dehydroalanine will occur to the same extent as with dehydroalanine. Yet, the possibility should not be overlooked. Further work is needed to assess the role and behavior of substituted threonyl residues in processed high protein foods.

ACKNOWLEDGMENT

This work was supported in part by National Institutes of Health Grant HD 00122 and Food and Drug Administration Grant FD 00568-02. Parts of this material were presented at the 67th Meeting of the American Society of Biological Chemists, June 6–10, 1976, San Francisco, Calif., and at the Centennial Meeting of the American Chemical Society, Aug 1976, San Francisco, Calif. This material constitutes part of the thesis of Honson S. Lee to be submitted to the Graduate Division of the University of California at Davis in partial fulfillment of the requirements for a Ph.D. degree in Nutrition. Appreciation is due to Clara Robison and Gail Nilson for typing of the manuscript and Chris Howland for editorial assistance.

LITERATURE CITED

Adams, J. B., *Biochem. J.* **94**, 368 (1965).
Asquith, R. S., Booth, A. K., Skinner, J. D., *Biochim. Biophys. Acta* **181**, 164 (1969).
Bohak, Z., *J. Biol. Chem.* **239**, 2878 (1964).
Carter, C. E., Greenstein, J. P., *J. Biol. Chem.* **165**, 725 (1946).
Carubelli, R., Bhavanandan, V. P., Gottschalk, A., *Biochim. Biophys. Acta* **101**, 67 (1965).
Corfield, M. C., Wood, C., Robson, A., Williams, M. J., Woodhouse, J. M., *Biochem. J.* **103**, 15c (1967).
DeVries, A. L., Vandenheede, J., Feeney, R. E., *J. Biol. Chem.* **246**, 305 (1971).
Donovan, J. W., Davis, J. G., Wiele, M. B., *J. Agric. Food Chem.* **20**, 223 (1972).
Downs, F., Pigman, W., *Methods Carbohydr. Chem.* **7**, 200 (1976).
Feeney, R. E., *Am. Sci.* **62**, 712 (1974).
Fiat, A.-M., Alais, C., Jolles, P., *Eur. J. Biochem.* **27**, 408 (1972).
Gehrke, C. W., Stalling, D. L., *Sep. Sci.* **2**, 101 (1967).
Gelpi, E., Koening, W. A., Gibert, J., Orõ, J., *J. Chromatogr. Sci.* **7**, 604 (1969).
Gottschalk, A., in "Glycoproteins. Their Composition, Structure, and Function", Vol. 5A, Gottschalk, A., Ed., Elsevier, Amsterdam, 1972, p 470.
Gross, E., Chen, H. C., Brown, J. H., *Fed. Proc., Fed. Am. Soc. Exp. Biol.* **34**, abstr. no. 3392 (1975).
Heller, E., Raftery, M. A., *Biochemistry* **15**, 1194 (1976).
Horn, M. J., Jones, D. B., Ringel, J. J., *J. Biol. Chem.* **138**, 141 (1941).
Jones, D. B., Caldwell, A., Horn, M. J., *J. Biol. Chem.* **176**, 65 (1948).
Kato, A., Nakamura, R., Sato, Y., *Agric. Biol. Chem.* **34**, 1009 (1970).
Kato, A., Sato, Y., *Agric. Biol. Chem.* **36**, 831 (1972).
Komatsu, S. K., Ph.D. Thesis, University of California, Davis, 1969.
Komatsu, S. K., DeVries, A. L., Feeney, R. E., *J. Biol. Chem.* **245**, 2909 (1970).
Mayo, J. M., Carlson, D. M., *Carbohydr. Res.* **15**, 300 (1970).
Nashef, A. S., Osuga, D. T., Lee, H. S., Ahmed, A. I., Whitaker, J. R., Feeney, R. E., *J. Agric. Food Chem.* **25**, 245 (1977).
Neiderhiser, D. H., Plantner, J. J., Carlson, D. M., *Arch. Biochem. Biophys.* **145**, 155 (1971).
Nutr. Rev. **34**, 120 (1976).
Osuga, D. T., Feeney, R. E., in "Fundamental Aspects of Proteins Basic to Foods", Whitaker, J. R., Tannenbaum, S., Ed., Avi, Westport, Conn., 1977, p 209.
Patchornik, A., Sokolovsky, M., *J. Am. Chem. Soc.* **86**, 1860 (1964).
Pigman, W., Moschera, J., *Adv. Chem. Ser.* **117**, 220 (1973).
Plantner, J. J., Carlson, D. M., *Anal. Biochem.* **65**, 153 (1975).
Price, V. E., Greenstein, J. P., *J. Biol. Chem.* **171**, 477 (1947).
Rando, R., *Science* **185**, 320 (1974).
Sen, L. C., Gonzalez-Flores, E., Feeney, R. E., Whitaker, J. R., *J. Agric. Food Chem.*, **25**, 632 (1977).
Simpson, D. L., Hranisavljevic, J., Davidson, E. A., *Biochemistry* **11**, 1849 (1972).
Smith, M. B., Reynolds, T. M., Buckingham, C. P., Back, J. F., *Aust. J. Biol. Sci.* **27**, 349 (1974).
Smyth, D. S., Utsumi, S., *Nature (London)* **216**, 332 (1967).
Spiro, R. G., *Methods Enzymol.* **28**, 3 (1972).
Spiro, R. G., *Adv. Protein Chem.* **27**, 349 (1973).
Tanaka, K., Bertolini, M., Pigman, W., *Biochem. Biophys. Res. Commun.* **16**, 404 (1964).
Tanaka, K., Pigman, W., *J. Biol. Chem.* **240**, Pc1488 (1965).
Vandenheede, J. R., Ahmed, A. I., Feeney, R. E., *J. Biol. Chem.* **247**, 7885 (1972).
Williams, M. J., Woodhouse, J. M., in "The 5th Colloquium in Amino Acid Analysis", Technicon International Division, Domont, France, 1967, p 96.
Ziegler, K., Melchert, I., Lürken, C., *Nature (London)* **214**, 404 (1967).

