Similarity Searching in Databases of Flexible 3D Structures Using Smoothed Bounded Distance Matrices

John W. Raymond*

Pfizer Global Research and Development, Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, Michigan 48105

Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom

Received January 9, 2003

This paper describes a method for calculating the similarity between pairs of chemical structures represented by 3D molecular graphs. The method is based on a graph matching procedure that accommodates conformational flexibility by using distance ranges between pairs of atoms, rather than fixing the atom pair distances. These distance ranges are generated using triangle and tetrangle bound smoothing techniques from distance geometry. The effectiveness of the proposed method in retrieving other compounds of like biological activity is evaluated, and the results are compared with those obtained from other, 2D-based methods for similarity searching.

INTRODUCTION

Similarity searching in databases of 2D chemical structures is widely used for virtual screening in lead-discovery programs. The similarity measure that is used to quantify the degree of structural resemblance between the target structure and each of the database structures is usually based on fingerprint representations of chemical structure (i.e., bit-strings encoding the presence of 2D fragment substructures), with the similarity between pairs of such representations being computed using the Tanimoto coefficient.[1,2] However, we have recently described an efficient new algorithm for the detection of 2D maximum common subgraphs (MCS)[3,4] and demonstrated that this can be used to provide facilities for MCS-based 2D similarity searching that are complementary to those in existing, fingerprint-based search systems.[5]

Similarity searching in databases of 3D structures is less well-established than in the case of 2D structures, with measures based on geometric substructural features and on molecular fields under active investigation.[6] Much of the work in the former area has involved similarity measures based on 3D pharmacophore patterns (see, e.g., refs 7−9), but we focus here on the use of 3D MCSs for similarity searching (see, e.g., refs 10−12). Specifically, we describe a new method for MCS-based 3D similarity searching that extends previous studies by including conformational flexibility in the matching algorithm, while still being sufficiently rapid in execution for similarity searching of databases of nontrivial size.

SIMILARITY METHODOLOGY

Chemical Graphs. All graphs referred to in the following text are assumed to be simple, undirected graphs. For an introduction to graph related concepts and notation, the reader is referred to an introductory text on graph theory.[13] The chemical applications of the maximum common subgraph concept have been reviewed by Raymond and Willett.[14,15] In a 2D chemical structure, the graph vertices represent the atoms, and the graph edges denote the bonds connecting each pair of covalently bonded atoms. In a 3D chemical graph, the vertices of the graph again denote the atoms but the edges here indicate the geometric distance or range of distances between a pair of atoms (vertices). An edge in one graph is compatible with an edge in another graph if their two vertex endpoints and edge labels are compatible. In a 3D chemical structure graph, compatibility of edge labels may involve the specification of some allowable distance tolerance. Figure 1 illustrates both a simple 2D and 3D chemical graph representation of acetic acid using CONCORD for interatomic distance generation.

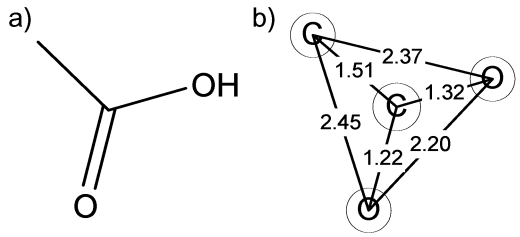

Figure 1. (a) 2D chemical graph and (b) 3D chemical graph representations of acetic acid.

In this paper, we focus on establishing the similarity between two molecules represented as 3D graphs, where the similarity between two molecules G_1 and G_2 is defined using the coefficient of Wallis et al.,[16] which is simply the graph-based equivalent of the Tanimoto coefficient that is commonly used for fingerprint-based similarity searching.[16] This coefficient is of the form

$$S_{12} = \frac{|G_{12}|}{|G_1| + |G_2| - |G_{12}|}$$

* Corresponding author e-mail: john.raymond@pfizer.com.

10.1021/ci034002p CCC: $25.00
Published on Web 04/04/2003

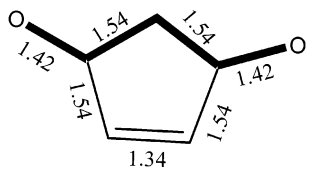

Figure 2. Example shortest path upper bound calculation.

where $|G_1|$ and $|G_2|$ are the numbers of atoms in two molecules, respectively, and $|G_{12}|$ is the number of atoms in the 3D MCS resulting from a graph matching between the two molecules. The 3D MCS consists of the largest set of atoms common to both molecules that preserves all pairwise distance constraints in both molecules.

One difficulty that arises using this definition of similarity is that most molecules exhibit some degree of flexibility, and thus the distances between atoms are not fixed. One approach to cope with this drawback is to generate several conformations for each structure under consideration and then to compare all possible pairs of the resulting conformations. This approach, however, is computationally demanding if a nontrivial sampling of the molecules' conformational space is to be achieved and still cannot guarantee that the optimal similarity has been identified. The approach taken here is to encode a molecule's conformational flexibility within a single graphical representation. Specifically, each edge of a 3D molecular graph is represented by a range of distances spanning the maximum and minimum allowable distance between two atoms.[17,18]

Distance Geometry. To establish the maximum and minimum possible distances between all pairs of atoms in a molecule, we implement the bound smoothing procedures of *distance geometry*.[19] Given an initial set of upper and lower bounds between all atom pairs in a molecule, the bound smoothing procedures serve to eliminate geometric inconsistencies by iteratively sharpening the upper and lower bounds. For a molecule with N atoms, the initial set of upper and lower bounds are represented by two symmetric $N \times N$ matrices, U and L, where the elements $U(i,j)$ and $L(i,j)$ correspond to the upper and lower distance bounds between atoms i and j, respectively.

