Journal of Computer-Aided Molecular Design, 5 (1991) 455–474
ESCOM

J-CAMD 128

Techniques for the calculation of three-dimensional structural similarity using inter-atomic distances

Catherine A. Pepperrell and Peter Willett*

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

Received 9 November 1990
Accepted 30 January 1991

Key words: Inter-atomic distances; Similarity searching; Structure matching algorithms; Three-dimensional chemical structure databases

SUMMARY

This paper reports a comparison of several methods for measuring the degree of similarity between pairs of 3-D chemical structures that are represented by inter-atomic distance matrices. The methods that have been tested use the distance information in very different ways and have very different computational requirements. Experiments with 10 small datasets, for which both structural and biological activity data are available, suggest that the most cost-effective technique is based on a mapping procedure that tries to match pairs of atoms, one from each of the molecules that are being compared, that have neighbouring atoms at approximately the same distances.

1. STRUCTURAL SIMILARITY

Current computer-based chemical information systems utilise 2-D connection tables as the primary means of representing chemical structure information [1,2]. Such systems have traditionally provided two main types of structure searching facility, these being structure search, which involves the search of a database for the presence or absence of a specified query compound, and substructure search, which involves the retrieval of all molecules in a database containing some specified query substructure, irrespective of the environment in which the query substructure occurs. More recently, there have been several reports of systems for similarity searching, in which those molecules are retrieved that are most similar to an input query molecule [3].

There are very many ways in which the similarity between a pair of molecules might be defined [4]: current similarity searching systems use an approach, first suggested by Adamson and Bush

* To whom correspondence should be addressed.

[5], in which the similarity between a pair of molecules is based on the fragment substructures that they have in common (these fragments typically being those used for the screening stage of a conventional substructure searching system). The rapid development of techniques for fragment-based similarity searching that has taken place over the last few years has led to such facilities becoming available in many chemical information systems as a complement to substructure searching [6–8]. A similarity search involves the user inputting a *target* compound, i.e., a molecule of interest. The similarity is calculated between the target and each compound in the database by comparing the corresponding sets of fragment screens to identify those in common; the structures are then sorted into order of decreasing similarity with the target. Interesting compounds from this ranking can then be used as the basis for subsequent searches. A similarity searching system can also be used to support drug discovery programmes; given a target with known activity or property characteristics, the system allows the identification of structurally related molecules that may be expected to exhibit similar property or activity characteristics [5,6,9–11].

Recent developments in molecular modelling mean that it is now relatively easy to generate sets of 3-D atomic coordinates from 2-D connection tables and several companies have created in-house databases of 3-D structures from their databases of 2-D structures. Several software systems have also been developed for substructure searching in these 3-D databases: examples include the in-house systems developed at Abbott Laboratories [12], Lederle Laboratories [13] and Pfizer Central Research (U.K.) [14], and the commercial systems developed by Molecular Design Limited [15] and by Chemical Design Limited [16]. This work is reviewed by Martin et al. [17]. There is, accordingly, a need to complement these substructure searching systems by the development of methods for determining the degree of similarity between pairs of 3-D molecules. In this paper, we describe several ways in which inter-atomic distance information can be used to calculate 3-D structural similarity, using 10 small sets of structures for which both 3-D coordinate data and biological activity data are available.

2. SIMILARITY MEASURES

2.1. Introduction

In this section, we describe techniques for determining the degree of similarity between a pair of structures for which atomic 3-D coordinates are available. Given such data, it is possible to visualise determining molecular similarity using either distance information or angular information, or both. Our initial studies have focussed on the first of these, since we have considerable previous experience of the use of inter-atomic distances for 3-D substructure searching (see, e.g., Ref. 18); the use of angular information for similarity searching is considered elsewhere [19].

In the work reported here, the degree of similarity between a pair of molecules, A and B, is calculated by a comparison of their distance matrices. Let A contain $N(A)$ non-hydrogen atoms; then the distance matrix for A, DA, is an $N(A) \times N(A)$ matrix such that the IJ-th element, $DA(I,J)$, contains the distance between the I-th and J-th atoms in A (and similarly for the matrix DB representing the $N(B)$-atom molecule B). Given the matrices DA and DB, four types of similarity measure have been defined as described in the remainder of this section (in order of increasing sophistication and of increasing computational complexity). The first two measures provide a global measure of similarity between A and B, i.e., a measure that denotes the overall degree of similarity without any attempt being made to specify which parts of them are structurally related; the second two measures, conversely, provide a local measure of similarity in which the structurally related features are specified. The advantages of local measures in the determination of similarity have been noted recently by Klopman and Raychaudhury [20].

2.2. Distance distribution

The basic idea of the first method is to approximate the overall topography, i.e., the 3-D shape, of a structure by the distribution of the inter-atomic distances; an alternative way of using distance distributions in the calculation of 3-D similarity has been described recently by Mitchell et al. [21]. The algorithm involves calculating all of the $N(A)(N(A)-1)/2$ distinct inter-atomic distances for an $N(A)$-atom molecule, A, and then calculating the frequency distribution of these distances; this procedure is repeated for the molecule B with which A is to be compared. The degree of similarity between A and B is then measured by the extent of the agreement between the two frequency distributions, FA and FB. If the two molecules are structurally similar then we would expect the two distributions to be similar in shape; however, this is unlikely to be the case if A and B are structurally disparate. The degree of fit between the two frequency distributions is measured by means of the sum of the squared differences; thus, if there are M elements in FA and FB, the similarity is given by

$$\sum_{I=1}^{M} (FA(I)-FB(I))^2$$

The method requires that a distance parameter, R, be specified that determines the size of each of the elements in the frequency distributions. Thus, if R is set at 1.0 (all distances in this paper are assumed to be in Å) then the distance range categories will be 0.00→0.99, 1.00→1.99, 2.00→2.99, etc., the categories starting at 0.00 and increasing in equal-sized steps. As each inter-atomic distance is calculated, it is used to increment the appropriate element of the frequency distribution, denoting the occurrence of an inter-atomic distance falling between the minimum and maximum values of the category.

The use of a single frequency distribution to characterise a molecule means that all of the inter-atomic distances are assumed to consist of the same pairs of atoms so that, e.g., a C-O distance is regarded as being equivalent to an N-O distance if they fall in the same distance range category. More useful results may be expected if different frequency distributions are specified for the carbon-carbon, carbon-heteroatom and heteroatom-heteroatom distances. We shall denote these three types of distribution for the molecule A by FA_{CC}, FA_{CX} and FA_{XX}, respectively (and similarly for the molecule B with which A is to be compared). Then, given some goodness-of-fit criterion $G(FA, FB)$, we can define the overall degree of similarity between A and B by

$$\alpha G(FA_{CC}, FB_{CC}) + \beta G(FA_{CX}, FB_{CX}) + \gamma G(FA_{XX}, FB_{XX})$$

where α, β and γ are user-defined weights reflecting the relative degree of importance attached by the user to these three types of distribution. In the absence of real users, we have tested a range of values for these three parameters.

