Journal of Research of the National Bureau of Standards
Volume 90, Number 6, November-December 1985

Pattern Recognition Studies of Complex Chromatographic Data Sets

P. C. Jurs, B. K. Lavine, and T. R. Stouch

The Pennsylvania State University, University Park, PA 16802

Accepted: June 24, 1985

Chromatographic fingerprinting of complex biological samples is an active research area with a large and growing literature. Multivariate statistical and pattern recognition techniques can be effective methods for the analyisis of such complex data. However, the classification of complex samples on the basis of their chromatographic profiles is complicated by two factors: 1) confounding of the desired group information by experimental variables or other systematic variations, and 2) random or chance classification effects with linear discriminants. We will treat several current projects involving these effects and methods for dealing with the effects.

Complex chromatographic data sets often contain information dependent on experimental variables as well as information which differentiates between classes. The existence of these types of complicating relationships is an innate part of fingerprint-type data. ADAPT, an interactive computer software system, has the clustering, mapping, and statistical tools necessary to identify and study these effects in realistically large data sets.

In one study, pattern recognition analysis of 144 pyrochromatograms (PyGCs) from cultured skin fibroblasts was used to differentiate cystic fibrosis carriers from presumed normal donors. Several experimental variables (donor gender, chromatographic column number, etc.) were involved in relationships that had to be separated from the sought relationships. Notwithstanding these effects, discriminants were developed from the chromatographic peaks that assigned a given PyGC to its respective class (CF carrier vs normal) largely on the basis of the desired pathological difference. In another study, gas chromatographic profiles of cuticular hydrocarbon extracts obtained from 179 fire ants were analyzed using pattern recognition methods to seek relations with social caste and colony. Confounding relationships were studied by logistic regression. The data analysis techniques used in these two example studies will be presented.

Previously, Monte Carlo simulation studies were carried out to assess the probability of chance classification for nonparametric and parametric linear discriminants. The level of expected chance classification as a function of the number of observations, the dimensionality, and the class membership distributions were examined. These simulation studies established limits on the approaches that can be taken with real data sets so that chance classifications are improbable.

Key words: classification effects; multicomponent spectra; pattern recognition.

Profiling of complex biological materials with high performance chromatographic methods is an active research area with a large and growing literature, e.g., [1–10][1]. Such chromatographic experiments often yield chemical profiles containing hundreds of constituents. These chromatograms can be viewed as chemical fingerprints of the complex samples. Objective analysis of the profiles depends upon the use of multivariate statistical methods. In this regard pattern recognition techniques have been found to be of utility.

Pattern recognition methods have been used to distinguish between individuals in a particular diseased state and normal individuals [7–10]. These methods attempt to classify a sample according to a specific property (e.g., diabetic vs normal) by using measurements that are indirectly related to that property. Mea-

About the Authors, Paper: P. C. Jurs, B. K. Lavine, and T. R. Stouch are with the Department of Chemistry at The Pennsylvania State University. The work they describe received support from the National Science Foundation and the Environmental Protection Agency.

[1] Bracketed numbers indicate literature references.

surements related to the property in question are made. An empirical relationship is then derived from a set of data for which the property of interest and the measurements are known (a training set). Such a relationship or classification rule may be used to infer the presence or absence of this property in objects that are not part of the original training set.

For pattern recognition analysis, each chromatogram is represented as a point, $X=(x_1, x_2, x_3, ..., x_d)$ where component x_j is the area of the jth peak. A set of chromatograms is represented by a set of points in a d-dimensional Euclidean space. The expectation is that the points representing chromatograms from one class will cluster in one limited region of the space separate from the points corresponding to the other class. Pattern recognition is a set of methods for investigating data represented in this manner in order to assess the degree of clustering and general structure of the data space. The four main subdivisions of pattern recognition methodology are mapping and display, discriminant development, clustering, and modelling [11-14]. The ADAPT computer software system [15] has routines in all these areas, and many were used in the two example studies below.

