(32) Levi, G. "A Note on the Derivation of Maximal Common Subgraphs of the Two Directed or Undirected Graphs". *Calcolo* **1972**, *9*, 341-352.
(33) Barrow, H. G.; Burstall, R. M. "Subgraph Isomorphism, Matching Relational Structures and Maximal Cliques". *Inf. Process. Lett.* **1976**, *4*, 83-84.
(34) Raznikov, V. V.; Talroze, V. L. "Automatic Generation of Complete Set of Structural Isomers With a Given Molecular Composition and Molecular Weight". *Zh. Struct. Khim.* **1970**, *11*, 357-360 (in Russian).
(35) Arlazarov, V. L.; Zuev, I. I.; Uskov, A. V.; Faradgev, I. A. "An Algorithm for Reduction of Finite Undirected Graphs to a Canonical Form". *Zh. Vych. Math. y Math. Phys.* **1974**, *14*, 737-743.
(36) Randič, M. "On Canonical Numbering of Atoms in a Molecule and Graph Isomorphism". *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 171-180.
(37) Golender, V. E.; Rozenblit, A. B. "Logico-Structural Approach to Computer-Assisted Drug Design". *Med. Chem. (Academic)* **1980**, *11*, 300-337.

Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures

GEORGE W. ADAMSON*† and DAVID BAWDEN‡

Postgraduate School of Librarianship and Information Science, Sheffield University, Sheffield, S10 2TN, England

Received April 13, 1981

Several hierarchical cluster analysis methods were applied to a set of benzenoid compounds by using structural features automatically derived from Wiswesser Line Notation. Comparisons of the differences in classification, due to choice of clustering algorithm and data standardization technique, were made.

INTRODUCTION

Cluster analysis, and similar techniques of numerical taxonomy, may be applied to descriptors of chemical structure to provide automatic classifications of sets of structures. Such classifications could be of value in information storage and retrieval, structure–property studies, and various areas of chemometrics.

Cluster analyses of sets of chemical substances of substituents have been shown to be of value in studying biological activity spectra[1,2] and selecting appropriate substituents for physicochemical property–biological activity studies.[3,4] These have, however, all used some molecular properties as variables in the clustering procedure.

Fewer examples have been reported of cluster analyses using variables directly representing chemical structure. Sneath described a classification of amino acids, based on both physical property and structural descriptor,[5] while Chu used augmented atom fragments in a structure–property study.[6] Adamson and Bush used various atom-and bond-centered fragments, automatically derived from connection tables, in clustering sets of amino acids[7] and diverse anaesthetic compounds.[8] This type of procedure, with automatic generation of variables from computer-readable representations of structure, could enable automatic classification to become a routine procedure within computerized chemical information systems. However, it is known that widely differing classifications, which may be equally valid representations of the data, are obtained by using different clustering algorithms.[9] The purpose of this work was to study this effect in the context of automatic classification of chemical structures and also to consider the related effect of using raw as against standardized data.

A small data set was constructed for this purpose, consisting of substituted benzene structures. This was chosen so that the effect of the choice of clustering methodology on the final result could be studied more easily than with "real", complicated data sets. Also it allowed classifications on the basis of "conventional chemical" structural features, in this case ring substituents, of the sort readily derived algorithmically from Wiswesser Line Notation (WLN).[10] These classifications could be compared with those obtained by using fixed-size atom- or bond-centered fragments as structural descriptors.[7,8]

EXPERIMENTAL SECTION

The structures were encoded in WLN[11] and descriptors derived algorithmically, as described previously.[10] The descriptors were counts of structural features (i.e., substituent type and relative position) of the kind readily derived from notation representation of structure.

Cluster analyses were carried out by using the CLUSTAN package.[12] Techniques of cluster analysis are fully described elsewhere,[9,13] and only an outline of significant points will be given here.

The methods used here fall into the category of hierarchical, agglomerative clustering techniques. By "hierarchical", it is meant that the classes, or clusters, are themselves classified into larger groups. Repetition of this process at different levels of similarity leads to the representation of the data set by a dendrogram or classification tree. By "agglomerative", it is meant that groups are formed, at each level of similarity, by fusions of existing groupings.

The first stage in such an analysis is the generation of a similarity or dissimilarity (or distance) matrix by computation of a similarity or dissimilarity coefficient between each pair of objects. A variety of such coefficients have been used.[9,13] However, earlier studies of chemical structure classification indicated that the choice of coefficient made little difference to the overall classification produced.[8] For the work reported here, therefore, it was decided to use a single coefficient. The Euclidean distance measure was chosen, because of its ready visualization, computational simplicity, and wide use in other areas.[13]

The Euclidean distance is defined as

$$d_{ij} = \sum_{k=1}^{n} [(X_{ik} - X_{jk})^2]^{1/2}$$

for the distance between objects i and j, where X_{ik} is the value of the kth variable for the ith object, and there are n variables

†ICI Pharmaceuticals Division, Alderley Park, Macclesfield, Cheshire, England.
‡Pfizer Central Research, Sandwich, Kent, England.

