Journal of Computer-Aided Molecular Design, 6 (1992) 607–628
ESCOM

J-CAMD 178

A fast and efficient method for 2D and 3D molecular shape description

Guy W. Bemis and Irwin D. Kuntz*

Department of Pharmaceutical Chemistry, University of California, San Francisco, CA 94143-0446, U.S.A.

Received 7 April 1992
Accepted 20 July 1992

Key words: Molecular similarity; Database clustering; Molecular comparison

SUMMARY

A new formalism for molecular shape description is described. The formalism, based on considering each molecule as a collection of its 3-atom submolecules, is applied to both the graph theory and geometrical coordinate representations of molecules. The timing results for shape description of several databases indicate that this new method is applicable to large databases. Furthermore, results from clustering a small database show good agreement with clustering results obtained by a distance-matching algorithm.

INTRODUCTION

There has been a renewed interest in computer-assisted searching of databases as part of the efforts in structure-based design [1–3], automated synthetic programs [4] and substructure retrieval [5–9]. Databases of both 2D and 3D information are growing rapidly, and while there have been major increases in computational power, the combinatorial aspect of proper retrieval makes some form of divide and conquer strategy desirable. A widespread need for shape clustering algorithms is apparent in the field of medicinal chemistry where it is often necessary to organize a large database into a number of groups from which a relatively small number of compounds might be chosen for initial testing [10–12]. The need for clustering algorithms based on 3D molecular properties is particularly acute. Our goal for this work is to develop a new method for molecular shape description. We pursue this goal by merging concepts from graph theoretical descriptors with difference distance-matching techniques in order to form useful and efficient new molecular shape descriptors.

Graph theoretical molecular descriptors, based on intra-molecular connectivity, have been use-

* To whom correspondence should be addressed.

ful in molecular property description [13–15]. These descriptors are of 2 main types: path number indices and molecular connectivity indices. Path numbers indices are based on the properties of various unique paths through a molecular graph. Perhaps the best known of the path number indices is the Wiener number [16], defined as the sum of the path lengths between all unordered pairs of atoms in the hydrogen-suppressed structural graph of a molecule. Path number indices, being based on the characteristic polynomial of the molecular graph, suffer from the well-known problem of isospectral graphs – different molecules may have the same characteristic polynomial [17,18]. Molecular connectivity indices are derived by considering bonds as a numerical combination of the valences of the atoms connected to each side of the bond. One of the most popular of these indices is the Randić branching index or path-one molecular connectivity index [19]. This index is the sum over all bonds of the inverse of the square root of the product of the valences of the 2 attached atoms. Simple connectivity indices have drawbacks similar to those of path number indices, and recent work has proposed various methods to address these problems [20]. Both types of indices have been used to correlate molecular structure with activity [13–15]. These correlations have been made for a wide variety of molecular properties varying from boiling point [16] to mutagenicity [21]. The uses and limitations of several of these descriptors have been reviewed recently [22].

Difference distance methods have proven useful for molecular structure matching of entire structures as well as for substructure matching [23–27]. The difference distance method reduces molecular similarity to a problem of graph labelling. The distance matrices for 2 molecules are calculated; one is fixed, and the labelling of the second is permuted so as to minimize the sum of elements in the difference distance matrix. The difference distance matrix, D, has the elements $D_{ij} = d1_{ij} - d2_{ij}$ where $d1_{ij}$ is the ij-th element of the distance matrix of the first molecule and $d2_{ij}$ is the ij-th element of the distance matrix of the second molecule. After labelling has been accomplished, the molecules may be oriented by their respective atom pairings, and an RMS score may be calculated. Molecules with differing numbers of atoms can be matched by incorporation of null points into the distance matrices [26,27]. Due to the local minimum and combinatorial aspects of the structure matching problem, it has been necessary to rely on the method of simulated annealing to find matches for molecules with larger numbers of atoms [23–28]. Other approaches to this problem have been described [29–32]. We reasoned that graph theoretical descriptions, which are essentially a means of representing a 3D molecule as a series of 1D descriptors, could describe molecular shapes more fully if they were expanded to a series of 2D descriptors. In contrast, difference distance comparisons describe 3D geometry for a molecule of n atoms in terms of an n-dimensional hyperspace, where n is typically much larger than the 3 dimensions necessary for shape description, so we thought that the difference distance method could be drastically simplified. We set the following goals for a molecular representation:

(1) The method should be applicable to both connectivity (2D) and coordinate (3D) descriptions of molecules;

(2) It should permit use of the normal clustering algorithms;

(3) It should lead to a simple shape-based hash code.

METHODS

We begin by considering only the hydrogen-suppressed form of a molecule. The 2D and 3D information contained in a molecular structure may be abstracted by considering 2 matrices: the all-pairs-shortest-paths matrix, and the distance matrix. The all-pairs-shortest-paths matrix, A, is composed of all intra-atomic distances in the molecule where distances are measured along bonds – consequently these distances are measured in integral units of bond lengths. The distance matrix, D, is composed of all intra-atomic distances measured in the familiar Euclidean sense. These matrices are symmetrical, and have zeroes along the diagonals, so calculations generally need only involve the subdiagonal elements. We next subject both matrices to the same formalism [33, 34]. We transform each matrix into a series of 3-atom submatrices, one for each unordered triplet of atoms [35]. In the general case, for a molecule with n heavy atoms, there will be a total of $n(n-1)(n-2)/3!$ submatrices. For each submatrix, we sum the squares of the subdiagonal elements and store the results in a histogram. This process reduces the 2D and 3D information for each molecule into 2 histograms; one describing the connectivity of a molecule, and the other describing the shape. Both representations are independent of the particular atom labelling scheme used and are invariant with respect to rotation, translation, and global handedness of the molecular coordinate system. The process of transforming a distance or all-pairs-shortest-paths matrix into its 3-atom submatrices is interesting and leads to several physical interpretations. Most simply, a 3-atom submatrix is a constituent triangle from the original molecule, and the sum of squares of the elements in the distance matrix is the sum of the squares of the side lengths of the triangle (we have chosen to sum the squares of the distances rather than the distances themselves for computational expediency). Additionally, if we make the approximation that the lengths of the 3 sides of the triangles can be represented as 3 orthogonal variables, then it follows that the sum of squares is the square of the magnitude of the vector representing the triangle, i.e.

$$\text{sum of squares} = |V|^2 = a^2 + b^2 + c^2$$

where *a*, *b*, and *c* are the lengths of the sides of the triangle.

For the all-pairs-shortest-paths submatrix, the sum represents the length of the shortest self-returning walk that includes the 3 sub-atoms. Alternatively, these sums can be thought of as modified Wiener numbers for each 3-atom subgraph. The Wiener number [16] for the subgraph is the sum of the subdiagonal elements in its all-pairs-shortest-paths matrix, and the 3D-Wiener number [33,34] for the subgraph is the sum of the subdiagonal elements in its distance matrix, whereas our numbers are the sum of the squares of the corresponding elements.

The ‘triangle’ interpretation then leads to a simple hash coding scheme, in which we calculate a weighted sum of the elements of the histogram. The elements that represent the largest triangles are weighted more heavily than the relatively less important smaller triangles. Note that for our sum-based hash coding scheme some manner of weighting is necessary, because an unweighted sum of the histogram elements is a constant $n(n-1)(n-2)/3!$, for any molecule with n atoms.

Computational details

To construct the histograms we first made the arbitrary choice of using 64 total bins for both the connectivity and the distance-based histograms. Once this choice had been made, we calculated numerical ranges for the individual bins that would allow a reasonable spread of data among the bins. We required that each value be mapped onto only 1 bin; that is, no value could be mapped onto 2 different bins. For each bin we calculated the minimum allowed value, $\text{bin}_i^{\text{min}}$, leaving the maximum value implicit in the minimum for the next larger bin, $\text{bin}_{i+i}^{\text{min}}$. In other words, the i-th bin, bin_i, contains all values such that $\text{bin}_i^{\text{min}} \leq \text{value} < \text{bin}_{i+i}^{\text{min}}$. The top bin is therefore left having no maximum value – this allows for arbitrarily large distances to be stored in our histogram. Our bin range assignment scheme works well for large heterogeneous databases, however, it could also be optimized for more homogeneous molecular databases.

