GENETIC ALGORITHMS FOR FEATURE SELECTION AND CLASSIFICATION OF COMPLEX CHROMATOGRAPHIC AND SPECTROSCOPIC DATA

By

NIKHIL SURESH MIRJANKAR

Bachelor of Technology
Institute of Chemical Technology
Mumbai, India
2004

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
DOCTOR OF PHILOSOPHY
December, 2012

GENETIC ALGORITHMS FOR FEATURE

SELECTION AND CLASSIFICATION OF COMPLEX

CHROMATOGRAPHIC AND SPECTROSCOPIC DATA

Dissertation Approved:

Dr. Barry K. Lavine

Dissertation Adviser

Dr. Nicholas F. Materer

Dr. Ziad El Rassi

Dr. Richard A. Bunce

Dr. A. K. Kalkan

Name: NIKHIL SURESH MIRJANKAR

Date of Degree: DECEMBER, 2012

Title of Study: GENETIC ALGORITHMS FOR FEATURE SELECTION AND CLASSIFICATION OF COMPLEX CHROMATOGRAPHIC AND SPECTROSCOPIC DATA

Major Field: CHEMISTRY

Abstract: A basic methodology for analyzing large multivariate chemical data sets based on feature selection is proposed. Each chromatogram or spectrum is represented as a point in a high dimensional measurement space. A genetic algorithm for feature selection and classification is applied to the data to identify features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. A good principal component plot can only be generated using features whose variance or information is primarily about differences between classes in the data. Hence, feature subsets that maximize the ratio of between-class to within-class variance are selected by the pattern recognition genetic algorithm. Furthermore, the structure of the data set can be explored, for example, new classes can be discovered by simply tuning various parameters of the fitness function of the pattern recognition genetic algorithm. The proposed method has been validated on a wide range of data.

A two-step procedure for pattern recognition analysis of spectral data has been developed. First, wavelets are used to denoise and deconvolute spectral bands by decomposing each spectrum into wavelet coefficients, which represent the samples constituent frequencies. Second, the pattern recognition genetic algorithm is used to identify wavelet coefficients characteristic of the class. In several studies involving spectral library searching, this method was employed. In one study, a search pre-filter to detect the presence of carboxylic acids from vapor phase infrared spectra which has previously eluted prominent researchers has been successfully formulated and validated. In another study, this same approach has been used to develop a pattern recognition assisted infrared library searching technique to determine the model, manufacturer, and year of the vehicle from which a clear coat paint smear originated. The pattern recognition genetic algorithm has also been used to develop a potential method to identify molds in indoor environments using volatile organic compounds. A distinct profile indicative of microbial volatile organic compounds was developed from air sampling data that could be readily differentiated from the blank for both high mold count and moderate mold count exposure samples. The utility of the pattern recognition genetic algorithm for discovery of biomarker candidates from genomic and proteomic data sets has also been shown.

TABLE OF CONTENTS

Chapter Page

LIST OF TABLES

Table Page

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

LIST OF ABBREVIATIONS

ADAPT	Advanced Data Analysis and Pattern Recognition Toolkit
ASCII	American Standard Code for Information Interchange
BE	Barrett's esophagus
BL	Burkitt's lymphoma
BPNN	Back propagation neural network
cDNA	Complementary Deoxyribonucleic acid
CVA	Canonical variate analysis
CVC	Cluster validity coefficient
DG-18	Dichloran glycerol-18
DMS	Differential mobility spectrometry
DNA	Deoxyribonucleic acid
EAC	Esophageal adenocarcinoma
EPA	Environmental Protection Agency
EWS	Ewing family of tumors
FCV	Fuzzy c-varieties
FTIR	Fourier transform infrared
GA	Genetic algorithm
GBSS	Granule bound starch synthase
GC	Gas chromatogram
GUI	Graphical user interface
HGD	High-grade dysplasia
ID	Identity

IMS	Ion mobility spectrometry
IR	Infrared
K-NN	K-nearest neighbors
LDA	Linear discriminant analysis
MALDI	Matrix-assisted laser desorption ionization
MEA	Malt extract agar
MS	Mass spectrometry
MVOC	Microbial volatile organic compound
NB	Neuroblastoma
NC	Normal controls
NIR	Near-infrared
PC	Principal component
PCA	Principal component analysis
PDQ	Paint Data Query
PLS	Partial least squares
QDA	Quadratic discriminant analysis
RCMP	Royal Canadian Mounted Police
RDA	Regularized discriminant analysis
RIP	Reactant ion peak
RMS	Rhabdomyosarcoma
SEC	Standard error of calibration
SIM	Selective ion monitoring
SIMCA	Soft independent modeling of class analogy
SPME	Solid phase microextraction
SRBCT	Small round blue cell tumors
SVD	Singular value decomposition
TOF	Time of flight

VOC	Volatile organic compound

CHAPTER I

INTRODUCTION

Profiling of complex samples using high performance chromatographic and spectroscopic methods is an active area of research with a large and growing literature [1-1 to 1-10]. The object of profile analysis is to correlate characteristic fingerprint patterns in a chromatogram or spectrum with the properties of a sample or in biomedical studies with the presence or absence of disease in a patient or animal from which the sample was taken. Fingerprint experiments of this type often yield profiles containing hundreds of components. Objective analysis of the profiles depends upon the use of multivariate statistical methods. However, there has been little research directed towards the development of methods to analyze data generated in such experiments.

Pattern recognition methods are well suited for analyzing chromatographic and spectroscopic data because of the characteristics of the procedures. Methods are available that assume no mathematical model but rather seek relationships that provide definitions of similarity between groups of data. Pattern recognition methods are also able to deal with high dimensional data where more than three measurements are used to describe each sample. Finally, techniques are available for selecting important features

from a large set of measurements. Thus, studies can be performed on systems where the exact relationships are not fully understood.

Problems arise when applying pattern recognition methods to large data sets. First, classification success rates often vary with the pattern recognition method employed. Second, low classification success rates for the prediction set are obtained despite a linearly separable training set. Automation of these techniques can be difficult.

The basic premise underlying the research described in this thesis is that all pattern recognition methods work well when the classification problem is simple. By identifying the appropriate features, a "hard" problem can be transformed into a "simple" one. For pattern recognition analysis, the goal is feature selection, in order to increase the signal to noise ratio of the data by discarding measurements or features that are not characteristic of the profile of each class in the data set. To ensure identification of all relevant features, it is best that a multivariate approach to feature selection be employed. This approach should take into account the existence of redundancies in the data.

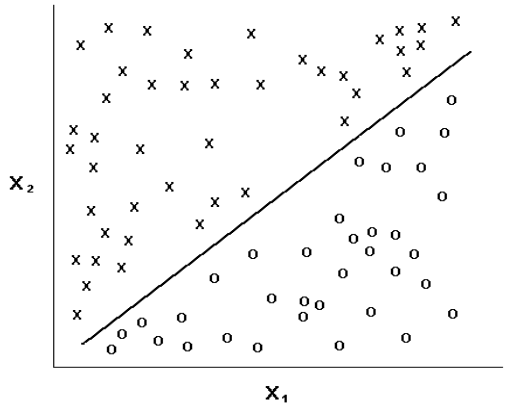

Figure 1.1. Data set consisting of two classes: circles = acceptable, and x = unacceptable. Each sample is characterized by two measurements: x_1 and x_2. Univariate criteria would rank x_1 and x_2 as uninformative variables.

The experience of the Lavine group in pattern recognition has shown that irrelevant features can introduce so much noise that a good classification of the data cannot be obtained. When these irrelevant features are removed, a clear and well-separated class structure can often be found. The deletion of irrelevant variables is therefore an important goal of feature selection. For averaging techniques such as partial least squares and discriminant analysis, feature selection is important since signal is averaged with noise over a large number of variables with a loss of discernible signal amplitude when noisy features are not removed from the data. With neural networks, the presence of irrelevant measurement variables may cause the network to focus its attention on the idiosyncrasies of individual samples due to the net's ability to approximate a variety of complex functions in higher dimensional space, thereby causing it to lose sight of the broader picture, which is essential for generalizing any relationship beyond the training set.

Feature selection is also necessary because of the sheer enormity of many classification problems. For example, consider DNA array data, which consists of thousands of descriptors per observation but only 50 or 100 observations distributed equally between two classes. Feature selection improves the reliability of a classifier because noisy variables increase the chances of false classification and decrease classification success-rates on new data. It is important to identify and delete features from the data set that contain information about experimental artifacts or other systematic variations in the data not related to legitimate chemical differences between the classes represented in the study. Feature selection can also lead to an understanding of the essential features that play an important role in governing the behavior of the process

under investigation. It can identify those measurements that are informative and those measurements that are uninformative. For all of these reasons, feature selection should be the principal focus of research on new methodology in data mining.

Clearly, one major goal of feature selection is the removal of noisy or irrelevant features from the data. Another is dimensionality reduction. In many pattern recognition problems, it is necessary to use artificial neural networks (i.e., nonlinear discriminants) to perform a classification. When artificial neural networks are used to classify data, the pattern recognition problem is often divided into two parts. The first part involves simplifying the problem through feature selection, which then facilitates implementation of the pattern classifier in the second part. The fewer features used, the easier is the classification task and the more reliable are the results. To ensure that random or chance separation is not a problem, the number of observations in the training set should be greater than the number of features. A ratio of 3 or greater for the object to descriptor ratio (n/d) in the training set is considered sufficient to minimize chance classification [1-11 to 1-13] but a larger object to descriptor ratio, for example, $n/d > 10$ is preferred [1-14].

No general theory has been developed to tackle the problem of feature selection because the results obtained must be evaluated in the classification step, which requires the user to consider both the transduction and preprocessing steps used, which in turn introduces additional complexity. The only approach that is guaranteed to identify the optimum set of features in a data set is for the analyst to examine all possible combinations of features. An exhaustive search of this type is only feasible with small data sets.

To understand the importance of feature selection, one must realize that data analysis problems encountered by chemists are often poorly defined from a mathematical standpoint. This occurs because of the way the data is taken. Consider a chromatogram or absorbance spectrum. Each sensor channel (i.e., a peak in the chromatogram or the absorbance at a particular wavelength) is often related to the next channel, and the net result is that the desired information is obscured by redundancies in the data. Furthermore, the data sets are usually underdetermined, i.e., there are more features than observations. Therefore, the major problem in chemistry and the physical sciences, which is to isolate the information of interest from the large amount of redundant or irrelevant data, is a different problem from the one encountered by statisticians who typically work with a small number of variables with information rich data sets generated in well designed experiments.

The procedure used to isolate the desired information from the large amount of redundant or irrelevant data is known as feature selection. Feature selection can be performed using statistical parameters (e.g., means and variances) computed directly from the data for each class to identify the most important features. Alternatively, feature selection can be performed by selecting those descriptors that an analyst believes are the most important based on his or her understanding of the problem. A feature selection step based on knowledge of the problem is often more effective than one based on heuristic methods. Feature selection can also be performed from the results of a classification using the classifier to identify the most informative features. Clustering methods (e.g., hierarchical and K-means clustering) can be used to replace a group of similar variables with a single variable that most closely corresponds to the centroid of

the cluster. Methods most frequently used to perform feature selection include filters to rank the variables or wrapper and embedded methods to assess subsets of features as to their usefulness for a given predictor.

Filter methods select variables by ranking them according to their individual predictive power. Examples of filter methods include the Fisher ratio and the variance weights [1-15] which measure the dichotomization power of each variable for a particular classification problem by computing the between class and within class variance of each feature. Alternatively, the value of the variable itself can be used as a discriminant by setting a threshold on the value with the predictive power of the variable measured in terms of error rate [1-16]. Information theory has also used to rank variables by developing empirical estimates of the information shared between each variable and the predictor. One criticism of filter methods is that it leads to the selection of redundant features even though better class separation may be obtained due to noise reduction (via the Central Limit Theorem) by averaging independent and identically distributed features [1-17]. Another criticism of filters is that variables are scored independent of each other and cannot determine the combination of variables that would give the best prediction (see Figure 1.1).

Wrappers and embedded methods utilize the classifier to score subsets of variables according to their classification power. Wrappers use the performance of the classifier to assess feature subsets of the data. However, a drawback of most wrapper methods is that an exhaustive search of the data cannot be performed on large data sets as the search quickly becomes computationally intractable. A wide range of search strategies have been employed [1-18] including forward selection, backward elimination,

branch-and-bound, and simulated annealing. All of these strategies are greedy as decisions to include or exclude variables during the early part of the search cannot be reversed in light of new information developed during the latter part of the search.

Embedded methods circumvent many of the problems associated with wrappers by incorporating variable selection in the development of the classifier, which allows for better use of the available data as it is not necessary to divide the data set into a training set and validation set. Examples of embedded methods include decision trees such as CART [1-19] that have routines for feature selection, and finite difference calculations [1-20], quadratic approximations of cost functions [1-21], and sensitivity of the objective function calculation [1-22] which have been used to predict the change in the object function of the discriminant and thereby identify variables for addition or removal from classifier such as linear support vector machines [1-23]. The drawback of embedded methods (as well as wrappers) is the dependence on the classification method used to identify informative features in the data. As the goal of feature selection in a pattern recognition problem is the identification of a variable subset that maximizes the ratio of between class variance to within class variance, it is evident that both wrappers and embedded methods do not directly utilize this figure of merit for variable selection.

The development and application of a genetic algorithm (GA) for pattern recognition analysis of chemical data is the subject of this thesis. The pattern recognition GA selects features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. A good principal component plot can only be generated using features whose variance or information is primarily about differences between the classes in the data. Thus, a feature subset is selected that maximizes the ratio

of between groups to within groups' variance. This criterion dramatically reduces the size of the search space since it limits the search to these types of feature subsets. In addition, the GA focuses on those classes and/or samples that are difficult to classify as it trains by boosting the relative importance of those classes and samples that consistently score poorly. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one -pass procedure for feature selection and pattern classification (using principal component analysis). The efficacy and efficiency of the pattern recognition GA is demonstrated using problems from spectral pattern recognition, e.g., infrared and differential ion mobility library searching, genotyping of wheat using near infrared spectroscopy, and biomarker candidate discovery, e.g., detection of molds, analysis of DNA microarrays.

REFERENCES

1-1. J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine, and A.M. Harper, "Applications of pyrolysis/gas chromatography/pattern recognition to the detection of CF heterozygotes," Anal. Chem. 1985, 57, 295-304.

1-2. B. K. Lavine, A. Faruque, P. Kroman, and H. T. Mayfield, "Source Identification of Fuel Spills by Pattern Recognition Analysis of High Speed Gas Chromatograms," Anal Chem., 1995, 67, 3846-3852.

1-3. B. K. Lavine, C. E. Davidson, and D. J. Westover, "Spectral Pattern Recognition Using Self Organizing Maps," J. Chem. Inf. Comp. Science, 2004, 44, 1056-1064.

1-4. B.B. Wall, M.T. Kachman, and S.Y. Gong, "Isoelectric focusing nonporous RP HPLC: A two-dimensional liquid-phase separation method for mapping of cellular proteins with identification using MALDI-TOF mass spectrometry", Anal Chem., 2000, 72, 1099–111.

1-5. M.P. Washburn, D. Wolters, and J.R. Yates, "Large-scale analysis of the yeast proteome by multidimensional protein identification technology", Nature Biotech., 2001, 19, 242–247.

1-6. H.J. Lee, E.Y. Lee, and M.S. Kwon, "Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics", Current Opinion in Chemical Biology, 2006a, 10, 42-49.

1-7. S. Sheng, D. Chen, and J.E. Van Eyk, "Multidimensional liquid chromatography separation of intact proteins by chromatographic focusing and reversed phase of the human serum proteome-Optimization and protein database", Molecular and Cellular Proteomics, 2006, 5, 26–34.

1-8. R. Aebersold, and M. Mann, "Mass spectrometry-based proteomics", Nature, 2003, 422, 198–207.

1-9. P.L. Ferguson, and R.D. Smith, "Proteome analysis by mass spectrometry". Annual Review of Biophysics and Biomolecular Structure, 2003, 32, 399-424.

1-10. I. Feuerstein, M. Rainer, and K. Bernardo, "Derivatized cellulose combined with MALDI-TOF MS: A new tool for serum protein profiling", J. Proteome Res., 2005, 4, 2320–2326.

1-11. B. K. Lavine, D. R. Henry, and P. C. Jurs, "Chance Classifications by Nonparametric Linear Discriminants," J. Chemom. 1988, 2, 1-10.

1-12. B. K. Lavine and D. R. Henry, "Monte Carlo Studies of Nonparametric Linear Discriminant Functions," J. Chemom. 1988, 2, 85-90.

1-13. B. K. Lavine, P. C. Jurs, D. R. Henry, R. K. Vander Meer, J. Pino, and J. McMurry, "Pattern Recognition Studies of Complex Chromatographic Data Sets: Design and Analysis of Pattern Recognition Experiments," Chemolab. 1988, 3, 79-89.

1-14. M. Sjostrom and B. R. Kowalski, "A Comparison of Five Pattern Recognition Methods based on the Classification Results from Six Real Data Bases," Anal. Chim. Acta, 1979, 112, 11-30.

1-15. M.A. Sharaf, D.L. Illman, and B.R. Kowalski, "Chemometrics (Chemical Analysis Series, Vol. 82)", John Wiley & Sons, NY, 1986.

1-16. A.J. Stuper, W.E. Brugger, and P.C. Jurs, "A Computer System for Structure-Activity Studies Using Chemical Structure Information Handling and Pattern Recognition Techniques," in Chemometrics: Theory and Application, B. R. Kowalski (Ed.), ACS Symposium Series 52, Washington DC, 1977.

1-17. I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," J. Machine Learning, 2003, 3, 1157-1182.

1-18. R. Kohavi and G. John, "Wrappers for Feature Selection," Artificial Intelligence, 1997, 97(1-2), 273-324.

1-19. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, Wadsworth and Brooks, 1984.

1-20. B. Schoelkopf and A. Smola, "Learning with Kernels", MIT Press, Cambridge, MA, 2002.

1-21. V. Vapnik, Estimation of Dependencies based on Empirical Data, Springer Series in Statistics, Springer, 1982.

1-22. S. Perkins, K. Lacker, and J. Theiler, "Grafting: Fast Incremental Feature Selection by Gradient Descent in Function Space," J. Machine Learning, 2003, 3, 1333-1356.

1-23. V. Vapnik, Statistical Learning Theory, John Wiley & Sons, NY, 1998.

CHAPTER II

PATTERN RECOGNITION

2.1. INTRODUCTION

Many relationships in chemical data cannot be expressed in quantitative terms. These relationships are better expressed in terms of similarity and dissimilarity among diverse groups of data [2-1]. The task confronting a scientist when investigating these types of relationships is twofold: (1) Can a useful structure based on distinct groups of data be discerned, and (2) Can a sample be classified into one of these groups for the prediction of some property? The first task is addressed using principal component analysis (PCA) [2-2] or cluster analysis [2-3], whereas the second task is addressed using classification methods [2-4]. For the second task, a set of known samples is used to separate information and noise sources, with the information sources combined to develop a discriminant that is used to predict the class membership of samples not part of the original training set. The development of suitable models to isolate groups of data according to their properties is known as classification. The basic premise underlying the use of PCA, cluster analysis or classification methods is that clustering of the data into less similar subgroups is associated with some underlying property of the data.

PCA [2-5] is the most widely used multivariate analysis technique in science and engineering. It is a method for transforming the original measurement variables into a new set of variables called principal components. Each principal component is a linear combination of the original measurement variables. Often, only 2 or 3 principal components are necessary to explain all of the information present in the data. By plotting the data in a coordinate system defined by the 2 or 3 largest principal components, it is possible to identify key relationships in the data, that is, find similarities and differences among samples (e.g., chromatograms or spectra) in a data set.

Cluster analysis [2-6] is the name given to a set of techniques that seek to determine the structural characteristics of a data set by dividing the data into groups, clusters or hierarchies. Samples within the same group are more similar to each other than samples in different groups. Cluster analysis is an exploratory data analysis procedure. Hence, it is usually applied to data sets for which there is no *apriori* knowledge about the class membership of the samples.

Pattern recognition [2-7] is a name given to a set of techniques developed to solve the class membership problem. In a typical pattern recognition study, samples are classified according to a specific property using measurements that are indirectly related to that property. An empirical relationship or classification rule is developed from a set of samples for which the property of interest and the measurements are known. The classification rule is then used to predict this property in samples that are not part of the original training set. The property in question may be the type of fuel responsible for a spill, and the measurements are the areas of selected gas chromatographic peaks from the recovered fuel. Classification is synonymous with pattern recognition, and scientists have

turned to it and principal component analysis and cluster analysis to analyze the large data sets typically generated in monitoring studies that employ computerized instrumentation.

In this chapter, pattern recognition methods are discussed. A summary of the techniques used in the studies described in this thesis are included in the following sections. Special emphasis is placed on the application of these techniques to problems in spectral and chromatographic pattern recognition.

2.2. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is probably the oldest and best known of the techniques used in multivariate analysis. The overall goal of PCA is to reduce the dimensionality of a data set, while simultaneously retaining the information present in the data. Dimensionality reduction or data compression is possible with PCA because chemical data sets are often redundant. That is, chemical data sets are not information rich. Consider a gas chromatogram of a JP-4 fuel (see Figure 2.1), which is a mixture of alkanes, alkenes, and aromatics. The gas chromatogram of a JP-4 fuel is characterized by a large number of early eluting peaks, which are large in size. There are a few late eluting peaks, but their size is small. Clearly, there is a strong negative correlation between the early and late eluting peaks of the JP-4 fuel. Furthermore, many of the alkane and alkene peaks are correlated, which should not come as a surprise as alkenes are not constituents of crude oil but instead are formed from alkanes during the refining process. In addition, the property of a fuel most likely to be reflected in a high resolution gas chromatogram is its distillation curve, which does not require all 85 peaks for characterization.

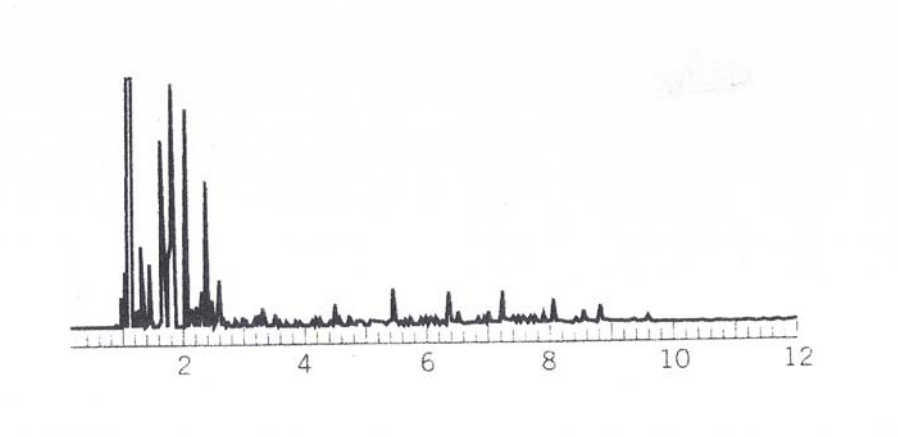

Figure 2.1. Gas chromatogram of JP-4 fuel

Redundancy in data is due to collinearity (i.e., correlations) among the measurement variables. Collinearity diminishes the information content of the data [2-8]. Consider a set of samples characterized by two measurements, X_1 and X_2. Figure 2.2 shows a plot of these data in a 2-dimensional measurement space, where the coordinate axes (or basis vectors) of this measurement space are the variables X_1 and X_2. There appears to be a relationship between these two measurement variables, which suggests that X_1 and X_2 are correlated, since fixing the value of X_1 limits the range of values possible for X_2. If the two measurement variables were uncorrelated, the enclosed rectangle in Figure 2.2 would be fully populated by the data points. Because information is defined as the scatter of points in a measurement space, it is evident that correlations between the measurement variables decrease the information content of this space. The data points, which are restricted to a small region of the measurement space due to correlations among the variables, could even reside in a subspace if the measurement variables are highly correlated. This is shown in Figure 2.3. X_3 is perfectly correlated

with X_1 and X_2 because X_1 plus X_2 equals X_3. Hence, these six sample points lie in a plane even though each data point has three measurements associated with it.

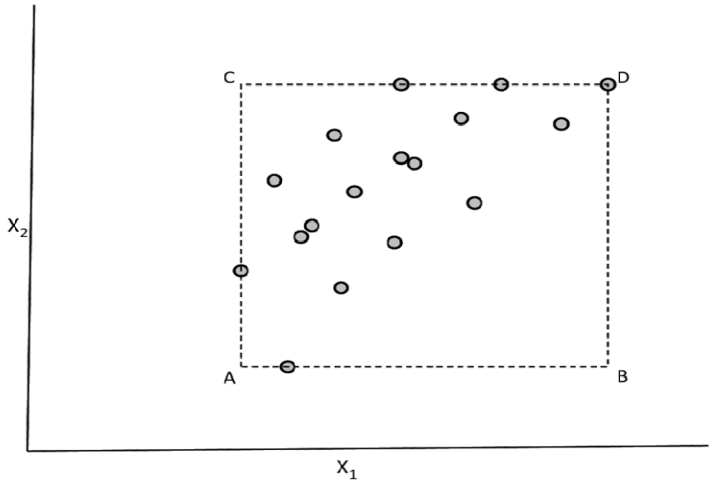

Figure 2.2. Seventeen hypothetical samples projected onto a 2-dimensional measurement space defined by the measurement variables X_1 and X_2. The vertices, A, B, C, and D, of the rectangle represent the smallest and largest values of X_1 and X_2. (Adapted from *NBS J. Res*., 1985, 190(6), 465-476)

$$X = \begin{pmatrix} 0.5 & 0.5 & 1.0 \\ 1.9 & 0.7 & 2.6 \\ 2.0 & 1.0 & 4.0 \\ 0.3 & 1.8 & 2.1 \\ 1.9 & 1.7 & 3.6 \\ 1.2 & 0.2 & 1.4 \end{pmatrix}$$

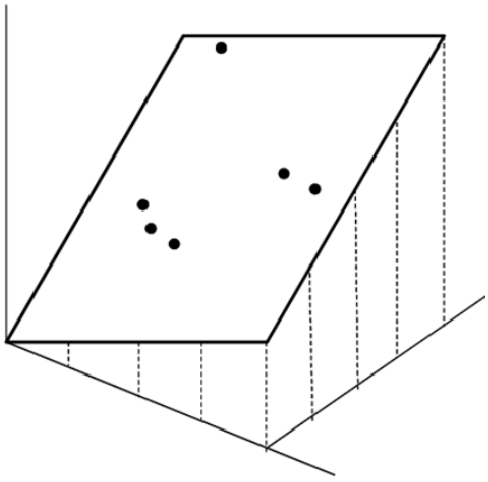

Figure 2.3. Six hypothetical samples projected onto a 3-dimensional measurement space. Because of strong correlations among the 3 measurement variables, the data points reside in a 2-dimensional subspace of the original measurement space. (Adapted from *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science Publishers, Amsterdam, 1992)

2.2.1. Variance Based Coordinate System. Variables that have a great deal of redundancy or are highly correlated are said to be collinear. High collinearity between variables is a strong indication that a new set of basis vectors can be found that will be better at conveying the information content present in data than axes defined by the original measurement variables. The new basis set which is linked to variation in the data can be used to develop a new coordinate system for displaying the data. The principal components of the data define the variance-based axes of this new coordinate system.

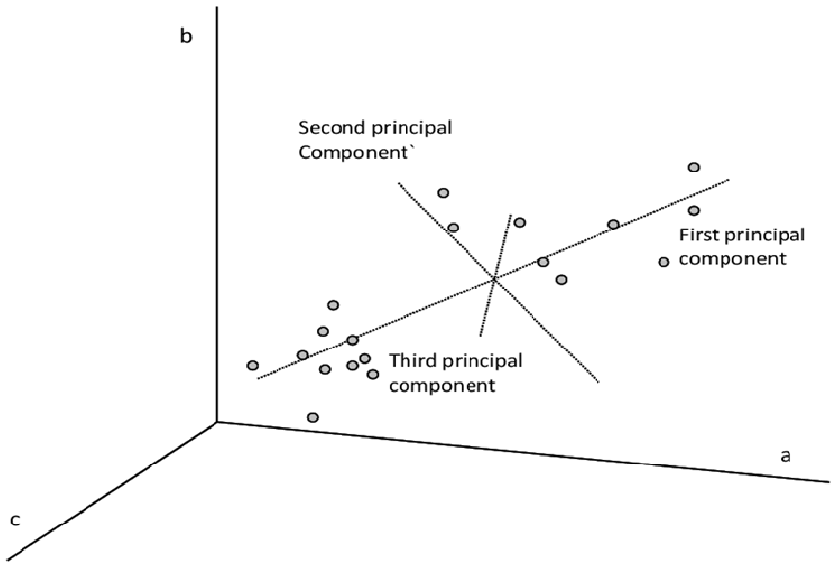

Figure 2.4. Principal component axes developed from the measurement variables a, b, and c. (Courtesy of Applied Spectroscopy, 1995, 49(12), 14A-30A)

The largest or first principal component is formed by determining the direction of largest variation in the original measurement space and modeling it with a line fitted by linear least squares (see Figure 2.4), which passes through the center of the data. The second largest principal component lies in the direction of next largest variation: It passes through the center of the data and is orthogonal to the first principal component. The third largest principal component lies in the direction of next largest variation: It also

passes through the center of the data; it is orthogonal to the first and second principal components, and so forth. Each principal component describes a different source of information because each defines a different direction of scatter or variance in the data. (The scatter of the data points in the measurement space is a direct measure of the data's variance.) Hence, the orthogonality constraint imposed by the mathematics of PCA ensures that each variance-based axis will be independent.

2.2.2. Information Content of Principal Components. One measure of the amount of information conveyed by each principal component is the variance of the data explained by the principal component. The variance explained by each principal component is expressed in terms of its eigenvalue. For this reason, principal components are usually arranged in order of decreasing eigenvalues or waning information content. The most informative principal component is the first and the least informative is the last. The maximum number of principal components that can be extracted from the data is the smaller of either the number of samples or number of measurements in the data set, as this number defines the largest number of independent variables in the data.

If the data are collected with due care, one would expect that only the first few principal components would convey information about the signal, since most of the information in the data should be about the effect or property of interest being studied. However, the situation is not always this straightforward. Each principal component describes some amount of signal and some amount of noise in the data because of accidental correlation between signal and noise. The larger principal components primarily describe signal variation, whereas the smaller principal components essentially describe noise. When smaller principal components are deleted, noise is being discarded

from the data, but so is a small amount of signal. However, the reduction in noise more than compensates for the biased representation of the signal that result from discarding principal components that contain a small amount of signal but a large amount of noise. Often plotting the data in a coordinate system defined by the two or three largest principal components provides more than enough information about the overall structure of the data. This approach to describing a data set in terms of important and unimportant variation is known as soft modeling in latent variables.

PCA takes advantage of the fact that a large amount of data is usually generated in monitoring studies when sophisticated chemical instrumentation, which is commonly under computer control, is used. The data have a great deal of redundancy and therefore a great deal of collinearity. Because the measurement variables are correlated, 85 peak gas chromatograms do not necessarily require 85 independent axes to define the position of the sample points. Utilizing PCA, the original measurement variables that constitute a correlated axis system can be converted to a system that removes correlation by forcing the new axes to be independent and orthogonal. This requirement greatly simplifies the data because the correlations present in the data often allow us to use fewer axes to describe the sample points. Hence, the gas chromatograms of a set of JP-4 and Jet-A fuel samples may reside in a subspace of the 85-dimensional measurement space. A plot of the two or three largest principal components of the data can help us to visualize the relative position of the Jet-A and JP-4 fuel samples in this subspace.

2.2.3. Soft Modeling in Latent Variables. The approach of describing multivariate data in terms of important and unimportant variation is known as soft modeling in latent variables. This approach is possible because instrumental data often

contains a large number of interrelated measurement variables. All of the variation in the data can be explained by a small number of surrogate variables, which are often principal components, because of the redundancies in the data. By examining these principal components, it is possible to identify important relationships in the data, that is, find similarities and differences among the samples in a data set, since each principal component captures a different source of information, i.e., variation in the data. The principal components that describe important variation in the data (i.e., signal) can be identified and regressed against the desired property variable via linear least squares to develop a soft calibration model. Surrogate variables which describe the property of interest are called latent variables.

With PCA, multivariate data can be plotted in a new coordinate system based on variance. The origin of the new coordinate system is the center of the data, and the axes of the new coordinate system are the principal components of the data that primarily contain signal. With this new coordinate system, relationships among the samples can be uncovered in the data. PCA is actually using the data to suggest the model which is a new coordinate system for the data. The model is local since the model center and the principal components will be different for each data set. The focus of PCA and all soft modeling methods is signal, not noise.

Plotting the scores corresponds to plotting the samples in the principal component space. The corresponding plots are called score plots. They are the "work horse" of PCA and other soft modeling techniques since score plots can be used to identify outliers, groups and trends in data and explore similarities among samples. The loadings, which describe the relationship between each principal component and the original

measurement variables from which the principal components are computed, can provide a clue as to the chemical information explained by each principal component. The fraction of the data that is not explained by the set of principal components judged to contain the signal is contained in a residual matrix which is principally noise.

PCA and soft modeling methods are generally not used to analyze data sets that contain a large number of samples and a few measurement variables as each variable probably describes a different source of information. Data sets with a large number of variables and relatively few samples usually contain redundant variables. For this type of data, PCA and soft modeling methods can isolate the various sources of information in the data.

2.2.4. Implementation of PCA. The procedure used to for implement PCA is as follows. First, the samples or data vectors are arranged in the form of a table, which is known as a data matrix. Each row of the matrix corresponds to a sample, and each column of the matrix is a measurement variable. It is crucial for each variable to encode the same information. If the fifth column of the data matrix is the area of a gas chromatographic (GC) peak for benzene in sample 1, then it must also be the GC peak area of benzene in sample 2, sample 3, … sample *N*. The data matrix is usually centered about the mean. This is accomplished by subtracting the mean of the variable from each entry in the corresponding column. Mean centering of a data matrix adjusts the means of each column in the matrix to zero. To ensure that each variable has equal weight in the analysis, the data are usually auto-scaled. In other words, after mean centering, each entry of a column is divided by the standard deviation of the column. Autoscaling adjusts the value of the measurements such that each variable has a mean of zero and a standard

deviation of one. Autoscaling removes inadvertent weighting of the variables that otherwise would occur due to differences in magnitude among the measurement variables.

Principal components are directly computed from the mean centered or auto-scaled data matrix using the singular value decomposition (SVD) algorithm [2-9]. SVD generates the loading matrix, the eigenvalues of each principal component, and a third matrix from which the sample scores, which define the coordinates of the data points in the principal component space, are obtained. The loading matrix defines the relationship between the original measurement variables and the new basis vectors describing variation. Principal components of the data can be reconstructed from the original measurement variables using information contained in the loading matrix. The eigenvalues define the amount of variation in the data contained in each principal component. The principal component representing the direction of largest variance in the data has the largest eigenvalue; the principal component representing the direction of next largest variance in the data has the second-largest eigenvalue, and so forth. The amount of information contained in a principal component relative to the original measurement space, i.e. the fraction of the total cumulative variance explained by the principal component, is equal to the eigenvalue of the principal component divided by the sum of all eigenvalues.

2. 3. CLUSTER ANALYSIS

Cluster analysis is a popular technique whose basic objective is to discover class structure within data. The technique is encountered in many fields, e.g., biology, geology, and geochemistry, under such names as unsupervised pattern recognition and

numerical taxonomy. Clustering methods are divided into three broad categories: hierarchical, object-functional, and graph theoretical. We will concentrate on hierarchical and object functional methods, as they are the most popular.

For cluster analysis, each sample is treated as a point in an n-dimensional measurement space. The coordinate axes of this space are defined by the measurements used to characterize the samples. Cluster analysis assesses the similarity between samples by measuring the distances between the points in the measurement space. Samples that are similar will lie close to one another, whereas dissimilar samples are distant from each other. The choice of the distance metric to express similarity between samples in a data set depends on the type of measurements used.

Typically, three types of measurement variables – categorical, ordinal, and continuous - are used to characterize chemical samples. Categorical variables denote the assignment of a sample to a specific category. Each category is represented by a number, e.g., 1, 2, 3, etc. Ordinal variables are categorical variables, in which the categories follow a logical progression or order, e.g., 1, 2, and 3 denoting low, middle, and high. Continuous variables, on the other hand, are quantitative. The difference between two values for a continuous variable has a precise meaning. If a continuous variable assumes the values 1, 2, and 3, the difference between the values 3 and 2 will have the same meaning as the difference between the values 2 and 1, since they are equal.

Usually, the measurement variables are continuous. For continuous variables, the Euclidean distance is the best choice for the distance metric, because inter-point distances between the samples can be computed directly. However, there is a problem with using the Euclidean distance, which is the so-called scaling effect. It arises from inadvertent

weighing of the variables in the analysis that can occur due to differences in magnitude among the measurement variables. For example, consider a data set where each sample is described by two variables: the concentration of Na and the concentration of K as measured by an ion selective electrode. The concentration of Na varies from 50 to 500ppm, whereas the concentration of K in the same samples varies from 5 to 50ppm. A 10% change in the Na concentration will have a greater effect on Euclidean distance than a 10% change in K concentration. The influence of variable scaling on the Euclidean distance can be eliminated by auto-scaling the data, which involves standardizing the measurement variables (see Equation 2.1), so each variable has a mean of zero and a standard deviation of 1, that is

$$x_{i,standardized} = \frac{x_{i,original} - m_{i,original}}{s_{i,original}} \tag{2.1}$$

where $x_{i,original}$ is the original measurement variable i, $m_{i,original}$ is the mean of the original measurement variable i, and $s_{i,original}$ is the standard deviation of the original measurement variable i. Thus, a 10% change in K concentration has the same effect on the Euclidean distance as a 10% change in Na concentration when the data is auto-scaled. Clearly, auto-scaling ensures that each measurement variable has an equal weight in the analysis. For cluster analysis, it is best to auto-scale the data, since similarity is directly determined by a majority vote of the measurement variables.

Defining a cluster is a problem in cluster analysis (see Figure 2.5) as there is no figure of merit that can serve as a reliable measure of cluster validity for a proposed partitioning of the data. For this reason, clusters are not defined mathematically but rather intuitively

depending on the nature of the problem investigated, the goals of the study, the number of clusters sought in the data, and previous experience. When using these methods, prior knowledge of the problem is often essential.

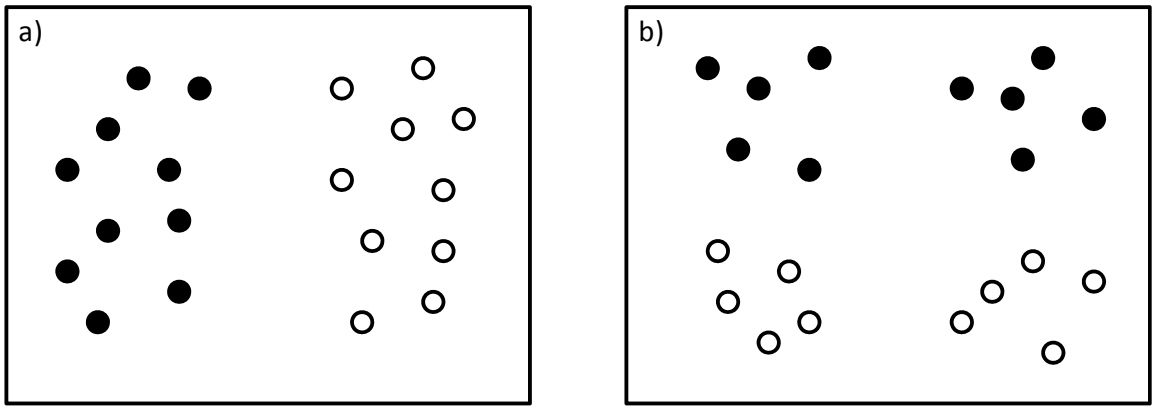

Figure 2.5. Defining a cluster can be a problem. Are there two or four clusters in the data? (Adapted from *Multivariate Pattern Recognition in Chemometrics*, Elsevier Science Publishers, Amsterdam, 1992)

2.3.1. Hierarchical Clustering. Hierarchical cluster analysis [2-10] is based on the principle that distances between pairs of points (i.e., samples) in the measurement space are inversely related to their degree of similarity. The starting point for a hierarchical clustering experiment is the similarity matrix. This matrix is formed by first computing the distances between all pairs of points in the data set. Each distance is converted into a similarity value using Equation 2.2

$$s_{ik} = 1 - \frac{d_{ik}}{d_{max}} \tag{2.2}$$

where s_{ik} is the similarity between samples i and k which varies from 0 to 1, d_{ik} is the Euclidean distance between samples i and k, and d_{max} is the largest distance in the data set which corresponds to the two most dissimilar samples. The similarity values are

organized in the form of a table or matrix which is then scanned to identify the most similar point pair (i.e., largest value). The two samples that comprise the point pair are combined to form a new point located midway between the two original points. Both the rows and columns corresponding to the old data points are removed from the matrix. The similarity matrix is then recomputed for the data set. In other words, the matrix is updated to include information about the similarity between the new point and every other point in the data set. The new nearest point pair is identified, and combined to form a single point. This process is repeated until all points have been linked.

There are a variety of ways to compute the distances between data points and clusters in hierarchical clustering (see Figure 2.6). The nearest linkage method assesses similarity between a point and a cluster of points by measuring the distance to the closest point in the cluster. The farthest linkage method assesses similarity by measuring the distance to the point furthest away in the cluster. Mean linkage assesses the similarity by computing the distances between all point pairs where a member of each pair belongs to the cluster. The mean of these distances is used to compute the similarity between the data point and the cluster.

The results of a hierarchical clustering study are usually displayed as a dendogram, which is a tree shaped map of the inter-sample distances in the data set. The dendogram shows the merging of samples into clusters at various stages of the analysis and the similarities at which the clusters merge, with the clustering displayed hierarchically. Interpretation of the results is intuitive, which is the major reason for the popularity of these methods.

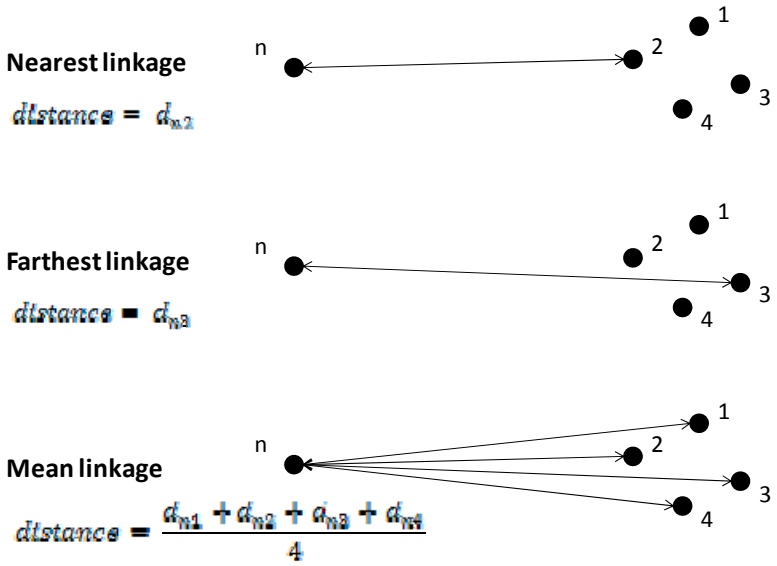

Figure 2.6. The distance between a data cluster and a point using (a) nearest linkage, (b) farthest linkage, and (c) mean linkage.

2.3.2. FCV Clustering

The FCV clustering algorithm [2-9 to 2-16] attempts to fit each of the c clusters in the data set to a principal component model of the form given by Equation 2.3

$$x = v + \sum_{j=1}^{PC} t_j \, d_j \qquad (2.3)$$

where x denotes a sample representative of the cluster, ν is the center of the cluster in the p-dimensional space, R_p, the vectors d_j are an orthonormal set (i.e., loadings) spanning a subspace of R_p, and t_j (scores) are the coordinates of the sample vector in the subspace. An interesting feature of the FCV clustering algorithm is that each data vector in the data set is assumed to contribute to the modeling of each of the clusters within the data. The

actual algorithm consists of solving four equations simultaneously (see Equations 2.4 to 2.7) using a Picard iteration [2-17].

$$v_i = \frac{\sum_{k=1}^{n} (\mu_{ik})^m \, x_k}{\sum_{k=1}^{n} (\mu_{ik})^m} \tag{2.4}$$

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{D_{ik}}{D_{jk}}\right)^{\frac{1}{(m-1)}}} \tag{2.5}$$

$$S_i = \sum_{k=1}^{n} (\mu_{ik})^m \, (x_k - v_i) \, (x_k - v_i)^T \tag{2.6}$$

$$D_{ik} = \left(| \, x_k - v_i \, |^2 \; - \sum_{j=1}^{r} \langle \, x_k - v_i, d_{ij} \, \rangle^2 \right)^{\frac{1}{2}} \tag{2.7}$$

μ_{ik} is the membership value of sample k with respect to cluster i where i = 1, 2, 3….c and is subject to the boundary conditions $0 < \mu_{ik} < 1$ and $\sum \mu_{ik} = 1$. D_{ik} is the distance between sample k and the center of cluster i, D_{jk} is the distance between sample k and the center of cluster center j, v_i is the center of cluster i, d_{ij} is a unit eigenvector (principal component) corresponding to the jth largest eigenvalue of the fuzzy covariance matrix S_i for cluster i, m is a fixed weighting exponent which is usually assigned a value of 2, and r defines the shape of the cluster (r = 0 for spherical clusters, r = l for linear varieties, and so forth).

To obtain a solution to this set of four equations, the user must provide the starting centers. Class membership values, the within cluster scatter matrix of each cluster, and the distance between each sample and each cluster are then computed in succession. New cluster centers are calculated in the last step of the first iteration, and the algorithm uses these new cluster centers as a starting point for a second iteration through the same set of equations. This process continues until convergence is achieved. The number of iterations required to achieve convergence is determined by the minimum prespecified change criterion used for the class membership values which is set by the user.

Usually, one chooses m to be equal to 2 but by increasing m less weight is attached to the importance of samples with small class membership values. As the value of m is increased, the algorithm becomes fuzzier. Data points whose membership values are uniformly low throughout the iterative procedure are less influential in defining a cluster at higher m values. It is this attribute of the FCV clustering algorithm that is appealing when one suspects the data may not exist, as well separated clusters. The ability to tune out noise by adjusting the value of m can be advantageous in obtaining meaningful clustering results. Alternatively, the value of m can be set close to unity and the clustering algorithm then functions as a hard clustering algorithm yielding results similar to those obtained by k-means.

Another advantage of the FCV clustering algorithm is the possibility of using the fuzziness of a given cluster configuration as an indicator of its quality. This is achieved by computing the cluster validity coefficient (CVC), which is a measure of the separation between clusters. The CVC is determined by computing the ratio of the distance between

cluster centers to the weighted scatter of the clusters [2-18]. The larger the value of this coefficient, the better the separation between the two clusters. By successively increasing the value of m, the effect of samples with poor class membership values can be filtered out. An indication of cluster quality can be obtained by comparing CVC values where m is increased stepwise. If the change in the CVC value is small, the conclusion is that distinct sample clusters exist within the data. A marked increase in the value of the CVC as m is increased could be interpreted as significant overlap between data clusters.

When using the FCV clustering algorithm to investigate data, one can search for round clusters in the data by specifying r = 0, find the best fit of the data to linear chain like clusters by specifying r = 1, or fit the data to other cluster shapes by setting $r \geq 2$. This feature allows the data to be fitted to a variety of distinct cluster shapes. Hierarchical clustering discussed in the previous section possesses a similar attribute as nearest linkage favors the formation of large linear clusters and complete linkage favors the formation of small spherical clusters.

2.3.3. Practical Considerations. All clustering procedures yield the same results for data sets with well-separated clusters. However, the results will differ when the clusters overlap. For this reason, it is a good idea to use at least two different clustering algorithms, e.g., single and complete linkage, when studying a data set. If the dendograms are in agreement, then a strong case can be made for partitioning the data into distinct groups as suggested by the dendograms. If the cluster memberships differ, the data should be further investigated using PCA or the FCV clustering algorithm. The results from FCV and PCA can be used to gauge whether nearest linkage or complete linkage is the better solution.

All hierarchical clustering techniques suffer from so-called space distorting effects. Nearest linkage, for example, favors the formation of large linear clusters instead of the usual elliptical or spherical clusters. As a result, poorly separated clusters are often chained together. Because of these space-distorting effects, all clustering methods should be used in tandem with PCA to detect clusters in multivariate data sets.

All clustering methods will always partition data, even randomly generated data, into distinct groups or clusters. Hence, it is important to ascertain the significance level of the similarity value selected by the user. For this task, the author proposes a simple three-step procedure. First, a random data set is generated with the same correlation structure, the same number of samples, and the same number of measurements as the real data set that is currently being investigated. Second, the same clustering technique(s) is applied to the random data. Third, the similarity or class membership value for the same number of clusters identified in the real data set is determined from the FCV results or the dendogram of the random data. If the similarity or the class membership value is substantially larger for the real data set, the likelihood of having inadvertently exploited random variation in the data to achieve clustering is probably insignificant.

2.3.4. Conclusion. Cluster analysis should be used in conjunction with mapping and display techniques such as PCA to identify similar samples, detect outliers, and summarize overall trends in a multivariate data set. Although clustering methods can identify distinct sample subgroups in a data set, they are not sufficient for developing a classification rule that can accurately predict the class membership of an unknown sample. For this reason, classification methods have been developed and are discussed in the following sections of this chapter.

2. 4. CLASSIFICATION METHODS

So far, only exploratory data analysis techniques, i.e., cluster analysis and PCA, have been discussed. These techniques attempt to analyze data without directly using information about the class assignment of the samples. In this section, pattern recognition techniques will be discussed. These techniques which were originally developed to solve the class membership problem, categorize a sample on the basis of regularities in observed data. The first applications of pattern recognition to chemistry were studies involving low-resolution mass spectrometry [2-19]. Since then, pattern recognition techniques have been applied to a wide variety of chemical problems, e.g., chromatographic fingerprinting [2-20 to 2-22], spectroscopic imaging [2-23 to 2-25], and data interpretation [2-26 to 2-28].

Pattern recognition techniques fall into one of two categories: (1) minimum distance classifiers and (2) non-parametric discriminants. Minimum distance classifiers, e.g., k-nearest neighbor [2-29] and SIMCA [2-30 to 2-33], treat each chromatogram or spectrum as a data vector $x = (x_1, x_2, x_3 \ldots x_j \ldots x_p)$ where component x_j is the area of the j^{th} peak or the absorbance of the j^{th} wavelength. Such a vector can also be viewed as a point in a high-dimensional measurement space. A basic assumption is that distances between points in the measurement space will be inversely related to their degree of similarity. Using a minimum distance classifier we can determine the class membership of a sample by examining the class label of the data point closest to it or by determining the class centroid that lies closest to the sample in the measurement space. In analytical chemistry, minimum distance classification rules are developed using either K-nearest neighbor or statistical discriminant analysis.

Non-parametric discriminants [2-34 to 2-36], e.g., neural networks and support vector machines, attempt to divide a data space into different regions. In the simplest case, that of a binary classifier, the data space is divided into two regions. Samples that share a common property (e.g., fuel type) will be found on one side of the decision surface, while those samples comprising the other category will be found on the other side. Non-parametric discriminants have provided insight into relationships contained within sets of chemical measurements. However, random or chance classification [2-39] can be a serious problem for data sets that are not sample-rich.

2.4.1 K-Nearest Neighbor (K-NN). For its simplicity, K-NN is a powerful classification technique. A sample is classified according to the majority vote of its k-nearest neighbors, where k is an odd integer, e.g., one, three, or five. For a given sample, Euclidean distances are first computed from the sample to every other point in the data set. These distances arranged from smallest to the largest are used to define the sample's k-nearest neighbors. A poll is then taken by examining the class identities among the point's k-nearest neighbors. Based on the class identity of the majority of its k-nearest neighbors, the sample is assigned to a class in the data set. If the assigned class and the actual class of the sample match, the test is considered a success. The overall classification success rate, calculated over the entire set of points, is a measure of the degree of clustering in the set of data. Clearly, a majority vote of the k-nearest neighbors can only occur if the majority of the measurement variables concur since the data is usually autoscaled.

K-NN cannot furnish a statement about the reliability of a classification. For training sets that have a large number of samples in each class, the 1-nearest neighbor

classification rule will have an error rate that is twice as large as the Bayes classifier [2-37] which is the optimum classifier for any set of data. To implement the Bayes classifier, one must have knowledge about all the statistics of the data including the underlying probability distribution function of each class. Usually, this knowledge is not available. Any other classification method, no matter how sophisticated, can at best only improve on the performance of K-NN by a factor of two. For this reason, K-NN is often used as a benchmark against which to measure other classification methods.

2.4.2. Linear and Quadratic Discriminant Analysis. Linear and quadratic discriminant analyses [2-38 to 2-40] develop classification rules based upon the statistical properties of the data. Classifiers are developed from prior knowledge of class membership, from *apriori* assumptions about the statistical distribution of the data, and from the mean vectors and covariance matrices of the classes. For linear and quadratic discriminant analysis, each class is assumed to possess a multivariate normal distribution. This is a reasonable assumption as most of the distribution functions encountered in chromatographic and spectroscopic data sets possess elliptical probability contours due to correlations among the measurement variables. They only differ in the rate at which the probability decreases away from the mean. In linear discriminant analysis (LDA), a sample is assigned to the class with the lowest discriminant score, (see Equation 2.8) where $d_k(x)$ is the discriminant score for class k, x is the data vector of the sample, and C_k is the pooled or average class covariance matrix of the data. Because LDA presumes that all class covariance matrices are equal, the classification rule is a Mahalanobis distance, which is similar to a Euclidean distance with a correction to taking into account

correlations among the measurement variables using the inverse of the covariance matrix (C_k^{-1}) when computing the distance between a sample and the centroid of a class.

$$min[\, d_k(x) \;=\; (x - m_k)^T \, C_k^{-1} \, (x - m_k) \,] \tag{2.8}$$

In quadratic discriminant analysis (QDA), class covariance matrices are not assumed to be equal. The classification rule is given by Equation 2.9 where ln $|C_k|$ is the determinant of the class covariance matrix which corresponds to the volume of space occupied by the sample points representing the class. This classification rule is not limited to the Mahalanobis distance computed for each class but is specific for the covariance matrix of each class.

$$min[\, d_k(x) \;=\; (x - m_k)^T \, C_k^{-1} \, (x - m_k) + \; ln|C_k| \,] \tag{2.9}$$

LDA and QDA are guaranteed to produce an optimal classification surface. However, LDA and QDA are seldom applied to problems in pattern recognition as there are usually too few samples in most data sets to reliably estimate C_k^{-1} [2-41]. The issue of covariance stabilization in discriminant analysis for data sets with a low object to descriptor ratio is discussed in the next section of this chapter. Both SIMCA and regularized discriminant analysis [2-42] have been used successfully to develop classifiers from data where the number of dimensions is large compared to the size of the training set. Both methods provide the benefit of increased stability without being

restrictive as to ignore essential differences in the covariance that may be present in the data.

2.4.3 SIMCA and Regularized Discriminant Analysis. The problem of covariance stabilization in discriminant analysis was first tackled by Wold in 1976 [2-43]. He addressed the problem of covariance stabilization by developing a biased estimate of the covariance matrix using a method called SIMCA (Soft Independent Modeling of Class Analogy). For each class in the data set, the inverse of the covariance matrix is approximated by a principal component representation involving the so-called secondary eigenvectors of the data. The inverse of the class k covariance matrix, C_k^{-1}, is decomposed by a process known as spectral decomposition [2-44], see Equation 2.10, where v_{jk} is the j^{th} principal component of C_k, λ_{jk} is the corresponding eigenvalue and p is the dimensionality of the data. It is the smaller eigenvalues, not the larger ones that are most important when reconstructing C_k^{-1}. However, it is the smaller eigenvalues that are difficult to estimate where the dimensionality of the data is large compared to the size of the training set. By taking the average of these smaller eigenvalues, more reliable estimates of them can be obtained (see Equation 2.11) where A is the number of principal components necessary to describe class k, which is determined directly from the data using a procedure known as cross validation [2-45].

$$C_k^{-1} = \sum_{j=1}^{p} \left[\frac{v_{jk}\, v_{jk}^{T}}{\lambda_{jk}} \right] \qquad (2.10)$$

$$C_k^{-1} = \frac{\sum_{j=A+1}^{p}\left[v_{jk}\ {v_{jk}}^T\right]}{\sum_{j=A+1}^{p}\left[\lambda_{jk}\right]} \tag{2.11}$$

In SIMCA, a principal component analysis is performed on each class in the data set, and a sufficient number of principal components are retained to account for most of the variation within each class. Thus, each class in the data set is represented by a principal component model. The number of principal components retained for each class is usually different. Determining the number of principal components that should be retained for each class is important as retention of too few components can distort the signal or information content contained in the class model, whereas retention of too many principal components will diminish the signal to noise. Using cross validation [2-46] model size can be determined directly from the data. To perform cross validation, segments of the data are omitted during the principal component analysis. Using one, two, three, etc., principal components, omitted data are predicted and compared to the actual values. This procedure is repeated until every data element has been omitted once. The principal component model that yields the minimum prediction error for the omitted data is retained. Using cross validation the number of principal components necessary to describe the signal in the data can be determined while ensuring high signal to noise by not including the so-called secondary or noise laden principal components in the class model.

The variance explained by the class model is called the modeled variance, which describes the signal, where as the noise in the data is described by the residual variance or the variance not accounted for by the model which is explained by the secondary principal components, which have been truncated or omitted from the class model. By

comparing the residual variance of an unknown to the average residual variance of those samples that comprise the class, a direct measure of the similarity of the unknown to the class can be obtained. This comparison also serves as a measure of the goodness of fit of the sample to a particular principal component model. Usually, the F-statistic is used to compare the residual variance of a sample with the mean residual variance of the class [2-47]. Employing the F-statistic, an upper limit for the residual variance can be calculated for those samples belonging to the class. The final result is a set of probabilities of class membership for each sample.

One advantage of SIMCA is that an unknown is only assigned to the class for which it has a high probability. If the residual variance of a sample exceeds the upper limit for every modeled class in the data set, the sample is not assigned to any of the classes because it is either an outlier or comes from a class that is not represented in the data set. Another advantage of SIMCA is that it is sensitive to the quality of the data used to generate the principal component models. As a result, there are diagnostics to assess the quality of the data, e.g., the modeling power [2-48] and the discriminatory power [2-49]. The modeling power describes how well a variable helps the principal components to model variation, and discriminatory power describes how well the variable helps the principal components to classify the samples in the data set. Variables with low modeling power and low discriminatory power are usually deleted from the data because they contribute only noise to the principal component models.

Friedman and Frank [2-50] also addressed the problem of covariance stabilization in discriminant analysis by developing a pooled estimate of the class covariance matrix, which they called regularized discriminant analysis (RDA). The class covariance matrix

is first shrunk towards the pooled covariance matrix used in LDA to estimate the common covariance of the data (see Equation 2.12) where $C_k(\lambda)$ is the shrunken estimate of the covariance matrix of class k, C_k is the covariance matrix of class k estimated directly from the data as in QDA, C is the pooled covariance matrix used in LDA, and λ, which varies from 0 to 1, regulates the variance bias trade-off. Equation 12 is then shrunk towards the identity matrix using the parameter γ which also varies between 0 and 1 (see Equation 2.13) where $C_k(\lambda, \gamma)$ is the biased estimate of the covariance matrix of class k, (γ which varies from 0 to 1 is another user provided shrinkage parameter, trace$[C_k(\lambda)]$ is a diagonal matrix consisting of the eigenvalues for class k, p is the number of measurement variables for each sample, and I is the identity matrix. LDA and QDA are special cases of RDA as λ = 0 and γ = 0 gives rise to QDA and λ = 1 and γ = 0 gives rise to LDA. λ and γ are determined by computing the cross validated error rate of the discriminant using the training set data on a unit square defined by λ and γ. λ and γ are varied by 0.1 on the grid, and a vector of misclassifications as a function of these shrinkage parameters is generated with the values of λ and γ that yield the lowest error rate selected.

$$C_k(\lambda) = (1-\lambda)C_k + \lambda C \tag{2.12}$$

$$C_k(\lambda,\gamma) = (1-\gamma)C_k(\lambda) + \gamma\, trace\left[\frac{C_k(\lambda)}{p}\right] * I \tag{2.13}$$

Both RDA and SIMCA use a bias estimate of the class covariance matrix to deemphasize the effect of the lower eigenvalues. The major difference between these two methods is the approach used to perform the necessary variance-bias trade-off. Although the approach used in RDA to estimate the class covariance matrix has been shown to superior to the approach used in SIMCA, RDA lacks the necessary tools to optimize the classification model, many of which are found in SIMCA e.g., the modeling and discriminatory power of the measurement variables used to develop the principal component models for each class in the data set.

2.4.4 Neural Networks. A large number of papers have been published on the advantages of using feed- forward neural networks to classify chromatographic and spectroscopic data. The flexibility of the distributed model defined by the weights of the network allows both linear and nonlinear classifiers to be defined. The addition of a hidden layer using the appropriate transfer function converts a simple, two-layer (input and output) linear neural network into a three-layer network capable of describing any continuous nonlinear surface defined on a p-dimensional space [2-51]. Although linear decision surfaces for classification can also be developed using linear transfer functions to connect the layers, linear feed forward neural networks have received only scant attention as LDA, QDA, SIMCA, and RDA have been repeatedly shown to be very effective in dealing with linear classification problems.

During training, a feed-forward neural network learns the relationship between independent variables (e.g., absorbance values at specific wavelengths), which serve as inputs to the network, and dependent variables (e.g., sample class membership) which are designated as outputs of the network. Learning occurs when a training set consisting of

spectra or chromatograms and the class labels of the samples are presented to the network, and the network weights are adjusted to minimize differences between the output of the network and the known class membership of the samples. Once the network weights have been adjusted using the training set, the network can be used to predict the class membership of unknown samples from their spectra or chromatograms.

The configuration of a three-layer feed-forward neural network for classification is shown in Figure 2.7. The input layer serves as a buffer to store the values of the input variables. In the hidden layer, neurons are arranged in parallel with each neuron corresponding to a hyperplane (i.e., linear boundary) or decision surface in the measurement space formed from the linear combination of the measurement variables as defined by the weights for each neuron. Using this network configuration, nonlinear decision surfaces can be developed from the combinations of hyperplanes formed in the construction of the class boundaries.

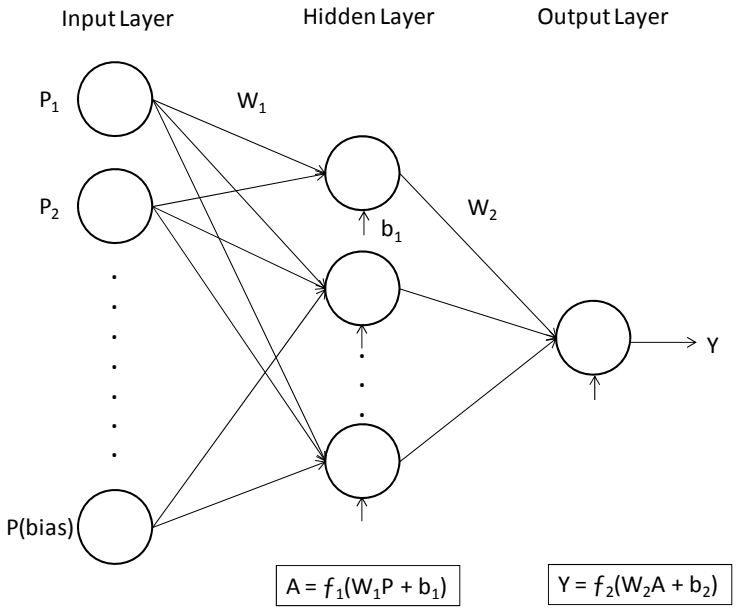

Figure 2.7. Configuration of a three-layer feed-forward neural network

The weights associated with each hidden layer neuron correspond to a weight vector that defines a specific hyperplane in the pattern space. The output layer weights define combinations of the hyperplane tests that can determine the class membership of the samples. Feed-forward neural networks that contain a finite number of neurons in the hidden layer connected to the outer layer by a sigmoid transfer function are capable of modeling any continuous nonlinear decision surface in the pattern space.

The optimization of the weights in a 3-layer feed-forward neural network amounts to an iterative search for an acceptable minimum as the samples comprising the training set are processed individually by the network to encode the relationship between the input and output variables in the network model. Back propagation of error [2-52 and 2-53], a nonlinear optimization method, is the most popular method for training the weights of feed-forward neural networks that contain hidden layers. The training of the network is initiated by initializing the weights at small random values that naturally grow to larger magnitudes through error feedback. As the weights grow in magnitude during training, the network model becomes increasingly nonlinear since larger weights imply greater nonlinear character while small weights result in projections of the network inputs onto linear portions of the sigmoid transfer function. The use of large numbers of measurement variables as inputs increases the nonlinear character of the neural network classifier in the early stages of training as a linear combination of a large number of terms is more likely to yield summations with large magnitudes that project on the nonlinear portions of the sigmoid curve.

The goal of classification using a neural network is to obtain a weight solution that produces a neural network model that fits the training set data and permits good

prediction of the data outside of the training set. The ability to accurately classify both the training and test set reflects good generalization by the network model. Neural networks that do not exhibit good generalization are usually overtrained. Overtraining occurs because the model is fitting noise or there are too many network weights to be estimated from the data. The tendency to fit noise in the data can be limited by controlling the flexibility of the network or by applying noise reduction techniques to the input data before the initiation of training. The number of neurons in the hidden layer, the number of network weights, and the magnitude of the weights should be limited while ensuring the necessary flexibility for modeling nonlinear decision surfaces. It is well known that classification should not be attempted when the degrees of freedom exceed the number of training set samples [2-54 and 2-55]. For this reason, the number of samples and the number of weights in the network should be carefully examined to avoid an over-determined model when analyzing data using multi-layer feed-forward neural network. The effects of noisy input data can be diminished by using a larger training set. However, the use of small, noisy training sets can result in a neural network model that fits the noisy data very well while generalizing poorly on data that has not been used in weight training. From these considerations, the obvious strategy when modeling noisy data using a multilayer feed forward neural network is to keep the number of network weights small and use large, representative data sets. In cases where fitting noise remains a concern, feature selection or wavelets (see Section 2.4.6) can be used to remove noise, thereby reducing the tendency of the network to model the noise.

2.4.5 Support Vector Machines. Support vector machines identify decision surfaces or hyperplanes with the widest margin to separate samples from different classes

in a data set (see Figure 2.8). This is done using only some samples in the data set which are known as support vectors. The input data are mapped from the original measurement space to a higher dimensional space using kernel functions to simplify the classification problem. There are a variety of kernel functions including linear kernels for linear hyperplanes, polynomial, Gaussian, and sigmoidal kernels for nonlinear decision surfaces. Kernel functions simplify the classification problem by allowing us to directly compute the dot product of the weight vector which defines the distance between the hyperplane and each support vector in the original measurement space. For a linearly separable data set, there will be a large number of linear hyperplanes that can be developed to separate samples into their respective classes. The hyperplane best able to generalize the data (i.e., accurately classify both the training and test sets) is the one with the widest margin. For data sets that are not linearly separable, a nonlinear decision surface will be used. The kernel function that yields the best classification for the samples is the one selected for discriminant development. When developing a classification rule for a data set that is not separable, the optimization problem is reformulated to allow for samples but as few as possible to be present in the margin.

Support vector machines have good generalization ability because the optimization problem when using the appropriate kernel functions is less prone to overfitting as there are fewer model parameters to compute from the data. Another advantage of support vector machines is that it is easier to train. The optimization search space of a support vector machine with the appropriate kernel will have a global solution without local minima. For a nonlinear classification problem, the selection of the parameter values for the kernel function becomes crucial as they have an effect on the

shape of the separating hyperplane. Larger values of the kernel parameters for a Gaussian kernel incorporate linearity to the model whereas smaller values tend to diminish the generalization ability by making the kernel more sensitive to noise in the data. It is also important to remember that a properly trained feed-forward multi-layer neural network will perform better than a support vector machine for nonlinear classifications as a neural network with a hidden layer is capable of modeling any continuous nonlinear decision surface in the pattern space.

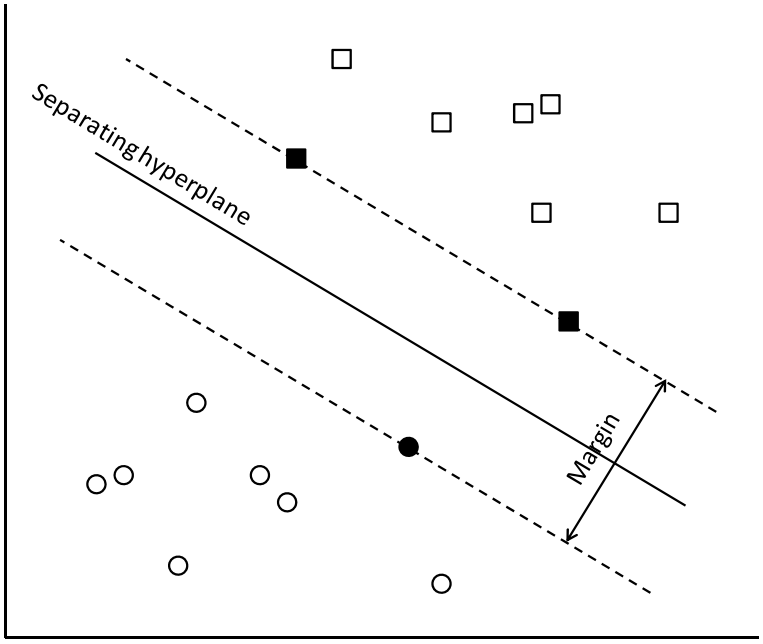

Figure 2.8. Decision surface from a support vector machine for a binary classification problem

2.4.6 Data Preprocessing. Classification of complex samples on the basis of their spectroscopic or chromatographic profiles can be confounded by noise. If the basis of classifications for samples in a data set is other than legitimate differences between the desired groups, unfavorable classification results can be obtained for the prediction set despite a linearly separable training set. The application of noise reduction techniques to

the input data before initiation of training can obviate this problem. In this section, two preprocessing techniques will be discussed for noise removal: feature selection and wavelets.

Feature selection is crucial in many pattern recognition studies as not all of the measurements taken on each sample are meaningful. For data sets containing a large number of measurement variables, irrelevant features can introduce so much noise that a good classification of the data cannot be obtained. A clear and well separated class structure can be uncovered when irrelevant features are removed from the data due to the elimination of noise variance unrelated to class structure. With averaging techniques such as discriminant analysis or PCA, feature selection is crucial since signal is averaged with noise over a large number of variables with a loss of signal amplitude when noisy features are not removed. For neural networks, the presence of irrelevant features can prevent the network from generalizing the data beyond the training set because the network is fitting the noise in the samples, not the signal. Feature selection will improve the reliability of any classifier because noisy variables increase chances of false classification and lower classification success-rates on new data. It is important to identify and delete features from the data set that contain information about experimental artifacts or other systematic variations in the data not related to legitimate chemical differences between the source profiles of the classes represented in the study. Feature selection is also important because of the sheer enormity of many classification problems, e.g., DNA microarray data that often consists of thousands of measurements per observation but only 50 or 100 observations distributed between two classes. For all of

these reasons, feature selection is often the principal focus of many pattern recognition studies.

Wavelets [2-56 and 2-57] can be used to separate noise from signal in multivariate chemical data. Spectra and chromatograms are characterized by peaks and other local features. However, important key features are often hidden or buried in noise. Wavelets can extract this hidden local information by analyzing the original data at different levels of resolution. Using wavelets, the data is transformed into a new set of variables called wavelet coefficients that are better at conveying information than the original measurement variables. Wavelets are scalable mathematical functions of localized waveforms that have the following properties: (1) they oscillate with varying frequency, (2) they decay rapidly, and (3) they have an average value of zero. The template of a typical wavelet basis function, the so-called “mother wavelet,” is shown in Figure 2.9.

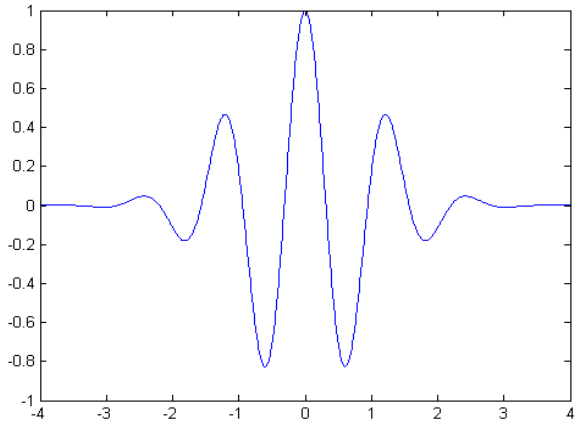

Figure 2.9. Template of a typical Wavelet basis function

The property of scalability of wavelet basis functions is used to extract information from signals. Removing noise and deconvoluting overlapping bands in a

spectrum is more efficient than using the Fourier transform since local information is extracted with wavelets as opposed to only global information which is extracted using the Fourier transform. The Fourier transform is limited to global information since it uses an infinite series of sine and cosine wave functions to describe the signal as opposed to wavelets where each function is defined for a limited region of the chromatogram or spectrum.