Received for review November 10, 1976. Accepted April 15, 1977.

Extraction of Important Molecular Features of Musk Compounds Using Pattern Recognition Techniques

William E. Brugger and Peter C. Jurs*

The relationships between molecular structure and the musk odor quality were investigated using pattern recognition techniques. A data set consisting of 60 musk odorants and 240 nonmusk compounds were coded with computer generated structural descriptors and then analyzed using a linear learning machine. After determining that the data set was linearly separable, a subset of 13 descriptors was identified and subsequently employed to predict the odor quality of nine, previously unused, musk odorants: all were correctly classified. The results of this work demonstrated the usefulness of pattern recognition techniques for studying structure–activity relationships of oflactory stimuli and elucidated some structural parameters common among musk odorants.

The perception of odors occurs in humans when airborne molecules of a volatile substance interact with some type of receptors in the olfactory region of the nose. Although the detailed mechanism of these interactions as well as the composition of the receptors remain unknown, several theories have been proposed which attempt to correlate different molecular properties with the perceived odor quality of a substance. For example, Wright (1954)

Department of Chemistry, The Pennsylvania State University, University Park, Pennsylvania 16802.

suggested that odor quality could be predicted on the basis of low-energy molecular vibrations occurring below 600 cm^{-1}. Good correlations were found in a study of substances with nitrobenzene like odor (Wright and Serenius, 1954) as well as in a study of musk odorants (Wright and Burgess, 1969). On the other hand, Amoore (1970) has reported that molecular shape, size, and electronic nature of a molecule are correlated to odor quality. Likewise, intermolecular interaction forces (Dravnieks and Laffort, 1972), molecular profiles (Beets, 1957), and functional groups in molecules (Henning, 1915; Brower and Schafer, 1975) have all been found to be related to some odor qualities. Unfortunately, none of these molecular parameters alone can predict the odor quality of a large collection of olfactory stimuli.

Nevertheless, the possibility exists that a collection of several different molecular parameters can account for odor quality. Such an approach was taken by Schiffman (1974) who used multidimensional scaling techniques and a newly developed weighting procedure to reproduce an odor space of 39 odorants using 25 physicochemical parameters. A correlation of 0.76 was found between the calculated and experimentally determined odor spaces. In another study, Dravnieks (1974) used 14 structural features and multiple linear regression to determine linear equations which would fit measured intensity, threshold, and odor quality data. Again, good correlations were found for the molecules used in this study. In a more recent study, the Hansch approach, which was developed for application to structure–activity relationship studies in medicinal chemistry, was used by Boelens (1976) to study compounds with bitter almond and musk odors. Using only the 1-octanol/water partition coefficients, gas chromatographic retention times, and molecular shape and volume parameters of the odorant molecules, good regression equations were found for the 14 compounds regressed in each odor category. Of the parameters tested, the partition coefficients were found to be the most important in each study.

The ability to obtain reproducible and relatively error free odor property measurements from human observers on a large number of chemical compounds is a major problem in olfactory research. Consequently, any attempt to fit quantitative data of this nature is limited by the inherent error. Therefore, until methods are developed to obtain good quantitative olfactory data, studies using regression analysis and other parametric methods will be restricted by this limitation.

Although a large amount of qualitative data exists on a large number of olfactory stimulants, studies of relationships between molecular structure and odor quality have been done only on limited data set using simple correlation techniques with a few variables. One factor limiting the investigations of qualitative data has been the lack of techniques for handling data of this type. In this paper, the usefulness of pattern recognition techniques for investigating molecular parameters which can predict odor quality will be demonstrated.

Pattern recognition involves the perception and recognition of significant features or attributes which can categorize input data into identifiable classes. In olfactory research, the specific objective of using pattern recognition is to find "common properties" which are shared by all members of a particular class of odorants and which can serve to classify a new compound into an odor class with similar properties.

Pattern recognition techniques are well established with initial research efforts dating back into the early 1950's when computer technology started its growth. Since then, these techniques have grown and expanded into several fields of application including computer and information science, statistics, biology, physics, and medicine. Pattern recognition has been applied to a large variety of chemical problems (see reviews: Jurs and Isenhour, 1975; Isenhour et al., 1974; and Kowalski, 1975) including applications to structure–activity relationships in pharmacology by Hiller et al. (1973), Chu et al. (1975), and Stuper and Jurs (1975).

