Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Chromatography A, 1096 (2005) 69–75
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Identification of Africanized honeybees

Barry K. Lavine [a,*], Mehul N. Vora [b]

[a] *Department of Chemistry, Oklahoma State University, Stillwater, OK 74078-3071, USA*
[b] *Mathematics & Computer Science, Clarkson University, Potsdam, NY 13699, USA*

Available online 11 July 2005

Abstract

Gas chromatography and pattern recognition methods were used to develop a potential method for differentiating European honeybees from Africanized honeybees. The test data consisted of 237 gas chromatograms of hydrocarbon extracts obtained from the wax glands, cuticle, and exocrine glands of European and Africanized honeybees. Each gas chromatogram contained 65 peaks corresponding to a set of standardized retention time windows. A genetic algorithm (GA) for pattern recognition was used to identify features in the gas chromatograms characteristic of the genotype. The pattern recognition GA searched for features in the chromatograms that optimized the separation of the European and Africanized honeybees in a plot of the two or three largest principal components of the data. Because the largest principal components capture the bulk of the variance in the data, the peaks identified by the pattern recognition GA primarily contained information about differences between gas chromatograms of European and Africanized honeybees. The principal component analysis routine embedded in the fitness function of the pattern recognition GA acted as an information filter, significantly reducing the size of the search space since it restricted the search to feature sets whose principal component plots showed clustering on the basis of the bees' genotype. In addition, the algorithm focused on those classes and/or samples that were difficult to classify as it trained using a form of boosting. Samples that consistently classify correctly are not as heavily weighted as samples that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network. The pattern recognition GA integrates aspects of artificial intelligence and evolutionary computations to yield a "smart" one-pass procedure for feature selection and classification.

Keywords: Pattern recognition; Genetic algorithms; Classification; Chemical taxonomy; Feature selection

1. Introduction

Africanized honeybees are descendants of African bees imported into Brazil by scientists attempting to breed a honeybee better adapted to the South American tropics. The variety of honeybee that resulted from the interbreeding of the established European bee with the newly imported African types, referred to as the Africanized bee, has since dominated the bee fauna of much of South and Central America. In 1990, the Africanized honeybee appeared outside the small south Texas town of Hildago [1]. In the past 14 years, Africanized honeybees have spread to southern California, Arizona, New Mexico, Nevada, and Oklahoma. The success of Africanized honeybees in supplanting the European honeybee population has been attributed to a variety of biological and behavior factors and is one of the most successful introgressions ever documented.

Africanized honeybees have received considerable attention in the popular press. Many stories have stressed the aggressive behavior of this bee and the inherent danger that Africanized bees pose for both man and domestic animals. In addition, Africanized honeybees also have the potential to alter agricultural practices and significantly increase the cost of bee-pollinated food products. Honeybees account for 80% of all insect pollination activity in the United States. They pollinate more than 100 different agricultural products including many fruits and vegetables, forage plants, which are important in production of meat and dairy products, and oil seed crops. The United States Department of Agriculture

* Corresponding author. Tel.: +1 405 744 5945; fax: +1 405 744 6007.
E-mail address: bklab@chem.okstate.edu (B.K. Lavine).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.06.049

estimates that 20 billion dollars worth of agricultural products are dependent on the European honeybee for pollination [2]. If Africanized honeybees appear in the United States in the same form in which they dispersed throughout Brazil, they could have deleterious effects on all aspects of the US agricultural economy influenced by bee pollination.

To control the spread of Africanized bees in the United States, it will be necessary to develop a program of stock certification. This program can only be implemented if a reliable and easy to use method for the identification of Africanized honeybees is developed. Currently, the method used by the United States Department of Agriculture for Africanized honeybee identification is morphometric analysis [3]. This procedure employs a linear discriminant developed from approximately 20 body measurements to identify individual bee specimens as Africanized or European. However, morphometric analysis cannot determine if a given bee population is in the initial stages of becoming Africanized, which is of great interest to Federal and State regulatory officials. Although a polymerase chain reaction based assay has recently been developed [4], more selective primers are needed to ensure accurate genotyping when using this method.

