Open Access

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

Livio Antonielli[1], Vincent Robert[2], Laura Corte[1], Luca Roscini[1], Rita Ceppitelli[3] and Gianluigi Cardinali[*,1]

[1]*Department of Applied Biology – Microbiology Division, University of Perugia, Italy*

[2]*CBS-KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands*

[3]*Department 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*

*Address correspondence to this author at the University of Perugia-Department of Applied Biology – Microbiology Division, Borgo 20 Giugno, 74, I – 06121 Perugia, Italy; Tel: +39 075 585 6478; Fax: +39 075 585 647; E-mail: gianlu@unipg.it

*"**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 "Centrality" 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].

   **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.
5. **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.

2. The distance matrix is transformed in an object of the class matrix (i.e. a square matrix).
3. The sum of values contained in each row (or column) is calculated.
4. The vector of the sum figures is sorted in ascending order.
5. The sum vector is normalized with *formula 1* to give index values ranging from 0, most central object, to 1, most peripheral object.

*Formula 1 NV=(N-min)/(max-min)*

where *NV* and *N* indicate the normalized and the input value respectively; *min* and *max* designate the minimum and maximum value of the sum vector.

Centrality of objects including the central point of the distribution is calculated according to an algorithm based on the Principal Coordinate Analysis (PCoA) [18], called "PCoA Distance Algorithm (PDA). PDA works as outlined below:

1. PDA calculates the distance matrix from the input *m* matrix.
2. The distance matrix is introduced in the PCoA algorithm, yielding the eigenvalues of each object.
3. The sum of the eigenvalues of each object is calculated.
4. The vector of the sum figures is sorted in ascending order.
5. The sum vector is normalized with *formula 1* to give index values ranging from 0, most central object, to 1, most peripheral object.

PCoA algorithm has been employed because it can process both qualitative and quantitative data.

The function *classify* has the following arguments:

1. **m.** As in *center.*
2. **ob.** This argument (object) calls the input of the object to use as center of the distribution.
3. **bin=F.** As in *center.*
4. **method=**"*Euclidean*" As in *center.*
5. **binmethod=2** As in *center.*

The function *classify* produces a plot reporting the distance of single objects from the object designated as central reference.

## RESULTS

### The Central Point is the Most Representative of the Distribution

The central point of a distribution is defined as the element whose coordinates are the arithmetic means of the objects' descriptors. Of course it is not necessarily an object of the distribution, but it is the most representative point of the system.

By applying simple mathematical tools, we will prove that the central point coincides with the vector that is the nearest to the other objects of the distribution.

Let $x^1, x^2, \ldots, x^m$ be m vectors in $\mathbb{R}^n$, $x^i = (x_1^i, x_2^i, \ldots, x_n^i)$, $i = 1, 2, \ldots, m$ and let $x = (x_1, x_2, \ldots, x_n)$ be another arbitrarily fixed point.

Put $f_i(x) = d(x, x^i) = \sum_{k=1}^{n} (x_k - x_k^i)^2$ the Euclidean distance from $x^i$ to $x$, let

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x) = \frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{n} (x_k - x_k^i)^2 .$$

$f(x)$ 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

$$\begin{cases} \dfrac{\partial}{\partial x_1} f(x) = \dfrac{2}{m} \sum_{i=1}^{m} (x_1 - x_1^i) = 0 \\ \dfrac{\partial}{\partial x_2} f(x) = \dfrac{2}{m} \sum_{i=1}^{m} (x_2 - x_2^i) = 0 \\ \quad \ldots \\ \dfrac{\partial}{\partial x_n} f(x) = \dfrac{2}{m} \sum_{i=1}^{m} (x_n - x_n^i) = 0 \end{cases} \Leftrightarrow \begin{cases} x_1 = \dfrac{1}{m} \sum_{i=1}^{m} x_1^i = \bar{x}_1 \\ x_2 = \dfrac{1}{m} \sum_{i=1}^{m} x_2^i = \bar{x}_2 \\ \quad \ldots \\ x_n = \dfrac{1}{m} \sum_{i=1}^{m} x_n^i = \bar{x}_n \end{cases}$$

