W. W. Moss,[1] Ph.D.; F. T. Posey,[2] M.S.; and P. C. Peterson,[3] Ph.D.

A Multivariate Analysis of the Infrared Spectra of Drugs of Abuse

REFERENCE: Moss, W. W., Posey, F. T., and Peterson, P. C., **"A Multivariate Analysis of the Infrared Spectra of Drugs of Abuse,"** *Journal of Forensic Sciences*, JFSCA, Vol. 25, No. 2, April 1980, pp. 304-313.

ABSTRACT: The infrared (IR) spectra of 15 drugs of abuse were analyzed for similarity by using techniques of numerical taxonomy. The study included six barbiturates (amobarbital, barbital, butabarbital, pentobarbital, phenobarbital, and secobarbital), four amphetamine-related compounds (amphetamine, ephedrine, methamphetamine, and phentermine), and five other drugs (cocaine, heroin, phencyclidine, phendimetrazine, and diazepam). Three character sets were based on increasing numbers (10, 24, and 36) of IR peaks. The cluster analysis, principal component analysis, and nonmetric multidimensional scaling elements of the program system NT-SYS were used to structure taxonomic distances between drugs. Best results were obtained from the 36-peak data set; ordination diagrams proved to be more visually informative than phenograms. Preliminary results from our analysis of this set of drugs indicate that an expanded multivariate approach to drug classification may be useful.

KEY WORDS: toxicology, spectroscopic analysis, multivariate analysis, classification

The number of drugs and drug combinations submitted to forensic science laboratories for analysis has increased rapidly during the past decade. The Drug Enforcement Agency of the U.S. Department of Justice now lists over 10 000 controlled drugs and drug combinations on the market [*1*]. In addition to controlled drugs, new clandestinely prepared drug mixtures appear on the streets each month.

Identification of these drugs has become an increasingly difficult analytical problem, eased somewhat by the development of new technology. Infrared (IR) spectrophotometry and gas chromatography/mass spectroscopy are useful tools for the characterization of drug samples, but not all laboratories possess the necessary facilities, such as complete reference files for standard drug spectra. Even when large numbers of spectra are available for comparison, a manual search of voluminous card files is time-consuming and tedious. Attempts have been made to develop simple, rapid systems for the identification of IR spectra of the most commonly encountered drugs [*2,3*]. Such systems involve the comparison of three to eight of the most intense absorption bands of an unknown spectrum with a table or reference spectra arranged by decreasing intensity of absorption bands.

Received for publication 11 May 1979; revised manuscript received 18 Oct. 1979; accepted for publication 19 Oct. 1979.

[1]Associate curator, Academy of Natural Sciences, Philadelphia, Pa.; presently, Systematic Biology Program, National Science Foundation, Washington, D.C.

[2]Director, Tri-State Laboratories, Inc., Youngstown, Ohio.

[3]Professor, Department of Biological Sciences, Youngstown State University, Youngstown, Ohio.

These systems are helpful, but they are often difficult to use and always require a final, visual comparison with a standard reference spectrum.

Automated file-searching programs have been developed for use with high-speed computers capable of rapidly searching files containing as many as 10 000 IR spectra coded on cards or magnetic tape [*4–16*]. The end result of this search is a match, near match, or correlation of an unknown with one or more reference compounds. These systems are oriented strictly toward providing an *identification* rather than a general-purpose *classification* [*17–20*].

The classification of drugs is a problem complementary to that of identification. To classify for general purposes one gathers as much data as possible from the objects (for example, drugs) to be classified. Evaluation of the data by operational techniques optimally yields a classification into groups (for example, barbiturates) whose members have characteristics in common. These groups or taxa should be robust enough to survive the discovery and incorporation of new data. Further, the jointly held features of such taxa should facilitate their subsequent identification. Greater ease of identification would be a significant benefit; identification should benefit most when the classification is based on a firm set of data.

The present study describes a system in which the IR spectra of numerous drugs were characterized not only for identification but also for classification, that is, to recognize and portray the relationships between several groups of more or less closely related drugs. We studied 15 commonly encountered drugs of abuse and used techniques of multivariate analysis originally developed for biological taxonomy [*17*]. Our goals were to determine the quantity of data necessary for an acceptable classification and to study the feasibility of using such an approach on a much larger scale.

