Articles

Pattern Recognition Used to Investigate Multivariate Data in Analytical Chemistry

PETER C. JURS

Pattern recognition and allied multivariate methods provide an approach to the interpretation of the multivariate data often encountered in analytical chemistry. Widely used methods include mapping and display, discriminant development, clustering, and modeling. Each has been applied to a variety of chemical problems, and examples are given. The results of two recent studies are shown, a classification of subjects as normal or cystic fibrosis heterozygotes and simulation of chemical shifts of carbon-13 nuclear magnetic resonance spectra by linear model equations.

REVOLUTIONARY CHANGES HAVE TAKEN PLACE IN ANALYTICAL chemistry over the past 20 years. With sophisticated instrumentation—commonly under computer control—chemists can routinely gather great quantities of data that characterize the systems being investigated (*1*). Armed with these capabilities, analytical chemists have attacked ever more complex problems, for example, environmental monitoring and biological analyses. To analyze the data generated during such complex experiments, multivariate data analysis techniques must be used. Multivariate methods used by analytical chemists include pattern recognition, classification, discriminant analysis, clustering, modeling, and others.

Pattern recognition methods have been applied to a wide variety of chemical problems over the past 15 years, and a number of books (*2–4*) and reviews (*5–7*) have appeared. The biannual *Reviews* issue of *Analytical Chemistry* includes a review on chemometrics with a section on pattern recognition (*8*).

The general objective of pattern recognition studies is to predict an obscure property of an object (its origin or the class to which it belongs) on the basis of a set of indirect measurements. To be suitable, a data set defining a problem must fulfill the following conditions. The data set must be well designed. that is, it must be homogeneous and must not be dominated by extraneous effects. For example, experimental variations must be controlled or otherwise accounted for, and experimental artifacts must be minimized. Each class of interest must be well represented in the data set. For each object a number of variables must be available that are relevant to the classification problem at hand.

Data Representation

The data to be studied by pattern recognition methods are represented in a particular way—as points in a high-dimensional space. A single observation is represented as

$$\mathbf{X} = (x_1, x_2, \ldots, x_d)$$

where x_i is an individual variable, and d is the number of dimensions of the space and corresponds to the number of descriptors (variables or measurements) that are available for each object or experiment. Many types of chemical data can be represented in this way, and some examples from recent papers include (i) particulate samples of polluted air characterized by their trace metal concentrations, (ii) trace level concentrations of organic acid in human body fluids, and (iii) molecular structures of antitumor drugs. Data represented in this manner cannot be examined visually, so pattern recognition methods have been developed to investigate problems in this domain.

A basic assumption is that the degree of similarity between pairs of objects or experiments will be reflected by the proximity of the n-dimensional points representing those objects or experiments. The through-space distance between pairs of points is inversely related to their degree of similarity.

An aspect of data representation that is not strictly part of the pattern recognition analysis, but definitely affects the results obtained, is preprocessing and normalization of the data. The natural units for the variables involved in a pattern recognition analysis problem are often different. For example, one variable in an analysis could have a range of 0.2 to 2 (for example, the percentage by weight of carbon in steel) and another a range of 2 to 10 ppm (a trace constituent). The normalization operation called autoscaling is often used to convert each raw variable in a data set to a standardized variable with zero mean and unit variance. A recent paper presents a discussion of the effects of normalizations on pattern recognition analyses (*9*).

Pattern recognition studies often involve a set of data called the training set, a set of observations with known class memberships. The discriminants or predictive models are developed from the training set by what are called supervised methods. From the training set, relations that tie together the available variables and an obscure property of the observations are found. Additional data are used for assessing the predictive ability of the discriminants or models—a process called validation. Alternatively, some pattern recognition methods use a set of data without class membership labels; they seek relationships among the data directly by unsupervised methods, such as clustering.

Methods

Pattern recognition methods can be divided into the following major categories (*10*): mapping and display, discriminant development, clustering, and modeling.

Mapping and display. The multidimensional data under study can

The author is a professor in the Department of Chemistry, Pennsylvania State University, University Park, PA 16802.

be displayed as a graph for direct viewing. This allows the scientist to seek visual patterns in the simplified two-dimensional display. A number of useful and imaginative graphical techniques have been developed for direct display of multivariate data, for example, metroglyphs, linear and circular profiles, and Andrews plots (*11*). A circular profile consists of a polygonal line connecting the points located on evenly spaced rays, where the distance from the center represents the value for each of the variables. These display methods have not been widely used in chemistry to date.

An alternative strategy is to map the multidimensional data points onto a two-dimensional plane and then display the results. One common way for selecting a suitable plane is to use eigenanalysis to find the principal components (PC's) of the data set. The data set being transformed is represented by an $n \times d$ data matrix $\mathbf{X}$ with elements x_{ij}, where n is the number of objects and d is the number of variables. Let μ_k be the mean value for variable k. Then the individual elements (c_{jk}) of the $d \times d$ covariance matrix are defined as

$$c_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \mu_k)(x_{ij} - \mu_j)$$

The eigenvectors and eigenvalues are extracted from $\mathbf{C}$ by diagonalization. The two eigenvectors corresponding to the two largest eigenvalues represent the two orthogonal axes with greatest variance. These two PC's define a plane, which can be thought of as a window into the high-dimensional data space. The high-dimensional points are mapped onto the plane defined by the two PC's, and then the plane is displayed. The resulting plot is usually called a PC plot. Each point in the PC plot is a linear combination of the original variables. This mapping operation has been called the Karhunen-Loève transformation in the pattern recognition literature (*10*).