An assumption in pattern recognition is that the ability to categorize the data into the proper classes is meaningful. Successful classification is thought to imply that a relationship between the measurements or features and the property of interest exists. However, classification based on random or chance separation can be a serious problem. For example, the probability of fortuitously obtaining 100% correct classification for a two class problem using a nonparametric linear discriminant can be calculated from the following equation

$$P=2\sum_{i=0}^{d} C_i^{n-1}/2^n \qquad (1)$$

where $C_i^{n-1}=(n-1)!/[(n-1-i)!i!]$, n is the number of objects in the data set, and d is the dimensionality or number of descriptors per object [16,17]. Figure 1 shows a plot of P versus the ratio of the number of objects to the number of descriptors per object (n/d) for $n=50$. The only assumption made concerning the data is that it be in general position, that is, none of the $d+1$ data points should be contained in a $(d-1)$-dimensional hyperplane. When n/d is large, the probability of achieving complete separation due to chance is small. As the number of descriptors approaches the number of objects used in the study, the probability of such an occurence increases. When $n/d=2$, the probability of complete separation is one-half. Such classifications arise due to chance and are not due to any relationship between the objects in the data set. A linear discriminant function developed with an inappropriately small n/d will probably have no predictive ability beyond random guessing.

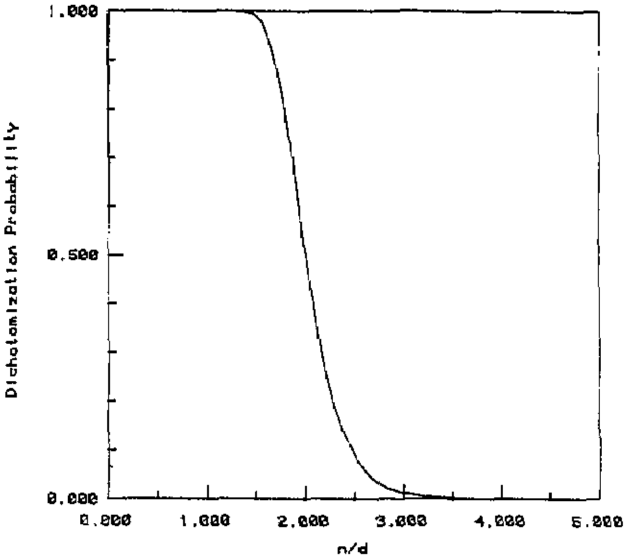

Figure 1–The probability of complete separation into classes by a nonparametric linear discriminant function versus the ratio of the number of objects to the number of descriptors per object.

If $n/d>3$, the probability of complete separation due to chance is small [18, 19]. However, classification rules using linear discriminants are often developed using training sets that are not completely linearly separable. Recently, Stouch and Jurs have reported Monte Carlo simulation studies [20] assessing the degree of fortuitous classification for such situations. Figure 2 is a plot of results obtained in hundreds of Monte Carlo experi-

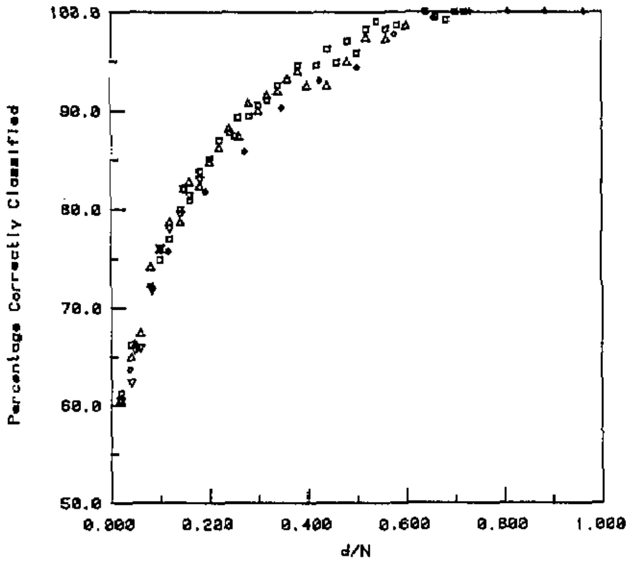

Figure 2–Plot of the percentage of correctly classified patterns versus the ratio of the number of descriptors per pattern to the number of patterns. Each plotting character represents the mean of a number of Monte Carlo experiments.