0095-2338/81/1621-0204$01.25/0

defined for the set of structures.

Standardization of variables involves dividing the value of each variable for every data point by a measure of the variance for that variable, usually the standard deviation. It has the effect of leaving the distance measure unaffected by changes of scale in the variables. Since the Euclidean distance measured is considerably affected by scaling factors, the classifications obtained using raw (i.e., unstandardized) and standard data are usually different when this coefficient is used. Standardization is common practice, although it may have the property of reducing between-group discrimination.

Although the variables being used here had an invariant scale, i.e., simple counts of occurrence, it was thought worthwhile to investigate the comparison between classifications with raw and standard data, in particular, the extent to which the clustering could be influenced to reflect the variables with different ranges of occurrence.

The agglomerative clustering techniques used here all create clusters by fusing the pairs of groups (which may be single individuals) which are most similar at each stage. The definition of "similarity", in this case the Euclidean distance, although straightforward for two individuals, may differ for the case of two groups or a group and an individual. It is this difference which distinguishes the clustering algorithms used here.

The *nearest neighbor* (or single link) technique defines the distance between two groups as the distance between their closest members. For the *furthest neighbor* (or complete linkage) technique the distance is taken as that between the most remote pair of members, and for the *group average* method the average of the distances between all pairs of members in the two groups is used.

Ward's method considers the potential joining of every possible pair of clusters. An "information loss" is defined as the total sum of squares deviations of every point from the mean of its cluster, i.e., the "looseness" of the cluster. The groupings are then constructed to minimize this information loss.

McQuitty's similarity analysis allows clusters to join on the basis of reciprocal similarity, i.e., the clusters to be joined must each resemble the other most closely.

Two other clustering methods were considered. These were the *centroid* and *median* techniques, both of which define the intercluster distance as the distance between cluster centers. However, in some cases, these techniques gave overlapping clusters and consequently undefined dendrograms. This made it very difficult to include the results in the kind of qualitative comparison envisaged here, and therefore these techniques were not further considered.

Comparison of classifications was carried out by inspection of the dendrograms. One criterion was whether well-defined clusters were produced or whether "chaining", i.e., the addition of entities to existing clusters one by one, gave a confused picture. The reasonableness, in chemical terms, of the classifications produced, and hence their potential usefulness, was also assessed. However, it is important to note that there can be no absolute measure of the "correctness" of a classification, since any discrepancy between the results produced by the various methods may reflect alternative, and equally valid, views of the data. Although the appropriateness of a classification for a particular purpose, e.g., property prediction, may be assessed objectively,[8] it seemed more appropriate for this study to rely on a wider intuitive assessment.

RESULTS

The compounds in this set of substituted benzenes are shown in Figure 1. Descriptors used were simply counts of substituent type; there were therefore seven variables, the oc-

Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures

Figure 1 Data-set

1. unsubstituted	19. 1, 3-Me, 4-Br
2. 1-Me	20. 1, 3-Me, 4-F
3. 1, 3-Me	21. 1, 2, 3, 5-Me, 6-Cl
4. 1, 2, 4-Me	22. 1, 2, 3, 5-Me, 6-Br
5. 1, 2, 3, 5-Me	23. 1, 2, 3, 5-Me, 6-F
6. 1, 2, 3, 4, 5-Me	24. 1-Me, 4-OMe
7. 1, 2, 3, 4, 5, 6-Me	25. 1-OMe, 4-Cl
8. 1-OMe	26. 1-OMe, 4-Br
9. 1, 3-OMe	27. 1-OMe, 4-F
10. 1-Me, 2-OMe	28. 1-OH
11. 1, 3-Me, 4-OMe	29. 1-Me, 2-OH
12. 1-Cl	30. 1, 3-Me, 4-OH
13. 1-Br	31. 1-NMe_2
14. 1-F	32. 1-Me, 2-NMe_2
15. 1, 3-diCl	33. 1, 3-Me, 4-NMe_2
16. 1, 3-diBr	34. 1-NMe_2, 4-Cl
17. 1, 3-diF	35. 1-NMe_2, 4-Br
18. 1, 3-Me, 4-Cl	36. 1-NMe_2, 4-F

12 August 1981

Comparison of Hierarchical Cluster Analysis Techniques for Automatic Classification of Chemical Structures

Figure 1. Data set.

