M-2001

Chemometrics and Intelligent Laboratory Systems, 3 (1988) 79–89
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

Pattern Recognition Studies of Complex Chromatographic Data Sets: Design and Analysis of Pattern Recognition Experiments

DANIEL WOJCIK

B.K. LAVINE *

Department of Chemistry, Clarkson University, Potsdam, NY 13676 (U.S.A.)

P.C. JURS

Department of Chemistry, The Pennsylvania State University, University Park, PA 16802 (U.S.A.)

D.R. HENRY

Molecular Design Limited, 2132 Farallon Dr., San Leandro, CA 94577 (U.S.A.)

R.K. VANDER MEER

United States Department of Agriculture, Gainesville, FL 32604 (U.S.A.)

J.A. PINO and J.E. McMURRY

Baker Laboratory, Cornell University, Ithaca, NY 14853 (U.S.A.)

ABSTRACT

Lavine, B.K., Jurs, P.C., Henry, D.R., Vander Meer, R.K., Pino, J.A. and McMurry, J.E., 1988. Pattern recognition studies of complex chromatographic data sets: design and analysis of pattern recognition experiments. *Chemometrics and Intelligent Laboratory Systems*, 3: 79–89.

Chromatographic fingerprinting of complex biological and environmental samples is an active research area with a large and growing literature. Multivariate statistical and pattern recognition techniques can be effective methods for the analysis 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. Several interesting projects involving these effects and methods for dealing with the effects are discussed.

Complex chromatographic data sets often contain information dependent on experimental variables as well as information which differentiates 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 from cultured skin fibroblasts was used to differentiate cystic fibrosis carriers from presumed normal donors. Several experimental variables (donor gender, chromatographic column, etc.) were observed to contribute to the overall classification process. Notwithstanding these effects, discriminants were developed from the chromatographic peaks that assigned a given pyrochromatogram to its respective class (cystic fibrosis carrier versus normal) largely on the basis of the desired pathological difference. In

0169-7439/88/$03.50

another study gas chromatographic profiles of cuticular hydrocarbon extracts obtained from 170 red fire ant samples were analyzed using pattern recognition methods. Clustering according to the biological variables of social caste and colony was observed.

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

INTRODUCTION

Profiling of complex biological materials with high-performance chromatographic methods is an active research area with a large and growing literature [1–10]. 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 inividuals 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. Measurements 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, $X = (x_1, x_2, x_3, \ldots, x_j, \ldots, x_d)$ 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] and many were used in the two example studies below.

CLASSIFICATION

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, P, of fortuitously obtaining 100% correct classification for a two class problem using a nonparametric linear discriminant can be calculated from the following equation

$$P = 2 \sum_{i=0}^{d} C_i^{N-1}/2^N \quad \text{for} \quad N > d+1 \qquad (1)$$

$$P = 1 \quad \text{for} \quad N \leqslant d+1$$

where $C = (N-1)!/[(N-1-i)!i!]$, N is the number of objects in the data set, and d is the dimensionality or number of descriptors per object [16,17]. Fig. 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 they 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

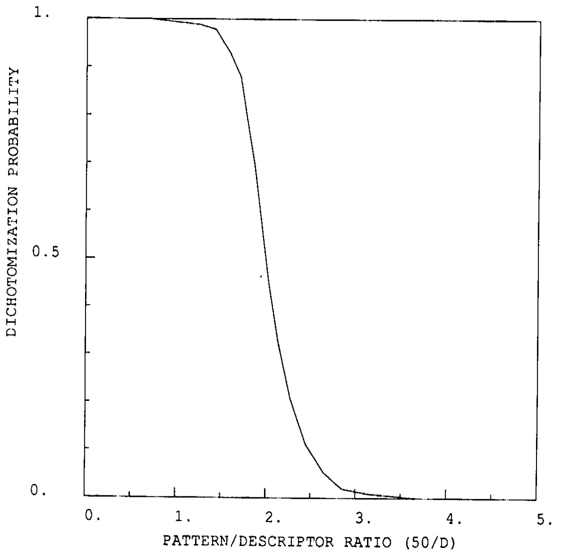

Fig. 1. Probability of dichotomization (complete linear separation) of fifty patterns as a function of n/d.