Nonlinear mapping methods develop a correspondence where each point in the original data set is mapped onto a point in a special two-dimensional plane. The interpoint distances in the two-dimensional plane are intended to mimic the interpoint distances in the original space, but such a mapping inevitably involves error. Operationally, nonlinear mapping is done by iteratively minimizing an error function by means of standard function-minimization techniques, for example, steepest descent.

Discriminant development. Many problems studied by pattern recognition involve category data. That is, each observation or event is tagged by its membership in a discrete category, for example, petroleum type, smelter that is the source of particulate emissions, or the clinical status of a patient. In such cases, pattern recognition methods can be used to examine the points of the data set as a whole to see whether they can be subdivided into meaningful categories. Placing a discriminant surface through the space and observing that members of one category are on one side separated from members of another category is one way to gain understanding of a data set.

Both parametric and nonparametric methods are used for the development of discriminants. Parametric pattern recognition methods use the mean vectors and covariance matrices (or other statistical measures) describing the members of the two classes as their basis for development of discriminants. An example of a parametric method for the development of discriminants is the Bayesian method, which is also known as linear discriminant analysis (LDA). The linear discriminant function developed from this approach is

$$s = \ln p_1 - \ln p_2 + \mathbf{X}'\mathbf{C}^{-1}(\mathbf{m}_1 - \mathbf{m}_2) - \mathbf{m}_1'\mathbf{C}^{-1}\mathbf{m}_1 + \mathbf{m}_2'\mathbf{C}^{-1}\mathbf{m}_2$$

where s is >0 for one class and s is <0 for the other class, $\mathbf{X}$ is the pattern being classified, p_k is the a priori probability for class k, $\mathbf{C}$ is the covariance matrix of the data set, and $\mathbf{m}_k$ is the mean vector for class k. In this approach a multivariate normal distribution for the data and equal covariance matrices for the members of the two classes are assumed.

Nonparametric pattern recognition programs develop their discriminants from the training set of patterns to be classified rather than from statistical measures of their distributions. Examples of nonparametric methods for the development of discriminants include error-correction feedback linear learning machines (perceptrons) (*12*), iterative least-squares methods (*3, 13*), and simplex optimization methods of searching for separating classification surfaces (*14*). Each of these methods searches for a separating discriminant by an iterative procedure designed to improve classification performance as experience increases. Error-correction feedback directly corrects errors when they are committed, whereas iterative least-squares and simplex methods minimize error functions iteratively.

The classification results obtained with linear discriminants are strongly affected by the ratio of the training set size, n, and the number of descriptors per observation, d. This point has been discussed (*10, 12*) and examined for real cases in recent papers (*15*). The probability of correctly classifying 100% of the members of a training set due to chance is low for $n/d > 3$, but substantial classification success above the random expectation of 50% can still be obtained. For example, for $n/d = 5$ the probability is one-half that 77% of the members of the training set will be correctly classified, as a result of chance alone. These results place limits on the problems that can be attempted by pattern recognition problems, and they provide measures by which classification results can be judged.

Clustering methods. Clustering methods are unsupervised, in that class labels are not used. These methods attempt to determine structural characteristics of a set of data by organizing the data into subgroups, clusters, or hierarchies. Hierarchical clustering is a widely used method, by which one measures the distances between all pairs of points, identifies the nearest pair, combines them into a new point midway between them, recalculates the distances from this new point to every other point in the data set, finds the new nearest pair, combines them, and so on, until all points have been linked. The resulting structure can be displayed as a dendrogram, which shows at what degree of similarity each pair of points was combined. Clustering methods have been applied to many types of data (*16*). These methods are exploratory and are used in seeking insights and suggestions contained in large data sets.

Modeling methods. The construction of mathematical models can be used for pattern recognition. The methods are closely allied to statistical modeling. A well-known method is soft independent modeling of class analogy (SIMCA) (*7, 17*), which models the members of each class separately in terms of PC's. Unknowns to be classified are fitted to the class models, and the classifications are made according to the goodness of the fits. An unknown pattern is assigned a probability for each class, and, if all probabilities are low, then the pattern may be an outlier in that it belongs to none of the original classes.

The K–nearest neighbor method is a direct classification scheme: an unknown pattern is assigned to the class to which the majority of its nearest neighbors belong.

Selected Applications of Pattern Recognition

Application studies of chemical problems based on the use of pattern recognition techniques have been reported in many areas (*2, 3, 5–7*). In this section I will summarize some areas of application and provide references to a sampling of the primary publications.

