Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures

PETER WILLETT

Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

Received June 7, 1983

Seven relocation clustering procedures were used to cluster eleven small sets of chemical structures characterized by augmented atom fragments. The effectiveness of the classifications was assessed by means of simulated property prediction experiments, and quite noticeable differences were observed between the performance of the various algorithms. Studies were made of the extent to which the classifications depended upon the initial partition chosen and upon the order in which the data sets were processed.

INTRODUCTION

Automatic classification, or cluster analysis, is a multivariate technique that seeks to identify groups, or clusters, of related objects in a multidimensional space.[1-5] Forsythe et al.[6] and Dunn et al.[7] have used the technique as a means of grouping substituents in drug design programs, the substituents being characterized by physicochemical parameters. An alternative approach, and the one that is investigated in this report, has been suggested by Harrison[8] and Adamson and Bush,[9,10] who have sought to group collections of structures on the basis of substructural similarities, with the degree of similarity between the compounds in a data set being determined by the numbers of common fragment substructures.

An important factor which must be considered when using cluster analysis is the choice of clustering method since there are very many different algorithms available, and different algorithms may well not give identical classifications when applied to the same data set. There are two main types of classification algorithm which may be used for the identification of the clusters present in a data set, these being hierarchal and relocation clustering. Hierarchal algorithms require the calculation of an intermolecular similarity matrix as a precursor to the generation of the classification. Initially, each compound in a data set is in a cluster on its own, and the hierarchal classification is built up by a series of fusions in which the most similar pairs of clusters are merged until all of the compounds are in a single cluster. Rather than this agglomerative approach, a hierarchal divisive algorithm may be used in which all of the molecules are initially in a single cluster that is progressively subdivided. Relocation, or partitioning, methods seek to divide a data set into some number of disjoint clusters such that related compounds will all be in the same cluster, with compounds unrelated to that cluster being distributed among the other clusters in the set. An optimal partition may be obtained by selecting some criterion which measures the "goodness", in some statistical sense, of a particular partition, and then systematically searching through all possible partitions for that one which optimizes the chosen criterion. Since this approach is computationally infeasible for all but trivially small data sets, the concept of relocation is introduced by which a local optimum may be found at little computational cost. Relocation involves the movement of compounds between clusters in such a way as to increase the homogeneity of the individual clusters, the degree of intercluster similarity being measured by some similarity or distance function.

Of these two classes of algorithm, the hierarchal methods have probably been the most widely used, and three recent papers have reported evaluations of the merits of the various hierarchal algorithms for grouping chemical structures.[11-13] In most applications areas, clustering algorithms can be evaluated only in qualitative terms on the basis of the perceived utility of the classifications resulting from their use; thus Adamson and Bawden[11] have studied the dendrograms arising from the application of several agglomerative algorithms to a group of substituted benzenes. However, Adamson and Bush[9,10] have described a more quantitative means of evaluating classification methods in cases where property data are available for the chemical structures that are being clustered. Their experiments measured the extent to which classifications based only upon fragment similarities also reflected similarities in biological activity, with the predicted property value for a compound being taken as the average of the values for the other molecules in the cluster containing that compound. This simulated property prediction approach was used for evaluating both agglomerative[12] and monothetic divisive[13] algorithms and also forms the basis for the work reported here which considers relocation clustering techniques.

RELOCATION CLUSTERING

A simple relocation algorithm is as follows.

(i) Initialization. The initial set of clusters is obtained by randomly assigning integers in the range 1 to c, where c is the desired number of clusters, to each of the compounds in a data set and then taking all molecules with the same index as being in the given cluster.

(ii) Relocation. Each of the structures is matched against the mean vector of each of the clusters and assigned to that cluster which results in the smallest (or largest) value for the chosen dissimilarity (or similarity) measure, and then the mean vectors of the new clusters are computed. The relocation is repeated for some fixed number of iterations, or until no further relocation of compounds takes place: this will correspond to a local, but not necessarily global, minimum in the clustering criterion. If a hierarchal classification is desired, the two most similar clusters are fused, the relocation phase repeated to obtain a local optimum for the new set of clusters, and the procedure iterated until all of the molecules are in a single cluster. A detailed description of relocation clustering is given by Dubes and Jain,[4] who describe the many modifications which have been suggested to this basic procedure.