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 $f(x)$

$$\begin{pmatrix} \dfrac{\partial^2 f(\bar{x})}{\partial x_1^2} & \cdots & \dfrac{\partial^2 f(\bar{x})}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f(\bar{x})}{\partial x_n \partial x_1} & \cdots & \dfrac{\partial^2 f(\bar{x})}{\partial x_n^2} \end{pmatrix} = \begin{pmatrix} 2 & 0 & \ldots & 0 \\ 0 & 2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & 2 \end{pmatrix}$$

proves that $\bar{x} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n)$ 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 *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 *ranmat* 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 discerniblein 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

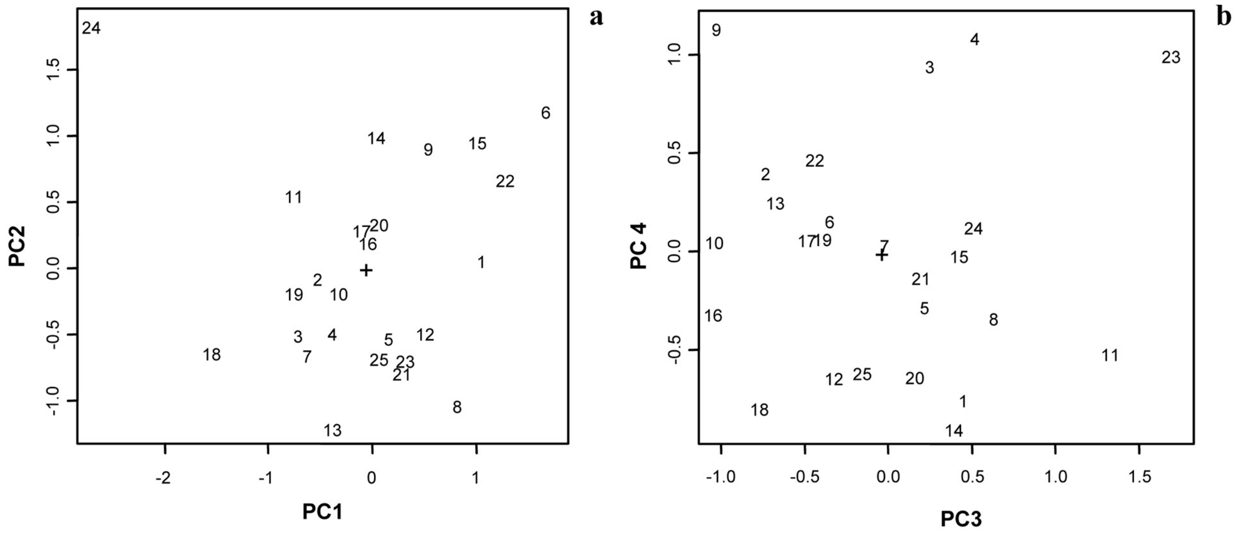
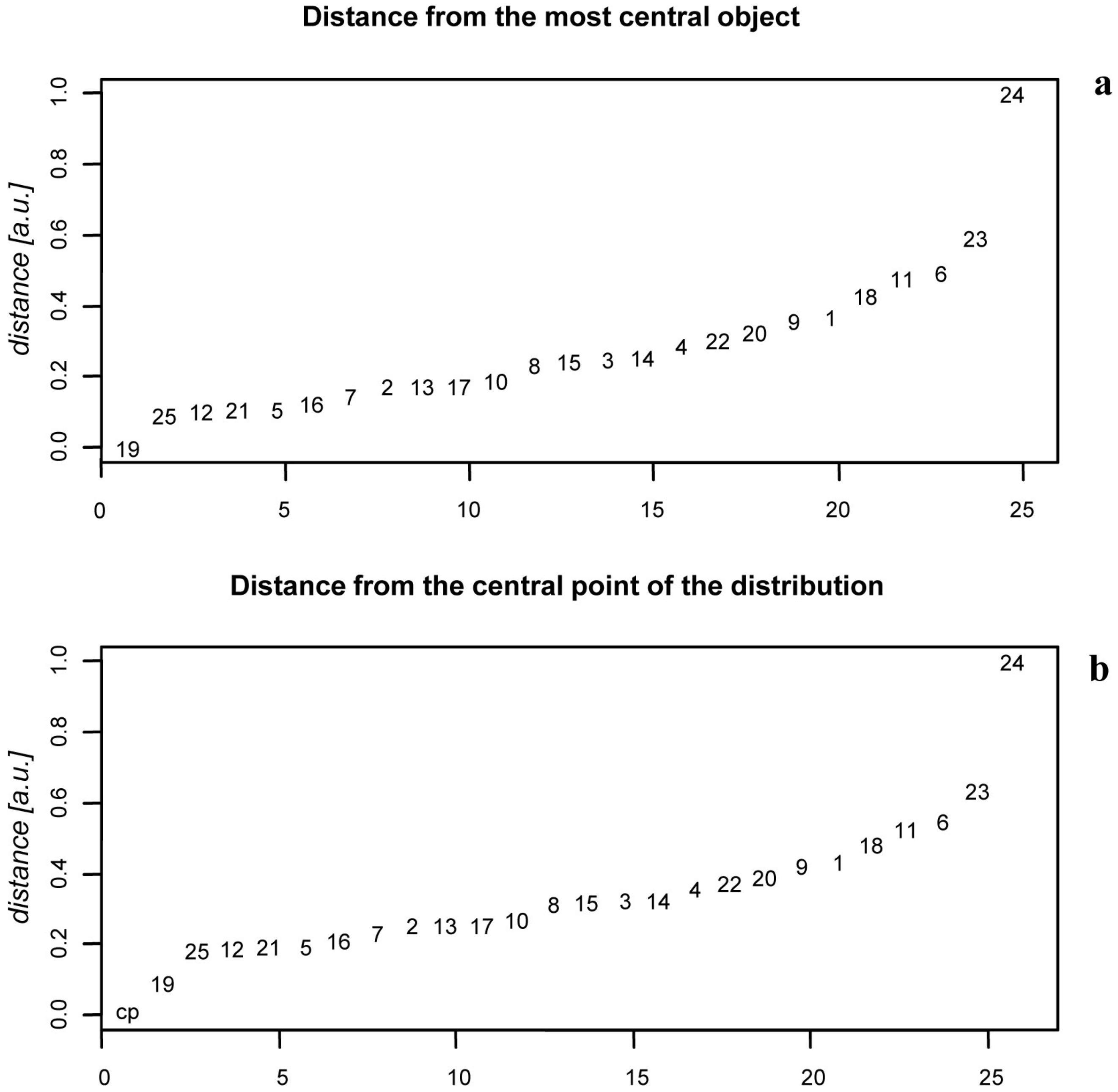