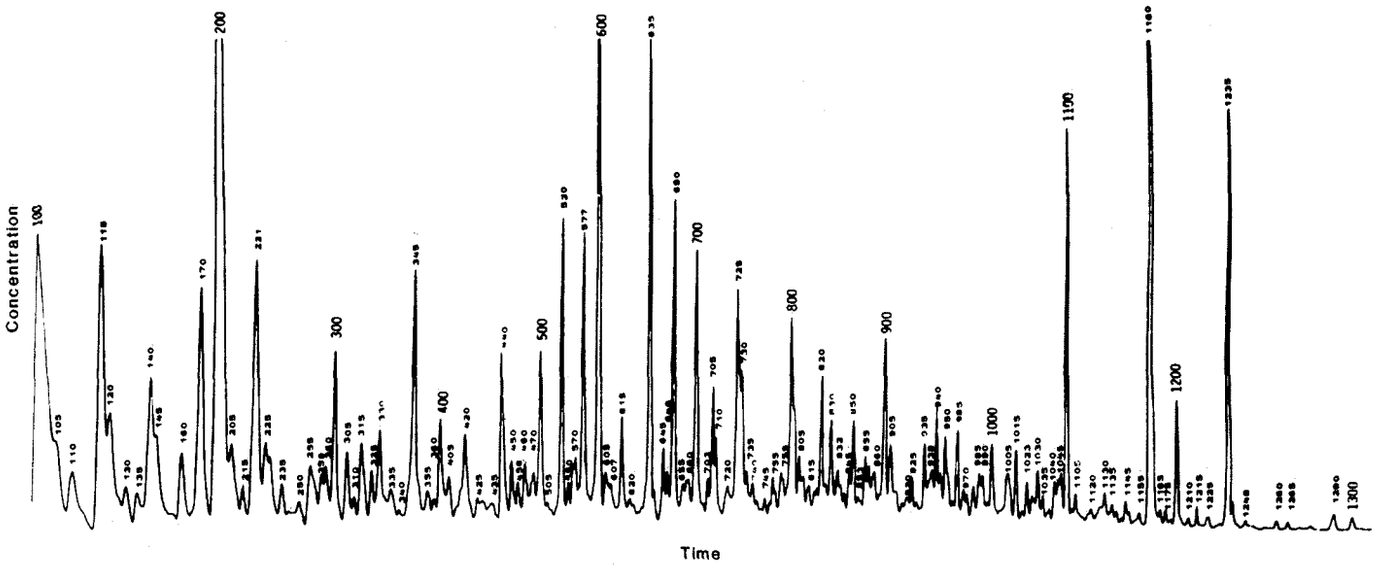

Fig. 1. Representative PyGC from the CF heterozygote study. The numbers marking each peak were assigned on the basis of the peak-matching software. The major peaks defining the regions are those with assignments that are multiples of 100. [Reprinted from (*33*) with permission of the American Chemical Society]

Spectral data analysis. The elucidation of chemical structure information from spectral data is a long-standing problem of chemistry. This is the area first studied and most intensively studied through the use of pattern recognition. Studies have been reported on mass spectra (*18*), infrared spectra (*19*), nuclear magnetic resonance (NMR) spectra (*20*), and electrochemical data (*21*).

Classification of complex mixtures. Materials or mixtures characterized by many measurements can be classified into categories, for example, origin, by pattern recognition methods. Examples of identification or classification problems drawn from diverse areas are found in the literature: manufacturers and grades of papers (*22*), quarry sites of archeological artifacts (*23*), sources of atmospheric particulate matter (*24*), classification of wines (*25*), determination of the origin of olive oil samples (*26*), identification of crude oil samples (*27*), selection of adsorbates for chemical sensor arrays (*28*), determination of the clinical status of patients from urine samples (*29*), classification of cancer cells (*30*), the study of acute lymphocytic leukemia (*31*), classification of human brain tissues (*32*), detection of cystic fibrosis heterozygotes (*33*), and the classification of bacteria (*34*).

Prediction of properties from molecular structure. Pattern recognition and associated multivariate methods can be used to predict physicochemical properties of compounds. Structural descriptors used can be physicochemical parameters or calculated structural descriptors. A few representative structure-property studies are as follows: gas chromatographic retention indices (*35*), liquid chromatographic retention indices (*36*), and chemical shifts of ^{13}C NMR spectra (*37, 38*).

Molecular structure–biological activity relations. Studies of the application of pattern recognition to the problem of searching for relations between molecular structure and biological activity have been reported. A large fraction of this type of research is involved with the generation of appropriate descriptors from the molecular structures available. Areas of study include drug structure-activity relations (SAR), studies of chemical communicants (for example, olfactory stimulants), and studies of structure-toxicity relations. Early applications of pattern recognition to drug design have been reviewed by Kirschner and Kowalski (*39*); a book describing one approach to SAR research has appeared (*4*). A few representative SAR studies include a study of 200 drugs for anticancer activity (*40*), a study of 9-anilinoacridines for antitumor selectivity (*41*), studies of drugs of accepted therapeutic value (*42*), structure–carcinogenic potential (*43*), olfactory quality of organic compounds (*44*), and structure–carcinogenic potential of polycyclic aromatic hydrocarbons (*45*).

Cystic Fibrosis Heterozygotes Versus Normal Subjects

Profiling of complex biological materials with high-performance chromatographic methods is an active research area with a large and growing literature (*6, 29–31, 46*). Such chromatographic experiments often yield chemical profiles containing hundreds of constituents, which are chemical fingerprints of the complex samples. Analysis of such profiles depends on the use of multivariate methods, and pattern recognition techniques have been useful.

