MICROCHEMICAL JOURNAL 41, 288–295 (1990)

False Color Data Imaging: A New Pattern Recognition Technique for Analyzing Chromatographic Profile Data

BARRY K. LAVINE,*,[1] ROBERT K. VANDER MEER,† LAURENCE MOREL,† ROBERT W. GUNDERSON,‡ JIAN HWA HAN,* AND ARTHUR STINE*

Department of Chemistry, Clarkson University, Potsdam, New York 13676; †USDA-ARS, Insects Affecting Man and Animals Laboratory, Gainesville, Florida 32604; and ‡Department of Electrical Engineering, Utah State University, Logan, Utah 84322

Received November 24, 1989; accepted December 15, 1989

Fuzzy pattern recognition techniques were used to analyze the gas chromatograms of hexane soaks obtained from red carpenter ants. The soaks were found to convey information about the caste and social experience of the ants. In this study a new mapping and display technique, which depends upon high-resolution computer graphics for the presentation of results, was used to develop these relationships. © 1990 Academic Press, Inc.

INTRODUCTION

Profiling of complex biological materials with high-performance chromatographic methods has been an active area of research over the past 15 years (*1–10*). The object of profile analysis is to correlate a characteristic fingerprint pattern in a chromatogram with a specific property of a sample. Chromatographic fingerprinting experiments often yield chemical profiles containing hundreds of constituents; objective analysis of the profiles depends upon the use of multivariate statistical methods. However, there has been little research focusing on the development of new methods to handle data generated in such experiments, as evidenced by the few papers on this subject that have appeared in the chemical literature in recent years (*11–15*).

Recently, our laboratories applied a new pattern recognition technique to the analysis of chromatographic profile data, FCV–false color data imaging (*16, 17*). This mapping and display technique utilizes high-resolution computer graphics for the presentation of results. The technique provides information to a scientist about trends present in multivariate data by transforming the data matrix into a picture which shows the relationship(s) between samples and their measurements in the data set.

This paper describes a study recently completed in our laboratories dealing with the cuticular chemistry of *Camponotus floridanus*. The goals of the study were twofold: (i) to assess the utility of FCV–false color data imaging for studying complex chromatographic data sets, and (ii) to identify fingerprint patterns in the chromatograms which are characteristic of the caste and social experience of the red carpenter ants. In this study capillary column gas chromatography was used

[1] To whom correspondence should be addressed.

0026-265X/90 $1.50

to analyze soaks obtained from 59 red carpenter ants. The GC traces of the soaks were then analyzed using FCV–false color data imaging and fuzzy pattern recognition techniques (i.e., the FCV clustering algorithm). The analytical methodology employed in this study with particular emphasis on the pattern recognition techniques which were used to solve this rather interesting classification problem is the focus of this report.

EXPERIMENTAL

Each ant was soaked in 150 μl of high-performance liquid chromatographic grade hexane for 3 h. The soaks from each ant were concentrated to about 20 μl under a stream of nitrogen. A Varian 3700 gas chromatograph equipped with a flame ionization detector was used to analyze the soaks. A representative gas chromatographic trace of the extract obtained from a forager is shown in Fig. 1. The chromatographic experiments were performed with a 30-m DB-1 fused silica capillary column, which was temperature programmed from 50 to 285°C at 5°C per minute. Further details regarding the collection of the gas chromatographic data can be found elsewhere (*18*).

Soaks were obtained from 59 red carpenter ants (see Table 1) which were procured from a single laboratory colony maintained in the USDA–ARS Fire Ant