In our implementation, initial upper-bound values, $U(i,j)$, for each hydrogen-suppressed molecular graph are determined by calculating the shortest bond length path between atoms i and j. This is a simple consequence of the fact that the farthest apart any two atoms in a molecule could possibly be is the sum of the distances for the shortest string of bonds between those two atoms. For example, in Figure 2, the upper-bound between the two oxygen atoms is calculated as 5.92 (i.e., $1.42 + 1.54 + 1.54 + 1.42$). Efficient shortest path algorithms are readily available.[20] Initial lower-bound values, $L(i,j)$, involving atoms i and j that are separated by four or more bonds are calculated as the sum of their van der Waals radii to serve as a "bump-checking" constraint. The upper and lower-bounds for all (i,j) atom pairs separated by two or fewer bonds are fixed using standard bond lengths and valence angles, so that $U(i,j) = L(i,j)$ for all atom pairs where i and j are separated by one or two bonds. In addition, all atom pair distances contained within aromatic rings are fixed using standard bond lengths, and no rotation about double and triple bonds is allowed. Any lower-bound values, $L(i,j)$, not addressed by the aforementioned constraints are set to zero as the sum of van der Waals radii was found to be too restrictive for some atom pairs separated by fewer than four bond lengths.

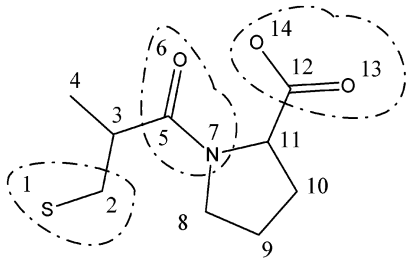

Figure 3. Captopril example.

The bound smoothing scheme used here consists of two stages, *triangle smoothing* and *tetrangle smoothing*, which can be readily implemented from their published descriptions.[19,21,22] Triangle smoothing, the simpler of the two procedures, considers all atom 3-tuples (i, j and k) and enforces the triangle-inequality limits $U(i,j) \leq U(i,k) + U(j,k)$ and $L(i,j) \geq L(i,k) - U(j,k)$ by iteratively decreasing and increasing the upper and lower-bounds, respectively. Triangle smoothing is described in detail by Crippen and Havel[19] and has a time complexity of $O(N^3)$. Tetrangle smoothing can be used to further sharpen the upper and lower-bound values by considering all 4-tuples and enforcing the tetrangle inequalities by means of Cayley-Menger determinants. Unlike triangle smoothing, no polynomial complexity algorithm is known for tetrangle smoothing. In fact, it has been conjectured to be an NP-hard problem for which no polynomial complexity algorithm is known.[19] For this reason, the tetrangle smoothing implementation used here includes a variable parameter which indicates the maximum number of times to cycle over all 4-tuples during tetrangle smoothing. The effect of increasing the threshold parameter on the effectiveness of 3D chemical similarity searching is investigated in subsequent sections.

Figure 3 depicts the structure of the ACE inhibitor, captopril, and Figure 4 lists all atom to atom distances bounds for its 14 atoms using the triangle and tetrangle bound smoothing previously described. There is not a notable difference for most of the upper and lower bounds between the three bounding schemes for this case since captopril contains only 14 non-hydrogen atoms. However, as the size and complexity of the molecules under consideration increase, the differences can become quite significant. Figure 5 depicts two example structures whose tetrangle-smoothed bounds are a significant improvement over the triangle-smoothed bounds. A measure of distance bound improvement is defined as

$$\Delta d = 2 \cdot \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} (U^{tri}(i,j) - U^{tet}(i,j)) + (L^{tet}(i,j) - L^{tri}(i,j))}{N(N-1)}$$

where $U^{tri}(i,j)$ and $L^{tri}(i,j)$ indicate the upper and lower triangle-smoothed bounds and $U^{tet}(i,j)$ and $L^{tet}(i,j)$ indicate the upper and lower tetrangle-smoothed bounds between atoms i and j. Δd then represents the average distance each atom-to-atom pair distance range was contracted using the tetrangle-smoothing procedure after triangle-smoothing. Using this measure of bound improvement, Figure 5 demonstrates that tetrangle-smoothing can appreciably sharpen

Triangle Bounds (Å)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		1.82	2.74	4.26	4.25	5.12	5.20	6.67	7.65	7.65	6.67	7.64	8.86	8.96
2	1.82		1.53	2.49	2.48	3.71	3.81	4.98	6.26	6.26	4.98	6.25	7.35	7.43
3	2.74	1.53		1.53	1.51	2.38	2.46	3.93	4.91	4.91	3.93	4.90	6.12	6.22
4	1.21	2.49	1.53		2.48	3.71	3.81	4.98	6.26	6.26	4.98	6.25	7.35	7.43
5	1.52	2.48	1.51	2.48		1.23	1.33	2.50	3.78	3.78	2.50	3.77	4.87	4.95
6	2.75	1.25	2.38	1.25	1.23		2.22	3.69	4.67	4.67	3.69	4.66	5.88	5.98
7	2.84	1.23	2.46	1.23	1.33	2.22		1.47	2.45	2.45	1.47	2.44	3.66	3.76
8	2.88	2.70	1.17	2.70	2.50	1.27	1.47		1.54	2.41	2.36	3.87	4.73	4.82
9	2.88	2.70	2.70	2.70	1.34	2.57	2.45	1.54		1.54	2.41	3.92	4.78	4.86
10	2.88	2.70	2.70	2.70	1.34	2.57	2.45	2.41	1.54		1.54	2.50	3.72	3.82
11	2.88	2.70	1.19	2.70	2.50	1.27	1.47	2.36	2.41	1.54		1.51	2.37	2.45
12	2.88	2.70	2.70	2.70	1.35	2.57	2.44	1.35	1.35	2.50	1.51		1.22	1.32
13	2.75	2.57	2.57	2.57	2.57	2.45	1.24	2.57	2.57	1.28	2.37	1.22		2.20
14	2.75	2.57	2.57	2.57	2.57	2.45	1.24	2.57	2.57	1.18	2.45	1.32	2.20	

