Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures†

J. M. Barnard* and G. M. Downs

Barnard Chemical Information Ltd., 46 Uppergate Road, Stannington, Sheffield S6 6BX, U.K.

Received June 29, 1992

The different methods which can be used to cluster files of chemical structures and their application in recently published work and in commercially-available software are reviewed. The effects of user-defined parameters on the operation of the Jarvis–Patrick algorithm are discussed, and experience in its use is briefly described.

TYPES OF CLUSTERING ALGORITHM

Clustering is a process whereby a file of objects may be divided into several classes, the members of each class being in some way "similar" to each other and "different" from the members of other classes. A large variety of measures of similarity and difference have been devised, based both on structural and on nonstructural features of chemical structures, and are discussed extensively in other papers in this symposium, and elsewhere.[1-4]

Given some measure of (dis)similarity between chemical structures, there remain several ways in which the compounds in a file may be divided into clusters;[4-6] Figure 1 shows an hierarchical classification (or clustering) of these clustering methods. This should not be regarded as definitive and omits some distinctions (such as that between algorithms which do or do not allow overlapping clusters—i.e., compounds which can belong to more than one cluster). Indeed the classification illustrates an important point about clustering: that there is no one "correct" set of clusters for a particular set of objects.

The main distinction is between hierarchical and nonhierarchical clusters, though even here some of the algorithms normally used for nonhierarchical can be used to produce hierarchical clusters.

Hierarchical Clustering Methods. These have been used to a greater extent traditionally, though more recent work with chemical structure classifications has tended to concentrate on nonhierarchical methods. Two basic approaches can be used for the generation of hierarchical clusters: **Agglomerative** algorithms build the clusters from the bottom up, first by merging individual compounds into clusters, and then by merging clusters into superclusters, until the final merge brings all the compounds into a single cluster. **Divisive** algorithms operate in a top–down manner, successively dividing the file into smaller subsets, by binary splitting.

All the agglomerative methods use essentially the same algorithm, in which a matrix of similarity (or dissimilarity) measures between each pair of compounds is used.[7,8] The most similar pair of compounds is combined in a cluster, and the similarity matrix recalculated to take account of the new cluster, which is now regarded as a single unit; the process is repeated until all the clusters have been merged together. The difference between the methods lies in the way in which the similarity matrix is recalculated. In single-linkage clustering, for example, the similarity between a pair of existing clusters is defined to be the closest similarity between any pair of compounds from each cluster: in complete-linkage clustering it is the least similarity between any such pair. The other methods use various forms of averaging.

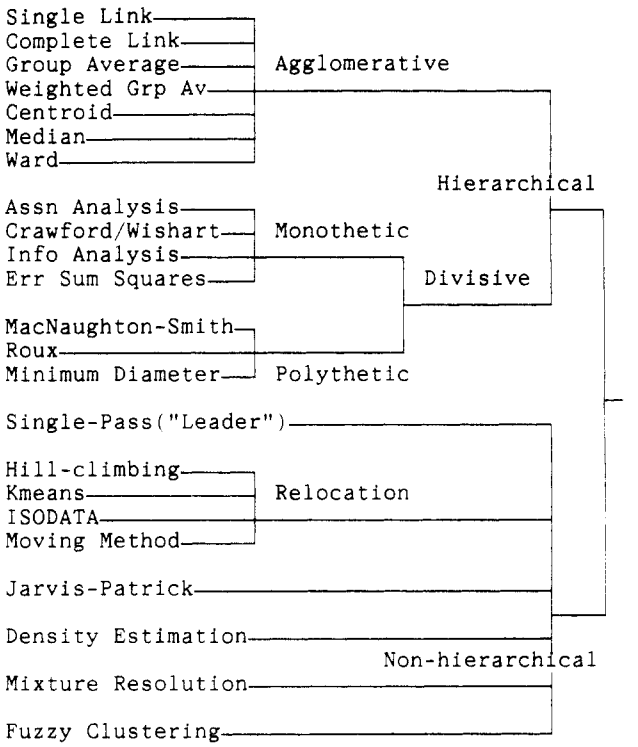


Figure 1. Classification of the major clustering methods available. Further discussion in text.

In the divisive methods, some criterion is used to separate the compounds in the file into subsets. The main problem is the identification of which of the $2^{N-1} - 1$ ways of dividing N objects into two classes is the most appropriate, since this quantity can be very large if N is other than trivially small. In **polythetic** divisive methods, the division may be made on the basis of several attributes of a compound simultaneously; the more commonly-used **monothetic** division methods are based on only a single attribute at a time. Both approaches have been studied, though the computational requirements of polythetic methods generally render them infeasible for all but the smallest data sets.[9,10] Nevertheless, a recent polythetic divisive algorithm,[11] based on analysis of the "diameters" of the clusters, is much more efficient and may find useful application in clustering of chemical structures as well as in other areas.

