Application of Multidimensional Scaling in Numerical Taxonomy: Analysis of Isoenzyme Types of Candida Species*

DAVID A. LACHER, M.D., M.ED.‡ and PAUL F. LEHMANN, PH.D.§

‡Department of Pathology
and §Department of Microbiology,
Medical College of Ohio,
Toledo, OH 43699-0008

ABSTRACT

Multidimensional scaling (MDS) was applied to the numerical taxonomy of Candida species based on isoenzyme profiles. Multidimensional scaling uses proximity measures to generate a spatial configuration of points in multidimensional space where distances between points reflect similarity among types. The biochemical profiles of 35 types of Candida species based on 26 tests consisting of isoenzymes of α-glucosidase, alkaline phosphatase, glucose-6-phosphate dehydrogenase, malate dehydrogenase, isocitrate dehydrogenase, and superoxide dismutase were analyzed. Cluster analysis of MDS, using the Euclidean distance as a proximity measure, separated C. tropicalis and C. paratropicalis from C. albicans and C. stellatoidea. Stepwise multiple linear regression revealed the isoenzyme tests which influenced each of the MDS dimensions. MDS was able to reduce the dimensionality of the test profile.

Introduction

Analysis of variation among related microorganisms, for the purpose of constructing numeric taxonomic schemes, is useful in epidemiologic studies. By using a profile of tests, typing schemes can help differentiate person-to-person spread from reinfection of a microorganism. In addition, the analysis of the distribution of types of organisms during different time intervals can be of interest.

Many statistical techniques have been used to classify microorganisms including Bayesian analysis, likelihood models and discriminant analysis.8,9,10,30 These techniques are based on "training" sets of known data and are used to establish rules (equations) to identify unknown organisms. In numerical taxonomy, "unsupervised" statistical techniques are used to establish meaningful relationships between unknown organisms.

Statistical solutions to numerical taxonomy include differential shading of
similarity matrix, cluster analysis and principal component analysis (PCA).\textsuperscript{3,7,12,13,26,31} Previously, multidimensional scaling (MDS) has been used to discriminate four species of Enterobacteriaceae.\textsuperscript{19} Both PCA and MDS examine the relationship between tests (variables) and seek lower dimensional representation of the original data.\textsuperscript{11,16,29} This is done by forming "factors" which are linear combinations of variables that reflect features of the data. Geometrically, a data set of isolates analyzed by multiple tests can be represented as a cloud of points in hyperellipsoid space. To visualize the relationship of the isolates, multidimensional scaling can form a "map" of the organisms in lower dimensional space, and this reflects the similarity of the organisms. Furthermore, multidimensional scaling can reduce the number of tests used to type organisms while preserving the majority of information contained in the original data set.

In this study, the data obtained in a previous study on isoenzyme patterns found within a population of strains of Candida albicans and C. tropicalis have been utilized.\textsuperscript{21} In that study, heterogeneity in the population of Candida species was observed. Strains that failed to assimilate sucrose were included. Because of the use of carbohydrate assimilation patterns in yeast taxonomy, these sucrose-negative variants of Candida have been given distinct names. There are two variants of C. albicans named C. stellatoidea type I and type II,\textsuperscript{17} and C. paratropicalis is the sucrose-negative variant of C. tropicalis.\textsuperscript{1} Using this data set, multidimensional scaling has been combined with cluster analysis to identify groups of Candida types derived from isoenzyme patterns. Regression techniques were applied to identify significant tests which form clusters. Finally, MDS was used to visualize clusters in lower dimensional models and reduce the number of tests needed to form clusters of organisms.

**Materials and Methods**

**Data Base**

Multidimensional scaling was applied to isoenzyme profiles that described the 22 types of C. albicans (38 strains), five types of C. tropicalis (13 strains), four types of C. paratropicalis (10 strains), two types of C. stellatoidea type I (11 strains), and two types of C. stellatoidea type II (three strains). Full details of the protein extraction, gel electrophoresis and isoenzyme detection have been described previously.\textsuperscript{21} The isoenzyme profile for each isolate was scored from the positions of six α-glucosidases (αG), three alkaline phosphatases (ALP), two lactic dehydrogenases (LDH), two sorbitol dehydrogenases (SDH), four glucose-6-phosphate dehydrogenases (G6PDH), seven malate dehydrogenases (MDH), four isocitrate dehydrogenases (ICDH), and 11 superoxide dismutases (SOD).

