Open Access

Searching for Related Descriptors Among Different Datasets: A New Strategy Implemented by the R Package "Dadi"

Livio Antonielli[1], Vincent Robert[2], Laura Corte[1], Luca Roscini[1], Ambra Bagnetti[1], Fabrizio Fatichenti[1] and Gianluigi Cardinali[*,1]

[1] *University of Perugia - DBA - Microbiology, Via Borgo 20 Giugno 74, 06121 Perugia, Italy*

[2] *Centraalbureau voor Schimmelcultures, Uppsalalaan 8, 3584 CT Utrecht, The Netherlands*

Abstract: *Background*: The increasing number of techniques introduced to describe organisms and taxa produce multivariate datasets, often composed of relatively independent descriptors. Handling several descriptors can be laborious and often unnecessary when their information is not congruent to that of other datasets used in the same study. On the other hand, different levels of correlation between single descriptors to a whole dataset may suggest useful scientific hints. The DADI (Distance-based Analysis for (optimal) Descriptor Identification) algorithm is proposed to allow a rapid and complete analysis among descriptors coming from two different datasets with the same number of objects. DADI was employed to select FTIR (Fourier Transform Infrared Spectroscopy) spectral wavelengths according to their correlation with the 26S rDNA sequences of strains belonging to a yeast genus.

Results: This procedure allowed to define a set of optimal wavelengths with an overall increase of the correlation between FTIR and 26S data.

Conclusions: DADI can identify the FTIR wavenumbers best fitting to the chosen reference defining the descriptors to be used in FTIR and possibly in other metabolomic analyses.

Keywords: Dataset, correlation, software, method, statistics, yeast.

BACKGROUND

Complex datasets, consisting of several descriptors, as DNA banding patterns, microarray intensity values, DNA sequencing or quantitative data from spectral analyses obtained with metabolomic techniques [1-5], are used to describe organisms and taxa in great detail. In general, all modern high through-output strategies provide increasingly more complex sets of data in shorter times, posing the problem of interpreting the biological meaning and the statistical significance of the dataset as a whole and of its single descriptors, when they can be handled independently.

The study of the relations among different single descriptors of the same dataset, also referred to as "R mode analysis", is normally carried out with association coefficients, among which the various correlation indexes are the most popular [6, 7]. The same indexes can be used to describe the correlation between a whole dataset and single descriptors (or groups of descriptors) of another dataset.

Biological studies focused on different approaches and techniques to characterize a given group of organisms are typical examples of the situation studied in this article.

A typical, but not exhaustive, example of this problem is represented by the analysis of FTIR spectra in order to extract the wave numbers (descriptors), whose information fit better with other taxonomic descriptions of the same microbial strains. This study has been traditionally carried out by empirical attempts or on the basis of case by case considerations.

Aim of this work is to develop an algorithm able to identify the most congruent descriptors of one matrix to a reference dataset, regardless of their data format.

MATERIALS AND METHODOLOGY

Implementation

The difference in data format and descriptor numbers of the datasets obtained from different experiments impedes a direct comparison of the single descriptors, however, distance matrices produced with a whatever number of descriptors will always have the same dimension allowing the comparison.

A basic assumption of the proposed algorithm is that one of the two datasets (M', hereinafter referred to as reference matrix) is taken as a whole, in order to individuate the single descriptors of the M" matrix (referred to as test matrix) whose signals fits better with the among objects distances of the M' matrix. Which of the two matrices should be considered the reference should be carried out on a case by case basis, considering that in many instances the role of the two matrices is reversible.

Once the reference matrix has been designed, the algorithm acts as follows:

1. A distance matrix (D') is calculated from M'.

[*] Address correspondence to this author at the University of Perugia - DBA - Microbiology, Via Borgo 20 Giugno 74, 06121 Perugia, Italy; Tel: 075 5856458; Fax: 075 5856470; E-mail: gianlu@unipg.it

2. A distance matrix (D1) is calculated using the first descriptor of the M” matrix.
3. The correlation value C1 between D’ and D1 is calculated.

Steps 2. And 3. are reiterated for all descriptors of the M” matrix.

All the Ci values are plotted against the series of descriptors, obtaining a graph like that showed in Fig. (**2**). In order to build a matrix with only the M” descriptors more coherent to M’ a threshold value (T) of correlation can be defined such that all descriptors of the new optimized matrix (M*) satisfy the condition Ci≥T.

This algorithm has been written in R language (http://cran.r-project.org/, 2010) and named DADI as acronym of Distance-based Analysis for optimal Descriptor Identification. The package allows for a number of choices to match different experimental and analytical situations, as elucidated in the brief list below and as depicted in Fig. (**1**).

1. **Importation Functions.** Input data can be imported as rectangular matrices, distance matrices or DNA alignment
2. **Distance Functions.** Distances can be calculated with the *dist* function (according to the six methods available in “R”, see Table **1b**) or with correlation based distances calculated as described below, where (T) is the transpose of the rectangular matrix and *corr* stands for correlation, producing a correlation square matrix among objects:

$$dcor1= 1-((1+corr(T))/2)$$

$$dcor2= 1-|corr(T)|$$

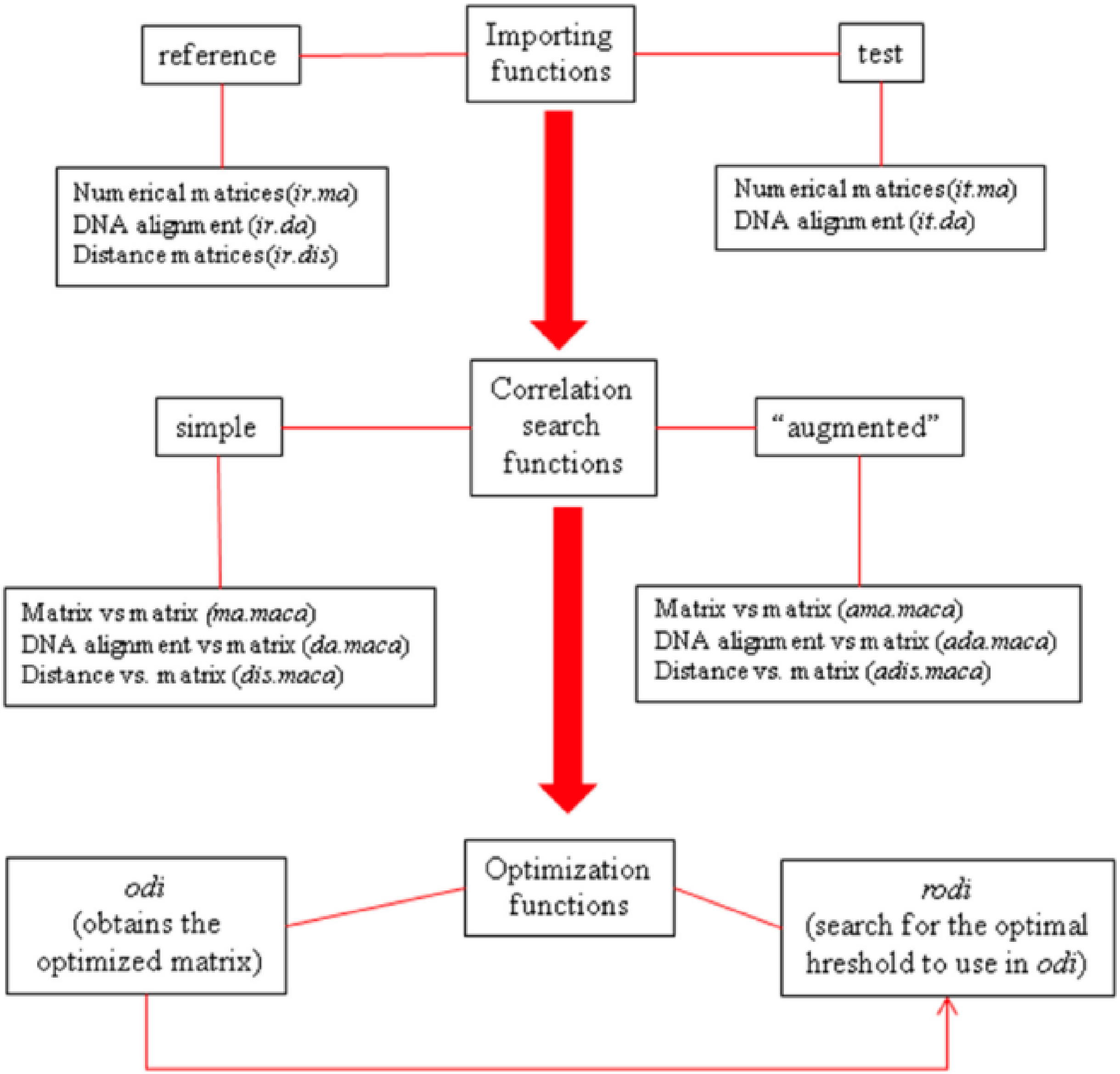
Importing functions
reference
test
Numerical matrices (ir.ma)
DNA alignment (ir.da)
Distance matrices (ir.dis)
Numerical matrices (it.ma)
DNA alignment (it.da)
Correlation search functions
simple
“augmented”
Matrix vs matrix (ma.maca)
DNA alignment vs matrix (da.maca)
Distance vs. matrix (dis.maca)
Matrix vs matrix (ama.maca)
DNA alignment vs matrix (ada.maca)
Distance vs. matrix (adis.maca)
Optimization functions
odi
(obtains the optimized matrix)
rodi
(search for the optimal hreshold to use in odi)

