3  | 4  | 2  |

**Table 5.** Four objects described by three parameters with several states.

The unrooted trees are the same, the rooted ones also if we again assume “0” as the ancestral state, and thus object O as root, but the tick marks are more complicated to read. In Fig. 8 we have adopted the following notation: cn.x means character cn goes from x to x-1 or reversely.



**Fig. 8.** Trees obtained with the sample data given in Table 5. From left to right, the scores of the trees are 11, 10 and 12 steps.

Cladistics is not more complicated than that. Computers come to our help

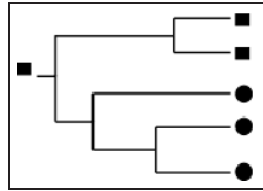
because searching for the simplest tree is rather tedious. Nevertheless, there are some important questions to address before performing analyses with real data.

### 3.3 Parameters as Characters

Cladistics compares objects from innovations inherited from a common ancestor. Characters are parameters that can trace this process of transmission with modification. States are discrete values taken by the characters, they supposedly describe the evolutionary stages of the character. To keep trace of an innovation, characters must have the right evolutionary behaviour.

#### Synapomorphy

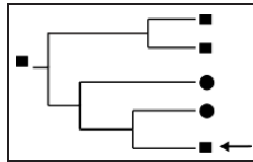
A state is specific to a clade (an ancestor and all its descendants). On the example to the right, the value “disk” of the character “shape” defines a clade gathering objects that inherited this property from a common ancestor. The five objects derive from a common ancestor, and thus define a clade, but this should be justified by other parameters than the shape.



Synapomorphies are the ideal situation for a phylogenetic reconstruction. The following behaviours (called homoplasies) must be avoided because they can destroy a tree:

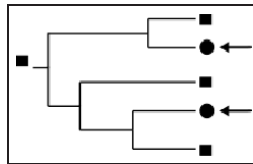
#### Reversal

An innovative state is lost again in favour of a previous character state. Here, the disk value is lost and the character evolves back to the square shape.



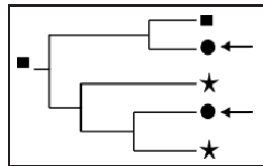
#### Parallel Evolution

The same evolutionary sequence occurs on two different paths, like the square to disk state evolution here.



#### Convergence

The same character state results from different evolutionary sequences. Here, the disk value has occurred both from square and from star states.



Some homoplastic characters are allowed if they are not too many. In practice it is difficult to avoid them. Most often the behaviours of characters can be determined once the phylogeny is established, so that there is no other way to find the synapomorphies than a trials and errors approach.

The Maximum Parsimony search minimizes the number of homoplasies and maximizes the number of synapomorphies.

### 3.4 Counting the steps: the cost of evolution

To evaluate the complexity of trees, we count the number of steps. Up to now, we have assumed a very simple algebra, the difference between two consecutive integer states being one step. However, is it possible to define more complex (realistic?) character transformations. Here are some common hypotheses ( $i$  and  $j$  are states of character  $k$ ,  $f_k(i)$  being the value of state  $i$ ). Usually  $f_k(i) = i$ :

- **Ordered or Wagner:** states are reversible and additive:  $d = |f_k(i) - f_k(j)|$ .
- **Unordered or Fitch:** states are reversible and non-additive:  $d = 1$  if  $i \neq j$ ,  $0$  if  $i = j$ .
- **Dollo:** each state occurs only once: no reversal, no parallel evolution.
- **Camin-Sokal:** states are irreversible:  $(\forall i > j, f_k(i) > f_k(j))$  or  $(\forall i > j, f_k(i) < f_k(j))$ .

### 3.5 The most Parsimonious Tree

The score of a tree is the total number of steps (change of parameters values of states). This is the total cost of evolution  $S = \sum_k \sum_{i>j} d(f_k(i), f_k(j))$ , given the choice of  $d$  made previously.

Among all possible trees, we choose the one with the lowest score, called the most parsimonious tree (Ockham's razor). This is why cladistics is also called the Maximum Parsimony method.

It is frequent that several equally most parsimonious trees are found. To synthesize the result, we build consensus Tree. Each node of this tree is valued with the percentage of occurrence of this node among all the most parsimonious trees. A example of a strict consensus tree (100% nodes only) is shown in Fig. 9.

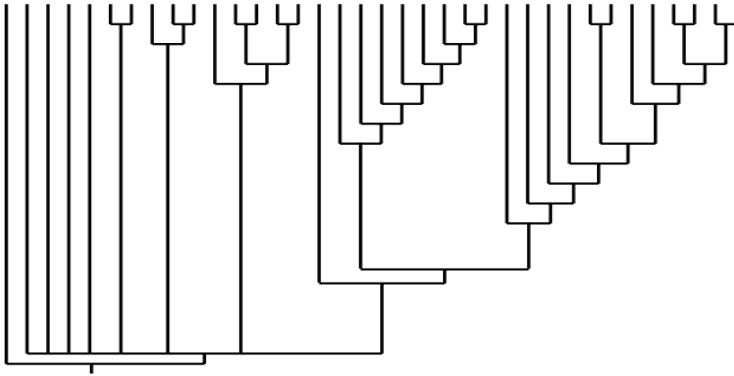
Bifurcating nodes (from which three branches emerge) are robust nodes since they occur in all possible most parsimonious trees. They are said to be resolved nodes. Multifurcating nodes indicate a lack of information or constraint from the data. This may also be due to too many homoplastic characters. It is commonly assumed that nodes with a consensus percentage higher than 70% are quite safe, and lower 50% not significant at all. Note that the consensus of two perfectly resolved but different trees is not perfectly resolved, meaning that there are two possible robust solutions. Only the interpretation can decide which is the most interesting.

It might be interesting at this level to compare the MST and MP techniques since they both search for a minimum path (Table 6).

Interestingly, the score of a MP tree is bounded by the total weight of the MST tree:

$$\frac{w_{min}}{2} < S_{min} < w_{min}$$





**Fig. 9.** Strict consensus tree (100% nodes only).

| MST   | Cladistics  |
|---|---|
| Labelled tree<br>No internal node<br>Polynomial time<br>$w(T) = \sum_e w(e)$          | Semi-labelled tree<br>More general<br>NP-hard<br>$S = \sum_k \sum_{i>j} d(f_k(i), f_k(j))$      |
| For a l1 norm:  |   |
| $w(e) = \sum_k  f_k(i) - f_k(j) $<br><br>$w(T) = \sum_{i>j} \sum_k  f_k(i) - f_k(j) $ | $d(f_k(i), f_k(j)) = \sum_k  f_k(i) - f_k(j) $<br><br>$S = \sum_k \sum_{i>j}  f_k(i) - f_k(j) $ |

**Table 6.** Comparison between the MST and the MP techniques.

The two techniques are thus rather similar, especially for the l1 norm (Manhattan distance), the main differences being the supplementary internal nodes in the MP approach. This implies that MP is more general and can depict more complicated relationships.

### 3.6 Robustness Assessment

Basically, the assessment of the reliability of a tree is based on consensus trees. There is no rigorous mathematical tool, but rather some practical approaches:

- **Trials and errors:** Number of most parsimonious trees, consensus tree resolution, analyses with subsets of characters and objects.... and the interpretation.
- **Bootstrap:** Random draw of characters with replacement.
- **Decay or Bremer degeneracy index:** Number of supplementary steps necessary to have the node disappear.
- **Characters indices:** Retention Index (RI), Consistency Index (CI), Rescaled CI (RCI) measure how much each character supports the tree structure.

### 3.7 *Continuous Parameters*

#### 3.7.1 Discretization

In cladistics, we count the scores of trees using the evolutionary states of the characters. How can we deal with continuous parameters?