**Fig. (1).** Dispersion of multivariate objects (matrix a) according to the first four coordinates. Fig. (**1a**) and (**1b**) display the scattering of the 25 objects in the first four dimensions. The cross indicates the center of the distribution space.



Distance from the most central object

Distance from the central point of the distribution

**Fig. (2).** Distance plot generated by *center* for the **a** matrix. Fig. (**2a**) displays the distance of each object from the most central object, as obtained with the algorithm MDA. Fig. (**2b**) displays the distance of each objects from the central point according to the PDA algorithm.

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 a**

| MDA | | | PDA | | |
|---|---|---|---|---|---|
| **obj** | **cl** | **cln** | **obj** | **clp** | **clnp** |
| - | nd | nd | **cp** | 0.991 | 0.000 |
| **19** | 0.951 | 0.000 | **19** | 1.107 | 0.096 |
| **25** | 1.037 | 0.095 | **25** | 1.214 | 0.184 |
| **12** | 1.046 | 0.104 | **12** | 1.225 | 0.193 |
| **21** | 1.049 | 0.108 | **21** | 1.228 | 0.196 |
| **5** | 1.053 | 0.112 | **5** | 1.232 | 0.199 |
| **16** | 1.068 | 0.128 | **16** | 1.250 | 0.214 |
| **7** | 1.086 | 0.149 | **7** | 1.273 | 0.233 |
| **2** | 1.109 | 0.173 | **2** | 1.300 | 0.255 |
| **13** | 1.110 | 0.174 | **13** | 1.303 | 0.257 |
| **17** | 1.112 | 0.176 | **17** | 1.303 | 0.258 |
| **10** | 1.125 | 0.191 | **10** | 1.320 | 0.271 |
| **8** | 1.167 | 0.237 | **8** | 1.372 | 0.314 |
| **15** | 1.176 | 0.247 | **15** | 1.382 | 0.322 |
| **3** | 1.181 | 0.252 | **3** | 1.387 | 0.327 |
| **14** | 1.184 | 0.256 | **14** | 1.391 | 0.330 |
| **4** | 1.213 | 0.288 | **4** | 1.426 | 0.359 |
| **22** | 1.227 | 0.304 | **22** | 1.444 | 0.374 |
| **20** | 1.247 | 0.325 | **20** | 1.467 | 0.392 |
| **9** | 1.277 | 0.358 | **9** | 1.503 | 0.423 |
| **1** | 1.289 | 0.371 | **1** | 1.518 | 0.435 |
| **18** | 1.341 | 0.428 | **18** | 1.581 | 0.487 |
| **11** | 1.386 | 0.478 | **11** | 1.635 | 0.531 |
| **6** | 1.404 | 0.497 | **6** | 1.657 | 0.549 |
| **23** | 1.491 | 0.593 | **23** | 1.760 | 0.635 |
| **24** | 1.861 | 1.000 | **24** | 2.203 | 1.000 |

**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.

of nature and cannot probably be considered an acceptable approach to classification [1, 4]. However, typology can be regarded as a practical approach to the understanding and to the representation of natural structures [21] and is particularly necessary in microbiology, given the statements of the two codes of botanical and bacteriological nomenclature. Moreover, microbiology is still trying to overcome the major problem of collecting enough biodiversity to better understand the structures in the microbial world. It is obvious that with this limitation, and with the relative poorness of morphological traits, microbiological taxonomy needs some simplifications and hypotheses to work with. In other words, the current problem is that without enough knowledge on microbial biodiversity any theoretical assumptions would be hardly proven or disproven, but without a good theoretical framework it is difficult to define and handle the species, which are widely regarded as the units of biodiversity [22]. This sort of vicious circle must be broken to avoid a conceptual paralysis. Typology revisited as a practical tool could be regarded as good principle to use it to gain enough knowledge necessary to formulate better classification theories.

With this intention, this paper aimed to develop and test a new tool to streamline one of the major problems inherent to topology, i.e. the choice of the type. We could demonstrate with simple mathematical tools that in a set of objects defined by quantitative descriptors (i.e. in ℝ) the central point minimizes the distances with other members of the set, because its coordinates are the means of each descriptor. Similarly, the most central object is that closest to the central point and shares with the central point the property of minimizing the distances with the other objects. The same approach could not be taken with qualitative data (binary in our case), because it is impossible to calculate the mean of these data formats. However, distance matrices can be produced with whatever data type and the concept could be extended to the binary formats in the numeric experiments. As a matter of fact the simple R script presented in this work

