Discriminant Function Analysis Based on Principal Components for Rapid Discrimination of Metabolic Capabilities of New Isolates

B. Sariyar-Akbulut*
Department of Bioengineering, Marmara University,
Kadikoy 34722, Istanbul, Turkey

Original scientific paper
Received: October 13, 2008
Accepted: September 8, 2009

The growing need for new microorganisms with novel metabolic capabilities has urged scientists to search for life in extreme environments. With the rapid progress in experimental methods, it is possible to isolate new microorganisms at high speeds but the bottleneck in this process is the biochemical characterization due to time and financial limitations. Inferential hierarchical clustering of new isolates may help to overcome this problem. In this work, discriminant function analysis, used in conjunction with principal component analysis (PCA) was able to rapidly discriminate eight new isolates using metabolic footprints (spectral data from electrospray injection mass spectrometry) and the results were compared with clustering based on the Euclidean distances computed both in the domain of original data and in the domain of PCA-transformed data. The presence of the replicates on the adjacent leaf nodes of dendrograms obtained by hierarchical cluster analysis confirmed the reliability of the method. This attractive tool is applicable to a chemical/biological system, which involves complex samples that can be analyzed by high-throughput instruments.

Key words:
Hierarchical clustering, discriminant function analysis, principal component analysis, metabolic capability

Introduction

Extremophiles that live under extreme saline conditions are called halophilic microorganisms. In view of the increasing importance of extremozymes, halophiles are gaining special interest. Halophilic microorganisms are important sources of enzymes including DNAses, lipases, amylases, gelatinizes, and proteases, which are capable of functioning under high concentrations of salt that lead to precipitation or denaturation of most proteins. Moreover, halophilic microorganisms are sources of metabolites such as ectoine and betaine, which are used as stabilizers for biomolecules and whole cells. Their ability to oxidize hydrocarbons in the presence of salt makes them important microorganisms for biological treatment of saline ecosystems contaminated with petroleum products.[1] Within this group of microorganisms, moderate halophiles receive special interest for their metabolic capabilities owing to their growth in a wide range of salt mass fractions.[2]

The growing demand for stabilizer metabolites and proteins stable under high salt mass fractions in different industrial applications has focused attention on a number of new salinity environments for production by novel microorganisms.[3,4] Microorganisms are being isolated from new environments at an unpredictable speed enabled by the rapid progress in experimental methodologies. However, biochemical characterization is essential for understanding the metabolic capabilities of the isolates to evaluate their industrial importance and move on to bio-process optimization. Unfortunately as the number of isolates to be analyzed increases exponentially, it becomes more difficult to perform each individual biochemical analysis due to time and financial limitations. For routine purposes, it would be ideal first to cluster the new isolates based on their metabolic similarities prior to extensive characterization. The substrates and the products of the complex network of many metabolic reactions are present in the culture broth. Therefore the analysis of the cocktail of metabolites left behind or secreted into the medium, *footprints*, is the closest indicator of metabolic state of a microorganism.[5,6] Its analysis may be used for *inferential* clustering of newly isolated microorganisms. Once the clusters are identified, only the selected analyses for each group of microorganisms will be carried out. This will help to reduce the time and money required for extensive experimental labor.

Footprints are complex biological samples, therefore their analysis necessitated developments

*Address of correspondence
Marmara University, Faculty of Engineering, Bioengineering
Department, Goztepe Campus, Kadikoy 34722 Istanbul, Turkey
Tel: + 90 (216) 348 02 92/721, Fax: + 90 (216) 345 01 26
e-mail: berna.akbulut@marmara.edu.tr

in analytical instrumentation and physical-chemical spectroscopic methods. The advent of soft ionization techniques, such as matrix-assisted laser desorption/i ionization and electrospray ionization (ESI), has enabled the mass spectrometric analysis of both large molecular weight compounds such as proteins and low molecular mass metabolites such as amino acids, nucleotides etc.[7,8] Electrospray ionization mass spectrometry (ESI-MS) in which the samples are in liquid phase, is a tool for the reproducible analysis of complex biological samples.[2,9,10] ESI-MS is very attractive for rapid and high throughput automated analysis since it does not require an analyte separation or clean-up step.

Multidimensional spectra from such high--throughput instruments give quantitative information about the total biochemical composition of a sample but their interpretation requires unsupervised pattern recognition methods.[11] Goodacre *et al.*[10] have shown that with the combination of unsupervised- and supervised-learning methods, such as principal component analysis (PCA), independent component analysis, factor analysis, discriminant function analysis (DFA), artificial neural networks and hierarchical cluster analysis (HCA), it is possible to seek for clusters and group complex biological samples. This approach has been proved to be useful for clustering with great accuracy, selectivity, and sensitivity.[10−16]

Even though much work has been done on the analysis of footprints, these research efforts have been confined to type strains, microorganisms that have already been characterized. In this work, using their footprint analysis from ESI-MS, eight new isolates (from Çamaltı Saltern in West cost of Turkey) were inferentially clustered using three different methods and the results were compared in terms of interpretability of the dendrograms obtained via HCA. Based on their salt tolerance, the isolates were initially characterized as moderately halophilic microorganisms and therefore moderately halophilic type strains *Halomonas salina* and *Halomonas halophila* were taken as precursors. The hierarchical tree constructed helped visualize the fine differentiation of the new isolates. This methodology that is based on metabolic differences proved to be rapid, automated and inexpensive for preliminary analysis and structuring.

Experimental

Soil samples collected from Çamaltı Saltern in West cost of Turkey were diluted serially in $w = 20$ % NaCl on agar plates containing yeast extract (0.5 %), sodium citrate (0.3 %), $MgSO_4 \cdot 7H_2O$ (0.2 %), KCl (0.2 %), NaCl (20 %). Colonies were selected after 1 week of growth (39 °C) based on shape, size and pigmentation. Type strains *H. salina* (DSM 5928) and *H. halophila* (DSM 4770) were purchased from DSMZ. *E. coli* K12, grown in LB, was from our laboratory stock. Chemicals used were from Merck AG (Darmstadt, Germany) and Sigma Chem. Ltd. (USA).