We next addressed the question of assigning the actual numerical ranges for the bins.

(A) Connectivity histograms. For the connectivity histogram, small bin ranges (bins 1–45) were calculated corresponding exactly to the first 45 possible sums of squares of the 3-atom sub-graphs. That is, we found the 45 smallest sum of squares of 3 integers in which the 3 integers obey the triangle inequality:

$$S1 \leq S2 + S3 \text{ and } S2 \leq S1 + S3 \text{ and } S3 \leq S1 + S2$$

where S1, S2 and S3 correspond to the integer subdiagonal elements of a 3-atom distance matrix (the sides of the triangle). Note that this provides an exact histogram representation for these small connectivity submatrices – these histogram bins have no range. In order to retain information provided by the larger 3-atom sub-graphs but at the same time keep the number of total bins reasonable, we then assigned the larger bin ranges (bins 46–64) by an exponentially scaled bin increment. The minimum value to be stored in the bin was calculated as:

$$\text{bin}_n = \text{bin}_{n-1} + (1.38)^{n-46}$$

The minimum values stored in each of the 64 bins in the connectivity histogram are shown in Table 1.

Our choice of cutoff for the exact bins vs. the approximated bins was based on our observation that in a typical molecule the small 'triangle circumferences' are densely populated while the larger ones are more sparsely populated. In terms of information content, much less information is lost by approximating the larger bin ranges than the smaller ones. The exponentially scaled part of our bin weighting scheme emphasizes the gross structural features of molecules at the expense of the finer details. Different bin weighting schemes for similar problems have been described [7].

(B) Distance histograms. Distance histogram minimum bin values were also calculated in 2 segments; a linear segment, and an exponential segment. Smaller bins (bins 1–10) started with a minimum value of 6 for the first bin, with the minimum values for the other bins calculated as:

$$\text{bin}_n = \text{bin}_{n-1} + 1$$

We then assigned the larger bin ranges (bins 11–64) by an exponentially scaled bin increment. For

TABLE 1
CONNECTIVITY HISTOGRAM BIN VALUES

3	6	9	12	14	17	19	22	24	26	27	29	33	34	36	38	41	42	43	45	48	50	51	54	56	57	59	61	62	65	66	68	70	73	74
75	76	77	78	81	83	86	88	89	90	94	95	96	97	99	102	107	113	122	135	153	178	212	259	324	414	539	712	950						

TABLE 2
DISTANCE HISTOGRAM BIN VALUES

6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 27 29 31 33 35 38 41 44 48 52 56 61 66 72 78
85 92 100 109 118 128 139 151 164 178 194 211 229 249 271 295 321 349 379 412 448 487 529 575 625
679 738 802 871

bins 11–64, the minimum value to be stored in the bin was:

$$\text{bin}_n = \text{bin}_{n-1} + (1.085)^{n-11}$$

The minimum values stored in each of the 64 bins in the distance histogram are shown in Table 2.

Hash codes and distances

An index number (or hash code) for each histogram was calculated by summing the bin values multiplied by the square of the bin number – this has the effect of weighting the larger bins (representing larger triangles) at the expense of the smaller ones. We envision these hash codes as a crude method for quickly retrieving molecules of similar shape from very large databases [36–38]. An indication of the quality of these hash codes is given in the Results section. The histograms we have calculated may also be considered as vectors and the 'distance' between vectors calculated. Then the compounds may be clustered by any one of the well-known clustering algorithms. In our case, 'distances' between histograms were calculated by the Euclidean metric, Eqn. 1, even though the vector components are not genuinely orthogonal.

$$\text{Euclidean distance} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_i - b_i)^2} \tag{1}$$

where a_i represents the i-th histogram bin of molecule a, and b_i is the i-th histogram bin of molecule b. A number of other methods for 'distance' measurement could be used [39]. Clustering was performed using the complete linkage method, although a number of other methods for hierarchical or non-hierarchical clustering could be used [40–42]. An example showing the generation of connectivity-based information for the molecule *n*-butane is shown in Appendix A.

TABLE 3
PROGRAM PERFORMANCE ON DIFFERENT DATABASES

Database	No. of molecules	CPU time (s)	Molecules/s
CMC	4817	233.8	20.6
MDDR	12048	913.2	13.2
FCD	57128	1788.2	31.9

Calculations performed on Silicon Graphics 4D/25 GT.

RESULTS

Performance

To generate program performance information, we applied our algorithm to the complete sets of 3 commonly used databases: the CMC database [43,44], the MDDR database [45] and the FCD database [46]. Access to these databases was supplied under an academic license from Molecular Design Limited, San Leandro, CA. Timing for connectivity and distance histograms and hash code calculation and disk storage for these databases is shown in Table 3 (calculations performed on Silicon Graphics 4D/25 GT) [47].

There are in the order of 10 000 000 compounds in the Chemical Abstracts database [48]. If we make the approximation that their size distribution is approximately equivalent to that of the CMC database, the processing time for the set of 10 000 000 molecules would be 135 h (5.6 days) with no further attention to program efficiency, computer speed, or parallel processing.

To explore how program performance depends on molecule size, we processed a series of databases each containing 100 molecules with the same numbers of heavy atoms. As Table 4 shows, over-all calculation scales roughly quadratically with the number of heavy atoms (calculations performed on Sun SPARCstation 1) [49].

Verification

We applied this method to selected subsets of the CMC database, which contains connectivity information along with CONCORD-generated coordinates [50,51] for structures from the 1990 edition of Comprehensive Medicinal Chemistry [43]. Since the total number of unordered triplets in a molecule with n heavy atoms is $n(n-1)(n-2)/3!$, exact comparisons can be made among molecules having the same number of heavy atoms. We will return to the question of comparing molecules with differing numbers of heavy atoms in the Discussion section.

We have explored several different ways to map the correspondence between 2 molecules with the same number of non-hydrogen atoms:

(1) Best-fit (RMS) orientation of the Cartesian coordinates;

(2) Triple atom histogram based on connectivity and hash code;

(3) Triple atom histogram based on coordinates and hash code.

We select the direct overlay of molecular coordinates as the measure that corresponds most closely to a chemist's intuition of shape similarity and ask how the other procedures compare.

TABLE 4
PROGRAM PERFORMANCE ON DIFFERENT MOLECULAR SIZES

Number of heavy atoms	Time (s)	Molecules/s
12	4.6	21.7
25	11.4	8.8
50	40.7	2.5
75	111.7	0.9
95	208.5	0.5
110	643.8	0.2

Calculations performed on Sun SPARCstation 1.

We made minor modifications to the existing DOCK [52,53] code to allow molecules with the same number of atoms to be overlaid, and their matching scored by an RMS-like function. For 2 molecules with the same number of atoms, the problem of calculating the best fit can be separated into 2 parts: (1) establishing an optimum equivalence or labelling of the atom sets, and (2) overlaying the 2 sets of atoms. DOCK provides a method for the overlay of unlabelled atoms. Since the final orientation of molecules in DOCK is based on an iterative least-squares minimization [54] of a subset of the atoms from each molecule, it was necessary to extend the procedure by assigning intramolecular atom pairings for all atoms in the molecules, then to reorient the molecules based on these pairings. The RMS-like score was then calculated by summing the squares of the distances between paired atoms. The following algorithm was used to assign atom pair matches for each DOCK orientation:

```
(1) For each atom in molecule A:
    – measure distance to each atom in molecule B;
    – store minimum distance.
(2) Select atom in molecule A with the largest minimum distance:
    – assign specified paired atom from molecule B;
    – remove assigned atoms from further calculations.
(3) Repeat from step 1 until all atoms are assigned.
```

Appendix B gives the FORTRAN coding of this algorithm.