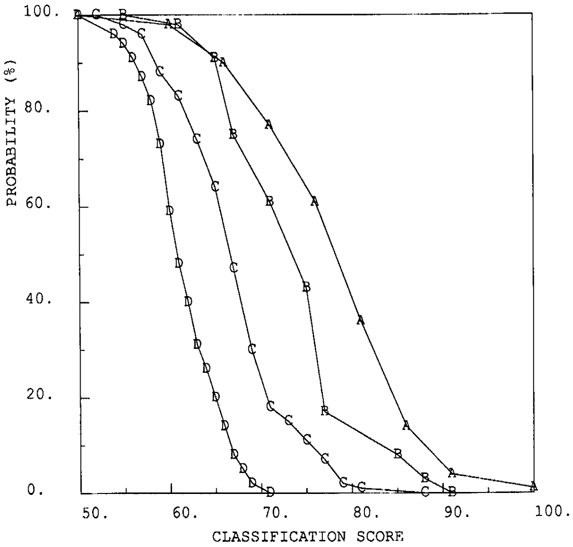

Fig. 2. Probability of achieving any degree of separation due to chance for several different values of n/d. A: 4:1; B: 6:1; C: 10:1; D: 20:1.

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 occurrence 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.

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 from training sets that are not linearly separable. Recently, Lavine [20] has reported Monte-Carlo simulation studies assessing the degree of fortuitous classification for such situations. Fig. 2 shows the cumulative probability of achieving any degree of separation due to chance for evenly-divided classes for several different values of n/d. The descriptors were uncorrelated, and the number of descriptors per pattern was set at five. At $n/d = 10$ (50 patterns, 5 descriptors), the probability is 0.5 that 66.7% of the patterns will be correctly classified. At $n/d = 4$ (20 patterns, 5 descriptors), the probability is 0.5 that 79% of the objects will be correctly classified due to chance. For each study, the apparent separability of the data increased as the n/d ratio decreased, even though there was no information contained within the descriptors and no possible relationship between the random patterns. Furthermore, the range of classifications in these studies also increased as the n/d ratio decreased. These results were obtained by using Gaussian distributed variables. Studies using uniformly distributed random numbers yielded essentially identical results.

If $n/d = 4$ (20 patterns, 5 descriptors, equal class size), the probability of achieving complete separation due to chance is approximately 3%. However, Stuper and Jurs [18] have previously reported in the literature that the probability of complete separation due to chance is small (< 1%) if $n/d > 3$. How can we explain this discrepancy? In Table 1, the probability that a dichotomy chosen at random will be linearly implementable was computed using eq. 1. The n/d ratio for each entry in the table is 3. As the number of descriptors increases, the probability of achieving com-

TABLE 1

The probability that a dichotomy chosen at random is linearly implementable

Number of descriptors	Number of patterns	Linear separability (%)
4	12	27.44
5	15	21.19
10	30	6.80
20	60	0.92
30	90	0.14

plete separation due to chance decreases. The three to one rule is, therefore, only applicable for those studies involving a large number (e.g., 30–40) of descriptors. For studies involving only a few descriptors (e.g., 5–10) a larger *n*/*d* ratio will be necessary to achieve the same.

The probability of achieving complete separation due to chance is dependent not only upon the *n*/*d* ratio but also upon the dimensionality of the data. It follows that for studies involving nonseparable training sets, the degree of separation will also be dependent upon both of these factors. Monte-Carlo simulation studies were performed to address this issue. Fig. 3 shows the cumulative probability of achieving any degree of separation due to chance for several different descriptor values at a fixed *n*/*d* ratio of 3. The patterns used to develop these curves were random, and equal class sizes were used. As the dimensionality of the uncorrelated data increased, the mean classification success rate also increased. However, the range of classifications expected from chance decreased. These results suggest the following: for a fixed value of *n*/*d*, the degree of separation in the data due to chance approaches a limiting value as the number of descriptors per pattern increases. This is contrary to what has been previously reported [21].