Pattern recognition techniques are uniquely suited for doing qualitative structure–activity relationship studies because of various characteristics of the procedures. First, heuristic methods are available which assume no mathematical model, but rather relationships are sought which provide definitions of similarity between diverse groups of data. Pattern recognition techniques are also able to deal with high-dimensional data where more than three measurements are used to describe each object or event. Furthermore, pattern recognition techniques can handle data in which the relationships are discontinuous as well as multisource data where each measurement can be the result of an independent experiment. This attribute is very important since structure–activity relationship studies involve data of this type. Finally, techniques are available for selecting important features from a large set of parameters. Thus, studies can be done on systems where the exact relationships are not fully understood.

The purpose of this paper is to report our investigation into the applicability of pattern recognition techniques for performing qualitative structure–activity relationship studies of olfactory stimuli as well as to determine structural features which can be used to predict musk odorants. The premises for applying pattern recognition to this type of study are: (1) molecular structure and odor quality are related, (2) the structure of a compound can be adequately represented by a set of molecular descriptors, (3) a relationship can be discovered between the structure's molecular descriptors and their odor quality by applying pattern recognition analysis to a set of tested compounds, and (4) any relations discovered can be extrapolated to predict the odor quality of untested compounds.

PROCEDURES

The procedures for doing any computer aided structure–activity relationship study can be broken down into the following general steps: (1) identify the data set and transform the compound's structural diagrams into computer compatible files, (2) generate molecular descriptors from the structures for each data set member, and (3) analyze the descriptors by searching for any relationships. Although there are several different ways to execute the above steps, the following discussion will describe only the procedures used in this study of musk odorants.

The class of compounds commonly referred to as musks was chosen for this initial study primarily because musk is a characteristic odor quality which perfumers rarely confuse with other odor qualities. Therefore, a data set composed of this class of odorants should be relatively free of misclassified compounds. Such a well-characterized data set is important for providing a fair test of the capabilities of pattern recognition techniques for performing structure–activity relationship studies in olfaction. However, this factor was not the only one to play a role in the selection of this class of compounds.

In recent years the perfume manufacturers have done considerable research in developing synthetic musk odorants to replace the diminishing supply of natural musk odorants. Consequently, a large amount of information on musk odorants exists (e.g., Beets, 1973; Theimer and Davies, 1967; and Beets, 1971). This is not true for any other major class of odorants. Furthermore, this class of compounds is structurally interesting since it contains a variety of different structural types including some steroids. In general, the structure–activity relationship study of musk odorants using pattern recognition techniques presented itself as a challenging problem with a high probability of success.

For this study, a data set consisting of 300 unique compounds was obtained from the list of odorants given by Amoore (1970). Sixty of these compounds were musk odorants and included 23 macrocyclic, 19 polynitrobenzenes, 11 steroids, 5 γ-butyrolactones plus two other structural types. Although 16 of the musk compounds were classified as weak musks or as having other odor overtones, enough strong musks were present to assure a fair representation of musk odorants.

To represent the nonmusk class of odorants, 240 compounds were randomly selected from the other odor categories given by Amoore (1970). The nonmusk class included 49 camphoraceous, 44 floral, 32 ethereal, 41 mint, 51 pungent, and 23 putrid compounds. In this nonmusk class, a large number of different functional group types as well as structural types were present to assure a good representation of the nonmusk class.

The structures of these 300 compounds served as the input to a computerized pattern recognition system called ADAPT (see Stuper and Jurs, 1976, for a complete description of the system). The two-dimensional representation of these compounds were encoded by drawing them on a graphics display terminal with an interactive program (cf. Brugger and Jurs, 1975), which converted the graphical representation of the structures into computer compatible connection tables. The set of connection tables for the 300 compounds were then stored on the system's disc files and were subsequently used to generate molecular descriptors.

Fragment, substructure, and geometric descriptors were all employed in this study of musk odorants. Since Brugger et al. (1976) contains a detailed discussion of these descriptor types as well as the methods employed in calculating them, only a brief introduction to each descriptor type will be presented here.

Any chemical structure can be broken down into its basic atom and bond components which are called "fragment" descriptors. Although these descriptors do not contain any structural information, they do reflect the chemical nature of the molecule. The total number of atoms, bonds, carbon atoms, oxygen atoms, nitrogen atoms, single bonds, double bonds, aromatic bonds, and a weighted summation of the four basic bond types were all generated for the 300 members of the musk data set.

Substructure descriptors were generated by searching the connection tables of the molecular structures for the presence of functional groups and other explicitly defined larger atom and bond fragments. These descriptors contain information about the compounds' chemical functionality and some structural information which was lost in the formation of fragment descriptors. Fifty-one different substructure descriptors were generated for this study.

In order to generate geometric descriptors, three-dimensional structures of the compounds are required. Since x-ray data was incomplete for the data set, and measuring space filling models constructed by hand would have been too tedious and inaccurate, a computer program was implemented to calculate low-strain, three-dimensional chemical structures. In this program a molecule is viewed as a collection of spherical atoms held together by a simple harmonic or elastic forces which are defined by a potential energy function. The independent variables of this function are the three-dimensional coordinates of the atoms in the molecule. By minimizing this function, a strain-free model of the molecule is obtained. From each molecule's coordinate matrix the structure's three principal moments of rotation (X = longest, Y = intermediate, Z = shortest), three ratios of the moments (X/Y, X/Z, and Y/Z), and the molecular volume can be calculated. All seven of these geometric descriptors were generated for all members of this data set.