2.3. Individual distances

Rather than using the distance distributions, which take no account of the actual distances, the second method calculates the degree of similarity between A and B on the basis of the number of components of DA and DB that are identical, i.e., it compares each distinct element of DA in turn with each of the distinct elements of DB to see if any of them match; if so, then the matching element of DB is deleted from further consideration and the next element of DA used to initiate a further scan of DB. A distance match occurs if the distances are the same (to within any allowed tolerance such as ± 0.5 Å) and if the elemental types associated with this inter-atomic distance are the same; this definition of a distance match also applies to the third and fourth similarity methods further discussed below. The overall degree of similarity between A and B is then calculated from the number of distinct distances in common, COMMON. The specific similarity coefficient used here was the Tanimoto coefficient [3]: given two objects characterised by A and B attributes, C of which are in common, the Tanimoto coefficient is defined to be

$$\frac{C}{A+B-C}$$

In the present context, the Tanimoto coefficient is thus

$$\frac{\text{COMMON}}{N(A)(N(A)-1)/2+N(B)(N(B)-1)/2-\text{COMMON}}$$

where, as previously, $N(A)$ and $N(B)$ are the numbers of non-hydrogen atoms in A and B.

2.4. *Atom mapping*

The third method attempts to map the individual atoms from A onto those atoms from B that are most similar to them. We assume, without loss of generality, that $N(A) \leqslant N(B)$. Consider the I-th atom in A, $A(I)$: then the I-th row of the distance matrix DA contains the distances from I to all of the other atoms in A. This set of distances is compared with each of the rows, J, from the distance matrix DB to identify the distance matches and hence to calculate the similarity using the Tanimoto coefficient as defined in the previous section; however, the similarity in this case is based just upon the distance matches that involve the atoms $A(I)$ and $B(J)$ (rather than all of the atoms in A and B as previously). In this way, we can calculate an $N(A) \times N(B)$ matrix, S, the IJ-th element of which contains the similarity, $S(I, J)$, between the pair of atoms $A(I)$ and $B(J)$. This matrix is referred to subsequently as the *atom match* matrix.

Once the atom match matrix has been created, the inter-atomic similarities contained within it are used to determine that atom from B which is most similar to each atom from A, i.e., the matrix is used to establish a set of equivalences of the form $A(I) \equiv B(J)$ where $B(J)$ is that atom in B that lies at the centre of the most similar area of 3-D space, as defined by the atoms surrounding it, as does the atom $A(I)$. The basic matching algorithm that is used to establish these equivalences is as follows:

(1) Sort the elements of the atom match matrix into order of decreasing similarity.
(2) Scan the atom match matrix to find the remaining pair of atoms, one from A and one from B, that have the largest calculated value for $S(I, J)$.
(3) Store the resulting equivalences as a tuple of the form $\{A(I) \equiv B(J); S(I, J)\}$.
(4) Remove $A(I)$ and $B(J)$ from further consideration.
(5) Return to Step 2 if there are atoms in A that have not yet been mapped to an atom in B.

The overall degree of similarity between A and B is then calculated as the sum of the similarities over all of the atoms in A using the information in the tuples that were stored in Step 3 of the algorithm above. Thus, the similarity is given by

$$\sum_{I=1}^{N(A)} S(I, J)$$

2.5. *Maximal common substructure*

The final similarity method that we have used is the maximal common substructure (MCS), where the MCS is the largest pattern of atoms in 3-D space that is isomorphic, i.e., structurally identical, to the two molecules that are being compared. Given two molecules, A and B containing $N(A)$ and $N(B)$ atoms and with an MCS containing $MCS(A, B)$ atoms, we have used a definition of structural similarity, analogous to the version of the Tanimoto coefficient suggested by Vladutz in the context of reaction similarity systems [22], in which the similarity between A and B is given by

$$\frac{MCS(A, B)}{N(A)+N(B)-MCS(A, B)}$$

The similarities calculated from this measure are determined solely by the MCS, and take no account of the very many smaller substructures that A and B have in common (the vast numbers of small common substructures resulting from the comparison of a pair of 3-D molecules are discussed by Brint and Willett [23]). An alternative similarity measure that takes account of all 3-D common substructures was thus defined as follows: let there be $N(X)$ substructures common to A and B of size X atoms, $1 \leqslant X \leqslant MCS(A, B)$. Then the similarity between A and B is given by

$$\sum_{X=1}^{MCS(A, B)} X \times N(X)$$

A characteristic of this method is that when some molecule, I, is chosen as a query to search against a database that contains it, I will not necessarily be the most similar structure to the query. This seemingly counter-intuitive result arises from the fact that cliques overlap to a great extent, and it is thus possible for a large molecule to have sufficiently many non-maximal cliques to enable it to be a better match with the query than I itself, even though I will, of course, have the largest MCS in common with the query.

3. EXPERIMENTAL DETAILS

The inherently subjective nature of similarity means that it is often difficult to compare different types of similarity measure. If, however, the measures are to be used in a quantitative structure–activity relationship (QSAR) environment, then there is a simple evaluation technique that has been employed in several previous investigations of molecular similarity [3,5,7], viz. the use of the measures for simulated activity prediction. The basic idea underlying this approach is that the utility of a similarity measure can be evaluated by the extent to which similarities in structure mirror similarities in activity; thus, if an active molecule is searched against a database containing both active and inactive structures, a good similarity measure will be one in which the similarity of the query molecule with the actives will tend to be greater than its similarity with the inactives. It is thus possible to compare the merits of different similarity procedures by the extent to which they are able to rank sets of structures for which both structural and activity data are available.

The similarity methods described in the previous section were applied to ten small datasets from the medicinal chemistry literature that have been used in previous QSAR studies and in earlier work in Sheffield on the comparison of fragment weighting schemes for substructural analysis [24]. These datasets cover a wide range of structural types, including both structurally homogeneous and structurally heterogeneous sets of molecules, and are as follows:

- A: 112 compounds of which 80 were mutagenic and 32 non-mutagenic in the sister chromatid exchange assay [25].
- B: 196 compounds of which 121 were carcinogenic and 75 were non-carcinogenic [26].
- C: 145 nitrosamines of which 112 were carcinogenic and 33 were non-carcinogenic [27].
- D: 141 aromatic amines of which 98 were carcinogenic and 43 were non-carcinogenic [28].
- E: 112 nitrobenzoic compounds of which 53 were musk odorants and 59 were non-musks [29].
- F: 114 organic compounds of which 66 were mutagenic and 48 non-mutagenic in the Ames reversion test [30].
- G: 109 cyclic nitrogeneous compounds of which 63 were mutagenic and 46 were non-mutagenic [31].
- H: 209 9-anilinoacridines of which 150 showed anti-tumour activity and 59 were inactive [32].
- I: 113 steroids of which 69 were potent and 44 non-potent in the McKenzie–Stoughton human vasoconstrictor assay for anti-inflammatory behaviour [33].
- J: 147 barbiturates for which duration of action data were available. Following earlier studies by Stuper and Jurs [34], the 110 compounds with a duration of action of less than 200 min were classified as inactives and the remaining 37 as actives.