In the discussion that follows, it is assumed that each of the N molecules in a data set is characterized by a set of M different fragment types, the actual fragment employed here being the *augmented atom* which consists of an atom together with the immediately adjacent atoms and bonds. Then the data set may be represented by an N by M matrix, each element of which, x_{ij}, contains the number of occurrences of the jth fragment in the ith molecule. As the classification proceeds, individual molecules within the data set become clustered, and the various relocation algorithms tested here differ primarily in the measure that is used for the calculation

0095-2338/84/1624-0029$01.50/0

of the distances between compounds and cluster centers.

Let n_p be the number of molecules in the pth cluster; then let u_{jp} be the mean value for the jth fragment for the molecules in the pth cluster, so that

$$u_{jp} = \frac{1}{n_p} \sum_{i \in p} x_{ij}$$

Let $\bar{U}_p$ be the average of the mean scores for the pth cluster, so that

$$\bar{U}_p = \frac{1}{M} \sum u_{jp}$$

Then the seven similarity measures used for the calculation of the similarity between two clusters, p and q, are as follows (the summations are over all j unless specified otherwise):

(i) Euclidean distance (DIS)

$$d_{pq}^2 = \frac{1}{M} \sum (u_{jp} - u_{jq})^2$$

(ii) Average distance (AVD)

$$\frac{1}{n_p n_q} \sum_{i \in p} \sum_{j \in q} d_{ij}^2$$

This is the average of the squares of the distances between all members of one cluster and all members of another.

(iii) Error sum of squares (ESS)

$$\frac{n_p n_q}{n_p + n_q} d_{pq}^2$$

ESS is the sum of the distances from each compound in a cluster to the center of that cluster, and it thus measures the extent of the scatter about the cluster center; with DIS, this is probably the most familiar and widely used similarity measure for relocation clustering.

(iv) Similarity ratio (SIM)

$$\frac{\sum u_{jp} u_{jq}}{\sum u_{jp}^2 - \sum u_{jp} u_{jq} + \sum u_{jq}^2}$$

In the case of binary data, this reduces to the well-known Jaccard coefficient.

(v) Size difference (SIZ)

$$(\bar{U}_p - \bar{U}_q)^2$$

This is the square of the difference between the average fragment values in the two clusters: this is probably the simplest of all of the measures.

(vi) Shape difference (SHA)

$$d_{pq}^2 - \frac{1}{M^2} (\sum u_{jp} - \sum u_{jq})^2$$

This is basically the variance of the M differences between the mean fragment occurrences in the two clusters. It has been suggested that this measure results in partitions that are strikingly different from those produced by DIS and SIZ.

(vii) Variance (VAR)

$$\frac{1}{M(n_p + n_q)} \sum_j \sum_{i \in p,q} (x_{ij} - u_{jp+q})^2$$

This is the sample variance of a cluster formed from the union of the two clusters, p and q, divided by the total number of fragment types.

There is a clear distinction between SIZ and the other similarity measures. SIZ measures the distance, or dissimilarity, between molecules or clusters by considering the mean fragment frequencies, regardless of the actual chemical nature of the substructures; the other measures, conversely, all take the fragment types into consideration when evaluating the degree of similarity between a pair of clusters or molecules.

EXPERIMENTAL DETAILS

The data sets are those involved in earlier experiments[12,13] using the hierarchal methods. Eleven small collections of compounds for which associated property data were available were selected from the structure–property literature. The data sets include physical, chemical, and biological properties and were as follows: (A) pI values of 20 naturally occurring amino acids,[9] (B) local anaesthetic activity of 37 diverse structures,[15] (C) inhibition of complement by 105 benzamidines,[14] (D) serum binding activity of 79 penicillins,[10] (E) tadpole narcosis activity of 34 diverse structures,[16] (F) molar refractivity of 65 aliphatic ethers, amines, alcohols, and halides,[17] (G) chymotrypsin hydrolysis by 72 *N*-acyl esters,[18] (H) mouse toxicity of 25 aliphatic and carbocyclic ethers,[19] (I) antimicrobial activity of 28 phenyl propyl ethers,[20] (J) inhibition of dihydrofolate reductase by 46 quinazolines,[21] and (K) heats of vaporization for 126 alkenes, alcohols, ketones, benzenes, and pyridines.[22]

The connection table representing each of the structures in a data set was analyzed to identify the substructural fragments present: as noted earlier, the fragment type used here was the well-known augmented atom. Each structure was described by a list of the augmented atoms contained within it, and these lists comprised the data matrix for the generation of the classifications. It has been suggested that the attributes used for characterizing the objects in an automatic classification study should be standardized to compensate for variations in scaling. However, previous experiments[12] have shown that doing this with fragment attributes leads to a noticeable drop in the utility of the subsequent classifications: accordingly, the raw, unstandardized data was used for the experiments.

