Pharmacophoric pattern matching in files of three-dimensional chemical structures: Implementation of flexible searching

David E. Clark and Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield, UK

Peter W. Kenny

Zeneca Pharmaceuticals, Cheshire, UK

The conformational space of a flexible three-dimensional (3D) molecule can be represented for searching purposes by a smoothed bounded distance matrix. Such matrices provide an effective way of carrying out flexible searching, but search times can be very long when compared with rigid searches that take no account of conformational flexibility. This paper considers two techniques for minimizing the computational requirements of flexible searching. In the first part, we compare four different indices that have been suggested for the quantification of molecular flexibility, and demonstrate that they produce comparable rankings of sets of molecules in order of decreasing flexibility. We then demonstrate that the prioritization of a set of structures in order of increasing flexibility can result in substantial reductions in the time requirements of flexible searching if some fixed number of hit structures is desired. In the second part, we report an analysis of the 3D crystal structures in the Cambridge Structural Database to generate distance ranges that are tighter than those produced by distance geometry. Experiments with a set of six query pharmacophores demonstrate that use of these tightened ranges results in substantial reductions in search times, but that this may also lead to a reduction in the number of hit molecules obtained from the final conformational search.

Keywords: bounds tightening, conformational flexibility, flexible searching, pharmacophoric pattern matching, rigid searching, three-dimensional substructure searching

Address reprint requests to Prof. Willett at Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.
Received 26 January 1993; revised 25 February 1993; accepted 3 March 1993

INTRODUCTION

Techniques for pharmacophoric pattern searching in databases of three-dimensional (3D) chemical structures have become well established in the last few years,[1,2] and are increasingly being used for the identification of novel active compounds.[3,4] Early 3D searching systems represented structures by single conformations, these being generated using programs such as CONCORD[5] that provide a rapid means of creating low-energy conformations for a large fraction of the molecules in a typical corporate database. The use of a single-conformation search is only appropriate for the representation of rigid molecules, since such a *rigid search* may fail to identify flexible structures that are able to adopt a conformation that contains the query pharmacophore (as is clearly demonstrated[6] in recent work by Haraki et al.). This is an inherent limitation of rigid-searching systems: the problem can be alleviated, but not eliminated, by the selection of a range of different conformations for each rigid molecule[7] or by careful processing of the query.[8]

A general solution to the problem of conformational flexibility requires an explicit representation of the full range of conformations that a flexible molecule can assume, together with search algorithms that are appropriate for the chosen representation. We have recently suggested[9] that flexible molecules can be represented using graph-theoretic methods analogous to those that are used in two-dimensional (2D) and rigid 3D substructure searching systems.[10] In the case of a 2D molecule, the edges are integers that represent the orders of the bonds linking pairs of atoms, while the edges in a rigid 3D molecule are real values that represent the interatomic distances in Å. For a flexible 3D molecule, we suggest that each edge in a molecule should contain two real values, these being the lower and upper bounds to the interatomic distance that are obtained when a distance geometry,

triangle inequality bound-smoothing procedure is carried out on the molecule.[11] This bounded-distance matrix representation of a conformationally flexible molecule can be searched using a three-stage *flexible searching* algorithm. The first stage is the *screen* or *key search,* which is based on a set of interatomic distance screens[12] that are generated from the bounded-distance matrix. The screen search rapidly eliminates large numbers of structures that cannot possibly match the query. It is followed by the *geometric search,* which uses a subgraph isomorphism algorithm[13,14] to determine the presence or absence of the subgraph representing a query pharmacophore in the graph representing the flexible molecule. This search thus yields a list of potential hits, subject to the approximations inherent in the bounded-distance matrix representation of conformational space. If a high-throughput screen is available,[15] then this hit list may not require further refinement; alternatively, the final, time-consuming *conformational search* examines the candidates from the geometric search for conformations that match the query and that are of a reasonable energy. Our initial experiments demonstrated the effectiveness of this algorithm for flexible searching.[9] However, an analysis of it revealed that flexible searching would be far slower than rigid searching: in this paper, we describe two techniques we have studied to increase the efficiency of flexible searching.

QUANTIFYING MOLECULAR FLEXIBILITY

Our previous paper[9] showed that the final conformational search will dominate the computational requirements of a flexible search (unless the query can be defined in sufficiently restrictive terms to ensure that all but a very small fraction of the search file is removed by the screening and geometric searching stages). Thus, any reductions in the time taken to effect this final procedure will result in substantial decreases in the overall time required to search a 3D database. In this section, we shall examine how a simple index of molecular flexibility can help towards this end.

From the viewpoint of 3D database searching, there are at least two reasons why the development of a simple, yet chemically intuitive, measure of molecular flexibility would be of utility. Firstly, if one is seeking to generate a set of representative conformations for a given structure (as occurs with some of the available 3D substructure searching systems), it would seem sensible to increase the number of conformations selected in proportion to the flexibility of the structure. Secondly, and more importantly for our work which has eschewed such an approach, an index of flexibility can be employed as a screening parameter in a flexible search. Highly flexible molecules bind less strongly at a receptor site, and will also be very expensive in the conformational search.[16] An index of flexibility could be used to rank the output from the geometric search so that the most rigid molecules are processed first by the conformational search routine, which could then continue until the requisite number of hits had been obtained (thus ensuring that the minimum possible number of highly flexible structures are considered). Such a procedure can, of course, be of little help if the search file contains large numbers of highly flexible molecules or if complete recall is required from the search.

Flexibility indices

We have compared four approaches to the quantification of molecular flexibility, as detailed below.

The simplest, but yet intuitively reasonable, index of flexibility is obtained by counting the number of rotatable bonds in a structure. This number, N_{Rot}, is the first of our four measures. A referee noted that this index suffers from the fact that it considers only the number of rotatable bonds, without any consideration of the extent to which their precise location affects the flexibility. For example, a single rotatable bond near the center of a molecule will affect a greater number of the interatomic distances than one located near to the edge of the molecule. However, we believe that the ease of calculation and ready comprehensibility of this measure makes it worthy of consideration.