The sum of the squares of these atom-by-atom distances provides a floating point number representing a DOCK-based measure of similarity between 2 molecules with the same number of atoms. By preparing a similarity matrix for a small database of molecules – calculating the similarity of every molecule to every other molecule – we have enough information to perform a hierarchical clustering [40–42] for the molecules which we can then use for 'calibrating' our new histogram method for molecular similarity description. We examined the behavior of our algorithm and the DOCK algorithm in detail for molecules from the CMC database with 12 heavy atoms, making similarity comparisons to phenylalanine where relative comparisons were necessary. We also studied the behavior of our histogram algorithms on molecules from the same database with 36 heavy atoms.

CMC 12 heavy atom dataset

The 108-non-salt* CMC compounds with 12 heavy atoms (CMC12) were clustered [55]. Due to the large amount of data, we present these results in several different fashions. The complete linkage clustering of compounds around phenylalanine is shown in Figs. 1–3 for connectivity histograms, distance histograms, and DOCK scoring, respectively. These figures are sub-dendrograms of the original clustering containing approximately the same number of compounds. The nature of hierarchical clustering makes it difficult to obtain clusters with identical sizes for comparison. Perhaps the most valuable information can be gained from these sub-dendrograms by looking at Fig. 4, which shows the set of compounds common to all 3 similarity methods, the consensus cluster. Figure 5 shows the additional compound, pholedrine, which is common to the distance and connectivity histogram methods, but not to the DOCK method.

* Only non-salt molecules were considered because our connectivity-histogram method treats salts as independent fragments while our distance-histogram method treats them as a single molecule.

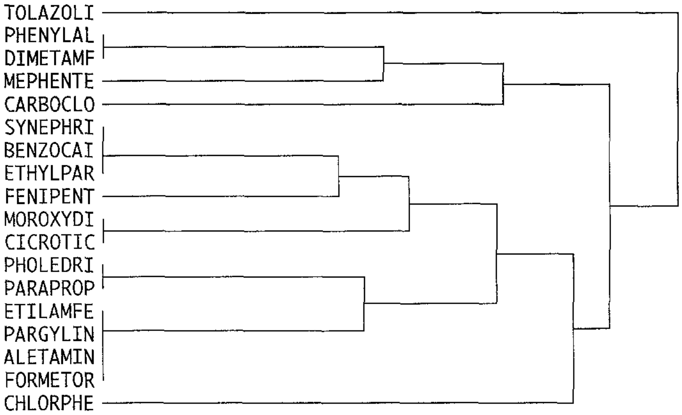


Fig. 1. Connectivity clustering of CMC12, phenylalanine root.

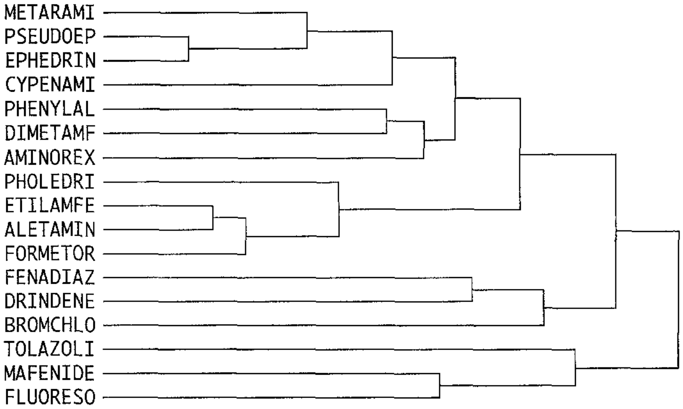


Fig. 2. Distance clustering of CMC12, phenylalanine root.

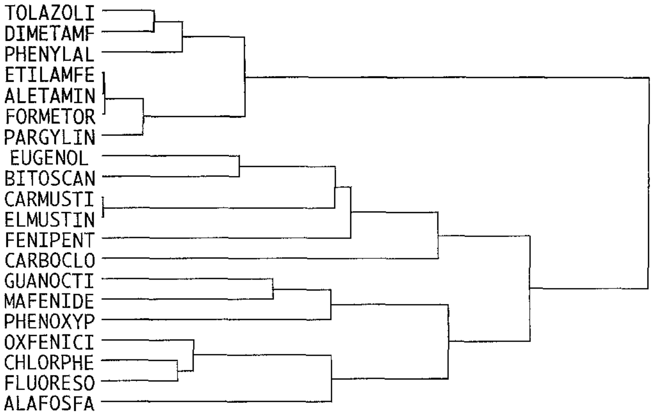
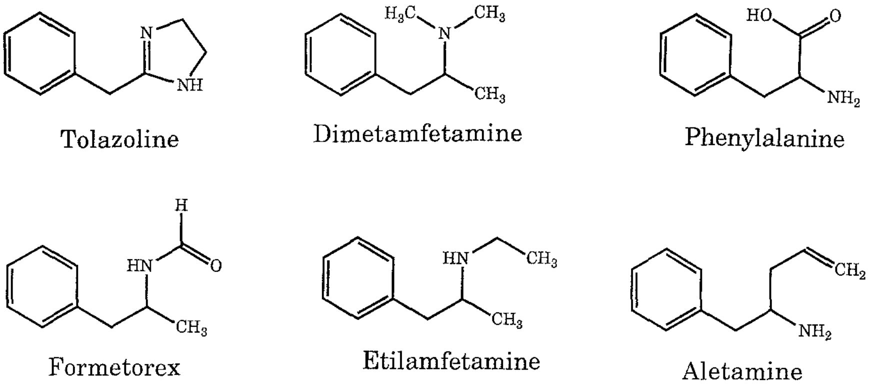


Fig. 3. DOCK clustering of CMC12, phenylalanine root.



Tolazoline

Dimetamfetamine

Phenylalanine

Formetorex

Etilamfetamine

Aletamine

Fig. 4. Consensus cluster of CMC12.

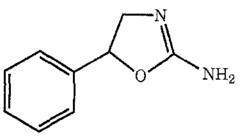
The results of Fig. 4 show that there is a good degree of overlap between all 3 similarity measurement methods. As a means of calibrating the relative histogram similarity scores, Figs. 6 and 7 show graphs comparing the connectivity and distance histogram similarity to phenylalanine with the DOCK similarity to phenylalanine.

As expected, distance histogram similarity shows a much better correlation ($r^2 = 0.50$) with DOCK similarity than does connectivity histogram similarity ($r^2 = 0.18$). Both the distance histogram and the DOCK methods deal with geometric 3D space, whereas the connectivity histogram method deals with 'connectivity space'. An indication of the correlation of the hash code similarity to the histogram similarity for both connectivity and distance histograms is shown in Figs. 8 and 9.

Hash code similarities were calculated by taking the absolute value of the difference between hash codes. It can be seen that the distance histogram similarity is well correlated with the hash coding ($r^2 = 0.63$), but that the connectivity hash codes seem to have 2 independent correlations. The comparison of our hash codes with the direct comparison of connectivity similarity with distance similarity is shown in Fig. 10.

The modest degree of correlation ($r^2 = 0.56$) indicates that these histograms provide both similar and complementary information. We should emphasize that each of the Figs. 6–10 shows the correlation of the radial distributions of the various scoring procedures about some reference molecule, in this case phenylalanine; hence, high levels of correlation are not expected.

To get some indication of how our newly described hash codes compare with similarly-based hash codes described in the literature, we compared them with both the Wiener number and the



Pholedrine

Fig. 5. Pholedrine.