Tetrangle Bounds (Å), 1 Pass

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		1.82	2.74	4.12	4.10	5.11	5.20	6.59	7.64	7.64	6.59	7.64	8.86	8.96
2	1.82		1.53	2.49	2.48	3.64	3.73	4.98	6.18	6.18	4.98	6.17	7.28	7.38
3	2.74	1.53		1.53	1.51	2.38	2.46	3.87	4.91	4.91	3.87	4.90	6.06	6.16
4	2.66	2.49	1.53		2.48	3.63	3.73	4.98	6.18	6.18	4.98	6.17	7.28	7.37
5	2.65	2.48	1.51	2.48		1.23	1.33	2.50	3.74	3.74	2.50	3.73	4.81	4.90
6	2.75	2.66	2.38	2.66	1.23		2.22	3.61	4.67	4.67	3.61	4.66	5.81	5.91
7	2.84	2.70	2.46	2.70	1.33	2.22		1.47	2.45	2.45	1.47	2.44	3.58	3.68
8	2.88	2.70	2.96	2.70	2.50	2.82	1.47		1.54	2.41	2.36	3.60	4.67	4.77
9	2.88	2.70	2.70	2.70	2.78	2.57	2.45	1.54		1.54	2.41	3.69	4.74	4.83
10	2.88	2.70	2.70	2.70	2.78	2.57	2.45	2.41	1.54		1.54	2.50	3.64	3.74
11	2.88	2.70	2.96	2.70	2.50	2.82	1.47	2.36	2.41	1.54		1.51	2.37	2.45
12	2.88	2.70	2.70	2.70	2.80	2.57	2.44	2.42	2.40	2.50	1.51		1.22	1.32
13	2.75	2.57	2.57	2.57	2.57	2.45	2.64	2.57	2.57	2.68	2.37	1.22		2.20
14	2.75	2.57	2.57	2.57	2.57	2.45	2.68	2.57	2.57	2.71	2.45	1.32	2.20	

Tetrangle Bounds (Å), 2 Pass

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1		1.82	2.74	4.12	4.10	5.11	5.20	6.57	7.64	7.64	6.57	7.64	8.78	8.87
2	1.82		1.53	2.49	2.48	3.64	3.73	4.98	6.18	6.18	4.98	6.17	7.28	7.38
3	2.74	1.53		1.53	1.51	2.38	2.46	3.87	4.91	4.91	3.87	4.90	6.04	6.14
4	2.66	2.49	1.53		2.48	3.63	3.73	4.98	6.18	6.18	4.98	6.17	7.28	7.37
5	2.65	2.48	1.51	2.48		1.23	1.33	2.50	3.74	3.74	2.50	3.73	4.81	4.90
6	2.75	2.66	2.38	2.66	1.23		2.22	3.61	4.67	4.67	3.61	4.66	5.80	5.89
7	2.84	2.70	2.46	2.70	1.33	2.22		1.47	2.45	2.45	1.47	2.44	3.58	3.68
8	2.88	2.70	2.96	2.70	2.50	2.82	1.47		1.54	2.41	2.36	3.60	4.67	4.77
9	2.88	2.70	2.70	2.70	2.78	2.57	2.45	1.54		1.54	2.41	3.69	4.74	4.83
10	2.88	2.70	2.70	2.70	2.78	2.57	2.45	2.41	1.54		1.54	2.50	3.64	3.74
11	2.88	2.70	2.96	2.70	2.50	2.82	1.47	2.36	2.41	1.54		1.51	2.37	2.45
12	2.88	2.70	2.70	2.70	2.80	2.57	2.44	2.42	2.40	2.50	1.51		1.22	1.32
13	2.75	2.57	2.57	2.57	2.57	2.45	2.64	2.57	2.57	2.68	2.37	1.22		2.20
14	2.75	2.57	2.57	2.57	2.57	2.45	2.68	2.57	2.57	2.71	2.45	1.32	2.20	

Figure 4. Distance bound smoothing example. Upper (lower) diagonal values indicate upper (lower) bound distances, respectively.

the atom-to-atom distance ranges. In this case, the average atom-to-atom distance range contraction ranged from 2.01 to 4.22 Å.

Clark et al.[17,18] have previously investigated the effectiveness of using triangle smoothing on small 3D substructure queries. This work seeks to determine whether distance geometry bounds can also be used effectively in 3D similarity searching applications. This study also investigates the extent to which the increased resolution obtained from tetrangle smoothing affects search performance. For instance, given an atom 4-tuple (*hijk*), the upper and lower bounds ($U^{tet}(h,k)$ and $L^{tet}(h,k)$) calculated between atoms *i* and *l* using tetrangle smoothing are equal to the trans and cis distance limits.[19]

Graph Matching. Once the set of structures to be queried has been preprocessed and stored using the upper- and lower-bound values calculated using the previously described distance bounding techniques, similarity searching can then be performed by using an MCS graph matching routine. An efficient MCS algorithm has recently been published which has been used successfully to determine the MCS in 2D chemical graphs[3,4] but which is directly amenable to any arbitrary type of graph. Here we implement the algorithm using the 3D graphs just described. The algorithm is a clique-based method which computes the MCS by determining the *maximum clique* in a correspondence graph (also known as the *modular product*[23] or *association graph*[24]), where a *maximum clique* in a graph *G* is the largest subset of vertices in the graph such that each pair of vertices in the subset is connected by an edge in the graph *G*. This process is described in detail in the published account.[3]

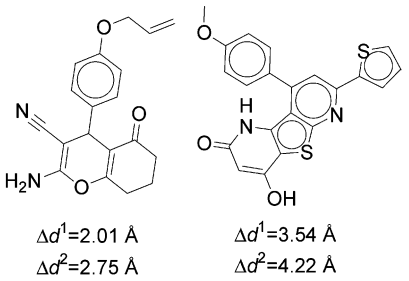

Figure 5. Example structures where tetrangle-smoothed bounds are a significant improvement over triangle-smoothed bounds. Δd^1 and Δd^2 represent the average atom-to-atom improvement in the upper and lower bounds (as described in the text) for the 1- and 2-pass tetrangle bound options, respectively, as compared to the simpler triangle smoothing.