Wavelet analysis is implemented by a non-redundant decomposition of the spectrum or chromatogram. The scaled and shifted version of the "mother wavelet" is projected onto the spectrum or chromatogram and a comparison is made between the wavelet and the original data using a set of approximation and difference functions [2-58]. During wavelet analysis, the signal is decomposed into sets of various signal component sets. Each set corresponds to a different time (or wavelength) and frequency scale. Each frequency component of the signal is analyzed separately using a resolution that matches its scale [2-59]. Changing the time scale of the wavelet also allows for a closer examination of local information of the signal at different resolutions of the chromatogram or spectrum. Scaling the wavelet corresponds to either dilating or compressing the wavelet basis function along its time axis by a scaling factor to fit different frequency scales of the signal. The dilated or compressed versions of the wavelet basis function are then shifted on the time axis to be projected on all parts of the signal to generate wavelet coefficients. These wavelet coefficients measure correlation of different sections of the signal with the scaled versions of the wavelet basis function. Higher scales correspond to highly stretched (or dilated) wavelets (see Figure 2.10), which are compared with longer portions of the signal to capture its smoother features.

These long range slowly changing coarse features form the lower frequency components of the signal. Similarly, lower scales correspond to compressed wavelets (see Figure 2.10) that measure short range rapidly changing features, which provide information pertaining to high frequency components of the signal.

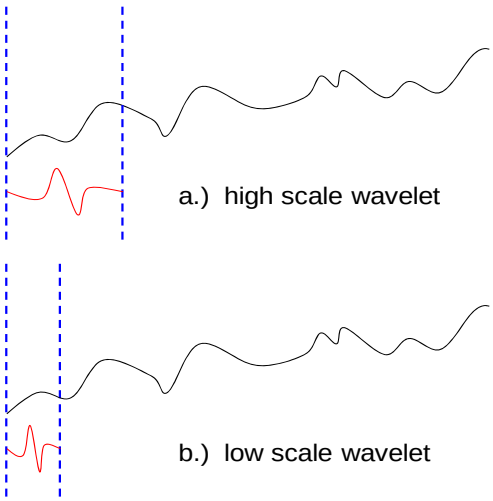

Figure 2.10. A comparison of: a.) high scale wavelet and b.) low scale wavelet for representation of signal

For the wavelet decomposition of data, a complementary pair of high-pass and low-pass wavelet scaling filters [2-59] is used as shown in Figure 2.11. The high-pass filter will allow only the high frequency component of the signal to be measured as a set of wavelet coefficients called “approximation”. The low-pass filter will measure the low frequency coefficient set called “detail”. The detail coefficients usually correspond to the noisy part of the data. This process of decomposition is continued with different scales of the wavelet filter pair in a step-by-step manner to separate the noisy components from the signal. Figure 2.12 shows the first and second levels of wavelet filtering applied to a noisy sine wave.

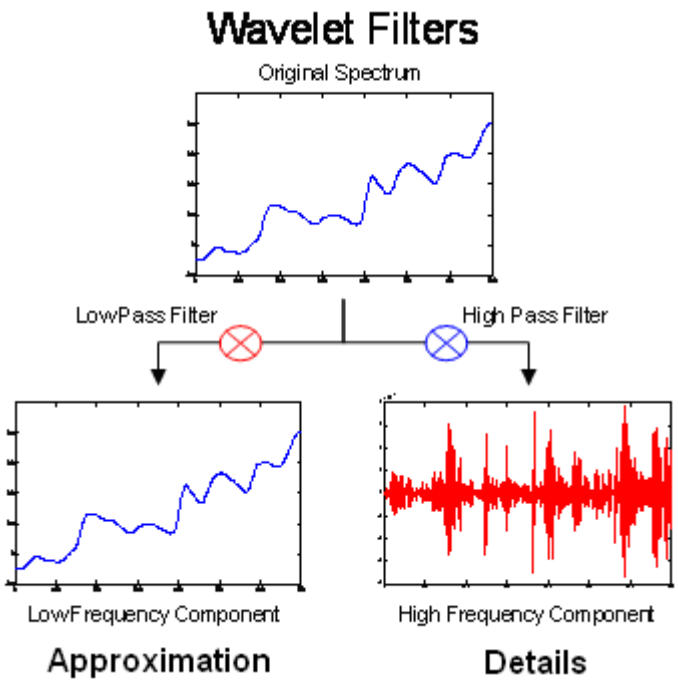

Figure 2.11. Decomposition of the spectrum using wavelet filters.

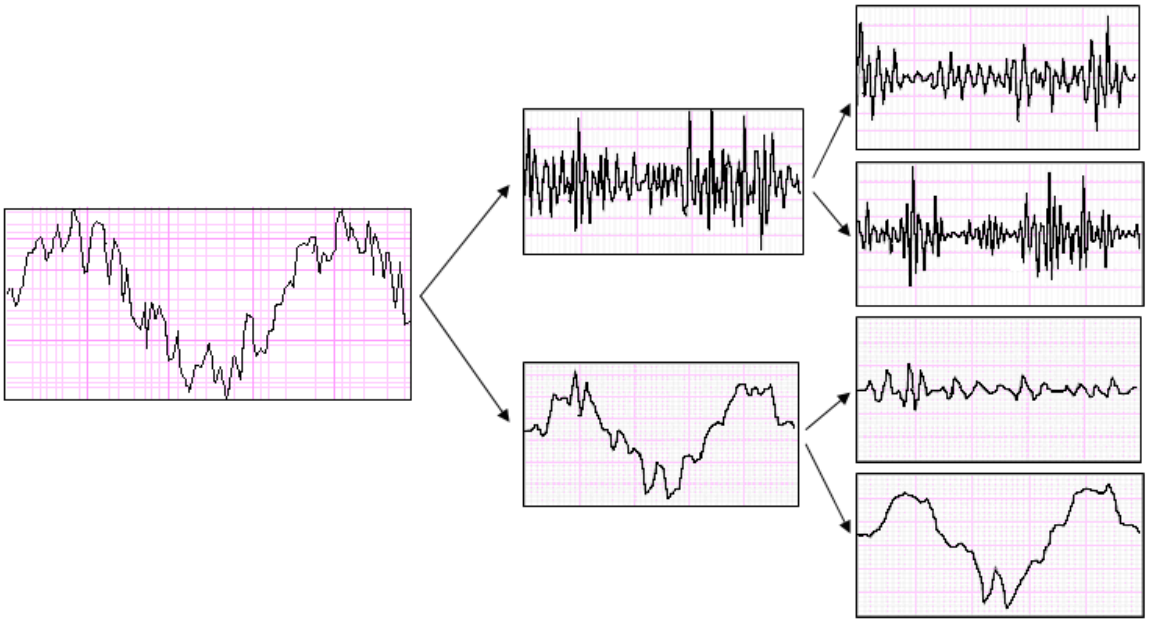

Figure 2.12. Second level decomposition of a noisy sine wave using wavelet filters

There are several ways to apply the wavelet transformation: continuous, discrete, fast, and complex wavelet transform, and the wavelet packet transform. The discrete

wavelet transform is implemented by decomposing the signal and all of its successive approximations until the desired level of signal decomposition is achieved as shown in Figure 2.13. The wavelet packet transform performs a much richer analysis by decomposing both the approximations and details at each level to give what is called a wavelet packet tree (see Figure 2.13). It can be used in difficult cases where the signal is highly convoluted. All the wavelet coefficient sets (approximations and details) generated at each level are organized in a specific order to form a data vector. Every sample in the data set after the wavelet transform is represented by such an array of wavelet coefficients.

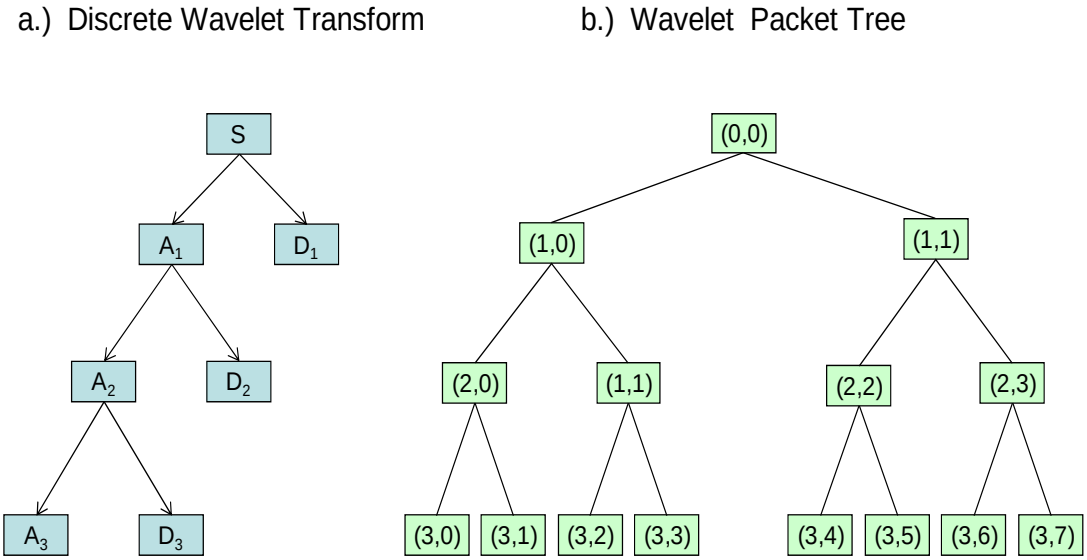

Figure 2.13. Two different types of Wavelet transform are shown: a.) Discrete wavelet transform of original signal S to give approximations A_n and details D_n where n is the decomposition level; b.) Wavelet packet tree where each packet (l,n) is represented by the level of decomposition (l) and its number (n) in that level.

There are many different types of mother wavelets: Daubechies, Symlet, Coiflet, Haar and Biorthogonal. The Haar wavelet is the simplest wavelet. It is one period of a square wave. A major drawback of using the Haar wavelet is that it is not continuous and

therefore not differentiable. Daubechies are compactly supported orthonormal wavelets suitable for discrete wavelet analysis. Symlet are nearly symmetrical wavelets. They are related to the Daubechies family as they share similar properties. Figure 2.14 shows the basic templates of the mother wavelets belonging to these families.

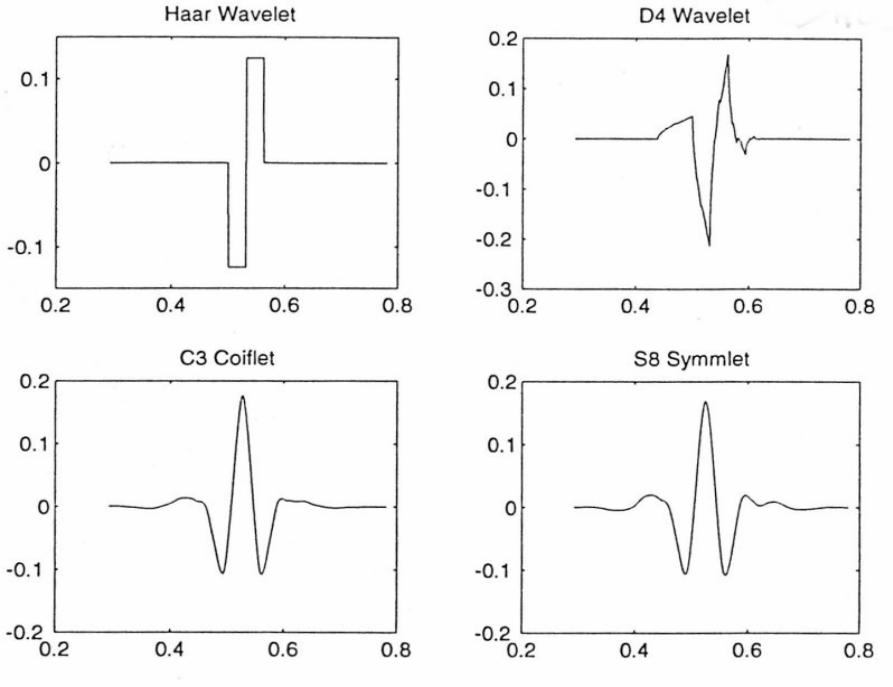

Figure 2.14. Templates of several “mother” wavelets

The selection criterion for a suitable choice of mother wavelet is based on the ability of the wavelet to denoise and deconvolute spectral or chromatographic data. The wavelet chosen should be similar to the type of features and attributes present in the spectra or chromatogram. The Daubechies and Symlet wavelets generally work well with chromatographic and spectral data. For signals with sharp peaks or discontinuities, mother wavelets such as Daubechies 2 to 4 also with sharp and abrupt features should be useful. If the signal comprises smoother or broad peaks than smoother wavelets such as Daubechies 5 through 12 or even larger may be employed depending on the type of

features that are needed to be extracted. We generally use a range of mother wavelets on the data set to get the best result. The implementation of wavelet transform was performed by MATLAB using its wavelet toolbox [2-60].

2.5. ADAPT

The pattern recognition analyses performed in the studied described in this thesis were performed using a tool kit written for MATLAB (MATLAB 7.6.0.324) called ADAPT (Advanced Data Analysis and Pattern Recognition Tool Kit). It was developed with the help of Graphical User Interface Development Environment (GUIDE) of MATLAB. GUIDE provides an easy to use framework for designing GUIs with simplified ways to attach sub-routines and functionalities.

The Main GUI panel can be invoked by entering the command ADAPTv5 and the toolkit can be used in a straightforward and intuitive manner with simple instructions provided. All of the pattern recognition analysis routines described in this thesis with the exception of support vector machines can be accessed from this main panel. The panel consists of a toolbar with menus at the top with the panels remaining area divided into two parts that display information about the training set (top half) and the prediction set (bottom half). The information displayed for both sets consists of the total number of samples in a data set and the misclassified samples listed in each class after training and prediction is performed using a particular pattern recognition method. It also displays the name of the data set uploaded, the number of descriptors, the type of preprocessing performed on the data, and the status of both training and prediction. The tags of individual samples after training and prediction are also displayed along with their actual and predicted class-labels to identify misclassified samples. The toolbar menus consist of

several menus: File, Edit, Analysis, Train, Predict, Calculate Error Rate and Display. Each menu has a sub-menu or it opens up another GUI panel with simple instructions.

The File menu allows for importing of data files (both training and prediction sets), saving classification models, training-prediction results, and uploading previously trained models for use in future studies. The data file that can be uploaded needs to have data in a table or matrix form with a delimited ASCII format. The rows of the data should correspond to sample data vectors and the columns should be the measurement variables (descriptors). The data columns should be preceded by label columns for sample and class id.

The Edit menu is used to remove or retain specific samples, classes, or descriptors and to perform the necessary preprocessing of the data, e.g., normalizing data vectors to unit length or to one. Autoscaling or mean centering is performed by each pattern recognition analysis routine when desired by the user.

The Analysis menu includes various methods for pattern recognition analysis such as PCA and canonical variate analysis (CVA), clustering methods like Hierarchical and FCV clustering, and variance and Fisher weights for feature selection. Each submenu opens a new sub-panel/window for selecting various parameters specific to those methods and for changing display settings. More information is also displayed on these sub-panels. For example, the cumulative variance for as many as 10 principal components is shown in PCA. Also choices are provided for displaying the score and loading plots, selecting the principal components for plotting, selecting sample or class label tags for display in the score plots, and projection of the prediction set samples. The

plots generated have provisions for focusing into a specific region by zooming, panning, and 3D rotation, which ensures a careful analysis of the data.

The Train menu leads to different pattern recognition and classification methods for developing classifiers. The available methods are LDA, QDA, RDA, RDA with self optimized parameters, K-NN, and a Back Propagation Neural Network. Each application opens a self explanatory sub- panel for selecting method parameters. The Predict menu enables the user to apply a stored classification model on the prediction set. Other menu functions are available for calculating the error rates of training (by boot-strapping or cross-validation) and for displaying the training and prediction sets results of individual samples.

2.6 CASE STUDIES

Pattern recognition is about reasoning, using the available information about the problem to uncover information contained within the data. Autoscaling, feature selection, and classification are an integral part of this reasoning process. Each plays a role in uncovering information contained within the data.

Pattern recognition analyses are usually implemented in four distinct steps: (1) scaling, (2) data preprocessing, (3) classification, and (4) mapping and display. However, the actual process is iterative, with the results of a classification or display often determining a further preprocessing step and reanalysis of the data. Although the procedures selected for a given problem are highly dependent upon the nature of the problem, it is still possible to develop a general set of guidelines for applying pattern recognition techniques to real data sets. In the last section of this chapter, a framework is

presented for solving the class membership problem by way of two published studies. In the first study, a set of microbial volatile organic compound (MVOC) profiles were developed with corresponding bioaerosol measurements as input-output pairs for a discriminant to predict the presence or absence of mold contamination in indoor environments [2-61]. In the second study, GC profiles of cuticular hydrocarbon extracts obtained from individual and pooled ant samples were analyzed using pattern recognition techniques [2-62]. Clustering was observed according to the biological variables of social caste and colony of origin. Pattern recognition methods were used to separate three temporal (age dependent) social castes (foragers, reserves, and broods) and to visualize colony cuticular hydrocarbon changes that occurred with time. The dynamic nature of these heritable characters in relationship to nestmate recognition confirmed previous studies [2-63].

2.6.1 Prediction of Mold Contamination from VOCs. The data consisted of 145 gas chromatograms of VOCs collected by solid phase microextraction (SPME) from homes and office buildings in Northern New York. Spore collection to characterize the indoor air quality of the residences and buildings investigated as part of this study was simultaneously performed using Anderson N6 impactors. The volatile organic signatures that molds emit as reflected by the GC profiles were compared to the impactor data collected from each building. By comparing the bioaerosol data to the volatile organic profiles, a discriminant could be trained to classify a residence by potential mold growth based on VOCs.

Bioaerosol data was collected during 33 sampling events from 16 locations in the village of Potsdam, New York from July 2006 until August 2007. Sampling of airborne

fungi onto agar plates required a high-volume vacuum pump and an Anderson N6 impactor, which was used in conjunction with malt extract agar (MEA) and dichloran glycerol 18 (DG18) in Petri dishes to collect viable mold samples. DG18 is a xerophilic agar whereas MEA is a mesophilic agar. By using two agar types, a broader range of fungi was cultured, providing a better representation of the site's fungal ecology.

At each residence or building, 4 or 5 samples of each type of agar were collected with an associated field blank for each sampling event. The samples were cultured for 6 days and counted. All counts were blank corrected, and a positive-hole correction was applied. These biological characterizations were converted into values after taking indoor-outdoor ratios for each sampling event. These values were used to tag each sampling site with a specific class label: low (ratio is less than 1.2), medium (ratio is between 1.2 and 3) and high (ratio is greater than 3) mold contamination. This allowed for a direct comparison between the volatile organic profile characterizations (input) and the relative bioaerosol concentrations in each building (output) using discriminant analysis. The class membership distribution of the sampling events using DG18 and MEA growth media is shown in Table 2-1.

Table 2-1. Class Membership Distribution of Bioaerosol Sampling Data

Mold Count	Number of Samples	
	DG18	MEA
Low	56	56
Medium	34	27
High	55	62

Concurrently, air sampling for VOCs was performed by SPME. An 85μm stableflex carboxen/polydimethylsiloxane fiber equipped with a commercial holder for manual operation was chosen for air sampling to reduce fiber breakage. Other advantages associated with this fiber include its high affinity for low molecular weight volatiles due

and its ability to extract a broad range of analytes. Multiple SPME fibers were used at each location, and the sampling time for each experiment was 2 hours. The SPME fibers were conditioned for 30 to 60 minutes prior to sampling. After sampling, the SPME fibers were injected into a HP5890 Series GC equipped with a HP 5971 mass selective detector using Chemstation software (Agilent). A splitless injection was performed with a purge of 1.5 minutes. The injector temperature of the GC was set at 250 degrees Centigrade, the detector temperature was set at 280 degrees Centigrade, and the column oven temperature was 40 degrees Centigrade for 5 minutes, followed by a 5 degree ramp to 200 degrees Centigrade, which was held for 5.5 minutes. Separation of the VOCs was performed on a VOCOL (Supelco) capillary column (60 meters L x 0.32mm id x 1.8μm film thickness). MS scan range was from 35 to 350amu.

For pattern recognition analysis, each gas chromatogram was represented as a data vector $X = (x_1, x_2, x_3 \ldots x_{29})$ using as descriptors the area of each of the VOCs identified by GC/MS. Peak areas were normalized to the percentage of the total integrated area for each gas chromatogram. Each gas chromatogram contained 29 standardized retention time windows. 43 GC profiles were common to the low mold count class, 14 profiles were common to the medium mold count class, and 42 were common to the high mold count class for both DG18 and MEA (see Table 2.1).

The 145 gas chromatograms were divided into three classes on the basis of impactor data obtained for DG18. Prior to CVA, each class was analyzed for outliers using PCA. Figure 2.15 shows a plot of the two largest principal components of the low mold count gas chromatograms of the VOCs. The first two principal components explain 50% of the total cumulative variance of the data. It is evident from this score plot and the

generalized distance test [2-64] that 4 gas chromatograms are outliers - they are very different from the other gas chromatograms. A visual analysis of these 4 gas chromatograms suggested that peak matching could be the source of the problem as several peaks were poorly resolved and there were variations in retention time from the established time windows for several peaks presumably due to a sloping baseline. Outlier analysis using PCA was also performed for the medium and high mold count gas chromatograms but no discordant observations in this data were detected. A plot of the two largest canonical variates of the GC profiles with the outliers removed is shown in Figure 2.16. There is considerable overlap between the low mold count and medium mold count gas chromatograms and some overlap between the medium and high mold count gas chromatograms, and this does not augur well for the viability of the proposed method.

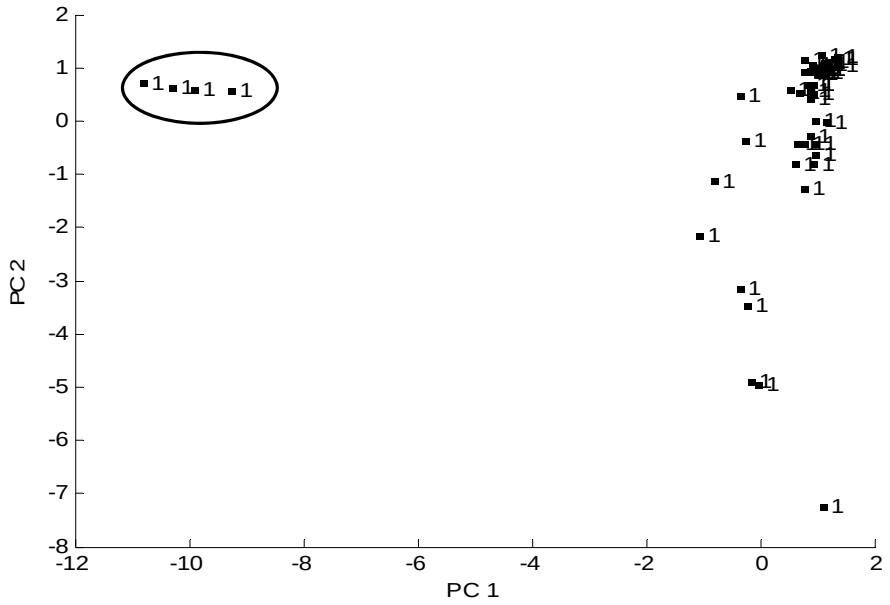

Figure 2.15. PC plot of the two largest principal components of the low mold count gas chromatograms as determined by the DG18 impactor data. 4 chromatograms enclosed by an ellipse are outliers in the PC plot of this data.

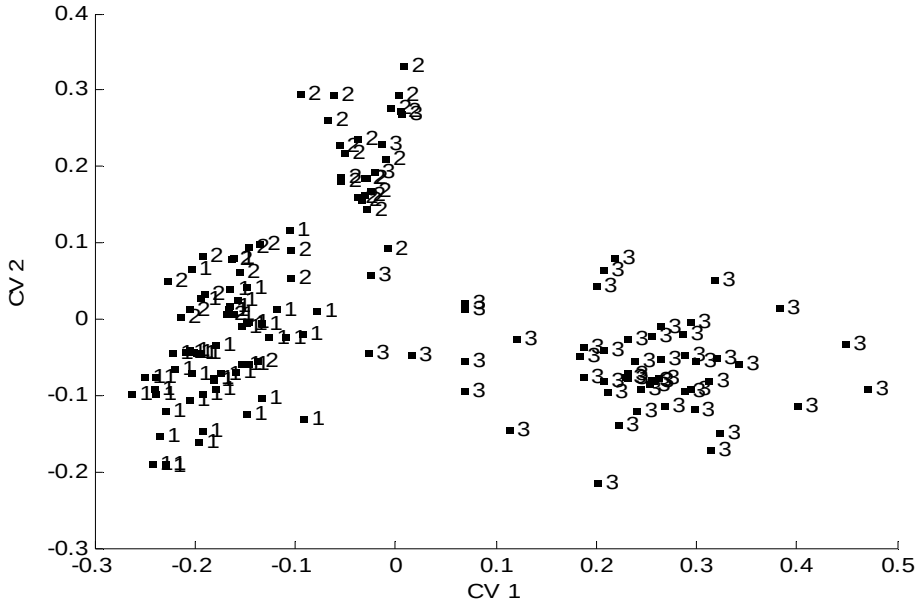

Figure 2.16. A plot of the two largest canonical variates of the GC profiles of the MVOCs with the 4 outliers removed. Each gas chromatogram is represented as a point in the plot. 1 = low mold count, 2 = medium mold count, and 3 = high mold count.

The same study for the 145 gas chromatograms was repeated using the MEA impactor data. Principal component plots of each class revealed the presence of outliers in the data (see Figures 2.17, 2.18, and 2.19). Again, the outliers were attributed to the poor quality of the GC data. A plot of the two largest canonical variates of the GC profiles with the 12 outliers removed (see Figure 2.20) revealed that indoor environments with high mold counts can be readily differentiated from indoor environments with medium and low mold counts. The overlap of VOC profiles from low and medium mold count environments could be due to chemical interferences from cooking and cleaning activities in many of these locations prior to sampling.

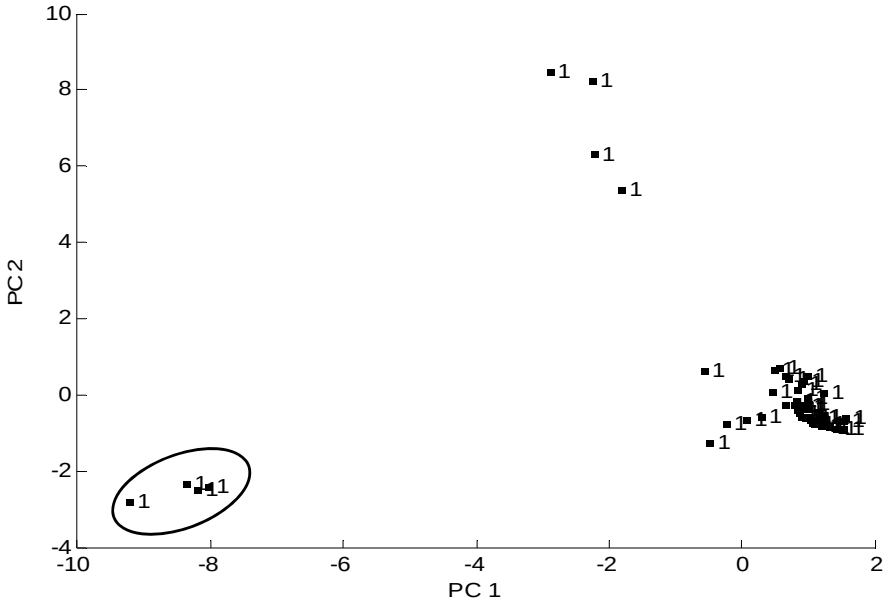

Figure 2.17. PC plot of the two largest principal components of the low mold count gas chromatograms as determined by the MEA impactor data. Each gas chromatogram is represented as a point in the plot. 4 chromatograms enclosed by an ellipse are outliers in the PC plot of this data.

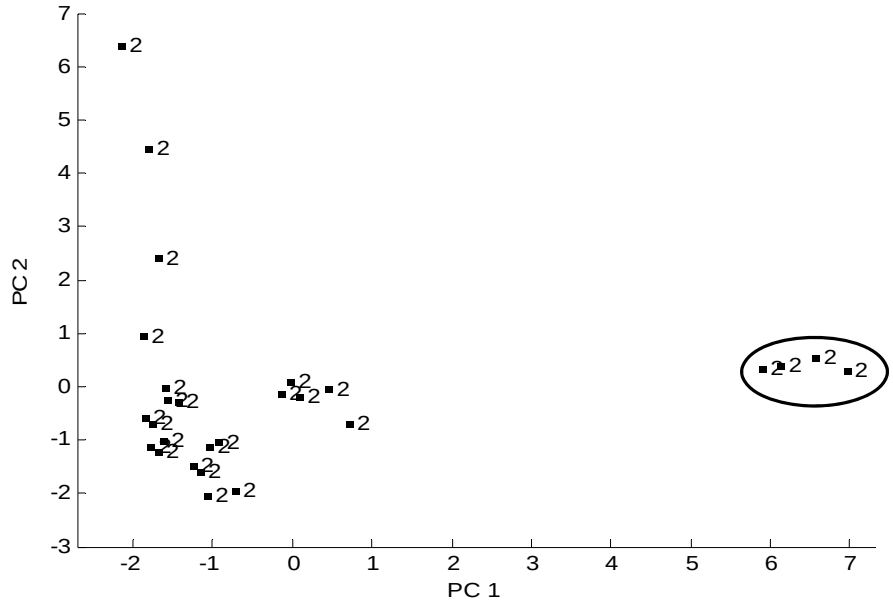

Figure 2.18. PC plot of the two largest principal components of the medium mold count gas chromatograms as determined by the MEA impactor data. Each gas chromatogram is represented as a point in the plot. 4 chromatograms enclosed by an ellipse are outliers in the PC plot of this data.

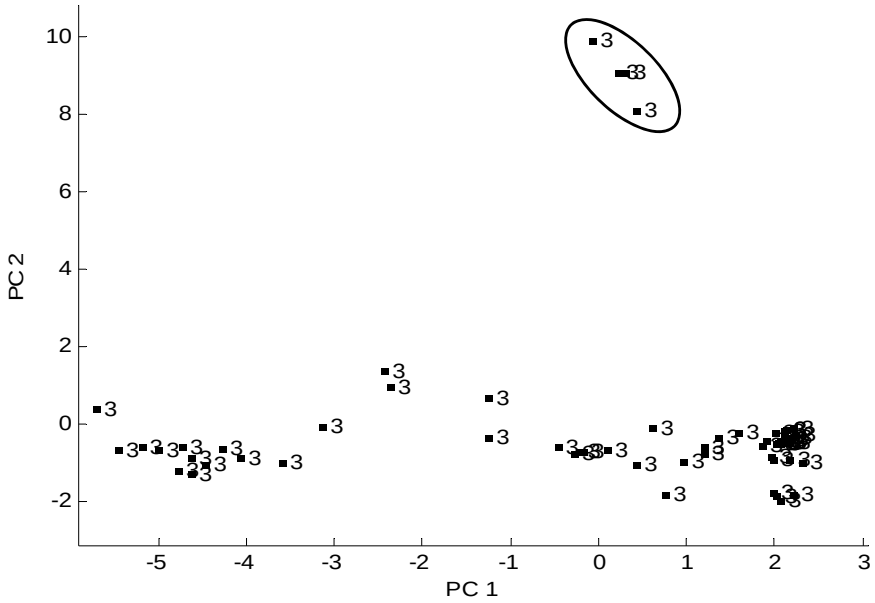

Figure 2.19. PC plot of the two largest principal components of the high mold count gas chromatograms as determined by the MEA impactor data. Each gas chromatogram is represented as a point in the plot. 4 gas chromatograms enclosed by an ellipse are outliers in the PC plot of this data.

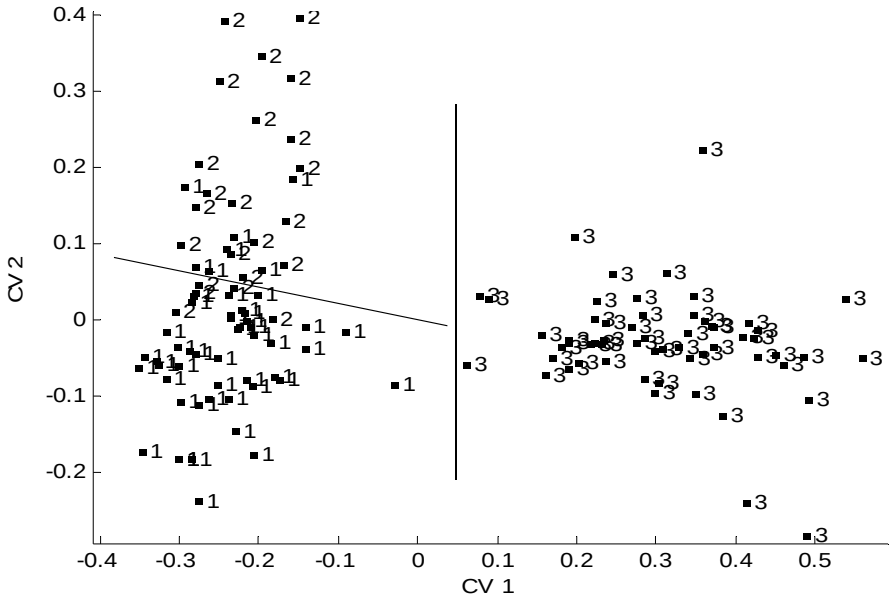

Figure 2.20. A plot of the two largest canonical variates of the GC profiles of the MVOCs with the 4 outliers removed. Each gas chromatogram is represented as a point in the plot. 1 = low mold count, 2 = medium mold count, and 3 = high mold count.

Because the MEA impactor data was more closely correlated to the GC profiles of the VOCs, it was selected for further study using a technique known as cross validation to simulate the ability of a classifier to predict the class membership of an unknown sample. 13 sets of VOC profiles were developed by random selection, where each training set consisted of 122 or 123 gas chromatograms and the corresponding prediction set contained the remaining 11 or 10 gas chromatograms. (In this study, the outliers were deleted from the analysis.) Each gas chromatogram was only present in one of the 13 prediction sets generated. Each training set was analyzed by LDA, QDA, and a 3-layer back propagation neural network (BPNN) using a sigmoid transfer function. The class membership of the volatile profiles in the corresponding prediction set samples was determined using these trained models.

Table 2-2 summarizes the results of the validation study. High classification success rates were obtained for the high mold count gas chromatograms suggesting that a distinct volatile profile representative of the MVOCs was identified by the pattern recognition methodology. For the moderate and low mold count indoor environments, classification success rates were lower. Misclassified GC profiles of VOCs from the moderate mold count were assigned to the low mold count class and vice versa. The overlap between these two classes can probably be attributed to background VOCs, which is obscuring the MVOC profile for the low and moderate mold count data. Because the GC data has been normalized to constant sum, the focus of this pattern recognition analysis study is the concentration pattern of the compounds present in the GC profile, not the total amount of VOCs captured in air sampling by the SPME fiber.

Table 2-2. Validation Set Results

Method	Class	Total	Missed	Success (%)
LDA	1	52	9	82.7
	2	23	11	52.2
	3	58	1	98.3
	Total	133	21	84.2
QDA	1	52	6	88.5
	2	23	18	21.7
	3	58	4	93.1
	Total	133	28	78.9
BPNN (29-3-3)	1	52	3	94.2
	2	23	9	60.9
	3	58	1	98.3
	Total	133	13	90.2

The classification results obtained for LDA were consistent with the CVA plot of the data (see Figure 2.20). LDA outperformed QDA because more model parameters are required to be estimated from the data using the more complex QDA model. The higher classification success rates obtained for BPNN can probably be attributed to the more robust nature of the decision surface generated by the neural network for the profile data. The architecture used for the neural network was based on previous experience with GC data [2-65].

From this study, it can be concluded that GC profile data of MVOCs can be correlated to MEA Impactor data. Gas chromatograms from buildings and residences with high mold counts can be reliably differentiated from moderate and low mold count indoor environments using SPME and GC/MS.

2.6.2 Analysis of Chemical Signals in Red Fire Ants. The test data consisted of gas chromatograms from 125 individual and 235 pooled ant samples obtained from laboratory colonies maintained at the USDA-ARS Fire Ant Project Laboratory in Gainesville, FL. Ants from each colony were fed with sugar-water (1:1) and crickets. Three temporal worker categories were represented in the data: foragers, reserves, and

brood tenders. Brood tenders were identified by disturbing a colony and observing the workers that were carrying broods.

Cuticular hydrocarbons were obtained by soaking individual or pooled ant samples for at least 10 minutes at room temperature in enough hexane (with n-$C_{26}H_{56}$ added for quantitation as an internal standard) to just cover them. After the rinses were complete, the soaks were processed using an Agilent 6890N Network Gas Chromatograph System (Palo, Alto, CA). The Agilent System was equipped with a split-splitless injector, a flame ionization detector, and a DB-1 fused silica capillary column (30 m, 0.25 mm id, 0.25 μm film thickness, J&W Scientific Inc., Folsom, CA). The injector and detector were set at 300^{o}C, and the oven temperature was programmed from 150^{o} to 285^{o}C at 10^{o}/min and then held at 285^{o}C for 4 min. Hydrogen was used as the carrier gas and nitrogen was used as the makeup gas. The chromatographic data **(**see Figure 2.21**)** were processed using Agilent Technologies GC Chemstation G2071AA A.10.01 (Agilent Technologies, Palo Alto). Peak retention times were compared to standard cuticular hydrocarbons from *S. invicta*. If there was ambiguity in a peak assignment, then mass spectra were obtained on an Agilent 5973 Network Mass Selective Detector US10480853 using Agilent 6890N Network Gas Chromatography System US10124023. For the GC/MS runs, the injector was set at 300^{o}C and the oven temperature was programmed from 100^{o} to 285^{o}C at 10^{o}/min, and then held at 285^{o}C for 10 min with the transfer line set at 285^{o}C. Helium was used as the carrier gas for the column. GC/MS data were processed using Agilent Enhanced GC/MS Chemstation software G1701DA version D.00.00.38.

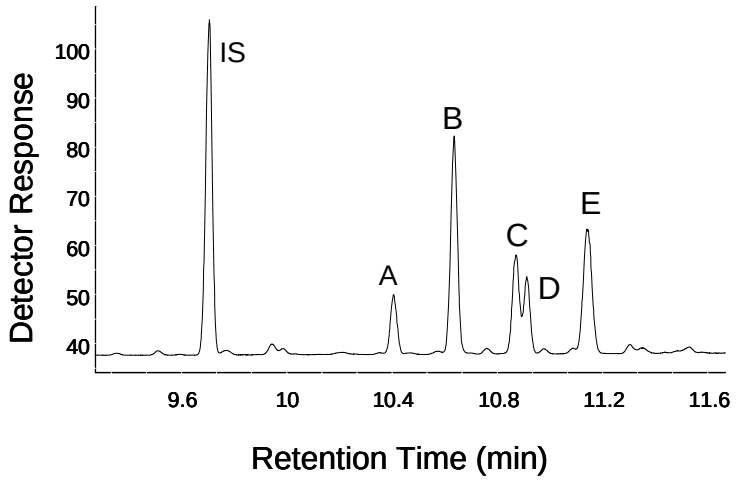

Figure 2.21. Gas chromatographic trace of cuticular hydrocarbons from *S. invicta.* The compounds eluting off the capillary column were identified and quantified by GC/MS: (a) heptacosane, (b) 13-methylheptacosane, (c) 13, 15-dimethylheptacosane, (d) 3-methylheptacosane, and (e) 3, 9-dimethylheptacosane. Hexacosane was added for quantitation as an internal standard (IS).

Several questions were addressed in this study. First, is there an advantage to analyze the cuticular hydrocarbon extracts of pooled versus individual ant samples? Second, is the cuticular hydrocarbon patterns of *S. invicta* workers correlated with their age-linked temporal caste? Third, do the hydrocarbon profiles of *S. invicta* significantly differ for each laboratory colony? And fourth, can the same methods used to distinguish colony of origin be used to track colony cuticular hydrocarbon changes over time? Previous studies [2-66 to 2-72] performed on differences in cuticular hydrocarbon profiles for carpenter ants and for *C. niger* have shown separation by colony, and temporal caste. For *S. invicta*, it has been previously reported, albeit in a preliminary study, that cuticular hydrocarbon patterns were consistent within colonies for a given sampling time, but they varied sufficiently from colony to colony. The cuticular hydrocarbon profiles of *S. invicta* colonies also changed over time [2-73 to 2-75]. In these studies, the results were reported on a subset of the data collected and/or the multivariate methods used were limited in

their ability to extract information from the cuticular hydrocarbon profiles. Furthermore, there was no attempt to deconvolve the confounding effects of the biological variables investigated. For these reasons, a more exhaustive investigation of the biological variables that influence the cuticular hydrocarbon profiles of *S. invicta* was undertaken.

To answer the first question, gas chromatograms of cuticular hydrocarbon extracts obtained from 65 pooled, reserve ant samples from five laboratory colonies were collected and analyzed using LDA, QDA, and RDA. The goal was to separate one colony from another. The results of this study are summarized in Figure 2.22a.

Because there was no validation set, Monte Carlo simulation studies were performed to assess the statistical significance of the classification scores. The goal was to estimate the separation in the data due to chance using LDA, QDA, and RDA. For these studies, data sets comprised of random numbers were generated. Both Gaussian and uniform distributions were employed. A method described in previous publications [2-76 and 2-77] was used to compute the expected level of chance classification for both the pooled and individual ant samples. For each chance classification study, 100 data sets consisting of random numbers were generated. The statistical properties of the simulated data (i.e., dimensionality, number of samples, class membership distribution, and covariance structure) were identical to the actual data for which we wish to determine its degree of classification due to chance. For each random data set, its degree of separability was assessed. The number of occurrences of several degrees of separation (e.g., at least 70% of the patterns were correctly classified or at least 80% of the patterns were correctly classified) was noted and the fraction of the total number of occurrences (cumulative probability) for each degree of separation was plotted against the percentage

of patterns correctly classified. These cumulative distribution curves provide information about the likelihood that a particular classification result is due to chance. For example, if the classification score obtained for real data is 80% but the mean classification success rate for the simulated data is only 37% and the probability of achieving 65% correct classification due to chance is zero (see Figure 2.22b), the score obtained using the real data (80%, see Figure 2.22a) would be considered statistically significant.

Results from these Monte Carlo simulation studies are summarized in Figures 2.22a and 2.22b. For the QDA classification study involving the pooled ant samples (see Figure 2.22b), one hundred data sets consisting of random numbers were generated. The statistical properties of the simulated data (i.e., dimensionality, number of samples, class membership distribution, and covariance structure) were identical to those of the 65 pooled ant samples. The separabililty of each random data set was assessed using QDA and a cumulative probability plot was generated for the random data. The mean classification score of the 100 random data sets was also computed and compared to the classification score obtained in the QDA study for the GC data. Since the mean classification success rate of the simulated data was only 57.3%, the classification score obtained for QDA using GC data expressed in nanograms was judged to be statistically significant (see Figure 2.22b).

Figure 3a summarizes the results obtained for 125 individual ant samples collected from the same laboratory colonies as the pooled ant samples. Each colony is represented by 25 reserve workers. Figures 2.23a and 2.23b summarizes the results of the chance classification studies for this data. Results for RDA were not reported because

the values of γ and λ that gave the best classification for colony were 0, 0 which corresponds to QDA.

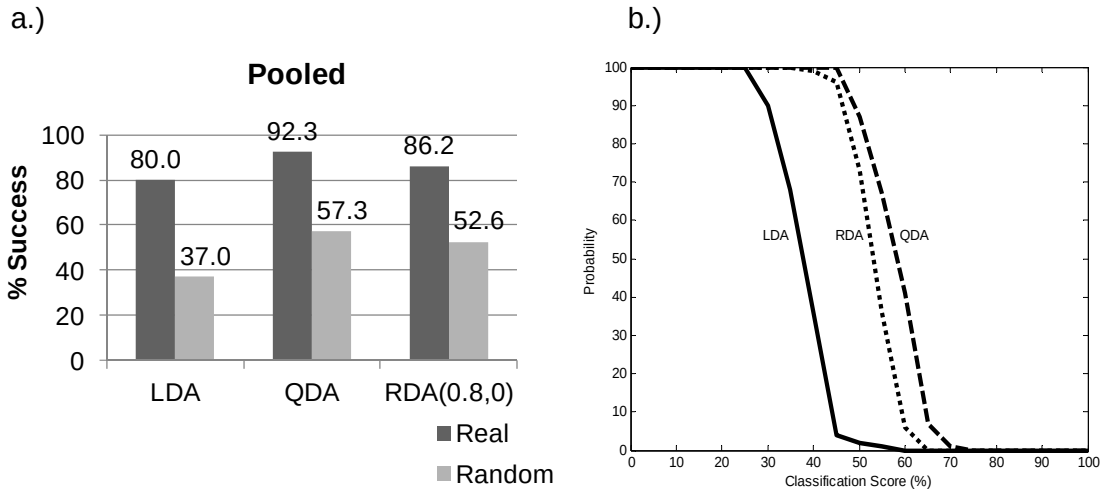

Figure 2.22. a) Comparison of the classification scores for the pooled ant samples versus the average degree of separation in the data due to chance. b) Probability of achieving any degree of separability due to chance for the pooled ant samples with RDA (0.8, 0), LDA, and QDA.

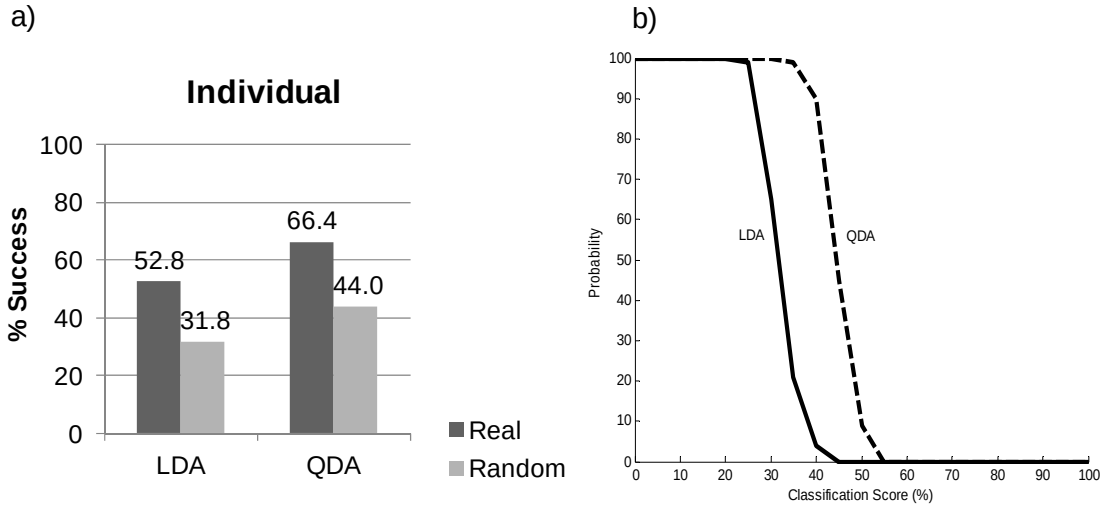

Figure 2.23. a) Comparison of the classification scores for the individual ant samples versus the average degree of separation in the data due to chance. b) Probability of achieving any degree of separability due to chance for the pooled ant samples with LDA, and QDA.

For the pooled ant samples, there were 65 independent samples equally distributed among 5 classes. For the individual ants, there were 125 independent samples distributed equally among 5 classes. As the number of objects in a data set increases, the degree of separation due to chance will decrease. For this reason, chance classifications are lower for the individual ants than for the pooled ant samples.

An examination of Figures 2.22 and 2.23 reveals that differences between the classification-success rates obtained for real data versus random data are smaller for the individual ant samples. This suggests that pooling the samples enhances the recognition of patterns indicative of colony in the cuticular hydrocarbon profiles of *S. invicta* when pattern recognition techniques are used to analyze the data. Evidently, the contribution of the ant's individual pattern to the overall hydrocarbon profile pattern obscures information about colony of origin in GC traces obtained from cuticular hydrocarbon extracts. For these reasons, it is suggested that cuticular hydrocarbon profiles from pooled ant samples, not individual ant samples be studied to seek meaningful relationships between cuticular hydrocarbon profiles and biological variables such as colony of origin and temporal (social) caste.

To address the question about patterns in the hydrocarbon profiles indicative of temporal caste, it was necessary to collect additional data. A set of 170 gas chromatograms of cuticular hydrocarbon extracts were obtained from 170 *S. invicta* samples. Each ant sample contains hydrocarbons extracted with hexane from the cuticle of 100 individual ants. The ant samples were obtained from 5 laboratory colonies (which were not the same laboratory colonies used in the pooled versus individual ant sample study), 3 temporal castes (foragers, reserves, and brood tenders), and the colonies were

sampled at four different time periods (three in the spring and summer, and one in the winter).

The first step was to analyze the data using PCA. Figure 2.24 is a plot of the two largest principal components of the 170 pooled *S. invicta* samples and the five GC peaks that characterize each sample. Each pooled ant sample is represented as a point in the principal component map of the data. It is evident from the plot that sample 31 (colony 1) is an outlier, and this sample was subsequently deleted from the analysis because of the adverse effect that outliers can have on the performance of pattern recognition methods.

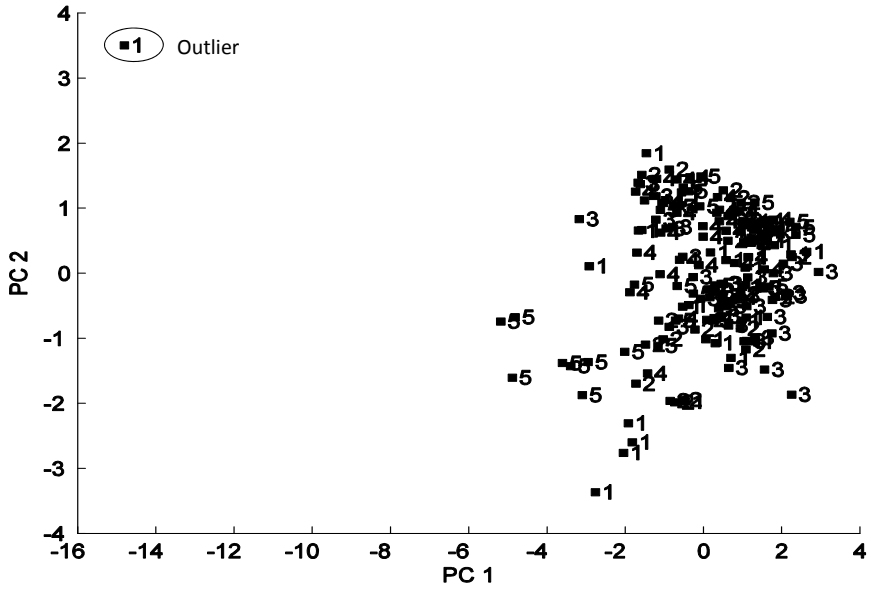

Figure 2.24. A plot of the two largest principal components of the 170 pooled red fire ant samples and the five high molecular weight hydrocarbon compounds that characterize the cuticle of *S. invicta*. Each ant sample is represented as a point in the principal component map of the data. 1 is a pooled ant sample from colony 1; 2 is a pooled ant sampled from colony 2; 3 is a pooled ant sample from colony 3; 4 is a pooled ant sample from colony 4; 5 is a pooled ant sample from colony 5.

For each laboratory colony, the data were divided into three categories according to temporal caste. Previous analyses of the cuticular hydrocarbons [2-74 and 2-75] using PCA to analyze the GC profiles of the hydrocarbon soaks revealed patterns indicative of the temporal caste of the *S. invicta* samples in only one of the five laboratory colonies investigated. Therefore, CVA was performed to separate the pooled ant samples in each colony by temporal caste. The results of this study are summarized in Figures 2.25, 2.26, 2.27, 2.28, and 2.29. Each pooled ant sample is represented as a point in the CVA map of the data. Foragers, which are represented by the symbol 1, could be readily differentiated from brood tenders (represented by the symbol 2) and reserves (represented by the symbol 3) in four of the five laboratory colonies (colonies 1, 2, 4, and 5) investigated. Because reserves can assume the role of brood tenders, it is plausible that both reserves and the brood tenders could have similar hydrocarbon profiles.

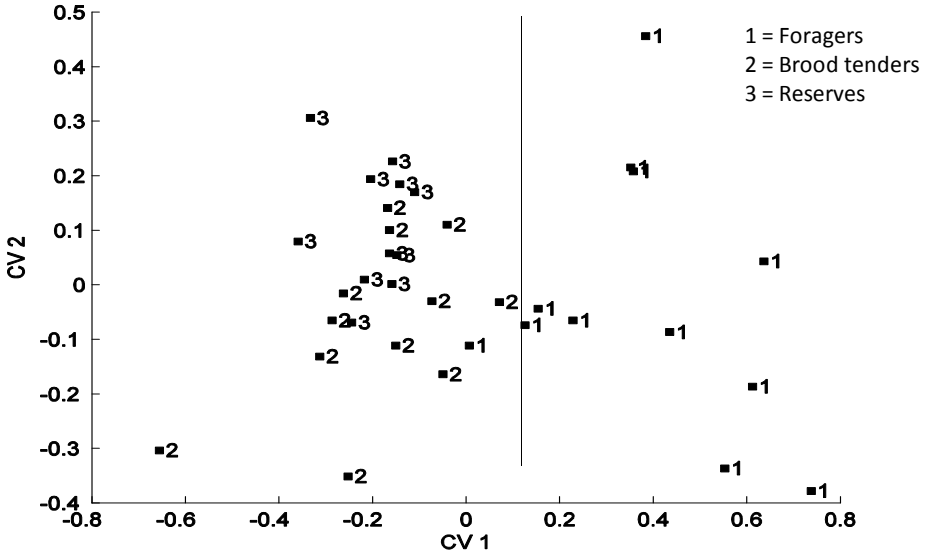

Figure 2.25. A plot of the two largest canonical variates of the pooled ant samples obtained from colony 1. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled forager ant sample; 2 is a pooled reserve ant sample; and 3 is a pooled brood tender ant sample. Separation of the foragers from brood tenders and reserves in the plot is evident.

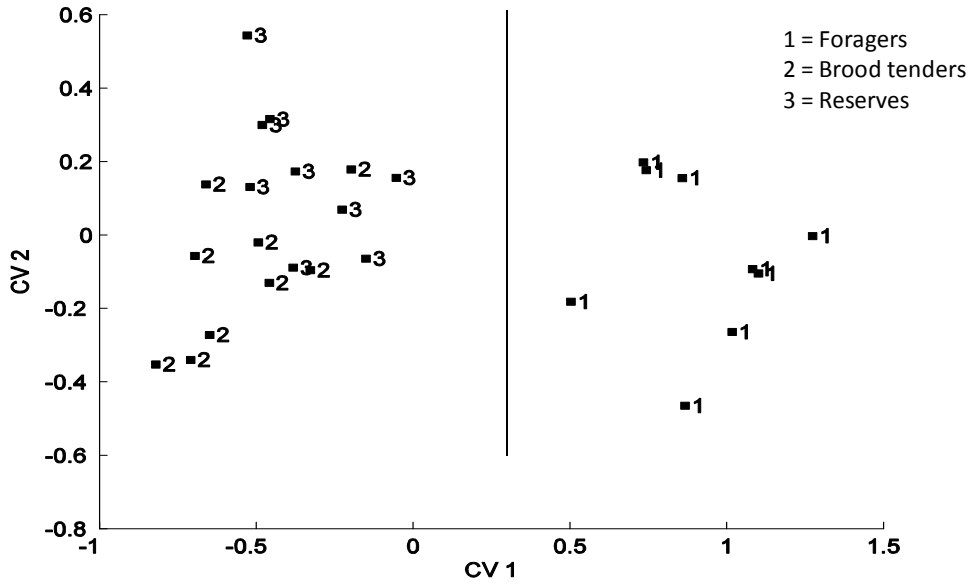

Figure 2.26. A plot of the two largest canonical variates of the pooled ant samples obtained from colony 2. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled forager ant sample; 2 is a pooled brood tender ant sample; and 3 is a pooled reserve ant sample. Separation of the foragers from brood tenders and reserves in the plot is evident.

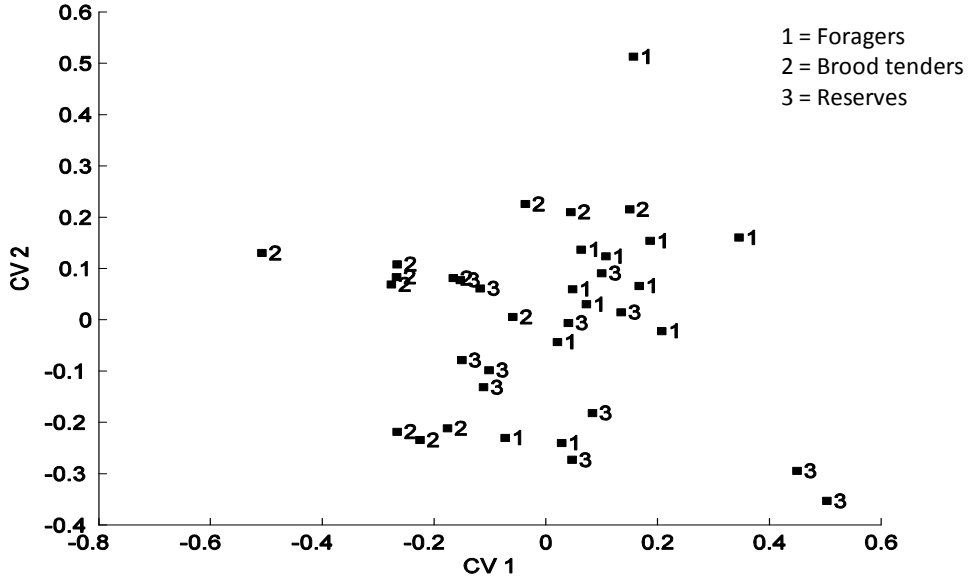

Figure 2.27. A plot of the two largest canonical variates of the pooled ant samples obtained from colony 3. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled forager ant sample; 2 is a pooled brood tender ant sample; and 3 is a pooled reserve ant sample. Clustering of the pooled ant samples on the basis of social caste is not observed in this plot.

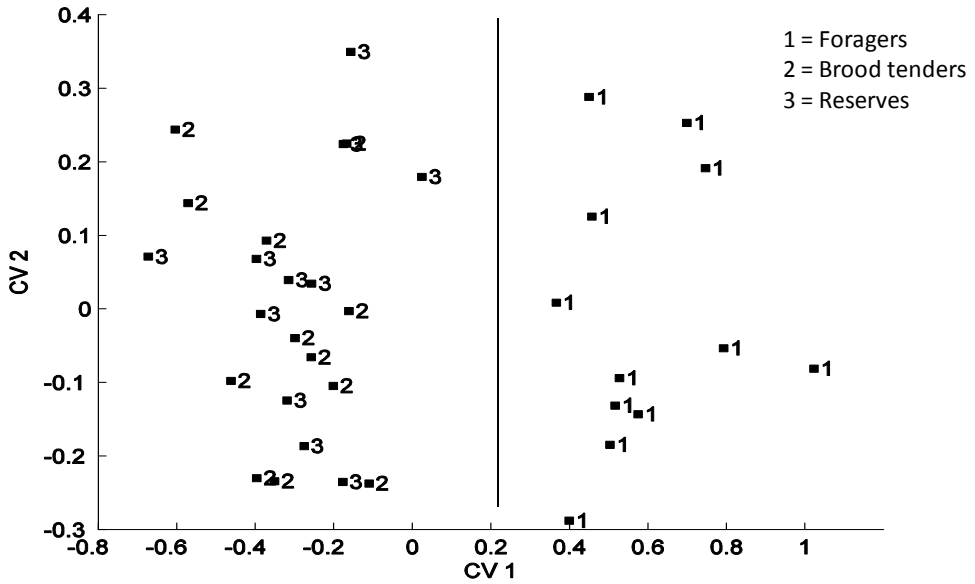

Figure 2.28. A plot of the two largest canonical variates of the pooled ant samples obtained from colony 4. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled forager ant sample; 2 is a pooled brood tender ant sample; and 3 is a pooled reserve ant sample. Separation of the foragers from brood tenders and reserves in the plot is evident.

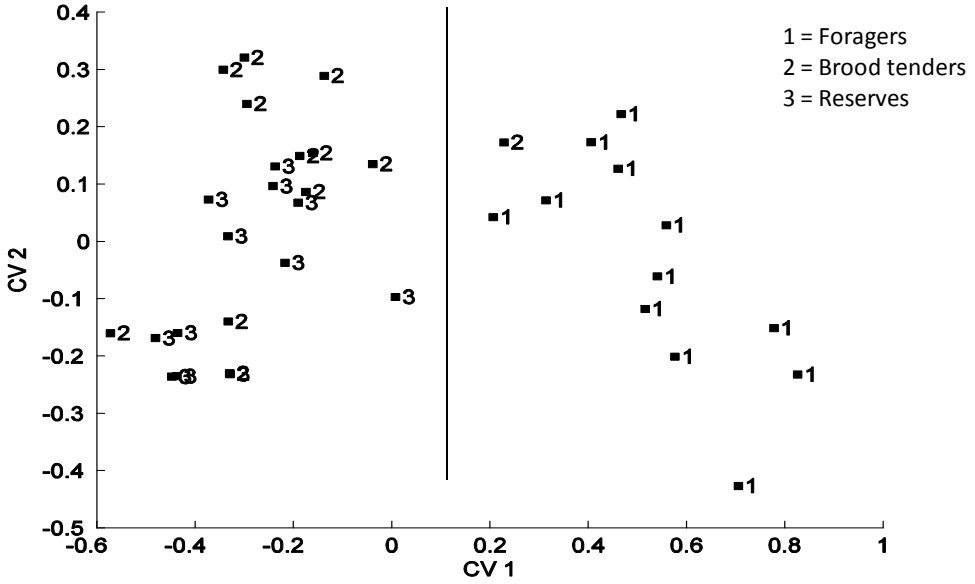

Figure 2.29. A plot of the two largest canonical variates of the pooled ant samples obtained from colony 5. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled forager ant sample; 2 is a pooled brood tender ant sample; and 3 is a pooled reserve ant sample. Separation of the foragers from brood tenders and reserves in the plot is evident.

Figure 2.30 shows a CVA plot of the GC data from colonies 1, 2, 4, and 5. The data was divided into three classes according to social caste. Again, separation of the foragers from the reserves and brood tenders is evident. When social caste is investigated on a per colony basis, separation of the foragers from the reserves and the brood tenders occurred on the first canonical variate. Upon investigating social caste as the class variable using GC data from colonies 1, 2, 4, and 5, separation of the foragers from the reserves and brood tenders occurred on the second canonical variate. These results (see Figures 2.25 thru 2.29 versus Figure 2.30) confirm that patterns correlated to temporal caste are present in the cuticular hydrocarbon profiles of *S. invicta*, but are not the major source of variation in the hydrocarbon profiles obtained from pooled ant samples.

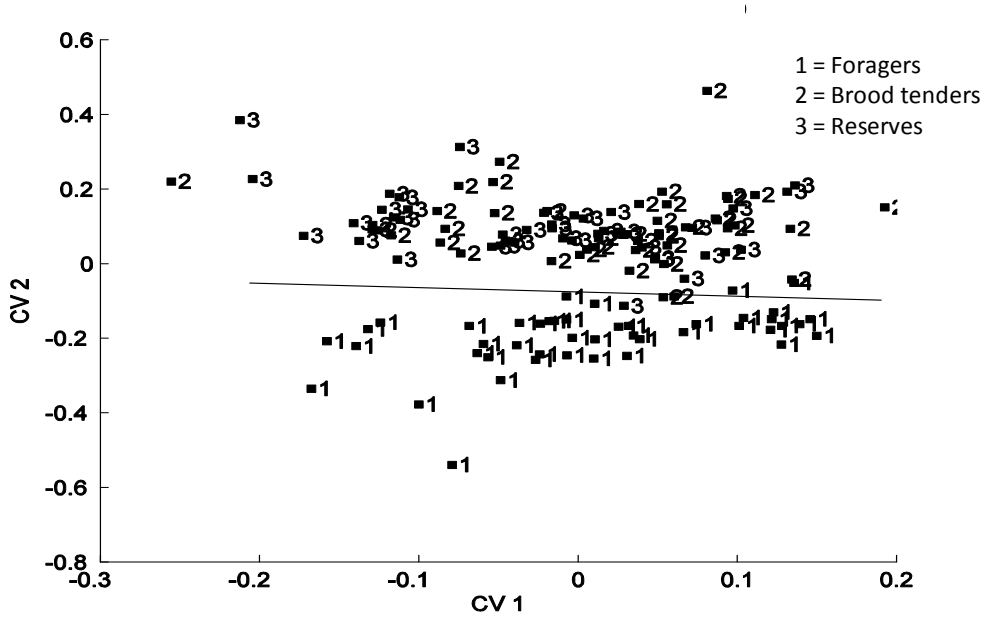

Figure 2.30. A plot of the two largest canonical variates of the pooled ant samples obtained from all five colonies. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled forager ant sample; 2 is a pooled brood tender ant sample; and 3 is a pooled reserve ant sample. Separation of the foragers from the brood tenders and reserves in the plot is evident.

Temporal changes in the cuticular hydrocarbon profiles were also investigated. For each colony, the data was divided into four categories according to the time period of sampling. CVA was performed to separate pooled ant samples in each colony by time period. Monte Carlo simulation experiments were also performed in tandem to assess the degree of separation in the data due to chance. One hundred data sets comprised of random numbers with Gaussian distributions that had statistical properties (i.e. dimensionality, number of samples, class membership distribution, and covariance structure) identical to those of the real data were generated. CVA was performed on a data set that was an average of the 100 random data sets generated. The results are summarized graphically, see Figures 2.31 through 2.40. It is evident from the Monte Carlo simulation experiments that separation of the pooled ant samples by time period in the CVA plots cannot be attributed to chance.

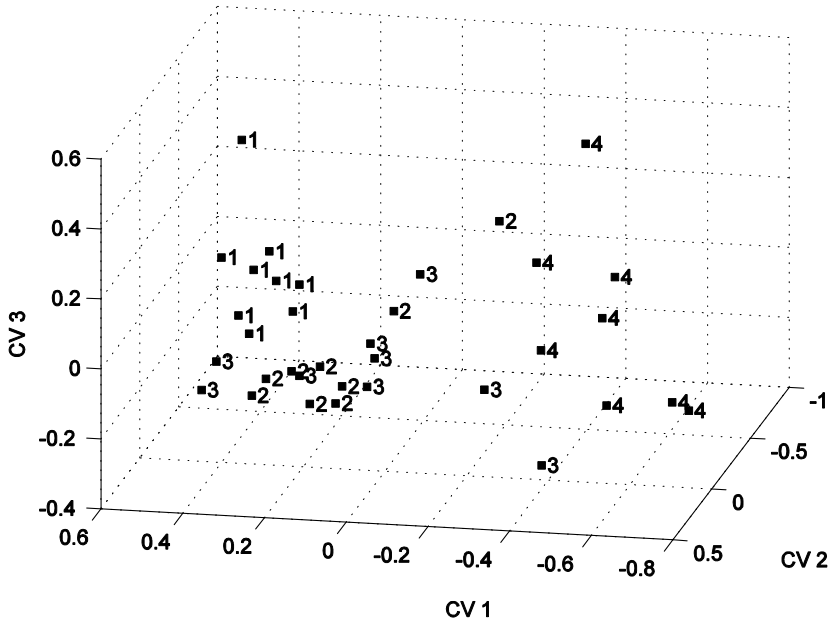

Figure 2.31. A plot of the three largest canonical variates of the pooled ant samples obtained from colony 1. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled ant sample from time period 1; 2 is a pooled ant sample from time period 2; 3 is a pooled ant sample from time period 3; and 4 is a pooled ant sample from time period 4. Clustering of the pooled ant samples by time period is evident in this plot.

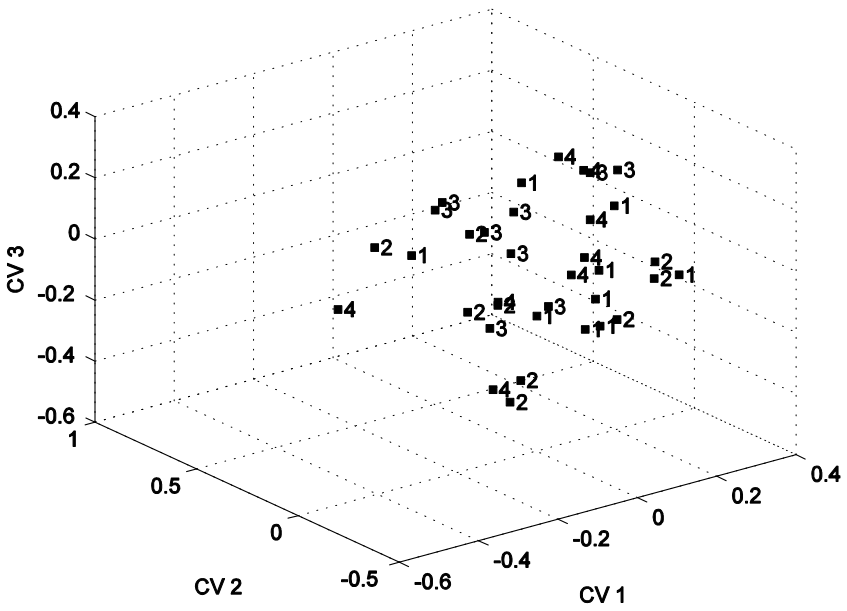

Figure 2.32. A plot of the three largest canonical variates of the simulated data sets for colony 1. Clustering of the pooled ant samples by time period is not evident in this plot.

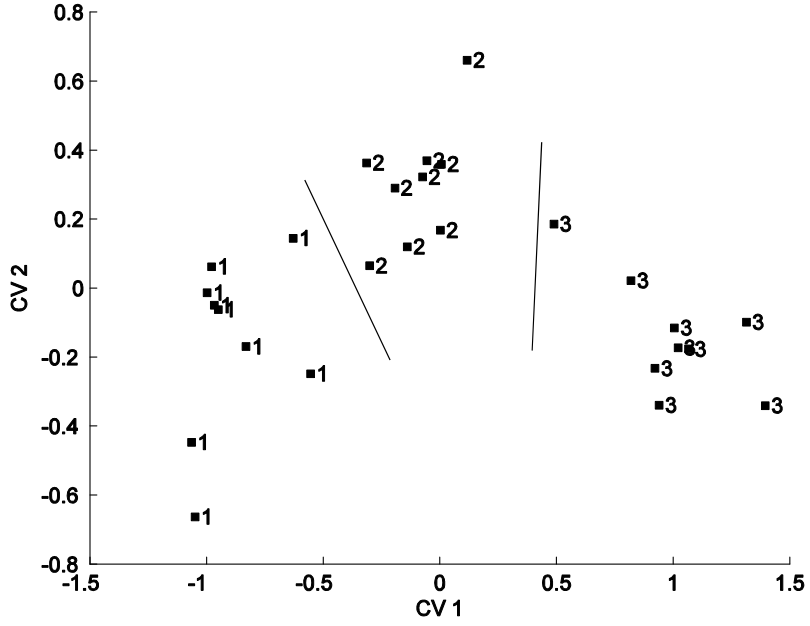

Figure 2.33. A plot of the three largest canonical variates of the pooled ant samples obtained from colony 2. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled ant sample from time period 1; 2 is a pooled ant sample from time period 2; 3 is a pooled ant sample from time period 3; and 4 is a pooled ant sample from time period 4. Clustering of the pooled ant samples by time period is evident in this plot.

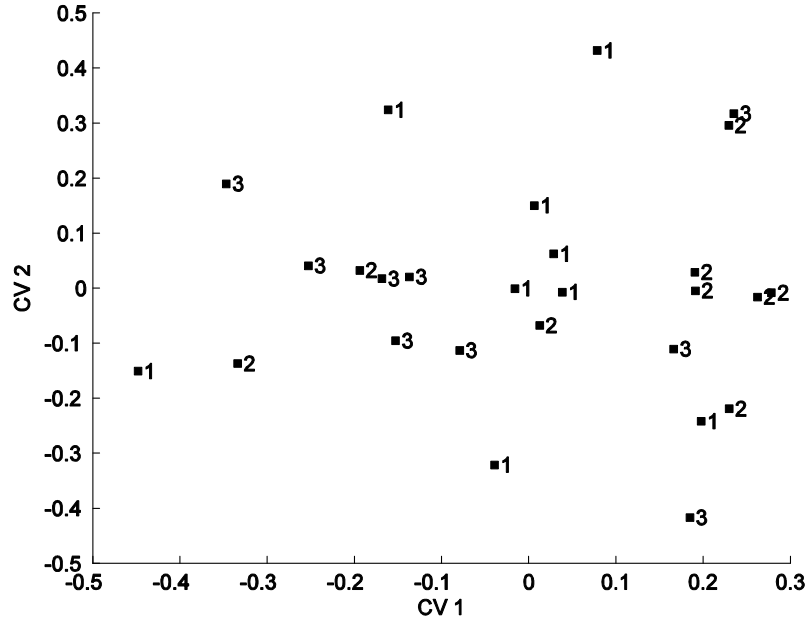

Figure 2.34. A plot of the three largest canonical variates of the simulated data sets for colony 2. Clustering of the pooled ant samples by time period is not evident in this plot.