Unfortunately, no algorithm has been developed to perform Maximum Parsimony in this case. This is probably because the complexity of the search depends on the number of the character states which is essentially infinite in the continuous case. We must however mention the software program TNT (Goloboff et al., 2008) that allows for 65 000 states, but it assumes some evolutionary model for the character changes that may not be appropriate for astrophysical objects.

The only solution, often used in biology, and the one that has been used up to now in astrocladistics, is to discretize the continuous parameters. Yes but how? This is a long story and a very difficult question. The simplest way is to divide each character into equal bins. But should they be equal? How many bins? These two questions have no answer in the absolute, and the result depends more or less on the choices made at this level. Another typical question in astrophysics is whether to consider the variables themselves or their logarithms. Probably, the best approach is to try several choices and compare.

In astrocladistics so far, between 10 to 32 equal bins have been used to discretize the parameters. At least, the result should not depend on the precise number of bins and I have found that this is the case between 15 and 32.

Nevertheless, research is going on about this problem. One interesting direction is to optimize the number and the nature of the bins to ensure to get the best and the most robust tree as allowed by the data. For instance, we begin to understand the mathematical relationships between distance-based and character-based approaches: Thuillard & Fraix-Burnet (2015) have shown that the two approaches are identical if a multistate character can be reduced to a binary one according to a precise rule (the four-gamete rule).

#### 3.7.2 Continuous Parameter Evolution

Finding a tree is never guaranteed since it depends so much on the nature of the sample data and the parameters. Selecting parameters or discovering disturbing

objects requires many trials and errors computation. Graph theory may bring interesting help, especially in the case of continuous characters. For instance, when the phylogeny is perfect (in the sense that characters are all synapomorphies), some mathematical rules (the Kalmanson inequalities) holds. It is thus possible to search for the best order (arrangement on a tree) to be as close as possible to the fulfilment of the Kalmanson inequalities. This implies some constraints on the behaviour of the characters, with some convexity property and quite stringent parameter evolution along the tree (Thuillard & Fraix-Burnet, 2009).

We are considering here only mathematical criteria to select the parameters that could yield a robust phylogenetic tree. Physical considerations should be avoided as much as possible at this level since our astrophysical a priori conception would bias this choice.

### 3.8 Interpretation

Representating astrophysical objects on a tree is quite unusual, except for the evolution of Dark Matter Haloes represented on a graph known as a merger tree (Fig. 10a, from Stewart et al., 2008). In these cosmological trees, the first seeds of Dark Matter were very small haloes (small both in mass and dimension) that merged under the action of gravity to form bigger and bigger haloes. Hence, from the top of the figure, we obtain downward the genealogy of the bottom big halo.

However, there are many such big haloes, and many haloes with all kinds of sizes. The merger tree representation is unable to show them all, and even to show the evolution of Dark Matter haloes as a whole, as a population. With this kind of graph, how could you depict the evolution of the size of the haloes as a function of time or evolutionary stage of the Universe? This is impossible.

In Fig. 10b, the cladogram we would obtain with a sample of Dark Matter haloes using one parameter (mass or dimension) is shown. This is a phylogeny depicting how the species represented by each mass appeared in the course of the evolution of the Universe: most primitive haloes were small, the bigger ones occurred progressively by successive mergers of smaller ones.

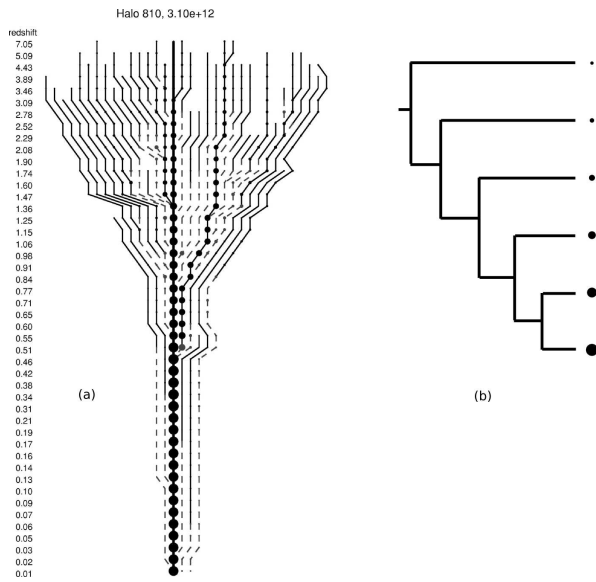
## 4 Clustering vs phylogenetic approaches

### 4.1 A simple example

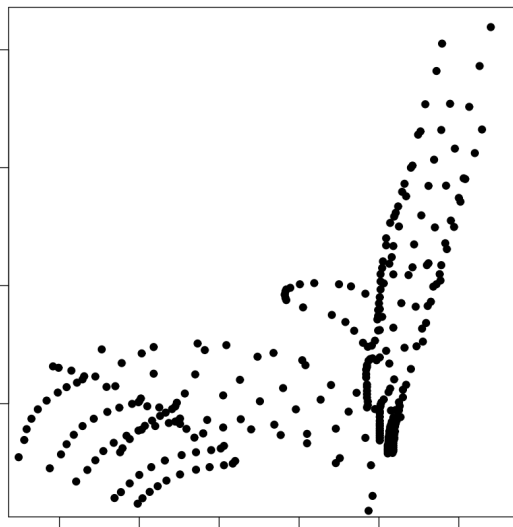
How many groups can you identify in the plot of Fig. 11?

I bet you have said two groups as depicted in Fig. 12, each one corresponding to different correlation properties.

The real answer is that there are six populations in this Hertzsprung-Russell (HR) diagram showing the luminosity  $\log L$  as a function of the surface temperature of the star  $\log T_{eff}$  (Fig. 13). The data come from the Geneva stellar evolutionary models (Schaller et al., 1992; Charbonnel et al., 1993; Schaerer et al., 1993a,b; Mowlavi et al., 1998) that compute the stellar parameters for stars of



**Fig. 10.** Two representation of the diversification of Dark Matter Haloes. (a) is the usual merger tree similar to a genealogy. (b) a cladogram made with one parameter, the size or the mass of the haloes, and represents a phylogeny.



**Fig. 11.** A bivariate diagram.

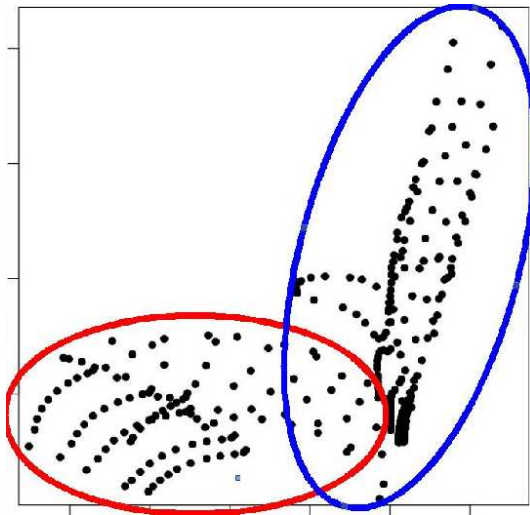


Fig. 12. A bivariate diagram where two groups are easily seen.

masses from  $M = 0.8$  to  $120$  Mo and metallicities  $Z$  from  $= 0.001$  to  $0.1$ . In the present sample, each of the six tracks chosen is represented by 51 time steps.

