Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information Systems: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output

PETER WILLETT* and VIVIENNE WINTERMAN
Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, England

DAVID BAWDEN
Research Information Services, Pfizer Central Research, Sandwich, Kent CT13 9NJ, England

Received February 7, 1986

The use of cluster analysis techniques with machine-readable files of chemical compounds characterized by substructural fragments is reported. The first part of the paper describes a comparative evaluation of several nonhierarchic clustering methods, the evaluation involving simulated property prediction experiments using 14 small sets of compounds for which associated physical, chemical, or biological property data are available. The experiments suggest that one particular nonhierarchic clustering method, that due to Jarvis and Patrick, performs consistently better than the other methods that were tested. The second and third parts of the paper consider the use of the Jarvis–Patrick method for clustering data sets of many hundreds or thousands of chemical compounds, typical of the types of structure encountered in the chemical information systems of pharmaceutical research organizations. The first application considered is the selection of trial compounds for activity testing in biological screens, thus providing a more systematic method of compound selection than is used in current empirical screening systems. The second application is the clustering of the molecules retrieved by chemical substructure search systems so as to permit rapid identification of the main structural classes present in the search output. These applications have been tested and implemented in an industrial pharmaceutical research environment.

INTRODUCTION

Cluster analysis, or automatic classification, is the name given to a range of methods that can be used to detect groupings present within multivariate data sets. Cluster analysis methods were first used extensively for the study of biological species but are now being applied to a wide range of subject areas where there is a need to identify and group together items that are similar to each other: good reviews of the field are given by Everitt,[1] by Dubes and Jain,[2] and by Gordon.[3] Cluster analysis methods have been extensively used for the clustering of a range of types of chemical entities:[4] in this paper, we are concerned specifically with the clustering of compounds characterized by small substructural fragments.

An early example of such work is that of Harrison,[5] who described the use of clustering techniques to group compounds using a probabilistic measure of intermolecular structural similarity. It was found that the groups of structurally related compounds that were identified also corresponded well with the activities observed in biological tests, and it was suggested that such an approach could form the basis for a nonempirical method of selecting compounds for activity testing; a development of the probabilistic model has been reported by White and Lewinson.[6] Adamson and Bush[7] reported the use of the single-linkage clustering method to classify the 20 naturally occurring amino acids and evaluated the classifications by means of simulated property prediction experiments using the amino acid pK_a values; a similar approach was adopted in an evaluation of similarity measures for the clustering of 39 structurally diverse compounds with local anaesthetic activity.[8] The single-linkage method produces binary tree-like classifications and is characteristic of a range of *hierarchic* clustering methods in which small clusters of very similar molecules are nested within larger and larger clusters of less closely related molecules. Such clustering methods can be either *agglomerative* or *divisive* in nature. Hierarchic agglomerative classifications are produced in a bottom-up manner by the fusion of individual compounds into clusters and then the fusion of these clusters into larger clusters, the process finally resulting in a single cluster that represents the entire data set. Adamson and Bawden[9] and Willett[10] have described comparative studies using a range of hierarchic agglomerative methods and showed that the single-linkage method, which has many theoretical advantages, is by no means the most effective method for grouping chemical structures. Divisive classifications are constructed by a top-down procedure that progressively subdivides a data set into smaller and smaller subsets: the use of such methods for the clustering of chemical compounds has been studied by Rubin and Willett.[11]

Although the hierarchic methods are very widely used, nonhierarchic classifications may also be generated in which a data set is partitioned into a set of (generally) nonoverlapping groups having no hierarchical relationships between them. Willett[12] studied the use of nonhierarchic relocation methods for the clustering of chemical molecules and found that they gave results in simulated property prediction tests that were comparable with those given by the hierarchic agglomerative methods. In many applications, such as those discussed later in this paper, there is no obvious need for a hierarchic structuring of the data; moreover, nonhierarchic methods are generally much less demanding of computer resources than are hierarchic methods, and these would thus seem to be more appropriate for the clustering of large files of chemical structures, such as those encountered in chemical industry.

In this paper, we present the results of a project to investigate the use of nonhierarchic clustering methods to enhance the facilities available in computerized chemical information systems. In the next section we describe an empirical comparison of nonhierarchic clustering methods based upon simulated property prediction experiments using 14 small sets of compounds for which associated property data are available. The results of these tests suggest that one method, that due to Jarvis and Patrick, is noticeably superior to the others tested, and in the second half of the paper we describe the implementation of this method in the chemical information system of a major pharmaceutical organization. The first application considered is the selection of trial compounds for activity testing in biological screens, this providing a more systematic method of selection than is used in many current empirical screening systems. The second application is the clustering of the molecular structures retrieved by chemical substructure

0095-2338/86/1626-0109$01.50/0

searches, so as to permit a rapid identification of the main types of compound that are present in the output of a search. The paper concludes with a summary of our findings.

EVALUATION OF NONHIERARCHIC CLUSTERING METHODS USING SIMULATED PROPERTY PREDICTION EXPERIMENTS

Methodology. There are many types of nonhierarchic clustering methods, and new methods are constantly being reported in the literature: good reviews are given by Everitt[1] and Dubes and Jain[2] while detailed discussions of algorithms are presented by Hartigan[13] and Spath.[14] This project has studied three main classes of clustering method; in order of increasing computational complexity, these are *single pass* clustering, *relocation* clustering, and *near-neighbor* clustering. The comparative evaluation involves simulated property prediction experiments using sets of compounds for which associated property data are available.[7,8,10-12]

In all, 14 data sets were selected from the structure–activity literature for the evaluation. The data sets are those used in an earlier evaluation of measures for the determination of intermolecular structural similarity,[15] with the exception of two data sets that have subsequently been shown to exhibit clustering behavior that is not significantly different from that of randomly generated data.[16] The data sets are as follows:

(A) p*I* values of 20 naturally occurring amino acids

(B) local anaesthetic activity of 37 structurally diverse compounds

(C) inhibition of ribonucleotide reductase by 28 substituted benzohydroxamic acids

(D) toxicity to mice of 25 aliphatic and carbocyclic ethers

(E) chromatographic retention indices for 56 aliphatic alcohols, ketones, ethers, and esters

(F) inhibition of complement by 105 benzamidines

(G) inhibition of monoamine oxidase by 24 hydrazides

(H) chymotrypsin hydrolysis of 72 *N*-acyl esters

(I) serum binding activity of 79 penicillins

(J) inhibition of dihydrofolate reductase by 46 quinazolines

(K) inhibition of chlorphentermine binding in rat lung by 20 structurally diverse compounds

(L) molar refractivity of 65 aliphatic ethers, amines, alcohols, and halides

(M) tadpole narcosis activity of 34 structurally diverse compounds

(N) heats of vaporization for 129 alkanes, alkenes, alcohols, ketones, and benzene and pyridine derivatives

It will be seen that the data sets span a wide range of chemical, physical, and biological properties and include both homogeneous and heterogeneous sets of compounds. The assumption is made that the use of such a wide range of types of compound and property will ensure that the results are of some generality and not conditioned by the characteristics of a particular data set.