The agglomerative procedures studied earlier all involve the fusion of individual molecules to form clusters, and the subsequent fusion of these clusters to form larger groupings: accordingly, the smallest cluster for each compound contained only a few structures, typically two or three. For a comparison with this earlier work, the number of initial clusters for each relocation was set to half of the number of compounds in a data set, and the compounds were then assigned to these initial clusters by using a random number generator.

The classifications were obtained from the initial partitions using the CLUSTAN package[23] which contains FORTRAN routines for a very wide range of clustering methods. These classifications formed the basis for simulated property prediction using the "leave-one-out" approach described above. The property value for each molecule, *i*, in turn within a data set was assumed to be unknown, and the classification was scanned to identify the cluster containing *i*: the predicted property value for *i* was taken to be the mean of the observed values for the other structures in that cluster. The correlation between the observed and predicted values was then determined by means of the product moment correlation coefficient.

Relocation clustering has two principal defects, owing to the manner in which the algorithms operate. First, the final classification will depend upon the initial assignment of compounds to clusters, with the possibility that different clusterings will be obtained from different initial partitions. Second, the final classifications are dependent upon the order in which the structures are processed, owing to the serial nature of the relocation. These two factors imply that one can never be certain that a better classification might not be obtained from an alternative partition or from a reordering of the data file.

The problem of the initial partition was investigated by using a set of five different assignments for each combination of clustering algorithm and data set and then calculating the mean and standard deviation for each such combination. The

Table I. Correlation Coefficients Using a Variable Initial Partition and a Fixed Order for Each Data Set[a]

	DIS	AVD	ESS	SIM	SIZ	SHA	VAR
A	0.81, 0.17	*	*	*	*	0.77, 0.15	*
B	0.70, 0.01	0.82, 0.01	0.83, 0.01	0.76, 0.02	0.94, 0.00	0.69, 0.01	0.72, 0.02
C	0.88, 0.01	0.86, 0.01	0.87, 0.01	0.84, 0.01	0.89, 0.01	0.88, 0.01	0.88, 0.00
D	0.76, 0.01	0.71, 0.05	0.77, 0.02	0.77, 0.01	0.51, 0.06	0.77, 0.01	0.75, 0.04
E	0.75, 0.07	0.78, 0.01	0.75, 0.01	0.66, 0.03	0.94, 0.00	0.73, 0.07	0.73, 0.04
F	0.61, 0.02	0.61, 0.03	0.61, 0.02	0.64, 0.02	0.84, 0.00	0.62, 0.01	0.58, 0.01
G	0.91, 0.01	0.89, 0.03	0.87, 0.01	0.90, 0.02	0.90, 0.01	0.91, 0.02	0.90, 0.02
H	0.52, 0.09	0.69, 0.08	0.62, 0.12	*	0.87, 0.08	0.46, 0.08	0.61, 0.11
I	0.52, 0.04	0.48, 0.06	0.49, 0.07	0.53, 0.05	*	0.50, 0.03	*
J	0.76, 0.03	0.70, 0.10	0.73, 0.16	0.77, 0.06	0.39, 0.02	0.76, 0.05	0.69, 0.09
K	0.85, 0.05	0.86, 0.05	0.83, 0.06	0.83, 0.04	0.53, 0.01	0.85, 0.02	0.85, 0.02

[a] The first figure quoted in each case is the mean value when averaged over five computer runs, and the second figure is the deviation. An asterisked entry means that one or more of the correlations was not significant at the 0.05 level of statistical significance.

