Journal of Chromatography, 349 (1985) 31–38
Elsevier Science Publishers B.V., Amsterdam — Printed in The Netherlands

CHROMSYMP. 640

USE OF COMPUTERIZED PATTERN RECOGNITION IN THE STUDY OF THE CUTICULAR HYDROCARBONS OF IMPORTED FIRE ANTS

I. INTRODUCTION AND CHARACTERIZATION OF THE CUTICULAR HYDROCARBON PATTERNS OF *SOLENOPSIS INVICTA* AND *S. RICHTERI*

JEFFREY H. BRILL*, HOWARD T. MAYFIELD, TOM MAR and WOLFGANG BERTSCH
Department of Chemistry, University of Alabama, University, AL 35486 (U.S.A.)

SUMMARY

A method is described in which gas chromatographic (GC) data obtained from cuticular hydrocarbons are treated by methods of pattern recognition. Based on a recently described sample preparation procedure, GC data are normalized to eliminate slight variations in chromatographic conditions and converted into the proper format for discriminant analysis by computer. The results of several methods of data treatment and display are discussed, based upon the chemometric system package, ARTHUR. The approach has the advantage of largely removing operator bias.

INTRODUCTION

Recently the cuticular hydrocarbons of insects have been receiving much attention in the literature. Besides being involved in preventing desiccation, the cuticular hydrocarbons also play a significant role in chemical communication[1–4]. The semiochemical functions of the cuticular hydrocarbons include territory marking, recruitment and alarm pheromones, kairomones and defensive secretions, as well as sex pheromones[5].

The two species of imported fire ants, *Solenopsis invicta* (red imported fire ant) and *S. richteri* (black imported fire ant), are serious pests in the southern United States. The cuticular hydrocarbons of these two species have been investigated and characterized[3,6]. It has been suggested that the cuticular hydrocarbons of insects are also involved in species and caste recognition[5,7]. When ants encounter each other, recognition occurs by the one ant brushing its antennae over the cuticle of the other ant[8]. This suggests that the cuticle acts as a source of semiochemicals which are species- and colony-specific. Besides the use of different substances as chemical messengers, different species or colonies may also use the same compounds, but in different mixtures, to communicate chemical messages[9]. If this is the case, the cuticular hydrocarbon profiles could become important for understanding ant communication.

Variations in the cuticular hydrocarbon patterns between different samples,

0021-9673/85/$03.30 © 1985 Elsevier Science Publishers B.V.

i.e. individual ants, colonies, or species, have to be investigated by statistical procedures, which take into consideration the variations within the same sample and between samples from different sources. Such methods are termed pattern recognition and are considered to fall into the domain of chemometrics[10]. Properties of different samples can be related to each other basically in two ways. In certain cases it is necessary to investigate how a particular property changes as a function of some external variable. This method is termed continuous property analysis. The change in profile in the cuticular hydrocarbons of an insect as a function of geographic location or season is an example. In cases where it is important to point out systematic differences between different types of samples, discontinuous property analysis is applied. The desired result is a clustering of the most distinguishing properties of the sample sets, and the task essentially is treated by principles of information theory[11].

Pattern recognition is a process whereby a hidden property of a collection of objects (in this case species, colonies, etc.) can be detected and/or predicted by using indirect measurements on the individual objects[12].

In most cases, a single, discriminating measurement cannot be found. Only a combination of measurements provides sufficient information. When dealing with a small number of measurements (three or less), the human perception is the best pattern recognizer. However, when the number of objects and measurements greatly exceeds three or four, the problem can only be handled successfully by using computerized pattern recognition procedures.

Pre-packaged computer programs are available for such purposes. The program used in this application, termed ARTHUR (available from Kowalski*, at nominal charge), consists of several subprograms, some of which are particularly suitable for cluster analysis. Although ARTHUR runs on main frame computers, software designed for microcomputers is currently being introduced, bringing these techniques well within the range of small, applications-oriented laboratories.