The compounds in each data set were encoded in Wiswesser line notation and the notations converted automatically to redundant connection tables: in some cases, the number of compounds listed is marginally smaller than the original data set, the differences arising from limitations in the connection table software available to us. The connection tables were then used for the generation of lists of fragment substructures: these lists provided the numerical characterization of each of the compounds for clustering. The substructure used in the experiments reported in this section was the *augmented atom* fragment, which consists of a non-hydrogen atom together with the atoms and bonds that are immediately adjacent to it. Such a fragment was generated centered upon each atom within a structure, and then a record was created containing the frequencies of occurrence of each fragment type present. Assume that a total of F different augmented atom types is identified during the characterization of a data set containing N compounds. Then, the basic data matrix D, which formed the input to the various clustering procedures, is an $N \times F$ integer array, in which the element $D[I,J]$ contains the frequency of occurrence of the Jth augmented atom in the Ith compound. The data was not standardized since previous work has suggested that this is not appropriate for fragment-occurrence data.[9,10]

The use of any clustering method implies the use of some quantitative definition of interobject similarity, specifically the similarity between pairs of structures or between a structure and a cluster. Recent work on similarity measures for chemical information systems[15,17] has shown that the Tanimoto coefficient performs well with chemical compounds characterized by fragment substructures, and this measure was used for all of the work reported in this paper. The value for the Tanimoto coefficient measure of similarity between two structures K and L is given by

$$\frac{\sum D[K,J]D[L,J]}{\sum D[K,J]^2 + \sum D[L,J]^2 - \sum D[K,J]D[L,J]}$$

where the summations are from $J = 1$ to F. The experiments reported in the later sections of the paper used a binary fragment representation in which each molecule is represented by the presence or absence of each fragment, i.e., $D[I,J] = 0$ or 1 only. In this case, the expression can be simplified; for two molecules containing A and B nonzero fragment occurrences, C of which are common, the Tanimoto coefficient is given by

$$C/(A + B - C)$$

The effectiveness of the classifications for property prediction was assessed with a "leave-one-out" approach. For each molecule I in a data set of size N, the set of clusters resulting from some classification procedure was scanned to identify the cluster containing I. The predicted property value for compound I, $P[I]$, was then set equal to the mean of the observed property values, $O[J]$, for each of the molecules J ($J \neq I$) in the chosen cluster. The overall measure of agreement between the N observed and predicted property values, i.e., the success of the method in grouping related chemical structures, was determined by means of the product moment correlation coefficient:

$$\frac{\sum(P[I] - \mu_P)(O[I] - \mu_O)}{[\sum(P[I] - \mu_P)^2 \sum(O[I] - \mu_O)^2]^{1/2}}$$

where μ_O and μ_P are the means of the observed and predicted values and where the summations are from $I = 1$ to N. The significance of the resulting coefficients may then be determined with a t test.

Implementation of Clustering Methods. The first procedure tested was the single-pass clustering method, hereafter SPCM, in which, as the name implies, the objects are clustered in the course of a single pass through the data. An important feature of this, and many other nonhierarchic clustering methods, is the use of a *cluster centroid*, this being some sort of average description of the set of structures contained within some cluster. The basic single-pass algorithm is as follows, with SIM being a user-defined threshold similarity:

(1) the first molecule becomes the first cluster

(2) the next molecule is compared with all current cluster centroids to identify the most similar one; call the resulting similarity MAXSIM

(3) if MAXSIM > SIM, the molecule joins the corresponding cluster and the centroid is recomputed; otherwise, it initiates a new cluster

(4) go to (2) if there are still molecules to be processed

The algorithm is clearly very fast in operation, the time complexity being governed primarily by the numbers of clusters that are formed, this in turn being determined by the threshold SIM. If SIM is low, it will be easy for a new molecule to join an existing cluster, and only a few clusters will be created during the processing of a data set. A high value for SIM, conversely, will result in the creation of large numbers of small clusters. The expected time requirement for the clustering of a file of N compounds is proportional to MN, where M reflects the number of clusters formed.

The SPCM suffers from being totally dependent upon the order in which the compounds are processed, with clusters identified early having the opportunity to become very much larger than those identified near the end of the file; several procedures have been suggested to minimize these limitations.[18] Such methods have been used extensively in information retrieval,[19,20] where very large data sets need to be handled, and to specify the initial centroids for relocation clustering.[13]

The centroid definition that is generally used to characterize an individual cluster is the arithmetic mean of the fragment lists for all of the compounds contained within that cluster. Such a definition is appropriate to compounds that are characterized by fragment lists containing the actual frequencies of occurrence of each fragment within a structure, and this definition forms the basis for the simulated property prediction experiments. However, it can be quite time consuming, as well as extremely demanding of internal storage, to calculate such cluster centroids when very large numbers of molecules need to be clustered. A simpler procedure involves the creation of a centroid by selecting those fragment substructures that occur in more than some threshold number of compounds within a cluster: such an approach is used to generate centroid bit strings for the large-scale clustering tests reported later in this paper.

Initial tests showed, as expected, that the efficiency and the effectiveness of the clustering was crucially dependent upon the threshold similarity, SIM, that was used. In particular, high values for SIM often resulted in large numbers of singletons, i.e., clusters containing only a single compound, and long run times owing to the number of molecule-centroid similarity calculations required. It was decided to choose that value of SIM for each data set that gave a final number of clusters approximately equal to $N/4$; this number was chosen arbitrarily but has been found to yield acceptable results in all of our experiments. Singleton clusters were then eliminated by assigning each singleton to its most closely related cluster, i.e., the cluster with a centroid most similar to the fragment list describing the molecule. This assignment procedure was adopted to eliminate singletons resulting from all of the different clustering methods described here.

A theoretical basis for a SPCM has been described by Can and Ozkarahan.[21] Although rather more complicated than the basic single-pass algorithm, this method provides a rational approach to the selection of the cluster centers and also eliminates the order dependence inherent in most SPCMs. In essence, the method involves an analysis of the data to identify those molecules that have the greatest numbers of discriminating fragments in common with other compounds. Such molecules can then be used to select the desired number of cluster centroids, $N/4$ in this case, by selecting the first $N/4$ compounds after ranking to act as the cluster centers. It will thus be realized that the Can–Ozkarahan clustering method, hereafter COCM, involves the identification of the initial clusters prior to clustering, as with the relocation methods described below, but involves only a single processing of the data, as with SPCMs.

A relocation clustering method, hereafter RCM, involves the repeated assignment of compounds to clusters and has proved to be a popular and effective means for grouping data. A simple relocation algorithm is that described by Forgy:[22]

(1) select some number of clusters, M

(2) initialize these clusters by selecting M molecules at random and assigning one of them to each cluster to act as the cluster centroid

(3) assign each of the molecules to the most similar cluster centroid

(4) calculate all of the resulting centroids

(5) if none, or only some small number, of the molecules have changed their assignment during the previous pass then stop; otherwise, go to (3)

An inspection of this algorithm shows that its time requirement is again proprotional to MN, as with SPCMs; however, relocation clustering is normally more demanding of computational resources owing to the need to process the entire data set several times. The name relocation clustering comes from the repeated assignments as steps 3 and 4 are iterated, the iteration implying that molecules may be transferred from one cluster to another as the classification proceeds.

