ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 66 (2003) 51–62

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemometrics

Genetic algorithms for deciphering the complex chemosensory code of social insects

Barry K. Lavine[a,*], Charles Davidson[a], Robert K. Vander Meer[b], Sigal Lahav[c], Victoria Soroker[c], Abraham Hefetz[c]

[a] *Department of Chemistry, Clarkson University, Box 5810, Potsdam, NY 13699-5810, USA*

[b] *Center for Medical Agricultural and Veterinary Entomology, United States Department of Agriculture, Agricultural Research Service, 1600 SW 23rd Drive, Gainesville, FL 32608, USA*

[c] *Department of Zoology, George S. Wise Faculty of Life Sciences, Tel-Aviv University, Ramat Aviv 69978, Israel*

Received 25 June 2002; received in revised form 19 December 2002; accepted 3 January 2003

Abstract

Chemical communication among social insects is often studied with chromatographic methods. The data generated in such studies may be complex and require pattern recognition techniques for interpretation. Presently, we are analyzing gas chromatographic (GC) profiles of hydrocarbon extracts obtained from the cuticle and postpharyngeal gland (PPG) of 400 *Cataglyphis niger* ants using a genetic algorithm (GA) for pattern recognition analysis to identify the factors influencing colony odor. The pattern recognition GA identifies features (i.e., chromatographic peaks) whose principal component plots show clustering of the samples on the basis of class. Because the largest principal components capture the bulk of the variance in the data, the peaks chosen by the GA primarily convey information about differences between the classes in a data set. As it trains, the pattern recognition GA focuses on those classes and/or samples that are difficult to classify by boosting their class and sample weights. Samples or classes that consistently classify correctly are not as heavily weighted as samples or classes that are difficult to classify. Over time, the algorithm learns its parameters in a manner similar to a neural network. The proposed algorithm integrates aspects of artificial intelligence and evolutionary computations to yield a “smart” one-pass procedure for feature selection and pattern recognition. Utilizing the pattern recognition GA, two specific questions were addressed in this study: (1) Does the overall hydrocarbon profile of the colony change with time? (2) Does the queen influence the hydrocarbon pattern of the colony?
© 2003 Elsevier Science B.V. All rights reserved.

Keywords: Nestmate recognition; Chemical communication; Genetic algorithms; Pattern recognition analysis; Feature selection; Principal component analysis; Boosting

1. Introduction

Nestmate recognition is defined as the ability of a worker ant to discriminate fellow nestmates from alien conspecific workers. Studies pertaining to nestmate recognition have focused on the nature of the recognition cues used and the mode by which colony odor is obtained. Correlative [1] and direct evidence exists to support a role for cuticular hydrocarbons as recognition cues [2,3]. However, the discovery that cuticular hydrocarbon composition is similar to that of the postpharyngeal gland (PPG) has focused research on

* Corresponding author. Tel.: +1-315-268-2394; fax: +1-315-268-6610.

E-mail address: bklab@clarkson.edu (B.K. Lavine).

0169-7439/03/$ - see front matter © 2003 Elsevier Science B.V. All rights reserved.
doi:10.1016/S0169-7439(03)00006-6

the role of this gland in nestmate recognition [4]. It has been shown that PPG extracts can modify aggressive behavior in two phylogenetically remote species of ants [4,5]. Furthermore, the PPG can act as a "gestalt organ" which enables the admixing and rapid integration of odors from various sources [6,7]. These studies as well as others provide empirical evidence supporting a "gestalt model" for the creation and acquisition of a general colony odor [8].

Earlier studies have focused on the role of both the cuticular and PPG hydrocarbons in nestmate recognition for *Cataglyphis niger*, a highly evolved social insect. In this study, the influence of time and the impact of the queen on the hydrocarbon profiles of *C. niger* worker ants have been investigated. Capillary column gas chromatography was used to analyze PPG and cuticular hydrocarbon soaks obtained from 400 *C. niger* worker ants. A genetic algorithm (GA) for pattern recognition analysis was used to analyze the gas chromatographic (GC) data. The pattern recognition GA was used as a data microscope to sort, probe, and to look for hidden relationships in the GC data. The focus of this paper is on the analytical methodology used to solve this rather interesting classification problem, with particular emphasis on the pattern recognition techniques used to identify the various fingerprint patterns within the GC data.

2. Experimental

Two polygynous colonies of *C. niger* ants were collected in Tel Aviv, Israel. Colony A had 11 queens and colony B had 18 queens. In the laboratory, these two field colonies were transferred to artificial nests and reared under controlled conditions (diet, temperature, lighting).

Three monogynous (one queen) fragments and one queenless group (daughter colonies) were created from each of two polygynous laboratory colonies (mother colonies). Each mother colony consisted of 3000 ants and each daughter colony consisted of 250