Isoenzyme bands were coded as 0 if absent or 1 if present. All Candida isolates had complete biochemical profiles. No isoenzyme test was uniformly present or absent for all isolates; thus, all tests had some discriminating power. However, there were tests which correlated perfectly (Pearson $r$ equal to 1 or $-1$) with other isoenzyme tests (table I) and this redundancy allowed the elimination of 13 tests. Therefore, statistical analysis was performed on 26 isoenzyme tests consisting of all of the α-glucosidase and alkaline phosphatase isoenzymes; G6PDH (bands 2 to 4); malate dehydrogenase (bands 2 to 6); isocitrate dehydrogenase (bands 1, 3, and 4); and superoxide dismutase (bands 1 to 5 and 10).

**Statistical Analysis**

Multidimensional scaling was performed using the SPSS-X ALSCAL program.\textsuperscript{32}
TABLE I
CORRELATED ISOENZYME TESTS

<table>
<thead>
<tr>
<th>Pearson r = 1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6PDH₂ : LDH₂, MDH₁, SOD₆, SOD₈</td>
</tr>
<tr>
<td>SOD₁ : SOD₁₁</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pearson r = -1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>G6PDH₂ : LDH₁, SDH₁, SDH₂, G6PDH₁, MDH₁, SOD₇, SOD₉</td>
</tr>
<tr>
<td>ICDH₄ : ICDH₂</td>
</tr>
</tbody>
</table>

G6PDH = glucose-6-phosphate dehydrogenase
LDH = lactic dehydrogenase
MDH = malate dehydrogenase
SOD = superoxide dismutase
ICDH = isocitrate dehydrogenase

Multidimensional scaling requires the input of a matrix of proximity measures. These were generated from the original data using the SPSS-X PROXIMITIES program. Several proximity measures were compared in the MDS analysis including the simple matching coefficient, Euclidean distance, and the phi coefficient. Although there were some minor differences in the MDS results (data not shown), the Euclidean distance was chosen as a proximity measure between each pair of isolates. The range of the Euclidean distance was from zero (indicating isolates with identical isoenzyme patterns) to a maximum of the square of the number of isoenzyme tests which indicated a completely nonidentical pattern. Also, nonmetric MDS analysis was performed by treating the proximity matrix data as ordinal and using Kruskal’s least-square monotonic transformation. There was no difference in the MDS results between the nonmetric analysis of data and the metric analysis using classical regression techniques where proximity measures were treated as ratio data.

Multidimensional scaling of the Candida data base was analyzed in one-, two-, and three-dimensions. The optimal number of dimensions for the MDS analysis was determined by Kruskal’s stress formula and the squared correlation. The MDS scores obtained for each isolate were plotted (figures 1, 2, and 3). Stepwise multiple regression, using MDS dimensional scores as dependent variables and isoenzyme test results as independent variables, was used to determine the tests which best predicted each MDS dimension score. Multiple linear regression was performed using the SAS STEPWISE program. In addition, classification and regression tree analysis (CART) was performed on the original data to determine the important tests for separating the five Candida groups.

Cluster analysis of the original data and MDS scores was used to find clusters among the Candida isolates. Cluster analysis was compared for the original Candida data and the MDS dimensional scores for a three-dimensional nonmetric model. Several cluster methods were compared including average linkage, centroid, complete linkage, flexible-beta, McQuitty’s similarity, median, single linkage, and Ward’s minimum variance. The optimal number of clusters for each cluster method was determined by the cubic clustering criterion and the Pseudo F and t² statistics. Dendrograms of cluster analysis were plotted to determine the structure and origin of each cluster. Cluster analysis was performed using the SAS CLUSTER program.

Results and Discussion

By eliminating 13 perfectly correlated tests (table I), the data consisted of binary results of 26 tests on 35 Candida types. Several statistical approaches to numerical taxonomy of the Candida were then considered. Principal component analysis (PCA), a form of factor analysis, was applied to the Candida data, but a solution could not be
obtained as a result of high correlations among some tests (data not shown); however, PCA indicated that three tests (G6PDH₂, MDH₄, SOD₄) accounted for a significant amount of variance among the Candida isolates.