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. Furthermore, dependencies among the descriptors can affect the classification process. Fig. 4 is a plot of the results that were obtained in two Monte-Carlo simulation studies for 72 patterns. The first study involved uncorrelated random data, and the second study involved correlated random data. For each study,

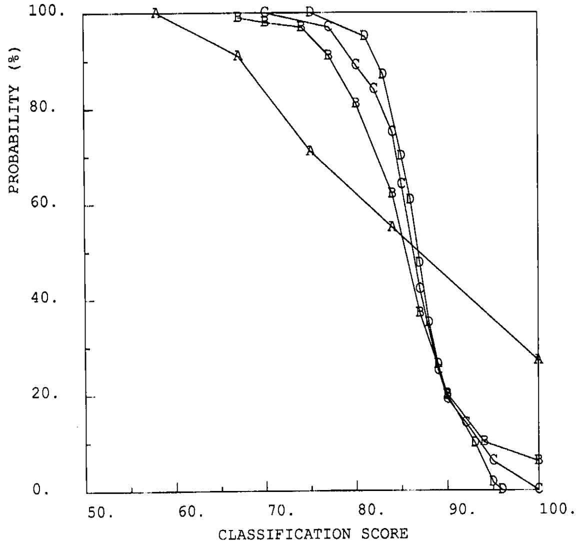

Fig. 3. Cumulative probability of achieving any degree of separation due to chance for several different descriptor values at a fixed *n*/*d* ratio of 3. A: 4 descriptors; B: 10 descriptors; C: 20 descriptors; D: 30 descriptors.

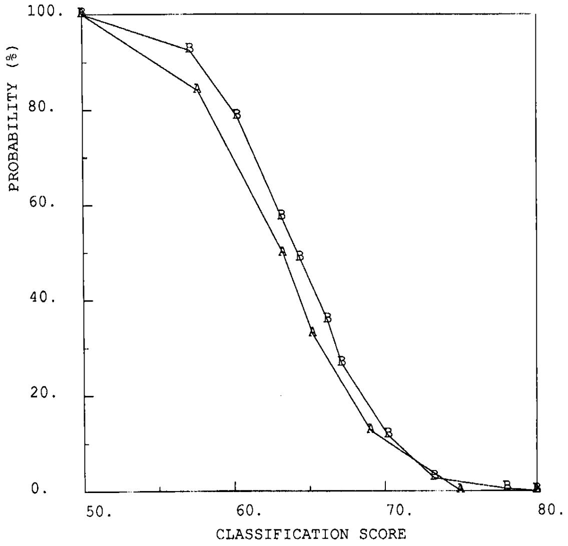

Fig. 4. Effects of the dependence structure of the data on random classification results. A: uncorrelated random data; B: correlated random data.

the number of descriptors per random pattern was set at five, and the classes were evenly divided. The correlation matrix used to generate the random data vectors was obtained from a study involving red fire ants. The purpose of that study was to separate red fire ants of one colony from another (colony E vs. colony R in this case) on the basis of their cuticular hydrocarbon profiles. It is apparent from the plot that the degree of separation due to chance is greater for correlated data than uncorrelated data.

Change classification can be a serious problem in linear discriminant analysis of chromatographic fingerprint data. The degree of separation in the data due to chance is dependent upon several factors (e.g. number of descriptors, *n/d* radio, etc.) Furthermore, our studies show that these factors do not act independently of one another. This complicates the interpretation of results that are obtained in discriminant analysis studies involving nonseparable training sets. Therefore, we recommend the following procedure for assessing the significance of classification scores obtained in real studies. For a given classification problem, 100 data sets consisting entirely of random numbers should first be generated. The statistical properties of the simulated data (i.e., dimensionality, number of objects, class membership distribution, and covariance structure) should be identical to that of the real training set. Next, the separability of each random data set should be 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, etc.) should be noted, and the fraction of the total number of occurrences (cumulative probability) for each degree of separation should be plotted against the percentage of patterns correctly classified. This curve as well as the mean classification success rate for the random data sets can then be used for comparisons.

As an example, if the classification score obtained in a real study was 80% but the mean classification success rate for the simulated random data was only 55% and the probability of achieving 65% correct classification due to chance was zero, the score obtained in the real study would be judged to be significant. On the other hand, if the classification score obtained in a real study was only 60% but the mean classification success rate for the random data was again 55% and the probability of achieving 60% correct classification was greater than 1%, the score obtained from this particular study would be judged not to be significant.

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.