Our previous work using packed column gas chromatography and the linear learning machine [5] to analyze cuticular hydrocarbons of insects has shown that bees which are fully Africanized can be differentiated from European honeybees on the basis of their hydrocarbon profiles [6]. We have also shown that it is possible to identify the African genotype in F1 hybrids [7]. Using gas chromatography and the linear learning machine, we have also demonstrated that heavily Africanized (i.e., bees that are fully Africanized) and moderately Africanized honeybees (bees that are not yet fully Africanized but possess many of the African traits) can be differentiated from European honeybees based on differences in their hydrocarbon profiles [8].

In the present study, hydrocarbon extracts obtained from the wax glands, cuticle, and exocrine glands of 238 European and Africanized honeybees were analyzed by capillary column gas chromatography. A genetic algorithm (GA) for pattern recognition [9–11] was used to identify features in the gas chromatograms characteristic of the African genotype. The pattern recognition GA searched for features in the chromatograms that optimized the separation of the European and Africanized honeybees in a plot of the two or three largest principal components [12] of the data. Because the largest principal components capture the bulk of the variance in the data, the peaks identified by the pattern recognition GA primarily contained information about differences between European and Africanized honeybees. The principal component analysis routine embedded in the fitness function of the pattern recognition GA acted as an information filter, significantly reducing the size of the search space since it restricted the search to feature sets whose principal component plots showed clustering on the basis of genotype. In addition, the algorithm focused on those classes and/or samples that were difficult to classify as it trained using a form of boosting. Samples that consistently classify correctly are not as heavily weighted as samples that are difficult to classify. Over time, the algorithm learns its optimal parameters in a manner similar to a neural network.

2. Experimental

2.1. Bee specimens

Hydrocarbon extracts were obtained from 294 adult worker bees. Of the 294 foragers, 128 were Africanized honeybees and the other 166 were European. The Africanized honeybees were collected from colonies in Costa Rica, Peru, Ecuador, Peru, Honduras, and Mexico. Many of the colonies were designated as moderately or heavily Africanized by workers at these sites based on a field test for colony defense behavior. European honeybees were collected from managed colonies maintained in the United States. They represented a variety of commercially available US stocks.

2.2. Sample preparation

Hydrocarbons were extracted from the wax gland, cuticle, and exocrine gland of individual whole bee specimens by first soaking individual specimens in pesticide grade hexane for 72 h. The hydrocarbons were isolated from the soak by means of a silica gel syringe column (silica Sep-Pac, Millipore) using pesticide grade hexane as the eluent. The hydrocarbon fraction was collected and concentrated to dryness under a stream of nitrogen. It was reconstituted with 50 μl of hexane prior to analysis by capillary column gas chromatography or gas chromatography–mass spectrometery (GC–MS).

2.3. Gas chromatographic analysis

Hydrocarbon extracts obtained from individual whole bees were analyzed on a 25-m 5% phenyl methyl silicone fused silica capillary column (Hewlett-Packard Ultra 2, i.d. = 0.32 mm), which was temperature programmed from 50 to 200 °C at 7°/min and then from 200 to 250° C at 1°/min. The gas chromatographic experiments were performed on a HP5890A instrument equipped with a flame ionization detector. GC–MS analysis was also performed in this study using a Finnigan OWA 1020 automated GC–MS. The presence of normal and branched chain alkanes, alkenes, and dienes was revealed in the extract. GC peaks corresponding to the *n*-alkanes were used as retention standards in the capillary column gas chromatographic experiments. Kovat retention indices were assigned to the compounds eluting from the column and these indices (as well as data from the GC–MS experiment) were used for peak identification. A typical gas chromatographic trace of the hydrocarbon extract from an Africanized honeybee is shown in Fig. 1. Each gas chromatogram contained 65 peaks corresponding to a set

