ESTHER: A "R Package" Implementing a Novel Approach to Bidimensional Display of Multidimensional Binary Data

G. Cardinali*, L. Antonielli, P. Rellini and F. Fatichenti

DBVBAZ - Microbiology, University of Perugia, Italy

Abstract: The R package ESTHER implements two novel algorithms designed to dispose in a reduced space multidimensional objects defined by binary descriptors. The approach is to assign discrete and fix positions to all possible combinations obtained with the descriptors employed. One of the two algorithms, called clock, position objects on a circle at regular intervals, whereas the other, star, maintains the angular position as in clock, but defines the distance of the object from the center of the circle proportionally to the number of descriptor in state"1". Comparisons with Principal Coordinate Analysis (PCoA) showed that the three methods perform differently according to the number of objects and descriptors and to the distance method employed to carry out the PCoA. The algorithm clock produced the best object clustering in a validation carried out with a matrix generated by molecular fingerprint of yeast isolates.

INTRODUCTION

Several biological experiments produce outputs as binary data; a non exhaustive list includes DNA fingerprinting (banding patterns), DNA microarrays, many morphological traits, Mendelian phenotypes, assimilation patterns, some ecological parameters. Each object (species, individual, ecological locale, etc.) being individuated by several binary descriptors is a point in a multidimensional space with as many dimensions as the number of descriptors employed. The large number of descriptors, and thus of dimensions, imposes to develop techniques to represent the objects as points in a bidimensional space. Among the several approach proposed, one of the best performing methods for binary data is that of Principal Coordinate Analysis, from now on referred to as PCoA (Legendre 1998) [1]. Principal Component Analysis is not considered a method of choice for multivariate binary matrices because it is especially designed for continuous data and uses only Euclidean distances (Gower JOC 1986) [2], whereas PCoA can use any resemblance metrics among the several proposed in the past decades.

The approaches to display multidimensional objects in reduced space suffer necessarily of some limitations. One is the loss of distance relationships among objects in the bidimensional space (Legendre 1998) [1], which increases with the number of descriptors employed. This problem is traditionally visualized by calculating the Shepard diagram (Legendre 1998) [1] obtained by plotting the objects in a diagram reporting in the abscissa the distances calculated in the original matrix, whereas in the ordinate the corresponding distances calculated among the points in the bidimensional space are reported. A good preservation of the original distances has been obtained when the points fall close to the y=x diagonal. Another problem of displaying objects in reduced space is the difficulty to visualize the contribution of single descriptors, especially when several variables are brought into play.

*Address correspondence to this author at DBVBAZ – Microbiologia, Università di Perugia, Via Borgo 20 giugno 74, I 06132 Perugia, Italy; Tel: + 39 075 585 6478; Fax: +39 075 585 6470; E-mail: gianlu@unipg.it

An additional problem, which is at the origin of the algorithm proposed in this paper, is the impossibility to show in the PCoA graph the proportion of the descriptors in use. In fact, n binary variables can describe a maximum of 2n possible objects, although the number of biological objects studied is normally well below 2n. In normal conditions, all objects (excluding the repetitions) are represented by one of the 2n combinations obtained with the n descriptors in use, meaning that only a few combinations are actually present in the matrix. The relationships among these few combinations and their positioning within the complex of all possible combinations are further aims of the present algorithms, named ESTHER after the ancient Persian word, meaning star, for some similarity between the point scattering and the stylized depiction of the star.

METHODS

General Presentation of the Algorithms Included in ESTHER

ESTHER includes seven functions, two of them (clock and star) are designed to position in a binary space object described by several descriptors, both of them include an option to evaluate the quality of the point scattering in the bidimensional graph. The other four functions are auxiliary; one, import, is designed to directly import binary matrices, three to produce binary matrices: fullmat generates matrices with all 2n combinations, partmat yields only a defined portion of all combinations and finally ranmat generates random combinations with the desired number of objects and descriptors. The function shepard returns a double plot with the Shepard diagrams of ESTHER and PCoA.

Definition of the *Clock* Algorithm

The clock algorithm has been so named because object points will be disposed as the hours on a clock, for this reason, the object position will be indicated with time notation such as h0, h3,h6 etc.

The rationale of point positioning in clock is that each point is a combination of descriptor and occupies a defined position in a circle with unitary radius. The algorithm to assign positions (Fig. **1a**) will be described for simplicity in the

1874-1363/07

case of 8 objects defined by 3 descriptors: 000, 001, 010, 011, 100, 101, 110, 111. Firstly the circle is subdivided in equal portion, each of $2\pi/2n$ amplitude, in this case $\pi/8$ (45). The first object (000) is placed at h0, then the second (001) at h1:30, the third (010) at h3 and the fourth at h4:30. The second half of the points is located considering that each point is opposite to its complementary. This means that h6 will be the place of 111, 110 will dwell at h7:30 and finally 101 and 100 will be placed at h9 and h 10:30, respectively.

The algorithm generating the clock disposition works as described below.

1. The size of the imported matrix is calculate. The number of objects (rows) is called *N* and that of descriptors (columns) *D*.