Coordinate data for 3-D structures are usually obtained from molecular or quantum mechanical calculations or from the publicly available Cambridge Structural Database of X-ray crystal structure determinations. No such data were available to describe the geometry of the compounds in the various datasets; instead, these were calculated using the CONCORD program, which has been developed by Dr. R. Pearlman and his associates at the University of Texas and which is marketed commercially by Tripos Associates. CONCORD is a sophisticated rule-based program that can generate reasonably accurate 3-D coordinates from a 2-D connection table record for most types of chemical structure [35]. The coordinates produced by CONCORD correspond to a single low-energy conformation and these were used to calculate a distance matrix for each molecule in each dataset. Pairs of these matrices then acted as the input to the similarity procedures that have been described in the second section of the paper. It should be noted that, in a few cases, it was not possible to generate a set of coordinates from the input connection table and thus the number of compounds in a dataset, as listed above, is sometimes slightly less than in the original paper.

Similarity searching methods presume the availability of at least one active molecule [3, 15] that can be used as the target molecule to identify other potentially active molecules. An active molecule in a dataset was selected and its similarity calculated with each of the other molecules in that dataset. The molecules were then ranked in decreasing order of similarity and cut-offs applied to retrieve the top 5, top 10 and top 20 compounds from the ranking. These were then checked to determine whether they were active or inactive in the particular biological test system associated with that dataset. This procedure was repeated for each of the active molecules in a dataset in turn (with the exception of dataset H where every fifth active molecule was used and dataset I where every second active molecule was used). The overall utility of a similarity method was calculated by taking the mean numbers of active molecules at the three cut-off positions when averaged over all of the active compounds that had been used to generate a ranking.

The experimental procedures used here provide a quantitative basis for the comparison of the various similarity measures. We should, however, point out the limitations of the particular approach we have adopted:

- The measures consider only the structural characteristics of molecules, as represented by their inter-atomic distance matrices, and thus take no account of the many other factors that may be determinants of activity, e.g., lipophilicity or charge density.
- The measures also take no account of the precise mode of action at the molecular level that is responsible for the observed activity; indeed, a referee noted that several different molecular mechanisms are involved in some of the carcinogenicity datasets studied here. However, similarity methods are of greatest applicability when little information is available as to the mode of action; given this information, then more sophisticated QSAR methods may be brought to bear on the problem.
- We have considered only a single conformation for each of the molecules, specifically the low-energy conformation produced by the CONCORD program. Other programs are available that allow a more detailed exploration of the conformational space of molecules, e.g., that described by Dolata et al. [36]. Systems are starting to appear that take account of conformational flexibility in the context of 3-D substructure searching [16] and we hope to extend this work to the 3-D similarity context in the future.

4. COMPARISON OF METHODS

Each of the four similarity methods can be implemented in several different ways, the precise similarity that is calculated being determined by the specification of one or more parameter values. These parameters are as follows:

- The Distance Distribution method requires the specification of α, β and γ, i.e., the relative importance of the carbon-carbon, carbon-heteroatom and heteroatom-heteroatom distances in determining the degree of similarity between a pair of frequency distributions. Experiments were carried out with the ratio α:β:γ set to 1:1:1, 1:1:5 and 1:1:10. In addition, this method requires the specification of R, i.e., the range of distances included within each of the elements of the frequency distributions. Experiments were carried out with R set to 0.5, 1.5 and 2.5 Å.
- The other three methods require the specification of the tolerance, ε, that is allowed for a distance match to take place. Experiments were carried out in which a distance $DA(I, J)$ $(1 \leqslant I, J \leqslant N(A))$ was accepted as being a match for a distance $DB(K, L)$ $(1 \leqslant K, L \leqslant N(B))$ if

$$DA(I, J) = DB(K, L) \pm \varepsilon$$

with ε being set to 0.25, 0.50 and 1.00 Å.
- Experiments were carried out in which all of the inter-atomic distances were used and in which all but the 1:2 and 1:3 distances were used, since these are determined by the 2-D structure of a molecule, i.e., by the pattern of its atom connectivities, rather than by its 3-D structure.

It was not possible, given this range of parameters, to test every combination of values on all of the 10 datasets; instead, exhaustive sets of runs were carried out using just datasets F, G and H and the results from these used to limit the runs carried out for the other 7 datasets. These exhaustive runs, which are discussed by Pepperrell [37], suggest that:

- There is very little difference between the three sets of values, 1:1:1, 1:1:5 and 1:1:10, that were used for the α:β:γ ratio in the Distance Distribution experiments, and similar comments apply to the three values, 0.5, 1.5 and 2.5 Å, that were used for *R*.
- There is very little difference between the three values, 0.25, 0.50 and 1.00 Å, that were used for ε in the Individual Distance, Atom Mapping and MCS experiments.
- When the 1:2 and 1:3 distances are excluded from consideration, the results are either very little different from when they are included or noticeably inferior; the latter behaviour is especially so when small molecules are being considered, since the exclusion of the distances in this case means that the molecular structure is only poorly described.

These conclusions are supported by the results shown in Tables 1 and 2, which detail the effects of variations in these parameters on the effectiveness of the anilinoacridine rankings (dataset H). Accordingly, the main experimental results discussed below are based on the use of all of the distances, on settings of 1:1:5 and 1.5 Å for the α:β:γ ratio and for *R*, respectively, in the Distance Distribution experiments, and on a setting of 0.50 Å for ε in the Individual Distance, Atom Mapping and MCS experiments.

The mean numbers of actives retrieved in the top-ranked 5, 10 and 20 structures for each of the datasets are listed in Table 3. An inspection of this table shows that the All Common Substructures method often does very well or does very badly in comparison with the other methods; however, the sheer volume of data means that it is not immediately obvious which is the most generally effective of the measures. This has been investigated using Kendall's Coefficient of

TABLE 1

EFFECT OF VARIATIONS IN THE α:β:γ RATIO AND IN *R* ON THE MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES FOR THE ANILINOACRIDINE DATASET

α:β:γ ratio	Number of top-ranked molecules	*R* (Å)					
		0.5		1.5		2.5	
1:1:1	5	4.43	4.43	4.37	4.40	4.30	4.20
	10	8.23	8.37	8.20	8.23	8.17	8.20
	20	16.00	16.03	16.47	16.47	16.40	16.53
1:1:5	5	4.53	4.47	4.43	4.37	4.43	4.40
	10	8.53	8.67	8.67	8.60	8.30	8.40
	20	16.17	16.27	16.53	16.60	16.33	16.47
1:1:10	5	4.50	4.50	4.47	4.40	4.37	4.37
	10	8.47	8.53	8.56	8.60	8.30	8.30
	20	16.23	16.17	16.50	16.63	16.17	16.20

The first and second figure in each of the main elements of the table correspond to the mean number of actives when the 1:2 and 1:3 distances are included in, and excluded from, the similarity calculation, respectively.