Pattern recognition methods have been used to distinguish between individuals in a particular diseased state and normal individuals (*29, 31, 32*) on the basis of fingerprint chromatographic data. By these methods the researcher attempts to classify a sample according to a specific property (for example, diabetic versus normal) from measurements that are indirectly related to that property. An empirical relation is then derived from a set of data for which the property of interest is known and the measurements are available (a training set). Such a relation or classification rule may be used to infer the presence or absence of this property in objects that are not part of the original training set.

One recently reported study involved the application of pyrolysis gas chromatography and pattern recognition methods to the problem of identifying carriers of the cystic fibrosis (CF) defect (*6, 33, 47*). The biological samples used in this experiment were 48 cultured skin fibroblasts grown from 24 samples obtained from parents of children with CF and from 24 presumed normal donors. The pyrolyzed fibroblasts were analyzed in triplicate on fused silica capillary columns with temperature programming, typically yielding pyrochromatograms (PyGC's) with more than 150 resolved peaks (Fig. 1). Each PyGC was normalized on the basis of the total area of its peaks.

The 144 PyGC's were standardized and peak-matched by means

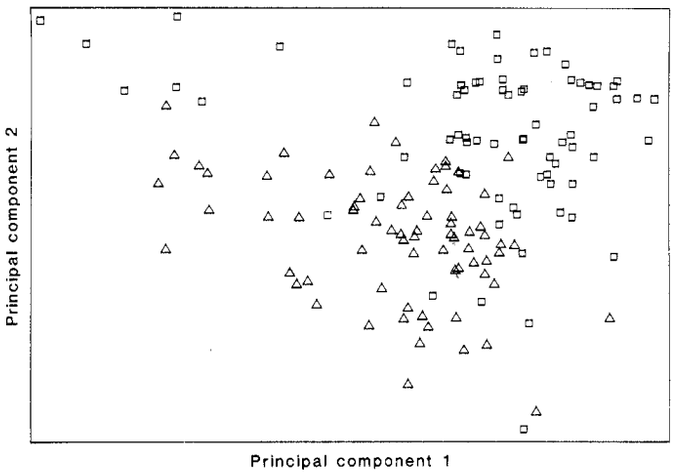

Fig. 2. Principal components (PC) plot of the 144 PyGC's as represented by six peaks. The squares represent the CF heterozygotes, and the triangles represent the normal subjects. Both PC1 and PC2 are linear combinations of the original six PyGC peaks with coefficients determined by eigenanalysis of the covariance matrix of the original data. The two PC's represent 59% of the total variance of the data. Some separation is observed but considerable overlap is present.

of an interactive computer program (*47*). Each PyGC was divided into 12 intervals bounded by 13 prominent peaks that were present in every PyGC. The retention times of the peaks within the intervals were scaled linearly for best fit with respect to a reference PyGC. The peak-matching procedure yielded 214 standardized retention time windows. The set of chromatographic data was autoscaled so that each peak had a mean of zero and a SD = 1 within the entire set of PyGC's. The training set thus consisted of 144 PyGC's of 214 peaks each, a 144 × 214 data matrix. Our objective was to use pattern recognition analysis to uncover relations buried in this mass of data that would support separation of the CF heterozygotes from normal subjects.

To apply pattern recognition methods to this overdetermined data set, it was first necessary to select a feature by objective means that would reduce the number of peaks per PyGC to reduce the probability of separation due to chance and that would render the analysis tractable. The data set contains 48 independent PyGC's, so we analyzed at one time 16 peaks at most to keep the probability of separation due to chance at a low level.

Experimental variables (cell culture, batch number, passage number, donor gender, and chromatographic column identity) contributed to the overall classification process. For example, one discriminant function was developed on the basis of only the 12 peaks contained in interval three. The CF PyGC's were completely separable from the PyGC's of the presumed normal donors. However, when the points from this 12-dimensional space were mapped onto a plane that best represented the pattern space (the plane defined by the two largest PC's), we observed groupings related to chromatographic column identity. Furthermore, classifiers developed from these 12 peaks yielded favorable classification results for several of the experimental variables. The information pertinent to the pathological alteration characteristic of CF heterozygotes must be isolated from the large amount of qualitative and quantitative data resulting from experimental conditions that is also contained in the complex capillary PyGC's.

We identified a set of six PyGC peaks that separated the PyGC's of CF heterozygotes from those of presumed normal subjects on the basis of valid chemical differences. Then we analyzed the 65 peaks that were present in at least 90% of the PyGC's. We assessed the usefulness of each of these 65 peaks alone for discrimination between PyGC's with respect to gender, passage number, and column identity. We selected for further analysis 12 peaks that had better classification success rates for CF versus normal subjects than for any other dichotomy. This procedure identified those peaks containing the most information about CF versus normal subjects as opposed to the extraneous experimental variables. A classification rule developed from these 12 peaks by means of the *K*–nearest neighbor procedure correctly classified 90% of the PyGC's in the data set. We used variance feature selection (*3*) combined with the linear learning machine and the adaptive least-squares methods (*3, 13*) to remove six of the peaks found to be least relevant to the classification problem. A final set of six PyGC peaks remained.