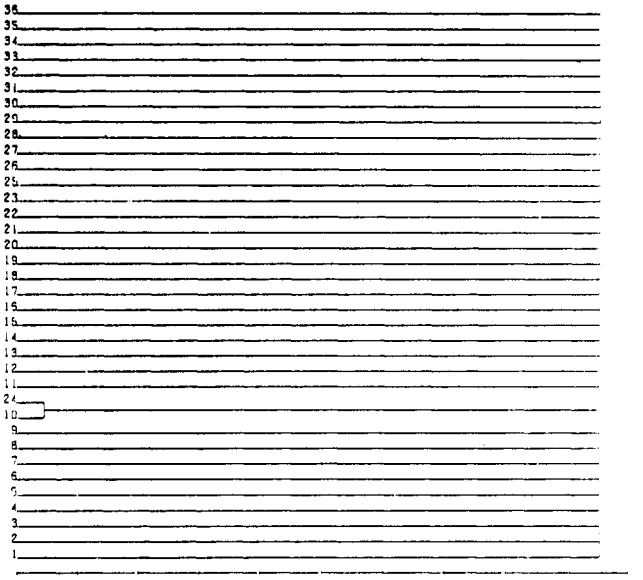

Figure 2. Single link, raw data.

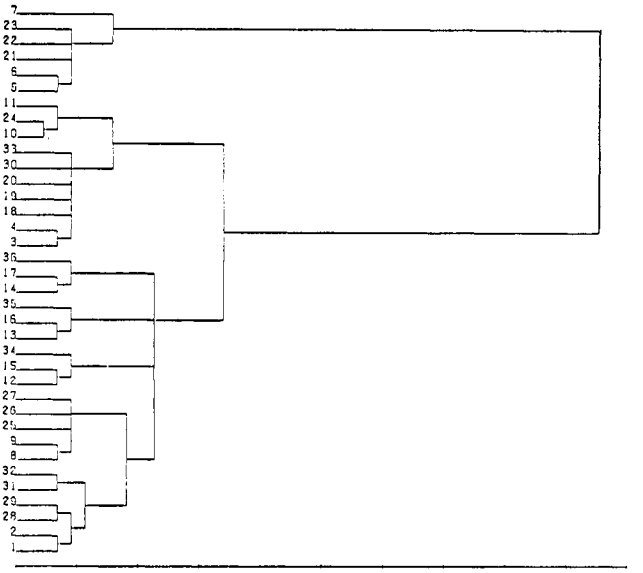

Figure 3. Complete link, raw data.

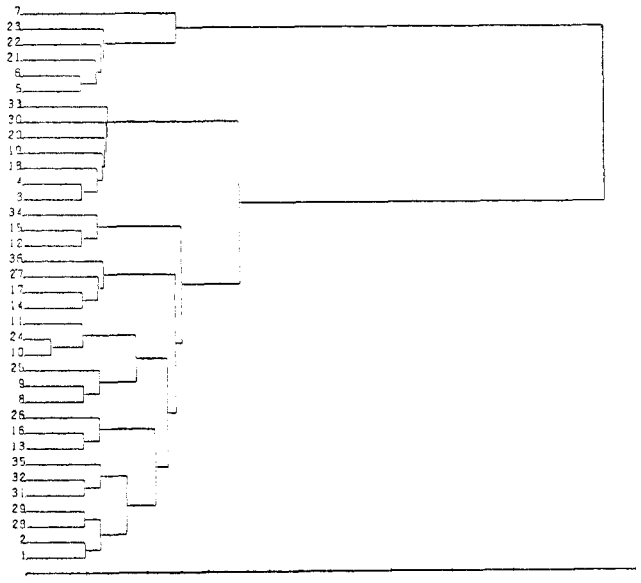

Figure 4. Group average, raw data.

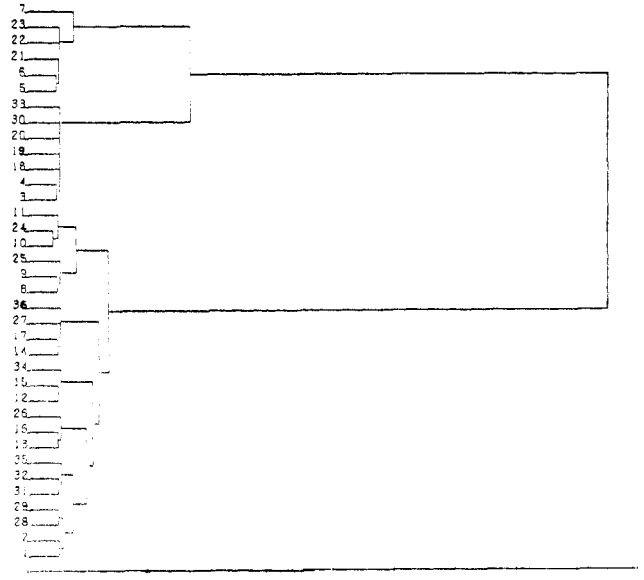

Figure 5. Ward's method, raw data.