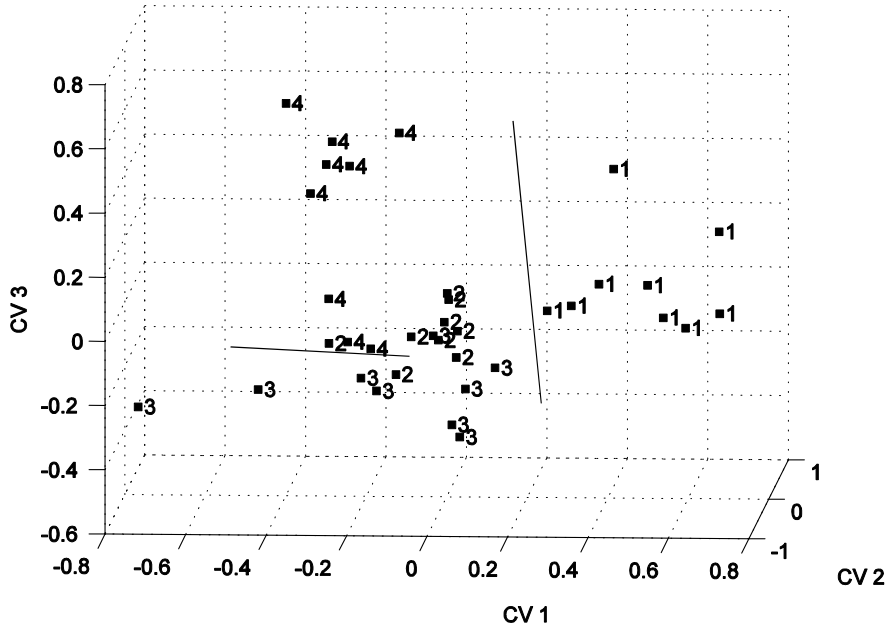

Figure 2.35. A plot of the three largest canonical variates of the pooled ant samples obtained from colony 3. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled ant sample from time period 1; 2 is a pooled ant sample from time period 2; 3 is a pooled ant sample from time period 3; and 4 is a pooled ant sample from time period 4. Clustering of the pooled ant samples by time period is evident in this plot.

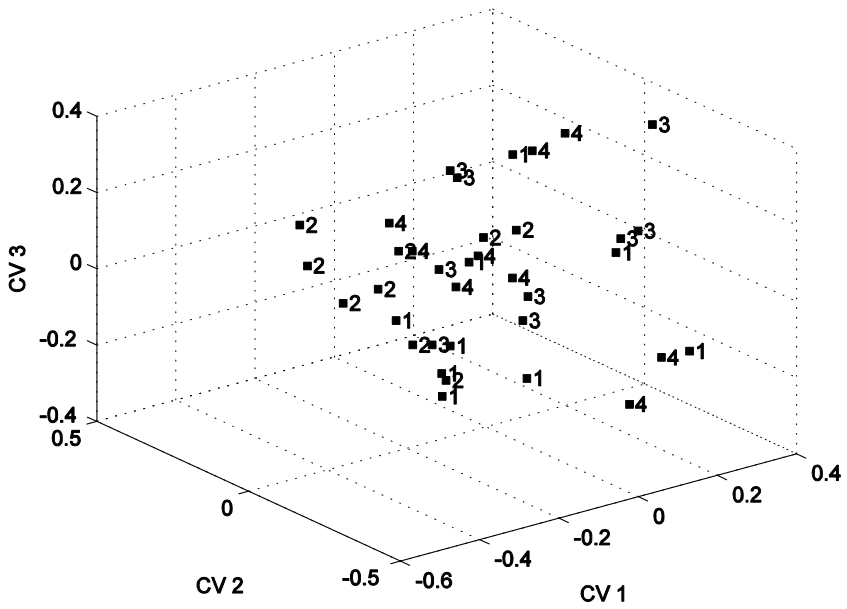

Figure 2.36. A plot of the three largest canonical variates of the simulated data sets for colony 3. Clustering of the pooled ant samples by time period is not evident in this plot.

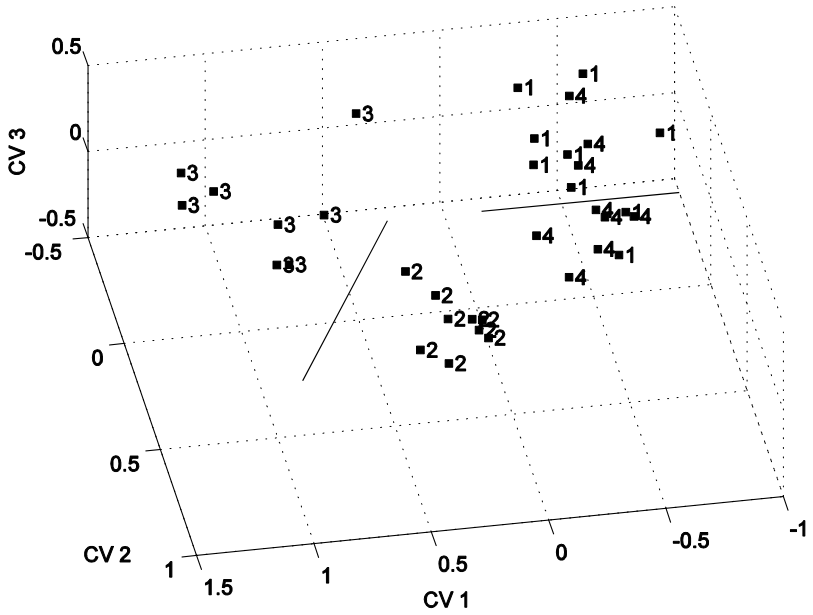

Figure 2.37. A plot of the three largest canonical variates of the pooled ant samples obtained from colony 4. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled ant sample from time period 1; 2 is a pooled ant sample from time period 2; 3 is a pooled ant sample from time period 3; and 4 is a pooled ant sample from time period 4. Clustering of the pooled ant samples by time period is evident in this plot.

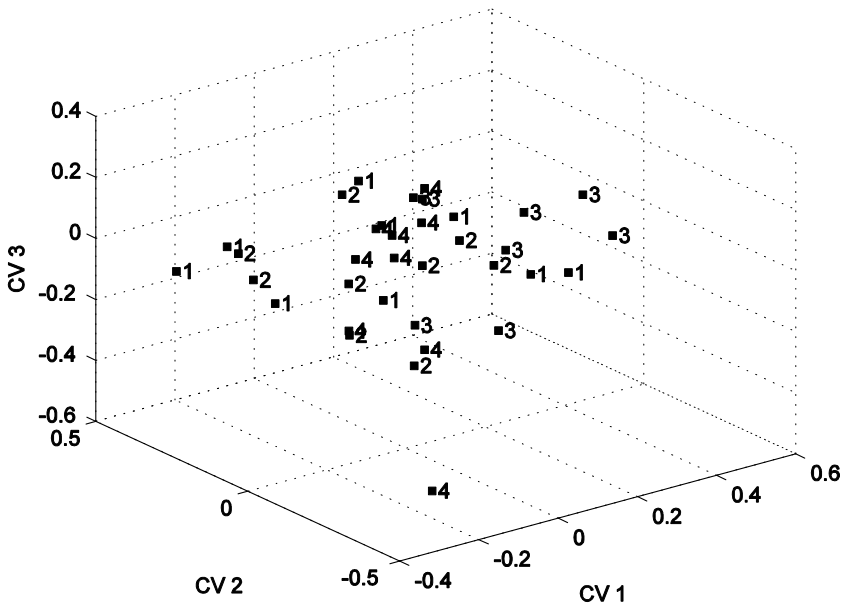

Figure 2.38. A plot of the three largest canonical variates of the simulated data sets for colony 4. Clustering of the pooled ant samples by time period is not evident in this plot.

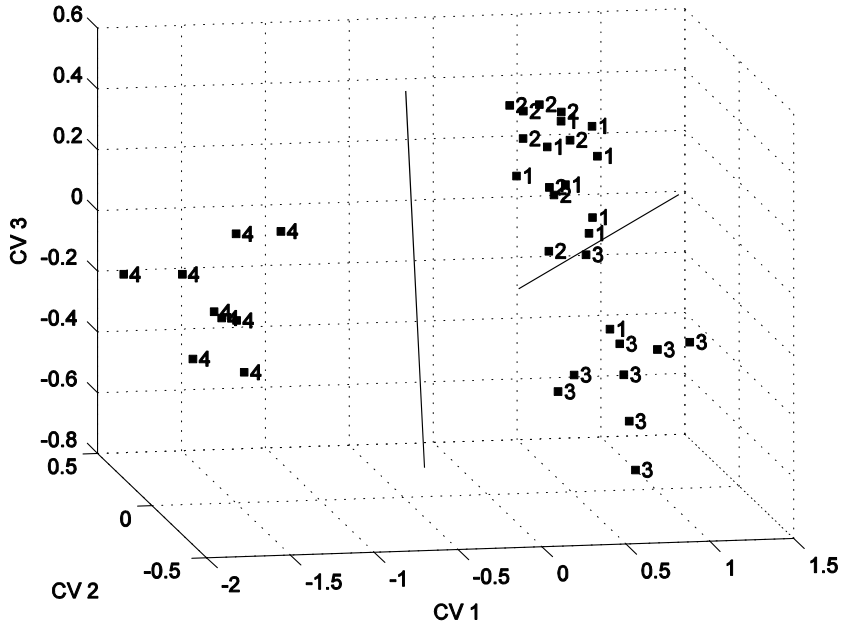

Figure 2.39. A plot of the three largest canonical variates of the pooled ant samples obtained from colony 5. Each pooled ant sample is represented as a point in the CVA map of the data. 1 is a pooled ant sample from time period 1; 2 is a pooled ant sample from time period 2; 3 is a pooled ant sample from time period 3; and 4 is a pooled ant sample from time period 4. Clustering of the pooled ant samples by time period is evident in this plot.

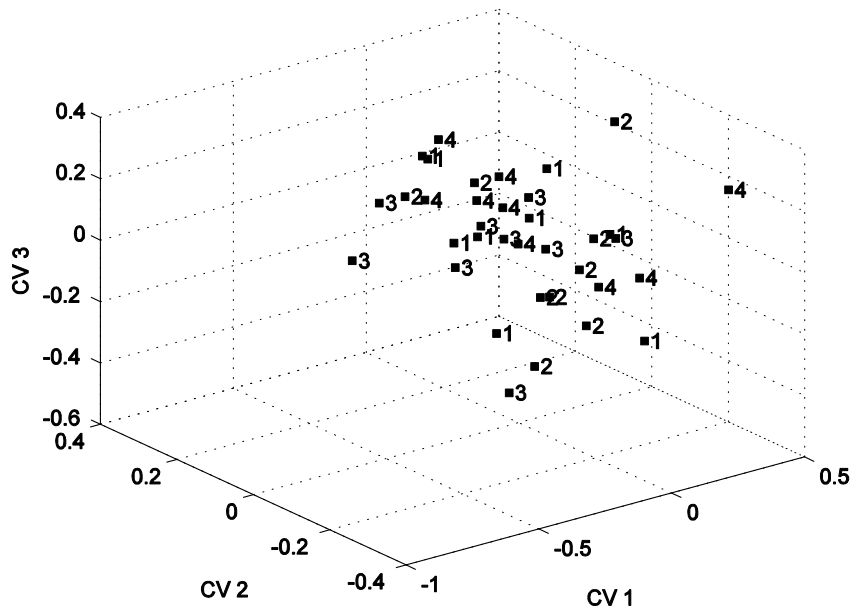

Figure 2.40. A plot of the three largest canonical variates of the simulated data sets for colony 5. Clustering of the pooled ant samples by time period is not evident in this plot.

Each laboratory colony exhibited a different pattern of change with time. In our previous studies [2-74 and 2-75], we were able to determine that *S. inivcta* cuticular hydrocarbon profiles from time period four were different from the cuticular hydrocarbon profiles of the other time periods and that only one laboratory colony exhibited a systematic change in its cuticular hydrocarbon profile over time. The results obtained in the present study indicate that cuticular hydrocarbon profiles of each *S. invicta* colony change with time. However, the pattern of change as shown in each CVA plot is different for each colony. In some instances, all of the time periods are well separated whereas in other instances only two of the four time periods are well separated. This should not come as a surprise for the cuticular hydrocarbon profiles of S. invicta may be a dynamic system that undergoes changes with time and the nature of this change will be different for each colony.

The cuticular hydrocarbon profiles of *S. invicta* were also found to be characteristic of the individual colony. For each time period, the data was divided into five categories according to the colony of origin of the pooled ant samples. Again, decision surfaces were developed from the five major hydrocarbon components. QDA was used to classify the data by colony for each time period. Monte Carlo simulation studies were also performed to assess the degree of separation in the data due to chance. The results of these studies are summarized in Figure 2.41. Clearly, the cuticular hydrocarbon profiles of the red fire ants are characteristic of the colony of origin for a given time period. However, it was surprising that our Monte Carlo simulations revealed high chance classification success rates for the case of 45 samples distributed equally among 4 classes with each sample characterized by 5 measurements using QDA. Chance classification may be a more serious problem with QDA than was previously thought.

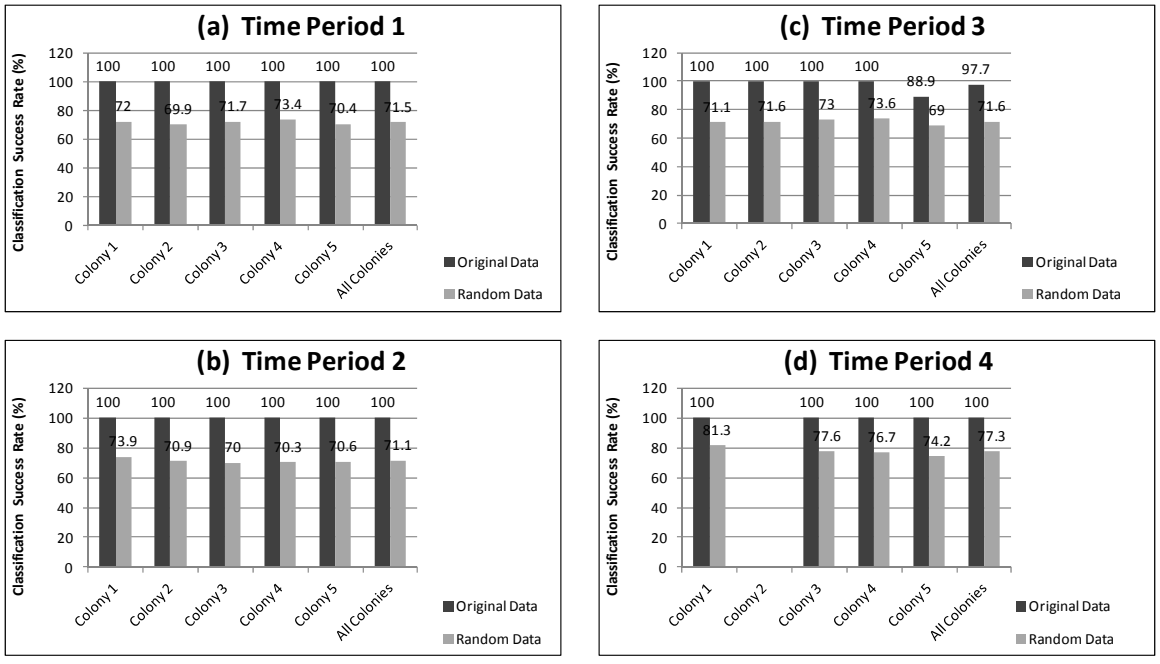

Figure 2.41. A comparison of the classification scores for colony versus the degree of separation in the data due to chance at (a) time period 1, (b) time period 2, (c) time period 3, and (d) time period 4.

QDA was also used to classify the data by colony across all time periods. The results of this study are summarized in Figure 2.42. To assess the significance of these classifications, Monte Carlo simulation studies were performed. The results of these studies are summarized in Figures 2.42 and 2.43. Differences in chance classification across all time periods versus individual time periods were due to the larger number of samples involved in colony classification across all time periods. Using the Monte Carlo simulation studies as a benchmark, it is evident that classifications obtained in the QDA study across all time periods are significant for four of the five laboratory colonies. When the classifications for colony from each time period (see Figure 2.41) are compared to the classifications for colony across all time periods (see Figure 2.42), it is evident that cuticular hydrocarbon profiles of *S. invicta* change with time, which can confound the classification of GC profile data by colony using pattern recognition techniques. This is most evident in the cuticular hydrocarbon profiles of pooled samples of *S. invicta* from colonies 4 and 5. The changes in the cuticular hydrocarbon profiles that occurred in laboratory colonies 4 and 5 over time caused their cuticular hydrocarbon profiles to overlap. For example, cuticular hydrocarbon profiles from colony 5 at time period 1 were similar to those of colony 4 at time period 3.

It has been previously reported [2-74 and 2-75] that 4 of 5 laboratory colonies could be differentiated on the basis of their cuticular hydrocarbon profiles. These studies were carried out by formulating the problem as a series of binary classifications using the linear learning machine and related linear nonparametric methods of classification. In the current study, better multivariate analysis methods have been used and the analysis of the cuticular hydrocarbon data was more detailed in its scope. From the current study, we

have learned that all five colonies could be separated on the basis of their cuticular hydrocarbon profiles when data from each time period is analyzed separately. When colonies are analyzed using data from all of the time periods, the classifications become confounded which considerably strengthens the previously stated conclusion that cuticular hydrocarbon profiles of red fire ants change over time.

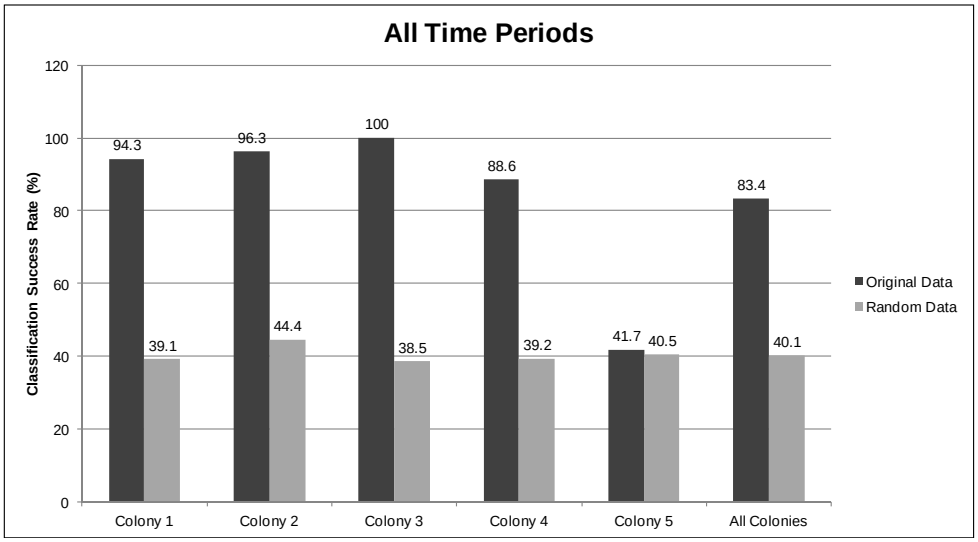

Figure 2.42. A comparison of the classification scores for colony across all time periods versus the degree of separation in the data due to chance.

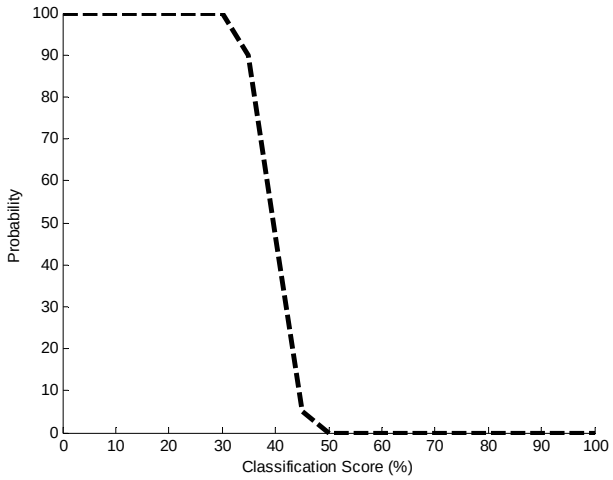

Figure 2.43. Probability of achieving any degree of separation in the data due to chance for all 5 laboratory colonies using QDA. There is a 50% probability of achieving a classification score of 40.1%.

The GC traces representing ant cuticle extracts can be related to colony of origin and temporal caste. These results support a correlative role for cuticular hydrocarbons in nestmate recognition. However, it remains for specific behavioral bioassays with purified hydrocarbons to determine if cuticular hydrocarbons are in fact used by *S. invicta* in nestmate recognition. In addition, the re-analysis of temporal caste and time on cuticular hydrocarbon patterns demonstrates that sampling time and social caste must be taken into account to avoid unnecessary variability and possible confounding. This and the fact that foragers could not be separated from reserves and brood-tenders in all five laboratory colonies suggests that cuticular hydrocarbons as a class of compounds cannot model every facet of nestmate recognition in *S. invicta* which in turn suggests a potential role for other compounds in the discrimination of alien conspecifics from nestmates.

It is truly remarkable that all of this information (social caste, colony of origin, and time period) is contained in the concentration pattern of five high molecular weight hydrocarbons which comprise a dynamic system that changes with time with the nature of these changes being different for each colony. Neither colony of origin, social caste, nor time period is the major source of variation in the data although distinct patterns in the concentration profiles of the five hydrocarbons characteristic of these biological variables can be identified.

This study also demonstrates the importance of using pattern recognition methods to analyze complex chromatographic data sets and to seek meaningful relations between chemical constitution and biological variables. The classification of complex biological samples on the basis of their GC profiles can be complicated by two factors: (1)

confounding of the desired group information by other systematic variations present in the data and (2) random or chance classification effects. The existence of these complicating relationships is an inherent part of fingerprint type data.

REFERENCES

2-1. B. K. Lavine and J. R. Workman, “Chemometrics: Past, Present, and Future,” in B. K. Lavine (Eds.) Chemometrics and Chemoinformatics, ACS Symposium Series 894, Oxford University Press, 2005.

2-2. I. T. Jolliffee, “Principal Component Analysis”, Springer-Verlag, New York, 1986.

2-3. D. L. Massart, and L. Kaufman, “The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis”, John Wiley & Sons, New York, 1983.

2-4. R. G. Brereton (Eds.), “Multivariate Pattern Recognition in Chemometrics”, Elsevier, Amsterdam, 1992.

2-5. S. D. Brown, “Chemical systems under indirect observation: latent properties and chemometrics,” Appl. Spec., 1995, 49, 14A-31A.

2-6. S. J. Haswell (Eds.), “Practical Guide to Chemometrics”, Marcel Dekker, NY. 1992.

2-7. G. L. McLachlan, “Discriminant Analysis and Statistical Pattern Recognition”, John Wiley & Sons, New York, 1992.

2-8. J. Mandel, “The regression analysis of collinear data,” J. Res. NBS, 1985, 90(6), 465-476.

2-9. G. Golub, C. Van Loan, “Matrix Computations”, Johns Hopkins University Press, Baltimore, 1971.

2-10. L. Kaufman, and P. J. Rousseeuw, “Finding Groups in Data”, Wiley-Interscience, NY 1990.

2-11. J. C. Bezdek, C. Coray, R. Gunderson, and J. Watson, “Detection and characterization of cluster substructure II: fuzzy c-variaties and convex combinations thereof,” SIAM J. Appl. MATH, 1981, 40, 358-372.

2-12. R. W. Gunderson, “An adaptive FCV clustering algorithm,” Int. J. Machine Studies,” 1983, 19, 97-104.

2-13. K. E. Thrane and R. W. Gunderson, “Source allocation of organic air pollutants by application of fuzzy c-variaties pattern recognition,” Anal. Chim. Acta, 1986, 191, 309-317.

2-14. T. Jacobson, and R. W. Gunderson, “Cluster analysis of beer flavor components. II. A case study of yeast strain and brewery dependency,” American Society of Brewery Chemists, 1981, 41, 73-77.

2-15. B. K. Lavine, A. Stine, and H. T. Mayfield, “Gas chromatography-pattern recognition techniques in pollution monitoring,” Anal. Chim. Acta, 1993, 227, 357-367.

2-16. B. K. Lavine, L. Morel, R. K. Vander Meer, R. W. Gunderson, K. H. Han, A. Bonanno, and A. Stine, “Pattern recognition studies in chemical communication: nestmate recognition in camponotus floridanus,” Chem. Intell. Lab. Syst., 1990, 9, 107-114.

2-17. E. Kreyszig, “Advanced Engineering Mathematics”, 4th ed., John Wiley & Sons, NY, 1979.

2-18. R. W. Gunderson and D. L. Denq, A Manual for Using the Program FCVPC, Research Report, Department of Mathematics, Utah State University, 1988.

2-19. P. C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, “Computerized learning machines applied to chemical problems: molecular structural parameters from low resolution mass spectrometry,” Anal. Chem., 1970, 42, 1387-1394.

2-20. J. A. Pino, J. E. McMurry, P. C. Jurs, B. K. Lavine, and A. M. Harper, “Application of pyrolysis/gas chromatography/pattern recognition to the detection of cystic fibrosis heterozygotes,” Anal. Chem., 1985, 57, 295-302.

2-21. A. B. Smith, A. M. Belcher, G. Epple, P. C. Jurs, and B. K. Lavine, “Computerized pattern recognition: a new technique for the analysis of chemical communication,” Science, 1985, 228, 175-177.

2-22. B. K. Lavine and D. Carlson, “European bee or africanized bee? Species identification through chemical analysis,” Anal. Chem., 1987, 59, 468A-470A.

2-23. P. Geladi and H. Grahn, “Multivariate Image Analysis”, John Wiley and Sons, New York, 1996.

2-24. W.H.A. van den Broek, D. Wienke, W. J. Melssen, R. Feldhoff, T. Huth-Fehre, T. Kantimm, L.M.C. Buydens, “Application of a spectroscopic infrared focal plane array sensor for on-line identification of plastic waste,” Appl. Spectrosc., 1997, 51, 856-865.

2-25. P. Robert, D. Bertrand, M.F. Devaux, and A. Sire, “Identification of chemical constituents by multivariate near-infrared spectral imaging,” Anal. Chem., 1992, 64, 664-667.

2-26. D. Coomans and D. I. Broeckaert, “Potential Pattern Recognition in Chemical and Medical Decision-Making”, Research Studies Press LTD., Letchworth, England, 1986.

2-27. B. K. Lavine, “Pattern recognition,” Critical Reviews in Analytical Chemistry, 2006, 36, 153-161.

2-28. F. W. Pjipers, “Failures and successes with pattern recognition for solving problems in analytical chemistry,” Analyst, 1984, 109, 299-303.

2-29. J. T. Tou and R. C. Gonzalez, “Pattern Recognition Principles”, Addison-Wesley Publishing, Reading, MA, 1974.

2-30. M. Sjostrom and B. R. Kowalski, “A comparison of five pattern recognition methods based on the classification results from six real data bases,” Anal. Chim. Acta, 1979, 112, 11-30.

2-31. S. Wold, “Pattern Recognition: Finding and Using Regularities in Multivariate Data,” in H. Martens and H. Russwurm, Editors, Food Research and Data Analysis, Applied Science, Essex, England, 1983.

2-32. B. R. Kowalski and S. Wold, “Pattern Recognition in Chemistry,” in Classification, Pattern Recognition and Reduction of Dimensionality, P. R. Krishnaiah and L. N. Kanal, Eds., North Holland, Amsterdam, 1982.

2-33. B. Soderstrom, S. Wold, and G. Blomquist, “Pyrolysis gas chromatography combined with SIMCA pattern recognition for classification of fruit-bodies of some ectomycorrhizal suillus species,” J. Gen. Microbiol., 1982, 128, 1783-1794.

2-34. P. D. Wasserman, “Neural Computing”, Van Nostrand Reinhold, New York, 1989.

2-35. J. Zupan, and J. Gasteiger, “Neural Networks for Chemists”, VCH Publishers, New York, 1993.

2-36. J. Shawe-Taylor and N. Cristianini, “An Introduction to Support Vector Machines”, Cambridge University Press, Cambridge, UK, 2000.

2-37. B. K. Lavine, D. R. Henry, and P. C. Jurs, "Chance classifications by nonparametric linear discriminants," J. Chemom., 1988, 2, 1-10.

2-38. M. James, “Classification Algorithms”, John Wiley & Sons, NY 1985.

2-39. C. J. Huberty, “Applied Discriminant Analysis”, John Wiley & Sons, NY, 1994.

2-40. B. S. Everitt and G. Dunn, “Applied Multivariate Data Analysis”, 2nd Edition, John Wiley & Sons, NY, 2001.

2-41. I. E. Frank and S. Lanteri, “Classification models: discriminant analysis, SIMCA, CART,” Chem. Intell. Lab. Syst., 1989, 5, 247-256.

2-42. J. H. Friedman, “Regularized discriminant analysis,” J. Amer. Statist. Assoc. 1989, 84, 165-175.

2-43. W. J. Dunn III, S. Wold, and D. L. Stalling, “Simple modeling by chemical analogy in pattern recognition,” in Environmental Applications of Chemometrics, J. J. Breen and P. E. Robinson (Eds.), ACS Symposium Series 292, Washington DC, 1985.

2-44. I. E. Frank, “DASCO – a new classification method,” Chem. Intell. Lab. Syst., 1988, 4(3), 215-22.

2-45. S. Wold, “Pattern recognition by means of disjoint principal component models,” Pattern Recognition, 1976, 8, 127-139.

2-46. S. Wold, “Cross validatory estimation of the number of components in factor and principal components models,” Technometrics, 1978, 20, 397-406.

2-47. S. Wold and M. Sjostrom, “SIMCA, a method for analyzing chemical data in terms of similarity and analogy,” in Chemometrics, Theory and Application (B.R. Kowalski, Ed.), American Chemical Society, Symposium Series 52, Washington DC, 1977.

2-48. S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjostrom “Multivariate Data Analysis in Chemistry,” in Chemometrics, Mathematics and Statistics in Chemistry, B. R. Kowalski (Ed.), D. Reidel Publishing Company, 1984.

2-49. G. Blomquist, E. Johansson, B. Soderstrom, and S. Wold, “Data analysis of pyrolysis chromatograms by means of SIMCA pattern recognition,” J. Analyt. Appl. Pyrolysis, 1979, 1, 53-65.

2-50. I. E. Frank and J. H. Workman, “Classification: oldtimers and newcomers,” J. Chemom., 1989, 3, 463-476.

2-51. G. Cybenko, “Mathematics of control,” Signals and Systems, 1989, 2, 303-310.

2-52. D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning Internal Representations by Error Propagations,” in Parallel Distributed Processing, Volume 1, MIT Press, Cambridge, MA, 1986.

2-53. R. Watrous, “Learning Algorithms for Connectionist Networks: Applied Gradient Methods of Nonlinear Optimization (Tech Report MS-CIS-87-51)”, University of Pennsylvania, Philadelphia, PA, 1987.

2-54. J.R.M. Smits, P. Schoenmakers, A. Stehmann, F. Sijstermans, and G. Kateman, "Interpretation of infrared spectra with modular neural-network systems", Chemom. Intell. Lab. Syst., 1993, 18, 27-39.

2-55. H.J. Lohninger, "Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters", J. Chem. Inf. Comp. Science, 1993, 33, 736-744.

2-56. B.B. Hubbard, "The World According to Wavelets", 2nd Edition, A. K. Peters, Natick, MA, 1998.

2-57. J. S. Walker, "A Primer on Wavelets and their Scientific Applications", Chapman & Hall/CRC, NY 1999.

2-58. K. Kawagoe, and T. Ueda, "A similarity search method of time series data with combination of Fourier and wavelet transforms", Proceedings, Ninth International Symposium on Temporal Representation and Reasoning, IEEE, 2002, pp. 86-92.

2-59. A. Graps, "An introduction to wavelets", Comp. Sci. Eng., IEEE, 1995, 2, 50-61.

2-60. MATLAB R 2006a Wavelet Toolbox, Natick, MA

2-61. B.K. Lavine, N. Mirjankar, R. LeBouf, and A. Rossner, "Prediction of mold contamination from microbial volatile organic compound profiles using solid phase microextraction and gas chromatography/mass spectrometry," Microchem. J., 2012, 103, 37-41.

2-62. B.K. Lavine, N. Mirjankar, and R. K. Vander Meer, "Analysis of chemical signals in red fire ants by gas chromatography and pattern recognition techniques," Talanta, 2011, 83, 1308-1316.

2-63. R.K. Vander Meer and L. Morel, "Nestmate recognition in ants," in R.K. Vander Meer, M. Breed, M. Winston, and K.E. Espelie (editors), Pheromone Communication in Social Insects, Westview Press, Boulder, CO, 1998.

2-64. M.A. Stapanian, F.C. Garner, K.E. Fitzgerald, G.T. Flatman, and J.M. Nocerino, "Finding suspected causes of measurement error in multivariate environmental data", J. Chem., 1993, 7, 165-176.

2-65. B.K. Lavine, A. Faruque, P. Kroman, and H. T. Mayfield, "Source identification of fuel spills by pattern recognition analysis of high speed gas chromatograms", Anal Chem., 1995, 67, 3846-3852.

2-66. L. Morel R.K. Vander Meer, and B.K. Lavine, "Ontogeny of nestmate recognition cues in the red carpenter ant (*Camponotus floridanus*)--behavioral and chemical evidence for the role of age and social experience", Behav. Ecol. Sociobiol., 1988, 22, 175-183.

2-67. B.K. Lavine., L. Morel, R.K. Vander Meer, R.W. Gunderson, J.H. Han, A. Bonanno, and A., Stine, "Pattern recognition studies in chemical communication: Nestmate recognition in *Camponotus floridanus*", Chem. Intell. Lab. Syst., 1990, 9, 107-114.

2-68. B.K. Lavine, R.K. Vander Meer, L. Morel, R. W. Gunderson, J.H. Han, and A. Stine, "False color data imaging: A new pattern recognition technique for analyzing chromatographic profile data", Microchem. J., 1990, 41, 288-295.

2-69. B.K. Lavine, C. Davidson, R.K. Vander Meer, S. Lahav, V. Soroker, and A. Hefetz, "Genetic algorithms for deciphering the complex chemosensory code of social insects. Chem. Intell. Lab. Syst.", J. Chem., 2003, 66, 51-62.

2-70. C.A. Johnson, H. Topoff, R.K. Vander Meer, and B. Lavine, "Host queen killing by slave-maker ant queen: When is a host queen worth attacking?", Anim. Behav., 2002, 64, 807-815.

2-71. C.A. Johnson, H. Topoff, R.K. Vander Meer, and B.K. Lavine, "Do these eggs smell funny to you? Egg discrimination by *Formica* hosts of the slave-making ant, *Polyergus breviceps* (Hymenoptera: Formicidae)", Behav. Ecol. Sociobiol., 2005, 57, 245-255.

2-72. C.A. Johnson, R.K. Vander Meer and B. Lavine, "Changes in the cuticular hydrocarbon profile of the slave-maker ant queen, *Polyergus breviceps* Emery, after killing a *Formica* host queen (Hymenoptera: Formicidae)", J. Chem. Ecol., 2001, 27, 1787-1804.

2-73. R.K. Vander Meer, D. Saliwanchik, and B. Lavine, "Temporal changes in colony cuticular hydrocarbon patterns of *Solenopsis invicta*: Implications for nestmate recognition", J. Chem. Ecol., 1989, 15, 2115-2125.

2-74. B.K. Lavine, P.C. Jurs, D.R. Henry, and R.K. Vander Meer, J.A. Pino, and J.E. McMurry, "Pattern recognition studies of complex chromatographic data sets: design and analysis of pattern recognition experiments", Chem. Intell. Lab. Syst., 1988, 3, 79-89.

2-75. P.C. Jurs, B.K. Lavine, and T.R. Stouch, "Pattern recognition studies of complex chromatographic data sets", J. Res. Nat Bur. Stand., 1985, 90, 543-549.

2-76. B.K. Lavine, P.C. Jurs, and D.R. Henry, "Chance classifications by non-parametric linear discriminant functions", J. Chem., 1988, 2, 1-10.

2-77. B.K. Lavine and D.R. Henry, "Monte Carlo studies of non-parametric linear discriminant functions", J. Chem., 1988, 2, 85-89.

CHAPTER III

GENETIC ALGORITHMS FOR PATTERN RECOGNITION AND FEATURE SELECTION

A genetic algorithm (GA) for classification of multivariate chemical data has been developed as part of the research described in this thesis. The pattern recognition GA identifies features that convey information about differences between sample classes in a plot of the two or three largest principal components of the data. The GA is also equipped to perform advanced tasks such as outlier detection, identification of training set samples that do not have the proper class label, and elucidation of clustering trends and data structures in large multivariate chemical data sets. The design of the pattern recognition GA incorporates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection and classification of chemical data. The general concept and methodology of genetic algorithms are discussed in this chapter followed by a detailed discussion of the pattern recognition GA used in the studies described in this thesis for feature selection.

3.1 GENETIC ALGORITHMS

Genetic algorithms were pioneered by John Holland [3-1]. They are adaptive heuristic search algorithms that simulate the process of evolution to solve optimization problems. Genetic algorithms use a population of strings to encode potential solutions for an optimization problem. A search of the solution space is conducted by exploiting knowledge contained in the population while simultaneously utilizing randomized operators to generate new and potentially better solutions. Each string or point is tested individually and desirable features from existing points are combined to form a new population of strings that often yield better solutions to the problem. A GA only requires knowledge about the quality of the solution generated by each string or parameter set.

Genetic algorithms have several advantages over conventional search algorithms. They work with the entire parameter set whereas conventional optimization techniques manipulate the parameters independently. This can pose a problem if an object function becomes overly sensitive to one parameter as the optimization function will tend to focus its effort on the troublesome parameter at the expense of the other parameters.

Genetic algorithms consider simultaneously many points in the search space. More of the response surface is probed reducing the change of convergence to a local minimum since genetic algorithms utilize parallelism as a large number of candidate solutions are searched simultaneously.

Genetic algorithms make no assumption about the geometry of the response surface. They do not require any knowledge of the search space beyond the fitness of individual solutions Discontinuities or singularities in the response surface which prevent the use of calculus (derivative) or simplex based methods will not cause a problem for the

GA. The computational environment offered by a GA can be readily adjusted to match a particular application allowing genetic algorithms to be tailored for individual problems.

When applying genetic algorithms, there are two decisions that must be made: (1) How to code the parameters as chromosomes (potential solutions), and (2) How to evaluate the fitness of each chromosome? Implementation of a genetic algorithm requires a population of candidate solutions and heuristics to manipulate them. The optimization procedure used by a GA consists of five interrelated steps.

- An initial population of strings (candidate solutions) is generated with each string representing a potential solution to the problem.
- The strings are decoded yielding the actual parameter set sent to the fitness function for evaluation. (Each string is assigned a value by the fitness function that is a measure of the quality of the proposed solution)
- The fitness is used to select strings for the reproduction operator, which produces a new population of strings through recombination and mutation
- The new population of solutions, which often yields better results for the problem, is evaluated by the fitness function.
- This simple procedure is repeated until an optimal solution is found or a specified number of generations have been achieved

During each generation, each solution in the population is evaluated by a fitness function. Solutions with a high fitness value have a high probability of being selected for crossover (the operator often used for reproduction) and mutation. The power of the GA arises from crossover, which causes a structured yet randomized exchange of information between solutions with the possibility that good solutions can lead to better ones.

Crossover enables the GA to investigate large regions of the solution space making the GA more robust to local optima.

Figure 3.1 provides an outline of the main steps involved in the operation of a simple genetic algorithm. The procedure begins with the generation of an initial population of strings. A genetic algorithm uses binary strings of uniform length known as chromosomes where each binary bit is analogous to a gene. The identity of a gene depends on its location in the chromosome. A bit-value of 0 or 1 stands for the absence or presence of a particular feature in the solution with 0 indicating its absence and 1 indicating the presence of the feature. Bit-values for all chromosomes in the initial population are randomly chosen.

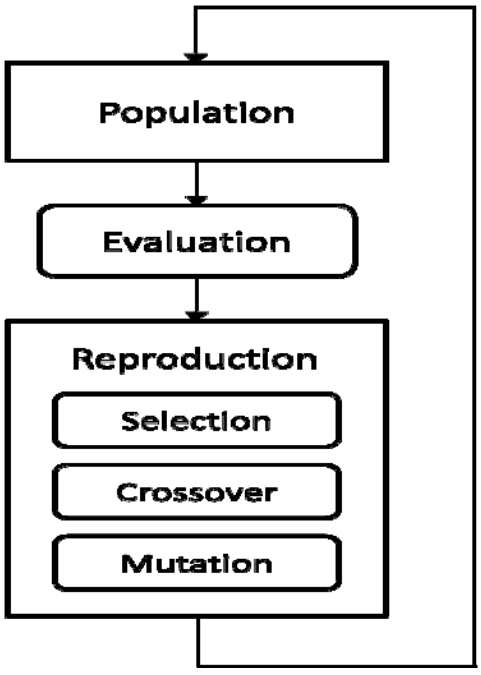

Figure 3.1. Processes involved in the operation of a simple genetic algorithm.

Each chromosome in the population is scored. The objective function used to score the chromosomes is called the fitness function. After evaluation, reproduction operators are used to generate a new population of chromosomes. The process of reproduction is implemented using three operators: selection, cross-over and mutation. During selection, chromosomes are chosen for reproduction with the selection probability proportional to their fitness. Chromosomes with higher fitness scores are more likely to be selected. Crossover involves a structured yet randomized recombination of the selected chromosomes (see Figure 3.2). Two parent chromosomes selected for reproduction undergo an exchange of binary bits at a randomly selected crossover point to yield two new off-spring chromosomes. Mutation is the random alteration of a single bit in a string and occurs with a predefined probability which is usually very low. Mutation is used to explore regions of the solution space above and below those that are being probed. The new generation of chromosomes replaces the older chromosomes in the population. The process of evaluation and reproduction of is repeated until a feasible solution is found or a certain number of generations have been exceeded.

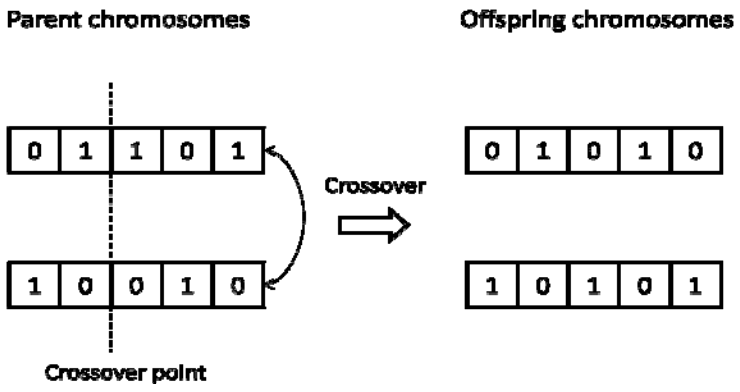

Figure 3.2. An example of one-point crossover

The selection criteria for reproduction exhibits bias towards higher ranking chromosomes. However, crossover and mutation operators ensure a significant degree of diversity in the solution population. As the population evolves by sampling more often from regions of the search space with higher average fitness, the average fitness of the population is expected to improve over successive generations. Although the GA has access only to the chromosomes of the current population and their fitness, the information extracted from the potential solutions is about the relationship between the chromosomes and their fitness. Highly fit chromosomes after they are identified often guide the search. This vast wealth of information explored by the GA is quantified by a framework called schema or similarity template.

A schema also referred to as a similarity template [3-2] represents a set of chromosomes. For example, the schema {1 * * * * * * 0} will match all chromosomes with eight bits that start with 1 and end with 0 with either 0 or 1 in positions 2 thru 7. Genetic algorithms use schema implicitly. For a genetic algorithm operating on a population of chromosomes of fixed length l, there are 3^l unique schema or patterns. Each chromosome is a member of 2^l of them. For example, {0.1} is a member of the following schema: {01}, {*1}, {0*}, and {**}. When the fitness function of a genetic algorithm is evaluating a chromosome, it is also evaluating many schemas. In a population of identical chromosomes, there are 2^l schema present, and in a population of n unique chromosomes, there can be as many as $n2^l$ schema represented. Evaluating different chromosomes which are members of the same schema can be thought of as estimating the average value of that pattern. Even though these averages are not

explicitly calculated, the survival of the pattern and the number of representative chromosomes can be expressed in terms of these averages.

Let S be a schema present in the population at generation g. Its multiplicity, $m(S, g)$ is defined as the number of instances of *S* in the population at generation *g*. The expected number of chromosomes which represents *S* in the next generation is given by Equation 3.1 (the schema theorem):

$$m(S, g+1) \;\geq\; m(S, g)\,\frac{f(S)}{\bar{f}}\left(1 - p_c\,\frac{d(S)}{l-1}\right)(1 - p_m)^{o(S)} \qquad (3.1)$$

where $m(S, g+1)$ is the expected number of chromosomes representing schema *S* in the next generation (g + *1*) based on the number of chromosomes representing schema *S* in the current generation, $m(S, g)$. The ratio of the fitness of the chromosome representing schema *S* to the average fitness of the population is given by $\frac{f(S)}{\bar{f}}$, and collectively $m(S, g)\frac{f(S)}{\bar{f}}$ is the likelihood that schema S is represented in the population. Due to selection pressure alone, schema will grow or decay depending on their fitness [3-3]. However, chromosomes selected for reproduction will undergo crossover and mutation. These operators can disrupt the schema S such that S is not present in the next generation. The second factor in Equation 3.1, $\left(1 - p_c\,\frac{d(S)}{l-1}\right)$ accounts for the probability that *S* survives crossover; p_c is the probability that a chromosome undergoes crossover. The last factor in Equation 3.1, $(1 - p_m)^{o(S)}$, accounts for the probability that S survives the

mutation operator. Here, p_m is the probability that a given bit is flipped, and $o(S)$ is the order of *S* or the number of non-* bits in *S*.

The consequence of a genetic algorithms use of schema is an implicit parallelism [3-4]. At each evaluation, the genetic algorithm is aware of a particular point in the fitness landscape because of the chromosomes it is evaluating. According to the schema theorem, the genetic algorithm makes observations about areas of the search space based on the schema, which allows the genetic algorithm to focus its attention on "hot spots" or areas likely to have a high fitness in the solution space similar to that of a gradient descent search. (In a gradient descent search, the value at a random position is calculated. The points around it are also inspected to calculate the direction and magnitude of greatest local descent. A new point in that direction is sampled and the process is repeated until the minimum is reached.) An obvious difference between these two methods (genetic algorithm versus gradient descent) is the number of points sampled per iteration. Even when the gradient descent method is modified to sample multiple points per iteration, the next point or set of points is near the last in the solution space, whereas genetic operators such as crossover and mutation produce points which are near or are distant from the parents in the solution space depending on the bits that are exchanged or flipped in the chromosome. Consequently, a genetic algorithm is less likely to get stuck in a local minimum in the solution space.

Genetic algorithms are probabilistic, neither random nor deterministic. This is demonstrated in the selection process where a chromosome's chances of being selected are weighted against its fitness. It is preferable that offspring are not produced in the same way each time. This is addressed by assigning a probability to each reproductive

function. When two chromosomes are selected for reproduction, a mechanism is chosen according to its probability (P). The user can assign p_m equal to 0.01 and p_c equal to 0.5. This mixing of reproductive operations preserves a certain amount of variation in the population

3.2 GENETIC ALGORITHM FOR FEATURE SELECTION AND PATTERN RECOGNITION

The genetic algorithm for feature selection and pattern recognition uses principal component plots to characterize the information contained in feature subsets of the data. For a classification problem, the amount of information in a set of features about class differences is directly proportional to the magnitude of the class separation achieved in a plot of the two or three largest principal components of the feature subset. An improvement in the separation of the classes in a principal component plot corresponds to an increase in the amount of information about class differences captured by a specific feature subset. PCA is incorporated into the fitness function of the pattern recognition GA to provide an information filter which reduces the size of the search space by limiting the search to feature subsets that enhance the separation between classes in the data set.

The approach to feature selection for classification described in this chapter is based on a very simple idea - identify a set of measurement variables that optimize the separation of the classes in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these features is about differences between classes in the data set. This idea is demonstrated in Figure 3.3, which shows a plot of the two largest principal components of a data set prior to feature selection. The data set consists of 30 samples distributed

between 3 classes (good, better, and best). Each sample is characterized by 10 measurements. However, only 4 of these measurements contain information about the classification problem. When a principal component map of the data is developed using only these 4 measurements, sample clustering on the basis of class is evident.

Using this approach to feature selection, an eigenvector projection of the data is developed that discriminates classes in the data set by maximizing the ratio of between- to within-group variance. This approach to feature selection has a number of advantages. It avoids overly complicated solutions, which do not perform as well on the prediction set because of over-fitting. Although a principal component plot is not a sharp knife for discrimination, if we have a principal component plot that shows clustering, then our experience is that we will be able to predict robustly using this set of descriptors. Furthermore, the principal component plot displays the variability between large numbers of samples and show the major clustering trends present in the data; the user can visually identify the presence of confounding relationships in the data, thereby gaining insight into how the decision is made for a classification.

A block diagram that illustrates the general operation of the pattern recognition GA for feature selection is shown in Figure 3.4. The fitness functions used, the reproduction operators and the mechanism to adjust internal parameters of the GA for guiding the search in the right direction via adjustment of the fitness function are important aspects of the pattern recognition GA that make it unique. The operators unique to the pattern recognition GA are described below

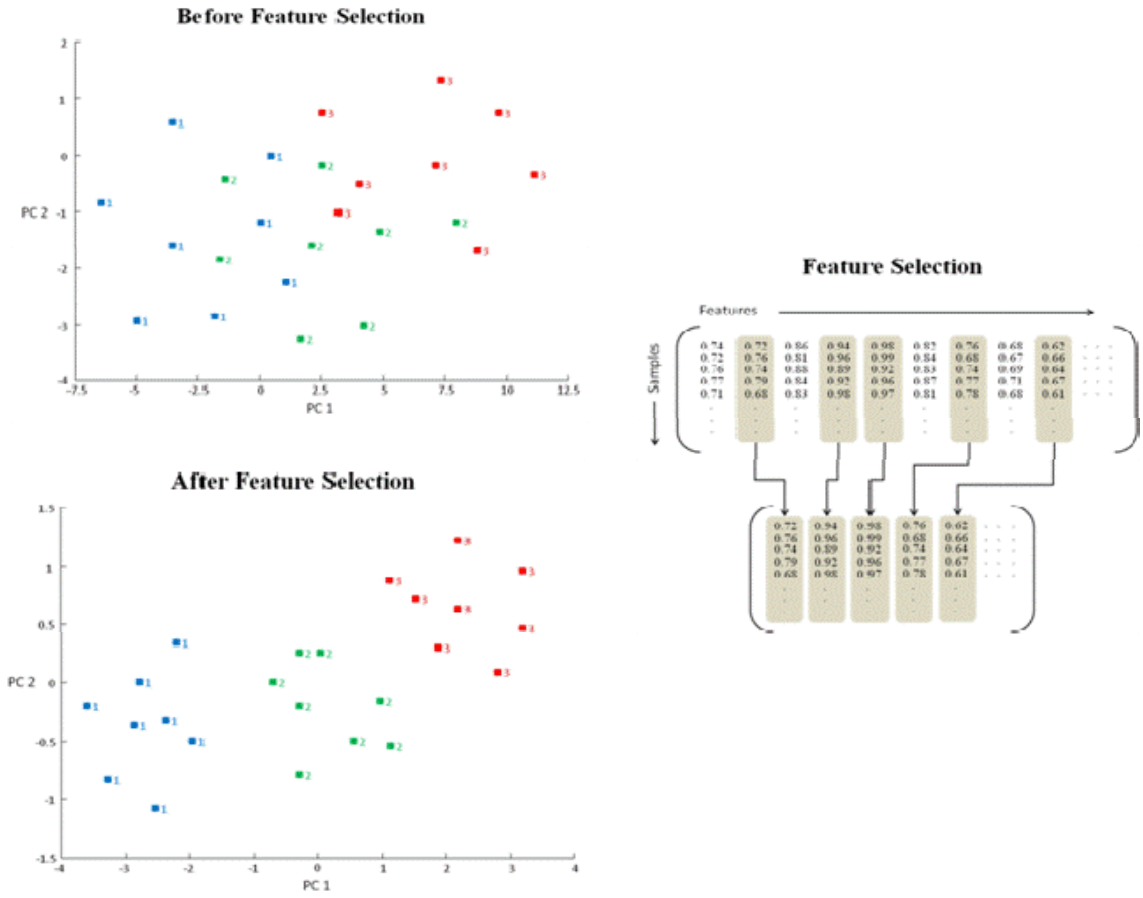

Figure 3.3. A plot of the two largest principal components of 10 features in the data set does not show class separation. When principal components are developed from features that contain information about class, clustering on the basis of the sample's class label (1= low, 2 = medium, and 3 = high) is evident.

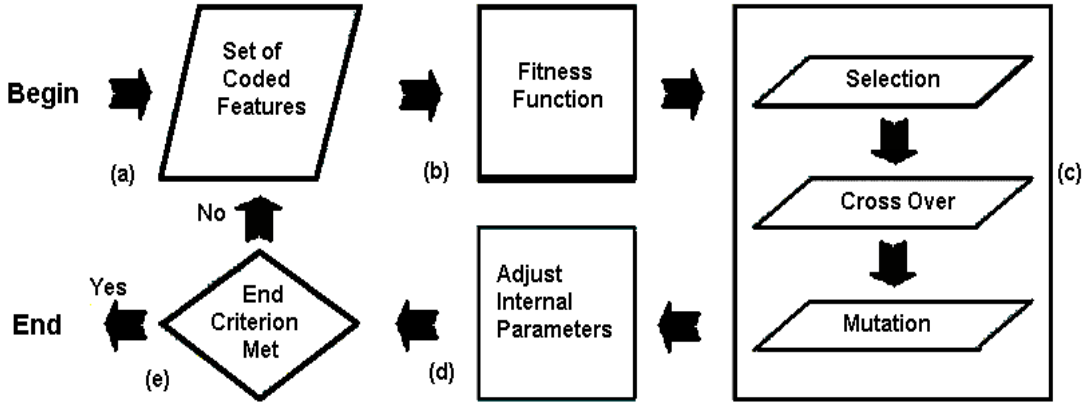

Figure 3.4. Block diagram of the pattern recognition GA used for feature selection.

3.2.1 Fitness Function. To evaluate and compare different chromosomes, an object function that quantifies the fitness of individual chromosomes needs to be formulated. One fitness function that is used by the pattern recognition GA for feature selection is PCKaNN [3-5 – 3-11]. PCKaNN utilizes both PCA and K-NN to score each feature subset in the population. For each chromosome in a population, PCA is used to plot the corresponding data using the two or three largest principal components of the data. The degree of class separation in the principal component plot of the data is assessed using the K-NN classification method. Class and sample weights which are an integral component of PCKaNN are computed using Equation 3.2 and Equation 3.3, where $CW(c)$ is the weight of class c, and $SW(s)$ is the weight of sample s in class c.

To evaluate and compare different chromosomes an objective function that quantifies the fitness of individual chromosomes needs to be formulated. The fitness function used by the pattern recognition GA for feature selection is PCKaNN [3-5 to 3-11]. PCKaNN is a combination of principal component analysis (PCA) and K-nearest neighbor (K-NN) method. For each chromosome of a given population, PCA is first used for extracting the information content of the feature subset and mapping it in a plot of its two or three largest principal components. Then the K-NN method is used for quantitatively characterizing the amount of class separation achieved in the PC-plot. Class and sample weights are computed as shown by Equations (3.1) and (3.2) respectively, where $CW(c)$ is the weight of class c, and $SW(s)$ is the weight of sample s in class c.

$$CW(c) = 100\frac{CW(c)}{\sum_{i=1}^{c} CW(c)} \tag{3.2}$$

$$SW_c(s) = CW(c)\frac{SW_c(s)}{\sum_{s\in C} SW_c(s)} \tag{3.3}$$

The sum of the sample weights for objects assigned to a particular class is equal to the class weight, and the sum of all class weights in the data set is equal to 100. For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest. A poll is then taken of the point's k_c-nearest neighbors. (K_c is set by the user, and for the most rigorous classification, K_c equals the number of samples in the class to which the point belongs. The number of K_c-nearest neighbors with the same class label as the sample point in question, called the sample hit count (SHC), is computed ($0 \leq SHC(s) \leq K_c$). It is then a simple matter to score a principal component plot (see Equation 4).

$$F = \sum_{c}\sum_{s\in c}\frac{SHC(s)}{K_c}SW(s) \tag{3.4}$$

To better understand the scoring of the principal component plots, consider a data set with two classes, which have been assigned equal weights. Class 1 has 10 samples, and class 2 has 20 samples. At generation 0, the samples in a given class will have the same weight. Thus, each sample in class 1 has a sample weight of 5, whereas each sample in class 2 has a weight of 2.5. Suppose a sample from class 1 has as its nearest

neighbors 7 class one samples. Hence, SHC/K = 0.7, and (SHC/K)*SW = 0.7*5, which equals 3.5. By summing (SHC/K_c)*SW for each sample, each principal component plot can be scored.

3.2.2 Reproduction. The process of creating a population with a higher average fitness score is called reproduction. The steps involved in reproduction are: (1) *Selection*, (2) *Crossover*, and (3) *Mutation*. Selection of chromosomes for reproduction should take into account their fitness while ensuring the existence of sufficient diversity in the population. Selection in the pattern recognition GA is implemented by ordering the population of strings, i.e. potential solutions, from best to worst, while simultaneously generating a copy of the same population and randomizing the order of the strings in this copy with respect to their fitness (see Figure 3.5). A fraction of the population is then selected as per the selection pressure, which is set at 0.5. The top half of the ordered population undergoes reproduction with strings from the top half of the random population, guaranteeing the best 50% are selected for reproduction, while ensuring that every string in the randomized copy has an equal probability of being selected. If a purely biased selection criterion were used to select strings, only a small region of the search space would be explored. Within a few generations, the population would consist of only copies of the best strings in the initial population. By using this two step approach for reproduction, a randomized selection criterion as well as one based on fitness is imposed on the strings (i.e., potential solutions) from the population ensuring that sufficient genetic diversity is always maintained. Maintaining genetic diversity is necessary to allow the population to evolve to better solutions by widening the search

space and preventing premature convergence. This has been a problem with some of the more popular reproduction operators, such as the roulette wheel and tournament selection.

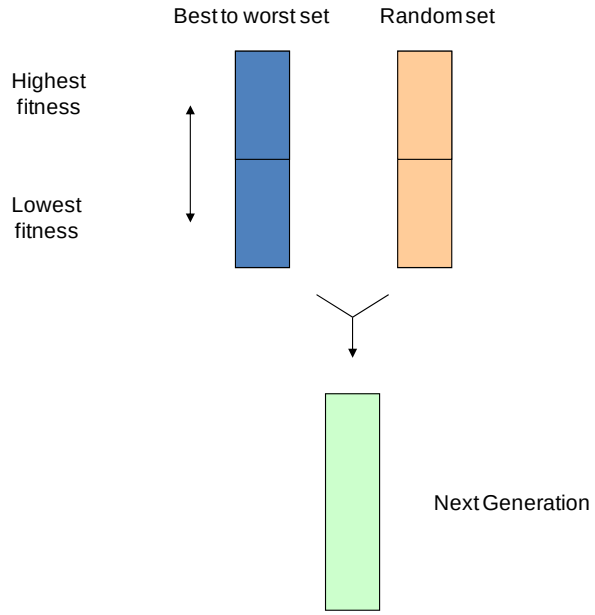

Figure 3.5. The top half of the ordered population is mated with strings from the top half of the random population, guaranteeing the best 50% are selected for reproduction, while every string in the randomized copy has an equal chance of being selected.

For each pair of strings selected for mating, two new strings are generated using three-point crossover. A mutation operator is then applied to the new strings. The mutation probability of the operator is usually set at 0.01, so 1% of the feature subsets are selected at random for mutation. A chromosome marked for mutation has a single random bit flipped, which allows the GA to explore other regions of the parameter space. The resulting population of strings, both the parents and children, are sorted by fitness, with the top ϕ strings retained for the next generation. Because the selection criterion used for reproduction exhibits bias towards the higher-ranking strings, the new population is expected to perform better on average than its predecessor. The

reproductive operators used, however, also assure a significant degree of diversity in the population, since the crossover points of each chromosome pair is selected at random.

3.2.3 Boosting (Adjusting Internal Parameters). Using the information gained from the solutions (feature subsets) of the previous generation, the genetic algorithm can focus on those classes and samples that are more difficult to classify by adaptively changing their weights. This process is called boosting [3-12 to 3-18]. (Boosting the weights is referred to as adjusting the internal parameters in the block diagram of the pattern recognition GA shown in Figure 3.4.) The first step in boosting is to compute the sample and class hit-rates using Equations 3.5 and 3.6 to characterize the degree of difficulty of classifying a particular sample or class using the entire population of feature subsets. The sample hit-rate, *SHR(s)*, is the mean value of the mean value of SHC/K_c for sample *s* over all the feature subsets of the population, and the class hit-rate, *CHR(c)*, is the mean sample hit-rate of all samples in a class. ϕ in Equation 3.5 is the number of chromosomes in the population, and AVG in Equation 3.6 refers to the average or mean value for the sample hit rate.

$$SHR(s) = \frac{1}{\varphi} \sum_{i=1}^{\varphi} \frac{SHC_i(s)}{K} \tag{3.5}$$

$$CHR_g(c) = avg\big(SHR_g(s): \forall_{s \in c} \big) \tag{3.6}$$

Sample and class hit-rates from the previous generation are used to update the sample and class weights for the next generation using Equations 3.7 and 3.8, where *g* is the previous generation, *g+1* is the current generation, and *P* is the momentum or learning rate (which is set by the user). Sample and class weights after adjustment using a perceptron algorithm are renormalized using Equations 3.2 and 3.3.

$$CW_{g+1}(c) = CW_g(c) + P\left(1 - CHR_g(c)\right) \tag{3.7}$$

$$SW_{g+1}(s) = SW_g(s) + P\left(1 - SHR_g(s)\right) \tag{3.8}$$

Classes and samples that are difficult to classify will have lower hit-rates. They will be more heavily weighted in the next generation. Their higher impact on the fitness function provides a driving force for the pattern recognition GA to search for feature subsets that can correctly classify difficult samples and/or classes. Unlike support vector machines or back propagation neural networks, all samples and classes contribute to the overall fitness score.

Boosting is performed in two stages. In the initial stage, the learning rate *P* is set at 0.5 to facilitate learning of the optimal class weights. Once the class weights become stable, i.e., the change in the class weights falls below some threshold tolerance, the class weights are fixed and Equation 3.7 is turned off for the remaining generations. During the second stage of boosting, *P* is set to 0.25. These values for P have been chosen in part because they facilitate learning by the genetic algorithm but do not cause a particular

sample or class to dominate the calculation, which would result in the other samples or classes not contributing to the scoring by the fitness function.

Boosting is crucial for the successful operation of the pattern recognition GA because it modifies the fitness landscape by adjusting values of both class and sample weights. This helps to minimize the problem of convergence to a local optimum. Hence, the fitness function of the pattern recognition GA changes as the population is evolving towards a solution using information from the population to guide these changes. The cycle of evaluation, reproduction, and boosting of potential solutions is repeated until a feasible solution is found or a specified number of generations are attained.

There are a number of parameters that affect the performance of the pattern recognition GA including the choice of the crossover and mutation rate and the configuration of the initial population. The experience of our research group using the pattern recognition GA has shown that 3-point crossover works. However, the number of features in each feature subset of the initial population should also be treated as an important parameter. If the feature sets are initially sparse, the probability of including features, which are neither good nor bad, is low since the principal component based fitness function does not provide additional points for adding them. Conversely, the probability of removing these features from less sparse feature subsets is also low since there is no advantage in deleting them. For data sets with a large number of good features, it is probably best not to employ sparse feature subsets in the initial population. Otherwise, it may take thousands of generations to ensure the inclusion of all good features in the solution.

To ensure removal of features, which are neither good nor bad, the corresponding loading plot that is generated with each principal component plot can be examined by the pattern recognition GA. If the loadings for a particular feature are near zero with both principal components, the descriptor is a likely candidate for removal since its contribution to the principal component score plot is negligible. Even good feature subsets often contain some features that are uninformative. The information derived from such a feature subset is equivalent to the subset with the uninformative features removed. As there is no explicit benefit gained by removing these features due to the attributes of PCA, they tend to remain lodged in the feature subset and could be misinterpreted as important features.

As part of the pattern recognition GA, a culling routine has been developed to remove these irrelevant features by careful examination of the corresponding principal component loading plots. Features with loadings near zero will have negligible contributions towards the computation of the principal components and are removed if they lie within a user specified radius of the origin in a plot of the loadings. Other parameters that can be controlled by the user for the implementation of culling are the *culling frequency* and the *culling pressure*. The culling frequency is the number of generations after which the culling is repeated. The fraction of the population to which culling is applied is indicated by the culling pressure. Culling is often implemented every 25 generations to check for features that are neither good nor bad. During the generation when culling is implemented, crossover is not performed on the strings.

Varying the composition of the initial population or the mutation rate can prove beneficial in optimizing a solution but this fact should not be viewed negatively as

suggested by some workers since it allows the user to vary the search of the solution space ensuring a more careful analysis of the data. Given the small number of iterations required for a solution (usually less than 100), the advantages of using these two GA parameters as search variables outweighs any disadvantage that might be incurred due to increased complexity.

A drawback of a genetic algorithm is that one cannot control the rate of convergence, but convergence is not what we are seeking. A genetic algorithm can evade local optima, but this does not mean that convergence necessitates an optimal solution. Convergence as a benchmark for the success of a GA would suggest that any genetic algorithm provides a deficient solution. However, the quality of the best solution found – and how quickly and reproducibly it is found – is the guide being used to determine the success of this method. The ease, speed, and reproducibility of our pattern recognition GA have been demonstrated on a variety of data sets. The success of the pattern recognition GA can be attributed to the large number of optimum solutions that exist in the data as a result of the high degree of collinearity between measurement variables in the data set.

3.3 MODIFICATIONS TO PCKaNN

By incorporating various modifications to PCKaNN, the fitness function of the pattern recognition GA has been generalized to tackle a variety of pattern recognition problems that are difficult to solve. Three different modifications have been made to PCKaNN: (1) combination of the Hopkins statistic with PCKaNN to perform transverse learning, (2) combination of a modified Hopkins statistic with PCKaNN to implement

transverse learning, and (3) canonical variate analysis (CVA) in lieu of PCA in PCKaNN. Transverse learning has been incorporated into the pattern recognition GA by coupling PCKaNN with the Hopkins statistic and the modified Hopkins statistic. The Hopkins statistic and the modified Hopkins statistic search for features that increase the clustering of the data whereas PCKaNN identifies feature subsets that create class separation. We will be able to explore the structure of a data set, for example, discover new classes, by simply tuning the relative contribution of the Hopkins statistic and the original fitness function to the overall fitness score. For training sets with small amounts of labeled data and large amounts of unlabeled data, this approach will perform better than a learning model developed from a set of features using only the labeled data points since information in the unlabeled data is used by the fitness function to guide feature selection which will prevent overfitting. This approach to feature selection is semi-supervised learning as it incorporates aspects of both supervised and unsupervised learning to develop a new paradigm for multivariate data analysis where classification, clustering, feature selection, and prediction can be combined in a single step enabling a more careful analysis of the data.

CVA [3-19] is similar to PCA except that a map of each feature subset is being developed with the focus on optimization of "between groups variability" not total variability. If information about class differences lies in the directions of maximum variance, then PCA and CVA produce similar scatter plots. That is to say, PCA is usually all that is needed when the variability in the data is small, except for that induced by class differences. When information about class differences cannot be related to total

variability, then techniques such as CVA must be be used in lieu of PCA to assess the information content of each feature subset.

3.3.1 Hopkins Statistic for Transverse Learning. The Hopkins statistic does not make any assumptions about the data to assess clustering. As it is a fast and simple method to use, it can be easily integrated with PCKaNN. The Hopkins statistic is defined as

$$H = \frac{\sum u}{\sum u + \sum w} \tag{3.9}$$

where U are the distances between randomly selected locations and their nearest neighbors in the PC plot and W are the distances between randomly selected data points and their nearest neighbors in the same PC plot. Usually 10% of the samples in the data are chosen. The same number of random locations in the projected data space of the principal component plot are chosen, and the distances (u) from these points to their nearest neighboring samples are also determined. The value of the Hopkins statistic, H is then computed as a ratio of the sum of the distances as shown by Equation 3.9. This process is repeated several times, and H is averaged to get a more accurate estimate of the clustering tendency.

The Hopkins statistic probes the volume occupied by samples (using w) in comparison to the volume of the projected data space (using u) to assess the degree of clustering. Uniformly distributed random data will not cluster and will have similar u and w distances with an H value of 0.5. If the data contain closely packed clusters, the volume occupied by the samples is only a small fraction of the entire data space. For this

reason, an increase in the clustering tendency is reflected by a decrease in *w* relative to *u* with the value of *H* approaching 1 well clustered data.

If a data set has more features than samples, then even multivariate normally distributed data that does not contain outliers will have variables that produce principal component plots containing points that appear as outliers. The Hopkins statistic will generate higher scores for these principal component plots as outliers tend to increase the value of *H*. This prevents the GA from searching for more meaningful feature subsets. For this reason, it was necessary to robustify the Hopkins statistic. An influence function [3-20] that can detect the presence of outliers in principal component plots by computing the leverage of each sample on the largest principal components. This information can then be used to de-weight the Hopkins statistic for outliers (i.e., samples with high leverage). The deweighted Hopkins statistic $\widetilde{H}$ is given by Equation 3.10, where H is the scaled Hopkins statistic and $max\left(Influence_j\right)$ is the influence value of the sample with the highest leverage on the j^{th} principal component where j = 1, 2,3.

$$\widetilde{H} = H - H\sum_{j=1}^{PC} max\left(Influence_j\right) \tag{3.10}$$

The influence function for the de-weighted Hopkins statistic is provided by Equation 3.11, where t_{ij} is the influence of the i^{th} sample on the j^{th} principal component, y_{ij} is the score of the i^{th} sample on the j^{th} principal component, λ_j is the eigenvalue of the j^{th} principal component, and *n* is the number of samples. Influence values are normalized across all samples to sum to 1 (Equation 3.12).

$$t_{ij} = \frac{y_{ij}^2}{(n-1)\lambda_j} \tag{3.11}$$

$$\sum_{i=1}^{n} t_{ij} = 1 \tag{3.12}$$

A sigmoid transfer function was also used to scale the range of the deweighted Hopkins statistic from 0 to 1. The value of *H* for clustered data seldom approaches 1, while *H* for data with an outlier is often close to 1. Increasing the range of *H* with a sigmoid transfer function (similar to the one used in neural networks) facilitates detection of a broader range of clustering configurations.

The scaled and deweighted Hopkins statistic is more robust and can be used to search for feature subsets that exhibit sample clustering. The fitness function of the pattern recognition GA for feature selection can be modified by coupling this robust Hopkins statistic to PCKaNN. The new fitness function $\hat{F}$ is given by a weighted average of the fitness scores *F* (PCKaNN) and $\tilde{H}$ (Hopkins statistic) as shown in Equation 3.13. The value of the weighting factor r, which is between 0 and 1, determines the degree of emphasis placed on searches for features that will display clustering.

$$\hat{F} = (1-r)F + r\tilde{H} \tag{3.13}$$

Smaller values of r will emphasize class separation because greater emphasis is placed on PCKaNN, whereas sample clustering will gain precedence as the value of r is increased. The fitness function is generalized to provide the pattern recognition GA with the capability to perform two tasks in tandem: classification and clustering. By varying the contributions of PCKaNN and the deweighted Hopkins statistic, it is possible to tune the fitness function, enabling the pattern recognition GA to search for different types of structure that may be present in the data. For example subclasses within a data set could be uncovered. Another advantage of using the fitness function described in Equation 3.13 is transverse learning. Often, chemical data sets contain a large number of samples without class labels as compared to the number of samples with class labels that can be used for training. A classical learning model which only uses samples with class labels to develop a classification rule may lack sufficient information to correctly classify the samples in a larger prediction set. Transverse learning [3-21] makes use of the information available in the prediction set with the training set samples in a semi-supervised manner. PCKaNN is a supervised learning method (i.e., uses class information) whereas the deweighted Hopkins statistic is an example of unsupervised learning method as it does not require class information. Transverse learning can be implemented using the modified PCKaNN fitness function which is a combination of PCKaNN and the deweighted Hopkins statistic. Feature subsets are simultaneously evaluated for class separation using only the labeled samples and for clustering using both the labeled and unlabeled samples. Searches for feature subsets capable of classifying the labeled samples are performed, while simultaneously allowing similar samples (both labeled and unlabeled) to cluster. This semi-supervised learning

methodology will produce higher classification success rates for the prediction set than one obtained using a classical learning model by minimizing the chances of overfitting the training set data.

3.3.2 Modified Hopkins Statistic for Transverse Learning. The unlabeled samples in the training set can be included into the boosting routine of the pattern recognition GA to further minimize the probability of convergence to a local optimum. The deweighted Hopkins statistic can be modified to allow the boosting routine to track the unlabeled samples by monitoring and adjusting their sample weights during each generation. The modified deweighted Hopkins score, *MH*, of a feature subset is given by Equation 3.14.

$$MH = \sum_{j=1}^{u} \frac{1}{1 + d_{ij}} USW_j \tag{3.14}$$

where u is the number of unlabelled samples, USW_j is the weight of the j^{th} unlabeled sample, and d_{ij} is the distance between the j^{th} unlabeled sample and the labeled sample that is its nearest neighbor in the principal component plot of the i^{th} feature subset. Each unlabelled sample is initially assigned a sample weight of $100/u$. To monitor the weight of troublesome samples, an average distance vector is computed for the entire population. The average distance vector $avgD(j)$ for unlabeled sample j is given by Equation 3.15, where φ is the number of chromosomes in the population. The sample weights are adjusted for boosting using Equation 3.16.

$$avgD(j) = \frac{1}{\varphi}\sum_{i=1}^{\varphi}\frac{1}{1+d_{ij}} \tag{3.15}$$

$$USW_{g+1}(j) = USW_g(j) + P\left(1 - avgD_g(j)\right) \tag{3.16}$$

This modification to the deweighted Hopkins statistics enables the GA for pattern recognition to focus on unlabelled samples that are difficult to cluster by making full use of the boosting routine of the pattern recognition GA for transverse learning.

3.3.3 Comparison of PCKaNN Fitness Functions. A data set of 98 Raman spectra was used to evaluate the efficacy of the three PCKaNN fitness functions discussed in this section of the thesis. Wood identification is usually accomplished by forestry experts who employ visual microscopy, hardness testing, and/or leaf analysis [3-22]. Vibrational spectroscopy offers another means of elucidating the structure of wood and characterizing wood types. Early studies [3-23 to 3-25] focused on the use of mid-infrared techniques, e.g., transmission using pressed alkali halide disks or diffuse reflection. More recent studies have focused on Raman spectroscopy, which offers the ability to analyze wood specimens nondestructively.

