Pharmacophoric pattern matching in files of three-dimensional chemical structures: use of bounded distance matrices for the representation and searching of conformationally flexible molecules

David E. Clark and Peter Willett

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

Peter W. Kenny

ICI Pharmaceuticals, Mereside, Alderley Park, Macclesfield, UK

This paper discusses the use of bounded distance matrices for the representation of conformationally flexible three-dimensional (3D) molecules. It is shown that pharmacophoric pattern searches of databases of flexible 3D molecules represented in this way can be carried out using screen and geometric searching algorithms that are analogous to those used for searching databases of rigid 3D structures. Molecules matching a query pattern after the geometric search must then undergo a final conformational search to determine whether they can, in fact, adopt a conformation that matches the query. An analysis of this three-stage searching procedure shows that searching databases of flexible 3D molecules is extremely demanding of computational resources.

Keywords: Bound smoothing, conformational flexibility, conformational search, distance geometry, geometric search, pharmacophoric pattern matching, screen search, bounded distance matrix

INTRODUCTION

The last three decades have seen the development of computer-based systems for the storage and retrieval of information pertaining to chemical molecules.[1,2] Till recently, the main focus of interest in these systems has been the processing of databases of two-dimensional (2D) structures, with much less account being taken of databases of three-dimensional (3D) structures.

The 3D structure is of crucial importance in determining the biological activity of a molecule, and this has led to the development of sophisticated molecular modeling systems, which facilitate the detailed study of molecular conformations.[3] However, the computational requirements of modeling systems mean that they can be used to study only small numbers of molecules that have been identified previously by alternative, less detailed approaches: there is thus clear scope for increasing the power of modeling systems by providing them with searching facilities analogous to those that are already available for the processing of databases of 2D structures. The last few years consequently have seen an explosion of interest in the development of techniques for 3D database searching. These techniques allow the identification of the presence of pharmacophores, or pharmacophoric patterns, in 3D molecules, where a pharmacophoric pattern is the geometrical arrangement of structural features in a drug molecule that is necessary for biological activity at a receptor site.

The basis for much of the subsequent work on 3D database searching was established in the pioneering studies of Gund and his coworkers on the MOLPAT program.[4,5] Conventional 2D chemical database systems represent molecules by connection tables, i.e., labeled graphs in which the nodes and edges of a graph correspond to the atoms and bonds of a chemical structure diagram. Gund et al. recognized that a 3D molecule could be represented by a graph

Address reprint requests to Prof. Willett at the Dept. of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.
Received 19 September 1991; revised 6 December 1991; accepted 10 December 1991

in which the nodes and edges correspond to the atoms and to the interatomic distances, respectively. Given such a representation, a search for a set of atoms and the corresponding interatomic distances can be carried out using a subgraph isomorphism algorithm, in just the same way as a search for a pattern of atoms and bonds in a 2D connection table. The MOLPAT program allowed a chemist to specify a query pattern and a drug molecule, and the program then searched for the occurrence of the pattern in the molecule. The time-consuming subgraph isomorphism search, which is now referred to as *geometric searching* in the 3D context, was preceded by a simple screening step to ensure that the molecule under consideration contained at least the numbers and types of atoms specified in the query pharmacophore.

Substantial increases in the speed of searching may be obtained by the use of efficient screening and geometric searching algorithms. Work in Sheffield has suggested the general applicability of screens that involve interatomic distance ranges, these ranges being identified by a statistical analysis of the frequencies of occurrence of each type of interatomic distance in a sample of the database that is to be searched.[7,8] Analogous approaches can be used to screen searches that contain angular information.[9] Molecules that match the query pharmacophore at the screening level then undergo the time-consuming geometric search.[6,10] The first operational 3D substructure searching system was developed by Jakes et al.[7,10] in collaboration with Pfizer Central Research (UK), where it has since been implemented as part of the company's SOCRATES information system.[11] Several commercial and proprietary systems have been described subsequently.[12–19] Reviews of this rapidly developing field are presented by Martin et al.[20] and by Willett.[21]

A major limitation of current 3D database systems is that they often take little account of the inherent conformational flexibility of molecules, storing only a single low-energy conformation for each chemical compound in a database. However, it has long been realized that biological molecules can exist in a variety of different conformations and can shift among these conformations, depending on energetic and environmental conditions.[22] This fact has considerable implications for 3D searching systems, since, in general, there is no simple relationship between a single conformation of a molecule and the receptor-bound conformation sought when one searches for a pharmacophore.[23] For example, acetylcholine bound to the nicotinic receptor adopts a conformation distinctly different from its conformation in solution or in the crystal state[24] Thus, if pharmacophoric pattern matching is to increase in effectiveness, the abilities both to represent and to search conformationally flexible molecules need to be developed. In this paper, we discuss the use of bounded distance matrices for the representation of conformationally flexible 3D molecules, and then demonstrate how these representations can be searched using screening and geometric searching algorithms analogous to those that are already available for searching rigid 3D molecules.