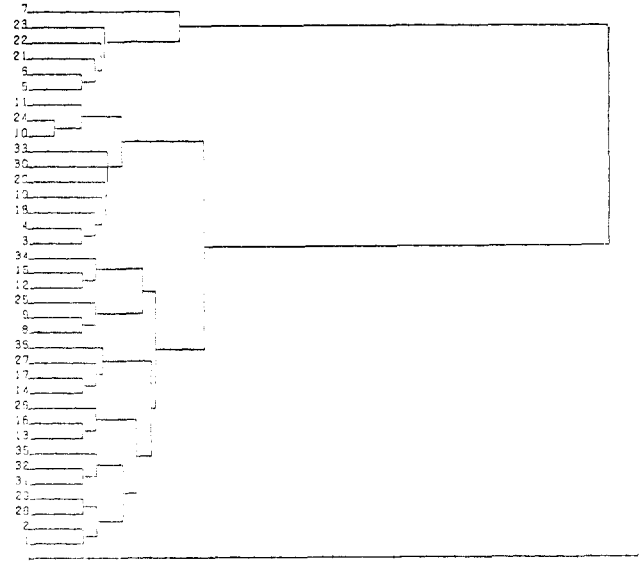

Figure 6. McQuitty's method, raw data.

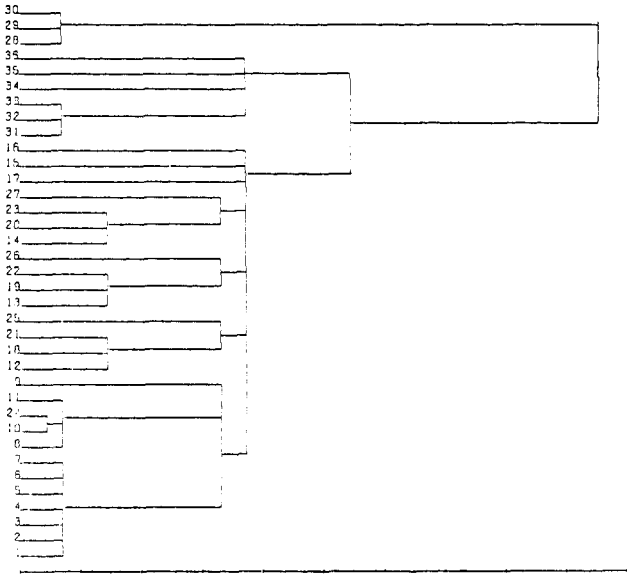

Figure 7. Single link, standard data.

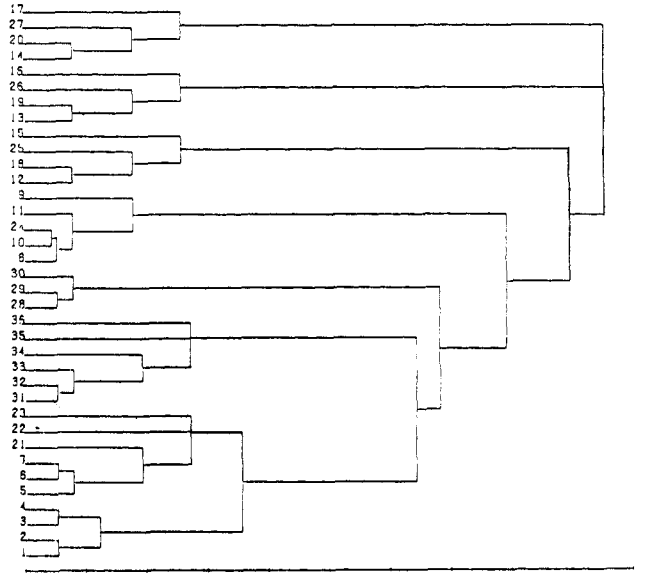

Figure 8. Complete link, standard data.

currences of the seven types of substituents (Me, OMe, Cl, Br, F, OH, NMe_2). The dendrograms resulting from the analyses are shown in Figures 2–11, as indicated:

	raw data	standardized data
single link	Figure 2	Figure 7
complete link	Figure 3	Figure 8
group average	Figure 4	Figure 9
Ward's method	Figure 5	Figure 10
McQuitty's method	Figure 6	Figure 11

With unstandardized data, the single link clustering was almost completely featureless and thereby valueless. The other four methods all gave reasonably clear classifications, although with some chaining. Ward's method gave a particularly distinct pattern of "flat" clusters. All four techniques gave classifications based both on number of substituents and on substituent type; thus the compounds containing four or more methyls, regardless of the other substituents present, are brought together in all four cases. The complete link classification differed considerably from the other three in showing a less well-defined overall structure, with considerable mixing of varying types of compound.