A more general method involves the bounded-distance matrices that are used for the representation of flexible molecules. A flexibility index can be obtained by a direct comparison of the sets of lower and upper bounds, with a high (low) similarity implying structural rigidity (flexibility). There are several means of calculating measures of similarity from such matrices; we have chosen to use the difference distance matrix *DDM* described by Dean.[17] Given an N-atom molecule, the sets of lower and upper bounds may be represented by $N \times N$ distance matrices, L and U, respectively. The elements of *DDM* are given by the modulus of the difference between the corresponding elements in L and U, i.e.,

$$DDM_{ij} = |u_{ij} - l_{ij}|$$

and the statistical difference s between the two matrices is then given by

$$s = \frac{1}{N}\sqrt{\sum_{i}^{N-1}\sum_{j=i+1}^{N} DDM_{ij}^2}$$

Kier has developed a series of *kappa indices,* which quantify the factors that can lessen the flexibility of a molecule, e.g., the presence of rings or chain branching.[18–21] For the full derivation of these indices, the reader is referred to the cited references, but a simple outline will be given in what follows. The first-order kappa shape index ${}^1\kappa$ encodes the number of atoms in a molecule and the relative cyclicity. It is calculated from the number of atoms N in the hydrogen-suppressed graph of the molecule in question and the number of paths of length one, ${}^1p$, i.e., the number of bonds. Specifically,

$${}^1\kappa = \frac{N(N-1)^2}{({}^1p)^2}$$

The second-order kappa index ${}^2\kappa$ encodes the branching and spatial density of a molecule. It is calculated from the number of atoms and a count of the number of paths of length two in the molecule, ${}^2p$. Specifically,

$${}^2\kappa = \frac{(N-1)(N-2)^2}{({}^2p)^2}$$

The last structural attribute that affects flexibility, as defined by Kier, is the presence of atoms other than sp^3 carbons. These are accounted for by means of a weighting factor α,

which is calculated from the covalent radii of the atoms relative to a reference sp^3 carbon. Thus, an atom i of type x makes a contribution of

$$\alpha_{xi} = \frac{r_x}{r_{C\,sp^3}} - 1$$

and the overall weighting α is then given by

$$\alpha = \sum_{i=1}^{N} \alpha_{xi}$$

Including this weighting factor in the equations for the first- and second-order kappa indices gives

$${}^1\kappa_a = \frac{(N + \alpha)(N + \alpha - 1)^2}{({}^1p + \alpha)^2}$$

and

$${}^2\kappa_\alpha = \frac{(N + \alpha - 1)(N + \alpha - 2)^2}{({}^2p + \alpha)^2}$$

and the overall flexibility index Φ is then given by

$$\Phi = \frac{1}{N} {}^1\kappa_\alpha \times {}^2\kappa_\alpha$$

Fisanick et al.[16] have recently introduced a flexibility index similar to that of Kier. They consider the four following structural attributes to be of importance in determining molecular flexibility: the type of bonds present (acyclic single bonds are generally considered to be flexible while other types are less so); the extent of branching at atoms (rotation around single bonds is sterically hindered if the atoms are highly substituted); the nature of the atoms (as in Kier's work, the presence of atoms other than sp^3 carbon is considered to affect flexibility by altering C-*X* bond lengths) and the location of rigid bonds (unlike Kier, Fisanick et al. consider the position of the structural attributes, and not just their presence, to be an important factor). The *global* flexibility index *GS* for a structure is obtained by calculating the mean of all the *local* flexibility indices LS_{ij} where i and j are a given pair of atoms. The *LS* index is given by

$$LS_{ij} = SPN - \frac{1}{2}(NRB + 0.75BX + 0.5BY)$$

where *SPN* is the number of nodes in the shortest topological path between the pair of atoms under consideration, *NRB* is the number of rigid bonds in the path (all alicyclic, aromatic and multiple bonds are considered to be inflexible), *BX* is the number of atoms with four nonhydrogen atoms attached, and *BY* is the number of atoms with three nonhydrogen atoms attached. When more than one possible shortest path exists between two atoms, average values are taken over all of the paths for these quantities (at present, our implementation of this index considers only the first-found shortest path). For any given structure with N nonhydrogen atoms, there will be $N(N - 1)/2$ such local indices and the mean of these gives the global flexibility index for the structure, i.e.,

$$GS = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} LS_{ij}$$

Results

The four flexibility indices were calculated for the 20 heterogeneous structures shown in Figure 1, and the resulting sets of values are listed in Table 1. It should be noted that the N_{Rot} values in this table (and the values used elsewhere in this paper) are calculated using the SEARCH SETUP functionality within the SYBYL molecular modeling package.[22] Thus, all nonterminal, acyclic single and double bonds are defined as rotatable. No rings were defined as being conformationally active.

An inspection of the four sets of values in Table 1 inevitably reveals some degree of variation. For example, structures 563 and 827 have equal values for N_{Rot}, implying that they are of equal flexibility, the *s* and *GS* values imply that 563 is the more flexible, while the Φ value implies that 827 is the more flexible. However, the overall impression from the figures in the table is that the various indices lead to comparable rankings of the molecules in order of decreasing flexibility. This is confirmed by a statistical analysis of the rankings using the Kendall coefficient of concordance W which measures the extent to which k rankings of the same set of M objects are in agreement with each other[23] and which shows that the indices produce very similar rankings.

In the present context, M is 20, the number of molecules in the ranking, and k is 4, the number of indices that are being used to rank the molecules. The Kendall coefficient can be used to test the null hypothesis H_0 that there is no significant difference in the rankings of the molecules in order of decreasing flexibility that are produced by the four flexibility indices. To compute W the calculated index values are first used to rank the 20 molecules; these rankings then yield the sum of the ranks R_{ij} in each column of the overall $k \times M$ table of results. The R_{ij} values are summed and then divided by M to obtain the mean value of the R_{ij}: each of the R_{ij} is then expressed as a deviation from the mean value. The sum of squares of the deviations S is calculated from

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

and W is then calculated as

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

The significance of the calculated value for W can be established using the χ^2 test, since

$$\chi^2 = k(M - 1)\,W$$

(for $M > 7$) with $M - 1$ degrees of freedom.