CYSTIC FIBROSIS HETEROZYGOTES VS. NORMAL SUBJECTS

The first study involves the application of pyrolysis–gas chromatography (Py–GC) and pattern recognition methods to the problem of identifying carriers of the cystic fibrosis (CF) defect [22]. 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 Fig. 5. The pyrolyzed 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 [23]. 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

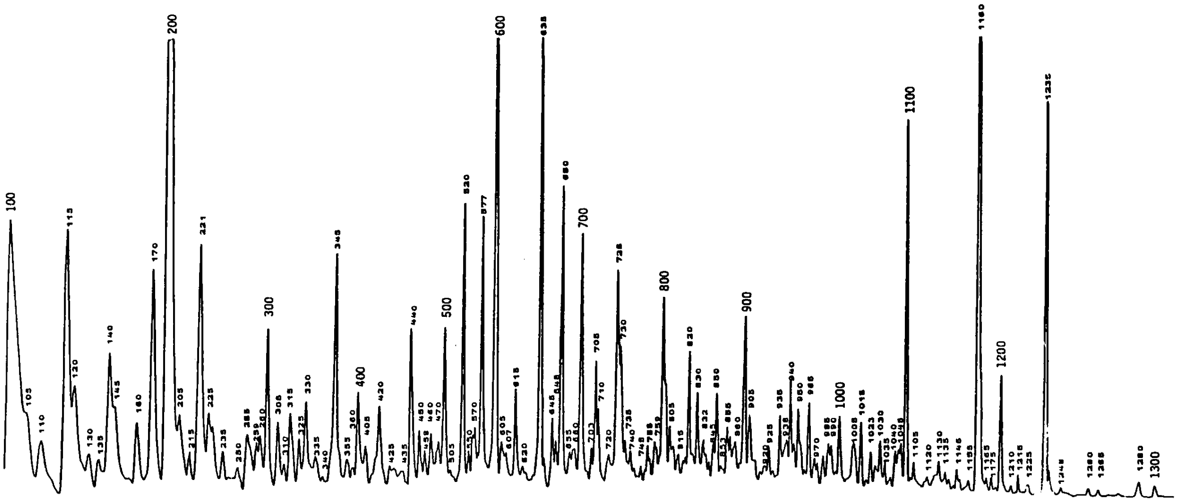

Fig. 5. A representative pyrochromatogram from the cystic fibrosis study. The peak identities are those assigned using the peak-matching software. The major peaks are those with assignments that are multiples of 100.

windows. Each pyrochromatogram was also normalized using the total area of the 214 peaks. This set of chromatographic data — 144 pyrochromatograms of 214 peaks each — was autoscaled so that each Py–GC 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 pyrochromatograms 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 comprising interval three. The CF pyrochromatograms were linearly separable from the pyrochromatograms 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 Py–GC 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 a discriminant is described in detail below.

The 65 peaks that were present in at least 90% of the pyrochromatograms were used as a starting point for the analysis. We assessed the ability of each of these 65 peaks alone to discriminate between pyrochromatograms 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 the 12 peaks using the *k*-nearest neighbor procedure correctly classified 90% of the pyrochromatograms in the data set. Variance features selection [24], combined with the linear learning machine [25] and the adaptive least-squares method [26] 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 six peaks was assessed by reordering experiments. The set of pyrochromatograms 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 Py–GC 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 pyrochromatograms 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. 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 170 red fire ant (*Solenopsis 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 [27]. A gas chromatographic trace of the cuticular hydrocarbons from a *S. invicta* sample is shown in Fig. 6. 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 gas chromatography–mass spectrometry 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

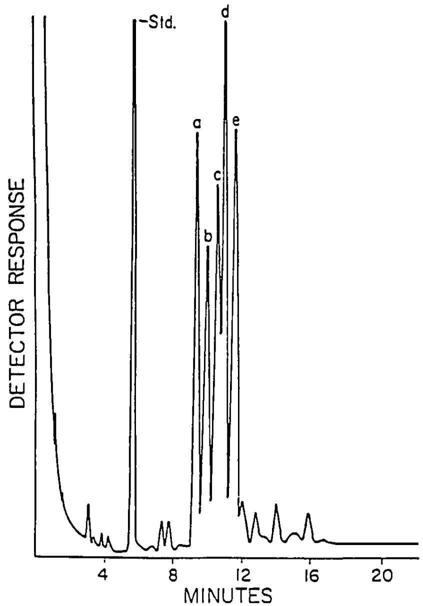

Fig. 6. Gas chromatographic trace of cuticular hydrocarbons from *S. invicta*. Peaks: a: heptacosane; b: 12-methylheptacosane; c: 13,15-dimethylheptacosane; d: methylheptacosane; e: 3,9-dimethylheptacosane.