2. Binary strings describing the objects are considered as a binary code and transformed in decimal values.

3. A correction factor (*cf*) is calculated to put complementary objects in opposite position, according to the *Formula 1*:

   Formula 1 $cf = floor(bin/(2^{D-1})$

   where *bin* is the decimal value calculated from the binary string as mentioned in the previous step and *floor* is the "R" function that returns the largest integers not greater than the argument. The correcting factor will be 0 for values ranging from 0 to 2^{D-1} (the first half in a ordered series) and 1 for the other.

4. The position (*pos*) of all objects is calculated, considering that the spacing between objects are identical, with *Formula 2*:

   Formula 2 $pos=(bin*|cf-1|) +((0.5*2^{D}-1-bin) *cf)$

   The first term of the *Formula 2* annihilates for all points (*bin*) from 2^{D-1} to 2^{D} (the second half of an ordered series) and is equal to 1 for the objects from 0 to 2^{D-1}-1. Vice versa, the second part of the formula annihilates for objects with *bin* ranging from 0 to 2^{D-1}-1. This means that the first half of the objects are positioned according to the first part of the formula, whereas the other points depend on the second element of *Formula 2*. Altogether this means that the position of all objects belonging to the first half of an ordered series is the *bin* value itself. For the other objects, the formula starts from h6 ($0.5*2^{D}$) and than moves anticlockwise for as many position as the *bin* value. This implies that the first value of the second half in an ordered series, say "5" when D=3 and the objects =8, moves 5 position anticlockwise from h6 (π). Since the amplitude between contiguous positions is $2\pi/8 = \pi/4$ (45°) this movement implies a position at $\pi - 5\pi/4 = -1/4\ \pi$ (315°= -45°) which is in fact the first position in anticlockwise order from h0.

5. The position values are transformed in the corresponding angular value (α)

6. The coordinates of each object(*i*) are calculates as

7. ***Formula 3*** $x_i = \sin \alpha_i$; $y_i = \sin \alpha_i$.

Definition of the *Star* Algorithm

The star algorithm is very similar to clock with the major difference that, whereas in clock the distance between each point and the center is always 1, in star this distance is proportional to the number of "1" in the binary string defining the object. This means that the angular position of the points is identical in both algorithms, but objects defined by several "0" will be placed close to the center, whereas those with many"1" will have a distance close to 1. This means that the two limit objects defined by only "0" and "1" will be placed in the center (coordinates 0,0) and at the maximum distance of 1 unit from the center (coordinates 0,-1), respectively.

Analysis of Correlation

Within both clock and star a correlation is implemented to show the relationship between point scattering and the original distances among objects. This function (compare) returns two correlations: one between the distance matrix (dm) obtained from the binary matrix and the distances among points scattered according to star or clock (dg), the other between dm and the first two coordinates of the points calculated according to the PCoA (dc). Correlation can be carried out according to the Pearson, Spearman or Kendall methods (Legendre 1998) [1]. Considering that linearity cannot be always out with the dist.binary function from the ADE4 package.

Auxiliary Algorithms

In order to avoid the hassle of writing the command string to import a text file with the matrix, ESTHER includes an import function requiring only the name of the matrix to import and to set header = TRUE (default is FALSE) in case the first row reports the descriptor names. The function assumes the first column reports the object names, otherwise set obj.names=0 (default=1).

Binary matrices can be generated with three auxiliary functions fullmat, partmat and ranmat.

The first produces "complete" binary matrices i.e. matrices with all 2n combinations obtained from n descriptors. partmat requires a p parameter indicating how many times the 2n number of objects is halved, therefore returning 2n-1 objects. Finally ranmat returns a binary matrix composed by the number of objects and descriptors specified in the command string.

The Shepard diagram is obtained by plotting on the abscissa the distances as calculated from the original matrix and on the ordinate those obtained from the scattered points on a bidimensional plot; optimal points distributions cluster around the y=x diagonal (Legendre 1998) [1]. The function shepard produces a plot with two diagrams overlaid one for ESTHER (clock or star) with points represented by black crosses and one for the PCoA with red squares.

RESULTS

Features of *Clock* and *Star*

In binary matrices the most distant object pairs are those with no common element, hereinafter referred to as "complementary", such as 1100 and 0011 (Hamming BSTJ 1950) [3]. The clock algorithm places complementary objects in