Morphological, cultural, physiological and biochemical properties of the isolates were determined as suggested by Sneath *et al.*[17] Isolates were grown in the above medium containing 0–25 % NaCl to determine salt tolerance. Eight microorganisms that have grown at salt mass fractions between 5–20 % were selected. Gram staining was performed using Gram Staining Kit (Biomerieux).[18] Effect of temperature was monitored by growth at $\theta = 37$, 45, 55, and 65 °C. Carbohydrate utilization was determined as described by Oren *et al.*[19] Oxidase,[20] catalase[21], DNase, urease, Tween 80, Tween 20 and indol productions in addition to gelatin degradation and starch hydrolyses were tested using the procedures described earlier.[17,18,22,23] Substrate mass fractions used were $w = 1$ % gelatin, 1 % soluble starch, and 0.1 % Tween 80 and Tween 20, respectively. Reduction of nitrate to nitrite was determined in tubes containing 10 mL of liquid medium supplemented with 1 % KNO_3. Durham tubes were used to examine the presence of gas and nitrite accumulation.[17] Medium with 0.001 ‰ phenol red and yeast extract reduced by $w = 0.5$ % was used to determine acid production from different sugars ($w = 1$ % final fraction). Sodium citrate was not added to this medium.

To obtain the footprints, cells were grown in DSM 593 medium (37 °C, 180 rpm). Supernatant from the harvested cells were passed through 0.45 μm filter units and then diluted 10-fold in 30 % methanol containing 0.1 % formic acid. Mass spectra were collected by ESI-MS (Waters/Micromass) in positive ion mode from m/z 65 to 1500 for 50 scan cycles with a flow rate of $Q = 50$ μL min^{-1}. MS optimization led to the following conditions: capillary voltage 3000 V, source temperature 80 °C, desolvation temperature 300 °C, desolvation gas flow rate at $Q = 200$ L h^{-1}, sample cone voltage 30 V and extraction cone voltage 2 V. Samples were analyzed in duplicate.[24]

Statistical methodology

The data consisted of spectral results as *intensity* vs. m/z values for the isolates and the type strains. The data were labeled as M_i^A and M_i^B ($i = 1,\ldots,10$) where the index i designates microorganisms and the letters A and B indicate duplicate analysis of identical samples. The assignment of the letters A and B was done at random; i.e. M_i^A could also be labeled as M_i^B. In the raw spectroscopic data, the *m*/*z* values ranged from 50 to 1500 *m*/*z* for all samples. Since there were no peaks in the *m*/*z* < 100 and *m*/*z* > 800 regions, these sections of the data were removed and, for all samples, only the range 100 < *m*/*z* < 800 was used. Four representative samples of these 20 spectra are given in Fig. 1.

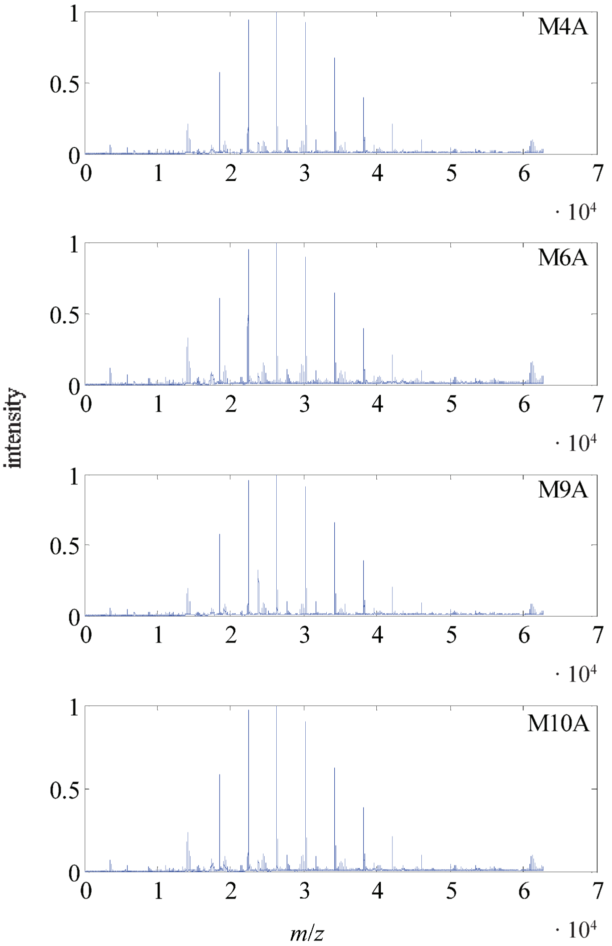

Fig. 1 – *Normalized raw spectroscopic data of footprints from ESI-MS of isolates M_4, M_6 and H. salina (DSM 5958, M_9) and H. halophila (DSM 4770, M_{10})*

The row size of the raw spectroscopic data (maximum number of rows) ranged from 62780 to 66808 and each column had irregularly spaced *m*/*z* values. In order to perform PCA, the columns needed to have equal lengths and common *m*/*z* values. However, both the number of data points and the *m*/*z* values in the chromatograms were different for different microorganisms. This problem was solved using a resampling technique, which transformed all columns of the data to an equal length with unique *m*/*z* values. Each column of the data was normalized to 0–1 range and Matlab's *interp1* function with the nearest-neighbor interpolation option was used to resample each column to 62780 rows. Regularly spaced *m*/*z* values between 100 and 800 were used as interpolation points identically for all the columns. Thus, the data could be represented as a matrix, $\mathbf{X} = [\ M_1^A \ldots M_{10}^A \mid M_1^B \ldots M_{10}^B\ ]$, with 62780 rows and 20 columns. This method preserved the peak positions and heights.