Materials and Methods

Operational Taxonomic Units

The operational taxonomic units (OTUs) in this study were samples of 15 standard drugs. Six barbiturates (amobarbital, barbital, butabarbital, pentobarbital, phenobarbital, and secobarbital) were used; scans were prepared from free acids in KBr pellets. Four amphetamine-related compounds (amphetamine, ephedrine, methamphetamine, and phentermine) were prepared as the hydrochloride salts in KBr pellets. Four basic drugs (cocaine, heroin, phencyclidine [PCP], and phendimetrazine) were also prepared in KBr pellets as hydrochloride salts. The 15th drug, diazepam, was used as the free base in KBr.

The KBr pellets were prepared with Beckman RIIC 13-mm dye and scanned on a Beckman Model IR 4240 IR spectrophotometer at 600 wave numbers per minute. The standard drug or generic name and structural formula for each OTU are shown in Fig. 1.

Characters

Three sets of data with increasing numbers of characters were obtained for the 15 drugs by using their recorded IR spectra. Data Set 1 (the 10-peak study) consisted of the wave numbers for the ten most intense absorption bands between 2000 and 300 wave numbers. Data Set II (the 24-peak study) was obtained by subdividing the spectrum into five regions and then recording the most intense peaks in each region. Three peaks each were taken from Region 1 (4000–2000), Region 2 (1999–1500), and Region 5 (699–300); ten peaks were taken from Region 3 (1499–1200) and five peaks from Region 4 (1199–700). Data Set III (the 36-peak study) was obtained by subdividing the spectrum as in Data Set II and then adding eight additional peaks to Region 4 (1199–700) and four peaks to Region 5 (699–300). There were no missing elements in the data tables.

