Classification of Mycobacteria by HPLC and Pattern Recognition

Mycobacteria include a number of respiratory and non–respiratory pathogens for humans, such as *M. tuberculosis*, the causative agent of the disease for which it is named. Identification of the responsible bacterium is, therefore, a critical first step in public health regulation or medical treatment.

Traditional methods have relied on classification based on morphology and enzymatic tests, which are subjective and can be time-consuming (typical turnaround from sample collection to results is measured in weeks). Recently, it has been demonstrated that the fatty acids of the cell walls of these bacteria are diagnostic for some species (1). These fatty acids are mainly comprised of high molecular weight (C_{70}–C_{90}) α-branched, β-hydroxy mycolic acids, which, due to their involatility, are not as amenable to gas chromatographic procedures. However, liquid chromatography of these mycolic acids can be utilized (2), and such an approach can provide a more rapid and reproducible method for the identification of target species of Mycobacteria.

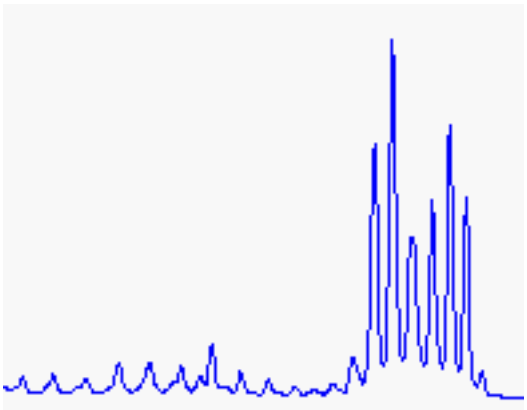

Figure 1. Representative chromatogram showing distribution of mycolic acids.

In this note, chromatographic results from the HPLC method are analyzed with chemometric methods. Principal Components Analysis is shown to aid in visual classification and determination of the presence of outlying or aberrant samples. Predictive models were built for use in rapid, routine classification using the K-Nearest Neighbors approach and SIMCA, the latter based on principal component models of individual classes.

Experimental

All laboratory work was performed at the Centers for Disease Control, in Atlanta, GA, where an extensive project is underway to perfect the classification system. The goal of the CDC work is to optimize the speed of analysis without sacrificing accuracy of classification.

Prior to chromatographic analysis, samples were derivatized to facilitate their detection. Samples were first prepared by basic saponification, then an acidic extraction, using $CHCl_3$, isolated the mycolic acids. Derivatization followed, converting the analytes to their p-bromophenacyl esters.

A Beckman System Gold HPLC instrument was used to analyze the resulting esters, using a 3 µm C_{18}, 4.6 mm x 7 cm cartridge column to separate the analytes. The mobile phase was applied in a gradient, running from 80% CH_3OH in CH_2Cl_2, to 35%, during 10 minutes, with detection at 260 nm. Of the peaks which eluted, 22 were integrated and stored in a results file for later multivariate analysis. Figure 1 is representative of the results of the chromatographic method, showing some of these peaks. The Infometrix *Pirouette* software was used to perform the pattern recognition and classifications.

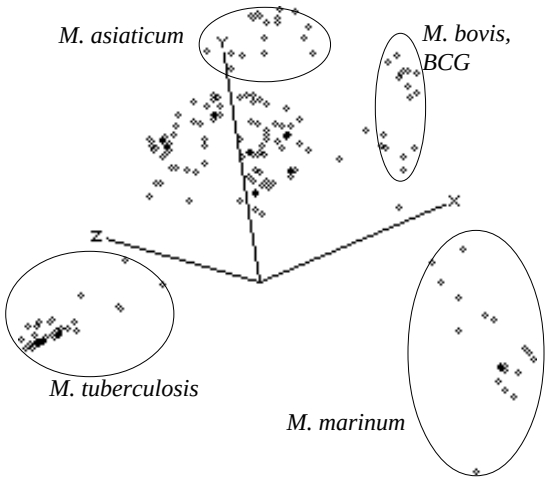

Figure 2. PCA scores of Mycobacteria samples

Results and Discussion

In all, some 50 different Mycobacteria species are under study. The mycolic acids which elute under this chromatographic regime occur throughout the 10 minute gradient window. However, main peak clusters can be observed for the majority of species, with two significant clusters occurring from 5 to 7 minutes and from 7.5 to 9 minutes. Those species which exhibit significant peaks in only the latter cluster are referred to as the Single Cluster species and are the focus of this note.