In pattern recognition, patterns of profiles such as gas chromatograms are examined. The data measured, *i.e.* gas chromatographic (GC) peaks, are called features. The same measurements are performed on each individual sample, giving rise to a series of chromatograms which differ by the magnitude and/or the nature of the peaks. Samples which are closely related usually share the same peaks and only differ in their relative magnitudes. Pattern recognition studies are usually conducted in a series of steps. In the first step, known as preprocessing, a calibration set containing objects of known class are characterized numerically. This is done by converting the GC data (*i.e.* retention times and peak areas) into ordered vectors, called data vectors or pattern vectors[13,14]. The resulting data matrix is known as a training set. The objects of unknown class(es) are then characterized in a similar fashion, to form a test set. The second step involves deriving a mathematical model from each of the training sets. This then allows the test set to be classified. The whole process is called supervised learning.

The basic goal of all pattern recognition analyses is to associate or recognize unobservable properties of samples with a set of observable properties provided by

* B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, WA, U.S.A.

the data vectors[12]. In this case, the unobservable property is the unknown identity or class of an ant and the observable properties are the data vectors resulting from the GC profile of the insect.

It is advantageous in this process to favor the features which carry the largest amount of information. If additional features are used, noise or information unrelated to the problem of interest is introduced into the system. This increases the difficulty of the analysis and may also detract from its reliability. The process of choosing which features to use is called feature selection[10,13,14]. The principal feature selection methods used for this work were autoscaling and Fisher weighting. Autoscaling prevents the data set from being biased by the average sizes or magnitudes of the features. Fisher weighting assigns greater importance to those features which vary little within a given category, but vary a great deal over the entire data set[13].

It is practically impossible for a person to visualize the data, when displayed in n-dimensional space (where $n > 3$). Therefore, computers are used to project an approximation of the points from n-space into two-dimensional space, to permit visual inspection of the data. This procedure is known as non-linear mapping and is achieved by taking non-linear combinations of the n coordinates of the n data vectors[15].

The ARTHUR package provides the four classifications methods used in this application. These methods are the Bayes method, the K-nearest neighbor (KNN) method, the linear learning machine (LLM) method and the SIMCA method (named for statistical isolinear multicomponent analysis).

In the Bayesian classification, for each class, as well as over all objects in the training set, the frequency distributions of each feature are determined. This allows a probability measure, which describes the fit of an object to a class, to be estimated, based upon how well the data vector elements of the object fit the class frequency distributions. The probability of each object is calculated for each class and the object is assigned to the class with the highest probability.

In the KNN procedure, classification is based upon the distance of a sample to its K-nearest neighbors. The objects are considered as points in an n-dimensional hyperspace, where n is equal to the number of measurements made on each object. This procedure is based on the assumption that nearness in space between two points is a good measure of similarity between the corresponding objects[12].

The LLM procedure determines $(Q-1)$ hyperplanes with a dimensionality of $(M-1)$, using a feedback procedure in such a way that the different classes fall on different sides of the hyperplanes[16].

The SIMCA method is related to methods of factor analysis and principal components analysis. Additional details are available elsewhere[10,14].

EXPERIMENTAL

Samples

The fire ants used in these studies were collected from nests located in Pickens county, in west central Alabama, U.S.A.

Samples were prepared using a recently described dynamic headspace analysis procedure. The details of this method are described elsewhere[17]. In each study, a single ant was placed in a quartz sample tube in a Pyroprobe® (CDS 100, Chemical

Data Systems, Oxford, PA, U.S.A.), which was then inserted into the inlet of a gas chromatograph. The cuticular hydrocarbons were desorbed from the specimen by rapid heating of the Pyroprobe to 300°C and maintaining that temperature for 5 s. Besides the cuticular hydrocarbons, other compounds (such as the venom alkaloids) are desorbed as well. However, these components either lie outside the diagnostic region of the chromatogram, or they can be eliminated from the chromatogram by using a selective detector such as a mass spectrometer[17].