Multidimensional scaling, using a non-metric, ordinal analysis based on Euclidean distances, was applied to the Candida isoenzyme data to form one-, two-, and three-dimensional models (figures 1, 2, and 3). The values of Kruskal's stress statistic were 0.322, 0.188, and 0.112 for the one-, two-, and three-dimensional models, respectively, indicating an improvement in the "goodness of fit" for higher dimensional models. This was also indicated by the squared correlation which were 0.762, 0.861, and 0.930 for the one-, two-, and three-dimensional solutions, respectively. When the results were plotted (figures 1, 2, and 3), MDS
was capable of separating *C. paratropicalis* and *C. tropicalis* from *C. albicans* and *C. stellatoidea* I and II in each of the three models.

Analysis of the original data and MDS scores using several cluster methods suggested that there were two clusters based on the cubic clustering criterion and the Pseudo F and $t^2$ statistics. There were some differences in the dendrograms produced by the different clustering methods, but all these methods suggested that only two main clusters were present. The dendrogram from Ward's minimum variance cluster analysis of a three-dimensional MDS model is seen in figure 4B. Again, there was clear separation of *C. paratropicalis* and *C. tropicalis* from *C. albicans* and *C. stellatoidea* I and II. For comparison, Ward's cluster analysis was performed on the original data using the original 39 tests (figure 4A). The same two major clusters were found, as had been reported earlier, but the detailed dendrogram of the original data was different from the dendrograms of the MDS model.

Stepwise multiple linear regression of the three-dimensional MDS model was then performed to determine which iso-
enzyme tests best predicted each MDS dimension. The G6PDH$_2$ test group (table I), which best predicted dimension one scores, was negative for C. tropicalis and C. paratropicalis and positive for C. albicans and C. stellatoidea I and II (table II). For dimension two, the best predictor test was G$_3$, which grouped some C. albicans with C. stellatoidea I and II. Finally, for dimension three, G6PDH$_3$ was the best predictor test which was positive for C. tropicalis and negative for C. paratropicalis except for one C. tropicalis type (J) which segregated with C. paratropicalis.

Classification tree analysis (CART) of the original data (figure 5) revealed significant tests similar to the stepwise
multiple linear regression of the MDS three-dimensional model. Again, the G6PDH\(_2\) test group separated \(C.\) \textit{albicans} and \(C.\) \textit{stellatoidea} from \(C.\) \textit{tropicalis} and \(C.\) \textit{paratropicalis}. Both \(\alpha G_3\) and G6PDH\(_3\) separated \(C.\) \textit{paratropicalis} from \(C.\) \textit{tropicalis} types with exception of the \(C.\) \textit{tropicalis} type J. The \(\alpha G_1\) segregated \(C.\) \textit{stellatoidea I} from \(C.\) \textit{albicans} and \(C.\) \textit{stellatoidea II}. It was the second best predictor for dimension one of the three-dimensional MDS model as determined by stepwise regression analysis. For CART, \(\alpha G_2\) separated \(C.\) \textit{stellatoidea II} and three types of \(C.\) \textit{albicans} from 19 other \(C.\) \textit{albicans} types. The CART analysis indicated that \(\alpha G_3\) was a good surrogate test for \(\alpha G_2\). These results supported the findings of Kwon-Chung et al who suggested that \(C.\) \textit{stel-
TABLE II
Stepwise Regression Analysis of Multidimensional Scaling Scores of 3-Dimensional Model

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Best Predictor Test or Group</th>
<th>Candida albi cans</th>
<th>Candida stellatoidea I</th>
<th>Candida stellatoidea II</th>
<th>Candida tropicalis</th>
<th>Candida paratropicalis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>G6PDH2 *</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>Alpha-glucosidases</td>
<td>77</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>G6PDH3</td>
<td>36</td>
<td>0</td>
<td>50</td>
<td>80</td>
<td>0</td>
</tr>
</tbody>
</table>

* Glucose-6-phosphate dehydrogenases: Test group consists of lactic dehydrogenases, sorbitol dehydrogenases, malate dehydrogenases, and superoxide dismutases (see table I).