Fig. (1). Flow chart of the main functions implemented in DADI.

1. Functions are reported in italic.

2. In the description of correlation functions the term on the left in the reference, that in the right is the test. E.g. DNA alignment *vs* matrix means that the DNA alignment serves as reference and the matrix as test.

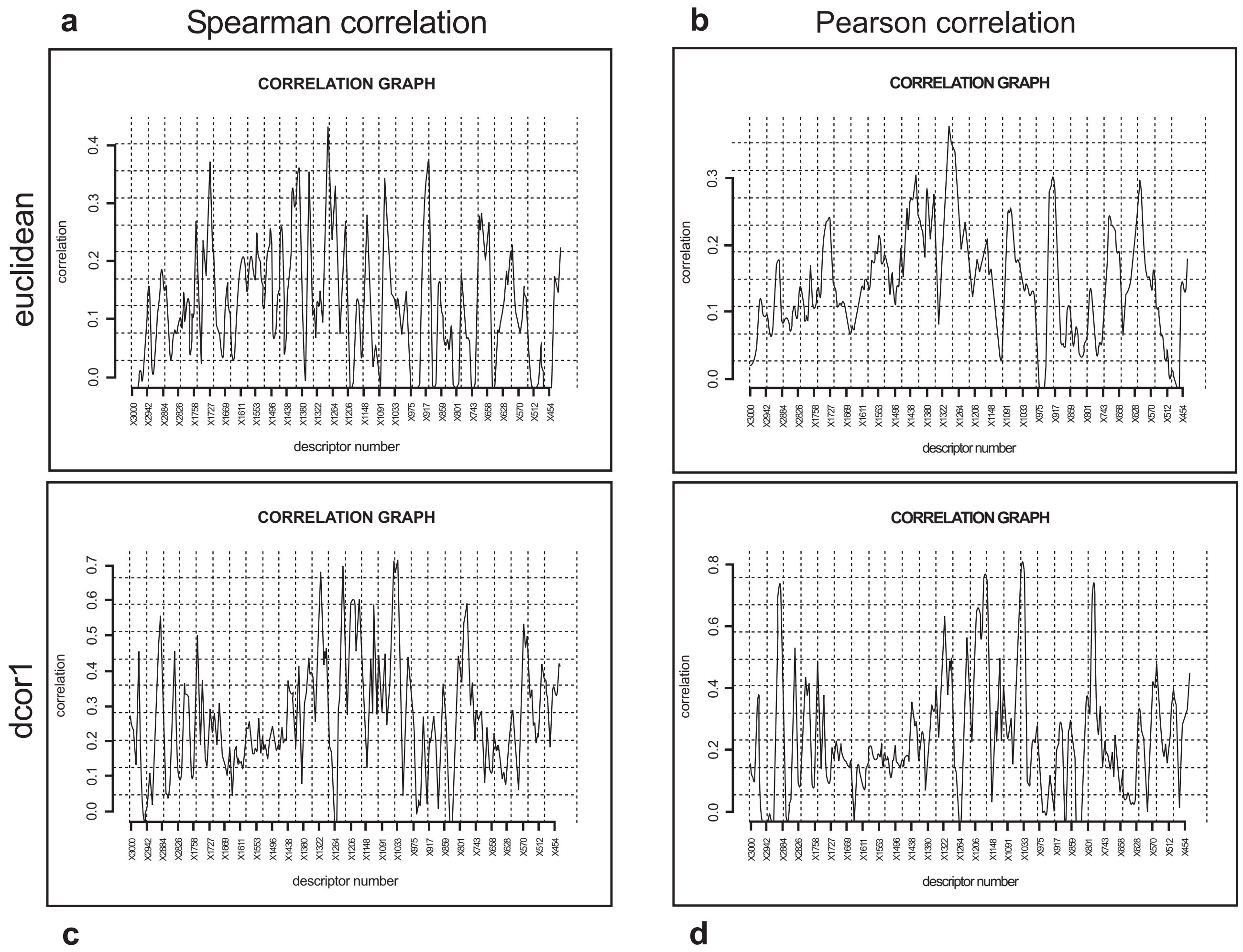

Fig. (2). Comparison of the effects caused by different distance and correlation algorithms (*Debaryomyces* species).

Plots in panel **a** and **b** were obtained with the Euclidean distance algorithm, **c** and **d** with the *dcor1* function described in the material and methods section. In panels **a** and **c** the correlation was calculated with the Spearman algorithm, in **b** and **d** with the Pearson's. Other parameters were: function *ada. maca*, grid=30, sf=20.

dcor3= 1-(corr(T)) for all corr(T)≥0; if corr(T)<0 then dcor3=1

Distances from DNA alignment are calculated according to the *dist.dna* function from the APE R - package (http://ape.mpl.ird.fr/, 2010/)

Table 1a. Nomenclature of the Available Algorithms According to their Fields of Application

Reference Input	Test Input (Numerical Matrix)	
	Simple Algorithms	Sliding Windows Algorithms
Matrix	*ma.maca*	*ama.maca*
Distance	*dis.maca*	*adis.maca*
DNA Alignment	*da.maca*	*ada.maca*

3. **Correlation Searching Functions.** These functions calculate the correlation among single or grouped descriptors of the test matrix *vs* the reference matrix. A summary of the results and a plot (Fig. **2**) are returned. Three basic correlation searching functions are available as shown in Table **1a**: *ma.maca* (a reference matrix *vs* a test matrix), *da.maca* (a DNA alignment *vs* a test matrix) and *dis.maca* (a distance matrix *vs* a test matrix). In these functions the whole range of descriptors is subdivided in equal portions of *sf* (sf stands for sampling frequency) size, in order to group together all descriptors in each *sf* range. The value *sf* can range from 1 to the number of descriptors of the test matrix. The number of spectral portions (*SP*) sampled is then *SP= n/sf* where *n* indicates the number of descriptors of the test matrix.

A different sampling of descriptors is available in the "extended functions ", distinguished by an "a" in front of the name of the corresponding basic functions, thus producing *ama.maca, adis.maca* and *ada.maca* (Table **1a**). In these functions a "sliding-window" sampling is used: the first group of descriptors range from 1 to *sf*, the second from 2 to *sf*+1 and so on. With the extended functions, the number of portion sampled does not decrease dramatically with the increase of *sf* and in general *PS= n+1-sf.*

With all the above described functions the correlation can be carried out according to the Pearson, Spearman and Kendal coefficients, already available with the *cor* function included in the base distribution of "R".