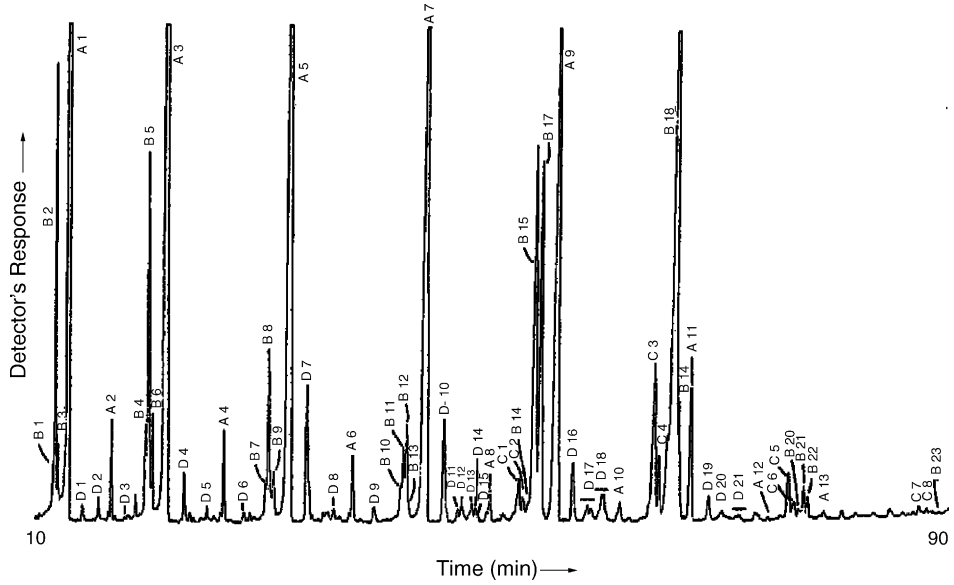

Fig. 1. Gas chromatographic trace of the hydrocarbon extracts obtained from the wax gland, cuticle, and exocrine gland of a heavily Africanized forager. A: normal alkanes; B: alkenes; C: dienes; and D: branched chain alkanes. Reprinted with kind permission from Lavine et al. [8].

of standardized retention time windows. The 65 gas chromatographic peaks selected for pattern recognition analysis were at least moderately resolved and computer integration of these peaks always yielded reliable results. These peaks were also readily identifiable in all the chromatograms by visual analysis so peak matching was not a problem.

3. Pattern recognition analysis

For pattern recognition analysis, the gas chromatographic data was divided into a training set (consisting of 130 European honeybees and 108 heavily and moderately Africanized honeybees, see Table 1) and a prediction set (consisting of 56 European and Africanized honeybees of which the honeybees from San Diego, Tampa, Berkeley and Mexico were not correctly classified by morphometric analysis, see Table 2). Each gas chromatogram was initially represented as a data vector $X = (x_1, x_2, x_3, \ldots x_j, \ldots x_{65})$ where x_j is the area of the jth peak. Such a vector can also be considered as a point in an n-dimensional Euclidean space. A set of chromatograms is therefore represented as a set of points in an n-dimensional Euclidean space. (In this study, n is equal to 65.) A basic assumption is that distances between points in this space are inversely related to their degree of similarity. The expectation is that points representing chromatograms from honeybees possessing the African genotype should cluster in a limited region of this space separate from the points corresponding to the European honeybees.

A genetic algorithm for pattern recognition was used to select features from the training set data characteristic of the bees' genotype. A block diagram of the pattern recognition GA is shown in Fig. 2. During each generation, a population of binary strings of fixed length is generated, each of which represents a potential solution to the Africanized/European honeybee classification problem. For a GC peak to be included, it is necessary for the corresponding bit in the string to be set at 1. If the bit is set to 0, the corresponding GC peak is not included. The strings are decoded yielding the subset of the 65 GC peaks sent to the fitness function for evaluation. Each string is assigned a value by the fitness function, which is a measure of the degree of separation between the European and Africanized honeybees in a principal component plot of the data defined by the extracted feature subset. The fitness (i.e., the quality of the proposed feature subset for bee classification) is used to select potential solutions for recombination, which produces a new population of strings. The power of the GA arises from recombination [13,14],

Table 1
Training set

Specimen type	Number of honeybees
European foragers from United States	130
Heavily Africanized foragers	64
Moderately Africanized foragers	44
Total	238

Table 2
Prediction set

Specimen type	Number of honeybees
San Diego (European)	7
Tampa (European)	22
Berkeley (European)	7
Mexico (Africanized)	6
French Guinea (Africanized)	7
Peru (Africanized)	7
Total	56