The use of an RCM requires some means for the specification of M, and several ways of doing this have been suggested in the literature;[23] in this work, we have chosen to use $N/4$ initial clusters as with the SPCM. A limitation of the basic Forgy algorithm is the use of random selection in step 2, this implying that different final partitions may be obtained from different initial assignments. Several means have, hence, been suggested for nonrandom assignments that would eliminate this arbitrary characteristic of relocation clustering, and two of these approaches were tested here. The first of these, hereafter RCM_1, uses the clusters resulting from the SPCM to give the centroids, which formed the basis for the subsequent relocation. Alternatively, the centroids for the relocation were derived by the procedure described by Hartigan and Wong.[24] In this, the centroid of the entire data set is calculated, and then, the individual structures are ranked in order of decreasing similarity from this centroid. A set of $N/4$ initial cluster seeds may then be obtained by one of the two following procedures: the first four compounds in the ranked list were allocated to the first cluster, the second four compounds to the second cluster, etc.; the first $N/4$ compounds were used as the cluster seeds. These two procedures will be referred to as RCM_2 and RCM_3, respectively.

The third, and most computationally demanding, group of nonhierarchic clustering methods derives from the work of Jarvis and Patrick[25] and involves the use of a near-neighbor table. This is an $N \times K$ integer array, NNTABLE, containing the identifiers for the K nearest neighbors for each of the N molecules, with NNTABLE$[I,J]$ containing the identifier for the Jth nearest neighbor of the Ith molecule. The nearest-neighbor for some molecule I is that compound which is most similar to I under the chosen definition of intermolecular similarity.[26] Jarvis and Patrick suggest that two objects, I_1 and I_2, should be placed in the same cluster if I_1 is a near neighbor of I_2, I_2 is a near neighbor of I_1, and I_1 and I_2 share at least S near neighbors in common, where S is a user-defined parameter. Although the procedure is designed to create nonhierarchic classifications, a progressive incrementing of S allows the production of a hierarchic sequence of partitions, an idea that has been developed in subsequent work.[27,28] The time requirement of the Jarvis–Patrick clustering method, hereafter JPCM, is proportional to N^2 if it is assumed that, as is normally the case, each of the nearest-neighbor searches has a time requirement proportional to N. The JPCM is thus more demanding of computer resources than SPCMs and RCMs: however, the near-neighbor table can be generated without too much difficulty by the efficient nearest-neighbor search algorithms that have been developed recently.[4,26,29]

Table I. Correlation Coefficients for Simulated Property Prediction Using SPCM and RCM_1

	SPCM		RCM_1	
data set	mean	SD	mean	SD
A	*[a]	*	*	*
B	0.83	0.11	0.78	0.10
C	0.67	0.15	0.68	0.14
D	0.77	0.24	0.73	0.37
E	0.68	0.07	0.71	0.03
F	0.91	0.02	0.91	0.02
G	0.76	0.05	0.75	0.06
H	0.94	0.02	0.94	0.02
I	0.79	0.06	0.78	0.06
J	0.61	0.00	0.61	0.00
K	0.68	0.15	0.72	0.08
L	0.80	0.07	0.80	0.07
M	0.81	0.09	0.78	0.10
N	0.87	0.04	0.87	0.05

[a] An asterisk denotes insignificant value (see text).

Two means are available for implementing the JPCM. $JPCM_1$ is the simple procedure described above, where a pair of compounds that occurs in each other's NNTABLE must share at least S near neighbors in common if they are to be grouped into the same cluster. This *unweighted* approach assumes that all of the near neighbors are of equal importance. Alternatively, the *weighted* approach, $JPCM_2$, takes account of the relative orderings of the neighbors in the two near-neighbor lists. If some molecule occurs at the Pth position in one list and the Qth position in the other and each such list contains NUM near neighbors, then the contribution of that molecule to S, the overall degree of similarity between I and J, is given by $(\mathrm{NUM} + 1 - P)(\mathrm{NUM} + 1 - Q)$. Thus, if NUM = 20 and if the same compound occurs at the first and fifth positions in the near-neighbor lists of I and J, a contribution of 320 is made to S.

For both variants of the basic JPCM, a range of different thresholds was tested, and those that gave a final number of clusters that was approximately equal to $N/4$ was selected so as to allow comparison with the other methods that were being tested.

Results. The correlation coefficients between the observed and predicted property values are listed in Tables I and II: asterisks denote correlation coefficients that were not significant at the 0.05 level of statistical significance.

Table I contains the results obtained from the use of the SPCM and from the use of RCM_1, the relocation method with an initial partition deriving from the SPCM. As noted previously, these methods result in classifications that are dependent on the ordering of the compounds in the data set that is being clustered. To determine the degree of order dependence, a total of five classifications was generated, each with a different ordering of the compounds, and the correlation coefficients listed are mean values averaged over the five runs; also listed are the corresponding standard deviations. An asterisk is listed if one or more of the coefficients was not significant. The results suggest that no obvious benefit accrues from implementing a relocation procedure after the SPCM method since there is no substantial difference between the two sets of results. Moreover, many of the standard deviations are quite large, this implying that the final classifications are strongly dependent upon the order in which the set of compounds is processed: examples of this behavior are the benzohydroxamic acid (C) and mouse toxicity (D) data sets. Such a strong dependence is quite unsatisfactory for any practical clustering system since it implies that a superior classification might be obtained simply by a reordering and reclustering of the data.

Table II contains the correlation coefficients for the COCM, RCM_2, RCM_3, $JPCM_1$, and $JPCM_2$ experiments. All of these methods produce classifications that are invariant under a reordering of the data set, and thus, only a single coefficient is reported in the table. Considering the COCM, RCM_2, and RCM_3 results first, the theoretically based COCM frequently gives results inferior to those obtained with the other two methods, and in some cases, the difference is very marked, e.g., the amino acid (A), penicillin (I), and tadpole narcosis (M) data sets inter alia. Of the two RCM methods, RCM_2 usually gives rather better results than the alternative RCM_3 procedure. With one or two exceptions, there is little difference between the results obtained from the use of the weighted and the unweighted versions of the Jarvis-Patrick clustering method, and in general, these classifications gave the best overall levels of predictive performance across the 14 data sets. They would thus seem to be the most appropriate of the methods tested for the clustering of chemical compounds characterized by fragment descriptors: this conclusion forms the basis for the work reported in the remainder of this paper.

Table II. Correlation Coefficients for Simulated Property Prediction Using COCM, RCM_2, RCM_3, $JPCM_1$, and $JPCM_2$

data set	RCM_2	RCM_3	COCM	$JPCM_1$	$JPCM_2$
A	0.62	0.62	*[a]	0.89	0.88
B	0.76	0.43	0.79	0.78	0.78
C	0.63	0.60	0.50	0.76	0.77
D	0.75	0.51	0.69	0.47	0.65
E	0.76	0.62	0.74	0.81	0.72
F	0.94	0.89	0.93	0.94	0.94
G	0.84	0.81	0.66	0.86	0.87
H	0.94	0.88	0.91	0.93	0.93
I	0.75	0.71	0.47	0.86	0.90
J	0.72	0.58	0.42	0.87	0.68
K	0.48	0.64	*	0.56	0.67
L	0.80	0.72	0.73	0.84	0.84
M	0.87	0.73	0.32	0.84	0.84
N	0.90	0.72	0.76	0.85	0.87

[a] An asterisk denotes insignificant value (see text).