The *sort=T* command in the correlation functions produces a correlation graph with correlation data sorted in ascending order.

4. **Optimization Functions.** Aim of the two optimization functions is to produce an optimized test matrix obtained by eliminating all descriptors of the original test matrix that do not correlate well with the reference matrix. The *odi* (optimal descriptors identification) function requires the analyst to choose *a priori* a minimum value of correlation above which the reference matrix will be retained in the optimal matrix. This function, in fact, takes all descriptors whose correlation with the reference matrix is not lower than the threshold chosen by the analyst and merges them in a "optimal matrix". The function, *rodi* (reiterated *odi*) does not require any predefined threshold, but computes *odi* for several values of threshold included in a range chosen by the analyst. The operator is required to input the minimum and maximum values of the range and the interval between two successive threshold values. This function calculates for each threshold the corresponding correlation between the optimal matrix and the reference input data. In addition, this function calculates the gain in correlation, i.e. the increase in correlation obtained with each single optimal matrix in comparison to the original reference matrix. Gains in correlation values are shown in the ordinate of a graph reporting the thresholds on the abscissa.

5. **Cross Correlation.** The *dcc* (descriptors cross-correlation) function is available only when both the reference and test data are represented by matrices. It carries out a cross-correlation of all descriptors of the two rectangular matrices, returning a list of descriptor pairs whose correlation exceeds the chosen threshold and a pseudo heat-map. As a graphic output, *dcc* produces a pseudo- heatmap in which the correlation between descriptors of the two rectangular matrices is shown with different colors.

RESULTS AND DISCUSSION

Experimental Rationale and Taxonomic Models Used in the Validation of DADI

In order to evaluate the functionality of the proposed algorithms, we considered an increasingly interesting situation in modern microbial taxonomy as the classification or characterization of taxa by means of totally different approaches and descriptors. One such case is represented by the sequencing of the D1/D2 domain of the 26S rDNA and the FTIR analysis of whole cells. The molecular approach is largely accepted since its first proposal as a tool in yeast taxonomy [8] and has the advantage of an extended and expanding collection of data contained in GenBank (http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html, 2010) or in similar databases. The decreasing costs of automatic DNA sequencing is another advantage of this largely adopted technique. On the other hand, FTIR has been shown to be able to differentiate between bacterial [9-11], and fungal taxa [12-15], although public spectral libraries are not available yet. The main advantages of the FTIR approach are its low cost in terms of consumables, the relative rapidity of the procedure and - most importantly - the capability of giving a complete picture of the cellular metabolome. A typical case in taxonomic paper dealing with FTIR is a laborious search of spectral regions whose data correlate with the accepted taxonomy of the group under study. This complex work is time consuming and is not necessarily able to pinpoint all optimal descriptors (in this case wavenumbers) with the necessary resolution. The basic rationale of DADI is to define a set of taxonomically well identified strains and to obtain both D1/D2 26S rDNA sequences and FTIR spectra from each of them. The D1/D2 domain alignment has been chosen as reference, since it represents a widely accepted taxonomic classification. The complex of the FTIR data has been arranged in a matrix representing the test input data. The alignment dimensions and FTIR matrix have in common the same objects (the strains in this case) but are constituted by very different numbers of descriptors (respectively nucleotides and wavenumbers in the alignment and in the FTIR matrix). The triangular distance matrices obtained from each data set will share the same dimension, in fact both will be $(N^2-N)/2$ where N is the number of strains. These triangular distance matrices, having the same dimensions, can be subject to Mantel analysis, obtaining an overall value of correlation and a statistical significance value indicating the probability that the calculated correlation can be obtained by chance.

The validation has been carried out with three independent and different sets of taxa: one is represented by members of the yeast genus *Debaryomyces*, the second is constituted by the species of the genus *Kazachstania*. Finally, the third set is composed by species of a group of three yeast genera: *Saccharomyces*, *Zygosaccharomyces* and *Lachancea* (Table **2**).

Searching for FTIR Descriptors Correlating with DNA Alignments

There is a wide consensus in using a specific DNA sequence of the D1/D2 domain of the 26S rDNA gene as marker at taxonomic level [8]. As a matter of fact, most of the yeast identification is now carried out using this system which takes advantage of large free databases such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/, 2010). DADI has been employed to define the correlation of the 26S D1/D2 rDNA alignments with the single FTIR descriptors (wave-numbers) produced by the strains of the genus *Debaryomyces*, using the *ada.maca* function with the

Table 1b. Synopsis of the Methods Employable in the Distance Algorithms

Function	Distal	Method	Model	Cormethod
explanation	Distance algorithm	Distance method when distal=dist	DNA Alignment distance	Method of correltion
Options	dist	Euclidean	K80	Pearson
	dcor1	Manhattan	Raw	Spearman
	dcor2	Minkowski	JC69	Kendall
	dcor3	Canberra	K81	
			F84	
			BH87	
			T92	
			TN93	
			GG95	
			Lodgment	
			paralin	

following settings: K80 DNA distance model, FTIR distance method calculated with the Euclidean or *dcor1*algorithm, the correlation method was according to the Spearman or the Pearson algorithm. The Spearman algorithm does not assume the linear relation between the two series of data and is therefore preferable if the analyst does not want to take this assumption or has no evidence of linearity. The Spearman correlation used (Fig. **2a**, **c**) gave outputs similar to the default Pearson correlation (Fig. **2b**, **d**) although the correlations, were somehow lower. The distance algorithm had a much stronger impact, in fact the distance based on the correlation between data (*dcor1*) produced much higher figures of correlation between the FTIR and the DNA data, than when Euclidean distances were used. This can be explained by the fact that the *dcor1* algorithm calculates the changes in the spectral curve trend rather than the absolute point-by point distance as the Euclidean distance. Interestingly, all the four combinations of Fig. (**2**) confirm the high correlation of the wavenumbers in the 1300-1200 cm^{-1} region, identified as the spectral range of the amide III and of the phospholipids [16]. Similarly, an area of high correlation between 1300 and 1100 cm^{-1} was obtained when the species of the *Kazachstania* genus were used (Fig. **3a**) and in the joint analysis of strains belonging to the genera *Saccharomyces*, *Lachancea*, *Zygosaccharomyces* and *Kazachstania* (Fig. **3b**). The area between 900 and 700 cm^{-1}, already individuated as important in species identification and strain typing [17], showed several interesting peaks with relatively high correlation.

From these analyses resulted that some descriptors (wavenumbers) were more interesting than others to correlate FTIR with DNA alignments. Reducing the original FTIR matrix to these wavenumbers requires to define a threshold of correlation such that only wavenumbers with correlation over the threshold will remain in the optimal matrix. The analyst can carry out this operation with the function *odi* by defining the desired threshold and observing the “gain in correlation”, i.e. the increase of correlation of the optimal matrix *vs* the original dataset. In case no threshold can be defined *a priori*, the function *rodi* carries out an automatic analysis of the gain in correlation obtainable at different thresholds. In this validation. the *rodi* function obtained three very different profiles for the three taxonomic groupings (Fig. **4**). The analysis on *Debaryomyces* species yielded a marked increase of the gain in correlation with a threshold around 0.75, triggering the correlation between DNA alignment and optimized FTIR matrix from 0.5 to 0.8 (Fig. **4a**). In the other two cases there is a slower increase of correlation which occurs with threshold values from 0.50 to 0.65 when the *Kazachstania* species were analyzed (Fig. **4b**) and between 0.35 and 0.45 in the four yeast genera group (Fig. **4c**). In this last case it is remarkable that when the threshold increases from 0.45 to 0.55 the correlation drops from 0.64 to 0.52. This behavior can be explained considering that the increase of the threshold decreases the number of descriptors and that these few wavenumbers produce a good correlation only for some strains, but not for the others, causing a decrease of the global correlation. As a matter of fact this phenomenon occurred in the more heterogeneous groups represented by strains from four yeast genera. The differences of the gain in correlation observed in these three taxonomic situations indicate that no general rule can be formulated regarding the increase of correlation when the threshold raises. However it should be noticed that very high thresholds increase the correlation, but can reduce too much the number of descriptors. These plots suggest that the best strategy to produce optimal matrices consists in defining threshold values producing high correlation retaining a sufficient number of descriptors that in our cases should be at least of some tents.

