Korean J. Chem. Eng., **26**(6), 1645-1651 (2009)
DOI: 10.1007/s11814-009-0243-2

Rapid differentiation of new isolates with MALDI-TOF mass spectrometry via discriminant function analysis based on principal components

Berna Sariyar-Akbulut†

Department of Bioengineering, Marmara University, Kadikoy 34722, Istanbul, Turkey
(*Received 17 December 2008 • accepted 29 March 2009*)

Abstract−Discriminant function analysis based on principal components was applied to the spectral outputs of whole cell suspensions of nine isolates from matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) mass spectrometry. First, based on the salt tolerance and whole cell proteins, the similarity of the isolates to moderate halophiles was established. Intact microorganisms were then inferentially clustered by MALDI-TOF mass spectroscopy taking four type strains as precursors. Two of these type strains were moderate halophilic bacteria (*Halomonas salina* and *Halomonas halophila*), one was a mesophilic bacteria (*Escherichia coli*), and one was a halophilic archaea (*Haloarcula vallismortis*). Results showed that the isolates were significantly similar to halophiles but were different from a mesophile. This investigation demonstrates the feasibility of using whole cell suspensions for rapid differentiation prior to extensive experimentation.

Key words: Hierarchical Clustering, Discriminant Function Analysis, Principal Component Analysis, Intact Microorganisms, MALDI-TOF-MS

INTRODUCTION

The growing demand for stabilizer metabolites and proteins stable under high salt concentrations in different industrial applications has focused attention on halophiles, microorganisms which live under extreme saline conditions. Within this group, moderate halophiles receive special interest owing to their growth in a wide range of salt concentrations.

The importance of halophilic microorganisms has increased investigations on different extreme environments. New microorganisms can be isolated at high speeds from these environments, but the time and financial limitations hamper the efforts for the characterization and clustering of the isolates [1]. The bottleneck in this process is the requirement for various biochemical tests and molecular diagnostic tools. Even with a molecular identification approach, multiple tests may be necessary. For routine purposes, the ideal method of microbial characterization requires minimum sample preparation and direct analysis, which is rapid, automated and inexpensive. The developments in analytical instrumentation, physico-chemical spectroscopic methods fulfilled these requirements [2]. The advent of soft ionization techniques such as matrix-assisted laser desorption ionization, and electrospray ionization has enabled the mass spectrometric analysis of both large molecular weight compounds [3] and low molecular weight metabolites [4].

Matrix-assisted laser desorption ionization-time-of-flight mass spectrometry (MALDI-TOF-MS) has been reported as a simple, reproducible, and inexpensive tool for characterization, identification, and typing of bacteria [5-11]. Briefly, a sample is mixed with a matrix solvent and co-crystallized onto a MALDI target plate. The matrix absorbs energy at a wavelength of a laser, releasing it into the sample as heat. This causes the sample to vaporize and form singly charged ions, which can then be mass analyzed, typically in a TOF mass analyzer, which determines an ion's mass based on the time required to reach the detector. The result is shown as a mass spectrum, in which the mass-to-charge is plotted against the intensity [11,12]. This technique allows the observation of only high-abundant peptides and proteins that are of low mass and ionize readily [13].

Different environmental conditions, mutations and natural variations cause differences in the large number and type of peptides and proteins synthesized, but they differ reproducibly even in closely related strains [14]. Therefore, for identification purposes the mass profiles have been used as fingerprints to compare with those other organisms (without knowing which components were ionized from the cell.) [15,16]. This challenging method is appreciated in microbiology due to its simple sample treatment and the easily interpretable spectra [5]. The analysis of intact cells renders the method even simpler since it does not require any separation or extraction steps.

Multidimensional spectra from these high-throughput instruments give quantitative information about the total biochemical composition of a sample, but their interpretation requires unsupervised pattern recognition methods. Goodacre et al. [17] have shown that the combination of unsupervised- and supervised-learning methods, such as principal component analysis (PCA), independent component analysis, linear discriminant function analysis (DFA or LDA), artificial neural networks and hierarchical cluster analysis (HCA), is a powerful method to overcome this problem. With the combination of DFA and principal components, it is possible to seek for clusters and group the microorganisms based on their perceived closeness for microbial similarities. This approach has been shown to be useful for rapid characterization of microorganisms with greater accuracy, selectivity, and sensitivity [2,17-21].

The present report examined the potential of intact-cell MALDI-

†To whom correspondence should be addressed.
E-mail: bsariyar@eng.marmara.edu.tr

TOF-MS for a rapid differentiation of halophilic isolates. Nine microorganisms, isolated from soil samples collected from Çamaltí Saltern, were first characterized as moderately halophilic based on their salt tolerance and whole cell proteins on 2-dimensional electrophoretic (2-DE) gels. Then taking *Halomonas salina*, *Halomonas halophila*, *Haloarcula vallismortis* and *E. coli* K12 as precursors, the spectral data of the microorganisms from MALDI-TOF-MS were used to inferentially classify the isolates based on their location in the hierarchical tree coupling HCA with DFA and PCA.