ments. It shows the percentage of objects correctly classified versus the d/n ratio. The patterns used to develop this curve were random, and equal class sizes were used. The percentage correctly classified for a given d/n value can only be due to chance. Although the probability of obtaining 100% correct classification for $n/d > 3$ is small, chance classification success rates range between 85% and 95%. The influence of the class membership distribution upon chance classification was also investigated, and unequal class sizes lead to even higher success rates due to chance. Figure 3 shows the cumulative probability of achieving any degree of separation due to chance for evenly-divided classes for three values of n/d. At $n/d = 5$, the probability is 50% that 77% of the objects will be correctly classified due to chance. Chance classifications can be a serious problem in linear discriminant analysis of chromatographic fingerprint data. Hence, the results obtained with real data sets must be compared to the results achievable by chance in order to assure that meaningful relations have been discovered.

A second complicating aspect of the classification of complex samples on the basis of their chromatographic profiles is the confounding of the desired group information by experimental variables or other systematic variations. If the basis of classification for patterns in the training set is other than the desired group difference, unfavorable classification results for the prediction set will be obtained despite a linearly separable training set. The existence of these types of complicating relationships is an inherent part of fingerprint-type data. We will discuss several current projects involving these effects and methods for dealing with them.

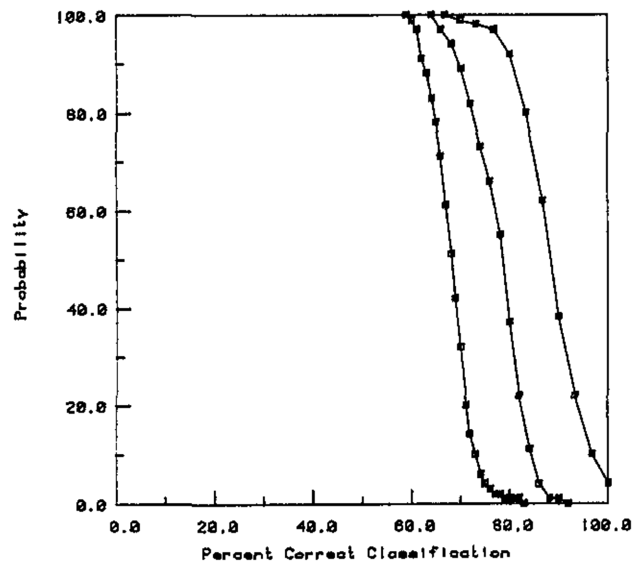

Figure 3–Plot of the cumulative probability of achieving any degree of separation due to chance versus that degree of separation. Three values of n/d are shown. Evenly divided training sets were used.

Cystic Fibrosis Heterozygotes vs Normal Subjects

The first study involves the application of pyrolysis gas chromatography (PyGC) and pattern recognition methods to the problem of identifying carriers of the cystic fibrosis (CF) defect [21]. The biological samples used in this experiment were cultured skin fibroblasts grown from 24 samples obtained from parents of children with CF and from 24 presumed normal donors. A typical CF heterozygote pyrochromatogram is shown in figure 4. The pyrolysed fibroblasts were analyzed on fused silica capillary columns with temperature programming. For each subject, triplicate pyrochromatograms were taken.

The 144 pyrochromatograms were standardized using an interactive computer program [22]. Each pyrochromatogram was divided into 12 intervals defined by 13 peaks that were always present. The retention times of the peaks within the intervals were scaled linearly for best fit with respect to a reference pyrochromatogram. This peak matching procedure yielded 214 standardized retention time windows. Each pyrochromatogram was also normalized using the total area of the 214 peaks. This set of chromatographic data—144 PyGCs of 214 peaks each—was autoscaled so that each PyGC peak had a mean of zero and a standard deviation of one within the entire set of pyrochromatograms.