Descriptors originating from experimental data were not included for two reasons: (1) obtaining this type of information from the literature for a large and diverse data set is extremely difficult if not impossible; and (2) the possibility of using any classifier developed in this study as a prescreen for new odorants would be ruled out since the compounds would have to be synthesized to obtain the data. Therefore, only the computer derived molecular descriptors mentioned previously were used in this study.

Since each descriptor has its own origin, scale, range, mean, and distribution of values, some form of preprocessing was required to alleviate any potential scaling problems. In this study the variables were standardized by adjusting the means to zero and the standard deviations to unit for each descriptor over all of the data set members.

The descriptors were subsequently combined into pattern vectors with the ith compound being represented by the vector: $\mathbf{X}_i = (x_1, x_2, ..., x_n, x_{n+1})$ where n equals the number of descriptors chosen to describe the compounds. Each component of the vector represents one observation or measurement. For example, x_1 could be the molecular weight of the compound, x_2 could be the molecular volume, and x_n could be the number of oxygen atoms in the compound. The $n + 1$ component of the vectors is a constant value added to all the vectors for computational convenience during the analysis of the data.

Data represented as vectors can be thought of as either points in an n-dimensional Euclidean space or as vectors pointing from the origin to those points. There is a one-to-one correspondence between the points and the compounds represented in this way. Experience has shown that points representing patterns with common characteristics cluster in limited regions of the n-dimensional space. Thus, one might expect the points representing compounds that are musks to cluster in one region of the space and the nonmusk compounds to cluster elsewhere. A way to investigate the structure of the set of points is to separate the clusters from one another by decision surfaces, with the simplest surface being an n-dimensional plane. Two clusters of points which can be completely separated by such a plane are said to be linearly separable.

Any n-dimensional plane has associated with it a normal vector called here a weight vector. The weight vector consists of an ordered sequence of components which correspond with the components of the pattern vectors used to describe the data set. Any pattern vector can be classified with respect to a decision surface by taking the dot product of the pattern vector and the decision surface's weight vector ($\mathbf{W}$):

$$\mathbf{S} = \mathbf{W} \cdot \mathbf{X} = w_1x_1 + w_2x_2 + \cdots + w_nx_n + w_{n+1}x_{n+1} = |\mathbf{W}|\ |\mathbf{X}| \cos \theta$$

Since $|\mathbf{W}|$ and $|\mathbf{X}|$ are always positive, the angle between the vectors, θ, determines the sign of the dot product. For patterns on one side of the plane, the dot product is always positive while the patterns on the opposite side have negative dot products. Therefore, any compound in the data set can be classified into one of the two classes by obtaining the sign of the dot product.

The remaining problem in this analysis is finding useful decision surfaces. Although parametric methods, such as multivariate discriminant analysis, could be used to obtain decision functions, we selected the nonparametric linear learning machine from the area of pattern recognition analysis for this work. The linear learning machine develops an effective decision surface by using a training set of pattern vectors whose correct classification are known. The members of this training set are presented to the learning machine one at a time and the decision surface's weight vector, which has been arbitrarily initialized, is used to classify each compound in turn. When an incorrect classification is made, the weight vector is altered in such a manner that will correctly classify the missed compound. This process continues until all of the training set members are correctly classified. One method for altering the weight vector is to move the decision surface so that after correction the misclassified vector is the distance on the correct side of the surface as it was previously on the incorrect side. The details of this procedure as well as other approaches can be found in Tou and Gonzalez (1974) or Nilsson (1965).

Although the ultimate use of any decision surface developed in the linear learning machine is to predict the class of an unknown compound, it can also be used to aid in the removal of unnecessary descriptors which might have been initially included into each pattern vector. Variance feature selection is a nonparametric method developed for use with the linear learning machine to accomplish this task. Given a linearly separable data set, the variance feature selection method can produce a list which ranks the descriptors under consideration according to importance. Using this list, unimportant descriptors can be removed until the minimal set, sufficient for separation, remains (cf. Zander et al., 1975). Thus, both a classifier and the intrinsic descriptors necessary for separation are obtained.

RESULTS

The initial test was to determine if a decision surface could be found to separate the 60 musk odorants from the other 240 compounds. Using the training procedure described above and the 68 descriptors generated for this study, a decision surface was found which correctly classified the entire data set. Knowing this, several studies were conducted to determine which of the 68 available descriptors were most important for the separation.

Instead of using the variance feature selection method to reduce the initial 68 descriptors, it was decided to test the geometric, fragment, and substructure descriptors individually as to their ability to separate the data set. In doing this, it was found that neither the fragment nor the geometric descriptors alone were able to completely separate the musks from the nonmusks. Even the combination of these descriptors was unsuccessful in finding a decision surface. However, it was found that only a few compounds were preventing linear separability. Using the seven geometric descriptors alone, only the ten compounds shown in Figure 1 were misclassified out of the entire data set. Upon the inclusion of the ten fragment descriptors into each pattern vector, only compounds a, b, and c in Figure 1 prevented linear separability. Since these three compounds have been characterized as weak musks, it was not surprising that they were confusing the linear learning machine and thus preventing linear separability. Instead of removing these compounds from the data set, it was decided to test more descriptors on the entire data set.