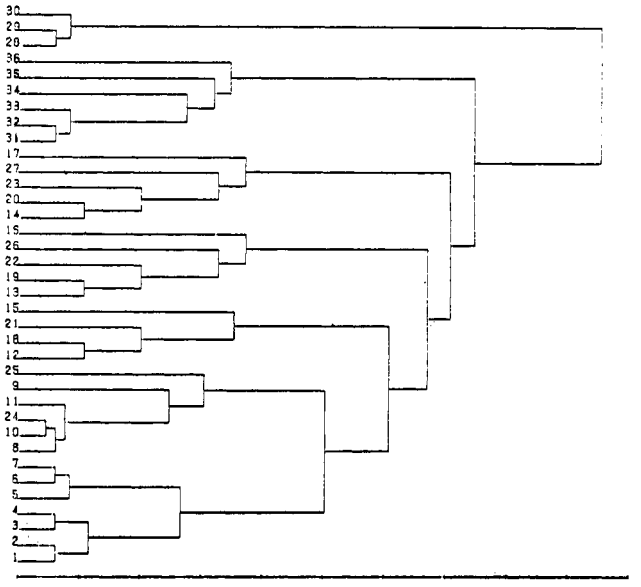

Figure 9. Group average, standard data.

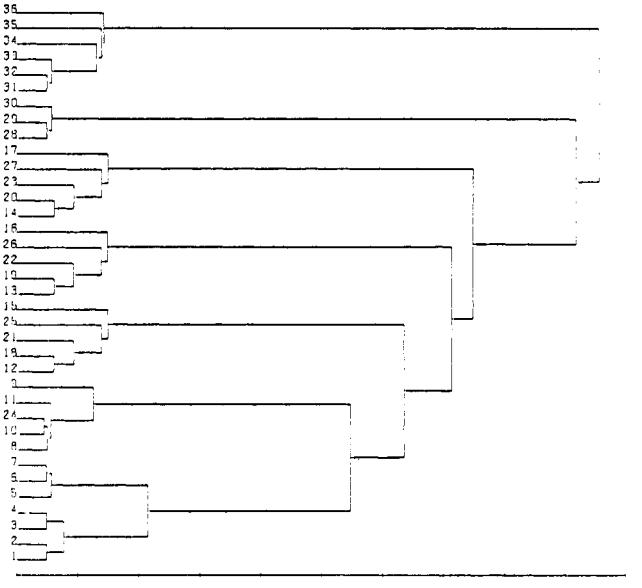

Figure 10. Ward's method, standard data.

The three remaining classifications (group average, Ward's, and McQuitty's methods) showed very similar classifications, though with some differences in detail. Ward's method, for example, differed from the other two in bringing the cluster of compounds with four or more methyls together with those with two or more methyls; the other two techniques gave the first as an outlying group. McQuitty's method differs, for example, in its placing of structures with methoxy substituents.

With standardized data, all the clustering techniques, including single link gave well-structured classifications, based upon substituent type although all differ in detail. The regular clustering pattern reflects very closely the intuitive view of this data set, as the labeling of the clusters indicates. Structures including NMe_2 and OH substituents are clustered first and then each of the halogen groups (except those which also contain NMe_2, which are in the NMe_2 group). The remaining clusters include firstly OMe alone and mixed OMe/Me (the mixed halogen/OMe compounds being in the halogen groups) and secondly compounds with Me alone and the unsubstituted compound (other Me containing structures being in the appropriate group).