To apply pattern recognition methods to this overdetermined data set, the necessary first step was feature selection. The number of peaks per chromatogram must be reduced to at least one-third the number of independent PyGCs in the data set, so at most 16 peaks could be analyzed at one time. For the final results of the analysis to be meaningful, this feature selection must be done objectively, that is, without using any class membership information.

For experiments of the type that we are considering here it is inevitable that there will be relationships between sets of conditions used in generating the data and patterns that result. One must realize this in advance when approaching the task of analyzing such data. One must isolate the information pertinent to the pathological alteration characteristic of CF heterozygotes from the large amount of qualitative and quantitative data due to experimental conditions that is also contained in the complex capillary pyrochromatograms.

We have observed that experimental variables (cell culture, batch number, passage number, donor gender, and column identity) can contribute to the overall classification process. For example, a decision function or classification rule was developed from the 12 peaks

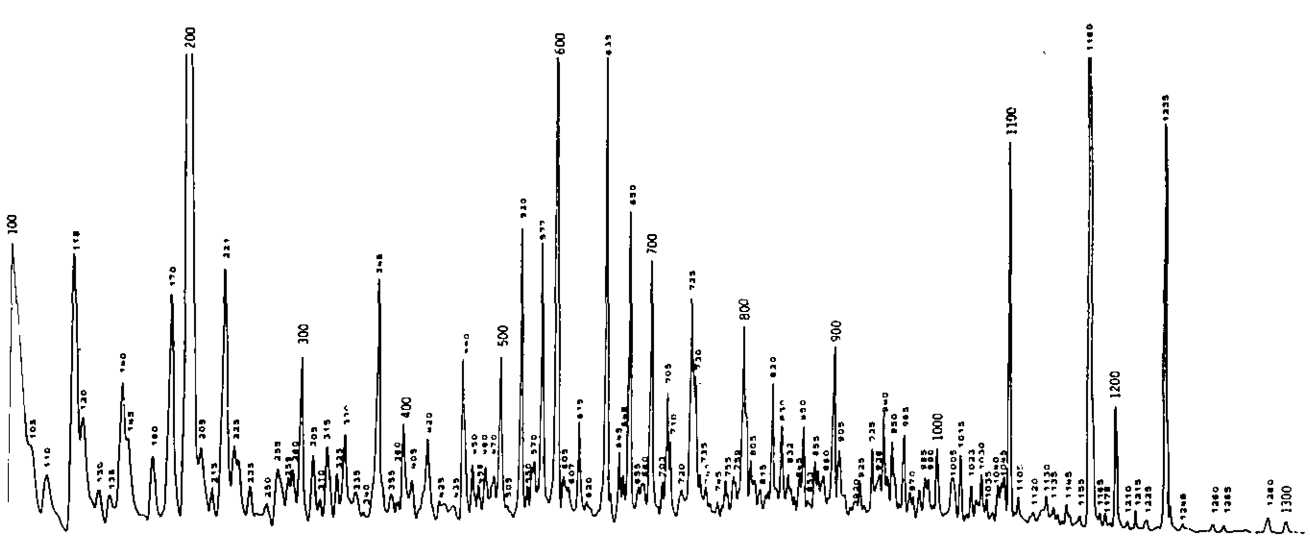

Figure 4–A representative pyrochromatogram from the CF study. The peak identities are those assigned using the peak-matching software. The major peaks are those with assignments that are multiples of 100.

comprising interval three. The CF PyGCs were linearly separable from the PyGCs of the presumed normal donors. However, when the points from this 12-dimensional space were mapped onto a plane that best represents the pattern space (the plane defined by the two largest principal components), groupings related to column identity were observed. Furthermore, classifiers could be developed from these 12 peaks that yielded favorable classification results for many of the experimental variables.

