Centrality of Objects in a Multidimensional Space and its Effects on Distance-Based Biological Classifications

Livio Antonielli¹, Vincent Robert², Laura Corte¹, Luca Roscini¹, Rita Ceppitelli³ and Gianluigi Cardinali¹

¹Department of Applied Biology – Microbiology Division, University of Perugia, Italy
²CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands
³Department of Mathematics and Informatics, University of Perugia, Italy

Abstract: Typology is based on the concept that one individual (the type) can represent the whole group. In yeast taxonomy the type strain is the representative of the whole species and is considered an important tool for classification. Although the evolutionary, phylogenetic and biological species concepts are in contrast with this approach, the International Codes of nomenclature still use typology, which remains one of the most operative systems. These incongruities demand a multidisciplinary investigation on the nature of the type, its characteristics and the possibility of the type to be defined on the basis of a shareable criterion. In this paper we present an approach to the problem of type centrality based on mathematical demonstrations and numerical examples. This paper intended to show the possibilities offered to bioinformatics research by the implementation of multidisciplinary approaches in biology and in proposing a general approach to the definition of the type representing any sort of set, described with multiple descriptors.

Keywords: Species, microbial species, typological species, type strain.

INTRODUCTION

Taxonomy aims to classify microbes on the basis of their overall similarity or by reconstructing their evolutionary history [1]. There are several approaches at the problem, supported by different schools of thought that are phenetics, cladism and evolutionary classification [2]. Phenetics aims to discriminate among microbes according to the distances observable in the present moment with numerical taxonomy methods [1]. Cladism bases the classification on the reconstruction of the possible evolutionary relationships, i.e. on the genealogy; whereas evolutionary (or traditional) classification is based on the comparison of similarity and differences in the light of the supposed evolutionary history. In recent years the cladism/phylogenetic approach seems to gain popularity [3] at a level that some exponents of this school claim for the eradication of phenetics [4].

As a matter of fact microbiologists cope with several problems at least partially absent in botany and zoology: the limited knowledge of the biodiversity and the poorness of characters. In fact, actual microbial biodiversity has been estimated to be several fold higher than that is already isolated and classified. Moreover, sophisticated morphological analyses are hampered by the simplicity of microbial shapes and structures. Altogether, microbial taxonomy is evolved from morphological analysis (under the light microscope) to an increasingly intense use of physiological characters [5-7]. More recently molecular characters were introduced with a remarkably beneficial effect of the times necessary to identify strains and in the character stability [8-10]. At the same time an evolution occurred in the character analysis with a succession of the hierarchical dichotomic system (the taxonomic keys) to multivariate analysis and phylogenetic inference [11-14].

Bacteria and yeast are currently identified on the basis of the observation that strains are likely to be conspecific when differing in a sequence encoding for the RNA of a large ribosomal subunit by less than a given threshold, established in 1% yeast taxonomy [9], and in 2.5-3% bacterial taxonomy [15, 16]. This system is clearly nominalistic and could be questioned on the type of genes to use or on the thresholds to employ, although none of these issues would affect the basic fact that this practical approach is the only one that microbiologists can currently undertake without a consensus on the microbial species concept [15]. Moreover, microbial taxonomists work in the mainframe of two codes of nomenclature that indeed requires a type strain to be designated in order to describe a new species. The International Code of Botanical Nomenclature (Vienna Code 2006), ruling for Fungi, at the point 7.2 states: “A nomenclatural type (typus) is that element to which the name of a taxon is permanently attached, whether as the correct name or as a synonym. The nomenclatural type is not necessarily the most typical or representative element of a taxon.” (http://ibot.sav.sk/icbn/main.htm). Similarly, the International Code of Nomenclature of Bacteria at Rule 15-Section 4 “A taxon consists of one or more elements. For each named taxon of the various taxonomic categories (listed below), there shall be designated a nomenclatural type. The nomenclatural type, referred to in this Code as
"type, " is that element of the taxon with which the name is permanently associated. The nomenclatural type is not necessarily the most typical or representative element of the taxon. The types are dealt with in Rules 16–22." (http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=icnb&part=A184).