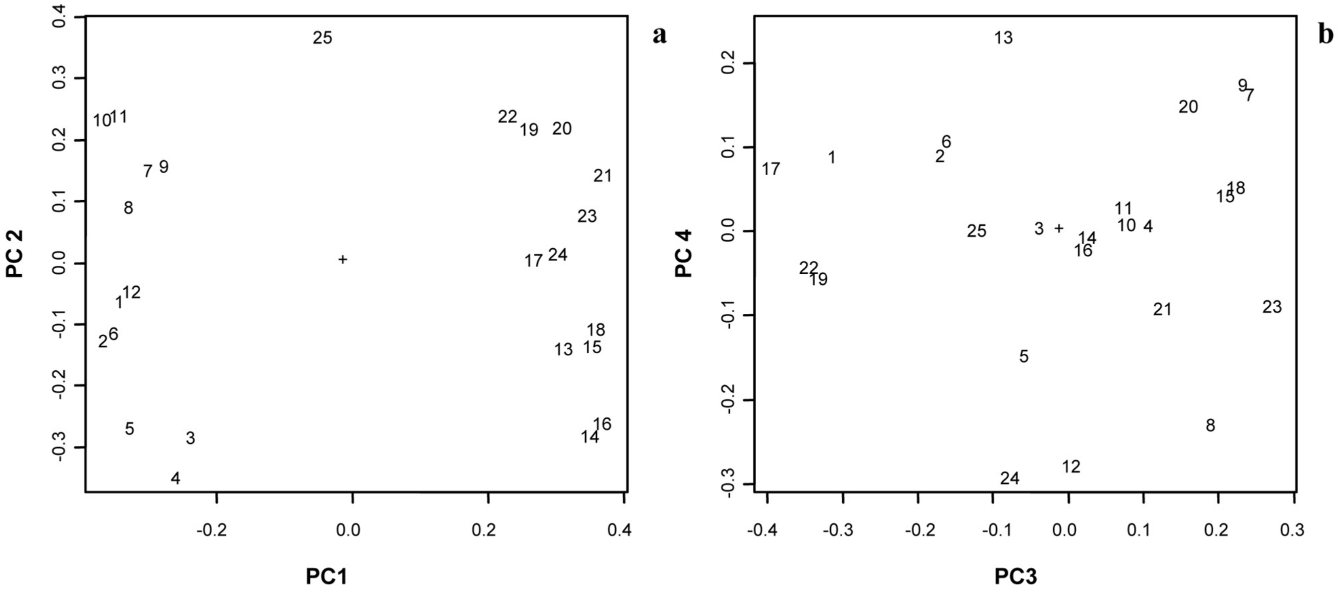


**Fig. (3).** Dispersion of multivariate objects (matrix c) according to the first four coordinates. Fig. (**3a**) and (**3b**) display the scattering of the 25 objects in the first four dimensions. The cross indicates the center of the distribution space.