The distance of the optimized matrices were subject to Mantel test with the distances of the original FTIR matrices and of the DNA alignment. Interestingly, there was a low correlation between the optimized matrix and the FTIR matrix from which it was derived and a large increase in correlation with the DNA alignment that in the three taxonomic panels was 0.613, 0.404 and 0.156 (Table **3**). The fact that the group of four yeast genera produced the smallest increase in correlation indicates that this method has a

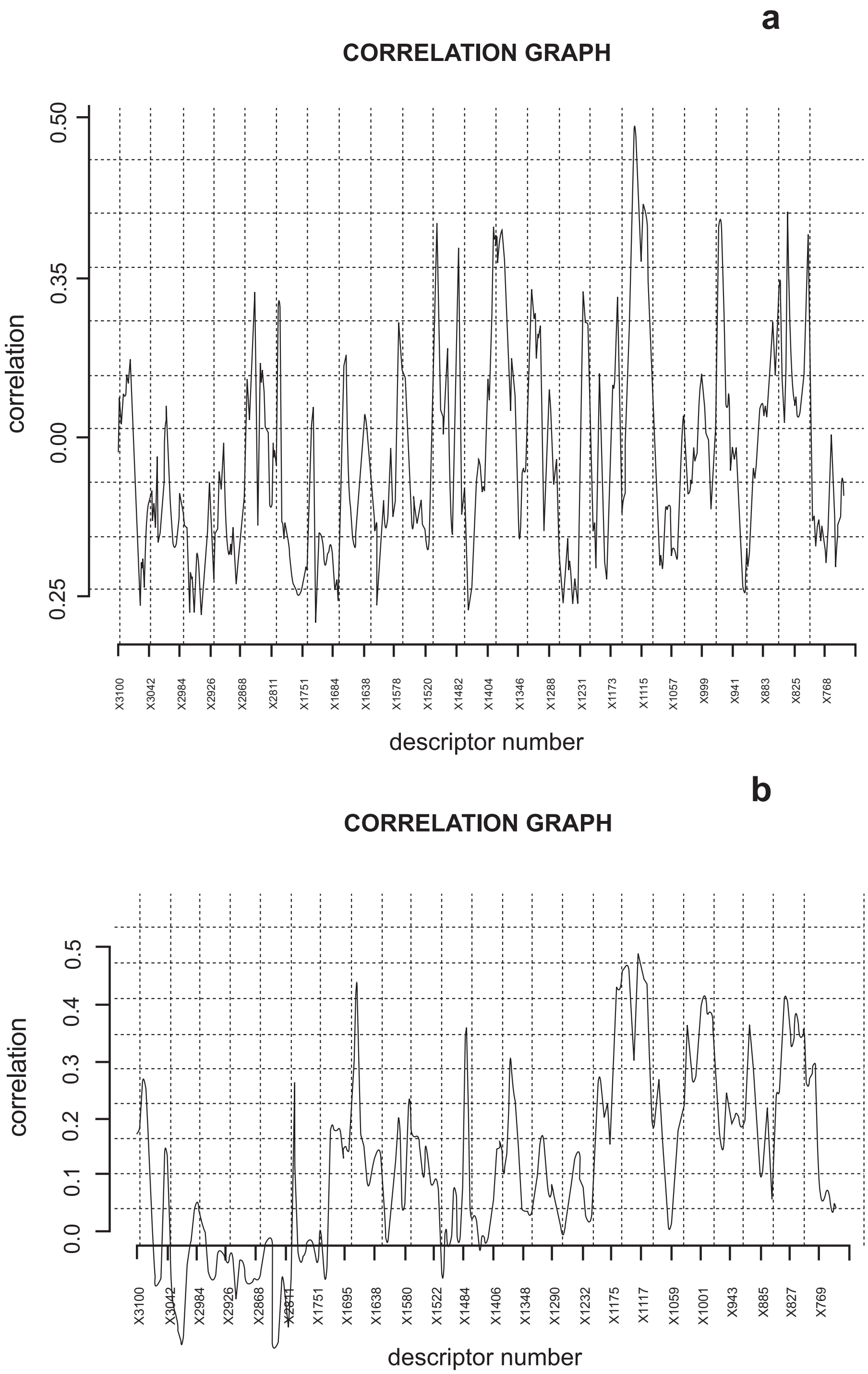

Fig. (3). Correlation plots between D1/D2 domain DNA sequences and whole FTIR spectra.

Panel **a** shows the correlation plot obtained with species of the genus *Kazachstania*, panel **b** reports the correlation plot obtained with some species of the four yeast genera *Saccharomyces, Zygosaccharomyces, Kazachstania* and *Lachancea*.

Parameters were: function *ada. maca*, grid=30, sf=20, distance method = *dcor1*, correlation according to Pearson's algorithm.

natural limit with heterogeneous groups in which the single wavenumbers were easily optimal for some members and not for others.

The increase of correlation obtained by the optimal matrix led to an arrangement of the strains in the dendrogram more similar to that obtained with the D1/D2 26S alignments. In fact the DNA marker indicates *K. transvaalensis* as the most distant species of the group and places together into two separate clusters the strains of *K. exigua* and of *K. unispora*. (Fig. **5a**). The FTIR dendrogram retained only the clustering of the *K. unisporai* strains, while

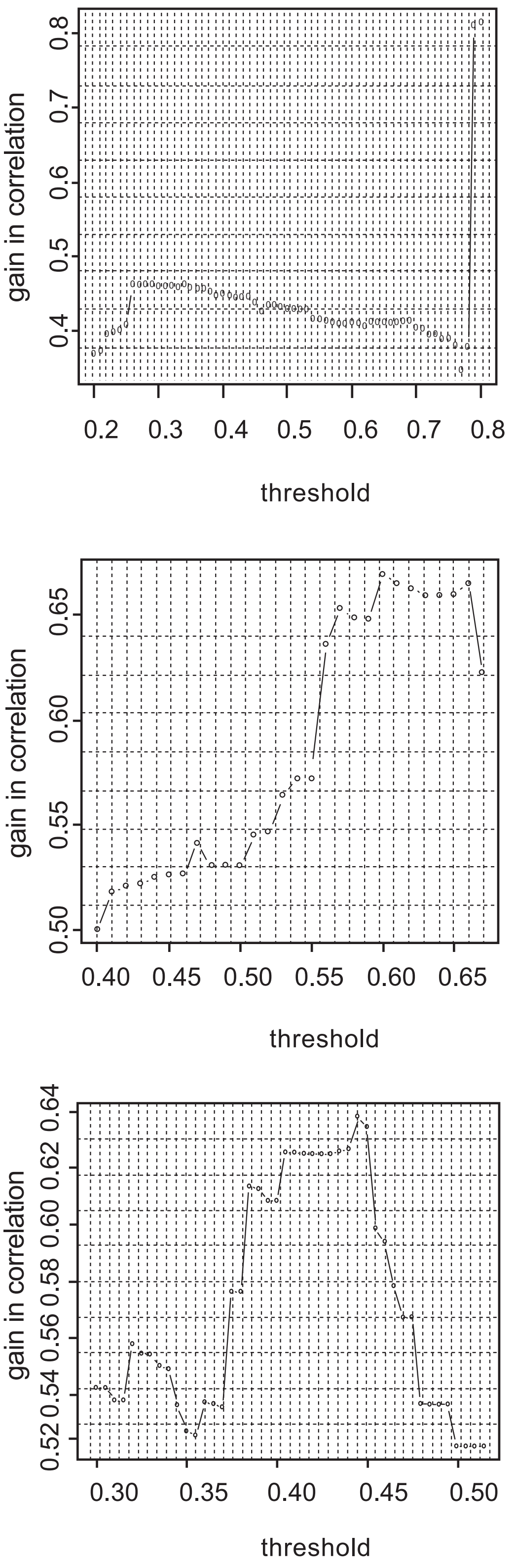

Fig. (4). Gain in correlation plots.