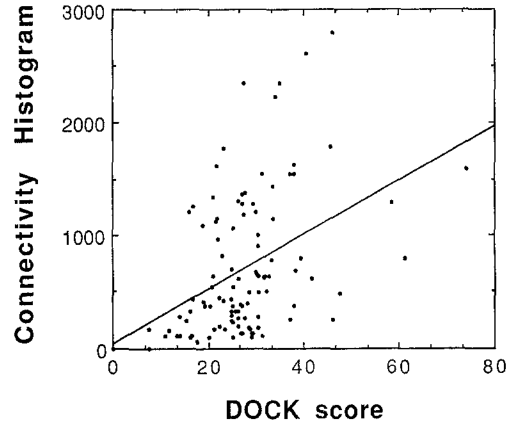


Fig. 6. Connectivity vs. DOCK, similarity to phenylalanine. Correlation equation: y = 41.082 + 23.950x (R^2 = 0.180).

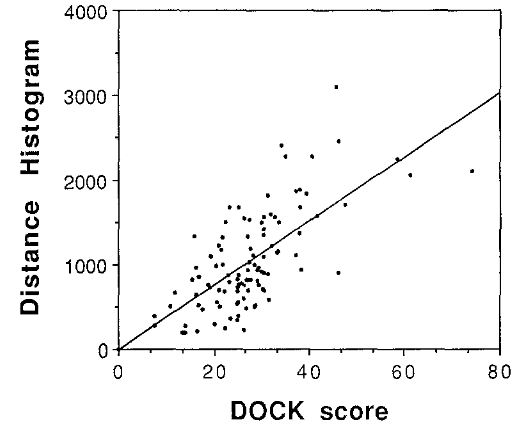


Fig. 7. Distance vs. DOCK, similarity to phenylalanine. Correlation equation: y = − 17.825 + 38.129x (R^2 = 0.499).

3D Wiener number. As Fig. 11 shows, our connectivity-based hash code correlates well (R^2 = 0.96) with the Wiener number for the CMC12 dataset.

Somewhat surprisingly, using the same CMC12 database, our distance-based hash code correlates extremely well (R^2 = 1.00) with the 3D Wiener number, as is shown in Fig. 12.

Similarity links. Having clustered the CMC12 dataset, we can now obtain additional information by making similarity comparisons to subsets of the CMC database containing different numbers of heavy atoms. We calculate the distance from phenylalanine to each compound in the CMC11 and CMC13 databases (we refer to subsets of the CMC database as CMCn, where n is the number of heavy atoms in each molecule in the subset) using the Euclidean metric, Eqn. 1. This provides us with similarity links from subset to subset. Next, we start from the most similar compound found in CMC11, and find the most similar in CMC10. We also start from the most

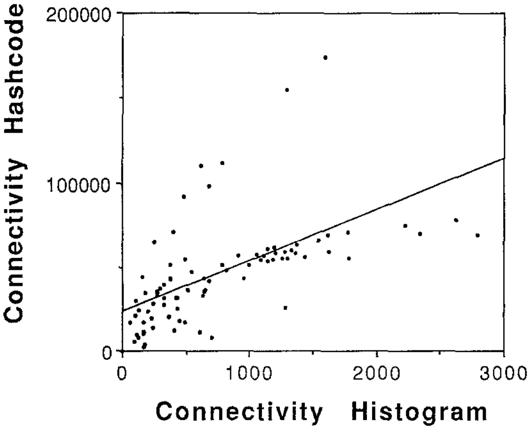


Fig. 8. Connectivity histogram vs. hash code, similarity to phenylalanine. Correlation equation: y = 23093 + 30.464x (R^2 = 0.356).

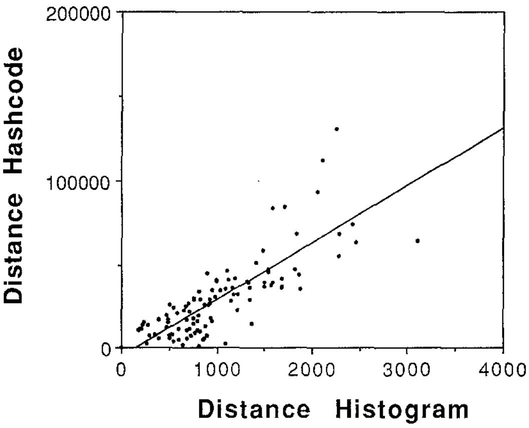


Fig. 9. Distance histogram vs. hash code, similarity to phenylalanine. Correlation equation: y = − 5720.2 + 34.056x (R^2 = 0.625).

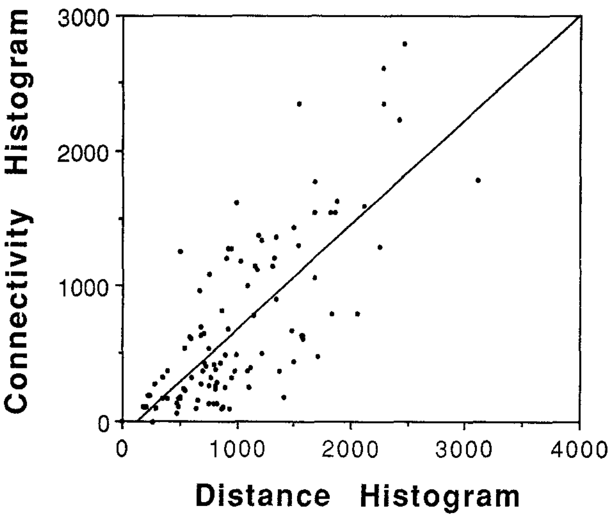


Fig. 10. Distance histogram vs. connectivity histogram, similarity to phenylalanine. Correlation equation: y = − 105.80 + 0.77966x (R^2 = 0.556).

similar compound from CMC13, and find the most similar in CMC14. A symbolic representation of these comparisons for both distance- and connectivity-based histograms is shown graphically in Fig. 13.

The advantages of clustering a database based on this method will be discussed below.

CMC 36 heavy atom dataset

The connectivity-histogram- and distance-histogram-clustered dendrograms for all 36 of the non-salt molecules in the CMC database with 36 heavy atoms (CMC36) are shown in Figs. 14 and

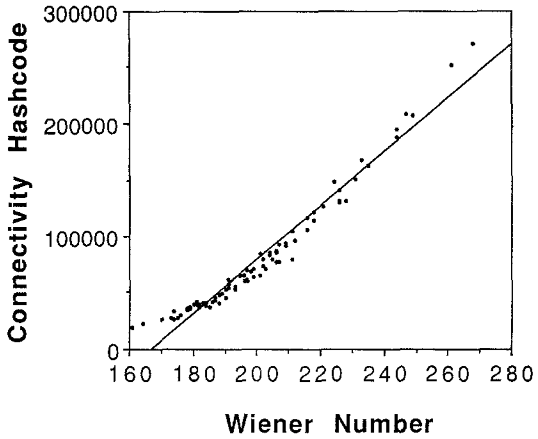


Fig. 11. Wiener number vs. connectivity hash code. Correlation equation: y = − 400080 + 2391.4x (R^2 = 0.958).

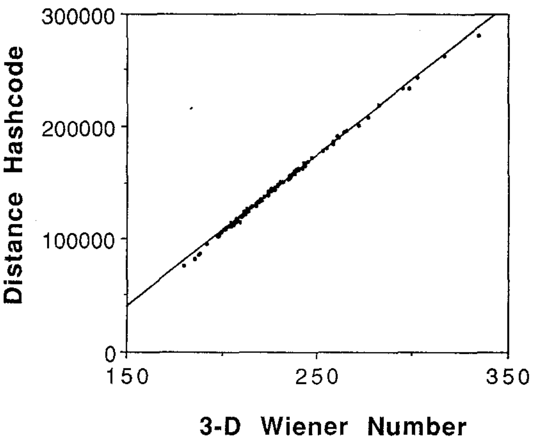


Fig. 12. 3D Wiener number vs. distance hash code. Correlation equation: y = − 161580 + 1342.2x (R^2 = 0.998).