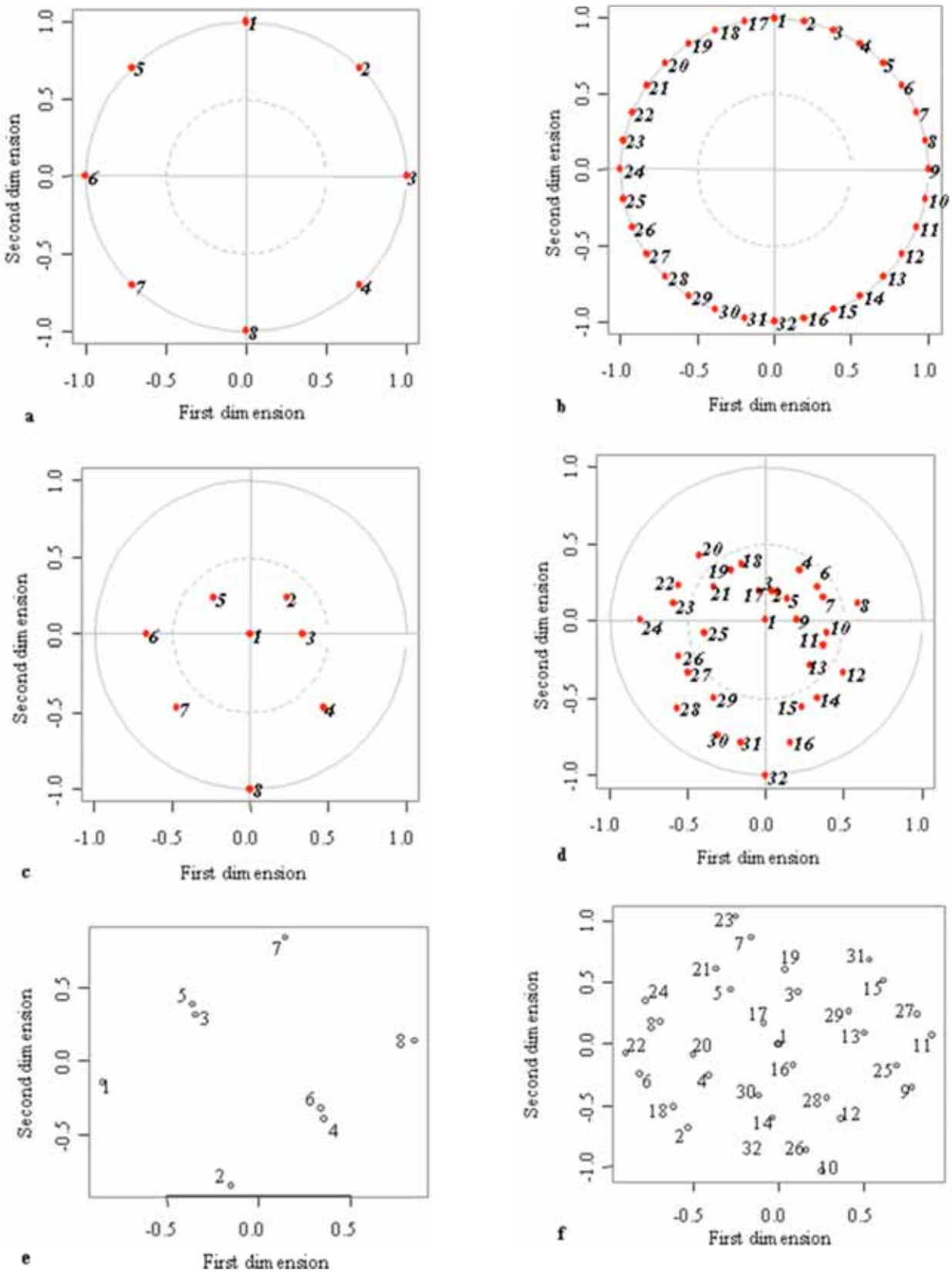

Fig. (1) Examples of object distribution with *clock*, *star* and PCoA. Panels (**a**) and (**b**) *clock* object scattering, panels (**c**) and (d) *star* object disposition, panel (**e**) and (**f**) PCoA diagram. Left panels refer to a 3 descriptors (8 object) complete matrix, those on the right to a 5 descriptor array.

diametrically opposite positions at the distance of two radii, thus maintaining the original relationship between such objects. This feature is maintained in star, where the maximum distance is one radius.

Another characteristic of clock is that the position of the points in the circle gives some indications:

a. The upper part of the circle includes objects with more descriptors in the state "1" than "0", vice versa for the lower part.

b. The points around h3 and h9 have approximately as many "1" as "0", although the disposition of the character states is different (Figs. **1a**,**2a**).

c. Invariant descriptors, especially in the first positions of the strings, produce asymmetric dispositions (Fig. **2**).

d. Since all possible combinations are evenly spaced, the lack of some of them is easily visualized, allowing for an easy detection of possible lacks of some of them. This concept can be readily seen in Fig. (**1**). The clock disposition (Fig. **1a**) poses in contiguous positions six out of objects only objects 2 and 3 and objects 6 and 7 are neighbor, although separated by two differences. The PCoA disposition (Fig. **1c**) produces two stringent aggregations, objects 3 and 5 and object 4 and 6, whose elements are separated by two differences. On the contrary similar objects such 1 and 2 or 7 and 8 are more spaced.

The algorithm star is quite similar to clock, but the fact that the distance between the center and the object point depends on the number of "1" produces some differences such as.

PCoA has been chosen as the algorithm to compare with *clock* and *star*, because it allows the choice of the distance metrics, as often required in biology and molecular biology studies.

Differential Efficiencies with Complete Matrices

The efficiency of *clock*, *star* and PCoA with complete matrices of different size was tested considering the cases from 2 up to 12 descriptors, i.e. from 4 to 4096 objects, with four different *dist.binary* methods (Table **2**). The disposition of objects in complete matrices can be seen in Table **1**, showing the matrix as generated by *fullmat*, with 3 descriptors.

In the four cases studied, *star* correlations with the original matrix are low with less than four descriptors, then grow rapidly up to the levels of *clock* and PCoA when distance methods 2 and 4 (called symmetric for the presence of both a and *d* in the numerator) are employed, whereas remain around zero with methods 1 and 3. On the contrary, PCoA correlations are high with less than three or four descriptors, than stabilize on the levels of *clock*, with the peculiarity that symmetric distances confer a little advantage to *clock*, which shows a slightly better correlation than PCoA throughout the whole range of the analysis. In general, all the three algorithms of disposition in reduced space seem to take advantage from symmetric distances. Whether the occurrence of double "1" (*a*) should be considered as that of double "0" (*d*) in binary distances is matter of discussion and of the specificities in the various situations analyzed, however the effect of the choice in the quality of the point representation should be considered.