Gas chromatography

The gas chromatograph used was a Hewlett-Packard 5830A, fitted with an injector port suitable to accept the Pyroprobe insert. The gas chromatograph was controlled by a 18850A GC terminal. The column used was a 16 m × 0.25 mm I.D. WCOT glass capillary, coated with a 0.25-μm film of immobilized OV-1. The column was temperature-programmed from 80 to 300°C at 8°C/min. The carrier gas was helium, at a flow-rate of *ca.* 1.0 ml/min. The split ratio was set to *ca.* 100:1. Each sample was treated with 250 ng of dotriacontane ($C_{32}H_{66}$) as an internal standard.

Data analysis

The retention time and area of each hydrocarbon peak was encoded onto a computer card. All of the data were then transferred to a UNIVAC mainframe computer for data handling. This process, however, can (and has been) done on-line[18]. At present, we are in the process of adapting our system so that the data can be collected directly from the gas chromatograph by a microcomputer and then transferred to the mainframe computer for data handling.

The data were entered in a format compatible with the SETUP GC transducing program[19], which also served to adjust chromatographic retention times for proper feature assignment. The card-image data were saved on disk and magnetic tape. Seven marker peaks were assigned from peaks which occurred in all of the chromatograms. Marker peaks are a form of internal standardization. Peaks designated as markers may be internal standards, or peaks which are common to all the chromatograms. The data were transduced into a multivariate form by the SETUP program using the conditions given in Table I. The resulting data set consisted of 49 data vectors, each containing 52 features. *S. richteri* was represented by 29 data vectors, obtained from 4 queens, 6 alate males, and 19 workers. This species is designated category 1 in the pattern recognition treatment. *S. invicta* was represented

TABLE I

ADJUSTABLE PARAMETERS USED FOR DATA SET TRANSDUCTION BY SETUP

Parameter	
Maximum allowed retention time error for matching peaks	0.10 min
Number of marker peaks per chromatogram	7
Minimum retention time distance between non-redundant features*	0.05 min.
Minimum frequency of occurrence of acceptable features*	0%

* Features are based on normalized peak areas.

by 20 data vectors, obtained from 1 queen and 19 workers. *S. invicta* was designated category 2.

This data set was submitted to the ARTHUR chemometric package for analysis[20]. The features were autoscaled, and an investigation was made to determine the number of features which gave the best classification results, as judged by the separation of the two clusters. Fisher weighting was used as a basis for the selection of data sets consisting of the two, three, four and six most discriminating features of the original data set. From each of these data sets non-linear maps were generated, in order to visualize the separations in the resulting data spaces. The maps generated from the three most discriminating features that had the best separation and were chosen for further processing.

The ability of this three-dimensional data space to classify properly chromatographic patterns of species/classes unknown to the computer, but known to the investigator, was examined using four discrete category classification methods contained in the ARTHUR chemometrics package. The four methods used were: the Bayes method, the KNN method, the LLM method and the SIMCA method.

RESULTS AND DISCUSSION

The optimum number of features was determined by plotting non-linear maps containing the two, three, four and six most significant features. The optimum was observed to be at around three or four features. The effect of including more features was that the noise level apparently increased without concomitant improvement in information. On the other hand, two features did not carry as much information as three or four. A non-linear map of the three-dimensional set, shown in Fig. 1, shows fairly tight clustering. It should be noted that the cluster formed by the *S. richteri* data set is more diffuse than that of the *S. invicta* set. This can be attributed to the