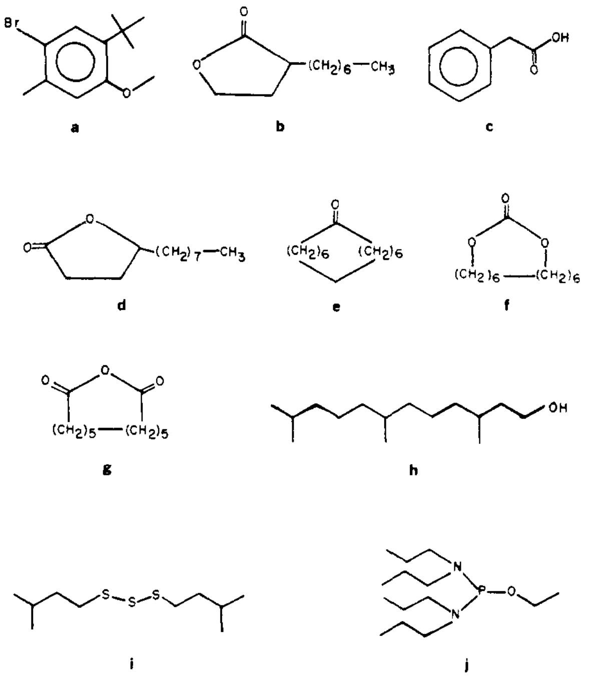

Figure 1. The ten compounds misclassified using only the seven geometric descriptors: (a) 2-bromo-4-*tert*-butyl-5-methoxytoluene, (b) α-heptyl-γ-butyrolactone, (c) phenylacetic acid, (d) γ-octyl-γ-butyrolactone, (e) cyclotetradecanone, (f) dodecamethylene carbonate, (g) dodecanedicarboxylic acid anhydride, (h) norhexahydrofarnesol, (i) diisoamyltrisulfide, (j) phosphorus acid ethyl ester bisdipropylamide.

Table I. Distribution of the Data Set Members into Training and Prediction Sets

	Training sets		Prediction sets		Number sets
Group	Musks	Nonmusks	Musks	Nonmusks	generated
A	50	200	10	40	20
B	48	192	12	48	20
C	45	180	15	60	20
D	42	168	18	72	20

When the 51 substructure descriptors were tested alone, a decision surface was found which correctly classified the entire data set. In order to have a measure of the predictive ability of these descriptors, the data set was subdivided into a series of training and prediction sets. Four groups of 20 training and prediction sets each were generated from the initial data set of 300 compounds. The distribution of musk and nonmusk compounds for the training and prediction sets in each group is given in Table I. The selection of the individual compounds for each set was done randomly so that no training and prediction sets were identical. Each training set was used to develop a decision surface which was then used to classify the compounds in the prediction set. The compounds in the prediction set served as unknowns since they were not used to develop the decision surface which was subsequently used to classify them. By using the 20 randomly selected sets in each group, an average predictive ability can be calculated. Table II contains the results of the predictive ability studies using all 51 substructure descriptors.

The training set which gave the highest predictive ability in each group using all 51 descriptors was subsequently used for feature selection. The number of substructure descriptors remaining after feature selection and their average predictive abilities for each group's training sets

Table II. Predictive Ability Studies Using Only the 51 Substructure Descriptors

Group	Initial[a] predictive ability	Final predictive ability	Final no. of descriptors
A	95.1	96.2	14
B	94.6	95.3	15
C	94.6	95.8	16
D	94.6	95.1	14

[a] Fifty-one substructure descriptors used in each pattern vector.

are also given in Table II. In each case at least two-thirds of the initial 51 descriptors were removed during feature selection while the average predictive ability increased slightly. In all cases the entire data set was linearly separable with the reduced number of descriptors. In Table III a listing of the substructure descriptors remaining after feature selection for each group of training and prediction sets is given. The asterisks indicate the substructures which were used to obtain the final predictive abilities listed in Table II. As can be seen, there is considerable overlap of descriptors retained from group to group with over one-third of these 27 descriptors being retained as important features in at least three-fourths of the groups. Since 24 of the 51 substructures were always excluded during the feature selection, it was assumed that they were not essential for the separation of this data set. Therefore, they were not employed in any of the subsequent studies.

The 27 substructures found to be useful in the foregoing study were then combined with the ten fragment and seven geometric descriptors to form pattern vectors containing 44 descriptors each. As in the substructure study, the 80 training and prediction sets were again used to obtain an average predictive ability for each group. Feature selection was then carried out using the best training set in each group. The results of these predictive ability studies are given in Table IV. In each group about two-thirds of the 44 descriptors could be removed while maintaining a high predictive ability. Table V contains the list of descriptors retained during feature selection. (The substructure numbers in this table correspond to the index numbers used in Table III.) As can be seen, some of the substructure descriptors found to be useful in the previous study were replaced by fragment and geometric descriptors with an increase of the predictive ability for each group resulting from these exchanges (compare results in Tables II and IV).