TABLE 2

EFFECT OF VARIATIONS IN ε ON THE MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES FOR THE ANILINOACRIDINE DATASET

ε (Å)	Number of top-ranked molecules	Individual distances	Atom mapping	Common substructures	
				Maximal	All
0.25	5	4.37	4.53	4.60	3.70
	10	8.10	8.80	8.47	7.73
	20	16.50	16.50	15.83	15.97
0.50	5	4.43	4.50	4.60	3.67
	10	8.33	8.63	8.47	7.90
	20	16.53	16.83	15.97	16.17
1.00	5	4.30	4.33	4.60	2.87
	10	8.40	8.63	8.47	6.70
	20	16.40	16.87	15.97	15.47

Concordance, W, which tests the extent to which κ rankings of the same set of M objects are in agreement with each other [38]. In the present context, M corresponds to the 5 different similarity methods and κ to the 10 different datasets, each of which gives an ordering of the similarity methods using the mean numbers of actives. In fact, W can be calculated in three ways, using the top 5, the top 10 or the top 20 structures in the rankings.

The first step in the calculation of W involves the use of the sets of mean numbers of actives to rank the similarity methods; the sum of the ranks, R_{ij}, in each column of the overall $\kappa \times M$ table of results is then determined. The R_{ij} values are summed and divided by M to obtain the mean value of the R_{ij}: each of the R_{ij} is then expressed as a deviation from the mean value. Finally, S, the sum of squares of the deviations is found and W calculated as

$$W=\frac{S}{\frac{1}{12}\kappa^2(M^3-M)}$$

where S is given by:

$$S=\sum\left(R_{\mathrm{ij}}-\frac{\sum R_{\mathrm{ij}}}{M}\right)^2$$

The significance of the calculated value for W can then be established using the χ^2 test, since $\kappa(M-1)W$ is approximately distributed as χ^2 with $M-1$ degrees of freedom. If a statistically significant degree of agreement between the κ rankings is found to be present, then the best overall ranking of the M objects is their mean ranks when averaged across the κ rankings.

We have already noted that the All Common Substructures method gives very inconsistent results. Thus, in some cases, e.g., datasets F and G, it is the best of all of the methods whereas in others, e.g., datasets E and J, it performs very badly. We have thus carried out two sets of analy-

TABLE 3
MEAN NUMBER OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES WITH THE α:β:γ RATIO SET TO 1:1:5, R SET TO 1.5 Å AND ε SET TO 0.50 Å

Dataset	Number of top-ranked molecules	Distance distribution	Individual distances	Atom mapping	Common substructures	
					Maximal	All
A	5	3.96	4.02	4.27	3.81	4.11
	10	7.66	7.70	8.22	7.47	8.09
	20	14.28	15.22	16.15	14.64	15.56
B	5	3.90	4.01	4.12	3.78	3.92
	10	7.21	7.45	7.60	7.03	8.44
	20	13.36	14.17	14.60	13.36	16.87
C	5	4.29	4.46	4.43	4.59	2.07
	10	8.36	8.64	8.40	8.90	5.23
	20	16.56	16.66	16.25	17.35	12.68
D	5	4.03	4.27	4.33	4.47	3.44
	10	7.62	7.94	7.82	8.24	5.96
	20	14.71	15.19	14.87	15.46	11.82
E	5	3.30	3.62	3.77	3.64	0.36
	10	6.23	6.30	7.09	6.68	1.83
	20	11.62	11.64	12.34	11.34	7.49
F	5	3.77	3.95	4.11	3.91	4.41
	10	6.86	7.27	7.39	7.26	8.65
	20	12.06	12.80	13.58	12.65	16.71
G	5	4.25	4.40	4.51	4.11	4.63
	10	7.89	8.38	8.57	7.05	9.40
	20	14.17	15.19	15.73	13.62	17.83
H	5	4.43	4.43	4.50	4.60	3.67
	10	8.67	8.33	8.63	8.47	7.90
	20	16.53	16.53	16.83	15.97	16.17
I	5	3.88	4.26	4.35	4.12	4.35
	10	7.29	7.79	8.44	7.82	8.03
	20	13.76	15.03	16.24	14.68	15.74
J	5	2.81	2.68	2.68	3.11	0.95
	10	5.00	4.86	4.19	5.27	1.95
	20	8.76	8.53	6.81	10.59	3.57

ses, the first of which includes this method and the second of which does not. The calculated R_{ij} values are listed in Table 4 and the corresponding W and χ^2 values in Table 5. The critical values for χ^2 with four degrees of freedom at the 0.05 level of statistical significance is 9.49 and that with

TABLE 4

R_{ij} VALUES WHEN THE SIMILARITY METHODS ARE RANKED IN ORDER OF DECREASING EFFECTIVENESS OVER THE FULL SET OF 10 DATASETS

Number of top-ranked molecules	Distance distribution		Individual distances		Atom mapping		Common substructures			
							Maximal		All	
5	39.5	34.5	29.0	25.0	19.0	16.5	29.0	24.0	33.5	–
10	37.0	32.0	30.0	25.0	21.0	18.0	30.0	25.0	32.0	–
20	38.0	32.0	26.5	23.5	21.0	17.0	33.5	27.5	31.0	–

The two values listed in each element of the table correspond to the inclusion and to the exclusion of the All Common Substructures method.

three degrees of freedom is 7.82; thus the only significant value for W in Table 5 is that corresponding to the top-ranked 5 structures when the All Common Substructures results are omitted. In this case, the average ranks for the four methods are 3.45, 2.50, 1.65 and 2.40 for the Distance Distribution, Individual Distances, Atom Mapping and Maximal Common Substructure methods, respectively, thus demonstrating that the best overall level of performance is obtained from use of the Atom Mapping method. In the other cases, it is not possible to give a statistically significant overall ranking of the similarity methods; however, the consistently low R_{ij} values for the Atom Mapping method suggest that it is the most generally useful of the methods that we have tested.