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 social castes (foragers, reserves, and broods), and for four different time periods (the first three in the spring and summer, the fourth in the winter).

The first step was to use mapping and display methods [12,14] to examine the structure of the data. Methods used included principal components mapping and nonlinear mapping. In Figs. 7 and 8 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 Figs. 7 and 8. 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.

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. (These dichotomies reflect the choices facing the red fire ant.) The results of these discriminant analysis experiments are summarized in Table 2. The first row of the table shows that colony E could be separated from colony J by a discriminant that achieved 98% correct classifications (62 correct out of 63 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).

In order to assess the significance of these classification scores, Monte-Carlo simulation studies were performed. The purpose was to estimate the degree of separation in the data due to chance. For these studies data sets comprised of random numbers were generated. Both Gaussian and uniform distributions were employed. Results from these Monte-Carlo experiments are summarized in Table 3. As an example, for the colony E vs. colony J classification study, 100 data sets consist-

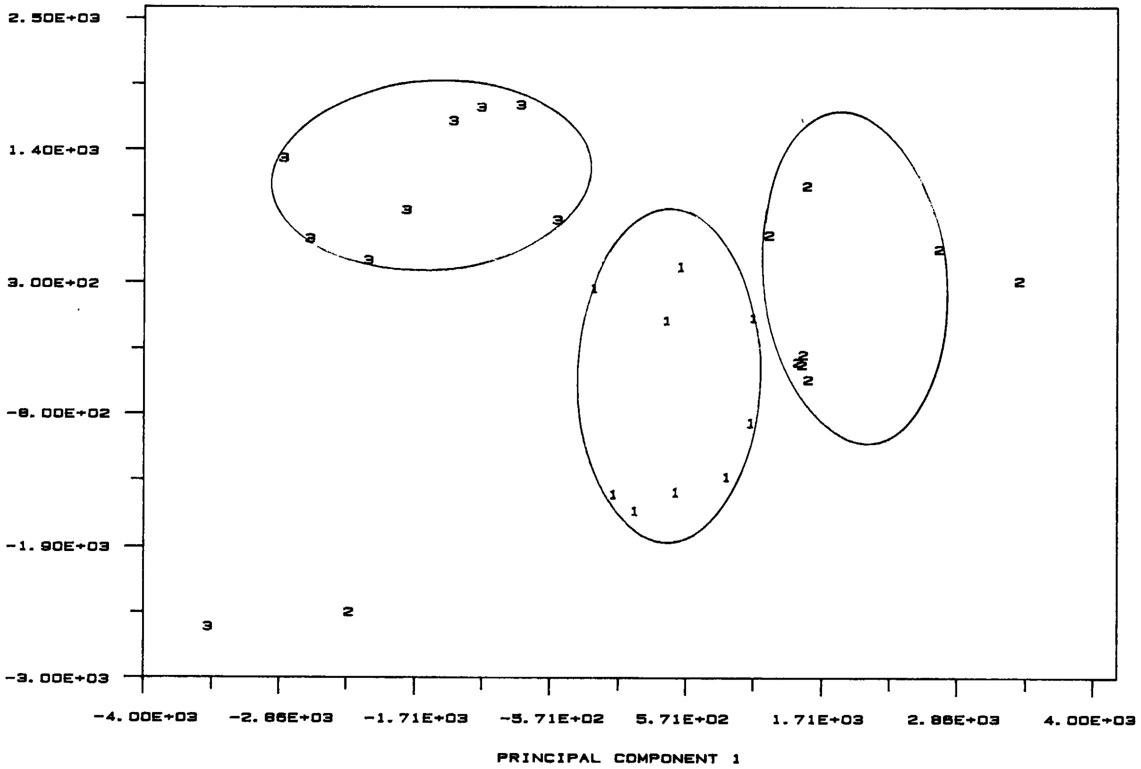

Fig. 7. Plot of the two principal components of the five gas chromatographic peaks for colony J. The ellipses show groupings of samples by time period. 1: time period 1; 2: time period 2; 3: time period 3.