Notwithstanding the effects of the experimental variables described above, a discriminant or decision function has been developed from the PyGC peaks that separates the pyrochromatograms of CF heterozygotes from those of presumed normal subjects, by and large, on the basis of valid chemical differences. The development of such discriminant is described in detail below.

The 65 peaks that were present in at least 90% of the PyGCs were used as a starting point for the analysis. We assessed the ability of each of these 65 peaks alone to discriminate between PyGCs with respect to gender, passage number, and column identity. Twelve peaks that had larger classification success rates for the CF vs normal than for any other dichotomy were selected for further analysis. This procedure identifies those peaks that contain the most information about CF vs normal as opposed to the experimental variables. We were attempting to simultaneously minimize both the probability of chance separation and that of confounding with unwanted experimental details. A classification rule developed from these 12 peaks using the *k*-nearest neighbor procedure correctly classified 90% of the PyGCs in the data set. Variance feature selection [23], combined with the linear learning machine and the adaptive least-squares methods [24], was used to remove 6 of the 12 peaks found to be least relevant to the classification problem. A discriminant that misclassified only eight of the pyrochromatograms (136 correct of 144, 94%) was developed using the final set of only six peaks.

The contribution of the experimental parameters to the overall dichotomization power of the decision function based on the six peaks was assessed by reordering experiments. The set of PyGCs was first reordered in terms of donor gender, and classification results indistinguishable from random were obtained. Similar studies were done for passage number and column identity, and comparable results were obtained. The results of the reordering tests suggest that the decision function based on the six PyGC peaks incorporates mainly chemical information to separate the pyrochromatograms of the CF heterozygotes from those of the normals.

The ability of the decision function to classify a simulated unknown sample was tested using a procedure known as internal validation. Twelve sets of pyrochromatograms were developed by random selection where the training set contained 44 triplicates and the validation set contained the remaining 4 triplicates. Any particular triplicate was only present in one validation set of the 12 generated. Discriminants developed for the training sets were tested on the PyGCs that were held out. The average correct classification for the held-out pyrochromatograms was 87%. This same internal validation test was repeated except that members of the held-out sets included triplicate samples analyzed on the same column or grown in the same batch of growth medium. The average correct classification for the held-out pyrochromatograms in this set of runs was 82%. Although the classification success rate of the decision function was diminished when we took into account these confounding effects, favorable results were still obtained.

Recognition of Ants by Caste and Colony

Chemical communication among social insects can be studied with chromatographic methods. For example, evidence regarding the role of cuticular hydrocarbons in nestmate recognition came from a study of the Myrmecophilous beetle [25]. The data generated in such studies can be complex and may require multivariate statistical or pattern recognition methods for interpretation. Presently, we are analyzing gas chromatographic profiles of high molecular weight hydrocarbon extracts obtained from the cuticles of 179 red fire ant (*Solenopsin invicta*) samples. We are using pattern recognition methods to seek relations with social caste and colony. Each sample contains the hydrocarbons extracted with hexane from the cuticles of 100 individual ants. The hydrocarbon fraction analyzed by gas chromatography was isolated from the concentrated hexane washings by means of a silicic acid column. Evidence regarding the role of cuticular hydrocarbons in nestmate recognition came from a study of the Myrmecophilous beetle [25]. A gas chromatographic trace of the cuticular hydrocarbons from a *S. invicta* sample is shown in figure 5. The hydrocarbon extract was analyzed on a glass column packed with 3% OV-17 using temperature programming.

Five major hydrocarbon compounds were identified and quantified by GC/MS analysis: heptacosane ($n-C_{27}H_{56}$), 13-methylheptacosane, 13,15-dimethylheptacosane, 3-methylheptacosane, and 3,9-dimethylheptacosane in the order of elution from the OV-17 column used. An internal standard was used for quantification. Each chromatogram was normalized using the weight of the collected ants.