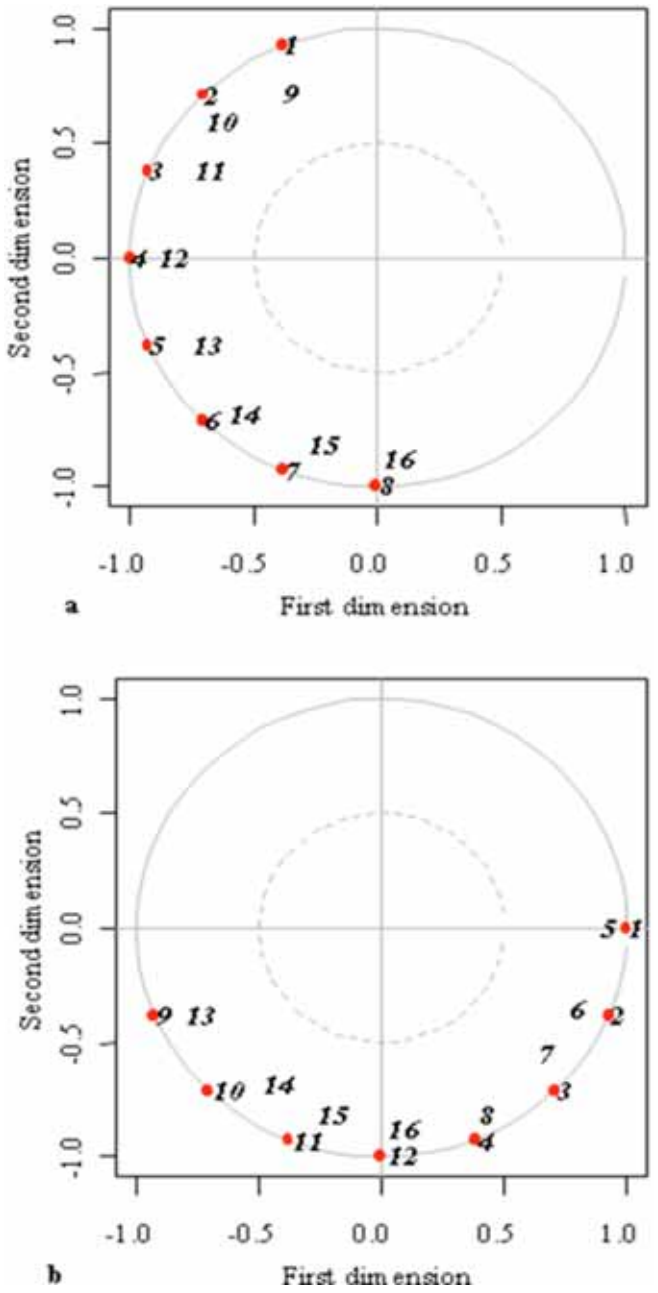

Fig. (2). Asymmetric distribution of points in *clock*.

Objects refer to a 4 descriptor (16 objects) complete matrix. Panel (**a**): the asymmetry is due to invariance of the first descriptor. Panel (**b**): asymmetry generated by invariance of the second descriptor.

a. Objects with more descriptors in the "0" state are posed in the upper part of the circle, close to the center. This feature produces a point disposition moved toward the lower part of the circle.

b. The points with approximately as many "1" as "0"are placed in the equatorial part on the inner circle of radius 0.5 (Figs. **1b**,**2b**).

c. The lack of combinations is much less visible in the star than in the clock disposition.

Table 1. Example of a Complete Matrix Generated by *fullmat*

	D1	D2	D3
obj 1	0	0	0
obj 2	0	0	1
obj 3	0	1	0
obj 4	0	1	1
obj 5	1	0	0
obj 6	1	0	1
obj 7	1	1	0
obj 8	1	1	1

Since complete matrices are not frequently encountered in biostatistics, we considered the hypothesis of matrices with a random distribution of "0" and "1". Another reason to undertake this analysis is that the number of objects and descriptors in complete matrices are tied by the relation $objects = 2^{descriptors}$, although in general cases these two parameters vary rather independently. In order to carry out these comparisons, the function *ranmat* was used to produce matrices of the desired number of objects and descriptors with random attribution of the descriptors state ("0" or "1"), hereinafter referred to as random matrices. Each combination of objects and descriptors was tested with three independently generated random matrices. These simulations were carried out considering three cases with 10, 20 and 30 descriptors (Fig. **3a-c**), and objects ranging from 10 to 500. Results of the simulations show that, whatever the number of descriptors, when the objects are relatively few PCoA performs better than *star* or *clock*, although this gap reduces with the increase of the objects considered (Fig. **3**). Moreover, in the case of 10 descriptors (Fig. **3a**) PCoA performance decrease more rapidly than with 20 and 30 descriptors (Fig. **3b**,**c**). On the contrary, the correlation between *star* or *clock* point scattering and the distances among objects in the multidimensional space remain rather invariant with the increase of the objects considered. All three systems appear to perform better with less objects, in fact the correlations tend to values around 0.40 with 10 descriptors and close to 0.30 and to 0.25 with 20 and 30 descriptors. Altogether these observations lead to the conclusion that the rate between objects and descriptors strongly influences the performance of the three methods compared. When this rate is high (several objects and relatively few descriptors) *star* and particularly *clock* perform better than PCoA, the reverse when the number of objects is relatively higher than that of descriptors. The simulations confirm the results of the complete matrices, which showed better performances of *star* and *clock* over PCoA with the increase of descriptors and therefore with the quadratic increase of objects. In all cases, the correlation between bidimensional and multidimensional distances decreases rather rapidly with the increase of objects and descriptors, confirming that point scattering in reduced space may not retain much of the original relationships among objects, suggesting to choose the better performing method according to different combinations of objects and descriptors. There are in fact cases in which relatively few descriptors (normally immunological or molecular markers) are used to define thousands of objects such as in epidemiological studies (Incardona MJ 2007) [4] (Lawson JCM 1999) [5] (Elliot BJS 1979) [6]. Another point to consider is the degree of independence between descriptors. Although this point has not been investigated specifically, ESTHER algorithms seem to perform better when the independence is high.