In this instance, S was found to be 9994.95; thus, with $k = 4$ and $M = 20$, W is calculated to be 0.94 and χ^2 is hence 71.44, which has an associated probability of ≤ 0.001 under the null hypothesis that there is no significant measure of agreement between the four rankings. We hence conclude that there is a very strong probability that the rankings are in agreement, and thus that the four indices studied here yield analogous rankings of a set of molecules. Accordingly, any of them may be used to rank a set of molecules in order of decreasing flexibility; given its simplicity, the number of rotatable bonds N_{Rot} is clearly the index of choice, and we

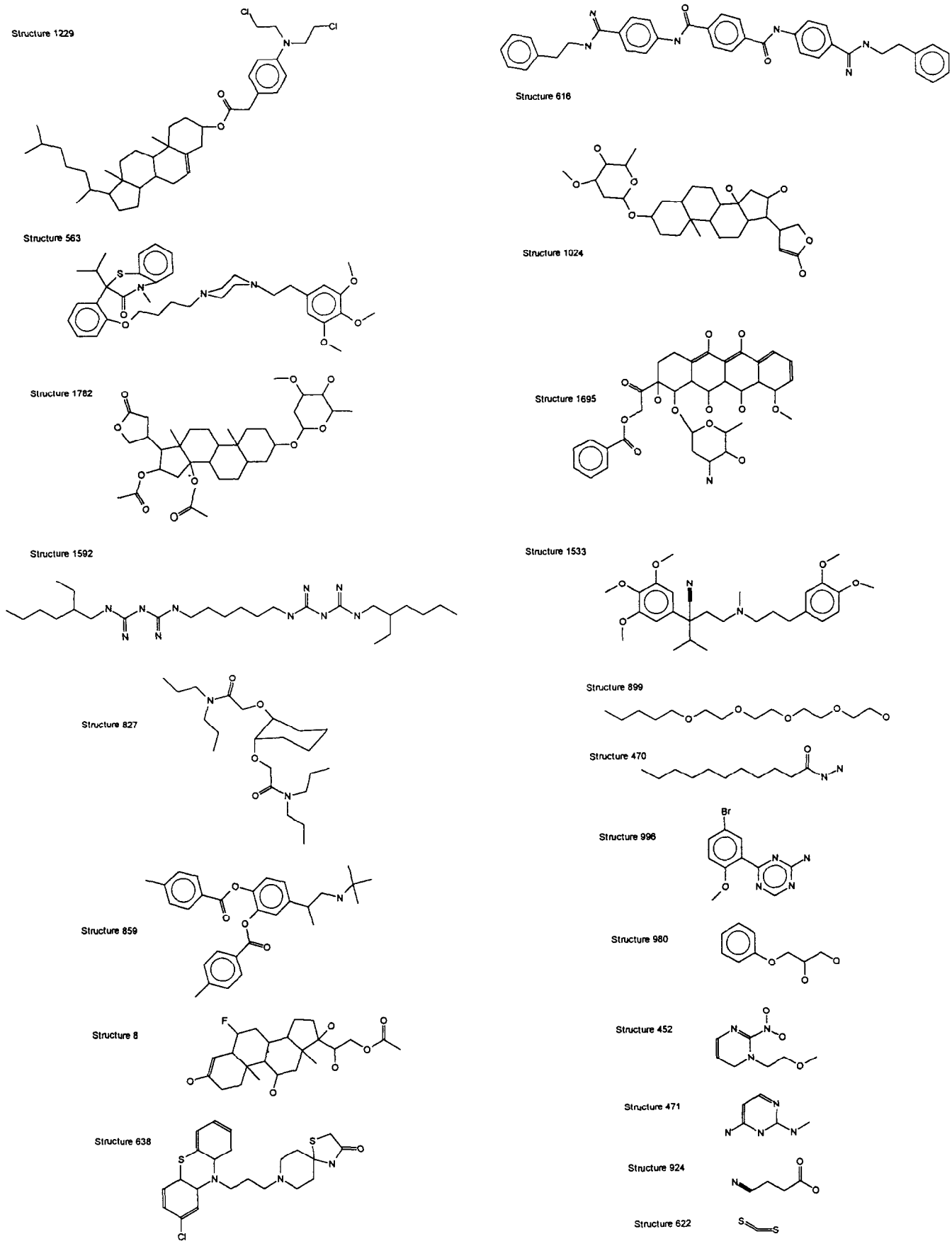

Figure 1. Heterogeneous molecules used in the flexibility index experiments.

Table 1. Values of four flexibility indices for the set of 20 database structures in Figure 1. *N* is the number of atoms, N_{Rot} the number of rotatable bonds, *s* the statistical difference between two DDMs, Φ Kier's flexibility index and *GS* the global index of Fisanick *et al.*

Structure	*N*	N_{Rot}	*s*	Φ	*GS*
1229	103	19	11.65	11.68	6.06
563	95	20	11.83	12.07	7.34
1782	94	15	8.50	8.87	4.27
1592	92	35	14.43	22.78	10.26
616	88	20	13.83	12.39	8.19
1024	84	11	8.02	6.98	3.95
1695	80	15	8.24	8.78	4.57
1533	75	23	8.60	11.32	6.21
827	70	20	8.36	13.58	6.03
859	65	16	8.06	8.12	5.24
8	59	9	5.31	4.95	3.13
638	53	4	5.98	5.84	4.26
899	46	17	8.09	16.61	7.33
470	38	12	5.54	10.36	5.59
996	27	5	3.19	3.43	2.54
980	24	6	3.14	3.73	3.51
452	21	5	2.59	3.03	3.20
471	17	3	2.01	1.72	2.38
924	12	4	1.26	2.90	3.01
622	3	0	0.02	2.38	2.33

have thus used this for the ranking experiments reported below.

Ranking of geometric search output

Assume that the geometric searching stage of a flexible search has been completed and the search system has a subset of the dataset that must now undergo the computationally demanding conformational search. Assume further that the user wishes to see some number *X* of the final hits, i.e., structures that match the query pharmacophore in the conformational search, and that this number is expected to be less than the total number of hits *Y* that would be obtained if all of the *Z* molecules in the geometric search output were to undergo the conformational search. We have shown previously that both *Y* and *Z* can be much larger in a flexible search than in a rigid search;[9] here, we describe a way of minimizing the computational resources that are required to identify *X* hits.

