Clustering and Phylogenetic Approaches to Classification: Illustration on Stellar Tracks

Didier Fraix-Burnet, Marc Thuillard

Clustering and Phylogenetic Approaches to Classification: Illustration on Stellar Tracks

D. Fraix-Burnet[1], M. Thuillard[2]

[1] Univ. Grenoble Alpes, CNRS, IPAG, 38000 Grenoble, France
email: didier.fraix-burnet@univ-grenoble-alpes.fr
[2] La Colline, 2072 St-Blaise, Switzerland

February 7, 2018

This pedagogical article was written in 2014 and is yet unpublished. Part of it can be found in Fraix-Burnet (2015).

Abstract

Classifying objects into groups is a natural activity which is most often a prerequisite before any physical analysis of the data. Clustering and phylogenetic approaches are two different and complementary ways in this purpose: the first one relies on similarities and the second one on relationships.

In this paper, we describe very simply these approaches and show how phylogenetic techniques can be used in astrophysics by using a toy example based on a sample of stars obtained from models of stellar evolution.

We first perform some cladistic analyses to understand how the evolutionary behaviours of the parameters may affect the clustering process. We then show the mathematical principles that connect four different algorithms: one partitioning method, k-medoids, and three phylogenetic methods, Minimum Spanning Tree, Neighbor Joining Tree Estimation and Maximum Parsimony (cladistics). We finally use a challenging sample of stars to assess their performances.

The phylogenetic methods naturally perform better than the partitioning method to retrieve the stellar lineages, and Maximum Parsimony, being more general, surpasses all approaches.

Keywords: Methods: statistical – Stars: evolution – Hertzsprung-Russell and C-M diagrams

1 Introduction

Clustering objects into synthetic groups is a natural activity of any science. Astrophysics is not an exception but has always adopted specific strategies to classify a relatively modest amount of diversity and has much counted on the physics to define the discriminant parameters. This discipline is now facing the need for sophisticated statistical tools to tackle the astronomical number of observed and catalogued objects and the increasing number of observed properties that describe them.

Clustering approaches gathers objects according to their similarities through the choice of a distance metrics. There is a huge class of techniques that partition the data into a pre-defined number of groups. A well-known algorithm is the k-means (MacQueen, 1967; Ghosh and Liu, 2010). Obviously, these kind

of approaches are suited for distinct and more or less spherical groups but not to find very elongated structures like filaments or lineages. In addition, the discriminant usefulness of distances is lost in high dimension parameter spaces since distances tend to become similar (one of the aspects of the "curse of dimensionality").

Another family of clustering techniques uses a hierarchical representation of the pairwise distances, through a bottom-up algorithm that constructs a tree by relating the closest objects together before relating these first clustering to closest clusters or objects, and so forth until the whole sample is exhausted. The final number of groups is then chosen by cutting the tree at a fixed distance level. The branches of the tree, called a dendrogram, may or may not represent relationships between the objects.

Originally, phylogenetic methods are designed to build a graph representing the evolutionary relationships between species (see reviews in Felsenstein, 2003; Makarenkov et al., 2006). Each node of the graph indicates a transmission with modification mechanism that creates two or more species inheriting from a common ancestor.

More generally, a phylogenetic approach can be viewed as a clustering approach in which relationships are provided. As a consequence, phylogenetic techniques are particularly versatile and powerful methods for building classification trees. They can be understood in the frame of the graph theory (Semple and Steel, 2003).

There are two kinds of phylogenetic methods, using either the parameters describing the objects or pairwise distances (or dissimilarities) computed from these parameters. The parameter-based methods evaluate all possible trees that can be constructed with the objects, and select the tree(s) corresponding to an optimising criterion. The process is thus based on the distribution of the parameter values. The distance-based methods build the tree entirely from the distances, putting forward the global similarities between the objects. Parameter-based methods can describe a larger variety of evolutionary scenarios and are thus more general that the distance-based methods. But this is at the cost of a larger computation time which quickly becomes prohibitive. Mathematically, formal connections between parameter- and distance-based methods are developed in the case of continuous parameters (e.g. Thuillard and Fraix-Burnet, 2009; Thuillard and Fraix-Burnet, 2015), explaining why both kinds of methods are successfully used in phylogenetic studies.

Among distance-based treelike techniques, the friend-to-friend algorithm is relatively famous in astrophysics (e.g. More et al., 2011, and references therein). Also known as the single linkage or nearest neighbor algorithm, it is mathematically related to the Minimum Spanning Tree technique which looks for the simplest graph connecting the objects under study (Gower and Ross, 1969; Feigelson and Babu, 2012). A more sophisticated approach used in phylogenetic studies is the Neighbor-Joining Tree technique (Saitou and Nei, 1987; Gascuel and Steel, 2006). Many other algorithms exist but we consider only these two that illustrate the general methodologies.

Among the parameter-based techniques, cladistics (or Maximum Parsimony) is the most famous one. Invented in the 1950's by William Hennig (Hennig, 1965), its principle looks simple: two (or more) objects are related if they share a common history, that is they possess properties inherited from a common ancestor. In practice, a cladistic analysis asks for the objects under study to be described by evolutionary characters (parameters or descriptors) for which at least two states are defined: one is said to be ancestral, the other one is said to be derived. The derived state corresponds to an innovation in the evolution and is assumed to have been acquired by an unidentified ancestor. This is the transmission phase of inheritance making descendants look similar to their parents. The accidents in this process are called modifications and generate diversity. This transmission with modification process was invoked by Darwin to explain the observed hierarchical organisation of the biological diversity. But in essence,

any entity, be it biological or not, evolving with a transmission with modification process can be a priori studied by Maximum Parsimony, provided evolutionary states can be defined for the characters.

In this paper, we show how these approaches can be applied in astrophysics. Using some very simple examples from the stellar evolution models, we explain and compare the four methods indicated above (k-means or k-medoids which is more robust to outliers, Minimum Spanning Tree, Neighbor-Joining Tree, and Maximum Parsimony or cladistics).

Most of the diversity in the Universe is due to evolution. Astronomical objects, such as stars, globular clusters or galaxies, evolve continuously according to physical laws governing their internal physics and chemistry, and also through interactions with their environment. Evolution in astrophysics is a continuous transformation, even during violent events. At each stage, this continuous transformation can be seen as a transmission with modification mechanism that generates lineages and branching diversification. This idea was the base for the development of astrocladistics (Fraix-Burnet et al., 2006a,b,c) which has been successfully used in the case of galaxies, globular clusters and Gamma-Ray Bursts (e.g. Fraix-Burnet et al., 2009, 2010, 2012; Cardone and Fraix-Burnet, 2013).

At first glance, stellar evolution does not appear to be a good case for a phylogenetic study. On one hand, the evolution of a star depends on only two parameters (mass and metallicity), so that a lineage can be easily defined as the evolutionary path of stars having initially the same two parameters. But on the other hand, the branching pattern is not obvious since there is no interaction between stars. However, i) there is an indirect physical relationship between the lineages through the explosions of the most massive stars the gas of which forms the new stars (transmission with modification), and ii) the link could be merely conceived mathematically as a change of the parameters in the continuum of possible values. In this paper, we do not address the question i) and we simply assume point ii) for a formal exercise of classifying stellar tracks.

The paper is organised as follows. The data from stellar evolutionary models and general ideas on the evolution of stars are presented in Sect. 2. The four methods used in this paper are described in Sect. 3. Then, we use the most powerful of these methods, Maximum Parsimony, on several samples to understand the roles and behaviours of the parameters (Sect. 4). This step is a prerequisite in any clustering analysis, since the parameters explain many features found in the results. In the next section, we choose a sample as a challenge to get the correct clustering according to the known lineages and compare the four clustering and phylogenetic approaches (Sect. 5). Some applications are briefly commented in Sect. 6. We finally explain in Sect. 7 the mathematical interconnections between the methods, explaining similarities and differences, and propose some possible strategies to improve the clustering analyses in the case of real data sets. The conclusions are given in Sect. 8.

2 Data and elements of stellar evolution

The data are taken from grids of stellar models, from 0.8 to 120 solar masses ($M_{\odot}$) at metallicities Z=0.001 (Schaller et al., 1992), 0.004 (Charbonnel et al., 1993), 0.008 (Schaerer et al., 1993b), 0.02 (Schaller et al., 1992), 0.04 (Schaerer et al., 1993a) and 0.10 (Mowlavi et al., 1998), available on VizieR of the CDS[1]. The 15 parameters given by the grid models are described in Table 1. In practice, the age cannot be estimated from the observations except when the distance to the star is known and if it is a member of a star cluster. The mass can be measured directly only in binary systems. The (absolute) luminosity requires the distance to be known, which is generally not the case for an isolated object. The temperature is derived from the color of the stars. The remaining eleventh quantities are element