The classification by group average and McQuitty's method are very similar, differing in the relative positions of some

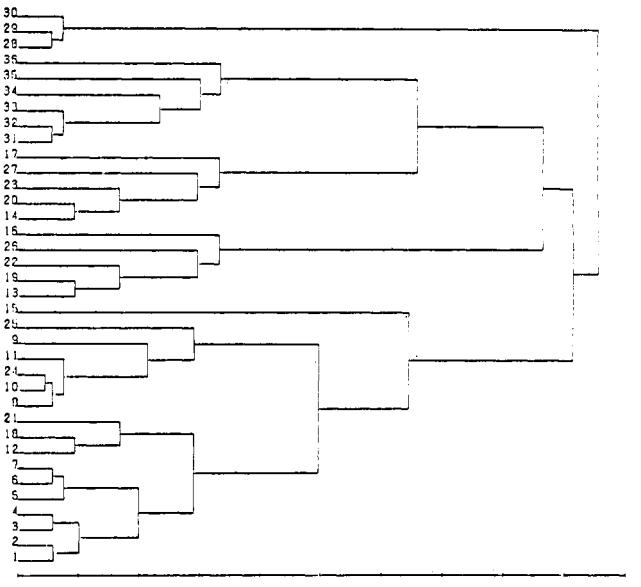

Figure 11. McQuitty's method, standard data.

groups and in the positioning of the Cl derivatives. The furthest neighbor shows a less regularly ordered clustering, with formation of subgroups; for example the compounds with one halogen and four methyl substituents form a group close to the methyl cluster, rather than being incorporated in the halogen clusters.

The single link clustering gave a classification with much less within-cluster structure. The same sort of clusters are seen as with the other methods: the dihalo and dimethoxy substituents are separated.

One noticeable effect of standardization of data was the introduction of a greater degree of within-cluster structure because of the splitting of coincident similarity levels caused by the integral values in the unstandardized data. This may be clearly seen in the comparison of the Ward's clusterings and also in the production of an intuitively reasonable single link clustering from the complete chaining with raw data.

DISCUSSION

The work described here shows that application of hierarchical clustering algorthims with descriptors representing "conventional" structural moieties, algorithmically derived, can give intuitively sensible classifications of structure.

The type of structure descriptors used here allows classifications based on substituent type. Other studies of classifications based on WLN fragments[15] indicate that other conventional descriptors, such as aliphatic functionalities and relative position of substituents, can also be used to give intuitively sensible classifications. Such classifications could usefully complement those based on atom- and bond-centered fragments,[7,8] which are more likely to represent overall structural composition. As noted earlier, these two types of descriptor are most readily derived from linear notations and connection tables, respectively.

The use of descriptors at varying levels of specificity can give differing clustering patterns.[15] A similar effect has been noted with atom- and bond-centered fragment descriptors.[14] This indicates the great advantages of flexible, automatic descriptor generation programs, perhaps using several structural representations, if clustering is to be used to best effect for exploratory analyses in sets of structures. The differences in classifications, based on the same set of descriptors, brought about by using several clustering algorithms, and by standardization of the data, are noteworthy. Although it is true that, in most cases, the broad features of the set of structures

were brought out by all techniques, the differences are likely to be of considerable importance in practical application. These effects have also been noted in studies of other data sets, including aliphatic structures.[15]

Over all these analyses, each of the clustering algorithms showed, in some cases, distinct differences from the others. These differences were seen in the overall pattern of clusters and in detail; only for the most clearly defined cases was identical clustering produced by all five methods. The centroid and median techniques, as has been noted, gave uninterpretable results in several cases. For the rest, it is difficult to categorize the alternative views of the data produced as good or bad per se. However, some generalizations may be made, on the basis of the "chemical sense" of the results and the likely usefulness.

The single link method gave a totally chained, and therefore uninformative, classification in one case and in general showed a less-structured clustering than the other techniques. The furthest neighbor classifications were in some cases poorly structured, with a rather muddled, and not apparently useful, clustering. Of course these cannot necessarily be rejected out of hand, since they may be perfectly valid representations of the data sets.

The remaining three techniques consistently gave clear and potentially useful classifications. Group average and McQuitty's method tended to give similar clusters, with Ward's method (with its characteristic "flat" distinct clusters) sometimes giving a complementary view of the data.

These results suggest clearly that it is advisable, where possible, to use several clustering techniques on these sorts of data sets. A useful combination could be Ward's method, one of group average or McQuitty's method, and one of single link or furthest neighbor.

It is interesting to note a recent summary of some comparisons of clustering algorithms on a variety of data sets:[16] "In general the results of such studies indicated that (1) no single method is best in every situaton, (2) the mathematically respectable single linkage is, in most cases, the least successful for the data used, and (3) group average clustering and Ward's method do fairly well overall". The results of the work reported here are entirely in accord with these findings.

The use of raw and standardized data for classification can also give complementary information with data sets of this sort. Standardization tends to weight clustering in favor of less commonly occurring features. In this work, this is seen as standardization giving classification primarily by substituent type, while with raw data the number of substituents is of equal importance. Other studies have shown a similar effect with aliphatic sets:[15] standardized data gives classifications primarily by functionality, while raw data includes carbon skeleton information. It is not possible to say a priori which of these will be more desirable.