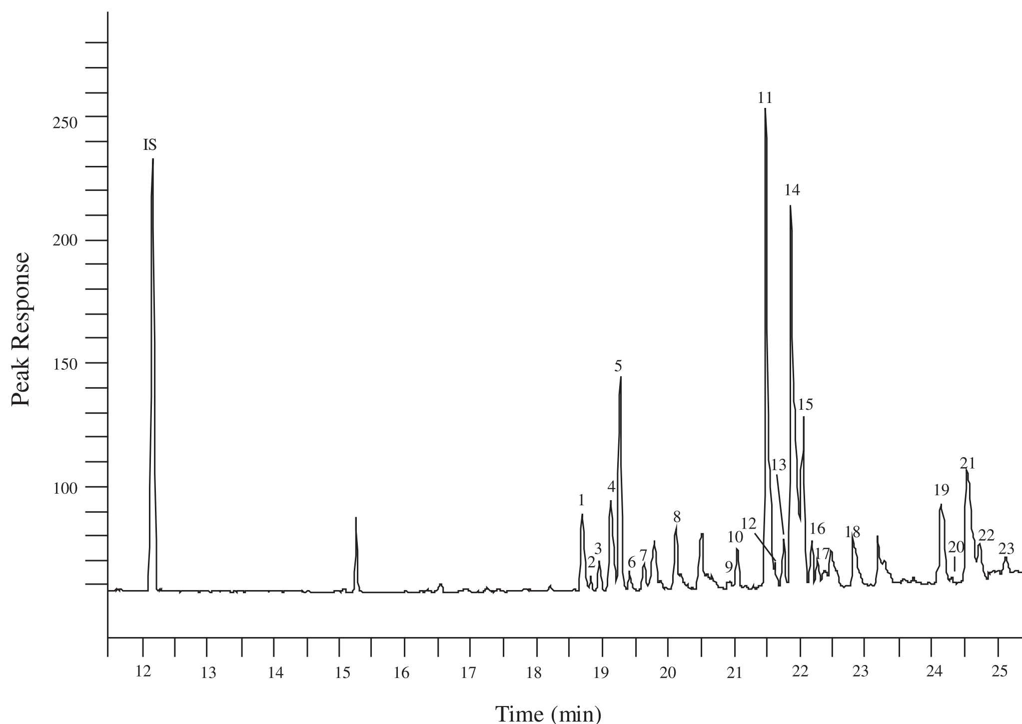

Fig. 1. Gas chromatogram of PPG secretion from a *C. niger* worker. Numbered peaks correspond to mono- and dimethyl branched hydrocarbons. See Ref. [9] for representative structures.

ants. The mother and daughter colonies were separated from each other for 3 months. Before separation, a random sample of 10 ants was taken from each mother colony. After separation, 10 ants were collected from each mother and one daughter colony once a month. All other daughter nests were sampled after 3 months of separation.

Hydrocarbons were extracted from individual ants by immersing the dissected PPG or the isolated thorax, which is representative of the cuticle, in 100 μl of pentane. The thorax was extracted for a time period of only 5 min to avoid contamination from internal compounds. Eicosane (750 ng) was added to the wash as an internal standard. Samples were analyzed by a Varian 3700 Gas Chromatograph (Varian Analytical Systems, Walnut Creek, CA) equipped with a split/splitless injector, a flame ionization detector, a Leap Technologies autosampler, and a fused silica DB-1 capillary column (J&W Scientific, Folsom, CA) that is 30 m in length, 0.32 mm in internal diameter, and with a film thickness of 0.25 mm. Hydrogen was used as the carrier gas and nitrogen was the make-up gas. The column was temperature programmed from 120 to 285 °C at 5 °C/min. Fig. 1 shows a representative gas chromatographic (GC) profile of a PPG secretion of a *C. niger* worker. Twenty-three of the 72 peaks in the gas chromatogram could be accurately and reliably quantified. These peaks were previously identified by gas chromatography/mass spectrometry [9]. Some peaks were composed of two or more inseparable compounds.

3. Genetic algorithm for pattern recognition analysis

For pattern recognition analysis, each gas chromatogram was initially represented as a 23-dimensional data vector, $x=(x_1, x_2, x_3 \ldots x_j \ldots, x_{23})$, where x_j is the area of the *j*th peak. The gas chromatograms were peak matched using the integration reports and Kovat's indices. Each data vector was normalized to constant sum using the total integrated peak area. A genetic algorithm (GA) for pattern recognition analysis [10–14] was used to analyze the GC data. The pattern recognition GA identifies feature sets (i.e., chromatographic peaks) whose principal component plots show clustering of the samples on the basis of class. Because the largest principal components capture the bulk of the variance in the data, the peaks chosen by the GA primarily convey information about differences between the classes in the data. As the pattern recognition GA trains, the algorithm focuses on those classes and/or samples that are difficult to classify by boosting their class and sample weights. Samples or classes that consistently classify correctly are not as heavily weighted as samples or classes that are difficult to classify. Over time, the algorithm learns its parameters in a manner similar to a neural network. The proposed algorithm integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection and pattern recognition.

The GA builds a population of binary strings called chromosomes, each of which represents a feature subset. For a GC peak to be included in a feature subset, the corresponding bit in the string must be set at 1. The length of each binary string is equal to the number of GC peaks in the data set. The number of chromosomes in the initial population, ϕ, is usually set at 100. The chromosomes comprising the initial population are generated at random to minimize potential bias.

During each generation, the strings are decoded, yielding the actual set of features that are sent to a fitness function for evaluation. Each string is assigned a score by the fitness function, which is a measure of the quality of the proposed solution for the pattern recognition problem. Chromosomes with a higher score have a higher probability of being selected for crossover. The power of the GA arises from crossover, which causes a structured yet randomized exchange of information between solutions, with the expectation that good solutions can generate even better ones. Additional variability in the population is achieved through the mutation operator, which flips the state of single bits based on a certain probability supplied by the user. Mutation allows the GA to explore other regions of the solution space. If the GA finds a better feature subset through mutation, the optimization will continue in a new direction.

The selection operator used by the pattern recognition GA orders the population of strings, i.e., the feature subsets, from best to worst fitness while simultaneously generating a copy of the same population and randomizing the order of the strings in this

copy with respect to fitness. A fraction of the population is then selected as per the selection pressure, which is usually set at 0.5. The top half of the ordered population is mated with the top half of the random population, guaranteeing that the best 50% are selected for reproduction, while ensuring that every string in the randomized copy has a uniform chance of being selected due to the randomized selection criterion imposed on the strings in this population.