The 98 Raman spectra consisted of 39 tropical woods (Brazil and Honduras), 28 softwoods (United States), and 31 hardwoods (United States). Raman spectra of the woods were measured on a Perkin Elmer System 2000 FT spectrometer fitted with a standard Perkin Elmer Raman attachment and a modified Spectron 301 Nd^{3+} laser (λ = 1064nm). The spectra were baseline corrected using a three point baseline to minimize the effect of fluorescence. All spectra were measured at a resolution of $4cm^{-1}$. Each

Raman spectrum was an average of 500 scans with the data stored from 3600 to 250 cm^{-1}. All spectra were de-resolved to 4 cm^{-1} yielding 3352 point spectra.

For pattern recognition analysis the Raman spectra were normalized to unit length to adjust for variations in the scattering cross-section of each sample. Each wood sample was represented by a data vector $x = (x_1, x_2, x_3 \ldots x_j \ldots x_{3352})$ where x_j is the Raman intensity of the j^{th} point in the normalized Raman spectrum. The data were autoscaled so that each variable had a mean of zero and a standard deviation of one within the entire set of 98 Raman spectra.

The first step in this study was to apply PCA to the entire data set. Figure 3.6 shows a plot of the scores of the two largest principal components of the 3352-point spectra that comprise this data set. The two largest principal components of the data explain 41% of the total cumulative variance. Each spectrum in the score plot is represented as a point (1 = soft, 2 = hard, and 3 = tropical). There is overlap between the tropical woods, hard woods, and soft woods in the score plot of the data.

Feature selection was the next step as deletion of uninformative features would ensure that discriminatory information about wood type would be the major source of variation in the data. PCKaNN was used to uncover features characteristic of the Raman spectra of each wood-type. In this study, the population consisted of 5000 chromosomes, and the mutation rate was 0.2. Three point cross-over was used, and K_c for each class was set equal to the number of spectra in the class. PCKaNN identified informative features in the data set by sampling key feature subsets, scoring their principal component plots, and tracking those samples or classes that were most difficult to classify. The boosting routine used this information to steer the population to an optimal solution.

After 300 generations, the genetic algorithm identified 11 wavelengths whose principal component plot showed clustering of the Raman spectra according to wood type (see Figure 3.7). The hardwoods, softwoods, and tropical woods are well separated from each other in the score plot. For these 11 features, between group differences are large compared to within group differences. This would suggest that all classification methods will work well with this data. An advantage of using a score plot to display the classification results instead of submitting the 11 features to linear or quadratic discriminant analysis for development of a classifier is that it allows the user to better understand how a classification decision is made for a particular sample.

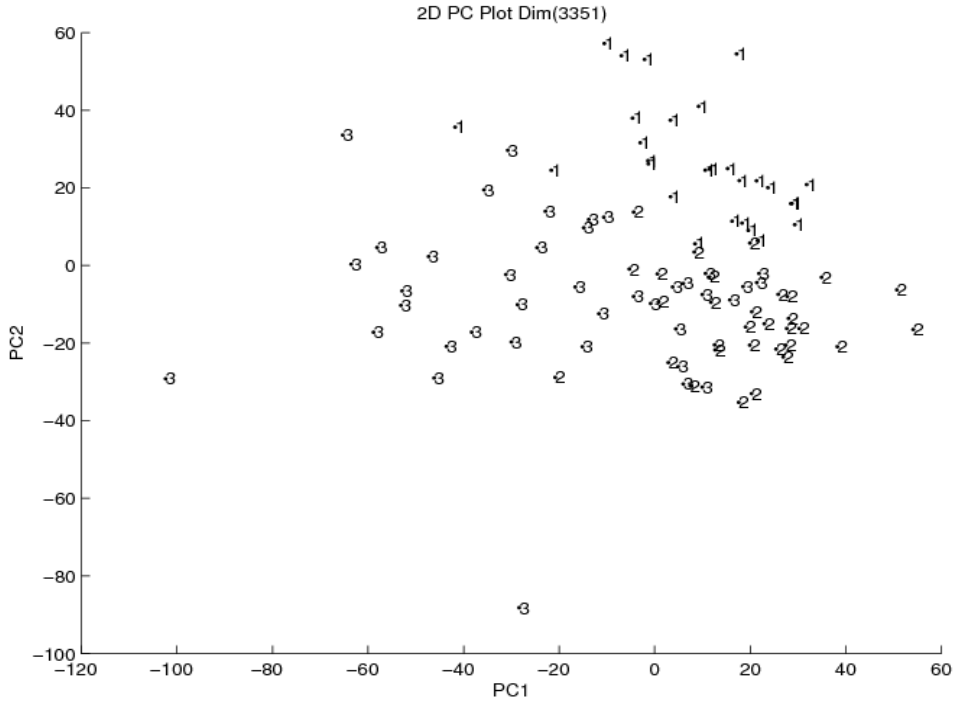

Figure 3.6. A score plot of the two largest principal components of the 3352 wavelengths. Each spectrum is represented as a point in the plot (1 = soft, 2 = hard, and 3 = tropical).

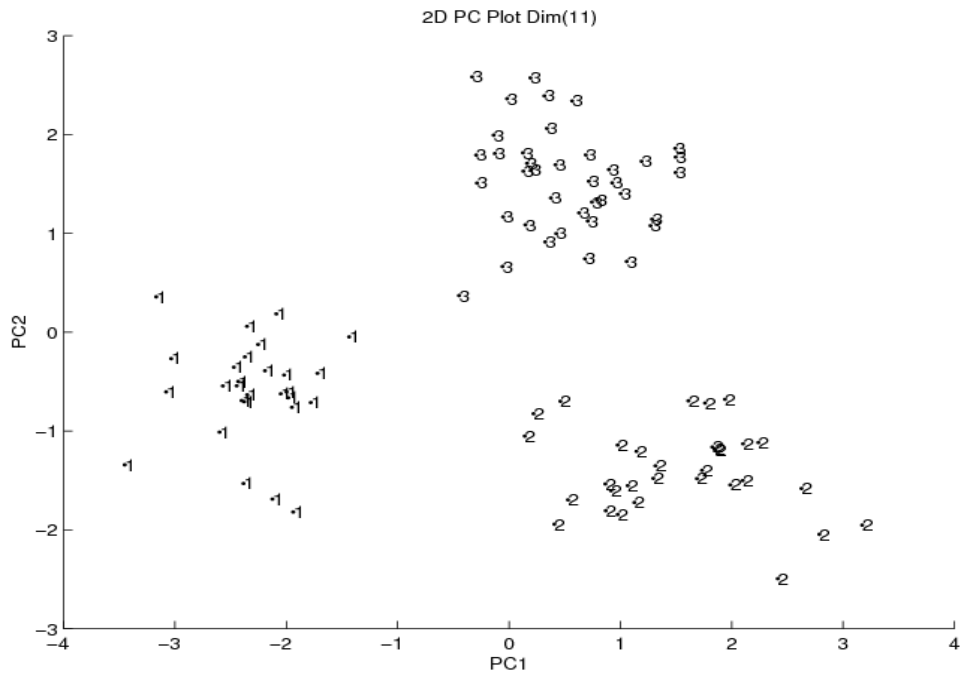

Figure 3.7. A score plot of the two largest principal components of the 11 wavelengths identified by the pattern recognition GA. Each spectrum is represented as a point in the plot (1 = soft, 2 = hard, and 3 = tropical).

The ability of a classifier to predict the class membership of a simulated unknown wood sample was tested using a procedure known as segmented cross validation [3-26]. The data set was divided into N training set prediction set pairs. For each training set, the pattern recognition GA identified informative features by sampling key feature subsets, scoring their PC plots, and tracking those classes and samples that were difficult to classify. For each training set, a classifier is developed using the features identified by the pattern recognition GA and then tested on the corresponding prediction set. Each sample was present in only one of the N prediction sets generated.

The cross validation procedure used in this study differed from the procedure used by other workers [3-26]. In this study, the features selected for each training set are different. In most cross validation studies reported in the literature, the features selected for each training set are the same and are identified using the entire data set prior to

dividing the data into training set/prediction set pairs. For this reason, cross validation usually gives overly optimistic estimates of the error rate. In this study, the validation set samples did not influence the features selected for each training set. Hence, the error rate reported with the cross validation procedure described in this study will be less biased.

For this study, two training set prediction set pair combinations were investigated: 80%/20% (5 training set prediction pairs with 80% of the samples in each training set and the remaining 20% in each prediction set), and 20%/80% (5 training set prediction set pairs with 20% of the samples in each training set and the remaining 80% in each prediction set). Classification success rates for LDA, RDA, and K-NN are shown in Table 3-1 for the features selected by the pattern recognition GA using the three fitness functions. They are comparable to each other. Because the classical learning model (i.e., the PCKaNN fitness function) performed well (100%) with LDA), differences between PCKaNN and the two fitness functions that employ transverse learning are expected to be negligible. We attribute this to the large number of samples (80% of the entire data set) in each training set.

Table 3-1. Discriminant Analysis Results for 80%/20% Cross Validation Study

Method	Average Tset % classification			Average Pset % classification		
	PCKaNN	Modified Hopkins	Hopkins	PCKaNN	Modified Hopkins	Hopkins
LDA	100	100	100	98	99	100
RDA	100	100	100	97	98	99
1-NN	99.25	100	98.75	97	98	98

Table 3-2 summarizes the results from the 20%/80% segmented cross validation study (5 training set prediction set pairs with 20% of the samples in each training set and the remaining 80% in each prediction set) using LDA, RDA, and K-NN for the features selected by the pattern recognition GA with transverse learning and without transverse learning. From this table, it is evident that classifiers developed from features selected by the pattern recognition GA using transverse learning performed better than classifiers developed from features selected by the pattern recognition GA using only PCKaNN. The classification success rate for the modified Hopkins statistic was approximately 10% higher than that achieved by PCKaNN, which is consistent with the improvement that is anticipated when transverse learning [3-21] is used compared to an inductive inference learning model which is developed using only samples that have class labels. The superior performance of classifiers developed from features selected by the pattern recognition GA using the modified Hopkins statistic (as compared to classifiers developed from features selected by the pattern recognition GA using PCKaNN with the Hopkins statistic) can be attributed to boosting used in all facets of the problem, i.e., for both the labeled and unlabeled data points.

Table 3-2. Discriminant Analysis Results for 20%/80% Cross Validation Study

Method	Average Tset % classification			Average Pset % classification		
	PCKaNN	Modified Hopkins	Hopkins	PCKaNN	Modified Hopkins	Hopkins
LDA	100	100	100	72	85.5	75
RDA	100	100	100	68.22	79.25	70
1-NN	100	100	98	78.5	88	81.6

3.3.4 Canonical Variate Analysis. CVA was implemented in the PCKaNN fitness function using a routine developed by Norgaard [3-27] to analyze underdetermined multivariate chemical data sets. This modification to PCKaNN is called CVAKNN. The number of canonical variates required for mapping the data is one less than then number of classes in the data. Hence a data set with three classes is mapped onto a plot of its two largest canonical variates and a data set containing four classes is mapped onto the three largest canonical variates.

3.4 SOFTWARE DESIGN AND IMPLEMENTATION

The software used to implement the pattern recognition GA provides an easy to use graphical user interface that allows the user to smoothly and intuitively navigate through the entire application. The software is portable and can be installed on a broad range of platforms (e.g. personal computers to high performance workstations). It supports additional functionality e.g., file import/export and visualization of multidimensional data. The software has a modular design that allows for the addition of new objects or functionality without requiring major revisions to the existing code. The pattern recognition GA has been implemented in two different versions built using MATLAB or JAVA.

MATLAB is a programming environment that provides an easy to use framework for algorithm development, numerical computations, data analysis and visualization. MATLAB is well designed to handle matrix operations and comes with a multitude of built in functions and toolboxes. However, there are some disadvantages to the MATLAB implementation of the pattern recognition GA. Routines developed in MATLAB can

become outdated as MATLAB is upgraded to newer versions where certain segments or the whole application may stop functioning when run under a newer version of MATLAB. The JAVA implementation of the pattern recognition GA, on the other hand, runs independently on any platform. The Java programming environment unlike MATLAB is freely available and more robust to changes. The JAVA version is also designed for parallel computing to handle larger tasks quickly and efficiently by their distribution.

The implementation of the pattern recognition GA comprises three modules: (1) *Data Preprocessing*; (2) O*peration of the Genetic Algorithm for Pattern Recognition*; and (3) *Visual Analysis of Results*. The first step in data preprocessing is acquiring and formatting the raw data to make it compatible to the program. The raw data is arranged in the form of a table or matrix using a spreadsheet and numerical labels are attached for the identification of samples and classes. Predictions set samples are marked as NaN for the class label. Mathematical transformation such as normalization and auto-scaling are applied to the data. The second module implements the operation of the pattern recognition GA by conducting the search for the best feature subsets using one of several fitness functions that can be chosen by the user. All other GA parameters are also set by the user before starting the pattern recognition GA. At the end of the GA operation, the third module facilitates the graphical display of results for visual analysis. For example, a principal component plot of the best feature subset from the final generation is displayed. The control panel of the visualization module provides several functionalities such as selecting the sample labeling (sample number, class number, etc.), scrolling through the results of different generations, projection of the prediction set samples onto the PC plots

of specific feature subsets identified by the pattern recognition GA, 3D rotation and magnification of a specific region of the plot.

The pattern recognition GA uses value encoding where chromosomes are arrays of positive integers such that each integer represents a distinct feature. Each chromosome is directly encoded as a unique subset of features that represents a potential solution to the problem of optimal feature selection for pattern recognition analysis. The chromosomes for the initial population are generated randomly because no information is available about the features before the start of the genetic algorithm operations. However, it is important to mention certain practical concerns. If certain features are not present in any of the chromosomes of a randomly generated population, then the crossover operator will be unable to reintroduce these features into the general population. Although there is a possibility of these missing features reintroduced into the general population using mutation, there is no guarantee of the mutation operator providing the desired result. Absence of important features in the population imposes significant constraints in the search that prevents the GA from identifying regions in the solution space that contain an optimal solution. For these reasons, it is important that all features are present in the initial population.

The MATLAB version of the pattern recognition GA displays the progress of the run from the first generation to the last generation in terms of the highest and average fitness of the population. It also displays the PC plot and fitness score of the best feature subset identified in each generation along with a plot of the corresponding class and sample weights for the best feature subset. It is neither a simple or easy task to add new functionality or methods as in the JAVA version. The MALAB implementation, unlike

the JAVA implementation has several limitations. The MATLAB implementation is cumbersome to operate since there are more command prompt interaction with different GUI's needed to be invoked for different applications, e.g., data import, GA run parameters and display of results. The JAVA version has a simple user friendly GUI which is very straightforward and all applications are integrated into a series of panels. The MATLAB version is not suitable for large datasets due to a limitation on the size of population that can be handled and the speed of operation. The JAVA version, however, is scalable and can run with larger populations to handle larger data sets very quickly and efficiently. In addition, the JAVA version is graphically well equipped for visual analysis of the results with extended functionalities available. The user can analyze the best results from each generation uncovering additional aspects of the pattern recognition problem. Built-in functions, such as displaying loading plots, and saving score plot images, are also a big advantage.

REFERENCES

3-1. J.H. Holland, “Adaptation in Natural and Artificial Systems”, 6th edition, MIT Press, Cambridge, MA, 2001.

3-2. D.E. Goldberg, “Genetic Algorithms in Search, Optimization, and Machine Learning”, Addison-Wesley Publishing Company, Reading, MA, 1989.

3-3. M. Mitchell, “An Introduction to Genetic Algorithms”, MIT Press, Cambridge, MA, 1998.

3-4. J.H. Holland, “An introduction to intrinsic parallelism”, In Proceedings of the Tenth Anniversary Convocation for IMMD. University of Erlangen, FRG, 1976, 47–55.

3-5. B.K. Lavine, D. Brzozowski, A. J. Moores, C. E. Davidson, and H. T. Mayfield, “Genetic algorithm for fuel spill identification”, Anal. Chim. Acta, 2001, 437, 233–246.

3-6. B.K. Lavine, C. E. Davidson, A. J. Moores, and P. R. Griffiths, “Raman spectroscopy and genetic algorithms for the classification of wood types”, Appl. Spectrosc., 2001, 55(8), 960–966.

3-7. B.K. Lavine, C. E. Davidson, and A. J. Moores, Chemom. Intell. Lab. Syst., 2002, 1(2), 161–171.

3-8. B.K. Lavine, C. E. Davidson, A. J. Moores, and P. R. Griffiths, Vib. Spectrosc., 2002, 28, 83–95.

3-9. B.K. Lavine, C. E. Davidson, R. K. Vander Meer, S. Lahav, V. Soroker, and A. Hefetz, Chemom. Intell. Lab. Syst., 2003, 66, 51–62.

3-10. B.K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, J. Chem. Inf. Comput. Sci., 2003. IN PRESS.

3-11. B.K. Lavine, C. E. Davidson, C. Breneman, and W. Katt, “Genetic algorithms for clustering and classification of olfactory stimulants”, In J. Bajorath (Ed.), Chemoinformatics: Methods and Protocols. Humana Press, Totowa, NJ, 2004, 399-426

3-12. Y. Freund, R. E. Schapire, “A decision theoretic-generalization of online learning and an application to boosting”, Journal of Computer and System Sciences, 1997, Vol. 55 (1), 119-139.

3-13. Y. Freund, R. E. Schapire, “Experiments with a new boosting algorithm”, Proceedings of Machine Learning: Thirteenth International Conference, 1996, 148-156.

3-14. R.E. Schapire, “The boosting approach to machine learning: An overview”, in Nonlinear Estimation and Classification, Springer, 2003.

3-15. R.E. Schapire, M. Rochery, M. Rahim, N. Gupta, “Boosting with prior knowledge for call classification”, IEEE Transactions on Speech and Audio Processing, 2005, 13.

3-16. C. Rudin, I. Daubechies, R.E. Schapire, “On the dynamics of boosting”**,** In Advances in Neural Information Processing Systems 16, 2004.

3-17. A.J. Moores, “The learning genetic algorithm: a hybrid learning system”, Master’s thesis, Clarkson University, Potsdam, NY, 2000.

3-18. S.N. Mukherjee, P. Sykacek, S.J. Roberts, S.J. Gurr, “Gene ranking using bootstrapped P-values”, SIGKKD Explorations, 2003, 5, 16-22.

3-19. W.J. Krzanowski, “Principles of Multivariate Analysis”, Oxford University Press, New York, 2000.

3-20. B.K. Lavine, C.E. Davidson, and W.T. Rayens, “Machine learning based pattern recognition applied to microarray data,” Combinatorial Chemistry & High Throughput Screening,” 2004, 7, 115-131.

3-21. V. Vapnik, “Statistical Learning Theory”, John Wiley & Sons: New York, 1998.

3-22. A.J. Panshin and C. de Zeeuw, “Textbook of Wood Technology: Structure, Identification, Properties, and use of the Commercial Woods of the United States and Canada”, McGraw-Hill, New York, 1980.

3-23. T.P. Schultz, M.C. Templeton, and G.D. McGinnis, “Rapid determination of lignocelluloses by diffuse reflectance Fourier transform infrared spectrometry”, Anal. Chem., 1985, 57, 2867-2869.

3-24. O. Faix, J.H. Bottcher, and E. Bertelt, Proc. 8th Int. Conf. On Fourier Transform Spectroscopy (8th ICOFTS), Lubek (Edited by H. M. Heise, E. H. Korte, and H. W. Siesler). Proc. SPIE 1575, 428, 1992.

3-25. N.L. Owen, D.W. Thomas, “Infrared studies of hard and soft woods”, Appl. Spectrosc, 1989, 43, 451-455.

3-26. H. Martens and T. Naes, “Multivariate Calibration”, John Wiley & Sons, NY, 1989.

3-27. L. Nørgaard, R. Bro, F. Westad, and S.B. Engelsen. “A modification of canonical variates analysis to handle highly collinear multivariate data”, J. Chemom., 2006, 20, 425–435.

CHAPTER IV

WAVELETS AND GENETIC ALGORITHMS FOR SPECTRAL PATTERN RECOGNITION

4.1. INTRODUCTION

A two step procedure for pattern recognition analysis of spectral data is proposed. First, the wavelet packet transform is used to denoise and deconvolute spectral bands by decomposing each spectrum into wavelet coefficients, which represent the samples constituent frequencies. Second, the pattern recognition GA is used to identify wavelet coefficients characteristic of the class. The pattern recognition GA employs both supervised and unsupervised learning to identify features that optimize clustering of the spectra by sample type in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by the selected wavelet coefficients is about differences between the classes in the data set. The principal component analysis routine embedded in the fitness function of the pattern recognition GA serves as an information filter, significantly reducing the size of the search space since it restricts the search to feature sets whose principal component plots show clustering on the basis of class. In addition, the algorithm focuses on those samples and or classes that are difficult to classify as it trains using a form of

boosting to adjust both the sample and class weights. Samples or classes that are consistently classified correctly are not as heavily weighted as samples that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a smart one pass procedure for feature selection, classification and prediction

The advantages of the proposed methodology have been demonstrated in four studies which are the subject of this chapter. In the first study [4-1], differential mobility spectra of alkanes, alcohols, ketones, cycloalkanes, substituted ketones, and substituted benzenes with carbon numbers between 3 and 10 were obtained from gas chromatography-differential mobility spectrometry (GC-DMS) analyses of mixtures in dilute solution. Spectra were produced in a supporting atmosphere of purified air with 0.6–0.8ppm moisture, gas temperature of 120°C, sample concentrations of <0.2–5ppm, and ion source of 2mCi (74 MBq) ^{63}Ni. Multiple spectra were extracted from chromatographic elution profiles for each chemical providing a library of 390 spectra from 39 chemicals. The spectra were analyzed for structural content by chemical family using the pattern recognition GA. The wavelet packet transform was used to denoise and deconvolute the DMS data by decomposing each spectrum into its wavelet coefficients, which represent the sample's constituent frequencies. The wavelet coefficients characteristic of the compound's structural class were identified using the genetic algorithm for pattern recognition analysis.

In the second study [4-2] the feasibility of near-infrared (NIR) spectroscopy to identify waxy wheat and differentiate it from partial waxy and wild-type phenotypes was

investigated. The effectiveness of NIR reflectance spectroscopy for classification of waxy (low amylase) wheat into its four possible alleles was undertaken. The four alleles were wild type, waxy, and two intermediate states which correspond to partially waxy. In the two intermediate states, a null allele occurs at either of the two homologous genes (*Wx-1A* and *Wx-1B*) that encodes for production of the enzyme, granule bound starch synthase (GBSS), which controls amylose synthesis.

The third study [4-3 to 4-4] involved the development of an IR search prefilter for carboxylic acids using the two step procedure. Carboxylic acid search prefilters developed from 463 vapor phase IR spectra were successfully validated (100% correct classification) using an external prediction set of 92 IR spectra. Recognition rates for carboxylic acids search prefilters previously reported in the literature for vapor phase spectra have varied from 81% to 92% [4-5 to 4-9]. This is comparable to a scientist somewhat familiar with IR. These search prefilters were developed using raw absorbance values from selected spectral regions and did not include information about band shape and band width. By using wavelet coefficients generated from the wavelet packet tree to represent the features of the IR spectrum, information about band shape and band width could be encoded in the search prefilters developed in this study.

In the fourth study, prefilters for searching IR spectral libraries of the Paint Data Query (PDQ) automotive database to differentiate between similar but nonidentical Fourier transform infrared (FTIR) paint spectra were developed. Currently, the identification of the make, model and year of a motor vehicle involved in a hit and run collision from only a clear coat paint smear left at the crime scene is not possible. Applying the wavelet packet transform, FTIR spectra of clear coat paint smears were

denoised and deconvolved by decomposing each spectrum into wavelet coefficients. The GA for pattern recognition was used to identify wavelet coefficients characteristic of the model and manufacturer of the automobile from which the spectra of the clear coats were obtained. Even in challenging trials where the samples evaluated were all the same manufacturer (Chrysler) with a limited production year range, the respective models and manufacturing plants were correctly identified. Search prefilters for spectral library matching are necessary for forensic databases to extract investigative lead information from a clear coat paint smear. Information obtained from these searches can also serve to quantify the general discrimination power of original automotive paint comparisons encountered in casework, and to succinctly communicate the significance of the evidence to the courts.

4.2 PATTERN RECOGNITION ANALYSIS OF DIFFERENTIAL MOBILITY SPECTRA WITH CLASSIFICATION BY CHEMICAL FAMILY

Differential mobility spectrometry (DMS) is an emerging technology for characterizing volatile organic compounds (VOCs) by measuring differences in the gas phase ion mobility of these VOCs through application of varying electric fields (between ~800 and 20,000 Vcm^{-1} or higher) to generate analytically useful information [4-5 to 4-7]. In DMS, the sample is ionized by the same methods or reactions as found in IMS. However, ions derived from the sample are moved in a gas flow of $<4ms^{-1}$ between two metal plates separated by a distance of 0.1mm or larger. An electric field, called the separation field, is applied between the plates as a <1MHz asymmetric waveform with amplitudes of ≤ -500 and $\geq 20{,}000 Vcm^{-1}$. Applications of DMS include explosive detection [4-8] and ion filtering before mass spectrometry to reduce chemical noise in

biological measurements [4-9 and 4-10]. DMS instrumentation is simpler with a smaller footprint and costs less than conventional ion mobility spectrometry (IMS) analyzers [4-5]. Using DMS, both positive and negative ions can be simultaneously characterized which is a distinct advantage for chemical analysis. Typical features of differential mobility spectra are shown in Figure 4.1, and include peaks for the reactant ions (at ≈−22V), a protonated monomer (at ≈−7.5V) and a proton bound dimer (at ≈+2V). These ions arise from the chemistry of gas phase reactions through proton transfer found with all β sources at ambient pressure providing both qualitative and quantitative information about sample vapors [4-11 to 4-13].

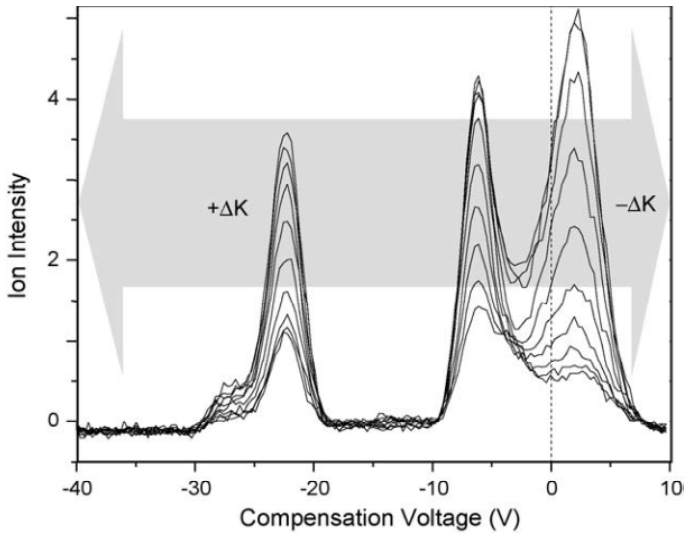

Figure 4.1. DMS spectra of octanone showing a reactant ion peak (−22 V) and peaks for product ions from chemical ionization of sample vapors, in positive polarity. The product ions are a protonated monomer (−6V) and a proton bound dimer (+2 V) form in purified air with moisture of ~0.2ppm. The relationship between ΔK (ion mobility difference) for the product ion and compensation voltage from the DMS measurement is shown using arrows. The ion source was 1 mCi of ^{63}Ni. (Courtesy of *Anal Chim. Acta* 2006, 579, 1–10)

Mobility spectra from conventional IMS analyzers are known to contain both structural information [4-14] and chemical class specific information [4-11 and 4-12].

The possibility of using pattern recognition methods to categorize differential mobility spectra by chemical family can be attributed to the spectral region near the reactant ion peak. Ions at low intensity and small mass observed in ion mobility spectra are understood to be class specific fragment ions formed during the ionization step in air at ambient pressure. Although models of ion separation using IMS have been proposed [4-15 and 4-16] and supported by subsequent studies [4-17], there has been little exploration of the detailed features in DMS and the information content available in their profiles.

The major objective of this study was to ascertain whether differential mobility spectra contain information about chemical family and the exploitation of this information by pattern recognition methods. A two-step procedure was used in this study to analyze DMS data. In the first step, the wavelet packet transform [4-18 to 4-20] was used to denoise and deconvolute the DMS data by decomposing each spectrum into its wavelet coefficients, which represent the sample's constituent frequencies. In the second step, the pattern recognition GA [4-21 to 4-25] was used to identify wavelet coefficients characteristic of chemical family. The pattern recognition GA employed supervised learning to identify coefficients that optimize the clustering of spectra in a plot of the two or three largest principal components of the data. A spectral library was prepared using DMS which was run under uniform experimental conditions. Thirty nine VOCs from six chemical classes (Table 4-1) were obtained in high purity from various manufacturers and stock solutions were prepared using glass distilled grade hexane. Standards were prepared by diluting stock solutions until DMS analysis produced depletion of the reactant ion peak (RIP) to 20–30% of the original RIP intensity. This avoided sample saturation of the ion source and corresponded to concentrations from 50 to 200 ng μL^{-1}.

Spectra were obtained continuously throughout a GC/DMS measurement. DMS spectra were produced in a supporting atmosphere of purified air with 0.6–0.8ppm moisture, gas temperature of 120°C, sample concentrations of <0.2–5ppm, and ion source of 2mCi (74 MBq) ^{63}Ni. Further details about the experimental conditions used to generate this data can be found elsewhere [4-1].

Spectra found in chromatographic peaks corresponding to authentic chemicals, confirmed by GC/MS measurements, were extracted into spreadsheets of intensity at 150 columns for the compensation voltage axis from −40 to +10V. Eight to fifteen spectra were available for each chemical throughout the dynamic range of the analyzer from limit of detection to maximum response without saturation of the ion source. In all, 390 spectra from six chemical families were included in the spectral library (Table 4-1). These compounds included cycloalkanes, alkanes, alcohols, ketones, substituted ketones, and substituted benzenes.

Table 4-1. Composition of the DMS spectral data set

Chemical family	Number of chemicals	Number of all spectra
Alkanes	9	90
Cycloalkanes	7	70
Alcohols	6	60
Ketones	5	47
Substituted ketones	8	83
Substituted benzenes	4	40
Total	39	390

The differential mobility spectra produced in these studies were consistent with previous spectra using the same or comparable DMS analyzers [4-6, 4-7, and 4-17]. Since the DMS analyzer exhibited fast response to effluent composition and also fast restoration to baseline levels on the falling edge of an eluting peak [4-6], the spectral files extracted from single chromatographic peaks provided profiles for authentic chemicals, free of impurities from manufacture or interferences from nearby eluting chemicals. Although spectra for chemicals with retention times near the solvent peak might be prone to contamination, hexane presented little response in positive polarity with chemical ionization at ambient pressure. Thus, the spectral library was considered free of complications that can arise from either impurities or various contaminations. Since, concentration dependence was known to be a necessary facet and a practical requirement for any robust classification method from prior studies of pattern recognition techniques with ion mobility spectra [4-11 and 4-12], a range of concentrations were used to produce differential mobility spectra in this study. The differential mobility spectra were comprised of an RIP and peaks for the protonated monomer ($MH^+(H_2O)_n$) and commonly, but not uniformly, the proton bound dimer ($M_2H^+(H_2O)_n$). The use of whole spectra (compensation voltages from −40 to +10V) provided the classifier with full content of peak width and shape, and not simply a peak location.

The Daubechies 12 wavelet up to the sixth level of decomposition was used to denoise and deconvolute the DMS data. Criteria used in this study for selection of a suitable wavelet is based on the ability of the wavelet to extract chemical information from the data, which can then be exploited by the pattern recognition GA for classification of the spectra into their respective chemical families. There was no

improvement in the ability of the pattern recognition GA to correctly classify the spectra when other wavelets such as the Daubechies 6 or 18 were used to denoise and deconvolute the data.

For pattern recognition analysis, each DMS vector was initially represented by 3690 wavelet coefficients. The first step in this study was to apply PCA to the data. Figure 4.2 shows a principal component (PC) plot developed from the 390 spectra and 65 wavelet coefficients identified by the pattern recognition GA. The alkanes (1's) and cycloalkanes (2's) are dispersed throughout the entire PC plot, whereas the alcohols (3's), ketones (4's), substituted ketones (5's) and substituted aromatics (6's) cluster in specific regions of the plot.

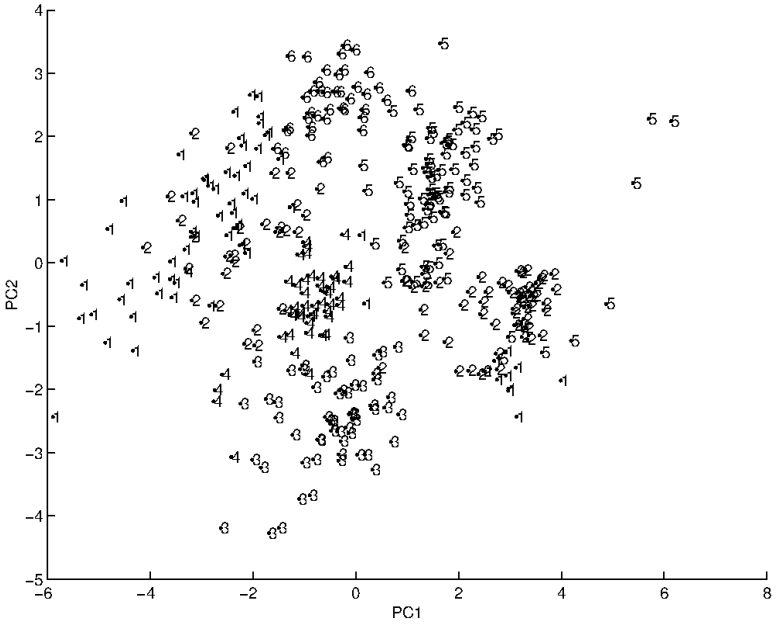

Figure 4.2. A plot of the two largest principal components of the 390 spectra that comprise the entire data set and the 65 wavelet coefficients identified by the pattern recognition GA. (1) alkanes, (2) cycloalkanes, (3) alcohols, (4) ketones, (5) substituted ketones, and (6) substituted aromatics. (Courtesy of *Anal Chim. Acta* 2006, 579, 1–10)

The failure of the alkanes and cycloalkanes to form a well defined cluster suggests that information characteristic of chemical family is not present in their spectra. This is probably due to their low proton affinity. Consequently, the response obtained for these chemicals is low since the origination of the response, displacement by the molecule of a water adduct on a hydrated proton, is poor. It is true that a response can be observed if moisture is low and charge exchange reactions occur with H_2O^+ or NO^+, both of which are found in dry atmospheres in a beta source with air or nitrogen. However, the conditions used in this study were not too dry, and these would have been very minor contributions. Therefore, the poor response exhibited by the alkanes and cycloalkanes is due to low spectral intensity where it counts, which is in their fragment ion intensity.

For these reasons, both the alkanes and cycloalkanes were removed from the analysis. Figure 4.3 shows a PC plot developed using spectra from the remaining four chemical families and 53 features identified by the pattern recognition GA. The substituted ketones (5's) and the aromatics (6's) are well separated from each other and from the other two chemical families, whereas there is some degree of overlap between the alcohols and the simple ketones. This conclusion was reinforced by performing two additional comparisons of the spectra using the pattern recognition GA to identify wavelet coefficients for discrimination of the chemical families in these comparisons: (1) classification (see Figure 4.4) of the spectra of the alcohols, substituted ketones, and substituted aromatics, and another classification study (see Figure 4.5) involving the simple ketones, substituted ketones, and the substituted aromatics.

The classification results obtained for the simple ketones and substituted ketones were unexpected in view of the fact that ketones undergo little or no fragmentation,

which is the basis for their chemical class recognition, although the response to form a product ion $MH^+(H_2O)$, can be good or very good. The ketone functional group can form a stable association with the gas phase proton with the ketone acting as a base. Nevertheless, the lack of fragmentation has been seen in past studies as a detriment to classification since the fragment ions are either of very low intensity or not detectable.

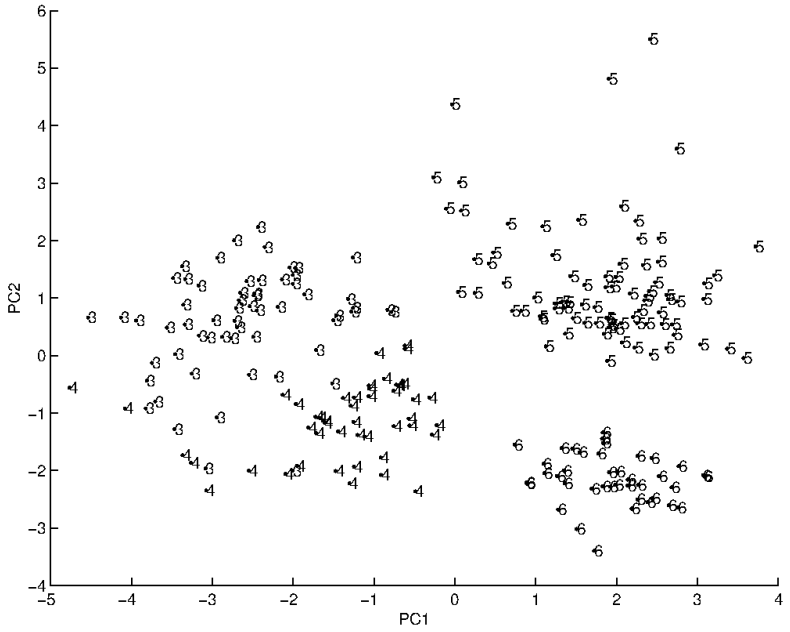

Figure 4.3. A plot of the two largest principal components of the 230 spectra from the remaining four chemical families and the 53 features identified by the pattern recognition GA. (3) alcohols, (4) ketones, (5) substituted ketones, and (6) substituted aromatics. (Courtesy of *Anal Chim. Acta* 2006, 579, 1–10)

Other analyses that were performed on the DMS spectral data included the use of back propagation neural networks to develop classifiers for characterizing spectra by chemical families. The neural network classifiers produced results similar to those obtained using the wavelets. Further details about the neural network studies can be found elsewhere [4-1]. The performance of the neural network also indicated that

chemical class information was largely associated with the region of the spectrum between −26.6 V and −20 V.

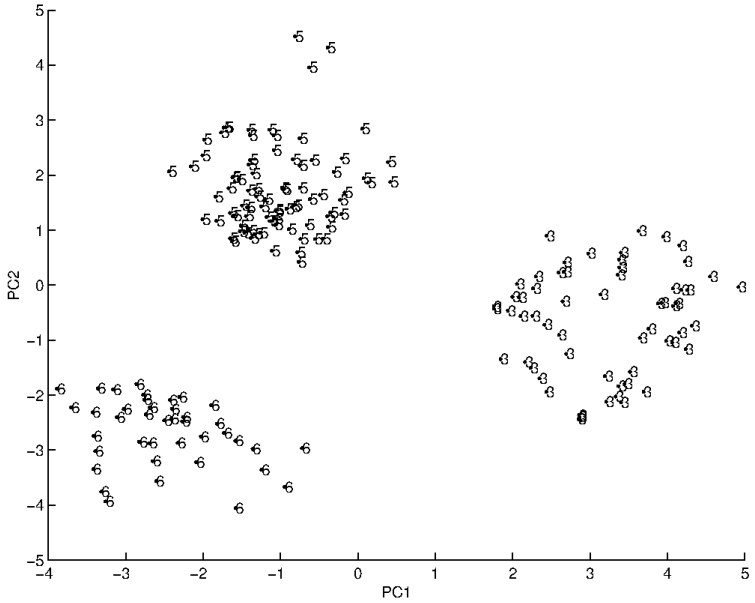

Figure 4.4. A plot of the two largest principal components of the 183 spectra from three chemical families and the 67 features identified by the pattern recognition GA. (3) alcohols, (5) substituted ketones, and (6) substituted aromatics. (Courtesy of *Anal Chim. Acta* 2006, 579, 1–10)

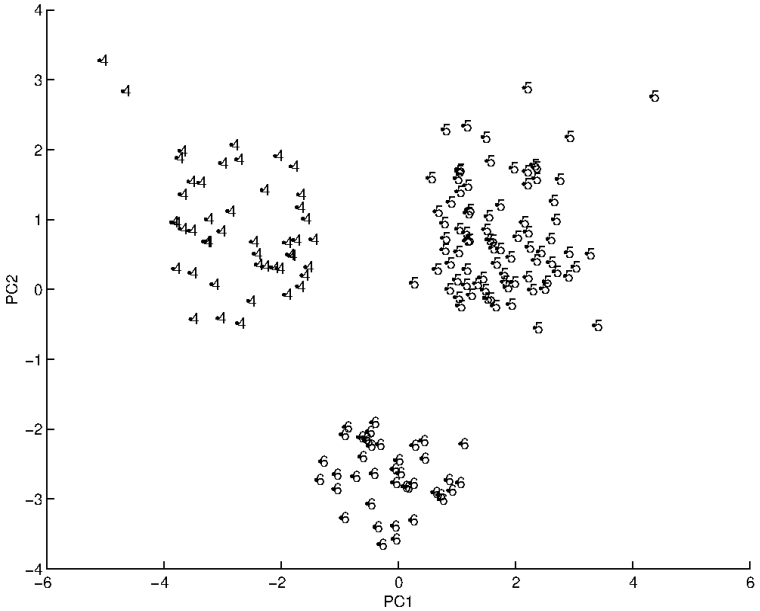

Figure 4.5. A plot of the two largest principal components of the 170 spectra from three chemical families and the 50 features identified by the pattern recognition GA. (4) ketones, (5) substituted ketones, and (6) substituted aromatics. (Courtesy of *Anal Chim. Acta* 2006, 579, 1–10)

Use of wavelets for spectral deconvolution followed by the pattern recognition GA for identification of informative features related to chemical class also showed that categorization of spectra by chemical families is observed in DMS for some functional groups but not others. The results of the wavelet study also indicate that chemical class information is available in the DMS spectra. These findings also demonstrate that DMS analyzers, like IMS instruments, can be valuable for the identification of a chemical not present in spectral libraries. The first step in any identification, which is assignment of the chemical class, is plausible using DMS based on the results from this study.

4.3 IDENTIFICATION OF WAXY WHEAT ALLELES BY NEAR INFRARED REFLECTANCE SPECTROSCOPY

During the past decade there has been renewed interest in the breeding of 'waxy' and 'partially waxy' wheat due to a multitude of potential applications including the development of stock flour material for blending by millers, flour for Asian noodle-making, and a substitute for waxy maize starch in the paper making and adhesive industries [4-26 to 4-30]. The 'waxy' condition of wheat is related to its amylose content. An enzyme called granule bound starch synthase (GBSS), which is also known as 'waxy' protein, is primarily responsible for amylose in wheat plants [4-30]. Absence of GBSS causes near zero amylose content in wheat, which is commonly referred to as the 'waxy' condition. Under natural conditions, active isoforms of GBSS in wheat are encoded by two waxy genes, *Wx-A* and *Wx-B*, for tetraploid (durum) wheat and three waxy genes, *Wx-A*, *Wx-B*, and *Wx-D*, for hexaploid wheat. In the native wild-type state, the wheat possesses all GBSS isoforms. The partially waxy condition occurs in a wheat line by natural mutation or through conventional breeding practices, when at least one

(but not all) of the waxy genes is a null allele. With the growing demand for waxy and partially waxy wheat, there is a need to develop a reliable and rapid test to authenticate the waxy condition.

Identification of waxy seeds is presently done using iodine-binding blue complex colorimetry for the determination of amylose content. However, this method is not definitive for identification of partially waxy lines [4-31]. Furthermore, the procedure is time consuming, has poor precision and is not suitable for commercial grades of wheat that have a narrower range of amylose content. Rather than quantifying the proportion of amylose in wheat by current chemical methods, an alternative approach is required to characterize samples according to the number of active GBSS genes. Detection of the different GBSS isoforms is typically performed by SDS-PAGE [4-32], ELISA [4-33], or multiplex PCR techniques [4-34]. These methods are expensive, complicated and not readily amenable to the various stages of wheat breeding, marketing and production. Near-infrared (NIR) spectroscopy is a simple, fast and inexpensive methodology that is well-established and widely used for determination of protein, moisture and other properties of cereals at grain processing facilities.

Previous attempts to characterize wheat genotypes of ground meal and whole kernel samples using NIR reflectance spectroscopy to characterize phenotype have not been successful [4-2 and 4-35]. These studies were performed using PCA and LDA to analyze the NIR data. Although the classification models were able to identify the waxy genotype, identification of the three other genotypes (*wx-A1* null, *wx-B1* null, and wild type) was not possible. The low classification success rate obtained in these studies,

which was approximately 50%, can be attributed to the inability to identify partial waxy lines.

In this study, the wavelet packet transform was applied to NIR spectra, followed by the use of the pattern recognition GA to identify informative wavelet coefficients that can be used to characterize spectra according to the four genotypes: waxy, *wx-A1* null, *wx-B1* null and wild type. NIR spectra of wheat from a previously published study [4-2] were used. The objective of this study was to evaluate the feasibility of NIR reflectance spectroscopy to genotype wheat samples. The confounding of chemical information with the expression level of the genes was also investigated by analyzing the selected wavelet coefficients for correlation with amylose and protein content.

Wheat samples were collected from various breeding programs. The number and type of active GBSS genes in the wheat samples were determined using SDS-PAGE. The samples were separately ground on a lab scale cyclone grinder and NIR spectra were generated using a reflectance spectrophotometer (Foss-NIR System Model 6500). An average spectrum of each sample was generated using 32 scans/spectrum with a wavelength range from 1100-2498nm and 2nm resolution. Amylose content of wheat samples was measured by iodine-binding blue complex colorimetry. Further details about the conditions used to collect the data can be found elsewhere [4-2]. Figure 4.6 shows a typical NIR spectrum of a wheat sample of the waxy type.

Ninety five NIR spectra from four different wheat genotypes were available for this study, see Table 4-2. Each spectrum was represented as a data vector, $x = (x_1, x_2, x_3 \ldots x_j \ldots x_{700})$ where x_j is the absorbance of the j^{th} point of the NIR spectrum. All spectra were normalized to unit length to nullify the effect of differences in the optical path

length between samples. The first step in this study was to perform PCA on the 95 NIR spectra. All spectral features were auto-scaled prior to performing PCA to remove any inadvertent weighing of the variables that otherwise would occur due to differences in magnitude among the spectral features comprising the data set. Figure 4.7 shows a plot of the two largest principal components of the 95 wheat samples using all 700 spectral features. Each spectrum (i.e., sample) is represented by a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type). The overlap of NIR spectra from the different genotypes in the principal component plot of the data is evident. One sample in the plot was identified as an outlier and was deleted from the analysis as its spectrum was very different from the other spectra in the data set.

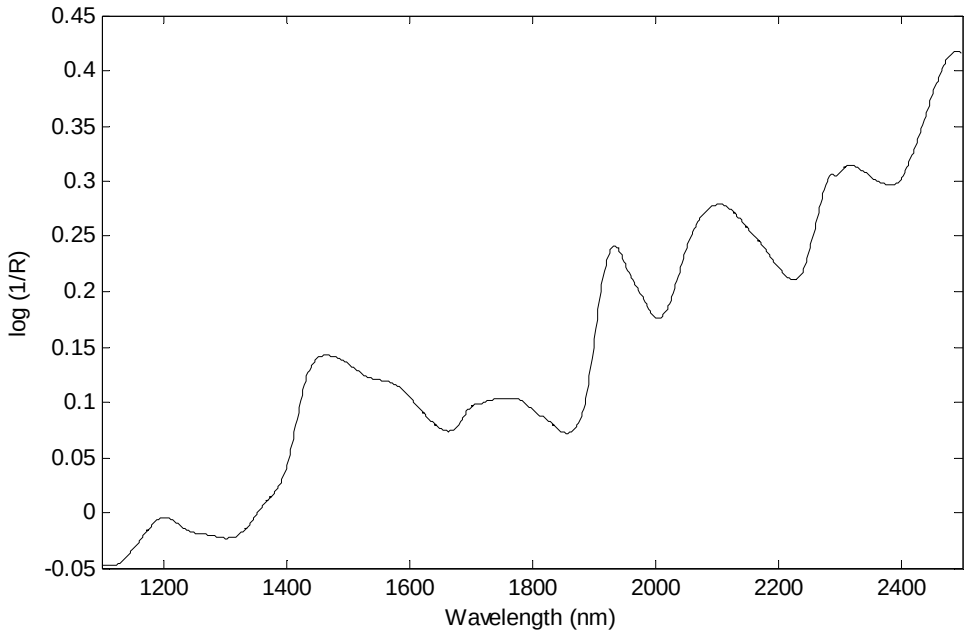

Figure 4.6. Typical spectrum of waxy wheat obtained by NIR diffused reflectance spectroscopy

Table 4-2. Wheat Data Set

Genotype	Number of NIR spectra
Waxy	24
Wx-A1	25
Wx-B1	24
Wild type	22
All	95

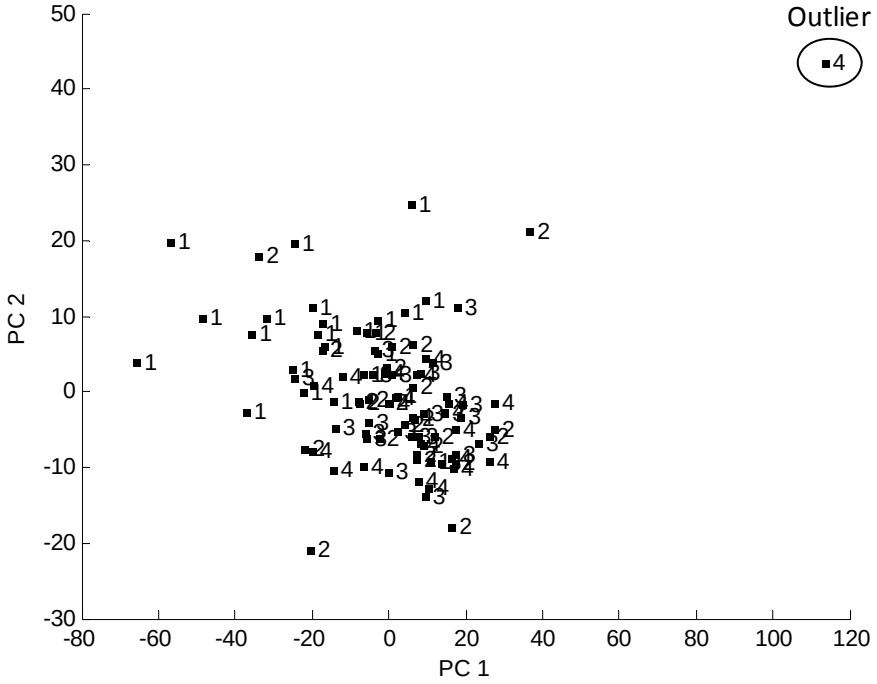

Figure 4.7. Plot of the two largest principal components of the 95 NIR spectra and 700 points that comprise the wheat data set. Each NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

The next step in this study was feature selection. The pattern recognition GA using the PCKaNN fitness function attempted to identify features in the NIR spectra of the wheat samples characteristic of genotype. The pattern recognition GA identified spectral features that optimized the separation of NIR spectra by genotype in a plot of the

two or three largest principal components of the data by sampling key feature subsets, scoring their PC plots, and tracking wheat samples or genotypes that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 300 generations, the GA identified 6 spectral features (see Figure 4.8). From the PC plot of these 6 spectral features, it is evident that information about genotype cannot be directly obtained from the original data. Further preprocessing of the original data is necessary.

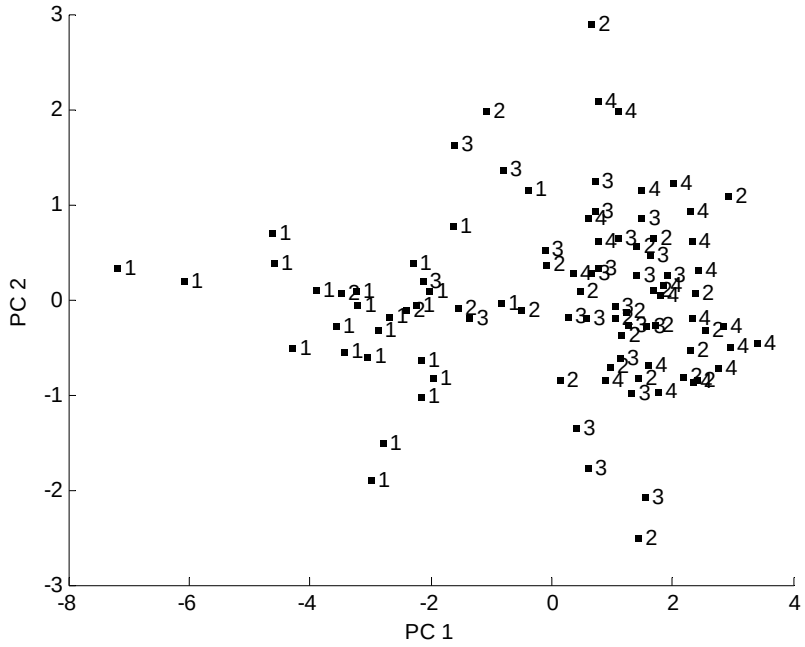

Figure 4.8. Plot of the two largest principal components of the 94 NIR spectra and 6 wavelengths identified by the pattern recognition GA. Each NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

For this reason, the second derivative function was applied to each spectrum using a 7-point Savitzky-Golay filter. The second derivative was used because it eliminates sloping baselines and offsets, as well as deconvolutes overlapping spectral bands. Figure

4.9 shows a plot of the two largest principal components of the 94 second derivative NIR spectra and 686 features. Again, separation of the samples by wheat genotypes is not observed. Applying the PCKaNN fitness function of the pattern recognition GA to the second derivative spectra, 17 features were identified that contained some information about wheat genotype (see Figure 4.10). Although the PC plot shows separation of waxy wheat from the other wheat samples, the partially waxy and wild type wheat samples overlap. Although the ability to discriminate waxy wheat has been improved by the second derivative, this preprocessing method alone cannot extract sufficient information from the spectral data about the different wheat genotypes.

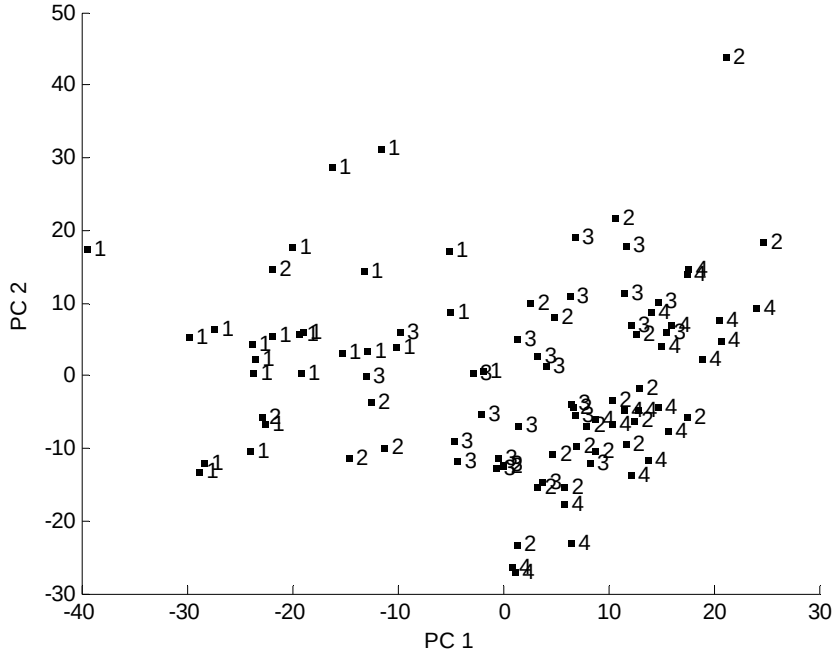

Figure 4.9. Plot of the two largest principal components of the 94 second derivative spectra and 686 points. Each second derivative NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

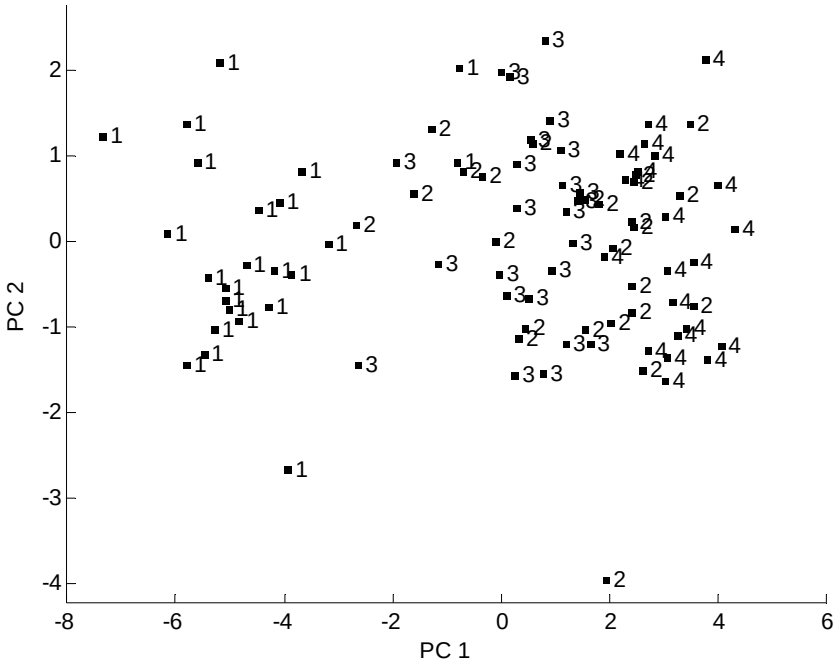

Figure 4.10. Plot of the two largest principal components of the 94 second derivative spectra and 17 features identified by the pattern recognition GA. Each second derivative NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

More powerful spectral preprocessing methods are required to denoise and deconvolve the spectral bands of these samples and to resolve information related to wheat genotype. For this reason, the wavelet packet transform was applied to the data. The Daubechies 4 wavelet was selected as the other member of the Daubechies family had either sharper or broader features compared to the NIR spectra. Furthermore, the Daubechies 4 wavelet does not suffer from oscillations, which is the case with the higher Daubechies mother wavelets. The Daubechies 4 mother wavelet at the 8^{th} level of decomposition was used to denoise and deconvolute each NIR spectrum into wavelet coefficients. Wavelet decomposition at the 4^{th} or 6^{th} level could not provide sufficient resolution of the signal in the data with respect to information about genotype.

To identify the informative wavelet coefficients, the pattern recognition GA was applied to the data. The pattern recognition GA identified 55 wavelet coefficients which contained information about wheat genotype. Figure 4.11 shows a plot of the two largest principal components of the 94 NIR spectra comprising the data set and the 55 wavelet coefficients identified the pattern recognition GA. Separation of NIR spectra by wheat genotype is evident.

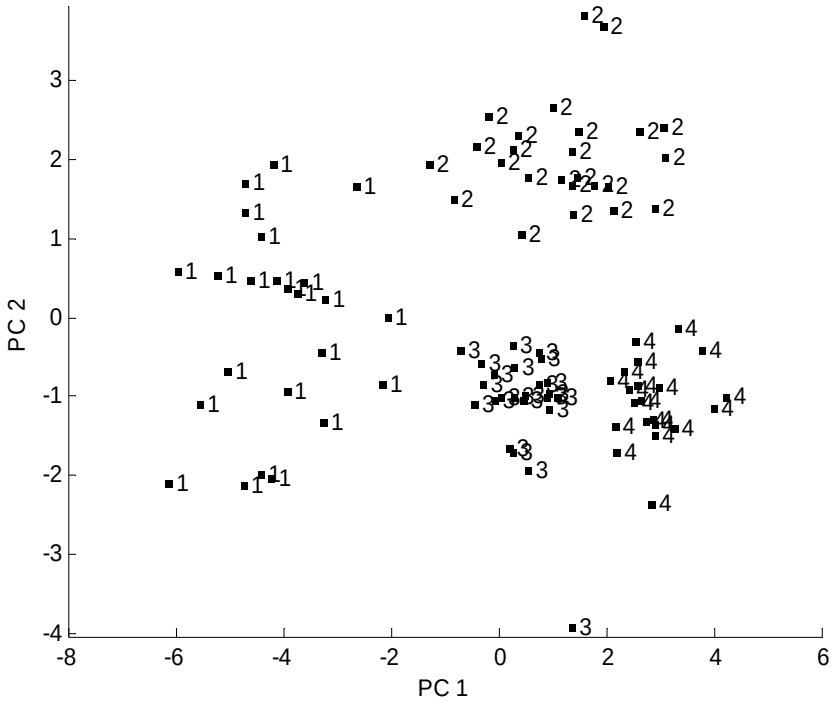

Figure 4.11. Plot of the two largest principal components of the 94 wavelet transformed NIR spectra and 55 wavelet coefficients identified by the pattern recognition GA. Each wavelet transformed NIR spectrum is represented as a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

To learn more about the information content of the 55 wavelet coefficients, partial least squares (PLS) regression was used to develop a correlation between the wavelet coefficients and the amylose or protein content of the wheat. PLS calibration models were developed for amylose and protein using the wavelet transformed NIR spectra and

the 55 wavelet coefficients identified by the pattern recognition GA. In PLS, the design space of the wavelet coefficients is approximated with one of lower dimension whose basis vectors are defined in terms of linear combinations of the original variables. Since the latent variables in PLS are developed simultaneously along with the regression model, each latent variable which is generated from an eigen analysis of the data is a linear combination of the original measurement variables rotated to ensure maximum correlation with the concentration information provided by the response variable. Latent variables produced by PLS are better at capturing information relevant to a calibration than a corresponding principal component analysis regression model.

Table 4-3 shows the results of the PLS analysis including the standard error of calibration (SEC), the range of amylose and protein content spanned by the 94 wheat samples, and the correlation coefficients for the amylose and protein calibrations. It is evident from Table 4-3 that information about amylose content and protein content is also present in the 55 wavelet coefficients identified by the pattern recognition GA. A histogram of the mean amylose and mean protein content of each genotype is shown in Figure 4.12. The standard deviation of the amylose and protein content for the wheat samples from each genotype is also listed in parentheses in the histograms. From Table 4-3, Figures 4.11 and 4.12, one can conclude that classification of the wheat samples by genotype is not strongly influenced by the protein or amylose content of the wheat.

Table 4-3. PLS Results

Y-block	PLS components	SEC (%)	Correlation
Amylose (14.6%-21.2%)	3	3.47	0.94
Protein (1.8% - 31.9%)	3	0.38	0.94

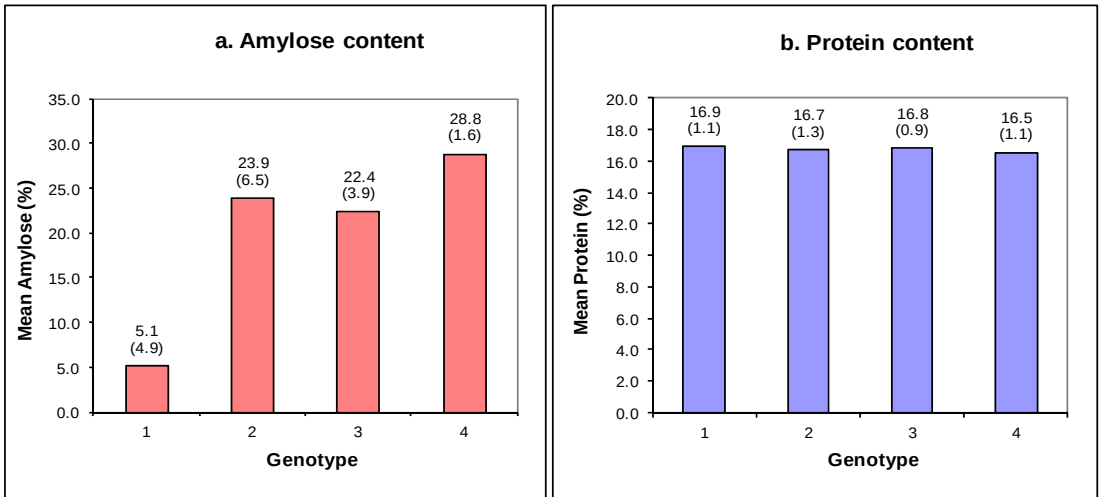

Figure 4.12. Plot of (a) amylose and (b) protein content (mean %, standard deviation) for each genotype in the wheat data set (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

To assess the predictive ability of the NIR methodology for genotyping wheat, object validation was performed. First, the 94 NIR spectra were divided into a training set of 86 wheat samples and a prediction set of 8 samples. The spectra comprising the prediction set were chosen by random lot. During the course of this study, two training set samples were identified as outliers since their removal from the training set allowed the pattern recognition GA to converge towards a solution. These two wheat samples were waxy but their amylose content was approximately 13% and was found to be substantially higher than the other waxy wheat samples whose amylose content was less than 5%. Therefore, these two samples were deleted from the study. Table 4-4 shows the composition of the prediction set.

The pattern recognition GA was applied to the 84 spectra of the training set. The GA identified informative wavelet coefficients for the training set samples by sampling key feature subsets, scoring their PC plots, and tracking those genotypes and samples that

were difficult to classify. After 300 generations, the GA identified 32 wavelet coefficients whose PC plot showed clustering of the training set samples on the basis of genotype. The prediction set of 8 NIR spectra was then employed to assess the predictive ability of the 32 wavelet coefficients identified by the pattern recognition GA. We chose to map the 8 spectra directly onto the principal component map defined by the 84 spectra and 32 wavelet coefficients. Figure 4.13 shows the prediction set samples projected onto the principal component map developed from the training set. All but one (*wx-A1*) wheat sample was projected in a region of the map near wheat samples of the same genotype.

Additional object validation studies were performed. Twenty three training set/prediction set pairs were generated by random selection where the training set consisted of 88 samples and the corresponding prediction set contained 4 samples. Any particular wheat sample was present in only one of the 23 prediction sets generated. For each training set, wavelet coefficients whose PC plots showed clustering on the basis of genotype were identified by the pattern recognition GA using PCKaNN with the modified Hopkins statistic. The ability of the wavelet coefficients selected by the pattern recognition GA to classify NIR spectra was further assessed using the corresponding prediction set. The object validation procedure used here differs from the procedure used by other workers. In this study, the features selected for each training set are different. In most object validation studies, the features selected for each training set are the same and are identified using the entire data set prior to dividing the data into training set/prediction set pairs. For this reason, object validation generally gives overly optimistic estimates of the error rate. In this study, the validation set samples do not

influence the features selected for each training set. Hence, the error rate reported with the cross validation procedure described in this study is less biased.

Table 4-4. Prediction Set

Genotype	Number of NIR spectra
Waxy (1)	2
Wx-A1 (2)	2
Wx-B1 (3)	2
Wild type (4)	2

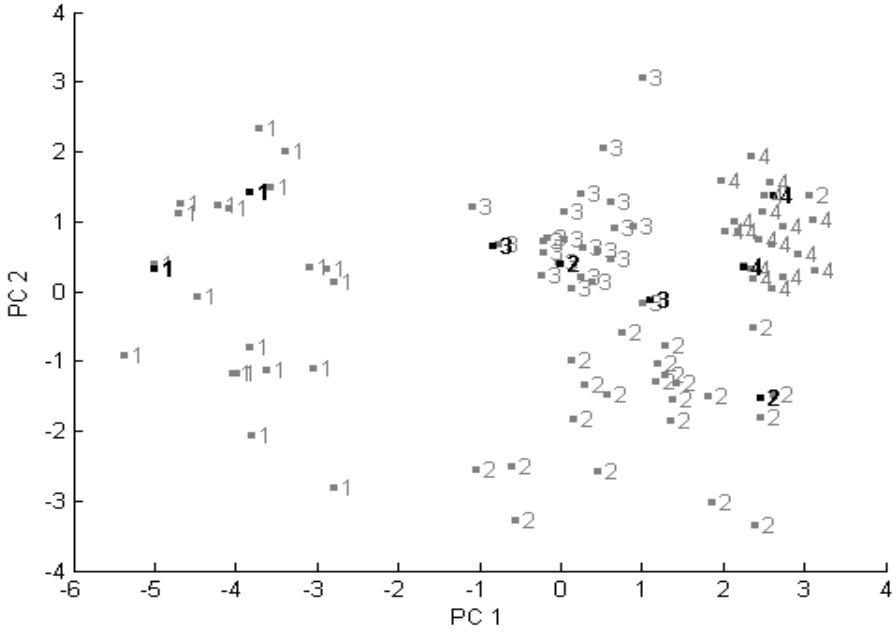

Figure 4.13 Projection of the prediction set samples onto the PC plot of the 84 wavelet transformed NIR spectra and 32 wavelet coefficients identified by the pattern recognition GA. Each wavelet transformed NIR spectrum in the training set (grey) and prediction set (black) is represented as a point in the plot (1 = waxy type, 2 = *wx-A1* null, 3 = *wx-B1* null, and 4 = wild type).

A summary of the object validation results is given in Table 4-5. The misclassified Wx-A1 samples were assigned to Wx-B1, and the misclassified WX-B1 samples were assigned to Wx-A1. This is not surprising as Wx-A1 and Wx-B1 are both single nulls but differ in the location of the GBSS gene on the chromosome. The 3

misclassified Wild types were assigned to Wx-A1 or Wx-B1. These results suggest that NIR reflectance spectroscopy has the potential to genotype wheat. Using wavelets and the pattern recognition GA, it was possible to differentiate the 4 genotypes including the partial waxy types with a recognition rate of 80% versus 50% which was reported in a previous study [4-2]. Amylose and protein content of the wheat samples did not appear to be a significant covariate that would confound the classification of the NIR spectra by genotype.

Table 4-5. Summary of Object Validation Results

Genotype	Number of samples	Number of correct classifications	Classification success rate (%)
Waxy (1)	22	22	100.0
Wx-A1 (2)	25	12	48.0
Wx-B1 (3)	24	20	83.3
Wild type (4)	21	18	85.7
Total	90	72	80.0

4.4 SEARCH PRE-FILTERS FOR INFRARED LIBRARY SEARCHING FOR CARBOXYLIC ACIDS

There has been renewed interest in infrared (IR) spectral library matching due to the wealth of information in an IR spectrum, improvements in computing power, the higher quality and larger amounts of IR data, and workers who are less skilled in the art of interpreting IR spectra [4-36 and 4-37]. However, a concern in the use of reference spectra for identification is the approach taken by most commercial search algorithms for spectral matching which involves some form of point by point numerical comparison

between the unknown and spectra in the library. These algorithms lack interpretative ability, are sensitive to band shifting, and often ignore bands of low intensity, which can be quite informative.

Many of the problems encountered in IR spectral library matching can be addressed using search prefilters. A search prefilter is a quick test to identify library spectra that are dissimilar to an unknown [4-38 and 4-39]. Prefilters allow for more sophisticated and correspondingly more time-consuming algorithms for spectral library searching since the size of the library is culled down for a specific match. Search prefilters also increase the selectivity of searches by precluding library spectra that do not contain the prescribed functional group. It is well known among workers in the field of IR spectroscopy that the number of false matches obtained between an unknown and reference spectra in a library increases with the size of the library. The likelihood of obtaining these types of matches due to chance is diminished when search prefilters are used to reduce the size of the library for a match.

A two-step procedure is proposed to develop search prefilters. The first step involves preprocessing the IR library spectra by wavelets to enhance subtle but significant features in the spectral library data [4-40 and 4-41]. Wavelet coefficients characteristic of specific functional groups are identified in the second step by the pattern recognition GA [4-42 to 4-50] for separation of spectra by functional group. Search prefilters for carboxylic acids were developed as part of this study to demonstrate the feasibility of the proposed methodology. For liquid samples, carboxylic acids are difficult to distinguish from large libraries of organic materials due to their somewhat indistinct band shapes. They contain elements of carbonyl and hydroxyl functional groups that are

often indistinguishable from noncarboxylic acids. The spectra of carboxylic acids exhibit a strong carbonyl stretching vibration, combined with a broad O-H stretching vibration. The O-H stretching band often appears in the spectrum of other compounds containing O-H groups or mixtures containing such molecules. The presence of both carbonyl and hydroxyl groups is quite common in mixtures of organic compounds that do not contain the organic acid functionality. For these reasons, searching liquid phase compound libraries for carboxylic acids often gives positive identification for organic compound mixtures containing both a broad O-H group and a narrow C=O functionality.

However, a clarification statement is needed here for a comparison of liquid phase versus vapor phase carboxylic acid spectra. Liquid phase carboxylic acids have a broad O-H stretching vibration that is possibly confused with other compounds containing O-H and carbonyl groups. In this study, vapor phase spectra of carboxylic acids where the O-H group is not hydrogen bonded was used. The O-H stretch in vapor phase spectra is relatively narrow and it is unusual to have the carboxylic acid O-H overlapped by the O-H stretching band of the hydroxylic compounds. For this study, some of the problems encountered when searching liquid phase spectral libraries for carboxylic acids were obviated.

Recognition rates for carboxylic acids search prefilters previously reported in the literature for vapor phase spectra have varied from 81% to 92% [4-51 to 4-55]. This is comparable to a scientist somewhat familiar with IR. However, these search prefilters were developed using raw absorbance values from selected spectral regions and did not include any information about band shape and band width. If search prefilters are to be of value to researchers, they must perform better. For this reason, information about

band shape and band width was included in the search prefilters developed in this study by using wavelet coefficients generated from the wavelet packet tree.