Following this pre-processing, the initial stage involved the reduction of the row dimension of the ESI-MS data by PCA. PCA is a statistical technique for finding patterns in data matrix of high dimensions to highlight their similarities and differences using simple mathematical concepts such as standard deviation, covariance, eigenvectors and eigenvalues. PCA is a linear projection method that defines a new dimensional space, the variables of which are called principal components (PCs). PCs are linear combinations of the original variables and each one is orthonormal to the others. PCA maximizes the variance along the PC axes which are aligned in the directions of significant variances in the original data. Thus, PCA reduces the dimension of the data by capturing the most significant variations (information) along the first few PCs. As the number of PCs increases, a larger fraction of the total information content is accounted for.[25,26]

DFA is commonly used to classify cases into different groups by determining a variable by which members in a group differ through its mean. DFA then uses that variable to predict group membership. For DFA within a group, first, the discriminating (independent) variables are found. Then using a linear combination of these discriminating variables such that $L = b_1x_1 + b_2x_2 + \ldots + b_nx_n + c$, where the *b*'s are discriminant coefficients, the *x*'s are discriminating variables, and c is a constant, a discriminant function (DF) is created. Those variables with the largest (standardized) discriminant coefficients are the ones that contribute most to the prediction of group membership. DFs will be independent or orthogonal, that is, their contributions to the discrimination between groups will not overlap. Computationally, a canonical correlation analysis that will determine the successive DFs and canonical roots (the term root refers to the eigenvalues that are associated with the respective canonical function) is performed. There is one DF for a 2-group discriminant analysis but in general the maximum number of functions will be equal to the number of groups minus one, or the number of discriminating variables in the analysis, whichever is smaller.[27,28]

The first DF maximizes the differences between the values of the dependent variables. This first function will be the most powerful differentiating dimension. The second function is orthogonal to the first one (uncorrelated with it) and maximizes the differences between values of the dependent variables, and so on.[27]

When interpreting multiple DFs, which arise from analyses with more than two groups and more than one variable, one would first test the different functions for statistical significance, and only consider the significant functions for further examination. Then the standardized ***b*** coefficients should be analyzed. The larger the standardized ***b*** coefficient, the larger is the respective variable's unique contribution to the discrimination specified by the respective discriminant function. However, these coefficients do not tell between which of the groups the respective functions discriminate. The nature of the discrimination for each DF can be identified by looking at the means for the functions across groups. How the two functions discriminate between groups can also be visualized by plotting the individual scores for the two discriminant functions.[28]

DFA is not performed on the original spectra because one cannot feed co-linear variables into DFA. The starting point for DFA is the inverse of the pooled variance-covariance matrix within *a priori* groups. This inverse can only exist when the matrix is non-singular.[29,30] Generally

$$(N_\mathrm{s} - N_\mathrm{g} - 1) > N_\mathrm{v}$$

where N_s is the number of samples, N_g is the number of groups, and N_v is the number of inputs (variables; i.e., mass intensities for MS). Singularity can be caused by collinearity, and PCA removes collinearities while also reducing the number of inputs to the DFA algorithm (as explained in Timmins *et al.*[14] and Goodacre *et al.*[31]).

DFA was used to discriminate the microorganisms on the basis of the retained 8 PCs and the prior knowledge of which spectra were replicates. In the computations, the Matlab toolbox for DFA developed by Goodacre *et al.* was used.[32]

Finally HCA was used to construct a similarity measure. In distance-based clustering, the similarity criterion is the distance: two or more objects belong to the same cluster if they are close according to a given distance. The agglomerative HC algorithm[33,34] finds the closest (most similar) pair of clusters and merges them into a single cluster; decreasing the number of clusters one by one. The clusters created are viewed graphically on a dendrogram. For HCA, using Matlab's Statistics Toolbox, the similarity or dissimilarity between every pair of objects (columns) in the data set was found by calculating the distance between objects using the *pdist* function with the Euclidean distance (ED) metric option. Then to group the objects into a binary HC tree, the pairs of microorganisms that were in close proximity were linked together using the *linkage* function (with the *average linkage* option) that used the above distance information. A more detailed Matlab-based application of HCA can be found in Akman *et al.*[35]

In the statistical part of this work, the isolates were clustered using three different methods and the results are compared in terms of interpretability of the dendrograms obtained via HCA:

1) HC based on the ED in the space of entire original raw spectral data (62780 rows corresponding to *m*/*z* values and 20 columns corresponding to intensities);

2) HC based on the ED in the space of the first two PCs of row-reduced spectral data (20 rows corresponding to PC scores of the intensities and 2 columns corresponding to first two PCs);

3) HC based on the ED in the DF space of row-reduced spectral data by the PCA (20 rows corresponding to PC scores of the intensities projected onto two DFs and 2 columns corresponding to first two DFs).

Results and discussion

Morphological and physiological characteristics of isolates

The morphological and physiological characteristics of the isolates are shown in Table 1.