[1] http://vizier.u-strasbg.fr/viz-bin/VizieR

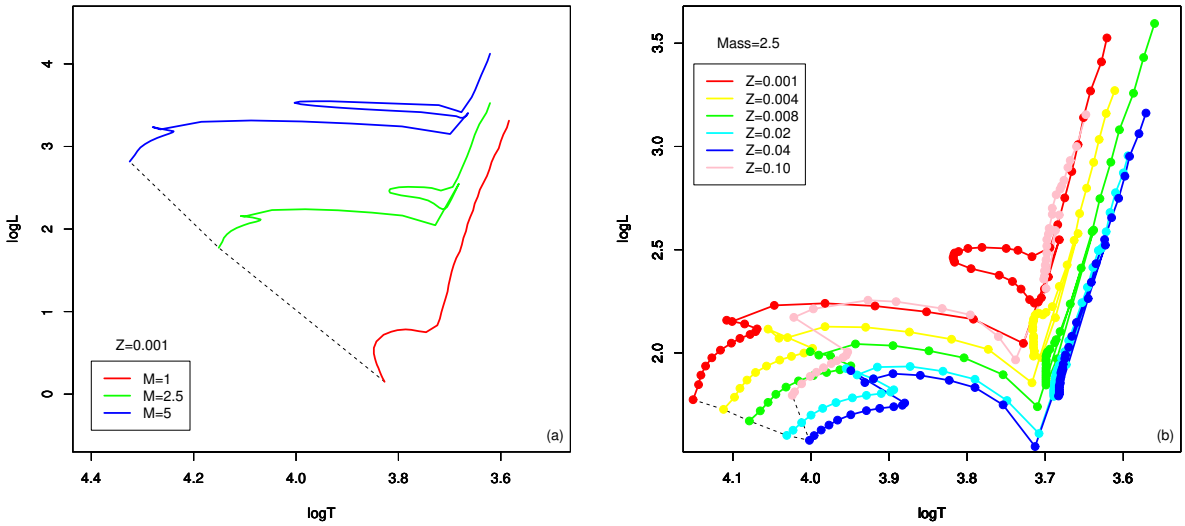

Figure 1: Evolutionary paths of stars in the HR diagram from the stellar evolutionary models, for a single metallicity Z=0.001 (a) and for a single mass M=2.5 $M_\odot$ (b). The dashed line is the Main Sequence (time zero of the models). Note the different scales of the two figures and the reversed $\log T$ axis.

abundances but are quite difficult to measure from the spectrum. In this work, we leave aside the observational question of whether the parameters are/can be measured or not, and rather concentrate on their roles regarding the clustering process and evolution. We always exclude the age from the analysis since it would be of too much help to retrieve the correct chronologies. We thus perform analyses either with two parameters ($\log T, \log L$), or with 12 to 14 parameters depending on whether the mass or metallicity is constant, and whether we include the mass in the analysis.

The diversity and evolution of stars are most generally represented in the HertzsprungRussell diagram (hereafter HR diagram) which is a scatter plot between the luminosity ($\log L$) versus their effective surface temperature ($\log T$), both in logarithm scales (Fig. 1). $\log T$ corresponds closely to the spectral types that are the basis of the classification of stars. After an initial formation phase which is quite complicated to follow on the HR diagram, and is relatively short, stars reach the so-called Main Sequence where they remain most of their life at a position that is imposed by their mass and metallicity. The Main Sequence is considered as the starting point in the grid models used here. From this point, the evolution of stars is quite well understood up to a certain limit. Both the evolutionary details and this limit depend mainly on the initial mass and metallicity. It is thus possible to trace the evolutionary tracks of stars in the HR diagram as shown on Fig. 1 which plots some examples from the grid models.

All stars with the same initial mass and metallicity thus define a lineage. The time line along the lineage varies greatly with the mass. The data of the stellar evolutionary models that we use have 51 points (time steps) for each lineage. The ending point is determined by the limit of validity of the model to describe the physics of the stars.

Since the mass has a much larger effect on the position of the track in the diagram than the metallicity (see Fig. 1), it is easier to study details of a cladistic analysis with a sample of stars with same metallicity (Sect. 4). Afterwards, we can propose a clustering challenge with a sample of stars having the same mass and compare several approaches (Sect. 5),

Table 1: The 15 parameters given by the grid models. Note that by definition: $X + Y + Z = 1$.

Parameter	Unit	Description
Age	yr	Age
Mass	$M_\odot$	Actual mass in solar masses
$\log L$	$L_\odot$	log(luminosity) in solar units
$\log T$	K	log(effective temperature)
Z		Star metallicity (mass fraction)
X		H surface abundance (mass fraction)
Y		He surface abundance (mass fraction)
C12		^{12}C surface abundance (mass fraction)
C13		^{13}C surface abundance (mass fraction)
N14		^{14}N surface abundance (mass fraction)
O16		^{16}O surface abundance (mass fraction)
O17		^{17}O surface abundance (mass fraction)
O18		^{18}O surface abundance (mass fraction)
Ne20		^{20}Ne surface abundance (mass fraction)
Ne22		^{22}Ne surface abundance (mass fraction)

3 Methods

In this section, we present the four clustering techniques that will be used in Sect. 5 and discussed in Sect. 7. Maximum Parsimony will also be used in Sect. 4.

3.1 Multivariate clustering: k-medoids

The k-means algorithm (MacQueen, 1967; Ghosh and Liu, 2010) is a partitioning approach that starts with k centroids, k corresponding to the number of clusters given a priori. It then assigns each data point to the closest centroid and when the clusters are built, the new k centroids are computed and the process iterates until convergence. The result depends very much on the initial centroids. Repeating the analysis with several initial choices is always a good idea, but consistency is not guaranteed if the data do not contain distinguishable and roughly spherical clusters. Some strategies have been devised to guess the best initial choice for the centroids (e.g. Sugar and James, 2003; Tajunisha and Saravanan, 2010).

The k-means algorithm has already been used in the context of stars (e.g. Gratton et al., 2011; Simpson et al., 2012), galaxies (e.g. Fraix-Burnet et al., 2010; Sánchez Almeida et al., 2010; Fraix-Burnet et al., 2012) or Gamma-Ray Bursts (Chattopadhyay et al., 2007).

In the present paper, we use the k-medoids algorithm (Kaufman and Rousseeuw, 1987; Reynolds et al., 2006), generally implemented through the Partitioning Around Medoids algorithm (PAM, Kaufman and Rousseeuw, 1990, used here through the command *pam* of the *cluster* library in the R-package). It chooses data points as centers (medoids) and is known to be more robust to noise and outliers.

These two partitioning approaches gather objects according to a distance matrix that is computed pairwise from the parameter matrix. They are thus well adapted to hyperspheres in the parameter space. In this paper, we have used both the euclidean distance (L2 norm) or the Manhattan distance (L1 norm) without noticing any difference in the results.

3.2 Minimum Spanning Tree

The Minimum Spanning Tree (MST) is mathematically related to the single linkage clustering, known to astronomers as the friends-of-friends algorithm or nearest-neighbor algorithm (Gower and Ross, 1969; Feigelson and Babu, 2012). A spanning tree is an acyclic, connected graph $G = (V, E)$ together with a function $w : E \to \mathbb{R}$ that assigns a weight $w(e)$ to each edge e in E. The minimum spanning tree of G, is the spanning tree T minimising the function :

$$w(T) = \sum_{e \in T} w(e) \tag{1}$$

If the weights $w(e)$ are distinct, then the solution is unique. A number of algorithms have been developed to solve exactly the Minimum Spanning Tree problem. The first algorithm is attributed to Boruvka (1926). Other popular algorithms are Prims, Krukals and the Reverse-Delete algorithms that all find solutions in polynomial time. The above algorithms also work at higher dimensions. At high dimensions, the Euclidean L2 distance or the L1 distance is generally used.

Minimum spanning trees have found applications in phylogeny, computer vision, and cytology just to name some domains. It has been used in astrophysics (Ascasibar and Sanchez-Almeida, 2011), and maybe very early since a large number of constellations defined by early civilisations are also shown to correlate well with a Minimum Spanning Tree (Dry et al., 2009).

For this paper, we used the *mst* command of the *ape* library in the R-package.

3.3 Neighbor Joining Tree Estimation

The Neighbor Joining Tree Estimation (NJ, Saitou and Nei, 1987; Gascuel and Steel, 2006) is based on a distance (or dissimilarity) matrix such as MST. This method is a bottom-up hierarchical clustering methods. It starts from a star tree (unresolved tree). A "corrected" distance $Q(i, j)$ between objects i and j from the dataset of n objects, is computed from the distances $d(i, j)$:

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

The branches of the two objects with the lowest $Q(i, j)$ are linked together by a new node u on the tree. This node replaces the pair (i, j) in the subsequent iterations through the distance between any other object k:

$$d(u, k) = \frac{1}{2}\left[d(i, k) - d(i, u)\right] + \frac{1}{2}\left[d(j, k) - d(j, u)\right] \tag{3}$$

Like most phylogenetic methods, NJ minimises a tree length, according to a criteria that can be viewed as a Balanced Minimum Evolution (Gascuel and Steel, 2006). For a tree metrics, Neighbor-Joining furnishes a simple algorithm to reconstruct a tree from the distance matrix. There is a large literature on how to best approximate a metrics by a tree metrics (see for instance Fakcharoenphol et al., 2003). Note that for points in $\mathbb{R}^2$, the distortion between the distance on a tree metrics and the Euclidean

distance between 2 points is zero if the points are collinear. Neighbor-Joining is justified if the difference between the original distance matrix and the distance matrix describing the X-tree obtained with Neighbor-Joining is not too large.

An important difference between MST (Sect. 3.2) and NJ is that the latter creates new internal nodes to build the tree. Such trees are called Steiner trees as opposed to MST trees. Most phylogenetic methods use Steiner trees with the advantage of a greater flexibility in the tree topologies but with the obvious disadvantage of a far larger computing time. However, the NJ algorithm is relatively fast as compared to other phylogenetic approaches.

In phylogeny, the internal nodes are assumed to be hypothetical ancestors from which the descendants inherited some specific properties. To allow for the discovery of intermediate species, the internal nodes always remain unlabelled. Steiner trees are thus also called semi-labelled trees.

For this paper, NJ computations were performed with the command *nj* of the *ape* library in the R-package.

3.4 Cladistics (Maximum Parsimony)

Cladistics seeks to establish evolutionary relationships between objects. It is a non-parametric character-based phylogenetic method (i.e. a non-parametric parameter-based approach), also called a maximum parsimony method. It does not use distances, and there is no assumption about the metrics of the parameter space. The "characters" are traits, descriptors, observables, or properties, which can be assigned at least two states characterising the evolutionary stage of the objects for that character. The maximum parsimony algorithm looks for the simplest arrangement of objects on a bifurcating tree. The complexity of the arrangement is measured by the total number of "steps" (i.e. changes in all character states) along the tree. The simplest tree supposedly depicts the simplest evolutionary scenario.