The 555 IR spectra used in this study were obtained from the Nicolet vapor phase library. Spectra comprising this library included the EPA gas phase IR collection, Bayerische Julius Maximilian Universitat Wurzburg and from Aldrich using their products as samples. Each IR spectrum was collected in a heated cell or in a light pipe connected to the outlet of a gas chromatograph. Vapor phase spectra were selected for this study because they exhibit somewhat simpler spectral band-shapes than condensed phase spectra. Spectra were originally acquired at 0.5 – 2.0cm^{-1} spectral resolution. All spectra (4000cm^{-1} to 455cm^{-1}) were mathematically deresolved to 8cm^{-1} resolution by apodization of the original interferograms before application of the Fourier Transform during conversion to Omnic Library format. Each IR spectrum contained 460 points. For pattern recognition analysis, each IR spectrum was initially represented as a data vector, $x = (x_1, x_2, x_{3,\ldots}x_j,\ldots x_{460})$ where x_j is the infrared absorbance of the j^{th} point. All IR spectra were normalized to unit length to correct for differences in optical path length.

The IR spectra were divided into a training set of 463 compounds and a validation set of 92 compounds. (See Appendix for the names of the compounds that comprise both the training and prediction set.) Spectra in the validation set were chosen by random lot. The training set (see Table 4-6) consisted of 146 carboxylic acids, 220 noncarbonyls (phosphates, alkenes, alkynes, and alkanes), 24 aldehydes, 25 ketones, 20 esters, 2 anhydrides, 2 acid chlorides, and 24 amides. As for the validations set (see Table 4-7), there were 25 carboxylic acids, 25 noncarbonyls (phosphates, alkenes, alkynes, and alkanes), 2 aldehydes, 12 ketones with 5 containing the OH functionality, 19 esters with 3

containing an OH group, and 9 amides with 2 containing an OH group. Aldehydes, ketones, amides, and esters, as well as compounds that contained both the carbonyl and OH functionality, were selected for this study to make this problem challenging and meaningful. The set of data selected for this study is also representative of the complexity of the modeling problem that must be solved in order to discriminate spectra of carboxylic acids from non-carboxylic acids in real world samples.

Table 4-6. Description of the Training Set

Functional Group	Number of Compounds
Carboxylic acids	146
Phosphates, alkenes, alkynes, and alkanes	220
Aldehydes	24
Ketones	25
Esters	20
Anhydrides and acid chlorides	4
Amides	24
Total Number of Compounds	463

Table 4-7. Description of the Validation Set

Functional Group	Number of Compounds
Carboxylic acids (two contain ester functionality)	25
Phosphates, alkenes, alkynes, and alkanes	25
Aldehydes	2
Ketones (5 contain OH)	12
Esters (3 contain OH)	19
Amides (2 contain OH)	9
Total Number of Compounds	92

The first step in this study was to apply PCA to the normalized raw spectra in the training set. Prior to PCA, the data were auto-scaled to ensure that each wavelength had equal weight in the analysis. Figure 4.14 shows a plot of the two largest principal components of the 463 IR spectra that comprised the training set. Each spectrum is represented as a point in the PC plot (1 = carboxylic acid and 2 = noncarboxylic acid). The overlap between the carboxylic acids and noncarboxylic acids in the PC plot of the data is evident.

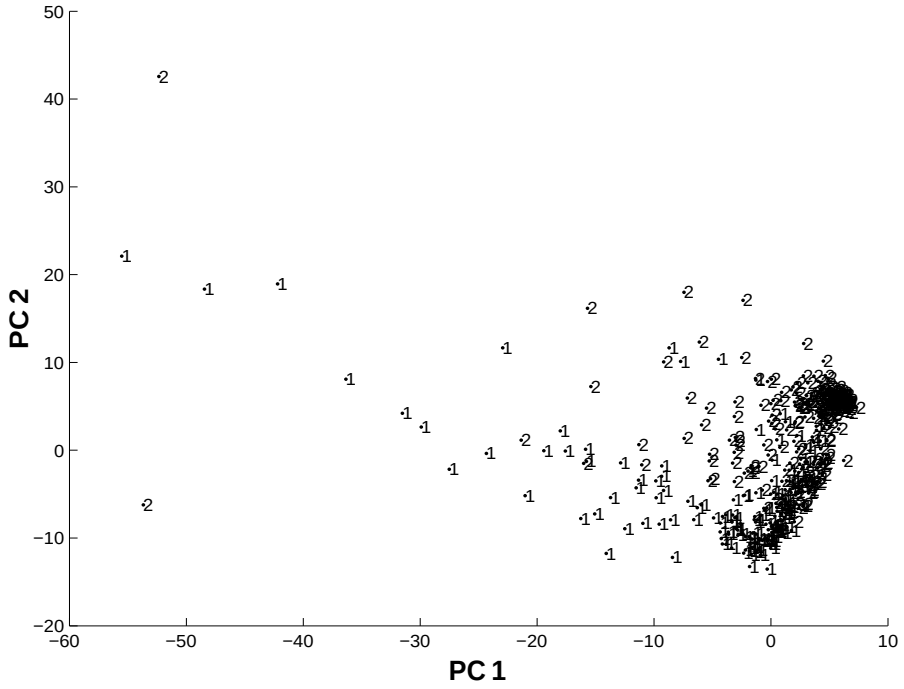

Figure 4.14. Plot of the two largest principal components of the 463 IR spectra that comprised the training set. Each spectrum is represented as a point in the PC plot (1 = carboxylic acid and 2 = noncarboxylic acid).

The pattern recognition GA (PCKaNN fitness function) was used to identify wavelengths characteristic of the IR absorption profile of carboxylic acids. Informative wavelengths were identified by sampling key feature subsets, scoring their PC plots, and tracking those classes and/or spectra that were most difficult to classify. The boosting

routine used this information to steer the population to an optimal solution. After 300 generations, the pattern recognition GA identified 22 wavelengths whose PC plot (see Figure 4.15) showed a limited degree of clustering of the IR spectra on the basis of functional group. The overlap between carboxylic acids and noncarboxylic acids in the PC plot of the data can be explained from an examination of Figure 4.16, which has examples of spectra that comprise the training set. Butyric acid has several characteristic carboxylic acid bands in its IR spectrum, whereas the spectrum of cyclopropanedicarboxylic acid more closely resembles the spectrum of the noncarboxylic acids propionic anhydride and octanoyl chloride.

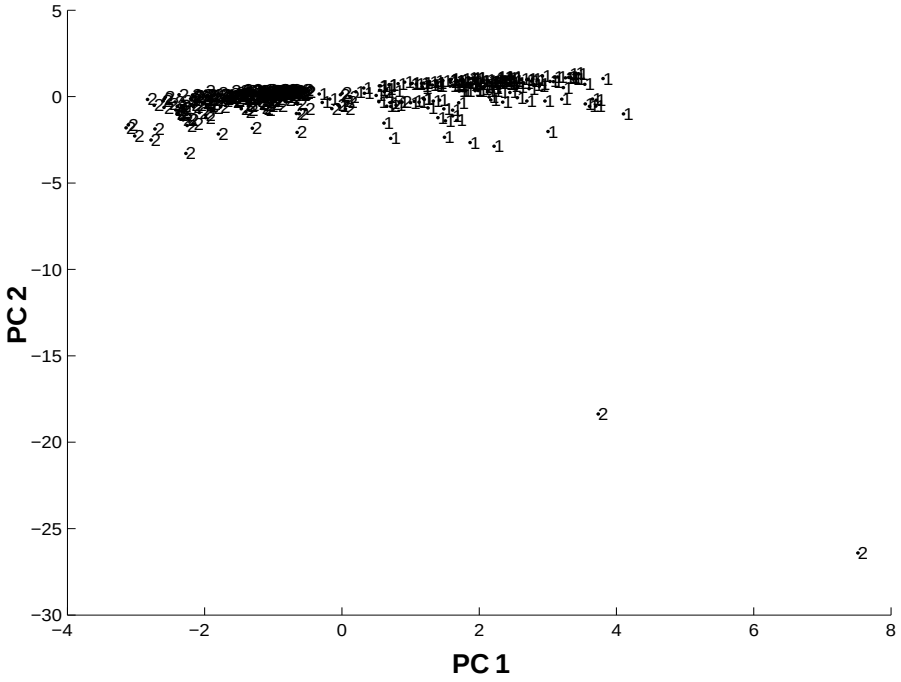

Figure 4.15. Plot of the two largest principal components of the 463 IR spectra that comprised the training set and the 22 wavelengths identified by the pattern recognition GA. Each spectrum is represented as a point in the PC plot (1 = carboxylic acid and 2 = noncarboxylic acid).

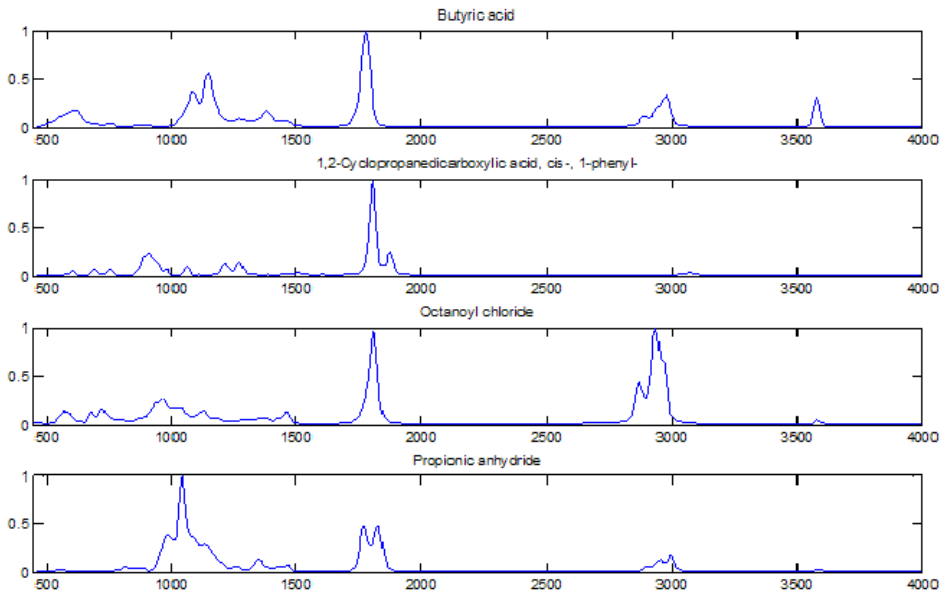

Figure 4.16. IR spectra of butyric acid, 1, 2-cyclopropanedicarboxylic acid *cis*-1-phenyl, octanoyl chloride, and propionic anhydride

The symlet 6 wavelet at the 8th level of decomposition was applied to the IR spectra in the training set to deconvolve overlapping spectral bands and to capture both bandwidth and band-shape information. Figure 4.17 shows a plot of the two largest principal components of the 463 wavelet transformed spectra and the 9398 wavelet coefficients used to represent each IR spectrum. The wavelet transform enhanced the separation of the carboxylic acids and noncarboxylic acids in a PC plot of the data.

The pattern recognition GA was applied to the wavelet transformed spectra to further enhance the separation of the carboxylic acids from the noncarboxylic acids in the PC plot of the training set data. For this problem, the fitness function used consisted of both PCKaNN and the modified Hopkins statistic. This was necessary because of so-called "outliers" present in PC plots of symlet 6 mother wavelet coefficient subsets of the data. Figure 4.18 shows a PC plot of the 463 IR spectra and 41 wavelet coefficients

identified by the pattern recognition GA. Each IR spectrum is represented as a point in the PC plot (1 = carboxylic acid and 2 = noncarboxylic acid). The carboxylic acids are well separated from the noncarboxylic acids in the plot.

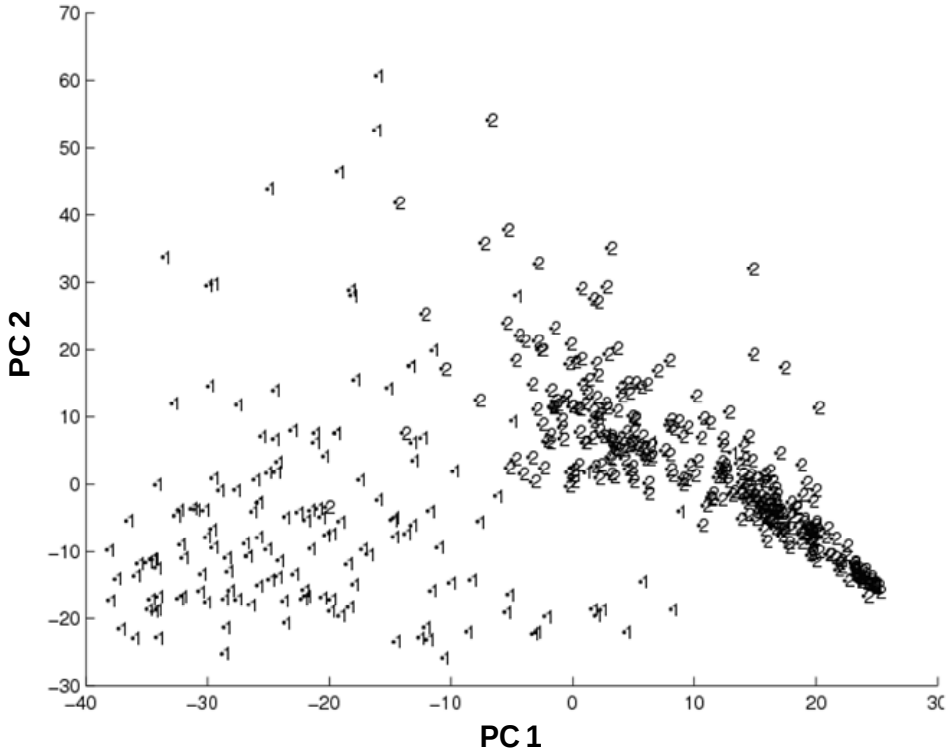

Figure 4.17. Plot of the two largest principal components of the 463 wavelet transformed IR spectra and the 9398 wavelet coefficients that comprised the training set. Each spectrum is represented as a point in the PC plot (1 = carboxylic acid and 2 = noncarboxylic acid).

To assess the predictive ability of the 41 wavelet coefficients identified by the pattern recognition GA, a validation set of 92 IR spectra was used (see Table 4-7). Spectra from the validation set were projected directly onto the PC map developed from the 463 IR spectra and 41 wavelet coefficients identified by the pattern recognition GA. Figure 4.19 shows the projection of the validation set spectra onto the PC map of the training set data. Each projected infrared spectrum lies in a region of the map occupied by IR spectra possessing the same class label.

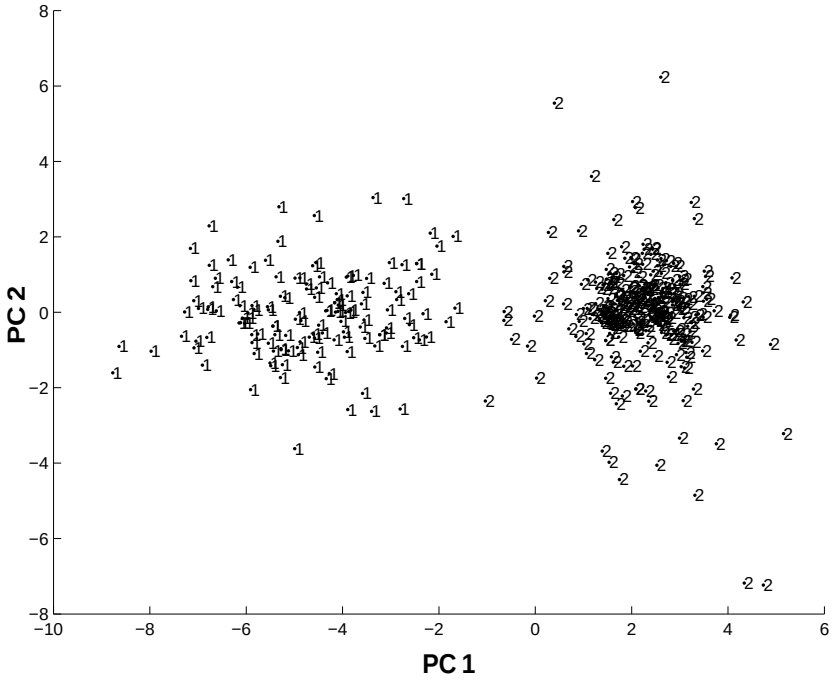

Figure 4.18. Plot of the two largest principal components of the 463 IR spectra and 41 wavelet coefficients identified by the pattern recognition GA. Each IR spectrum is represented as a point in the PC plot (1 = carboxylic acid and 2 = noncarboxylic acid). The carboxylic acids are well separated from the noncarboxylic acids in the plot.

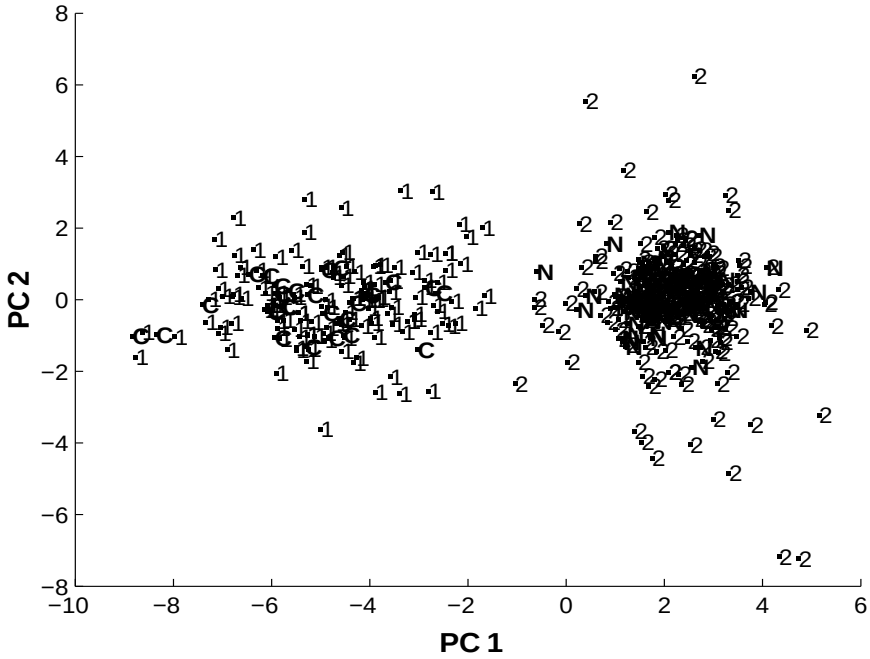

Figure 4.19. Projection of the validation set spectra onto the PC map of the 463 IR spectra and 41 wavelet coefficients identified by the pattern recognition GA. Each projected infrared spectrum lies in a region of the map occupied by spectra possessing the same class label. (1 = carboxylic acid from the training set, 2 = noncarboxylic acid from the training set, C = carboxylic acid from the validation set, and N = noncarboxylic acid from the validations set).

Discriminant analysis was also used to classify the 463 spectra in the training set. The data were divided into two classes: carboxylic acids and noncarboxylic acids. LDA, QDA, RDA, SIMCA, and back propagation neural networks (BPNN) were used to develop classifiers to differentiate carboxylic acids from noncarboxylic acids using the 41 wavelet coefficients identified by the pattern recognition GA as descriptors for these four classification algorithms. Again, a classification success rate of 100% was achieved using each of these classification algorithms.

To further test the predictive ability of these descriptors and the discriminants associated with them, a validation set of 92 IR spectra (see Table 4-7) was employed. Again, a classification success rate of 100% was achieved for the spectra in the validation set using LDA, QDA, RDA, SIMCA, or BPNN. Evidently, the pattern recognition GA can identify wavelet coefficients that are characteristic of the carboxylic acid functional group. This suggests that wavelet analysis coupled to the pattern recognition GA can be used to extract structural information from IR data. The feasibility of the proposed methodology to develop IR search prefilters for spectral library matching (e.g., identification of carboxylic acids) is evident.

4.5 PATTERN RECOGNITION ASSISTED INFRARED LIBRARY SEARCHING FOR PDQ DATABASE

Paint samples are often recovered from collisions where injury or death to a pedestrian and damage to a vehicle have occurred. Studies [4-56 and 4-57] conducted over 30 years ago by the Royal Canadian Mounted Police (RCMP) showed that vehicles can be differentiated by comparing the color, layer sequence and chemical composition of each individual layer in paint. In order to make these comparisons, a comprehensive

database was developed as well as a means of searching and retrieving information from it. Today, the Paint Data Query (PDQ) database contains over 19,000 samples (street samples and factory panels), that correspond to over 72,000 individual paint layers, representing the paint systems used on most domestic and foreign vehicles marketed in North America.

Automotive paint [4-58] consists of several layers: a clear coat over a color coat which in turn is over one or two undercoats (primers). Each paint layer, with the exception of the clear coat, contains pigments and fillers, and all layers contain binders. Automotive manufacturers often use a unique combination of pigments and binders. It is this unique combination that allows forensic scientists to determine the possible make, model, and year of a vehicle from a paint chip left at the scene of a crime.

PDQ is a database of the chemical composition and physical attributes (i.e. color and layer sequence) of each layer of the original manufacturer's paint. The database also contains digital libraries of IR spectra of each of the layers. PDQ is accessed through a general text-based search and retrieval system. Text-based coding of both chemical and physical characteristics serves as a pre-screen for a manual infrared spectral search of materials that tend to be chemically very similar to one another. The capability to directly search FTIR spectra in the database does not exist. Commercial spectral search algorithms cannot distinguish subtle differences between spectra from one vehicle model to the next. The coding used for IR in PDQ is generic, based on functional group recognition, which often leads to non-specific search criteria that can result in a large number of spurious hits that a scientist must work through and eliminate.

Currently, modern automotive paints use thinner undercoats and color coat layers which is protected by a thicker clear coat layer. All too often, only a clear coat paint smear is the only layer of paint left at the crime scene. The text based system of PDQ will not allow for effective searching of clear coats because all modern clear coats applied to any painted automotive parts have only one of two possible formulations: acrylic melamine styrene, or acrylic melamine styrene polyurethane. They contain no inorganic fillers or color with which to further discriminate the sample. In these cases, the motor vehicle cannot be identified using the text based portion of the PDQ database.

The inability to access information about clear coats and search clear coat FTIR spectra is a significant limitation to the current text-based PDQ database. Library searching algorithms incorporated directly into the database have the potential for more specific searches by relying less on subjective text-based characteristics. However, searches of the database's associated IR libraries using commercial software have met with only some success. Automotive paint libraries are composed of a large number of similar spectra. Commercial search algorithms are not able to distinguish subtle but significant features in the data, such as shoulders, unique shapes and patterns, and minor peaks. Since most commercial search algorithms involve some form of numerical comparison between the spectrum of an unknown and each library spectrum [4-59], these algorithms are unable to handle peak shifting and ignore peaks of low intensity, which may be informative [4-60].

By using search prefilters [4-61 and 4-62], many of the problems encountered in spectral searching of IR automotive paint libraries can be addressed. A search prefilter is a quick test to identify library spectra that are dissimilar to an unknown. Prefilters

increase the selectivity of the search as the size of the library is culled down for a specific match using pattern recognition techniques which increases the accuracy of the search. However, it is important that information contained in the search prefilter be based on the relationship between the composition of the clear coat layer and the manufacturer and model of the vehicle. The high quality of FTIR data in the PDQ spectral libraries, and the comprehensiveness of this database, makes it an excellent source of data for the development and validation of search prefilters.

To assess the utility of this approach for spectral library matching, IR spectra of clear coats were collected using a BioRad 40A or BioRad 60 FTIR spectrometer. Each clear coat paint sample, which was between 3µg and 4µg, was run from 2000 to 200 cm^{-1} between diamond windows. IR spectra of paint samples from six Chrysler plants (see Table 4-8) were obtained from the PDQ database. Each plant (BRA, STL, JFN, STH, SAL, and NEW) was represented by at least 10 paint samples obtained from a variety of automobile parts (see Table 4-9). With the exception of the STL plant, all of the paint samples were from the same production year (see Table 4-8). This made the problem more challenging as the paint samples evaluated were all the same make (Chrysler) with a limited production year range. The IR spectra were divided into a training set of 88 spectra (see Table 4-10) and a validation set of 3 spectra (see Table 4-11). Samples from the validation set were chosen by random lot. Further details about the experimental conditions used can be found elsewhere [4-63].

As the overall goal of this study was to differentiate IR spectra of clear coats by manufacturing plant, the initial focus of the pattern recognition analysis was the training set data. The first step in the study was to apply PCA to the autoscaled IR spectral data.

For PCA, each sample was initially represented as a data vector, $x = (x_1, x_2, x_3, ...x_j,...x_{1944})$ where x_j is the absorbance of the j^{th} point from the IR spectrum. In Figure 4.20, a plot of the two largest principal components of the 88 spectra and 1944 features comprising the training set is shown. Each spectrum (i.e., sample) is represented by a point in the plot (1 = BRA, 2 = STL, 3 = JFN, 4 = STH, 5 = SAL, 6 = NEW). The overlap of clear coats from the different manufacturing plants in the principal component of the data is evident.

Table 4-8. Clear Coat Paint Data Set
(Courtesy of *Talanta*, 2011, 87, 46-52

	Plant	Code	Vehicle Model	Number of Spectra
1	Bramalea, Canada	BRA	*Intrepid, Concorde, LHS and 300M (1999)	25
2	**St. Louis, USA	STL	Dodge Ram Trucks (1999-2000) Chrysler/Plymouth SUV's (1999)	21
3	Jefferson North Plant, USA	JFN	Jeep Cherokee (1999)	13
4	Sterling Heights, USA	STH	Dodge Stratus (1999)	9
5	Saltillo, Mexico	SAL	Dodge Ram Trucks (1999)	12
6	Newark, USA	NEW	Durango SUV (1999)	11

*Chrylser Intrepid, Chrysler Concorde, Chrysler LHS and Chrysler 300M are mechanically similar vehicles as they are interrelated models. Both the Concorde and Intrepid are built on the same identical (LH) platform.
**St. Louis plant has two distinct production lines: Dodge Ram Trucks (North Plant) and Chrysler/Plymouth SUV's (South Plant)

Table 4-9. Automobile Parts Used in the Data Set
(Courtesy of *Talanta*, 2011, 87, 46-52)

Part	Number of samples
Roof	68
Hood	9
Fender	10
Door	2
Hatchback	1
Trunk	1

Table 4-10. Training Set
(Courtesy of *Talanta*, 2011, 87, 46-52)

	Plant	Number of Spectra
1	Bramalea (BRA)	23
2	St. Louis (STL)	21
3	Jefferson North Plant (JFN)	13
4	Sterling Heights (STH)	9
5	Saltillo (SAL)	11
6	Newark (NEW)	11

Table 4-11. Validation Set
(Courtesy of *Talanta*, 2011, 87, 46-52)

Sample (PDQ Number)	Manufacturing Plant
M0057OT2	Bramalea (BRA)
W0001OT2	Bramalea (BRA)
P0093OT2	Saltillo (SAL)

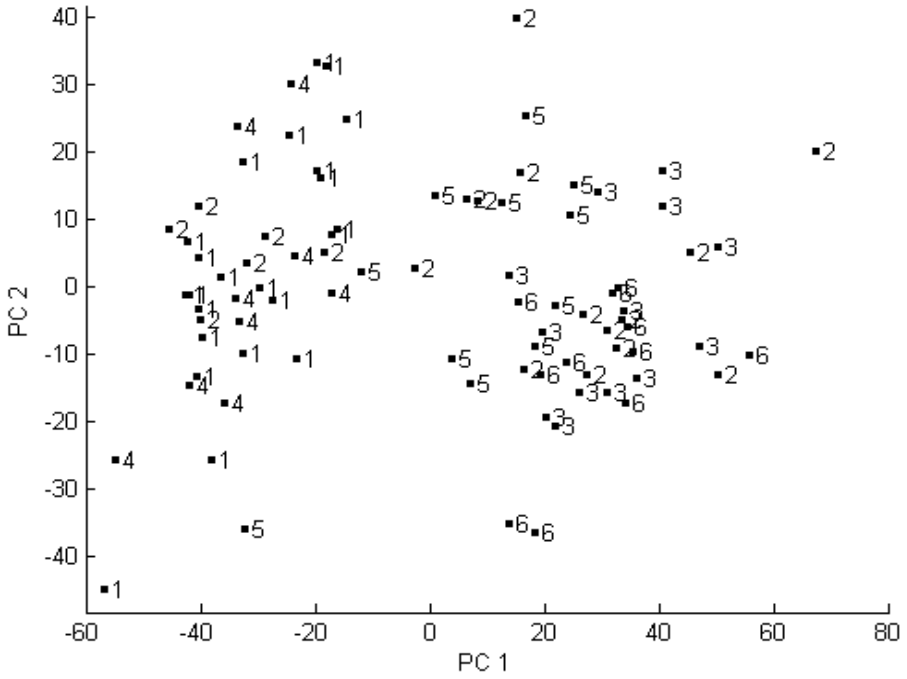

Figure 4.20. Plot of the two largest principal components of the 88 clear coat IR spectra and 1944 points that comprise the training set. Each IR spectrum is represented as a point in the plot (1 = BRA, 2 = STL, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW). (Courtesy of *Talanta***,** 2011, 87, 46-52)

The next step was feature selection. A genetic algorithm for pattern recognition analysis [4-64 to 4-74] was used in this study to identify features from the IR spectra of the clear coats characteristic of the profile of each manufacturing plant. The pattern recognition GA identified features that optimized the separation of the IR spectra by manufacturing plant in a plot of the two or three largest principal components of the data. Because principal components maximize variance, the bulk of the information encoded by these features is about plant identity. A principal component plot that shows separation of the spectra by manufacturing plant can only be generated using features whose variance or information is primarily about differences between the plants. This fitness criterion reduces the size of the search space as it limits the search to these types

of features. In addition, the pattern recognition GA focuses on those classes and/or samples that are difficult to classify by boosting the relative importance of those classes and/or samples that consistently score poorly as it trains. Over time, the algorithm learns its optimal parameters in a manner similar to that of a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one -pass procedure for feature selection and pattern classification.

The pattern recognition GA identified spectral features by sampling key feature subsets, scoring their principal component plots, and tracking clear coat samples or plants that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 300 generations, the GA identified 8 wavelengths whose PC plot showed clustering on the basis of Plant ID (see Figure 4.21). Plant 5 is well separated from the other manufacturing plants, whereas Plants 1 and 4 are separated from each other and Plants 3 and 6 overlap with each other. Plant 2 (STL) appears to be composed of three distinct clusters indicative of three different types of clear coats. Two of these clusters overlap with Plants 1 and 4, and Plants 3 and 6, respectively. The principal component map of these 8 spectral features suggests that information about manufacturing plant is present in the IR spectra of clear coats.

A PC map of the STL clear coat samples was developed to investigate the clustering. Figure 4.22 shows a plot of the two largest principal components of these 21 STL IR spectra and 1944 spectral features. Clustering in three distinct sample groups is again evident, which corresponds to the clustering of STL in the original six-class study. Each cluster has a distinctive IR spectrum (see Figure 4.23). Group A are SUV's (Plymouth Voyager, Dodge Grand Caravan, and Chrysler Town and Country), whereas

Groups B and C are Dodge Ram trucks (1500, 2500, and 3500 for Group B and 1500 and 3500 for Group C). Chrysler had made a change in the clear coat formulation used at the St. Louis North Plant in 2000. Group B falls under the BASF supplied Duraclear II clear coat and has a chemistry of acrylic, melamine, styrene, and polyurethane. Group C falls under DuPont supplied Gen IV AW clear coat and has the chemistry acrylic, melamine, and styrene. One can conclude from an examination of this PC plot that information about model and specific production line can be obtained from an IR spectrum of a clear coat paint smear.

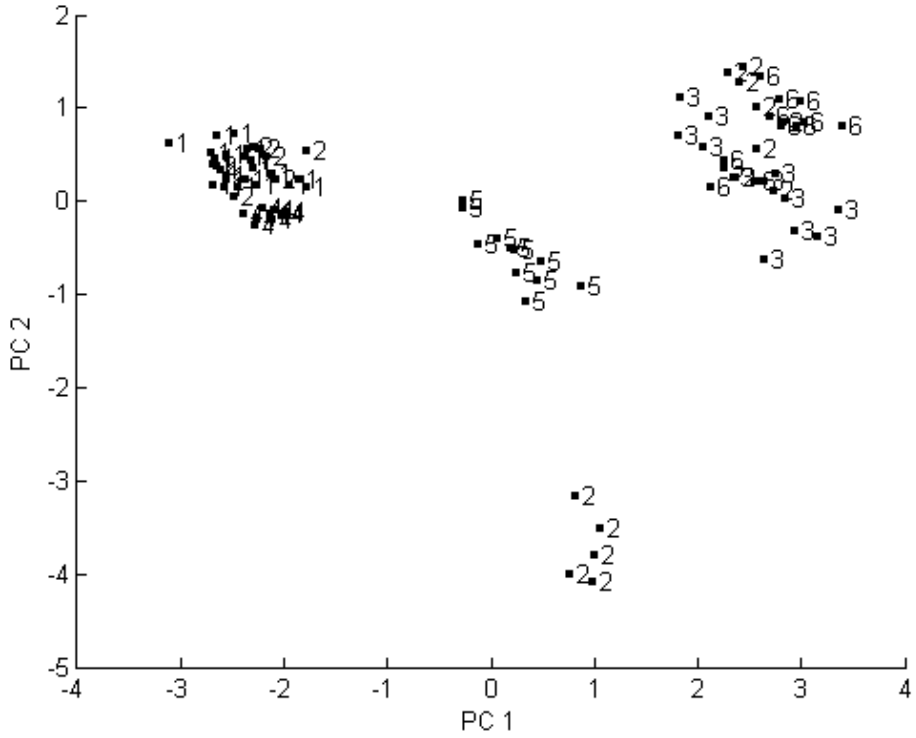

Figure 4.21. Plot of the two largest principal components of the 88 clear coat IR spectra of the training set and 8 wavelengths identified by the pattern recognition GA. Each IR spectrum is represented as a point in the plot (1 = BRA, 2 = STL, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW). (Courtesy of *Talanta*, 2011, 87, 46-52)

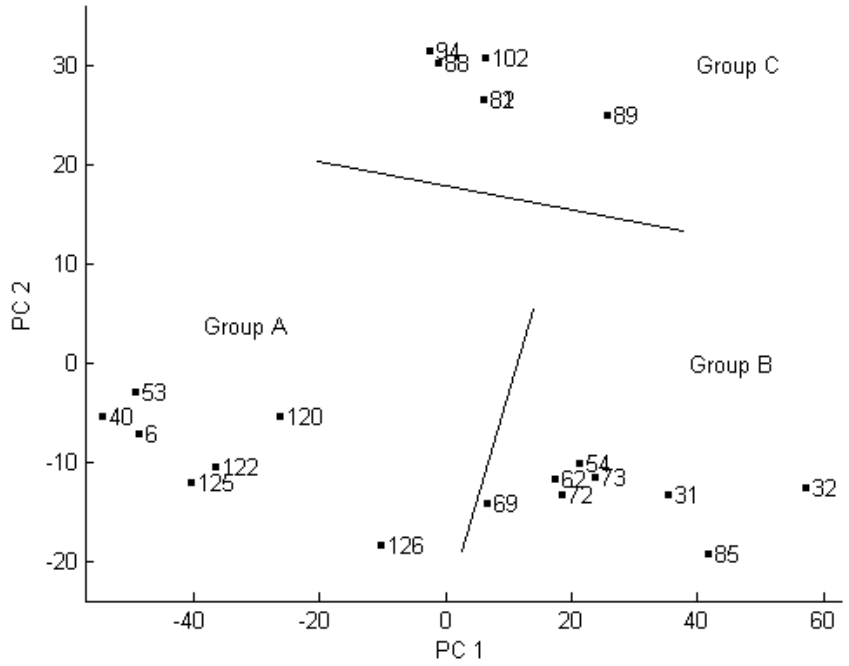

Figure 4.22. Plot of the two largest principal components of 1944 points of the 21 STL clear coat IR spectra. Each IR spectrum is represented by its sample ID in the plot. (Courtesy of *Talanta***,** 2011, 87, 46-52)

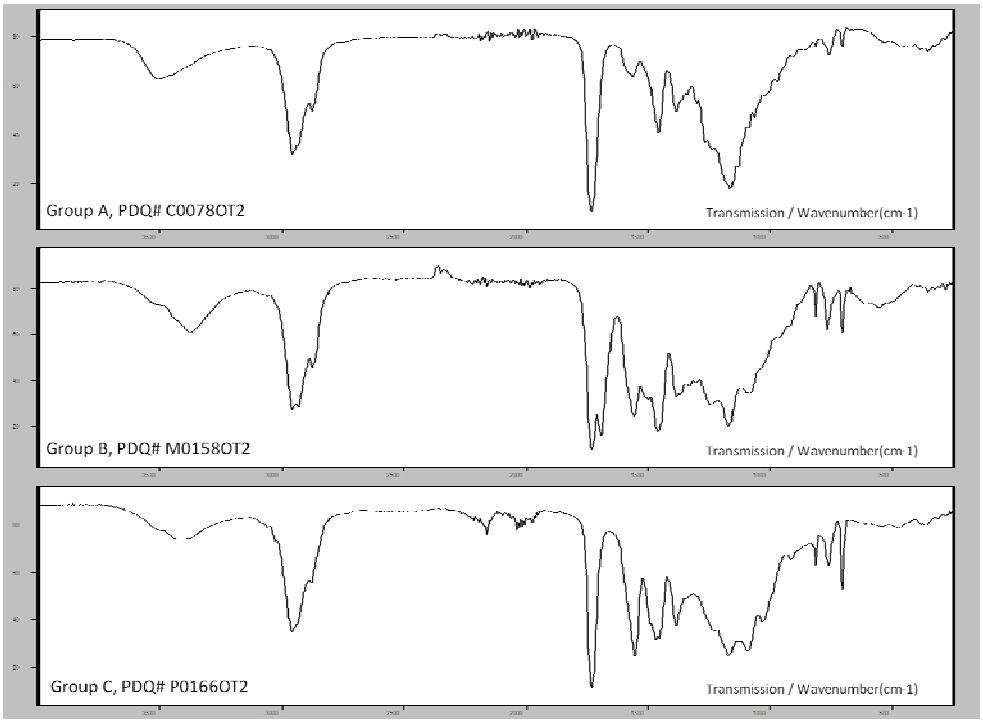

Figure 4.23. Prototypical IR spectrum representative of each STL clear coat cluster. (Courtesy of *Talanta***,** 2011, 87, 46-52)

STL clear coats were removed from the training set and the pattern recognition analysis was again repeated. Figure 4.24 shows a plot of the two largest principal components of the 67 IR spectra and 1944 features. Applying the pattern recognition GA to the data, 10 wavelengths were identified that contained information about the manufacturing plant of these clear coats. Figure 4.25 shows a plot of the two largest principal components developed from the 10 wavelengths selected by the pattern recognition GA. The same trends observed in the PC plot for the larger training set (which contained STL samples) are again reported. Plant 5 is well separated from the other plants, whereas Plants 1 and 4 (BRA and STH) and Plants 3 and 6 (JFN and NEW) overlap.

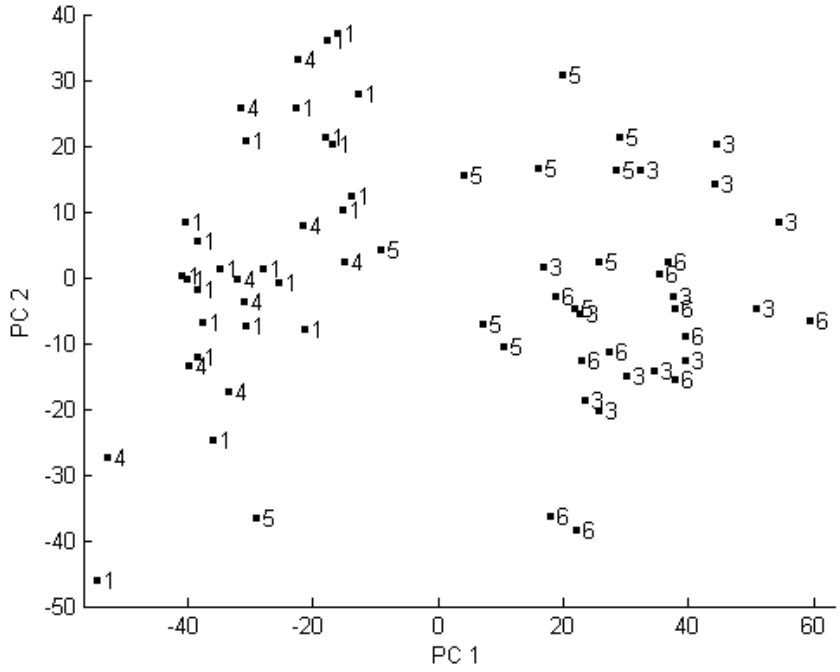

Figure 4.24. Plot of the two largest principal components of the 67 clear coat IR spectra and 1944 points that comprise the training set used for prediction. Each IR spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW). (Courtesy of *Talanta*, 2011, 87, 46-52)

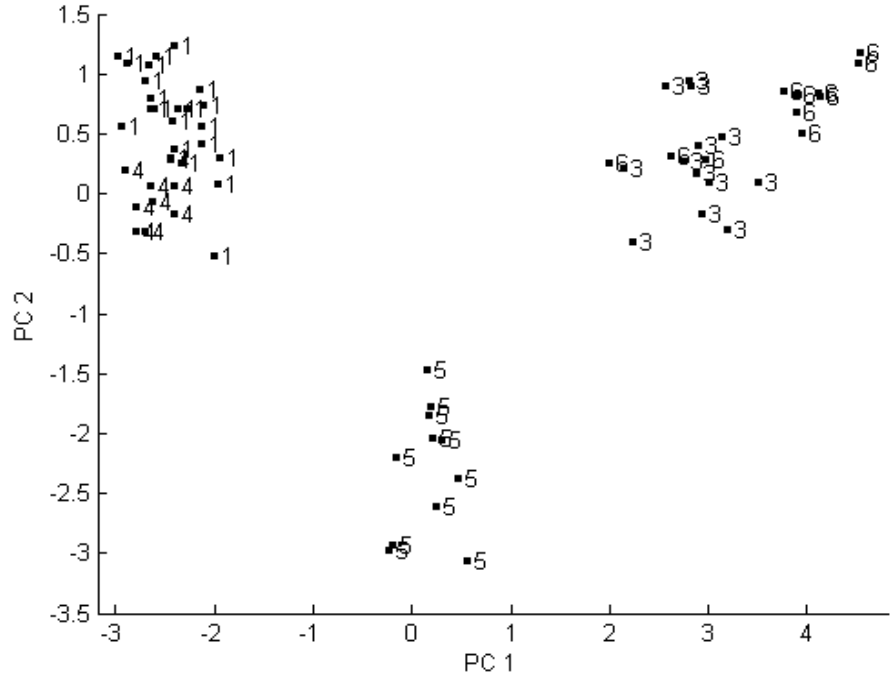

Figure 4.25. Plot of the two largest principal components of the 67 clear coat IR spectra from the training set and 10 wavelengths identified by the pattern recognition GA. Each IR spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW). (Courtesy of *Talanta*, 2011, 87, 46-52)

More powerful preprocessing methods were needed to extract information about manufacturing plant from the IR spectra of these clear coats. For this reason, wavelets were applied to this data. To identify informative wavelet coefficients, it was necessary to use the pattern recognition GA. The Daubechies 12 mother wavelet at the 8th level of decomposition was used to denoise and deconvolute each IR spectrum into 16362 wavelet coefficients. Figure 4.26 shows a plot of the two largest principal components of the wavelet transformed IR spectra of the clear coats. Each IR spectrum was represented by 16362 wavelet coefficients. Because this plot was uninformative, the pattern recognition GA was used to identify informative wavelet coefficients. Using the pattern recognition GA, 36 wavelet coefficients which contained information about the manufacturing plant were identified. Figure 4.27 shows a plot of the two largest principal

components of the 67 clear coat spectra comprising the training set and the 36 wavelet coefficients identified the pattern recognition GA. Separation of IR spectra by manufacturing plant is evident.

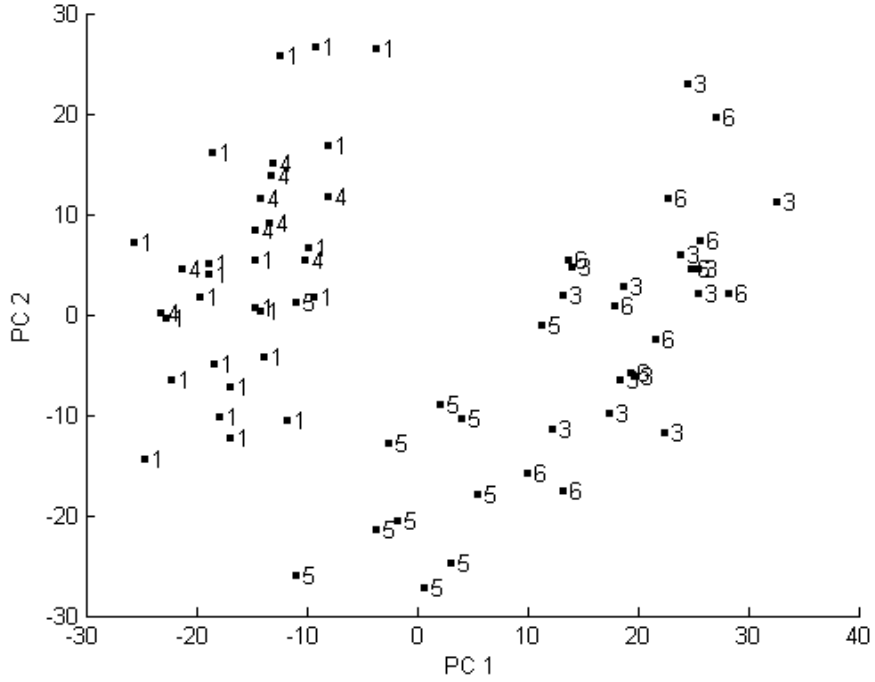

Figure 4.26. Plot of the two largest principal components of the 67 wavelet transformed clear coat IR spectra and 16362 wavelet coefficients that comprise the training set used for prediction. Each IR spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW). (Courtesy of *Talanta*, 2011, 87, 46-52)

A prediction set of 3 spectra was employed (see Table 4-11) to assess the predictive ability of the 36 wavelet coefficients identified by the pattern recognition GA. Figure 4.28 shows the projection of the prediction set samples onto a PC map developed from the 67 clear coat spectra and the 36 wavelet coefficients. Each projected sample lies in a region of the map near a clear coat with the same class label. Evidently, the pattern GA can identify wavelet coefficients characteristic of the manufacturing plant of the clear coat sample. This suggests that IR spectra of clear coats can be used to characterize paint smears by manufacturing plant and production line.

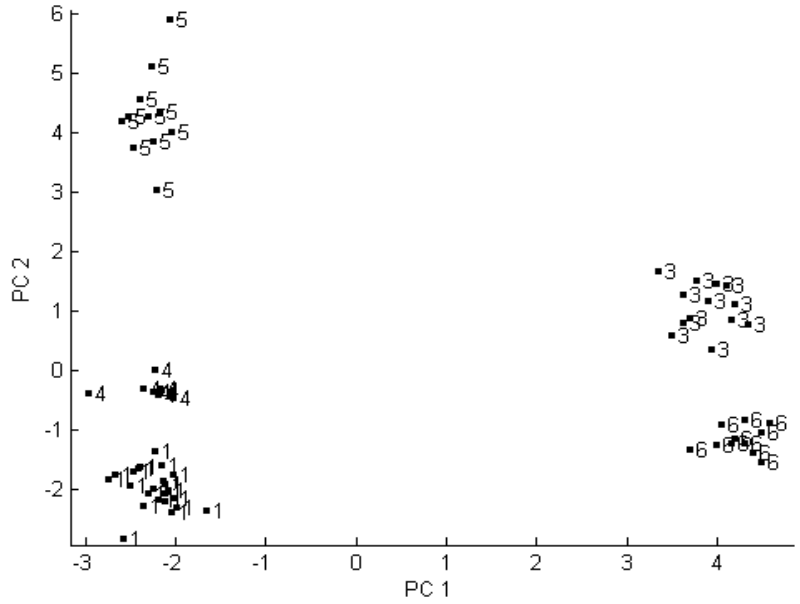

Figure 4.27. Plot of the two largest principal components of the 67 wavelet transformed clear coat IR spectra from the training set and 36 wavelet coefficients identified by the pattern recognition GA. Each IR spectrum is represented as a point in the plot (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW). (Courtesy of *Talanta*, 2011, 87, 46-52)

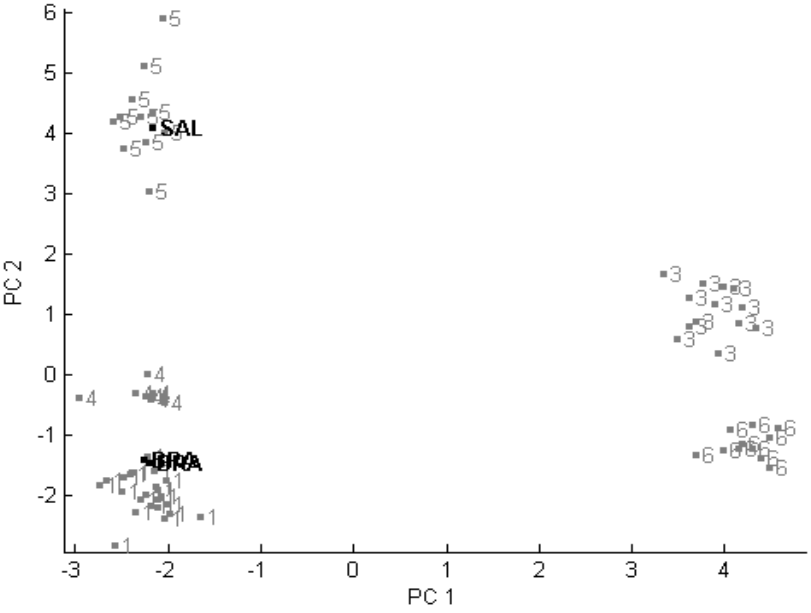

Figure 4.28. Projection of the prediction set samples onto the PC plot of the 67 wavelet transformed IR spectra and 36 wavelet coefficients identified by the pattern recognition GA. Each IR spectrum in the training set (1 = BRA, 3 = JFN, 4 = STH, 5 = SAL, and 6 = NEW) and prediction set (BRA and SAL) is represented as a point in the plot. All projected samples lie in a region of the map near clear coats with the same class label. (Courtesy of *Talanta*, 2011, 87, 46-52)

REFERENCES

4-1. G.A. Eiceman, M. Wang, S. Prasad, H. Schmidt, F.K. Tadjimukhamedov, B. K. Lavine, N. Mirjankar, "Pattern recognition analysis of differential mobility spectra with classification by chemical family", Analytica Chimica Acta, 2006, 579, 1–10.

4-2. S.R. Delwiche and R.A. Graybosch, "Identification of waxy wheat by near-infrared reflectance spectroscopy", J. Cereal Sci., 2002, 35, 29–38.

4-3. B.K. Lavine, K. Nuguru, N. Mirjankar, J. Workman Jr., "Pattern recognition assisted infrared library searching", App. Spectr., 2012, 66-8.

4-4. B.K. Lavine, N. Mirjankar, S. Ryland, M. Sandercock, "Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics", Talanta, 2011, 87, 46–52.

4-5. R.A. Miller, G.A. Eiceman, E.G. Nazarov, A.T. King, "A novel micromachined high-field asymmetric waveform-ion mobility spectrometer", Sens. Actuators B Chem., 2000, 67, 300–306.

4-6. G.A. Eiceman, E.V. Krylov, B. Tadjikov, R.G. Ewing, E.G. Nazarov, R.A. Miller, "Differential mobility spectrometry of chlorocarbons with a micro-fabricated drift tube", Analyst, 2004, 129, 297–304.

4-7. G.A. Eiceman, B. Tadjikov, E. Krylov, E.G. Nazarov, R.A. Miller, J. Westbrook, P. Funk, "Miniature radio-frequency mobility analyzer as a gas chromatographic detector for oxygen-containing volatile organic compounds, pheromones and other insect attractants", J. Chromatogr., 2001, 917, 205–217.

4-8. "EGISTM Defender Portable Lightweight Desktop Explosives Trace Detection System" ©2005 Thermo Electron Corporation, Part Number: 41841800-Revision A., available at www.sionex.com, 2006.

4-9. R. Guevremont, "High-field asymmetric waveform ion mobility spectrometry: a new tool for mass spectrometry", J. Chromatogr. 2004, 1058, 3–19.

4-10. K. Venne, E. Bonneil, K. Eng, P. Thibault, "Improvement in peptide detection for proteomics analyses using NanoLC-MS and high-field asymmetry waveform ion mobility mass spectrometry", Anal. Chem., 2005, 77, 2176–2186.

4-11. G.A. Eiceman, E.G. Nazarov, J.E. Rodriguez, “Chemical class information in ion mobility spectra at low and elevated temperatures”, Anal. Chim. Acta, 2001, 433, 53–70.

4-12. S.E. Bell, E.G. Nazarov, Y.F. Wang, G.A. Eiceman, “Classification of ion mobility spectra by functional groups using neural networks”, Anal. Chim. Acta, 1999, 394, 121–133.

4-13. C.A. Veasey, C.L.P. Thomas, “Fast quantitative characterization of differential mobility responses”, Analyst, 2004, 129, 198–204.

4-14. M.D. Wessel, J.M. Sutter, P.C. Jurs, “Prediction of reduced ion mobility constants of organic compounds from molecular structure”, Anal. Chem., 1996, 68, 4237–4243.

4-15. E. Krylov, E.G. Nazarov, R.A. Miller, B. Tadjikov, G.A. Eiceman, “Field dependence of mobilities for gas-phase-protonated monomers and proton-bound dimers of ketones by planar field asymmetric waveform ion mobility spectrometer (PFAIMS)”, J. Phys. Chem., 2002, 106, 5437–5444.

4-16. N. Krylova, E. Krylov, G.A. Eiceman, J.A. Stone, ”Effect of moisture on the field dependence of mobility for gas-phase ions of organophosphorus compounds at atmospheric pressure with field asymmetric ion mobility spectrometry”, J. Phys. Chem., 2003, 107, 3648–3654.

4-17. G.A. Eiceman, E.V. Krylov, N.S. Krylova, E.G. Nazarov, R.A. Miller, “Separation of ions from explosives in differential mobility spectrometry by vapor-modified drift gas”, Anal. Chem., 2004, 76, 4937–4944.

4-18. F. Chau, Y. Liang, J. Gao, X. Shao, “Chemometrics – From Basics to Wavelet Transform, John Wiley & Sons, NY, 2004.

4-19. G. Chen, P.B. Harrington, “Real-time two-dimensional wavelet compression and its application to real-time modeling of ion mobility data”, Anal. Chim. Acta, 2003, 490, 59-69.

4-20. B. Walczak, D.L. Massart, “Wavelet packet transform applied to a set of signals: A new approach to the best-basis selection”, Chemomet. Intell. Lab. Syst., 1997, 36, 81-94.

4-21. J. Karasinski, S. Andreescu, O.A. Sadik, B. Lavine, M.N. Vora, “Multiarray sensors with pattern recognition for the detection, classification, and differentiation of bacteria at subspecies and strain levels”, Anal. Chem. 2005, 77, 7941-7949.

4-22. B.K. Lavine, C.E. Davidson, W.T. Rayens, “Machine learning based pattern recognition applied to microarray data”, Comb. Chem. High Throughput Screening, 2004, 7, 115-131.

4-23. B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, M.C. Sundling, “Electronic van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases”, J. Chem. Inf. Comp. Sci., 2003, 43, 1890-1905.

4-24. B.K. Lavine, C.E. Davidson, R.K. Vander Meer, S. Lahav, V. Soroker, A. Hefetz, “Genetic algorithms for deciphering the complex chemosensory code of social insects”, Chem. Intelligent Lab. Instr., 2003, 66, 51–62.

4-25. B.K. Lavine, C.E. Davidson, A.J. Moores, “Genetic algorithms for spectral pattern recognition”, Vib. Spec., 2002, 28, 83–95.

4-26. R.A. Graybosch, “Waxy wheats: origin, properties, and prospects”. Trends in Food Science and Technology, 1998, 9, 135-142.

4-27. I. Reddy, and P.A. Seib, “Modified waxy wheat starch compared to modified waxy corn starch”. J. Cereal Sci., 2000, 31, 25-29.

4-28. T. Hoshino, R. Yoshikawa, S. Ito, K. Hatta, T. Nakamura, M. Yamamori, K. Hayakawa, K. Tanaka, H. Akashi, S. Endo, S. Tago, and S. Ishigami, “Flour blends for breads, cakes, or noodles, and foods prepared from the flour blends”, United States Patent, 2000, 6, 042, 867.

4-29. J. Epstein, C.F. Morris, and K.C. Huber, “Instrumental texture of white salted noodles prepared from recombinant inbred lines of wheat differing in the three granule bound starch synthase (waxy) genes”, J. Cereal Sci., 2002, 35, 51-63.

4-30. T. Nakamura, and M. Yamamori, “Production of waxy (amylose-free) wheats”. Mol. Gen. Genet., 1995, 248, 253-259.

4-31. R.A. Graybosch, C.J. Peterson, L.E. Hansen, S. Rahman, A. Hill, and J.H. Skerritt, “Identification and characterization of U.S. wheats carrying null alleles at the wx loci”. Cereal Chem., 1998, 75, 162-165.

4-32. X.C. Zhao and P.J. Sharp, “An improved 1-D SDS-PAGE method for the identification of three bread wheat ‘waxy’ proteins”, J. Cereal Sci., 1996, 23, 191-193.

4-33. S. Rahman, B. Kosarhashemi, M.S. Samuel, A. Hill, D.C. Abbott, J.H. Skerritt, J. Preiss, R. Appels, and M.K. Morell, “The major proteins of wheat endosperm starch granules”, Aust. J. Plant Physiol., 1995, 22, 793-803.

4-34. T. Nakamura, P. Vrinten, M. Saito, and M. Konda, “Rapid classification of partial waxy wheats using PCR-based markers”, Genome, 2002, 45, 1150-1156.

4-35. S.R. Delwiche, R.A. Graybosch, L.E. Hansen, E. Souza, F.E. Dowell, “Single kernel near-infrared analysis of tetraploid (durum) wheat for classification of the waxy Condition”, Cereal Chemistry, 2006, 83, 287-292.

4-36. R. Karoui, G. Downey, and C. Blecker. “Mid-Infrared Spectroscopy Coupled with Chemometrics: A Tool for the Analysis of Intact Food Systems and the Exploration of the Their Molecular Structure-Quality Relationships – A Review”. Chem. Rev., 2010, 110, 6144-6168.

4-37. P.R. Griffiths, “Fourier Transform Infrared Spectrometry”, Wiley Interscience, New York, 2nd ed., 2007.

4-38. S.R. Lowry, D.A. Huppler, and C.R.J. Anderson, “Database development and search algorithms for automated infrared spectral identification”. J. Chem. Inf. Comput. Sci., 2010, 25, 235-241.

4-39. C.P. Wang, and T.L. Isenhour, “Infrared Library Search on Principal Component-Analyzed Fourier Transform Absorption Spectra”, Appl. Spec., 1987, 41, 185-195.

4-40. J.S. Walter. A Primer on Wavelets and their Scientific Applications, New York: Chapman & Hall/CRC, 1999.

4-41. B.B. Hubbard, “The World According to Wavelets”, Natick, MA: A. K. Peters, 2nd ed., 1998.

4-42. B.K. Lavine, and A.J. Moores. “Genetic algorithms for pattern recognition analysis and fusion of sensor data,” in K. Siddiqui and D. Eastwood editors., Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, Proceedings of SPIES. Bellingham, WA: SPIE, 1999, 103-112.

4-43. B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, and H.T. Mayfield. “Source identification of underground fuel spills by solid phase micro-extraction/high-resolution gas chromatography/genetic algorithms”, Anal. Chem., 2000, 72, 423-431.

4-44. B.K. Lavine, C. E. Davidson, A.J. Moores, and P. R. Griffiths. “Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types”. Appl. Spec., 2001, 55, 960-966.

4-45. B. K. Lavine, C. E. Davidson, and A.J. Moores, “Innovative genetic algorithms for chemoinformatics”, Chemom. Intell. Lab. Instr., 2002, 60, 161-171.

4-46. B.K. Lavine, C.E. Davidson and A.J. Moores, “Genetic algorithms for spectral pattern recognition”, Vibrat. Spec., 2002, 28, 83-95.

4-47. A. Karasinski, S. Andreescu, O.A. Sadik, B. Lavine, and M.N. Vora, “Multiarray sensors with pattern recognition for the detection, classification, and differentiation of bacteria at subspecies and strain levels”. Anal. Chem., 2005, 77, 7941- 7949.

4-48. G.A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F.K. Tadjimukhamedov, B.K. Lavine, and N. Mirjankar. "Pattern recognition analysis of differential mobility spectra with classification by chemical family", Anal. Chim. Acta, 2006, 579, 1-10.

4-49. J. Karasinski, L. White, Y. Zhang, E. Wang, S. Andreescu, O.A. Sadik, B. Lavine, and M.N. Vora. "Detection and identification of bacteria using antibiotic susceptibility and a multi-array electrochemical sensor with pattern recognition", Biosensors and Bioelectronics, 2007, 22, 2643-2649.

4-50. B.K. Lavine and C.E. Davidson, "Multivariate Approaches to Classification Using Genetic Algorithms," in: S. Brown R. Tauler R, and R. Walczak R editors. Comprehensive Chemometrics 3, Amsterdam, The Netherlands: Oxford- Elsevier 2009, 619-646.

4-51. L. Damokos, I. Frank, G. Matolcsy, and G. Jalsovszky, "Pattern recognition applied to vapor phase infrared spectra", Anal. Chim. Acta, 1983, 154, 181-189.

4-52. D.S. Frankel. "Pattern recognition of Fourier transform infrared spectra of organic compounds", Anal. Chem., 1984, 56, 1011-1014.

4-53. H.B. Woodruff, S. R. Lowry, and T. L. Isenhour, "Comparison of two Discriminant Functions for Classifying Binary Infrared Data", Appl. Spec., 1975, 29, 226-230.

4-54. H.B. Woodruff, "Novel Advances in Pattern Recognition and Knowledge-Based Methods in Infrared Spectroscopy," in L.C. Meuzelaar, and T.L. Isenhour editors, Computer-Enhanced Analytical Spectroscopy I, Plenum Press, New York, 1987, 201-219.

4-55. R.J. Anderegg and D.J. Pyo, "Selective reduction of infrared data." Anal. Chem., 1987, 59, 1914-1917.

4-56. J.L. Buckle, D.A. MacDougal, and R.R. Grant, "PDQ-paint data queries: the history and technology behind the development of the royal Canadian mounted police forensic science laboratory services automotive paint database", Can. Soc. Forens. Sci. J., 1997, 30, 199-212.

4-57. N.S. Cartwright and P.G. Rodgers, "A proposed data base for the identification of automotive paint," Can. Soc. Forens. Sci. J., 1976, 9(4), 145-154.

4-58. G. Fettis, (Editor), "Automotive Paints and Coatings", VCH Publications, New York, 1995.

4-59. S.R. Lowry, D.A. Huppler, and C.R. Anderson, "Database development and search algorithms for automated infrared spectral identification," J. Chem. Inf. Computer Sci., 1985, 25, 235-241.

4-60. J.C.W. Bink and H.A. Van'T Klooster, "Classification of organic compounds by infrared spectroscopy with pattern recognition and information theory," Anal. Chim. Acta, 1983, 150, 53-59.

4-61. G.W. Small, "Automated spectral interpretation," Anal. Chem., 1987, 59, 535-546.

4-62. C.P. Wang and T.I. Isenhour, "Infrared library search on principal component-analyzed Fourier transformed absorption spectra," Appl. Spec., 1987, 41, 185-194.

4-63. A. Beveridge, T. Fung, and D. MacDougall, "Use of Infrared Spectroscopy for the Characterization of Paint Fragments," In: Forensic Examination of Glass and Paint Analysis and Interpretation, B. Caddy, ED., Taylor and Francis, NY, 2001, 220-233.

4-64. B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, and H.T. Mayfield, "Source identification of underground fuel spills by solid phase micro-extraction/high-resolution gas chromatography/genetic algorithms," Anal. Chem., 2000, 72, 423-431.

4-65. B.K. Lavine, D. Brzozowski, J. Ritter, A.J. Moores, and H.T. Mayfield, "Fuel spill identification by selective fractionation prior to gas chromatography I. Water soluble components," J. Chromat. Sci., 2001, 39, 501-506.

4-66. B.K. Lavine, D. Brzozowski, A.J. Moores, C.E. Davidson, and H.T. Mayfield, "Genetic algorithm for fuel spill identification," Anal. Chim. Acta, 2001, 437, 233-246.

4-67. B.K. Lavine, C.E. Davidson, A.J. Moores, and P.R. Griffiths, "Raman Spectroscopy and Genetic Algorithms for the Classification of Wood Types," Applied Spectroscopy, 2001, 55, 960-966.

4-68. B.K. Lavine, A. Vesanen, D.M. Brzozowski, and H.T. Mayfield, "Authentication of fuel standards using gas chromatography/pattern recognition techniques", Anal Letters, 2001, 34, 281- 294.

4-69. B.K. Lavine, C.E. Davidson, and A.J. Moores, "Innovative genetic algorithms for chemoinformatics", Chemom. Intell. Lab. Instr., 2002, 60, 161-171.

4-70. B.K. Lavine, C.E. Davidson, and A.J. Moores, "Genetic algorithms for spectral pattern recognition," Vibrat. Spec., 2002, 28, 83-95.

4-71. B.K. Lavine, C.E. Davidson, C. Breneman, and W. Katt, "Electronic Van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases," J. Chem. Inf. Science, 2003, 43, 1890-1905.

4-72. B. K. Lavine, C. E. Davidson, and D. J. Westover, "Spectral Pattern Recognition Using Self Organizing Maps," J. Chem. Inf. Comp. Science, 2004, 44, 1056-1064.

4-73. G.A. Eiceman, M. Wang, S. Pradad, H. Schmidt, F. K. Tadjimukhamedov, B.K. Lavine, and N. Mirjankar, "Pattern recognition analysis of differential mobility spectra with classification by chemical family," Anal. Chim. Acta, 2006, 579, 1-10.

4-74. J. Karasinski, L. White, Y. Zhang, E. Wang, S. Andreescu, O.A Sadik, B.K. Lavine, and M.N. Vora, "Detection and identification of bacteria using antibiotic susceptibility and a multi-array electrochemical sensor with pattern recognition," Biosensors & Bioelectronics, 2007, 22, 2643-2649.

CHAPTER V

DISCOVERY OF BIOMARKER CANDIDATES USING THE GENETIC ALGORITHM FOR PATTERN RECOGNITION ANALYSIS

5.1 INTRODUCTION

Biomarkers are compounds that are used as indicators of biological states in a wide range of applications which include medicine, cell biology, genetics, geology and astrobiology. In the context of medicine, a biomarker can serve as an indicator of the presence of disease in a subject. For example, changes in the expression levels of particular proteins in serum have been correlated with the progression of cancer [5-1 to 5-3]. Biomarkers can also indicate the presence of harmful micro-organisms in indoor environments. The absence of clinically proven biomarkers has limited efforts to improve detection of cancer and pathogenic micro-organisms.

During the past decade, significant progress in air sampling, micro-array technology, and mass spectrometry has offered the possibility of using genetic, proteomic, or chemical fingerprinting to predict the progression of cancer in patients or the presence of harmful micro-organisms in homes or office buildings. Recent advances in air sampling technology, e.g., the development of an integrated chemical and microbiological approach to characterize the fungal load of a contaminated area in a

building, have the potential to discriminate mold contamination from mold free areas in buildings in an economic and timely manner [5-4]. Advances in microarray technology during the past decade offer the opportunity to perform comparative transcriptional profiling as large amounts of genetic and proteomic expression data is routinely generated [5-5 to 5-7]. The development of new mass spectrometric based techniques, e.g., techniques for ionization of proteins and peptides such as matrix-assisted laser desorption ionization (MALDI) and electrospray ionization combined with time of flight mass spectrometry and new hybrid mass spectrometers, are becoming routine tools for protein characterization [5-8 to 5-10]. The advantage of using these new methods to facilitate the discovery of clinically relevant biomarker candidates is demonstrated in several studies that are the subject of this chapter.

In the first study [5-11], a set of volatile organic compound (VOC) profiles were developed with corresponding bioaerosol measurements as input-output pairs for a discriminant to predict mold exposure in indoor environments. A novel air sampling methodology was used to collect whole air grab samples while viable spores were collected concurrently using an Andersen impactor in conjunction with malt extract agar and dichloran glycerol 18. By comparing the bioaerosol data to VOC profiles obtained using a GC/MS equipped with a cold trap preconcentrator, a discriminant was developed to classify a residence as to potential mold growth based on its microbial VOC profile. The pattern recognition GA was used to identify features in the GC/MS profile of the VOCs correlated with spore count.

In the second study [5-12], biopsy material of small round blue cell tumors analyzed by cDNA microarrays was identified as to type (Ewings sarcoma, Burkitt's

lymphoma, neuroblastoma, and rhabdomyo sarcoma through supervised learning implemented using the pattern recognition GA. Protein profiling of serum and tissue extracts using mass spectrometry is a popular method to detect biomarker patterns in cancer research. As each sample is very complex and is described by thousands of features, an important issue is the methodology used to extract the information present in these mass spectral profiles that distinguish disease from the controls. The remaining five studies in this chapter focus on using the pattern recognition GA to simplify complex proteomic data and to facilitate the visualization of data through identification of the mass spectral features related to the disease state of the subjects.

5.2 PREDICTION OF MOLD CONTAMINATION FROM MICROBIAL VOC PROFILES

The goal of this study is to determine the degree of mold infestation in indoor environments. Mold contamination in indoor environments has become a public safety concern. Mold infestations have closed schools, condemned houses, caused lost revenue in industrial settings, rendered crops unfit for human consumption, and caused allergic reactions in sensitive populations [5-13 to 5-21]. Current mold sampling techniques are not effective at elucidating the fungal load in a contaminated area nor are they economical and timely in providing results. Therefore, there is interest in developing new sampling and analytical approaches to detect molds in homes and buildings.

A new analytical methodology to characterize the fungal load of a contaminated area in a building has been developed as part of this study using an integrated chemical and microbiological approach. A set of VOC profiles were developed with corresponding bioaerosol measurements as input-output pairs for a discriminant to predict the presence

or absence of mold contamination in indoor environments. The VOC signatures that molds emit as reflected by the gas chromatographic profiles were compared to impactor data collected from each sampling site. By comparing the bioaerosol data to VOC profiles, a discriminant was trained to classify a residence as to potential mold growth based on its microbial volatile organic compound (MVOC) profile.

VOC air sampling was performed using Entech Bottle-Vacs [5-22]. Sampling volatile organics in air using evacuated containers, which can be shipped out for analysis, has been shown to be a viable low-cost method that can be performed in seconds by building occupants or homeowners [5-23]. The air samples were analyzed using GC/MS equipped with a cold trap preconcentrator.

Bioaerosol and chemical sampling data were collected concurrently at 10 locations in Northern New York during 17 sampling periods from July 2006 to August 2007 with the majority of samples collected during the summer. Because cooking and cleaning activities can produce considerable chemical interference, occupants at these locations agreed to abstain from these activities for 12 hours prior to sampling and during sampling events. Whole-air grab samples for MVOCs were collected using Entech Bottle-Vacs. The bottles were checked by an analogue pressure gauge prior to sampling to ensure that a proper vacuum was maintained. Each sample was analyzed by an Agilent 6890/5973N GC/MS (Agilent Technologies, Santa Clara, CA) equipped with a 7500 Autosampler attached to a 7100A Extended Cold Trap Dehydration Preconcentrator (Entech Instruments, Inc, Simi Valley, CA). Separation of the MVOC mixture was performed by a DB-1 column (60 meters by 0.32 mm ID with a film thickness of 1 μm). Calibration standards used in the analysis included bromochloromethane, 1, 4-

difluorobenzene and deuterated chlorobenzene. Relative response factors were calculated for several MVOCs. All samples were analyzed in both SIM and Scan mode. Eighteen chemicals were selected as representative MVOCs (see Table 5-1) known to be emitted by molds during metabolic activity. All concentrations determined were blank corrected.

Bioaerosol data was collected using an Anderson N6 Impactor [5-24] in conjunction with malt extract agar (MEA) and dichloran glycerol (DG18) in Petri dishes to obtain viable mold samples. Since MEA is a mesophilic agar whereas DG18 is a xerophilic agar, a broader range of fungi can be cultured giving a better representation of the fungal ecology by using both types of agar. During each sampling event, 6 samples of each agar type were collected with an associated field blank. The samples were cultured for 6 days with colony counts blank corrected with a positive-hole correction applied [5-25]. The mold count values, which were expressed as a ratio using the field blank, were divided into three categories: low (less than 1.2), medium (1.2 to 3.0) and high (greater than 3.0). These values were used to assign the MVOC gas chromatographic (GC) profiles with the appropriate class label.

Table 5-1. MVOC Compounds

2-Methylfuran	1-Pentanol
2-Butanone	2-Hexanone
3-Methylfuran	2-Heptanone
2-Methyl-1-propanol	1-Octen-3-ol
3-Methyl-2-butanol	3-Octanone
2-Pentanol	2-Pentylfuran
1, 4 Dioxane	3-Octanol
3-Methyl-1-butanol	2-Ethyl-1-hexanol
2-Methyl-1-butanol	1-Octanol

For pattern recognition analysis, each gas chromatogram was represented as a data vector $X = (x_1, x_2, x_3 \ldots x_{18})$ where the components of the data vector are the

concentrations of the VOCs identified by GC/MS. All profiles were normalized to constant sum and the data were autoscaled to ensure that each compound had equal weight in the analysis. Because of the preprocessing methods used, the focus of the pattern recognition analysis is the **concentration pattern present in the GC profiles**, not the total amount of VOCs captured in the whole air grab samples by the Entech Bottle-Vacs.