Two data sets containing only single cluster samples were combined to determine the feasibility of a pattern recognition approach to classification. Represented in these combined data were examples of the eight species listed below, for a total of 188 samples analyzed.

M. asiaticum	*M. bovis, BCG*
M. gastri	*M. gordonae*
M. kansasii	*M. marinum*
M. szulgai	*M. tuberculosis*

Exploratory Data Analysis

Two multivariate methods, Hierarchical Cluster Analysis and Principal Components Analysis, were utilized to provide an overview of the distinguishability of species, based on the chromatographic data. In the clustering method, samples are intercompared based on their multivariate distances in a 22-coordinate space (each coordinate axis corresponds to one of the variables or chromatographic peaks). Following PCA, the sample points in 22-space are transformed to a new data space: the new coordinates, the principal component axes, are determined as those which contain the most variance (information) in the data set. Axes which represent mostly noise can be ignored, resulting in a reduction in the dimensionality of the data.

These transformed data projections, referred to as scores (see Figure 2), showed that two species formed into distinct clusters; another two species could be distinguished clearly. Samples of the remaining four species were somewhat overlapped.

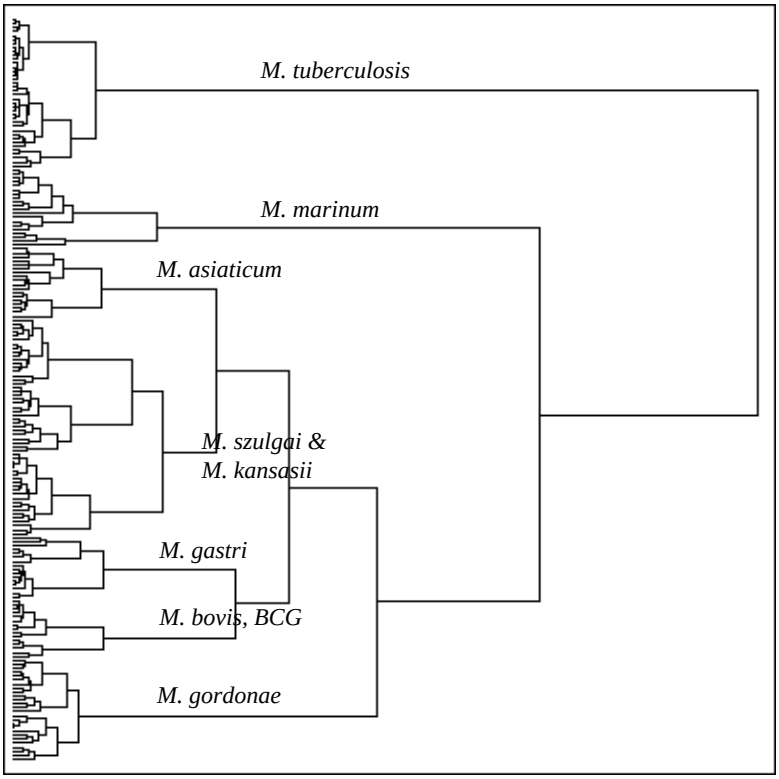

Figure 3. HCA dendrogram for 188 Mycobacteria samples

The results from HCA can best be viewed in a dendrogram, as that in Figure 3. Here, we can see that good separation is possible, and in most subbranches, the species present were homogeneous. In fact, only 2 samples appeared in subbranches to which they did not clearly belong.