Although the descriptors found to be useful for each group could have been used to predict the class designation of unknown odorants, a better method is to develop a classifier on the basis of the entire data set. When feature selection was performed on all 300 compounds, the 13 descriptors listed in Table VI remained out of the initial 44 descriptors. Although linear separability was still maintained when descriptors two and five in Table VI were removed, the predictive ability decreased for each group indicating a removal of some information (see last two

Table III. Substructure Descriptors Retained during Feature Selection

Hydrogens in substructures					
	Search[a] type	Included in groups			
Substructure		A	B	C	D
1. $-CH_3$	S	*	*	*	*
2. $\overset{\cdots}{-}CH\overset{\cdots}{-}$	G	*	*	*	*
3. $-CH_2-$	G	*	*	*	*
4. $-CH_2-$	S	*	*	*	*
5. $-O-$	S	*	*	*	*
6. $-OH$	S		*	*	*
7. $\overset{\cdots}{-}\overset{\vert}{C}\overset{\cdots}{-}CH\overset{\cdots}{-}$	G	*	*	*	
8. $\overset{\cdots}{-}\overset{\Vert\vdots}{C}-\overset{\vdots\Vert}{C}\overset{\cdots}{-}$	G	*	*		*
9. $-\overset{\Vert}{C}-$	S	*		*	*
10. $-\overset{\vert}{\underset{\vert}{C}}-$	S		*	*	*
11. $-CH_2\text{-}CH_3$	S		*		*
12. $\overset{\cdots}{-}\overset{\vert}{C}\overset{\cdots}{-}CH\overset{\cdots}{-}CH\overset{\cdots}{-}$	G		*	*	
13. $-\overset{\vert}{\underset{\vert}{C}}-CH-$	G	*			*
14. $-\overset{\vert}{\underset{\vert}{C}}-CH_3$	G	*			*
15. $-(CH_2)_4-$	G	*		*	
16. $CH_3\text{-}\overset{\vert}{\underset{\vert}{C}}\text{-}CH_3$	G		*	*	
17. $-CH_2\text{-}O-$	G		*		*
18. $\overset{\cdots}{-}\overset{\vert}{C}\overset{\cdots}{-}\overset{\vert}{C}\overset{\cdots}{-}$	G			*	
19. $-(CH_2)_2-$	S	*			
20. $-\overset{\Vert}{C}-$	G			*	
21. $-(CH_2)_3-$	G	*			
22. $-CH_2\text{-}\overset{\Vert}{C}-$	G				*
23. $-O-$	G	*			
24. $CH_3\text{-}\overset{\vert}{\underset{\vert}{C}}\text{-}CH_3$	S	*			
25. $-\overset{\Vert}{C}\text{-}\overset{\vert}{C}H_3-$	G		*		
26. $-CH_2\text{-}\overset{\vert}{\underset{\vert}{C}}-$	G		*		
27. $-CH_2\text{-}\overset{\vert}{C}H\text{-}CH_3$	G			*	

[a] G = general search, i.e., the substructure unit can appear anywhere in the molecule and a match will be made. S = specific search, i.e., the substructure unit can be matched only with acyclic atoms. ($\overset{\cdots}{-}$) is an aromatic bond type.

Table IV. Predictive Ability Studies Using the 44 Combined Descriptors

Group	Initial[a] predictive ability	Final predictive ability	Final no. of descriptors	Final[b] 13 descriptors	11[c] descriptors
A	95.3	96.6	16	97.5	96.4
B	96.1	95.9	15	97.6	96.8
C	95.6	96.4	15	96.8	96.6
D	95.7	97.1	14	97.8	96.7

[a] The 27 substructure descriptors plus the seven geometric and the ten fragment descriptors were included in each pattern vector. [b] All 13 descriptors listed in Table VI were used in each pattern vector. [c] All descriptors in Table VI except for numbers 2 and 5 were included in each pattern vector.

Table V. Descriptors Retained during Feature Selection of the 44 Combined Descriptors

Descriptor[a]	Included in groups			
	A	B	C	D
1. Number of oxygen atoms	*	*	*	*
2. Substructure number 3	*	*	*	*
3. Substructure number 9	*	*	*	*
4. Substructure number 5	*	*	*	*
5. *X* moment of rotation	*	*	*	*
6. Number of single bonds	*	*		*
7. Number of double bonds		*	*	*
8. Substructure number 1	*		*	*
9. Substructure number 8	*	*		*
10. Substructure number 15	*		*	*
11. Substructure number 27		*	*	*
12. Number of aromatic bonds		*		*
13. Substructure number 6		*	*	
14. Substructure number 18	*		*	
15. Substructure number 13		*	*	
16. Substructure number 23	*	*		
17. *Y* moment of rotation	*		*	
18. Number of carbon atoms			*	
19. Substructure number 2	*			
20. Substructure number 12	*			
21. Substructure number 21				*
22. Substructure number 26	*			
23. Substructure number 25	*			
24. Substructure number 16				*
25. Substructure number 17		*		
26. *X* moment/*Z* moment		*		
27. *Y* moment/*Z* moment			*	

[a] The substructure numbers used in this table correspond to the index numbers in Table III.