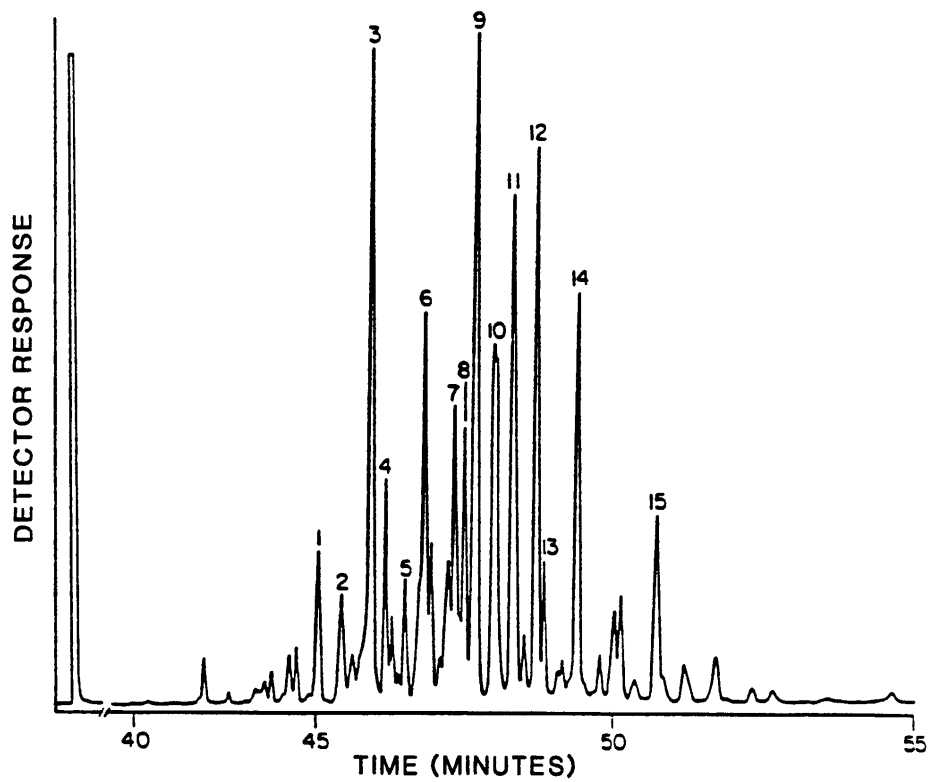

FIG. 1. A representative gas chromatographic trace of a forager. The 15 peaks used in the pattern recognition analysis are shown.

Project Laboratory in Gainesville, Florida. The ants were housed in petri dishes and were fed regularly with honey–water (1:1) and immature insects. Four different worker categories were represented in the data set: (i) foragers, (ii) nurses, (iii) normal callow workers, and (iv) naive callow workers (Table 1). Callows were removed from each colony as they emerged from their cocoon and were maintained in separate petri dishes with nurses (normal callows) or without nurses (naive callows).

DATA PREPROCESSING

Each chromatogram was initially represented by a data vector $X = (x_1, x_2, x_3, x_4, \ldots\ldots x_j, \ldots\ldots x_n)$, where component x_j is the area of the jth peak. In this study each chromatogram was normalized to constant sum using the area of all 40 GC peaks. Therefore, each peak was expressed as percentage of total area to indicate the relative concentration of the hydrocarbons. Of the 40 peaks comprising each chromatogram, only 15 were considered for pattern recognition analysis (see Fig. 1). The 15 peaks possessed a common set of attributes: (i) each peak had an area representing more than 1% of the total; (ii) each peak had a reliable area count; and (iii) each peak was well resolved and readily identifiable in all of the chromatograms so peak matching was not a problem. The feature selection process was made on the basis of objective criteria, not class information. The probability of exploiting random variation in the data was, therefore, minimized.

Of the 59 samples, 51 make up a training set. There were 15 foragers, 11 normal callows, 10 naive callows, and 15 nurses. The prediction set consisted of 8 ant samples (six foragers, and two naive callows). Members of the prediction set were chosen by random lot.

PATTERN RECOGNITION ANALYSIS

In this study FCV–false color data imaging was used to seek relationships between the GC profiles of the ants and the biological variables: caste and social experience. This technique utilizes false color data imaging (*19*) and the FCV clustering algorithm (*20–22*). In a false color data imaging experiment, each data vector is projected onto a three-dimensional coordinate system. Each of the three axes is assigned a primary color, i.e., red, yellow, or blue. A given data point can then be assigned a color which is a combination of the three primary colors. The intensity of each primary color is inversely scaled according to the distance of the

TABLE 1
The *Camponotus floridanus* Data Set

Worker's category	Number of samples
Nurses	15
Foragers	21
Normal callows	11
Naive callows	12
Total	59

data point from the particular color axis in question. Thus, data projected onto a line equidistant from all three coordinate axes would be assigned a color composed of equal intensities of the three primary colors, i.e., some shade of white.