It is intuitively reasonable to suppose, other things being equal, that the more flexible a molecule is, the greater the probability that it will be shown to contain the query pharmacophore during the conformational search, and the greater the amount of time that this conformational search will take. Accordingly, the time taken to identify *X* hit structures will be minimized if the conformational search is carried out in order of increasing molecular flexibility, so that the highly flexible, and very time-consuming structures will be considered if, and only if, fewer than *X* hits have been obtained during the processing of the less flexible structures. The validity of this simple idea was established by carrying out two series of conformational searches on the hits from the geometric search. In the first set, structures were submitted to the conformational search in the order that they had been identified in the geometric search; as there was, to our knowledge, no explicit order in the file of 3D flexible molecules used here, the structures were ordered randomly. In the second set, the hits from the geometric search were sorted into increasing order of N_{Rot} (as defined above) before being submitted to the conformational search. We shall refer to these two types of search as the *unordered search* and the *ordered search*, respectively. Let the run times required to identify *X* hit structures in these two types of search be $T(X)_{Unordered}$ and $T(X)_{Ordered}$, respectively. If the two suppositions above are correct, then we would expect that

$$\frac{T(X)_{Unordered} - T(X)_{Ordered}}{T(X)_{Unordered}} \geq 0$$

with the ratio decreasing with increases in *X*, and tending towards zero as *X* tends towards the limiting value of *Y*. In what follows, we shall refer to this ratio as the *reduction ratio* R_T.

Our experiments used the same set of 1538 bounded distance matrices as were employed in our previous studies of flexible searching.[9] Screening and geometric searches were carried out on this file for the six pharmacophoric patterns shown in Figure 2. In our previous study, the conformational search was implemented using distance-geometry embedding; the experiments here used the CSEARCH algorithm embodied within the SYBYL software.[22,24] Both unordered and ordered conformational searches were carried out. The experimental parameters for these runs are detailed in the appendix.

The behavior for an individual query is exemplified by the

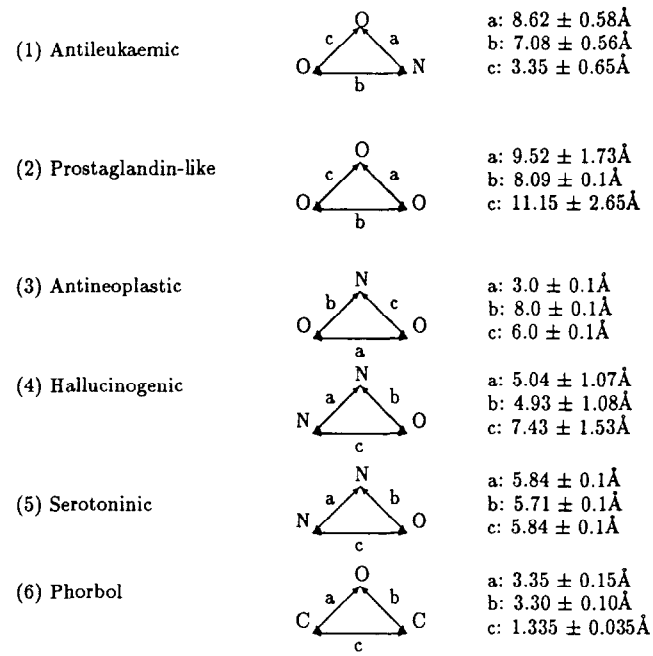

Figure 2. Query pharmacophoric patterns used in the flexible searches.

sixth pattern, which retrieves a total of 14 hits in a systematic search that uses a 30° torsion increment as detailed in Table 2. The R_T values in the righthand column of the table are initially very high, with a reduction of 90% in the time required for the flexible search. This figure decreases, but only slowly, until nearly all of the 14 hits have been identified. The R_T values are lower in some of the other searches; even so, the mean values shown in Table 3 demonstrate the same general trends, showing the mean R_T values when 25%, 50%, 75% and 100% of the total number of hits has been identified. Note that the 50% and 75% values of the ratio for the second pattern are negative: these correspond to cases where $T(X)_{Unordered} < T(X)_{Ordered}$ at the point where 50% and 75% of the hits had been identified. With the exception of these two points, the observed values for the reduction ratio demonstrate clearly that nontrivial improvements in search time can be achieved by this very simple expedient if it is not necessary to retrieve all of the potential matches for a query pharmacophore.

Table 2. Effect of variations in X on the CPU times $T_{Unordered}$ and $T_{Ordered}$ and on the reduction ratio R_T for the 14 hits for query pharmacophore number 6. The CPU times are given as minutes:seconds on an ESV 3

X	$T_{Unordered}$	$T_{Ordered}$	R_T
1	20:55	2:11	0.90
2	25:04	2:30	0.90
3	25:06	2:30	0.90
4	25:07	3:45	0.85
5	31:21	3:52	0.88
6	40:28	7:56	0.80
7	41:54	7:56	0.81
8	41:54	8:31	0.80
9	54:42	9:21	0.83
10	72:18	25:02	0.65
11	80:25	33:09	0.59
12	80:30	33:18	0.59
13	80:31	50:25	0.37
14	80:38	54:26	0.33

Table 3. Values of the reduction ratio R_T when 25%, 50%, 75%, and 100% of the total hits for a systematic conformational search have been identified

	R_T			
Query	25%	50%	75%	100%
1	0.64	0.29	0.20	0.01
2	0.57	−0.14	−0.22	0.08
3	0.85	0.63	0.32	0.36
4	0.91	0.80	0.48	0.07
5	0.60	0.33	0.18	0.07
6	0.88	0.81	0.62	0.33
Mean	0.74	0.45	0.26	0.15

We make two further points to conclude this section of the paper. The first is that the ranking procedure described here could also be used to prioritize the structures that are processed in the geometric search: we have considered only the final, conformational analysis stage since this is the most time-consuming part of a flexible search. The second point is that this approach is completely general in that any type of procedure can be used to order the structures prior to the conformational search: we have considered N_{Rot} but any appropriate measure could be used that was related to the expected computational requirements. For example, one could visualize more complex flexibility indices that described not only the overall flexibility of a structure but also the precise location of the flexible features.

TIGHTENING OF DISTANCE BOUNDS

Estimation of interatomic distance ranges

Conformational searching is extremely time consuming, and it is hence vital that as large a percentage of the database as possible is screened out in the computationally less-demanding screening and geometric stages of a flexible search.[9] These stages involve matching operations that are based on bounded-distance matrices and the overall speed of a flexible search will thus be maximized if we can characterize the distance ranges within a structure as accurately as possible. Reliable techniques for estimating the minimum and maximum interatomic distances are thus of crucial importance in the development of database systems for flexible searching.