SELECTION OF COMPOUNDS FOR BIOLOGICAL TESTING

Introduction. In the past, drug discovery programs have often included empirical (random) screening of very large numbers of compounds. This approach has largely fallen from favor, being replaced by more rational procedures, which involve the testing of a much smaller number of compounds, believed to have potential activity on the basis of knowledge of the receptor site or the mode of action, computer modeling, structure activity relationships, etc.[30,31] Nevertheless, there are still occasions when large-scale empirical testing is appropriate, e.g., upon initiation and chemical characterization of a new, high-throughput in vitro biological screen.

Pfizer Central Research (Sandwich, U.K.) has for some years maintained a Structural Representatives File (SRF), a subset of approximately 5% of the total compound collection. The SRF comprises compounds for which sufficient material is available for use in screening, chosen to be representative of the structural variation present in the total collection. The method of selection has to date been intellectual, and somewhat ad hoc, generally amounting to the inclusion of compounds when they contain previously unrepresented ring systems or functionalities. It was felt that a clustering procedure could be used to generate the SRF by automatic means. There are several advantages associated with the use of such an approach. A complex and time-consuming intellectual operation that involves highly trained staff is replaced by a cheap automatic procedure; an effective clustering procedure should help to ensure that no classes of compounds are overlooked when selecting structures for testing and that the selection is consistent and free of bias; the existence of a classification can help to dictate which compounds are tested next in a program, since the identification of one active compound would suggest that the other members of that compound's cluster should also be investigated. Finally, the possibility of altering various parameters within the clustering program allows the creation of different types of SRF to suit different screening requirements. For example, when there is screening capacity available for many compounds to be tested, it should be possible to select molecules from a classification in which there are large numbers of small clusters; conversely, a classification that comprised a smaller number of relatively large clusters would be more appropriate when only limited screening capacity is available.

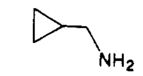
The Pfizer chemical stores file, a collection of ca. 8500 compounds available on-site in usable quantities, was used to test the effectiveness of the JPCM for clustering large files of structures, rather than the small data sets studied in the previous section. This stores file is represented for search purposes as a bit matrix, B, in which the bit $B[I,J]$ is set if the Jth screen has been allocated to the Ith molecule. The screens are small atom-centered or bond-centered fragments that have been selected on the basis of their frequencies of occurrence in the main Pfizer structure file: in all, there are some 1315 such screens. It should be noted that whereas the molecules in the property prediction experiments were characterized by fragments at a single level of substructural description, the augmented atom, the molecules here and in the next section are represented by fragments at many different levels; moreover, only the presence or absence of the fragment in a structure is noted, without regard to its frequency of occurrence.

Clustering of Pfizer chemical stores file. The stores file was clustered with JPCM in two stages. In the first, the 20 nearest neighbors for each of the compounds in the file were identified with the efficient nearest-neighbor search algorithm that has been described previously.[26,32] This near-neighbor table formed the input to the interactive clustering program in which the user can specify that either a weighted or an unweighted classification should be produced and can specify the particular similarity thresholds that are to be used (although default values are adopted if required). The user of the program is given the option of eliminating singleton clusters by assigning them to that nonsingleton cluster to which they are most similar, as described previously. The cluster centroids for this assignment are created by a thresholding procedure so as to obtain bit strings entirely analogous to those that are used to represent each of the individual molecules. Theoretical work by Croft[33] suggests that a feature should be included in a centroid if it occurs in rather less than $T/2$ of the compounds in a cluster of size T; in the work reported here, the bit corresponding to some particular screen is set in the centroid bit string if it occurs in at least $T/4$ of the compounds in a cluster.

In the initial testing of the system, unweighted classifications were generated with thresholds of 2, 3, 4, ... 9, and the resulting sets of clusters were studied in detail to determine the effectiveness of the classifications. Obvious criteria by which a classification may be evaluated are the number of clusters produced and the distribution of the cluster sizes: classifications that exhibit a highly skewed distribution of sizes, i.e., classifications consisting in large part of very small clusters together with a small number of very large clusters, are unlikely to be of very much practical use in the context of an SRF. The distribution of sizes obtained in the test runs is shown in Table IIIA.

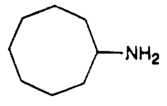
An increase in the similarity threshold corresponds to the identification of clusters that contain more and more strongly similar structures. It is thus to be expected that the use of a high threshold will result in the generation of large numbers of small, tightly bound clusters while a low threshold will result in the generation of a few clusters that contain many compounds and that have a low degree of internal cohesion. That this is indeed so can be seen from an inspection of Table IIIA, where there is a monotonically increasing relationship between the threshold that is used and the number of clusters that is formed. The number of singleton clusters increases in a comparable manner: these singletons are allocated to the most similar nonsingleton cluster as described previously.

An analysis of the clusters formed at the higher threshold levels shows that, in general, they have arisen from the splitting of a cluster generated at a lower threshold into two or more smaller subsets. As an example, the cluster containing the compound

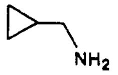


I

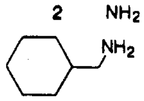
in the unweighted threshold 3 classification is



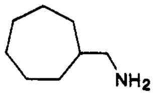
1



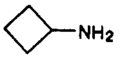
2



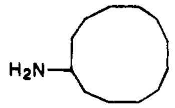
3



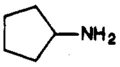
4



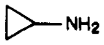
5



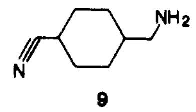
6



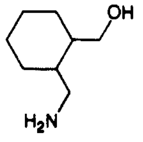
7



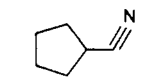
8



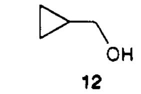
9



10

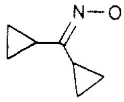


11



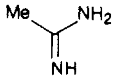
12

It may be noted in passing that no less than 9 of the other 11 compounds in this cluster are included in the set of 20 nearest neighbors for I. The cluster containing I in the unweighted threshold 9 classification contains just compounds 2–4 from the threshold 3 cluster, together with the molecule



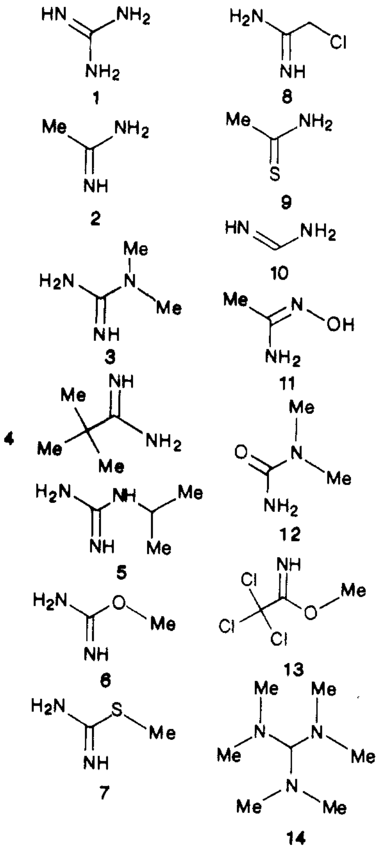
which joins the cluster when singletons are allocated to the most similar nonsingleton cluster.