Nonhierarchical Clustering Methods. In these methods, the file is simply divided into a number of, normally nonoverlapping, subsets, giving a **partition** in which there are no hierarchical relationships. The simplest method for doing this is the single-pass or "leader" algorithm in which each compound is compared with the clusters formed so far and

† Paper presented at the Beilstein Institute Workshop on Similarity in Organic Chemistry, Schloss Korb, Bozen, Italy, May 25–29, 1992.

0095-2338/92/1632-0644$03.00/0

either added to the nearest or used to start (lead) a new cluster if it is insufficiently close to any of the existing clusters.

Relocation algorithms involve an initial assignment of compounds to clusters, which is then iteratively refined by moving (relocating) certain compounds from one cluster to another. Various methods have been used both for the initial assignments and as criteria for relocation.[5] A recent algorithm, the "moving method" of Zhang and Boyle,[12] appears to have a number of advantages over earlier methods in this class.

Density estimation involves plotting the compounds in an *n*-dimensional space and estimating the local density around each by counting the compounds enclosed by a hypersphere or hypercube of specified dimensions around it; local density maxima can then be used as cluster centers. The method depends rather heavily on the size of hypersphere specified and also requires a fairly high ratio of data points to dimensions. **Mixture resolution** entails making statistically-based estimates of the parameters of what is assumed to be a finite mixture of separate distributions. **Fuzzy clustering** allows objects to belong to several different clusters, with a numerical "degree of belongingness" to each.

A slightly different nonhierarchical clustering algorithm, which has received considerable attention for chemical structure applications, is that due to Jarvis and Patrick.[13] This is based on the establishment of a list of the *k* nearest-neighbors for each compound in the file and based on a rule that for two compounds to appear in the same cluster they must appear in each other's nearest-neighbor lists and have a specified number of near-neighbors in common. A related approach has also been used for agglomerative clustering by Gowda and Krishna.[14]

Choice of Clustering Algorithm. This may depend on a variety of factors. In the first place, the computing resources required (time and space) vary considerably, and some algorithms become impractical for all but the smallest data sets. For example, the hierarchical agglomerative algorithms normally require $O(N^2)$ space to store the $N \times N$ (dis)similarity matrix and $O(N^3)$ time for the clustering process itself [in addition to the $O(N^2)$ time required to generate the (dis)-similarity matrix], though more efficient algorithms are known. In contrast, nonhierarchical relocation clustering does not use a (dis)similarity matrix, and the clustering process itself may require only $O(NC)$ time, where C is the number of clusters; though again there are considerable variations between the available algorithms.

Secondly, some algorithms are inherently better at identifying certain types of cluster. The single-link hierarchical agglomerative algorithm is particularly good at identifying long stringy clusters, since a compound will join a cluster which contains only a single near-neighbor. Other algorithms will tend to identify more compact globular clusters. The question of which is "better" will depend not only on the nature of the file to be clustered but also on the purpose for which the file is being clustered. In a file of monofunctional alkanes, should the various compounds with the same functional group but different chain lengths be clustered together, or should clusters be formed for each chain length? The determination of this question will, of course, also depend on the basis for the similarity measures established within and between the homologous series.

The results obtained from some algorithms are dependent on the order in which the file is processed. This is especially true of the single-pass leader algorithm for nonhierarchical clustering, where large clusters tend to form around the first few compounds processed and much smaller ones around those occurring later in the file. A similar problem occurs with some relocation algorithms and some hierarchical agglomerative algorithms when equal-valued similarities are encountered during processing of the similarity matrix; in many cases the problem can be avoided by ensuring that the file is processed in some appropriate nonrandom order.

Validity of Clusters Formed. Some efforts have been made to find objective measures of the validity of the clusters formed. In the first place, there is the question of whether or not the clusters which can be identified are significantly different from random. This is discussed by Dubes and Jain[15] and by Willett;[16] Jain et al.[17] have studied the extent to which hierarchical agglomerative methods will find clusters in random data and found that the complete linkage method has the least tendency to do so. Lawson and Jurs[18] have suggested a modified Hopkins statistic to measure the clustering tendency of a data set, though Hodes[19] has recently expressed reservations about it. Hodes and Feldman[20] have also drawn attention to the occurrence in many clustering experiments of "singleton" clusters (clusters with only one member), which may place limits on the ability to cluster a file and upset attempts to measure the clustering tendency of the remaining structures in it.

In practice, the best measure of the validity of clusters (and hence the choice of clustering algorithm) has been the usefulness of the resulting classification, which in most experiments has meant the ability of the classification to predict property values.[1,10,21-25]

IMPLEMENTATION OF CLUSTERING METHODS

A number of groups have published work on clustering of chemical structure databases, and clustering software has been incorporated into some commercially-available packages.