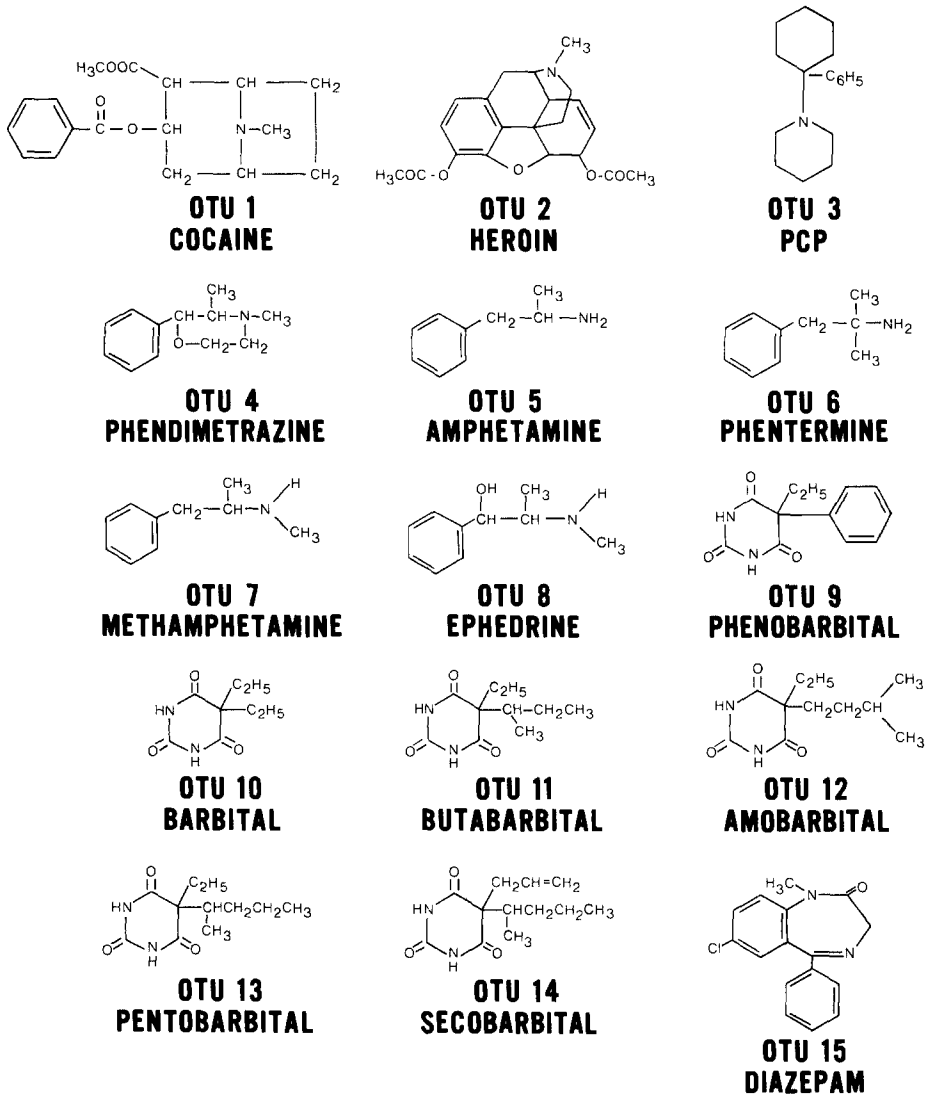


FIG. 1—*Structural formulas for the 15 standard drugs.*

Data Processing

Computations were carried out in the batch mode via a remote job-entry station to an IBM 370/168 VS2 computer by using the Stony Brook numerical taxonomy system NT-SYS (Version 3, Modification Level 2 [*1*]). Both Q- and R-mode analyses were applied to the data [*22*].

Character readings were standardized by variance [*17*]. For Q-mode analysis, both taxonomic distances and correlations were found and structured by the cluster analysis, minimum spanning tree (MST), and "subsets" components of NT-SYS. Cluster analysis was done by the unweighted pair group method using average linkage (UPGMA), the MST was done by the method described by Prim [*23*], and the subsets were formed by using the method of Sale [*24*]. Phenograms obtained from cluster analysis were dia-

grammed but used mainly to show general trends because of their inherent distortions. Clusters were designated by the method of Rohlf [*25*] in which three periods separate the first and last cluster members in the linear sequence in which they appear in the phenogram.

For R-mode analysis, correlations between characters were subjected to principal component analysis (PCA), with components extracted until eigenvalues became less than 1.0. Our choice of PCA over other techniques of factor analysis was determined by the results of Fisher [*26*], who found that PCA tended to find relatively few underlying factors of variation with high factor loadings and reasonable biological interpretation. A matrix of OTU projections in PCA space was obtained but used primarily as a starting point for nonmetric multidimensional scaling (MDS) [*27*]. Taking an initial configuration from a PCA solution helps to avoid both the possibility of MDS entrapment in local minima and the tendency for PCA illustrations to portray close-relative similarities less accurately than distant-relative ones [*28,29*]. The OTU configurations in MDS three-space were adjusted after scaling by performing a PCA on a variance-covariance matrix obtained from the MDS coordinates. This procedure realigns major trends of variation in the reduced configuration space with the coordinate axes while maintaining the accuracy of distances between OTUs in the ordination space [*21*]. Distances between OTUs in the MDS ordination space were compared with the Q-mode taxonomic distance matrix by using the matrix correlation coefficient. Two-dimensional ordination diagrams were constructed with superimposed MST and subsets connections, thus combining Q- and R-mode results in a visually informative manner.

Results

Data Set I

The phenogram based on ten readings (Fig. 2) includes clusters that are generally acceptable. The barbiturates join as a cohesive unit (OTUs 9 through 14), but this unit also includes heroin (OTU 2). The four amphetamines cluster together (OTUs 5 through 8), with the inclusion of phendimetrazine (OTU 4). The remaining drugs join the phenogram at relatively low levels. The inclusion of cocaine with the barbiturates seems undesirable; although one generally cannot place a great deal of faith in the lowest clustering levels of a phenogram, cocaine does link with amobarbital via the minimum spanning

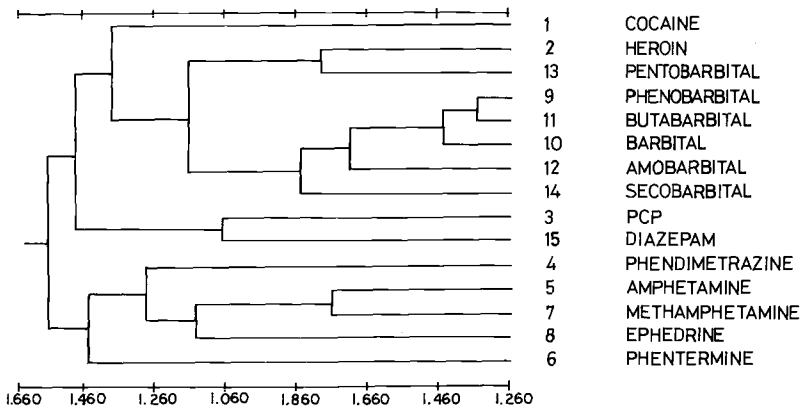


FIG. 2—*Phenogram based on cluster analysis of taxonomic distances from 10-peak infrared spectrum absorption bands of ephedrine hydrochloride. Matrix correlation is* $r = 0.844$.

tree (Fig. 3). The cluster formed by PCP and diazepam seems difficult to justify. One would expect the latter to cluster with the barbiturates (OTUs 9 through 14) on the basis of chemical structure.