C. stellatoidea type II was more similar to most C. albicans strains than was C. stellatoidea type I.17

When multidimensional scaling was applied to a profile of Candida isoenzyme tests, it was able to reduce the dimensionality of the test set from 26 to three dimensions. Cluster and regression analysis indicated that the three-dimensional MDS model preserved most of the information found in the original test set. Discriminant analysis or classification tree analysis of MDS scores or the original data could be used to classify unknown isolates. Both MDS and PCA can reduce a large number of variables to a few significant variables to simplify data analysis. The reduced set can be used to place related types of Candida into groups. The G6PDH2 test (group) was most important in separating Candida into two clusters by cluster analysis of both the original data and MDS scores. Use of further isoenzyme

\[ \text{G6PDH}_2 \text{ Neg.} \]
\[ \alpha-G_6 \text{ or G6PDH}_3 \text{ Neg.} \]
\[ \text{C. paratropicalis (4)} \]
\[ \text{C. tropicalis (1)} \]
\[ \text{C. stellatoidea (2)} \]

\[ \text{G6PDH}_2 \text{ Pos.} \]
\[ \alpha-G_6 \text{ or G6PDH}_3 \text{ Pos.} \]
\[ \text{C. tropicalis (4)} \]

\[ \alpha-G_2 \text{ Neg.} \]
\[ \alpha-G_2 \text{ Pos.} \]
\[ \text{C. albicans (19)} \]
\[ \text{C. stellatoidea II (2)} \]
\[ \text{C. albicans (3)} \]

**Figure 5.** Classification tree analysis of original data of Candida isolates. G6PDH2 refers to a group of 12 correlated tests (table I).
tests failed to subdivide the populations of these two species into distinct races; however, these tests can be used to type strains. 

While the current study applies MDS to the analysis of populations of yeast species with variant isoenzyme profiles, the same analytical procedures may be used to determine groupings of yeasts described using other data sets. Previously, both factor analysis and cluster analysis have been used to group species within yeast genera, for example in Pichia and Candida. However, these studies were based on comparisons either of a single strain of each species, or of the species’ taxonomic descriptions where most of the data had been obtained from physiological and biochemical tests.

In contrast, there have been relatively few studies involving numerical taxonomy that have been applied to variation within a yeast species. Variation in phenotypic characteristics has been studied for fairly large groups of C. albicans strains. In some of these studies, cluster analysis has suggested the presence of some structure within the yeast population with the development of distinct races of C. albicans. Using 64 physiological, biochemical, and morphological characteristics to analyze 56 strains of C. albicans, Kamiyama and colleagues reported that only the sucrose-negative isolates (C. stellatoidea) were separable into a distinct subgroup, whereas the remaining strains formed a tight cluster indicating no evidence for other races. Using fewer characteristics to analyze C. albicans populations, results from cluster analyses have suggested that distinct races may be present. However, both in our earlier study and the study by Odds et al., the methods used to generate data have been found to be fairly unreliable when repeated. In the latter study, the analysis of only nine characteristics may have created a highly artificial structure containing multiple races of C. albicans within the yeast population. Any such groupings of organisms that are based on a limited number of characteristics, even if they have a genetic basis such as restriction fragment length polymorphisms of DNA, should be treated with suspicion until they have been confirmed by more genetic characteristics.

In our study, many more types of each Candida species would be desirable to validate the scheme derived from MDS. The discovery of unusual variants can alter the scheme substantially, and it is desirable to have at least five distinct types for each variable being used in the analysis. The simple application of cluster analysis to the raw data, which was used by us previously, may artificially accentuate any differences found. In this regard, Candida types that were separable as a distinct group by cluster analysis, for example G, 047 and 069 (figure 4A), were no longer clearly separated in the MDS analysis (figure 4B).

The difficulty of finding large numbers of different types may limit the usefulness of MDS in the analyses of variation within unusual species, but MDS provides a further, valuable method for analyzing large populations of yeasts containing one or several species and for determining the affinities of species described by classical taxonomic criteria.

References