Much of the early work on chemical structure clustering was done at Sheffield University,[24-26] and during the 1980s Willett and his students evaluated several different clustering algorithms on the basis of simulated property prediction.[1,10,21-23] Among their conclusions was that while the relocation methods were slightly inferior to the hierarchical ones (as measured by the correlation of predicted and observed property values), this was in part compensated for by the much more efficient algorithms, which would allow the method to be applied to much larger data sets. The Jarvis–Patrick algorithm, which is discussed in more detail in the next section, was found to perform at least as well as any other algorithm, with computational requirements between those of the hierarchical and other nonhierarchical methods.

In cooperation with Pfizer Central Research in the U.K., clustering software using the Jarvis–Patrick algorithm has been implemented and is now used operationally both for the selection of sets of representative compounds from the internal compound file and for the clustering of the output of substructure searches carried out using the SOCRATES system.[23,27,28]

Hodes has described experiments in clustering a large number of compounds from the National Cancer Institute's Repository.[20,29,30] These used a leader algorithm with a similarity measure based on a weighted fragment description of each compound, the weights taking account of the fragment's size, frequency in the database, and multiplicity of occurrence within the compound. The compounds to be clustered were first ordered by total fragment weight, which allowed the algorithm to take advantage of the inherent order-dependence of leader algorithms. Overlapping clusters were permitted (i.e., each compound joined the cluster of every leader for which it exceeded a preset similarity threshold). In order to permit the clustering of the full file of 230 000 compounds,

the whole system was implemented on a massively parallel connection machine and resulted in the identification of some 116 000 leader compounds, which could then be selected for testing.

Work on the use of clustering for the selection of compounds for testing at the Upjohn Co. has been reported by Lajiness et al.[31-34] This has involved the use of a commercially-available clustering package to form a predetermined number of nonhierarchical clusters and the selection of one compound from each for testing. This approach has been compared[34] with entirely random selection of compounds and with selection based on a computationally more demanding procedure called **maximum dissimilarity selection**, in which compounds are selected iteratively to be as dissimilar as possible from those already chosen.[32] In the comparison, which simulated the analysis of a set of compounds of known activity, it was found that maximum dissimilarity selection generally performed better in finding the active groups of compounds, especially where there was considerable natural clustering of the compounds (both active and inactive) in the file. In a file of compounds maintained by a pharmaceutical company, such clustering may be the result of analog synthesis programs.

Lawson and Jurs[35] have used nonhierarchical relocation algorithms to cluster a group of 143 acrylates from the TOSCA inventory based on a set of eight descriptors, including topological, topographical, electronic, and physicochemical properties, which were selected on the basis of a principal components analysis. Dissimilarity measurements used Euclidean distance measurements in an eight-dimensional space. Analysis of successive clustering runs, with different user-specified numbers of clusters, enabled the authors to conclude that the set fell naturally into five clusters, though aspects of their analysis have recently been criticized by Hodes.[19]

Daylight Chemical Information Systems Inc. have recently introduced a clustering package to their commercial software system (Version 4.2).[36] This uses the Jarvis–Patrick algorithm, with a variety of user-specified parameters (discussed in the next section), including an option to relocate ("rescue") singletons to the cluster containing the bulk of its nearest-neighbors (subject to a minimum). Similarities are measured using a Tanimoto coefficient derived from the folded "fingerprint" bit string used for structural characterization in the Daylight system.[37] Various output options are provided, including identification of the compound closest to the cluster centroid. It is claimed that the program will work well on up to about 250 000 compounds on a desktop workstation, though there is a warning that very long running times may be needed for generating the nearest-neighbor lists.

Though it does not include a complete clustering program, the Power Search Module of Molecular Design's MACCS-II System[38] allows similarity searches to be carried out in a database, using a specified query molecule. By use of the Customization Module, applications can be created to select a series of leader compounds from the database, and by carrying out a similarity search on each in turn, a cluster can be built up around each leader compound. Once a potentially-active leader compound is identified, a similarity search can be used to find related molecules which may have been placed in different clusters during the initial clustering.

PARAMETERIZED CONTROL OF JARVIS–PATRICK CLUSTERING

The Jarvis–Patrick algorithm[13] has been studied extensively by Willett's group at Sheffield University, where it was found to perform particularly well in clustering chemical structures on the basis of two-dimensional fragment descriptors.[23] The algorithm provides considerable scope for user control of the clustering process, which can have a marked effect on the clusters obtained.

In the first place, an $N \times k$ matrix showing the k nearest-neighbors for each of the N compounds must be constructed. Not only must a suitable value for k be selected, but it is also possible to impose a **threshold** similarity: if the similarity coefficient between two compounds lies below this threshold, they cannot be considered for inclusion in each other's nearest-neighbor lists. Normally 10 or 20 is a suitable value for k, though if the threshold similarity is high, some compounds may have less than this number of near-neighbors. An alternative approach is to include in the nearest-neighbor lists all those compounds which exceed the threshold similarity, resulting in variable-length lists. The process of building the nearest-neighbor lists involves, in principle, the calculation of a similarity coefficient for all $N(N-1)/2$ pairs of compounds in the file and, thus, has a time complexity of $O(N^2)$; substantial improvements on a "brute-force" algorithm can, however, be obtained by use of appropriate heuristics.[39] Though the establishment of the nearest-neighbor lists is relatively slow, it needs only to be done once, since a large variety of clustering runs can be done from a single nearest-neighbor table. Appropriate design of the table also allows new compounds to be added to it, without the need to recreate it from scratch.