Several questions have been addressed in this study: 1) Are the hydrocarbon patterns characteristic of individual colonies? 2) Does the overall colony hydrocarbon pattern change with time? 3) Are the hydrocarbon patterns significantly different for the social castes? In this study, ant samples were obtained from five different colonies (E, J, P, Q, R), three different castes (foragers, broods, and reserves), and for four different time periods (the first three in spring and summer and time period four in the winter).

The first step was to use mapping and display methods [12,17] to examine the structure of the data set. Methods used included principal components mapping and nonlinear mapping [14]. In figures 6 and 7 the results of principal component mapping experiments for colonies J and Q are shown. Colony J includes samples from time periods one through three, whereas colony Q is represented by ants from all four time periods. Colony J has 9 and colony Q has 12 members from each social caste. Pattern groupings according to time period and caste can be seen in figures 6 and 7. The first two principal components account for 96.2% and 97% respectively of the total cumulative variance in the two plots shown. Mapping experiments of this nature were also carried out for samples from a particular caste or time period, and pattern groupings with respect to colony identity, social caste, and temporal period were observed.

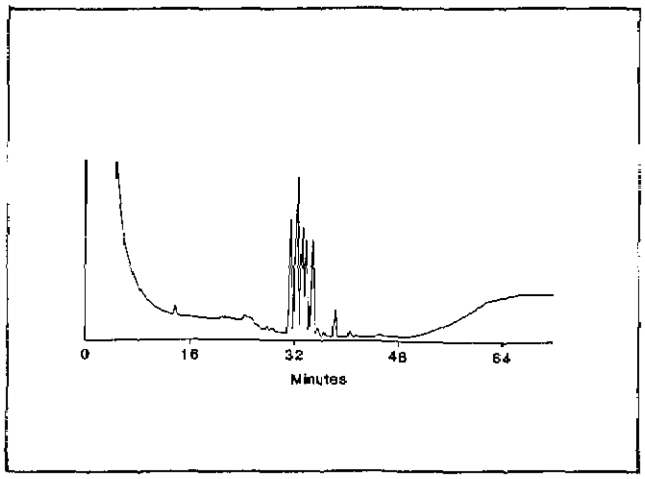

Figure 5–Gas chromatographic trace of cuticular hydrocarbons from *S. invicta* (Reprinted with permission from ref. [25]).

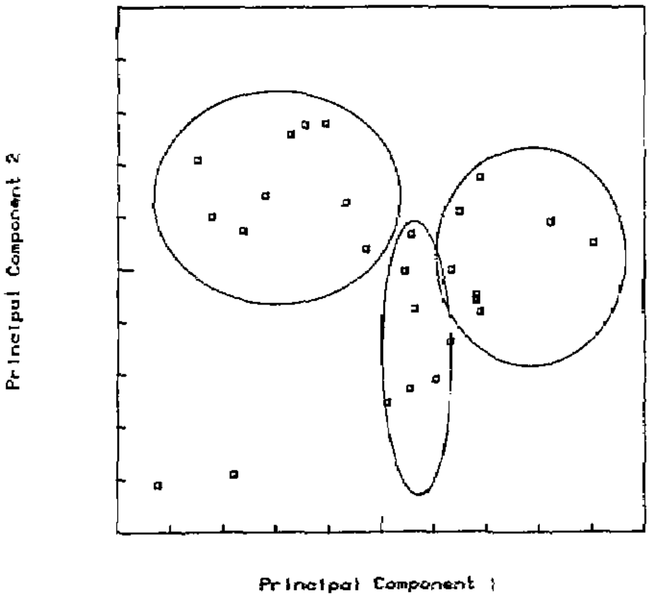

Figure 6–Plot of the two principal components of the five GC peaks for colony J. The elipses show groupings of samples by time period.