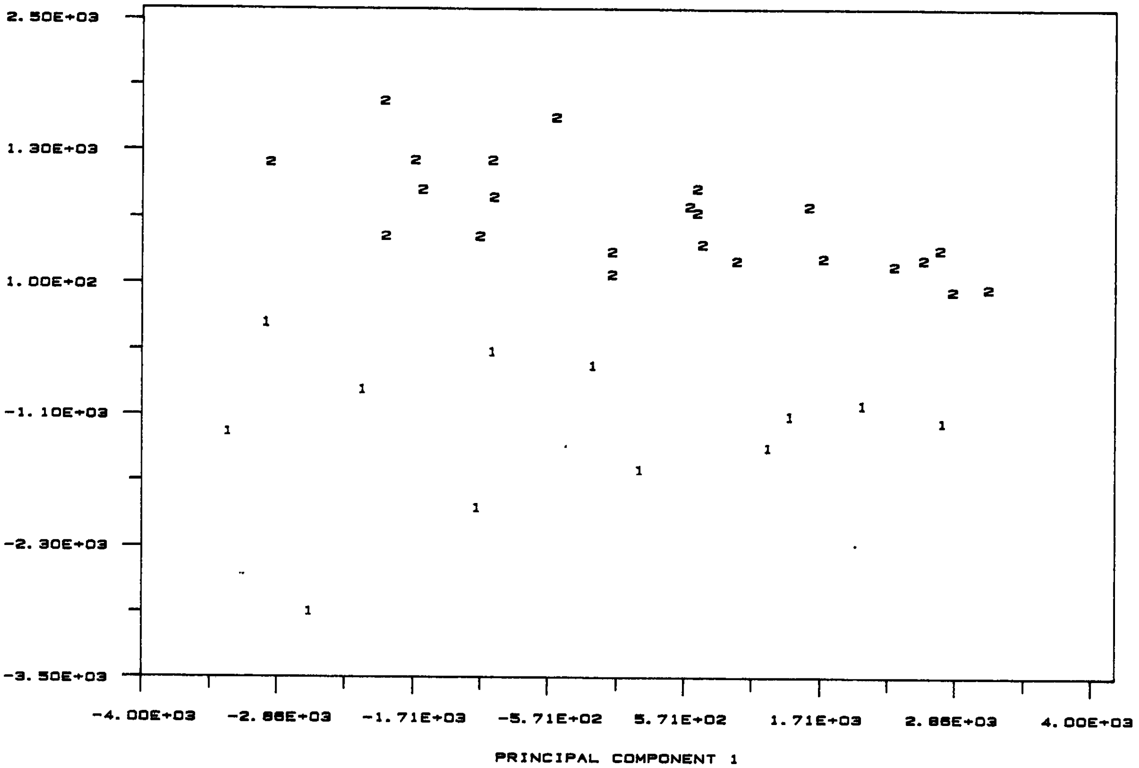

Fig. 8. Plot of the two largest principal components of the five gas chromatographic peaks for colony Q. The foragers are well separated from the reservers and broods in the principal component space. 1: foragers; 2: reservers and broods.

TABLE 2

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

Colony	Number 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

TABLE 3

Expected classification success rates for random data

Colony	Number in colony	Colony in second group (%)					
		E	J	P	Q	R	All
E	36	–	67.9	65.8	65.1	64.5	75.5
J	27		–	67.7	67.7	67.6	83.0
P	36			–	64.8	64.6	76.7
Q	35				–	64.2	76.0
R	36					–	76.6

ing entirely of random numbers were generated. The statistical properties of the simulated data (i.e. dimensionality, number of patterns, class membership distribution, and covariance structure) were identical to those of the colony E–colony J training set. The separability of each random data set was assessed using an iterative least-squares method. The mean classification success rate for the 100 data sets was computed and compared to the classification score obtained in the colony E–colony J study. Since the mean classification score of the simulated data was only 67.9% (versus 98% for the real training set), the classification score obtained in the colony E–colony J study was judged to be significant.

It is clear from the results of these Monte-Carlo experiments that the classification scores obtained in the linear discriminant analysis experiments listed in Table 2 are, by and large, significant. The only exceptions are those found in the following studies: colony Q versus colony R, colony Q versus all remaining colonies, and colony R versus all remaining colonies.