Each MVOC profile (air sample) was assigned two class labels as viable mold samples were collected using two different types of agar. One label was based on spore counts from DG18 agar and the other on spore counts from MEA agar. The two bioaerosol data sets (which are summarized in Table 5-2) have 20 MVOC profiles in common for low mold count exposure, 5 for medium mold count exposure, and 29 for high mold count exposure. The 58 MVOC profiles in the DG18 data set were divided into 3 classes on the basis of the impactor data. Figure 5.1 shows a PC plot of all 18 compounds for the DG18 data set. Each air sample is represented as a point in the plot. MVOC profiles of the high mold count air samples are well separated from medium and low mold count samples in the PC plot of the data.

The genetic algorithm for pattern recognition analysis was used to identify specific compounds in the gas chromatograms characteristic of the MVOC profile of each class. The GA sampled key feature subsets, scored their principal component plots, and tracked those samples and/or classes that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 100 generations, the GA identified 8 MVOCs whose PC plot showed clustering of the gas chromatograms on the basis of mold count (see Figure 5.2). This suggests that

information about mold count is contained within the gas chromatograms of these air samples.

Table 5-2. Bioaerosol Data

Mold Count	Number of Samples	
	DG18	MEA
Low	20	22
Medium	5	7
High	33	29

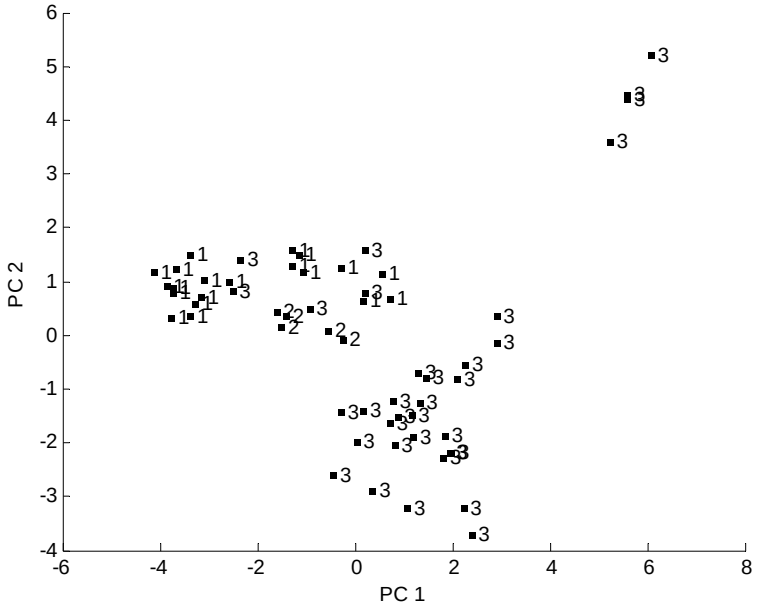

Figure 5.1. Plot of the two largest principal components of the 18 VOCs for DG18. Each air sample is represented as a point in the plot. 1 = low mold count exposure, 2 = moderate mold count exposure, and 3 = high mold count exposure. (Courtesy of *Microchem. J.* 2012, 103, 119-124.)

Table 5-3 shows the results of K-NN classification study for the 8 MVOCs identified by the pattern recognition GA. The overall classification success rate, calculated over the entire set of points using the 1-NN and 3-NN classification rule, indicated a high degree of clustering of the air samples based on the mold count. Only 2

samples were misclassified using the 1-NN classification rule and 5 samples by the 3-NN classification rule.

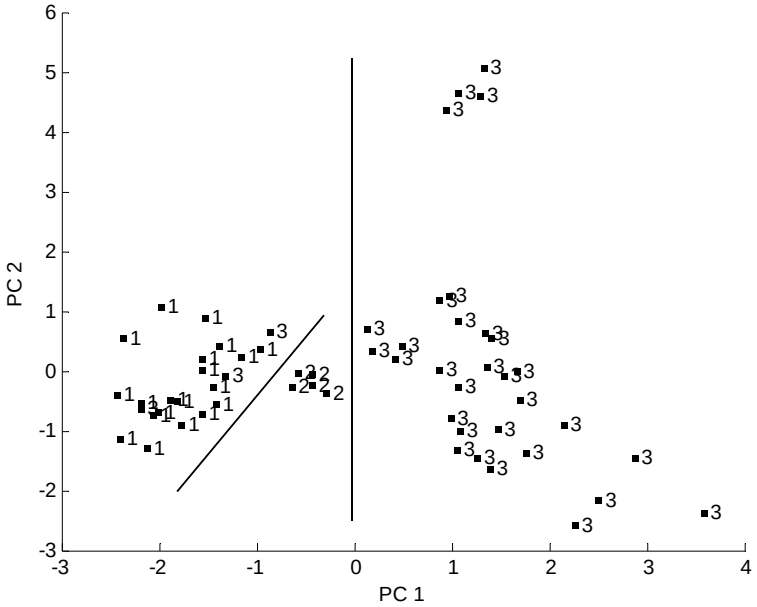

Figure 5.2. Plot of the two largest principal components of the 8 VOCs identified by the pattern recognition GA for DG18. Each air sample is represented as a point in the plot. 1 = low mold count exposure, 2 = moderate mold count exposure, and 3 = high mold count exposure. (Courtesy of *Microchemical J.,* 2012, 103, 119-124.)

Table 5-3. K-NN Results for DG18

K-NN	Total Number of Samples	Misclassifications
1-NN	58	2
3-NN	58	5

The 8 VOCs identified by the pattern recognition GA were selected for further study using cross validation to simulate the ability of the compounds to predict the mold count exposure of an unknown air sample. Twenty nine sets of VOC profiles were developed by random selection, where each training set consisted of 56 MVOC profiles and the corresponding prediction set contained the remaining 2 profiles. Each profile was only present in one of the 29 prediction sets generated. The training sets were analyzed

by three classification methods: LDA, QDA, and a 3-layer back propagation neural network (BPNN) using a sigmoid transfer function. The mold count exposure of the samples in the corresponding prediction set was determined using these trained models. Table 5-4 summarizes the results of the validation study. High classification success rates were obtained for both high and low mold count exposure suggesting that a distinct VOC profile representative of the MVOCs which could be differentiated from the blank was identified by the pattern recognition methodology. For moderate mold count exposure, the classification success rates were low with the misclassified VOC profiles assigned to the low mold count exposure class. This indicates that problems exist with differentiating background from VOCs of air samples collected in indoor environments with moderate mold counts.

Table 5-4. DG18 Cross Validation Set Results

Training method	Low mold count		Medium mold count		High mold count		Total	
	Missed	Success (%)	Missed	Success (%)	Missed	Success (%)	Missed	Success (%)
LDA	2	90	2	60	5	84.8	9	84.4
QDA	1	95	5	0	1	97	7	87.9
BPNN (10-3-3)	0	100	3	40	3	90.9	6	89.7

The 58 gas chromatograms from the MEA data set were also analyzed using pattern recognition methods. Figure 5.3 shows a plot of the two largest principal components of the 58 air samples and 18 compounds comprising this data set. Each sample is represented as a point in the PC plot of the data. There is overlap between the

three classes in the PC map of the data. Therefore, the pattern recognition GA was again used to identify features characteristic of the MVOC profile of each class. The pattern recognition GA identified informative descriptors by sampling key feature subsets, scoring their principal component plots and tracking those samples and or classes that were difficult to classify. After 100 generations, the pattern recognition GA identified 5 compounds whose PC plot showed separation on the largest principal component based on mold count exposure (see Figure 5.4). Table 5-5 shows the results of K-NN classification for the 5 MVOCs identified by the pattern recognition GA using the 1-NN and 3-NN classification rule. Table 5-6 summarizes the results of a cross validation study for the 5 MVOCs. Although K-NN yielded better results for DG18 than MEA, the results of the cross validation study for MEA were similar to DG18.

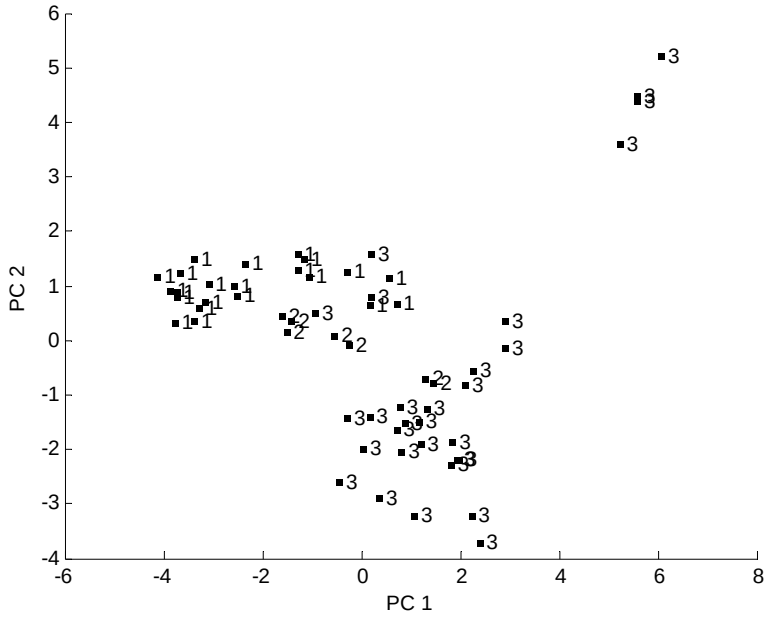

Figure 5.3. Plot of the two largest principal components of the 18 VOCs for MEA. Each air sample is represented as a point in the plot. 1 = low mold count exposure, 2 = moderate mold count exposure, and 3 = high mold count exposure. (Courtesy of *Microchemical J.*, 2012, 103, 119-124.)

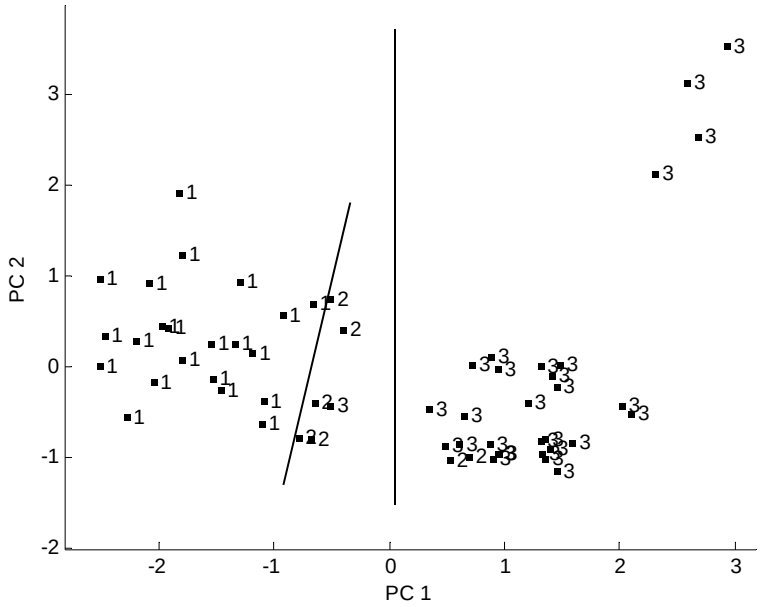

Figure 5.4. Plot of the two largest principal components of the 5 VOCs identified by the pattern recognition GA for MEA. Each air sample is represented as a point in the plot. 1 = low mold count exposure, 2 = moderate mold count exposure, and 3 = high mold count exposure. (Courtesy of *Microchemical. J.*, 2012, 103, 119-124.)

Table 5-5. K-NN Results for MEA

K-NN	Total Number of Samples	Misclassifications
1-NN	58	4
3-NN	58	10

Table 5-6. MEA Cross Validation Set Results

Training method	**Low mold count**		**Medium mold count**		**High mold count**		**Total**	
	Missed	Success (%)	Missed	Success (%)	Missed	Success (%)	Missed	Success (%)
LDA	2	90.91	4	57.14	3	89.7	9	84.75
QDA	0	100	5	28.57	1	96.6	6	89.7
BPNN (10-3-3)	2	90.91	2	71.43	2	93.1	6	89.7

To ascertain the validity of the PC map as an accurate representation of the 5-dimensional feature space for MEA, the FCV clustering algorithm was used. Four starting cluster centers were selected (Sample IDs 10, 13, 55, and 59) based on their location in the PC plot of the 5 compounds selected by the pattern recognition GA. A zero principal component model was used to characterize each fuzzy cluster in the data. The FCV clustering algorithm converged in 7 iterations with a class membership error of 0.0027. The four clusters identified by the FCV clustering algorithm, which are consistent with the PC plot of the data, are shown in Figure 5.5. Although FCV clustering was also used to assess the accuracy of the PC plot as an accurate representation of the 8-dimensional feature space identified by the pattern recognition GA for DG18, the results are not reported here as they were inconclusive.

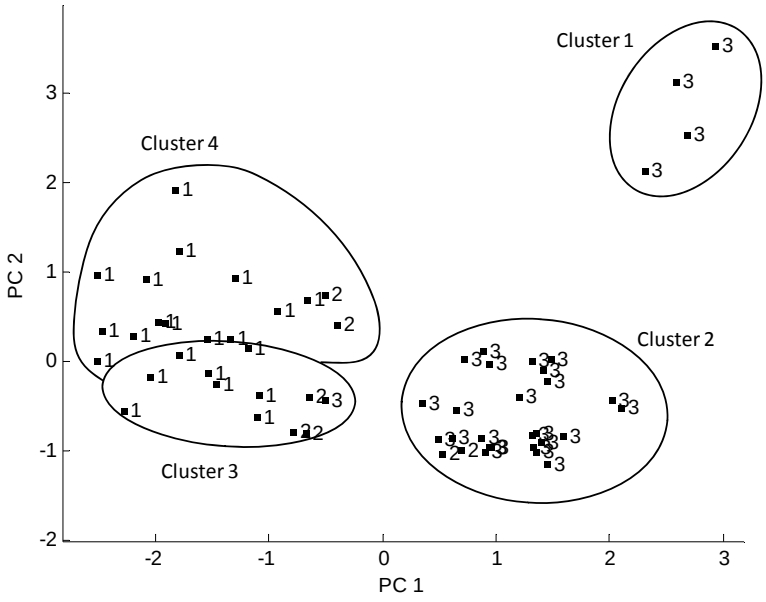

Figure 5.5. The samples comprising each cluster identified by the FCV clustering algorithm are circled and shown in the PC plot of the 5 VOCs that were identified by the pattern recognition GA for MEA. Each air sample is represented as a point in the plot. 1 = low mold count exposure, 2 = moderate mold count exposure, and 3 = high mold count exposure. (Courtesy of *Microchem. J.* 2012, 103, 119-124.)

DG18 and MEA were combined into a single data set to take full advantage of the broader range of fungi cultured by these two agars for assigning class labels to the MVOC profiles. GC class assignments remained unchanged for air samples that had the same mold counts in the two bioaerosol data sets. However, samples with different mold counts for DG18 and MEA agars were assigned the higher value. The GA for pattern recognition analysis was used to identify the informative compounds in the profile correlated to mold count exposure. Figure 5.6 is a PC plot of 8 VOCs identified by the pattern recognition GA as characteristic of the MVOC profiles. The 3 classes are well separated in the PC plot of these 8 compounds with the largest principal component containing the bulk of the discriminatory information about sample mold count.

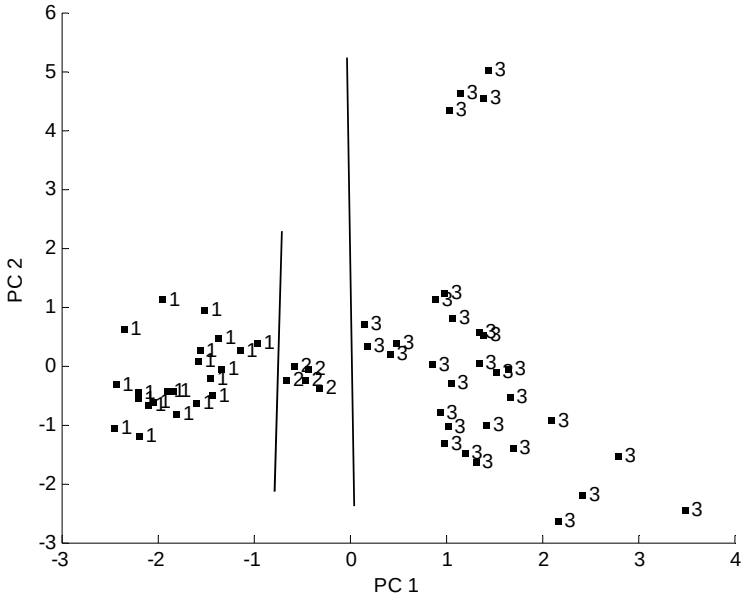

Figure 5.6. Plot of the two largest principal components of the 8 VOCs identified by the pattern recognition GA for DG18-MEA. Each air sample is represented as a point in the plot. 1 = low mold count exposure, 2 = moderate mold count exposure, and 3 = high mold count exposure. (Courtesy of *Microchem. J.* 2012, 103, 119-124.)

The 8 compounds identified by the pattern recognition GA from the combined DG18 and MEA data set were assessed for their ability to predict the mold count exposure of an unknown air sample. The same segmented cross validation design used for DG18 and MEA was used for the combined data set. Each training set was analyzed by LDA, and a 3-layer back propagation neural network (BPNN). QDA was not used to perform cross validation as there were only 5 chromatograms representing air samples with moderate mold count exposure, and the inverse of the class variance-covariance matrix cannot be directly computed from the data when the number of observations in a class is less than the number of measurements used to characterize the class. Table 5-7 summarizes the results of the validation study for LDA and BPNN. All GC profiles were correctly classified using a back propagation neural network. As for LDA, the assumption of equal class covariance, which was assessed by computing the determinant of the variance-covariance matrix for each class, did not hold true for this data. Therefore, it is not surprising that LDA did not perform as well as the neural network which does not utilize information about class covariance matrices in the development of decision surfaces to classify the VOC profile data.

Table 5-7. DG18-MEA Cross Validation Set Results

Training method	Low mold count		Medium mold count		High mold count		Total	
	Missed	Success (%)	Missed	Success (%)	Missed	Success (%)	Missed	Success (%)
LDA	2	90.91	1	80	0	100	3	94.9
BPNN (10-3-3)	0	100	0	100	0	100	0	100

A distinct profile indicative of MVOCs was developed from the air sampling data that could be readily differentiated from the blank for both high mold count and moderate mold count exposure samples. However, these results should be viewed as preliminary due to the small number of air samples obtained from moderate mold count exposure environments. Future studies will need to be undertaken in other locales to further assess the validity of the proposed method.

5.3 DIFFERENTIATION OF SMALL ROUND BLUE CELL TUMORS

The goal of this study was to differentiate between the different types of small round blue cell tumors (SRBCT), namely Neuroblastoma (NB), Rhabdomyosarcoma (RMS), Burkitt's lymphoma (BL), and the Ewing family of tumors (EWS), using gene expression data from cDNA microarrays. These cancers are difficult to distinguish by light microscopy, and currently there is no single test that can precisely identify these cancers. In clinical practice, several techniques are used for diagnosis including immunohistochemistry, cytogenetics, interphase fluorescence *in situ* hybridization, and reverse transcription.

The SRBCT data set, which consisted of 2308 genes across 83 samples, was divided into a training set of 63 samples and a prediction set of 20 samples as in the original study [5-26] published by Khan, see Tables 5-8 and 5-9. (In the original study there were 25 samples in the validation set. However, 5 of the samples were non SRBCT and were excluded from this study because they were not represented in the training set.) The training set data were autoscaled to remove any inadvertent weighting that might

otherwise occur due to differences in magnitude among the measurement variables. Further information about the collection of this data can be found elsewhere [5-26].

Table 5-8. Training Set for SRBCT

Cancer Type	Tumor Biopsy Material	Cell Lines	Total
EWS	13	10	23
BL	-	8	8
NB	-	12	12
RMS	10	10	20
Total	23	40	63

Table 5-9. Prediction Set for SRBCT

Cancer Type	Tumor Biopsy Material	Cell Lines	Total
EWS	5	1	6
BL	-	3	3
NB	4	2	6
RMS	5		5
Total	14	6	20

Three different fitness functions were employed: PCKaNN, PCKaNN with the Hopkins statistic and PCKaNN with the modified Hopkins statistic. The first step in this study was to apply PCA to the data. Figure 5.7 shows a plot of the two largest principal components developed from the 63 training set samples and 2308 genes. Each sample is represented as a point in the score plot (l = EWS, 2 = BL, 3 = NB, and 4 = RMS). There is overlap between the different types of small round blue cell tumors in the principal component map of the data indicating that feature selection is necessary. Therefore, the pattern recognition GA was used to identify features characteristic of the gene expression profile of each tumor class. The GA identified features by sampling key feature subsets, scoring their principal component plots, and tracking those classes and/or samples that were difficult to classify. The population consisted of 5000 chromosomes and the mutation rate of the genetic algorithm was set at 0.2. After 37 generations, PCKaNN identified 22 features whose principal component plot showed clustering of the training

set samples (grey) on the basis of tumor type (see Figure 5.8). The 20 prediction set samples (black) were then projected onto the principal component map developed from the 63 training set samples and the 22 genes identified by PCKaNN (see Figure 5.8). 15 of the 20 prediction set samples were projected onto a region of the map containing tumor samples with the same class label.

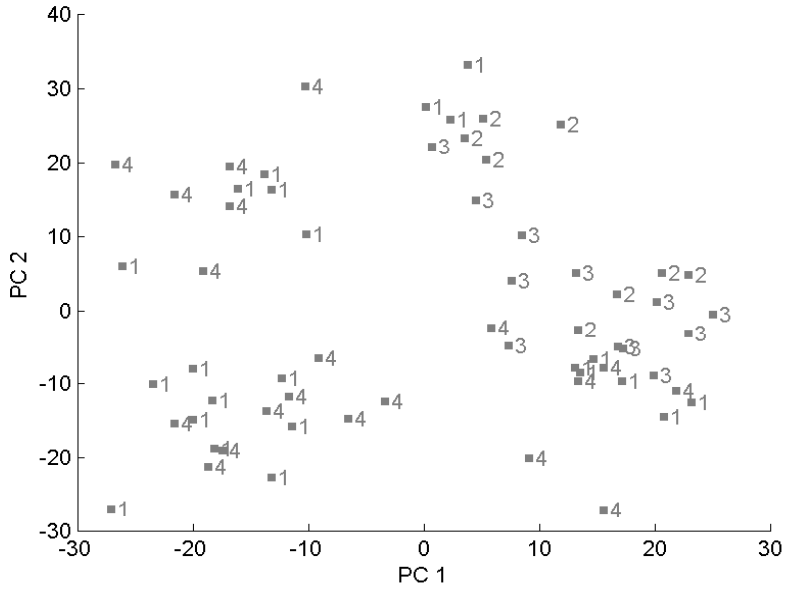

Figure 5.7. A plot of the two largest principal components developed from the 63 training set samples and the 2308 genes. Each sample is represented as a point in the score plot (1 = EWS, 2 = BL, 3 = NB, and 4 = RMS).

This classification problem was also tackled by incorporating transverse learning into the feature selection process. Figure 5.9 summarizes the results obtained for PCKaNN with the deweighted Hopkins statistic, and Figure 5.10 summarizes the results obtained for PCKaNN with the modified Hopkins statistic. In both studies, the population consisted of 5000 chromosomes, the mutation rate was 0.2, 200 generations were run, and the fitness function consisted of 90% PCKaNN and 10% Hopkins or 90% PCKaNN with 10% modified Hopkins statistic. For the run involving the deweighted Hopkins statistic (see Figure 5.9), 30 features were selected by the pattern recognition

GA. All of the training set samples (grey) were correctly classified, and 18 of the 20 prediction set samples (black) were located in a region of the map with SRBCT samples that had the same class label. The results obtained for the modified Hopkins statistic (see Figure 5.10) were more impressive. Thirty one features were identified by the pattern recognition GA. All of the training set samples (grey) were correctly classified, and all prediction set samples (black) were located in a region of the map with SBRCT samples that had the same class label.

Of the 83 features identified by the pattern recognition GA using the three fitness functions, 61 were unique. Only 5 features were selected by all three fitness functions, whereas 45 features were selected by a single fitness function and 11 features were selected by two fitness functions. The fitness functions PCKaNN and PCKaNN with deweighted Hopkins had 8 features in common, whereas the fitness functions PCKaNN and PCKaNN with modified Hopkins had 9 features in common. The two fitness functions incorporating transverse learning also had 9 features in common. Of the 45 features selected by only a single fitness function, 7 were selected by PCKaNN, 17 were selected by PCKaNN with Hopkins and 18 were selected by PCKaNN with modified Hopkins.

Clearly, a larger fraction of the features identified by the two fitness functions that utilize transverse learning are unique features and this has implications with regard to applying this methodology for the selection of biomarker candidates. These results also confirm that this set of gene expression data contains a wealth of information relevant to separating the samples by tumor type.

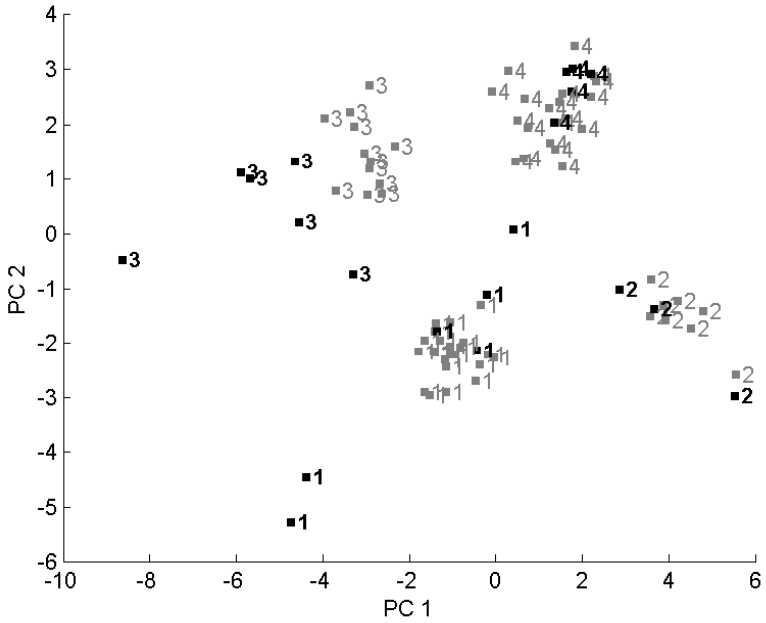

Figure 5.8. A plot of the two largest principal components developed from the 63 training set samples and 22 genes identified by PCKaNN. Each sample is represented as a point in the score plot (1 = EWS, 2 = BL, 3 = NB, and 4 = RMS). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

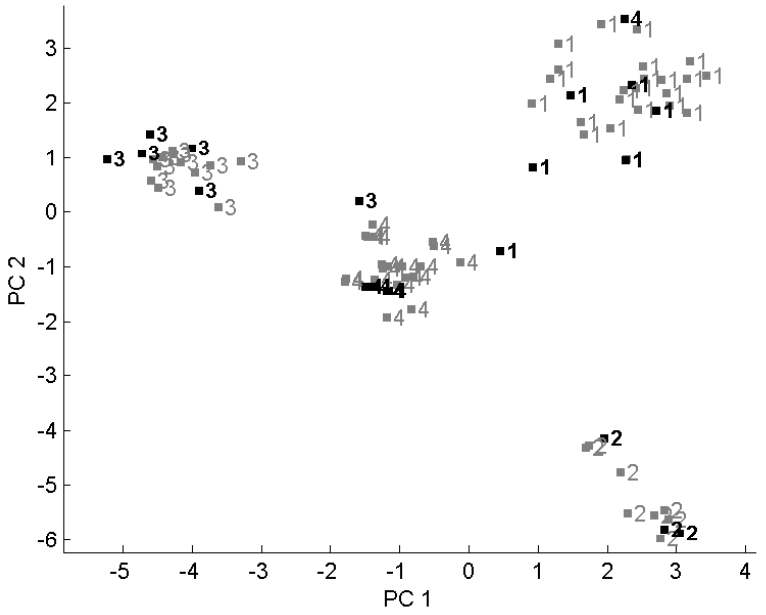

Figure 5.9. A plot of the two largest principal components developed from the 63 training set samples and 22 genes identified by PCKaNN and the Hopkins statistic. Each sample is represented as a point in the score plot (1 = EWS, 2 = BL, 3 = NB, and 4 = RMS). Training set samples (labeled points) are in grey and the prediction set samples (unlabeled points) are in black.

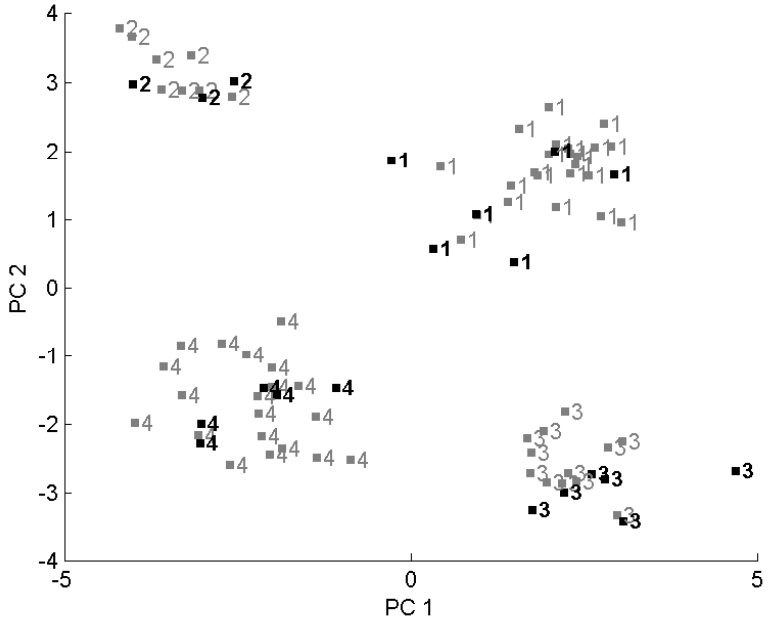

Figure 5.10. A plot of the two largest principal components developed from the 63 training set samples and 22 genes identified by PCKaNN and the modified Hopkins statistic. Each sample is represented as a point in the score plot (1 = EWS, 2 = BL, 3 = NB, and 4 = RMS). Training set samples (labeled points) are in grey and the prediction set samples (unlabeled points) are in black.

Features selected by the pattern recognition GA using a fitness function that incorporates transverse learning performed better for prediction of the validation set samples included in the training set (as unlabeled points) than features selected by PCKaNN. However, these same features and subsequently any model developed from them may not perform as well (compared to the features selected by PCKaNN) for future samples, i.e., those not included in the training set as unlabeled samples.

5.4 DISCOVERY OF BIOMARKER CANDIDATES FOR LIVER CANCER FROM MALDI-TOF DATA OF TISSUE *N*-LINKED GLYCANS

The goal of this study was to identify potential biomarkers for liver cancer from MALDI-TOF data of tissue *N*-linked glycans. *N*-linked glycans are polysaccharides or oligosaccharides that are attached to a nitrogen atom of an amino acid residue in

a glycoprotein. Glycans have received considerable attention as potential biomarkers for detection of cancer. Aberrant glycosylation has been linked to several cancers [5-27 to 5-36] and to specific structural changes occurring in metastatic cells, e.g., an increase in sialylation or fucosylation has been reported [5-37]. Incomplete or truncated structures originating from variations in the expression of glycosyltransferases often are also associated with diseased phenotypes [5-37]. Glycan characterization is challenging because glycans can form a vast number of different structures due to alternative branching and linkage possibilities, which leads to a large number of diverse molecules and multiple isomeric species [5-37 and 5-38].

For this study, tissue samples were collected from various donors, some of whom were healthy, while others were diagnosed with Hepatocellular Carcinoma or Liver Cirrhosis. Thirty three liver tissue samples were collected that comprised 5 normal controls, 5 cirrhosis, 16 hepatocellular carcinoma, and 7 uninvolved tissue samples from patients diagnosed with hepatocellular carcinoma. To generate the MALDI-TOF data, the tissue samples were first processed for enzymatic release of *N*-glycans from glycoproteins, followed by extraction and solid phase permethylation [5-39 and 5-40]. Permethylated glycans were spotted on a MALDI plate with DHB-matrix and vacuum dried for co-crystallization. Glycomics data was collected using Ultraflex II MALDI-TOF-TOF (Bruker Daltonics, Billerica, MA) in reflectron positive ion mode with an accelerating voltage of 23 kV. The m/z range of data collected was from 500-7000 Da.

Two data sets were developed from the raw MALDI-TOF data. The first data set consisted of 125 peaks that represented known *N*-glycans, while the second dataset was developed using all of the peaks in the MALDI-TOF data. For the first data set, raw mass

spectra were processed using flexAnalysis 2.4 (Bruker Daltonics, Billerica, MA) and 125 peaks in the m/z range of 1400-6000 Daltons were selected by a software tool called Glycoworkbench [5-41]. For the second data set, raw MALDI-TOF data was preprocessed for baseline correction, denoising, binning and peak alignment using the pkDACLAS software package [5-42]. After pre-processing, each mass spectrum consisted of 6500 points. The spectra in both data sets were normalized by taking the ratio of each peak with the largest peak in the spectrum. Both data sets were divided into a training set of 29 spectra and a validation set of 4 spectra as shown in Table 5-10. Spectra comprising the validation set were chosen by random lot.

Table 5-10. Training Set and Prediction Set

Tissue Type	Number of Samples	Number of Samples in Training Set	Number of Samples in Prediction Set
Normal	5	4	1
Cirrhosis	5	4	1
Hepatocellular Carcinoma	16	14	2
Uninvolved	7	7	-
Total	33	29	4

The 125 peak MALDI-TOF data set was first analyzed. For pattern recognition analysis, each spectrum was represented as a data vector $X = (x_1, x_2, x_3 \ldots x_{125})$, where the components of the data vector are the line intensities at specific m/z values that represent *N*-glycans in the MALDI-TOF data. The training set spectra were autoscaled to remove any inadvertent weighting that might occur due to differences in magnitude among the measurement variables. The first step in this study was to analyze 125 peak MALDI-TOF spectra using PCA. Figure 5.11 shows a plot of the three largest principal components developed from the 29 spectra and 125 peaks that comprised the training set

data. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, 3 = Hepatocellular Carcinoma, and 4 = Uninvolved). The PC plot does not show any separation between the four classes in the data. One sample that belonged to Hepatocellular Carcinoma was identified as an outlier and was removed from the study.

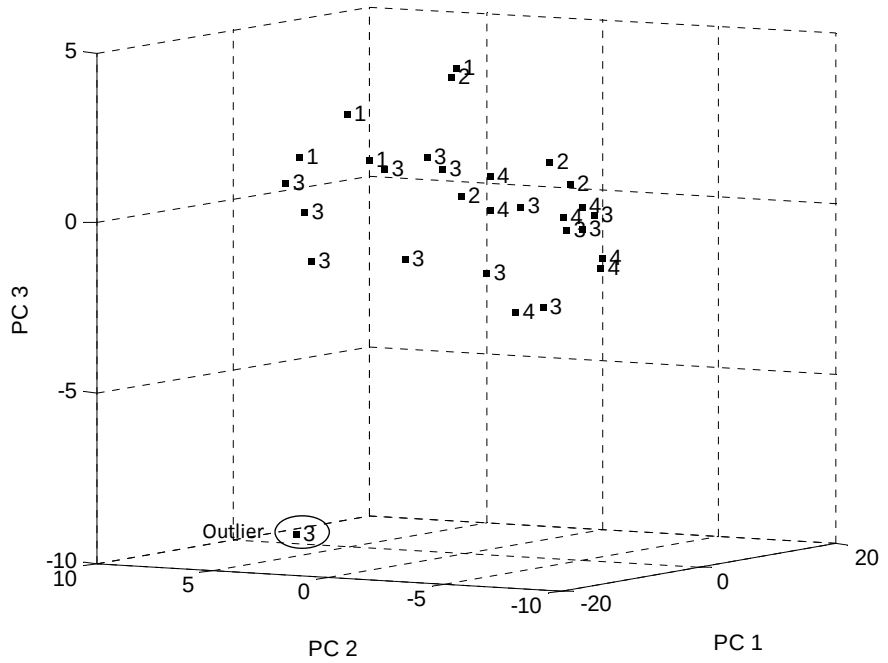

Figure 5.11. A plot of the three largest principal components developed from the 29 training set spectra and 125 peaks. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, 3 = Hepatocellular Carcinoma, and 4 = Uninvolved).

In the next step, the GA for pattern recognition analysis using PCKaNN was used to identify mass spectral peaks characteristic of each class. The pattern recognition GA identified features that could separate the four sample types by sampling key feature subsets, scoring their PC plots and tracking those classes and/or samples that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the GA identified 11 peaks that contained

discriminatory information about sample type. Figure 5.12 shows a plot of the three largest principal components developed from the 28 training set spectra and 11 peaks identified by the pattern recognition GA. The Normal and Cirrhosis samples were well separated from each other and the other sample types whereas Hepatocellular Carcinoma and Uninvolved tissue samples overlapped with each other. The 4 prediction set samples (black) were then projected onto the principal component map of the 28 training set samples (grey) and the 11 peaks identified by the GA (see Figure 5.13). All 4 prediction set samples were projected onto a region of the map that contained samples with the same class label. Table 5-11 lists the m/z values of the 11 features identified by the pattern recognition GA and the sample class for which they have the highest average intensity value.

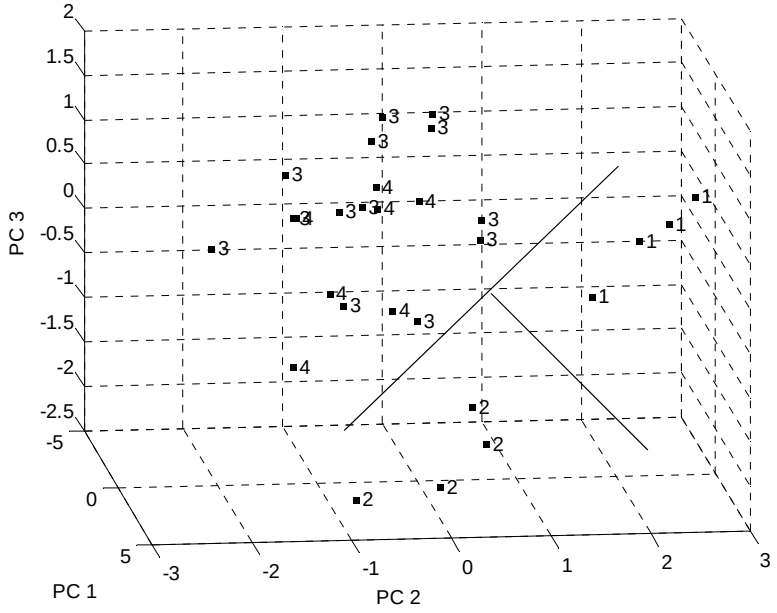

Figure 5.12. A plot of the three largest principal components developed from the 28 training set spectra and 11 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (1 = Normal, 2 = Cirrhosis, 3 = Hepatocellular Carcinoma, and 4 = Uninvolved).

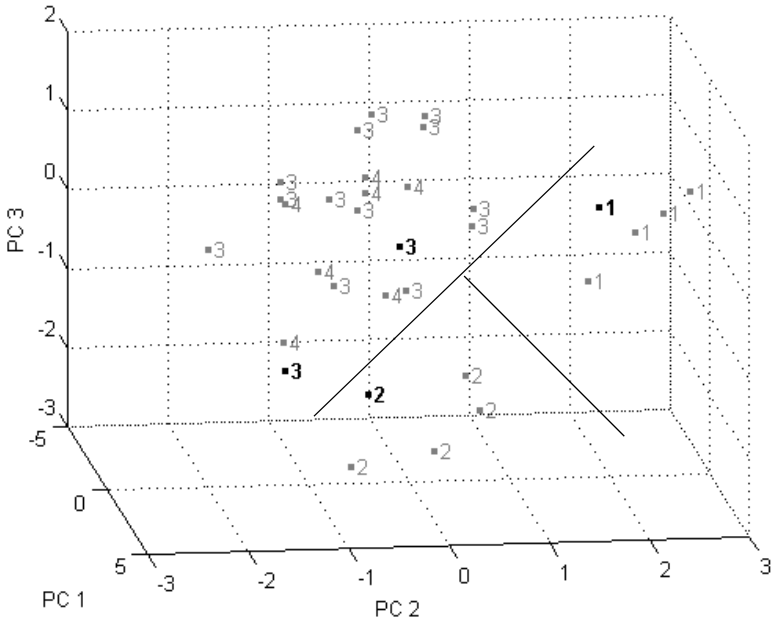

Figure 5.13. A plot of the three largest principal components developed from the 28 training set spectra and 11 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (1 = Normal, 2 = Cirrhosis, 3 = Hepatocellular Carcinoma, and 4 = Uninvolved). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

Table 5-11. Identity of the 11 Features Identified by the GA

m/z Value of Feature	Class with Highest Average Intensity Value
1763.7	Hepatocellular Carcinoma and Uninvolved
1783.9	Hepatocellular Carcinoma and Uninvolved
2591.3	Hepatocellular Carcinoma and Uninvolved
2736.3	Normals
2822.4	Normals
3602.8	Hepatocellular Carcinoma
3865.2	Normals
4056.9	Hepatocellular Carcinoma
4400.3	Hepatocellular Carcinoma
4443.3	Hepatocellular Carcinoma
4763.6	Hepatocellular Carcinoma

The overlap between Hepatocellular Carcinoma and Uninvolved tissue samples suggested that information about the identity of the patient is also contained in the MALDI-TOF profiles, as these samples were obtained from the same donors. To assess this hypothesis, other comparisons of the MALDI-TOF profiles involving samples from Hepatocellular Carcinoma and/or Uninvolved tissue were undertaken. For example, GA runs involving several three-way and two-way classifications of the data were undertaken: (1) Normal, Cirrhosis, and Hepatocellular Carcinoma; (2) Normal, Cirrhosis, and Uninvolved; (3) Normal, Hepatocellular Carcinoma and Uninvolved; (4) Cirrhosis, Hepatocellular Carcinoma and Uninvolved; and (5) Hepatocellular Carcinoma and Uninvolved. For the 2-way and 3-way classifications, Uninvolved tissue samples always overlapped with Heptacellular Carcinoma, whereas the other three sample types could be separated from each other (see Figures 5.14 to 5.18).

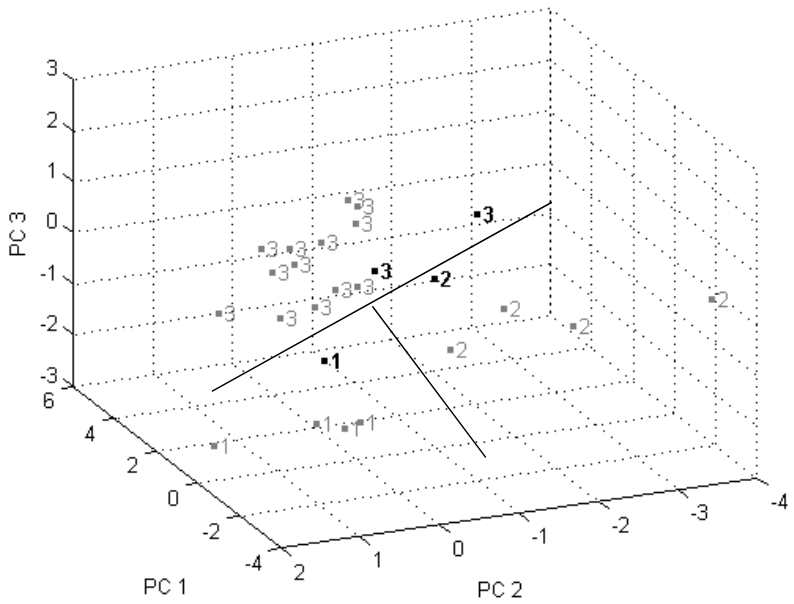

Figure 5.14. A plot of the three largest principal components developed from the 21 training set spectra and 12 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, and 3 = Hepatocellular Carcinoma). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

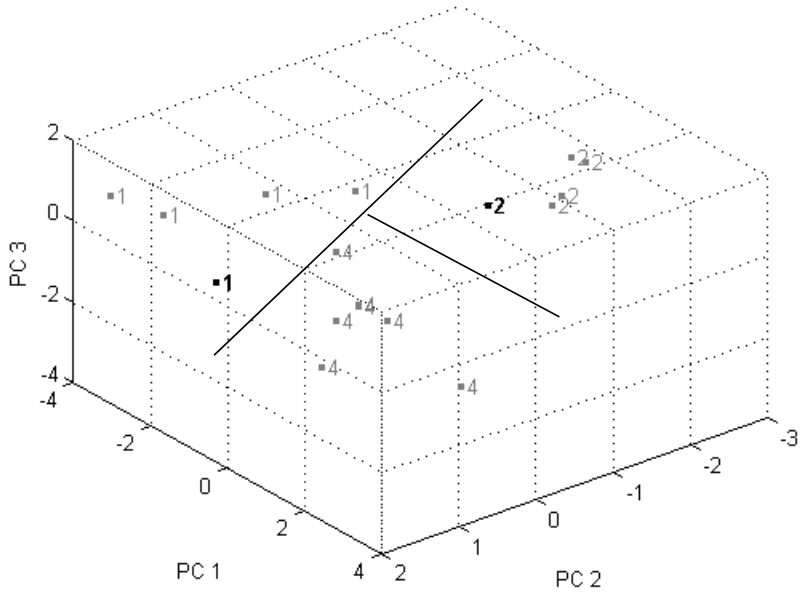

Figure 5.15. A plot of the three largest principal components developed from the 15 training set spectra and 10 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, and 4 = Uninvolved). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

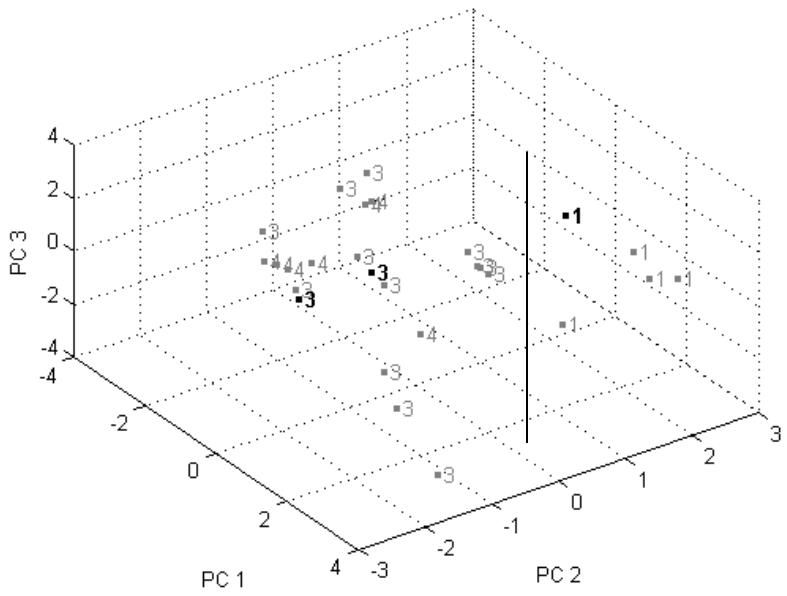

Figure 5.16. A plot of the three largest principal components developed from the 24 training set spectra and 10 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 3 = Hepatocellular Carcinoma, and 4 = Uninvolved). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

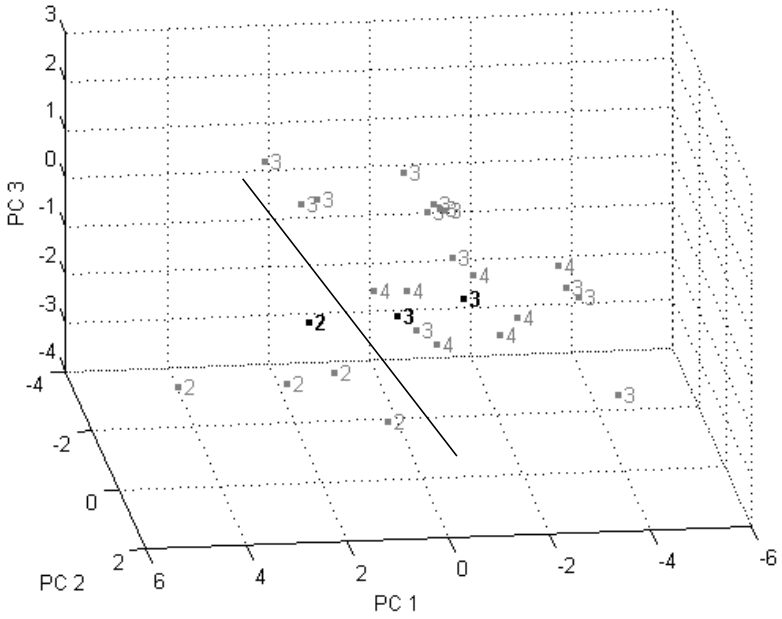

Figure 5.17. A plot of the three largest principal components developed from the 24 training set spectra and 12 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (2 = Cirrhosis, 3 = Hepatocellular Carcinoma, and 4 = Uninvolved). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

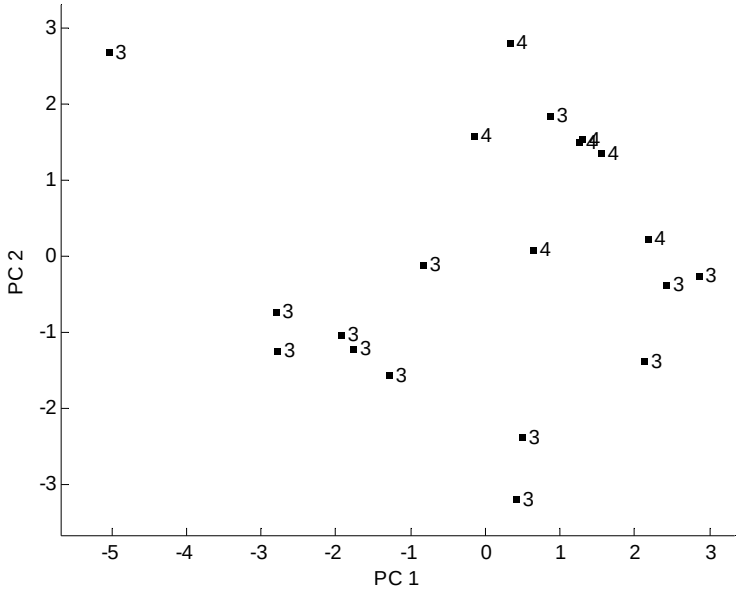

Figure 5.18. A plot of the two largest principal components developed from the 20 training set spectra and 13 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (3 = Hepatocellular Carcinoma and 4 = Uninvolved).

Because uninvolved tissue samples always overlapped with Heptacellular Carcinoma, both Hepatocellular Carcinoma and Uninvolved tissue samples were assigned to the same class and studied along with the other two classes (Normal and Cirrhosis) using transverse learning (80 % PCKaNN and 20 % modified Hopkins). The pattern recognition GA identified 6 peaks that contained discriminatory information about sample type. Figure 5.19 shows a plot of the three largest principal components developed from the 28 training set spectra and 6 mass spectral peaks identified by the pattern recognition GA. The PC map of these 6 features shows clustering of the samples on the basis of class. When the 4 prediction set samples (black) are projected onto the PC map developed from the 28 training set samples (grey) and the 6 peaks identified by the pattern recognition GA (see Figure 5.20), all 4 prediction set samples lie in a region of the map containing samples with the same class label. This was consistent with the results obtained from the 4-class study. Table 5-12 lists the m/z values for the 6 features identified by the pattern recognition GA for the 3-way classification involving Normals, Cirrhosis, and Heptacellular Carcinoma and Uninvolved tissue, and the sample class for which they have the highest average intensity value.

Table 5-12. Identity of the 6 Features identified by the GA

m/z Value of Feature	Class with Highest Average Intensity Value
1729.9	Hepatocellular Carcinoma
1763.9	Hepatocellular Carcinoma and Uninvolved
1982.0	Hepatocellular Carcinoma
2426.2	Hepatocellular Carcinoma
4057.0	Hepatocellular Carcinoma
4763.6	Hepatocellular Carcinoma

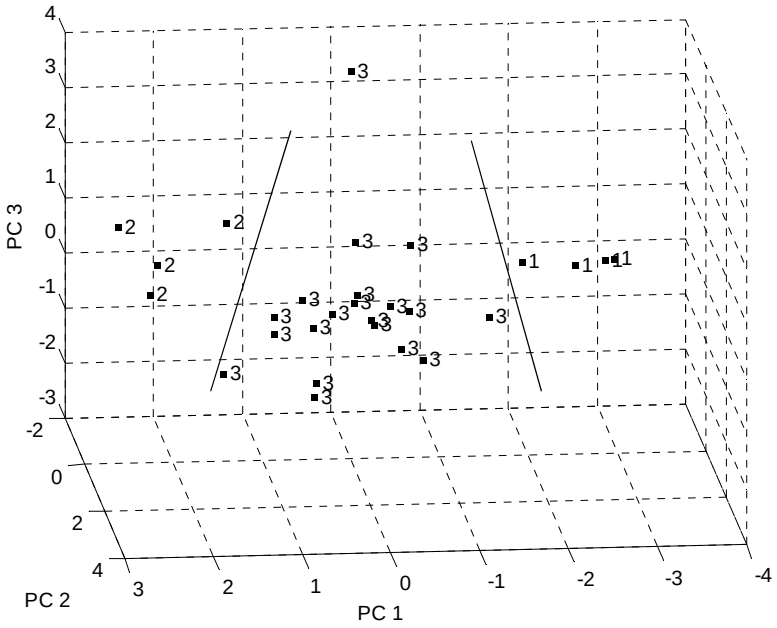

Figure 5.19. A plot of the three largest principal components developed from the 28 training set spectra and 6 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, and 3 = Hepatocellular Carcinoma or Uninvolved).

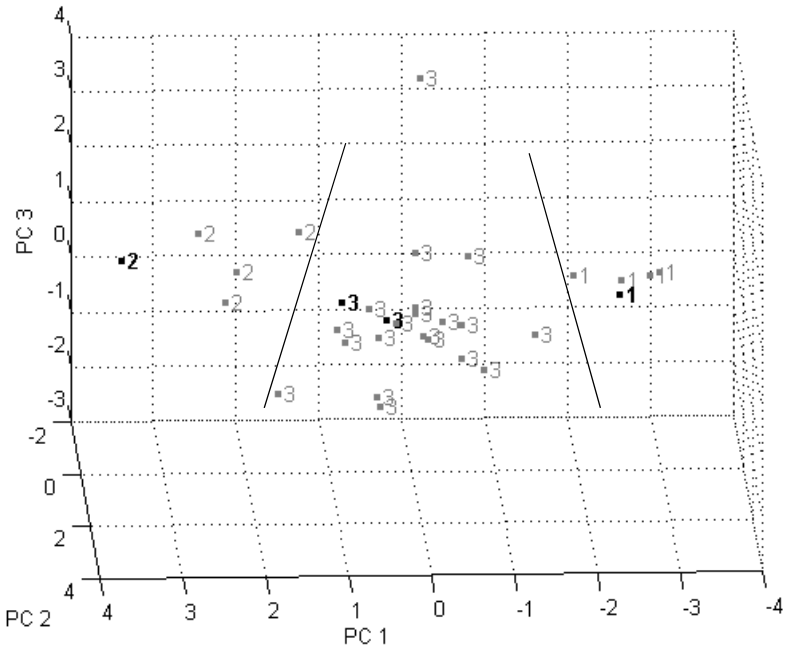

Figure 5.20. A plot of the three largest principal components developed from the 28 training set spectra and 6 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, and 3 = Hepatocellular Carcinoma or Uninvolved). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

The second MALDI-TOF data set with 6500 points per spectrum was also analyzed using the pattern recognition GA. Each mass spectrum was divided into three intervals based on visual analysis. Several GA runs were performed on each interval which revealed that only noise was contained in the spectral range 500 to 1000 m/z and 5000 to 7000 m/z. Therefore, the spectra were truncated and only the points in the range from 1000 m/z to 5000 m/z which corresponded to 4000 points were retained for further analysis. The truncated data set was reanalyzed using the pattern recognition GA with transverse learning (80 % PCKaNN and 20 % modified Hopkins). A 4-way classification of the data was unsuccessful. When the uninvolved tissue samples were combined with samples of Hepatocellular Carcinoma to form a single class, clustering was observed on the basis of sample class membership. Figure 5.21 shows a plot of the two largest principal components developed from the 29 training set spectra and the 12 points identified by the pattern recognition GA. Each point represents a different peak as these 12 points span the entire mass spectrum. When the 4 prediction set samples (black) are projected onto the PC map developed from the 29 training set samples (grey) and the 12 points identified by the pattern recognition GA (see Figure 5.22), all 4 projected samples lie in a region of the map with training set samples that have the same class label. Although a different transduction method was used to represent the mass spectra as data vectors (5000 points versus 125 peaks selected by the Glycoworkbench software tool), the same results were obtained. Evidently, the MALDI-TOF mass spectra contain information about the diseased state of the subject and the identity of the subject as the samples of Uninvolved tissue and Hepatocellular Carcinoma are from the same donors.

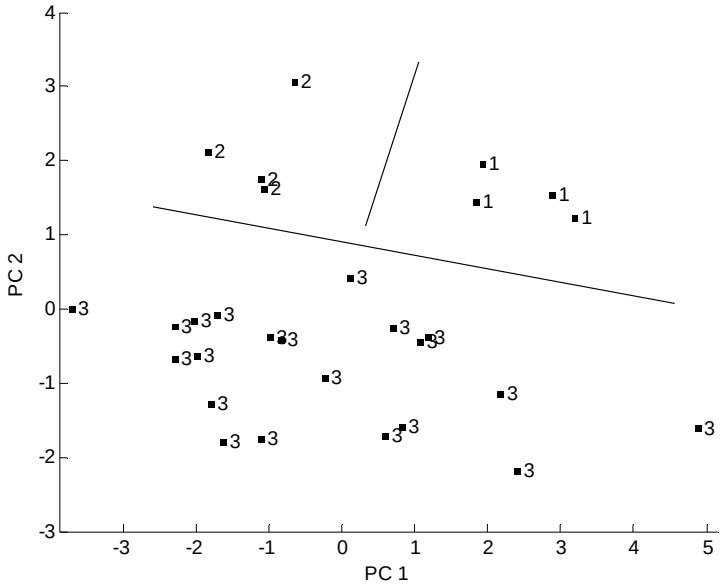

Figure 5.21. A plot of the two largest principal components developed from the 29 training set spectra and 12 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, and 3 = Hepatocellular Carcinoma or Uninvolved).

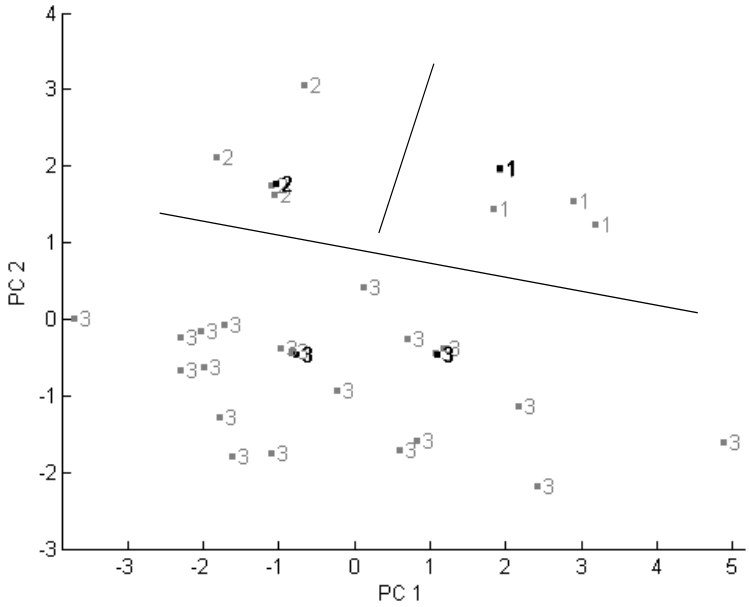

Figure 5.22. A plot of the two largest principal components developed from the 29 training set spectra and 12 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Cirrhosis, and 3 = Hepatocellular Carcinoma or Uninvolved). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

5.5 DISCOVERY OF BIOMARKER CANDIDATES FOR LIVER CANCER FROM IMS DATA OF SERUM *N*-GLYCANS

The goal of this study was to identify potential biomarkers for liver cancer from IMS data of serum *N*-glycans. The genetic algorithm for pattern recognition analysis was used to identify time tags from IMS data of serum *N*-glycans that could serve as potential biomarkers for liver cancer. 90 human serum samples were collected from donors to develop the *N*-glycan IMS data. The data set consisted of 30 normal control, 30 hepatocellular carcinoma, and 30 liver cirrhosis samples.

The sample preparation for developing IMS profiles included enzymatic release of *N*-glycans from serum glycoproteins, followed by solid phase extraction and permethylation. Further details about the sample processing can be found elsewhere [5-43]. The IMS spectra for the 90 glycan samples were developed using instrumentation built in-house. The ions were generated by electrospray ionization using a modified NanoMate (TriVersa, Advion, Ithaca, NY) auto injection system. The beam of electrosprayed ions were accumulated in an hourglass shaped ion funnel and a packet of ions was periodically released into the drift tube using an electrostatic gate. The drift tube was filled with 2.5 Torr of 300 K He buffer gas and the activation region was set at 20 V. Upon exiting the drift tube, ions were focused into the source region of a time-of-flight mass analyzer and detected by a multichannel plate (MCP) detector. Further details about the IMS instrument used for generating the data can be found elsewhere [5-44].

Each IMS spectrum was represented by 4972 time tags. The data were divided into a training set of 81 samples and a prediction set of 9 samples. Table 5-13 shows the number of samples in the training set and prediction set for each class. The prediction set samples were chosen by random lot. Each spectrum in the IMS data was normalized to

constant sum and the training set data were autoscaled prior to pattern recognition analysis. The training set spectra were analyzed using PCA. Figure 5.23 shows a plot of the two largest principal components developed from the 81 training set spectra and 4972 time tags. Each ion mobility spectrum is represented as a point in the PC score plot (l = Normal, 2 = Hepatocellular Carcinoma, and 3 = Cirrhosis).

Table 5-13. Training Set and Prediction Set

Tissue Type	Number of Samples	Number of Samples in Training Set	Number of Samples in Prediction Set
Normal	30	27	3
Hepatocellular Carcinoma	30	27	3
Cirrhosis	30	27	3
Total	90	81	9

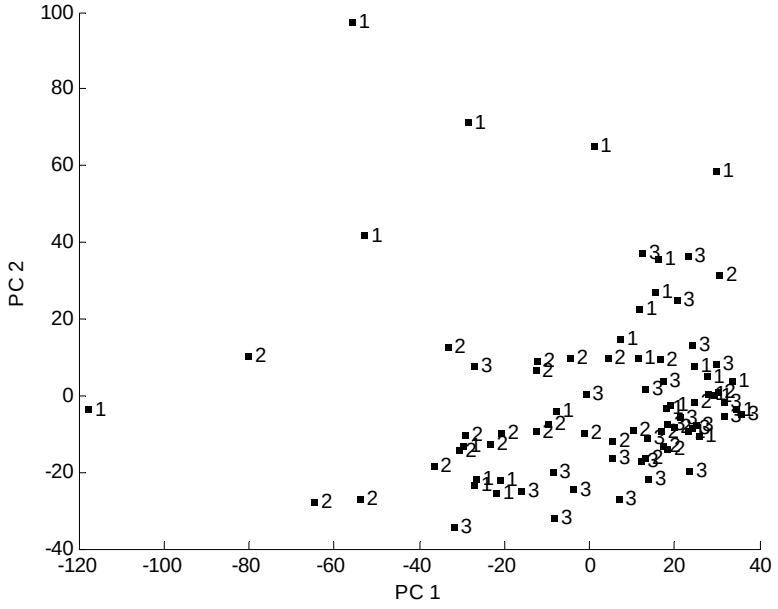

Figure 5.23. A plot of the two largest principal components developed from the 81 training set spectra and 4972 time tags. Each ion mobility spectrum is represented as a point in the PC score plot (l = Normal, 2 = Hepatocellular Carcinoma, and 3 = Cirrhosis).

The IMS data was analyzed using the pattern recognition GA with transverse learning (80% PCKaNN and 20% modified Hopkins). The pattern recognition GA identified potential biomarkers for liver cancer by sampling key feature subsets, scoring their principal component plots, and tracking those samples and/or classes that were most difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 100 generations, the genetic algorithm identified 20 time tags whose principal component plot showed clustering of the ion mobility spectra according to sample type. The normals, hepatocellular carcinoma, and cirrhosis samples are well separated from each other in the plot of the two largest principal components (see Figure 5.24).

The prediction set of 9 ion mobility spectra was employed to assess the predictive ability of the 20 time tags identified by the pattern recognition GA. Figure 5.25 shows the 9 spectra from the prediction set (black) that were projected onto the PC score plot defined by the 81 spectra of the training set (grey) and 20 time tags. Each projected spectrum is in a region of the map with samples that have the same class label. Evidently, the GA can identify time tags from the ion mobility spectra that are correlated to the disease state of the subject from which the IMS spectrum was obtained.

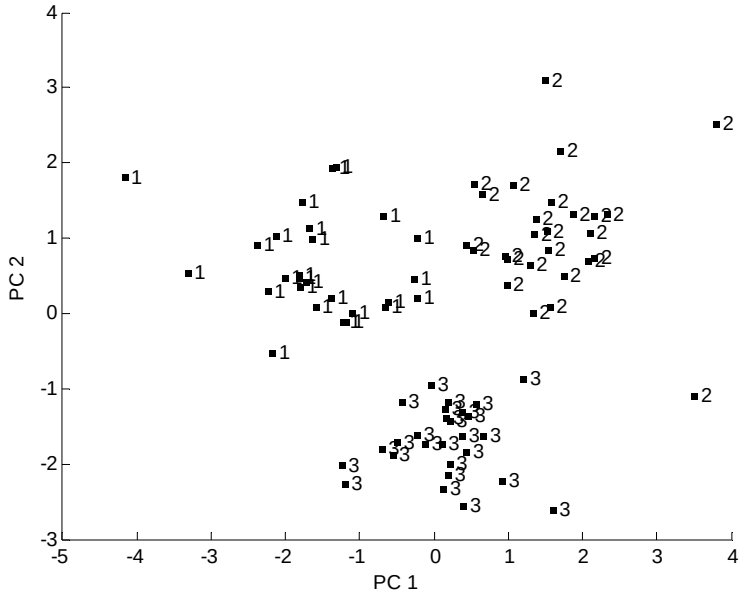

Figure 5.24. A plot of the two largest principal components developed from the 81 training set spectra and 20 time tags identified by the pattern recognition GA. Each ion mobility spectrum is represented as a point in the PC score plot (l = Normal, 2 = Hepatocellular Carcinoma, and 3 = Cirrhosis).

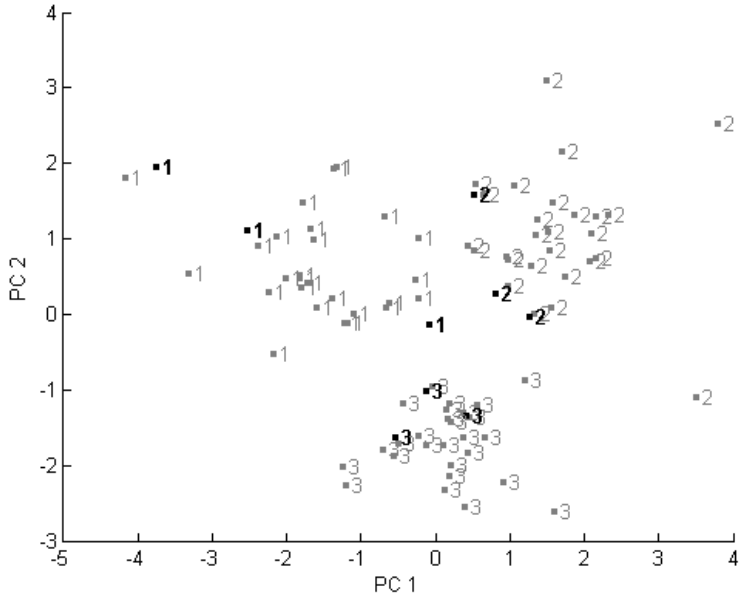

Figure 5.25. A plot of the two largest principal components developed from the 81 training set spectra and 20 time tags identified by the pattern recognition GA. Each ion mobility spectrum is represented as a point in the PC score plot (l = Normal, 2 = Hepatocellular Carcinoma, and 3 = Cirrhosis). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

5.6 DISCOVERY OF BIOMARKER CANDIDATES FOR PANCREATIC CANCER FROM MALDI-TOF DATA OF SERUM *N*-GLYCANS

The goal of this study was to identify potential biomarkers for pancreatic cancer from MALDI-TOF data developed from serum *N*-glycans. The genetic algorithm for pattern recognition analysis was used to identify peaks from MALDI-TOF profiles of serum *N*-glycans that were potential biomarkers for pancreatic cancer. The MALDI-TOF data was developed from 41 serum samples that consisted of 22 normal controls, 7 chronic pancreatis and 12 pancreatic adenocarcinoma samples. Pattern recognition analysis also performed on a second data set that served as a validation set and contained 32 MALDI-TOF profiles, which consisted of 8 normal controls, 8 chronic pancreatis and 16 pancreatic adenocarcinoma samples. The composition of the two data sets is summarized in Table 5-14.

Table 5-14. Training Set and Prediction Set

Sample Type	Number of Samples in First Data Set	Number of Samples in Second Data Set
Normal	22	8
Chronic Pancreatis	7	8
Pancreatic Adenocarcinoma	12	16
Total	41	32

The experimental conditions and procedure used for preparation of samples and procurement of MALDI-TOF data was similar to that used in a previous study and can be found elsewhere [5-45 to 5-47]. The spectral profiles consisted of 127 peaks that represented *N*-glycans. The data were normalized to unit length and autoscaled prior to pattern recognition analysis. PCA was used to analyze the 41 spectral profiles. Figure 5.26 shows a plot of the two largest principal components developed from the 41 spectra and 127 peaks. Each spectrum is represented as a point in the PC score plot (l = Normal,

2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). An examination of the score plot shown in Figure 5.26 reveals the presence of two outliers in the data. These two samples always behaved as outliers even in combination with other normalization (unit length, constant sum, and ratio of largest peak) and scaling (autoscaling and mean centering) routines. They were excluded from further analyses.

PCKaNN was used to identify mass spectral peaks that showed clustering based on the class label of the sample. The population consisted of 200 chromosomes, mutation rate was 0.2, three-point cross-over was used, and K for each class was set equal to the number of samples in the class. The GA identified informative peaks by sampling key feature subsets, scoring their principal component plots, and tracking those samples and/or classes that were most difficult to classify. After 50 generations, the GA identified 5 spectral features whose PC plot did not show clustering on the basis of sample type (see Figure 5.27).

Because separation between classes was not achieved, different combinations of normalization and scaling methods were investigated using PCKaNN. The normalization and transformation methods investigated were normalization to largest peak ratio, unit length, constant sum, and log transform. Scaling methods used to preprocess the data were autoscaling, mean centering, range scaling, level scaling, Paretoscaling and generalized Paretoscaling. In all, 24 preprocessing combinations were investigated for PCKaNN. From the results obtained using PCKaNN, it was obvious that a different eigenvector projection method was needed to display the information content of the data for the pattern recognition GA. For this reason, CVAKNN was substituted for PCKaNN in the pattern recognition GA.

Figure 5.28 shows a plot of the two largest canonical variates developed from the 39 spectra and 127 peaks. No clustering of the samples based on class was observed. Pattern recognition analysis was performed using CVAKNN. The population consisted of 200 chromosomes, mutation rate was 0.2, three-point cross-over was used, and K for each class was set equal to the number of samples in the class. The GA identified peaks by sampling key feature subsets, scoring their CVA plots, and tracking those classes and samples that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 50 generations, the GA identified 18 mass spectral features whose CVA plot showed clustering of the samples on the basis of the class label (see Figure 5.29).

Object validation was performed by dividing the data set into a training set (36 spectra) and a prediction set (3 spectra). One sample was chosen by random lot from each of the three classes for the prediction set. The pattern recognition GA was run on the training set with CVAKNN. The GA identified 22 mass spectral peaks that were correlated with the class membership of the samples in the training set. Prediction set samples were projected on the CVA plot developed from the 22 peaks identified by the pattern recognition GA (see Figure 5.30). All 3 prediction set samples were projected onto a region of the map containing samples with the same class label.

To reduce the occurrence of chance classification and select only those features which will have true potential as biomarkers, a method known as schema hunting was implemented. Several GA runs were performed on the training set with each run having different values for the GA parameters such as different mutation rates (0.1, 0.2, 0.3, and 0.4). A histogram depicting the number of times each spectral feature was selected by the

pattern recognition GA during each generation (number of hits) is shown in Figure 5.31. Features that have a number of hits equal to or larger than a specific threshold value were selected for classification and analyzed. The threshold used for selecting subsets of features was increased in a stepwise manner to obtain a decreasing number of features in the feature subsets. CVA plots were developed to evaluate the feature subsets. The minimum number of features required to separate the samples in the training set by sample type and correctly predict the class membership of the mass spectra in the prediction set was 24 peaks (number of hits $\geq$ 9). Further removal of uninformative features was achieved stepwise by removal of mass spectral peaks that had canonical loadings near zero. Fourteen mass spectral peaks that were correlated to the class membership of the samples were identified as potential biomarkers (see Figure 5.32).