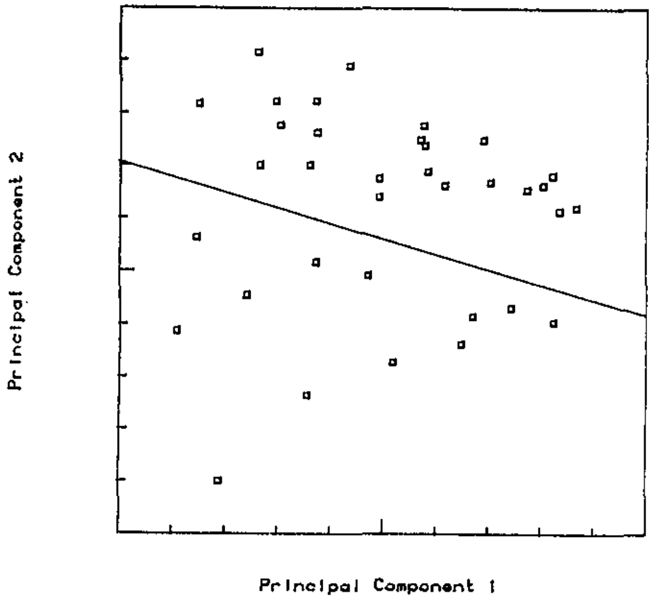

Figure 7–Plot of the two principal components of the five GC peaks for colony Q. The foragers are separated from the reserves and broods by the linear discriminant.

Discriminant analysis studies were also performed. In one study the data set was divided into three categories according to the social caste of the pooled ant sample. Linear discriminants were developed using the areas of the five GC peaks. The hydrocarbon patterns of the foragers were found to be very different from the broods and reserves. In fact, information necessary to discriminate foragers from broods and reserves was primarily encoded in the concentration pattern of the first GC peak. A similar study was undertaken for time period, and the fourth time period was found to be very different from time periods one, two, and three. During time period four the ants are in a state of hibernation, whereas time periods one, two, and three correspond to the spring and summer months.

The hydrocarbon profiles were also found to be characteristic of the individual colonies. Linear decision surfaces were developed from the five GC peaks, using an iterative least-squares method. The purpose was to separate one colony from another or one colony from all other colonies. The results of these discriminant analysis experiments are summarized in table 1. The first row of the table shows that colony E could be separated from colony J by a discriminant that achieved 98% correct classifications (63 correct out of 64 samples) and that colony E could be separated from all the remaining colonies by a discriminant that achieved 95% correct classifications (162 correct out of 170). Colonies Q and R could not be separated well by this method. In addition, multivariate statistical methods such as multivariate analysis of variance and stepwise logistic regression have been employed in this study. The results obtained using these techniques support the conclusions drawn from the pattern recognition experiments. In summary, the GC traces representing ant cuticle extracts could be related to colony identity, social caste, and time period using pattern recognition methods.

Table 1. Percentage of chromatograms correctly classified by colony for several two-way classifications.

Colony	No. in Colony	Colony in Second Group					
		E	J	P	Q	R	All
E	36	—	98	100	100	100	95
J	27		—	100	100	100	98
P	36			—	100	100	99
Q	35				—	73	85
R	36					—	82

The authors thank J. A. Pino, J. E. McMurry, and R. K. Vander Meer for their helpful contributions.

References

[1] Zlatkis A.; R. S. Brazell and C. F. Poole, The role of organic volatile profiles in clinical diagnosis, Clin. Chem. **27**, 789–797 (1981).

[2] Jellum, E. J., Profiling of human body fluids in healthy and diseased states using gas chromatography and mass spectrometry, with special reference to organic acids, Jour. Chromatog. **143**, 427–462 (1977).

[3] Reiner, E., and F. L. Bayer, Botulism: a pyrolysis-gas-liquid chromatographic study, Jour. Chrom. Sci. **16**, 623–629 (1978).

[4] Reiner, E., and J. J. Hicks, Differentiation of normal and pathological cells by pyrolysis-GLC, Chromatographia **5**, 525–528 (1972).