Given the nearest-neighbor lists, the clustering process itself is very rapid. As stated earlier, two compounds will appear in the same cluster provided that

(a) they occur in each other's nearest-neighbor lists

(b) a user-specified number (k_{min}) of their respective nearest-neighbor lists are in common

Variations of the algorithm allow the omission of condition (a) or take into account the position of each near-neighbor in the nearest-neighbor list, summing the values obtained for each common near-neighbor by giving a weight to its position in each list; this sum must exceed a user-specified value for the two compounds to be included in the same cluster.

A further point where the user is able to control the clustering process is in the handling of singleton clusters—clusters with only one member. They can either be left as they are or they can be relocated to the "nearest" cluster (a variety of methods can be used to determine which this is).

The effect of the user-specified parameters on the clusters obtained can be quite marked, and we have examined some of these effects in connection with work we have been involved with, using the Jarvis–Patrick algorithm. If the threshold similarity specified in generating the nearest-neighbor lists is quite high, this will result in many compounds having fewer than k near-neighbors identified. This will cause the generation of a large number of clusters, with relatively few compounds in each, as compared with a lower threshold value. Some compounds will have no near-neighbors listed at all, and these compounds will remain unclustered. These unclustered compounds will be the "odd" ones in the file and will be both very different from each other and very different from all others in the file. The higher the threshold value, the more unclustered structures there will be.

The value of k may also be used to affect cluster sizes, though we have not so far investigated this in detail. Inspection of the algorithm suggests that a larger value of k will increase the sizes of clusters, though the extent to which this occurs will depend on the similarity threshold being used.

Increasing the value of k_{min} has the effect of subdividing existing clusters (and introducing more singletons), and reclustering the same data with various values of k_{min} allows the production of an hierarchical clustering, though this is

Table I. Effect on Number of Unclustered Structures and Number of Clusters of Changing Values for Parameters to Jarvis–Patrick Clustering Method, When Clustering Results of an Online Substructure Search

similarity threshold	k	k_{min}	unclustered structures	no. of clusters
0.9	20	20	4	60
0.9	20	19	4	24
0.9	20	18	4	18
0.9	20	17	4	12
0.9	20	15	4	10
0.9	20	10	4	9
0.9	20	8	4	8
0.9	20	5	4	7
0.8	20	20	0	64
0.8	20	19	0	23
0.8	20	18	0	16
0.8	20	16	0	8
0.8	20	15	0	7
0.8	20	10	0	4
0.7	20	18	0	31
0.7	20	15	0	6
0.6	20	18	0	30
0.6	20	15	0	7

unlikely to result in the same clusters as would be obtained by an explicitly hierarchical method. The effect of k_{min} is slightly different from that of the threshold similarity where an increase does not simply subdivide existing clusters, but may move some compounds between clusters. In experiments we have carried out, we have often found that intuitively more satisfactory clusters can be obtained by increasing the threshold slightly and reducing k_{min}; this moves some compounds into the "unclustered" set and, thus, allows slightly smaller (and more compact) clusters, without yielding unmanageable numbers of singleton clusters.

In parallel with Hodes and Feldman's conclusion, using a nonhierarchical leader algorithm,[20] we have generally found singleton relocation to be rather unsatisfactory. After all, singletons are singletons precisely because they did not meet the algorithm's criteria for being included in a cluster. If, however, the "very different" compounds are put in the unclustered set, by increasing the threshold similarity, the remaining singletons may sometimes be usefully relocated to other clusters.

Of course, equally substantial changes to the clusters formed can be obtained by redefining the similarity measure used to generate the nearest-neighbor lists or the descriptors on which it is based.

CLUSTERING OF THE ECDIN DATABASE

As subcontractors to the Department of Information Studies at the University of Sheffield, we have recently been involved in a project with the European Communities Joint Research Centre at Ispra, Italy, the U.K. Health and Safety Executive, and the Instituto Superiore di Sanita in Rome. In this, the Jarvis–Patrick algorithm has been used to cluster subsets of the ECDIN database of chemical compounds which are in widespread use in Europe.[40] Property data, and in particular toxicological information, are available for only a small fraction of the 100 000 or so compounds in the database. The purpose of clustering them is to reveal any natural groupings in the database, to identify groups of compounds which may be expected to have similar properties and which may be suitable for more rigorous QSAR techniques, and to allow selection of representative compounds.