The plots show the gain in correlation by varying threshold values. Panel **a** *Debaryomyces* species, panel **b** species of the genus *Kazachstania*, panel **c** some species of the yeast genera *Saccharomyces, Zygosaccharomyces, Kazachstania* and *Lachancea*.

the two isolates of *K. exigua* resulted in separate groups. Moreover, the *K. transvaalensis* appeared clustered with two other species (Fig. **5b**). The optimized FTIR dendrogram appeared topologically very similar to that obtained with the DNA sequences with the only difference that the *K. barnetti* - *K. spencerorum* clade was split.

These data confirm that an increase of correlation improves the topological distribution of the strains in the dendrogram, although this effect was less visible in the four yeast genera group (data not shown).

Comparison of DADI with Other Statistical Methods

The efficacy of DADI in selecting the best correlating wavenumbers to a 26S DNA alignment was compared with that obtainable with QUANT2, a Partial Least Square (PLS) regression method developed for the specific use of spectroscopy and included in the package OPUS (Bruker - Germany). The vector of reference data to set up the regression in QUANT2 was obtained as the distances of each *Kazachstania* strains from *K. barnettii* as calculated according to the 26S rDNA alignment. The Quant2 method produced a regression curve with R^2 =0.245 using the regions from 2557.6 to 2196 and from759.9 to 399.3cm^{-1} (Fig. **6a**). These two wavenumbers ranges selected by QUANT2 were employed to build the dendrogram displayed in Fig. (**6b**). The dendrogram is inconsistent from a phylogenetic point of view, in fact separates the two *K. exigua* strains, which are indeed very similar at the 26S level. Furthermore, the region spanning from 2557.6 to 2196 cm^{-1} does not include any significant signal and is considered critical by the PLS procedure, whereas it gave very low correlation values when DADI was employed to search significantly correlated wavenumbers [18]. Another advantage of DADI is the possibility to set the threshold of correlation (*odi* function) to select the optimal wavelengths, whereas the QUANT2 algorithm gives a series of ranges without the possibility for the analyst to interact with the program.

Comparison with Random Matrices

The interpretation of the correlation graph could raise the question on whether the figures observed are really due to a solid correlation between the matrices or to random effects. In order to address the question the package DADI includes the function *randomtest,* which produces a matrix of random values (*ran*) with the same dimensions of the test matrix. The *ran* matrix can be further used in any correlation function (*ma.maca, ama.maca* etc) to visualize the difference between the correlation plot obtained with the actual test matrix and that yielded by random data. Using the *randomtest* function we carried out an *ada.maca* analysis similar to that shown in the validation of the algorithm using 26S alignment and FTIR data (Fig. **7a**). The same analysis with the same settings was repeated using a *ran* matrix (Fig. **7b**). In the former case the peaks are well shaped and discriminated, whereas in the latter are irregularly distributed. Moreover, the maximum correlation observed with real data was 0.6, whereas was 0.4 with random data. However, it seems more important to consider the peak distribution and their significance (biological or chemical) more than the maximum correlation value. Using the *sort* command within the *ada.maca* function, two rather different

Table 2. List of the Strains Employed in this Study

Genus	Species	Strain	Genus	Species	Strain
Debaryomyces	*D. carsonii*	CBS 2285	*Saccharomyces*	*S. bayanus*	CBS 380
Debaryomyces	*D. carsonii*	CBS 4050	*Saccharomyces*	*S. cariocanus*	CBS 8841
Debaryomyces	*D. castellii*	CBS 2923	*Saccharomyces*	*S. cerevisiae*	CBS 1171
Debaryomyces	*D. coudertii*	CBS 5167	*Saccharomyces*	*S. cerevisiae*	CBS 1508
Debaryomyces	*D. etchelsii*	CBS 2011	*Saccharomyces*	*S. cerevisiae*	CBS 2247
Debaryomyces	*D. hansenii var. fabryii*	CBS 789	*Saccharomyces*	*S. cerevisiae*	CBS 5835
Debaryomyces	*D. hansenii*	CBS 1793	*Saccharomyces*	*S. kudriavzevii*	CBS 8840
Debaryomyces	*D. hansenii*	CBS 2330	*Saccharomyces*	*S. mikatae*	CBS 8839
Debaryomyces	*D. marama*	CBS 1958	*Saccharomyces*	*S. paradoxus*	CBS 432
Debaryomyces	*D. marama*	CBS 4262	*Saccharomyces*	*S. pastorianus*	CBS 1538
Debaryomyces	*D. melissophilus*	CBS 6344			
Debaryomyces	*D. melissophilus*	CBS 6694	*Kazachstania*	*K. barnettii*	CBS 5648
Debaryomyces	*D. nepalensis*	CBS 1325	*Kazachstania*	*K. exigua*	CBS 2141
Debaryomyces	*D. nepalensis*	CBS 2334	*Kazachstania*	*K. exigua*	CBS 4660
Debaryomyces	*D. nepalensis*	CBS 5921	*Kazachstania*	*K. martiniae*	CBS 6334
Debaryomyces	*D. occidentalis occidentalis*	CBS 4516	*Kazachstania*	*K. spencerorum*	CBS 3019
Debaryomyces	*D. occidentalis parsoonii*	CBS 2169	*Kazachstania*	*K. transvalensis*	CBS 2186
Debaryomyces	*D. polymorphus africanus*	CBS 6741	*Kazachstania*	*K. unispora*	CBS 3004
Debaryomyces	*D. polymorphus polymorphus*	CBS 186	*Kazachstania*	*K. unispora*	CBS 399
Debaryomyces	*D. polymorphus polymorphus*	CBS 4346			
Debaryomyces	*D. pseudopolymorphus*	CBS 2008	*Zygosaccharomyces*	*Z. bailii*	CBS 680
Debaryomyces	*D. udenii*	CBS 7056	*Zygosaccharomyces*	*Z. florentinus*	CBS 746
Debaryomyces	*D. vanrijiae*	CBS 6454	*Zygosaccharomyces*	*Z. microellipsoides*	CBS 427
Debaryomyces	*D. vanrijiae*	CBS 6756			
Debaryomyces	*D. yamadae*	CBS 7035	*Lachancea*	*L. cidri*	CBS 4575
Debaryomyces	*D. yamadae*	CBS 7036	*Lachancea*	*L. fermentati*	CBS 707

curves were obtained (Fig. **7c**, **d**). The sorted curve from random data is quite similar to a uneven degree curve, whereas that from real data is much less smooth and presents some regions of irregularity. All together it seems that the shape of the sorted and usorted correlation graphs is a good parameter to determine whether the correlation values are similar to those yielded with a series of random data.

Table 3. Results of Mantel Test on DNA, FTIR and Optimized FTIR Matrix Distance Correlation

Table 3a.

	Dd	Fd	OFd
DNA distances (Dd)	■	0.055	<0.001
FTIR distances (Fd)	0.199	■	0.24
Optimized FTIR Distances (Ofd)	0.812	0.07	■

Table 3b.

	Dd	Fd	OFd
DNA distances (Dd)	■	0.117	0.024
FTIR distances (Fd)	0.261	■	0.021
Optimized FTIR Distances (OFd)	0.665	0.434	■

Table 3c.