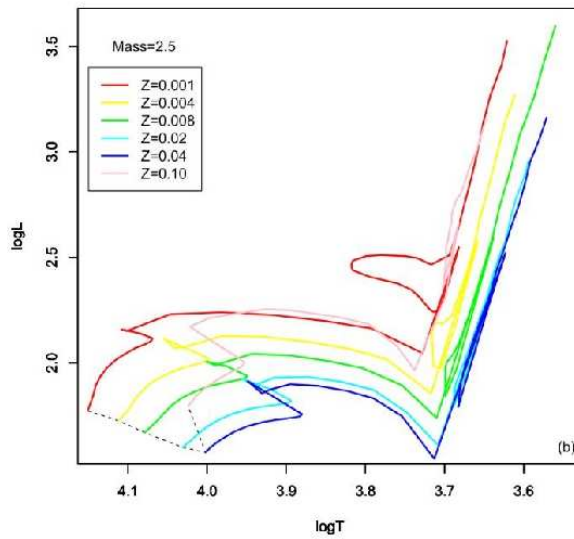


Fig. 13. Six stellar evolutionary tracks.

The fifteen parameters available are given in Table 7 (note that by definition  $X+Y+Z=1$ ). We will not use the age and the mass of the stars as these parameters are generally not known in practice. Only two ( $\log L$  and  $\log T_{eff}$ ) or thirteen (all except age and mass) parameters will be used in the following.

|    | Parameter | Description                          |
|----|-----------|--------------------------------------|
|    | Age       | Age                                  |
|    | Mass      | Actual mass                          |
| 1  | $\log L$  | $\log(\text{luminosity})$            |
| 2  | $\log T$  | $\log(\text{effective temperature})$ |
| 3  | Z         | Star metallicity                     |
| 4  | X         | H surface abundance                  |
| 5  | Y         | He surface abundance                 |
| 6  | C12       | $^{12}\text{C}$ surface abundance    |
| 7  | C13       | $^{13}\text{C}$ surface abundance    |
| 8  | N14       | $^{14}\text{N}$ surface abundance    |
| 9  | O16       | $^{16}\text{O}$ surface abundance    |
| 10 | O17       | $^{17}\text{O}$ surface abundance    |
| 11 | O18       | $^{18}\text{O}$ surface abundance    |
| 12 | Ne20      | $^{20}\text{Ne}$ surface abundance   |
| 13 | Ne22      | $^{22}\text{Ne}$ surface abundance   |

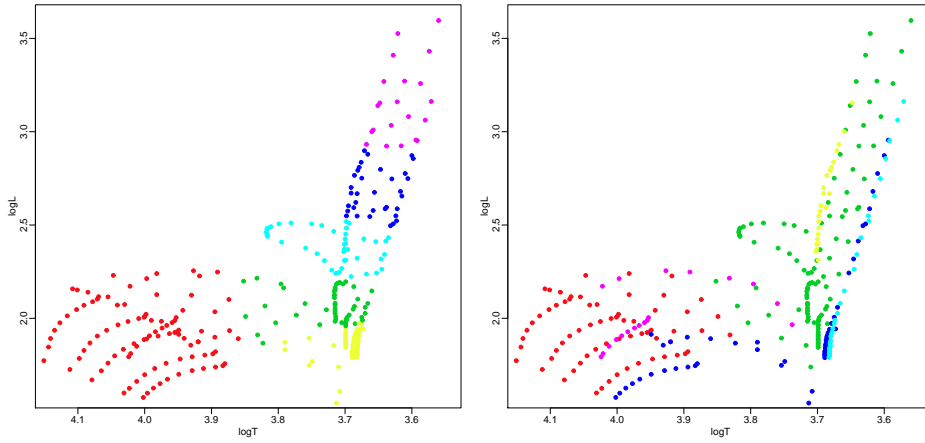
**Table 7.**

#### 4.1.1 Partitioning with six groups and phylogenetic analyses

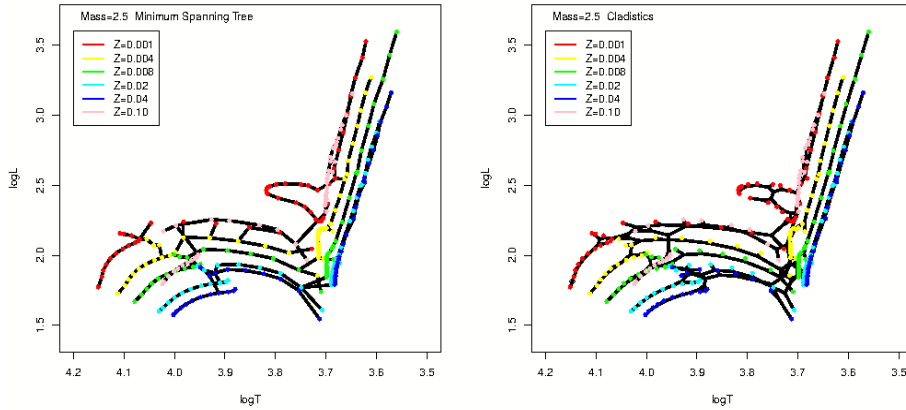
Knowing that we should find six groups, it is easy to perform a k-means analysis. The result is shown in Fig. 14.

With two parameters, it is clear that the partitioning approach looks for hyperspheres and cannot find elongated or one-dimensional structures. It is slightly better with thirteen parameters, but still far from satisfactory.

On the contrary, the phylogenetic methods, MST and MP, perform extremely well as can be seen in Fig. 15. However, this is still not perfect since the tracks are not connected on the lower left at the level of the Main Sequence of the stellar evolution. It is enlightning to investigate this point further as it shows that a clustering analysis always depend on the data and in particular on the available parameters.



**Fig. 14.** K-means analysis assuming six groups, with logL and logT only (left) or with the thirteen parameters (right).

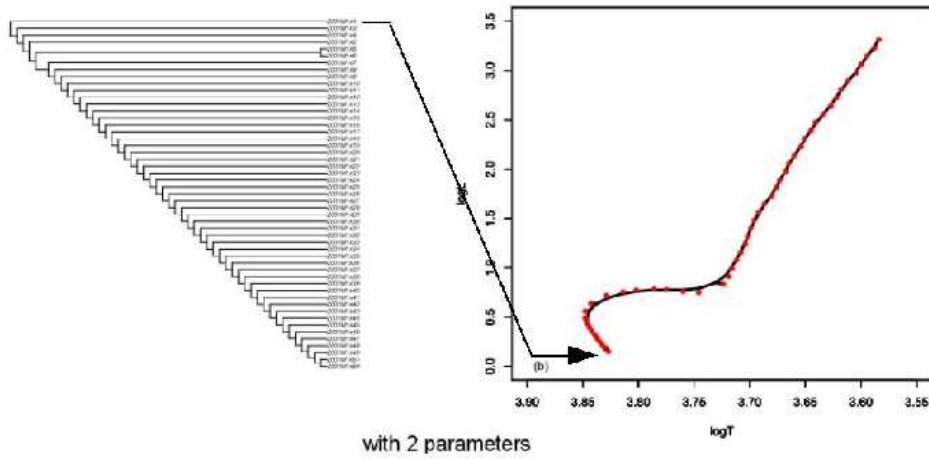


**Fig. 15.** Phylogenetic analyses of the stellar sample with thirteen parameters. Left: MST result. Right: MP result. These representations are projections of the trees on the two dimension diagram so that the lines are the branches of the trees.

#### 4.1.2 Role of parameter behaviour

Consider one evolutionary track ( $M=1 \text{ Mo}$ ,  $Z=0.001$ ). Only two parameters ( $\log L$  and  $\log T$ ) are enough here since the reconstruction is easy (Fig. 16).