On the contrary, genomic or proteomic analyses often yield the opposite situation with thousands of descriptors and relatively few objects (Figueroa PICSBC 2003) [7].

Table 2. Distance Methods Implemented by the *dist.binary* Function of ADE4 and Used in this Study

Method	Similarity Formula**	Notes
method=1	*a/(a+b+c+d)* ***	Jaccard index, in Legendre* S7
method=2	*(a+d)/(a+b+c+d)*	Sockal and Michener index, in Legendre* S1
method=3	*a/(a+2b+2c)*	Soakal and Sneath index, in Legendre * S10
method=4	*(a+d)/(a+2b+2c+d)*	Rogers and Tanimoto index

*Legendre is the cited book by Legendre and Legendre, pag 275.
**Distances (d) are calculated from similarities (s) with the formula $d= (1-s)^{0.5}$.
****a* is the number of descriptors in state "1" in both objects, *d* is the opposite case (both "0"), *c* and *d* represent the number of cases in which one object has descriptor "1" and the other "0". For more details refer to Legendre and Legendre pag. 254.

A Real-World Validation

In order to test the proposed methods in a real situation, we considered a binary matrix with 13 descriptors obtained applying a previously published classification method (Cardinali Bioinformatics 2003) [8] to RAPD profiles of 26 yeast isolates (Table **3**).

Results show that the three methods produce significantly different point scattering. Using the method 2 of dist.binary, PCoA correlation between bidimensional and multidimensional distances was 0.773, whereas star and clock correlations were 0.689 and 0.462, respectively. These figures indicate that there is an overall retention of objects relationships with PCoA and star, whereas clock correlates significantly less. However, a closer analysis of the object positioning is important to highlight the types of distribution produced and the amount of information immediately visible in each graph. The two least correlated objects are strains 415 and 420 (Table **3**) with a distance of 0.96. Although all three methods position these two strains far away, clock poses the two objects at a distance of 1.996, very close to the maximum of 2, whereas PCoA spaces them out at a distance of 0.63, much smaller than the maximum 0.799 found in the distance matrix; star performs in a intermediate way with a distance of 1.07 and a maximum 1.26.

PCoA places close to the lower left corner the strains 415, 409, 411, 551 and 552 (Fig. **5c**) although these objects are not particularly similar (Table **3**), in fact their binary profile include respectively 3,6,6 and 8 "0" out of 13 descriptors. As a matter of fact, star (Fig. **5b**) poses strain 409 close to the 415, and the isolates at larger distances, although the 411 results more distant than the 551. Similarly, clock position the isolate 551 far from the 415, although the 552 (with 6 differences out of 13) is very close.

Table 3. Binary Banding Patterns of 26 Yeast Isolates Characterized with the RAPD Primer m13

Strains	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13
373	0	1	1	1	0	1	0	1	0	0	0	0	1
369	0	1	0	1	0	1	0	1	0	0	0	0	0
387	1	1	1	1	0	0	0	1	0	0	0	1	0
409	1	1	1	1	1	0	0	1	0	1	1	1	1
411	0	0	1	1	0	1	0	0	0	1	1	1	1
415	1	1	1	1	1	1	1	1	1	1	1	1	1
420	0	0	0	0	0	0	0	1	0	0	0	0	0
423	0	1	0	1	0	1	0	1	0	0	0	0	0
445	0	1	1	0	0	0	0	1	0	0	0	0	0
454	0	0	1	0	0	1	0	1	0	0	0	0	0
546	1	0	0	0	0	0	0	0	0	0	0	0	1
547	1	1	0	0	0	0	0	0	0	0	0	0	1
549	1	0	0	0	0	0	0	0	0	0	0	0	1
550	1	1	0	0	0	0	0	0	0	0	0	0	1
551	0	0	1	1	1	0	0	0	0	1	0	1	0
552	0	1	1	1	1	0	0	0	0	1	1	1	0
553	1	0	0	0	0	0	0	0	0	0	0	0	1
554	0	1	0	1	1	0	0	1	0	0	0	0	0
555	0	1	0	1	1	0	0	1	0	0	0	0	0
557	0	0	1	1	0	0	0	1	0	0	0	1	0
558	1	1	1	1	0	0	0	1	0	0	0	1	0
560	1	0	0	0	0	0	0	0	0	0	0	0	0
561	0	0	1	1	0	0	0	0	0	1	0	0	0
562	0	0	1	1	0	0	0	0	0	1	0	0	0
953	0	1	0	1	0	1	0	1	0	0	0	0	0
319	0	1	0	1	1	0	0	1	0	0	0	0	0

1. Boldface accession numbers indicate peculiar strains discussed with special emphasis in the text.
2. Descriptors (B1 to B13) size: B1 2000 bp, B2 1500 bp, B3 1200 bp, B4 1000 bp, B5 900 bp, B6 800 bp, B7 700 bp, B8 600 bp, B9 500 bp, B10 400 bp, B11 300 bp, B12 200 bp, B13 100 bp.