To facilitate the tracking and scoring of the feature subsets, i.e., the principal component plots, class and sample weights, which are integral parts of the fitness function, are computed (see Eqs. (1) and (2), where CW(c) is the weight of class c and c varies from 1 to the total number of classes in the data set, and SW(s) is the weight of sample s in class c). The class weights sum to 100, whereas the sample weights in a class sum to a value equal to the corresponding class weight.

$$\mathrm{CW(c)} = 100 \frac{\mathrm{CW(c)}}{\sum_{\mathrm{c}} \mathrm{CW(c)}} \tag{1}$$

$$\mathrm{SW(s)} = \mathrm{CW(c)} \frac{\mathrm{SW(s)}}{\sum_{\mathrm{s} \in \mathrm{c}} \mathrm{SW(s)}} \tag{2}$$

$$F(d) = \sum_{\mathrm{c}} \sum_{\mathrm{s} \in \mathrm{c}} \frac{1}{K_{\mathrm{c}}} \times \mathrm{SHC(s)} \times \mathrm{SW(s)} \tag{3}$$

Each principal component plot generated for each feature subset is scored using the K-nearest neighbor (K-NN) classification algorithm [15]. For a given data point, Euclidean distances are computed between it and every other point in the principal component plot. These distances are arranged from smallest to largest. A poll is then taken of the point's K-nearest neighbors. For the most rigorous classification, K equals the number of samples in the class to which the point belongs. (K, which is assigned by the user, usually varies with the class.) The number of nearest neighbors with the same class label as the sample point in question, the so-called sample hit count (SHC), is computed ($0 \leq \mathrm{SHC(s)} \leq K_{\mathrm{c}}$), where K_{c} is the number of nearest neighbors to be calculated for each sample in class c. It then becomes a simple matter to score a principal component plot (see Eq. (3), where $F(d)$ is the fitness of the feature set being scored, SHC(s) is the number of nearest samples with the same class label as sample s, and SW(s) is the weight of sample s).

To better understand the scoring of the principal component plots, consider the hypothetical example of a data set with two classes having been initially assigned equal weights by the pattern recognition GA. Class 1 has 10 samples, and class 2 has 20 samples. Therefore, K_1 is 10 and K_2 is 20. At generation 0 (no children have yet been created), the samples in a given class will have the same weight. Therefore, each sample in class 1 has a sample weight of 5, whereas each sample in class 2 has a weight of 2.5. If sample 3, which is in class 1, has as its nearest neighbors 7 class 1 samples, then $\mathrm{SHC}/K_1 = 0.7$ and $(\mathrm{SHC}/K_1) \times \mathrm{SW}(3) = 0.7 \times 5$ or 3.5. By summing $(\mathrm{SHC}/K_{\mathrm{c}}) \times \mathrm{SW(s)}$ for each sample point in the plot, the principal component map of the feature subset is scored.

The fitness function of the GA is able to focus on samples and classes that are difficult to classify by boosting their weights over successive generations. In order to boost, it is necessary to first compute the sample hit rate (SHR), which is the mean value of $\mathrm{SHC}/K_{\mathrm{c}}$ over all feature subsets (ϕ) produced in a particular generation (see Eq. (4)) (ϕ is usually set at 100). SHR, which is calculated over the entire population of solutions in a particular generation, provides consistent information about the difficulty in classifying a particular sample.

$$\mathrm{SHR(s)} = \frac{1}{\phi} \sum_{i=1}^{\phi} \frac{\mathrm{SHC}_i(\mathrm{s})}{K} \tag{4}$$

Boosting is then performed in two steps. First, the class hit rate (i.e., average sample hit rate for all samples in a class) is computed (see Eq. (5), where $\mathrm{CHR_g(c)}$ is the class hit rate for class c during generation g, AVG is the average, and $\mathrm{SHR_g(s)}$ is the sample hit rate for each sample in class c during generation g). CHR, like SHR, provides consistent information about the difficulty in classifying a particular sample type. Classes and samples with low hit rates will be weighted more heavily, i.e., they will

have more influence in the fitness calculation, than classes or samples that score well.

$$CHR_g(c) = AVG(SHR_g(s) : \forall_{s \in c}) \quad (5)$$

Second, class and sample weights are adjusted during each generation using a perceptron (see Eqs. (6) and (7), where $CW_{g+1}(c)$ is the class weight for c during the current generation g+1, $CW_g(c)$ is the class weight for c during the previous generation g, P is the momentum, $CHR_g(c)$ is the class hit rate for c during generation g, $SW(s)_{g+1}$ is the sample weight for s during generation g+1, $SW(s)_g$ is the sample weight for s during generation g, and $SHR_g(s)$ is the sample hit rate for s during generation g). The user must set the momentum, P.

$$CW_{g+1}(c) = CW_g(c) + P(1 - CHR_g(c)) \quad (6)$$

$$SW_{g+1}(s) = SW_g(s) + P(1 - SHR_g(s)) \quad (7)$$

During each generation, class and sample weights are updated using the class and sample hit rates from the previous generation. After a certain number of generations, the class weights become fixed. Eq. (6) is turned off, P is halved, and sample weights are renormalized using Eq. (2). The GA then focuses on the troublesome samples (see Eq. (7)).

Boosting is crucial for the successful operation of the GA because it allows the values of the class and sample weights to change, thereby modifying the criteria for a good score, which can help to minimize the problem of convergence to a local optimum. Hence, the fitness landscape of the pattern recognition GA changes as the population evolves towards a solution.

4. Results and discussion

Two specific questions (i.e., hypotheses) were addressed in this study: (1) Does the overall hydrocarbon pattern of the colony change with time? (2) Is the queen influencing the hydrocarbon pattern of the colony? The hydrocarbon profiles, PPG and cuticle of both colonies were affected to the same degree by time and queen influence. Therefore, to more clearly visualize the data, only the results for the postpharyngeal gland hydrocarbons of colony B and its subunits will be discussed.