The two samples that behaved as outliers were reanalyzed using CVA to better understand their relationship to the other samples in the data set. Each outlier was reintroduced in the training set or both outliers were introduced in the training set, and pattern recognition analysis was performed by the GA using CVAKNN. The results obtained for these three trials are shown in Figures 5.33, 5.34, and 5.35. The outliers are marked by circles in the score plots. The presence of these outliers in the training set diminished the separation between the three classes (see Figures 5.33 and 5.35) and lowered the classification success rates obtained for the 3 validation set samples (see Figures 5.33, 5.34, and 5.35). From these three plots, it is evident that deleting the two outliers that were identified by PCA during the beginning of this study was the correct course of action to be taken.

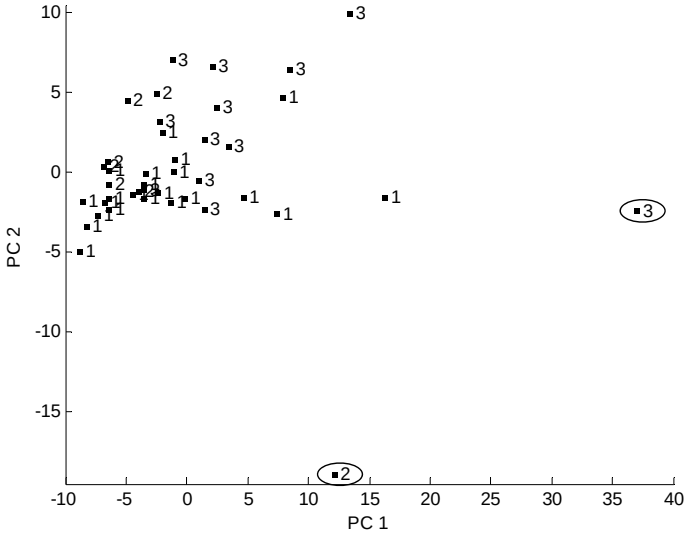

Figure 5.26. A plot of the two largest principal components developed from the 41 spectra (first data set) and 127 peaks. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Outliers are circled.

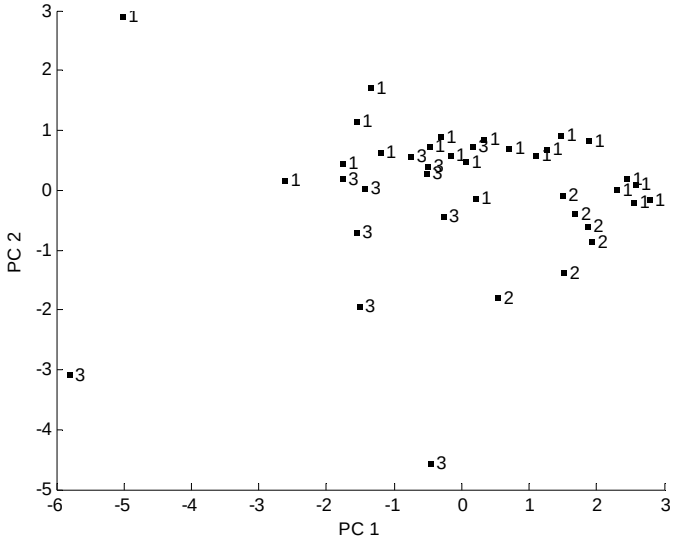

Figure 5.27. A plot of the two largest principal components developed from the 39 spectra (first data set) and 5 peaks identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma).

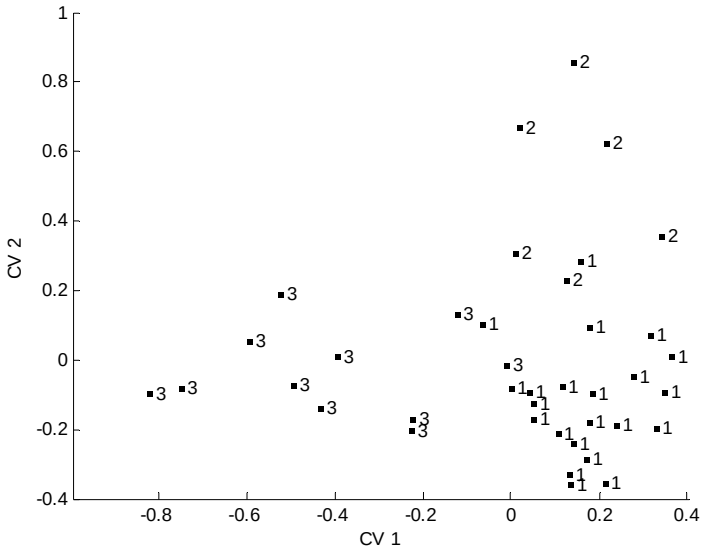

Figure 5.28. A plot of the two largest canonical variates developed from the 39 spectra (first data set) and 127 peaks. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma).

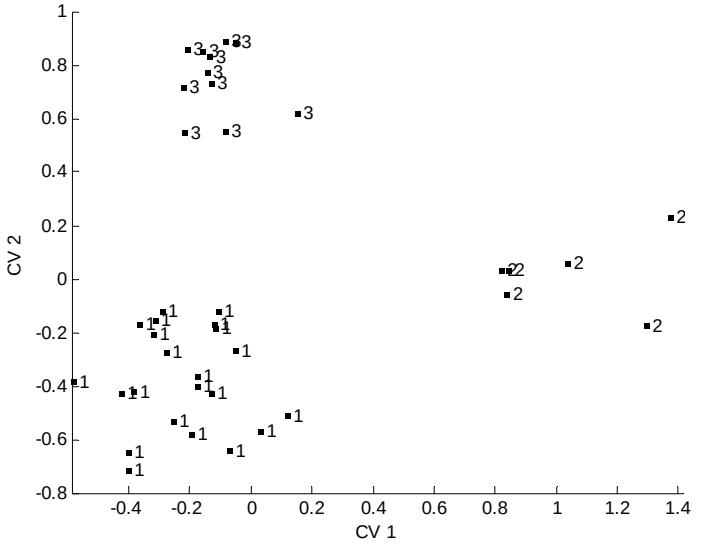

Figure 5.29. A plot of the two largest canonical variates developed from the 39 spectra (first data set) and 18 peaks selected by the pattern recognition GA. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma).

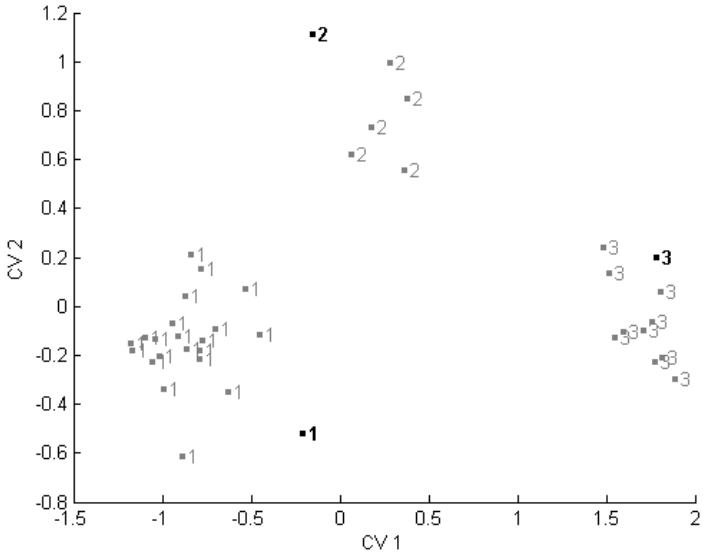

Figure 5.30. A plot of the two largest canonical variates developed from the 36 spectra (first data set) and 22 peaks selected by the pattern recognition GA. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black.

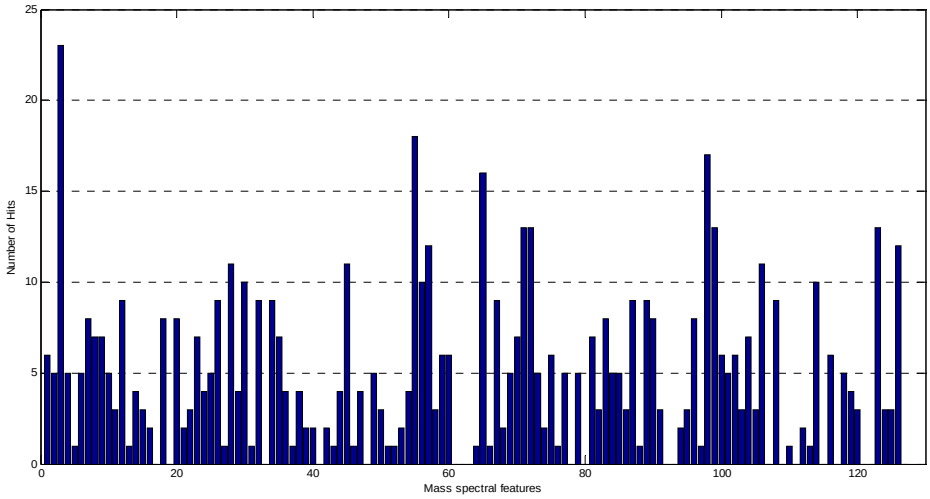

Figure 5.31. A histogram depicting the number of times each spectral feature was selected by the pattern recognition GA in each generation (number of hits) for schema hunting.

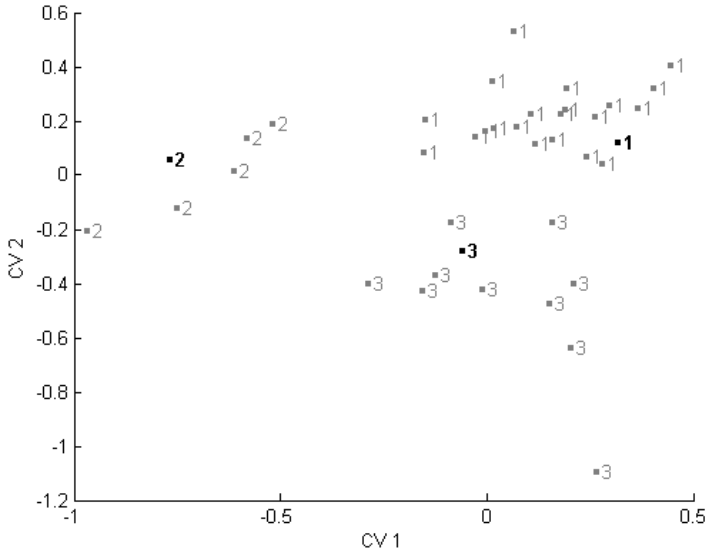

Figure 5.32. A plot of the two largest canonical variates developed from the 36 spectra (first data set) and 14 peaks identified by schema hunting and loading plots. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black.

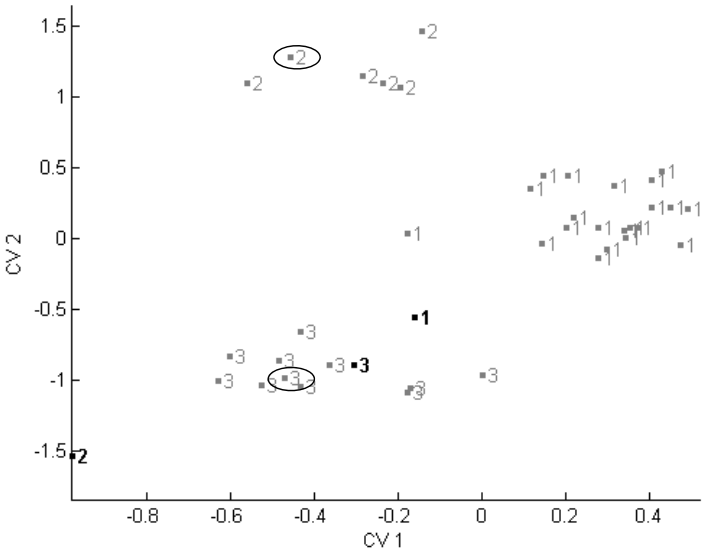

Figure 5.33. A plot of the two largest canonical variates developed from the 38 spectra (with the two outliers included) of the first data set and 23 peaks identified pattern recognition GA. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black. Outliers are circled.

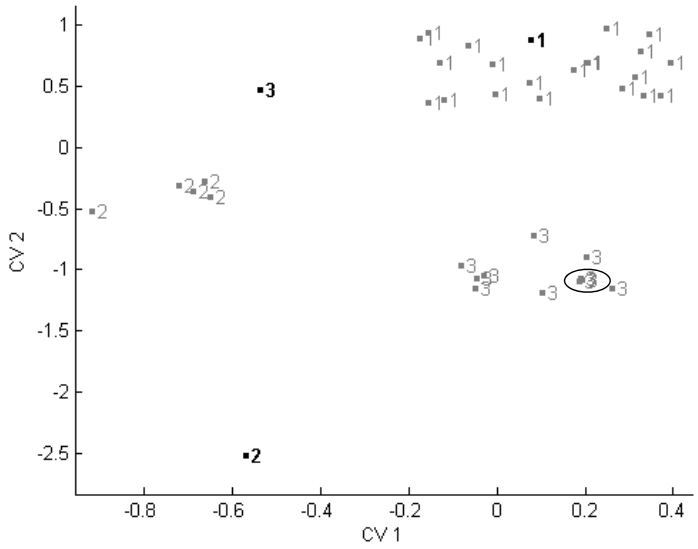

Figure 5.34. A plot of the two largest canonical variates developed from the 37 spectra (only one outlier included) of the first data set and 22 peaks identified pattern recognition GA. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black. Outlier is circled.

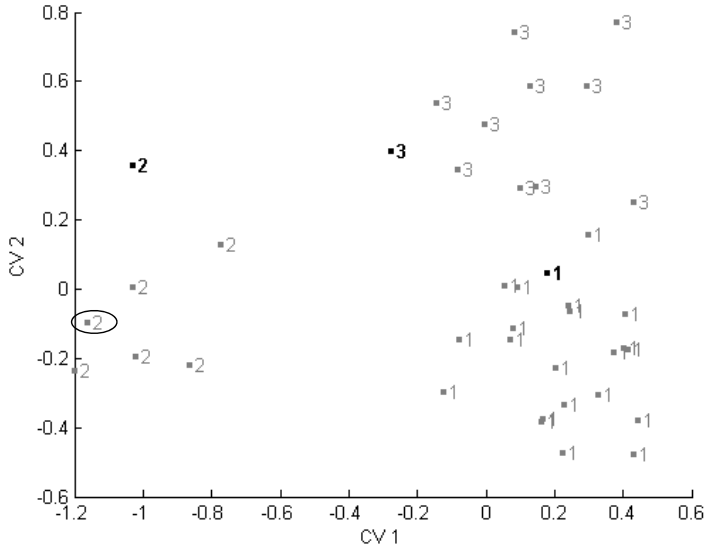

Figure 5.35. A plot of the two largest canonical variates developed from the 37 spectra (only one outlier included) of the first data set and 24 peaks identified pattern recognition GA. Each spectrum is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black. Outlier is circled.

The second data set of 32 MALDI-TOF profiles, which had 126 peaks, was compared with the first data set of 39 spectra (with the two outliers removed) using pattern recognition analysis. Spectra in both datasets were normalized to unit length and autoscaled prior to pattern recognition analysis. The two data sets were combined and the 71 samples were analyzed using both PCA and hierarchical clustering. The results of PCA and hierarchical clustering for the 71 samples and 126 peaks are shown in Figures 5.36 and 5.37 respectively. Two clusters were detected in the data. The first cluster consisted of the samples from the first data set and the second cluster consisted of the samples from the second data set. Evidently, the experimental conditions used to generate the first data set were not the same for the second data set. Although the same MALDI-TOF instrument was used to generate both data sets, it was not standardized during the generation of the second data set. This would explain the disparity in the MALDI-TOF profile of the samples from the two data sets.

The second data set was then investigated independent of the first data set using CVAKNN and object validation. For the 32 spectra, the pattern recognition GA identified 23 mass spectral features that provided discrimination between the three classes in the second MALDI-TOF data set (see Figure 5.38). The second data set was then divided into a training set of 28 spectra and a prediction set of 4 spectra. The prediction set samples were selected by random lot and included 1 sample that was a normal control, 1 chronic pancreatic sample, and 2 pancreatic adenocarcinoma samples. The 23 mass spectral features previously identified by the pattern recognition GA were used to classify the training set of 28 spectra and to predict the class membership of the 4 prediction set samples (see Figure 5.39) using CVA. Three of the 4 prediction set samples were

misclassified. A comparison was made with the first data set using this same approach. The 18 mass spectral features identified previously by the pattern recognition GA using 39 samples from the first data set (see Figure 5.29) were used to classify the 36 training set samples and to predict the class membership of the 3 prediction set samples (see Figure 5.40). All prediction set samples from data set 1 were correctly classified.

Several mass spectral features that could serve as potential biomarkers for pancreatic cancer were identified from the first data set. Each mass spectral data set was obtained under different conditions, which prevented these two data sets from being combined into a single data set or for the second data set to be used as a validation set for the first data set. There is also a problem with the quality of the mass spectra in the second data set which is probably due to the failure of the investigators to standardize the instrumental conditions during the mass spectral runs.

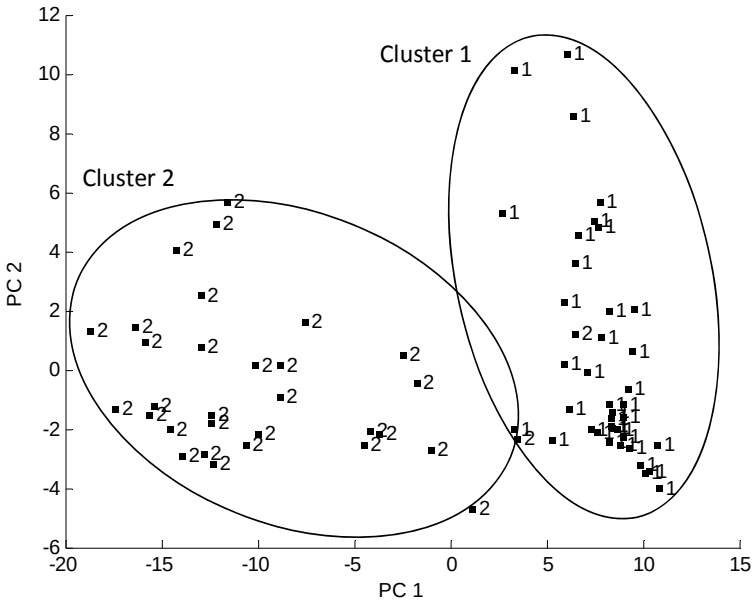

Figure 5.36. A plot of the two largest principal components developed from the 71 spectra (both data sets) and 126 peaks. Each spectrum is represented as a point in the PC score plot (l = First data set samples, 2 = Second data set samples).

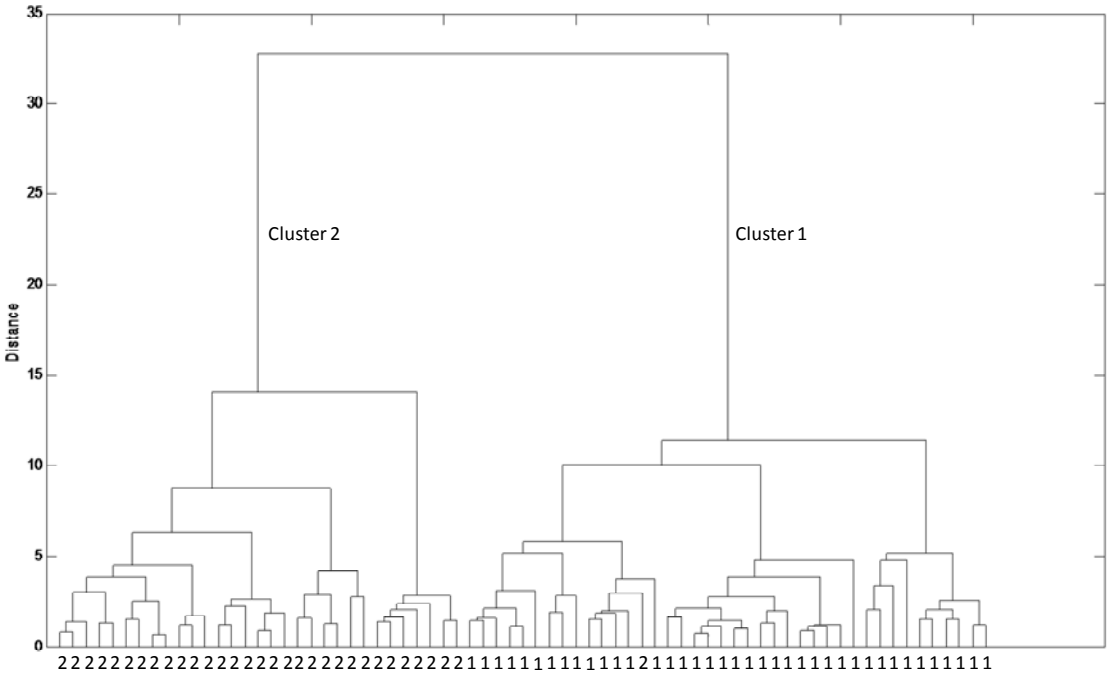

Figure 5.37. A hierarchical clustering (Wards) obtained by using 73 spectra (both data sets) and 126 peaks. (l = First data set samples, 2 = Second data set samples).

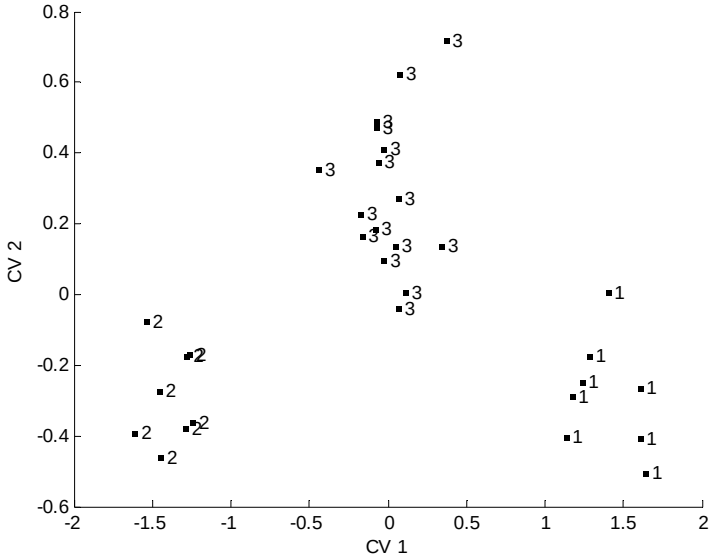

Figure 5.38. A plot of the two largest canonical variates developed from the 32 spectra (second data set) and 23 peaks selected by the pattern recognition GA. Each spectra is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma).

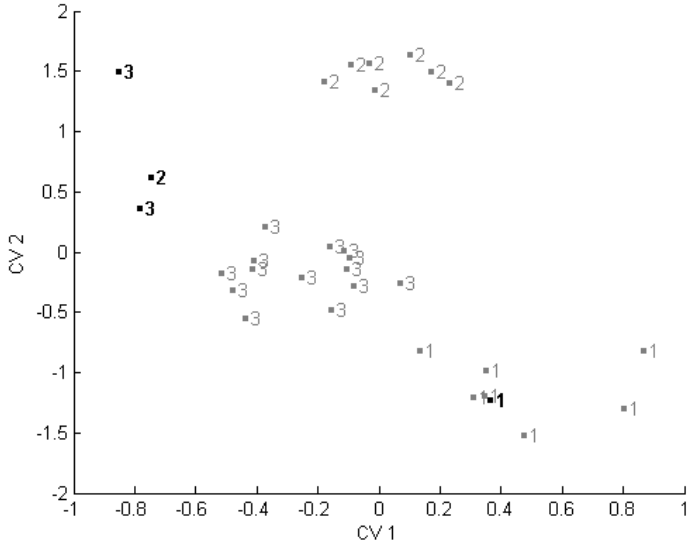

Figure 5.39. A plot of the two largest canonical variates developed from the 28 spectra (second data set) and 23 peaks selected previously by the pattern recognition GA. Each spectra is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black.

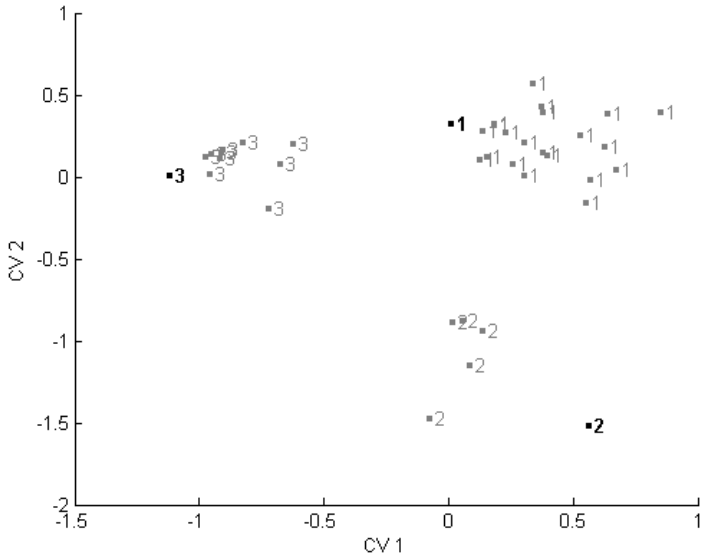

Figure 5.40. A plot of the two largest canonical variates developed from the 36 spectra (first data set) and 18 peaks selected previously by the pattern recognition GA. Each spectra is represented as a point in the CV score plot (l = Normal, 2 = Chronic Pancreatis, and 3 = Pancreatic Adenocarcinoma). Training set samples are in grey and the prediction set samples which are projected onto the CV map of the data are in black.

5.7 DISCOVERY OF BIOMARKER CANDIDATES FOR ESOPHAGEAL CANCER FROM IMS-MS DATA OF SERUM *N*-GLYCANS

The goal of this study was to identify potential biomarkers for esophageal cancer and its progression using IMS-MS data developed from serum *N*-glycans. Before the glycans data set was probed for biomarkers, the quality of the IMS-MS data was evaluated with respect to instrumental variability and variability associated with the procedure used for extraction of serum glycans.

The IMS-MS data collected for each sample were recorded as intensities for individual drift time bins and time-of-flight bins. A box extraction algorithm was used for extracting data within a specified range of drift time and m/z values and for peak alignment. The algorithm added intensities across a narrow m/z range around nominal m/z values for each drift time bin to generate drift time distributions. These ion mobility distributions are generated over the specified range of m/z and drift time bins. Each IMS-MS spectrum was represented as a data vector where the components of the vector are ion intensities obtained at specific drift times and m/z values. Ion mobility profiles were extracted for the m/z range 650-1600 Da across a drift time of 15-35 ms. Each IMS spectrum consisted of 5401 points. The spectra were normalized to constant sum of 1 and autoscaled before the analysis. Further details about the IMS-MS instrumental setup used to collect this data can be found elsewhere [5-48].

To assess the quality of the IMS-MS data, a test data set of 10 spectra was collected. This data set consisted of 5 replicate runs of the same sample worked up and run on the same day and 5 spectra of the same sample that was worked up and run on different days. A comparison of these spectra was performed using PCA. Each spectrum is represented as a point in the score plot. It is expected that replicate spectra would be

closer to each other and would cluster in a distinct region of the PC map, whereas the other 5 spectra that were prepared and run on different days would not form a well defined cluster due to variability from sources other than the instrument. Figure 5.41 shows a plot of the three largest principal components developed from the 10 spectra and 5401 peaks. Each spectrum is represented as a point in the PC plot (1 = replicates, 2 = samples worked up and collected on different days). The five replicates cluster in a distinct region of the PC plot whereas the other 5 samples occupy a larger region of the map.

Another data set was developed by extracting ion mobility profiles corresponding to 6 known glycans for these 10 spectra. Each spectrum was developed by combining ion mobility profiles that were obtained by using the box extraction algorithm for 6 regions of the IMS-MS data (m/z bin values 826.2, 903.5, 946.6, 1086.3, 1217, and 1408.2) that corresponded to these 6 glycans. Each mass spectrum consisted of 2831 peaks. Figure 5.42 shows a plot of the three largest principal components developed from the 10 spectra and 2831 peaks of the 6 glycans. Differences between the 5 replicates versus the other 5 spectra were not as pronounced when only peaks from known glycans were used, which indicates that experimental artifacts in the data are a less serious problem when the analysis is focused on the glycans. When analyzing these PC plots, one must realize in advance that variability is scaled to the samples that comprise the plot. A minor source of variability can be magnified depending on the samples used to develop the PC plot. For this reason, the design of the training set is crucial for assessing the magnitude of potential sources of variability in data.

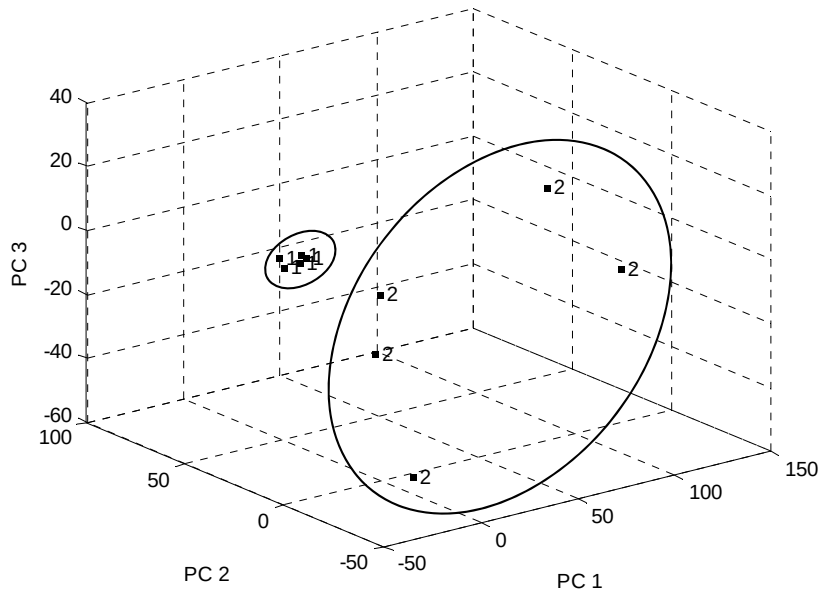

Figure 5.41. A plot of the three largest principal components developed from the 10 spectra and 5401 peaks. Each spectrum is represented as a point in the PC plot. (1 = replicates, 2 = samples worked up and collected on different days).

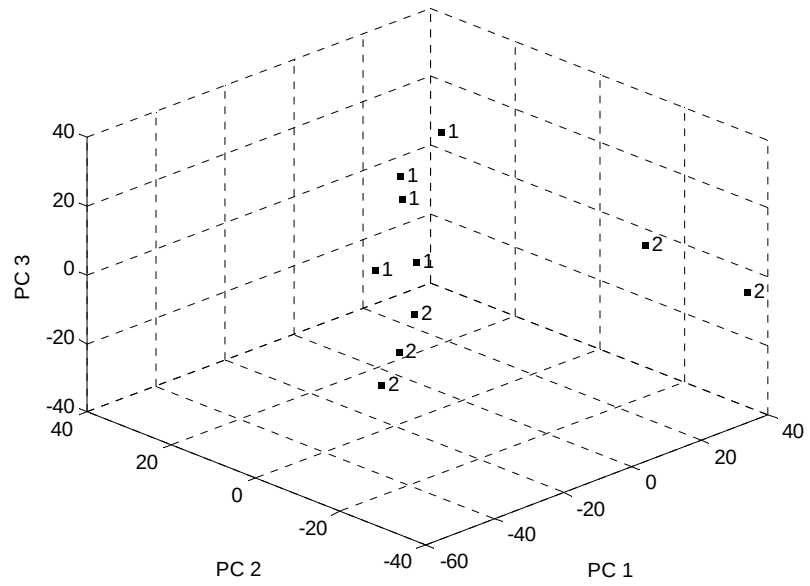

Figure 5.42. A plot of the three largest principal components developed from the 10 spectra and 2831 peaks of the 6 glycans. Each spectrum is represented as a point in the PC plot. (1 = replicates, 2 = samples worked up and collected on different days)

IMS-MS data of *N*-glycans was collected from serum samples for the identification of potential biomarkers associated with esophageal cancer and its progression. The data set was composed of 136 human serum samples collected from donors that were normal controls (NC) and those that expressed three different stages of esophageal cancer: Barrett's esophagus (BE), high-grade dysplasia (HGD), and esophageal adenocarcinoma (EAC). The composition of the glycan data set is shown in Table 5.15. *N*-glycans were extracted from the 136 serum samples followed by the generation of IMS-MS data. Each ion mobility spectra contained 1431 points that represented 11 *N*-glycans associated with the following m/z bin values: 763.4, 801.4, 825.7, 883.8, 946.1, 1004.2, 1095.2, 1216.3, 1227.1, 1274.3, and 1407.7.

Table 5-15. Composition of the IMS-MS Data Set

Sample Type	Number of Samples
Normal Controls (NC)	61
Esophageal Adenocarcinoma (EAC)	56
Barrett's Esophagus (BE)	7
High-grade Dysplasia (HGD)	12
Total	136

Outlier analysis was performed on each class in the data set using the generalized distance test using SCOUT [5-49], PCA (using PC score plots and sample leverages via ADAPT) and visual analysis of the IMS-MS spectra. Seven EAC samples were identified as outliers. Three of these samples were dirty and a visual analysis of the IMS spectra of the other 4 samples indicated the presence of artifacts in the data. As these 7 samples were from the same lot, this suggested that problems occurred during sample processing. For the NC samples, a plot of the three largest principal components developed with the 1431 spectral features (see Figure 5.43) indicated the presence of 3 distinct clusters in the

data. This clustering was also attributed to differences in the sample work up. Eight NC samples (cluster 1) were removed from the data as they were detected to be outliers from the generalized distance test. One BE sample was also identified as an outlier from the generalized distance test and from its leverage. HGD also exhibited clustering (see Figure 5.44) with cluster 1 containing the outliers (6 samples) as determined by a visual analysis of their spectra. Twenty two samples were judged to be outliers and were removed from the analysis due to the low quality of their spectra.

The pattern recognition GA was used to identify features for the discrimination of NC samples from EAC samples. The pattern recognition GA identified informative features by sampling key feature subsets, scoring their principal component plots, and tracking samples that were difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the GA identified 46 IMS features that showed separation between the samples of the two classes on a PC plot (see Figure 5.45). These 46 spectral features were found to be related to four glycans.

The results from this study indicate that a relationship exists between the conditions used to process the samples and the IMS-MS spectra obtained. It will be crucial in future studies to standardize the conditions used for glycan extraction as this will decrease variability within a class and also reduce the occurrence of outliers in the data.

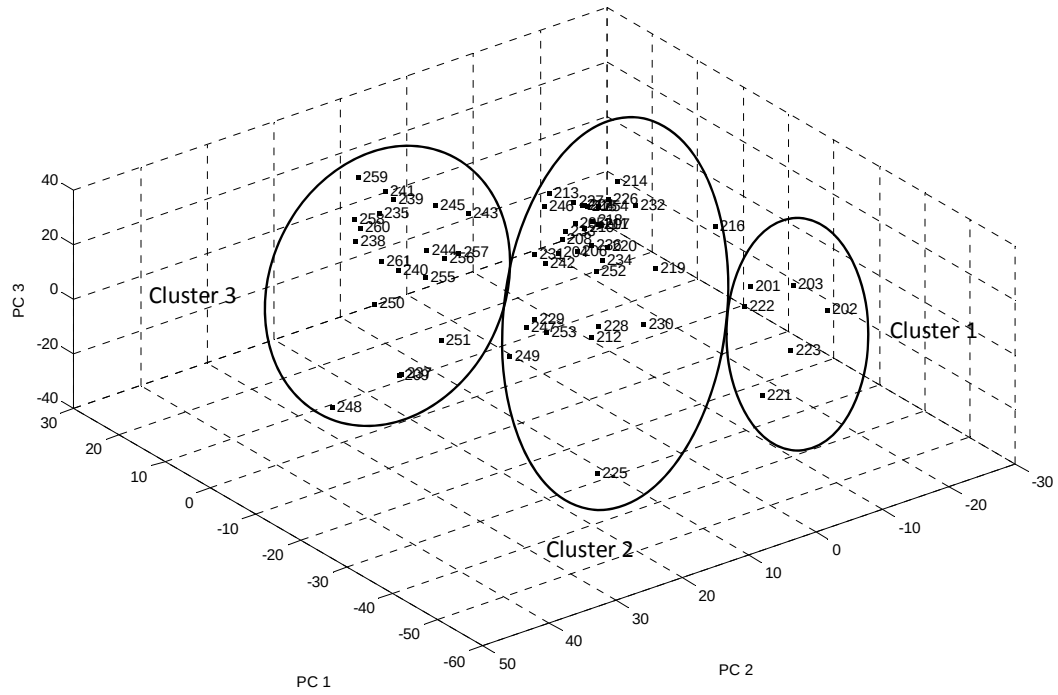

Figure 5.43. A plot of the three largest principal components developed from the 61 normal control (NC) samples and 1431 spectral features. 3 distinct clusters are observed.

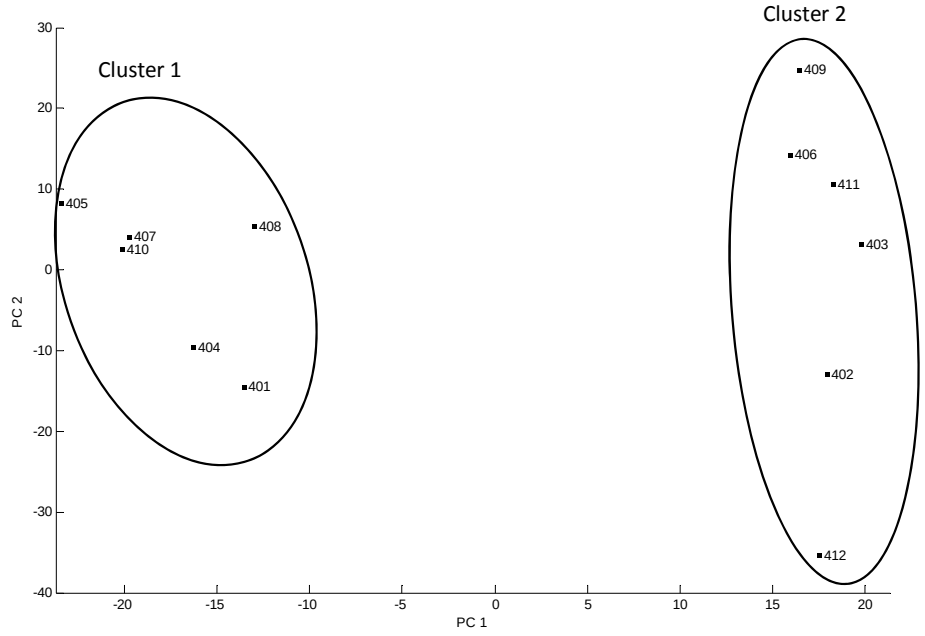

Figure 5.44. A plot of the two largest principal components developed from the 12 HGD samples and 1431 spectral features. 2 distinct clusters are observed.

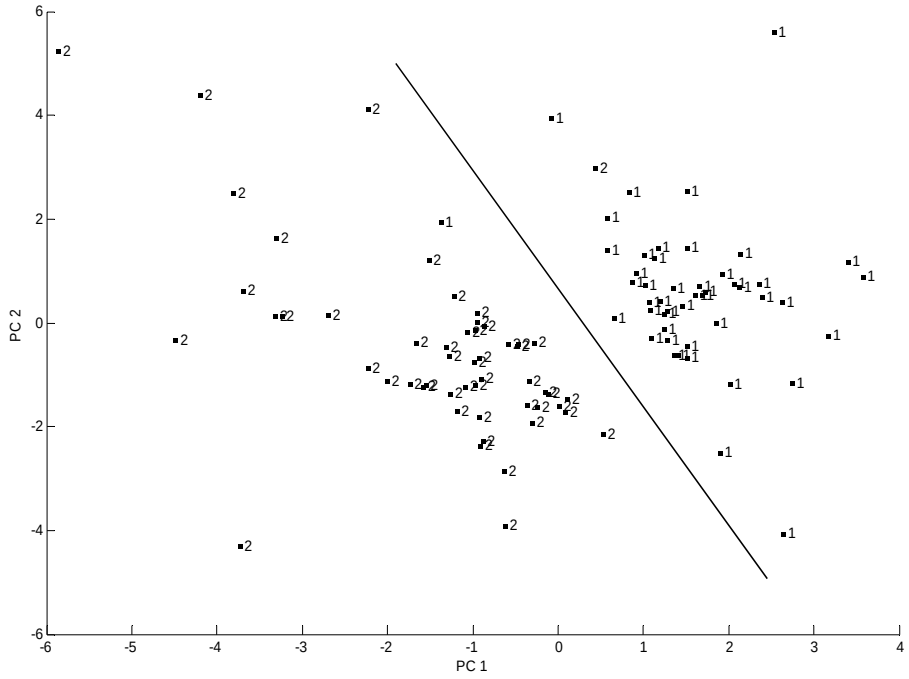

Figure 5.45. A plot of the two largest principal components developed from the 102 IMS-MS spectra and 46 spectral features identified by the pattern recognition GA. Each spectrum is represented as a point in the PC score plot (l = NC, 2 = EAC).

5.8 DISCOVERY OF BIOMARKER CANDIDATES FOR ESOPHAGEAL CANCER FROM MALDI-TOF DATA OF SERUM *N*-GLYCANS

The goal of this study was to identify potential biomarkers for esophageal cancer from MALDI-TOF data developed from serum *N*-glycans. The pattern recognition GA was used to identify features in the MALDI-TOF data that could serve as potential biomarkers for esophageal cancer. The mass spectral data set developed in this study consisted of 82 spectra that represented serum samples for normal controls and three different stages of esophageal cancer: Barrett's esophagus (BE), high-grade dysplasia (HGD), and esophageal adenocarcinoma (EAC). The composition of the MALDI-TOF data set is shown in Table 5.16.

The MALDI-TOF spectra were generated from *N*-glycans extracted and processed from serum samples. The experimental conditions and procedure used for preparation of samples and procurement of MALDI-TOF data was similar to that used in a previous study and can be found elsewhere [5-45 to 5-47]. Each spectrum was represented by 514 spectral features over the m/z range of 1580 – 5490 Da. The data set was divided into a training set of 75 spectra and a prediction set of 7 spectra, with the samples for the prediction set chosen by random lot (see Table 5.16).

Table 5-16. Composition of the MALDI-TOF Data Set

Sample Type	Number of Samples	Number of Samples in Training Set	Number of Samples in Prediction Set
Normal Controls (NC)	18	16	2
Esophageal Adenocarcinoma (EAC)	48	44	4
Barrett's Esophagus (BE)	5	5	-
High-grade Dysplasia (HGD)	11	10	1
Total	82	75	7

The mass spectral data was analyzed using PCA. Figure 5.46 shows a plot of the two largest principal components developed from the 75 spectra and 514 mass spectral features. Each spectrum is represented as a point in the PC score plot. No separation was observed between samples from the four classes (1 = NC, 2 = EAC, 3 = BE, and 4 = HGD) in the PC map of the data.

The pattern recognition GA was used for the analysis of the data. The GA identified potential biomarkers for esophageal cancer by sampling key feature subsets, scoring their principal component (PC) plots, and tracking those samples and/or classes that were most difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 200 generations, the genetic algorithm identified

25 spectral peaks whose PC plot showed clustering of the mass spectral data by sample type. The NC, EAC, and BE samples were all well separated from each other in a plot of the two largest principal components of the data (see Figure 5.47). However, two HGD samples overlapped with the normals.

The prediction set was employed to assess the predictive ability of the 25 mass spectral peaks identified by the pattern recognition GA. The 7 prediction set spectra (black) were projected directly onto the PC score plot developed from the 75 spectra of the training set (grey) and 25 mass spectral features (see Figure 5.48). Six of the 7 projected spectra lie in a region of the PC map with samples that have the same class label. Evidently, the GA can identify features from the mass spectra that are correlated to the diseased state of the subject.

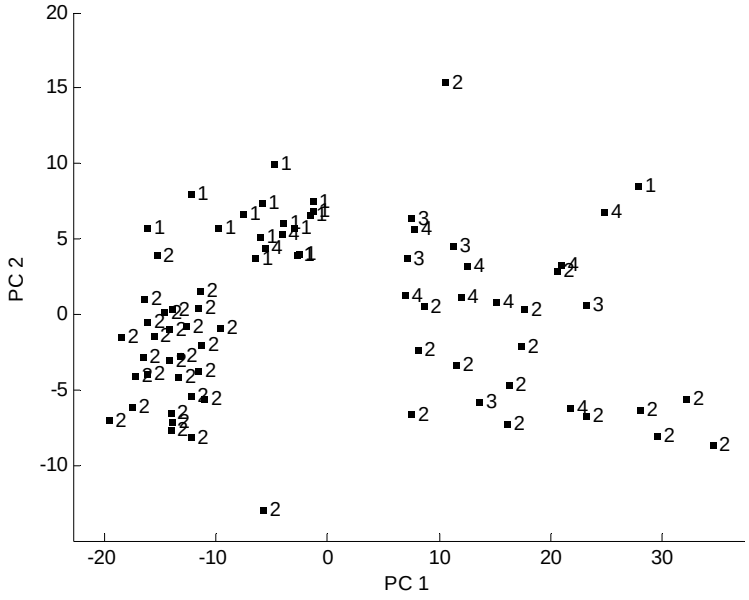

Figure 5.46. A plot of the two largest principal components developed from the 75 spectra and 514 mass spectral features. Each spectrum is represented as a point in the PC score plot (1 = NC, 2 = EAC, 3 = BE, and 4 = HGD).

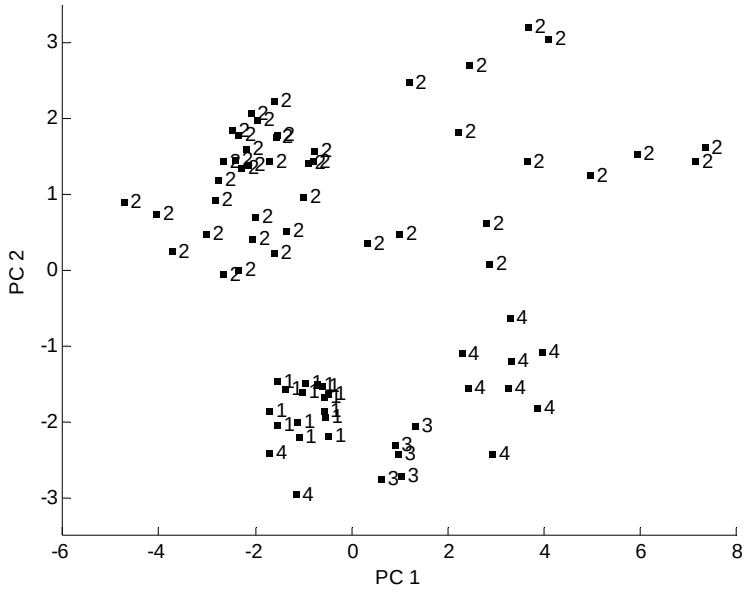

Figure 5.47. A plot of the two largest principal components developed from the 75 spectra and 25 mass spectral features. Each spectrum is represented as a point in the PC score plot (1 = NC, 2 = EAC, 3 = BE, and 4 = HGD).

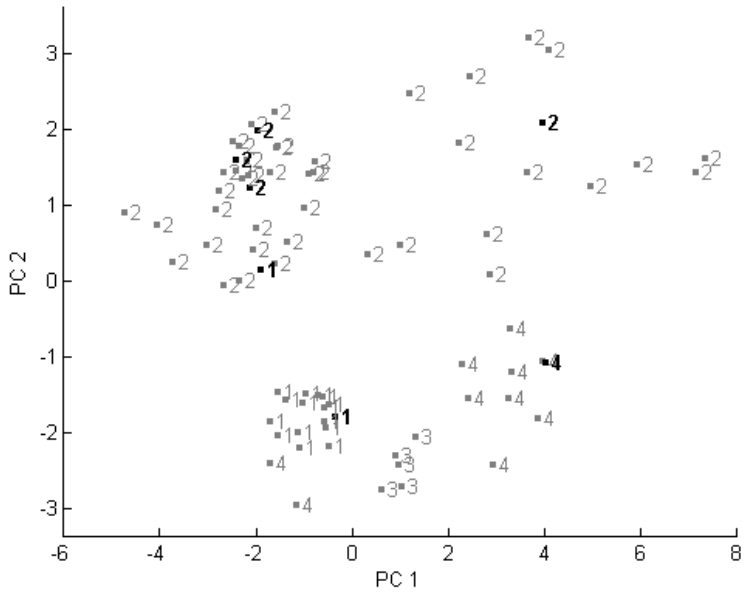

Figure 5.48. A plot of the two largest principal components developed from the 75 spectra and 25 mass spectral features. Each spectrum is represented as a point in the PC score plot (1 = NC, 2 = EAC, 3 = BE, and 4 = HGD). Training set samples are in grey and the prediction set samples which are projected onto the PC map of the data are in black.

REFERENCES

5-1. E.F. Petricoin, A.M. Adrekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, "use of proteomic patterns in serum to identify ovarian cancer", Lancet, 2002, 359, 572-577.

5-2. M.E. de Noo, B.J. Mertens, A. Ozalp, M.R. Bladergroen, M.P. van der Werff, C.J. van de Velde, A.M. Deelder, R.A. Tollenaar, "Detection of colorectal cancer using MALDI-TOF serum protein profiling", Eur. J. Cancer, 2006, 42, 1068-1076.

5-3. J. Villanueva, D.R. Shaffer, J. Phillip, C.A. Chaparro, H. Erdjument-Bromage, A.B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E.C. Holland, C. Cordon-Cardo, H.I. Scher, P. Tempst, "Differential exoprotease activities confer tumor-specific serum peptidome patterns", J. Clinical Investigation, 2006, 116, 271-284.

5-4. B.K. Lavine, N. Mirjankar, R. LeBouf, and A. Rossner, "Prediction of Mold Contamination from Microbial Volatile Organic Compound Profiles Using Solid Phase Microextraction and Gas Chromatography/Mass Spectrometry," Microchemical Journal, 2012, 103, 37-41.

5-5. J.D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis", Nat. Rev. Genet., 2006, 7, 200-210.

5-6. A. Balmain, J. Gray, B. Ponder, "The genetics and genomics of cancer", Nature Genet., 2003, 33, 238-244.

5-7. L.J. van′t Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, "Gene-expression profiling predicts clinical outcome of breast cancer", Nature, 2002, 415, 530-536.

5-8. P.A. Binz, D.F. Hochstrasser, and R.D. Appel, "Mass spectrometry based proteomics: Current status and potential use in clinical chemistry", Clinical Chem. Lab. Med., 41, 1540–1551.

5-9. R. Aebersold, and M. Mann, "Mass spectrometry-based proteomics", Nature, 2003, 422, 198–207.

5-10. P.L. Ferguson, and R.D. Smith, “Proteome analysis by mass spectrometry”. Annual Review of Biophysics and Biomolecular Structure, 2003, 32, 399-424.

5-11. B.K. Lavine, N. Mirjankar, R. LeBouf, A. Rossner, "Prediction of mold contamination from microbial volatile organic compound profiles using head space gas chromatography/mass spectrometry", Microchem. J., 2012, 103, 119-124.

5-12. B.K. Lavine, K. Nuguru, and N. Mirjankar, “One stop shopping - feature selection, classification, and prediction in a single step”, J. Chemom., 2011, 25, 116-129.

5-13. T. G. Lee, “Health symptoms caused by molds in a courthouse,” Archives of Environ. Health, 2003, 58, 442-446.

5-14. M. R. Gray, J. D. Thrasher, R. Crago, R. A. Madison, L. Arnold, A. W. Campbell, and A. Vojdani, “Mixed mold mycotoxicosis: immunological changes in humans following exposure in water-damaged buildings, “ Archives of Environ. Health, 2003, 58, 410-420.

5-15. F. Fung, D. Tappen, and G. Wood, “Alternaria-associated asthma,” Appl. Occ. Environ. Hyg., 2000, 15, 924-927.

5-16. P. Carrer, M. Maroni, D. Alcini, and D. Cavallo, “Allergens in indoor air: environmental assessment and health effects,” Sci. of Total Environ., 2001, 270, 33-42.

5-17. U. Gehring, J. Douwes, G. Doekes, A. Koch, W. Bischof, B. Fahlbusch, K. Richter, H. Wichmann, and J. Heinrich, “β(1-3)-Glucan in house dust of German homes: Housing characteristics, occupant behavior, and relations with endotoxins, allergens and molds” Environ. Health Perspect, 2001, 109, 139-144.

5-18. R. Savilahti, J. Uitti, P. Laippala, T. Husman, and P. Roto, “Respiratory morbidity among children following renovation of water-damaged school”, Arch. Environ. Health, 2000, 55, 405-410.

5-19. J. Bunger, G. Westphal, A. Monnich, B. Hinnendahl, E. Hallier, and M. Muller, “Cytotoxicity of occupationally and environmentally relevant mycotoxins, Toxicology”, 2004, 202, 199-211.

5-20. J. D. Spengler, J. K. Jaakkola, H. Parise, B. A. Katsnelson, L. I. Privalova, and A. A. Kosheleva, “Housing characteristics and children respiratory health in the Russian Federation”, Am J. Public Health, 2004, 94, 657-662.

5-21. H. A. Burge, “Fungi: toxic killers or unavoidable nuisance”, Ann. Allergy Asthma Immunol., 2001, 87, 52-56.

5-22. R. L. LeBouf, S. A. Schuckers, and A. Rossner, "Preliminary assessment of a model to predict mold contamination based on microbial volatile organic compound profiles," Sci. of Total Environ., 2010, 408, 3648-3653.

5-23. R. F. LeBouf, C. Casteel, and A. Rossner, "Evaluation of air sampling technique for assessing low-level volatile organic compounds in indoor environments," J. Air & Waste Management Assoc., 2010, 60, 156-162

5-24. C. Yang and P. A. Heinsohn, "Sampling and Analysis of Indoor Microorganisms", John Wiley & Sons, New York, 2007.

5-25. A. Andersen, "New sampler for the collection, sizing, and enumeration of viable airborne particles," J. Bacteriol., 1958, 76, 471-484.

5-26. J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine, 2001, 7, 673-679.

5-27. M.S. Bereman, T. I. Williams, D.C. Muddiman, "Development of a nanoLC LTQ Orbitrap Mass Spectrometric Method for Profiling Glycans Derived from Plasma from Healthy, Benign Tumor Control, and Epithelial Ovarian Cancer Patients", Anal. Chem., 2008, 81, 1130-1136.

5-28. D. Isailovic, R.T. Kurulugama, M.D. Plasencia, S.T. Stokes, Z. Kyselova, R. Goldman, Y. Mechref, M.V. Novotny, D.E. Clemmer, "Profiling of human serum glycans associated with liver cancer and cirrhosis by IMS-MS", J. Proteome Res., 2008, 7, 1109-1117.

5-29. Y. Mechref, A. Hussein, S. Bekesova, V. Pungpapong, M. Zhang, L.E. Dobrolecki, R.J. Hickey, Z.T. Hammond, M.V. Novotny, "Quantitative serum glycomics of esophageal adenocarcinoma, and other esophageal disease onsets", J. Protoeme Res., 2009, 8, 2656-2666.

5-30. M.S. Bereman, D.D. Young, A. Deiters, D.C. Muddiman, "Development of a robust and high throughput method for profiling N-linked glycans derived from plasma glycoproteins by nanoLC-FTICR mass spectrometry", J. Protoeme Res. 2009, 8, 3764-3770.

5-31. Z. Kyselova, Y. Mechref, M.M. Al Bataineh, L.E. Dobrolecki, R.J. Hickey, J. Vinson, C.J. Sweeney, M.V. Novotny, "Alterations in the serum glycome due to metastatic prostate cancer:, J. Proteome Res., 2007, 6, 1822-1832.

5-32. R. Saldova, L. Royle, C.M. Radcliffe, U.M. Abd Hamid, R. Evans, J.N. Arnold, R.E. Banks, R. Hutson, D.J. Harvey, R. Antrobus, S.M. Petrescu, R.A. Dwek, P.M. Rudd, "Ovarian cancer is associated with Changes in glycosylation in both acute phase proteins and IgG", Glycobiology, 2007, 17, 1344–56.

5-33. R. Goldman, H.W. Ressom, R.S. Varghese, L. Goldman, G. Bascug, C.A. Loffredo, M. Abdel-Hamid, I. Gouda, S. Ezzat, Z. Kyselova, Y. Mechref, M.V. Novotny, "Detection of hepatocellular carcinoma using glycomic analysis", Clin. Cancer Res., 2009, 15, 1808-1813.

5-34. J.N. Arnold, R. Saldova, U.M.A. Hamid, P.M. Rudd, "Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation", Proteomics, 2008, 8, 3284-3293.

5-35. J.N. Arnold, R. Saldova, M.C. Galligan, T.B. Murphy, Y. Mimura-Kimura, J.E. Telford, A.K. Godwin, P.M. Rudd, "Novel glycan biomarkers for the detection of lung cancer", J. Proteome Res., 2011, 10, 1755-1764.

5-36. C. Robbe-Masselot, A. Herrmann, E. Maes, I. Carlstedt, J.C. Michalski, C. Capon, "Expression of a core 3 Disialyl-Le(x) hexasaccharide in human colorectal cancers: A potential marker of malignant transformation in colon", J. Proteome Res., 2009, 8, 702-711.

5-37. A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart, J. Marth, (editors) "Essentials of Glycobiology", 2nd edition, CSHL Press, 2009.

5-38. C.B. Lebrilla, H.J. An, "The prospects of glycan biomarkers for the diagnosis of diseases". Mol. BioSyst., 2009, 5, 17-20.

5-39. M.B. West, Z.M. Segu, C.L. Feasley, P. Kang, I. Klouckova, C. Li, M.V. Novotny, C.M. West, Y.Mechref, and M.H. Hanigan, "Analysis of site-specific glycosylation of renal and hepatic γ-glutamyl transpeptidase from normal human Tissue", J. Biol. Chem., 2010, 285, 29511-29524.

5-40. P. Kang, Y. Mechref, I. Klouckova, M.V. Novotny, "Solid-phase permethylation of glycans for mass spectrometric analysis", Rapid Commun. Mass Spectrom., 2005, 19, 3421-3428.

5-41. A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell and S.M. Haslam, "GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans" J. Proteome Res., 2008, 7, 1650-1659.

5-42. M. Atlas, S. Datta, "A Statistical Technique for Monoisotopic Peak Detection in a Mass Spectrum", J. Proteomics Bioinform., 2009, 2, 202-216.

5-43. D. Isailovic, M.D. Plasencia, M.M. Gaye, S.T. Stokes, R.T. Kurulugama, V. Pungpapong, M. Zhang, Z. Kyselova, R. Goldman, Y. Mechref, M.V. Novotny, and D.E. Clemmer, "Delineating diseases by IMS-MS profiling of serum N-linked glycans", J. Proteome Res., 2012, 11, 576–585.

5-44. D. Isailovic, R.T. Kurulugama, M.D. Plasencia, S.T. Stokes, Z. Kyselova, R. Goldman, Y. Mechref, M.V. Novotny, D.E. Clemmer, "Profiling of human serum

glycans associated with liver cancer and cirrhosis by IMS-MS", J. Proteome Res., 2008, 7, 1109-1117.

5-45. H.W. Ressom, R.S. Verghese, L. Goldman, C.A. Loffredo, M. Abdel-Hamid, Z. Kyselova, Y. Mechref, M. Novotny, R. Goldman, " Analysis of MALDI-TOF mass spectrometry data for detection of glycan biomarkers", Pac. Symp. Biocomput., 2008, 216-227.

5-46. R. Goldman, H.W. Ressom, R.S. Varghese, L. Goldman, G. Bascug, C.A. Loffredo, M. Abdel-Hamid, I. Gouda, S. Ezzat, Z. Kyselova, Y. Mechref, and M.V. Novotny, "Detection of hepatocellular carcinoma using glycomic analysis", Clin. Cancer Res., 2009, 15, 1808-1813.

5-47. Z. Tang, R.S. Varghese, S. Bekesova, C.A. Loffredo, M.A. Hamid, Z. Kyselova, Y. Mechref, M.V. Novotny, R. Goldman, and H.W. Ressom, "Identification of *N*-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data", J. Proteome Res., 2010, 9, 104-112.

5-48. M. Gaye, S.J. Valentine, Y. Hu, Y. Mechref, B.K. Lavine, N. Mirjankar, and D.E. Clemmer, "Characterization of serum N-linked glycans associated with esophageal adenocarcinoma and its progression by IMS-MS", Anal. Chem., 2012, IN PRESS.

5-49. M.A. Stapanian, F.C. Garner, K.E. Fitsgerald, G.T. Flatman, J.M. Nocerino, J. Chemom., 1993, 7, 165-176.

CHAPTER VI

CONCLUSION

In the preceding chapters, a basic methodology for analyzing large multivariate chemical data sets based on feature selection is described. A chromatogram or spectrum is represented as a point in a high dimensional measurement space. Exploratory data analysis techniques (principal component analysis and clustering) are then used to investigate the properties of this measurement space. A genetic algorithm for feature selection and classification is then applied to the data to identify features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. A good principal component plot can only be generated using features whose variance or information is primarily about differences between classes in the data. Hence, feature subsets that maximize the ratio of between-class to within-class variance are selected by the pattern recognition GA. Furthermore, the structure of the data set can be explored, for example, we can discover new classes, by simply tuning K_c in PCKaNN and by varying the relative contribution of PCKaNN and the deweighted or modified Hopkins statistic to the overall fitness score, ensuring a careful analysis of the data.

The pattern recognition GA has been validated on a wide range of data. In three studies involving spectral library searching, the use of wavelets and the pattern recognition GA as a general solution to problems in spectral pattern recognition was demonstrated. In one study, differential mobility spectra of VOCs were analyzed for structural content by chemical family. In another study, a search prefilter to detect the presence of carboxylic acids from vapor phase IR spectra has been successfully formulated and validated. In a third study, this same approach has been used to develop a pattern recognition assisted infrared library searching technique to determine the model, manufacturer, and year of the vehicle from which a clear coat paint smear originated. Because modern automotive paints are using thinner undercoat and color coat layers, protected by a thicker clear coat layer, all too often only a clear coat paint smear is the only layer of paint left at the crime scene. In these cases, PDQ database, which is used to identify the automobile from which a paint sample originates, cannot be used to identify the motor vehicle. PDQ cannot differentiate between similar but nonidentical FTIR paint spectra. This result is a major breakthrough in the area of trace evidence.

In addition, the pattern recognition GA has been used to develop a potential method to identify molds in indoor environments using VOCs. A distinct profile indicative of microbial VOCs was developed from air sampling data that could be readily differentiated from the blank for both high mold count and moderate mold count exposure samples. The pattern or profile of the VOCs is more important than individual quantities or total quantities of VOCs, i.e. the profile itself is the unique identifier

Analytical methods often generate such complex data that multivariate and pattern recognition methods must be used for their analysis. The studies described in this thesis

afforded us an opportunity to work on a variety of problems with different characteristics which necessitated the development of new mathematical and statistical methods for their solution. Although the pattern recognition GA has proven successful in a variety of studies, further research and development will be necessary to ensure that this approach becomes part of the routine practice of analytical chemists for data interpretation. Improvements for the pattern recognition GA that need to be undertaken are listed below.

1. Applying FCV false color data imaging [6-1], a method for increasing the information content by a PC plot, further aiding in its interpretation.

2. Estimation of the uncertainty of the scores in PC plots using jackknife and boostrap resampling approaches [6-2] which will involve rotation of each resampled PC model to ensure that it matches with a reference model eliminating rotational ambiguity.

3. Fitness functions for the pattern recognition GA must be formulated that will allow searches of the data space for significant structure by identifying features that increase the clustering of the data using ideas taken from the field of thin positions for knots and 3-dimensional manifolds to detect clusters in data subspaces [6-3, and 6-4]. Advantages of using these clustering algorithms are two-fold: (1) data preprocessing will not be a critical issue which is not the situation when using PCA, and (2) both linear and nonlinear manifolds in data subspaces can be detected.

4. Fitness functions for the pattern recognition GA must be formulation that will allow for data fusion where data from multiple databases are combined and information in the form of actionable items is extracted, e.g., the class membership of a sample. Data fusion has the potential to increase both the robustness and the selectivity of a classification

5. A practical limitation of classification occurs when an existing model is applied to data measured under new sampling or environmental conditions or on a different instrument. Even if samples with identical amounts of analyte are measured, the variation that is captured by the model will differ because of the different contributions from the sample matrix, the instrumental functions, and the environment of the measurement. For this reason, a model developed using data from one instrument generally cannot be used on data from a second instrument to provide accurate estimates of calibrated property values. For this reason, new methods must be developed to transfer a classification model between different instruments.

Pattern recognition methods operate with well defined criteria and attempt to extract useful information from raw data. If the limitations of the methods are not fully understood, the danger of misinterpretation and misuse of costly measurements is significant. It is our opinion that multivariate analysis techniques such as principal component analysis and discriminant analysis should be used to extend the ability of human pattern recognition to uncover hidden relationships in chemical data. The pattern recognition GA described in this thesis relies heavily on graphics for the presentation of results. Although the computer can assimilate greater quantities of data at any given time

than can the chemist, it is the chemist, who in the end must make the decisions and judgments about the problem.

.

REFERENCES

6-1. B.K. Lavine, A. B. Stine, H. Mayfield, and R. Gunderson, "Application of High-Resolution Computer Graphics to Pattern Recognition Analysis," J. Chem. Inf. Comp. Sci., 1993, 33, 826-834.

6-2. O. Preisner, J. A. Lopes, and J. C. Menzes, “Uncertainty Assessment in FT-IR Spectroscopy Based Bacteria Classification Models,” Chem. Intell. Lab. Systems, 2008, 94, 33-42.

6-3. D. R. Heisterkamp and J. Johnson, “Pinch Ratio Clustering from a Topologically Intrinsic Lexicographic Ordering,” Proceedings of 2012 IEEE International Conference on Data Mining (ICDM 2012), Brussels, Belgium, 2012, submitted.

6-4. J. Johnson and D.R. Heisterkamp, "Topological Graph Clustering with Thin Position,” Geometriae Dedicata, 2012, submitted.

REFERENCES

1. R. Aebersold, and M. Mann, “Mass spectrometry-based proteomics”, Nature, 2003, 422, 198–207.

2. R.J. Anderegg and D.J. Pyo, “Selective reduction of infrared data.” Anal. Chem., 1987, 59, 1914-1917.

3. A. Andersen, “New sampler for the collection, sizing, and enumeration of viable airborne particles,” J. Bacteriol., 1958, 76, 471-484.

4. J.N. Arnold, R. Saldova, M.C. Galligan, T.B. Murphy, Y. Mimura-Kimura, J.E. Telford, A.K. Godwin, P.M. Rudd, “Novel glycan biomarkers for the detection of lung cancer”, J. Proteome Res., 2011, 10, 1755-1764.

5. J.N. Arnold, R. Saldova, U.M.A. Hamid, P.M. Rudd, “Evaluation of the serum N-linked glycome for the diagnosis of cancer and chronic inflammation”, Proteomics, 2008, 8, 3284-3293.

6. M. Atlas, S. Datta, “A Statistical Technique for Monoisotopic Peak Detection in a Mass Spectrum”, J. Proteomics Bioinform., 2009, 2, 202-216.

7. Balmain, J. Gray, B. Ponder, "The genetics and genomics of cancer", Nature Genet., 2003, 33, 238-244.

8. S.E. Bell, E.G. Nazarov, Y.F. Wang, G.A. Eiceman, “Classification of ion mobility spectra by functional groups using neural networks”, Anal. Chim. Acta, 1999, 394, 121-133.

9. M.S. Bereman, T. I. Williams, D.C. Muddiman, “Development of a nanoLC LTQ Orbitrap Mass Spectrometric Method for Profiling Glycans Derived from Plasma from Healthy, Benign Tumor Control, and Epithelial Ovarian Cancer Patients”, Anal. Chem., 2008, 81, 1130-1136.

10. M.S. Bereman, D.D. Young, A. Deiters, D.C. Muddiman, “Development of a robust and high throughput method for profiling N-linked glycans derived from plasma glycoproteins by nanoLC-FTICR mass spectrometry”, J. Protoeme Res., 2009, 8, 3764-3770.

11. A. Beveridge, T. Fung, and D. MacDougall, "Use of Infrared Spectroscopy for the Characterization of Paint Fragments," In: Forensic Examination of Glass and Paint Analysis and Interpretation, B. Caddy, ED., Taylor and Francis, NY, 2001, 220-233.

12. J.C. Bezdek, C. Coray, R. Gunderson, and J. Watson, "Detection and characterization of cluster substructure II: fuzzy c-variaties and convex combinations thereof," SIAM J. Appl. MATH, 1981, 40, 358-372.

13. J.C.W. Bink and H.A. Van'T Klooster, "Classification of organic compounds by infrared spectroscopy with pattern recognition and information theory," Anal. Chim. Acta, 1983, 150, 53-59.

14. P.A. Binz, D.F. Hochstrasser, and R.D. Appel, "Mass spectrometry based proteomics: Current status and potential use in clinical chemistry", Clinical Chem. Lab. Med., 41, 1540–1551.

15. G. Blomquist, E. Johansson, B. Soderstrom, and S. Wold, "Data analysis of pyrolysis chromatograms by means of SIMCA pattern recognition," J. Analyt. Appl. Pyrolysis, 1979, 1, 53-65.

16. L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, Wadsworth and Brooks, 1984.

17. R.G. Brereton (Eds.), "Multivariate Pattern Recognition in Chemometrics", Elsevier, Amsterdam, 1992.

18. S.D. Brown, "Chemical systems under indirect observation: latent properties and chemometrics," Appl. Spectrosc., 1995, 49, 14-31.

19. J.L. Buckle, D.A. MacDougal, and R.R. Grant, "PDQ-paint data queries: the history and technology behind the development of the royal Canadian mounted police forensic science laboratory services automotive paint database", Can. Soc. Forens. Sci. J., 1997, 30, 199-212.

20. J. Bunger, G. Westphal, A. Monnich, B. Hinnendahl, E. Hallier, and M. Muller, "Cytotoxicity of occupationally and environmentally relevant mycotoxins, Toxicology", 2004, 202, 199-211.

21. H.A. Burge, "Fungi: toxic killers or unavoidable nuisance", Ann. Allergy Asthma Immunol., 2001, 87, 52-56.

22. P. Carrer, M. Maroni, D. Alcini, and D. Cavallo, "Allergens in indoor air: environmental assessment and health effects," Sci. of Total Environ., 2001, 270, 33-42.

23. N.S. Cartwright and P.G. Rodgers, "A proposed data base for the identification of automotive paint," Can. Soc. Forens. Sci. J., 1976, 9, 145-154.

24. A. Ceroni, K. Maass, H. Geyer, R. Geyer, A. Dell and S.M. Haslam, "GlycoWorkbench: A tool for the computer-assisted annotation of mass spectra of glycans" J. Proteome Res., 2008, 7, 1650-1659.

25. F. Chau, Y. Liang, J. Gao, X. Shao, "Chemometrics – From Basics to Wavelet Transform, John Wiley & Sons, NY, 2004.

26. G. Chen, P.B. Harrington, "Real-time two-dimensional wavelet compression and its application to real-time modeling of ion mobility data", Anal. Chim. Acta, 2003, 490, 59-69.

27. D. Coomans and D. I. Broeckaert, "Potential Pattern Recognition in Chemical and Medical Decision-Making", Research Studies Press LTD., Letchworth, England, 1986.

28. G. Cybenko, "Mathematics of control," Signals and Systems, 1989, 2, 303-310.

29. L. Damokos, I. Frank, G. Matolcsy, and G. Jalsovszky, "Pattern recognition applied to vapor phase infrared spectra", Anal. Chim. Acta, 1983, 154, 181-189.

30. S.R. Delwiche and R.A. Graybosch, "Identification of waxy wheat by near-infrared reflectance spectroscopy", J. Cereal Sci., 2002, 35, 29-38.

31. S.R. Delwiche, R.A. Graybosch, L.E. Hansen, E. Souza, F.E. Dowell, "Single kernel near-infrared analysis of tetraploid (durum) wheat for classification of the waxy Condition", Cereal Chemistry, 2006, 83, 287-292.

32. M.E. de Noo, B.J. Mertens, A. Ozalp, M.R. Bladergroen, M.P. van der Werff, C.J. van de Velde, A.M. Deelder, R.A. Tollenaar, "Detection of colorectal cancer using MALDI-TOF serum protein profiling", Eur. J. Cancer, 2006, 42, 1068-1076

33. W. J. Dunn III, S. Wold, and D. L. Stalling, "Simple modeling by chemical analogy in pattern recognition," in Environmental Applications of Chemometrics, J. J. Breen and P. E. Robinson (Eds.), ACS Symposium Series 292, Washington DC, 1985.

34. G.A. Eiceman, E.V. Krylov, N.S. Krylova, E.G. Nazarov, R.A. Miller, "Separation of ions from explosives in differential mobility spectrometry by vapor-modified drift gas", Anal. Chem., 2004, 76, 4937-4944.

35. G.A. Eiceman, E.V. Krylov, B. Tadjikov, R.G. Ewing, E.G. Nazarov, R.A. Miller, "Differential mobility spectrometry of chlorocarbons with a micro-fabricated drift tube", Analyst, 2004, 129, 297-304.

36. G.A. Eiceman, E.G. Nazarov, J.E. Rodriguez, "Chemical class information in ion mobility spectra at low and elevated temperatures", Anal. Chim. Acta, 2001, 433, 53-70.

37. G.A. Eiceman, B. Tadjikov, E. Krylov, E.G. Nazarov, R.A. Miller, J. Westbrook, P. Funk, "Miniature radio-frequency mobility analyzer as a gas chromatographic detector for oxygen-containing volatile organic compounds, pheromones and other insect attractants", J. Chromatogr., 2001, 917, 205–217.

38. G.A. Eiceman, M. Wang, S. Prasad, H. Schmidt, F.K. Tadjimukhamedov, B. K. Lavine, N. Mirjankar, "Pattern recognition analysis of differential mobility spectra with classification by chemical family", Analytica Chimica Acta, 2006, 579, 1–10.