The Atom Mapping results listed in Table 3 have assumed that the importance of an atom mapping is independent of the type of atom that is being mapped. In some situations, heteroatom mappings may be felt to be of greater importance than carbon mappings; this may be encompassed by the simple expedient of weighting heteroatom mappings, in a manner that is similar to that used previously in the Distance Distribution experiments. Specifically, the similarity is calculated using

$$\sum_{I=1}^{N(A)} \theta \times S(I,J)$$

TABLE 5

CALCULATED VALUES FOR THE KENDALL COEFFICIENT OF CONCORDANCE, W, AND FOR χ^2 WHEN THE SIMILARITY METHODS ARE RANKED IN ORDER OF DECREASING EFFECTIVENESS OVER THE FULL SET OF 10 DATASETS

Number of top-ranked molecules	W		χ^2	
5	0.23	0.33	9.20	9.90
10	0.13	0.20	5.20	6.00
20	0.20	0.24	8.00	7.20

The two values listed in each element of the table correspond to the inclusion and the exclusion of the All Common Substructures method.

where θ is a user-defined parameter reflecting the relative importance attached to a heteroatom mapping as against a carbon mapping. If $\theta = 1$, i.e., if all of the mappings are considered to be of equal importance, then the formula above reduces to that given previously. If $\theta > 1$, then the mapping of a pair of heteroatoms, I and J, will contribute θ times as much to the overall similarity as this mapping would contribute if both I and J were carbons. The results obtained with values for θ of 1, 2, 5 and 10 are listed in Table 6. As would be expected, the effect of this heteroatom weighting scheme is strongly dependent on the particular dataset that is being considered; in some cases, however, it is possible to achieve a non-trivial increase in performance.

5. COMPUTATIONAL COMPLEXITY OF THE METHODS

The results in Tables 3–6 describe the *effectiveness* of the similarity methods; however, it is also necessary to consider the *efficiency* of the methods, i.e., how demanding they are of computational resources. In this section, we discuss their expected time complexities for the matching of a pair of molecules, A and B.

The Distance Distribution method involves the calculation of all of the distances in both A and B, this requiring $O(N(A)^2 + N(B)^2)$ operations, and a similar number to create the frequency distributions FA and FB. The actual comparison of the two distributions requires M operations, where M is the number of elements in FA or FB. M will be less, or very much less, than the number of distances in either of the molecules: the comparison stage can accordingly be neglected and thus the computation is dominated by the calculation of the distances, giving an expected time complexity of $O(N(A)^2 + N(B)^2)$.

The Individual Distances method requires the comparison of the $O(N(A)^2)$ distances from A with the $O(N(B)^2)$ distances from B; this is done most efficiently if the two sets of distances are first sorted into increasing order, this requiring $O(N(A)^2 \log N(A)^2 + N(B)^2 \log N(B)^2)$ operations if an appropriate sorting algorithm is used. Once the lists have been sorted in this way, the comparison to identify the matching distances can be accomplished in $O(N(A)^2 + N(B)^2)$ time; thus the computation is dominated by the sorting operation, giving an expected time complexity of $O(N(A)^2 \log N(A)^2 + N(B)^2 \log N(B)^2)$ (if the lists are not sorted, then the comparison operation dominates, with an expected time complexity of $O(N(A)^2 \times N(B)^2)$).

The Atom Mapping method involves the comparison of the entire set of distances for each atom in A with the entire set of distances for each atom in B. If the lists of distances associated with each individual atom are sorted, then the comparison of one atom with another requires $O(N(A) + N(B))$ time; the comparison of all of the $N(A) \times N(B)$ pairs of lists thus requires $O(N(A) \times N(B)(N(A) + N(B)))$ time, while the necessary sorting of the lists requires $O(N(A)^2 \log N(A) + N(B)^2 \log N(B))$ time, which will be rather less. Once these comparisons have been executed and the inter-atomic Tanimoto coefficients evaluated, the method involves repeatedly scanning the elements of the atom match matrix to find the next pair of atoms that should be regarded as being equivalent. The scanning step needs to be carried out $N(A)$ times, once for each atom in A, and the matrix contains $N(A) \times N(B)$ elements; this stage thus requires $O(N(A)^2 \times N(B))$ operations at most (although the time requirement is much less in practice since the atom match matrix is sorted into order of decreasing similarity before any of the equivalences are identified). Accordingly, the computation is dominated by the comparison stage with an expected time complexity of $O(N(A) \times N(B)(N(A) + N(B)))$ time.

TABLE 6
EFFECT OF VARIATIONS IN θ ON THE MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES USING THE ATOM MAPPING METHOD

Dataset	Number of top-ranked molecules	θ=1	θ=2	θ=5	θ=10
A	5	4.27	4.40	4.45	4.44
	10	8.22	8.39	8.56	8.66
	20	16.15	16.44	16.70	16.80
B	5	4.12	4.08	4.10	4.07
	10	7.60	7.68	7.67	7.69
	20	14.66	14.60	14.56	14.42
C	5	4.43	4.40	4.46	4.50
	10	8.40	8.47	8.55	8.52
	20	16.25	16.23	16.34	16.30
D	5	4.33	4.26	4.14	4.10
	10	7.82	7.82	7.89	7.94
	20	14.87	14.79	14.93	14.82
E	5	3.77	3.77	3.74	3.64
	10	7.09	7.08	7.19	7.17
	20	12.34	12.30	12.23	12.21
F	5	4.11	4.11	4.14	4.08
	10	7.39	7.27	7.35	7.26
	20	13.58	13.20	13.03	13.03
G	5	4.51	4.54	4.49	4.48
	10	8.57	8.52	8.48	8.41
	20	15.73	15.73	15.76	15.57
H	5	4.50	4.53	4.57	4.60
	10	8.63	8.73	8.83	8.83
	20	16.83	16.83	17.03	17.30
I	5	4.35	4.26	4.26	4.26
	10	8.44	8.44	8.38	8.35
	20	16.24	16.41	16.56	16.56
J	5	2.68	2.73	2.84	2.95
	10	4.19	4.19	4.46	4.49
	20	6.81	6.95	7.19	7.30

The reader should note that these complexities are for the comparison of two arbitrary molecules and take no account of the situation that would apply in an operational database system. Here, one need consider only the actual matching operations since any necessary pre-computation for the database structures, i.e., the preparation of the frequency distributions or the pre-sorting of lists of distances, can be carried out just once at database creation time (and the analogous computations for the target molecule can be done prior to the search). In this case, the complexities of the Distance Distribution and Individual Distances methods would be $O(M)$ and $O(N(A)^2+N(B)^2)$, with the Atom Mapping requirement remaining unchanged at $O(N(A) \times N(B)(N(A)+N(B)))$.