Initially, SIMCA pattern recognition was used to analyze the data. However, SIMCA was unable to identify trends in the data that could validate either hypothesis. The low classification success rates achieved by SIMCA for these data are attributed to the large amount of irrelevant information present in the GC data. Therefore, a new approach to analyze the data was taken in this study. The new approach requires a scientist to combine empirical data with careful analysis, prior knowledge and reasoning, and is more rigorous in some respects than simply formulating a hypothesis, i.e., model of the data, from a set of observations since a variety of techniques can be used to validate the so-called model, with predictive success being the most powerful. This new approach attempts to explore the implications of the data so that hypotheses are developed with a greater awareness of reality. It does not involve a thought ritual; rather it is a method that searches for significant structure in multivariate data. Mathematics is not used for modeling per se but more for discovery.

The first step was to apply principal component analysis to the data. Principal component analysis can reduce the dimensionality of the data while simultaneously preserving the information present in the data. Dimensionality reduction is possible using principal component analysis because in chromatographic data sets the measurement variables (areas of the selected GC peaks) are highly correlated. High collinearity between measurement variables is a strong indication that a new set of basis vectors can be found that are better at conveying the information present in the data than axes defined by the original measurement variables. This new basis set, which is linked to variation in the data, can be used to develop a new coordinate system for displaying the data. The first principal component is formed by determining the direction of largest variation in the data and modeling it by a line, which passes through the center of the data. The second principal component defines the direction of next largest variation; it passes through the center of the data and is orthogonal to the largest principal component. The third principal component lies in the direction of next largest variation; it passes through the center of the data and is orthogonal to the first and

second largest principal components, and so forth. Each principal component is a linear combination of the original measurement variables. Using principal component analysis, the original measurement variables, which constitute a correlated axis system, can be converted into an orthogonal system that removes correlations by forcing the new axes to be independent. This requirement dramatically reduces the dimensionality of the data because only a few independent axes are necessary to describe the data. Principal component analysis is routinely applied to high-dimensional data to affect dimensionality reduction and to search for structure in multivariate data.

Outliers can adversely influence the performance of principal component analysis and other pattern recognition techniques. Therefore, outlier analysis was performed on each class in the data set prior to principal component analysis using the generalized distance test [16], which was implemented by SCOUT [17]. Two gas chromatograms identified as outliers were removed from the data set. These two outliers were from the same daughter colony.

In Fig. 2, a plot of the two largest principal components of the 23 GC peaks obtained from the gas chromatograms for the 40 ants that belonged to a monogynous fragment (i.e., daughter nest) of colony B is shown. Each gas chromatogram or ant sample is represented as a point in the principal component plot (1 = 0-month ants, 2 = 1-month ants, 3 = 2-month ants, and 4 = 3-month ants). The ants at 0 and 1 months (1's and 2's) are well separated from each other and the other ants in the plot. However, there is overlap between the ants at 2 and 3 months (3's and 4's), which raises questions about the validity of the hypothesis that the overall hydrocarbon pattern of a colony changes with time.

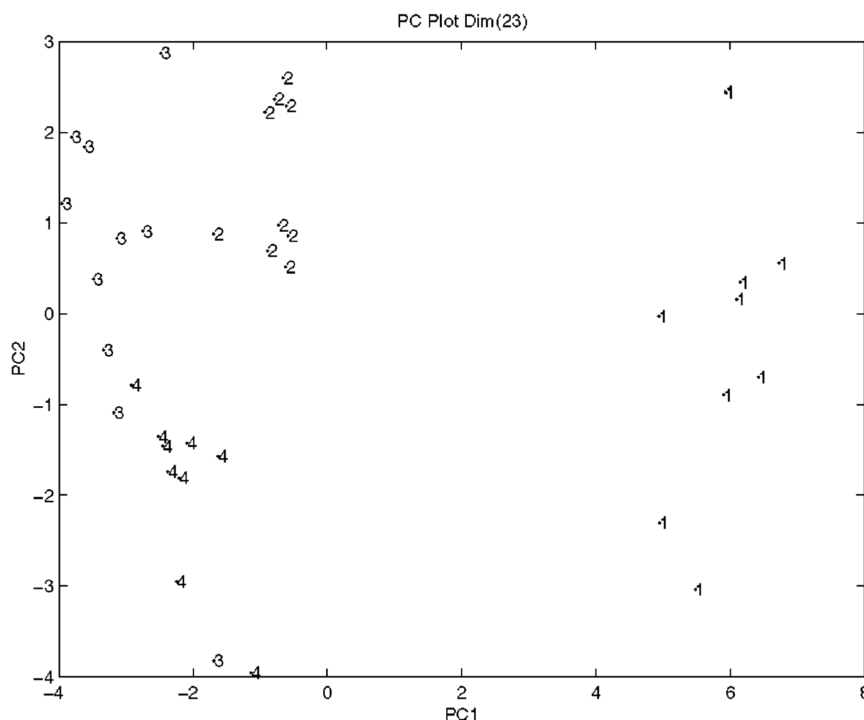

Fig. 2. A plot of the two largest principal components of the 23 GC peaks obtained from the gas chromatograms of the 40 ants that belonged to a daughter nest (i.e., monogynous fragment) of colony B is shown. Each ant sample is represented as a point in the principal component map (1 = 0-month ants, 2 = 1-month ants, 3 = 2-month ants, and 4 = 3-month ants). The ants at 0 and 1 months (1's and 2's) are well separated from each other and the other ants in the plot, whereas there is overlap between the ants at 2 and 3 months (3's and 4's), raising questions about the validity of the hypothesis that the overall hydrocarbon pattern of a colony changes with time.

The pattern recognition GA was used to identify peaks characteristic of the profile at each time period, thereby removing irrelevant information through judicious feature selection. The GA identified features by sampling key feature subsets, scoring their principal component plots, and tracking those samples or classes that were most difficult to classify. The boosting routine used this information to steer the population to an optimal solution. After 100 generations, the pattern recognition GA identified five standardized retention time windows, whose principal component map (see Fig. 3) shows clustering of the ants according to their sampling time (*K* set equal to the number of samples in each class). The results of this study indicate that PPG hydrocarbon profiles of *C. niger* ants change with time, as has been reported for some other ant species, e.g., *Solenopsis invicta* [18], *Leptothorax lichtensteni* [19], *Cataglyphis iberica* [20], *Formica truncorum* [21], and most recently *C. niger* [22].