The ordination diagram for this data set (Fig. 3) shows a generally good picture of barbiturate relationships, with the questionable inclusion of heroin (OTU 2) as the nearest neighbor of pentobarbital (OTU 13). The amphetamines (OTUs 5 through 8) are not as well shown, with phentermine (OTU 6) linking to PCP (OTU 3). The remaining drugs (OTUs 1, 3, 4, and 15) are shown as outliers.

Data Set II

The phenogram based on 24 spectral readings (Fig. 4) is a considerable improvement over Fig. 2. The barbiturates (OTUs 9 through 14) form a cluster that no longer includes heroin (OTU 2). Three of the amphetamines (OTUs 5 through 7) form a tight cluster, but ephedrine (OTU 8) clusters with phendimetrazine (OTU 4). Cocaine, heroin, diazepam, and PCP join the phenogram at low levels of similarity.

The ordination diagram (Fig. 5) indicates a tight cluster of barbiturates (OTUs 9 through 14) near the origin. Three of the amphetamines (OTUs 5 through 7) are on the right, with less satisfactory placement of ephedrine (OTU 8). As in the phenogram, the remaining OTUs are outliers. Diazepam (OTU 15) links with phendimetrazine (OTU 4) in closer proximity to the barbiturate cluster.

Data Set III

The phenogram based on 36 readings (Fig. 6) is generally comparable to that based on 24 (Fig. 4), but the UPGMA clusters are somewhat more diffuse. A barbiturate cluster is evident (OTUs 9 through 14), but it includes phendimetrazine. The same three amphetamines cluster (OTUs 5 through 7) but are again separated from ephedrine (OTU 8). Co-

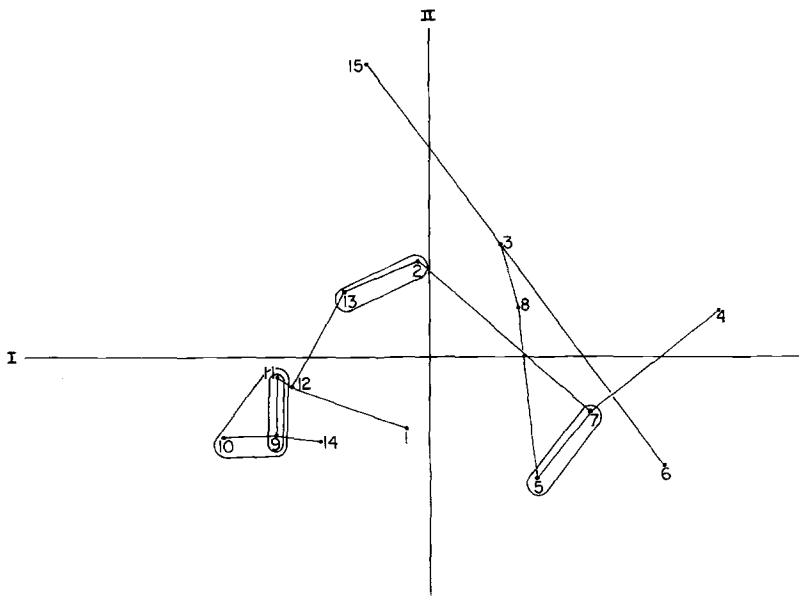


FIG. 3—*Ordination diagram based on nonmetric multidimensional scaling of taxonomic distances from 10-peak infrared spectrum absorption bands. Matrix correlation is* r = *0.877, stress* = *0.363.*