Table II. Correlation Coefficients Using a Fixed Initial Partition and a Variable Order for Each Data Set[a]

	DIS	AVD	ESS	SIM	SIZ	SHA	VAR
A	0.93, 0.01	0.84, 0.00	0.90, 0.06	0.87, 0.06	*	0.93, 0.01	0.64, 0.13
B	0.76, 0.03	0.82, 0.00	0.82, 0.00	0.79, 0.02	0.94, 0.00	0.70, 0.00	0.74, 0.10
C	0.84, 0.01	0.85, 0.00	0.85, 0.00	0.85, 0.00	0.89, 0.00	0.84, 0.01	0.88, 0.01
D	0.75, 0.02	0.74, 0.01	0.77, 0.01	0.77, 0.01	0.48, 0.06	0.77, 0.01	0.73, 0.07
E	0.79, 0.04	0.77, 0.01	0.77, 0.01	0.67, 0.04	0.95, 0.00	0.79, 0.04	0.70, 0.02
F	0.59, 0.02	0.63, 0.03	0.58, 0.01	0.64, 0.02	0.83, 0.01	0.57, 0.01	0.60, 0.01
G	0.93, 0.00	0.88, 0.00	0.92, 0.02	0.92, 0.01	0.89, 0.01	0.93, 0.00	0.90, 0.01
H	0.58, 0.05	*	0.63, 0.02	*	0.87, 0.08	0.56, 0.04	0.51, 0.11
I	0.55, 0.09	0.51, 0.03	0.53, 0.11	0.53, 0.01	*	0.55, 0.09	0.54, 0.04
J	0.75, 0.09	0.59, 0.16	0.70, 0.12	0.71, 0.10	0.39, 0.02	0.78, 0.03	0.59, 0.14
K	0.78, 0.02	0.78, 0.01	0.82, 0.04	0.80, 0.01	0.53, 0.01	0.79, 0.01	0.83, 0.04

[a] See footnote to Table I.

extent to which the classifications are dependent upon the order in which the data set is processed was studied by taking five different orderings of each of the data sets, with the same initial assignment in each case; the means and standard deviations were again noted. The fixed, initial partitions were obtained by using a method initially described by Dattola[24] in the context of automatic document classification for information retrieval purposes. The method consists of summing the frequencies of occurrence for each of the fragments in a data set and then selecting as the initial cluster centers those molecules with the highest sums of fragment frequencies: the initial partition was then generated by assigning the remaining molecules to their nearest center.

RESULTS

Tables I and II give the correlation coefficients between the observed and predicted property values for each of the 11 data sets listed above. Table I refers to the use of a fixed ordering for each of the data sets, but with different initial assignments, while Table II refers to the converse; the figures listed are the mean values and standard deviations when averaged over five computer runs. An asterisked entry in these tables means that one or more of the runs resulted in a correlation coefficient that was not significantly greater than zero at the 0.05 level of statistical significance.

For comparison with the figures in Tables I and II, Table III contains the correlation coefficients obtained by using Ward's hierarchal agglomerative algorithm on these data sets with Euclidean distance as the measure for the calculation of the intermolecular similarity matrix, a combination that has been shown to give reasonable results in previous work.[11,12]

Table III. Correlation Coefficients for Ward's Hierarchal Agglomerative Clustering Algorithm

data set	correlation	data set	correlation
A	0.94	G	0.87
B	0.86	H	0.35
C	0.87	I	0.65
D	0.77	J	0.79
E	0.78	K	0.86
F	0.65		

DISCUSSION

The fixed-order, variable initial partition results will be discussed first. Going through the data sets, one at a time, it will be seen that there is a very marked difference in the performance of the clustering algorithms when taken as a whole. Thus, in the case of the amino acid pI values, only two of the algorithms, DIS and SHA, consistently gave significant values, and the standard deviations show a large variation from one run to another. The ether toxicity and quinazoline data sets are also characterized by high standard deviations, and in such cases, little credence could be put upon a single classification since it is clear that the local optima identified by the algorithms differ significantly from one run to another. In the case of some of the other data sets, such as the benzamidines and local anaesthetics, the standard deviations are uniformly low, and in such cases, the local optima result in only slight variations in predictive ability; this does not, of course, necessarily mean that any of these do, in fact, coincide with the globally optimal partition for the data set.