C. niger workers exchange hydrocarbon material with other *C. niger* workers and the queen through social contact and regurgitation of hydrocarbons from the PPG. Therefore, ants in social contact are expected to have a similar hydrocarbon profile, whereas ants with limited social contact are not expected to influence each other as much. Because the queen is central to the normal functioning of the colony, we assessed whether or not she tracked the changes in worker profiles with time or deviated from the worker profiles. The worker hydrocarbon profiles of one of the queenright daughter colonies were monitored at 0 (same as polygyne mother colony before group separation), 1, 2, and 3 months. Random samples of 10 workers were sacrificed at each of these specific time intervals, as well as the queen at 3 months. Fig. 4 shows a plot of the two largest principal components defined by the 23 gas chromatographic peaks and the 40 *C. niger* workers. Each ant sample is represented as a point in the principal component map (1 = 0-

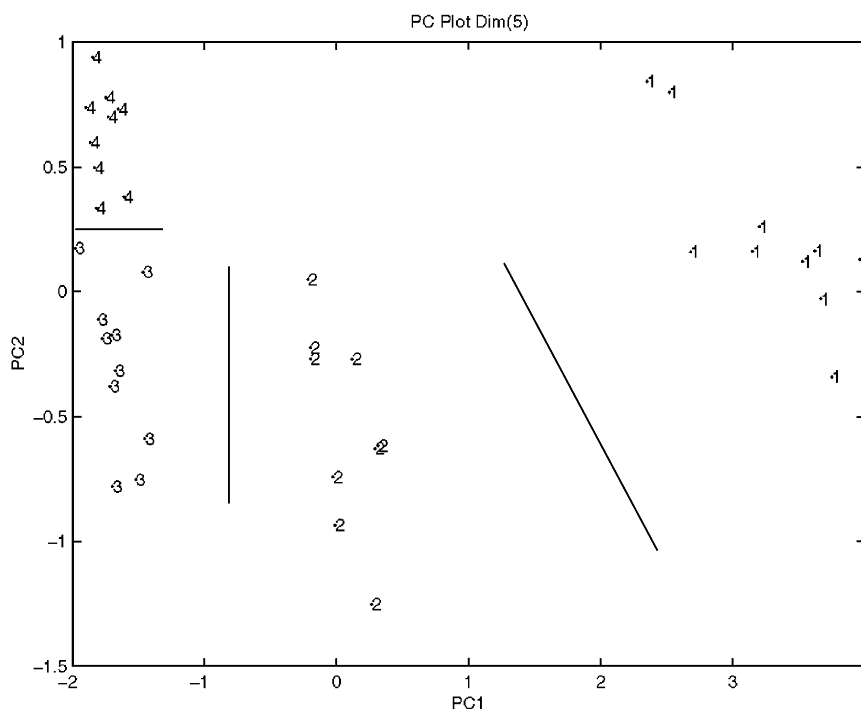

Fig. 3. A plot of the two largest principal components of the 5 GC peaks (9, 11, 13, 17, and 19 in Fig. 1) identified by the pattern recognition GA is shown for the 40 ants that belonged to a daughter nest (i.e., monogynous fragment) of colony B. Each ant sample is represented as a point in the principal component map (1 = 0-month ants, 2 = 1-month ants, 3 = 2-month ants, and 4 = 3-month ants). Clustering of the ants according to their sampling time is evident, indicating that PPG hydrocarbon profiles of *C. niger* ants change with time.

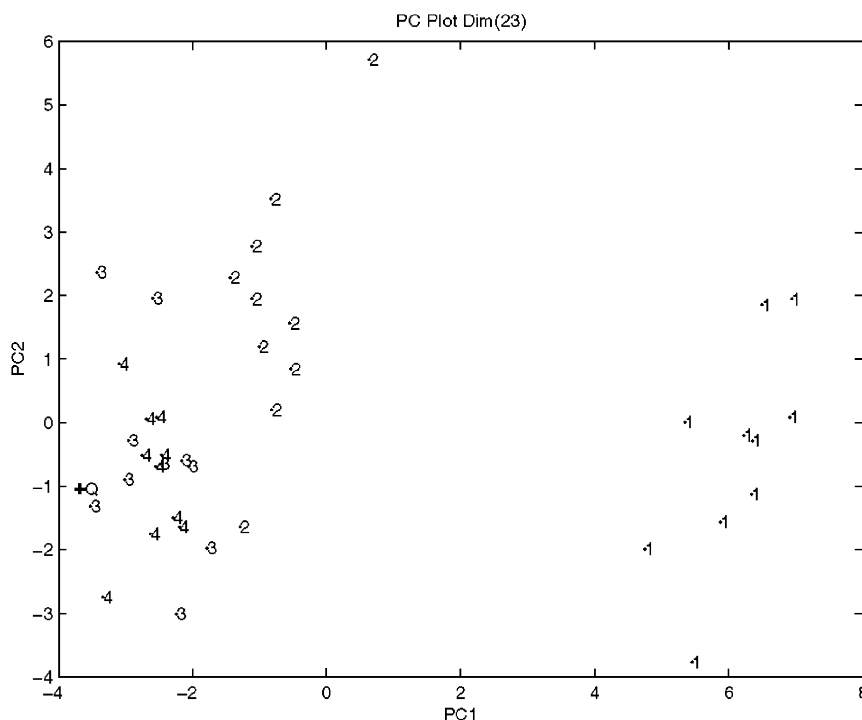

Fig. 4. A plot of the two largest principal components of the 23 GC peaks obtained from the gas chromatograms of 40 *C. niger* ants from a queenright daughter colony monitored over 0, 1, 2, and 3 months. Each ant sample is represented as a point in the principal component map (1 = 0-month ants, 2 = 1-month ants, 3 = 2-month ants, and 4 = 3-month ants). Q is the gas chromatogram of the queen projected onto the principal component map defined by the 23 gas chromatographic peaks and the 40 ant samples. The queen was sampled at the end of the third month; however, when the queen is projected onto the map it is unclear as to which time category she should be assigned.