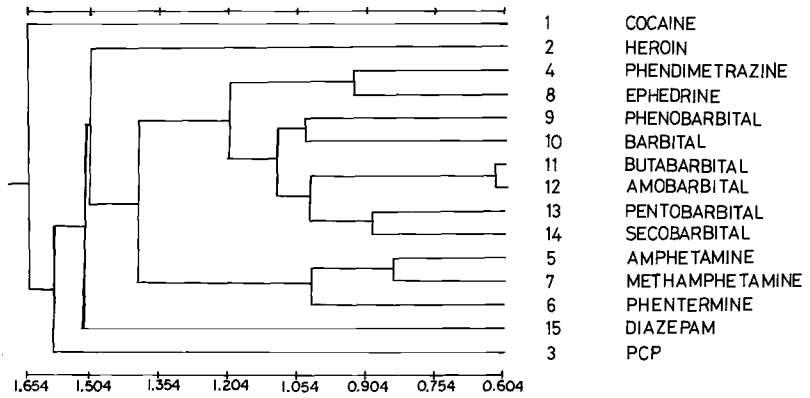


FIG. 4—*Phenogram obtained as in Fig. 2, based on 24-peak infrared spectrum absorption bands. Matrix correlation is* r = *0.876.*

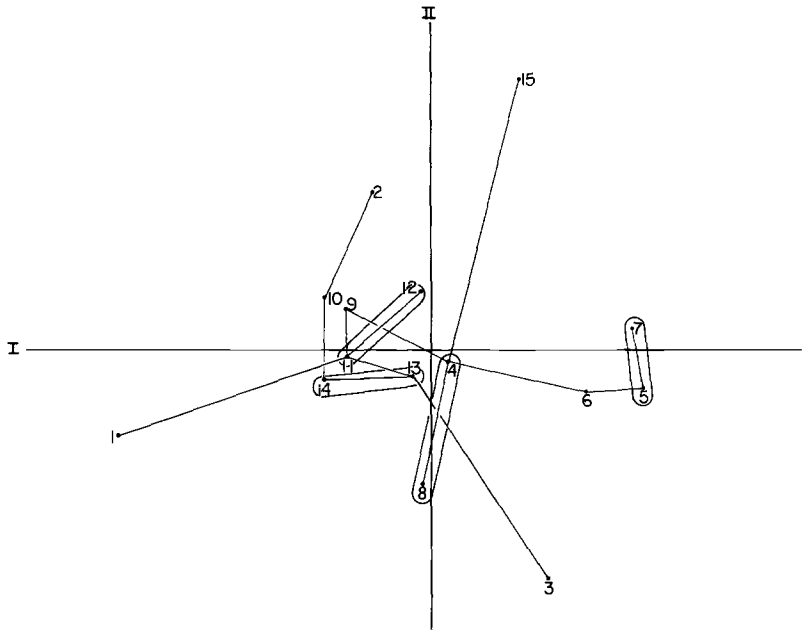


FIG. 5—*Ordination diagram obtained as in Fig. 3, based on 24-peak infrared spectrum absorption bands. Matrix correlation is* r = *0.914, stress* = *0.280.*

caine, heroin, and PCP join at the lowest level, while diazepam appears to be intermediate between the barbiturates and amphetamines.

The ordination diagram (Fig. 7) indicates a tight grouping of barbiturates (OTUs 9 through 14) near the origin. The amphetamines are placed toward the right of the diagram, as in Fig. 5. The outlying placement of cocaine, heroin, PCP, and diazepam is shown most clearly in Fig. 7; these drugs are shown to be roughly distant from all others in the study but more closely linked to the barbiturates than to the amphetamines. Diazepam now connects with amobarbital (OTU 12). The 36-peak phenogram (Fig. 6) depicts a somewhat surprising similarity between phendimetrazine and phenobarbital,

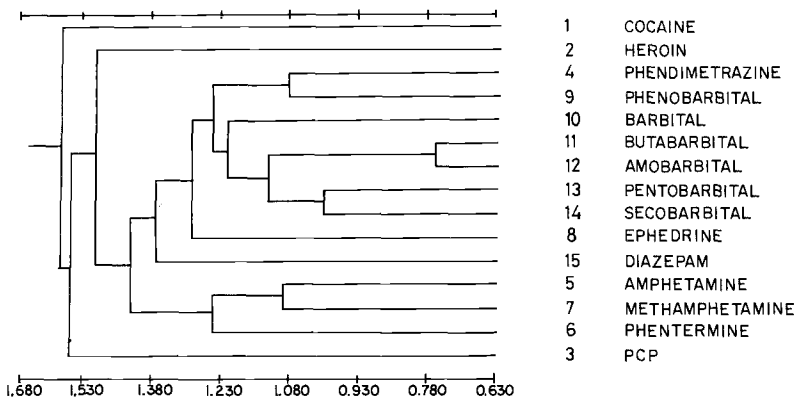


FIG. 6—*Phenogram obtained as in Fig. 2, based on 36-peak infrared spectrum absorption bands. Matrix correlation is* r = *0.855.*