When we evaluated the ability of each of the six peaks individually to differentiate between the PyGC's of CF and normal subjects, classification success rates ranged from 59 to 80%. A mapping of the six-dimensional space was done by means of the Karhunen-Loève transform described above, and Fig. 2 shows a plot of the two PC's. While some tendency toward separation is evident, there is considerable overlap between the classes. Classifiers were developed by several routines for parametric and nonparametric discriminant generation. The iterative least-squares method generated a discriminant that misclassified only eight of the PyGC's (136 correct of 144, 94%). A histogram shows the distances from the discriminant to each of the 144 PyGC's and a scatter plot of these distances against the sequence numbers of the sample (Fig. 3). The discriminant divided the six-dimensional space into two regions, with 136 of the 144 PyGc's located on the correct side.

We assessed the contribution of the experimental parameters to the overall dichotomization power of the decision function based on the six peaks by reordering experiments. The set of PyGC's was first

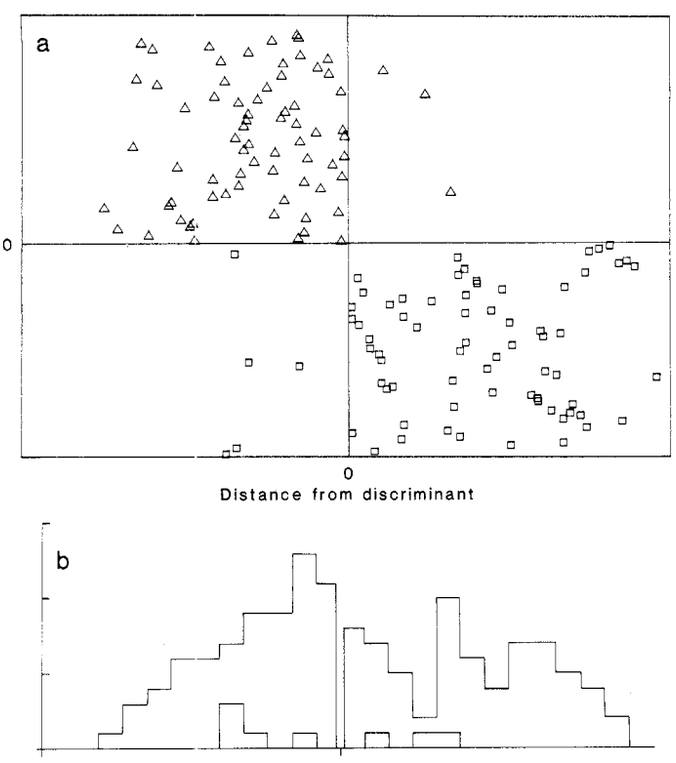

Fig. 3. (a) Plot of the distance between each of the 144 PyGC's (as represented by six peaks) and the separating discriminant (*x*-axis), against the sequence number of the PyGC (*y*-axis). The squares represent the CF heterozygotes and the triangles represent the normal subjects. Five CF heterozygote PyGC's are misclassified and three from normal subjects are misclassified. (b) Histogram showing the distribution of the distances. The small areas represent the eight misclassified PyGC's.

reordered in terms of donor gender, and the classification results obtained were indistinguishable from random. When we did similar studies for passage number and column identity, comparable results were obtained. The results of the reordering tests suggested that the decision function based on the six peaks incorporated mainly chemical information to separate the PyGC's of the CF heterozygotes from those of the normal subjects.

We tested the ability of the decision function to classify a simulated unknown sample by a procedure known as internal validation. Twelve sets of PyGC's were chosen by random selection where the training set contained 44 triplicates and the validation set contained the remaining four triplicates (held-out set). Any particular triplicate was present in only one validation set of the 12 generated. Discriminants developed for the training sets were tested on the PyGC's that were held out. The average correct classification percentage for the held-out PyGC's was 87%. This same internal validation test was repeated except that members of the held-out sets included triplicate samples analyzed on the same column or grown in the same batch of growth medium. The average correct classification for the held-out PyGC's in this set of runs was 82%. Although the classification success rate of the decision function was diminished, we obtained favorable results.

The set of 144 PyGC's was also studied by other pattern recognition methods (*6, 33*). It was shown that the PyGC's did indeed contain a great deal of information relevant to separating normal subjects from CF heterozygotes.

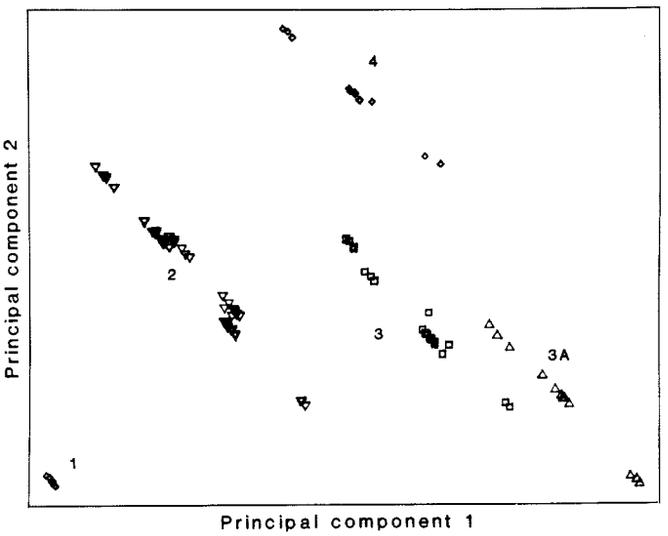

Fig. 4. Principal components plot of the 470 unique carbon centers contained in the 31 hydroxy steroids. Five clusters of carbon centers are observed. The points in clusters 1, 2, 3, 3A, and 4 correspond to primary, secondary, tertiary, tertiary with attached hydroxyl, and quaternary carbon centers, respectively.