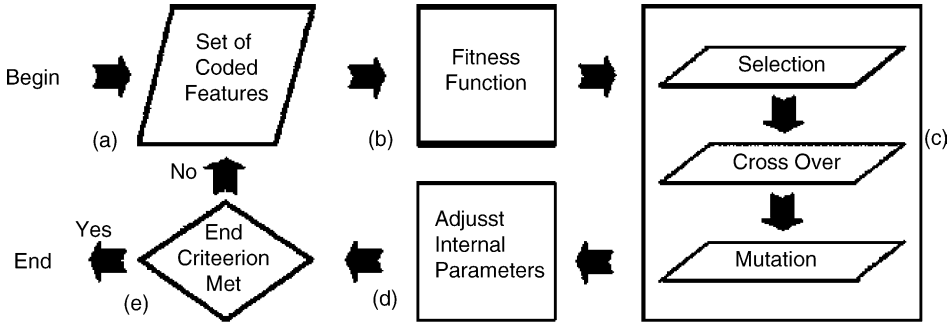

Fig. 2. Block diagram of the pattern recognition GA. Reprinted with kind permission from B.K. Lavine, A.J. Moores, H.T. Mayfield, A. Faruque, Fuel spill identification using gas chromatography/genetic algorithms-pattern recognition techniques, Analytical Letters 31 (1998) 2805.

which causes a structured yet randomized exchange of information between strings (i.e., potential solutions), with the expectation that good solutions can generate even better ones. In addition, some of the binary strings may undergo mutation, where one of the bits is randomly changed. (If a bit is zero or the feature is excluded, the mutation operator applied to the string in question causes the bit to change to one and forces its inclusion in the feature subset or vice versa.) This allows the GA to search adjacent regions of the solution space mitigating local convergence. The aforementioned processes: evaluation, selection, crossover, reproduction, and adjustment of internal parameters (which is discussed below), are repeated until a specified number of generations is achieved or a feasible solution is found. The operators comprising our pattern recognition GA are described below.

3.1. Evaluation

The pattern recognition GA emulates human pattern recognition through machine learning to score the principal component plots. To track and score the principal component plots, class and sample weights, which are an integral part of the fitness function, are computed (see Eqs. (1) and (2)) during each generation. Class weights sum to 100; the sample weights for samples of a particular class sum to a value equal to the corresponding weight of the class.

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

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

The principal component plot generated for each chromosome after the subset of features in the chromosome has been extracted is scored with 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, and a poll is taken of the point's K-nearest neighbors. For the most rigorous classification, K (which is a user defined parameter) equals the number of samples in the class to which the sample point belongs. The number of K-nearest neighbors with the same class label as the sample point in question, the so-called sample hit count (SHC), is computed ($0 \le \mathrm{SHC(s)} \le K_c$). Using Eq. (3), it is a simple matter to score each principal component plot, i.e., determine the degree of separation between classes in the plot.

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

To better understand the scoring of the principal component plots, consider a data set with two classes, which have been assigned equal weights. Class 1 (e.g., Africanized honeybees) has 20 samples, and class 2 (e.g., European bees) has 50 samples. At generation 0, all samples in a given class (European or Africanized) will have the same weight. Thus, each sample (honeybee) in class 1 has a sample weight of 2.5, whereas each sample (honeybee) in class 2 has a weight of 1. Suppose a sample (honeybee) from class 1 has as its 20 nearest neighbors, 14 Africanized honeybees and 6 European honeybees. Hence, $\mathrm{SHC}/K = 0.7$, and $(\mathrm{SHC}/K) \times \mathrm{SW} = 0.7 \times 2.5$, which equals 1.75. By summing $(\mathrm{SHC}/K_c) \times \mathrm{SW}$ for each bee sample, the principal component plot can be scored.

3.2. Adjusting internal parameters

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 the weights, it is necessary to first compute the sample hit rate, SHR(s), which is the mean value of SHC/K_c over all feature subsets in a particular generation. SHR(s) is a measure of the difficulty of classifying a particular sample. If a sample (e.g., honeybee) is difficult to classify, it has a low sample hit rate since it has a low SHC/K_c value in most feature subsets of the population. If a sample (e.g., honeybee) is easy to classify, it has a high sample hit rate since it has a high SHC/K_c value in most feature subsets of the population.