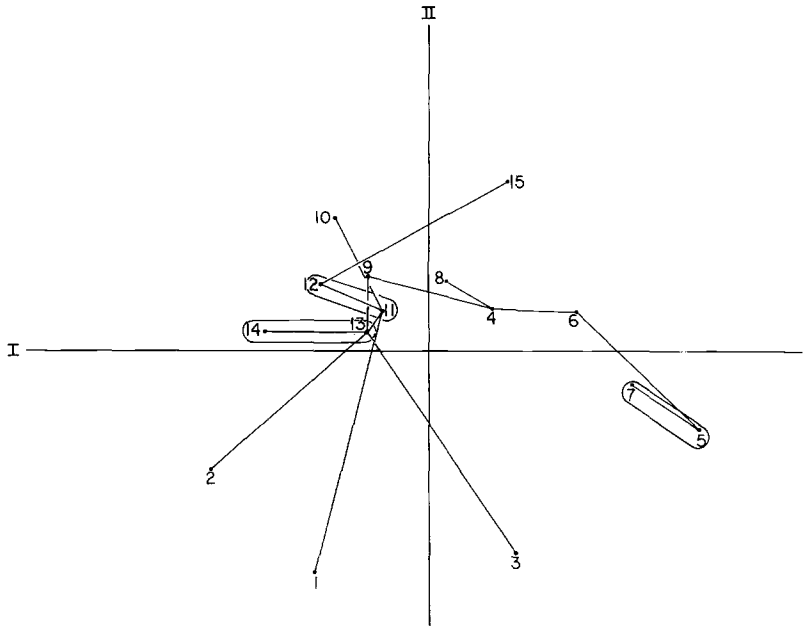


FIG. 7—*Ordination diagram obtained as in Fig. 3, based on 36-peak infrared spectrum absorption bands. Matrix correlation is* r = *0.860, stress* = *0.389.*

but the ordination diagram (Fig. 7) clearly places the former drug midway between the amphetamine and barbiturate groups.

Discussion

We have assumed that the currently recognized groups of barbiturates and amphetamines are valid. Our classification results were thus acceptable overall, with reasonably good definition of these two major drug groups, regardless of the number of characters used. However, a definite focusing of major groups and better separation of outliers

occurred in the ordination diagrams with an increase in the character set. It is possible that the incorporation of additional data might yield an additional improvement, although we suspect that the 36-peak approach may be adequate. Taxonomic distances based on 36 versus 24 peaks were correlated at $r = 0.898$.

The ordination diagrams yielded a better visual representation of results than did the phenograms. This was to be expected in view of the multidimensional nature of the data and the distortions to which phenograms are prone. However, it is often helpful to examine taxonomic distances by both ordination and cluster analysis [*22*].

The 10-peak data did not generate a classification entirely suitable to us. Only a small spectral area was sampled by using only 10 peaks and not subdividing the spectrum. This resulted in a poor heroin-phentobarbital cluster (Figs. 2 and 3) and a relatively diffuse definition of barbiturates and amphetamines. The sets of taxonomic distances based on 10 peaks were correlated at only $r = 0.425$ and $r = 0.380$ with distances based on 24 and 36 peaks, respectively. Principal component scores from the 24- and 36-peak studies showed that characters from the first three areas of the spectrum (4000–2000, 2000–1500, and 1500–1200) accounted for a large amount of variation among the drugs. Subdivision of the spectrum (24- and 36-peak studies) and the inclusion of the area between 4000 and 2000 wave numbers prevented any one area of the spectrum from being too heavily weighted in its contribution to the classification.

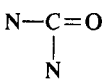
Twenty-four or, preferably, 36 peaks (Figs. 5 and 7) should be used to classify IR spectra in subsequent studies. Increasing the number of peaks between 1200 and 300 wave numbers produced an improvement in the results. Factor scores for characters in this area were low for the most part, indicating that they accounted for a small amount of variation between the samples; however, the extra effort required to collect twelve additional readings was justified by an improvement in visual representation of the 36-character results. It is likely that the increasing dimensionality of the distance matrix accounted for the drop in matrix correlation and increase in stress observed in progressing from the 24-character to the 36-character diagrams.