Table VI. 13 Descriptors Remaining after Feature Selection Using the Entire Data Set

Classification % correct	Descriptor[a]
1. 84.3	Total number of oxygen atoms/molecule
2. 82.3	Total number of double bonds/molecule
3. 80.0	Total number of aromatic bonds/molecule
4. 86.7	Longest principal moment of rotation
5. 80.0	Shortest principal moment of rotation
6. 80.0	$\cdots\overset{\vert}{C}\cdots CH\cdots CH\cdots$
7. 80.0	$CH_3-\overset{\vert}{\underset{\vert}{C}}-CH_3$
8. 86.0	$\cdots\overset{\vert}{C}\cdots\overset{\vdots}{C}\cdots$
9. 90.3	$-CH_2-$
10. 80.7	$-CH_3$
11. 80.0	$-\overset{\Vert}{C}-$
12. 80.0	$-O-$
13. 83.0	$-\overset{\vert}{C}H-$

[a] $(\cdots)$ indicates an aromatic bond.

columns in Table IV). Included in Table VI are the predictive abilities for each descriptor alone to classify the entire data set. These percentages should be compared to 80% which would be obtained by classifying the entire data set as nonmusks. The best single predictor of this list of features was the methylene substructure (feature 9, Table VI), which reflects the fact that macrocyclic musks contain a larger number of these substructure units.

To further test the predictive ability of these descriptors listed in Table VI and the decision surface associated with them, nine previously unused musk odorants were tested in the classifier. The odorants, shown in Figure 2, were entered into the ADAPT system and pattern vectors incorporating only the best 13 descriptors were generated for each compound. After preprocessing, these nine unknowns were classified as musk or nonmusk using the discriminant function trained using the entire data set of 300 odorants. All nine compounds were correctly classified as musk odorants. The correct classification of the five nitro musks and one macrocyclic musk in the data set of unknowns was expected since the training set contained structurally similar compounds. However, the correct prediction of the remaining three unknowns was most interesting since these were new structural types never used in training the discriminant function. Thus, the classifier was able to recognize new categories of musk odorants on the basis of a few molecular parameters which were derived from musk odorants of different structural types. Therefore, it appears that these parameters reflect the molecular properties which are common among musk odorants.

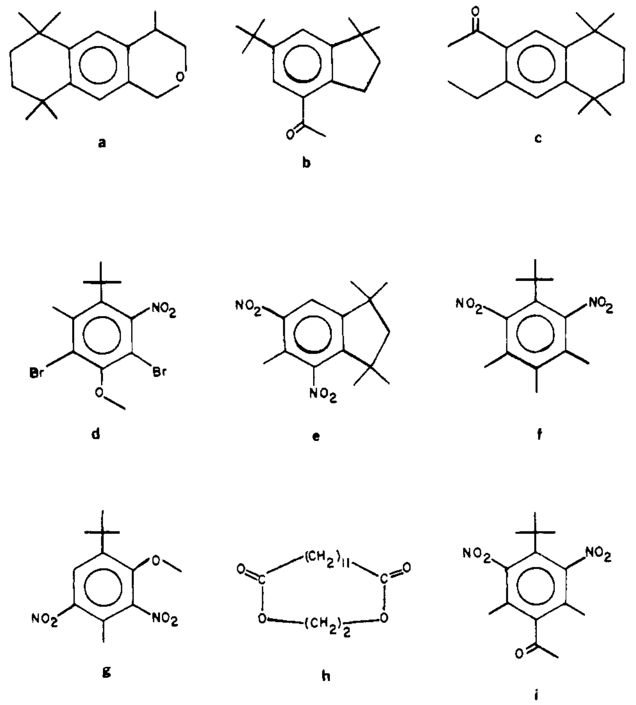

Figure 2. The nine musk compounds which were used as unknowns to test the best classifier obtained from these studies: (a) musk 89, (b) celestolide, (c) versalide, (d) musk alpha, (e) moskene, (f) musk tibetine, (g) musk ambrette, (h) astratone, (i) musk ketone.

DISCUSSION AND CONCLUSIONS

Indeed, pattern recognition techniques can be used to extract important features from a large collection of parameters for a given class which have a common property. Advantages of using pattern recognition techniques include: it is suitable for multicomponent data, the methodology exists for feature extraction, multisource data can be used, and the progress being made in the study can easily be measured.

Although musk odorants were the only compounds studied in this work, other groups of compounds which have the same characteristic odor could be studied in an analogous manner yielding a series of discriminant functions which could be used to predict the odor class of a new compound. By employing computer generated descriptors exclusively, as was done in this study, it would be possible to use these discriminant functions as prescreens for new odorants proposed for synthesis.

The ability of the fragments and geometric descriptors alone to classify all but three weak musks indicates that both information about the compound's structural shape as well as its chemical composition are necessary to separate musk odorants from nonmusks. Therefore, it was not surprising that the substructure descriptors alone were able to linearly separate the entire data set since these descriptors encode both structural and chemical information. As would be expected, the best set of descriptors for predicting musk odorants was found to be a combination of all three types of descriptors (see Table VI). Although these 13 descriptors were found to have the highest predictive ability and were able to correctly classify the nine unknown musks, they should not be considered the optimum set of descriptors for musk odorants since they were found through a heuristic search of a limited set of molecular descriptors.