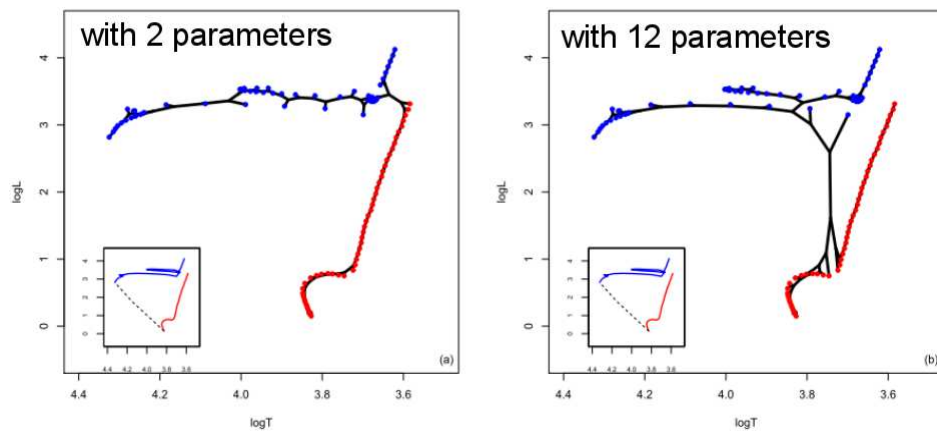
The tree is linear, and if rooted correctly, reproduces perfectly the chronology



**Fig. 16.** MP reconstruction of one stellar track with two parameters ( $\log L$  and  $\log T$ ). The tree on the left has been rooted with the star the closest to the Main Sequence corresponding to the initial step of the evolution.

of the stellar evolution. Note that this track is simple, with no loop, there is thus no reversal in any of the two parameters.

Consider now two tracks ( $M=1$  and  $5$  Mo,  $Z=0.001$ ) that are shown as they should be in the inlet of Fig. 17. Note that there are only twelve parameters since the metallicity  $Z$  is constant in the two tracks.



**Fig. 17.** MP reconstruction of two stellar tracks with two parameters ( $\log L$  and  $\log T$ ) and twelve ( $Z$  is the same for the two tracks).



There are several problems with the result. The first one is the connection between the two tracks that occurs at the end of them with two parameters, and in the middle with twelve parameters. This implies that the chronology is not respected at least on one of the branches of the tree. The second problem is the loop seen on the track on the top. This loop is totally missed with two parameters and is only sketched with twelve parameters.

What should we obtain? Ideally something that looks like the curves in the inlet with a connection at the main sequence (dotted line). Why this is not the case? Let us examine the evolution of the parameters along the track (Fig. 18).

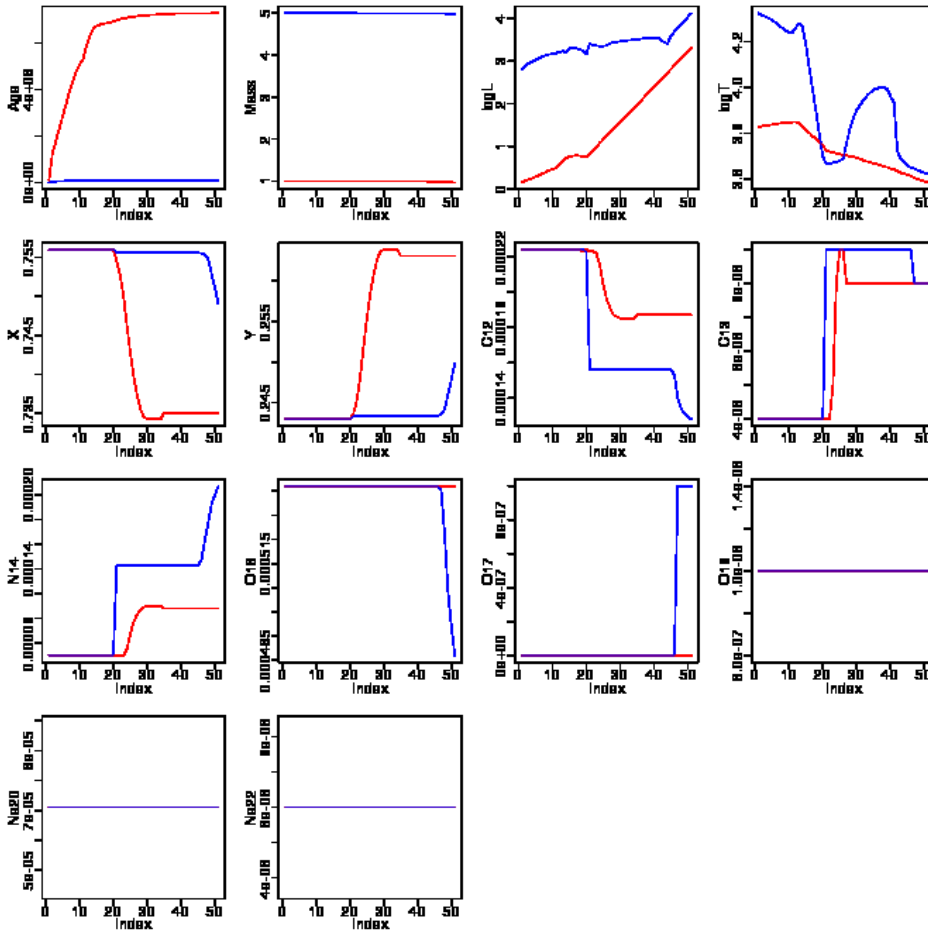


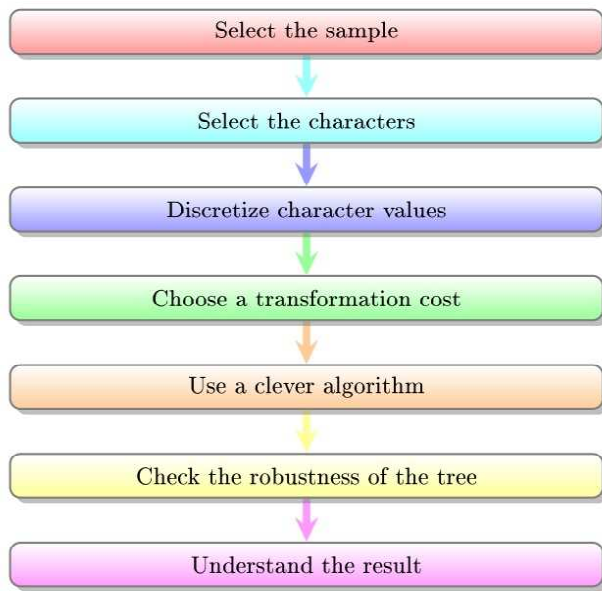
Fig. 18. Evolution of the thirteen parameters along the two tracks of Fig. 17. Note the strong reversal in  $\log T$  (upper right diagram) and the lack of evolution and often identical values of many parameters at the beginning of the tracks.

We can first notice one homoplasy, here a reversal, in logT. Then there are ranges (steps 1 to 20) where the tracks cannot be distinguished on most of the parameters. These behaviours explain why it is not possible to recover the perfect sequences, whatever the technique used.

## 5 Astrocladistics in Practice

### 5.1 Workflow

We have now seen all the aspects of a complete cladistic analysis. This is summarized in Fig. 19. It is important to recall that the selection of the sample and of the parameters must be made on statistical grounds, avoiding a priori and subjective choices that will heavily influence the results. For instance, if you choose only structural parameters, you will very certainly obtain a sequence according to what can be called the size.



**Fig. 19.** The workflow of a typical astrocladistic analysis.

### 5.2 Softwares

The last point to address is which software could we use to perform the tree search.

The bioinformaticians have worked for long on sophisticated algorithms to find the phylogenetic classification of living organisms. We can simply use them. All

softwares presented below are freely available and most are open source, except PAUP that is in the process of becoming so.