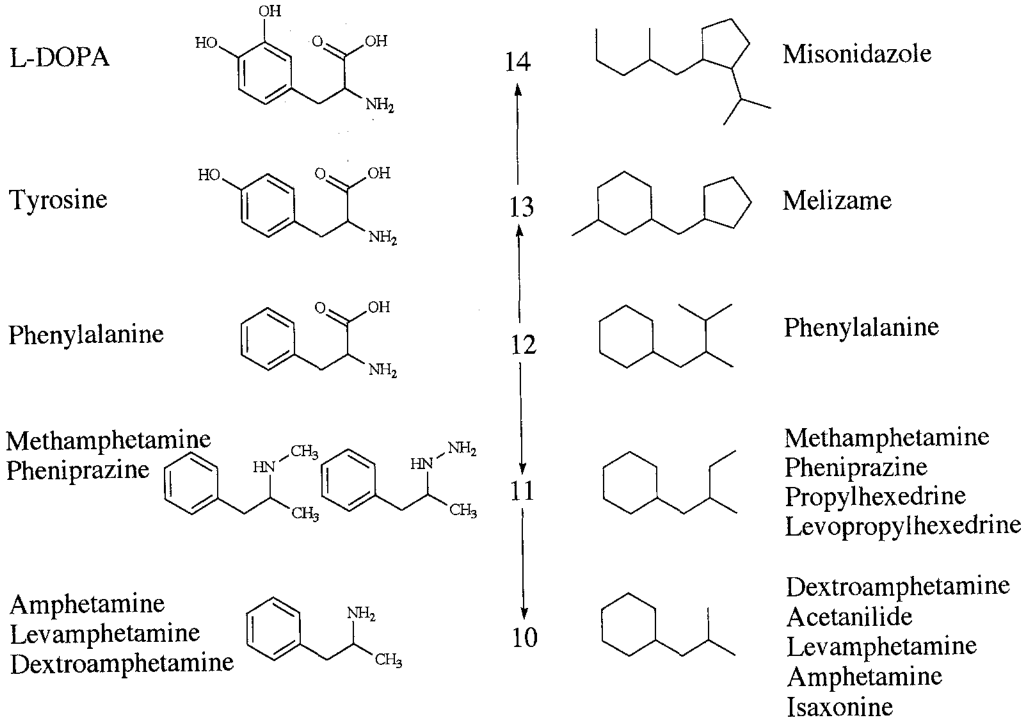


Fig. 13. Extended similarity links for distance histogram of phenylalanine.

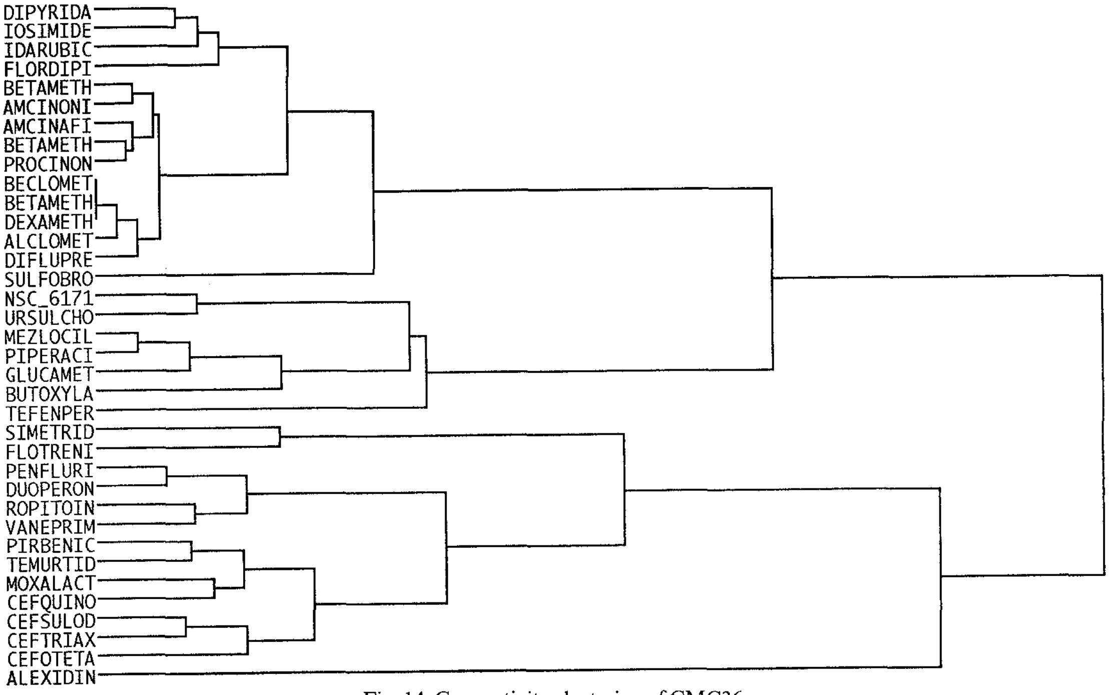


Fig. 14. Connectivity clustering of CMC36.

15. It is useful to look at the individual histogram bin values for some of the similar compounds that have been clustered together. Accordingly, Fig. 16 shows graphs representing the histogram bin values for the connectivity clustering of betamethasone acibutate, penfluridol, amcinonide, mezlocillin, piperacillin, and duoperone, compounds that are clustered into 3 different groups relatively early in the dendrogram. These examples are specific for the compounds listed; however, they also show the general features of similar vs. dissimilar compounds and connectivity information vs. distance information.

In this figure, bins containing smaller distances are on the left and larger distances are on the right. It can be seen that most of the differences between these compounds lie in the histogram bins of large size – approximately the top fifth. In contrast, Fig. 17 shows the distance clustering of betamethasone acibutate, difluprednate, ropitoin, duoperone, mezlocillin, and piperacillin. The differences for the distance histograms span a much wider range or bin sizes.

Conformational variations. For conformational variants of one molecule, differences will be present in the distance histograms but not in the connectivity histograms. This may be extremely useful where the 3D coordinates for a molecule may not be extremely reliable, or where conformational flexibility is expected to play a role in the biological activity of a compound. Table 5 shows connectivity hash code and distance hash code vs. the internal torsional angle for *n*-butane. These molecular geometries were calculated using SYBYL [56].

Our histogram bins provide for the differentiation of many of the various conformations and bonding patterns that 6-membered rings can adopt. Figure 18 shows the distance-histogram-based complete link clustering for 6-membered rings. Geometries for these rings were taken from the SYBYL fragment database.

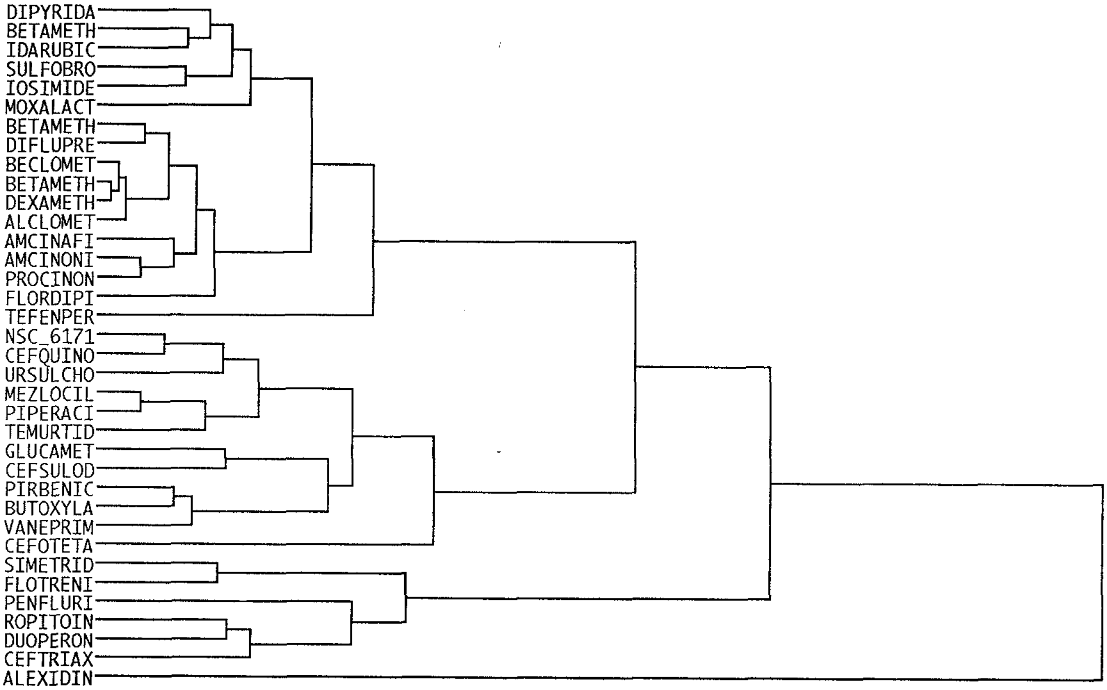


Fig. 15. Distance clustering of CMC36.