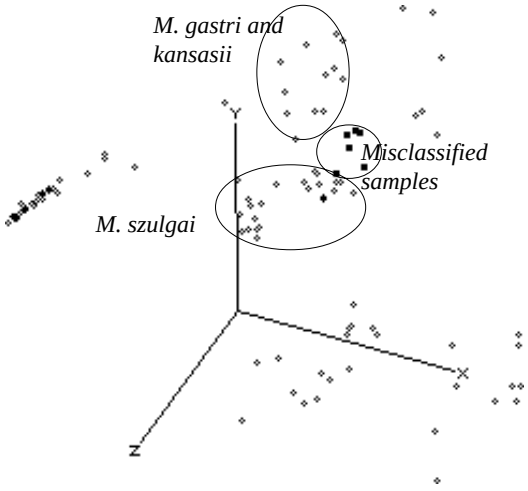

Figure 4. PCA scores plot highlighting six misclassified samples

Classification Analysis

A general approach for pattern recognition is to develop a reliable training set from which predictions of the classes of unknown samples can be accomplished. For this note, an evaluation was first done with the K-Nearest Neighbors method. In this method, the prediction of the class of an unknown or test sample is obtained by observing, in the 22-space defined above, the class of its nearest neighbors in the training set.

The first goal of KNN was to determine if outliers were present in training set. One of the two combined sets was designated as a training set (84 samples), with the other as a test set (104 samples). No significant outliers were found, therefore, KNN was repeated to predict the class of the second set. Because in this set the actual species were already known, it was used to validate the model developed during the training step.

Six samples were incorrectly classified: samples of *M. kansasii* were classified as *M. szulgai*. If many samples of a single species are misclassified, one should not necessarily assume that they are "bad" samples. Such a situation merits closer inspection to understand the cause. In Figure 4, the misclassified samples were highlighted so that they might better stand out in the scores plot. This 3D view of scores was rotated to best view the misclassified samples with respect to the other neighboring samples.

From this view, it is clear why these samples were missed—there is a subset of *M. szulgai* samples that are definitely closer to the misclassified samples than are the remaining *M. kansasii* samples. To probe deeper, another data subset was created which contained only samples of these two overlapped species, and the exploratory algorithms were run once more.

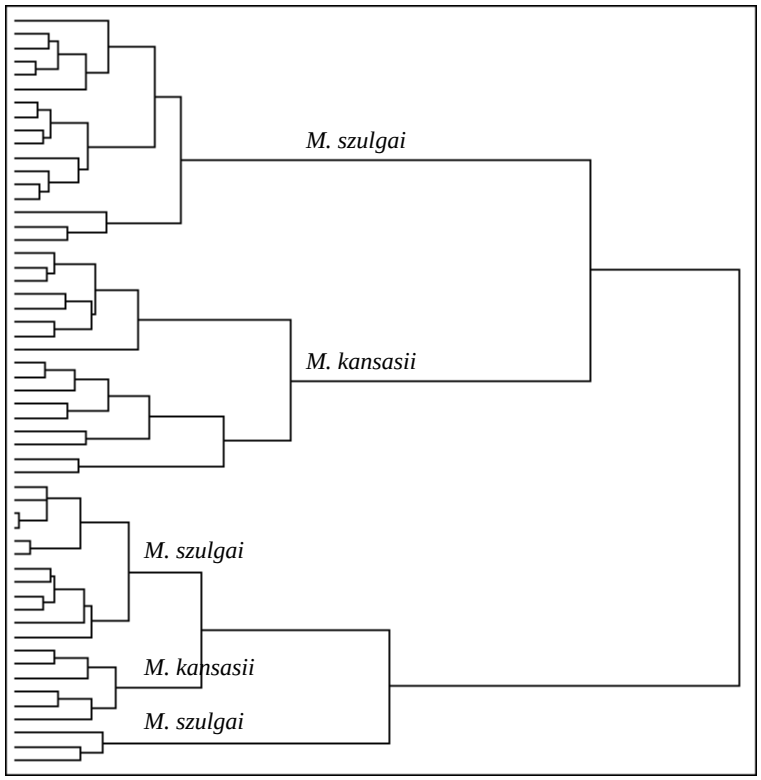

Figure 5. Dendrogram showing subclustering of M. kansasii and M. szulgai samples

In the dendrogram (Figure 5), the data points appear to cluster in 5 groups—two of *M. kansasii* and three of *M. szulgai*. Questions arise whether these two species are truly homogeneous. But other considerations include whether there are simply enough samples to adequately characterize the intra-species variation, especially if the samples are found to be properly identified by independent means.