Undoubtly, the most useful software package to undertake a maximum parsimony search on astrophysical data is **PAUP\*** (Phylogenetic Analysis Using Parsimony<sup>1</sup>, Swofford, 2003). There are many useful tools to analyse, manipulate, and modify trees. But an important point I think is the possibility to use up to 32 or 64 bins for the characters.

A very convenient environment to prepare the matrix and analyse the results is the statistical package **R** which sees a recent and considerable development for phylogenetic studies<sup>2</sup>. Unfortunately, it still lacks the ability to perform the MP search itself, even though the package *phangorn* allows this with up to eight bins, which I do not find enough.

With these two softwares, you have all that is necessary to perform a complete cladistic analysis of astrophysical objects. There are other tools you may find interesting.

**Mesquite**<sup>3</sup> (Maddison & Maddison, 2004) is a rich java environment with very convenient graphical visualisation of trees. There are more and more modules allowing many kinds of computations, and it is linkable with PAUP and R to perform other analyses.

**MacClade** is as popular and powerful as PAUP but it is only available for MACs and seems not to work anymore on new versions of the OS. They advise to use Mesquite instead. **Phylip**<sup>4</sup> is also very powerful but devoted mainly to genetic data which have no equivalent in astrophysics. Finally, **TNT** (Tree analysis using New Technology<sup>5</sup> Goloboff et al., 2008) which is more efficient for big data sets (say 500) and also for continuous data that can be binned into 65 000 bins! However there seems to have some assumptions for the changes of character states which may not be adapted to astronomical processes.

### 5.3 *An application of cladistics in astrophysics: the Globular Clusters of our Galaxy*

#### 5.3.1 Context

Globular Clusters are compact and self-gravitating clouds of stars. For each cluster, and in a first approximation, stars are formed together at the same time in the same physical and chemical conditions. Globular Clusters exist around and in all galaxies, and they are certainly strongly linked to the transformation processes that build the observed galaxies. Hence finding the different populations of globular clusters for a particular galaxy yields information on its history.

---

<sup>1</sup><http://paup.csit.fsu.edu/>

<sup>2</sup><https://www.cran.r-project.org/web/views/Phylogenetics.html>

<sup>3</sup><http://mesquiteproject.wikispaces.com/>

<sup>4</sup><http://evolution.genetics.washington.edu/phylip.html>

<sup>5</sup><http://www.lillo.org.ar/phylogeny/tnt/>

For our own Galaxy, several studies have agreed upon two or three populations. The first natural discriminant parameter is the metallicity ( $[\text{Fe}/\text{H}]$ ) that measures the amount of “heavy” atomic elements (indeed all except hydrogen). But it appears that another parameter is needed to explain the diversity of the globular clusters around our Galaxy. This mysterious variable is called the second parameter.

Finding the populations is here a problem of clustering with two parameters, one of them being unknown and not necessarily observable. Such a multivariate problem is usually tackled in astrophysics by defining arbitrary cuts in a scatterplot, i.e. a 2D parameter space. For the particular case of the Globular Clusters of our Galaxy, one possible second parameter is the distance to the galactic centre. A very good illustration is given in Table 1.6 by Harris (2001). Subgroups are defined by first splitting the metallicity at -1 and then the distance to galactic centre at 4, 8, 12 and 20 kpc. They also can be defined by several splits in the metallicity (-1.85, -1.65, -1.50, -1.32). For each subgroup, the mean of the rotation velocity and its dispersion is computed to conclude on the dynamical differences between these subgroups. It is however hard to believe that hard frontiers in distance from the galactic centre between the globular cluster populations result from a long evolution with many transformation processes of our Galaxy.

In reality, this is not this distance that matters to explain the different properties of the populations, but the environment where they formed.

### 5.3.2 The Cladistics Analysis

If the populations of globular clusters were formed in different environments, they probably formed during different stages of evolution of the environment in our Galaxy at different stages of development. In other words, the environments in which the globular cluster populations form, were different and are related by “evolutionary” relationships due to the changes in metallicities and/or physical conditions. As a consequence, the globular cluster populations are also related by some “evolutionary” relationships.

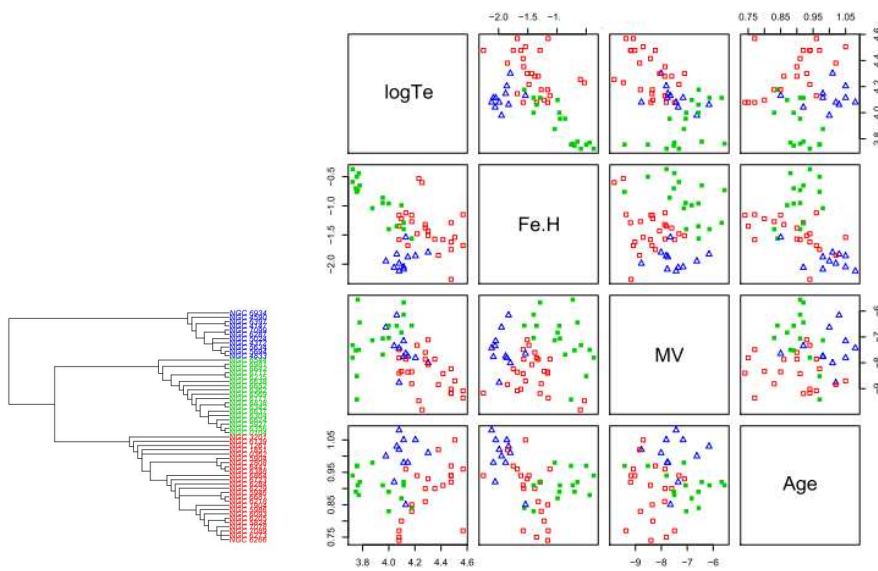
This problem has been reconsidered with cladistics by Fraix-Burnet et al. (2009). The data were retrieved from the literature on 54 globular clusters of our Galaxy, described by the three parameters selected from previous other studies :

- metallicity ( $[\text{Fe}/\text{H}]$ ),
- absolute V magnitude ( $M_V$ , more or less indicative of the total stellar mass of the cluster),
- maximum effective temperature ( $T_e$ ) on the horizontal branch.

Each parameter was discretized into 10 equal bins. The number of bins varied (from 3, 5, 8, 15 and 20), and an excellent agreement was found between the results except for 3 and 5 bins. The only reasonable evolutionary model here is the Wagner one (ordered states).

### 5.3.3 The tree

A very robust tree was found, with three obvious groups shown in Fig. 20. The tree is rooted with the less metallic objects (supposedly the oldest ones since the global metallicity in the Universe increases). The scatterplots in Fig. 20 shows the three parameters of the cladistic analysis plus the age that helps to date the groups and the transforming processes of the Galaxy during which the globular clusters formed. It is already easy to derive the different histories for the three groups.



**Fig. 20.** Left: the cladogram of the Globular Clusters with the three obvious groups. Right: Scatter plots involving the three parameters used in the analysis plus the age.

### 5.3.4 Interpretation

One can use any parameter available to interpret the result. For instance we find that the three groups have quite different average distances to the galactic centre, as well as different kinematics around our Galaxy. Note that the different distances to the galactic centre, not used in the cladistics analysis, is derived from our analysis, not hypothesized like in the work by Harris (2001) mentioned above.