An initial study was done on about 750 compounds, for which various physical property values were available. Structural descriptors were derived automatically from connection

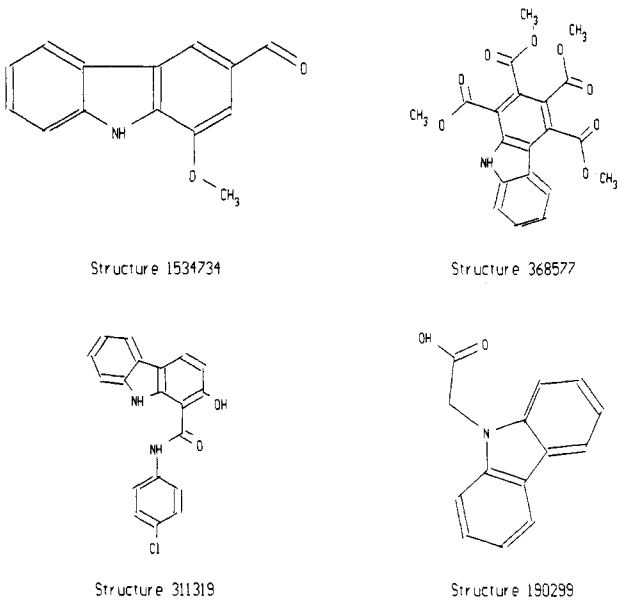


Figure 2. Four "unclustered" compounds from the set of 64 retrieved by substructure search.

tables for the compounds and included augmented atom, atom sequence, and bond sequence fragments of the type used in the Chemical Abstracts Service's substructure search system,[41] along with various ring system descriptors,[42] and a group of CROSSBOW[43] fragments, which mainly represent functional groups. Bit strings for each compound, showing the presence or absence of each fragment, were used to calculate Tanimoto coefficient similarity measures between pairs of compounds; earlier work by Willett[4] had suggested that the Tanimoto coefficient was as effective a similarity measure as any other, and easier to calculate than many.

Several different clustering runs were done, using different clustering parameters, and satisfactory correlations were obtained between observed property values and those predicted by taking the average property value for other compounds in the same cluster.[44] One significant result was that different clustering parameters were needed to give the best correlations for different properties, thus emphasizing the point made earlier concerning the absence of a "correct" set of clusters for a particular data set. Another conclusion from this study was that the "weighted" version of the algorithm did not give better results than the unweighted version; this was in contrast to the results of an earlier study with a larger data set.[23]

The results of this work were sufficiently encouraging to lead to a further study with a subset of 10 700 compounds, though the absence of sufficient property values in the database has not enabled the clusters formed to be evaluated on the same basis. The clusters formed are now being evaluated and compared with intellectually-based classifications of the same compounds, which have been made in connection with structure–activity analysis techniques being developed by Tosato et al.[45] Meanwhile the clustering software, along with graphics utilities for browsing through the clusters formed, has been installed at Ispra, with enhancements by the staff there enabling efficient implementation on workstations. The aim is to cluster all compounds from the database for which connection tables are available, for the purposes of property prediction, and to form a basic classification of the database.

A critical question in these studies is the selection of the structural descriptors used in the calculation of similarity measures. If cluster analysis is to be used for property prediction, fragment descriptors of the type developed for substructure searching, or at least the dictionaries of them, may not be appropriate. In principle, the clustering methods

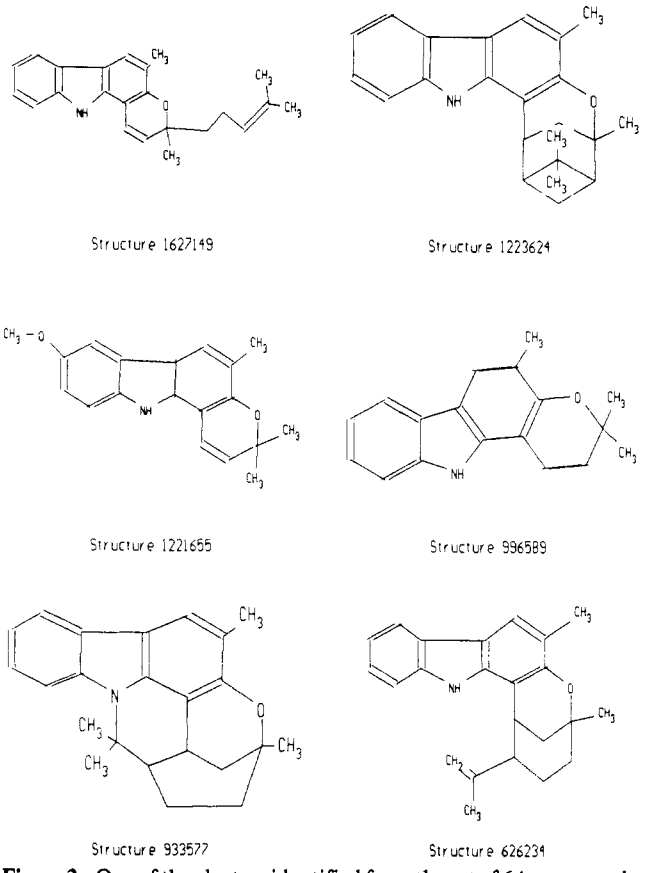


Figure 3. One of the clusters identified from the set of 64 compounds retrieved by substructure search. All compounds contain a six-membered ring with an oxygen atom fused in the same position to the target substructure, though there are differences in the degree of unsaturation and in the occurrence of additional fused rings. No other compounds from the retrieved set contained such a ring system.