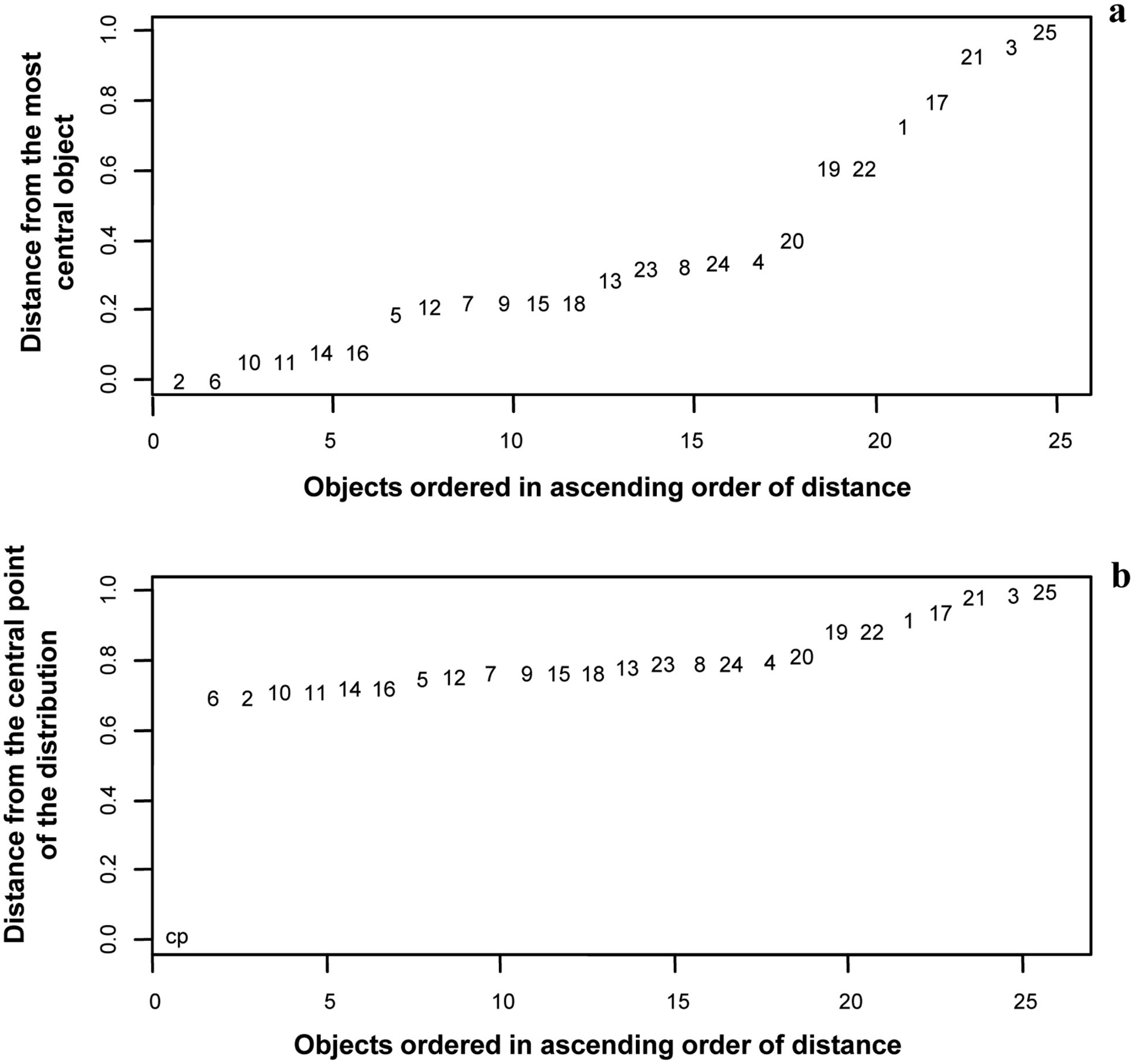


**Fig. (4).** Distance plot generated by *center* for the **c** matrix. Fig. (**4a**) displays the distance of each object from the most central object, as obtained with the algorithm MDA. Fig. (**4b**) displays the distance of each objects from the central point according to the PDA algorithm.

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 **c** 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.