The vertices of the correspondence graph are constructed by creating the set of all compatible pairs of atoms consisting of one atom from each molecular graph being compared such that the atom types are identical. This atom type condition could be relaxed using a more liberal definition of atom type compatibility to incorporate some "fuzziness" or chemical knowledge in the matching process. Edges are formed between the vertices of the correspondence graph if the two distance ranges specified by the two atom pairs in each molecular graph overlap. There exists an edge between two vertices in the correspondence graph, (i,x) and (j,y), if the distance bounds between atoms i and j in the first graph, $U_1(i,j)$ and $L_1(i,j)$, and the distance bounds, $U_2(x,y)$ and $L_2(x,y)$, between atoms x and y in the second graph obey both of the conditions $L_2(x,y) \leq U_1(i,j) + \epsilon$ and $U_2(x,y) \geq L_1(i,j) - \epsilon$, where the constant ϵ is a distance tolerance value. In addition to this compatibility criterion, we have found that requiring $\max\{U_1(i,j) + U_2(x,y)\} \leq 2 \cdot \min\{U_1(i,j) + U_2(x,y)\}$ helps to prevent unfavorable atom matchings and provides better results.

The published matching procedure[3] consists of two components, a screening and a rigorous matching procedure. However, the screening procedure requires that each label for an edge between two vertices (atoms) be unambiguous. This condition is obviously met for a 2D graph where each edge represents an unambiguous bond type, but for the 3D graphs used here this is not the case as it is typical for a given distance range to intersect with many other specified distance ranges. It is possible, however, to implement the screening procedure on 3D graphs when the interatomic distances are assumed to be fixed (i.e., fixed conformations) and can therefore be classified into bins representing unambiguous distance labels. All of the 3D similarity calculations performed here, however, consist only of the rigorous graph matching component of the originally published algorithm, but we have used a screening procedure based on our previously described procedure for 2D similarity searching. The procedure involves setting a 2D threshold similarity value below which graph-matching will not be carried out: specifically, the similarity has been set here at the (low) level of 0.50, thus allowing significant variations with respect to 2D connectivity while simultaneously preventing many meaningless and computationally expensive 3D matchings.

VALIDATION STUDIES

Efficiency Evaluation. The principal focus of this study is the potential effectiveness of the 3D MCS approach to chemical similarity, but we have also carried out some small-scale studies of the method's efficiency. The test-set used here contained a total of 250 compounds downloaded from the Web,[25] these containing a mean of 26.6 (standard deviation of 5.9) atoms per molecule. For this simulation, all 31125 pairwise comparisons were performed. The simulation consisted of a total of 60 trials consisting of five distance tolerance values (ϵ = 0.1, 0.2, 0.3, 0.5, and 0.75 Å) for each of the four distance representations: fixed distances (assuming a single conformation obtained from the CONCORD program); triangle smoothing bounds; and tetrangle smoothing bounds, with either one pass or two passes through all of the 4-tuples. Each of these 20 combinations of tolerance values and distance representations were run at three different 3D similarity thresholds (used to lower bound the size of the maximum clique[3]): the similarity was calculated using the coefficient of Wallis et al. with the thresholds being 0.20, 0.40, and 0.60. The preprocessed distance bound files were computed as described previously.

Table 1 lists the per comparison time results for each of the 60 time trials using Visual C++ 6.0, Windows 2000, 522 Mb RAM, and a 1 GHz Intel processor. Minimum similarity thresholds of 0.20, 0.40, and 0.60 were set for the 3D graph matching procedure and 0.50 for the 2D screening procedure (as previously discussed). As Table 1 demon-

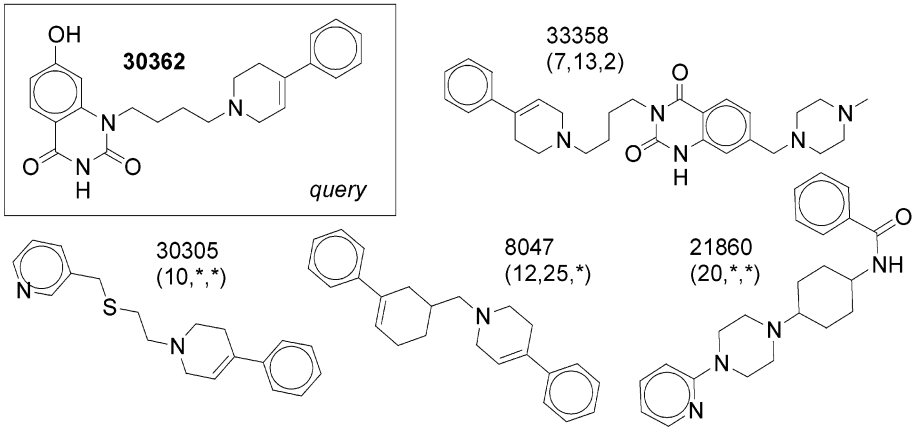

Figure 6. Query example 1. The notation (x,y,z) denotes the rank positions in the 25-member hit-list for that molecule in the hit-lists from the 3D MCS (Tet2, T1), 2D MCS, and Daylight fingerprints, respectively. * indicates not present in the hit-list.

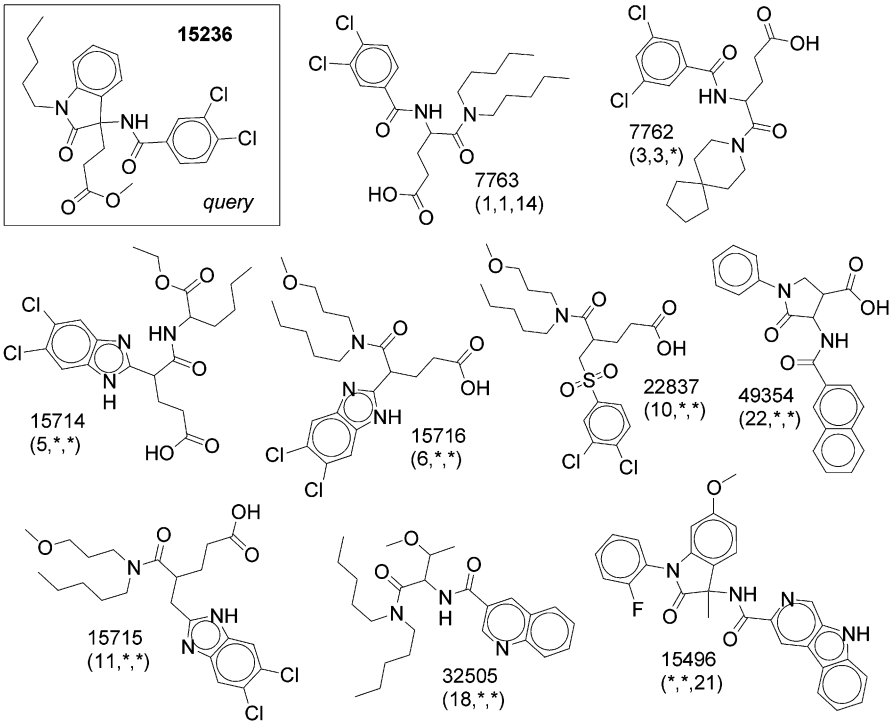

Figure 7. Query example 2. Notation (x,y,z) same as in Figure 6.