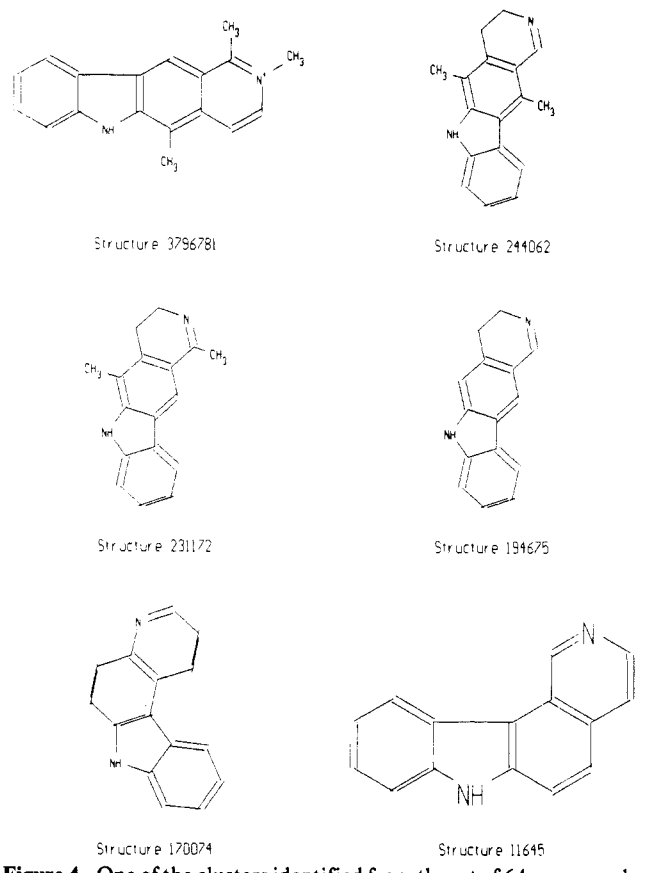


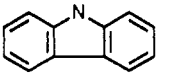
Figure 4. One of the clusters identified from the set of 64 compounds retrieved by substructure search. All compounds contain a six-membered ring with a nitrogen atom fused to the target substructure, though the fusion position, heteroatom position, and degree of unsaturation all differ. No other compounds from the retrieved set contained such a ring system.

described here are applicable with any quantitative similarity or distance measure between objects. Recent work at Sheffield University has concentrated on the use of three-dimensional fragments,[46,47] and clustering using similarity measures based on quantitative property values is currently under investigation in association with Chemical Abstracts Service.[48]

CLUSTERING OF SUBSTRUCTURE SEARCH OUTPUT

Another area in which structure-based clustering may be useful is in the analysis of the output of substructure searches of large databases; as mentioned above, a system to do this was implemented at Pfizer Central Research some years ago.[23] Even quite specific searches in databases of several million compounds, such as the CA Registry and Beilstein files, may result in several hundred hits, and clustering of these can allow the identification of the major structural types present. The clustering parameters available to control the operation of the Jarvis–Patrick algorithm can be adjusted to ensure that a suitable partition of the file is obtained, even when all the compounds contain a large common substructure.

In the implementation of the Beilstein database for substructure searching on the Dialog system,[49] the structures retrieved are transmitted to the user in the form of ROSDAL strings,[50] which are compact linear connection tables. If captured on a personal computer, the structures may be used to generate structural descriptors as described above, which can then be used to cluster the compounds using various parameter settings for the Jarvis–Patrick method. For example, a search of the Beilstein ONTAP File on Dialog (which contains some 50 000 substances selected from the full database) using the following substructure (any substitution permitted)



yielded 64 hits, for which the ROSDAL strings were captured on a PC. Table I shows the numbers of clusters and unclustered compounds for various settings of the Jarvis–Patrick parameters. In some cases, though a convenient number of clusters was formed, they were not very satisfactory as they consisted of one or two very large clusters and several very small ones. Examination of the clusters formed for different parameter settings suggested that the most satisfactory clusters (from an intuitive point of view) were obtained with a similarity threshold value of 0.9 and a k_{min} of 15. The four structures left unclustered (because they did not have a similarity coefficient of 0.9 with any other structures) are shown in Figure 2. The ten clusters formed included the following:

- Six compounds in which the target substructure was fused to a six-membered ring containing a nitrogen atom (Figure 3).
- Six compounds in which the target substructure was fused to a six-membered ring containing an oxygen atom (Figure 4).
- Eight compounds in which the target substructure had various aldehyde and ketone substituents.
- Nine compounds in which the target substructure had various nitro substituents.