Moderately halophilic bacteria may be gram(+) or gram(−), motile or non-motile, may have different morphological characters (rod, short rod or coc) and colors (yellow, cream, pink, brown, and white). They can grow at different temperatures (0–60 °C) and pH (4.5–11).[2,36] Ventosa *et al.*[2] also proposed that moderately halophilic bacteria are microorganisms which grow optimally in media containing 5–20 % NaCl. All isolates required minimum w = 5 % NaCl for growth. Optimum growth occurred at w = 10 % NaCl mass fraction at 37 °C. All isolates stained gram negative and grew at pH between 6.5 and 7.5. However five isolaters, M_2, M_4, M_5, M_6, and M_8 have the ability to grow at pH up to 8.5. The isolates M_1, M_2, M_7, and M_8 were motile. Based on the salt requirements and morphological characteristics, the isolates (abbreviated as M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8) were moderately halophilic. When compared with the moderately halophilic type strains, although the moderately halophiles *H. salina*, *H. halophila* and *H. elongate*[37,38] use nitrate for growth, none of the isolates grew anaerobically on nitrate and produced nitrite and gas from nitrate. *H. salina* and *H. halophila* are oxidase-positive but as *H. elongate*, the isolates were oxidase-negative. None of the isolates hydrolyzed caseine and gelatine. Isolates M_2, M_4, M_5, and M_8 hydrolyzed Tween-80, however the type strains and isolates M_1, M_3, M_6, and M_7 did not hydrolyze

Table 1 – *Differential phenotypic characteristics of isolates from Çamaltı Saltern Area and type strains (M_9, H. salina; M_{10}, H. halophila; M_{11}, H. elongata)*

Characteristics	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8	M_9	M_{10}	M_{11}
morphology	rod	rod	rod	rod	rod	rod	rod	rod	short rod	short rod	rod
pigmentation	cream	cream	cream	cream	cream	cream	cream	cream	cream-yellow	cream	white
motility	+	+	–	–	–	–	+	+	–	+	+
NaCl range, %	3–20	3–20	3–20	3–20	3–20	3–20	3–20	3–20	2–20	1–20	0–32
NaCl optimum, %	10	10	10	10	10	10	10	10	5	1–12	3–8
pH range	6.5–7.5	6.5–8.5	6.5–7.5	6.5–8.5	6.5–8.5	6.5–8.5	6.5–7.0	6.5–8.5	5–10	5.5–10	5–10
temperature range, °C	15–45	15–55	15–45	15–45	15–40	15–45	15–45	15–40	5–45	10–47	10–45
nitrate reduction	–	–	–	–	–	–	–	–	+	+	+
oxidase	–	–	–	–	–	–	–	–	+	+	–
Hydrolyses of:											
caseine	–	–	–	–	–	–	–	–	–	–	–
gelatin	–	–	–	–	–	–	–	–	–	–	+
tween 80	–	+	+	–	+	–	–	+	–	–	–
urea	–	–	–	–	–	–	–	–	+	+	+
Acid production from:											
D-glucose	+	+	+	+	+	+	+	+	–	+	+
maltose	+	+	+	+	–	+	+	–	–	+	+
lactose	+	+	+	+	–	+	+	–	–	–	+
arabinose	+	+	+	+	+	+	+	+	–	+	+
mannose	+	+	+	+	+	+	+	+	–	+	–

*Data obtained from Lee *et al.* (2005)

Tween-80. All isolates used glycerol (data not shown) and all isolates produced acid from D-glucose, arabinose and mannose. Only the isolates M_5 and M_8 did not produce acid from maltose and lactose. With these findings, the similarity of the isolates to moderately halophiles was clear.[37,38]

These are the basic tests required to determine the biochemical characteristics of new isolates, however they are not sufficient for comparing the metabolic capabilities since they reflect only a limited number of properties. The results presented in Table 1 clearly indicate that a more extensive experimental work is necessary to group the microorganisms based on their metabolic similarities. For this purpose up to 234, morphological, physiological, biochemical, nutritional and antimicrobial susceptibility tests may be required.[38] An alternative approach to this is simply the analysis of footprints. The complex mixture of metabolites found in the culture broth accurately reflects the metabolic state.[5,6] Therefore footprints of the isolates as spectral data from ESI-MS were used for inferential clustering.

Hierarchical clustering analysis

Since the isolates showed identity to moderately halophilic microorganisms, the two moderately halophilic type stains *H. salina* (M_9) and *H. halophila* (M_{10}), were used as precursors for inferential clustering based on the metabolome. To show that the isolates are distant to a mesophilic microorganism such as *E. coli*, the intension was to include its footprint analysis in the HC tree, but the optimal media for isolates had salt concentrations much higher than *E. coli* could tolerate. Even the slightest variations in growth culture would be reflected in the footprints, hence only the isolates and *H. salina* (M_9) and *H. halophila* (M_{10}) were chosen for further studies. The high similarity of the spectral data from ESI-MS analysis (Fig. 1) illustrated that visualization of the spectra was not sufficient for grouping of the closely related isolates and that there was a need for a multivariable analysis.

The frequency of sampling was machine dependent; therefore the recorded instances of *m*/*z* values varied significantly from run to run even for the same sample over the course of the analysis. For the statistical analysis of the data via PCA, the *m*/*z* values of all samples should be identical. Therefore, prior to statistical analysis, the data were resampled for regularization, which is the interpolation of the sample-specific *m*/*z* values to a unique set of *m*/*z* values common to all samples. This regularization did not cause any degradation of the vertical resolution of the spectra.

First, HCA using the ED between resampled spectra was performed. However, the HC-ED did not produce interpretable results due to misalignments of the default footprint signaling instances (Fig. 2). These unavoidable peak misalignments, although not noticeable to the eye, affected the computed ED values between the samples significantly due to the presence of narrow and high peaks and thus resulted in misleading HC results. As can be seen, even the replicates did not appear at the leaf nodes adjacent to each other.