Like the NJ method (Sect. 3.3), Maximum Parsimony use semi-labelled trees which are graphical representations of the relationships between the objects that are situated on the leaves. The other kind of vertices are the internal nodes which are not labelled. The edges (or branches) connect any two adjacent vertices. If $f_k : V \to \mathbb{R}$ is a function assigning a real value $f_k(i)$ at each vertex $i \in V$ for character k, and C the cost function that assigns a value to change $f_k(i)$ into $f_k(j)$, the score s of the tree is the total number of evolutionary changes or steps:

$$s = \sum_k \sum_{\substack{i \in V \\ j \in V \\ i > j}} C\left(f_k(i), f_k(j)\right) \tag{4}$$

In practice, $f_k(i)$ at the leaves is the value of the parameter k. Note that with this definition, the parameters can be discrete or continuous. For continuous parameters, the cost function C can be a distance and chosen as the L1-norm:

$$s = \sum_k \sum_{\substack{i \in V \\ j \in V \\ i > j}} |f_k(i) - f_k(j)| \tag{5}$$

Cladistics looks for the tree minimising s. This is why cladistics is also referred to as the Maximum Parsimony method.

Robinson (1973) has shown that for a tree defined by continuous characters, a maximum parsimony score is reached for values of the internal nodes belonging to the set of values (or states) defined on the leaves. For that reason, the parsimony approach is often applied to the set of character states defined on the leaves. Note that this approach is valid for the L1 norm (Equation 5), but not for a L2 norm (euclidean distance). In addition, for a metric, the length of the MST furnishes bounds to s (Robinson, 1973).

The success of a cladistic analysis much depends on the behaviour of the parameters. In particular, it is sensitive to redundancies, incompatibilities, too much variability (reversals), and parallel and convergent evolutions. It is thus a very good tool for investigating whether a given set of parameters can lead to a robust and pertinent diversification scenario.

Being non-parametric and based on the parameters, Maximum Parsimony can be seen as a more general method to find tree-like arrangements of objects than NJ and, of course, MST. The 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 analysed.

Maximum Parsimony can take uncertainties into account. This is an invaluable capability that is rather rare among clustering methods. The implementation is very simple since it suffices for $f_k(i)$ to be a range of values instead of a single value. The algorithm then evaluates the different possibilities allowed by the range of values and select among them the one that provides the smallest s. In the same way, undocumented parameters can be included, and the most parsimonious diversification scenario provides a prediction for the unknown values.

The use of this approach in astrophysics is known as astrocladistics (for details and applications, see Fraix-Burnet et al., 2006b,c, 2009, 2010, 2012; Cardone and Fraix-Burnet, 2013). Simply speaking, the characters here are the parameters which supposedly evolve with the level of diversification of the objects. The discretization of the (continuous) values of parameters represent the states of the characters even though in the absolute discretization is not necessary. but imposed by softwares.

In the present study, we used the same kind of analysis as in our previous papers on astrocladistics. We discretized the parameters into 30 equal-width bins, which play the role of discrete evolutionary states. This choice of 30 bins, the maximum allowed by the software we have used, is justified by a fair representation of diversity, a stability of the analysis in the sense that the result does not depend on the number of bins, and a bin width roughly corresponding in general to the typical order of magnitude of the uncertainties (i.e. 7-10%, see Fraix-Burnet et al., 2009). We also adopted the parsimony criterion, which consists in finding the simplest evolutionary scenario that can be represented on a tree. Our maximum parsimony searches were performed using the heuristic algorithm implemented in the PAUP*4.0b10 (Swofford, 2003) package, with the Ratchet method (Nixon, 1999). The results were interpreted with the help of the Mesquite software (Maddison and Maddison, 2004) and the R-package (used for graphics and statistical analyses).

A cladistic analysis yields a tree that is easier to interpret when rooted because it defines an arrow for evolution. In the present case, the obvious root is the time when the star leaves the Main Sequence, that is the first point of each of the lineages. This tree can be projected on the HR diagram by computing the values of the parameters at the node with a squared change parsimony. In this way, all branches can be plotted with the internal nodes and the leaves correctly placed on the scatter plot.

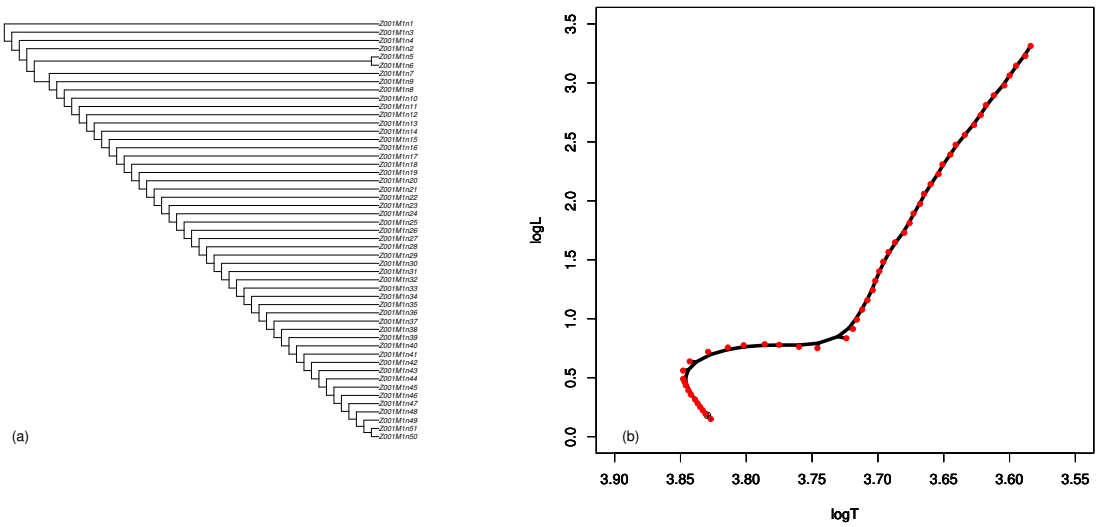

Figure 2: Cladistic analysis of the single lineage Z=0.001, M=1 $M_{\odot}$. (a) The tree obtained with two parameters $(\log T, \log L)$. (b) The projection of the tree on the HR diagram (black lines). Each red dot corresponds to one of the leaves of the tree.

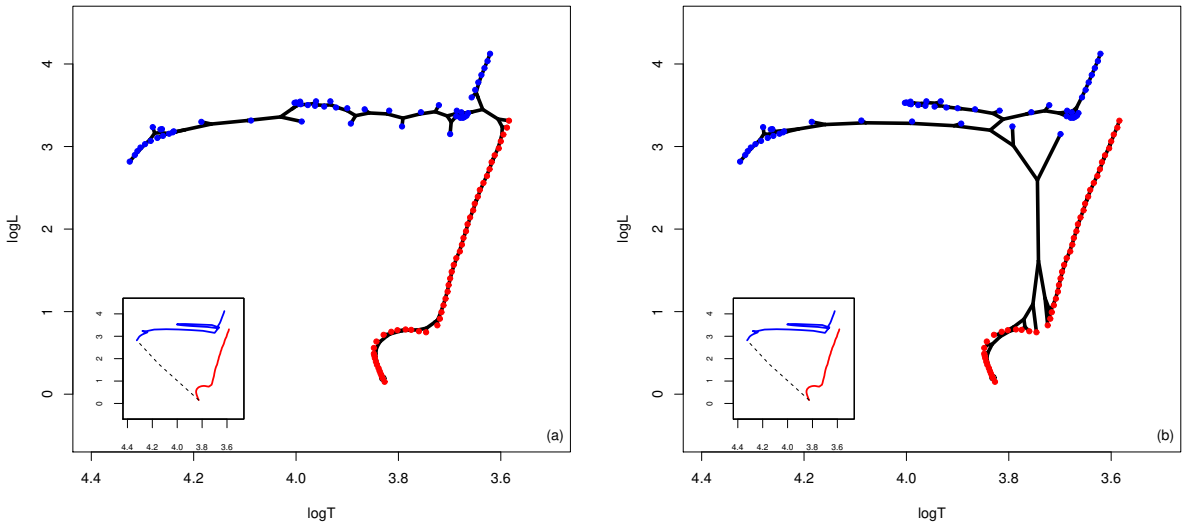

Figure 3: Projection of the tree obtained with Maximum Parsimony (black lines) for Z=0.001 M=1 $M_{\odot}$ (red) and 5 $M_{\odot}$ (blue). In the inlets, the theoretical paths are shown (see Fig. 1a) with the dashed line outlining the Main Sequence for this metallicity. (a) Cladistic analysis with two parameters $(\log T, \log L)$. (b) Cladistic analysis with 12 parameters (we exclude the mass and Z is constant).

4 Cladistic analyses of stellar evolution

We consider several configurations that are presented in subsections below: one lineage (single mass and single metallicity), two lineages with same metallicity, and nine lineages with three masses and three metallicities.

4.1 Single lineage

We first consider the very simple case of a single lineage. The tree obtained for Z=0.001 and M=1 $M_\odot$ using only the two parameters of the HR diagram, $\log T$ and $\log L$, is shown in Fig. 2a. It is perfectly linear, reproducing the correct chronology with the first point as root on top of this representation of the tree.

The projection of the tree on the HR diagram is shown in Fig. 2b. The lineage is perfectly recovered with only the two parameters $\log T$ and $\log L$, so that we do not show here the result with all parameters since it is identical.