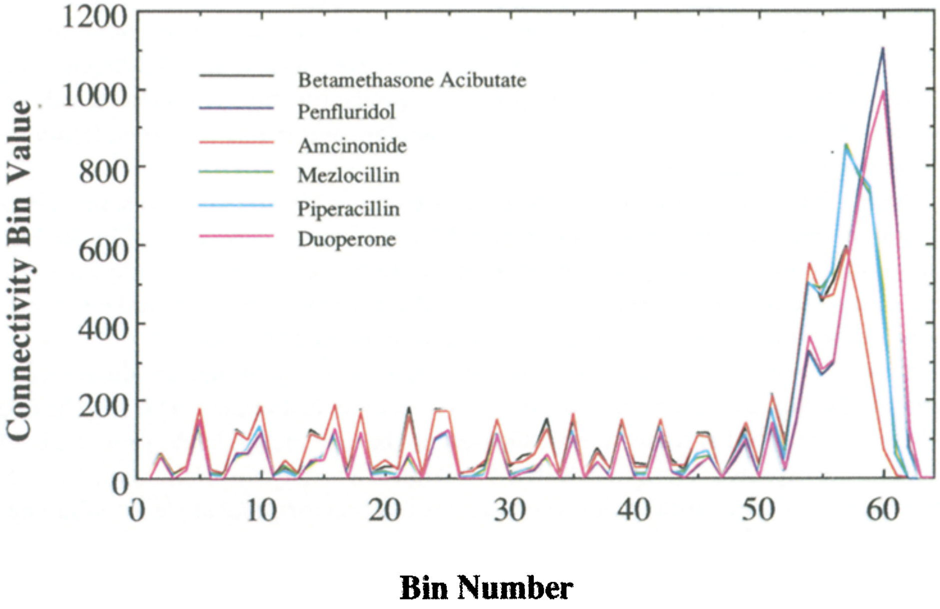


Fig. 16. Selected connectivity histogram profiles, CMC36.

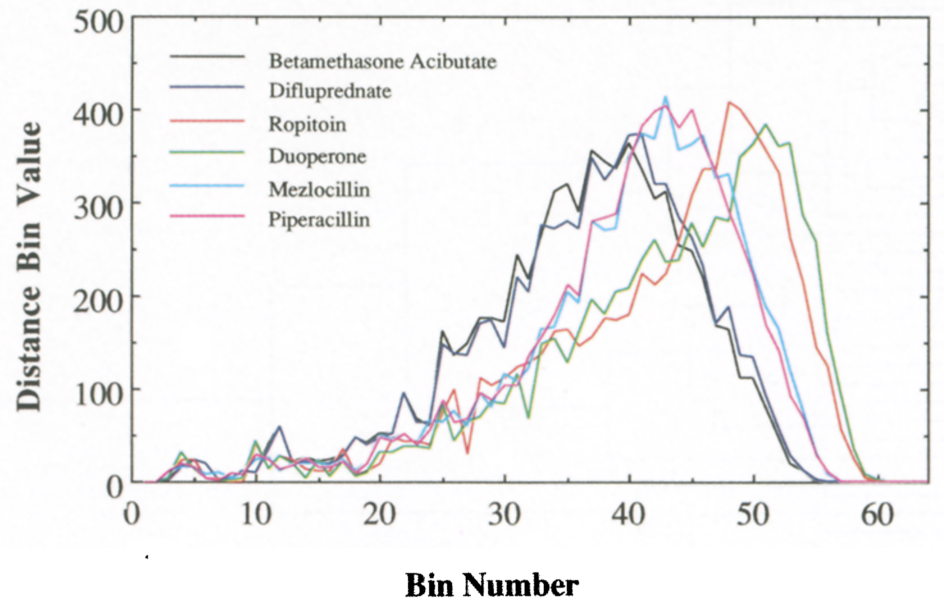


Fig. 17. Selected distance histogram profiles, CMC36.

TABLE 5
VARIATION OF DISTANCE HASH CODE FOR CONFORMERS OF *n*-BUTANE

Connectivity hash code	Distance hash code	Torsional angle (°)
58	272	0
58	314	45
58	464	90
58	650	135
58	720	180

DISCUSSION

While there is probably no best way to describe molecular shape nor to measure the similarity of 2 molecules, each of the methods discussed here is useful for specific purposes. We have explored molecular similarity clustering based on both connectivity and distance. The strengths of each approach lie in the ability to quickly generate a description of the complex shape of molecules by decomposing them into 3-atom 'submolecules'. The triplet atom subgraph concept is both geometrically appealing and, at the same time, readily calculable. In this unified approach, distance and connectivity information are represented by the same formalism. By encoding the entire set of triplets in a molecule, our approach avoids one of the difficulties associated with most fragment-based molecular description approaches – the determination of which fragments to encode. Additionally, our methods lead to a highly useful hash code which may prove valuable for dealing with arbitrarily large collections of molecules. The degree of correlation between our hash codes and the corresponding Wiener numbers is encouraging, and indicates that we have retained much of the original structural information. However, our approach has notable weaknesses. The amount of information stored for an individual molecule in some cases is larger than would be required to store the heavy atom coordinates themselves. Each molecule is used to generate 2 hash codes and 2 histograms with 64 values each, so for a molecule with fewer than 43 heavy atoms it is more compact to store just the 3 coordinates for each heavy atom. As for processing of the individual histograms, our method is less efficient than the bit-by-bit operations on key-coded strings popular in chemical information processing [57,58], although bit comparisons are easily imple-

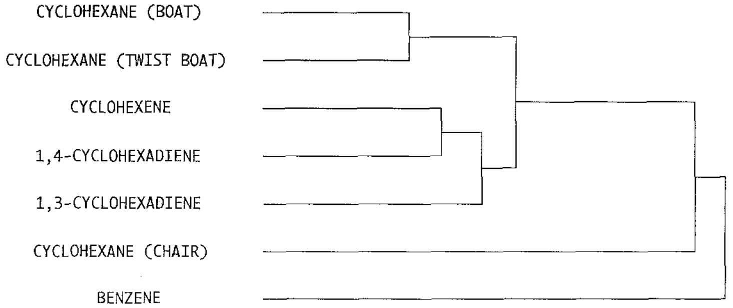


Fig. 18. Six-membered ring dendrogram.

mented if performance is an issue. Additionally, our method has the drawback that it uses a molecular description with a cubic dependence on the number of heavy atoms.

We have used for our comparisons hierarchical clustering methods. These involve a quadratic dependence on the number of molecules and are clearly very inefficient for large databases. However, our similarity measures are not dependent upon the method of clustering used, and they can readily accommodate more efficient clustering techniques applicable to much larger databases [59–62]. Further, if we make use of the hash codes in our shape description scheme, we can avoid many of the time constraints implicit in any clustering method. By sorting a list of compounds by their hash codes (either distance or connectivity based) we have essentially pre-clustered our database, and we may then go through a substantially reduced data set with one of the classical clustering algorithms to gain finely detailed information while at the same time weeding out any artifactual information introduced by our hash code approximation. Of course, as one reviewer has pointed out, the quality of the clustering will be highly dependent upon the quality of the hash codes. To improve this quality, it may be possible to use more than one hash code for this pre-clustering operation.