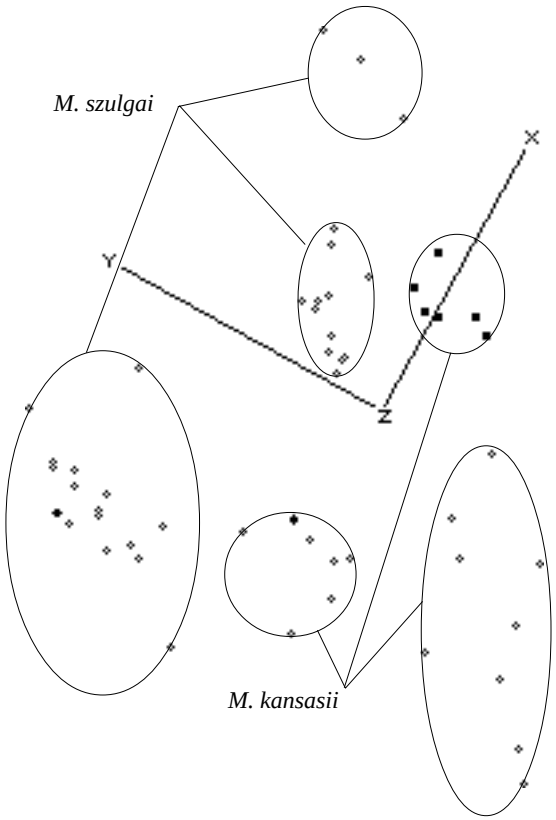

Figure 6. PC scores of M. kansasii and M. szulgai showing subclusters of each species

In the scores plot of these two species (Figure 6), the subclustering is clearly evident. However, by careful manipulation of the 3D view, a perspective can be achieved in which the two species might be separated by a linear discriminant. Such a view is given in Figure 7.

This implies that including specimens from each of the three subclusters in a training set might improve the prediction quality.

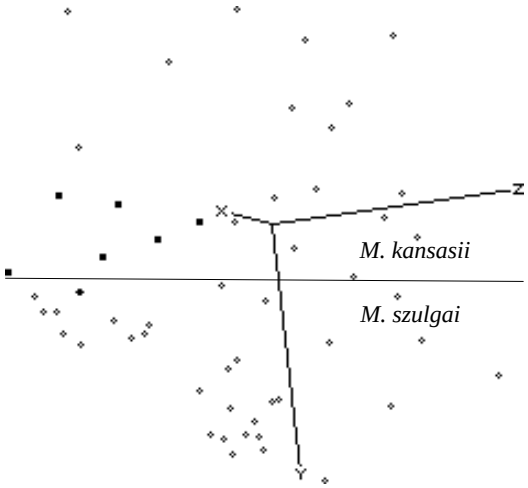

Figure 7. Discriminant view separating M. kansasii and M. szulgai

Thus, a new training set was created from among all of the species, but now including representatives from each of the subsets of *M. kansasii* and *M. szulgai* samples. This new training set of 65 samples was evaluated with KNN by predicting the species of all of the remaining samples. In this instance, 100% of the samples were correctly assessed.

Conclusions

From the results derived with this preliminary study, we can conclude the following:

Pattern recognition can be used on HPLC data to rapidly classify Mycobacteria to species and perhaps to strain within a species.

The algorithmic classification process can be automated by implementation into an instrumental method, removing operator dependence and therefore the subjectivity.

Information can be derived which is not obvious from the traditional approaches or even from a casual examination of the chromatographic profiles (as in the distinction of subclusters within M. kansasii and M. szulgai).

The M. kansasii and M. szulgai mixup also points out the importance of having a representative set of samples in the training set prior to predicting unknown species.

References

1. Minnikin, D.E., S. M. Minnikin, J.H. Parlett, M. Goodfellow, and M. Magnusson. *Arch. Microbiol.* (1984) *139*:225-231.
2. Butler, W. Ray and J. O. Kilburn. *J. Clin. Microbiol.* (1988) *26*: 50-53.

Infometrix, Inc., 10634 E. Riverside Dr., Bothell, WA Phone: (425) 402-1450, Fax: (425) 402-1040, email: info@infometrix.com