4.2 Two lineages of same metallicity

In the case of two lineages, the difficulty is both to reproduce each of them and to connect them in a "correct" manner. The latter point would deserve a long discussion, since it requires to find an adequate common ancestor to both lineages. Ideally the two lineages should be connected by the points on the Main Sequence in order to respect the correct chronology of the evolution. However, the concept of ancestorship is quite delicate for stars since no lineage of stars gives birth to another lineage directly. The transmission with modification process between lineages, necessary for phylogenetic tools, does indeed exist and is the following. A group of stars with a distribution of masses is formed from a cloud of gas, the metallicity of which is thus given to all these stars. Then, the stars process the gas and enrich it in heavier elements. The most massive stars eject this enriched gas. This increases the metallicity of the interstellar medium from which a new group of stars may form with a distribution of masses and a higher metallicity. A similar process explains why Maximum Parsimony has been successful in the study of globular clusters (Fraix-Burnet et al., 2009). In this paper, we simply show in Sect. 4.3 that in theory we could connect the lineages at the Main Sequence, but we first explain below why this is not possible with the available parameters.

We consider two lineages with the same metallicity Z=0.001 and two different masses M=1 and 5 $M_\odot$. With the two parameters $\log T$ and $\log L$, the cladistic analysis is able to reconstruct the two lineages (Fig. 3a). However, since the connection takes place at the end of the lineages, the chronology of one of them is necessarily reversed. This is easily explained by the fact that in the two-parameter space $\log T$ vs $\log L$, the two lineages are the closest at their end (upper right). The connection at this point costs less with the parsimony criterion. With these two parameters, the chronology of both lineages cannot thus be reproduced with a cladistic analysis.

In addition, the evolutionary track of the M=5 $M_\odot$ lineage shows a loop that is not reproduced at all. This is not a surprise either, since there is not enough physical information in our analysis with this two parameters to explain the reversals of $\log T$ that create the loop.

The result of the cladistic analysis with 12 parameters (Z is constant and we exclude the mass, Fig. 3b) is significantly better. The loop in the track of the M=5 $M_\odot$ lineage is sketched although not

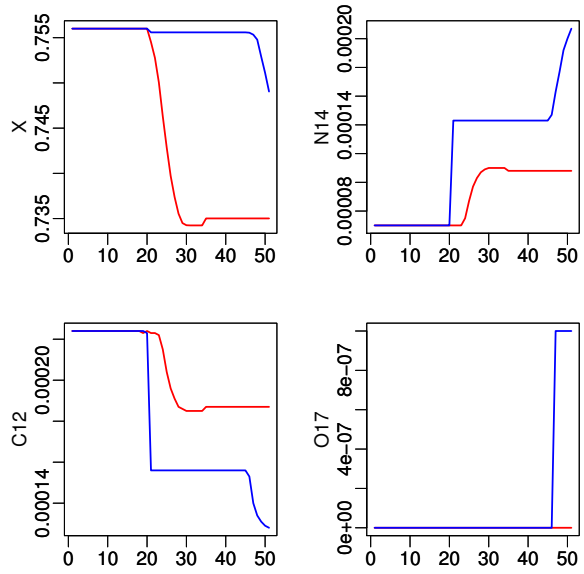

Figure 4: Examples of the evolutions of the abundance parameters as a function of the evolutionary stage for Z=0.001, M=1 $M_\odot$ (red) and 5 $M_\odot$ (blue).

fully reproduced. The connection between the two lineages takes place in the middle of each one. This is both a great improvement and disappointing.

The explanation lies in the evolutions of the parameters themselves (Fig. 4). All the abundance parameters (hence all available parameters except age, mass, $\log T$ and $\log L$) are constant and indistinguishable until the middle of the lineages. This means that the evolutionary tracks are closer at this point in the 12-dimension parameter space, in a manner similar to the case of the 2-dimension parameter space where the tracks are closer at the end (see above).

There is thus no hope to connect and separate the lineages at the Main Sequence with the available parameters.

4.3 Correct ancestorship and complex evolutionary tracks

We have seen previously that twelve parameters is better than two, but the informative power of these parameters should be adequate to retrieve the details of the evolutionary paths. In particular, what kind of parameters should we need to connect the lineages on the Main Sequence, and how can we reproduce the loop of the track for Z=0.001 and M=5 $M_\odot$?

We are compelled to create an artificial parameter that separates the lineages in the multidimensional parameter space right at the Main Sequence. After some trials and errors, we propose to replace the mass by an artificial parameter equal to the mass $\times$ the time step, and give it a weight of 3. This weight mimics an ideal situation where the information that the lineages should be distinct and connected at the Main Sequence (the phylogenetic signal) is present and hidden in several (here 3) compatible parameters.

In these conditions with 13 parameters, we get a much better result (Fig. 5), both for the connection and for the loop of the M=5 $M_\odot$ track.

Still, we can notice a small sub-branch on the M=1 $M_\odot$ track that goes backwards. This is because

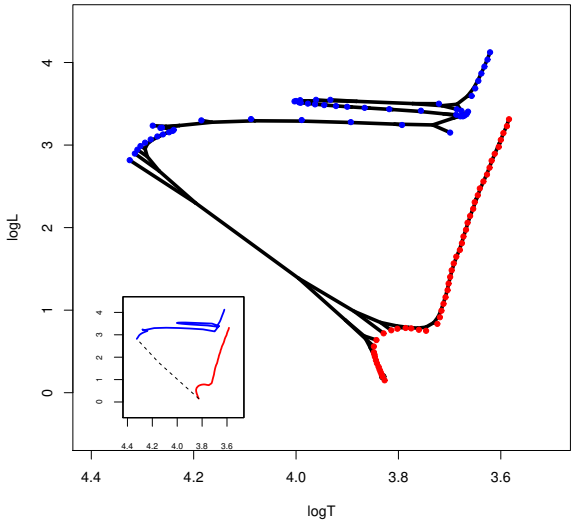

Figure 5: Projection of the tree obtained with Maximum Parsimony (black lines) for Z=0.001, M=1 $M_{\odot}$ (red) and 5 $M_{\odot}$ (blue), with 13 parameters (Z is constant) and replacing the mass by an artificial parameter equal to mass $\times$ the time step.

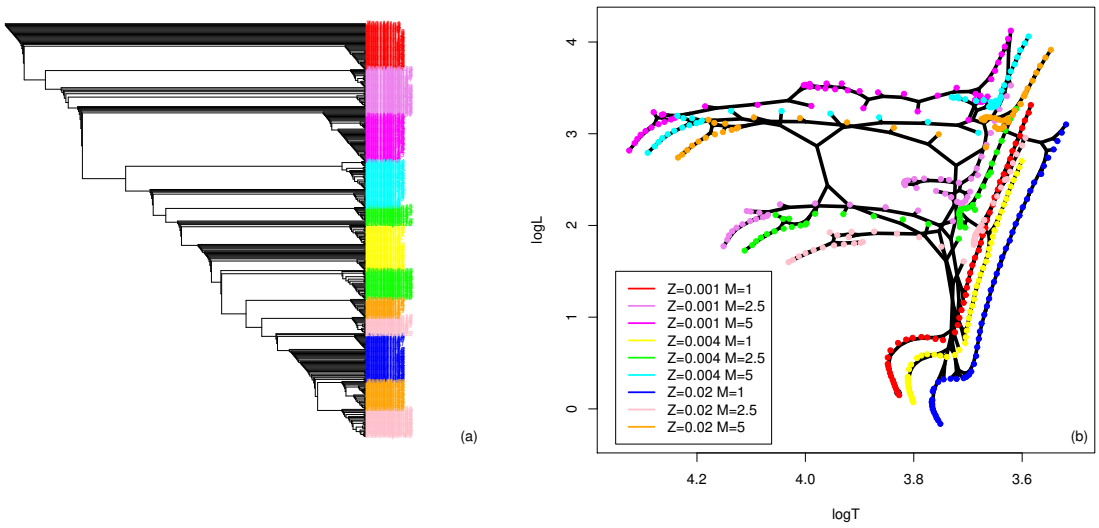

Figure 6: Cladistic analysis with 13 parameters (we exclude mass) of a sample with three masses (M=1, 2.5 and 5 $M_{\odot}$) and three metallicities (Z=0.001, 0.004 and 0.02). Colours indicate the theoretical lineages. (a) The tree obtained by Maximum Parsimony. (b) Projection of the tree on the HR diagram (black lines).

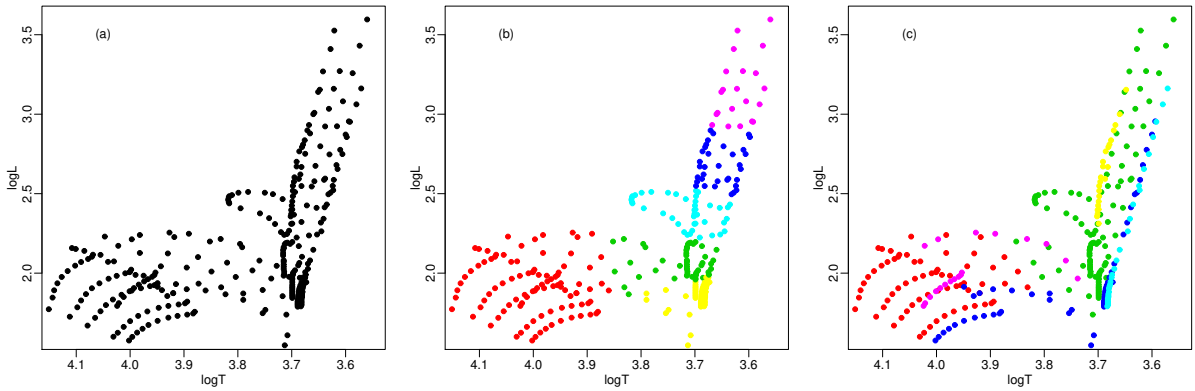

Figure 7: (a) All the points in the grid models for M=2.5 $M_\odot$ and Z=0.001, 0.004, 0.008, 0.02, 0.04 and 0.1. (b) A k-medoids analysis with six clusters using the two standardised parameters $\log T$ and $\log L$. (c) Same as (b) but with all parameters of Table 1 except age and mass. In (b) and (c), colors are arbitrary and represent the groups found by the k-medoids algorithm.

the $\log T$ parameter for this lineage first increases and then decreases again as a function of the evolutionary stage. Indeed, both $\log T$ and $\log L$ do not generally show an ideal behaviour for phylogenetic reconstruction since they sometimes show reversals. Too much of such behaviours, called homoplasies, can destroy a cladistic or any phylogenetic analysis. Fortunately in the present case, the effect is weak.