We view our reliance on comparisons between molecules with the same number of heavy atoms as both a strength and a weakness. Systems with the same number of atoms can be in exact correspondence. Further, the number of compounds that must be compared within a data set is reduced significantly. Clearly any reduction in the size of sets to be compared will lighten the computational burden of clustering a large database since the process scales quadratically with the number of compounds no matter what similarity method is chosen. The weakness lies in this: compounds with very similar shapes but differing numbers of heavy atoms (e.g. methyl vs. ethyl substituent) will be in disjoint sets. This is a problem of underclustering, not overclustering. As indicated, we deal with this situation by calculating the similarity between every molecule with n heavy atoms and every molecule with n + 1 heavy atoms in our database. We can thus form similarity links between heavy atom groupings while at the same time still avoiding the computational burden of calculating a similarity coefficient for all molecule pairings in the database. As an example of the time savings, imagine a database comprised of 10 000 molecules that can be divided into 100 heavy atom groupings. Exhaustive comparison of these molecules will result in 49 995 000 similarity measurements being performed ((n)(n − 1)/2!). If these molecules divide evenly so that there are 100 compounds in each heavy atom grouping, and similarity measurements are performed only between compounds with equal numbers of heavy atoms, it is only necessary to make 495 000 comparisons (100 * ((n/100)((n/100) − 1)/2!)). If we now make our suggested 'similarity link' comparisons between every compound with n atoms and every compound with n + 1 atoms, we only add on an additional 990 000 comparisons for a total of 1 485 000 comparisons (100 * ((n/100)((n/100) − 1)/2!) + 99 * (100 * 100)). This represents a speed up of an order of magnitude with little loss of information. For much larger databases (n $\sim$ 10^7), the savings can be quadratically greater.

Future directions

Extension of this method to include hydrogen atoms is straightforward. Adding hydrogens would have the advantage of providing a more complete molecular description, but it would have the drawback of approximately doubling the number of atoms that must be processed for each molecule. Alternatively, we may be able to provide a useful molecular description by considering the submolecule containing hydrogens only and ignoring heavy atoms. Representation of atomic properties may be usefully described by considering the submolecule containing only polar atoms. Such a set of submolecular graphs would permit coding elemental composition and spatial distribution. Additionally, comparisons between molecules containing different numbers of atoms may also be possible by incorporating some form of null correspondence [26,27] into our histogram generation procedure.

CONCLUSION

Our connectivity- and distance-based shape description methods have been compared with each other and with a distance-based DOCK [52,53] similarity score. The two histogram methods provide similar, but not completely correlated similarities, and as expected, the distance-based histogram correlated best with the distance-based DOCK score, while the connectivity-based histogram was poorly correlated. Compounds with differing conformations but similar connectivities are clustered together by this approach, thereby providing one simple solution to the multiple conformation problem [5–9]. Hash codes based on both histogram methods have been developed, and their usefulness for extremely rapid similarity searches has been shown for a small test case. The connectivity and distance triangle histogram measures of molecular similarity presented here offer a useful, efficient, and novel method for performing shape similarity searches and shape clustering for molecules that have equal numbers of atoms. The process of data abstraction from molecule to constituent triangles to histogram to hash code appears to provide much useful information. The histograms provide a detailed method of molecular shape comparison for dealing with small numbers of molecules, while the hash codes provide an approximate method of comparison for dealing with very large databases. Clearly there are many ways in which this work may be extended. Hydrogen atoms may be considered, atomic features may be incorporated, histogram length and bin sizes may be optimized for specific classes of compounds, and it may even be possible to incorporate these similarity methods usefully into some of the newer methods for automated structure generation [63]. We are currently refining a program for designing new molecules based on our connectivity-histogram scheme.

ACKNOWLEDGEMENTS

This work was supported by grants GM-31497 and GM-39552 (G.L. Kenyon, P.I.). We thank Molecular Design Limited and Tripos Associates for providing their software under an academic license. We thank Dr. Jeff Blaney for helpful discussions.

REFERENCES

1 Dean, P.M., Molecular Foundations of Drug Receptor Interaction, Cambridge University Press, Cambridge, 1987.
2 DesJarlais, R.L., Sheridan, R.P., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 29 (1986) 2149.
3 DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722.
4 Corey, E.J. and Wipke, W.T., Science, 166 (1969) 178.
5 Güner, O.F., Henry, D.R. and Pearlman, R.S., J. Chem. Inf. Comput. Sci., 32 (1992) 101.

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

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

8 Martin, Y.C., Danaher, E.B., May, C.S. and Weininger, D., J. Comput.-Aided Mol. Design, 2 (1988) 15.

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

10 Hodes, L., J. Chem. Inf. Comput. Sci., 29 (1989) 66.

11 Whaley, R. and Hodes, L., J. Chem. Inf. Comput. Sci., 31 (1991) 345.

12 Hodes, L. and Feldman, A., J. Chem. Inf. Comput. Sci., 31 (1991) 347.

13 Hansen, P.J. and Jurs, P.C., J. Chem. Ed., 65 (1988) 574.

14 Bawden, D., J. Chem. Inf. Comput. Sci., 23 (1983) 14.

15 Brugger, W.E., Stuper, A.J. and Jurs, P.C., J. Chem. Inf. Comput. Sci., 16 (1976) 105.

16 Wiener, H., J. Am. Chem. Soc., 69 (1947) 17.

17 Rücker, G. and Rücker, C., J. Chem. Inf. Comput. Sci., 31 (1991) 442 and references therein.

18 Lukovits, I., J. Chem. Inf. Comput. Sci., 31 (1991) 503.

19 Randić, M., J. Am. Chem. Soc., 97 (1975) 6609.

20 Randić, M., In Johnson, M.A. and Maggiora, G.M. (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990, pp. 77–145.

21 Tinker, J., J. Chem. Inf. Comput. Sci., 21 (1981) 3.

22 Johnson, M.A. and Maggiora, G.M. (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990.

23 Danziger, D.J. and Dean, P.M., J. Theor. Biol., 116 (1985) 215.

24 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 295.

25 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 317.

26 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1991) 107.

27 Papadopoulos, M.C. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 119.

28 Kirkpatrick, S., Gelatt, Jr., C.D. and Vecchi, M.P., Science, 220 (1983) 671.

29 Pepperrell, C.A. and Willett, P., J. Comput.-Aided Mol. Design, 5 (1991) 455.

30 Pepperrell, C.A., Poirrette, A.R. and Willett, P., Pestic. Sci., 33 (1991) 97.

31 Varkony. T.H., Shiloach, Y. and Smith, D.H., J. Chem. Inf. Comput. Sci., 19 (1979) 104.

32 Crandell, C.W. and Smith, D.H., J. Chem. Inf. Comput. Sci., 23 (1983) 186.

33 This concept has been described, see: Mihalić, Z., Nikolić, S. and Trinajstić, N., J. Chem. Inf. Comput. Sci., 32 (1992) 28 and references therein.

34 Mihalić, Z. and Trinajstić, N., J. Mol. Struct. (Theochem.), 232 (1991) 65.

35 For a similar descriptor that uses all atom pair distances, see Carhart, R.E., Smith, D.H. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 25 (1985) 64.

36 Freeland, R.G., Funk, S.A., O'Korn, L.J. and Wilson, G.A., J. Chem. Inf. Comput. Sci., 19 (1979) 94.

37 Bawden, D., Catlow, J.T., Devon, T.K., Dalton, J.M., Lynch, M.F. and Willett, P., J. Chem. Inf. Comput. Sci., 21 (1981) 83.

38 Dubois, J.-E., Carrier, G. and Panaye, A., J. Chem. Inf. Comput. Sci., 31 (1991) 574.

39 Willett, P., In Johnson, M.A. and Maggiora, G.M. (Eds.), Concepts and Applications of Molecular Similarity, Wiley, New York, 1990, pp. 43–63.