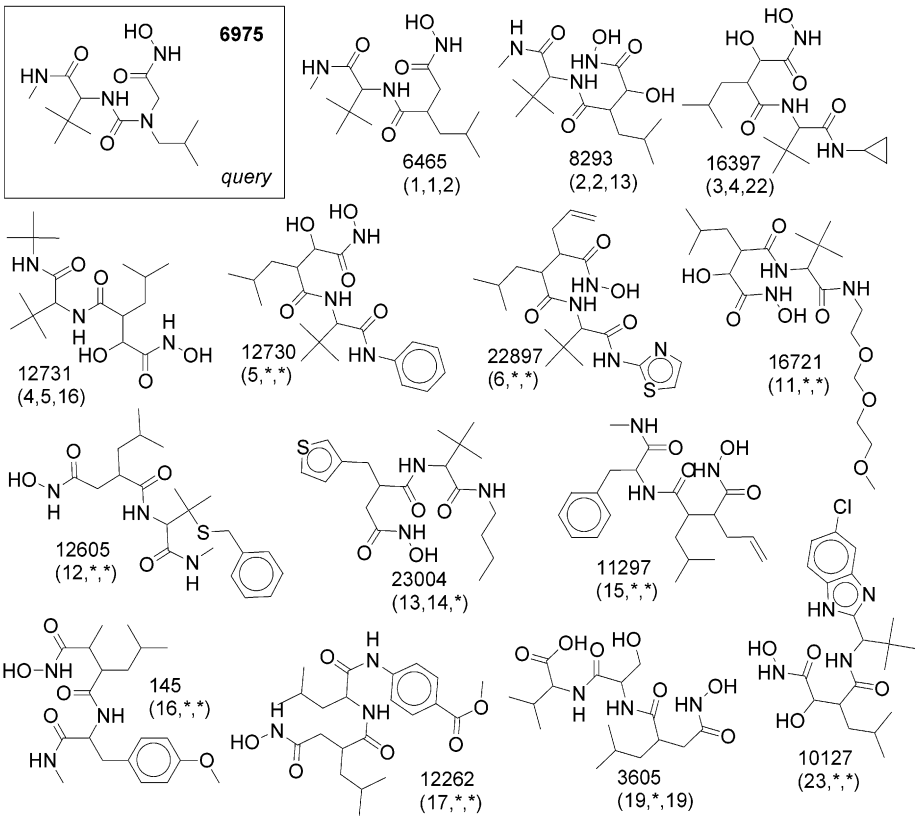

Figure 8. Query example 3. Notation (x,y,z) same as in Figure 6.

strates, increasing the threshold values decreases the time per comparison, especially at the lower distance tolerance values, and it might well be worth quantifying the nature of the tradeoff in effectiveness for such increases in efficiency in the future. The table also shows that the fixed atom-to-atom distances simulations ranged from one to three times faster than the conformationally inclusive distance ranges, with the largest differences occurring at the lower similarity thresholds. It also appears that the tetrangle smoothed bounds were slightly faster than the triangle smoothed bounds. On a related note, the currently implemented 2D similarity threshold of 0.5 excludes only 17 846 (57%) of the 31 125 possible pairwise comparisons from the more computationally expensive 3D graph matching procedure, thus providing

Table 1. Average Time Per Comparison for a Data Set of 250 Compounds (s)

		distance tolerance (Å)				
simulation	ST	0.1	0.2	0.3	0.5	0.75
Fix	0.2	0.004	0.004	0.005	0.007	0.007
Tri		0.013	0.015	0.015	0.015	0.015
Tet1		0.011	0.013	0.013	0.014	0.014
Tet2		0.011	0.012	0.013	0.013	0.013
Fix	0.4	0.003	0.003	0.003	0.004	0.004
Tri		0.011	0.015	0.015	0.015	0.015
Tet1		0.008	0.012	0.013	0.014	0.014
Tet2		0.007	0.010	0.011	0.012	0.013
Fix	0.6	0.003	0.003	0.003	0.003	0.003
Tri		0.005	0.005	0.006	0.006	0.006
Tet1		0.004	0.004	0.005	0.005	0.005
Tet2		0.004	0.004	0.004	0.005	0.005

considerable scope for further increases in search efficiency if required.

Effectiveness Evaluation. The effectiveness of the searches was measured using the cumulative recall measure, as described by Edgar et al.[26] Recall is the fraction of the active structures that are retrieved in the search (a) over the total number of active structures in the database (A), i.e.

$$R = \frac{a}{A}$$

with the cumulative recall being simply the recall at some fixed cutoff number of top-ranked molecules, e.g., the top-20 nearest neighbors. The cumulative recall achieved using the 3D MCS similarity method is compared here with the recall for corresponding search outputs generated with 2D fingerprints and 2D MCS-based similarities.[5]

The data set used in the following analysis is identical to one used to evaluate the relative effectiveness of 2D MCS and fingerprint similarities for the retrieval of biologically active compounds.[5] The original set contains 11 607 compounds from the ID Alert database that have been classified according to pharmacological activity, together with 100 target structures that were obtained as follows: 100 activity classes were selected at random, subject to there being at least 20 compounds with that activity; one of the compounds was then chosen at random from each of the selected activity classes; each such resulting compound was used as the target structure for a similarity search of the ID Alert file. For the present study, it was possible to generate 11 175 3D structures for these compounds using the CONCORD program, this figure including 96 of the original target structures. The 96 query structures contain 8.6 rotatable bonds on average (standard deviation of 7.1). Of the 96 query compounds, 62 of the structures had six or more rotatable bonds, and 28 had 10 or more rotatable bonds. The effectiveness of each similarity search was determined by seeing how many of the top-ranked molecules belonged to the same activity class as the target structure, i.e., would exhibit the same activity in a real screening program.