MATERIALS AND METHODS

1. Chemicals and Bacterial Strains

The nine microorganisms of this study were obtained from soil samples collected at Çamaltı Saltern. Soil samples were serially diluted in 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 (at 39 °C) based on shape, size and pigmentation. Type strains *Halomonas salina* (DSMZ 5928) and *Halomonas halophila* (DSMZ 4770), *Haloarcula vallismortis* (DSMZ 3756) were purchased from DSMZ. *E. coli* K12, was from our laboratory stock. Chemicals used were from Merck AG (Darmstadt, Germany) and Sigma Chem. Ltd. (USA).

2. Growth Conditions

The cells were grown in shaking cultures (at 37 °C and 180 rpm) in Brown medium containing tri-sodium citrate (0.3%), $MgSO_4$ (2%), KCl (0.2%), and yeast extract (0.5%). NaCl concentration was kept at 1% for *E. coli* K12 and 10% for the rest of the microorganisms. Initial pH was adjusted to 7.5. Cells were harvested at 12,000 rpm for five minutes. For the analysis of whole cell proteins, harvested cells were washed with 50 mM Tris buffer. For MALDI-TOF-MS analysis harvested cells were washed with 2% NH_4Cl and refrigerated until further analysis.

3. Analysis of Whole Cell Proteins

The proteins were extracted with Sigma, PROT-TOT kit, as described by the manufacturer. Grown cells were disrupted by multiple cycles of freeze-thaw and then the resulting cell suspension was centrifuged. Following reduction of the proteins in the supernatant by tributylphosphine, the proteins were alkylated by iodoacetamide. After pelleting the insoluble material, this protein extract was analyzed on 2-DE gels. High-resolution 2-DE gels were prepared with the nonequilibrium pH gel electrophoresis (NEPHGE) technique in the first dimension, where proteins move at different rates across the gel owing to their charge [22], and sodium dodecyl sulphate polyacrylamide gel electrophoresis (SDS-PAGE) in the second dimension, where migration is determined by molecular weight. [23]. To assure reproducible patterns the accumulated volt hours were kept consistent. pH range of the gels was optimized as 3-7 based on the acidic character of the samples. 80 μg of sample was loaded to the 0.9 mm thick first-dimension gels. Following electrophoresis the gels were stained with silver as described by Blum et al. [24].

4. Target Plate Preparation

A loopful of cells was resuspended to an optical density of 1.0 (at 600 nm) with 2% NH_4Cl and mixed at a 1 : 1 ratio with 30 : 70 (v/v) acetonitrile containing 0.1% trifluoroacetic acid (TFA) to prevent contamination. 1 μl samples were spotted on the target plate and allowed to air dry. Subsequently, 1 μl matix solution was applied on the dried spots. Two different matrices have been used for analysis, α-cyano-4-hydroxycinnamic acid (CHCA) and 5-choloro-2-mercaptobenzothiazole (CMBT), but since only CHCA gave reproducible results it was used in further studies. The matrix solution consisted of 1 : 1 : 1 ratio of acetonitrile: methanol containing 0.1% formic acid: 0.01 M crown ether (18-crown-6). Each sample preparation was spotted at least in triplicate to test reproducibility. After one hour, the target plate was analyzed.

5. Data Acquisition

The target plates were analyzed automatically with the MALDI-TOF-MS (Waters/Micromass, USA) for which the reflective option was chosen. A nitrogen laser (λ=337 nm) was used and the laser intensity was set just above the threshold for ion production. The mass spectrometer was used in the positive ion mode with an acceleration voltage of +15 kV. Mass spectra were obtained in a molecular mass range of 500 to 3,000 kDa. Final spectra were generated by summing 50 spectra per sample produced from 10 laser shots up to 20 places per well with an acquisition rate of 0.5 ns. The TOF measurements were converted to m/z values with corresponding intensities, i.e., raw spectra, using Masslynx 4.0 software. These raw spectra were exported as peak lists to be used in statistical applications.

6. Chemometric Methodology

Data consisted of spectral results as *intensity* versus *m/z* values for the 13 microorganisms. Data were labeled as M_i^k (i=1, …, 13), where the index i designates microorganisms and the letter k designates different analysis of identical samples (k=A, B, C, D). The assignment of the letters was random. In the raw spectral data, there were no significant peaks in the m/z<500 and m/z>1,200 regions. Therefore, sections of the data outside this region were removed. PCA requires columns to have equal lengths and common m/z values, but the number of rows of the spectral raw data was variable and each column had irregularly spaced m/z values. Matlab's *interp1* function with the *nearest-neighbor* interpolation option was used to interpolate and resample each column to produce regularly spaced 36,355 identical m/z values between 500 and 1,200. Baseline correction (de-trending) was done with Matlab's (version 2008a) *msbackadj* function using the following options: *WindowSize=2*, *StepSize=4*, *PreserveHeights=true*, *RegressionMethod=pchip*. Each column of the resampled and baseline-corrected data was normalized to 0-1 range. These methods applied preserved the peak positions and heights.

Thus, the data could be represented as a matrix, $X=[\, M_1^A \,..\, M_{13}^A | M_1^B \,..\, M_{13}^B | M_1^C \,..\, M_{13}^C | M_1^D \,..\, M_{13}^D]$, with 36,355 rows and 49 columns (for microorganisms i=5, 7, and 11 only three different (replicate) analyses (labeled as k=A, B, and C) were present).