The detection of an MCS belongs to the class of NP-complete problems since the identification of all subgraphs containing $MCS(A, B)$ atoms requires

$$O\left(\frac{N(A)! \times N(B)!}{MCS(A, B)!(N(A)-MCS(A, B))!(N(B)-MCS(A, B))!}\right)$$

operations in the worst case (although this number can be reduced very substantially by an appropriate MCS algorithm that can make use of any application-specific heuristics to reduce the number of atom-atom mappings that need to be considered) [39]. MCS detection is thus far more demanding of computational resources than the other approaches that have been described above and it is accordingly important to use as efficient an algorithm as possible for this purpose. Algorithms for MCS detection are discussed by Brint and Willett [23], who have shown the general utility of clique detection procedures for MCS detection in 3-D structures [40].

The complexity arguments suggest that the MCS procedure is far more demanding of computational resources than is the Atom Mapping procedure, which is, in its turn, more demanding than the Distance Distribution and Individual Distances methods. These conclusions are supported by the relative execution times listed in Table 7, which refer to FORTRAN 77 programs run on an IBM 3083 BX processor. The execution times listed for each dataset have been standardised by dividing each time by that for the Distance Distribution method, which is the fastest of the methods, when implemented on that dataset. This was done to remove differences in the sizes of the datasets, in the structural complexities of the molecules comprising a dataset, and in the numbers of active compounds that were used to generate a ranking.

TABLE 7
EXECUTION TIMES FOR THE SIMILARITY METHODS, RELATIVE TO THE TIME FOR THE IMPLEMENTATION OF THE DISTANCE DISTRIBUTION METHOD FOR EACH DATASET

Dataset	Distance distribution	Individual distances	Atom mapping	Maximal common substructure
A	1.00	8.43	17.23	56.06
B	1.00	4.40	15.09	48.12
C	1.00	1.95	12.61	21.93
D	1.00	3.69	15.56	57.10
E	1.00	4.48	16.94	41.29
F	1.00	2.76	13.76	29.97
G	1.00	3.89	14.09	42.80
H	1.00	14.15	33.91	153.20
I	1.00	36.50	95.21	537.89
J	1.00	3.50	14.89	33.13

6. IMPLEMENTATION OF THE ATOM MAPPING METHOD

A careful inspection of the Atom Mapping algorithm, as described previously, will reveal that it does not necessarily result in the best possible mapping of A onto B, i.e., the largest possible value for the overall similarity

$$\sum_{I=1}^{N(A)} S(I,J)$$

This arises from the sequential nature of the atom mapping algorithm, which first identifies the largest single $S(I, J)$ value, removes the atoms $A(I)$ and $B(J)$ from further consideration, and continues to find the next largest $S(I,J)$ value, etc.; the resulting set of $S(I,J)$ values is then substituted into the mapping formula above to determine the similarity between A and B. However, it is possible for a larger value of this similarity to be obtained if a non-maximal remaining $S(I, J)$ value is chosen at some point during the scanning of the atom match matrix such that this choice could allow a larger $S(I, J)$ value (or values) to be identified at a later point (or points) during the scanning, with a subsequent increase in the overall similarity that is calculated.

The identification of the maximal value for the mapping formula can be achieved only if an exhaustive search is carried out which considers all possible mappings of $A(I)$ onto $B(J)$. This inher-

TABLE 8

CALCULATED SIMILARITIES AND OBSERVED EXECUTION TIMES (IN CPU SECONDS ON AN IBM 3083) FOR THE ATOM MAPPING METHOD WHEN IT IS APPLIED TO PAIRS OF 3-D STRUCTURES USING AN EXHAUSTIVE ALGORITHM AND USING THE SIMPLE ALGORITHM DESCRIBED IN THE SECOND SECTION OF THE PAPER

Pair number	Molecular similarity			Execution time		
	Exhaustive	Simple	Ratio	Exhaustive	Simple	Ratio
1	3.98	3.98	1.00	0.68	0.10	6.80
2	4.23	4.19	1.01	0.16	0.14	1.14
3	3.17	3.17	1.00	2.70	0.02	135.00
4	4.58	4.51	1.02	0.06	0.04	1.50
5	4.90	4.84	1.01	0.28	0.03	9.33
6	4.72	4.67	1.01	0.12	0.11	1.09
7	1.26	1.26	1.00	16.74	0.02	837.00
8	5.12	5.12	1.00	0.10	0.08	1.25
9	5.09	5.04	1.01	0.11	0.11	1.00
10	1.33	1.33	1.00	56.91	0.02	2845.50
11	1.73	1.73	1.00	0.21	0.20	1.05
12	3.18	3.18	1.00	0.18	0.17	1.06
13	3.82	3.82	1.00	0.32	0.16	2.00
14	0.89	0.89	1.00	111.27	0.02	5563.50
15	4.00	4.00	1.00	50.39	0.07	719.86
16	2.83	2.83	1.00	0.37	0.35	1.06
17	3.81	3.65	1.04	0.04	0.03	1.33

ently combinatorial problem is extremely demanding of computational resources (as has been noted in related work by Danziger and Dean [41]), even if a depth-first branch-and-bound algorithm is used to minimise the number of correspondences that need to be considered [42]. Pepperrell reports experiments in which pairs of 3-D structures were compared using not only the simple atom mapping algorithm described previously but also an exhaustive algorithm that considered every possible mapping [37]. The results of these experiments are detailed in Table 8. This lists the similarities calculated from, and the execution times for, the simple and exhaustive algorithms, together with the ratios of the calculated similarities and of the observed execution times. The table shows clearly that the simple algorithm provides an extremely effective heuristic since the molecular similarities calculated from its use are only marginally less than those calculated from the use of the exhaustive algorithm; indeed, they are usually the same with the largest difference, that for pair number 17, being less than 5%. However, the running time of the simple algorithm is less, and sometimes orders of magnitude less, than that of the latter algorithm. We hence conclude that the simple algorithm described here is a much more cost-effective way of identifying atom mappings than is the exhaustive algorithm.

7. COMPARISON OF 2-D AND 3-D SIMILARITY METHODS

As noted in the introduction to this paper, current chemical similarity searching systems use fragment co-occurrence data to determine the degree of similarity between a pair of molecules. The fragments are topological in character, consisting of small patterns of atoms and bonds derived from a 2-D chemical structure diagram. A typical fragment descriptor is the *augmented atom*, which consists of a non-hydrogen atom together with the atoms that are bonded to it and the types of these bonds.

For comparison with the results in Table 3, the molecules in each of the datasets were represented by their constituent augmented atoms. The similarity between a pair of structures A and B, having $N(A)$ and $N(B)$ augmented atoms, respectively, and having COMMON of these in common was calculated using the Tanimoto coefficient [8] in the form

$$\frac{\text{COMMON}}{N(A)+N(B)-\text{COMMON}}$$

Active compounds were selected in turn and the top-ranked 5, 10 and 20 structures identified as in the 3-D experiments. The results of these runs are listed in Table 9.