Another interesting cluster of strains is that including the two identical objects 561 and 562 both characterized by 10 out of 13 "0". Whereas clock and star place them far away from the isolate 415, the PCoA distance is rather little.

Beyond the comparisons regarding the correlations, an important issue of clock and star is that of providing further information on the nature of the strains. For instance, strains 420, 546, 549, 553 and 560, all characterized by a large number of "0" and then correctly posed around h12 by clock and at h12 close to the center by star. The overall distribution of points in star is within the internal circle suggesting that most of the strains have less than half characters in the state "0", in fact the original matrix includes only 114 "1" out of 338 data (26 objects * 13 descriptors). The distribution with

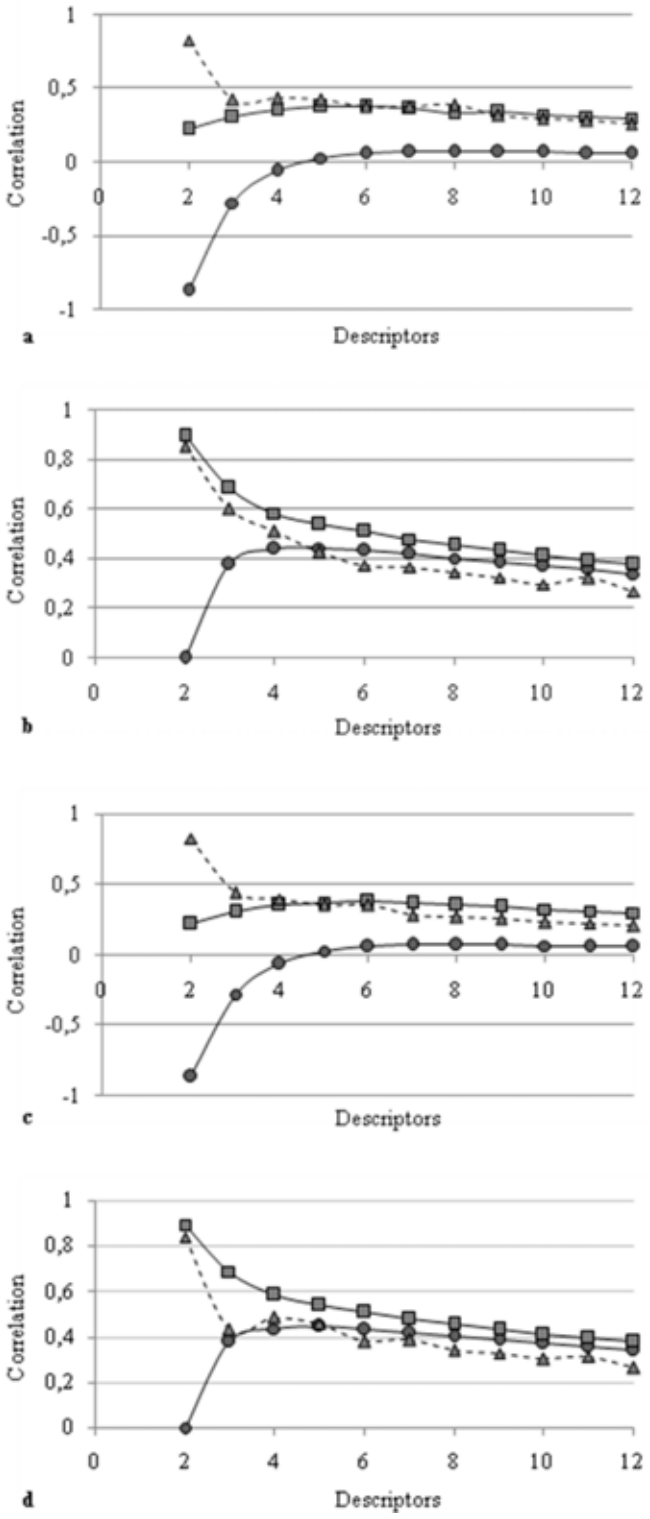

Fig. (3). Correlation between multidimensional and bidimensional distances. Correlation (Spearman) was carried out with complete matrices generated by *fullmat* with the indicated number of descriptors. Triangles indicate PCoA correlations, squares and circles correlations obtained with *clock* and *star* respectively. The four different panels were obtained by using four different methods of the ADE4 *dist.binary* function: panel (**a**) method1, panel (**b**) method 2, panel (**c**) method 3, panel (**d**) method 4. Methods 1 and 3 are asymmetric, 2 and 4 are symmetric.