40 Hartigan, J.A., Clustering Algorithms, Wiley, New York, 1975.

41 Lorr, M., Cluster Analysis for Social Scientists, Jossey-Bass, San Francisco, CA, 1983.

42 Romesburg, H.C., Cluster Analysis for Researchers, Wadsworth, Belmont, CA, 1984.

43 Hansch, C., Sammes, P.G. and Taylor, J.B. (Eds.), Comprehensive Medicinal Chemistry, Pergamon, Oxford, 1990.

44 CMC-3D version 90.1. Available from Molecular Design Ltd., San Leandro, CA.

45 MDDR-3D version 90.1. Available from Molecular Design Ltd., San Leandro, CA.

46 FCD-3D version 89.2. Available from Molecular Design Ltd., San Leandro, CA.

47 Silicon Graphics Computer Systems, Mountain View, CA.

48 Personal communication, Chemical Abstracts Service, Columbus, OH, 1992.

49 Sun Microsystems Inc., Mountain View, CA.

50 Rusinko III, A., Sheridan, R.P., Nilakantan, R., Haraki, K.S., Bauman, N. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 29 (1989) 251 and references therein.
51 CONCORD. Available from Tripos Associates Inc., St. Louis, MO, 63144.
52 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
53 Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comp. Chem., 13 (1992) 505.
54 Ferro, D.R. and Hermans, J., Acta Crystallogr., A33 (1977) 345.
55 Clustering programs MacDendro and GraphMu kindly provided by J. Thioulouse, Laboratoire de Biométrie, Génétique et Biologie des populations, U.R.A. CNRS 243 - Université Lyon 1, 69622 Villeurbanne Cedex, France. See Thioulouse, J., Comput. Appl., Biosci., 5 (1989) 287.
56 SYBYL, Molecular Modelling Software, Tripos Associates Inc., St. Louis, MO 63144, 1991.
57 Willett, P., Similarity and Clustering in Chemical Information Systems, Research Studies Press, Letchworth, Herts., U.K., 1987.
58 Wipke, W.T., Krishnan, S. and Ouchi, G.I., J. Chem. Inf. Comput. Sci., 18 (1978) 32.
59 Jarvis, R.A. and Patrick, E.A., IEEE Trans. Comput., C-22 (1973) 1025.
60 Adamson, G.W. and Bawden, D., J. Chem. Inf. Comput. Sci., 21 (1981) 204.
61 Willett, P., J. Chem. Inf. Comput. Sci., 24 (1984) 29.
62 Willett, P., Winterman, V. and Bawden, D., J. Chem. Inf. Comput. Sci., 26 (1986) 109.
63 Nilakantan, R., Bauman, N. and Venkataraghavan, R., J. Chem. Inf. Comput. Sci., 31 (1991) 527.
64 Cormen, T.H., Leiserson, C.E. and Rivest, R.L., Introduction to Algorithms, The MIT Press, Cambridge, 1990, pp. 558–565.

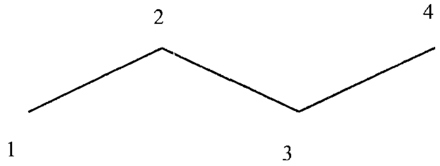
APPENDIX A

As a simple example, the connectivity-based histogram for *n*-butane is calculated as follows:

Start with molecular structure:

$CH_3CH_2CH_2CH_3$

Assign arbitrary numbering of heavy atoms:



Calculate 2 matrices: all-pairs-shortest-paths using Floyd–Warshall algorithm [64], and Euclidean distance. For Euclidean distance matrix, use real coordinates (e.g. CONCORD or X-ray).

	1	*2*	*3*	*4*
1	0	1	2	3
2	1	0	1	2
3	2	1	0	1
4	3	2	1	0

For all 3-atom subgraphs, sum the squares of the sub-diagonals.

3-atom subgraphs Σd^2
subdiagonal

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & & \\ 2 & 1 & 0 & \\ 3 & 2 & 1 & 0 \end{array} \quad =6 \qquad \begin{array}{c|ccc} & 1 & 3 & 4 \\ \hline 1 & 0 & & \\ 3 & 2 & 0 & \\ 4 & 3 & 1 & 0 \end{array} \quad =14$$

$$\begin{array}{c|ccc} & 1 & 2 & 4 \\ \hline 1 & 0 & & \\ 2 & 1 & 0 & \\ 4 & 3 & 2 & 0 \end{array} \quad =14 \qquad \begin{array}{c|ccc} & 2 & 3 & 4 \\ \hline 2 & 0 & & \\ 3 & 1 & 0 & \\ 4 & 2 & 1 & 0 \end{array} \quad =6$$

For each sum, increment the appropriate bin in a histogram.

Bin value	0	2	0	0	2	0	0
Bin range	3–5	6–8	9–11	12–13	14–16	17–18	...
Bin number	*1*	*2*	*3*	*4*	*5*	*6*	...

Use a weighted sum of the bin values to give a hash code.

$$\text{Connectivity hash code} = \sum_{i=1}^{i=64} (i^2\,(\text{bin value})_i) = (2{\cdot}2^2) + (2{\cdot}5^2) = 58.$$

The histograms of individual molecules may be considered as vector data, and molecules may be clustered using any one of a number of hierarchical clustering methods.

APPENDIX B

This routine takes the 'DOCK' matching of two molecules with the exact same number of atoms, and finds an atom by atom pairing of them.

it uses these parameters supplied by DOCK:

"pairs" is ordered list of pairs
sams as iasign(100,2)

"recept" is original receptor crdts
same as spcorr(3,maxpts)

"ligand" is rotated and translated ligand crdts
same as xatm(3,maxlig)

"natl" is number of ligand atoms
used as "number" in program

```
        subroutine match (pairs, recept, ligand, natl)


        implicit none

        include 'max.h'
        include 'chemscore.h'

        integer natl
        integer pairs(100,2)
        real recept(3,*)
        real ligand(3,*)

        integer i, j, k, l
        real dx, dy, dz
        real mindis (maxpts), dist (maxpts, maxpts)
        integer minnum (maxpts)
        logical 1_asn (maxpts), r_asn (maxpts)
        integer number, maxdis, maxnum

        number = natl
        maxdis = -9999999

initialize the arrays

"r_asn" indicates if receptor atom has been assigned yet

"1_asn" indicates if ligand atom has been assigned yet

"pairs" for receptor spheres is same as index so each is assigned

"mindis" is for minimum distance in each row

       do 100 i = 1, number
          r_asn (i) =.false.
          l_asn (i) =.false.
          mindis (i) = 9999999.0
          pairs (i, 1) = i
100    continue

calculate distance for each receptor-ligand pair

       do 300 i = 1, number
          do 200 j = 1, number
             dx = (recept (1, i)-ligand(1, j))**2
             dy = (recept (2, i)-ligand(2, j))**2
             dz = (recept (3, i)-ligand(3, j))**2
             dist (i, j) = sqrt (dx + dy + dz)
200       continue
300    continue

go thru each row, find minimum, then find maximum of minima
```

```
      do 1000 l = 1, number
          do 500 i = 1, number
              do 400 j = 1, number
                  if((dist(i, j).lt.mindis(i)).and.(.not.r_asn(i))
     &            .and.(.not.l_asn(j)))then
                      mindis (i) = dist (i,j)
                      minnum (i) = j
                  endif
400           continue
500       continue

calculate maximum distance

          maxdis = -9999999

          do 700 k = 1, number
              if ((mindis(k).gt.maxdis).and.(.not.r_asn(k)))then
                  maxdis = mindis (k)
                  maxnum = k
              endif
700       continue

assign match-> maxnum (receptor) to minnum[maxnum] (ligand)

          l_asn (minnum(maxnum)) =.true.
          r_asn (maxnum) =.true.
          pairs (maxnum,2) = minnum(maxnum)

set minimum distance for each receptor center to large value

          do 800 i = 1, number
              mindis (i) = 9999999.0
800       continue

1000 continue

      return
      end
```