This artificial parameter was introduced here for illustration only. It will not be considered anymore in this paper. We have to stick to the available parameters and the limitations they impose (Sect. 4.2).

4.4 Several lineages

Let us consider a more complex sample with three metallicities and three masses. There are now nine lineages and we keep 13 parameters leaving mass aside. Including the mass does not change the result significantly.

Fig. 6a shows that the tree has correctly gathered the lineages together sequentially, except for three of them which are split in two parts. By identifying groups as sub-branches of the tree, one can isolate many parts of most of the lineages. The projection of the tree on the HR diagram (Fig. 6b) confirms that Maximum Parsimony is able to retrieve most of the evolutionary tracks of the different lineages.

As discussed previously, the parameters do not allow for correct connections of the lineages, yielding some confusion in their identification on the HR diagram. This also, explains why the lineages do not appear as distinct sub-branches on the tree.

5 A clustering challenge

When observing many stars, the HR diagram may look like Fig. 7a, where we have plotted all the points in the grid models for M=2.5 $M_\odot$ and Z=0.001, 0.004, 0.008, 0.02, 0.04 and 0.1. This is the same as Fig. 1b but without indicating the six known lineages.

From Fig. 7a, how can we cluster stars into physically meaningful groups? In other words, how can

we gather the stars that had the same mass and metallicity on the Main Sequence, but have since evolved? How can we reconstruct the lineages without fitting a physical model?

The keen eye of an astronomer would see two groups, one extending along the $\log T$ axis on the left and another one extending along the $\log L$ axis on the right. It could thus be concluded that there are two kinds of stars, one with a low luminosity, and one with a low temperature. But we know that there are six groups.

5.1 K-medoids

A natural extension to the intuitive eye-led method is a distance-based approach. As said in Sect. 3.1, k-means and k-medoids methods require the number of clusters as an input. In the present case, we know that there are six groups.

Since these methods tend to build hyperspheres in the parameter space, there is little hope to obtain lineages with the two parameters of the HR diagram. This is clearly seen in Fig. 7b where only $\log T$ and $\log L$ are used. None of the lineages are revealed.

The result looks better when 13 parameters (mass is constant) are used (Fig. 7c). But the groups obtained are still far from being correct.

5.2 Minimum Spanning Tree

A potential progress could be made if we could take into account the information that we are looking for lineages. Among the approaches suited for these kinds of structures, the MST method is probably the first obvious choice.

With an euclidean distance and 13 parameters, we have obtained the tree shown on Fig. 8a (top) with a "nsca" representation (using the two first axes of a non-symmetric correspondence analysis or nsca Lauro and D'ambra, 1984). There are three principal branches and a few smaller ones. It is difficult to justify more than three groups from this tree. Nevertheless, when this tree is projected on the HR diagram (Fig. 8a bottom) it appears that the MST method does quite good a job in retrieving many parts of the lineages, a large improvement as compared to the k-medoids result.

5.3 Neighbor Joining Tree Estimation

A better approach still is to consider tools that are used in phylogenetic studies, such as the Neighbor Joining Tree method, since they allow for more general tree topologies. The tree obtained with an euclidean distance and the same 13 parameters (Fig. 8b) shows that the three lineages with the higher metallicity are well grouped together, even though they are split into several sub-branches.

The projection of the tree on the HR diagram is a little bit confused by the mixing up of the lineages of low metallicities at low $\log T$ and high $\log L$, but otherwise provides a much better result than the k-medoids, and very comparable to MST, although this is still not perfect. Nevertheless, this projection can be misleading in giving the feeling that the MST is better. This is not exactly true since the MST tree would probably lead to a number of lineages smaller than six, while the NJ tree would lead to more groups or sub-groups.

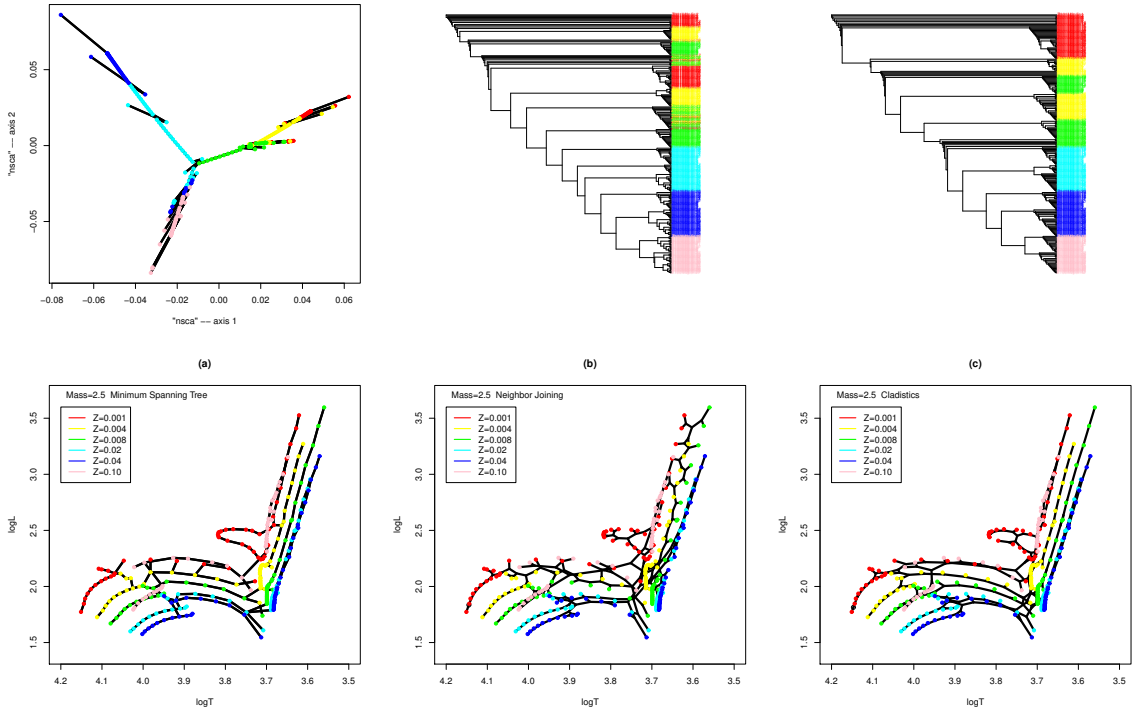

Figure 8: Phylogenetic analyses of the same sample with the same 13 parameters as in Fig. 7. The trees are shown on the top row and their projection (black lines) onto the HR diagram are shown on the bottom row. The colours of dots represent the theoretical lineages as in Fig. 1b. (a) Minimum Spanning Tree analysis. (b) Neighbour Joining Tree Estimation analysis. (c) Cladistic analysis.

5.4 Maximum Parsimony

The result of the cladistic analysis on the same sample and with the same 13 parameters is shown in Fig. 8c.

On the tree, it is clear that all the lineages are gathered together except two that are split into two parts. Indeed, all lineages are split into two sub-branches, which is the graphical translation of their connection near their middle due to the behaviour of the abundance parameters, as clearly illustrated in Sect. 4.2 and Fig. 3b.

The projection on the HR diagram looks far much better than the result from k-medoids analyses (Fig. 7). It also looks better than the NJ plot (Fig. 8b), and significantly better than the MST result (Fig. 8a). The loop of the $Z = 0.001$ track is less continuous than in the latter case, but it is yet entirely recovered.

6 Some applications

6.1 Isochrones and evolutionary tracks

In real life, we do not observe evolutionary tracks when observing a single cluster of stars since we take a snapshot as compared to the time scale of stellar evolution. Nevertheless, stars of same initial metallicity define an isochrone which is a linear sequence of evolution with mass. Consequently an isochrone can be considered as a lineage in the sense of a linear track in a parameter space.

This becomes more complicated in groups of stars like globular clusters and galaxies where several populations (defined as having same initial metallicity and same age) are present. Each population so defined builds an isochrone, but if several populations have the same metallicity and different ages, they also build parts of evolutionary tracks. The case with populations of different metallicities and ages have been analysed in Sect. 4.4. This result shows that the popullations can be identified and the lineages reconstructed, provided enough indicators of the chemical composition are available.

6.2 Maximum Parsimony on Globular Clusters

To a first approximation, globular clusters contain one population of stars, so that each cluster can be initially caracterised by its metallicity and its mass that both evolve with its age. Globular clusters formed in the same conditions should have very similar initial chemical compositions, with a range of masses which change strongly and non linearly with time. This ensemble of clusters thus define a lineage. Understanding the origin of the globular clusters requires to gather them into lineages.

Such a study has been performed on the globular clusters of our Galaxy using Maximum Parsimony (Fraix-Burnet et al., 2009). Three lineages were found, they correspond to three groups of clusters that formed during three phases of the assembly of our Galaxy. It cannot be excluded that one of the groups originates from a different galaxy than the other two.