As a further example, the unweighted threshold 3 cluster containing

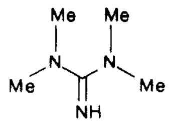


II

is



with all but the last three compounds being amongst the 20 nearest neighbors for the structure II; the corresponding threshold 9 cluster contains structures **1–4** and **6–8** from the threshold 3 classification, together with the molecule



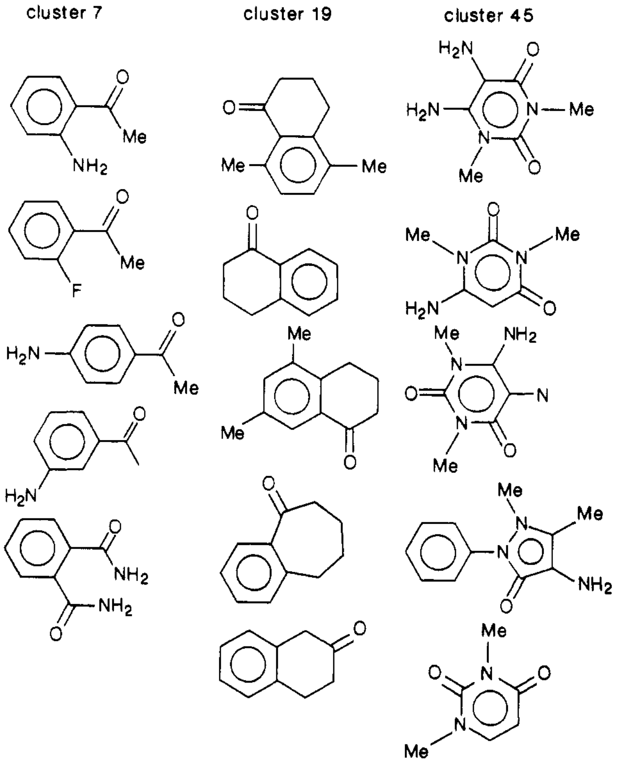
The ordering of the compounds in the clusters above has been obtained by generating a cluster centroid, i.e., a fragment bit string comparable to those that are used to characterize individual compounds. The similarity is then calculated between this centroid and the bit strings characterizing each of the structures within that cluster; the resulting similarities may then be sorted so as to obtain an ordering of the compounds. While the centroid summarizes the structural features present within a cluster, it may be useful in some circumstances to select an actual molecule to characterize the set of compounds within some cluster; a natural choice for such a *representative* is that compound which is most similar to the centroid, i.e., the molecule at the top of the ranking. Such representatives might be required, for example, when selecting one compound from each cluster for activity testing.

Table III. Distribution of Cluster Sizes for Classification of the Pfizer Stores File Using (A) $JPCM_1$ and (B) $JPCM_2$[a]

THRESH	CLUSNUM	1	2–5	6–10	11–20	21–30	>30
		(A) $JPCM_1$					
2	825	185	196	137	143	102	62
3	892	230	204	141	161	100	56
4	1006	299	236	151	174	94	51
5	1120	359	281	161	177	99	43
6	1297	476	335	161	192	92	41
7	1535	606	428	184	198	88	31
8	1843	838	473	238	191	84	19
9	2245	1128	553	289	194	67	14
		(B) $JPCM_2$					
200	855	201	198	149	147	102	58
300	909	225	218	149	165	99	53
400	981	271	231	147	189	97	46
500	1060	319	247	161	197	90	46
600	1150	381	271	164	203	86	45
700	1275	454	308	186	204	90	33
800	1439	564	342	212	203	89	29
900	1605	670	380	235	220	74	26

[a] THRESH is the similarity threshold used, and CLUSNUM is the total number of clusters identified

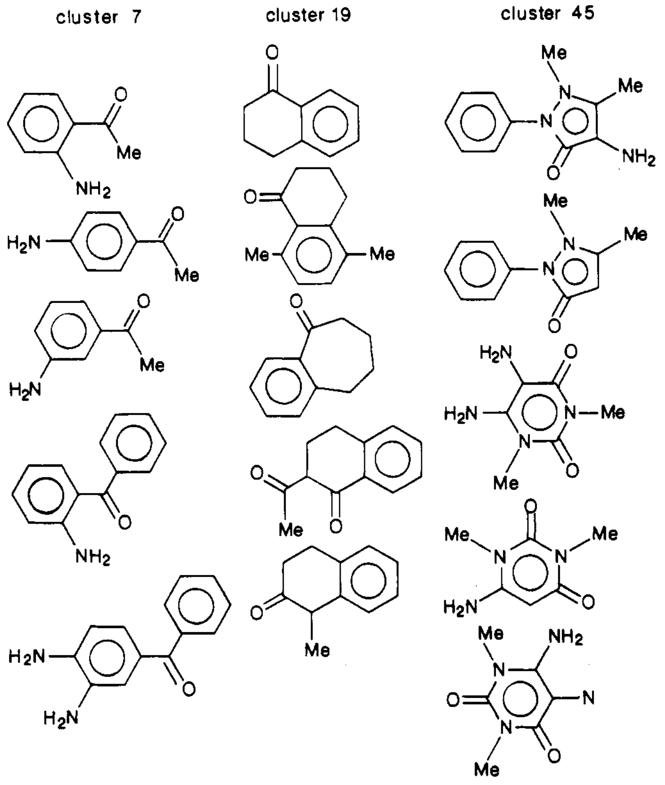
Chart I. Threshold 3 Unweighted



For the screening application considered here, relatively large clusters were felt to be of more general utility than the small clusters of highly similar molecules generated by the use of a high similarity threshold. For example, the threshold 3 classification corresponds to a mean cluster size of 12.8, when singletons are reassigned, as against 7.6 for the threshold 9 classification.

Following initial experiments using the unweighted thresholds, a corresponding series of classifications was generated with weighted thresholds of 200, 300, 400, ..., 900. The distribution of cluster sizes is shown in Table IIIB, and it will be seen that there is a more gradual increase in the number

Chart II. Threshold 300 Weighted



of clusters as the threshold is progressively raised than in the case of the unweighted classification. An inspection of the resulting clusters often reveals relatively little difference in composition when comparable classifications are compared. Thus, the threshold 3 unweighted and threshold 300 weighted classifications both contain ca. 900 clusters: the first five compounds, when ranked in decreasing order of similarity with the centroid, for the clusters corresponding to the compounds numbered 7, 19, and 45 in the stores file are shown in Charts I and II. It will be seen that there is a high degree of similarity between the two sets of clusters, and either classification might reasonably be used as a basis for the selection of compounds for testing purposes. A detailed inspection of the two types of classification suggested that the weighted clusters were slightly more satisfactory, on an intuitive basis, in bringing together closely related compounds, and thus, $JPCM_2$ was selected for the final series of experiments, which are described in the next section.