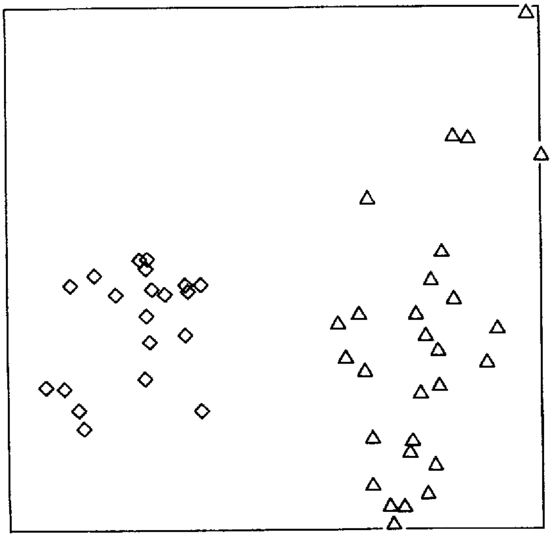

Fig. 1. Classification of the two imported fire ant species. Non-linear map of the three best features, by Fisher weight. △, *S. richteri*; ◇, *S. invicta*.

inclusion of males, queens, and workers, which clearly results in a more heterogeneous distribution.

The ability of this three-dimensional set to classify correctly an unknown sample was tested by dividing the original data set into five training set–test set combinations. The computer was given the proper classifications in the training sets. It was then given test patterns from samples known only to the observer and asked for a classification. All data from the original data set were used as test set patterns at least once. Table II reports on classification probabilities using the Bayes, KNN, LLM and SIMCA methods. No errors were encountered with Bayes, KNN and SIMCA methods, but a few misclassifications occurred with the LLM procedure. The overall results are quite acceptable.

Fig. 2 shows a typical chromatogram for each of the two species. The three most discrininating features are labelled. The chemical identities of these compounds have not yet been established, but are under study.

TABLE II

RESULTS OF CLASSIFICATION TEST METHODS

Number of runs: category 1 = 29; category 2 = 20. Category 1 = *S. richteri*; category 2 = *S. invicta.*

Method	*Category*	*Training set*		*Test set*		*Overall percent correct*
		No. of misses	*Percent correct*	*No. of misses*	*Percent correct*	
KNN	1	0	100	0	100	100
(K = 10)	2	0	100	0	100	100
LLM	1	0	100	3	89.6	97.9
	2	0	100	0	100	100
SIMCA	1	0	100	0	100	100
	2	0	100	0	100	100
Bayes	1	0	100	0	100	100
	2	0	100	0	100	100

The particular application presented here may not necessarily require a computer, since the difference between the cuticular hydrogen profiles of *S. invicta* and *S. richteri* are easily discernable by visual inspection. Rather, it is used to illustrate a principle and method which is generally applicable. There are, however, many applications examples where the computerized system clearly is necessary and superior to human perception. In cases where the different classes have very similar patterns, computerized pattern recognition procedures are required. The identification of discriminating features between different colonies of the same species, the examination of any seasonal variation with a colony, and the study of the profiles of hybrid species are examples of such situations.

We have proceeded further with these studies. Part II of these studies examines the cuticular hydrocarbon profiles of different colonies of the black imported fire ant, *Solenopsis richteri.*