Multivariate statistical methods such as multivariate analysis of variance and stepwise logistic regression have also been employed in this study. The results obtained using these techniques support the conclusions drawn from the pattern recognition experiments. In summary, GC traces representing ant cuticle extracts could be related to colony identity, social caste, and time period using pattern recognition methods.

REFERENCES

1 E.J. Jellum, Profiling of human body fluids in healthy and diseased states using gas chromatography and mass spectrometry, *Journal of Chromatography*, 143 (1977) 427–462.

2 E. Reiner and F.L. Bayer, Botulism: a pyrolysis–gas–liquid chromatography study, *Journal of Chromatographic Science*, 16 (1978) 623–629.

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

4 E.J. Jellum, R. Bjoernson, R. Nesbakken, E. Johansson and S. Wold, Classification of human cancer cells by means of capillary gas chromatography and pattern recognition analysis, *Journal of Chromatography*, 217 (1981) 231–237.

5 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, *Journal of General Microbiology*, 128 (1982) 1773–1784.

6 M.L. McConnell, G. Rhodes, U. Watson and M. Novotny, Application of pattern recognition and feature extraction techniques to volatile constituent metabolic profiles obtained by capillary gas chromatography, *Journal of Chromatography*, 162 (1979) 495–506.

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

8 H.A. Scoble, J.L. Fasching and P.R. Brown, Chemometrics and liquid chromatography in the study of acute lymphocytic leukemia, *Analytica Chimica Acta*, 150 (1983) 171–181.

9 G. Rhodes, 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, *Clinical Chemistry*, 27 (1981) 580–585.

10 O.M. Kvalheim, K. Ygard and O. Grahl-Nielsen, SIMCA multivariate data analysis of blue mussel components in environmental pollution studies, *Analytica Chimica Acta*, 150 (1983) 145–152.

11 P.C. Jurs and T.L. Isenhour, *Chemical Applications of Pattern Recognition*, Wiley-Interscience, New York, 1978, pp. 31–83.

12 M.A. Sharaf, D.L. Illman and B.R. Kowalski, *Chemometrics*, Wiley-Interscience, New York, 1986, pp. 216–219.

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

14 D.L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis*, Wiley-Interscience, New York, 1983, pp. 75–87.

15 A.J. Stuper, W.E. Brugger and P.C. Jurs, *Computer Assisted Studies of Chemical Structure and Biological Function*, Wiley-Interscience, New York, 1979, pp. 126–136.

16 N.J. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965, p. 25.

17 J.T. Tou and R.C. Gonzalez, *Pattern Recognition Principles*, Addison Wesley Publ. Co., Reading, MA, 1974, p. 61.

18 A.J. Stuper and P.C. Jurs, Reliability of nonparametric linear classifiers, *Journal of Chemical Information and Computer Science*, 16 (1976) 238–241.

19 E.K. Whalen-Pedersen and P.C. Jurs, The probability of dichotimization by a binary linear classifier as a function of training set population distribution, *Journal of Chemical Information and Computer Science*, 19 (1979) 264–266.

20 B.K. Lavine, *Pattern recognition studies of complex chromatographic data sets*, Ph.D. Thesis, The Pennsylvania State University, 1986.

21 T.R. Stouch and P.C. Jurs, Monte Carlo studies of the classification made by nonparametric linear discriminant functions, *Journal of Chemical Information and Computer Science*, 25 (1985) 45–50.

22 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, *Analytical Chemistry*, 57 (1985) 295–302.

23 J.A. Pino, *Pyrochromatography of human skin fibroblasts: normal subjects vs. cystic fibrosis heterozygotes*, Ph.D. Thesis, Cornell University, 1984.

24 G.S. Zander, A.J. Stuper and P.C. Jurs, Nonparametric feature selection in pattern recognition applied to chemical problems, *Analytical Chemistry*, 47 (1975) 1085–1093.

25 Sing-Tze Bow, *Pattern Recognition*, Marcel Dekker, New York, 1984, p. 57.

26 I. Moriguchi, K. Komatsu and Y. Matsushita, Adaptive least squares method applied to structure–activity correlations of hypotensive N-alkyl-N″-cyano-N″-pyridylguanidines, *Journal of Medicinal Chemistry*, 23 (1980) 20–26.

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