The time and effort required to record the locations of peaks manually and to punch these data on cards were significant components of the present study. It should be possible to connect the IR spectrophotometer to a device that will generate digitized or punched output suitable for direct read-in to the computer program used for classification [*14*]. The current limitation of NT-SYS to batch processing may require us in a subsequent study to process via punched paper tape from the reading device through a conversion program to disk or tape for storage, with subsequent output to cards for input to NT-SYS.

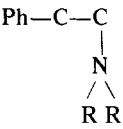
It is worth stressing that the aim of our study has been the *classification* of drugs rather than their *identification* (see Ref *17* for further discussion of this distinction). Better definition of the major areas of spectra to be emphasized in data capture and of the major clusters of drugs should lead us eventually to more efficient schemes of identification, including automated keys.

Correspondence between chemical structure and taxonomic placement of the drugs is noteworthy, especially as our approach involved observation of the entire IR spectrum rather than a restricted data base of "most informative" peaks. The eventual recognition of such peaks may allow us to pass from a structural to a functional study, for the greater part of a drug's structural formula seems to have little effect on its functional toxicity. If toxicologically active sites are differentially stimulated by IR, as seems likely, it should be possible to analyze and visualize toxic relationships between drugs by using the approach described.

The intermediate place of diazepam was noted above. This drug has the



group of OTUs 9 through 14, but it also has the extracyclic benzene ring of PCP and phendimetrazine. The latter in turn has a



arrangement as in OTUs 5 to 8, but in the 24- and 36-peak studies the possession of an αOH group apparently pulls OTUs 4 and 8 together. The different ways in which chemical structures may be drawn suggests that better information for data processing might be obtained from X-ray crystallographic studies.

The present study has shown the feasibility of classifying representatives from a few common drug groups, including amphetamines, barbiturates, and opiates. We consider it worthwhile to progress to a larger study involving a much larger set, with several examples from all of the important drug groups. Future investigations should also include mixtures prepared from two and three different drugs.

Acknowledgments

We thank Dr. F. J. Rohlf, SUNY Stony Brook, for making a copy of NT-SYS available. Thanks are extended also to Dr. Janet Del Bene, Department of Chemistry, Youngstown State University, and to Dr. James Mears and Daniel Bereza, Academy of Natural Sciences, for their helpful comments and suggestions.

References

[1] "Project Label," U.S. Department of Justice, Drug Enforcement Administration, Washington, D.C., 1976.

[2] Clarke, E. G. C., *Isolation and Identification of Drugs*, Vol. 1, The Pharmaceutical Press, London, 1969.

[3] "Analysis of Drugs Manual, Appendix 1, Infrared Search System," U.S. Department of Justice, Drug Enforcement Administration, Washington, D.C., 1973.

[4] Rogoff, M., "Automatic Analysis of Infrared Spectra," *Annals of the New York Academy of Sciences*, Vol. 69, 1957, pp. 27-37.

[5] Erley, D. S., "Fast Searching System for the ASTM Infrared Data File," *Analytical Chemistry*, Vol. 40, No. 6, May 1968, pp. 894-898.

[6] Sebesta, R. W. and Johnson, G. G., Jr., "New Computerized Infrared Substance Identification System," *Analytical Chemistry*, Vol. 44, No. 2, Feb. 1972, pp. 260-265.

[7] Tanabe, K., Saëki, S., and Tamura, T., "Computer Retrieval of Infrared Spectra," *Japan Analyst*, Vol. 23, No. 6, 1974, pp. 626-630.

[8] Tanabe, K. and Saëki, S., "Computer Retrieval of Infrared Spectra by a Correlation Coefficient Method," *Analytical Chemistry*, Vol. 47, No. 1, Jan. 1975, pp. 118-122.

[9] Mattson, J. S., "Computerized Infrared Spectroscopy," *Industrial Research*, Vol. 17, No. 2, Feb. 1975, pp. 56-59.

[10] Cubitt, J. M. and Wilkinson, J. G., "A Classification Study of Shales in Kansas Using Electron-Spin Resonance Spectroscopy," *Mathematical Geology*, Vol. 8, No. 3, 1976, pp. 337-344.