The FCV clustering algorithm attempts to fit each of the γ classes in the data set to a principal component model. An interesting feature of the FCV clustering algorithm is that each data vector in the training set is assumed to contribute to the modeling of each of the classes within the data. The actual algorithm consists of simultaneously solving the following set of equations:

$$u_{ik} = 1/\sum_{j=1}^{c} (D_{ik}/D_{jk})^{1/(m-1)} \tag{1}$$

$$v_i = \sum_{k=1}^{n} (u_{ik})^m x_k/(u_{ik})^m \tag{2}$$

$$S_i = \sum_{k=1}^{n} (u_{ik})^m (x_k - v_i)(x_k - v_i)^T \tag{3}$$

$$D_{ik} = \left(|x_i - v_i|^2 - \sum_{j=1}^{r} \langle x_k - v_{ij}, d_{ij} \rangle^2 \right)^{1/2} \tag{4}$$

The membership value of sample k with respect to class i ($i = 1, 2, 3, \ldots$) is u_{ik}, and these values are subject to the condition $0 < u_{ik} < 1$ and $\Sigma u_{ik} = 1$. D_{ik} is the distance of sample k from cluster center i, v_i is the center of cluster i (i.e., class i), d_{ij} is a unit eigenvector corresponding to the jth largest eigenvalue of the fuzzy within cluster scatter matrix S_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 round clusters, $r = 1$ for linear varieties, and so forth.)

To obtain an approximate solution to this set of four equations, the user must supply the starting cluster centers. The class membership values, the within cluster scatter matrix for each cluster, and the distance of each sample from each cluster are computed in rapid succession. New cluster centers are then computed for the samples in the final step of the first iteration. The algorithm uses these new cluster centers as the starting point for a second iteration through the same set of four equations. This process continues until convergence is achieved. The number of iterations required to achieve convergence depends upon the minimum prespecified change criterion for the class membership value.

The first step in a false color data imaging experiment is to project the original data (for our study points in a 15-dimensional space) onto a suitable 3-dimensional subspace. We used the first three principal components (*23*) of the data to define the subspace. Since this can be easily accomplished with microcomputer graphics for even large data sets, it is a reasonable first step in any analysis.

However, there is a problem associated with principal component maps. The

spatial relationships between individual data points and between groups of data points in the original measurement space are often distorted in the projective process. It is at this point that false color data imaging departs substantially from other graphical display techniques. We will take advantage of the membership coefficients generated by the FCV clustering algorithm to restore much of the missing spatial information through the use of color, a crucially important information dimension. Using the FCV clustering algorithm, the data will be fitted to a user-specified number of principal component models. Each model which represents a different cluster of points will then be assigned a different basic color: yellow, red, blue, and so forth. A given data point will then be displayed as a combination of these colors, with the amount of any one color determined by the sample's membership value for that particular class. Interpretation of the resulting color images will provide valuable insight into the data structure. For example, two projected data points do not lie close to one another in the high dimensional space (and hence are not similar) unless they are displayed in nearly the same color. A single cluster of data points in the three-dimensional principal component space, which is made up of two different basic colors, will be interpreted as two distinct clusters whose true multidimensional separation has been lost through projection. A solid group of data appearing in a nonbasic color suggests the presence of an unsuspected class, and so forth.