[5] Jellum, E.; I. Bjoernson, R. Nesbakken, E. Johansson, and S. Wold, Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis, Jour. Chromatog. **217**, 231–237 (1981).

[6] Soderstrom, B.; W. Wold and G. Blomquist, Pyrolysis-gas chromatography combined with SIMCA pattern recognition for classification of fruit-bodies of some ectomycorrhizal *suillus* species, Jour. Gen. Micro. **128**, 1773–1784 (1982).

[7] McConnell, M. L.; G. Rhodes, U. Watson, and M. Novotny, Application of pattern recognition and feature extraction techniques to volatile constitutent metabolic profiles obtained by capillary gas chromatography, Jour. Chromatog. **162**, 495–506 (1979).

[8] Wold, S.; E. Johansson, E. Jellum, I. Bjoernson, and R. Nesbakken, Application of SIMCA multivariate data analysis to the classification of gas chromatographic profiles of human brain tissues, Anal. Chim. Acta **133**, 251–259 (1981).

[9] Scoble, H. A.; J. L. Fashing and P. R. Brown, Chemometrics and liquid chromatography in the study of acute lymphocytic leukemia, Anal. Chim. Acta **150**, 171–181 (1983).

[10] Rhodes, G.; M. Miller, M. L. McConnell, and M. Novotny, Metabolic abnormalities associated with diabetes mellitus, as investigated by gas chromatography and pattern recognition analysis of profiles of volatile metabolites, Clin. Chem. **27**, 580–585 (1981).

[11] Jurs, P. C., and T. L. Isenhour, Chemical applications of pattern recognition, Wiley-Interscience: New York (1975).

[12] Varmuza, K., Pattern recognition in chemistry, Springer-Verlag: Berlin (1980).

[13] Kryger, L., Interpretation of analytical chemical information by pattern recognition methods—a survey, Talanta **28**, 871–887 (1981).

[14] Fukunaga, K., Introduction to statistical pattern recognition, Academic Press: New York (1972).

[15] Stuper, A. J.; W. E. Brugger and P. C. Jurs, Computer assisted studies of chemical structure and biological function, Wiley-Interscience: New York (1979).

[16] Nilsson, N. J., Learning machines, McGraw-Hill: New York (1965).

[17] Tou, J. T., and R. C. Gonzalez, Pattern recognition principles, Addison-Wesley Pub. Co.: Reading, MA (1974).

[18] Stuper, A. J., and P. C. Jurs, Reliability of nonparametric linear classifiers, Jour. Chem. Inf. Comp. Sci. **16**, 238–241 (1976).

[19] Whalen-Pedersen, E. K., and P. C. Jurs, The probability of dichotomization by a binary linear classifier as a function of training set population distribution, Jour. Chem. Inf. Comp. Sci. **19**, 264–266 (1979).

[20] Stouch, T. R., and P. C. Jurs, Monte Carlo studies of the classification made by nonparametric linear discriminant functions, Jour. Chem. Inf. Comp. Sci. **25**, 45–50 (1985).

[21] Pino, J. A.; 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. **57**, 295–302 (1985).

[22] Pino, J. A., Pyrochromatography of human skin fibroblasts: normal subjects vs cystic fibrosis heterozygotes, Ph.D. Thesis, Cornell University (1984).

[23] Zander, G. S.; A. J. Stuper and P. C. Jurs, Nonparametric feature selection in pattern recognition applied to chemical problems, Anal. Chem. **47**, 1085–1093 (1975).

[24] Moriguchi, I.; K. Komatsu and Y. Matsushita, Adaptive least-squares method applied to structure-activity correlation of hypotensive N-Alkyl-N″-cyano-N′-pyridylguanidines, Jour. Med. Chem. **23**, 20–26 (1980).

[25] Vander Meer, R. K., and D. P. Wojcik, Chemical mimicry in the myrmecophilous beetle *myrmecaphodius excavaticollis,* Science **218**, 806–808 (1982).