[11] Lewi, J., "Spectral Mapping, a Technique for Classifying Biological Activity Profiles of Chemical Compounds," *Arzneimittel-Forschung (Drug Research)*, Vol. 26, No. 7, Nov. 1976, pp. 1295-1300.

[12] Maugh, T. H., II, "Pyrolysis-Spectrometry: Automated Identification of Microbes," *Science* (Washington, D.C.), Vol. 194, No. 4272, 24 Dec. 1976, pp. 1403-1405, 1447.

[13] Milne, G. W. A. and Heller, S. R., "The MSDC/EPA/NIH Mass Spectral Search System," *American Laboratory*, Vol. 8, No. 9, June 1976, pp. 43-54.

[14] Sherman, R. E., "Computer-Assisted Infrared Spectral Identification," *American Laboratory*, Vol. 8, No. 9, Sept. 1976, pp. 91-95.

[15] Blaisdell, B. E., "Automatic Computer Construction, Maintenance, and Use of Specialized Joint Libraries of Mass-Spectra and Retention Indexes from Gas Chromatography/Mass Spectrometry Systems," *Analytical Chemistry*, Vol. 49, No. 1, Jan. 1977, pp. 180-186.
[16] Heller, S. R., Milne, G. W. A., and Feldmann, R. J., "A Computer-Based Chemical Information System," *Science* (Washington, D.C.), Vol. 195, No. 4275, 21 Jan. 1977, pp. 253-259.
[17] Sneath, P. H. and Sokal, R. R., *Numerical Taxonomy. The Principles and Practices of Numerical Classification*, Freeman & Co., San Francisco, 1973.
[18] Thornton, J. I., "The Use of an Agglomerative Numerical Technique in Physical Evidence Comparisons," *Journal of Forensic Sciences*, Vol. 20, No. 4, Oct. 1975, pp. 693-700.
[19] Rozett, R. W. and Petersen, E. M., "Classification of Compounds by Factor-Analysis of Their Mass Spectra," *Analytical Chemistry*, Vol. 48, No. 6, May 1976, pp. 817-825.
[20] Ritter, G. L. and Isenhour, T. L., "Minimal Spanning Tree Clustering of Gas Chromatographic Liquid Phases," *Computers and Chemistry*, Vol. 1, No. 3, 1977, pp. 145-153.
[21] Rohlf, F. J., Kishpaugh, J., and Kirk, D., "NT-SYS, Numerical Taxonomy System of Multivariate Statistical Programs," Technical Report, State University of New York, Stony Brook, 1972.
[22] Moss, W. W., Peterson, P. C., and Atyeo, W. T., "A Multivariate Assessment of Phenetic Relationships Within the Feather Mite Family Eustathiidae (Acari)," *Systematic Zoology*, Vol. 26, No. 4, Dec. 1977, pp. 386-409.
[23] Prim, R. C., "Shortest Connection Networks and Some Generalizations," *Bell System Technical Journal*, Vol. 36, No. 6, Nov. 1957, pp. 1389-1401.
[24] Sale, A. H. J., "Algorithm 65. An Improved Clustering Algorithm," *Computer Journal*, Vol. 14, 1971, pp. 104-106.
[25] Rohlf, F. J., "Classification of *Aedes* by Numerical Taxonomic Methods (Diptera: Culicidae)," *Annals of the Entomological Society of America*, Vol. 56, Nov. 1963, pp. 798-804.
[26] Fisher, D. R., "A Comparison of Various Techniques of Multiple Factor Analysis Applied to Biosystematic Data," *University of Kansas Science Bulletin*, Vol. 50, 1973, p. 162.
[27] Kruskal, J. B., "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis," *Psychometrika*, Vol. 29, No. 1, March 1964, pp. 1-27.
[28] Rohlf, F. J., "An Empirical Comparison of Three Ordination Techniques in Numerical Taxonomy," *Systematic Zoology*, Vol. 21, No. 3, Sept. 1972, pp. 271-280.
[29] Webster, R., "Intuition and Rational Choice in the Application of Mathematics to Soil Systematics," *Soil Science*, Vol. 119, No. 5, May 1975, pp. 394-404.

Address requests for reprints or additional information to
W. Wayne Moss
Systematic Biology Program
National Science Foundation
1800 G St. NW
Washington, D.C. 20550