The dimension of the preprocessed data was reduced by PCA, which uses simple mathematical concepts such as standard deviation, covariance, eigenvectors and eigenvalues. By linear projection PCA 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 information along the first few PCs. Clearly, as the number

of PCs increases, a larger fraction of the total information is accounted for [25,26]. PCA was performed on the rows of the 36,355×49 data matrix, and the 36,355 rows were represented by 30 PCs. Thus, the dimension of the raw data matrix was reduced to 30×49 by PCA-projection.

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 variables are found. Then using a linear combination of these discriminating variables such that $L=b_1x_1+b_2x_2+...+b_nx_n+c$, where the b's are discriminant coefficients (DCs), the x's are discriminating variables, and c is a constant, a discriminant function (DF) is created. Those variables with the largest DCs are the ones that contribute most to the prediction of group membership. DFs will be independent, 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* (eigenvalues 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 [28]. 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 DFs [28].

DFA is not performed on the original spectra because one cannot feed collinear 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

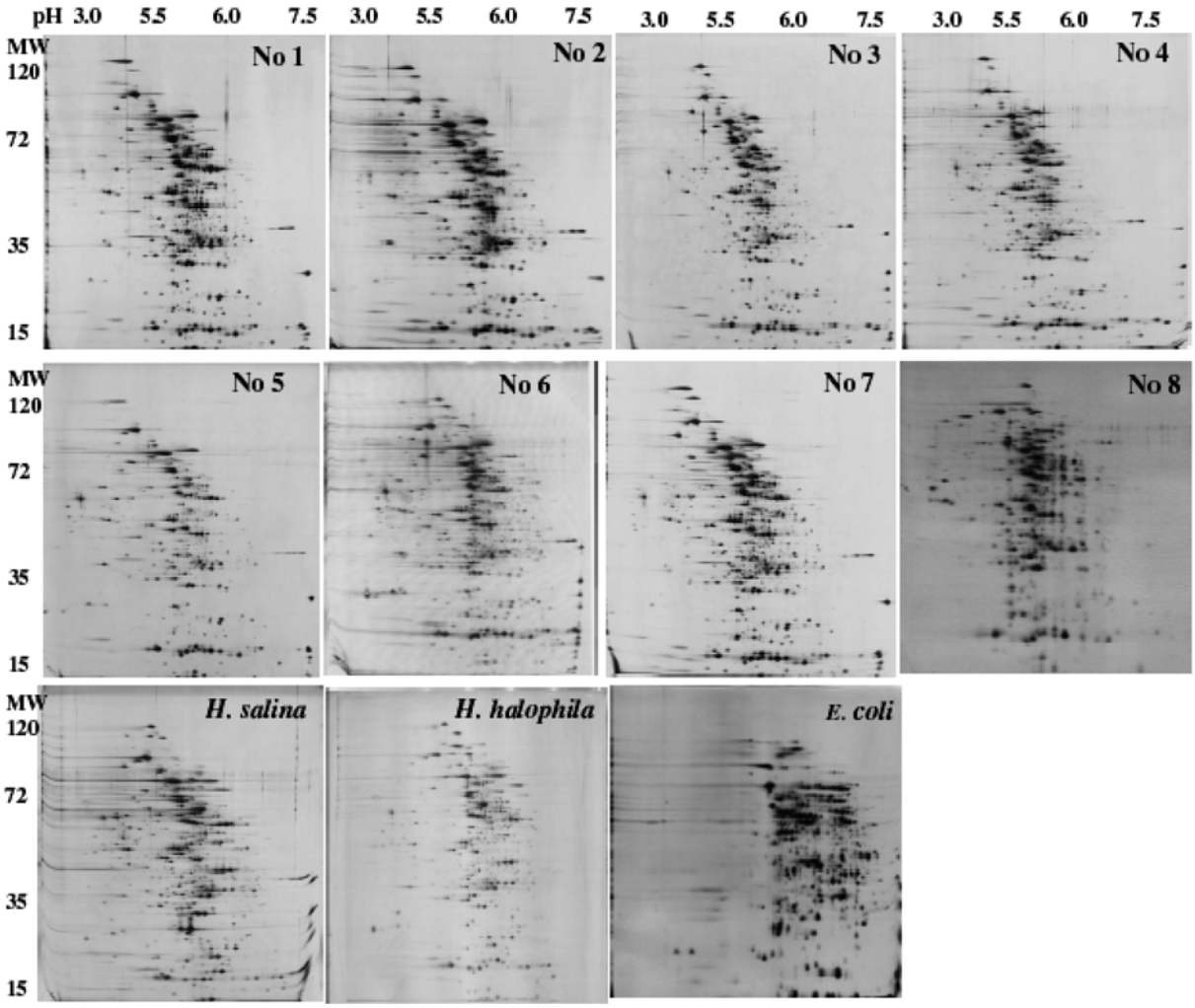

Fig. 1. Whole protein profiles of the isolates and *H. salina* (M_9), *H. halophila* (M_{10}), *E. coli*.

[29,30]. Generally, $(N_s-N_g-1)>N_v$, 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. [19] and Goodacre et al. [17]).