Because all the retrieved compounds contained a large common substructure, almost all had a high level of similarity with the other compounds in the set. This meant that a high

similarity threshold was needed to achieve satisfactory clusters; an even better partitioning might have been obtained using a threshold value of 0.95, though the present version of the software only allows the threshold to be incremented in steps of 0.1. Work is currently in hand to develop automatic means of evaluating the usefulness of the different sets of clusters formed, so that the program itself may select the most appropriate parameters for a particular data set.

ACKNOWLEDGMENT

We thank the organizers of the Beilstein Workshop on Similarity in Organic Chemistry for the invitation to present this paper. We also thank the following for providing information about their work on chemical similarity and clustering: Dr. W. G. Town (for Daylight Chemical Information Inc.), Drs. D. Hounshell and G. Grethe (Molecular Design Ltd.), Mr. O. Norager and Drs. B. Hansen and W. Karcher (EC JRC, Ispra), Dr. P. Walsh (Health and Safety Executive), and Drs. M. Johnson and M. Lajiness (Upjohn). Thanks are also due to the Beilstein Institute for permission to reproduce the structures shown in Figures 2–4 and to Professor Peter Willett (Sheffield University) for helpful comments and discussions.

REFERENCES AND NOTES

(1) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, 1987.
(2) Johnson, M. A.; Maggiora, G. M., Eds. *Concepts of Molecular Similarity Analysis*; Wiley Interscience: New York, 1990.
(3) Sneath, P. H. A.; Sokal, R. R. *Numerical Taxonomy*; Freeman: San Francisco, 1973.
(4) Willett, P.; Winterman, V. A comparison of some measures for the determination of inter-molecular structural similarity. *Quant. Struct.–Act. Relat.* **1986**, *5*, 18–25.
(5) Dubes, R.; Jain, A. K. Clustering methodologies in exploratory data analysis. *Adv. Comput.* **1980**, *19*, 113–228.
(6) Zupan, J. *Clustering of Large Data Sets*; Research Studies Press: Letchworth, 1982.
(7) Lance, G. N.; Williams, W. T. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Comput. J.* **1967**, *9*, 373–380.
(8) Podani, J. New combinatorial clustering methods. *Vegetatio* **1989**, *81*, 61–77.
(9) Roux, M. Basic procedures in hierarchical cluster analysis. In *Applied Multivariate Analysis in SAR and Environmental Studies*; Devillers, J., Karcher, W., Eds.; Kluwer: Brussels, 1991; pp 115–135.
(10) Rubin, V.; Willett, P. A comparison of some hierarchical monothetic divisive clustering algorithms for structure–property correlation. *Anal. Chim. Acta* **1983**, *151*, 161–166.
(11) Guénoche, A.; Hansen, P.; Jaumard, B. Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *J. Classif.* **1991**, *8*, 5–30.
(12) Zhang, Q.; Boyle, R. D. A new clustering algorithm with multiple runs of iterative procedures. *Pattern Recognit.* **1991**, *24*, 835–848.
(13) Jarvis, R. A.; Patrick, E. A. Clustering using a similarity measure based on shared nearest neighbours. *IEEE Trans. Comput.* **1973**, *C-22*, 1025–1034.
(14) Gowda, K. C.; Krishna, G. Agglomerative clustering using the concept of mutual nearest neighbourhood. *Pattern Recognit.* **1978**, *10*, 105–112.
(15) Dubes, R.; Jain, A. K. Validity studies in clustering methodologies. *Pattern Recognit.* **1979**, *11*, 235–254.
(16) Willett, P. Clustering tendency in chemical classifications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 78–80.
(17) Jain, N. C.; Indrayan, A.; Goel, L. R. Monte Carlo comparison of six hierarchical clustering methods on random data. *Pattern Recognit.* **1986**, *19*, 95–99.
(18) Lawson, R. G.; Jurs, P. C. New index for clustering tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 36–41.
(19) Hodes, L. Limits of classification. 2. Comment on Lawson and Jurs. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 157–166.
(20) Hodes, L.; Feldman, A. Clustering a large number of compounds. 3. The limits of classification. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 347–350.
(21) Willett, P. A comparison of some hierarchical agglomerative clustering algorithms for structure-property correlation. *Anal. Chim. Acta* **1982**, *136*, 29–37.
(22) Willett, P. Evaluation of relocation clustering algorithms for the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 29–33.
(23) Willett, P.; Winterman, V.; Bawden, D. Implementation of nonhierarchic cluster analysis methods in chemical information systems: selection of compounds for biological testing and clustering of substructure search output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109–118.
(24) Adamson, G. W.; Bush, J. A. A method for the automatic classification of chemical structures. *Inf. Storage Retr.* **1973**, *9*, 561–568.
(25) 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.