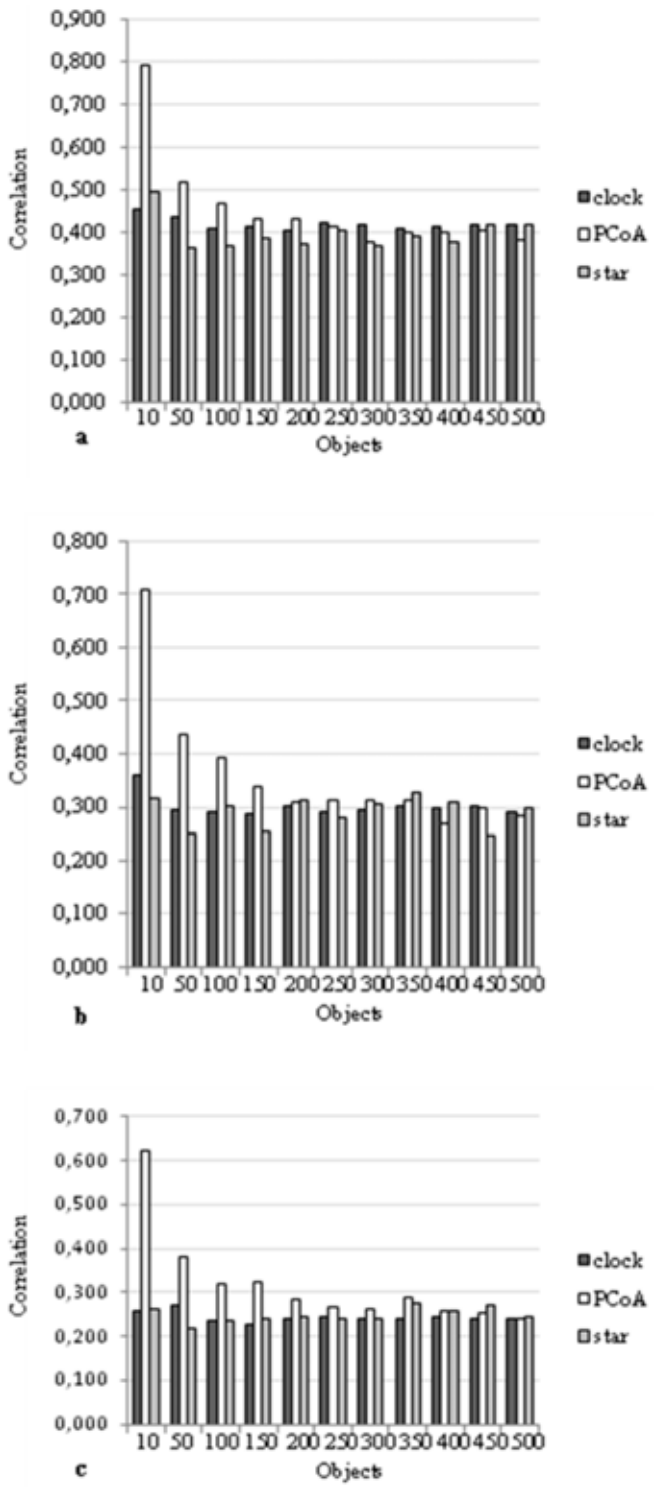

Fig. (4). Performance of the three methods with randomly generated binary matrices. The descriptor number vary in the three simulations: 10, 20 and 30 in panel (**a**), (**b**) and (**c**), respectively. Objects are reported in abscissa, Spearman correlation on the ordinate.

clock shows five clusters, respectively around h0, h2, h4, h6 and h9. Apart the group including two identical isolates (550 and 547), the group at h 6 shows 3 invariant descriptors, whereas the number of descriptors in state"1" ranges from 6 to 13. The group at h4 has 9n invariant descriptors, whereas all object except one have four"1". The cluster at h2 with 5 invariant variables has from 3 to 7 "1" per object. Finally the cluster at h 12 has obviously few "1" (1 or 2) and 10n invariant descriptors. In accordance with these observations the two clusters at h12 and h4 are the tightest, whereas that at h6 shows more variability.