Linear DFA was used to discriminate the isolates on the basis of the retained 30 PCs (which explained the 99.3% of the variance of the 36,355 rows) and the a priori knowledge of which spectra were replicates. DFA toolbox developed for Matlab by Goodacre et al. [31] was used on the PCA-transformed 49×30 matrix, defining the membership for 49 features.

Euclidean distance (ED) in the space of entire original spectral data or between a priori group centers in PCA or DFA space was used to construct a similarity measure, and the distance measures were processed by HCA to construct a dendrogram. In distance-based clustering, two or more objects belong to the same cluster if they are close according to a given distance. HC algorithm [32] finds the closest pair of clusters and merges them into a single cluster, decreasing the number of clusters. The clusters created are viewed on a dendrogram. For HCA, using Matlab's Statistics Toolbox, similarity or dissimilarity between every pair of columns in the data set was found by calculating the distance between columns using *pdist* function with the ED metric option. To group the objects into a binary HC tree, the pairs of microorganisms that were in close proximity were linked together by using *linkage* function (*average linkage*) with the above distance information. A detailed Matlab-based application of HCA is found in Akman et al. [33].

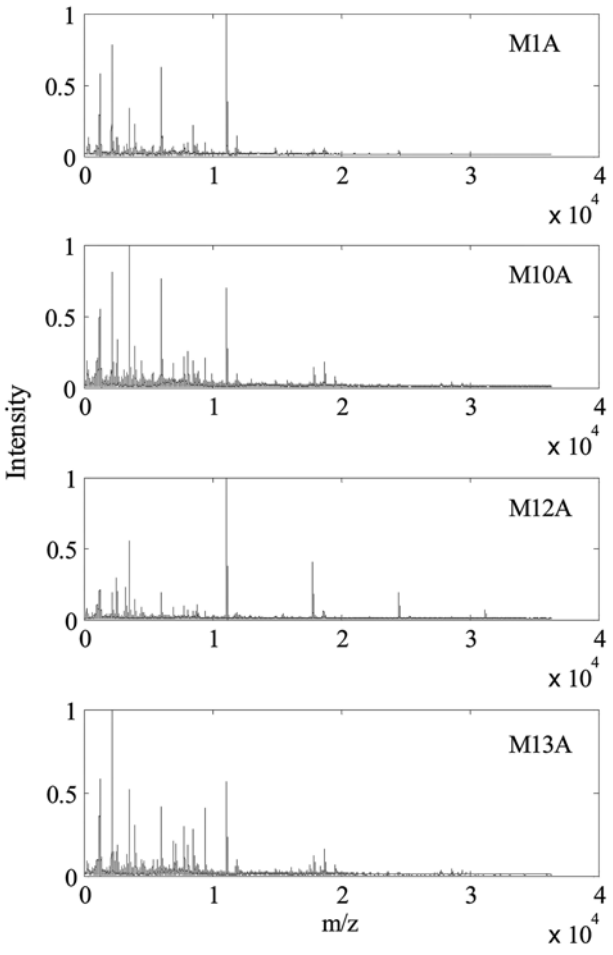

Fig. 2. Normalized raw spectral data from MALDI-TOF-MS of isolate M_1, *Halomonas salina* (DSMZ 5958, M_{10}), *E. coli* K12 (M_{12}) and *Haloarcula vallismortis* (DSMZ 3756, M_{13}).

RESULTS AND DISCUSSION

1. Salt Tolerance and Protein Profiles of Microorganisms

All of the isolates required salt for growth and could tolerate salt concentrations up to 20%. This was similar to the conditions for moderately halophilic type strains *H. salina* and *H. halophila* which could grow at salt concentrations between 2-20% and 1-20%, respectively. In addition, whole cell protein maps of *H. salina* and *H. halophila* were also similar to the new isolates (Fig. 1). The proteome analysis of the whole cell extracts of the isolates showed that more