We have suggested previously that a significant decrease in the output of the geometric search could be obtained if the distance bounds obtained from the triangle inequality bound-smoothing process were to be further "tightened," so as better to represent the true conformational space of the molecules that were being searched. The experiments upon which this suggestion was based were highly artificial, and involved simply reducing each of the distance bounds in a molecule by a specified percentage and then using the resulting matrices as a basis for the geometric searches.[9] Clearly, it would be desirable to investigate this result using a more realistic means of tightening the set of distance bounds that represent a structure. One way in which this might be achieved is by use of *tetrangle inequality bound smoothing*, which maintains the mathematical rigor of the distance geometry formalism and which can result in a substantial degree of bounds tightening.[25,26] However, this is achieved only at the expense of increased complexity of implementation and much greater computational expense, and the approach still suffers from the correlation effects and the unrealistically high energies that are associated with structures that are generated by the distance-geometry technique of embedding.

There has recently been interest in the use of interatomic path lengths (i.e., the numbers of bonds separating a pair of atoms) to estimate the corresponding interatomic distances.[16,27] We have chosen to study the work of Fisanick[16] et al. at Chemical Abstracts Service (CAS), who state that the "basic premise is that if one has access to a large enough and/or random database, Nature with its infinite chemical diversity has already performed a reasonable conformational analysis." The possible distance ranges between any given

atom pair should be reasonably represented within a large database of 3D structures, and realistic distance bounds may hence be obtained by means of a statistical analysis of the distribution of distances in a large 3D database. This analysis results in a look-up table, the *IJK*th element of which describes the distance ranges characterizing atoms of type *I* and *J* when separated from each other by a path of length *K* bonds (more complex functions than the simple path length can also be used, e.g., functions that take account of the flexibility of the path). This table permits the rapid construction of a bounded-distance matrix for any structure from a knowledge of its connectivity. Although this approach is very empirical, it has the advantages that the analysis will consider only energetically feasible structures and will take implicit account of correlation effects that are ignored by simple bounds smoothing. Fisanick et al. illustrate the use of their methods with distance ranges that have also been produced by distance geometry and conformational search, but present no detailed experimental studies of the use of the method; in what follows, we report a detailed statistical analysis of the 3D structures in the Cambridge Structural Database, and evaluate the effectiveness of the resulting distance ranges when they are used for flexible searching.

Analysis of the Cambridge Structural Database

The workers at CAS used a file of 3D structures produced by the CONCORD program. In the work reported here, we have carried out an analogous analysis of a large subset of the 3D structures in the Cambridge Structural Database (CSD), which is the main repository of experimentally derived 3D structural information on small organic molecules in the crystalline state.[28] The CSD contains a wide variety of structural types, and it was decided to use just those molecules that were deemed to be "pharmaceutically plausible," *viz* molecules containing the heavy atoms carbon, nitrogen, phosphorus, sulphur and the halogens. Following Allen et al.,[29] the molecules were subject to three further constraints: that the crystallographic *R* factor be ≤ 0.01 (an accuracy at which structures are suitable starting points for molecular modelling studies); that the 3D coordinates be error-free at the 0.02-Å level; and that the structure should contain no reported disorder. Application of these constraints to the January 1991 release of the CSD yielded 25,598 sets of coordinates in the CSD FDAT format;[28] these coordinate sets formed the basis for the four-stage analysis procedures described below.

(1) The fractional crystallographic coordinates were converted to orthogonalized coordinates, and the FDAT connectivity information converted into an adjacency matrix.

(2) The adjacency matrix was submitted to Floyd's shortest-path algorithm[30] to determine the shortest paths between all heavy-atom pairs in the structure.

(3) The coordinates were used to calculate the exact distances between all pairs of atoms in the structure; structures containing counter-ions or solvent molecules were processed in the normal way except that interatomic distances between these moieties and the "main" structure were discarded.

(4) The resulting distances were converted to descriptors that were written to an output file for subsequent processing. A typical descriptor is of the form Br 1 C 1.45, which represents a bromine atom separated by one bond from a carbon atom at a distance of 1.45 Å. The output file contained a total of 7,409,063 descriptors, which were then sorted and cumulated. The mean interatomic distance μ and the associated standard deviation σ were calculated for each of the 1543 unique descriptor types that resulted from this cumulation.

The μ, σ data were then used to generate tight bounds for the set of 1538 molecules from the POMONA 89 database that we have used in our previous studies of flexible searching[9,31] and that have bounded distance matrices produced by Smellie's DG program.[32] The precise form of the bounds that are generated is determined by a parameter n which is discussed further below and which is specified by the user at the start of a flexible search. Each atom pair in a molecule in turn is used to generate an atom pair/path length descriptor, and this descriptor is then used to access the sets of stored μ and σ values. If there are no values present for the descriptor under consideration, e.g., if the two atoms are separated by a number of bonds that had not been encountered in our subset of the CSD, the distance geometry lower and upper bounds for that atom pair are retained. If, as is generally the case, values are available and if the distance range given by $\mu \pm n\sigma$ is wholly contained within the original distance bounds (from the DG program), then the original lower bound and upper bounds are replaced by $\mu - n\sigma$ and $\mu + n\sigma$, respectively. The user-defined parameter n thus enables the searcher to control the width of the interatomic distance ranges, and thence the number of hits that is produced in a flexible search. Once all of the atom pairs have been processed in this way, the new bounded-distance matrix is resmoothed to eliminate geometric inconsistencies.

The effect of the suggested procedure is illustrated by reference to the molecule shown in Figure 3. Following the work of Fisanick et al.,[16] distance bounds for the structure were obtained by three methods. The first set of distance bounds were produced by a systematic search of conformational space using the CSEARCH algorithm of the SYBYL molecular modeling software.[22] The search increment was set at 30° and all of the rotatable bonds were defined as active. Using the distance_map option within the SEARCH SETUP function, it is possible to obtain the minimum and maximum separations attained by all atom pairs in a structure during the conformational search. The search over the four rotatable bonds yielded 9132 conformations in a time of 2.19 CPU seconds (on an ESV 3). Bounds were also generated using triangle inequality bounds smoothing, as implemented in Smellie's DG program suite,[32] and using the database analytic procedure described above with $n = 0.05$, 1.0, 1.5, and 2.0 (from the normal distribution these values

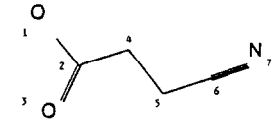

Figure 3. Sample molecule used to illustrate the bounds-tightening procedure.

of $\pm n$ are expected to encompass 38.3%, 64.2%, 86.6%, and 95.4%, respectively, of all of the possible observable distances). Since only some of the interatomic distances in a structure are affected by torsional motion, the results in Table 4 consider only these flexible distances.