The necessity of a “nomenclatural type”, hereinafter simply referred to as “type” poses the question on what happens if the type strain is not centric and if the species have not similar dimensions [17]. In fact, the above cited minimal distances for two strains to belong to two different species can be applied in two alternative ways:

1. Distance is calculated from whatever member of the species.
2. Distance is calculated from the type strain.

In the first case the species is likely to widen with the accumulation of successive identifications which include strains similar to peripheral strains of the species. The second strategy is highly affected by where the type strain falls within the species distribution, in fact if the type strain is acentric the identification procedure is likely to assign strains not highly similar to the species as a whole and to exclude others, which instead should be included. If the type strain has to work as the reference is therefore important that it is as representative as possible, implying the necessity to know how central it is within the distribution of the strains belonging to its species.

In this article we will deal with the centrality of the reference object within a multidimensional group of objects, from a merely mathematical and statistical point of view, leaving the specific taxonomical discussion for another instance.

The analysis of the objects described by many parameters is the subject of many studies and in particular of the group of techniques called “ordination in reduced space” [18]. Techniques such as the Principal Component Analysis (PCA) or the Principal Coordinate Analysis (PCoA) define a number of dimensions on which the objects are scattered according to their relative distances. The combination of two or three of these dimensions creates a space within which the points are distributed so that the distances of the graph correspond as closely as possible to the relative distances between objects [18]. Hierarchical cluster analysis produces dendrograms, widely used in taxonomy to give a synthetic view of the relative distances and relations among strains or species. None of these systems give a way to determine which of the objects of a putative group are “central” or “how much an object is close to the center”. These issues are the aims of the present work, which builds on the problem of the taxonomic type, but points to the general problem of defining the type in any sort of classification based on multidimensional objects.

**MATERIALS AND METHODS**

The script “Centralinity” was written in the open source statistical environment “R” (http://cran.r-project.org/) and is freely available from the Journal’s site as supplementary materials or upon request to the corresponding author.

**Outline of the Centrality Script**

The script consists of two functions: `center` and `classify`. The former defines the centrality of multivariate objects, the latter simulates a distance-based classification of the objects using a defined object as a central reference, i.e. as type. All software materials described in this paper can be obtained freely upon request to the corresponding author.

The function `center` has the following arguments:

1. **m.** “m” is the input matrix, which can be either of quantitative or qualitative descriptors. As for the latter descriptors, the current version (0.64) is designed only for qualitative binary characters. The possibility to work with DNA alignments will be implemented in the next future.
2. **bin=F.** bin is a logical argument can be FALSE (default) or TRUE for qualitative or quantitative matrices respectively.
3. **method=“euclidean”.** The argument “method” refers to the distance method used with quantitative characters. The options are the same of the `dist` function in the stats package of the R package base distribution. For more details, simply digit “?dist” to read the specific guide. The default value is “Euclidean”.
4. **binmethod=2.** This argument refers to the “method” to use in the `dist.binary` function described in the package ADE4 (http://cran.r-project.org/) [19].
5. **doubleplot=T.** The `center` function yields a graphical output consisting of two plots, one describing the level of centrality of single objects and a second including the center of the distribution. If the logical argument `doubleplot` is set to TRUE (default) both plots are displayed, otherwise when FALSE is chosen only the first plot is shown.
6. **labels=F.** Logical argument, if FALSE (default) object labels are not reported in the plot. This option is preferable when the plot reports several objects and labels are jammed.

The function `center` produces the following outputs:

1. Plot reporting the centrality of single objects.
2. Plot reporting the centrality of single objects and the central point of the distribution.
3. The series of absolute and relative distances from the central object.
4. The series of absolute and relative distances from the central point.
5. A series of parameters such as the mean distance.

**Algorithms Implemented in center**

The centrality of objects is calculated with the “minimal distance algorithm” (MDA), based on the rationale that the central object minimizes the distances from all other objects. MDA works as outlined below:

1. MDA produces a distance matrix from the input matrix.
Let \( x^i, x^2, \ldots, x^m \) be \( m \) vectors in \( \mathbb{R}^n \):
\[
x^i = (x_{1}^i, x_{2}^i, \ldots, x_{n}^i), \quad i = 1, 2, \ldots, m
\]
and let \( x = (x_{1}, x_{2}, \ldots, x_{n}) \) be another arbitrarily fixed point.