Table 2 lists recall values for hit-lists of 25, 50, 75, and 100 structures averaged over all 96 queries for each similarity measure. Both the Daylight 2D fingerprint and the 2D MCS measures exhibit comparable results. The 3D MCS results, however, show somewhat more variability. As one would expect, the searches employing fixed distances between atom pairs on a single conformation resulted in lower values of recall than the flexible approaches, although increasing the distance tolerance (0.1 to 0.5 Å) slightly increases recall values, making them somewhat comparable with the 2D similarity values. The highest average recall values were associated with the tetrangle smoothed bounds at a distance tolerance of 0.1 Å. A sign test over all of the sets of 96 recall values for each hit-list cutoff value (i.e., $R_{25} - R_{100}$) at each distance tolerance for the two respective tetrangle bound cases with the triangle smoothed bounds was used to test the hypothesis that there was a statistically significant difference in the numbers of actives retrieved between the triangle and tetrangle bounding scenarios. The sign test revealed that at the 0.1 Å tolerance value the differences between the tetrangle and triangle smoothed bounds were significant at a 0.05 significance level at all hit-list cutoff values. However, no such differences were observed at any of the other tolerance values. A sign test was also performed comparing the tetrangle smoothed bounds at a tolerance of 0.1 Å with both the Daylight and 2D MCS recall values. A statistically significant difference between the Daylight and tetrangle bound cases was observed at all hit-list cutoffs. In fact, the differences were significant to a 0.0025 significance level. The difference between the 2D MCS and the tetrangle bounds cases were also statistically significant. At a significance level of 0.05, a difference was distinguished for all hit-list cutoffs with the exception of R_{25} where the significance level was 0.06. This helps support the conclusion that the tetrangle bounds using a distance tolerance of 0.1 Å performed the best of the methods tested in these simulations.

Table 2. Ranking Results for Fingerprint-Based Similarity Coefficients[a]

ID	R_{25}	R_{50}	R_{75}	R_{100}	ID	R_{25}	R_{50}	R_{75}	R_{100}
Daylight	0.08	0.11	0.13	0.15	2D MCS	0.09	0.12	0.14	0.15
3D MCS Fix T1	0.07	0.10	0.11	0.12	3D MCS Tet1 T1	0.10	0.14	0.16	0.18
3D MCS Fix T2	0.08	0.11	0.12	0.13	3D MCS Tet1 T2	0.09	0.13	0.14	0.16
3D MCS Fix T3	0.08	0.11	0.12	0.13	3D MCS Tet1 T3	0.09	0.13	0.15	0.16
3D MCS Fix T5	0.08	0.11	0.13	0.14	3D MCS Tet1 T5	0.09	0.13	0.15	0.16
3D MCS Fix T75	0.08	0.11	0.13	0.14	3D MCS Tet1 T75	0.09	0.12	0.15	0.16
3D MCS Tri T1	0.09	0.13	0.15	0.16	3D MCS Tet2 T1	0.10	0.14	0.16	0.18
3D MCS Tri T2	0.09	0.12	0.14	0.16	3D MCS Tet2 T2	0.10	0.13	0.15	0.16
3D MCS Tri T3	0.09	0.13	0.15	0.16	3D MCS Tet2 T3	0.09	0.13	0.14	0.16
3D MCS Tri T5	0.09	0.13	0.15	0.16	3D MCS Tet2 T5	0.10	0.13	0.15	0.16
3D MCS Tri T75	0.09	0.13	0.14	0.16	3D MCS Tet2 T75	0.09	0.12	0.14	0.16

[a] The values listed are the mean recall for all 96 queries for hit-lists of sizes 25, 50, 75, and 100 compounds. Ty: distance tolerance value = 0.y Å.

Recall that a set of N points (atoms) has $3N-6$ degrees of freedom in three dimensions. Therefore, given a set of $N(N-1)/2$ fixed distances, the geometry will be overspecified for N of any meaningful magnitude as only a subset of distances is necessary to delineate the corresponding coordinate geometry.[27] It is conjectured that this is precisely why the proposed method has proven effective. Since molecular

structures possess some degree of flexibility, we were forced to introduce uncertainty into the atom-pair distances using maximum and minimum allowable distance ranges. This added uncertainty, however, was partially offset, by the combinatorics involved. The uncertainty introduced to any one atom-pair distance by providing for conformational flexibility is somewhat offset by the over-specification involved in having all pairwise distance ranges for a set of *N* atoms so that combinations of atom-pair distance ranges that are more discriminating can compensate for those which are found to be less effective.

Note that the variation of the distance tolerance with respect to the recall of biologically active compounds is an attempt to discover an equilibrium between maintaining the geometric integrity of localized substructure which may or may not be considered a pharmacophore and at the same time allowing for some global distance range variability between the substructural elements. These studies have indicated that maintaining the local substructure using lower distance tolerance values is significantly more important than allowing for global distance range variability between the substructural elements. This is one of the reasons 2D similarity methods such as fingerprints have proven so effective in practice. Our current research on this topic involves "layering" the matching process so that small distance tolerances are used for closely located atom-pairs and larger distance tolerances for more distant atom-pairs, thus allowing for some geometric imprecision between substructural elements. As an example of this paradox, relaxing the distance tolerance to a value of 1.25 Å may allow the location of pharmacophoric substructures to become geometrically compatible, but it typically would also prevent any geometric differentiation between every atom-pair within three bond lengths, dramatically reducing both the effectiveness and efficiency of the similarity comparison.