month ants, 2 = 1-month ants, 3 = 2-month ants, and 4 = 3-month ants). The results for 0-month workers were clearly separated from the other three time periods; however, the 2- and 3-month ant samples again were poorly separated. The queen was sampled at the end of the third month; however, when the queen was projected onto the map it was unclear as to which time category she fits the best (2 month or 3 month; see Fig. 4).

The pattern recognition GA was used to uncover features characteristic of the gas chromatographic profiles of each group. For this study, the GA was configured in the following manner. The number of "chromosomes" or binary strings in the population was set at 50, the momentum, P, was 0.5, the mutation rate was 0.1, and K_c was set at 10 for each class. The maximum number of iterations was set at 100. After 100 generations, the GA identified three standardized retention time windows whose principal component plot showed clustering of the ant samples according to time period (see Fig. 5). When the gas chromatogram of the queen was projected onto the principal component map defined by the three retention time windows and the 40 ant samples, it was evident that the queen profile is most similar to that of the 3-month group. This result was expected because the queen and the workers after 3 months should have similar profiles due to the exchange of material between the workers and the queen. The efficacy of the genetic algorithm to extract pertinent information from GC data is also demonstrated, and it shows that PPG hydrocarbons can be used as a model system to study the passage of materials such as semio-chemicals through an ant colony. As for the changes that occurred in the hydrocarbon profiles of the queen and her workers, we could not determine if either the queen or the workers had a greater influence in directing these changes. All that can be concluded from the data in this experiment

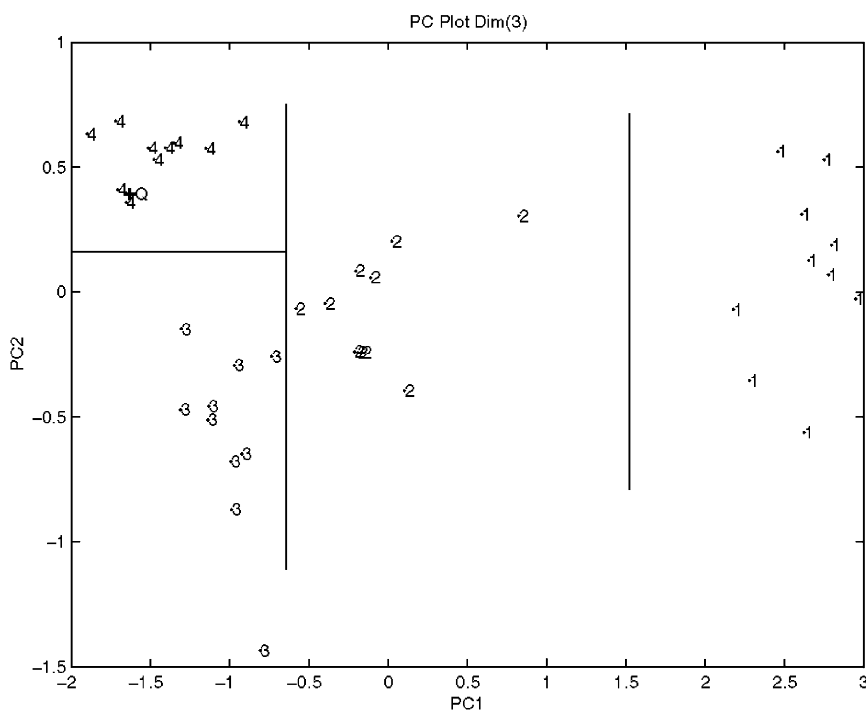

Fig. 5. A plot of the two largest principal components of the 3 GC peaks (8, 9, and 13 in Fig. 1) identified by the pattern recognition GA is shown for the 40 *C. niger* ants from a queenright daughter colony. Each ant sample is represented as a point in the principal component map (1 = 0-month ants, 2 = 1-month ants, 3 = 2-month ants, and 4 = 3-month ants). Q is the gas chromatogram of the queen projected onto the principal component map defined by the 3 GC peaks and the 40 ant samples. It is evident from the map that the queen profile is most similar to that of the 3-month-old group.

is that changes in the hydrocarbon profiles of the queen and her workers tracked each other.

To compare the effect of time and queen influence on the PPG profiles, a subset of 30 ants from colony B was analyzed. Ten ants from the original polygynous colony were sampled at 0 months; 10 ants from a queenless daughter colony were sampled at 1 month; and 10 ants from the original polygynous colony were sampled at 1 month. Fig. 6 shows a plot of the two largest principal components for the 30 *C. niger* ants. Again, each ant is represented as a point in the principal component plot (1 = original polygynous colony sampled at 0 months, 2 = queenless daughter colony sampled at 1 month, 3 = original polygynous colony sampled at 1 month). The ants from the original colony at 0 months are well separated from the other ants in the principal component plot. However, the ants that are sampled at the same time do overlap, which makes it difficult to draw any meaningful conclusions about the effect of time and queen influence on the PPG profiles.