Put \( f(x) = d(x, x^i) = \sum_{k=1}^{n} (x_{k} - x_{k}^i)^2 \) the Euclidean distance from \( x^i \) to \( x \), let
\[
f(x) = \frac{1}{m} \sum_{i=1}^{m} f(x^i) = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{n} (x_{k} - x_{k}^i)^2.
\]

\( f(x) \) measures the distance of the objects of the distribution from \( x \). The factor \( \frac{1}{m} \) allows us to normalize the function with respect to the number of the points.

To find a vector \( x \) that is the nearest to the other vectors \( x^1, x^2, \ldots, x^m \) it is equivalent to minimize the function \( f(x) \).

It is well known that the local maxima or minima of \( f(x) \) are found among the solutions of the system
\[
\begin{align*}
\frac{\partial}{\partial x_{1}} f(x) = & \frac{2}{m} \sum_{i=1}^{m} (x_{1} - x_{1}^i) = 0 \quad \Rightarrow \quad x_{1} = \frac{1}{m} \sum_{i=1}^{m} x_{1}^i = \tau_1 \\
\frac{\partial}{\partial x_{2}} f(x) = & \frac{2}{m} \sum_{i=1}^{m} (x_{2} - x_{2}^i) = 0 \quad \Rightarrow \quad x_{2} = \frac{1}{m} \sum_{i=1}^{m} x_{2}^i = \tau_2 \\
& \ldots \\
\frac{\partial}{\partial x_{n}} f(x) = & \frac{2}{m} \sum_{i=1}^{m} (x_{n} - x_{n}^i) = 0 \quad \Rightarrow \quad x_{n} = \frac{1}{m} \sum_{i=1}^{m} x_{n}^i = \tau_n
\end{align*}
\]

The unique solution is the vector \( \bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n) \), that is the central point of the distribution and the study of the eigenvalues of the Hessian matrix of \( f(x) \)
\[
\begin{pmatrix}
\frac{\partial^2 f(x)}{\partial x_{1}^2} & \ldots & \frac{\partial^2 f(x)}{\partial x_{1} \partial x_{n}} \\
\vdots & \ddots & \vdots \\
\frac{\partial^2 f(x)}{\partial x_{n} \partial x_{1}} & \ldots & \frac{\partial^2 f(x)}{\partial x_{n}^2}
\end{pmatrix} =
\begin{pmatrix}
2 & 0 & \ldots & 0 \\
0 & 2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 2
\end{pmatrix}
\]

proves that \( \tau = (\tau_1, \tau_2, \ldots, \tau_n) \) is the minimum point of \( f(x) \).

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].

**The Function center Applied to Quantitative Data Matrices**

The central point is not necessarily an object within the distribution. On the basis of what demonstrated above, we define “the most central object” as the object minimizing the distances with all other members of the set. This means that the most central object is the closest to the central point, as will be shown with some numerical experiments in the following chapters of this article.

One of the algorithms developed (PDA) is based on the objects coordinates calculated on the basis of the Principal Coordinates Analysis (PCoA), in which, every coordinate of the central point is zero [18].
The *center* function was tested with a matrix of quantitative random positive data consisting of 25 objects and 10 descriptors (matrix a). The scattering of these objects according to the first four dimensions resulting from the PCoA is depicted in Fig. (1). The object distribution was quite asymmetrical, in fact the object 24 was far from all other objects when considering the first two dimensions (Fig. 1a). The central point of the distribution was quite close to objects 16, 2 and 10 in the first two dimensions plot (Fig. 1a), whereas it resulted near to objects 7 and 21 in the third and fourth dimensions (Fig. 1b). This simple case demonstrated the difficulty to define the most central object visually considering different projections of the PCoA coordinates. In this case there were few coordinates and few objects and a series of visual inspections to determine the most central object that could be considered, but increasing the numbers of objects and descriptors this operation would be quite impractical. Moreover, visual inspections do not allow for the definition of “how much” each single object is close to the center. The *center* function was applied to the a matrix obtaining the plots of Fig. (2) in which the most central object was the 19 and the most peripheral 24 (Fig. 2a). Introducing the central point with the PDA algorithm the object distribution did not change at all (Fig. 2b).

Another output of *center* is a series of four vectors indicating distances from the center. Namely, cl lists the distances from the most central object (algorithm MDA) and cln the same distances normalized according to *Formula 1*. Similarly, clp and clnp report figures of distances from the most central object, as calculated with the PDA algorithm. These data showed a good agreement among corresponding series of data (Table 1). The clnp series allowed to consider that the most central object is rather close to the center of the distribution, in fact object 19 was almost equidistant from the central point and from the second most central object (25).