39. "EGISTM Defender Portable Lightweight Desktop Explosives Trace Detection System" ©2005 Thermo Electron Corporation, Part Number: 41841800-Revision A., available at www.sionex.com, 2006.

40. J. Epstein, C.F. Morris, and K.C. Huber, "Instrumental texture of white salted noodles prepared from recombinant inbred lines of wheat differing in the three granule bound starch synthase (waxy) genes", J. Cereal Sci., 2002, 35, 51-63.

41. B.S. Everitt and G. Dunn, "Applied Multivariate Data Analysis", 2nd Edition, John Wiley & Sons, NY, 2001.

42. O. Faix, J.H. Bottcher, and E. Bertelt, Proc. 8th Int. Conf. On Fourier Transform Spectroscopy (8th ICOFTS), Lubek (Edited by H. M. Heise, E. H. Korte, and H. W. Siesler). Proc. SPIE 1575, 428, 1992.

43. P.L. Ferguson, and R.D. Smith, "Proteome analysis by mass spectrometry". Annual Review of Biophysics and Biomolecular Structure, 2003, 32, 399-424.

44. G. Fettis, (Editor), "Automotive Paints and Coatings", VCH Publications, New York, 1995.

45. L. Feuerstein, M. Rainer, and K. Bernardo, "Derivatized cellulose combined with MALDI-TOF MS: A new tool for serum protein profiling", J. Proteome Res., 2005, 4, 2320–2326.

46. E. Frank, "DASCO – a new classification method," Chem. Intell. Lab. Syst., 1988, 4, 215-22.

47. E. Frank and S. Lanteri, "Classification models: discriminant analysis, SIMCA, CART," Chem. Intell. Lab. Syst., 1989, 5, 247-256.

48. E. Frank and J. H. Workman, "Classification: oldtimers and newcomers," J. Chemom., 1989, 3, 463-476.

49. D.S. Frankel. "Pattern recognition of Fourier transform infrared spectra of organic compounds", Anal. Chem., 1984, 56, 1011-1014.

50. Y. Freund, R. E. Schapire, “Experiments with a new boosting algorithm”, Proceedings of Machine Learning: Thirteenth International Conference, 1996, 148-156.

51. Y. Freund, R. E. Schapire, “A decision theoretic-generalization of online learning and an application to boosting”, Journal of Computer and System Sciences, 1997, 55, 119-139.

52. J. H. Friedman, “Regularized discriminant analysis,” J. Amer. Statist. Assoc. 1989, 84, 165-175.

53. F. Fung, D. Tappen, and G. Wood, “Alternaria-associated asthma,” Appl. Occ. Environ. Hyg., 2000, 15, 924-927.

54. M. Gaye, S.J. Valentine, Y. Hu, Y. Mechref, B.K. Lavine, N. Mirjankar, and D.E. Clemmer, “Characterization of serum N-linked glycans associated with esophageal adenocarcinoma and its progression by IMS-MS”, Anal. Chem., 2012, IN PRESS.

55. U. Gehring, J. Douwes, G. Doekes, A. Koch, W. Bischof, B. Fahlbusch, K. Richter, H. Wichmann, and J. Heinrich, “β(1-3)-Glucan in house dust of German homes: Housing characteristics, occupant behavior, and relations with endotoxins, allergens and molds” Environ. Health Perspect, 2001, 109, 139-144.

56. P. Geladi and H. Grahn, “Multivariate Image Analysis”, John Wiley and Sons, New York, 1996.

57. D.E. Goldberg, “Genetic Algorithms in Search, Optimization, and Machine Learning”, Addison-Wesley Publishing Company, Reading, MA, 1989.

58. R. Goldman, H.W. Ressom, R.S. Varghese, L. Goldman, G. Bascug, C.A. Loffredo, M. Abdel-Hamid, I. Gouda, S. Ezzat, Z. Kyselova, Y. Mechref, and M.V. Novotny, “Detection of hepatocellular carcinoma using glycomic analysis”, Clin. Cancer Res., 2009, 15, 1808–1813.

59. G. Golub, C. Van Loan, “Matrix Computations”, Johns Hopkins University Press, Baltimore, 1971.

60. A. Graps, “An introduction to wavelets”, Comp. Sci. Eng., IEEE, 1995, 2, 50-61.

61. R. Gray, J. D. Thrasher, R. Crago, R. A. Madison, L. Arnold, A. W. Campbell, and A. Vojdani, “Mixed mold mycotoxicosis: immunological changes in humans following exposure in water-damaged buildings, “ Archives of Environ. Health, 2003, 58, 410-420.

62. R.A. Graybosch, “Waxy wheats: origin, properties, and prospects”. Trends in Food Science and Technology, 1998, 9, 135-142.

63. R.A. Graybosch, C.J. Peterson, L.E. Hansen, S. Rahman, A. Hill, and J.H. Skerritt, “Identification and characterization of U.S. wheats carrying null alleles at the wx loci”. Cereal Chem., 1998, 75, 162-165.

64. P.R. Griffiths, “Fourier Transform Infrared Spectrometry”, Wiley Interscience, New York, 2nd ed., 2007.

65. R. Guevremont, “High-field asymmetric waveform ion mobility spectrometry: a new tool for mass spectrometry”, J. Chromatogr. 2004, 1058, 3–19.

66. R.W. Gunderson, “An adaptive FCV clustering algorithm,” Int. J. Machine Studies,” 1983, 19, 97-104.

67. R.W. Gunderson and D. L. Denq, A Manual for Using the Program FCVPC, Research Report, Department of Mathematics, Utah State University, 1988.

68. I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection,” J. Machine Learning, 2003, 3, 1157-1182.

69. S.J. Haswell (Eds.), “Practical Guide to Chemometrics”, Marcel Dekker, NY. 1992.

70. R. Heisterkamp and J. Johnson, “Pinch Ratio Clustering from a Topologically Intrinsic Lexicographic Ordering,” Proceedings of 2012 IEEE International Conference on Data Mining (ICDM 2012), Brussels, Belgium, 2012, submitted.

71. J.D. Hoheisel, "Microarray technology: beyond transcript profiling and genotype analysis", Nat. Rev. Genet., 2006, 7, 200-210.

72. J.H. Holland, “Adaptation in Natural and Artificial Systems”, 6th edition, MIT Press, Cambridge, MA, 2001.

73. J.H. Holland, “An introduction to intrinsic parallelism”, In Proceedings of the Tenth Anniversary Convocation for IMMD. University of Erlangen, FRG, 1976, 47–55.

74. T. Hoshino, R. Yoshikawa, S. Ito, K. Hatta, T. Nakamura, M. Yamamori, K. Hayakawa, K. Tanaka, H. Akashi, S. Endo, S. Tago, and S. Ishigami, “Flour blends for breads, cakes, or noodles, and foods prepared from the flour blends”, United States Patent, 2000, 6, 042, 867.

75. B.B. Hubbard, “The World According to Wavelets”, 2nd Edition, A. K. Peters, Natick, MA, 1998.

76. J. Huberty, “Applied Discriminant Analysis”, John Wiley & Sons, NY, 1994.

77. D. Isailovic, R.T. Kurulugama, M.D. Plasencia, S.T. Stokes, Z. Kyselova, R. Goldman, Y. Mechref, M.V. Novotny, D.E. Clemmer, “Profiling of human serum glycans associated with liver cancer and cirrhosis by IMS-MS”, J. Proteome Res., 2008, 7, 1109-1117.

78. D. Isailovic, M.D. Plasencia, M.M. Gaye, S.T. Stokes, R.T. Kurulugama, V. Pungpapong, M. Zhang, Z. Kyselova, R. Goldman, Y. Mechref, M.V. Novotny, and D.E. Clemmer, "Delineating diseases by IMS-MS profiling of serum N-linked glycans", *J. Proteome Res.,* 2012, 11, 576–585.

79. T. Jacobson, and R. W. Gunderson, "Cluster analysis of beer flavor components. II. A case study of yeast strain and brewery dependency," American Society of Brewery Chemists, 1981, 41, 73-77.

80. M. James, "Classification Algorithms", John Wiley & Sons, NY, 1985.

81. C.A. Johnson, H. Topoff, R.K. Van der Meer, and B. Lavine, "Host queen killing by slave-maker ant queen: When is a host queen worth attacking?", Anim. Behav., 2002, 64, 807-815.

82. C.A. Johnson, H. Topoff, R.K. Van der Meer, and B.K. Lavine, "Do these eggs smell funny to you? Egg discrimination by *Formica* hosts of the slave-making ant, *Polyergus breviceps* (Hymenoptera: Formicidae).", Behav. Ecol. Sociobiol., 2005, 57, 245-255.

83. C.A. Johnson, R.K. Van der Meer and B. Lavine, "Changes in the cuticular hydrocarbon profile of the slave-maker ant queen, *Polyergus breviceps* Emery, after killing a *Formica* host queen (Hymenoptera: Formicidae)", J. Chem. Ecol., 2001, 27, 1787-1804.

84. J. Johnson and D.R. Heisterkamp, "Topological Graph Clustering with Thin Position," Geometriae Dedicata, 2012, submitted.

85. T. Jolliffee, "Principal Component Analysis", Springer-Verlag, New York, 1986.

86. P.C. Jurs, B. R. Kowalski, T. L. Isenhour, and C. N. Reilley, "Computerized learning machines applied to chemical problems: molecular structural parameters from low resolution mass spectrometry," Anal. Chem., 1970, 42, 1387-1394.

87. P.C. Jurs, B.K. Lavine, and T.R. Stouch, "Pattern recognition studies of complex chromatographic data sets", J. Res. Nat Bur. Stand., 1985, 90, 543-549.

88. P. Kang, Y. Mechref, I. Klouckova, M.V. Novotny, "Solid-phase permethylation of glycans for mass spectrometric analysis", Rapid Commun. Mass Spectrom., 2005, 19, 3421-3428.

89. J. Karasinski, S. Andreescu, O.A. Sadik, B. Lavine, M.N. Vora, "Multiarray sensors with pattern recognition for the detection, classification, and differentiation of bacteria at subspecies and strain levels", Anal. Chem. 2005, 77, 7941-7949.

90. J. Karasinski, L. White, Y. Zhang, E. Wang, S. Andreescu, O.A. Sadik, B. Lavine, and M.N. Vora. "Detection and identification of bacteria using antibiotic susceptibility and

a multi-array electrochemical sensor with pattern recognition", Biosensors and Bioelectronics, 2007, 22, 2643-2649.

91. R. Karoui, G. Downey, and C. Blecker. "Mid-Infrared Spectroscopy Coupled with Chemometrics: A Tool for the Analysis of Intact Food Systems and the Exploration of the Their Molecular Structure-Quality Relationships – A Review". Chem. Rev., 2010, 110, 6144-6168.

92. L. Kaufman, and P. J. Rousseeuw, "Finding Groups in Data", Wiley-Interscience, NY 1990.

93. K. Kawagoe, and T. Ueda, "A similarity search method of time series data with combination of Fourier and wavelet transforms", Proceedings, Ninth International Symposium on Temporal Representation and Reasoning, IEEE, 2002, 86-92.

94. J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. Antonescu, C. Peterson, P. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks", Nature Medicine, 2001, 7, 673-679.

95. R. Kohavi and G. John, "Wrappers for Feature Selection," Artificial Intelligence, 1997, 97, 273-324.

96. B.R. Kowalski and S. Wold, "Pattern Recognition in Chemistry," in Classification, Pattern Recognition and Reduction of Dimensionality, P. R. Krishnaiah and L. N. Kanal, Eds., North Holland, Amsterdam, 1982.

97. E. Kreyszig, "Advanced Engineering Mathematics", 4th ed., John Wiley & Sons, NY, 1979.

98. E. Krylov, E.G. Nazarov, R.A. Miller, B. Tadjikov, G.A. Eiceman, "Field dependence of mobilities for gas-phase-protonated monomers and proton-bound dimers of ketones by planar field asymmetric waveform ion mobility spectrometer (PFAIMS)", J. Phys. Chem., 2002, 106, 5437–5444.

99. N. Krylova, E. Krylov, G.A. Eiceman, J.A. Stone, "Effect of moisture on the field dependence of mobility for gas-phase ions of organophosphorus compounds at atmospheric pressure with field asymmetric ion mobility spectrometry", J. Phys. Chem., 2003, 107, 3648–3654.

100. W.J. Krzanowski, "Principles of Multivariate Analysis", Oxford University Press, New York, 2000.

101. Z. Kyselova, Y. Mechref, M.M. Al Bataineh, L.E. Dobrolecki, R.J. Hickey, J. Vinson, C.J. Sweeney, M.V. Novotny, "Alterations in the serum glycome due to metastatic prostate cancer:, J. Proteome Res., 2007, 6, 1822-1832.

102. B.K. Lavine, “Pattern recognition,” Critical Reviews in Analytical Chemistry, 2006, 36, 153-161.

103. B.K. Lavine, D. Brzozowski, A. J. Moores, C. E. Davidson, and H. T. Mayfield, “Genetic algorithm for fuel spill identification”, Anal. Chim. Acta, 2001, 437, 233–246.

104. B.K. Lavine, D. Brzozowski, J. Ritter, A.J. Moores, and H.T. Mayfield, “Fuel spill identification by selective fractionation prior to gas chromatography I. Water soluble components,” J. Chromat. Sci., 2001, 39, 501-506.

105. B.K. Lavine and D. Carlson, “European bee or africanized bee? Species identification through chemical analysis,” Anal. Chem., 1987, 59, 468-470.

106. B.K. Lavine and C.E. Davidson, “Multivariate Approaches to Classification Using Genetic Algorithms,” in: S. Brown R. Tauler R, and R. Walczak R editors. Comprehensive Chemometrics 3, Amsterdam, The Netherlands: Oxford- Elsevier 2009, 619-646.

107. B.K. Lavine, C.E. Davidson, C. Breneman, and W. Katt, “Electronic Van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases,” J. Chem. Inf. Science, 2003, 43, 1890-1905.

108. B.K. Lavine, C.E. Davidson, C. Breneman, and W. Katt, “Genetic algorithms for clustering and classification of olfactory stimulants”, In J. Bajorath (Ed.), Chemoinformatics: Methods and Protocols. Humana Press, Totowa, NJ, 2004, 399-426.

109. B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, M.C. Sundling, “Electronic van der Waals surface property descriptors and genetic algorithms for developing structure-activity correlations in olfactory databases” J. Chem. Inf. Comput. Sci., 2003, 43, 1890-1905.

110. B.K. Lavine, C.E. Davidson, and A.J. Moores, “Innovative Genetic Algorithms for Chemoinformatics”, Chemom. Intell. Lab. Syst., 2002, 60, 161–171.

111. B.K. Lavine, C.E. Davidson, A.J. Moores, “Genetic algorithms for spectral pattern recognition”, Vib. Spec., 2002, 28, 83–95.

112. B.K. Lavine, C.E. Davidson, A.J. Moores, and P. R. Griffiths, “Raman spectroscopy and genetic algorithms for the classification of wood types”, Appl. Spectrosc., 2001, 55, 960–966.

113. B.K. Lavine, C.E. Davidson, and W.T. Rayens, “Machine learning based pattern recognition applied to microarray data,” Combinatorial Chemistry & High Throughput Screening,” 2004, 7, 115-131.

114. B.K. Lavine, C. Davidson, R.K. Van der Meer, S. Lahav, V. Soroker, and A. Hefetz, “Genetic algorithms for deciphering the complex chemosensory code of social insects. Chem. Intell. Lab. Syst.”, J. Chem., 2003, 66, 51-62.

115. B.K. Lavine, C.E. Davidson, and D.J. Westover, “Spectral Pattern Recognition Using Self Organizing Maps,” J. Chem. Inf. Comp. Science, 2004, 44, 1056-1064.

116. B.K. Lavine, A. Faruque, P. Kroman, and H. T. Mayfield, "Source Identification of Fuel Spills by Pattern Recognition Analysis of High Speed Gas Chromatograms," Anal Chem., 1995, 67, 3846-3852.

117. B.K. Lavine and D. R. Henry, "Monte Carlo Studies of Nonparametric Linear Discriminant Functions," J. Chemom. 1988, 2, 85-90.

118. B.K. Lavine, D. R. Henry, and P. C. Jurs, "Chance Classifications by Nonparametric Linear Discriminants," J. Chemom. 1988, 2, 1-10.

119. B.K. Lavine, P.C. Jurs, D.R. Henry, and R.K. Van der Meer, J.A. Pino, and J.E. McMurry, “Pattern recognition studies of complex chromatographic data sets: design and analysis of pattern recognition experiments”, Chem. Intell. Lab. Syst., 1988, 3, 79-89.

120. B.K. Lavine, N. Mirjankar, R. LeBouf, and A. Rossner, “Prediction of mold contamination from microbial volatile organic compound profiles using solid phase microextraction and gas chromatography/mass spectrometry,” Microchem. J., 2012, 103, 37-41.

121. B.K. Lavine, N. Mirjankar, R. LeBouf, A. Rossner, "Prediction of mold contamination from microbial volatile organic compound profiles using head space gas chromatography/mass spectrometry", Microchem. J., 2012, 103, 119–124.

122. B.K. Lavine, N. Mirjankar, S. Ryland, M. Sandercock, “Wavelets and genetic algorithms applied to search prefilters for spectral library matching in forensics”, Talanta, 2011, 87, 46–52.

123. B.K. Lavine, N. Mirjankar, and R. K. Van der Meer, “Analysis of chemical signals in red fire ants by gas chromatography and pattern recognition techniques,” Talanta, 2011, 83, 1308-1316.

124. B.K. Lavine, and A.J. Moores. “Genetic algorithms for pattern recognition analysis and fusion of sensor data,” in K. Siddiqui and D. Eastwood editors., Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, Proceedings of SPIES. Bellingham, WA: SPIE, 1999, 103-112.

125. B.K. Lavine, L. Morel, R. K. Van der Meer, R. W. Gunderson, K. H. Han, A. Bonanno, and A. Stine, “Pattern recognition studies in chemical communication: nestmate recognition in camponotus floridanus,” Chem. Intell. Lab. Syst., 1990, 9, 107-114.

126. B.K. Lavine, K. Nuguru, and N. Mirjankar, “One stop shopping - feature selection, classification, and prediction in a single step”, J. Chemom., 2011, 25, 116-129.

127. B.K. Lavine, K. Nuguru, N. Mirjankar, J. Workman Jr., “Pattern recognition assisted infrared library searching”, Appl. Spectrosc., 2012, 66-68.

128. B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, and H.T. Mayfield. “Source identification of underground fuel spills by solid phase micro-extraction/high-resolution gas chromatography/genetic algorithms”, Anal. Chem., 2000, 72, 423-431.

129. B.K. Lavine, A. Stine, and H. T. Mayfield, “Gas chromatography-pattern recognition techniques in pollution monitoring,” Anal. Chim. Acta, 1993, 227, 357-367.

130. B.K. Lavine, A. B. Stine, H. Mayfield, and R. Gunderson, "Application of High-Resolution Computer Graphics to Pattern Recognition Analysis," J. Chem. Inf. Comp. Sci., 1993, 33, 826-834.

131. B.K. Lavine, R.K. Van der Meer, L. Morel, R. W. Gunderson, J.H. Han, and A. Stine, “False color data imaging: A new pattern recognition technique for analyzing chromatographic profile data”, Microchem. J., 1990, 41, 288-295.

132. B.K. Lavine, A. Vesanen, D.M. Brzozowski, and H.T. Mayfield, “Authentication of fuel standards using gas chromatography/pattern recognition techniques”, Anal Letters, 2001, 34, 281- 294.

133. B.K. Lavine and J. R. Workman, “Chemometrics: Past, Present, and Future,” in B. K. Lavine (Eds.) Chemometrics and Chemoinformatics, ACS Symposium Series 894, Oxford University Press, 2005.

134. F. LeBouf, C. Casteel, and A. Rossner, “Evaluation of air sampling technique for assessing low-level volatile organic compounds in indoor environments,” J. Air & Waste Management Assoc., 2010, 60, 156-162.

135. L. LeBouf, S. A. Schuckers, and A. Rossner, “Preliminary assessment of a model to predict mold contamination based on microbial volatile organic compound profiles,” Sci. of Total Environ., 2010, 408, 3648-3653.

136. C.B. Lebrilla, H.J. An, “The prospects of glycanbiomarkers for the diagnosis of diseases”. Mol. BioSyst., 2009, 5, 17-20.

137. H.J. Lee, E.Y. Lee, and M.S. Kwon, “Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics”, Current Opinion in Chemical Biology, 2006, 10, 42-49.

138. T.G. Lee, “Health symptoms caused by molds in a courthouse,” Archives of Environ. Health, 2003, 58, 442-446.

139. H.J. Lohninger, “Evaluation of neural networks based on radial basis functions and their application to the prediction of boiling points from structural parameters”, J. Chem. Inf. Comp. Science, 1993, 33, 736-744.

140. S.R. Lowry, D.A. Huppler, and C.R.J. Anderson. “Database development and search algorithms for automated infrared spectral identification”. J. Chem. Inf. Comput. Sci., 2010, 25, 235-241.

141. G.L. McLachlan, “Discriminant Analysis and Statistical Pattern Recognition”, John Wiley & Sons, New York, 1992.

142. J. Mandel, “The regression analysis of collinear data,” J. Res. NBS, 1985, 90(6), 465-476.

143. H. Martens and T. Naes, “Multivariate Calibration”, John Wiley & Sons, NY, 1989.

144. L. Massart, and L. Kaufman, “The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis”, John Wiley & Sons, New York, 1983.

145. MATLAB R 2006a Wavelet Toolbox, Natick, MA.

146. Y. Mechref, A. Hussein, S. Bekesova, V. Pungpapong, M. Zhang, L.E. Dobrolecki, R.J. Hickey, Z.T. Hammond, M.V. Novotny, “Quantitative serum glycomics of esophageal adenocarcinoma, and other esophageal disease onsets”, J. Protoeme Res., 2009, 8, 2656-2666.

147. R.A. Miller, G.A. Eiceman, E.G. Nazarov, A.T. King, “A novel micromachined high-field asymmetric waveform-ion mobility spectrometer”, Sens. Actuators B Chem., 2000, 67, 300–306.

148. M. Mitchell, “An Introduction to Genetic Algorithms”, MIT Press, Cambridge, MA, 1998.

149. A.J. Moores, “The learning genetic algorithm: a hybrid learning system”, Master’s thesis, Clarkson University, Potsdam, NY, 2000.

150. L. Morel, R.K. Van der Meer, and B.K. Lavine, “Ontogeny of nestmate recognition cues in the red carpenter ant (*Camponotus floridanus*)--behavioral and chemical evidence for the role of age and social experience”, Behav. Ecol. Sociobiol., 1988, 22, 175-183.

151. S.N. Mukherjee, P. Sykacek, S.J. Roberts, S.J. Gurr, “Gene ranking using bootstrapped P-values”, SIGKKD Explorations, 2003, 5, 16-22.

152. T. Nakamura, P. Vrinten, M. Saito, and M. Konda, “Rapid classification of partial waxy wheats using PCR-based markers”, Genome, 2002, 45, 1150-1156.

153. T. Nakamura, and M. Yamamori, “Production of waxy (amylose-free) wheats”. Mol. Gen. Genet., 1995, 248, 253-259.

154. L. Nørgaard, R. Bro, F. Westad, and S.B. Engelsen. “A modification of canonical variates analysis to handle highly collinear multivariate data”, J. Chemom., 2006, 20, 425–435

155. N.L. Owen, D.W. Thomas, “Infrared studies of hard and soft woods”, Appl. Spectrosc, 1989, 43, 451-455.

156. A.J. Panshin and C. de Zeeuw, “Textbook of Wood Technology: Structure, Identification, Properties, and use of the Commercial Woods of the United States and Canada”, McGraw-Hill, New York, 1980.

157. S. Perkins, K. Lacker, and J. Theiler, “Grafting: Fast Incremental Feature Selection by Gradient Descent in Function Space,” J. Machine Learning, 2003, 3, 1333-1356.

158. E.F. Petricoin, A.M. Adrekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, “use of proteomic patterns in serum to identify ovarian cancer”, Lancet, 2002, 359, 572-577.

159. J.A. Pino, J.E. McMurry, P.C. Jurs, B.K. Lavine, and A.M. Harper, "Applications of pyrolysis/gas chromatography/pattern recognition to the detection of CF heterozygotes," Anal. Chem. 1985, 57, 295-304.

160. W. Pjipers, “Failures and successes with pattern recognition for solving problems in analytical chemistry,” Analyst, 1984, 109, 299-303.

161. O. Preisner, J.A. Lopes, and J. C. Menzes, “Uncertainty Assessment in FT-IR Spectroscopy Based Bacteria Classification Models,” Chem. Intell. Lab. Systems, 2008, 94, 33-42.

162. S. Rahman, B. Kosarhashemi, M.S. Samuel, A. Hill, D.C. Abbott, J.H. Skerritt, J. Preiss, R. Appels, and M.K. Morell, “The major proteins of wheat endosperm starch granules”, Aust. J. Plant Physiol., 1995, 22, 793-803.

163. I. Reddy, and P.A. Seib, “Modified waxy wheat starch compared to modified waxy corn starch”. J. Cereal Sci., 2000, 31, 25-29.

164. H.W. Ressom, R.S. Verghese, L. Goldman, C.A. Loffredo, M. Abdel-Hamid, Z. Kyselova, Y. Mechref, M. Novotny, R. Goldman, “ Analysis of MALDI-TOF mass spectrometry data for detection of glycan biomarkers”, Pac. Symp. Biocomput., 2008, 216-227.

165. C. Robbe-Masselot, A. Herrmann, E. Maes, I. Carlstedt, J.C. Michalski, C. Capon, “Expression of a core 3 Disialyl-Le(x) hexasaccharide in human colorectal cancers: A potential marker of malignant transformation in colon”, J. Proteome Res., 2009, 8, 702-711.

166. P. Robert, D. Bertrand, M.F. Devaux, and A. Sire, "Identification of chemical constituents by multivariate near-infrared spectral imaging," Anal. Chem., 1992, 64, 664-667.

167. C. Rudin, I. Daubechies, R.E. Schapire, "On the dynamics of boosting", Advances in Neural Information Processing Systems 16, 2004.

168. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagations," in Parallel Distributed Processing, Volume 1, MIT Press, Cambridge, MA, 1986.

169. R. Saldova, L. Royle, C.M. Radcliffe, U.M. Abd Hamid, R. Evans, J.N. Arnold, R.E. Banks, R. Hutson, D.J. Harvey, R. Antrobus, S.M. Petrescu, R.A. Dwek, P.M. Rudd, "Ovarian cancer is associated with Changes in glycosylation in both acute phase proteins and IgG", Glycobiology, 2007, 17, 1344–56.

170. R. Savilahti, J. Uitti, P. Laippala, T. Husman, and P. Roto, "Respiratory morbidity among children following renovation of water-damaged school", Arch. Environ. Health, 2000, 55, 405-410.

171. R.E. Schapire, "The boosting approach to machine learning: An overview", in Nonlinear Estimation and Classification, Springer, 2003.

172. R.E. Schapire, M. Rochery, M. Rahim, N. Gupta, "Boosting with prior knowledge for call classification", IEEE Transactions on Speech and Audio Processing, 2005, 13.

173. B. Schoelkopf and A. Smola, "Learning with Kernels", MIT Press, Cambridge, MA, 2002.

174. T.P. Schultz, M.C. Templeton, and G.D. McGinnis, "Rapid determination of lignocelluloses by diffuse reflectance Fourier transform infrared spectrometry", Anal. Chem., 1985, 57, 2867-2869.

175. M.A. Sharaf, D.L. Illman, and B.R. Kowalski, "Chemometrics (Chemical Analysis Series, Vol. 82)", John Wiley & Sons, NY, 1986.

176. J. Shawe-Taylor and N. Cristianini, "An Introduction to Support Vector Machines", Cambridge University Press, Cambridge, UK, 2000.

177. S. Sheng, D. Chen, and J.E. Van Eyk, "Multidimensional liquid chromatography separation of intact proteins by chromatographic focusing and reversed phase of the human serum proteome-Optimization and protein database", Molecular and Cellular Proteomics, 2006, 5, 26–34.

178. M. Sjostrom and B. R. Kowalski, "A Comparison of Five Pattern Recognition Methods based on the Classification Results from Six Real Data Bases," Anal. Chim. Acta, 1979, 112, 11-30.

179. G.W. Small, “Automated spectral interpretation,” Anal. Chem., 1987, 59, 535-546.

180. B. Smith, A.M. Belcher, G. Epple, P.C. Jurs, and B.K. Lavine, “Computerized pattern recognition: a new technique for the analysis of chemical communication,” Science, 1985, 228, 175-177.

181. J.R.M. Smits, P. Schoenmakers, A. Stehmann, F. Sijstermans, and G. Kateman, “Interpretation of infrared spectra with modular neural-network systems”, Chemom. Intell. Lab. Syst., 1993, 18, 27-39.

182. B. Soderstrom, S. Wold, and G. Blomquist, “Pyrolysis gas chromatography combined with SIMCA pattern recognition for classification of fruit-bodies of some ectomycorrhizal suillus species,” J. Gen. Microbiol., 1982, 128, 1783-1794.

183. J.D. Spengler, J.K. Jaakkola, H. Parise, B.A. Katsnelson, L.I. Privalova, and A.A. Kosheleva, “Housing characteristics and children respiratory health in the Russian Federation”, Am. J. Public Health, 2004, 94, 657-662.

184. M.A. Stapanian, F.C. Garner, K.E. Fitzgerald, G.T. Flatman, and J.M. Nocerino, “Finding suspected causes of measurement error in multivariate environmental data”, J. Chem., 1993, 7, 165-176.

185. J. Stuper, W. E. Brugger, and P.C. Jurs, “A Computer System for Structure-Activity Studies Using Chemical Structure Information Handling and Pattern Recognition Techniques,” in Chemometrics: Theory and Application, B. R. Kowalski (Ed.), ACS Symposium Series 52, Washington DC, 1977.

186. Z. Tang, R.S. Varghese, S. Bekesova, C.A. Loffredo, M.A. Hamid, Z. Kyselova, Y. Mechref, M.V. Novotny, R. Goldman, and H.W. Ressom, “Identification of *N*-glycan serum markers associated with hepatocellular carcinoma from mass spectrometry data”, J. Proteome Res., 2010, 9, 104–112.

187. K E. Thrane and R.W. Gunderson, “Source allocation of organic air pollutants by application of fuzzy c-variaties pattern recognition,” Anal. Chim. Acta, 1986, 191, 309-317.

188. J.T. Tou and R.C. Gonzalez, “Pattern Recognition Principles”, Addison-Wesley Publishing, Reading, MA, 1974.

189. W.H.A. van den Broek, D. Wienke, W. J. Melssen, R. Feldhoff, T. Huth-Fehre, T. Kantimm, L.M.C. Buydens, “Application of a spectroscopic infrared focal plane array sensor for on-line identification of plastic waste,” Appl. Spectrosc., 1997, 51, 856-865.

190. R.K. Van der Meer and L. Morel, “Nestmate recognition in ants,” in R.K. Vander Meer, M. Breed, M. Winston, and K.E. Espelie (editors), Pheromone Communication in Social Insects, Westview Press, Boulder, CO, 1998.

191. R.K. Van der Meer, D. Saliwanchik, and B. Lavine, “Temporal changes in colony cuticular hydrocarbon patterns of *Solenopsis invicta*: Implications for nestmate recognition”, J. Chem. Ecol., 1989, 15, 2115-2125.

192. L.J. van′t Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, "Gene-expression profiling predicts clinical outcome of breast cancer", Nature, 2002, 415, 530-536.

193. V. Vapnik, Estimation of Dependencies based on Empirical Data, Springer Series in Statistics, Springer, 1982.

194. V. Vapnik, “Statistical Learning Theory”, John Wiley & Sons: New York, 1998.

195. A. Varki, R. Cummings, J. Esko, H. Freeze, G. Hart, J. Marth, (editors) “Essentials of Glycobiology”, 2nd edition, CSHL Press, 2009.

196. C.A. Veasey, C.L.P. Thomas, “Fast quantitative characterization of differential mobility responses”, Analyst, 2004, 129, 198–204.

197. K. Venne, E. Bonneil, K. Eng, P. Thibault, “Improvement in peptide detection for proteomics analyses using NanoLC-MS and high-field asymmetry waveform ion mobility mass spectrometry”, Anal. Chem., 2005, 77, 2176–2186.

198. J. Villanueva, D.R. Shaffer, J. Phillip, C.A. Chaparro, H. Erdjument-Bromage, A.B. Olshen, M. Fleisher, H. Lilja, E. Brogi, J. Boyd, M. Sanchez-Carbayo, E.C. Holland, C. Cordon-Cardo, H.I. Scher, P. Tempst, “Differential exoprotease activities confer tumor-specific serum peptidome patterns”, J. Clinical Investigation, 2006, 116, 271-284.

199. B. Walczak, D.L. Massart, “Wavelet packet transform applied to a set of signals: A new approach to the best-basis selection”, Chemomet. Intell. Lab. Syst., 1997, 36, 81-94.

200. S. Walker, “A Primer on Wavelets and their Scientific Applications”, Chapman & Hall/CRC, NY, 1999.

201. B.B. Wall, M.T. Kachman, and S.Y. Gong, “Isoelectric focusing nonporous RP HPLC: A two-dimensional liquid-phase separation method for mapping of cellular proteins with identification using MALDI-TOF mass spectrometry”, Anal Chem., 2000, 72, 1099–1111.

202. J.S. Walter. A Primer on Wavelets and their Scientific Applications, New York: Chapman & Hall/CRC, 1999.

203. C.P. Wang, and T.L. Isenhour. “Infrared Library Search on Principal Component-Analyzed Fourier Transform Absorption Spectra”, Appl. Spectrosc., 1987, 41,185-195.

204. M.P. Washburn, D. Wolters, and J.R. Yates, “Large-scale analysis of the yeast proteome by multidimensional protein identification technology”, Nature Biotech., 2001, 19, 242–247.

205. P.D. Wasserman, “Neural Computing”, Van Nostrand Reinhold, New York, 1989.

206. R. Watrous, “Learning Algorithms for Connectionist Networks: Applied Gradient Methods of Nonlinear Optimization (Tech Report MS-CIS-87-51)”, University of Pennsylvania, Philadelphia, PA, 1987.

207. M.D. Wessel, J.M. Sutter, P.C. Jurs, “Prediction of reduced ion mobility constants of organic compounds from molecular structure”, Anal. Chem., 1996, 68, 4237-4243.

208. M.B. West, Z.M. Segu, C.L. Feasley, P. Kang, I. Klouckova, C. Li, M.V. Novotny, C.M. West, Y.Mechref, and M.H. Hanigan, “Analysis of site-specific glycosylation of renal and hepatic γ-glutamyl transpeptidase from normal human Tissue”, J. Biol. Chem., 2010, 285, 29511-29524.

209. S. Wold, “Pattern recognition by means of disjoint principal component models,” Pattern Recognition, 1976, 8, 127-139.

210. S. Wold, “Cross validatory estimation of the number of components in factor and principal components models,” Technometrics, 1978, 20, 397-406.

211. S. Wold, “Pattern Recognition: Finding and Using Regularities in Multivariate Data,” in H. Martens and H. Russwurm, Editors, Food Research and Data Analysis, Applied Science, Essex, England, 1983.

212. S. Wold, C. Albano, W. J. Dunn III, U. Edlund, K. Esbensen, P. Geladi, S. Hellberg, E. Johansson, W. Lindberg, and M. Sjostrom “Multivariate Data Analysis in Chemistry,” in Chemometrics, Mathematics and Statistics in Chemistry, B. R. Kowalski (Ed.), D. Reidel Publishing Company, 1984.

213. S. Wold and M. Sjostrom, “SIMCA, a method for analyzing chemical data in terms of similarity and analogy,” in Chemometrics, Theory and Application (B.R. Kowalski, Ed.), American Chemical Society, Symposium Series 52, Washington DC, 1977.

214. H.B. Woodruff, “Novel Advances in Pattern Recognition and Knowledge-Based Methods in Infrared Spectroscopy,” in L.C. Meuzelaar, and T.L. Isenhour editors, Computer-Enhanced Analytical Spectroscopy I, Plenum Press, New York, 1987, 201-219.

215. H.B. Woodruff, S. R. Lowry, and T. L. Isenhour. “Comparison of two Discriminant Functions for Classifying Binary Infrared Data”, Appl. Spectrosc., 1975, 29, 226-230.

216. C. Yang and P.A. Heinsohn, “Sampling and Analysis of Indoor Microorganisms”, John Wiley & Sons, New York, 2007.

217. X.C. Zhao and P.J. Sharp, “An improved 1-D SDS-PAGE method for the identification of three bread wheat ‘waxy’ proteins”, J. Cereal Sci., 1996, 23, 191-193.

218. J. Zupan, and J. Gasteiger, “Neural Networks for Chemists”, VCH Publishers, New York, 1993.

APPENDIX I

Thesis Publications

1. Barry K. Lavine, David J. Westover, Leah Oxenford, Nikhil Mirjankar, and Necati Kaval, "Construction of an Inexpensive Surface Plasmon Resonance Instrument for Use in Teaching and Research," **Microchemical Journal**, 2007, 86, 147-155.
   The construction of an inexpensive SPR instrument that can be used for both teaching and research is described. Using a 2'x 2' optical table to construct this instrument allows both scientists and students full access to the operation of the spectrometer. Furthermore, the use of open platform instrumentation has the advantage of maintaining the focus on the relationship between emerging technology and analytical chemistry as well as allowing the user to modify the instrument to enhance the measurement process for a particular application. This is a change from the learning paradigm used in most research and teaching laboratories where commercial instrumentation is treated as a black box due to its complexity. Three studies, which were performed using this instrument, are presented to demonstrate the suitability of this instrument for both teaching and research. These studies include measuring the refractive index of alcohols, investigating the partitioning of ruthenium (II) trisbipyridine chloride into Nafion, and understanding the mechanism controlling metal ion adsorption by polyacrylamide hydrogels.

2. B. K. Lavine, D. J. Westover, N. Kaval, N. Mirjankar, L. Oxenford, and G. Mwangi, "Swellable Molecularly Imprinted Poly N-(N-propyl)acrylamide Particles for Detection of Emerging Organic Contaminants Using Surface Plasmon Resonance Spectroscopy," **Talanta**, 2007, 72, 1042-1048.
   Lightly crosslinked theophylline imprinted polyN-(N-propyl)acrylamide particles (ca. 300nm in diameter) that are designed to swell and shrink as a function of analyte concentration in aqueous media were spin coated onto a gold surface. The nanospheres responded selectively to the targeted analyte due to molecular imprinting. Chemical sensing was based on changes in the refractive index of the imprinted particles that accompanied swelling due to binding of the targeted analyte, which was detected using surface plasmon resonance (SPR) spectroscopy. Because swelling leads to an increase in the percentage of water in the polymer, the refractive index of the polymer nanospheres decreased as the particles swelled. In the presence of aqueous theophylline at concentrations as low as 10-6 M, particle swelling is both pronounced and readily detectable. The full scale response of the imprinted particles to template occurs in less than ten minutes. Swelling is also reversible and independent of the ionic strength of the solution in contact with the polymer. Replicate precision is less than 10-4 RI units.

By comparison, there is no response to caffeine which is similar in structure to theophylline at concentrations as high as 1x10-2 M. Changes in the refractive index of the imprinted polymer particles, as low as 10-4 RI units could be readily detected. A unique aspect of the prepared particles is the use of light crosslinking rather than heavy crosslinking. This is a significant development as it indicates that heavy crosslinking is not entirely necessary for selectivity in molecular imprinting with polyacrylamides.

3. Barry K. Lavine, Nikhil Mirjankar, and Robert K. Vander Meer, "Analysis of Chemical Signals In Red Fire Ants by Gas Chromatography and Pattern Recognition Techniques," **Talanta**, 2011, 83, 1308-1316.
   Gas chromatographic (GC) profiles of cuticular hydrocarbon extracts obtained from individual and pooled ant samples were analyzed using pattern recognition techniques. Clustering according to the biological variables of social caste and colony were observed. Pooling individual extracts enhanced the recognition of patterns in the GC profile data characteristic of colony. Evidently, the contribution of the ant's individual pattern to the overall hydrocarbon profile pattern can obscure information about colony in the GC traces of cuticular hydrocarbon extracts obtained from red fire ants. Re-analysis of temporal caste and time period data on the cuticular hydrocarbon patterns demonstrates that sampling time and social caste must be taken into account to avoid unnecessary variability and possible confounding. This and the fact that foragers could not be separated from reserves and brood-tenders in all 5 laboratory colonies studied suggests that cuticular hydrocarbons as a class of sociochemicals cannot model every facet of nestmate recognition in S. invicta which in turn suggests a potential role for other compounds in the discrimination of alien conspecifics from nestmates.

4. Barry K. Lavine, Kadambari Nuguru, and Nikhil Mirjankar, "One Stop Shopping - Feature Selection, Classification, and Prediction in a Single Step," **Journal of Chemometrics**, 2011, 25, 116-129
   We report on the application of a genetic algorithm (GA) for pattern recognition that uses both supervised and transverse learning to mine spectroscopic and proteomic data. The pattern recognition GA selects features that optimize the separation of the classes in a plot of the two or three largest principal components of the data. For training sets with small amounts of labeled data (i.e., data points tagged with a class label) and large amounts of unlabeled data (i.e., data points that are not tagged with a class label), this approach is preferred, as our results show since information in the unlabeled data is used by the fitness function to guide feature selection. The advantages of incorporating transverse learning into the fitness function of the pattern recognition GA have been evaluated in two recently published studies by our group. In one study, Raman spectroscopy and the pattern recognition GA were used to develop a potential method to discriminate hardwoods, softwoods and tropical woods. In a second study, biopsy material of small round blue cell tumors analyzed by cDNA microarrays was identified as to type (Ewings sarcoma, Burkitt's lymphoma, neuroblastoma, and rhabdomyo sarcoma) through supervised learning implemented by the pattern recognition GA.

5. J. Bowen, J. Meecham, M. Hamlin, B. Henderson, M. Kim, N. Mirjankar, B. K. Lavine, "Development of Field-Deployable Instrumentation Based on "Antigen-Antibody" Reactions for Detection of Hemorrhagic Disease in Ruminants," **Microchemical Journal**, 2011, 99(2), 415-420.
   Development of field-deployable methodology utilizing antigen-antibody reactions and the surface plasmon resonance (SPR) effect to provide a rapid diagnostic test for recognition of the blue tongue virus (BTV) and epizootic hemorrhage disease virus (EHDV) in wild and domestic ruminants is reported. A Spreeta® chip, which utilizes microelectronic technology to implement the SPR effect, is shown to possess sufficient sensitivity and operating speed to detect either BTV and EHVD antigens or antibodies in real time. The biosensor consists of an outer active surface layer comprised of either an antibody or antigen immobilized by covalent bonds through several other organic layers including a self assembled monolayer to a gold surface. Parallel experiments were run on the biosensor surface using either a home-built high resolution SPR instrument or a low resolution solid state Spreeta® SPR chip. Both instruments were capable of monitoring the antigen-antibody reaction used to selectively detect the presence of BTV and EHDV viral pathogens. Results for the antibody and antigen reactive layers with antigen or antibody solutions as well as the modeling of these layers are discussed. The characteristics of these biosensors – specificity and time of reaction – were assessed. The antibody surface biosensors exhibited a high degree of specificity, even when using low resolution instrumentation. The time of analysis was under 20 minutes, which was the arbitrary exposure time. Results indicate the potential of even shorter times of analysis.

6. Barry K. Lavine, Nikhil Mirjankar, Scott Ryland, and Mark Sandercock, "Wavelets and Genetic Algorithms Applied to Search Prefilters for Spectral Library Matching in Forensics," **Talanta**, 2011, 87, 46-52.
   Currently, the identification of the make, model and year of a motor vehicle involved in a hit and run collision from only a clear coat paint smear left at the crime scene is not possible. Search prefilters for searching infrared (IR) spectral libraries of the Paint Data Query (PDQ) automotive database to differentiate between similar but nonidentical Fourier transform infrared (FTIR) paint spectra are proposed. Applying wavelets, FTIR spectra of clear coat paint smears can be denoised and deconvolved by decomposing each spectrum into wavelet coefficients which represent the sample's constituent frequencies. A genetic algorithm for pattern recognition analysis is used to identify wavelet coefficients characteristic of the model and manufacturer of the automobile from which the spectra of the clear coats were obtained. Even in challenging trials where the samples evaluated were all the same manufacturer (Chrysler) with a limited production year range, the respective models and manufacturing plants were correctly identified. Search prefilters for spectral library matching are necessary to extract investigative lead information from a clear coat paint smear; which, like the undercoat and color coat paint layers, exhibits chemical features in its FTIR spectrum unique to the automobile manufacturing plant at which it was applied. Information obtained from these searches can also serve to quantify the general discrimination power of original automotive paint comparisons

encountered in casework, and to succinctly communicate the significance of the evidence to the courts.

7. Barry K. Lavine, Kadambari Nuguru, Nikhil Mirjankar, and Jerome Workman, "Development of Carboxylic Acid Search Prefilters for Spectral Library Matching," **Microchemical Journal**, 2012, 103, 21-36
   435 infrared (IR) absorbance spectra of 140 carboxylic acids and 295 noncarboxylic acids which included aldehydes, ketones, esters, amides as well as compounds containing both carbonyls and alcohols were preprocessed using the wavelet packet tree to enhance subtle but important features in the data. Wavelet coefficients that optimized the separation of the spectra by functional group in a plot of the two largest principal components of the data were identified using a genetic algorithm (GA) for pattern recognition analysis. Because principal components maximize variance, the bulk of the information encoded by the wavelet coefficients selected by the pattern recognition GA is characteristic of the carboxylic acid functional group. The carboxylic acid search prefilter developed as part of this study was successfully validated using two external validation sets. The first validation set consisted of 24 carboxylic acids and 61 noncarboxylic acids and the second validation set consisted of 264 carboxylic acids and 72 noncarboxylic acids.

8. Barry K. Lavine, George Mwangi, Nikhil Mirjankar, and Mariya Kim, "Characterization of Swellable Molecularly Imprinted Polymer Particles by Surface Plasmon Resonance Spectroscopy," **Applied Spectroscopy**, 2012, 66(4), 440-446.
   Surface plasmon resonance (SPR) has been used to investigate template binding at sites in theophylline imprinted poly N-(N-propyl) acrylamide particles. At concentrations as low as 10-6M theophylline, particle swelling is detected as a shift in the angle of minimum reflectance. The binding constant of theophylline estimated from the inflection point of the theophylline calibration curve is approximately 10,000. The imprinted polymer particles do not respond to caffeine or theobromine (which differs from theophylline by a single methyl group) at concentrations as high as 10-2M. Full scale response of the imprinted polymer particles to theophylline (template) occurs in less than 15 minutes, and swelling is reversible. The immobilized imprinted polymer particles can undergo approximately 20-25 swelling and shrinking cycles before there is significant loss in functionality. A unique aspect of these imprinted polymer particles is that template binding causes the angle of minimum reflectance to decrease, not to increase in magnitude. Adsorption, which causes an increase in the angle of minimum reflectance, can be readily discriminated from template binding.

9. Barry K. Lavine, Nikhil Mirjankar, Ryan LeBouf, and Alan Rossner, "Prediction of Mold Contamination from Microbial Volatile Organic Compound Profiles Using Solid Phase Microextraction and Gas Chromatography/Mass Spectrometry," **Microchemical Journal**, 2012, 103, 37-41
   An integrated chemical and microbiological approach was used to develop a new sampling and analytical methodology to characterize the fungal load of a contaminated area or building. A set of microbial volatile organic compound

(MVOC) profiles were developed with corresponding bioaerosol measurements as input-output pairs for a discriminant to predict the presence or absence of mold contamination in indoor environments. Spore collection to characterize the indoor air quality of the residences and buildings was performed using an Anderson N6 impactor. Simultaneously, solid phase microextraction was used as a passive sampling device to collect VOCs from the air for GC/MS analysis. The volatile organic signatures that molds emit as reflected by the gas chromatographic profiles were compared to the impactor data collected from each sampling site. By comparing the bioaerosol data to the volatile organic profiles, a discriminant could be trained to classify a residence with potential mold growth based on its MVOCs.

10. B. K. Lavine, N. Mirjankar, R. LeBouf, and A. Rossner, "Prediction of Mold Contamination from Microbial Volatile Organic Compound Profiles Using Head Space Gas Chromatography/Mass Spectrometry," **Microchemical Journal**, 2012, 103, 119-124.
    An integrated chemical and microbiological approach was used to develop a new sampling and analytical methodology to characterize the fungal load of a contaminated area or building. A set of microbial volatile organic compound (MVOC) profiles were developed with corresponding bioaerosol measurements as input-output pairs for a discriminant to predict the presence or absence of mold contamination in indoor environments. Spore collection to characterize the indoor air quality of the residences and buildings was performed using an Anderson N6 impactor. Simultaneously, solid phase microextraction was used as a passive sampling device to collect VOCs from the air for GC/MS analysis. The volatile organic signatures that molds emit as reflected by the gas chromatographic profiles were compared to the impactor data collected from each sampling site. By comparing the bioaerosol data to the volatile organic profiles, a discriminant could be trained to classify a residence with potential mold growth based on its MVOCs.

11. Barry K. Lavine, Kadambari Nuguru, Nikhil Mirjankar, and Jerome Workman, "Pattern Recognition Assisted Infrared Library Searching," **Applied Spectroscopy**, 2012, IN PRESS.
    Pattern recognition methods have been used to develop search prefilters for infrared (IR) library searching. A two step procedure has been employed. First, the wavelet packet tree is used to decompose each spectrum into wavelet coefficients that represent both the high and low frequency components of the signal. Second, a genetic algorithm for pattern recognition analysis is used to identify wavelet coefficients characteristic of functional group. Even in challenging trials involving carboxylic acids, compounds that possess both carbonyl and hydroxyl functionalities can be readily differentiated from carboxylic acids. The proposed search prefilters allow for the use of more sophisticated and correspondingly more time-consuming algorithms in IR spectral library matching because the size of the library can be culled down for a specific match using information from the search prefilter about the presence or absence of specific functional groups in the unknown.

APPENDIX II

Compounds Used in the Study for Developing Search Pre-filters for Infrared Library Searching of Carboxylic Acids (Section 4.4)

Training Set Compounds

(E)-2-Butene	1,5-Hexadiene, 2-methyl-
(E)-2-Pentene, 4,4-dimethyl-	1,9-Decadiene
(E)-3-Heptene, 2,2-dimethyl-	10-Undecenoic acid
(E)-3-Hexene	1-Butene
(E)-3-Hexene, 2,5-dimethyl-	1-Butene, 2-ethyl-
(E)-3-Octene	1-Butene, 2-methyl-
(E)-4-Octene	1-Butene, 3,3-dimethyl-
(E)-5-Decene	1-Cyclohexene, 4-vinyl-
(Z)-3-Hexene, 2,5-dimethyl-	1-Cyclohexene-1-acetic acid
1,2-Cyclobutanedicarboxylic anhydride, cis-	1-Decene
1,2-Cyclopropanedicarboxylic acid, trans-	1-Dodecene
1,3,4-Thiadiazole, 2,5-dimethyl-	1-Heptene
1,3-Butadiene	1-Hexadecene
1,3-Cyclooctadiene	1-Hexanol, 2-ethyl-, phosphite
1,3-Dithiane	1-Hexene (liquid)
1,4-Benzoquinone, 2,6-dimethyl-	1-Hexene, 2,3-dimethyl-
1,4-Cyclohexadiene, 1-methyl-	1-Hexene, 2-ethyl-
1,4-Dithiane	1-Naphthoic acid
1,4-Piperazinedicarboxylic acid, diethyl ester	1-Octadecene
1,5-Heptadiene, 3-methyl-	1-Octene
1,5-Hexadiene, 2,5-dimethyl-	1-Oxa-6-thiacycloheptadecan-17-one

1-Pentene, 2,4,4-trimethyl-	2H-1,2-Benzothiazin-3(4H)-one, 2-ethyl-, 1,1-dioxide
1-Pentene, 2-methyl-	2-Heptenoic acid
1-Pentene, 4-methyl-	2-Hexanone, 5-methyl-
1-Piperazinecarboxylic acid, methyl ester	2-Hexanone, 5-methyl-, oxime
1-Pyrrolidinepropionitrile, b-oxo-	2-Hexene
1-Tetradecene	2-Hexene, 2,5-dimethyl-
1-Tridecene	2-Hexene, 2,5-dimethyl-
1-Undecene, 2-methyl-	2H-Pyran-2-carboxaldehyde, 3,4-dihydro-2,5-dimethyl-
2,4,6-Cycloheptatrien-1-one, 2-hydroxy-4-isopropyl-	2-Imidazolidinone, 1-allyl-
2,4-Hexadiene	2-Isoindolineacetic acid, 1,3-dioxo-a-isopropyl-
2,4-Hexadiene, 2,5-dimethyl-	2-Isoindolineacetic acid, 1,3-dioxo-a-methyl-
2,5-Norbornadiene	2-Nonadecanone
2,6-Octadien-1-ol, 3,7-dimethyl-, formate	2-Norbornaneacetic acid
2,6-Pyridinedicarboxylic acid, 4-methoxy-,	2-Norbornanecarboxylic acid
2-Azetidinone, 1,4-diphenyl-3-ethyl-	2-Norbornene
2-Benzofurancarboxaldehyde, 3-methyl-	2-Octanone oxime
2-Benzofurancarboxaldehyde, 3-methyl-	2-Octene
2-Benzofurancarboxylic acid	2-Pentanone oxime
2-Butanone, 3,3-dimethyl-, oxime	2-Pentanone, 4-methyl- oxime
2-Butanone, 3-hydroxy-	2-Pentene
2-Butene, 2,3-dimethyl-	2-Pentene, 2,4,4-trimethyl-
2-Butene, 2-methyl-	2-Pentene, 2-methyl-
2-Decenoic acid	2-Pentene, 4-methyl-
2-Dodecenoic acid	2-Pyrrolidinone, 1-(3-aminopropyl)-
2-Ethylhexylphosphonic acid, bis(2-ethylhexyl) ester	2-Thiazoline, 2-methyl-
2-Furaldehyde	2-Thiophenebutyric acid
2-Furaldehyde, 5-(acetoxymethyl)-	

2-Thiophenecarboxaldehyde, 5-chloro-	Acetic acid, (2-ethoxyphenyl)-
2-Tridecenoic acid	Acetic acid, (4-chloro-o-tolyloxy)-
3,5-Heptanedione, 2,2,6,6-tetramethyl-	Acetic acid, 2-(2-methoxyphenyl)-
3,8-Diazabicyclo[3.2.1]octan-2-one, 8-methyl-3-(3-methyl-2-butenyl)-	Acetic acid, 2-methoxy-2-phenyl-
3-Cyclohexene-1-carboxylic acid	Acetic acid, 4-bromophenyl-
3-Decyne	Acetic acid, bromo-
3-Decyne	Acetic acid, bromo-, pentachlorophenyl ester
3-Furoic acid, 5-formyl-2-(trifluoromethyl)-, ethyl ester	Acetic acid, dichloro-
3-Heptene, 2,2,4,6,6-pentamethyl-	Acetic acid, ethoxy-
3-Heptene, 2,6-dimethyl-	Acetic acid, mercapto-
3-Quinolinecarboxaldehyde	Acetic acid, mercapto-, 2-methoxyethyl ester
4,4'-Stilbenediol, a,a'-diethyl-, diacetate	Acetic acid, methoxy-
4-Octanone, 5-hydroxy-	Acetic acid, phenoxy-
4-Pentenoic acid	Acetic acid, trifluoro-
5-Norbornene-2-carboxaldehyde	Acetic anhydride
5-Pyrimidinecarboxylic acid, 2,4-bis(methylthio)-, ethyl ester	Acetic anhydride
8-Hexadecanone, 9-hydroxy-	Acetophenone, 2,2-dichloro-
Acetaldehyde, tribromo-	Acetophenone, 2'-amino-
Acetamide, 2-hydroxy-N-phenyl-	Acetophenone, 4'-hydroxy-
Acetamide, N,N-diisopropyl-	Acetyl chloride (liquid)
Acetamide, N-butyl-N-(p-tolylsulfonyl)-	Acrylamide, N-(1,1-dimethyl-3-oxobutyl)-
Acetanilide, 2,2',4',5'-tetrachloro-	Acrylic acid
Acetanilide, 2,2',4',5'-tetrachloro-	Acrylic acid, 2-ethyl-3-propyl-
Acetanilide, 2,4'-dichloro-	Acrylic acid, 2-methyl-
Acetic acid	Adipic acid, monomethyl ester
Acetic acid, (2,4-dichlorophenoxy)-	Allyl sulfide
	Ammonium thiocyanate, tetrapentyl-
	Ammonium thiocyanate, tetrapentyl-

Amyl disulfide	Benzoic acid, 4-chloro-3-nitro-
Anthranilic acid, 5-chloro-	Benzoic acid, 4-fluoro-
Arachidic acid	Benzoic acid, 4-hydroxy-
Azelaic acid	Benzoic acid, 4-nitro-
Azelaic acid, monomethyl ester	Benzoic acid, 4-t-butyl-
Barbituric acid, 5-ethyl-1-methyl-5-phenyl-	Benzoic acid, cmpd with dibutylamine
Behenic acid	Benzophenone, 2-amino-2',5-dichloro-
Benzaldehyde, 2,3,4-trimethoxy-	Benzophenone, decafluoro-
Benzaldehyde, 2,4-dimethyl-	Bicyclo[3.1.1]hept-2-ene, 2,6,6-trimethyl-
Benzaldehyde, 2-ethoxy-	Bicyclo[3.1.1]hept-2-ene, 2,6,6-trimethyl-
Benzaldehyde, 4-hydroxy-3,5-dimethoxy-	Bicyclo[4.4.0]decane
Benzene, tetrahydro-	Bicyclohexyl
Benzo[b]thiophene-2-carboxaldehyde, 7-	Butane
Benzo[b]thiophene-2-carboxaldehyde, 7-methyl-	Butane, 1,4-diiodo-
Benzoic acid	Butane, 1-iodo-
Benzoic acid, 2,4-dichloro-	Butane, 2,3-dimethyl-
Benzoic acid, 2,5-dichloro-	Butane, 2-methyl-
Benzoic acid, 2-benzyl-	Butyl phosphate
Benzoic acid, 2-bromo-	Butyl phosphite
Benzoic acid, 2-chloro-	Butyric acid
Benzoic acid, 2-ethoxy-	Butyric acid, 2-bromo-
Benzoic acid, 2-nitro-	Butyric acid, 2-bromo-3-methyl-
Benzoic acid, 3-chloro-	Butyric acid, 2-chloro-
Benzoic acid, 3-chloro-4-hydroxy-	Butyric acid, 2-hydroxy-2-methyl-
Benzoic acid, 3-hydroxy-	Butyric acid, 2-methyl-
Benzoic acid, 3-methyl-4-nitro-	Butyric acid, 3,3-dimethyl-
Benzoic acid, 3-nitro-	Butyric acid, 3-bromo-
Benzoic acid, 4-acetoxy-3-methoxy-	Butyric acid, 3-chloro-

Butyric acid, 3-methyl-	Cyclohexane, butyl-
Butyric acid, 3-methyl-2-phthalimido-	Cyclohexane, ethyl-
Butyric acid, 4-(2,4-dichlorophenoxy)-	Cyclohexane, iodo-
Butyric acid, 4-(4-methoxyphenyl)-	Cyclohexane, isobutyl-
Butyric acid, 4-(ethylthio)-	Cyclohexane, pentyl-
Butyric acid, 4-acetyl-	Cyclohexane, propyl-
Butyrophenone, 4'-hydroxy-	Cyclohexane, t-butyl-
Capric acid	Cyclohexane, vinyl-
Caprylic acid	Cyclohexaneacetic acid
Carbamic acid, allyl-, ethyl ester	Cyclohexanecarboxylic acid
Carbamic acid, diallyl-, ethyl ester	Cyclohexanecarboxylic acid, 1-methyl-
Carbamic acid, diallyl-, ethyl ester	Cyclohexanepropionic acid
Carbamic acid, dibutyl-, ethyl ester	Cyclohexanol, 2,5-dimethyl-
Carbamic acid, dimethyl-, 3-nitrophenyl ester	Cyclohexanone, 4-ethyl-
Carvomenthene	Cyclohexene, 1-methyl-
Chloral	Cyclohexene, 4-isopropenyl-1-methyl
Chloroformic acid, hexyl ester	Cyclohexene, 4-isopropenyl-1-methyl-
Cinnamic acid	Cyclooctane
Cinnamic acid, 2-bromo-a-cyano-, ethyl ester	Cyclooctene
Crotonic acid, 4-phosphono-, triethyl ester	Cyclopentane
Cyclobutane, octafluoro-	Cyclopentane, butyl-
Cyclobutane, perfluoro-1,2-dimethyl-	Cyclopentane, methyl-
Cyclobutanecarboxylic acid	Cyclopentaneacetic acid, a-phenyl-
Cyclododecene	Cyclopentanecarboxylic acid
Cyclohexane	Cyclopropanecarboxylic acid, trans-2-nitro-,
Cyclohexane, 1,1-dimethyl-	Decane
Cyclohexane, 1,4-dimethyl-	Decane, 1,10-diiodo-
Cyclohexane, 1-hexyl-4-tetradecyl-, trans-	Decane, 1-fluoro-

Decane, 1-iodo-	Formamide, N-(a-methylbenzyl)-
Decanedioic acid	Formamide, N-(a-methylbenzyl)-
Decyl disulfide	Formamide, N-ethyl-
Decyl disulfide	Formic acid
Dibutyltin diacetate	Formic acid, isopentyl ester
dimethyl ester	Formic acid, propyl ester
Dimethyl hydrogen phosphite	Fumaric acid, methyl-
Disiloxane, hexamethyl-	Furan, 2-acetyl-
Docosane	Glutaric acid, 3-oxo-, diethyl ester
Dodecane	Glutaric acid, methyl ester
Dodecanoic acid	Glycolic acid, ethyl ester
Eicosane	Hendecane
Enanthic acid	Hendecanoic acid
Ethane	Heptadecane
Ethane, 1,1-difluoro-	Heptadecane, 6,9,12-tripropyl-
Ethane, 1,2-bis(ethylthio)-	Heptadecanoic acid
Ethane, hexafluoro-	Heptane
Ethane, iodo-	Heptane, 1-iodopentadecafluoro-
Ethyl disulfide	Heptane, 1-iodopentadecafluoro-
Ethyl phosphite	Heptane, 2,2-dimethyl-
Ethyl phosphonate	Heptane, 3-(iodomethyl)-
Ethyl phosphorothioate	Heptane, 3,3-dimethyl-
Ethyl sulfite	Heptanoic acid, 2-bromo-
Ethylene, 1,1-difluoro-	Heptanoic acid, 3-ethyl-3-methyl-
Ethylene, fluoro-	Hexacosane
Ethylene, tetrafluoro-	Hexadecane
Ethylphosphonic acid, diethyl ester	Hexadecane, 6,11-dipentyl-
Flavanone	Hexadecanoic acid

Hexane	Isonicotinaldehyde, O-propyloxime
Hexane, 1,6-diiodo-	Isonicotinaldehyde, O-propyloxime
Hexane, 2,2,4-trimethyl-	Isooctane (so-called)
Hexane, 2,2,5-trimethyl-	Isopentyl disulfide
Hexane, 2,4-dimethyl-	Isopropyl disulfide
Hexane, 2,4-dimethyl-	Itaconic acid, monomethyl ester
Hexane, 2,5-dimethyl-	Lactamide
Hexanoic acid, 2-bromo-	Lactic acid
Hexanoic acid, 3,5,5-trimethyl-	Levulinic acid
Hexanoic acid, 6-phenyl-	Maleic acid, (2-acetyl-1,2-dimethylhydrazino)-, dimethyl ester
Hexyl phosphite	Malonic acid, piperonyl-, diethyl ester
Hippuric acid, methyl ester	m-Anisaldehyde, 2-hydroxy-
Hydantoin, 1-acetyl-3,5-dimethyl-2-thio-	m-Anisic acid
Hydantoin, 3-benzyl-5,5-dimethyl-	Mercury, diethyl-
Hydantoin, 5,5-dimethyl-2-thio-	Mercury, diethyl-
Hydratropic acid	Methane, fluoro-
Hydrocinnamic acid	Methane, iodo-
Hydrocinnamic acid, 4-hydroxy-	Methane, trifluoro-
Hydrocinnamic acid, a-isopropyl-2-methyl-	Methyl disulfide
Hydrocinnamic acid, b-methyl-	Methyl phosphite
Imidazole-2-carboxaldehyde, 1-benzyl-	Methyl phosphorothioate
Indene, 3a,4,7,7a-tetrahydro-	Methyl sulfate-d6
Iodoform	Methyl sulfide (liquid)
Isobutane	Morpholine, 4-acetyl-
Isobutyl disulfide	m-Toluic acid
Isobutyl sulfide	Myristic acid
Isobutylene	Nicotinaldehyde
Isocaproic acid	Nicotinic acid, hydrazide

Nonadecane	Pentyl sulfite
Nonane	Phosphine oxide, dimethylhexadecyl-
Nonane, 1,9-diiodo-	Phosphine oxide, dimethylhexadecyl-
Nonane, 2,2,4,4,6,8,8-heptamethyl-	Phosphine oxide, dimethyltetradecyl-
Nonanoic acid	Phosphine oxide, dimethyltetradecyl-
Nonanoic acid, 2,2,3,3,4,4,5,5,6,6,7,7,8,8,9,9-hexadecafluoro-	Phosphonic acid, butyl-, dibutyl ester
Octadecane	Phosphonic acid, vinyl-, diethyl ester
Octadecanoic acid	Phosphoramidic acid, cyclohexyl-, diethyl ester
Octane	Phosphoric acid, diethyl ester
Octanoic acid, pentadecafluoro-	Phthalimide, N-(1-formylethyl)-
Octanoyl chloride	Pilocarpine, mononitrate
Oleic acid	Pivalic acid
o-Veratric acid	p-Mentha-1,4(8)-diene
Paraldehyde	Propane
Pentacosane	Propane, 1,2-dibromo-
Pentadecane	Propane, 1,3-diiodo-
Pentadecane, 2,6,10,14-tetramethyl-	Propane, 1-iodo-
Pentadecanoic acid	Propane, 2-iodo-
Pentane	Propane, 3-iodo-1,1,1,2,2-pentafluoro-
Pentane, 1-iodo-	Propanoic acid
Pentane, 2,2,3-trimethyl-	Propene
Pentane, 2,3,3-trimethyl-	Propionic acid, 2-(2,4,6-trichlorophenoxy)-
Pentane, 2,3,4-trimethyl-	Propionic acid, 2-(2,5-dichlorophenoxy)-
Pentane, 2,4-dimethyl-	Propionic acid, 2-bromo-
Pentane, 3,3-dimethyl-	Propionic acid, 2-chloro-
Pentane, 3-ethyl-	Propionic acid, 2-methyl-2-(m-tolylthio)-
Pentane, 3-methyl-	Propionic acid, 3-(p-tolylthio)-
Pentanoic acid	Propionic acid, 3,3-diphenyl-

Propionic acid, 3-mercapto-	Sulfide, ethyl isopropyl
Propionic acid, a-(2,4-dichlorophenoxy)-	Sulfoxide, dimethyl-d6
Propionic acid, a-(4-chloro-o-tolyloxy)-	Tartaric acid, diethyl ester
Propionic anhydride	Tetracosane
Propionic anhydride	Tetradecane
Propyl disulfide	Thiophene, 2-acetyl
p-Toluic acid	Thiophene, 2-iodo-
Pyridine, 2-acetyl-	Thiopyran, tetrahydro-
Pyridine, 3-acetyl	Tin chloride, trimethyl-
Pyrrole, 1-methyl-	Tin dichloride, diisobutyl-
Pyrrole-2-carboxaldehyde	Tin dichloride, diisobutyl-
Rhodanine, 3-methyl-	Tin trichloride, butyl-
Salicylaldehyde, 5-methoxy-	Tin, tetraethyl-
Salicylic acid	Tin, tetramethyl-
Salicylic acid, 5-bromo-	Tridecane
Salicylic acid, 5-chloro-	Tridecanoic acid
Salicylic acid, 5-fluoro-	Tropolone
Salicylic acid, 5-t-butyl-	Urea, tetramethyl-
Silane, tetrafluoro-	Urea, tetramethyl-
Silane, tetramethyl-	Valeraldehyde
Silane, triethoxyethyl-	Valeric acid, 2,2-dimethyl-
Spiro[5.5]undecane-3-carboxylic acid	Valeric acid, 4-hydroxy-3-mercapto-, g-lactone
Succinic acid, methyl-	Valeric acid, 5-chloro-
Sulfide, allyl sec-butyl	Vanillic acid
Sulfide, allyl sec-butyl	Vanillin, 6-bromo-
Sulfide, butyl ethyl	Vanillin, acetate
Sulfide, butyl ethyl	Veratric acid
Sulfide, ethyl isopropyl	

Validation Set Compounds

1(6H)-Pyridazineacetic acid, 3-chloro-6-oxo-, ethyl ester	Anthranilic acid, methyl ester
1,2,4-Triazine-3,5(2H,4H)-dione, 2-methyl	Benzoic acid, 2-iodo-
1,4-Benzoquinone, 2,5-dihydroxy-	Benzoic acid, 2-trifluoromethyl-
1,4-Naphthoquinone, 3-methyl-2,5,8-trihydroxy-	Benzoic acid, 3,4-dihydroxy-, ethyl ester
1,5-Cyclooctadiene	Benzoic acid, 3,4-dimethyl-
10-Eicosanone, 11-hydroxy-	Benzophenone, 2-hydroxy-5-methyl-
1-Butene, 3-methyl-	Boric acid, tris(2-ethylhexyl) ester
2(10)-Pinene	Butyl sulfite
2,4-Pentadienoic acid, 5-phenyl-	Butyric acid, 2-ethyl-
2-Azepinone, hexahydro-	Butyric acid, 4-(2,5-xylyl)-
2-Cyclohexen-1-one, 3-amino-5,5-dimethyl	Butyric acid, ethyl ester
2-Naphthoic acid	Butyric acid, hydrazide
2-Octenoic acid	Caproic acid
2-Penten-4-one, 2-hydroxy-5-methoxy-	Carbamic acid, diphenyl-, ethyl ester
2-Propen-1-one, 1,3-di-2-thienyl-	Carbamic acid, isopropyl ester
2-Thiophenecarboxaldehyde	Carbonic acid, diethyl ester
3-Decenoic acid	Chloroformic acid, phenyl ester
5-Undecyne	Cinnamic acid, 4-methyl-
Acetamide, N-ethyl-N-(p-tolylsulfonyl)-	Cyclododecane
Acetanilide, 2-chloro-4'-nitro-	Cycloheptanecarboxylic acid
Acetic acid, carvacryl-	Cyclohexane, 1-dodecyl-4-octyl-, trans-
Acetic acid, m-tolyl-	Cyclohexane, cis-1,3-dimethyl-
Acetophenone, 3'-fluoro-4'-methoxy	Cyclohexane, methyl-
Adamantane, 1,3-dimethyl-	Cyclohexane, trans-1,2-dimethyl-
Adipic acid	Cyclopentaneacetic acid
Anthranilic acid, cyclohexyl ester	Cyclopropane, benzoyl-
	Cyclopropanecarboxamide, N-ethyl-2-phenyl-, cis-

Cyclotetrasiloxane, octamethyl-	Piperidine, 1-(trichloroacetyl)-
Dicyclopentadiene	Propionic acid, 3-(6-hydroxy-m-anisoyl)-, methyl ester
Ethyl phosphate	Pyridine, 2-acetyl-6-methyl
Fluorene, dodecahydro-	Salicylic acid, benzyl ester
Formic acid, methyl ester	Salicylic acid, isopentyl ester
Furan, 3,4-bis(acetoxymethyl)-	Suberic acid
Furoin	Succinimide, N-(4-chloroanilinomethyl)-
Glutaric acid, 2-methyl-	Succinyl chloride
Heptadecane, 6,12-diethyl-9-pentyl-	Tricosane
Heptane, 2,2,4,6,6-pentamethyl-	Valeric acid, 2-bromo-
Hexanoic acid, 2-butyl-	Valeric acid, 2-bromo-4-methyl-, dl-
Hexanoic acid, 2-ethyl-	
Hydantoin, 1-ethyl-3-methyl-2-thio-	
Isobutyric acid	
Ketone, 4-methyl-2-pyridyl 2-thienyl	
Ketone, di-2-pyridyl,	
Mercury, chloroethyl-	
Methacrylic acid, 2-dimethylamino-, ethyl ester	
Methacrylic acid, 2-hydroxyethyl ester	
Naphthalene, decahydro-, cis-,	
Nicotinamide, N,N-dipropyl-	
Octadecane, 1-iodo-	
Octyl disulfide	
o-Toluic acid	
Phenoxyacetic acid, isobutyl ester	
Phenylacetic acid	
Phosphine oxide, diethyltetradecyl-	
Phthalimide, N-(2-hydroxyethyl)-	

VITA

Nikhil Suresh Mirjankar

Candidate for the Degree of

Doctor of Philosophy

Thesis: GENETIC ALGORITHMS FOR FEATURE SELECTION AND CLASSIFICATION OF COMPLEX CHROMATOGRAPHIC AND SPECTROSCOPIC DATA

Major Field: Chemistry

Biographical:

Education:

Completed the requirements for the Doctor of Philosophy in Chemistry at Oklahoma State University, Stillwater, Oklahoma in December, 2012.

Completed the requirements for the Bachelor of Technology in Fibers and Textile Chemistry at University of Mumbai, Institute of Chemical Technology, Mumbai, India in 2002.

Experience:

Worked as a Marketing and Technical Services Executive at Biocon, India's premier biopharmaceutical company from August 2002 to June 2004.