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

$$\mathrm{CHR}_g(c) = \mathrm{AVG}(\mathrm{SHR}_g(s) : \forall_{s\in c}) \tag{5}$$

Next, the class-hit rate (see Eq. (5)), which is the average sample hit rate for all of the samples in a class, is computed. Class and sample weights are then adjusted using a perceptron (see Eqs. (6) and (7)). Classes with a low class hit rate and samples with a low sample hit rate are weighted more heavily than classes or samples that score well. The user must set the momentum, P. The value of P should be high enough to facilitate learning while ensuring that a particular sample or class does not dominate the calculation, which would result in other samples and/or classes not contributing to the fitness function. For this reason, P is usually set at 0.5. After a certain number of generations, the class weights will not change. Eq. (6) is then turned off and the GA focuses exclusively on the troublesome samples via Eq. (7). P is then halved. During each generation, class and sample weights are updated (i.e., boosted) using the class and sample hit-rates from the previous generation. ($g + 1$ is the current generation, whereas g is the previous generation.) Boosting of sample and class weights is crucial because it modifies the fitness landscape, as the population evolves towards a better solution, thereby ensuring that convergence to a local optimum does not occur.

$$\mathrm{CW}_{g+1}(s) = \mathrm{CW}_g(s) + P(1 - \mathrm{CHR}_g(s)) \quad (6)$$

$$\mathrm{SW}_{g+1}(s) = \mathrm{SW}_g(s) + P(1 - \mathrm{SHR}_g(s)) \quad (7)$$

3.3. Selection

Selection, crossover, and mutation operators are applied to the chromosomes to develop new and potentially better solutions. The selection operator is implemented by ordering the population of strings, i.e., the potential solutions, from best to worst by their fitness while simultaneously generating a copy of the same population and randomizing the order of the strings in this copy with respect to their 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 strings from the top half of the random population, guaranteeing the best 50% are selected for reproduction, while every string in the randomized copy has a uniform chance of being selected due to the randomized selection criterion imposed on the strings from this population. This selection process strikes a balance between genetic diversity and elitism with the data dictating the juxtaposition of these two opposing behaviors.

For each pair of strings selected for mating, two new strings are generated using two-point crossover. The resulting population of strings, both parents and children, are sorted by their fitness and the top ϕ strings are retained for the next generation. (Our previous studies have shown that inclusion of both parents and children as opposed to only children in the next generation improves algorithm performance.) The new population should perform better on average than its predecessor because the selection criterion used for reproduction exhibits bias for the higher-ranking strings. However, the aforementioned reproduction operators also assure a significant degree of diversity in the population, since the crossover points and reordering of exchanged string fragments of each chromosome pair is selected at random.

4. Results and discussion

The first step in this study was to apply principal component analysis to the data. Principal component analysis [16,17] is the most widely used multivariate analysis technique in science and engineering. It is a method for transforming the original measurement variables into new, uncorrelated variables called principal components. Each principal component is a linear combination of the original measurement variables. Using this method is analogous to finding a new coordinate system better at conveying information present in the data than axes defined by the original measurement variables. This new coordinate system is linked to variation in the data. The basis vectors of this new coordinate system are the principal components. Often, only the two or three largest principal components are necessary to explain all of the information present in a data set if the data contains a large number of interrelated measurement variables. Using principal component analysis, dimensionality reduction, classification of samples, and identification of outliers in high dimensional data is possible.

Fig. 3 shows a plot of the two largest principal components of the 238 European and Africanized honeybee specimens that comprise the training set. Each bee is represented as a point in the principal component plot: 1 represents European honeybees, and 2 represents moderately and heavily Africanized honeybees. The overlap between the

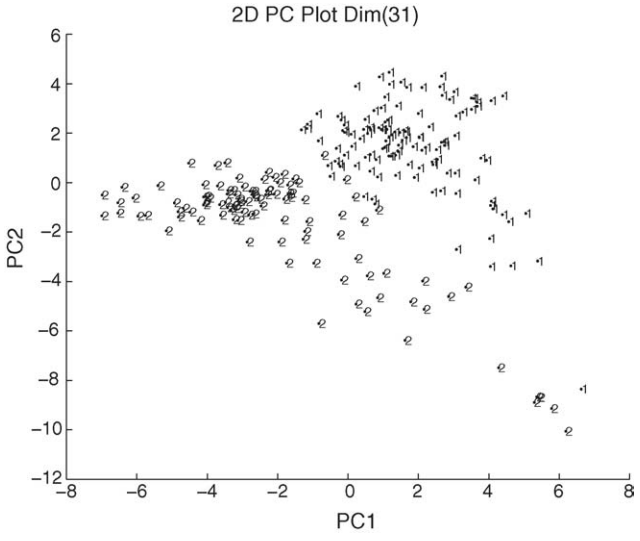

Fig. 3. Plot of the two largest principal components of the 65 GC peaks and 238 European and Africanized honeybee gas chromatograms that comprise the training set. Each bee is represented as a point in the principal component plot: 1 represents European honeybees, and 2 represents moderately and heavily Africanized honeybees.