The Function *center* Applied to Qualitative (Binary) Data Matrices

The same procedure reported in the previous chapter was repeated with a matrix of the same size (25 x 10) consisting of qualitative data (1/0). This data format is quite common in biology and is normally employed to indicate the presence (1) or the absence (0) of a specific character. Conceptually, this data format with two states is not different from any other data format with n states. DNA is a particular case in which four states (A, C, G, T) are used. The matrix to test the *center* function (matrix c) was produced with the *rmat* function from the ESTHER package [20]. This function generates a binary matrix of defined dimensions (number of objects and descriptors) in which the single figures have the same random probability of being “0” or “1”. The c matrix consisted of 25 strains and 10 descriptors. Two of such descriptors were manually modified assigning “0” to the first 12 objects and “1” to the others. This manipulation produced two partially distinct groups of objects as shown in Fig. (3a) plot where over 64% variability was displayed in the first two dimensions. This distinction, however, was not visible in the other two dimensions (Fig. 3b). This matrix has been designed to test the *center* function with a set of objects partially discernible in two subsets. This situation is mimic of many matrices obtained with taxonomic analyses in which some descriptors allow clear-cut discriminations, whereas the others do not.

The MDA algorithm of the *center* function produced a distribution of objects rather similar to that of Fig. (4a), indicating that neither the type of character nor the presence of two subsets could influence the distance distribution. The PDA algorithm generated a distribution in which roughly 50% of the whole distance range was covered by the distance between the center and the most central object (number 6), as shown in Fig. (4b). As already observed with the quantitative data, both MDA and PDA generated the same sequence of objects from the most central to the most peripheral.

The Choice of the Central Object Affects the Object Classification

The main aim of this article was to define multivariate objects in terms of distance from the most central object or from the distribution center, supposing that the use of...
acentric references could produce problems in a distance based classification. In order to check this hypothesis, we used the matrix $c$, which can be divided in two major subsets (1 to 12 and 13 to 24) plus one outgroup (object 25) (Fig. 5).

The two subsets were processed separately with the `center` function, finding that objects 2 and 14 were the most central of the first and second subset, respectively. Two different classifications were simulated with `classify`, using as a reference the most central and the outgroup object. When using object 2 (the most central) as reference, a classification was produced in which 12 objects of the first subset were clustered in four groups at distances spanning from 0 to 0.5. No cluster included members of the two subgroups (Fig. 5a). When the most peripheral object 25 was the reference, the distances of the first subset members spanned from 0.5 to 0.8 (Fig. 5b) and all the five groups in which the 12 objects of the first subset were clustered contained members of the other subset (Table 2).

**DISCUSSION**

All biological classifications can be considered as tentative theories on the organization of living being, and in this respect reflect the philosophical approach to the problem. Typology is not likely to give a good interpretation...
Table 1. Output of the center Function Applied to the Matrix

<table>
<thead>
<tr>
<th></th>
<th>MDA</th>
<th></th>
<th>PDA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>obj</td>
<td>cl</td>
<td>cln</td>
</tr>
<tr>
<td>-</td>
<td>nd</td>
<td>nd</td>
<td>cp</td>
</tr>
<tr>
<td>19</td>
<td>0.951</td>
<td>0.000</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>1.037</td>
<td>0.095</td>
<td>25</td>
</tr>
<tr>
<td>12</td>
<td>1.046</td>
<td>0.104</td>
<td>12</td>
</tr>
<tr>
<td>21</td>
<td>1.049</td>
<td>0.108</td>
<td>21</td>
</tr>
<tr>
<td>5</td>
<td>1.053</td>
<td>0.112</td>
<td>5</td>
</tr>
<tr>
<td>16</td>
<td>1.068</td>
<td>0.128</td>
<td>16</td>
</tr>
<tr>
<td>7</td>
<td>1.086</td>
<td>0.149</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td>1.109</td>
<td>0.173</td>
<td>2</td>
</tr>
<tr>
<td>13</td>
<td>1.110</td>
<td>0.174</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td>1.112</td>
<td>0.176</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>1.125</td>
<td>0.191</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>1.167</td>
<td>0.237</td>
<td>8</td>
</tr>
<tr>
<td>15</td>
<td>1.176</td>
<td>0.247</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>1.181</td>
<td>0.252</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>1.184</td>
<td>0.256</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>1.213</td>
<td>0.288</td>
<td>4</td>
</tr>
<tr>
<td>22</td>
<td>1.227</td>
<td>0.304</td>
<td>22</td>
</tr>
<tr>
<td>20</td>
<td>1.247</td>
<td>0.325</td>
<td>20</td>
</tr>
<tr>
<td>9</td>
<td>1.277</td>
<td>0.358</td>
<td>9</td>
</tr>
<tr>
<td>1</td>
<td>1.289</td>
<td>0.371</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>1.341</td>
<td>0.428</td>
<td>18</td>
</tr>
<tr>
<td>11</td>
<td>1.386</td>
<td>0.478</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>1.404</td>
<td>0.497</td>
<td>6</td>
</tr>
<tr>
<td>23</td>
<td>1.491</td>
<td>0.593</td>
<td>23</td>
</tr>
<tr>
<td>24</td>
<td>1.861</td>
<td>1.000</td>
<td>24</td>
</tr>
</tbody>
</table>