Then, PCA was conducted on the spectral data and the effectiveness of the PCs for clustering was tested. Plotting the PCA results showed that the grouping obtained was not very informative for detailed microbial clustering on the PC1-PC2 plane (Fig. 3a) and thus construction of an interpretable dendrogram failed (Fig. 3b). In a previous study, when PCA was used for the analysis of mass spectral data, the groupings were visualized on the score plot but otherwise PCA alone was not enough for HCA.[18]

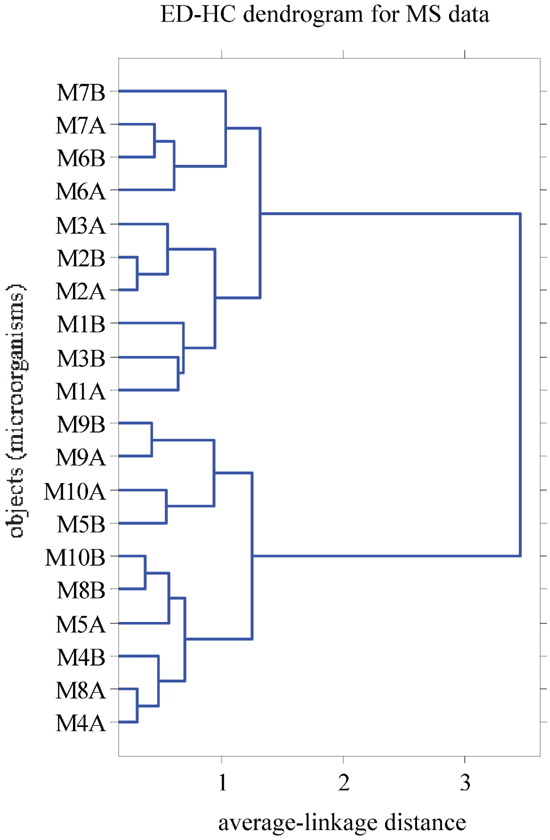

Fig. 2 – *HCA of the microbial footprint data using the ED between spectra as a similarity measure*

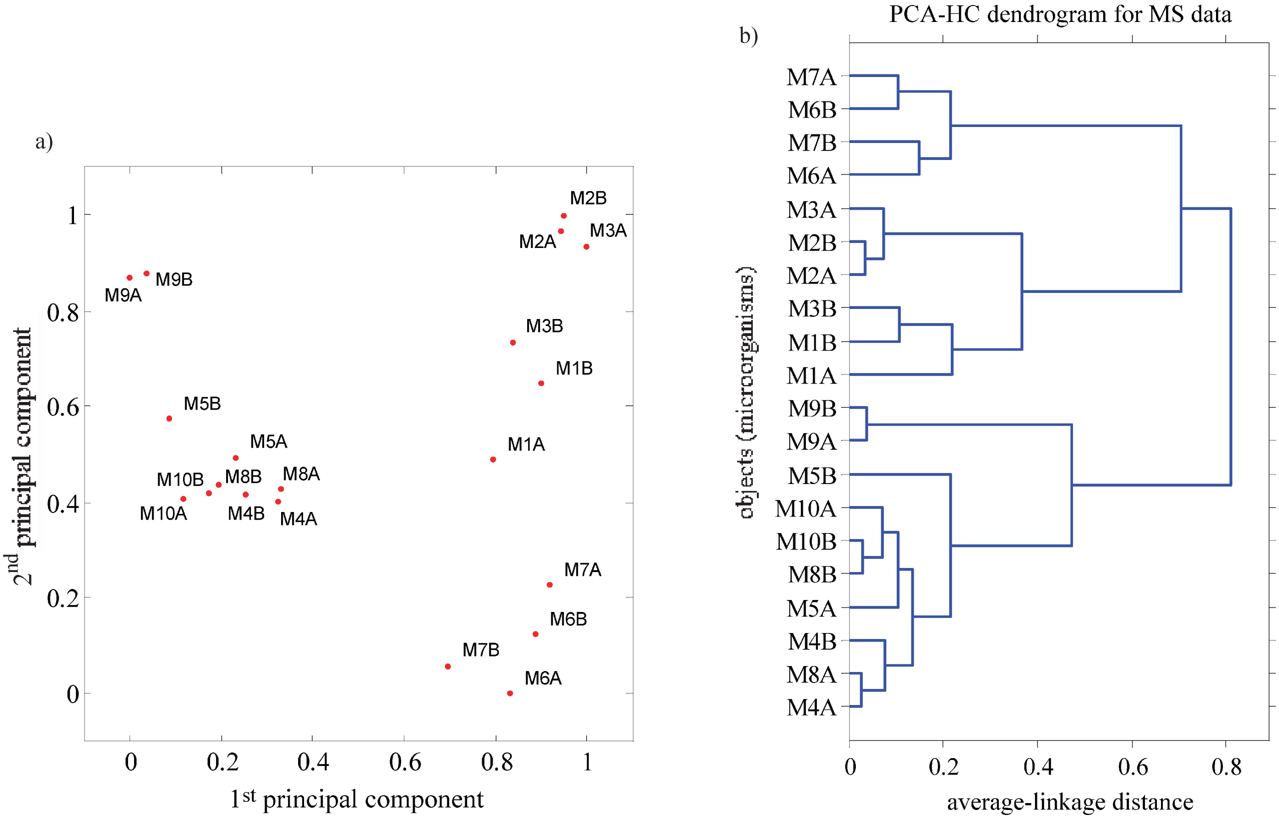

Fig. 3 – *a) PCA (normalized PCs) derived loading plots of microbial footprints of the isolates M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8 and H. salina (5958, M_9) and H. halophila (DSMZ4770, M_{10}) on the PC1-PC2 plane, b) HCA of the microbial footprint data using the similarity measure from PCA*

The first two PCs (PC1 and PC2), plotted on Fig. 3a, explained 95.60 % of the variance in the data set. In order to assess the possible contributions of other PCs they were also plotted as binary pairs (e.g. PC3 vs. PC1, PC3 vs. PC2, etc.), however no further significant clustering information was obtained. This was not unexpected since those further PCs contained only very little information, as evidenced on the Scree plot (Fig. 4). The reason for plotting PC loadings prior to cluster analysis was based on the knowledge that PCs are uncorrelated and ordered, and that the first few PCs contained most of the variation in the data set. However, as seen in Fig. 3a, these first two PCs do not necessarily capture most of the cluster structure.[38] As can be seen from Fig. 3b, most of the replicates did not appear at the leaf nodes adjacent to each other.