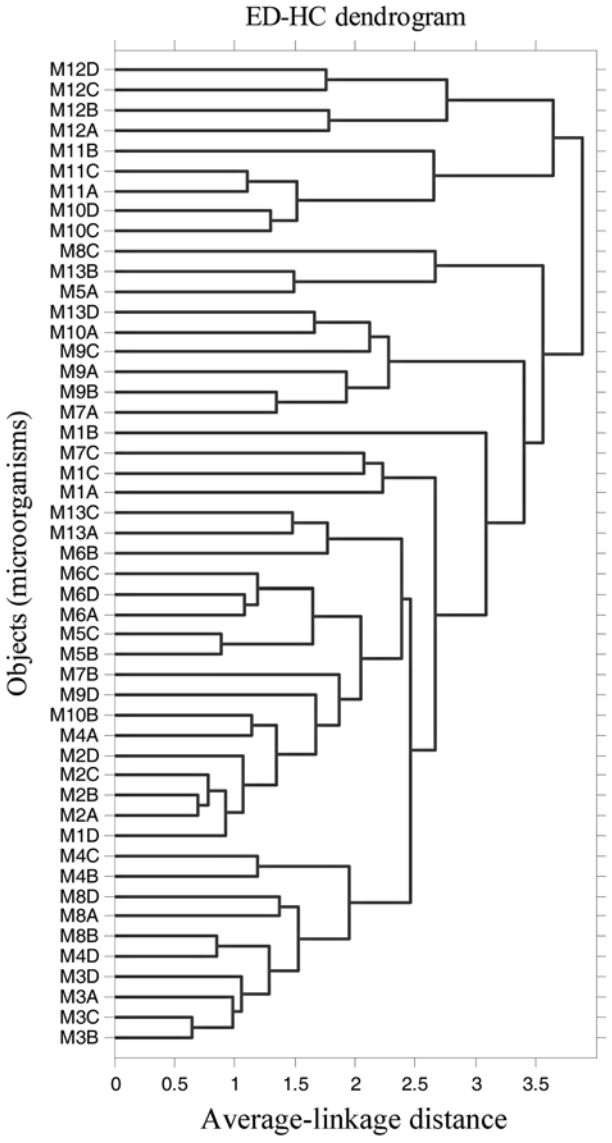

Fig. 3. HCA of the microbial fingerprint data using the ED between spectra as a similarity measure.

than 95% of their proteins were in the acidic region; between pH 4-6 on the 2-DE gels. As opposed to this finding, when *E. coli* K12 proteins were analyzed in the same pH region, all its proteins were found to be above the pH 5.5 region. This was not surprising since *E. coli* proteins are commonly analyzed on gels in the pH range of 4.5-8.5. Therefore the phylogenetic similarity of the isolates to the moderately halophiles was clear.

In addition to accumulating organic solutes such as ectoine, betaine, and glucosylglycerol to establish osmotic balance [34], proteins from halophilic microorganisms are in general acidic in order to tolerate high salt concentrations [35]. Therefore, it was not very surprising to find the halophilic isolates to have a large excess of acidic amino acids, compositions similar to those also found in halophilic archaea [36]. The results obtained in this work are in agreement with the reports on the characteristics of other halophilic microorganisms [37]. As the salt tolerance of a halophilic microorganism increases, the acidic character of the proteins in that organism increases with more aspartic acid than glutamic acid residues [38]. This relationship between acidity and salt interest is important for protein stabilization in such environments. The acidic amino acid residues found in abundance on the protein surfaces bind in a network of hydrated salt ions that cooperatively contribute to the stabilization of the protein [38].

2. Hierarchical Clustering Analysis

The isolates showed identity to moderately halophilic microorganisms. For this reason *H. salina* and *H. halophila* were taken as precursors for inferential clustering of the new isolates based on their protein mass profiles. Protein mass profiles of a halophilic archaea, *Haloarcula vallismortis*, and a mesophilic bacteria, *E. coli* K12, were also included in the chemometric analysis to compare the closeness of the isolates to such diverse microorganisms. The qualitative spectral difference was not significant especially among the halophiles (Fig. 2). This illustrated the need for a multivariable analysis since visualization of the spectra was not sufficient for clustering.

When simple ED in the space of entire original spectral data (in resampled form) was used to perform the HCA, the HC-ED did not produce interpretable results due to misalignments of the default footprint signaling instances (Fig. 3). The unavoidable peak misalignments of the replicates were not noticeable to the eye but 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. Most of the replicates on Fig. 3 did not appear at the neighboring leaf nodes.

To improve the clustering results, PCA was conducted on the resampled data since PCA is a statistical technique for finding patterns in data matrix of high dimensions to highlight similarities and differences. The knowledge that PCs are uncorrelated and ordered and that the first few PCs contain most of the variation in the data

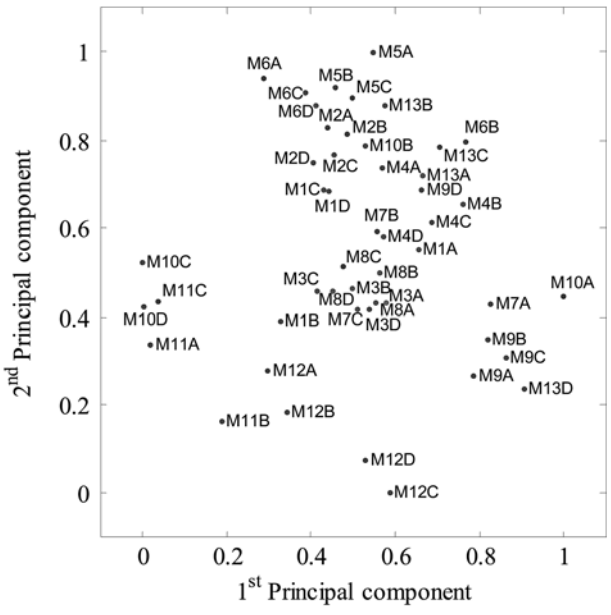

Fig. 4. PCA (normalized PCs) derived loading plots of microbial fingerprints of the isolates M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9 and *H. salina* (DSMZ 5958, M_{10}), *H. halophila* (DSMZ 4770, M_{11}), *E. coli* K12 (M_{12}) and *Haloarcula vallismortis* (DSMZ 3756, M_{13}) on the PC1-PC2 plane.