	Dd	Fd	OFd
DNA distances (Dd)	■	0.08	<0.001
FTIR distances (Fd)	0.182	■	0.004
Optimized FTIR Distances (OFd)	0.638	0.33	■

Mantel test were carried out on distance matrices among strains of the *Debaryomyces* genus (Tab. 3a), *Kazachstania* genus (Table **3b**) and of the *Saccharomyces, Zygosaccharomyces* and *LAchancea* genera (Table **3c**)

– Lower triangle figures are results of the mantel test between distance matrices. Upper triangle figures represent the probability error of the corresponding tests; low values indicate robust results.

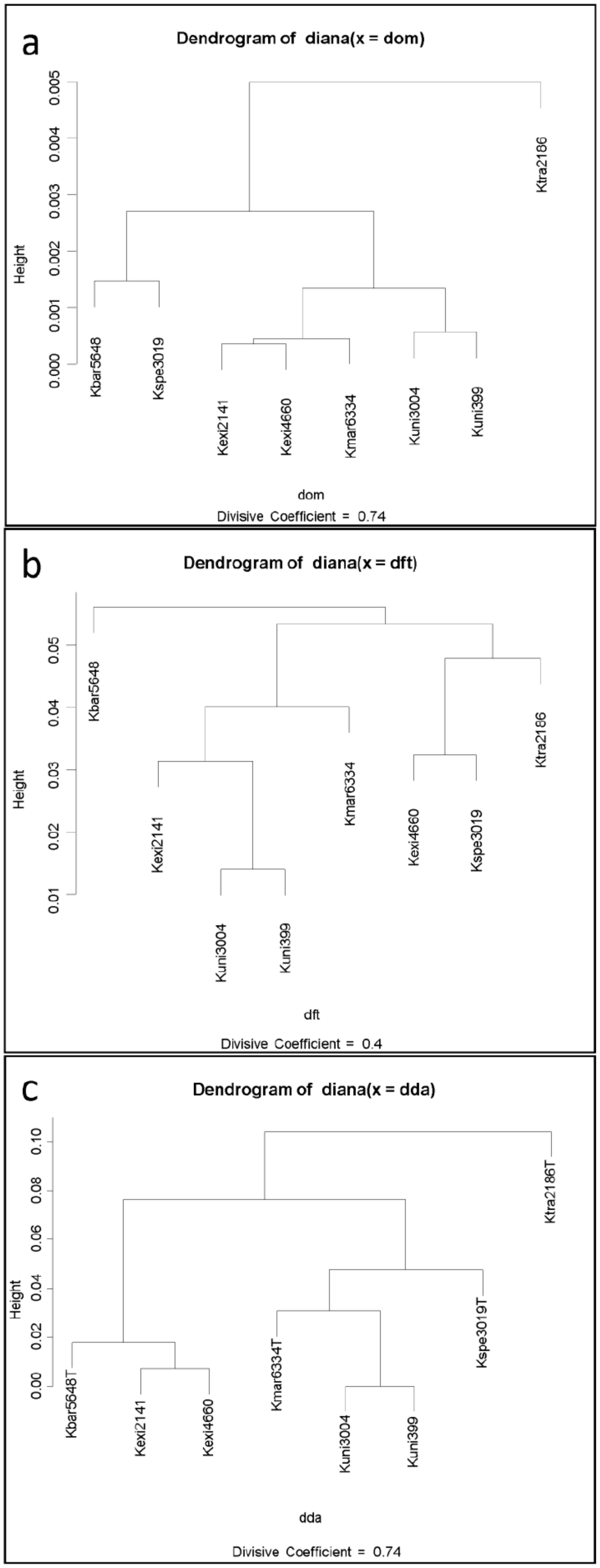

Fig. (5). Dendrograms of species belonging to the yeast genus *Kazachstania.*

Dendrograms were obtained with the *diana* function from the *cluster* R package. Panel **a** dendrogram calculated on the basis of the 26S D1/D2 domain alignments, panel **b** and **c** dendrograms obtained respectively from the whole and optimized FTIR matrices.

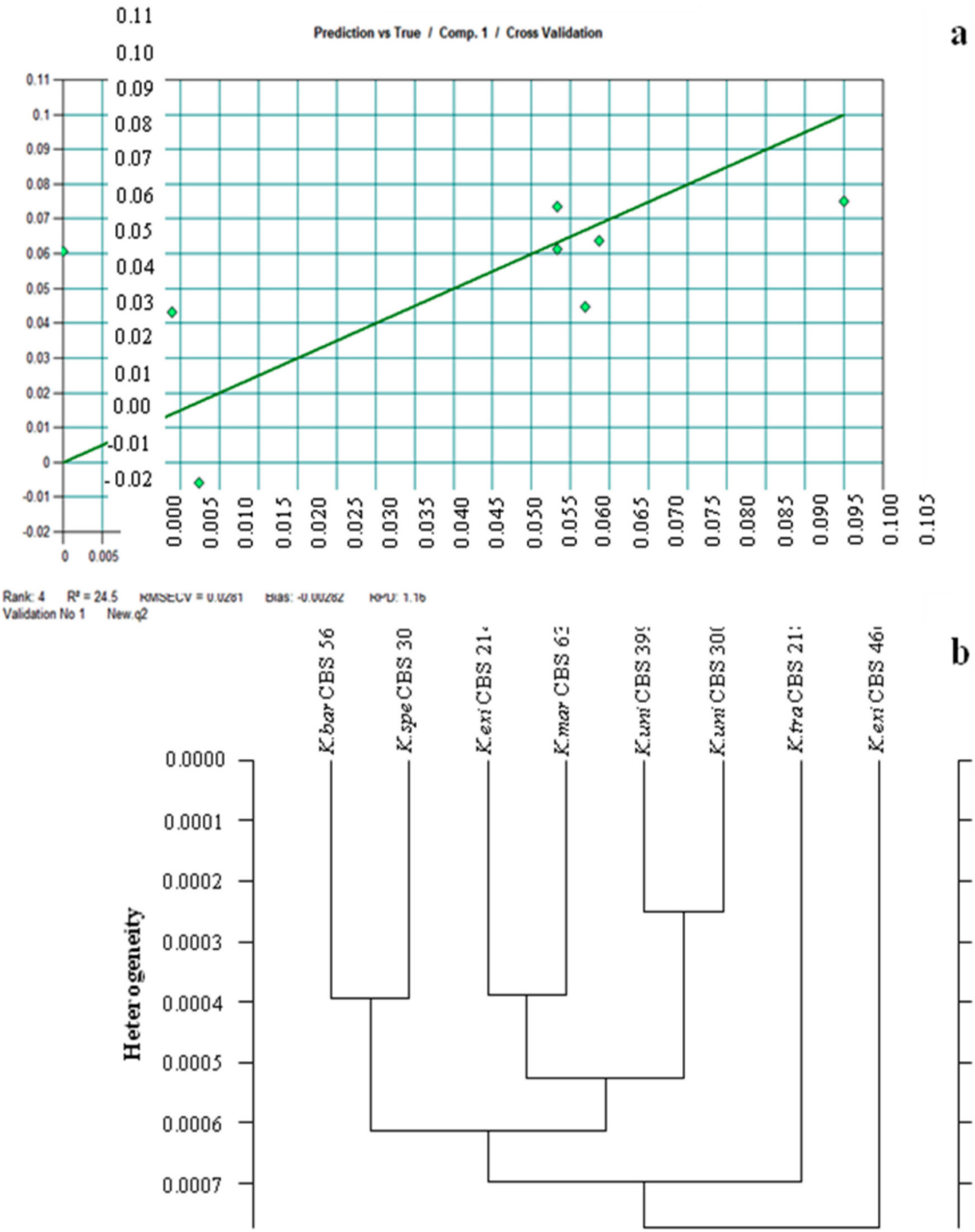

Fig. (6). QUANT analysis of FTIR data on the basis of the 26S rDNA alignments.

a) QUANT2 regression, **b**) dendrogram obtained with the FTIR regions indicated by the QUANT2 analysis.