Specifically, the confusions in the neighboring leaf nodes of both Fig. 2 and Fig. 3 expose the unreliability of the above approaches used in HCA.

Finally, DFA that allowed groups in the data to be defined has been applied on the retained 8 PCs, which explained 99.25 % of the variance (Fig. 4). Groups were assigned simply on the basis of the replicate number. By defining the groups in this way, the model has essentially been informed that each microorganism is different and those metabolic differences were preserved. Consequently, when the strains clustered together, this demonstrated the presence of real relationship between their metabolome.

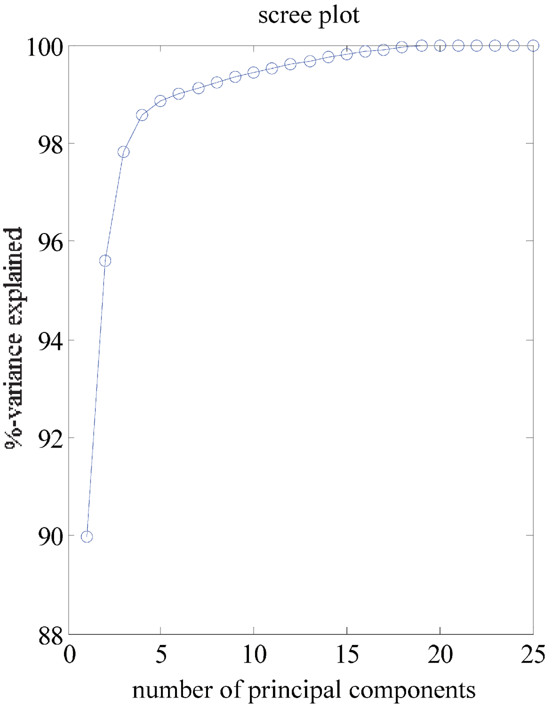

Fig. 4 – *Variance explained in the data set by different number of PCs*

Fig. 5 shows that DFA-based clustering of footprints on the DF1-DF2 plane serves to discriminate the isolates on the metabolome scale without any specific metabolite measurement. Isolates M_4,

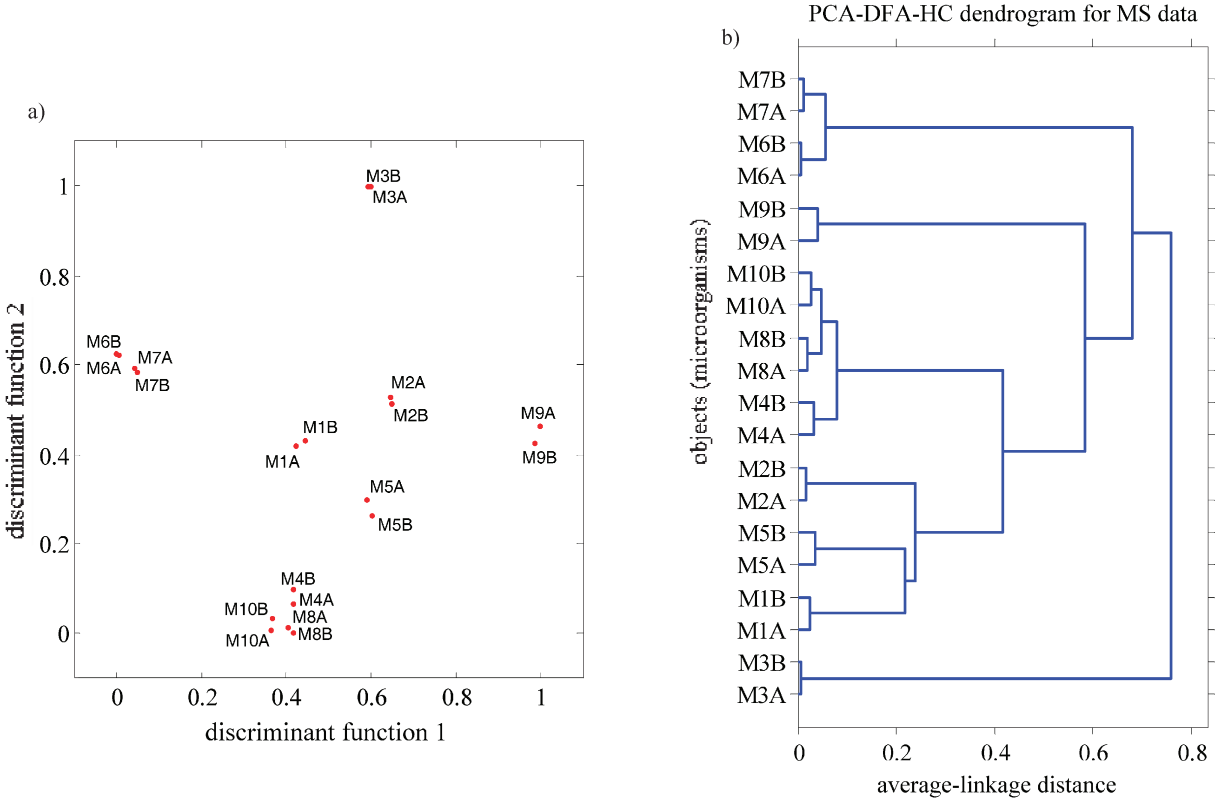

Fig. 5 – *a) Footprint data of the isolates M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8 and H. salina (M_9) and H. halophila (M_{10}) after DFA on the normalized DF1-DF2 plane, b) HCA of the microbial footprint data using the similarity measure from DFA*

M_8, were found to be very closely related to *H. halophila*, an indication for similar metabolic activities. Isolates M_6 and M_7 also shared significant similarities. From the plot of the DF plane, it is clear that these two isolates together with M_3 are the ones that have the greatest difference from the type strains, *H. salina* and *H. halophila*. Though the replicates always appeared close to each other, the rest of the isolates were scattered on the plot, not enabling us to deduce any finer clusters. Hence, DFA plot helped to visualize the rough relationships between the new isolates.