An inspection of Table 4 shows that, in all cases, the bounds derived from the SYBYL systematic search lie within those from the distance geometry procedure. This is as would be expected since the SYBYL search takes account of correlation effects and permits no distortions in either bond lengths or bond angles. This finding thus bears out one of the underlying hypotheses of our approach to conformational searching, *viz*, that the conformational space defined by the smoothed bounded-distance matrix for a molecule is a superset of the true conformational space that is available to that molecule.[9] Turning to the bounds derived from the database analytic procedure, the expected trend is observed, in that the bounds tend towards the distance geometry bounds as n is increased. With $n = 0.5$, all of the bounds lie within the distance geometry bounds, but by $n = 1.5$, all but one have assumed the default distance geometry value. This effect was also seen in the more extended experiments discussed below.

The effectiveness of the bound tightening procedure was evaluated by the extent to which the distance bounds were reduced, when compared with the original bounds from the DG program. Assume that the original lower and upper bounds for some atom pair in a molecule are L_{orig} and U_{orig}, respectively, and that the corresponding tightened bounds are L_{new} and U_{new}, respectively. The original and new distance ranges are $U_{orig} - L_{orig}$ and $U_{new} - L_{new}$, respectively, and the reduction in the distance range R_R for this atom pair is hence given by

$$\frac{(U_{orig} - L_{orig}) - (U_{new} - L_{new})}{(U_{orig} - L_{orig})}$$

The mean reduction was calculated by averaging over all distinct atom pairs in a molecule, and then by averaging over all of the molecules in the dataset.

Bound tightening was carried out for $0.50 \leq n \leq 2.00$ in steps of 0.50, so that four sets of new bounded distance matrices were produced. The mean reductions, when averaged over the 1538 structures in our sample database, are listed in Table 5, where it will be seen that substantial reductions in the bounded distances are obtained only with small values of n, i.e., if a very precise search is specified. However, as discussed in the next section, even quite small reductions in the bounded distances result in large reductions in the numbers of molecules matching a query pharmacophore (and hence large increases in the overall speed of searching). It should be noted that in a few cases, particularly with $n \leq 1.0$ and highly flexible molecules, the imposition of new distance ranges upon the original matrix caused erroneous distance bounds, i.e., $U_{new} < L_{new}$ to be generated. This appeared to be caused by triples of atoms consisting of a bonded pair and a single atom separated by a large distance (greater than 20 bonds). In general, the longer the bond path between an atom pair, the less likely it is to occur in the CSD and so the distribution is highly nonnormal. This means that the distance bounds for the atom pair are likely to be unrealistic and this fact, when combined with the tight bounds for the bonded pair of atoms, results in the erroneous distance bounds mentioned above. Structures wherein this was the case were removed from further consideration, and are thus not included in the figures in Table 5.

Flexible searches with new distance bounds

Given an input set of bounded distance matrices, the procedure described above can be used to generate tighter bounds for use in both the screen and the geometric search stages of the proposed three-stage, flexible searching algorithm. However, we believe that the procedure is more appropriate for improving the performance of the geometric search than

Table 5. Reduction in the distance bounds consequent on the choice of the value of n

n	Reduction in distance bounds
2.00	0.3
1.50	3.9
1.00	11.7
0.50	22.9

Table 4. Distance ranges for the molecule shown in Figure 3 calculated using the SYBYL systematic search procedure, using the DG program, and using the database analytic procedure described in Section 3 with n set to 0.5, 1.0, 1.5 and 2.0. The atom pair entries denote the pairs of heavy atoms for the molecule in Figure 3 for which the interatomic distances are being calculated

Atom pair	SYBYL	DG	$n = 0.5$	$n = 1.0$	$n = 1.5$	$n = 2.0$
1,5	2.69–3.71	2.68–3.72	3.06–3.46	2.86–3.66	2.68–3.72	2.68–3.72
1,6	3.03–4.87	3.00–4.90	3.87–4.44	3.58–4.72	3.00–4.90	3.00–4.90
1,7	3.38–5.89	2.90–5.91	4.17–5.04	3.73–5.47	3.30–5.91	2.90–5.91
2,6	2.88–3.77	2.51–3.78	2.88–3.78	2.63–3.78	2.51–3.78	2.51–3.78
2,7	3.61–4.86	3.10–4.95	4.05–4.55	3.80–4.80	3.10–4.95	3.10–4.95
3,5	2.66–3.63	2.65–3.64	3.06–3.46	2.65–3.64	2.65–3.64	2.65–3.64
3,6	3.00–4.79	3.00–4.83	3.87–4.44	3.58–4.72	3.00–4.83	3.00–4.83
3,7	3.38–5.81	2.90–5.83	4.17–5.04	3.73–5.47	2.90–5.83	2.90–5.83

of the screen search. There are two reasons for this belief. First, we have suggested previously that bound tightening has less effect on the performance of the screen search than it does on the performance of the geometric search (although, as noted above, our earlier experiments were far cruder than the method studied here).[9] Secondly, the implementation of bound tightening prior to the screen search involves the rebuilding of the screen file each time that a previously unused value of n is specified by a searcher. Even if only a limited number of values for n was to be allowed, a complete screen file would be required for each such value, which implies very substantial storage costs. For these reasons, we have chosen to apply bound tightening to just those molecules that match a query pharmacophore in the screen search (and that are thus candidates for the geometric search). Tightening does, of course, involve some additional computation prior to the geometric search; however, it is fast in operation (requiring simple table look-up operations followed by resmoothing) and results in substantial reductions (as discussed in detail below) in the numbers of molecules that must undergo the final conformational search, which is by far the most time-consuming part of a flexible search. Our searching experiments have hence focused on the extent to which the reduction in bounds is reflected in a reduction in the number of molecules that match a query pharmacophore in the geometric and conformational searches.