The other effect of standardization, as noted, is the introduction of a greater degree of within-cluster structure. This is particularly desirable for the single link and further neighbor techniques, where a clear cluster pattern sometimes cannot be obtained with unstandardized data. Standardization can also have the effect of giving different classifications from several techniques which give identical results with raw data.

It is evident from these results that it is possible to obtain a variety of complementary classifications of sets of chemical structures by varying descriptor type, clustering algorithm, and preprocessing. It is also possible, to some extent, to alter the factors so as to get the desired type of classification.

There are two main potential applications for clustering techniques of this sort as applied to the structures of sets of chemical compounds. The first is information retrieval. Because of the computational requirements of the techniques,[17] there is a very limited size of file which may be dealt with by hierarchical techniques of this sort. In practice this could mean that clusterings could be carried out on output from conventional searching systems to aid ranking of output, selection of some typical answers, etc. A variation of this last application is the selection of representative members from some larger set, e.g., for some testing procedure.

The feasibility of this sort of procedure is established by the work reported here and earlier studies.[7,8] The effect which the varying classifications, obtained by varying descriptors, clustering procedures, etc., would have on such applications is a matter for further research.

The second potential application is structure–property correlation. The prediction of some molecular property on the basis of similarity to compounds of known property has been demonstrated.[6-8] However, the very different, though equally valid, classification found in this work shows the problem of the choice of clustering technique, etc., for a particular predictive application. It may well be that other techniques, such as multiple regression or discriminant analysis, will prove more appropriate than cluster analysis for correlation and prediction in studies of structure–activity relationships (SAR).

The main value of cluster analysis is likely to lie in the exploratory data analysis phase of SAR as a tool for preliminary assessment of structure within data sets, perhaps in conjunction with other techniques such as multidimensional scaling, nonlinear mapping, or principal component analysis.[18,19] Its usefulness in such studies could be enhanced by the ability to apply a variety of clustering algorithms and preprocessing techniques; an interactive graphics facility could also be valuable for rapid display of results. An efficient, flexible procedure for generating structure descriptors, perhaps based on an ability to generate multiple structural representations, is essential if effective use is to be made of classification techniques in this way.

ACKNOWLEDGMENT

We thank Drs. J. A. Bush and P. Willet for helpful discussions and the Department of Education and Science (London) for the award of a Postgraduate Research Studentship to D.B.

REFERENCES AND NOTES

(1) Saggers, D. T. "The Application of the Computer to a Pesticide Screening Program". *Pestic. Sci.* **1974**, *5*, 341–352.
(2) Lewi, P. J. "Computer Technology in Drug Design". *Med. Chem. (Academic)* **1978**, *11* (Drug Design, Vol. 7), 209.
(3) Hansch, C., et al. "Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents". *J. Med. Chem.* **1973**, *16*, 1217–1222.
(4) Dunn, W. J., et. al. "Use of Cluster Analysis in the Development of Structure–Activity Relations for Antitumour Triazines". *J. Med. Chem.* **1976**, *19*, 1299–1301.
(5) Sneath, P. H. A. "Relations between Chemical Structure and Biological Activity in Peptides". *J. Theor. Biol.* **1966**, *12*, 157–195.
(6) Chu, K. C. "Applications of Artificial Intelligence to Chemistry". *Anal. Chem.* **1974**, *46*, 1181–1187.
(7) Adamson, G. W.; Bush, J. A. "A Method for the Automatic Classification of Chemical Structures". *Inf. Storage Retr.* **1973**, *9*, 561–568.
(8) Adamson, G. W.; Bush, J. A. "A Comparison of the Performance of Some Similarity and Dissimilarity Measures in the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 55–58.
(9) Everitt, B. "Cluster Analysis"; Heinemann: 1974.
(10) Adamson, G. W.; Bawden, D. "An Empirical Method of Structure–Activity Correlation for Polysubstituted Cyclic Compounds Using Wiswesser Line Notation". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 161–165.
(11) Smith, E. G.; Baker, P. A. "The Wiswesser Line-Formula Chemical Notation", 3rd ed.; Chemical Information Management Inc.: Cherry Hill, NJ, 1976.
(12) Wishart, D. "CLUSTAN User Manual"; University of Edinburgh: Edinburgh, Scotland, 1978.
(13) Sneath, P. H. A.; Sokal R. R. "Numerical Taxonomy"; W. H. Freeman: San Francisco, CA, 1973.
(14) Bush, J. A. Ph.D. Thesis, University of Sheffield, 1976.
(15) Bawden, D. Ph.D. Thesis, University of Sheffield, 1978.