CLUSTERING OF SUBSTRUCTURE SEARCH OUTPUT

Current chemical retrieval systems allow highly efficient and effective substructure searches to be carried out if the user of the system can specify the structural requirements of the query in reasonably specific and precise terms. It is more difficult for such systems to handle generic queries where many alternative atoms or substituent patterns may need to be specified to ensure that all structures of interest are retrieved. In such cases, the recall-oriented nature of the search may result in the retrieval of many compounds that, although satisfying the constraints imposed by the query, are of only marginal relevance to the problem at hand. We have described previously[17] a similarity matching procedure that requires the user to specify a typical compound of interest: this compound is then matched against the molecules retrieved by a sub-

Chart III

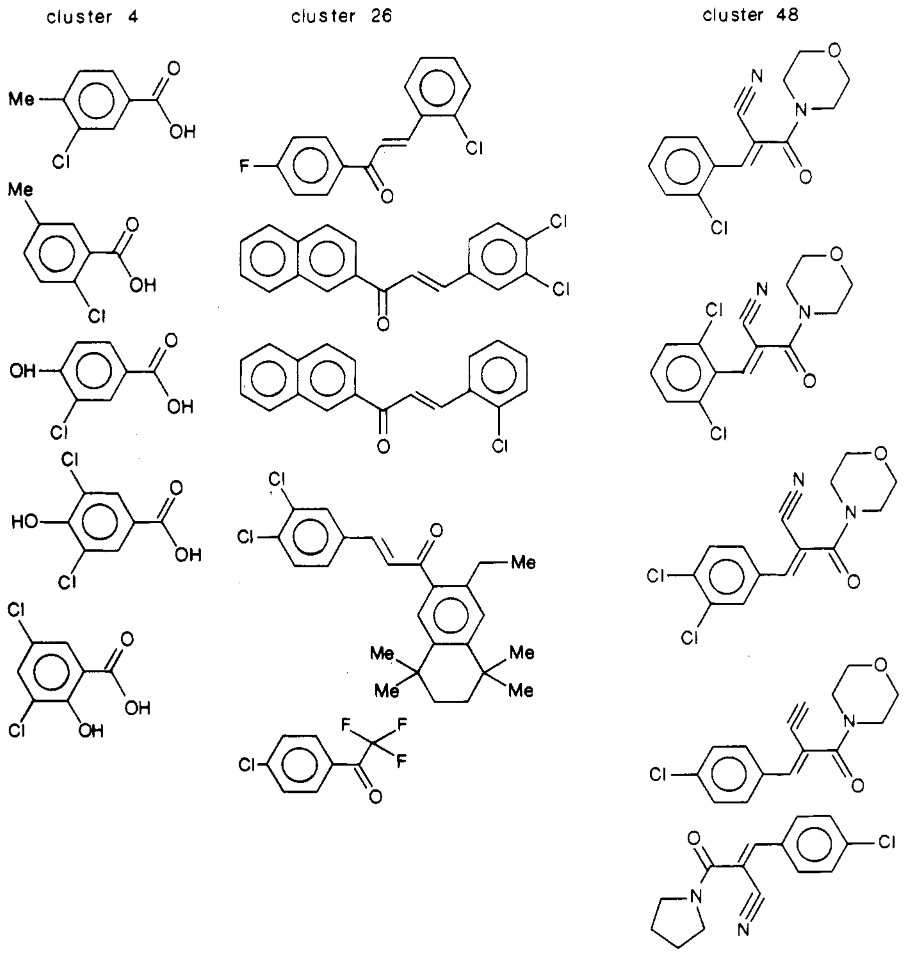
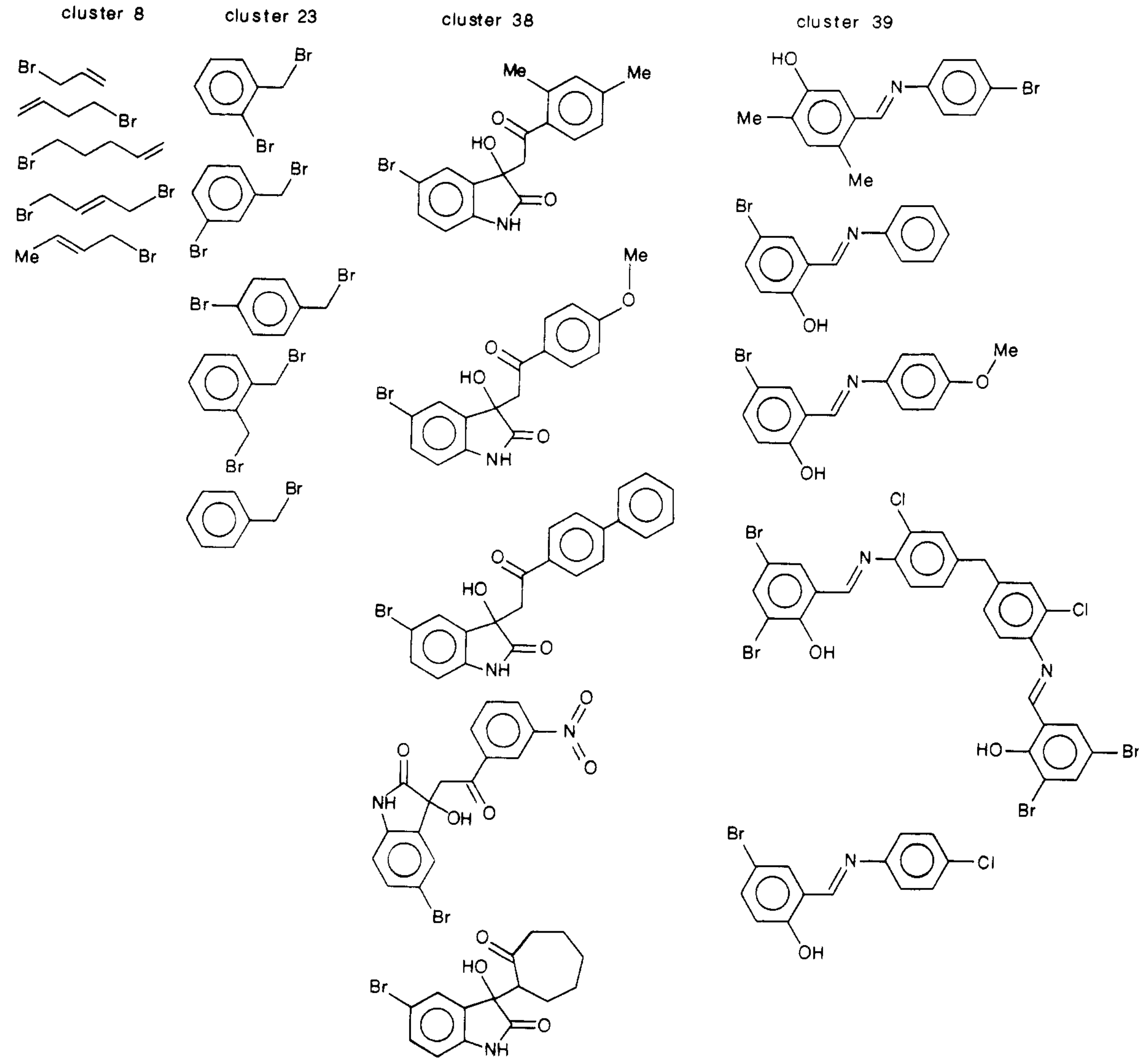


Chart IV



cluster 8 cluster 23 cluster 38 cluster 39

structure search so as to identify those that are most similar to it. In some cases, however, the chemist may not be able to specify such a target compound or may wish instead to gain an overview of the main structural classes that are present in the search output. The work reported in this section was undertaken to determine the extent to which a clustering procedure could be used to meet such search requirements.