(26) 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.
(27) Bawden, D. Browsing and clustering of chemical structures. In *Chemical structures: the international language of chemistry (Proceedings of an international conference at the Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, 31 May–4 June 1987)*; Warr, W. A., Ed.; Springer: Heidelberg, 1988; pp 145–150.
(28) Bawden, D. Applications of Two-dimensional chemical similarity measures to database analysis and querying. In *Concepts of Molecular Similarity Analysis*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley Interscience: New York, 1990; pp 65–76.
(29) Hodes, L. Clustering a large number of compounds. 1. Establishing the method on an initial sample. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 66–71.
(30) Whaley, R.; Hodes, L. Clustering a large number of compounds. 2. Using the connection machine. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 345–347.
(31) Johnson, M.; Lajiness, M.; Maggiora, G. Molecular similarity: a basis for designing drug screening programs. In *QSAR: Quantitative Structure–Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.: New York, 1989; pp 167–171.
(32) Lajiness, M.; Johnson, M. A.; Maggiora, G. M. Implementing drug screening programs using molecular similarity methods. In *QSAR: Quantitative Structure–Activity Relationships in Drug Design*; Fauchere, J. L., Ed.; Alan R. Liss Inc.: New York, 1989; pp 173–176.
(33) Lajiness, M. S. Molecular similarity-based methods for selecting compounds for screening. In *Computational Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science Publishers: New York, 1990.
(34) Lajiness, M. S. An evaluation of the performance of dissimilarity selection. In *QSAR: Rational Approaches to the Design of Bioactive Compounds*; Silipo, C., Vittoria, A., Eds.; Elsevier: Amsterdam, 1991.
(35) Lawson, R. G.; Jurs, P. C. Cluster analysis of acrylates to guide sampling for toxicity testing. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 137–144.
(36) Daylight Chemical Information Systems, Inc., 18500 Von Karman Ave., Suite 450, Irvine, CA 92715.
(37) Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.
(38) Molecular Design Ltd., 2132 Farallon Dr., San Leandro, CA 94577.
(39) Willett, P. Some heuristics for nearest-neighbor searching in chemical structure files. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 22–25.
(40) Norager, O. Poster Session: ECDIN, Environmental Chemicals Data and Information Network. In *Chemical structures: the international language of chemistry (Proceedings of an international conference at the Leeuwenhorst Congress Center, Noordwijkerhout, The Netherlands, 31 May–4 June 1987)*; Warr, W. A., Ed.; Springer: Heidelberg, 1988; pp 195–209.
(41) Dittmar, P. G.; Farmer, N. A.; Fisanick, W.; Haines, R. C.; Mockus, J. The CAS Online Search System. 1. General system design, and selection, generation and use of search screens. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 93–102.
(42) Downs, G. M.; Gillet, V.; Holliday, J. D.; Lynch, M. F. Theoretical aspects of ring perception, and development of the extended set of smallest rings (ESSR) concept. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 187–206.
(43) Eakin, D. The ICI CROSSBOW System. In *Chemical Information Systems*; Ash, J. E., Hyde, E., Eds.; Ellis Horwood: Chichester, 1975; Chapter 14.
(44) Downs, G. M.; Willett, P. The use of similarity and clustering techniques for the prediction of molecular properties. In *Applied Multivariate Analysis in SAR and Environmental Studies*; Devillers, J., Karcher, W., Eds.; Kluwer: Brussels, 1991; pp 247–279.
(45) Tosato, M. L.; Marchini, S.; Passerini, L.; Pino, A.; Eriksson, L.; Lindgren, F.; Hellberg, S.; Jonsson, J.; Sjostrom, M.; Skagerberg, B.; Wold, S. QSARs based on statistical design and their use for identifying chemicals for further biological testing. *Environ. Toxicol. Chem.* **1990**, *9*, 265.
(46) Pepperell, C. A.; Willett, P. Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455–474.
(47) Willett, P. Similarity searching in databases of three-dimensional molecules and macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, in this issue.
(48) Fisanick, W.; Cross, K. P.; Lillie, D. H.; Lipkus, A. H.; Rusinko, A. A comparison of similarity searching on 2D, 3D, and molecular property data for CAS registry substances. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, in this issue.
(49) Hartwell, I. O.; Haglund, K. A. An overview of Dialog. *ACS Symp. Ser.* **1990**, *No. 436*, 42–63.
(50) Barnard, J. M.; Jochum, C. J.; Welford, S. M. ROSDAL: a universal structure/substructure representation for PC-host communication. In *Chemical structure information: interfaces, communication and standards*; Warr, W. A., Ed.; ACS Symposium Series No. 400; American Chemical Society: Washington, DC, 1989; pp 76–81.