An inspection of the results in Table 9 shows that they are broadly comparable with those obtained from the Atom Mapping method; a comparison of the two sets of results (with $\theta=1$ in the Atom Mapping method) using the Sign test [38] reveals that neither gives a significantly better level of overall performance at the 0.05 level of statistical significance. However, although the mean numbers of actives are often very similar, an inspection of the actual top-ranked structures shows that the two methods result in the retrieval of noticeably different sets of compounds. This is not unexpected since very different types of descriptor are being used to characterise the molecules that are being compared; moreover, only the Individual Distances measure involves a calculation that is directly analogous to that used for the 2-D similarities. Disparate sets of structures are also obtained when the different rankings for the 3-D measures are compared.

TABLE 9
MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES USING AUGMENTED ATOM FRAGMENTS AND THE TANIMOTO COEFFICIENT

Dataset	Number of top-ranked molecules			Dataset	Number of top-ranked molecules		
	5	10	20		5	10	20
A	4.01	7.47	14.67	F	4.23	7.98	14.55
B	4.16	7.57	14.58	G	4.44	8.00	15.84
C	4.58	9.07	17.91	H	4.07	8.57	16.80
D	4.37	8.00	15.49	I	4.38	8.21	15.94
E	3.83	6.66	10.13	J	3.70	6.14	9.08

The finding that the 2-D and 3-D measures result in the identification of different sets of molecules suggests that superior results might be obtained by using some sort of combination match, in which the similarity between the target structure and a structure in the database is based on both the 2-D similarity, S_2, and the 3-D similarity, S_3. Specifically, an overall similarity, S_{32}, was defined as

$$S_{32} = \mu S_3 + \upsilon S_2$$

where μ and υ are user-defined weights reflecting the relative contributions of the 3-D and 2-D structural information, respectively, to the overall degree of similarity between a pair of molecules. Tests were run for all of the datasets with the μ:υ ratio set to 1:1, 1:2, 1:5, 2:1 and 5:1 as well as the two extreme values of 0:1 and 1:0, i.e. the whole range from complete 2-D similarity through to complete 3-D similarity. In what follows, we refer to the two extreme values as *non-combined results* and the remainder as *combined results*.

The results that were obtained for dataset H, the anilinoacridines, are shown in the upper part of Table 10. It will be seen that there are only slight variations from the two extreme values (corresponding to the non-combined similarities), and that none of the combined results are superior to those obtained from the better of the two non-combined results. Pepperrell [37] presents results for all of the datasets and shows that there are only a very few occasions when *any* of the combined results are superior to the better of the two non-combined results for a dataset. The only exception to this general behaviour is with dataset G, where nearly all of the combined results are better than the non-combined results as shown in the lower part of Table 10. In general, then, there are no obvious guidelines as to whether or how the two types of similarity should be combined.

A further comparison used the molecular shape index, $^2\kappa_\alpha$, described by Kier [43]. This is a single-valued parameter, which can be readily calculated from a 2-D connection table and covalent radii data and which provides an explicit and quantitative encoding of the 3-D shape of a molecule. The $^2\kappa_\alpha$ values were calculated for each structure in a dataset; each active molecule, I, was taken as the target compound in turn and the other molecules, J, ranked in order of increasing mod $(^2\kappa_\alpha(I) - {}^2\kappa_\alpha(J))$. The top ranking 5, 10 or 20 structures were then selected in the normal way.

TABLE 10
EFFECT OF VARIATIONS IN THE μ:υ RATIO ON THE MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES FOR THE ANILINOACRIDINE AND CYCLIC NITROGENEOUS DATASETS

Dataset	Number of top-ranked molecules	μ:υ ratio						
		1:1	1:2	1:5	0:1	2:1	5:1	1:0
H	5	4.23	4.20	4.07	4.07	4.37	4.40	4.50
	10	8.47	8.40	8.43	8.57	8.50	8.60	8.63
	20	16.67	16.63	16.53	16.80	16.67	16.53	16.83
G	5	4.54	4.60	4.44	4.44	4.62	4.62	4.51
	10	8.79	8.49	8.22	8.00	8.68	8.68	8.57
	20	16.37	16.16	15.95	15.84	16.51	16.51	15.73

The results obtained from these experiments are shown in Table 11, where it will be seen that the mean numbers of actives are always less than those resulting from the use of the Atom Mapping method, with the sole exception of dataset J, the barbiturates. A comparison of the two sets of results using the Sign test shows that there is a significant difference in performance at the 0.05 level of statistical significance.

The final set of results, Table 12, lists the numbers of actives that would be expected to be retrieved for each dataset if one was to select 5, 10 or 20 molecules at random. A comparison of the figures in this table with those in Table 3 shows that the Atom Mapping method always results in the retrieval of a greater number of active molecules than would be expected on the basis of random selection. This has an associated probability of occurrence under the Sign test of 0.005, thus validating the use of the method for ranking databases of 3-D structures. A more detailed comparison of the rankings resulting from the Atom Mapping method and from random selection is presented elsewhere [19].

TABLE 11
MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES USING KIER'S $^2\kappa_\alpha$ MEASURE

Dataset	Number of top-ranked molecules			Dataset	Number of top-ranked molecules		
	5	10	20		5	10	20
A	4.13	7.89	14.94	F	3.38	6.17	12.18
B	3.59	6.63	12.71	G	4.06	7.73	15.03
C	4.11	7.96	15.62	H	3.97	7.90	16.00
D	3.93	7.35	14.05	I	4.00	7.56	14.24
E	3.32	5.36	10.87	J	2.92	4.54	8.46

TABLE 12
MEAN NUMBERS OF ACTIVES IN THE TOP-RANKED 5, 10 OR 20 MOLECULES OBTAINED BY PICKING STRUCTURES AT RANDOM

Dataset	Number of top-ranked molecules			Dataset	Number of top-ranked molecules		
	5	10	20		5	10	20
A	3.85	7.41	14.52	F	3.30	6.18	11.93
B	3.46	6.54	12.69	G	3.30	6.17	11.91
C	4.08	7.94	15.65	H	3.87	7.45	14.61
D	3.77	7.24	14.16	I	3.43	6.46	12.54
E	2.87	5.22	9.90	J	1.99	3.22	5.68

8. CONCLUSIONS

In this paper, we have discussed several techniques that allow the calculation of the degree of structural similarity between pairs of 3-D molecules using inter-atomic distance information. Experiments with ten small datasets for which both biological activity and structural data are available suggest that the first two methods, the Distance Distribution method and the Individual Distances method, give broadly comparable levels of performance and that both are noticeably inferior to the Atom Mapping method. Overall, this method is the best of those tested, although a superior level of performance is achieved, in some cases, when the Maximal Common Substructure method is used. However, the computational requirements of the latter method can be very large indeed, and similar comments apply to the Atom Mapping method if it is implemented in a completely rigorous manner; we have, however, described a simple heuristic that allows this method to be implemented much more efficiently, whilst still providing accurate measurements of inter-molecular similarity. We thus conclude that the atom mapping procedure provides a cost-effective way of determining the degree of structural similarity between pairs of 3-D chemical structures. We have not been able to demonstrate that the 3-D similarities calculated by this method are superior to the computationally less demanding 2-D similarities that can be calculated using fragment occurrence data; however, the two approaches result in the retrieval of noticeably different sets of structures.