The procedure that has been developed operates in three main stages. In the first, the chemist who is using the system inputs a substructural query, of whatever degree of generality, which is then searched against a file of molecules in the normal manner. If it is felt that the output is too large for individual inspection and that clustering is called for, the JPCM is used to group the compounds that have been retrieved. The clustered output file resulting from this second stage may then be inspected by selecting a representative compound from each cluster, typically the compound most similar to the cluster centroid as described in the previous section, and displaying the set of such representatives on the terminal, or as paper copy. All of the compounds in a particular cluster may then need to be inspected by the chemist if the representative molecule proves to be of interest, or the compounds may be rejected if the representative appears to be inappropriate.

This procedure has been implemented initially on the chemical stores file that was used for the clustering experiments reported in the previous section. Queries are matched against this file with SOCRATES, the Pfizer Central Research (U.K.) in-house chemical searching system, which includes graphical query input, bit screen searching, and atom-by-atom searching. The bit strings of the compounds retrieved by this search are then passed to the clustering routine: thus, for each molecule retrieved in the search, the 20 nearest-neighbors within this subset are identified, and these nearest-neighbor lists are processed as described previously to generate the classification. Initial testing suggested that $JPCM_2$ with a threshold of 400 gave the most generally useful results, and this parameter setting has been adopted as the default option, together with the assignment of singleton compounds as described previously. The set of cluster representatives is then generated and displayed as the output from the search.

The effectiveness of this three-stage search procedure was tested with a range of highly generic substructural queries. The first search was for optionally substituted chlorobenzenes: this query substructure retrieved a total of 690 molecules from the file, and they were clustered to give a total of 55 nonsingleton clusters. As with the work described in the previous section, it is rather difficult to give the reader a feeling for the utility of the procedure without displaying very large numbers of structures. In what follows, we present examples of the membership of typical clusters, restricting attention to the five structures in each case that are most similar to the cluster centroid.

The first five members of three randomly selected clusters are shown in Chart III. Cluster 4 contains carboxyl-sub-

Table IV. Distribution of Cluster Sizes Using Threshold 400 JPCM$_2$ Classification of Output from Six Substructure Searches[a]

CPDNUM	CLUSNUM	1	2–5	6–10	11–20	21–30	>30
690	75	20	16	15	14	7	3
591	63	16	13	12	12	8	2
642	72	26	12	9	14	7	4
224	19	3	3	5	5	2	1
730	81	19	20	16	14	10	2
402	42	9	11	8	8	4	2

[a] CPDNUM is the total number of compounds retrieved in each search.

stituted monobenzenes, while the cluster 26 molecules all contain an α-unsaturated carbonyl grouping linking together two benzene rings, one of which contains the specified chloro substituent. The molecules in cluster 48 again all contain the α-unsaturated carbonyl, substituted by a cyanide group, but here one of the rings is a saturated N-heterocycle.

The second search was for all bromine-containing molecules. Such a query is, of course, much less specific than would generally be the case in an operational chemical substructure search system, and similar comments apply to the other queries that were used for the testing of the system. However, such highly generic queries were selected so as to provide a relatively large number of compounds for clustering, despite the small size of the file that was to be searched, and so as to provide a "worst case" test of system effectiveness. The search retrieved a total of 591 molecules, these being grouped into 47 nonsingleton clusters. The first five compounds from clusters 8, 23, 38, and 39 are shown in Chart IV. It will again be seen that there are stark differences in structure between the various clusters. Cluster 8 consists entirely of bromoalkenes while cluster 23 contains mono- and dibromobenzenes; all of the structures in cluster 38 contain the same moiety while the compounds of cluster 39 are characterized by the —CH=NH— linkage between a pair of benzene rings.

In all, six highly generic queries were input during the testing phase, these being the chlorobenzene and bromide queries discussed above, together with the following: all NO_2-containing molecules; all NH_2-containing molecules; all saturated five-membered monoheterocycles, optionally substituted; all bromo-, fluoro-, and iodobenzenes, optionally substituted. The distribution of cluster sizes for the six queries is shown in Table IV, where it will be seen that there is a relatively even spread of cluster sizes: the mean size averaged over the entire set of queries was 11.8 compounds per nonsingleton cluster. An inspection of the classifications resulting from these queries proved them to be just as effective as in the case of the examples above.

In view of the success of this work, clustering of substructure search output is to be made available as a standard retrieval facility of the SOCRATES system, which currently operates on a data base of some 200 000 structures.

CONCLUSIONS

In this paper, we have described the results of a detailed study of the use of nonhierarchic clustering methods for the clustering of compounds in computerized chemical information systems. The initial experiments studied the effectiveness of a range of such clustering methods for structure–activity studies and resulted in the identification of one particular clustering method, that due to Jarvis and Patrick, as the most suitable method for chemical structure applications. This clustering method was then used to generate several classifications of the internal stores file of Pfizer Central Research. The resulting classes were found to be intuitively reasonable, and to compare favorably with manual groupings of internal Pfizer files. It is intended that such an automatic classification will be used on a routine basis for the selection of compounds for activity testing in a range of biological screens. The Jarvis–Patrick method has also been applied to the clustering of search output from a graphics-based chemical substructure search system. The clusters produced have been shown to provide an efficient and effective characterization of the main classes of compound in the output resulting from highly generic substructural queries. This facility for the processing of search output is also to be a standard feature of the Pfizer chemical information system SOCRATES.

The clustering applications described here and the similarity matching applications described previously[17] by no means exhaust the possibilities for research into the uses of similarity and clustering techniques in chemical information systems. For example, one might wish to carry out similarity searches to identify molecules that are *dissimilar* to some query structure, to consider methods for the searching of clusters of compounds, and to investigate alternative forms of structural representation and of intermolecular structural similarity. We feel confident that the use of clustering and similarity techniques will be prove to be highly effective in providing computational support for pharmaceutical research.

ACKNOWLEDGMENT

We thank Prof. M. F. Lynch, Dr. T. K. Devon, and the referees for helpful comments on an earlier draft of this paper and the Science and Engineering Council and Pfizer Central Research (U.K.) for funding.