Using all the information at our hand, we find that the three groups can satisfactorily be interpreted in the frame of our evolutionary hypothesis in the following way (given here for the simple picture of a monolithic assembly of our Galaxy):

- Group 2 (top of the tree, triangle symbols in Fig. 20) formed early during

a dissipationless phase of the assembly of our Galaxy, a gentle collapse of a huge cloud of gas,

- Group 1 (middle of the tree, filled squares) formed later in a dissipational phase, implying more turbulent and more metallic gas,
- Group 3 (bottom of the tree, open squares) formed in the disk of our Galaxy and rapidly in an intermediate epoch.

#### 5.4 To go further

Astrocladistics is the implementation of phylogenetic techniques in astrophysics. Three fundamental papers (Fraix-Burnet et al., 2006a,b,c) explain in detail why and how cladistics can be used in the case of galaxies, using samples taken from cosmological numerical simulations or closeby Dwarf galaxies.

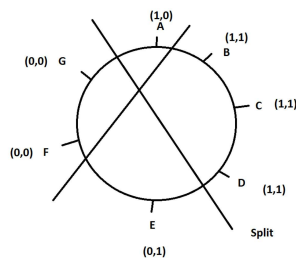
You can find other applications of cladistics on galaxies (Fraix-Burnet et al., 2010, 2012), globular clusters (Fraix-Burnet & Davoust, 2015) and Gamma Ray Bursts (Cardone & Fraix-Burnet, 2013).

## 6 Generalization: Networks

### 6.1 Split networks

Outer planar networks are generalizations of trees (Huson & Bryant, 2006) since they can represent simultaneously alternative trees with reticulations (generated for instance by hybridization). The link with phylogenetic trees is given by the definition of splits (see e.g. Fraix-Burnet et al., 2015).

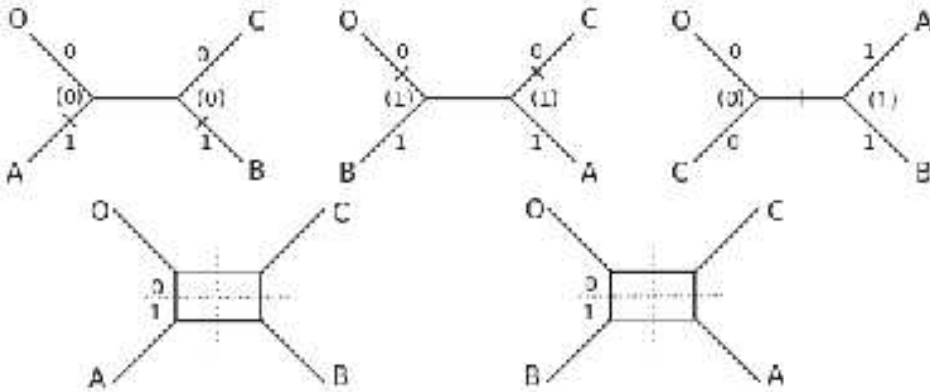
A split creates a partition of objects into two disjoint sets. Objects sharing a common property, as defined by splits, are consecutive in a circular order. This can be more easily visualized for binary characters (Fig. 21). Note that multistate characters can be transformed into binary characters.



**Fig. 21.** A circular order for objects A to G, with their pairs of binary states. The two lines show two different splits: (0,) vs (1,) and (,0) vs (,1).

In other words, splits in an outer planar network furnish neighboring relationships between objects. Figure 22 explains how several trees are represented by a

network and how splits separate the different trees. With four objects, there are three possible tree arrangements. Adding transversal branches allows to combine two of these trees, that is to merge two conflicting trees into a single diagram. A 3-dimensional representation could combine the three trees on a single scheme.



**Fig. 22.** Top: the three possible arrangements of four objects described by one binary character. Bottom: the two split networks combining two pairs of the three trees. The dotted lines indicate the split with the corresponding values of the character.

Networks can be very complicated (Fig. 23) and become rapidly quite difficult to read, and to interpret. However, this is probably the closest scheme to the true diversification scenario of living organisms, especially bacteria. It is not impossible that we also should use this kind of representation for galaxy diversification. You can find an example of its potential use in astrophysics in Thuillard & Fraix-Burnet (2009).

## 7 Conclusion

In a multivariate world, in the era of big data, we need adapted methodologies.

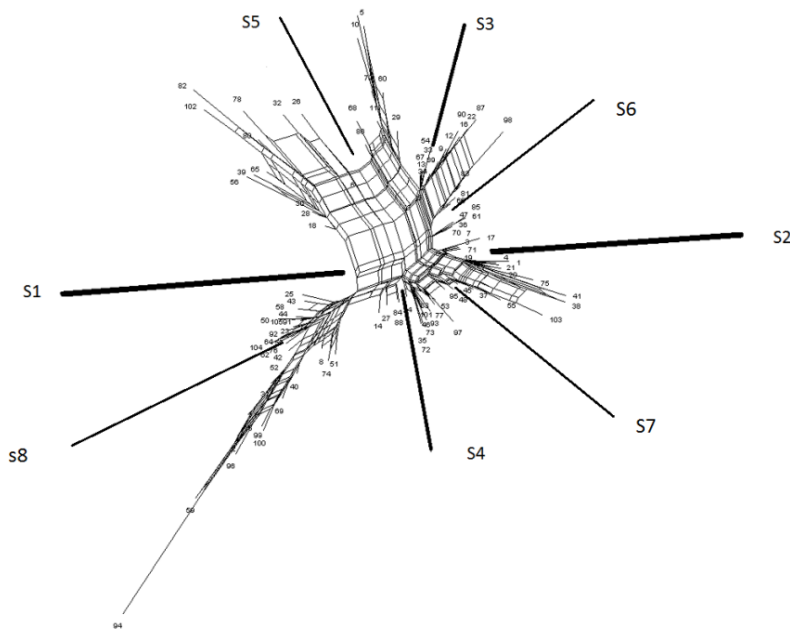
With the evolution of our knowledge and our observational/experimental technologies, we need to modify our tools for data analysis. In particular, our classifications must evolve accordingly.

Evolution, which is the continuous transformation of objects, is inherent to any astrophysical object and population.

Hence, the use of phylogenetic approaches is justified and must be pursued. This is the goal of astrocladistics<sup>6</sup> that started fifteen years ago.

Astronomers, as physicists, put physics above everything else. But before doing science with your data, don't you process them to correct for instrumental

<sup>6</sup><http://astrocladistics.org/>



**Fig. 23.** An example of an outer planar network showing the eight splits of the eight parameters  $s_1 \dots s_8$ .

distortions? Don't you transform them to get calibrated quantitative information? Don't you use signal processing techniques to diminish the noise, extract some hidden signal or translate it into something tractable? Doing advanced statistical analyses proceeds to the same kind of data management that aims at finding the information that can be used for the physical interpretation of the data.

In any statistical clustering study, be it phylogenetic or not, you need to be rigorous:

- Know your data (*statistics*)
- Select your parameters objectively (*statistics*)
- Compare several methods (*statistics*)
- Characterize your groups (*statistics*)
- And then, only then, interpret (*astrophysics*)

Yes astrophysics comes last, to interpret a statistical robust result, and to conclude whether the statistical analysis brings something useful. I like to compare the statistical tools as telescopes that help us explore an invisible world.



## Appendix

### A Exercises<sup>7</sup>

#### A.1 Presentation of the session

In this session you will compute and play with phylogenetic trees obtained by Maximum Parsimony (MP or cladistics), Minimum Spanning Tree (MST) and Neighbor Joining Tree Estimation (NJ).

The computation of Maximum Parsimony in R is not yet as powerful as it is in other dedicated softwares. However the package `phangorn`, mainly intended for genetic data, will be sufficient for this session.