Considering the mean values for the coefficients, rather than the standard deviations, differences between the individual algorithms are apparent. The most widely used criteria for relocation clustering would appear to be DIS and ESS, and both of these gave reasonable levels of predictive performance across the data sets; moreover, these levels were at least as good as those obtained by using the similarity measures SIM, AVD, SHA, and VAR. The results from SIZ tend to be either markedly superior to the other methods, as happens with the molar refractivity, tadpole narcosis, local anaesthetic, and ether toxicity data, or noticeably inferior, as with the quinazolines, penicillins, and heats of vaporization. It was noted above that the SIZ measure is very different in nature from the six other similarity functions tested here, in that it does not take the actual substructural nature of the fragments into consideration, and it would hence be expected to give low levels of predictive performance for those data sets where the property is strongly dependent upon the functionalities present. This is, of course, what happens since heats of vaporization are highly functional group dependent, while the biological activities of the penicillin and quinazoline data sets are dependent upon log P, and hence upon functionality. In the cases where the SIZ results are noticeably superior, the tadpole narcosis and local anaesthetic data sets are very diverse in size and complexity; molar refractivity is to some extent functionality dependent, but this effect might well be outweighed in this data set by sheer molecular size, that is, the number of fragments, while the ether toxicity data set, where all of the molecules share a common functionality, might also be strongly dependent upon molecular complexity.

Turning to the figures in Table II, those with a fixed initial partition and with different orderings of the file, similar trends are apparent. The magnitudes of the correlations are comparable with those in Table I, as are the standard deviations, and the SIZ measure again results in a pattern of behavior that is quite different from the other similarity measures. The main difference is the results for the amino acids, data set A, where a much better correlation is evident between structure and property. Since it is relatively easy to obtain a fixed initial partition, and thus to obviate the dependence of the final classification upon it without affecting predictive ability, it would seem to be well worth doing this.

The values for Ward's hierarchal agglomerative method are given in Table III, and these results may be compared with those in Tables I and II by determining whether the Ward value for a particular data set lies within the 95% confidence limits for a relocation method: if it does, it may be concluded that the two clustering methods do not give significantly different correlations for that data set. Such a comparison emphasizes the rather poor performance of the SIM, AVD, SHA, and VAR similarity measures and also shows that the other measures offer a level of predictive ability that is not noticeably inferior to that obtained from Ward's method.

The slight inferiority of the correlations obtained with the relocation methods is compensated for in part by the greater efficiency of relocation as against hierarchal clustering. Hierarchal agglomerative algorithms involve the calculation of an intermolecular similarity matrix as a precursor to the generation of the classification, and although relatively efficient algorithms exist for this purpose,[25] the computational demands are heavy, being of order $O(N^2)$ for a file of N structures. Polythetic divisive clustering is well-nigh infeasible on a large scale, and even the less demanding monothetic methods require a considerable amount of computation; additionally, the correlation between structure and property is often noticeably less than with the polythetic agglomerative methods when heterogeneous data sets are tested.[13] Relocation clustering as tested here involves computation of the order $O(N)$, although the constant of proportionality involves the product of the number of iterations and the number of clusters in the initial partition. Thus when large numbers of compounds need to be processed, as would be the case with an internal structure file, the relocation methods may be implemented on a very much larger scale than hierarchal clustering. Comparable problems of scale occur in applications of cluster analysis to the grouping of documents for information retrieval purposes, and there has accordingly been a considerable amount of interest in relocation methods for document classification.[24,26,27] It is hoped some of these procedures may be tested in a chemical context in the near future.

CONCLUSIONS

The experimental results show clearly that relocation clustering methods are quite successful in grouping molecules with similar biological activities upon the basis of substructural similarities: in this respect, at least some of the methods offer a viable alternative to the hierarchal agglomerative and divisive procedures investigated in earlier work. With some data sets, the correlation between activity and structure was highly dependent upon the initial partition chosen and upon the order in which the data set was processed. Since DIS is widely used, simple to calculate, and performs at least as well as AVD, ESS, SIM, SHA, and VAR, while SIZ offers a quite markedly different pattern of behavior, it is suggested from the results presented here that any investigation involving relocation clustering should consider the use of at least DIS and SIZ similarity measures. Additionally, the initial partition should be fixed, and it would be worth carrying out two or three runs to determine whether the data was such as to be dependent upon the order in which the compounds are processed.

ACKNOWLEDGMENT

My thanks are due to Dr. D. Bawden (Pfizer Central Research, Sandwich, Kent) for helpful discussions of an earlier draft of this paper.