RESULTS AND DISCUSSION

The starting centers for each FCV–false color data imaging experiment were determined by the single linkage hierarchical clustering method (*24*). In these imaging experiments, r was set equal to zero, m was set equal to 2, and the minimum prespecified change criterion for the class membership value was set at 0.005. The value c was varied from 2 to 6. (For GC profile data, we have learned from previous studies that r should usually be set equal to zero.) When c was set equal to 4 (see Fig. 2), an interesting result was obtained. Foragers, represented by the yellow triangles and yellow squares, are well separated from the nurses. They are represented by the 13 orange triangles and two of the white triangles. Normal callows, represented by 11 of the 15 white triangles, lie in the same region of the principal component map as most of the naive callows which are represented by blue triangles. (The other two naive callows are represented by white triangles in the principal component map.) Because the points in this region of the map are made up of two different colors, they actually represent two distinct data clusters (which are predominantly comprised of normal callow and naive callow samples) whose true multidimensional separation has been lost through projection. Evidently, much of the information which is lost when the points are projected onto a lower dimensional subspace can be restored through the use of color.

The FCV clustering algorithm was then used to develop a principal component model for each cluster. For this clustering experiment, again, c was set at 4, r was set at 0, and m was set at 2. The starting centers for the experiment were a normal callow, a naive callow, a nurse, and a forager. (Each of these samples was displayed as a different color in the principal component map.) The algorithm con-

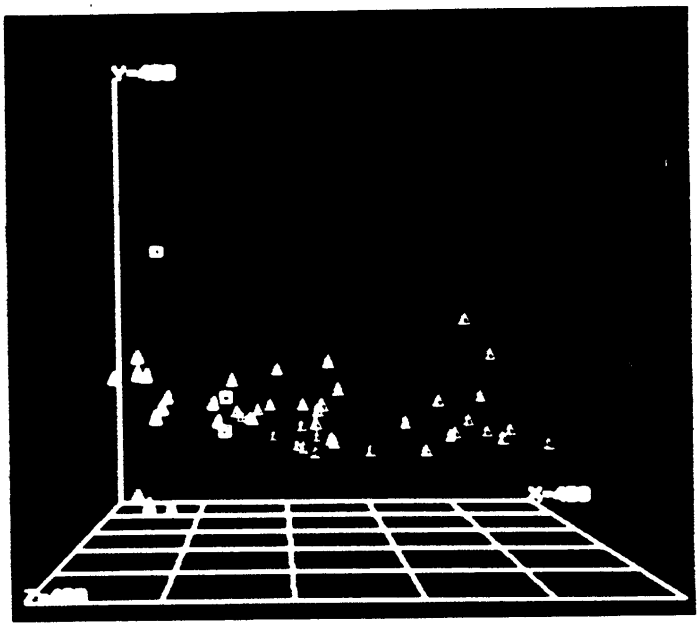

FIG. 2. A false color data image print of the 51 samples which comprised the training set.

verged after 20 iterations. The foragers had a large class membership value (greater than 0.60) for cluster one. None of the other samples had a large class membership value for that cluster. All of the normal callows had a large class membership value for cluster two. Only four other samples (two nurses and two naive callows) had a large membership value for cluster two. Eight of the ten naive callows had a large membership value for cluster three, and 13 of the 15 nurses had a large membership value for cluster four. When the ant samples were assigned to the cluster where they had the highest class membership value, it was apparent the data could be partitioned into different classes based on the ant's caste and social experience. These results are summarized in Table 2.

The principal component models developed in this experiment were validated using the samples from the prediction set. The class membership value for each

TABLE 2
Training Set Results

Number of samples	Worker's category	Maximum membership cluster
15	Foragers	1
11	Normal callows	2
2	Nurses	2
2	Naive callows	2
8	Naive callows	3
13	Nurses	4

prediction set sample was computed for each principal component model. The sample was then assigned to the cluster where it had the highest class membership value. A classification success rate of 100% was achieved for the prediction set (i.e., the six foragers were assigned to cluster one, and the two naive callows were assigned to cluster three). Evidently, the caste and social experience of red carpenter ants can be directly correlated to specific fingerprint patterns in the chromatograms.

The results of this study also demonstrate that FCV–false color data imaging is an important tool for uncovering obscure relationships which are often present in multivariate data sets. By visually examining a three-dimensional map which adequately represents the distribution of the data points in the high-dimensional measurement space, it will be possible for a scientist to assess the structural characteristics of a set of data by organizing the data into subgroups, clusters, or hierarchies. Furthermore, we wish to note that principal component analysis and the FCV clustering algorithm are able to handle data sets which have a low object to descriptor ratio. Therefore, FCV–false color data imaging will also be able to handle data sets which have a low object to descriptor ratio. This means FCV–false color data imaging will be especially attractive for tackling data sets generated in chemical profiling studies.