(16) Everitt, B. S. "Unresolved Problems in Cluster Analysis". *Biometrics* **1979**, *35*, 169–181.
(17) Williams, W. T.; Lance, G. N. "Hierarchical Classificatory Methods". In "Statistical methods for Digital Computers"; Enslein, K., et al., Eds.; Wiley: New York, 1977; Vol. III.
(18) Kowalski, B. R.; Bender, C. F. "Pattern Recognition. II. Linear and Nonlinear Methods for Displaying Chemical Data". *J. Am. Chem. Soc.* **1973**, *95*, 686–693.
(19) Everitt, B. S.; "Graphical Techniques for Multivariate Data"; Heinemann: 1978.

PULSAR: A Personalized Microcomputer-Based System For Keyword Search and Retrieval Of Literature Information

SCOTT F. SMITH,[1] WILLIAM L. JORGENSEN,*[2] and PHILIP L. FUCHS*[3]

Department of Chemistry, Purdue University, West Lafayette, Indiana 47907

Received March 9, 1981

A keyword-based storage and retrieval system for literature references has been developed with a TRS-80-II microcomputer. The system, called PULSAR, has been designed to provide and maintain rapid access to a personalized data base. Application of the PULSAR system to the literature of synthetic organic chemistry is described.

INTRODUCTION

When a scientist requires information, he will often draw first upon personal resources. These resources include all the specific contributions of that individual as well as any relevant literature information which he can recall. It is at this latter stage that major retrieval problems occur. A researcher acquires a rapidly increasing accumulation of data as his career proceeds. The specific manner in which these data are stored will determine their subsequent availability.

A variety of techniques have been traditionally employed for this task; most usually some variant of the well-known "card-file" system. In this case information is recorded on an index card, the file grows, and the cards are resorted as new categories are created. The system begins to become inefficient when the cards number in the thousands. At this point the categories have usually become too general, and it is apparent that substantial cross-indexing is necessary. The researcher might go through a temporary phase in which the cards can supposedly be sorted by passing a knitting needle through the edge of the cards. There are numerous inconveniences associated with all card-filing systems.

A number of general and extensive chemical information systems are now available to facilitate literature searching, including the Chemical Abstracts, Lockheed, NIH/EPA, and other systems.[4-7] Even with these systems readily accessible, there is strong justification for maintaining a personalized system restricted to the interests of an individual and based on the individual's own choice of keywords.[8,9] Due to the availability of microcomputers with diskettes, we have been able to develop such a system at reasonable cost capable of handling up to 20 000 references including keywords. The program is called "PULSAR" for Purdue University Literature Search and Retrieval system.[10]

PROGRAM FEATURES

The PULSAR system is implemented on a Radio Shack TRS-80-II computer with 64K bytes of memory, a printer, and from one to four 500K byte disk drives for a total storage capability of 2 Mbytes. The program is written in the BASIC language, and all 64K bytes of memory are used. The various options available in the PULSAR system are as follows:

I. Add Articles
II. Search for Keywords
III. Display Routines
(A) Display a single article
(B) Display a series of articles
(C) Display keywords alphabetically
(D) Display all journal book names
(E) Display system status information
(F) Display free space map
(G) Display disk directory
(H) Write a message on the printer
IV. Data Checking Routines
(A) Check keywords (start check at Keyword K)
(B) Check articles (starting with article No. *M*)
(C) Reformat (compact) link records
V. Editing Routines
(A) Edit keywords (rename/merge/delete)
(B) Combine keywords A and B to C
(C) Edit articles
(D) Edit journal names
VI. Disk File Management Routines
(A) Make a complete backup
(B) Format a new diskette
(C) Swap diskettes
(D) Move files from one disk to another
(E) Enable remote terminal

USING PULSAR

In the Add Articles routine (I), for each article reference, the following information is stored:

Entry number
Journal name
Volume, year, page
Type of article (Paper, Communication, Note, Thesis, Miscellaneous)
Keywords (1–8 keywords, ≤30 characters in length each)

The above information can later be retrieved by use of the Search for Keywords routine (II). For example, *A and B* will retrieve all article references containing both keyword A and keyword B. The logical expressions *or*, *xor*, and *not* are also usable; so *P and not* (*Q xor R*) *or* (*S and not T*) is a valid expression as well. Up to eight separate keywords may be referred to in any one search expression. Upon completing a search, the computer displays the number of matches upon a video screen. One may now elect to enter an alternate search

0095-2338/81/1621-0209$01.25/0