The pattern recognition GA was used to identify a set of features characteristic of the source profile of each group in the data. Fig. 7 shows a plot of the two largest principal components for the nine standardized retention time windows identified by the pattern recognition GA. Clustering of the ants by class is evident. Classes 1 and 3 share the same queen, which suggests that differences in their hydrocarbon profiles can be attributed to a change over time. For classes 2 and 3, the change in the hydrocarbon profiles is due to the presence or absence of a queen, whereas differences in the hydrocarbon profiles of classes 1 and 2 reflect a combination of these two factors (time and queen). We also observed a similar trend for the same queenless daughter colony sampled at 2 and 3 months, respectively. Evidently, changes in PPG profiles of *C. niger* worker ants over time exceed

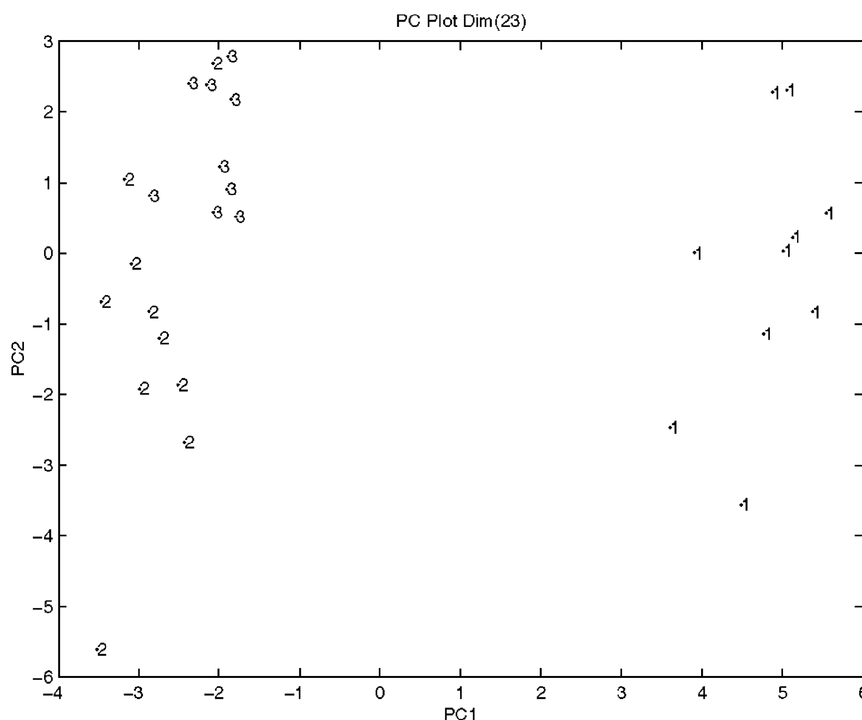

Fig. 6. A plot of the two largest principal components of the 23 GC peaks obtained from the gas chromatograms of 30 *C. niger* ants. Each ant sample is represented as a point in the principal component map (1 = original polygynous colony sampled at 0 month, 2 = queenless daughter colony sampled at 1 month, 3 = original polygynous colony sampled at 1 month). The ants from the original colony at 0 month are well separated from the other ants in the principal component plot, whereas the ants that are sampled at the same time do overlap, which makes it difficult to draw any meaningful conclusions about the effect of time and queen influence on the PPG profiles.

the queen's influence on these changes. (Queen influence is correlated to the second principal component in Fig. 7, whereas time influence is correlated to the first principal component, which accounts for a larger fraction of the total cumulative variance of the data than the second principal component.) The queen does not appear to play a dominant role in influencing the hydrocarbon profiles. Time appears to be the major factor in influencing the hydrocarbon profiles although the separation between the 1-month-old queenless nest and the 1-month-old queenright nest would indicate that the queen could have some influence on the hydrocarbon profiles. Through normal social interactions, the queen and the workers are expected to exchange hydrocarbons with each other; however, measurement of hydrocarbon cue exchanges between queens and workers revealed that a queen receives more hydrocarbons than she gives away [22]. This, coupled with shear quantity (there are many more workers than queens), appears to allow workers to overpower the possible effects of the queen.

Temporal chemical analyses indicate that hydrocarbon profiles of colonies of *C. niger* ants change over time. This was true irrespective of the number of queens in the original colony or daughter colonies sampled. However, divergence in hydrocarbon profiles did not affect nestmate recognition. Workers separated for 3 months still considered each other as nestmates despite measurable differences in their hydrocarbon profiles [23]. This suggests that observed differences in PPG profiles are still within the cue variation that normally occurs in *C. niger* colonies.

Rearing conditions used in this study were constant. Therefore, we can exclude the possibility that changes in the hydrocarbon profiles over time are attributable to either change in the physical environment and/or population changes. In all likelihood, the