#### A.2 Packages to install

You need to install these packages, either from the command line or from RStudio. If you know your repository, add the option: `repos="http://..."` .

```
install.packages("ape")      # the basis for most
                             # phylogenetic work
install.packages("phytools") # a new set of tools
install.packages("phangorn") # for computing Maximum
                             # Parsimony and other stuff
install.packages("igraph")  # the network analysis package
```

Then load the three first libraries:

```
library(ape)
library(phytools)
library(phangorn)
```

#### A.3 MP analysis

##### A.3.1 A simple example

The file `Data/phylo_data.txt` contains an artificial data set of 10 galaxies with 7 binary properties: “0” means either low or absent, “1” means high or present.

Load the matrix and transform it to the class `phyDat` for use in `phangorn`:

```
X <- read.table("./Data/phylo_data.txt", header = TRUE)
X
X <- as.phyDat(X, type = "USER", levels = c(0, 1))
X
```

<sup>7</sup>The corresponding R codes and the required data files can be found at <http://stat4astro2015.sciencesconf.org/>

First check all the possible trees with 4 and 5 objects:

```
trees <- allTrees(4) ; trees
op <- par(mfrow = c(1,3))
for(i in 1:length(trees)) plot(trees[[i]], type="unrooted",
                             cex=1.5) ; par(op)

trees <- allTrees(5) ; trees
op <- par(mfrow = c(3,5))
for(i in 1:length(trees)) plot(trees[[i]], type="unrooted",
                             cex=1.5) ; par(op)
```

How many trees are there for 6 objects? For 7?

We use the ratchet technique to compute the MP tree without exploring all possible arrangements. This is a very efficient method that avoids as much as possible to be trapped in a local minimum. Do not worry about the parameters of the command, it is not important for us here.

```
tree <- pratchet(X, method="sankoff", all=TRUE, maxit=100000,
                k=100)
op <- par(mfrow=c(2,2))
plotTree(tree) ; par(ask=F)
par(op)
```

Why are there several trees?

There are two ways to summarize the multiple trees. The first one is to build a consensus tree. It can be a strict consensus (nodes occurring in 100% of the trees) or a majority rule (always higher than 50%, 70% is a good compromise).

```
tcons <- consensus(tree, p=0.7)
plotTree(tcons)
tcons <- root(tcons, outgroup = "GAL1", resolve.root = TRUE)
plotTree(tcons)
```

Try to root with another galaxy.

The second way to summarize the multiple trees is to compute a split network.

```
tnet <- consensusNet(tree, prob=0.3)
plot(tnet, "2D")
plot(tnet, "2D", show.edge.label=TRUE)
plot(tnet) # in 3D you need the package rgl
```

To better understand the split network, compare with the consensus.

```
op <- par(mfrow=c(1,2))
plotTree(tcons)
plot(tnet, "2D")
par(op)
```

## A.3.2 Local dwarf galaxies

Load the data in the file dwarfs.txt and select only 14 galaxies for which we know a robust tree exists.

```
dwarfs <- read.delim("Data/dwarfs.txt", na.strings="?")
dwarfs14 <- dwarfs[c(10,7,35,33,15,9,11,16,30,1,25,5,4,24),]
str(dwarfs14)
```

To code this matrix, I provide you with a script that you have to source. We choose 8 bins since it is the maximum allowed in phangorn.

```
source("Data/CodeMatrix.R")
dwarfs14cod <- CodeMatrix(dwarfs14, bins=8)
Y <- as.phyDat(data.frame(t(dwarfs14cod)), type = "USER",
                 levels = c(0, 1,2,3,4,5,6,7,"?"))
```

Now perform the Maximum Parsimony Analysis as previously. The command to root the tree looks odd (write and read), this is to ensure that the order of the names corresponds to the tree order. This is not absolutely necessary but avoids some confusion in some plotting. The ladderize function organizes the tree so that it is easier to compare several trees.

```
treeMP <- pratchet(Y, method="sankoff", all=TRUE,
                  maxit=100000, k=100) ; treeMP
plotTree(treeMP) ; par(ask=F)
treeMP <- read.tree(text=write.tree(ladderize(root(treeMP,
                                                outgroup="SagDIG"))))
plotTree(treeMP) ; par(ask=F)
```

Generally, distance-based approaches do not accept NA values. In order to compare to MST and NJ, we now work with a reduced matrix where all parameters having NA are removed. How many parameters are left?

```
dwarfs14red <- dwarfs14[,-c(unique(which(is.na(dwarfs14),
                                       arr.ind=T)[,2]))]
dwarfs14redcod <- CodeMatrix(dwarfs14red, bins=8)
Z <- as.phyDat(data.frame(t(dwarfs14redcod)), type = "USER",
               levels = c(0, 1,2,3,4,5,6,7,"?"))
treeMPred <- pratchet(Z, method="sankoff", all=TRUE,
                     maxit=100000, k=100)
treeMPred <- read.tree(text=write.tree(ladderize(root(
  treeMPred, outgroup="SagDIG"))))
plot(treeMPred)
```

Plot side by side the two trees treeMP and treeMPred. Do you see a difference? How can you explain it?

#### A.4 Minimum Spanning Tree

We first need to compute the pairwise distance matrix, and then compute the MST. We can use the `mst` command of the `ape` package.

```
dis <- dist(scale(dwarfs14red), method="manhattan")

# ape
treemst <- mst(dis)
plot(treemst)
plot(treemst, graph="nsca")
PC <- prcomp(scale(dis)) # PCA analysis
plot.mst(treemst, x1 = PC$x[, 1], x2 = PC$x[, 2])
# in the V vs M diagram:
plot.mst(treemst, x1=dwarfs14red[, 3], x2=dwarfs14red[, 4])
```

The `igraph` package is more powerful for graphical representation of networks. We import the distance matrix into `igraph` and compute the MST. We can get informations on the list of vertices (nodes), edges (branches), their weights (distances) and the degrees of the vertices.

```
library(igraph)
net <- graph.adjacency(as.matrix(dis), mode="undirected",
                      weighted=TRUE, diag=F)

g <- minimum.spanning.tree(net) # or mst(net)
V(g) # gives the list of vertices
E(g) # gives the list of edges
E(g)$weight
degree(g) # gives the number of edges per vertex
```

There are many possibilities to plot the tree. In particular there are several “layout” that optimize the visibility of all vertices and edges on the graph.

```
plot(g, layout=layout_with_fr, vertex.size=4, vertex.
      label.dist=0.5, vertex.color="red",
      edge.width=E(g)$weight,
      edge.label=round(E(g)$weight, 2))

# try other layouts :
# layout_as_tree, layout_auto, layout_with_kk
plot(g, layout=layout_mds(g, dist=as.matrix(dis)),
      vertex.size=4, vertex.label.dist=0.5,
      vertex.color="red", edge.width=E(g)$weight,
      edge.label=round(E(g)$weight, 2))
```

We can also separate the MST by cutting the edges whose length is larger than say 5.

```
gp <- g-E(g)[which(E(g)$weight >=5)]
plot(gp, layout=layout_mds(gp, dist=as.matrix(dis)),
      vertex.size=4, vertex.label.dist=0.5,
```

```
vertex.color="red",edge.width=E(gp)$weight,
edge.label=round(E(gp)$weight,2))
```

We have finished with igraph, it is good to unload it:

```
detach("package:igraph", unload=TRUE)
```

### A.5 Neighbor Joining Tree Estimation

We use the same pairwise distance matrix as for the MST computation.

```
treenj <- nj(dis)
treenj <- read.tree(text=write.tree(ladderize(root(treenj,
outgroup="SagDIG"))))
plot(treenj, use.edge.length=FALSE)
plot(treenj, type="unrooted")
```