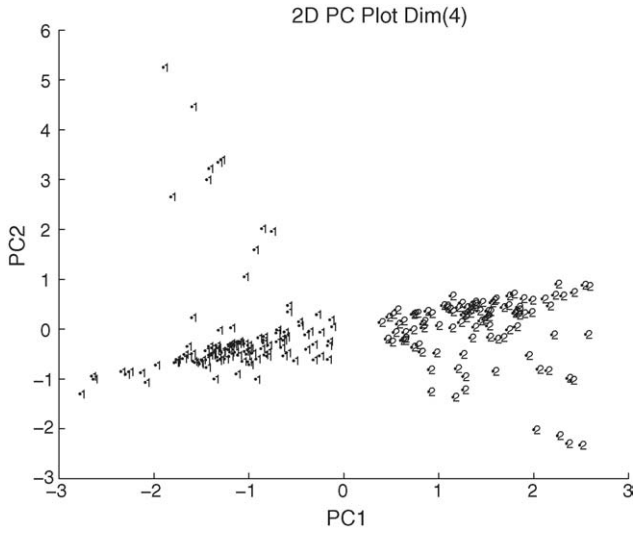

Fig. 4. Plot of the two largest principal components developed from the four GC peaks identified by the pattern recognition GA. Each bee is represented as a point in the principal component plot: 1 represents European honeybees, and 2 represents moderately and heavily Africanized honeybees. Clustering of the honeybees by genotype is evident.

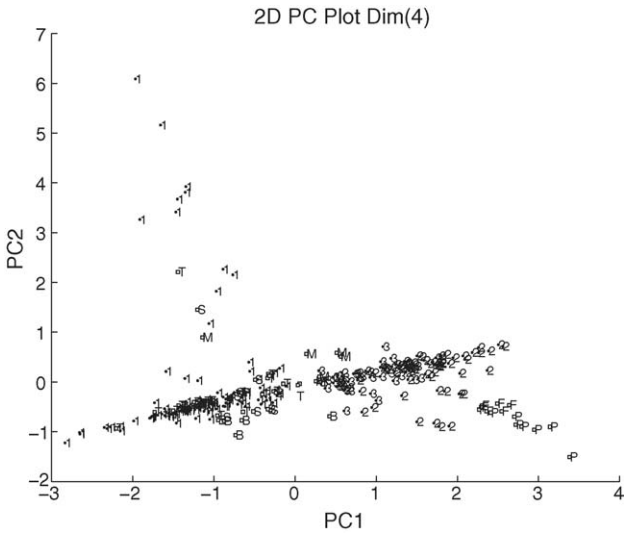

Fig. 5. Projection of the honeybees from the prediction set onto the principal component plot defined by the 238 training set samples and the four features identified by the pattern recognition GA. Each gas chromatogram from the training set is represented by a number: 1: European honeybees, 2: heavily Africanized honeybees, and 3: moderately Africanized honeybees. Each chromatogram from the prediction set is represented by a letter: T: Tampa, M: Mexico, B: Berkeley, S: San Diego, F: French Guinea, and P: Peru. The two circled prediction set samples are incorrectly classified in the principal component map developed from the training set data.

gas chromatograms of European and Africanized honeybee in the plot is evident.

The pattern recognition GA was used in this study to identify key features that are characteristic of the hydrocarbon profile of each class. Features were identified by sampling key feature subsets, scoring their principal component plots, and tracking classes and/or samples, which 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 four peaks (B11, B14, B15, and B22) whose principal component plot showed clustering on the basis of genotype (see Fig. 4). These four peaks correspond to the following alkenes: 9-C29:1, 6-C29:1, 10-C31:1, and 10-C35:1.