REFERENCES AND NOTES

(1) Hartigan, J. A. "Clustering Algorithms"; Wiley: New York, 1975.
(2) Spath, H. "Cluster Analysis Algorithms"; Ellis Horwood: New York, 1980.
(3) Everitt, B. "Cluster Analysis"; Heinemann: London, 1980.
(4) Dubes, R.; Jain, A. K. "Clustering Methodologies in Exploratory Data Analysis". *Adv. Comput.* **1980**, *19*, 113-228.
(5) Gordon, A. D. "Classification"; Chapman and Hall: London, 1981.
(6) Hansch, C.; Unger, S. H.; Forsythe, A. B. "Strategy in Drug Design. Cluster Analysis as an Aid in the Selection of Substituents". *J. Med. Chem.* **1973**, *16*, 1217-1222.
(7) Dunn, W. J.; Greenberg, M. J.; Callejas, S. S. "Use of Cluster Analysis in the Development of Structure-Activity Relations for Antitumour Triazenes". *J. Med. Chem.* **1976**, *19*, 1299-1301.
(8) Harrison, P. J. "A Method of Cluster Analysis and some Applications". *Appl. Stat.* **1968**, *17*, 226-236.
(9) Adamson, G. W.; Bush, J. A. "A Method for the Automatic Classification of Chemical Structures". *Inf. Storage Retr.* **1973**, *9*, 561-568.
(10) 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.
(11) Adamson, G. W.; Bawden, D. "Comparison of Hierarchical Cluster Analysis Techniques for the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 204-209.
(12) Willett, P. "A Comparison of some Hierarchal Agglomerative Clustering Algorithms for Structure Property Correlation". *Anal. Chim. Acta* **1982**, *136*, 29-37.
(13) Rubin, V.; Willett, P. "A Comparison of some Hierarchal Monothetic Divisive Clustering Algorithms for Structure Property Correlation". *Anal. Chim. Acta* **1983**, *151*, 161-166.
(14) Hansch, C.; Yoshimoto, M. "Structure-Activity Relationships in Immunochemistry. 2. Inhibition of Complement by Benzamidine". *J. Med. Chem.* **1974**, *17*, 1160-1167.
(15) Adamson, G. W.; Bush, J. A. "Evaluation of an Empirical Structure-Activity Relationship for Property Prediction in a Structurally Diverse Group of Local Anaesthetics". *J. Chem. Soc., Perkin Trans. I* **1976**, 168-172.
(16) Kier, L. B.; Murray, W. J.; Hall, L. H. "Molecular Connectivity. 4. Relationship to Biological Activities." *J. Med. Chem.* **1975**, *18*, 1272-1274.
(17) Kier, L. B.; Hall, L. H. "Molecular Connectivity. VII. Specific Treatment of Heteroatoms". *J. Pharm. Sci.* **1976**, *65*, 1806-1809.
(18) Hansch, C.; Gricco, C.; Silipo, C.; Vittoria, A. "Quantitative Structure-Activity Relationship of Chymotrypsin-Ligand Interactions". *J. Med. Chem.* **1977**, *20*, 1420-1435.
(19) DiPaolo, T. "Structure-Activity Relationships of Anaesthetic Ethers Using Molecular Connectivity". *J. Pharm. Sci.* **1978**, *67*, 565-566.
(20) Hall, L. H.; Kier, L. B. "Antimicrobial Activity of Substituted Phenyl Propyl Ethers". *J. Pharm. Sci.* **1978**, *67*, 1743-1747.
(21) Chen, B. K.; Horvath, C.; Bertino, J. R. "Multivariate Analysis and Quantitative Structure-Activity Relationships. Inhibition of Dihydrofolate Reductase and Thymidylate Synthetase by Quinazolines". *J. Med. Chem.* **1979**, *22*, 483-491.
(22) Adamson, G. W.; Bawden, D. "Automated Additive Modeling Techniques Applies to Thermochemical Property Estimation". *J. Chem. Inf. Comput. Sci.* **1980**, *20*, 242-246.
(23) Wishart, D. "CLUSTAN 1C User Manual"; Edinburgh University: Edinburgh, Scotland, 1978.
(24) Dattola, R. T. "Experiments with a Fast Algorithm for Automatic Classification". In Salton, G. "The SMART Retrieval System";

Prentice-Hall: Englewood Cliffs, NJ, 1971.

(25) Willett, P. "The Calculation of Intermolecular Similarity Coefficients Using an Inverted File Algorithm". *Anal. Chim. Acta* **1982**, *138*, 339-342.

(26) Crouch, D. "A File Organisation and Maintenance Procedure for Dynamic Document Collections". *Inf. Process. Manag.* **1975**, *11*, 11-21.

(27) Willett, P. "Document Clustering Using an Inverted File Approach". *J. Inf. Sci.* **1980**, *2*, 223-231.