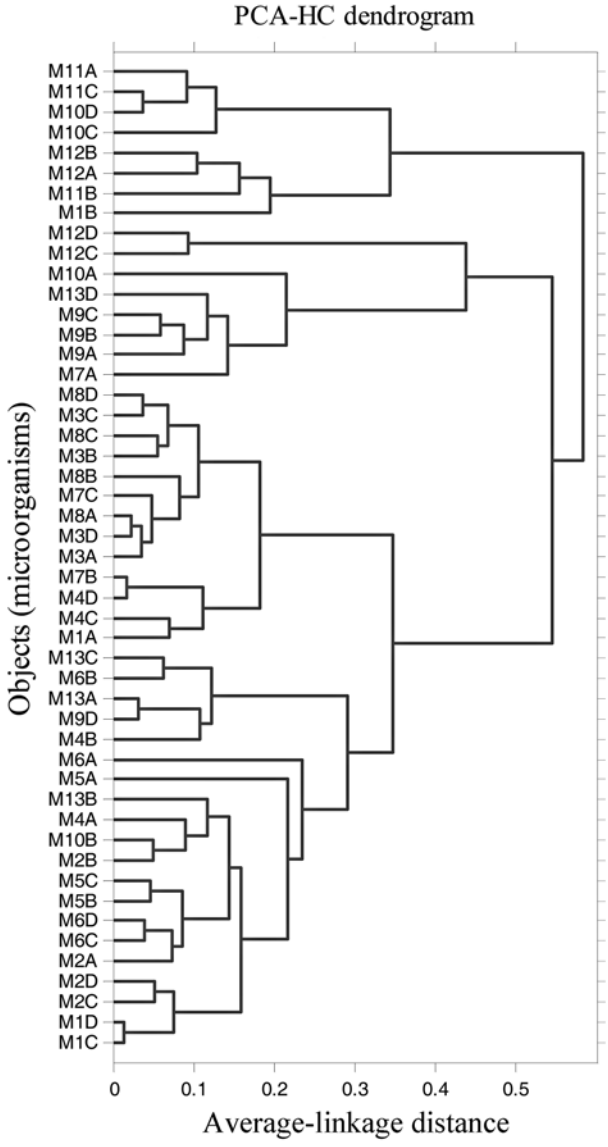

Fig. 5. HCA of the microbial fingerprint data using the similarity measure from PCA.

set has urged us to plot PC loadings prior to cluster analysis. However, the results showed that these first few PCs do not necessarily capture most of the cluster structure, as also reported by Yeung and Ruzzo [39]. The PCA results plotted on the PC1-PC2 plane (Fig. 4) showed that the grouping was still not very informative for detailed microbial clustering. Hence the attempts for construction of an interpretable dendrogram failed (Fig. 5).

Finally, DFA that allowed groups in the data to be defined was then applied on the retained 30 PCs. Groups were assigned simply on the basis of the replicate number. By defining the groups in this way, the model had essentially been informed that each microorganism was different and those differences in protein expression profiles were preserved. Consequently, when the strains clustered together, this demonstrated the presence of a real relationship between their mass spectral fingerprints.

Simple DFA-based clustering of fingerprints served to differentiate the intact cells on the DF plane (Fig. 6). The mesophilic microorganism *E. coli* K12 (M_{12}) is clearly different from the halophiles, *H. salina* (M_{10}), *H. halophila* (M_{11}) and *Haloarcula vallismortis* (M_{13}). This finding is not very surprising since the analysis of intact cells by MALDI-TOF-MS is based on the observation of ionizable high-abundant peptides and proteins which clearly reside on the surface of the cell membrane. Since the surface proteins of halophiles, no matter what the microbial origin is, archaea or bacteria, share the common feature of having a large number of acidic residues, the halophiles are grouped together away from *E. coli* K12. Within the halophilic group, *H. salina* and *H. halophila* were closer to each other than *Haloarcula vallismortis*, but the isolates were scattered between these three halophilic type strains and no exact information could be extracted from the plot of the DF1-DF2 plane.

When HCA was carried out on the DF1-DF2 plane, it was possible to make a finer discriminatory analysis to obtain the groupings between isolates and type strains. The dendrogram thus constructed is given in Fig. 7. Most of the replicates appeared on the neighboring leaf nodes, which shows the reliability of the approach. The analysis of the constructed dendrogram showed that *E. coli* K12 was distinctively different from the halophiles. On the other hand within the halophiles, there were three main groups that could not be deduced simply from the plot on the DF1-DF2 plane. The clustering in the first group suggested that *H. salina* is more similar to the isolates M_4, M_5, and M_9. The replicates were perfectly discriminated in this group. The second group encompassed *H. halophila* and the isolates M_1, M_2, and M_6. As opposed to the first group, the leaf nodes of the dendrogram were not very distinctive for discrimination of the replicates of the microorganisms. This finding suggests that they are highly related species. The final group had the isolates M_3, M_7, M_8, and *Haloarcula vallismortis*. This provides a clue for the similarity of the protein mass fingerprints of the isolates