Carbon-13 NMR Chemical Shifts

Carbon-13 NMR spectroscopy is one of the most widely used methods for organic structure elucidation, because it provides direct information about the skeletal carbon atoms in a molecule. The observed chemical shifts of the peaks are directly related to the environments of the carbon atoms, and each carbon center produces a peak. The development of computer-assisted methods for the interpretation of ^{13}C NMR spectra is an active area of research (*48*). Methods for file searching (*49*), spectral interpretation (*50*), chemometrics approaches (*51*), and additivity of contributions (*37, 38*) have appeared recently. Empirical studies in this area are justified because theoretical methods have not yet fully elucidated the origin of chemical shifts.

We have reported a project to simulate ^{13}C NMR shift (*38*). This research focuses on the simulation of ^{13}C NMR chemical shifts by linear model equations that relate the molecular environments of carbon centers to their observed chemical shifts. The linear models have the following form

$$S = b_0 + b_1x_1 + b_2x_2 + \ldots + b_px_p$$

where S is the predicted chemical shift of a given carbon center, x_i are numerical descriptors, derived directly from the molecular structure, that encode structural features of the chemical environment of the carbon atom of interest, b_i are coefficients determined through multiple linear regression analysis of a set of experimentally observed chemical shifts from compounds with assigned spectra, and p is the number of descriptors contained in the model.

The initial efforts focused on designing, implementing, and testing the system, which provides the user with the capability (i) to enter and to store the structures of compounds and their associated ^{13}C NMR spectra, and (ii) to build three-dimensional models of the compounds based on molecular mechanics. One can (iii) perceive unique and similar carbon centers within the compounds to predict the number of expected ^{13}C NMR peaks, and (iv) generate a wide variety of sophisticated carbon-center descriptors derived from the topology and geometry of the structures. From that basis, it is possible (v) to build and to evaluate linear predictive equations by means of multiple linear regression analysis and other multivariate statistical methods, (vi) to search spectral libraries to evaluate the quality of predicted spectra compared to authentic spectra, (vii) to store and to manipulate models derived from different sets of compounds, and (viii) to choose which of many stored models to use for predicting the ^{13}C NMR chemical shifts of the unique carbon centers in an unknown compound. The overall goal of the project is the development of a minicomputer-based interactive system for the simulation of ^{13}C NMR spectra of high quality for a variety of molecular structural classes.

Before a model is developed, the carbon centers in a set of compounds under investigation that are unique must be selected to facilitate removal of duplicate carbon centers that would unduly affect the statistical model building. We reported a practical method for assessing similarity (*52*) where the surroundings of each carbon center were represented by a six-dimensional vector whose terms represented the effects of the environment within five bonds of the carbon center. Each term was derived from known effects of structural moieties on ^{13}C NMR shifts in simple molecules. Since six variables cannot be viewed directly, we used PC plots, which showed strong clustering among carbon centers. This observation, coupled with other experiments, demonstrated the effectiveness of the similarity assessment method.

In structure-property studies such as this one, a major focus of the work is the development of relevant structural descriptors, here for individual carbon centers. Several classes of descriptors have been developed: (i) simple topological descriptors include nearest neighbor counts, valency counts, and connectivity indices, all derived solely from the topology of the structures. (ii) Topological electronic descriptors include partial charges from sigma electrons on the carbon center of interest and nearby atoms. (iii) Geometrical descriptors are derived from the three-dimensional molecular models, and they include counts of atoms within specified radial distances from the carbon of interest, energy parameters related to conformational strain energy, and distances from the carbon center to other nearby atoms. Details regarding the calculation of these descriptors have appeared (*53*).

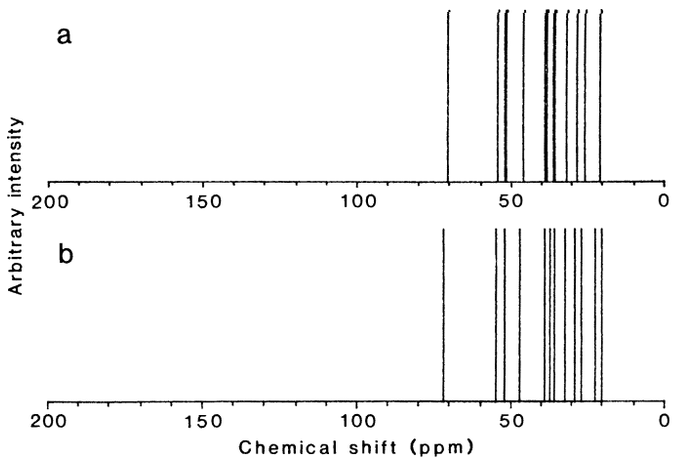

Fig. 5. Simulated (a) and observed (b) ^{13}C NMR spectra of 5α-androstan-16-α-ol. The standard error of the simulated spectrum is 0.885 ppm. There are 15 peaks in each spectrum, and some pairs of peaks are not distinguishably different on the scale of this plot.