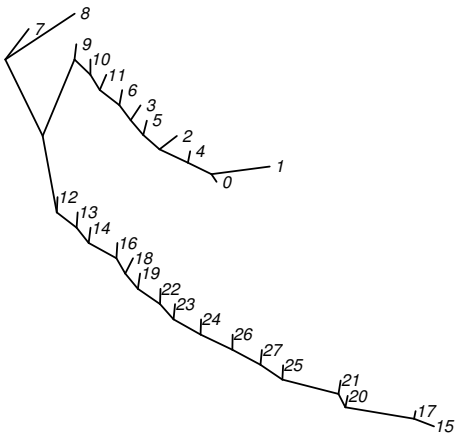

Figure 9: Tree obtained using Maximum Parsimony on the ASK classification by Sánchez Almeida et al. (2010). This tree should be compared with the MST result in Fig. 1 (bottom left) by Ascasibar and Sanchez-Almeida (2011).

6.3 MST and Maximum Parsimony on galaxy spectra

Ascasibar and Sanchez-Almeida (2011) performed a MST analysis of the 28 classes of galaxy spectra obtained with a k-means approach by Sánchez Almeida et al. (2010). The MST result made with the 3850 wavelengths of the template spectra and using a L1-norm (Manhattan distance) is shown in Fig. 1, bottom left, of Ascasibar and Sanchez-Almeida (2011).

This should be compared with the tree obtained through a Maximum Parsimony analysis using the same data as for the MST analysis and shown in Fig. 9. The two results are identical except that the Maximum Parsimony tree has added internal nodes. This illustrates the mathematical links between the two approaches as presented in Sect. 3.

Note that Maximum Parsimony is unable to analyze 700 000 spectra (containing several thousands wavelengths or parameters) in one row. MST could probably be used but it would be extremely difficult to determine groups from such a huge graph. Several strategies can be envisaged. Clustering the sample before performing a phylogenetic analysis is possible but the philosophies are different and a lot of care must be taken. As illustrated in Fig. 9 the final tree is heavily dependent on the previous classification. Another approach is to reduce the number of spectra by binning each wavelength in a very small number of bins (say 2 to 4). This will automatically reduce the number of different kinds of objects. Another approach is to proceed step by step as illustrated in the next section (Sect. 7).

6.4 Maximum Parsimony and Galaxy classification

The evolution and the diversity of galaxies cannot be summarized by a few parameters like for stars or even stellar clusters. Defining a lineage is therefore more difficult but still possible as detailed in Fraix-Burnet et al. (2006b,c). Of course these lineages cannot be easily visualised in a scatter plot like the CMD or HR diagram since the evolutionary tracks of galaxies take place in a multivariate parameter

space. The tree representation is certainly the best way to depict these "tracks".

However, very few works have been devoted to study the diversity of galaxies in this way. The most advanced study, using Maximum Parsimony, is given in Fraix-Burnet et al. (2012). Interestingly, these authors also use a clustering (k-means) analysis to compare the results. The agreement between the two approaches reinforces their conclusion, and can be explained by the careful objective selection of the parameters used to obtain the classifications. However, the clustering technique does not provide any relationship between the members of each group and between the groups themselves, and is thus not well adapted to establish the lineages.

7 Discussion

Before comparing the results of different clustering or phylogenetic methods (Sect. 5), we want to stress the importance of the quality of the parameters describing the objects and, consequently, their evolutionary status (Sect. 4). Because of the detailed limitations of the present data set, none of the approaches envisaged here can yield a perfect result.

The evolutionary paths of stars obviously do not form compact and distinct groups in the parameter space defined from Table 1. As a consequence we do not expect partitioning methods such as k-medoids to perform well in retrieving the lineages of stars defined by their properties on the Main Sequence. This is entirely confirmed the results presented in Sect. 5.1.

This remark should hold for most of the astrophysical objects since their properties change with time, triggered by secular or violent, internal or external, events, a global process that is called evolution. The purpose of classification is to group objects into physically pertinent classes, in order to identify lineages to distinguish between evolutionary transformation and diversification. For stars, this means distinguishing stars $\left\{X_{(M,Z)}(t), \forall t\right\}$ that have evolved from an ancestor $X_{(M,Z)}(0)$ with a mass-metallicity set (M,Z) from stars $\left\{X_{(M',Z')}(t), \forall t\right\}$ that evolved from an ancestor $X_{(M',Z')}(0)$ having a different set (M',Z'). This has motivated the development of astrocladistics (Fraix-Burnet et al., 2006a,b,c).

The eye-eld approach of classification from a scatter plot, which partitioning methods extend objectively to parameter spaces of dimension higher than 2 or 3, use global similarity to compare objects, and can thus be questioned on their relevance to classify objects in inevitable evolution.

The evolutionary paths of stars suggest graphs to be better suited to describe the clustering of these objects. We have considered three approaches, MST, NJ and Maximum Parsimony (or cladistics). All of these methods depict the relationships between the objects on a graph that is a tree. The main difference is that the MST tree have only labelled nodes while the other two kinds of trees are Steiner or semi-labelled trees and allows for unkown internal nodes. Obviously NJ and Maximum Parsimony provide a much wider variety of tree topologies, at the expense of the computing time.

Both MST and NJ are based on pairwise distances, while Maximum Parsimony uses the parameters themselves. However, Eq. 1 and Eq. 5 show a striking similarity for the L1 norm. On the contrary, the NJ uses a more sophisticated distance (Eq. 2) that is supposed to correct for different evolutionary rates along different branches.

The MST approach imposes the ordering of the objects along the tree, while with the NJ and cladistic methods, it is necessary to choose the root of the tree to interpret it with an arrow of evolution or diversification.

The tree in Fig. 2a is good example of a perfect MST tree. It has been obtained with Maximum Parsimony and may be described by a MST through a simple transformation. This tree is almost a caterpillar tree. A caterpillar tree is an unrooted binary phylogenetic tree that reduces to a path once the edges incident to the leaves are removed. By just removing one object, the tree in Fig. 2a is indeed a caterpillar tree. The order of the labels on the caterpillar tree is the same as the order of the labels on the MST (see Fig. 2b). In the situation of Fig. 2, it is equivalently justified to describe the data by a minimum spanning tree or a maximum parsimony tree. For more general trees, the two approaches may furnish quite different orders of the objects and there may be no circular order of the objects common to the MST and the cladistic trees. The MST trees can nevertheless be used as a first approximation to the cladistic tree. For a metric, the length of the MST furnishes bounds to the maximum parsimony score ($\frac{w_{min}}{2} < s_{min} < w_{min}$). While finding the maximum parsimony tree is NP-hard, determining the MST can be solved in polynomial time.

If the tree to find is thought to be MST, then the MST method is a good choice since it is very fast and finds the optimal tree. On the contrary, heuristic approaches with NJ and Maximum Parsimony are the only options, even if they can miss the optimal solution despite very long computation times.

We would like to finish the discussion by emphasising the role of the parameters. We have mentioned that $\log T$ and also $\log L$ are not ideal parameters for a phylogenetic approach since they show some reversals (Sect. 4.2, 4.3). It would seem strange to disregard them while projecting the evolutionary paths on the scatter plot they form. It could also seem whimsical to do that since they look so obvious relevant physical parameters to describe the evolution of stars.

However, it is clear that $\log T$ and $\log L$ evolve differently than the abundance parameters (Fig. 4) and they show reversals that explains some loop structures in the evolutionary paths (Fig. 1). In a real study with real data, detecting these reversals might be difficult, but the careful examination of the results should reveal some inconsistencies between the behaviours of all parameters. In any case, several trials should be attempted with several subsets of objects and subsets of parameters to assess the sensitivity of the result (its "robustness") on the input data.

In the present study, the results obtained with $\log T$ and $\log L$ only are clearly different from those obtained with all parameters. This is a clear indication that there are two kinds of parameters that play somewhat different roles. With this info, we may proceed in several steps. From Sect. 4, we can conclude the the results obtained with $\log T$ and $\log L$ only are not very satisfactory because they are not very informative. Then, we can make an analysis without these two parameters and with the 11 abundance parameters only. This is shown on Fig. 10 and Fig. 11 for MST and Maximum Parsimony respectively using the same sample as in Sect. 5.

For MST, the comparison of Fig. 10a with Fig. 8a (top) shows a striking difference. On the former figure, it is not possible to distinguish groups objectively, but the six lineages form a unique sequence in the "nsca" representation of the MST. When projected on the HR diagram (Fig. 10b), this becomes quite messy.

However, if we split the curve on the "nsca" representation in two parts, to the left and right of a vertical line at 0 on the axis 1, we obtain two curves which are somewhat monolithic. A further MST analysis with 13 parameters performed independently on each of the two subsets yields the results shown in Fig. 10c and Fig. 10d. The improvement as compared to Fig. 8a (bottom) is clear since the six lineages are nearly perfectly retrieved.