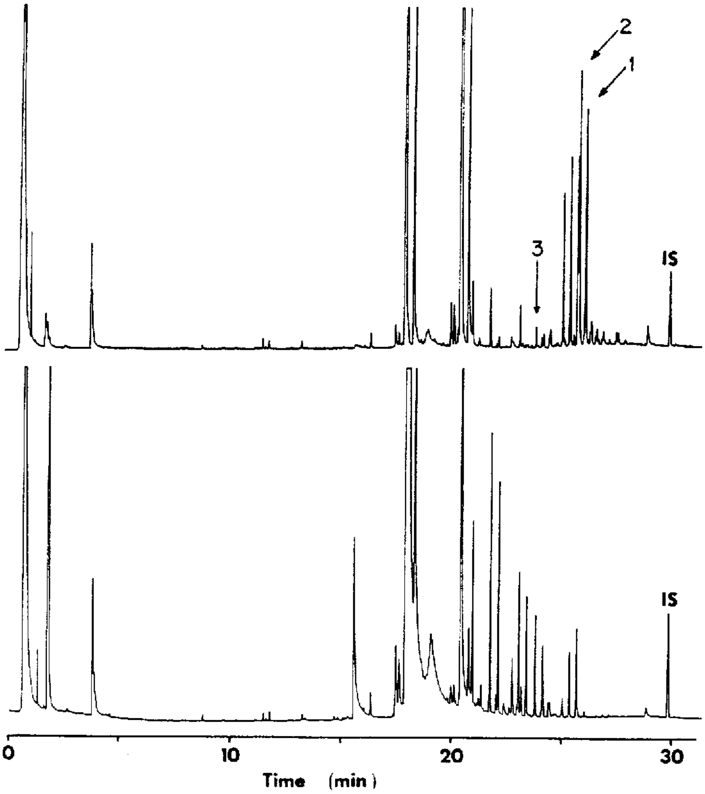

Fig. 2. Top, cuticular hydrocarbon profile of *S. invicta*; bottom, cuticular hydrocarbon profile of *S. richteri*; IS, internal standard. Peaks 1, 2 and 3 are the three most discriminating features, by Fisher weight, between the hydrocarbon profiles of the two species/classes. Note: peaks between 15 and 21 min represent venom alkaloids.

ACKNOWLEDGEMENTS

This work was supported by a BRS grant from the University of Alabama. We also wish to acknowledge the loan of a pyrolyzer from Chemical Data Systems.

REFERENCES

1 N. F. Hadley, G. L. Blomquist and U. N. Lanham, *Insect Biochem.*, 11 (1981) 173.

2 G. L. Blomquist and L. L. Jackson, *Prog. Lipid Res.*, 17 (1979) 319.

3 J. B. Lok, E. W. Cupp and G. L. Blomquist, *Insect Biochem.*, 5 (1975) 821.

4 D. R. Nelson, J. W. Dillwith and G. L. Blomquist, *Insect Biochem.*, 11 (1981) 187.

5 R. W. Howard and G. L. Blomquist, *Ann. Rev. Entomol.*, 27 (1982) 149.

6 D. R. Nelson, C. L. Fatland, R. W. Howard, C. A. McDaniel and G. L. Blomquist, *Insect Biochem.*, 10 (1980) 409.

7 R. K. Vander Meer and D. P. Wojcik, *Science (Washington, DC)*, 218 (1982) 806.

8 E. O. Wilson, *The Insect Societies*, Harvard University Press, Cambridge, MA, 1971, p. 272.

9 M. S. Blum, *Bull. Entomol. Soc. Amer.*, 20 (1974) 30.

10 B. R. Kowalski, *ACS Symp. Ser.*, 52 (1977) 14.

11 E. Shannon and W. Weaver, *The Mathematical Theory of Information*, University of Illinois Press, Urbana, IL, 1947.

12 B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 94 (1972) 5632.

13 D. L. Duewar, J. R. Koskinen and B. R. Kowalski, *Documentation for ARTHUR, Version 1-8-75, Chemometrics Society Report No. 2,* Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, WA, 1975.

14 P. C. Jurs and T. L. Isenhour, *Chemical Applications of Pattern Recognition,* Wiley, New York, 1975.

15 B. R. Kowalski and C. F. Bender, *J. Amer. Chem. Soc.*, 95 (1973) 686.

16 N. B. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965.

17 J. H. Brill and W. Bertsch, *Insect Biochem.*, 15 (1985) 49.

18 H. Engman, H. T. Mayfield, W. Bertsch and T. Mar, *J. Anal. Appl. Pyrol.*, 6 (1984) 137.

19 H. T. Mayfield and W. Bertsch, *Computer Applications in the Laboratory*, 2 (1983) 130.

20 D. L. Duewar, J. R. Koskinen and B. R. Kowalski, *ARTHUR*, obtained from B. R. Kowalski, Laboratory for Chemometrics, Department of Chemistry, University of Washington, Seattle, WA, 1975.