### A.6 Comparison of the three results

Analyse the differences between the trees of dwarf galaxies obtained with the three techniques.

```
op <- par(mfrow=c(1,3))
plot(treeMPred, main="MP", cex=1.3)
plot(g, layout=layout_as_tree, vertex.size=4,
vertex.label.dist=0.5, main="MST",
vertex.label.cex=1.5)
plot(treenj, use.edge.length=FALSE, main="NJ", cex=1.3)
par(op)
```

### A.7 Interpretation of the tree

There are many ways to project parameter values on the tree, or the tree on diagrams. Here is only two examples using the phytools package. You can now try to identify and characterize some groups from the tree. Try to plot the nine parameters using a loop and the function below for the trees treeMP, treeMPred and treenj.

```
interpret <- function(param, tree) {
  layout(matrix(c(1,2),1,2),c(0.7,0.3))
  plot(tree, mar=c(4.1,1.1,1.1,0), use.edge.length=FALSE)
  par(mar=c(4.1,0,1.1,1.1))
  barplot(dwarfs14red[tree$tip.label, param],
          horiz=TRUE, width=1, space=0,
          ylim=c(1, length(tree$tip.label)) - 0.5, names="")
  title(colnames(dwarfs14red)[param])
}
```

There is also a nice way to visualize the evolution of the parameters along the tree.

```
evol <- function(param, tree) {
  v <- dwarfs14red[[param]]
  names(v) <- rownames(dwarfs14red)
  tree$edge.length <- rep(1, length(tree$edge[,1]))
  contMap(tree, v)
  title(colnames(dwarfs14red)[param])
}
```

You can redo the exercises with the full data set in dwarfs.txt (36 galaxies).

## References

- Adanson M. 1763. *Famille Des Plantes*, chez Vincent, impr.-libraire de Mgr le Comte de Provence (Paris), num. BNF de l'éd. de Paris : INALF, 1961 (Fran-text ; R263R)
- Barrow J. D., Bhavsar S. P. & Sonoda D. H. 1985. Minimal spanning trees, filaments and galaxy clustering, *Monthly Notices of the Royal Astronomical Society* 216, 17
- Cardone V. F. & Fraix-Burnet D. 2013. Hints for families of GRBs improving the hubble diagram, *Monthly Notices of the Royal Astronomical Society* 434, 1930
- Charbonnel C., Meynet G., Maeder A., Schaller G. & Schaerer D. 1993. Grids of stellar models - part three - from 0.8 to 120-solar-masses at z=0.004, *Astronomy & Astrophysics Supplement Series* 101, 415
- Feigelson E. & Babu G. 2012. *Modern Statistical Methods for Astronomy: With R Applications*. Cambridge University Press
- Felsenstein J. 1984. *Cladistics: Perspectives on the reconstruction of evolutionary history*, Duncan T., Stuessy T., eds., Columbia University Press, New York, pp. 169–191
- Fraix-Burnet D., Chattopadhyay T., Chattopadhyay A. K., Davoust E. & Thuillard M. 2012. A six-parameter space to describe galaxy diversification, *Astronomy & Astrophysics* 545, A80
- Fraix-Burnet D., Choler P. & Douzery E. 2006a. Towards a Phylogenetic Analysis of Galaxy Evolution : a Case Study with the Dwarf Galaxies of the Local Group, *Astronomy & Astrophysics* 455, 845
- Fraix-Burnet D., Choler P., Douzery E. & Verhamme A. 2006b. Astrocladistics: a phylogenetic analysis of galaxy evolution I. Character evolutions and galaxy histories, *Journal of Classification*, 23, 31

- Fraix-Burnet D. & Davoust E. 2015. Stellar populations in  $\omega$  centauri: a multivariate analysis, *Monthly Notices of the Royal Astronomical Society* 450, 3431
- Fraix-Burnet D., Davoust E. & Charbonnel C. 2009. The environment of formation as a second parameter for globular cluster classification, *Monthly Notices of the Royal Astronomical Society* 398, 1706
- Fraix-Burnet D., Douzery E., Choler P. & Verhamme A. 2006c. Astrocladistics: a phylogenetic analysis of galaxy evolution II. Formation and diversification of galaxies, *Journal of Classification*, 23, 57
- Fraix-Burnet D., Dugué M., Chattopadhyay T., Chattopadhyay A. K. & Davoust E. 2010. Structures in the fundamental plane of early-type galaxies, *Monthly Notices of the Royal Astronomical Society* 407, 2207
- Fraix-Burnet D., Thuillard M. & Chattopadhyay A. K. 2015. Multivariate approaches to classification in extragalactic astronomy, *Frontiers in Astronomy and Space Sciences*, 2, 3
- Gascuel O. & Steel M. 2006. Neighbor-joining revealed, *Molecular Biology and Evolution*, 23, 1997
- Goloboff P. A., Farris J. S. & Nixon K. C. 2008. TNT, a free program for phylogenetic analysis, *Cladistics*, 24.5, 774
- Gower J. C. & Ross G. J. S. 1969. Minimum spanning trees and single linkage cluster analysis, *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 18, pp. 54
- Harris W. E. 2001. Globular Cluster Systems, Saas-Fee Advanced Course, Vol. 28, *Star Clusters*, Labhardt L., Binggeli B., eds., Springer-Verlag Berlin Heidelberg, p. 223
- Hennig W. 1965, Phylogenetic systematics, *Annual Review of Entomology*, 10, 97
- Huson D. H. & Bryant D. 2006, Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution*, 23, 254
- Maddison W. P. & Maddison D. R. 2004, Mesquite: a modular system for evolutionary analysis (<http://mesquiteproject.org>)
- Mowlavi N., Schaerer D., Meynet G., Bernasconi P. A., Charbonnel C. & Maeder A. 1998, Grids of stellar models. vii. from 0.8 to 60  $M_{\odot}$  at  $z = 0.10$ , *Astronomy & Astrophysics Supplement Series* 128, 471
- Saitou N. & Nei M. 1987, The neighbor-joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, 4, 406
- Schaerer D., Charbonnel C., Meynet G., Maeder A. & Schaller G. 1993a, Grids of stellar models - part four - from 0.8-solar-mass to 120-solar-masses at  $z=0.040$ , *Astronomy & Astrophysics Supplement Series* 102, 339

- Schaerer D., Meynet G., Maeder A. & Schaller G. 1993b, Grids of stellar models. ii - from 0.8 to 120 solar masses at  $z = 0.008$ , *Astronomy & Astrophysics Supplement Series* 98, 523
- Schaller G., Schaerer D., Meynet G. & Maeder A. 1992, New grids of stellar models from 0.8 to 120 solar masses at  $z = 0.020$  and  $z = 0.001$ , *Astronomy & Astrophysics Supplement Series* 96, 269
- Stewart K. R., Bullock J. S., Wechsler R. H., Maller A. H. & Zentner A. R. 2008, Merger histories of galaxy halos and implications for disk survival, *The Astrophysical Journal* 683, 597
- Swofford D. L. 2003, Paup\*: Phylogenetic analysis using parsimony (\*and other methods), nauer Associates, Sunderland, Massachusetts, <http://paup.csit.fsu.edu/>
- Thuillard M. & Fraix-Burnet D. 2009, Phylogenetic Applications of the Minimum Contradiction Approach on Continuous Characters, *Evolutionary Bioinformatics*, 5, 33
- Thuillard M. & Fraix-Burnet D. 2015, Phylogenetic trees and networks reduce to phylogenies on binary states: Does it furnish an explanation to the robustness of phylogenetic trees against lateral transfers?, *Evolutionary Bioinformatics*, 11, 213
- Williams T. L. & Moret B. M. 2003, An investigation of phylogenetic likelihood methods, in *Third IEEE Symposium on Bioinformatics and Bioengineering*, pp. 79–86