ACKNOWLEDGMENTS

This study was partially supported by Contract FO8635-90-C-0105 between Clarkson University and the United States Air Force. The authors thank Janet M. Lavine for many valuable discussions.

REFERENCES

1. Clark, H. A.; Jurs, P. C. *Anal. Chem.* 1975, **47**, 374–378.
2. Reiner, E.; Bayer, F. L. *J. Chrom. Sci.* 1978, **16**, 623–629.
3. McConnell, M. L.; Rhodes, G.; Watson, U.; Novotny, M. *J. Chromatogr.* 1979, **162**, 495–506.
4. Jellum, E. J.; Bjoernson, R.; Nebakken, R.; Johansson, E.; Wold, S. *J. Chromatogr.* 1981, **217**, 231–237.
5. Soderstrom, B.; Wold, S.; Blomquist, G. *J. Gen. Microbiol.* 1982, **128**, 1773–1784.
6. Scoble, H. A.; Fasching, J. L.; Brown, P. R. *Anal. Chim. Acta* 1983, **150**, 171–181.
7. Kvalheim, O. M.; Ygard, K.; Grahl-Nielsen, O. *Anal. Chim. Acta* 1983, **150**, 145–152.
8. Stenroos, L. E.; Siebert, K. J. *J. Amer. Soc. Brew. Chem.* 1984, **42**, 54–61.
9. Pino, J. A.; McMurry, J. E.; Jurs, P. C.; Lavine, B. K.; Harper, A. M. *Anal. Chem.* 1985, **57**, 295–302.
10. Kvalheim, O. M. *Chem. Lab.* 1987, **2**, 127–136.
11. Dunn, W. J.; Stalling, D. L.; Schwartz, T. R.; Hogan, J. W.; Petty, J. D.; Johansson, E.; Wold, S. *Anal. Chem.* 1984, **56**, 1308–1313.
12. Jurs, P. C.; Lavine, B. K.; Stouch, T. R. *J. Res. N. B. S.* 1985, **90**, 543–549.
13. Eide, M. O.; Kvalheim, O. M.; Telnaes, N. *Anal. Chim. Acta* 1986, **191**, 433–437.
14. Telnaes, N.; Bjorseth, A.; Christy, A.; Kvalheim, O. M. *Chem. Lab.* 1987, **2**, 149–153.
15. Lavine, B. K.; Jurs, P. C.; Henry, D. R.; Vander Meer, R. K.; Pino, J. A.; McMurry, J. E. *Chem. Lab.* 1988, **3**, 79–89.
16. Gunderson, R. W.; Thrane, K. E.; Nilson, R. D. *Chem. Lab.* 1988, **3**, 119–132.
17. Lavine, B. K.; Vander Meer, R. K.; Morel, L.; Gunderson, R. *Pattern Recognition Studies in Chemical Communication.* FACSS, Boston, 1988.

18. Morel, L.; Vander Meer, R. K.; Lavine, B. K. *Behav. Ecol. Sociobiol.* 1988, **22**, 175–183.
19. Ekftrom, M. P. *Digital Imaging Techniques*, Academic Press, Orlando, FL, 1984.
20. Bezdek, J. C.; Coray, C.; Gunderson, R.; Watson, J. *Siam J. Appl. Math.*, 1981, **40**, 358.
21. Gunderson, R. W.; Jacobson, T. *ASBC J.*, 1983, **41**(2), 73.
22. Jacobsen, T.; Gunderson, R. W. *Int. J. Man-Mach. Stud.* 1983, **19**, 105.
23. Jolliffe, I. T. *Principal Component Analysis*, Springer–Verlag, New York, 1986.
24. Massart, D. L.; Kaufman, L. *The Interpretation of Analytical Chemical Data By Use of Cluster Analysis*, pp. 78–80. Wiley, New York, 1986.