Then, on the DF1-DF2 plane, HCA was carried out for a finer discriminatory analysis and for obtaining the groupings between the isolates and the type strains, and a dendrogram has been constructed (Fig. 5b). The appearance of the replicates on the neighboring leaf nodes in the dendrogram constructed supported the reliability of the approach. As a result, DFA based on PCs was the most powerful approach for clustering without extensive experimental labor.

The analysis showed five clusters. In the first group is the isolate M_3, in the second group are the isolates M_6 and M_7, and in the third group is *H. salina* alone. In the fourth group are the isolates M_4 and M_8, which clustered together with *H. halophila*. Within this group M_8 is more similar to *H. halophila*. In the final group are the isolates M_1, M_2, and M_5. On a slightly rough scale, isolates M_1, M_2, M_4, M_5, and M_8 share a higher similarity with *H. halophila* than isolates M_3, M_6, and M_7. If the number of clusters is reduced to three, both type stains will be clustered with M_1, M_2, M_4, M_5, and M_8. Isolates M_6 and M_7, will be in a second group and isolate M_3 will still be single in its group. It was not possible to extract this specific information by simply investigating the DFA plots in Fig. 5a.

Although phenotypically very similar, *H salina* and *H. halophila* show slight differences in their metabolic capabilities, e.g. for acid production.[39] *H. halophila* produces acid from simple sugars such as glucose, fructose, and galactose but *H. salina* can not produce acid from neither these sugars nor from lactose, sucrose or mannitol.[38] On the other hand, all of the isolates produced acid from glucose, arabinose and mannose, as *H. halophila*. In terms of acid production from simple sugars, M_5 and M_8 were more similar to *H. halophila* than the others since the other isolates could produce acid from lactose. *H. salina* was the most distant to the isolates in this regard. Unfortunately, no fine phenotypic discrimination between the isolates can be obtained from such limited biochemical information. However the dendrogram obtained by performing a statistical analysis using PCA and DFA successfully clustered the isolates based on such metabolic differences.

This is a very attractive alternative to all the classical biochemical tests that would be needed to identify the metabolites/enzymes secreted to the extracellular medium or synthesized and retained within the cell. In the absence of any statistical tools, the above samples should further be analyzed for the presence and the fraction of the metabolites/enzymes. This experimental knowledge would make it possible to comment on which metabolic pathways are switched on/off or which are up/down regulated at a given time. Only then can one understand the metabolic differences and possibly regulatory mechanisms between the isolates. However this is not a trivial job and requires tedious experimentation. The high turnover rate of the metabolites makes the experimental process even more complicated.

Conclusions

The basic biochemical tests carried out with the new isolates were enough to propose that the isolates are very similar to *Halomonas sp.* type strains. However, as an alternative to more extensive biochemical tests, footprints analyzed with statistical approaches enabled us to appreciate the significant differences in the metabolic capabilities in a shorter period. Two of the isolates were found to be very similar to *H. halophila* whereas three were relatively distant from both *H. salina* and *H. halophila*.

This work presents a statistical approach for inferential clustering of new isolates based on the complex mixture of metabolites they leave in the culture broth. The methodology which is based on the application of DFA on the PC space for HCA of microbial footprints may be offered as a routine method for clustering to reduce the number of biochemical tests required for characterization of isolates and identification of their main metabolic differences. Since the outputs analyzed in this work are in the form of spectral data, this methodology will also be applicable to any chemical/biological system producing similar data and it will offer an attractive solution to the structuring of information from high-throughput instruments.

List of symbols

b	– discriminant coefficient
M	– label for the microorganisms
N	– number
Q	– volumetric flow rate, μL min^{-1}, L h^{-1}
w	– mass fraction, %
x	– discriminant variable
X	– matrix
ϑ	– temperature, °C