Legend. cp is the central point of the distribution. obj is the number of the matrix a objects. cl and cln are the actual and normalized distance values from the most central object. clp and clnp are the actual and normalized distance values from the central point.

Fig. (3). Dispersion of multivariate objects (matrix c) according to the first four coordinates. Fig. (3a) and (3b) display the scattering of the 25 objects in the first four dimensions. The cross indicates the center of the distribution space.
has been tested in two theoretical frameworks represented by objects described by real (quantitative) and binary (qualitative) data. The results presented have shown that the same algorithms can be applied to both areas. This fact allows to extend the approach to molecular markers as the DNA sequences, which share the same nature with binary characters, but differ only by the number of possible states (four instead than two).

The type object could be individuated with two different approaches: one that seeks the most central object (MDA) and another considering that the most representative object is that sitting close to the mass center, or central point (PDA). Both algorithms produced the same outputs in terms of ordinality of distance, i.e. both classified in the same way the objects according to the distance from the most central object. It is important to underline here that the most central object is a natural entity present in distribution of non identical objects with more than two members. In fact, in a set of two different objects, both will be equally distant from the center of the distribution and none can be considered more central than the other. The analysis of the binary matrix used to validate the script has shown a situation in which even the most central point is quite distant from the center.

The numerical experiments suggested that indeed the choice of a most central object as reference reduces the problems of misclassification and are likely to remove problems of extension of the set caused by the addition of new elements similar to the peripheral elements, but indeed rather different from the “type”. The validation of the centrality concept and script was carried out with simple datasets for reason of space and clarity of exposition. Even the matrix was divided in two rather similar subsets. These cases are rather realistic, although could not consider cases of subsets of different size, different variability, and at different distance from each other as considered in a previous theoretical paper [17]. Further work is therefore necessary along two research lines: deepening the theoretical aspects of this algorithm and the application of the centrality criterion to biological sets such as groups of species. Moreover, this approach can be likely applied as a classification tool in disciplines other than taxonomy and systematic.

**Fig. (4).** Distance plot generated by center for the c matrix. Fig. (4a) displays the distance of each object from the most central object, as obtained with the algorithm MDA. Fig. (4b) displays the distance of each objects from the central point according to the PDA algorithm.
**Table 2. Binary Matrix with Objects Clustered According to the PDA Algorithm**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Object</th>
<th>col1</th>
<th>col2</th>
<th>col3</th>
<th>col4</th>
<th>col5</th>
<th>col6</th>
<th>col7</th>
<th>col8</th>
<th>col9</th>
<th>col10</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>a</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>b</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>10</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>c</td>
<td>12</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>d</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>8</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>e</td>
<td>13</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>e</td>
<td>17</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>e</td>
<td>25</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>f</td>
<td>14</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>19</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>22</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>f</td>
<td>24</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>g</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>18</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>g</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>h</td>
<td>23</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>i</td>
<td>21</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Legend.** Boldface figures indicate characters state different from the corresponding character of cluster “a” members.
ACKNOWLEDGEMENTS

L.A. and L.R. were supported by a Grant of the Italian Ministry of University and Research.

CONFLICT OF INTEREST

None declared.

REFERENCES