The results of these studies neither confirm nor disprove any theory of human olfaction since the descriptors found to be useful in this study could be interpreted as substantiating any of the present theories. For example, the geometric descriptors could be used to reinforce Amoore's molecular shape theory; whereas, the substructure descriptors could be used to substantiate Beets' molecular profile theory. However, these results do indicate that several molecular parameters are necessary to predict the odor quality of odorants rather than a single parameter.

Although the actual meaning of each molecular descriptor, found to be important in this study, is not clear, the fact that they do fall into two categories (i.e., chemical composition and geometric shape) indicates that there may be a two-step process involved in producing the musk odor. One hypothesis which can be made is that the chemical nature is measuring the ability of the compound to pass from the air phase to the site of interaction and that the geometric shape of the molecule determines how well it fits into a receptor site. However, more data is needed before this conjecture can receive adequate verification.

In conclusion, these studies have demonstrated that pattern recognition techniques are well suited for finding invariant properties among a large data set of compounds which have the same odor quality. As for musk odorants, molecular shape seems to be an important factor for the accurate prediction of musks, but it is by no means the only factor. As was shown, a few molecular parameters are capable of producing a good classifier for predicting musk odorants. Although the correct classification of nine unknown compounds is not a comprehensive test of the predictive ability, it does give a strong indication that the descriptors used in the classifier are reflecting molecular properties common among musk odorants. The fact that three new structural types were recognized in the prediction study lends weight to this conclusion. Further prediction studies are planned for the future.

Although musk odorants were the focus of this work, other odor qualities can be studied in a similar manner and will be the topic of later papers. By comparing the parameters found to be useful in the different odorant class studies, trends may be found which may aid researchers in unlocking the mystery of olfaction.

ACKNOWLEDGMENT

The authors acknowledge the partial financial support of the National Science Foundation for the MODCOMP computer used in this project. We thank A. J. Stuper and G. S. Zander for their contributions toward the computerization of this research.

LITERATURE CITED

Amoore, J. E., "Molecular Basis of Odor", Charles C. Thomas, Springfield, Ill., 1970.

Beets, M. G. J., in "Molecular Structure and Organoleptic Quality", Society of Chemical Industry, Macmillian, New York, N.Y., 1957, pp 54–90.

Beets, M. G. J., in "Handbook of Sensory Physiology", Vol. IV, "Chemical Senses", Part 1, Olfaction, Beidler, L. M., Ed., Springer-Verlag, Berlin, 1971, pp 257–321.

Beets, M. G. J., in "Structure–Activity Relationships", Cavallito, C. J., Ed., Pergamon Press, New York, N.Y., 1973, pp 225–295.

Boelens, H., in "Structure–Activity Relationships in Chemoreception", Benz, G., Ed., Information Retrieval Limited, London, 1976, pp 197–206.

Brower, K. R., Schafer, R., *J. Chem. Educ.* **52**, 538–540 (1975).

Brugger, W. E., Jurs, P. C., *Anal. Chem.* **47**, 781–783 (1975).

Brugger, W. E., Stuper, A. J., Jurs, P. C., *J. Chem. Info. Comp. Sci.* **16**, 105–110 (1976).

Chu, K. C., Feldman, R. J., Shapiro, M. B., Hazard, G. F., Jr., Geran, R. I., *J. Med. Chem.* **18**, 539–545 (1975).

Dravnieks, A., *Ann. N.Y. Acad. Sci.* **237**, 144–163 (1974).

Dravnieks, A., Laffort, P., in "Olfaction and Taste", Vol. IV, Schneider, D., Ed., Wissens-Verlag-MBH, Stuttgart, Germany, 1972, pp 142–148.

Henning, H., *Z. Psychol.* **73**, 161–257 (1915).

Hiller, S. A., Golender, V. E., Rosenblit, A. B., *Comp. Biomed. Res.* **6**, 411–421 (1973).

Isenhour, T. L., Kowalski, B. R., Jurs, P. C., *Crit. Rev. Anal. Chem.* **4**, 1–44 (1974).

Jurs, P. C., Isenhour, T. L., "Chemical Applications of Pattern Recognition", Wiley-Interscience, New York, N.Y., 1975.

Kowalski, B. R., *Anal. Chem.* **47**, 1152A–1162A (1975).

Nilsson, N. J., "Learning Machines", McGraw-Hill, New York, N.Y., 1965.

Schiffman, S. S., *Science* **185**, 112–117 (1974).

Stuper, A. J., Jurs, P. C., *J. Am. Chem. Soc.* **97**, 182–187 (1975).

Stuper, A. J., Jurs, P. C., *J. Chem. Info. Comp. Sci.* **16**, 99–105 (1976).

Theimer, E. T., Davies, J. T., *J. Agric. Food Chem.* **15**, 6–14 (1967).

Tou, J. T., Gonzales, R. C., "Pattern Recognition Techniques", Addison-Wesley, Reading, Mass., 1974.

Wright, R. H., *J. Appl. Chem.* **4**, 611–615 (1954).

Wright, R. H., Serenius, R. S. E., *J. Appl. Chem.* **4**, 615–621 (1954).

Wright, R. H., Burgess, R. E., *Nature (London)* **224**, 1033–1035 (1969).

Zander, G. S., Stuper, A. J., Jurs, P. C., *Anal. Chem.* **47**, 1085–1093 (1975).

Received for review October 7, 1976. Accepted June 1, 1977.