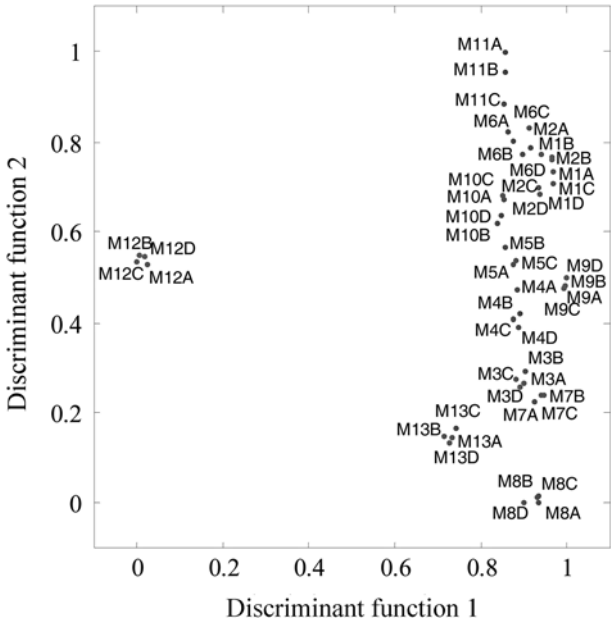

Fig. 6. Footprint data of the isolates M_1, M_2, M_3, M_4, M_5, M_6, M_7, M_8, M_9 and *H. salina* (DSMZ 5958, M_{10}), *H. halophila* (DSMZ 4770, M_{11}), *E. coli* K12 (M_{12}) and *Haloarcula vallismortis* (DSMZ 3756, M_{13}) after DFA on the normalized DF1-DF2 plane.

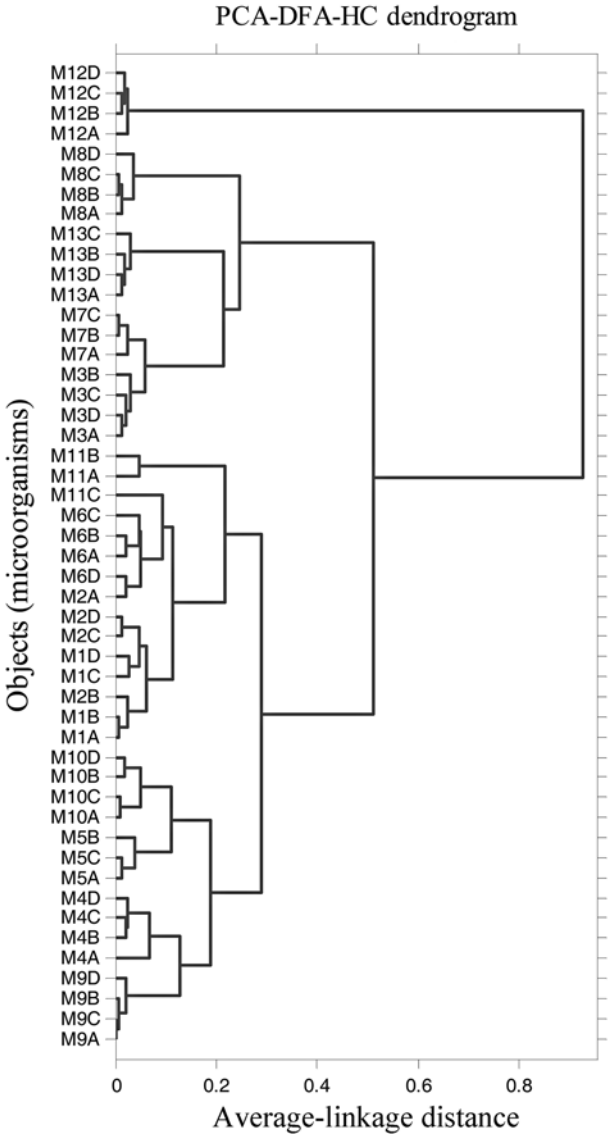

Fig. 7. HCA of the microbial fingerprint data using the similarity measure from DFA.

M_3, M_7, and M_8 to archaea. However, further biochemical analyses are definitely required for verification.

CONCLUSIONS

This work presents the application of PCA-DFA-HCA on the mass spectral microbial fingerprints from MALDI-TOF-MS for inferential clustering of new isolates. Since the effectiveness of the PCA-DFA-HC method has been shown in microbial clustering [17-20] we used this method to inferentially classify nine new isolates with confidence from earlier success. When HCA was performed neither with simple ED in the space of entire resampled data nor with PCs obtained from PCA of the resampled data, it was possible to obtain interpretable groupings. In addition, DFA could not be performed on the original data since the data formed a large set with high collinearity. On the other hand, the combined method of PCA-DFA-HCA could sufficiently discriminate the microorganisms. Therefore, this method has shown to be an attractive tool for rapid differentiation of new isolates based on their perceived closeness on the proteome level prior to extensive experimentation. In other words, it is a powerful preliminary dry-lab approach as opposed to the requirement of many biochemical tests. It cannot replace the wet-lab approach; however, this statistical approach may help to overcome the bottleneck in the process of characterization of new isolates.