It must be emphasised that there are very many possible ways in which one could use inter-atomic distance information to calculate 3-D molecular similarity. The Atom Mapping method appears to be the most appropriate method of those that we have tested in the particular context of activity prediction. We are now investigating the use of this method for similarity searching in large files of 3-D structures, as is routinely done with large files of 2-D structures [8].

ACKNOWLEDGEMENTS

This work has been funded by a Science and Engineering Research Council CASE Studentship in collaboration with ICI Agrochemicals. We thank Helen Grindley for providing her MCS program and for assistance with its use, Robin Taylor for helpful discussions and for providing the 3-D coordinate data, and the referees for their comments on an earlier draft of this paper.

REFERENCES

1 Barnard, J.M., Perspect. Inform. Manage. 1 (1989) 133.

2 Lipscombe, K.J., Lynch, M.F. and Willett, P., Annu. Rev. Inform. Sci. Technol., 24 (1989) 189.

3 Willett, P., Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, Herts., U.K., 1987.

4 Johnson, M.A., J. Math. Chem., 3 (1989) 117.

5 Adamson, G.W. and Bush, J.A., Inform. Stor. Retr., 9 (1973) 561.

6 Carhart, R.E., Smith, D.H. and Venkataraghavan, R., J. Chem. Inform. Comput. Sci., 25 (1985) 64.

7 Johnson, M.A. and Maggiora, G.M., Concepts and Applications of Molecular Similarity, John Wiley, New York, 1990.

8 Willett, P., Winterman, V. and Bawden, D., J. Chem. Inform. Comput. Sci., 26 (1986) 36.

9 Johnson, M.A., Lajiness, M. and Maggiora, G.M., Prog. Clin. Biol. Res., 291 (1989) 167.

10 Lajiness, M.S., Johnson, M.A. and Maggiora, G.M., Prog. Clin. Biol. Res., 291 (1989) 173.

11 Willett, P. and Winterman, V., Quant. Struct.-Act. Relat., 5 (1986) 18.

12 Van Drie, J.H., Weininger, D. and Martin, Y.C., J. Comput.-Aided Mol. Design, 3 (1989) 225.

13 Sheridan, R.P., Nilakantan, R., Rusinko, A., Bauman, N., Haraki, K.S. and Venkataraghavan, R., J. Chem. Inform. Comput. Sci., 29 (1989) 255.

14 Jakes, S.E., Watts, N.J., Bawden, D. and Fisher, J.D., J. Mol. Graphics, 5 (1987) 41.

15 Christie, B.D., Henry, D.R., Guner, O.F. and Moock, T.E., Online Information 90. Proceedings of the 14th International Online Information Meeting. Learned Information, Oxford, 1990, p. 137.

16 Murrall, N.W. and Davies, E.K., J. Chem. Inform. Comput. Sci., 30 (1990) 312.

17 Martin, Y.C., Bures, M.G. and Willett, P., In Lipkowitz, K.B. and Boyd, D.B. (Eds.) Reviews in Computational Chemistry, VCH, New York, 1990, pp. 213–263.

18 Jakes, S.E. and Willett, P., J. Mol. Graphics, 4 (1986) 12.

19 Pepperrell, C.A., Poirrette, A.R., Willett, P. and Taylor, R., Pesticide Sci., in press.

20 Klopman, G. and Raychaudhury, C., J. Chem. Inform. Comput. Sci., 30 (1990) 12.

21 Mitchell, E.M., Allen, F.H., Mitchell, G.F. and Rowland, R.S., Proc. 2nd International Conference on Chemical Structures, in press.

22 Vladutz, G.E., In Willett, P. (Ed.) Modern Approaches to Chemical Reaction Searching, Gower, Aldershot, 1986, pp. 202–220.

23 Brint, A.T. and Willett, P., J. Chem. Inform. Comput. Sci., 27 (1987) 152.

24 Ormerod, A., Willett, P. and Bawden, D., Quant. Struct.-Act. Relat., 8 (1989) 115.

25 Jurs, P.C., Stouch, T.R., Czerwinski, M. and Narvaez, J.N., J. Chem. Inform. Comput. Sci., 25 (1985) 296.

26 Jurs, P.C., Chou, J.T., and Yuan, M., J. Med. Chem., 22 (1979) 476.

27 Rose, S.L. and Jurs, P.C., J. Med. Chem., 25 (1982) 769.

28 Yuta, K. and Jurs, P.C., J. Med. Chem., 24 (1981) 241.

29 Chastrette, M., Zakarya, D. and Elmouaffek, A., Eur. J. Med. Chem., 21 (1986) 505.

30 De Flora, S., Koch, R., Strobel, K. and Nagel, M., Toxicol. Environ. Chem., 10 (1980) 157.

31 Walsh, D.B. and Claxton, L.D., Mutat. Res., 182 (1987) 55.

32 Henry, D.R., Jurs, P.C. and Denny, W.A., J. Med. Chem., 25 (1982) 899.

33 Stouch, T.R. and Jurs, P.C., J. Med. Chem., 29 (1986) 2125.

34 Stuper, A.J. and Jurs, P.C., J. Pharm. Sci., 67 (1978) 745.

35 Rusinko, A., Sheridan, R.P., Nilakantan, R., Haraki, K.S., Bauman, N. and Venkataraghavan, R., J. Chem. Inform. Comput. Sci., 29 (1989) 252.

36 Dolata, P.D., Leach, A.R. and Prout, K., J. Comput.-Aided Mol. Design, 1 (1987) 73.

37 Pepperrell, C.A., Ph.D. thesis, University of Sheffield, in preparation.

38 Siegel, S. and Castellan, N.J., Nonparametric Statistics, 2nd edn., McGraw-Hill, New York, 1988, p. 399.

39 Levi, G., Calcolo, 9 (1972) 341.

40 Brint, A.T. and Willett, P., J. Comput.-Aided Mol. Design, 2 (1988) 311.

41 Danziger, D.J. and Dean, P.M., J. Theoret. Biol., 116 (1985) 215.

42 McGregor, J.J., Software-Pract. Exp., 12 (1982) 23.

43 Kier, L.B., Med. Res. Rev., 7 (1987) 417.