This two step procedure can also be applied to Maximum Parsimony. The analysis with 11 parameters yields the tree shown in Fig. 11. The comb-shape of the subtrees for each lineages is due to the lack of information in the abundances parameters to organise the objects (see Fig. 4. Nevertheless, the

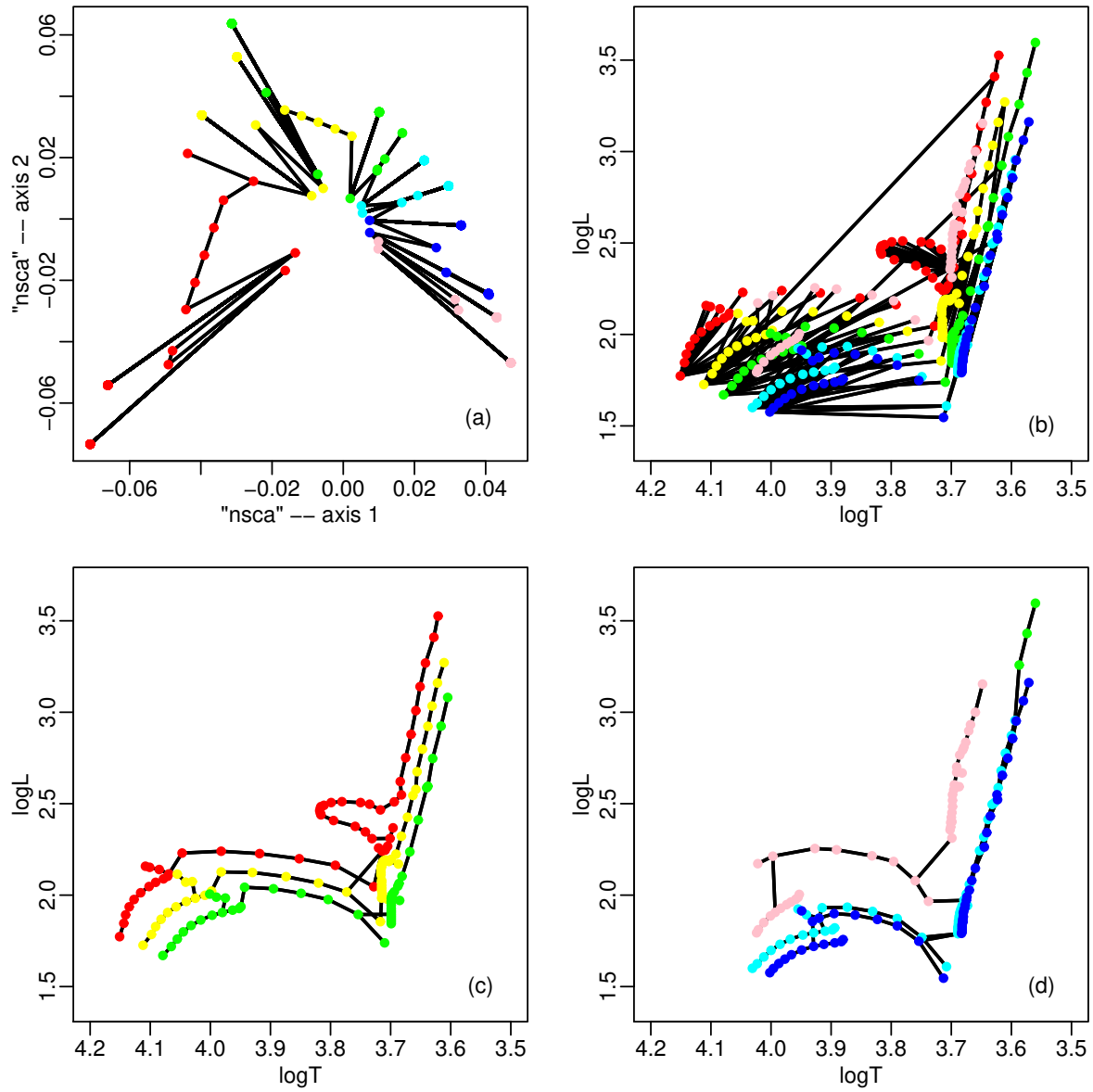

Figure 10: A MST analysis of the sample used in Sect. 5 in several steps. The colours represent the theoretical lineages as in Fig. 1b and Fig. 8. (a) The MST tree obtained with 11 parameters (excluding $\log T$ and $\log L$) in the "nsca" representation space. (b) The projection of the tree on the HR diagram. (c) The projection of the MST tree obtained with 13 parameters using the fist half of the points selected from (a). (d) Same with the second half.

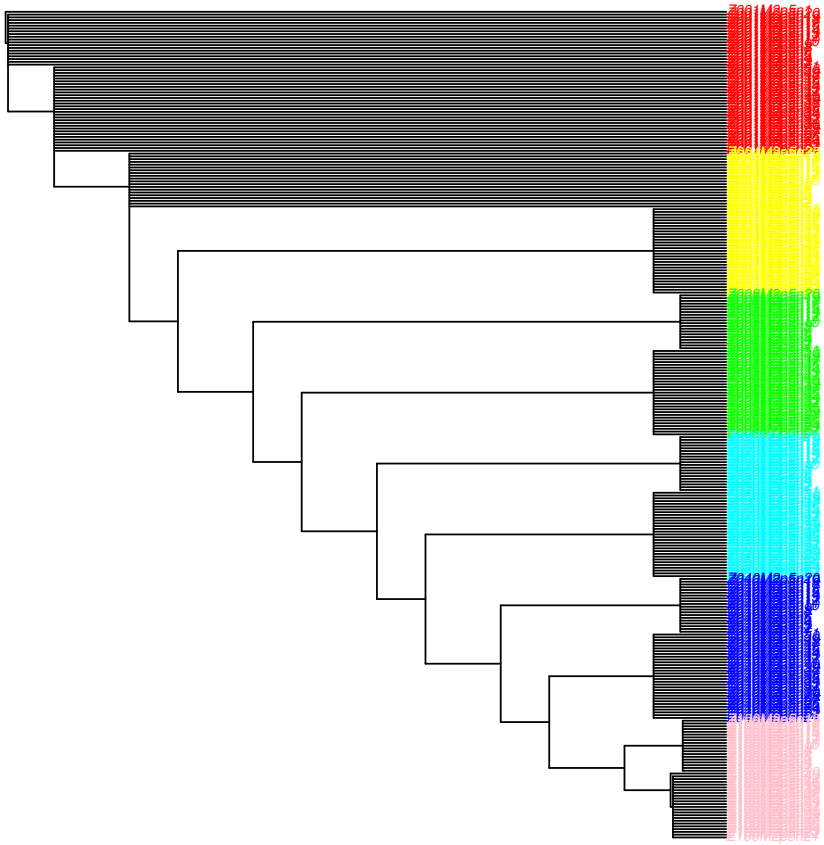

Figure 11: A cladistic analysis of the sample used in Sect. 5 with 11 parameters and without $\log T$ and $\log L$. The colours represent the theoretical lineages as in Fig. 1b, Fig. 8 and Fig. 10.

lineages are perfectly retrieved, even better than with 13 parameters (Fig. 8c), except for the fact that they are split in two parts due to the behaviour of the abundances parameters as already mentioned (Sect. 4.2). Since only $\log T$ and $\log L$ can provide an evolutionary information within each lineages, further analyses of subsamples drawn from the tree in Fig. 11 does not provide improvement as compared to a direct analysis of the full sample with 13 parameters (Sect. 5.4).

As a conclusion, an approach in several steps can be quite efficient in first determining the groups and second establishing the relationships between and inside the groups using subsamples if necessary. Even if the MST is very fast, its efficiency seems to depend very significantly on the parameters and the subsamples used. This is clearly much less the case for Maximum Parsimony.

8 Conclusions

In this paper, we have used model grids of stellar evolution to see how several partitioning (k-medoids) and phylogenetic (Minimum Spanning Tree, Neighbour-Joining Tree and Maximum Parsimony) methods can retrieve the lineages of stars depicted by the evolutionary tracks on the HR diagram ($\log T$ vs $\log L$). Mass and eleven abundances parameters are available in addition to $\log T$ and $\log L$.

A sub-sample of stars defining some simple and well separated tracks have first been analysed with Maximum Parsimony. It appears that the abundance parameters used here cannot distinguish the lineages up to the middle of the tracks, so that the analysis cannot connect them to each other at the Main Sequence as we might ideally hope. This analysis points out that the result of a statistical multivariate analysis much depends on the parameters that should be understood and investigated as much as possible.

With this information in mind, we then proceed to retrieve six lineages of stars with the same mass (M=2.5 $M_{\odot}$) and six different metallicities. Since the groups are made of evolutionary sequences in the parameter space, they are expected to be elongated, defining a lineage, a situation that is not adequate for partitioning methods like k-means or k-medoids used here. We confirm this point even when all 13 parameters are used.

We find that the MST is relatively efficient on the stellar evolution model grids. It retrieves the lineages quite correctly on the HR diagram. However, it would not be easy to derive that six groups are present. It seems that a two step analysis can be much more successful. This two step analysis takes into account the fact that $\log T$ and $\log L$ do not evolve in the same manner as the abundance parameters, something known from the models but that can be guessed from the comparison between other analyses made with different subsets of parameters. Hence, a first analysis with the abundance parameters suggests to divide the sample into two parts, while a subsequent analysis of each of them with 13 parameters give a very good result, especially when projected onto the HR diagram.

In the MST approach, each node of the tree is labelled, that is it corresponds to an object of the sample. By introducing unlabelled ("internal") nodes, one can greatly increase the possible topologies, of course at the expense of the computation time. Phylogenetic methods all use such semi-labelled trees. In this paper, we have considered two of them, the NJ and Maximum Parsimony. NJ, like MST, is based on the pairwise distances between the objects. This method here yields a fairly good result, slightly better than MST in the sense that the tree more easily points to the correct lineages.

Maximum Parsimony is based only on the parameters themselves. We show in this paper that for continuous parameters, and using a L1-norm (Manhattan distance), there is a mathematical similarity between MST and Maximum Parsimony. The main and important difference is that Maximum Parsimony uses semi-labelled trees. Hence, it can explore a larger variety of tree topologies and, unsurprisingly,

yields nearly perfect results on the samples used in this work.

Classification is a fundamental activity in astrophysics that plays crucial roles in our understanding of the Universe. It is only to think to the Hubble classification of galaxies, the stellar classification, or maybe more spectacularly, the Cepheid and Supernovae classifications that are so influencial in determining the scale of our Universe. However, classification is becoming a more difficult topic with the huge flow of data astrophysics is now facing. Sophisticated statistical tools must be used and developed, and the present paper tries to briefly review the two principal approaches to classification, clustering and phylogenetic, with a very simple illustration of their application in astrophysics.

Acknowledgments

This research has made use of the VizieR catalogue access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23. We thank Emmanuel Davoust for a careful reading of the manuscript, and many colleagues at IPAG for useful discussions.