REFERENCES

1. A. Oren, *J. Ind. Microbiol. Biot*., **28**, 56 (2002).
2. R. Goodacre, S. Trew, C. Wrigley-Jones, M. J. Neal, J. Maddock, T. W. Ottley, N. Porter and D. B. Kell, *Biotechnol. Bioeng*., **44**, 1205 (1994).
3. P. Roepstorff, *Curr. Opin. Biotech.*, **8**, 6 (1997).
4. G. E. Black and A. Fox (1996) In: P. A. Snyder (ed.) *Biochemical and Biotechnological Applications of Electrospray Ionization Mass Spectrometry*, Vol. DCXIX. American Chemical Society, Washington DC, 81-105 (1996).
5. E. Vanlaere, K. Sergeant, P. Dawyndt, W. Kallow, M. Erhard, H. Sutton, D. Dare, B. Devreese, B. Samyn and P. Vandamme, *J. Microbiological Methods*, **75**, 279 (2008).
6. R. Dieckmann, I. Graber, I. Kaesler, U. Szewzyk and H. Von Dohren, *Appl. Microbiol. Biotechnol.*, **67**, 539 (2005).
7. C. J. Keys, D. J. Dare, H. Sutton, G. Wells, M. Lunt, K. McKenna, M. McDowall and H. N. Shah, *Infect. Genet. Evol.*, **4**, 221 (2004).
8. M. P. Kumar, M. Vairamani, R. P. Raju, C. Lobo, N. Abumani, C. P. Kumar, T. Menon and S. Shanmugasundaram, *Indian J. Med. Res.*, **119**, 283 (2004).
9. P. Pribil and C. Fenselau, *Anal. Chem.*, **77**, 6092 (2005).
10. V. Ruelle, B. Moualij, W. Zorzi,, P. Ledent and E. D. Pauw, *Rapid Commun. Mass Spectrom.*, **18**, 2013 (2004).
11. B. L. Van Baar, *FEMS Microbiol. Rev.*, **24**, 193 (2000).
12. J. J. Dalluge, *Fresenius J. Anal. Chem.*, **366**, 701 (2000).
13. A. Fox, *J. Clin. Microbiol.*, **44**, 2677 (2006).
14. D. B. Kell, M. Brown, H. M. Davey, W. B. Dunn, I. Spasic and S. G. Oliver, *Nat. Rev. Microbiol*., **3**, 557 (2005).
15. T. J. Siegrist, P. D. Anderson, W. E. Huen, G. T. Kleinheinz, C. M. Mcdermott and T. R. Sandrin, *J. Microbiol. Methods*, **68**, 554 (2007).
16. R. E. Mandrell, L. A. Harden, A. Bates, W. G. Miller, W. F. Haddon and C. K. Fagerquist, *Appl. Environ. Microbiol.*, **71**, 6292 (2005).
17. R. Goodacre, J. K. Heald and D. B. Kell, *FEMS Microbiol. Lett*., **176**, 17 (1999).
18. J. Allen, H. M. Davey, D. Broashurst, J. K. Heald, J. J. Rowland and S. G. Oliver, *Nat. Biotechnol.*, **21**, 692 (2003).
19. E. M. Timmins, A. S. Howell, B. K. Alsberg, W. C. Noble and R. Goodacre, *J. Clin. Microbiol.*, **36**, 367 (1998).
20. S. Vaidyanathan, J. J. Rowland, D. B. Kell and R. Goodacre, *Anal. Chem.*, **73**, 4134 (2001).
21. H. Zhao, R. L. Parry, D. I. Ellis, G. W. Griffith and R. Goodacre, *Vib. Spectrosc.*, **40**, 213 (2006).
22. P. Z. O'Farrel, H. M. Goodman and P. H. O'Farrel, *Cell*, **12**, 1133 (1977).
23. U. K. Laemmli, *Nature*, **227**, 680 (1970).
24. H. Blum, H. Beier and H. J. Gross, *Electrophoresis*, **8**, 93 (1987).
25. B. R. Bakshi, *AIChE J*., **44**, 1596 (1987).
26. L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, *Multi- and megavariate data analysis; principles and applications*, Umetrics AB, Umea, Sweden (Chapter 3) (2001).
27. http://www2.chass.ncsu.edu/garson/pa765/discrim.htm.
28. http://www.statsoft.com/textbook/stathome.html.
29. W. J. Dixon, *Biomedical computer programs*, Los Angeles:University of California Press (1975).
30. H. J. H. MacFie, C. S. Gutteridge and J. R. Norris, *J. Gen. Microbiol.*, **104**, 67 (1978).
31. http://personalpages.manchester.ac.uk/staff/Roy.Goodacre.
32. A. K. Jain, M. N. Murty and P. J. Flynn, *ACM Comput. Surv*., **31**, 264 (1999).
33. U. Akman, N. Okay and O. Hortacsu, *Korean J. Chem. Eng*., **25**, 329 (2008).
34. E. A. Galinski and P. Louis, *Compatible solutes: ectoine production and gene expression*, In: A. Oren (ed), *Microbiology and biogeochemistry of hyper-saline environments*, CRC Press, Inc. Boca Raton, Fla., 187-202 (1999).
35. J. K. Lanyi, *Bacteriological Reviewa*, **38**, 272 (1974).
36. R Reistad, *Arch. Mikrobiol.*, **71**, 353 (1970).
37. A. Ventosa, J. J. Nieto and A. Oren, *Microbiol. Mol. Biol. R*., **62**, 504 (1998).
38. S. Fukuchi, K. Yoshimune, M. Wakayama, M. Moriguchi and K. Nishikawa, *J. Mol. Biol.*, **327**, 347 (2003).
39. K. Y. Yeung and W. L. Ruzzo, *Bioinformatics*, **17**, 763 (2001).