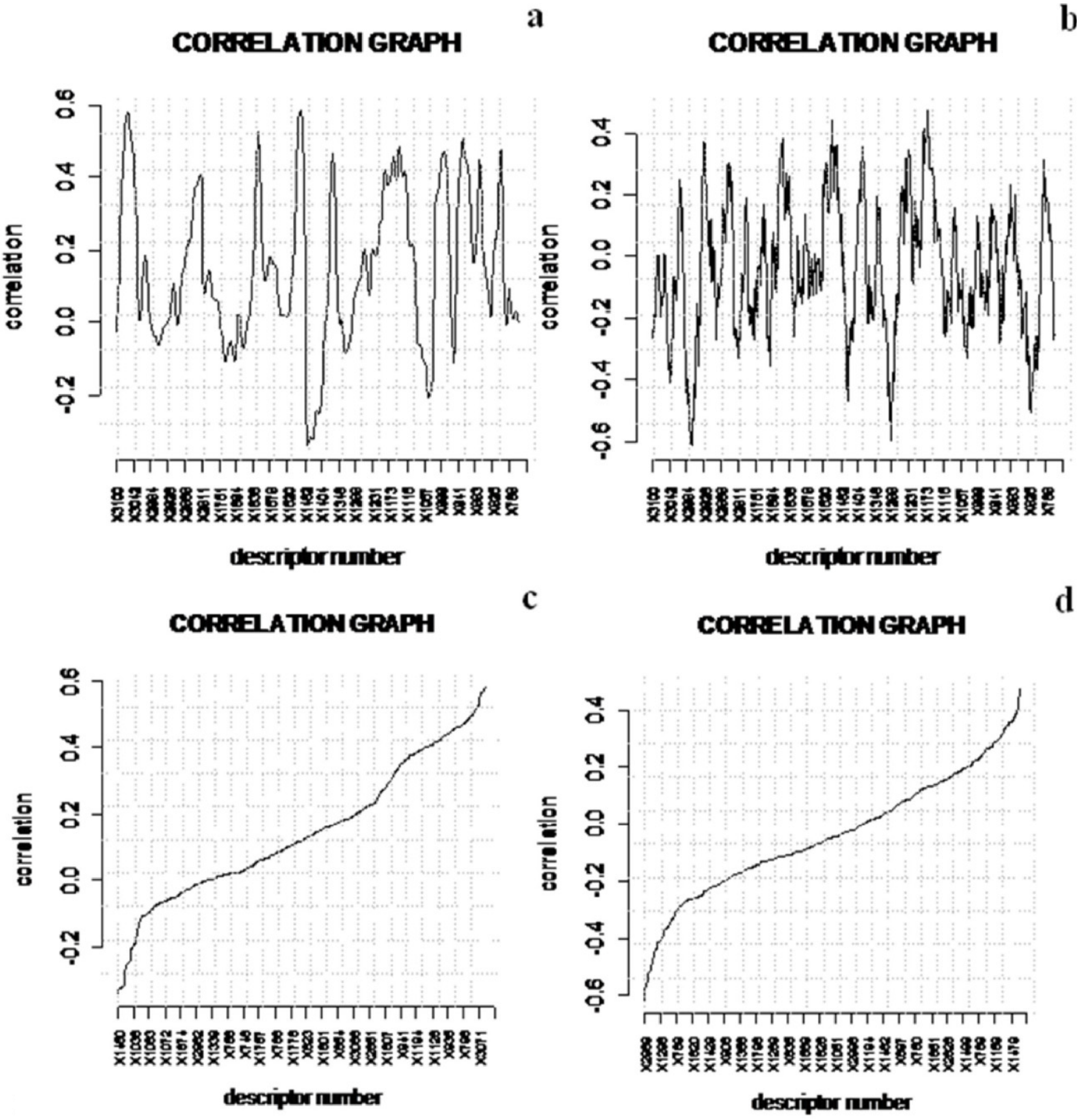

Fig. (7). Compsrison of correlation graphs derived from FTIR spectral data (Test matrix) and 26S rDNA sequences of eight yeast strains (reference matirx); **b**. Correlation graph obtained with a randomly generated test matrix and the same reference matrix as in **a**, **c**, and **d** are the same correlation graphs as **a** and **b**, respectively, with correlation values sorted in ascending order.

Cross Correlation

The cross-correlation function *dcc* allows for rapid analyses in diverse combination, of which three are illustrated below.

The *dcc* algorithm was firstly used to correlate the strains of the four genera group, according to the description provided by the whole FTIR matrix. This can be accomplished by introducing in the *dcc* function the transposed matrix of the FTIR matrix. Results in Fig. (**8**) showed that intermediate correlation values prevailed (green squares). The high correlation values (blue), beyond those of the ascending diagonal, are present within the two groups delimited by curly brackets. Interestingly, in the larger group were included species of all the four genera, indicating that FTIR, without optimization could not organize the taxa in a way similar to that of the DNA alignments.