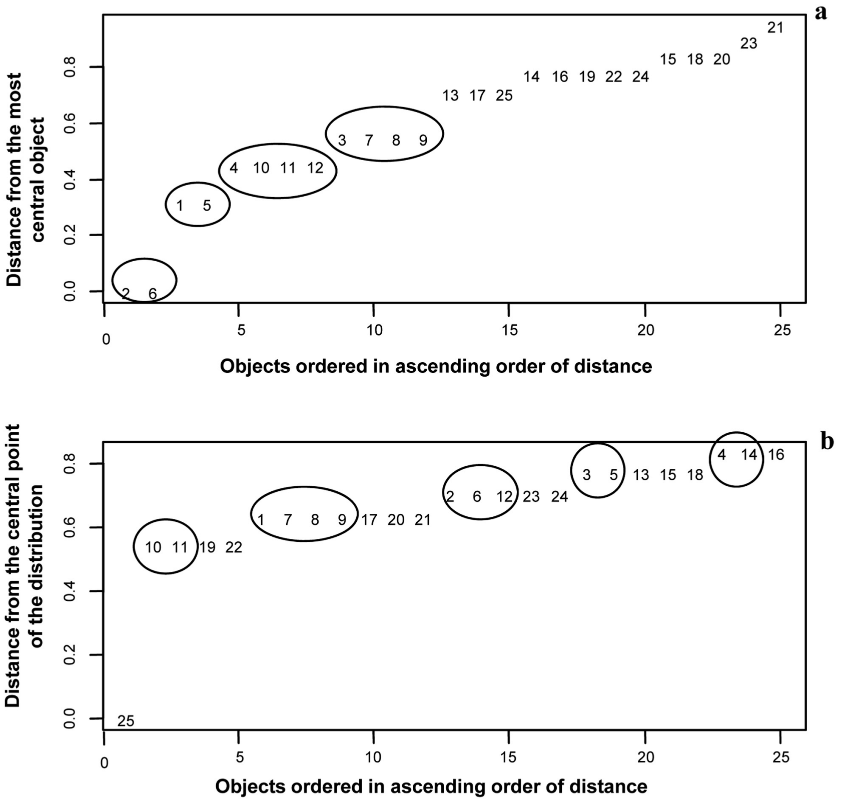


**Fig. (5).** Distance of objects calculated with the *classify* function. Fig. (**5a**) display the classification using the most central object. Fig. (**5b**) shows the classification using the most peripheral object as reference.

**Table 2. Binary Matrix with Objects Clustered According to the PDA Algorithm**

| Cluster | Object | col1 | col2 | col3 | col4 | col5 | col6 | col7 | col8 | col9 | col10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| a | 2 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| a | 6 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 |
| b | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | **1** | 1 | 0 |
| b | 5 | 1 | 0 | 1 | 1 | 0 | 0 | **0** | 0 | 1 | 0 |
| c | 4 | 1 | 0 | 1 | 1 | 0 | 0 | **0** | 0 | **0** | 0 |
| c | 10 | 1 | 0 | 1 | 1 | 0 | **1** | 1 | 0 | 1 | **1** |
| c | 11 | 1 | 0 | 1 | 1 | 0 | **1** | 1 | 0 | 1 | **1** |
| c | 12 | 1 | 0 | 1 | 1 | 0 | **1** | **0** | 0 | 1 | 0 |
| d | 3 | 1 | 0 | 1 | 1 | 0 | 0 | **0** | **1** | **0** | 0 |
| d | 7 | 1 | 0 | 1 | 1 | 0 | **1** | 1 | 0 | **0** | **1** |
| d | 8 | 1 | 0 | 1 | 1 | 0 | **1** | **0** | 0 | 1 | **1** |
| d | 9 | 1 | 0 | 1 | 1 | 0 | **1** | 1 | 0 | **0** | **1** |
| e | 13 | **0** | **1** | **0** | 1 | **1** | 0 | 1 | 0 | **0** | 0 |
| e | 17 | **0** | **1** | **0** | 1 | **1** | 0 | 1 | **1** | 1 | 0 |
| e | 25 | 1 | **1** | 1 | 1 | **1** | **1** | 1 | **1** | 1 | **1** |
| f | 14 | **0** | **1** | **0** | 1 | **1** | 0 | **0** | 0 | **0** | 0 |
| f | 16 | **0** | **1** | **0** | 1 | **1** | 0 | **0** | 0 | **0** | 0 |
| f | 19 | **0** | **1** | **0** | 1 | **1** | **1** | 1 | **1** | 1 | 0 |
| f | 22 | **0** | **1** | **0** | 1 | **1** | **1** | 1 | **1** | 1 | 0 |
| f | 24 | **0** | **1** | **0** | 1 | **1** | **1** | **0** | 0 | 1 | 0 |
| g | 15 | **0** | **1** | **0** | 1 | **1** | 0 | **0** | 0 | **0** | **1** |
| g | 18 | **0** | **1** | **0** | 1 | **1** | 0 | **0** | 0 | **0** | **1** |
| g | 20 | **0** | **1** | **0** | 1 | **1** | **1** | 1 | 0 | **0** | **1** |
| h | 23 | **0** | **1** | **0** | 1 | **1** | **1** | **0** | 0 | **0** | **1** |
| i | 21 | **0** | **1** | **0** | 1 | **1** | **1** | **0** | **1** | **0** | **1** |

**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

[1] R. J. Owen, "Bacterial taxonomics: finding the wood through the phylogenetic trees," *Methods Mol. Biol.,* vol. 266, pp. 353-383, 2004.

[2] E. Mayr, "Biological classification: toward a synthesis of opposing methodologies," *Science,* vol. 214, pp. 510-516, 1981.

[3] O. Rieppel, "Re-writing Popper's philosophy of science for systematics," *Hist. Philos. Life Sci.,* vol. 30, pp. 293-316, 2008.