While the recall values in Table 2 reveal that the 3D MCS with tetrangle smoothing bounds, the 2D MCS, and the Daylight approaches perform well when averaged over all queries, they provide little indication as to the variability in the types of active structures retrieved in response to each of the individual target structures. To ascertain whether the proposed 3D approach is able to retrieve active structures distinct from the two 2D methods, a cumulative hit-list similarity was calculated using the asymmetric coefficient $c/\min\{a,b\}$ where a is the number of actives retrieved using measure Tet2 T1 for a hit-list of 100 compounds, b is the number of actives retrieved using the Daylight or 2D MCS method for a hit-list of 100 compounds, and c is the number of actives common to both similarity measures over all hit-lists of 100 compounds. This measure is denoted S_a. Additionally the average hit-list similarity was also calculated using the asymmetric coefficient; the average similarity was obtained by determining the value of S_a for each individual query and then averaging them over all queries to give a value denoted by $\bar{S}_a$.

Table 3 lists the respective values of S_a and $\bar{S}_a$ as well as the total number of actives retrieved by each measure. It is evident from the values listed that the 3D MCS method (Tet2 T1) was not only able to retrieve more active structures but was also able to retrieve active structures distinct from the 2D methods. There is a fair degree of difference between the sets of retrieved active structures, even though we have used a 2D similarity threshold to minimize the computational requirements (thus increasing the similarity between the 2D and 3D hit-lists). This indicates that the 3D approach is able, to an appreciable degree, to attach significance to structural features distinct from the 2D methods. For instance, Table 3 shows that the 3D MCS method was able to retrieve 383 (889−723*0.7) active structures distinct from those retrieved using Daylight fingerprints. This may allow the proposed method to be used in "scaffold-hopping" applications where one is interested in compounds with similar geometric placement of pharmacophore elements but with different structural templates.[28−30]

Table 3. Hit-List Similarities (R_{100}) Using the Asymmetric Coefficient[a]

method	N_T	S_a	$\bar{S}_a$
Daylight	723	0.70	0.82
2D MCS	764	0.77	0.88
3D MCS Tet2 T1	889		

[a] N_T: total number of active compounds retrieved over all R_{100} queries.

Example Search. As Table 3 shows, there is a fair degree of difference between the search outputs for individual queries. For example, Figures 6−8 depict various example queries where the 3D MCS method (Tet2 T1) performed better than either of the 2D approaches We have not found it possible, thus far, to specify those cases where one method (2D or 3D) will perform better than the other using simple molecular descriptors such as the number of atoms, rings, and rotatable bonds. A simpler and potentially more successful approach may be to fuse the 2D and 3D rankings into a single ranking.[5,31]

CONCLUSIONS

In this paper, we have described a new method for the calculation of intermolecular structural similarity in searches of databases of 3D structures. Although taking full account of conformational flexibility, the method is sufficiently rapid to allow searches of databases of nontrivial size. Simulated virtual screening experiments demonstrate that the method is broadly comparable in effectiveness to published methods for 2D similarity searching and retrieves many active molecules that are not identified by the 2D methods.

There are many ways in which this work could be extended. For example, rather than requiring merely the presence of a nonzero overlap between a pair of distance ranges for the creation of an edge in the correspondence graph, one could require at least some minimal distance overlap or even adopt a weighted scheme in which the edge is labeled by the extent of the overlap, thus favoring strongly similar ranges. Another potentially useful approach involves implementing a bounding scheme in which the upper and lower bounds also take into account energetically unfavorable distance ranges without assuming any energetically optimal conformations: the supplemental distance range compatibility criterion, $\max\{U_1(i,j) + U_2(x,y)\} \leq 2 \cdot \min\{U_1(i,j) + U_2(x,y)\}$, was a crude step in this direction. A more sophisticated distance compatibility criterion should not only improve the effectiveness of the proposed method but also increase the efficiency by simplifying the clique detection

process by reducing the number of compatible distance ranges. Again, in the experiments presented here all atoms are considered equivalently, but the method can readily be tailored to the situation where it is known that certain atoms are appreciably more important than others. For example, the atoms corresponding to the ACE pharmacophore are known[32] and are outlined in Figure 3. Their corresponding atom to atom distance pairs are also highlighted in Figure 4. In instances such as this, these atoms can be weighted more heavily than the other secondary atoms in the similarity calculation. It would also be possible to use a 3D, rather than 2D, screening procedure prior to the graph-matching stage of the similarity method to improve efficiency. The current approach to screening is based on the 2D screening procedure involving a simple mapping of augmented atoms (i.e., an atom and its bonded neighbor atoms) in a molecule with another molecule, presuming that is unlikely that two molecules will exhibit a high degree of 3D similarity when their constituent augmented atoms show a low degree of correlation. However for general applicability, a 3D specific screening procedure based on a bit-string representation of interatomic distances[17,18,33,34] could provide a more effective approach by not imposing any restrictions with regard to 2D similarity. However, even the current version of the program is able to identify bioactive molecules that are not retrieved by established, 2D measures of structural similarity; it hence provides an effective extension to current approaches to virtual screening.

Aside from the demonstrated utility in virtual screening, the proposed approach may also be potentially useful in other applications such as scaffold or lead-hopping[28−30] where the objective is to discover new chemical entities whose structural templates are distinct from known lead compounds from a 2D perspective but still confer a similar biological profile based on the placement of pharmacophore atoms in 3D space. It could also be used effectively to automate the implementation of alignment techniques such as ensemble distance geometry[35] where one attempts to align several molecules by enforcing the superposition of key atoms, as the proposed 3D MCS procedure could be used to automatically select the requisite key atoms. Our methods would hence seem to provide a powerful tool for probing the structural relationships between pairs of 3D molecules in a range of applications.