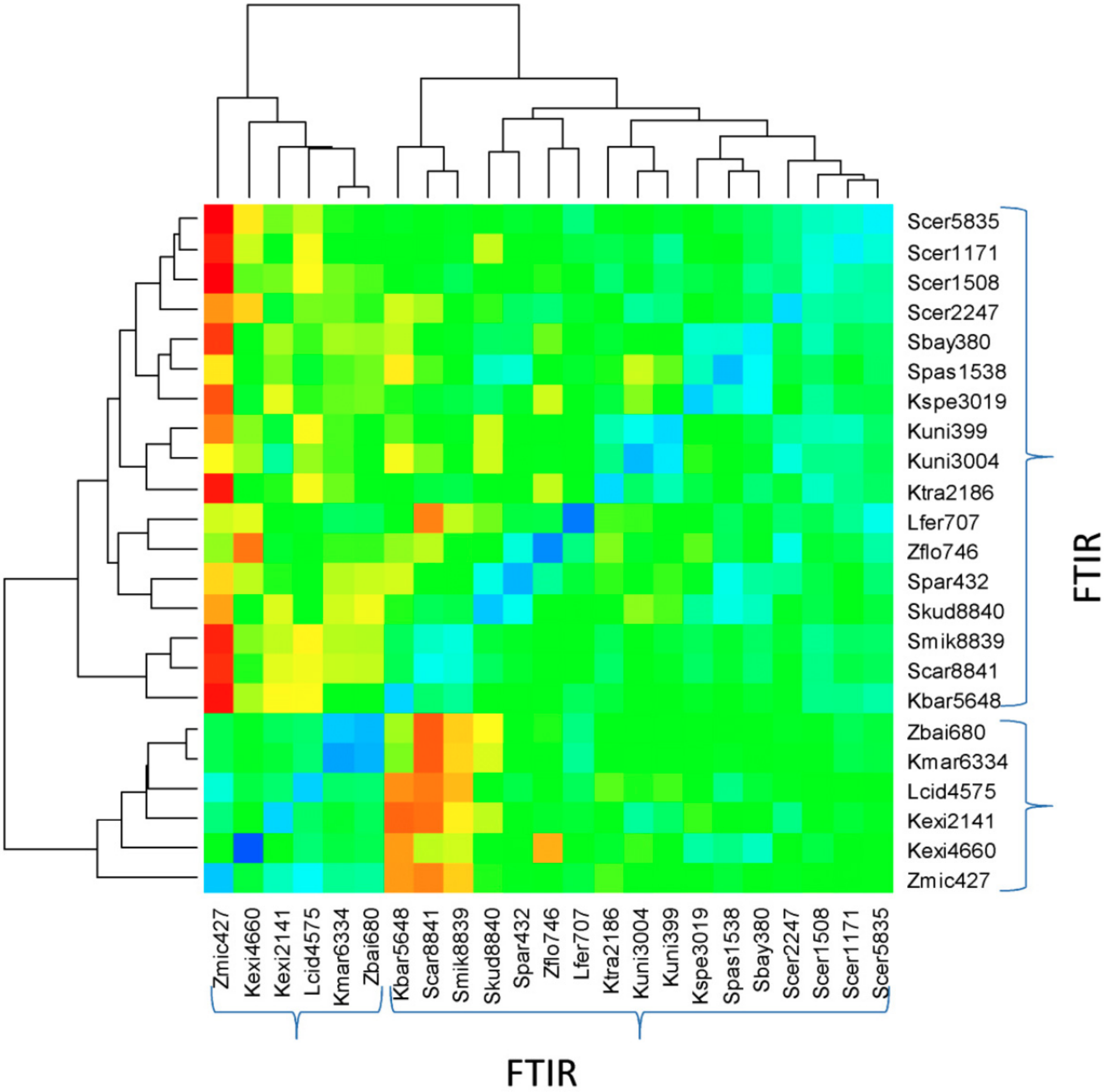

Fig. (8). Pseudo heathmap showing the correlation among some species of the genera *Saccharomyces, Zygosaccharomyces, Kazachstania* and *Lachancea*.

Another application of *dcc* was carried out by introducing the complete FTIR matrix, producing a cross comparison of the wavenumbers (Fig. **8**), which were maintained in decreasing order from left to right and from bottom to top, in order to facilitate the reading of the pseudo-heatmap. In this very dense map (2400 x 2400) the intermediate correlations prevail, alternated with red and yellow areas (very low and low correlation) and blue areas (high correlations). Interestingly, the size of the red-yellow and blue stains decreases with the decrease of the wavenumbers, resulting in a very dense web in the right-upper part. This observation suggests that most of the FTIR descriptors are independent from one another as shown by the prevalence of the green squares. However, in several parts of the map a combination of close red (low correlation) and blue (high correlation) stains could be found as emphasize by the dotted squares. These areas represent an alternation of descriptors linked by a strong correlation which alternated from positive (blue) to negative (red) within few cm^{-1} of wavenumbers. The right understanding of the meaning of these areas of high correlation needs further insight and could help in the interpretation of the FTIR spectra and in their use.

CONCLUSIONS

The proposed R package DADI was able to identify the FTIR wavenumbers best fitting to the chosen reference, supporting its use in defining the descriptors to be used in FTIR and possibly in other metabolomic analyses. Beyond taxonomy and phylogenetics, DADI can be employed in a number of applications. In general, it seems that DADI can be a helpful application to optimize complex datasets composed of independent variables. The main conceptual problem is maybe represented by the difficulty to define a significant reference model really representative of the situation under study. The possibility to consider alternatively each matrix as reference or as test and the cross correlation obtained with *dcc* are additional opportunities to identify the best descriptors of each matrix and to define the reference model carefully.

ACKNOWLEDGEMENTS

The authors thank Dr. Pierangelo Morini (BRUKER Optics Italia) for the many useful discussions on the topics treated in this paper. This work has been partially supported by FIRB (Italian Ministry of University and Technology) N° RBIP06H7H4_003.

Availability. DADI is freely available as a R package from the CRAN repository (http://cran.r-project.org/src/contrib/PACKAGES.html); Contact: gianlu@unipg.it

REFERENCES

[1] S. Vaidyanathan, G.G. Harrigan, and R. Goodacre, *Metabolome analyses: strategies for systems biology*, Springer Verlag, 2005.

[2] L. Corte, M. Lattanzi, P. Buzzini, A. Bolano, F. Fatichenti, and G. Cardinali, "Use of RAPD and killer toxin sensitivity in *Saccharomyces cerevisiae* strain typing", *J. Appl. Microbiol.*, vol. 99, pp. 609-617, 2005.

[3] A.M. Martins, W. Sha, C. Evans, S. Martino-Catt, P. Mendes and V. Shulaev, "Comparison of sampling techniques for parallel analysis of transcript and metabolite levels in *Saccharomyces cerevisiae*", *Yeast*, vol. 24, pp. 181-188, 2007.

[4] M. Imielinski, C. Belta, A. Halasz and H. Rubin, "Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities", *Bioinformatics*, vol. 21, pp. 2008-2016, 2005.

[5] V. Mapelli, L. Olsson, and J. Nielsen, "Metabolic footprinting in microbiology: methods and applications in functional genomics and biotechnology", *Trends Biotechnol.*, vol. 26, pp. 490-497, 2008.

[6] P. Legendre, and L. Legendre, *Numerical ecology*, Elsevier Science, 1998.

[7] D.P. Enot, W. Lin, M. Beckmann, D. Parker, D.P. Overy, and J. Draper, "Preprocessing, classification modeling and feature selection using flow injection electrospray mass spectrometry metabolite fingerprint data," *Nat. Protoc.*, vol. 3, pp. 446-470, 2008.

[8] 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.

[9] G.D. Sockalingum, W. Bouhedja, P. Pina, P. Allouch, C. Bloy, and M. Manfait, "FT-IR spectroscopy as an emerging method for rapid characterization of microorganisms" *Cell. Mol. Biol. (Noisy-le-grand)*, vol. 44, pp. 261-269, 1998.

[10] A. Oust, T. Moretro, C. Kirschner, J.A. Narvhus, and A. Kohler, "FT-IR spectroscopy for identification of closely related lactobacilli", *J. Microbiol. Methods.*, vol. 59, pp. 149-162, 2004.

[11] A. Oust, T. Moretro, K. Naterstad, G.D. Sockalingum, I. Adt, M. Manfait, and A. Kohler, "Fourier transform infrared and Raman spectroscopy for characterization of *Listeria monocytogenes* strains", *Appl. Environ. Microbiol.*, vol. 72, pp. 228-232, 2006.

[12] E.M. Timmins, S.A. Howell, B.K. Alsberg, W.C. Noble, and R. Goodacre, "Rapid differentiation of closely related Candida species and strains by pyrolysis-mass spectrometry and Fourier transform-infrared spectroscopy", *J. Clin. Microbiol.*, vol. 36, pag. 367-374, 1998.

[13] K. Tintelnot, G. Haase, M. Seibold, F. Bergmann, M. Staemmler, T. Franz, and D. Naumann, "Evaluation of phenotypic markers for selection and identification of *Candida dubliniensis*", *J. Clin. Microbiol.*, vol. 38, pag. 1599-1608, 2000.

[14] K. Maquelin, C. Kirschner, L.P. Choo-Smith, N. Van Den Braak, H.P. Endtz, D. Naumann, and G.J. Puppels, "Identification of medically relevant microorganisms by vibrational spectroscopy", *J. Microbiol. Methods.*, vol. 51, pp. 255-271, 2002.

[15] M. Essendoubi, D. Toubas, C. Lepouse, A. Leon, F. Bourgeade, J.M. Pinon, M. Manfait, and G.D. Sockalingum, "Epidemiological investigation and typing of *Candida glabrata* clinical isolates by FTIR spectroscopy", *J. Microbiol. Methods.*, vol. 71, pp. 325-331, 2007.

[16] M. Kummerle, S. Scherer, and H. Seiler, "Rapid and reliable identification of food-borne yeasts by Fourier-transform infrared spectroscopy", *Appl. Environ. Microbiol.*, vol. 64, pp. 2207, 1998.

[17] D. Naumann, D. Helm, and H. Labischinski, "Microbiological characterizations by FT-IR spectroscopy", *Nature*, vol. 351, pp. 81-82, 1991.

[18] C. Yu and J. Irudayaraj. "Spectroscopic characterization of microorganisms by Fourier transform infrared microspectroscopy", *Biopolymers*, vol. 77, pp. 368-77, 2005.

Received: January 13, 2010 Revised: March 4, 2010 Accepted: June 9, 2010