REFERENCES AND NOTES

(1) Everitt, B. *Cluster Analysis*; Heinemann: London, 1980.
(2) Dubes, R.; Jain, A. K. "Clustering Methodologies in Exploratory Data Analysis". *Adv. Comput.* **1980**, *19*, 113–228.
(3) Gordon, A. D. *Classification*; Chapman and Hall: London, 1981.
(4) Willett, P. *Similarity and Clustering Techniques in Chemical Information Systems*; Research Studies: Letchworth, U.K., in press.
(5) Harrison, P. J. "A Method of Cluster Analysis and Some Applications". *Appl. Stat.* **1968**, *17*, 226–236.
(6) White, R. F.; Lewinson, T. M. "Probabilistic Clustering for Attributes of Mixed Type with Biopharmaceutical Applications". *J. Am. Stat. Assoc.* **1977**, *72*, 271–277.
(7) Adamson, G. W.; Bush, J. A. "A Method for the Automatic Classification of Chemical Structures". *Inf. Storage Retr.* **1973**, *98* 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) 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.
(10) Willett, P. "A Comparison of Some Hierarchal Agglomerative Clustering Algorithms for Structure-Property Correlation". *Anal. Chim. Acta* **1982**, *136*, 29–37.
(11) Rubin, V.; Willett, P. "A Comparison of Some Hierarchal Monothetic Divisive Clustering Algorithms for Structure-Property Correlation". *Anal. Chim. Acta* **1983**, *151*, 161–166.
(12) Willett, P. "Evaluation of Relocation Clustering Algorithms for the Automatic Classification of Chemical Structures". *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 29–33.
(13) Hartigan, J. A. *Clustering Algorithms*; Wiley: New York, 1975.
(14) Spath, H. *Cluster Analysis Algorithms*; Horwood: New York, 1980.
(15) Willett, P.; Winterman, V. "A Comparison of Some Measures for the Determination of Inter-Molecular Structural Similarity". *Quant. Struct.-Activ. Relat. Pharmacol., Chem. Biol.* **1986**, *5*, 18–25.
(16) Willett, P. "Clustering Tendency in Chemical Classifications". *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 78–80.
(17) Willett, P.; Winterman, V.; Bawden, D. "Implementation of Nearest-Neighbor Searching in an Online Chemical Structure Search System". *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(18) Fritsche, M. *Automatic Clustering Techniques in Information Retrieval*; Commission of the European Communities: Brussels, 1974.
(19) van Rijsbergen, C. J. *Information Retrieval*; Butterworth: London, 1979; 2nd ed.
(20) Salton, G. *The SMART Retrieval System*; Prentice-Hall: Englewood Cliffs, NJ, 1971.
(21) Can, F.; Ozkarahan, E. A. "Two Partitioning Type Clustering Algorithms". *J. Am. Soc. Inf. Sci.* **1984**, *35*, 268–276.
(22) Forgy, E. "Cluster Analysis of Multivariate Data: Efficiency versus Interpretability of Classifications". *Biometrics* **1965**, *58*, 768.
(23) Everitt, B. "Some Unresolved Problems in Cluster Analysis". *Biometrics* **1979**, *35*, 169–181.

(24) Hartigan, J. A.; Wong, M. A. "A K-Means Clustering Algorithm". *Appl. Stat.* **1979**, *28*, 100-108.

(25) Jarvis, R. A.; Patrick, E. A. "Clustering Using a Similarity Measure Based on Shared Nearest Neighbours". *IEEE Trans. Comput.* **1973**, *C-22*, 1025-1034.

(26) Willett, P. "Some Heuristics for Nearest-Neighbor Searching in Chemical Structure Files". *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 22-25.

(27) Gowda, K. C.; Krishna, G. "Agglomerative Clustering Using the Concept of Mutual Nearest Neighbourhood". *Patt. Recognit.* **1978**, *10*, 105-112.

(28) Mizoguchi, R.; Shimura, M. "A Nonparametric Algorithm for Detecting Clusters Using Hierarchical Structure". *IEEE Trans. Patt. Anal. Mach. Intell.* **1980**, *PAMI-2*, 292-300.

(29) Murtagh, F. "A Review of Fast Techniques for Nearest Neighbour Searching". In International Association for Statistical Computing. "*COMPSTAT 1984*"; Physica-Verlag: Vienna, 1984.

(30) Olson, E. C.; Christoffersen, R. E. *Computer-Assisted Drug Design*; American Chemical Society: Washington, DC, 1979.

(31) Topliss, J. G. *Quantitative Structure-Activity Relationships of Drugs*; Academic: New York, 1983.

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

(33) Croft, W. B. "Organizing and Searching Large Files of Document Descriptions". Ph.D. Thesis, University of Cambridge, 1978.

Computer Storage and Retrieval of Generic Chemical Structures in Patents. 7. Parallel Simulation of a Relaxation Algorithm for Chemical Substructure Search

VALERIE J. GILLET, STEPHEN M. WELFORD, MICHAEL F. LYNCH,* PETER WILLETT, JOHN M. BARNARD, and GEOFF M. DOWNS

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

GORDON MANSON and JON THOMPSON

Department of Computer Science, University of Sheffield, Sheffield S10 2TN, U.K.

Received February 7, 1986

A relaxation algorithm for chemical substructure search is simulated for implementation on general-purpose multiprocessors. An improved relaxation algorithm is described and the inherent parallelism detailed. The general-purpose simulation package PASSIM is described, and the methods used to simulate the algorithm are given. A variably sized pool of processors was assumed. The simulation was run on 71 structure/query pairs, and an average maximum speedup of 5.5 over a single processor was found, for approximately 20 processors. A great variation is found for individual structure/query pairs. The overall factor limiting the performance is the serial bottlenecks in the algorithm.

INTRODUCTION

This paper describes the results obtained with the simulation package PASSIM[1] to simulate the use of parallel processors in substructure matching of specific chemical structures by a relaxation technique. The study involved specific chemical structures only, as an initial step toward understanding the application of parallelism to the relaxation searching of files of generic chemical structures.[2]

Exact substructure matching for specific chemical structures is recognized as being very computationally demanding, involving establishing an atom-by-atom correspondence between the query and the file structure.[3] Screening systems have therefore been developed where the number of structures requiring an atom-by-atom matching is rapidly reduced by an approximate method.[4] The problems are much greater when a file of generic structures is searched. The variable nature of the structures is likely to result in less effective fragment screening and a more complicated and time-consuming atom-by-atom search. The relaxation algorithm developed at Sheffield by von Scholley[5] is more discriminating than fragment screening but not as computationally expensive, or as discriminating, as an atom-by-atom search. It was developed for screening generic structures after fragment screening.

Relaxation refers to a class of iterative methods for pattern matching. An initial mapping is made between two patterns by establishing a correspondence between the components of one pattern and the components of the second, according to some attributes of the components. The initial mapping is approximate and is refined by iterations that extend the range of local similarity and are repeated until no further refinement is possible, or until some components of one pattern can no longer be mapped onto the other pattern. In the past, relaxation methods have been used in areas such as image segmentation[6] and breaking of substitution ciphers.[7] Relaxation was suggested as a method of chemical substructure searching by Kitchen and Krishnamurthy[8] and Kitchen and Rosenfeld.[9] Here, the problem is one of subgraph isomorphism, where the patterns to be matched are connected graphs and the components used for establishing a mapping between the graphs may be the nodes and edges of each graph (corresponding to the atoms and bonds of a chemical structure). von Scholley's algorithm is based on ideas from Kitchen and Krishnamurthy.

Relaxation searching is well suited to matching patterns characterized by areas of variability, e.g., generic structures, and the inherent parallelism of the algorithm can be exploited to make this a viable search method for the large number of candidates passing fragment screening. The parallelism of this class of algorithms has already been noted[8] although neither tested nor simulated for chemical substructure searching.

Parallelism has already been investigated for some areas of nonnumerical chemistry. The CAS ONLINE search system[4] uses parallelism whereby the database of over 7 million compounds is divided equally among a series of pairs of PDP-11 minicomputers. A query is broadcast to each pair, a section of the database is searched serially by one of the PDP-11's, and candidates are passed to the second for atom-by-atom searching. A constant search speed can be maintained by adding further pairs of computers as the file increases in size.

Ullman[10] has proposed a parallel asynchronous solution for a subgraph isomorphism problem by placing a vital part of