Development of a Computer Language and Compiler for Expressing the Rules of Infrared Spectral Interpretation

GRAHAM M. SMITH* and HUGH B. WOODRUFF

Merck Sharp & Dohme Research Laboratories, Rahway, New Jersey 07065

Received July 18, 1983

A specialized computer language, compiler, and interpreter (CONCISE: Computer Oriented Notation Concerning Infrared Spectral Evaluation) were developed to be used to express rules for IR spectral interpretation in the PAIRS[1a] system (Program for the Analysis of Infrared Spectra). The factors considered in designing this language will be discussed as will its final structure. The nature and operation of the associated compiler and interpreter will also be described.

INTRODUCTION

An experimental computer program (PAIRS) has been developed to provide assistance in the interpretation of infrared spectra. In a previous paper[1] the overall design and application of this program was described. The intent here is to provide technical detail on the nature of the specialized language (CONCISE) developed for describing the empirical rules of infrared spectroscopy and on the compiler and interpreter for this language.

The objective of this project was threefold: first, to develop a method to extract, efficiently and consistently, structural information from infrared spectra and to present this information to the user with a calculated level of confidence; second, to provide a means for organizing and storing the empirical rules of infrared spectroscopy; third, to provide a program which is easy to use and rules which are easy to modify. The methods and principles which have been developed will be reported.

An earlier program, CASE,[2] indicated that computer-assisted infrared interpretation was a viable approach; however, it had a major drawback in that the rules were encoded in the logic of the program, which made them difficult to identify or change. A similar problem had been faced by the developers of such programs as LHASA,[3] SECS[4] (organic synthetic planning), and MYCIN[5] (medical diagnosis), and the approach used in these systems was to separate out the rules from the mechanism for interpretation and supply them as data to the programs. In addition, the data in each case were represented in the form of statements of a specialized language which allowed higher level concepts to be directly expressed. This allowed easy understanding of the rules and increased the efficiency of rule growth and correction. It was decided to use this approach in PAIRS.

The IR interpreter program (PAIRS) could have been designed to read the text form of the language directly during the interpretation process; however, this would have caused additional overhead and slowed the response time of what is an interactive program. It was decided to have an intermediate step which would convert the text form of the language into a numerically encoded form which could be read and used quickly. In addition, the text form would only have to be reconverted in the event of modification of the rules. The resulting "conversion" program is in fact a compiler and was developed by using principles of compiler design.[6] The development of this project followed five major steps: (1) Development of a language, starting with the types of information it must be able to express and progressing through exact definitions of statements in a formal grammer[7] and ultimately a parse table. (2) Development of a compiler using the parse table and a definition of the numerical codes. This was followed by design of methods for syntactic analysis and text handling. (3) Development of the main IR interpreter program. This consists of three main parts: the interaction handler, to allow the user to supply information and obtain results; the internal data storage format; the analysis section which would take the spectrum, read the rules, and produce the analysis. (4) Production of the rules as provided by various sources of IR information by beginning with an array of facts about an individual functional group and conversion of these facts into a logical procedure for encoding in the language. (5) Testing of each component in stages and as a unit to ensure that all flaws had been found.

LANGUAGE DESIGN

Language development began as discussions of the process of IR analysis with an expert in this field in order to determine the types of information needed. In the very beginning it was clear that there would have to be a facility to ask questions about peak location, intensity, and width. However, there were many features of human IR interpretation which were not clearly defined to us. It was decided to follow the process of interpretation by an experienced spectroscopist who would verbalize his thinking and stop to answer questions. This was not done to gather exact information for the rules themselves but to understand the process of interpretation. From this analysis several principles were derived about the interpretation process. We also identified a difference between the approach used by the spectroscopist and that planned for use in the PAIRS program. The difference is that on viewing an IR spectrum, various peaks immediately suggest groups to the spectroscopist for which he or she then gathers support among the remainder of the peaks from their position, intensity, and width. During the process other groups may be suggested, and the process continues to a self-consistent conclusion. This appears to be a very complex process. In the case of the computer, however, we implemented a sequential process where each set of rules for each encoded group is applied in turn to the spectrum, giving an analysis for each group. This difference in approach should ultimately not be significant. Several of the principles which emerged from this process will now be discussed. An

0095-2338/84/1624-0033$01.50/0 © 1984 American Chemical Society