The computer software system was used to develop linear predictive models for a set of 31 hydroxy steroids (*54*). The 31 structures contained 589 carbon centers in all (neglecting alkyl side chains). When the carbon centers were each described by the six-dimensional representations of their surroundings and then displayed in a PC plot, Fig. 4 resulted. It showed the carbon centers forming five band-shaped regions. After elimination of duplicate carbon centers, we developed model equations for the sets of carbon centers comprising each band in Fig. 4: primary carbon centers (n = 48, p = 4, r = 0.973, SE = 0.74), secondary carbon centers (n = 224, p = 7, r = 0.990, SE = 1.07), tertiary carbon centers (n = 120, p = 5, r = 0.994, SE = 0.97), tertiary carbons with attached hydroxyls (n = 25, p = 4, r = 0.966, SE = 1.21), and quaternary carbon centers (n = 53, p = 6, r = 0.970, SE = 0.81), where n is the number of carbon centers used to derive each model, p is the number of terms in the equation, r is the multiple correlation coefficient, and SE is the standard error in parts per million. Full simulated spectra were generated from these equations and were compared with the experimentally observed spectra. Mean errors were in the 1-ppm range. As an example, Fig. 5 shows the simulated and observed spectra for 5α-androstan-16-α-ol. The spectra of compounds not used to derive the equations were also simulated, and similar results were obtained. When simulated spectra were used to search against libraries of authentic spectra, we found that the simulated spectra were excellent approximations to the actual spectra of these hydroxy steroids (*54*).

The sets of compounds studied to date were limited in their structural diversity, saturated, and had only the hydroxyl functional group (cyclohexanes, cyclohexanols, decalin alcohols, steroids, and alkyl chain–substituted analogs of them). The approach may be applied to more diverse molecular structures that are unsaturated or that contain additional functional groups. Pattern recognition may be used to investigate a set of substituted cyclopentanes and cyclopentanols to probe the applicability of these methods to structures less rigid than the six-membered ring systems studied previously. Another extension of this work could involve bicyclo and tricyclo compounds.

Conclusions

Pattern recognition methods are an effective way to investigate multivariate data in analytical chemical problems. Chemists can use such methods to study complex data, to seek obscure relationships for classifying objects or events into categories, or to build quantitative models for simulation.