References

1. *Margesin*, *R*., *Schinner*, *F*., Extremophiles **5** (2001) 73.
2. *Ventosa*, *A*., *Nieto*, *J. J*., *Oren, A*., Microbiol. Mol. Biol. Rev. **62** (1998) 504.
3. *Joo*, *W. A*., *Ki*, *C. W*., J. ChromatogrA. **815** (2005) 237.
4. *Oren*, *A*., J. Ind. Microbiol. Biot. **28** (2002) 56.
5. *Kell*, *D. B*., *Brown*, *M*., *Davey*, *H. M*., *Dunn*, *W. B*., *Spasic*, *I*., *Oliver*, *S. G*., Nat. Rev. Microbiol. **3** (2005) 557.
6. *Scholtz*, *M*., *Gatzek*, *S*., *Sterling*, *A*., *Fiehn*, *O*., *Selbig*, *J*., Bioinformatics **20** (2004) 2447.
7. *Roepstorff*, *P*., Curr. Opin. Biotech. **8** (1997) 6.
8. *Black*, *G. E*., *Fox*, *A*., in *Snyder, P. A.* (Ed.), Biochemical and Biotechnological Applications of Electrospray Ionization Mass Spectrometry, Vol. DCXIX, p. 81, American Chemical Society, Washington DC, 1996.
9. *Liu*, *C. L*., *Hofstadler, S. A., Bresson, J. A., Udseth, H. R., Sukuda, T., Smith, R. D., Snyder, A. P.,* Anal. Chem. **70** (1998) 1797.
10. *Goodacre*, *R*., *Heald*, *J. K*., *Kell*, *D. B*., FEMS Microbiol. Lett. **176** (1999) 17.
11. *Goodacre*, *R*., Microbiol. Eur. **2** (1994) 16.
12. *Goodacre*, *R*., *Trew*, *S*., *Wrigley-Jones*, *C*., *Neal*, *M. J*., *Maddock*, *J*., *Ottley*, *T. W*., *Porter*, *N*., *Kell*, *D. B*., Biotechnol. Bioeng. **44** (1994) 1205.
13. *Kaderbhai*, *N. N*., *Broadhurst*, *D. I*., *Ellis*, *D. I*., *Goodacre*, *R*., *Kell*, *D. B*., Comp. Funct. Genom. **4** (2003) 376.
14. *Timmins*, *E. M*., *Howell*, *A. S*., *Alsberg*, *B. K*., *Noble*, *W. C*., *Goodacre*, *R*., J. Clin. Microbiol. **36** (1998) 367.
15. *Vaidyanathan*, *S*., *Rowland*, *J. J*., *Kell*, *D. B*., *Goodacre*, *R*., Anal. Chem. **73** (2001) 4134.
16. *Zhao*, *H*., *Parry*, *R. L*., *Ellis*, *D. I*., *Griffith*, *G. W*., *Goodacre*, *R*., Vib. Spectrosc. **40** (2006) 213.
17. *Sneath*, *P. H. A*., *Mair*, *N. S*., *Sharpe*, *M. E*., *Holt*, *J. G*., Bergey's Manual of Determinative Bacteriology, Vol. 2, Williams and Wilkins, Baltimore, 1986.
18. *Tamer*, *A. Ü*., *Uçar*, *F*., *Ünver*, *E*., *Karaboz*, *İ*., *Busalıoğlu*, *M*., *Oğultekin*, *R*., Mikrobiyoloji Laboratuvarı Klavuzu, Vol. 55, Baskı Ege Üniversitesi Fen Fakültesi Baskısı, Teks., İzmir, 1989, pp 260.
19. *Oren*, *A*., *Litchfield*, *C. D*., FEMS Microbiol. Lett. **173** (1999) 353.
20. *Gerhardt*, *P*., *Murray*, *R. G. E*., *Costilow*, *R. N*., *Nester*, *E. W*., *Wood*, *W. A*., *Krieg*, *N. R*., *Philips*, *G. B*., Manual of Methods for General Bacteriology. American Society for Microbiology, Washington, D.C., 1981.
21. *Prescott*, *L. M*., *Harley*, *J. P*., *Klein*, *D. A*., Microbial Growth and Metabolism and Microorganisms and the Environment, in Microbiology. Iowa, 1993.
22. *Gonzales, C., Gutierrez, C., Ramirez, C.,* Can. J. Microbiol. **24** (1978) 710.
23. *Amoozegar*, *M. A*., *Malekzadeh*, *F*., *Malik*, *K. A*., J. Microbiol. Meth. **52** (2003) 353.
24. *Sariyar-Akbulut*, *B*., *Salman-Dilgimen*, *A*., *Ceylan*, *S*., *Perk*, *S*., *Denizci*, *A. A*., *Kazan*, *D*., Arch. Microbiol. **189** (2008) 19.
25. *Bakshi*, *B. R*., AIChE J. **44** (1998) 1596.
26. *Eriksson*, *L*., *Johansson*, *E*., *Kettaneh-Wold*, *N*., *Wold*, *S*., Multi- and Megavariate Data Analysis; Principles and Applications. Chapter 3, Umetrics AB, Umea, Sweden, 2001.
27. http://faculty.chass.ncsu.edu/garson/PA765/discrim.htm
28. http://www.statsoft.com/textbook/stathome.html
29. *Dixon*, *W. J*., Biomedical Computer Programs. University of California Press, Los Angeles, 1975.
30. *MacFie*, *H. J. J*., *Gutteridge*, *C. S*., *Norris*, *J. R*., J. Gen. Microbiol. **104** (1978) 67.
31. *Goodacre*, *R*., *Timmins*, *E. M*., *Burton*, *R*., *Kaderbhai*, *N*., *Woodward*, *A. M*., *Kell*, *D. B*., *Rooney*, *P. J*., Microbiol. **144** (1998) 1157.
32. http://personalpages.manchester.ac.uk/staff/Roy.Goodacre.
33. *Härdle, W., Hlávka, Z., Klinke, S*., XploRe – Application Guide, Springer, 2000. (Downloadable from www.quantlet.com/mdstat/scripts/xag/html).
34. *Jain*, *A. K*., *Murty*, *M. N*., *Flynn*, *P. J*., ACM Comput. Surv. **31** (1999) 264.
35. *Akman, U., Okay, N., Hortacsu, O*., Korean J. Chem. Eng. **25** (2008) 329.
36. *Sanchez-Porro*, *C*., *Martin*, *S*., *Mellado*, *E*., *Ventosa*, *A*., J. Appl. Microbiol. **94** (2002) 295.
37. *Lee*, *J.-C*., *Jeon*, *C. O*., *Lim*, *J.-M*., *Lee*, *S.-M*., *Lee*, *J.-M*., *Song*, *S.-M*., *Park*, *D.-J*., *Li*, *W.-J*., *Kim*, *C.-J*., Int. J. Syst. Evol. Micr. **55** (2005) 2027.
38. *Mata*, *J. A*., *Martínez-Cánovas*, *J*., *Quesada*, *E*., *Béjar*, *V*., Syst. Appl. Microbiol. **25** (2002) 360.
39. *Lim*, *J.-M*., *Yoon*, *J.-H*., *Lee*, *J.-C*., *Jeon*, *C. O*., *Park*, *D.-J*., *Sung*, *C*., *Kim*, *C. J*., Int. J. Syst. Evol. Microbiol. **54** (2004) 2037.