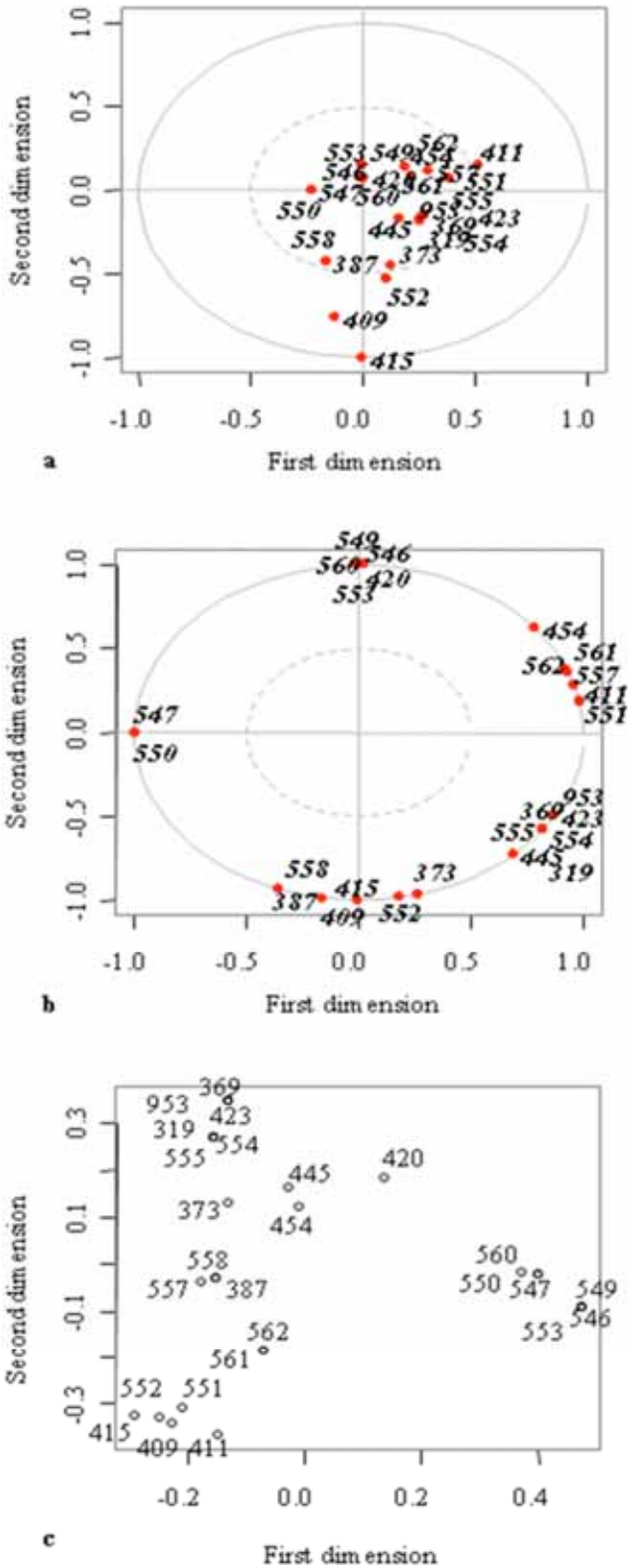

Fig. (5). Distribution of 26 yeast strains RAPD patterns according to the three methods. Data of the original matrix are reported in Table **3**. Distance method used = 4. Panels: (**a**) *star*, (**b**) *clock* and (**c**) PCoA object disposition.

Although ESTHER works better with many objects, these considerations suggest that star is rather good in positioning points according to their original relationships, but shows much less efficiently object clusters. Conversely, clock is more efficient in the clustering and less in positioning, although several points occupy positions in accordance with the original relationships among objects.

CONCLUSIONS

The R package ESTHER has been designed to implement a novel approach to ordination in reduced space of objects described with binary descriptors. The two algorithms, *clock* and *star*, perform differently both in terms of retention of the original relationships among objects and of overall topology of the point scattering. In comparison to a well tested procedure as PCoA, the two proposed algorithms perform better with several objects and relatively few descriptors. The ability of clock to give an immediate perception of the object clustering seems to be supported by the validation presented; moreover both ESTHER algorithms performed better than PCoA in giving an immediate series of information such as the number of "1" and "0" of the objects and the continuum between groups of objects. None of these algorithms has been designed as an alternative to PCoA, but rather as additional systems to figure out in a bidimensional space the multidimensional relationships among objects. Additional comparisons with other approaches, such as Correspondence Analysis (Benzecri 1992) [9] should be undertaken in future also to take into consideration the impact of differential correlations between descriptors. Further developments are under study, such as the extension to quaternary data, and therefore to DNA sequences, or to other discrete state systems used in biostatistics and bioinformatics.

REFERENCES

[1] P. Legendre and L. Legendre, *Numerical ecology*. Amsterdam: Elsevier., 1998.

[2] J. C. Gower and P. Legendre, "Metric and Euclidean Properties of Dissimilarity Coefficients", *Journal of Classification*., vol. 35, pp. 5-48. 1986.

[3] R. W. Hamming, "Error-detecting and error-correcting codes", *Bell System Technical Journal*., 29(2), pp. 147-160. 1950.

[4] S. Incardona, S. Vong, L. Chiv, *et al.*, "Large-scale malaria survey in Cambodia:novel insights on species distribution and risk factors". *Malar J*., vol 6, pp. 37. 2007.

[5] A. J. Lawson, J. M. Logan, G. L. O'neill, M. Desai, J. Stanley, "Large-scale survey of Campylobacter species in human gastroenteritis by PCR and PCR-enzyme-linked immunosorbent assay", *J Clin Microbiol*., 37(12), pp. 3860-4. 1999.

[6] M.S. Elliot and J.H. Louw, "*A 10-year survey of large bowel carcinoma at Groote Schuur Hospital with particular reference to patients under 30 years of age*. *Br J Surg*., 66(9), pp. 621-4. 1979.

[7] A. Figueroa, J. Borneman, T. Jiang, "Clustering binary fingerprint vectors with missing values for DNA array data analysis", *J Comput Biol*., 11(5), pp. 887-901. 2004.

[8] G. Cardinali, F. Maraziti, S. Selvi, "Electrophoretic data classification for phylogenetics and biostatistics", *Bioinformatics*., 19(16), pp. 2163-5. 2003.

[9] J.P. Benzecri, *Correspondence Analysis Handbook*. New York: Marcel Dekker, 1992.

Received: August 30, 2007 Revised: September 20, 2007 Accepted: September 27, 2007