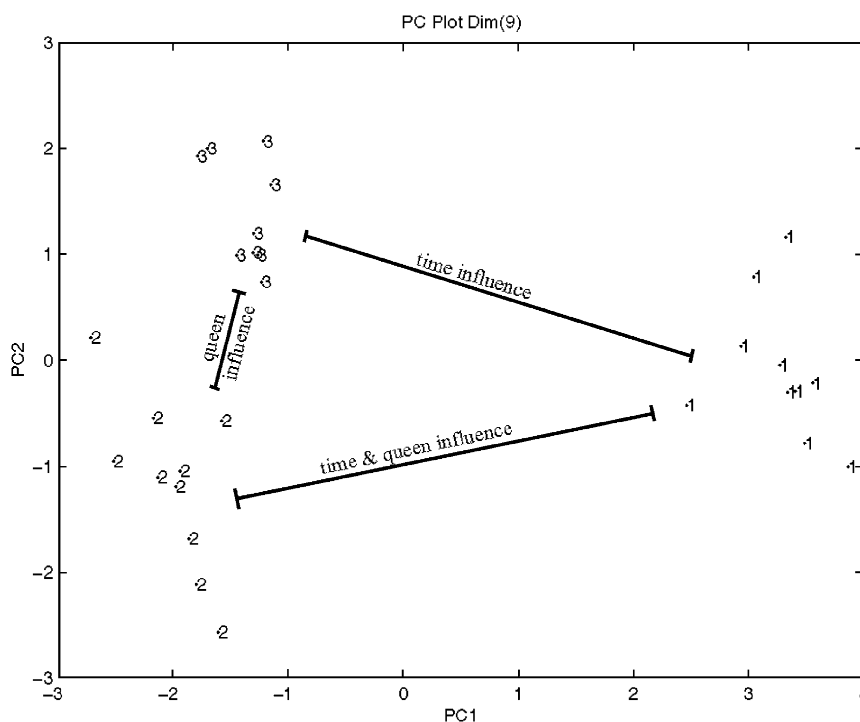

Fig. 7. A plot of the two largest principal components for the nine standardized retention time windows (2, 3, 4, 8, 9, 12, 13, 19, and 22 in Fig. 1) identified by the pattern recognition GA. Each ant is represented as a point in the principal component map (1 = original polygynous colony sampled at 0 month, 2 = queenless daughter colony sampled at 1 month, 3 = original polygynous colony sampled at 1 month). Classes 1 and 3 share the same queen, which suggests that differences in their hydrocarbon profiles can be attributed to a change with time. For classes 2 and 3, the change in the hydrocarbon profiles is due to the presence or absence of a queen, whereas differences in the hydrocarbon profiles of classes 1 and 2 reflect a combination of these two factors (time and queen).

changes in hydrocarbon expression are genetically based. Because genetic heterogeneity in polygynous *C. niger* colonies is large, it is predicted that individual differences in hydrocarbon composition will create divergent odor patterns in the individuals within a group. Nevertheless, within-group homogeneity for all groups was always observed. Furthermore, the differences in the hydrocarbon profiles were not qualitative, i.e., they did not involve disappearance or appearance of specific compounds, but were quantitative changes expressed as shifts in the relative concentration of existing compounds.

Finally, we have demonstrated the application of a genetic algorithm to the solution of problems involving complex multivariate chromatographic data. Principal component analysis alone generally gave poor resolution of the classes; however, principal component analysis when coupled to feature selection via the application of the pattern recognition GA proved crucial to uncovering structure in the data. We anticipate that this method will find application in basic research, as well as in quality control and forensic chemistry.

References

[1] R.K. Vander Meer, L. Morel, Nestmate recognition in ants, in: R.K. Vander Meer, M. Breed, M. Winston, K.E. Espelie (Eds.), Pheromone Communication in Social Insects, Westview Press, Boulder, CO, 1998, pp. 79–103.

[2] S. Lahav, V. Soroker, A. Hefetz, R. Vander Meer, Naturwissenschaften 86 (1999) 246–249.

[3] M.L. Thomas, L.P. Parry, R.A. Allan, M.E. Elgar, Naturwissenschaften 86 (1999) 87–92.

[4] A.G. Bagneres, E. Morgan, Experientia 47 (1991) 106–111.

[5] A. Hefetz, C. Errard, A. Chambris, A. Le Negrate, J. Insect Behav. 9 (1996) 709–717.

[6] V. Soroker, C. Vienne, A. Hefetz, J. Chem. Ecol. 21 (1995) 365–378.
[7] C. Vienne, V. Soroker, A. Hefetz, Insectes Soc. 42 (1995) 267–277.
[8] R.H. Crozier, M.W. Dix, Behav. Ecol. Sociobiol. 4 (1979) 217–224.
[9] V. Soroker, A. Hefetz, M. Cojocaru, J. Billen, S. Francke, W. Francke, Physiol. Entomol. 20 (1995) 323–329.
[10] B.K. Lavine, A.J. Moores, H.T. Mayfield, A. Faruque, Microchem. J. 61 (1999) 69–78.
[11] B.K. Lavine, A.J. Moores, Genetic algorithms for pattern recognition analysis and fusion of sensor data, in: K. Siddiqui, D. Eastwood (Eds.), Pattern Recognition, Chemometrics, and Imaging for Optical Environmental Monitoring, Proceedings of SPIES, Society for Optical Engineering, Bellingham, Washington, 1999, pp. 103–112.
[12] B.K. Lavine, J. Ritter, A.J. Moores, M. Wilson, A. Faruque, H.T. Mayfield, Anal. Chem. 72 (2) (2000) 423–431.
[13] B.K. Lavine, A. Vesanen, D.M. Brzozowski, Anal. Lett. 34 (2) (2001) 281–294.
[14] B.K. Lavine, D. Brzozowski, J. Ritter, A.J. Moores, C.E. Davidson, H.T. Mayfield, Anal. Chim. Acta 437 (2001) 233–246.
[15] M. James, Classification, Wiley, New York, 1992.
[16] S.J. Schwager, B.H. Margolin, Annu. Stat. 10 (1982) 943–953.
[17] M.A. Stapanian, F.C. Garner, K.E. Fitsgerald, G.T. Flatman, J.M. Nocerino, J. Chemom. 7 (1993) 165–176.
[18] R. Vander Meer, D. Saliwanchik, B.K. Lavine, J. Chem. Ecol. 15 (1989) 2115–2125.
[19] E. Provost, G. Rieviere, M. Roux, E.D. Morgan, A.G. Bagneres, Insect Biochem. Mol. Biol. 23 (1993) 945–957.
[20] A. Dahbi, A. Lenoir, Behav. Ecol. Sociobiol. 42 (1998) 349–355.
[21] J. Liebig, C. Peeters, N.J. Oldham, C. Markstadter, B. Holldobler, Proc. Natl. Acad. Sci. 97 (2000) 4124–4131.
[22] S. Lahav, V. Soroker, R.K. Vander Meer, A. Hefetz, Behav. Ecol. Sociobiol. 43 (1998) 203–212.
[23] S. Lahav, V. Soroker, R.K. Vander Meer, A. Hefetz, J. Chem. Ecol. 27 (2001) 927–943.