[4] W. F. Doolittle, "Eradicating typological thinking in prokaryotic systematics and evolution," *Cold Spring Harb. Symp. Quant. Biol.,* vol. 74, pp. 197-204, 2009.

[5] A. Guilliermond, *Les Levures*. Paris: Octave Doin, 1912.

[6] J. Lodder, "Die Anaskosporogenen Hefen, Erste Hälfte," *Verh. K. Acad. Wet. Afd. Naturk Tweede Reeks. II,* vol. 32, pp. 1-301, 1934.

[7] J. Lodder, and N. J. W. Kreger-van Rij, *The Yeasts, A Taxonomic Study*, 1 ed. Amsterdam: North-Holland, 1952.

[8] C. P. Kurtzman, and P. A. Blanz, "Ribosomal RNA/DNA sequence comparisons for assessing phylogenetic relationships," in *The Yeasts, A Taxonomic Study*, 4th ed., C. P. Kurtzman, and J. W. Fell, Eds. Amsterdam: Elsevier, pp. 69-74, 1998.

[9] C. P. Kurtzman and C. J. Robnett, "Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences," *Antonie van Leeuwenhoek,* vol. 73, pp. 331-371, 1998.

[10] R. Scheda and D. Yarrow, "The instability of physiological properties used as criteria in the taxonomy of yeasts," *Arch. Mikrobiol.,* vol. 55, pp. 209-225, 1966.

[11] G. Cardinali, and A. Martini, "Electrophoretic karyotypes of authentic strains of the *sensu stricto* group of the genus *Saccharomyces*," *Int. J. Syst. Bacteriol.,* vol. 44, pp. 791-797, 1994.

[12] R. J. Pankhurst, "A computer program for generating diagnostic keys," *Comput. J.,* vol. 12, pp. 145-151 1970.

[13] N. Saitou, and M. Nei, "The neighbor joining method, a new method for reconstructing phylogenetic trees," *Mol. Biol. Evol.,* vol. 4, pp. 406-425, 1987.

[14] J. P. van der Walt, "The yeast *Kluyveromyces africanus* nov. spec. and its phylogenetic significance," *Antonie van Leeuwenhoek,* vol. 22, pp. 321-326, 1956.

[15] E. Stackebrandt, and B. M. Goebel, "Taxonomic note: a place for DNA-DNA reassociation and 16s rRNA sequence analysis in the present species definition in bacteriology" *Int. J. Syst. Bacteriol.,* vol. 44, pp. 846-849, 1994.

[16] G. E. Fox, J. D. Wisotzkey, and P. J. Jurtshuk, "How close is close: 16s rRNA sequence identity may not be sufficient to guarantee species identity " *Int. J. Syst. Bacteriol.,* vol. 42, pp. 166-170, 1992.

[17] G. Cardinali, "Measure of species variability for a microbial taxonomy based on the relative resemblance," *Riv. Biol.,* vol. 96, pp. 271-291, 2003.

[18] P. Legendre and L. Legendre, *Numerical Ecology*, 2nd. English ed. Amsterdam: Elsevier Science B.V., 1998.

[19] S. Dray and A. B. Dufour, "The ade4 package: implementing the duality diagram for ecologists." *J. Stat. Softw.,* vol. 22, pp. 1-20, 2007.

[20] G. Cardinali, L. Antonielli, P. Rellini, and F. Fatichenti, "ESTHER: A "R Package" Implementing a novel approach to bidimen-sional display of multidimensional binary data," *Open Appl. Info.,* vol. 1, pp. 20-27, 2007.

[21] A. C. Love, "Typology reconfigured: from the metaphysics of essentialism to the epistemology of representation," *Acta. Biotheor.,* vol. 57, pp. 51-75, 2009.

[22] M. F. Claridge, H. A. Dawah, and M. R. Wilson, "*Species, The Units of Biodiversity*," London: Chapman & Hall, 1997.

Received: April 29, 2010 Revised: October 10, 2010 Accepted: March 27, 2011