ACKNOWLEDGMENT

We thank the following: Current Drugs Limited for provision of the ID Alert database; Daylight Chemical Information Systems Inc., the Royal Society, Tripos Inc., and the Wolfson Foundation for laboratory, software, and hardware support. We would also like to thank Patric Stenberg for helping with the sign tests of the recall data, Gordon Crippen for his helpful comments regarding the implementation of the distance bounding routines, and the reviewers for their helpful comments. The Krebs Institute for Biomolecular Research is a designated center of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Willett, P.; Barnard, J.; Downs, G. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(2) Sheridan, R. P.; Kearsley, S. K. Why Do We Need So Many Chemical Similarity Search Methods. *Drug Discov. Today* **2002**, *7*, 903−911.
(3) Raymond, J.; Gardiner, E.; Willett, P. RASCAL: Calculation of Graph Similarity Using Maximum Common Edge Subgraphs. *Comput. J.* **2002**, *45*, 631−644.
(4) Raymond, J.; Gardiner, E.; Willett, P. Heuristics for Rapid Similarity Searching of Chemical Graphs Using a Maximum Common Edge Subgraph Algorithm. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 305−316.
(5) Raymond, J.; Willett, P. Effectiveness of Graph-Based and Fingerprint-Based Similarity Measures for Virtual Screening of 2D Chemical Structure Databases. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 59−71.
(6) Willett, P. Structural Similarity Measures for Database Searching. In *Encyclopedia of Computational Chemistry*; Schleyer, P. R., Allinger, N. L., Clark, T., Gasteiger, J., Kollman, P. A., Schaefer, H. F., Schriener, P. R., Ed.; J. Wiley: 1998; pp 2748−2756.
(7) Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.
(8) Good, A.; Kuntz, I. Investigating the Extension of Pairwise Distance Pharmacophore Measures to Triplet-Based Descriptors. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 373−379.
(9) Mason, J., et al. New 4-Point Pharmacophore Method for Molecular Similarity and Diversity Applications: Overview of the Method and Applications, Including a Novel Approach to the Design of Combinatorial Libraries Containing Privileged Substructures. *J. Med. Chem.* **1999**, *42*, 3251−3264.
(10) Brint, A.; Willett, P. Upperbound Procedures for the Identification of Similar Three-Dimensional Chemical Structures. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 311−320.
(11) Moon, J. B.; Howe, W. J. 3D Database Searching and De Novo Construction Methods in Molecular Design. *Tetrahedron Comput. Methodol.* **1990**, *3*, 697−711.
(12) Ho, C. M. W.; Marshall, G. R. FOUNDATION: A Program to Retrieve All Possible Structures Containing a User-Defined Minimum Number of Matching Query Elements from Three-Dimensional Databases. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 3−22.
(13) Diestel, R. *Graph Theory*; Springer-Verlag: 2000.
(14) Raymond, J.; Willett, P. Maximum Common Subgraph Isomorphism Algorithms for the Matching Of Chemical Structures. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 521−533.
(15) Willett, P. Matching of Chemical and Biological Structures Using Subgraph and Maximal Common Subgraph Isomorphism Algorithms. *IMA Math. Appl.* **1999**, *108*, 11−38.
(16) Wallis, W. D.; Shoubridge, P.; Kraetz, M.; Ray, D. Graph Distances Using Graph Union. *Patt. Recog. Lett.* **2001**, *22*, 701−704.
(17) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Use of Smoothed Bounded Distances for Incompletely Specified Query Patterns. *J. Mol. Graph.* **1991**, *9*, 157−160.
(18) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Use of Bounded Distance Matrices for the Representation and Searching of Conformationally Flexible Molecules. *J. Mol. Graph.* **1992**, *10*, 194−204.
(19) Crippen, G.; Havel, T. *Distance Geometry and Molecular Conformation*; Research Studies Press: 1988.
(20) Syslo, M.; Deo, N.; Kowalik, J. Shortest Path Problems. In *Discrete Optimization Algorithms*; Prentice-Hall: 1983; pp 227−253.
(21) Easthope, P. L.; Havel, T. F. Computational Experience with an Algorithm for Tetrangle Inequality Bound Smoothing. *Bull. Math. Biol.* **1989**, *51*, 173−194.
(22) Kumar, N.; Deo, N.; Addanki, R. Empirical Study of an Improved Tetrangle-Inequality Bound-Smoothing Algorithm. *Congr. Numer.* **1996**, *117*, 15−31.
(23) Bessonov, Y. E. Generalized Modular Products and the Structural Similarity of Graphs (in Russian). *Vychisl. Sistemy* **1985**, *23*−*32*, 121.
(24) Pelillo, M.; Siddiqi, K.; Zucker, S. W. Matching Hierarchical Structures Using Association Graphs. *IEEE Trans. Pattern Anal. Mach. Intell.* **1999**, *21*, 1105−1120.
(25) This data set was downloaded from Nanosyn, Mountain View, CA, U.S.A. at URL http://www.nanosyn.com.
(26) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: A Review of Performance Measures. *J. Mol. Graph. Model.* **2000**, *18*, 343−357.
(27) Dong, Q.; Wu, Z. A Linear-Time Algorithm for Solving the Molecular Distance-Geometry Problem with Exact Inter-Atomic Distances. *J. Global Optim.* **2002**, *22*, 365−375.

(28) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 2894−2896.
(29) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1406.
(30) Cramer, R. D.; Jilek, R. J.; Andrews, K. M. dbtop: Topomer Similarity Searching of Conventional Structure Databases. *J. Mol. Graph. Model.* **2002**, *20*, 447−462.
(31) Wang, R.; Wang, S. How Does Consensus Scoring Work for Virtual Library Screening? An Idealized Computer Experiment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1422−1426.
(32) Dammkoehler, R. A.; Karasek, S. F.; Shands, E. F. B.; Marshall, G. R. Constrained Search of Conformational Hyperspace. *J. Comput.-Aided Mol. Des.* **1989**, *3*, 3-21.
(33) Cringean, J.; Pepperrell, C.; Poirette, A.; Willett, P. Selection of Screens for Three-Dimensional Substructure Searching. *Tetrahedron Comput. Methodol.* **1990**, *3*, 37−46.
(34) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D Databases. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.
(35) Sheridan, R. P.; Nilakantan, R.; Dixon, J. S.; Venkataraghavan, R. The Ensemble Approach to Distance Geometry: Application to the Nicotinic Pharmacophore. *J. Med. Chem.* **1986**, *29*, 899−906.

CI034002P