REFERENCES AND NOTES

1. F. W. McLafferty, *Science* **226**, 251 (1984).
2. K. Varmuza, *Pattern Recognition in Chemistry* (Springer-Verlag, Berlin, 1980); B. R. Kowalski, Ed., *Chemometrics: Theory and Application* (American Chemical Society, Washington, DC, 1977); B. R. Kowalski, Ed., *Chemometrics, Mathematics, and Statistics in Chemistry* (Reidel, New York, 1984).
3. P. C. Jurs and T. L. Isenhour, *Chemical Applications of Pattern Recognition* (Wiley-Interscience, New York, 1975).
4. A. J. Stuper, W. E. Brugger, P. C. Jurs, *Computer Assisted Studies of Chemical Structure and Biological Function* (Wiley-Interscience, New York, 1979).
5. K. Varmuza, in *Computer Applications in Chemistry*, S. R. Heller and R. Potenzone, Jr., Eds. (Elsevier, Amsterdam, 1983); B. R. Kowalski and S. Wold, in *Handbook of Statistics*, P. R. Krishnaiah and L. N. Kanal, Eds. (North-Holland, Amsterdam, 1982), vol. 2; L. Kryger, *Talanta* **28**, 871 (1981).
6. A. M. Harper, in *Pyrolysis and GC in Polymer Analysis*, S. A. Liebman and E. J. Levy, Eds. (Dekker, New York, 1985).
7. S. Wold *et al.*, in *Chemometrics, Mathematics and Statistics in Chemistry*, B. R. Kowalski, Ed. (Reidel, New York, 1984).
8. B. R. Kowalski, *Anal. Chem.* **52**, 112R (1980); I. E. Frank and B. R. Kowalski, *ibid.* **54**, 232R (1982); M. F. Delaney, *ibid.* **56**, 261R (1984).
9. E. Johansson, S. Wold, K. Sjodin, *ibid.* **56**, 1685 (1984).
10. J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles* (Addison-Wesley, Reading, MA, 1974).
11. P. H. Wang, Ed., *Graphical Representation of Multivariate Data* (Academic Press, New York, 1978).
12. N. J. Nilsson, *Learning Machines* (McGraw-Hill, New York, 1965).
13. I. Moriguchi, K. Komatsu, Y. Matsushita, *J. Med. Chem.* **23**, 20 (1980).
14. G. L. Ritter, S. R. Lowry, C. L. Wilkins, T. L. Isenhour, *Anal. Chem.* **47**, 1951 (1975).
15. T. R. Stouch and P. C. Jurs, *J. Chem. Inf. Comput. Sci.* **25**, 45 (1985); *ibid.*, p. 92.
16. J. A. Hartigan, *Clustering Algorithms* (Wiley, New York, 1975); D. L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis* (Wiley-Interscience, New York, 1983).
17. S. Wold and M. Sjostrom, *Am. Chem. Soc. Symp. Ser.* **52**, 243 (1977).
18. D. R. Burgard, S. P. Perone, J. L. Wiebers, *Biochemistry* **16**, 1051 (1977); H. Rotter and K. Varmuza, *Anal. Chim. Acta* **103**, 61 (1978).
19. H. B. Woodruff and M. E. Munk, *Anal. Chim. Acta* **95**, 13 (1977); D. S. Frankel, *Anal. Chem.* **56**, 1011 (1984).
20. U. Edlund and S. Wold, *J. Magn. Reson.* **37**, 183 (1980); T. R. Brunner, C. L. Wilkins, R. C. Williams, P. J. McCombie, *Anal. Chem.* **47**, 662 (1975).
21. W. A. Byers, B. S. Freiser, S. P. Perone, *Anal. Chem.* **55**, 620 (1983).
22. D. L. Duewer and B. R. Kowalski, *ibid.* **47**, 526 (1975).
23. J. R. McGill and B. R. Kowalski, *Appl. Spectrosc.* **31**, 87 (1977).
24. P. D. Gaarenstroom, S. P. Perone, J. L. Moyers, *Environ. Sci. Technol.* **11**, 795 (1977).
25. W. O. Kwan and B. R. Kowalski, *Anal. Chim. Acta* **122**, 215 (1980).
26. M. Forina and E. Tiscornia, *Ann. Chim.* **72**, 143 (1982).
27. H. A. Clark and P. C. Jurs, *Anal. Chem.* **51**, 616 (1979).
28. W. P. Carey *et al.*, *ibid.* **58**, 149 (1986).
29. M. L. McConnell, G. Rhodes, U. Watson, M. Novotny, *J. Chromatogr.* **162**, 495 (1979); G. Rhodes, M. Miller, M. L. McConnell, M. Novotny, *Clin. Chem.* **27**, 580 (1981).
30. E. Jellum, I. Bjornson, R. Nesbakken, E. Johansson, S. Wold, *J. Chromatogr.* **217**, 231 (1981).
31. H. A. Scoble, J. L. Fasching, P. R. Brown, *Anal. Chim. Acta* **150**, 171 (1983).
32. S. Wold, E. Johansson, E. Jellum, I. Bjornson, R. Nesbakken, *ibid.* **133**, 251 (1981).
33. J. A. Pino, J. E. McMurry, P. C. Jurs, B. K. Lavine, A. M. Harper, *Anal. Chem.* **57**, 295 (1985).
34. H. Engman *et al.*, *J. Anal. Appl. Pyrolysis* **6**, 137 (1984).
35. L. Buydens, D. L. Massart, P. Geerlings, *Anal. Chem.* **55**, 738 (1983); A. Sabljic, *J. Chromatogr.* **319**, 1 (1985).
36. A. S. Cohen and E. Grushka, *J. Chromatogr.* **318**, 221 (1985); K. Jinno and K. Kawasaki, *ibid.* **316**, 1 (1984).
37. O. Submeijer, A. E. Wilson, G. R. Hays, *Org. Magn. Reson.* **22**, 459 (1984); H. N. Cheng and S. J. Ellingsen, *J. Chem. Inf. Comput. Sci.* **23**, 197 (1983).
38. G. W. Small, T. R. Stouch, P. C. Jurs, *Anal. Chem.* **56**, 2314 (1984).
39. G. L. Kirschner and B. R. Kowalski, in *Drug Design*, E. J. Ariens, Ed. (Academic Press, New York, 1979), vol. 8.
40. B. R. Kowalski and C. F. Bender, *J. Am. Chem. Soc.* **96**, 916 (1974).
41. D. R. Henry, P. C. Jurs, W. A. Denny, *J. Med. Chem.* **25**, 899 (1982).
42. G. K. Menon and A. Cammarata, *J. Pharm. Sci.* **66**, 304 (1977).
43. S. L. Rose and P. C. Jurs, *J. Med. Chem.* **25**, 769 (1982).
44. P. C. Jurs, C. L. Ham, W. E. Brugger, *Am. Chem. Soc. Symp. Ser.* **148**, 143 (1981).
45. B. Norden, U. Edlund, S. Wold, *Acta Chem. Scand.* **B32**, 602 (1978).
46. A. Zlatkis, R. S. Brazell, C. F. Poole, *Clin. Chem.* **27**, 789 (1981); E. J. Jellum, *J Chromatogr.* **143**, 427 (1977).
47. J. A. Pino, thesis, Cornell University (1984).
48. N. A. B. Gray, *Prog. Nucl. Magn. Reson. Spectrosc.* **15**, 201 (1982).
49. R. W. Bally *et al.*, *Anal. Chim. Acta* **157**, 227 (1984).
50. C. A. Shelley and M. E. Munk, *Anal. Chem.* **54**, 516 (1982); C. W. Crandell, N. A B. Gray, D. H. Smith, *J. Chem. Inf. Comput. Sci.* **22**, 48 (1982).
51. D. Johnels *et al.*, *J. Chem. Soc. Perkin Trans.* **2**, 863 (1983).
52. G. W. Small and P. C. Jurs, *Anal. Chem.* **56**, 1314 (1984).
53. ———, *ibid.* **55**, 1121 (1983).
54. ———, *ibid.* **56**, 2307 (1984).

55. The research described here was supported under NSF grant CHE8202620.