The searches used the six query pharmacophores detailed in Figure 2. The effectiveness of the bound-tightening procedure was measured by the reduction in the number of structures matching the query in the geometric and conformational searches, when compared with the number matching the query when the original bounded matrices were used. The reductions obtained with each value of n are listed in the main body of Table 6, where the first and second entries in each column of the main body of the table are the fractional reduction in the number of matching structures in the geometric search and in the conformational search, respectively. Considering the geometric search figures first, it will be seen that, as expected, the reduction in the number of hits increases as n is decreased, and more importantly, that this reduction is much greater than the corresponding reduction in the distance bounds. Thus, even a small reduction in the distance ranges used to characterize a flexible molecule can bring about substantial increases in search performance. It is of interest to compare the results obtained here with those obtained from the simulated bound tightening procedure described in our previous paper,[9] which suggested a much smaller improvement in search performance than is demonstrated by the results in Table 6.

Although the magnitude of the reduction in the number of hits from the geometric search is of interest, the usefulness of the reduction cannot really be gauged without performing a conformational search on the remaining hits. It is quite conceivable, albeit most undesirable, that the structures removed by bound tightening are, in fact, those which yield the hits at the final stage of a flexible search. For the bounds-tightening technique to be validated, therefore, the percentage reduction in the number of hits from the conformational search should be equal to or (ideally) less than the percentage reduction in the number of hits from the geometric search.

Table 6. Reduction in the numbers of matching structures in flexible searches consequent on the choice of the value of n. The first and second entries in each column of the main body of the table are the reductions in the numbers of matching structures in the geometric search and in the conformational search, respectively

	n							
Query	0.5		1.0		1.5		2.0	
1	0.48	0.31	0.18	0.03	0.06	0.03	0.00	0.03
2	0.53	0.00	0.08	0.18	0.00	0.00	0.00	0.00
3	0.73	0.00	0.38	0.25	0.13	0.25	0.02	0.00
4	0.21	0.22	0.07	0.22	0.00	0.13	0.00	0.09
5	0.80	1.00	0.45	1.00	0.03	0.00	0.00	0.00
6	0.96	1.00	0.77	1.00	0.43	1.00	0.05	1.00*

An inspection of the second entries in each column of the main body of Table 6 reveals that the results are somewhat erratic with no real trends observable. The technique appears to be most successful for the first three queries; indeed, the $n = 0.5$ results for queries 2 and 3 show that it is possible to reduce the number of hits in the geometric search (and hence the number going on to the conformational search) by 53% and 73%, respectively. Thus, one could approximately double or treble, respectively, the search speeds for these two queries with no effect at all on the final search output. The situation is very different for the last three queries, where the conformational search often results in a proportionally smaller number of hits than would have been expected from the reductions in the geometric search output. There seems to be no obvious reason for the observed behavior: this inability to predict *a priori* how well the technique will work for a given query and a given value of n would seem to cast doubt on the usefulness of this bound-tightening process for expediting flexible searches. One interesting point to emerge from the experiments is that the changes to the structure's distance bounds can cause structures to become hits which were not previously. This is illustrated by the final query with $n = 2.0$, where the number of hits in the conformational search actually doubles (as represented by the starred value in the table).

CONCLUSIONS

In this paper, we have discussed two techniques for maximizing the efficiency of flexible searching in databases of 3D structures.

Measures of molecular flexibility have been reviewed and compared. Little difference was found between those examined for the purpose of ranking structures prior to a conformational search. This suggests that the simplest of the measures tested, *viz*, the number of rotatable bonds, will perform as well as any of the more complex procedures that we have evaluated, and we have used this measure to rank the outputs of several pharmacophoric pattern searches. It is shown that substantial reductions in search time can be achieved by output ranking if one requires only a sample of the hits from a flexible search.

We then described a database analytic procedure that can result in effective reductions in the distance bounds using data from a statistical analysis of a large chemical structure database. Since the magnitude of these reductions is dependent on the value of a single, user-defined parameter, the searcher is given an element of control over the volume of output from the geometric search. In flexible searching, where many potential hits are often found, this degree of control could be extremely valuable. In some cases, it is possible to achieve substantial reductions in the numbers of hits in the geometric search (and hence in the numbers of molecules undergoing the conformational search), without any effect on the final number of hits. In other cases, however, the reduction in the hits from the geometric search has a deleterious effect on the final number of hits obtained from the conformational search. Thus, the increase in the efficiency of flexible searching that results from bounds-tightening is sometimes at the expense of retrieval effectiveness.

ACKNOWLEDGMENTS

We thank Zeneca Pharmaceuticals and the Science and Engineering Research Council for the award of a CASE studentship to DEC, and TRIPOS Associates Inc. for the provision of hardware and software. We are grateful to Tad Hurst for useful discussions on the subject of conformational flexibility, to Yvonne Martin for suggesting the searching experiments, to Bill Fisanick for advice on the implementation of his flexibility index, to Scott Rowland and Peter Bath for assistance with the Cambridge Structural Database, and to the referees for helpful comments on an earlier draft of this paper. This paper is a contribution from the Krebs Institute for Biomolecular Research, which is a designated Centre for Molecular Recognition Studies under the Molecular Recognition Initiative of the Science and Engineering Research Council.