References

Ascasibar, Y., Sanchez-Almeida, J., Aug. 2011. Do galaxies form a spectroscopic sequence? MNRAS 415, 2417–2425.
URL http://adsabs.harvard.edu/abs/2011MNRAS.415.2417A

Boruvka, O., 1926. O jistem problemu minimalnim (about a certain minimal problem). Praca Moravske Prirodovedecke Spolecnosti 3, 37–58.

Cardone, Vincenzo, F., Fraix-Burnet, D., Jun. 2013. Hints for families of grbs improving the hubble diagram. Monthly Notices of the Royal Astronomical Society 434 (3), 1930–1938, 10 pages, 6 figures, 4 tables, accepted for publication on MNRAS.
URL http://hal.archives-ouvertes.fr/hal-00847156

Charbonnel, C., Meynet, G., Maeder, A., Schaller, G., Schaerer, D., Oct. 1993. Grids of stellar models - part three - from 0.8 to 120-solar-masses at z=0.004. A& AS01, 415–419.
URL http://adsabs.harvard.edu/abs/1993A%26AS..101..415C

Chattopadhyay, T., Misra, R., Naskar, M., Chattopadhyay, A., 2007. Statistical evidences of three classes of gamma ray bursts. Astrophysical Journal 667, 1017.

Dry, M., Navarro, D., Preiss, K., Lee, M., 2009. The perceptual organization of point constellations. In: Annual Meeting of the Cognitive Science Society.
URL http://csjarchive.cogsci.rpi.edu/proceedings/2009/papers/267/index.html

Fakcharoenphol, J., Rao, S., Talwar, K., 2003. A tight bound on approximating arbitrary metrics by tree metrics. In: Proceedings of the 35th Annual ACM Symposium on Theory of Computing. pp. 448–455.
URL http://www.dcg.ethz.ch/lectures/fs11/seminar/paper/jasmin-1.pdf

Feigelson, E., Babu, G., 2012. Modern Statistical Methods for Astronomy: With R Applications. Cambridge University Press.
URL http://books.google.fr/books?id=Gd8MywAACAAJ

Felsenstein, J., 2003. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fraix-Burnet, D., Chattopadhyay, T., Chattopadhyay, A. K., Davoust, E., Thuillard, M., 2012. A six-parameter space to describe galaxy diversification. Astronomy and Astrophysics 545, A80.
URL http://dx.doi.org/10.1051/0004-6361/201218769

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 and Astrophysics 455, 845–851.
URL dx.doi.org/10.1051/0004-6361:20065098http://hal.archives-ouvertes.fr/hal-0

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, 31–56.
URL dx.doi.org/10.1007/s00357-006-0003-5http://hal.archives-ouvertes.fr/hal-00

Fraix-Burnet, D., Davoust, E., Charbonnel, C., 2009. The environment of formation as a second parameter for globular cluster classification. MNRAS 398, 1706–1714.
URL http://hal.archives-ouvertes.fr/hal-00394425/en/

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, 57–78.
URL dx.doi.org/10.1007/s00357-006-0004-4http://hal.archives-ouvertes.fr/hal-00

Fraix-Burnet, D., Dugué, M., Chattopadhyay, T., Chattopadhyay, A. K., Davoust, E., Oct. 2010. Structures in the fundamental plane of early-type galaxies. MNRAS 407, 2207–2222.
URL http://hal.archives-ouvertes.fr/hal-00487853/en/

Gascuel, O., Steel, M., 2006. Neighbor-joining revealed. Molecular Biology and Evolution 23 (11), 1997–2000.
URL http://mbe.oxfordjournals.org/content/23/11/1997.abstract

Ghosh, J., Liu, A., 2010. The Top Ten Algorithms in Data Mining. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. Taylor & Francis, Ch. The k-means Algorithm, pp. 21–36.
URL http://books.google.fr/books?id=_kcEn-c9kYAC

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 (1), pp. 54–64.
URL http://www.jstor.org/stable/2346439

Gratton, R. G., Johnson, C. I., Lucatello, S., D'Orazi D'Orazi, V., Pilachowski, C., Oct. 2011. Multiple populations in omega centauri: a cluster analysis of spectroscopic data. A& A34, A72.
URL http://adsabs.harvard.edu/abs/2011A%26A...534A..72G

Hennig, W., 1965. Phylogenetic systematics. Annual Review of Entomology 10, 97–116.

Kaufman, L., Rousseeuw, P., 1987. Clustering by means of medoids. In: Dodge, Y. (Ed.), Statistical Data Analysis Based on the L1-Norm and Related Methods. Elsevier/North-Holland, Amsterdam, pp. 405–416.

Kaufman, L., Rousseeuw, P. J., 1990. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chinchester, New York, Weinheim.
URL http://www.amazon.com/Finding-Groups-Data-Introduction-Probability/dp/0471

Lauro, N., D'ambra, L., 1984. L'analyse non symtrique des correspondances. In: Diday, E. . C. e. (Ed.), Data Analysis and Informatics III. Elsevier, North Holland., p. 433446.
URL https://www2.lirmm.fr/lirmm/interne/BIBLI/CDROM/INFO/2001/NSDA_2001/NSDA/da

MacQueen, J. B., 1967. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. pp. 281–297.
URL http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html#mac

Maddison, W. P., Maddison, D. R., 2004. Mesquite: a modular system for evolutionary analysis.
URL http://mesquiteproject.org

Makarenkov, V., Kevorkov, D., Legendre, P., 2006. Phylogenetic network construction approaches. In: Dilip K. Arora, R. M. B., Singh, G. B. (Eds.), Bioinformatics. Vol. 6 of Applied Mycology and Biotechnology. Elsevier, pp. 61 – 97.
URL http://www.sciencedirect.com/science/article/pii/S1874533406800067

More, S., Kravtsov, A. V., Dalal, N., Gottlöber, S., Jul. 2011. The overdensity and masses of the friends-of-friends halos and universality of halo mass function. ApJS95, 4.
URL http://adsabs.harvard.edu/abs/2011ApJS..195....4M

Mowlavi, N., Schaerer, D., Meynet, G., Bernasconi, P. A., Charbonnel, C., Maeder, A., Mar. 1998. Grids of stellar models. vii. from 0.8 to 60 m_\odot at z = 0.10. A& AS28, 471–474.
URL http://adsabs.harvard.edu/abs/1998A%26AS..128..471M

Nixon, K. C., 1999. The parsimony ratchet, a new method for rapid parsimony analysis. Cladistics 15 (4), 407–414.
URL http://dx.doi.org/10.1111/j.1096-0031.1999.tb00277.x

Reynolds, A., Richards, G., Iglesia, B., Rayward-Smith, V., 2006. Clustering rules: A comparison of partitioning and hierarchical clustering algorithms. Journal of Mathematical Modelling and Algorithms 5, 475–504.
URL http://dx.doi.org/10.1007/s10852-005-9022-1

Robinson, D., 1973. Extending a function on a graph. Discrete Mathematics 6 (1), 89 – 99.
URL http://www.sciencedirect.com/science/article/pii/0012365X73900381

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4 (4), 406–425.
URL http://mbe.oxfordjournals.org/content/4/4/406.abstract

Sánchez Almeida, J., Aguerri, J. A. L., Muñoz-Tuñón, C., de Vicente, A., May 2010. Automatic unsupervised classification of all sloan digital sky survey data release 7 galaxy spectra. ApJ 714, 487–504.
URL http://arxiv.org/abs/1003.3186

Schaerer, D., Charbonnel, C., Meynet, G., Maeder, A., Schaller, G., Dec. 1993a. Grids of stellar models - part four - from 0.8-solar-mass to 120-solar-masses at z=0.040. A& AS02, 339–342.
URL http://adsabs.harvard.edu/abs/1993A%26AS..102..339S

Schaerer, D., Meynet, G., Maeder, A., Schaller, G., May 1993b. Grids of stellar models. ii - from 0.8 to 120 solar masses at z = 0.008. A& AS8, 523–527.
URL http://adsabs.harvard.edu/abs/1993A%26AS...98..523S

Schaller, G., Schaerer, D., Meynet, G., Maeder, A., Dec. 1992. New grids of stellar models from 0.8 to 120 solar masses at z = 0.020 and z = 0.001. A& AS6, 269–331.
URL http://adsabs.harvard.edu/abs/1992A%26AS...96..269S

Semple, C., Steel, M. A., 2003. Phylogenetics. Oxford University Press.

Simpson, J. D., Cottrell, P. L., Worley, C. C., Dec. 2012. Spectral matching for abundances and clustering analysis of stars on the giant branches of ω centauri. MNRAS27, 1153–1167.
URL http://adsabs.harvard.edu/abs/2012MNRAS.427.1153S

Sugar, C. A., James, G. M., 2003. Finding the number of clusters in a dataset. Journal of the American Statistical Association 98 (463), 750–763.
URL http://pubs.amstat.org/doi/abs/10.1198/016214503000000666

Swofford, D. L., 2003. Paup*: Phylogenetic analysis using parsimony (*and other methods).
URL http://people.sc.fsu.edu/~dswofford/paup_test/

Tajunisha, Saravanan, 2010. Performance analysis of k-means with different initialization methods for high dimensional data. International Journal of Artificial Intelligence & Applications (IJAIA) 1 (4), 44–52.
URL http://airccse.org/journal/ijaia/currentissue.html#october

Thuillard, M., Fraix-Burnet, D., 06 2009. Phylogenetic Applications of the Minimum Contradiction Approach on Continuous Characters. Evolutionary Bioinformatics 5, 33–46.
URL www.la-press.com/phylogenetic-applications-of-the-minimum-contradiction-ap

Thuillard, M., Fraix-Burnet, D., 10 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, 213–221.
URL www.la-press.com/phylogenetic-trees-and-networks-reduce-to-phylogenies-on-