The ability of the GC data to predict the class of an unknown sample was first tested using a procedure known as internal (cross) validation. The original training set of 238 gas chromatograms (samples) was divided into 8 training set/validation set pairs. Each pair consisted of 208 training set samples and 30 validation set samples. Each training set/validation set pair was generated by random selection. A particular bee specimen was present in at least one of the 8 validation sets. Features that correctly classified the bees were identified using the training set and were then tested on the corresponding validation set. The validation set samples were classified by computing their coordinates for a new coordinate system based on the two largest principal components developed from the 208 training set samples and the informative GC peaks identified by the pattern recognition GA. Both the training set and prediction set samples were then mapped onto the space defined by this new coordinate system. Validation set samples were correctly classified if they were projected into clusters containing training set samples that had the same class label. The classification success rate for this study was 100%. Furthermore, the same four features identified in the run involving the entire training set were also identified as key features in the cross validation study.

A set of 56 gas chromatograms (see Table 2) was used to assess the predictive ability of the four peaks identified by the pattern recognition GA. The prediction set samples were projected onto the principal component map developed from the 238 gas chromatograms and 4 gas chromatographic peaks. Fig. 5 shows the projection of the prediction set samples onto a principal component map defined by the four peaks selected by the pattern recognition GA. Fifty-four of the 56 projected samples lie in a region of the map occupied by bees possessing the same class label. One honeybee from Mexico was classified as European and one Tampa bee was classified as Africanized.

We consider these results to be significant since morphometric analysis could not correctly classify any of the honeybees from San Diego, Tampa, Berkeley, or Mexico in the prediction set. Evidently, the pattern recognition GA can identify features in the gas chromatograms characteristic of genotype. This suggests that concentration patterns of high molecular weight hydrocarbons convey taxonomic information about honeybees. Using gas chromatography and pattern recognition methods, an entomologist can correctly identify the subspecies of a bee specimen by simply measuring the concentration of only a few hydrocarbons. This approach to taxonomy places species identification on a firm chemical basis.

Acknowledgements

The authors thank Howell Daly (University of California at Berkeley) and Orley Taylor (University of Kansas) for obtaining the European and Africanized honeybee specimens used in this study and Roy Keith Smith for performing the gas chromatographic and GC–MS studies.

References

[1] E.A. Sugden, K.R. Williams, Glean. Bee Cult. 119 (1990) 18.
[2] R. McDowell, In Agricultural Economics Report No. 519, United States Department of Agriculture, Washington, DC, 1984.
[3] H.V. Daly, S.S. Balling, J. Kans. Entomol. Soc. 51 (1978) 857.
[4] M.A. Pinto, J.S. Johnston, W.L. Rubink, R.N. Coulson, J.C. Patton, W.S. Sheppard, Ann. Entomol. Soc. Am. 96 (2003) 679.
[5] T.L. Isenhour, P.C. Jurs, Anal. Chem. 43 (1971) 20A.
[6] B.K. Lavine, D. Carlson, Anal. Chem. 59 (1987) 468A.
[7] B.K. Lavine, D. Carlson, P.C. Jurs, J. Chemom. 2 (1988) 29.
[8] B.K. Lavine, A.J.I. Ward, R.K. Smith, O.R. Taylor, Microchem. J. 39 (1989) 308.
[9] B.K. Lavine, C.E. Davidson, A.J. Moores, Vib. Spectrosc. (2002) 83.
[10] B.K. Lavine, C.E. Davidson, C. Breneman, W. Katt, J. Chem. Inf. Sci. 43 (2003) 1890.
[11] B.K. Lavine, C.E. Davidson, W.T. Rayens, Comb. Chem. High Through. Screen. 7 (2004) 115.
[12] J. Edward Jackson, A User's Guide to Principal Component Analysis, John Wiley & Sons, NY, 1991.
[13] D.E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison Wesley, Reading, MA, 1989.
[14] Z. Michalewicz, Genetic Algorithms + Data Structures = Evolution Programs, Springer Verlag, NY, 1992.
[15] K. Fukunaga, Statistical Pattern Recognition, second ed., Academic Press, San Diego, 1990.
[16] I.T. Jolliffe, Principal Component Analysis, Springer Verlag, New York, 1986.
[17] E.R. Malinowski, Factor Analysis in Chemistry, second ed., John Wiley & Sons, New York, 1991.