REFERENCES

1 Martin, Y.C., Bures, M.G., and Willett, P. Searching databases of three-dimensional structures. In: *Reviews in Computational Chemistry*. (K.B. Lipkowitz and D.B. Boyd, Eds) VCH, New York, 1990, pp. 213–263
2 Willett, P. *Three-Dimensional Chemical Structure Handling*. Research Studies Press, Taunton, 1991
3 Bures, M.G., Black-Schaefer, C., and Gardner, G. The discovery of novel auxin transport inhibitors by molecular modeling and three-dimensional pattern analysis. *J. Comp.-Aided Mol. Design*. 1991, **5**, 323–334
4 Martin, Y.C. 3D database searching in drug design. *J. Med. Chem*. 1992, **35**, 2145–2154
5 Rusinko III, A., Skell, J.M., Balducci, R., McGarity, C.M., and Pearlman, R.S. *CONCORD: a program for the rapid generation of high quality approximate three-dimensional molecular structures*. The University of Texas at Austin and Tripos Associates, St. Louis, Missouri (1988)
6 Haraki, K.S., Sheridan, R.P., Venkataraghavan, R., Dunn, D.A., and McCulloch, D. Looking for pharmacophores in 3D databases: does conformational searching improve the yield of actives? *Tetrahedron Comput. Methodol*. 1990, **3**, 565–573
7 Murrall, N.W. and Davies, E.K. Conformational freedom in 3D databases. 1. Techniques. *J. Chem. Inform. Comput. Sci*. 1990, **30**, 312–316
8 Güner, O.F., Henry, D.R., and Pearlman, R.S. Use of flexible queries for searching conformationally flexible molecules in databases of three-dimensional structures. *J. Chem. Inform. Comput. Sci*. 1992, **32**, 101–109
9 Clark, D.E., Willett, P., and 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. Graphics*. 1992, **10**, 194–204
10 Ash, J.E., Warr, W.A., and Willett, P. (Eds.). *Chemical Structure Systems*. Ellis Horwood, Chichester (1991)
11 Havel, T.F., Kuntz, I.D., and Crippen, G.M. The theory and practice of distance geometry. *Bull. Math. Biol*. 1983, **45**, 665–720
12 Jakes, S.E. and Willett, P. Pharmacophoric pattern matching in files of three-dimensional chemical structures: selection of interatomic distance screens. *J. Mol. Graphics*. 1986, **4**, 12–20
13 Brint, A.T. and Willett, P. Pharmacophoric pattern matching in files of three-dimensional chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics*. 1987, **5**, 49–56
14 Ullmann, J.R. An algorithm for subgraph isomorphism. *J. Assoc. Comput. Machinery*. 1976, **23**, 31–42
15 Leichtfried, F.E. Novel approaches to high throughput screening automation. In: *Proceedings of the First Forum on Data Management Techniques in Biological Screening*. (C. Carter and K.R. Freter, Eds.) SRI International, Menlo Park CA (1992), pp. 81–86
16 Fisanick, W., Cross, K.P., and Rusinko III, A. Characteristics of computer-generated 3D and related molecular property data for CAS Registry substances. *Tetrahedron Comput. Methodol*. 1990, **3**, 635–652
17 Dean, P.M. Molecular recognition. In: *Concepts and Applications of Molecular Similarity*. (M.A. Johnson and G.M. Maggiora, Eds.) Wiley-Interscience, New York (1990), pp. 211–238
18 Kier, L.B. An index of molecular flexibility from kappa shape attributes. *Quant. Struct.-Act. Relat*. 1989, **8**, 218–221
19 Kier, L.B. A shape index from molecular graphs. *Quant. Struct.-Act. Relat*. 1985, **4**, 109–116
20 Kier, L.B. Shape indexes of order one and three from molecular graphs. *Quant. Struct.-Act. Relat*. 1986, **5**, 1–7
21 Kier, L.B. Distinguishing atom differences in a molecular graph shape index. *Quant. Struct.-Act. Relat*. 1986, **5**, 7–12
22 SYBYL. Tripos Associates Inc., St. Louis, Missouri, USA
23 Siegel, S. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Kogakusha, Tokyo (1956)
24 Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B.,

and Marshall, G.R. Constrained search of conformational hyperspace. *J. Comp.-Aided Mol. Design* 1989 **3,** 3–21
25 Easthope, P.L. and Havel, T.F. Computational experience with an algorithm for tetrangle inequality bound smoothing. *Bull. Math. Biol.* 1989, **51,** 173–194
26 Havel, T.F. An evaluation of the computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. *Prog. Biophys. Mol. Biol.* 1991, **56,** 43–78
27 Bradshaw, J. and Maliski, E.G. Use of most restrictive paths in 3D search strategy. Paper presented at the 4th Chemical Congress of North America, New York, 25–30 August 1991
28 Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M., and Watson, D.G. The development of Version 3 and Version 4 of the Cambridge Structural Database System. *J. Chem. Inform. Comput. Sci.* 1991, **31,** 187–204
29 Allen, F.H., Doyle, M.J., and Taylor, R. Automated conformational analysis from crystallographic data. 1. A symmetry-modified single-linkage clustering algorithm for three-dimensional pattern recognition. *Acta Crystallogr. B* 1991, **47,** 29–40
30 Floyd, R.W. Algorithm 97: shortest path. *Comm. Assoc. Comput. Machinery.* 1962, **5,** 345
31 Clark, D.E., Willett, P., and 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. Graphics.* 1991, **9,** 157–160
32 Smellie, A.S. *Distance geometry: new methods and applications,* PhD Thesis, University of Oxford, 1989

APPENDIX

The SEARCH SETUP parameters in SYBYL were specified as follows:

(1) Cleaning up CONCORD structures. All of the database structures were subjected to a single energy minimization using the default settings of the MAXIMIN2 facility, which serves to lessen any close nonbonded contacts in the CONCORD structure.
(2) Definition of rotatable bonds. This was accomplished by means of the command ROTATABLE_BOND DEFINE bond_expr. All rotatable bonds were specified by using '*' as the bond_expr option.
(3) Definition of search angle increments. The scan parameter of the conformational search is controlled by the command ANGLES INCREMENT step_size; experiments were carried out with step_size set to 60°, 40°, 30°, 20°, or 10°.
(4) Definition of distance constraints. The geometric requirements of the pharmacophore were imposed upon the candidate structure using the command CONSTRAINING_DISTANCES DEFINE atom1 atom2 min_dist max_dist, where the latter values were the matching ranges produced by the geometric search program.
(5) Energy calculations. The energies of the conformations produced during the search were calculated using the command ENERGY ENERGY max_energy no_electrostatics. No energy limit was imposed in the initial experiments.
(6) All other settings were left as the program defaults, i.e., the reference conformation from which all angle increments are to initiate was "zeroed" (all rotatable bond torsion angles set to 0°), no bump checking was performed, and the VDW scaling factors were as determined by the program.

The ordered and unordered searches for patterns 1 and 6 used a torsion increment of 30°, while the searches for the other four patterns used a torsion increment of 60°. A molecule was assumed not to contain the sought query pharmacophore if the CSEARCH routine had not terminated within a maximum of 240 CPU seconds on an Evans and Sutherland ESV 3 workstation. These parameter values are derived from an ongoing project that is evaluating a range of conformational-search algorithms when used for flexible 3D searching. The results of this comparative study will be reported elsewhere.