REPRESENTATION OF FLEXIBLE MOLECULES

Two main approaches have been used to date for the representation of conformationally flexible molecules. The first approach (the one that has been adopted since the earliest days of 3D searching) is to store a set of representative low-energy conformations for each flexible structure in a database. This approach is exemplified by Molecular Design Ltd's MACCS-3D system.[14] The main problems with such an approach are the large amount of storage that is required if many conformations are to be considered and the impossibility of ensuring that all of the important, biologically active conformations are present in the representative sample that is selected. A second, more sophisticated strategy has been adopted in Chemical Design Limited's ChemDBS-3D module.[15,16] Here, a set of representative low-energy conformations for each structure is generated, and screens are assigned to each conformation. The bit strings for the set of conformers are linked by a logical OR function to give a single bit string for each structure to be matched against the query in the screen search. The structures that match at the screening level are then subjected to a second conformational analysis, and a geometric search is carried out to determine which of the resulting conformations match the query. While this method obviates the need for storing a number of discrete conformations, there is still a reliance on the selection of some set of representative, low-energy conformations. Even so, Haraki et al. have demonstrated that this approach can result in the retrieval of substantially larger numbers of active compounds than if only rigid structures are used in a pharmacophoric pattern search.[25] Martin et al. [20] suggest that a precise definition of conformational flexibility might be achieved using techniques derived from distance geometry:[23,26–28] in this paper, we present the results of such an investigation. Specifically, we report the use of bounded distance matrices for the representation and searching of conformationally flexible molecules; a previous paper has discussed the use of such matrices for the representation of flexible query pharmacophores.[29]

We have noted previously that rigid molecules in a 3D database system are represented by graphs in which the edges of a graph denote the set of interatomic distances in a molecule. The screening and geometric searching algorithms that are currently available have been developed to process such representations,[21] and there would be clear advantages if similar methods could be used for flexible molecules. It is relatively simple to create an appropriate representation using distance geometry techniques, as we will now show. The procedure starts with an initial matrix that contains one upper and one lower bound for each interatomic distance, these bounds corresponding to the maximum and minimum separations that are possible for a given pair of atoms. Some of these distance bounds can be set using considerations of molecular connectivity, together with lists of standard bond lengths and bond angles; the remainder are set to predetermined default values, typically the sum of the van der Waals' radii and some arbitrarily large value, e.g., 500 Å, for the lower and upper bounds, respectively.[30,31] Thus, for an N-atom system, the matrix is of dimensions $N \times N$ and is composed of the upper and lower distance bounds (U_{IJ} and L_{IJ}, respectively) for each interatomic distance, separated by a leading diagonal of zeros. Since default values have been set for many of the interatomic distance bounds, it is certain that this initial matrix will contain some geometrical inconsistencies. These may be eliminated by an iterative process known as *triangle* *inequality bound smoothing*.[26,28] The triangle inequality demands that, given the distances D_{IJ}, D_{JK} and D_{IK} between three points, I, J, and K:

$$D_{IJ} - D_{JK} \leq D_{IK} \leq D_{IJ} + D_{JK}$$

and likewise for D_{IJ} and D_{JK}. The bound smoothing algorithm takes each possible set of three atoms in turn and repeatedly applies the triangle inequality to produce the final *bounded distance matrix B*. Since B now contains geometrically consistent upper and lower bounds for all of the interatomic distances in the molecule, the matrix in some way approximates its conformational space. Molecular modeling studies, however, suggest that the true form of such a conformational space is generally more complex than the space bounded by B. The discrepancy arises from the fact that the construction of B is based purely on geometric considerations, and takes no account of energetic considerations or of correlation effects. Therefore, while the space defined by B will contain all the geometrically feasible conformations, not every point in this space will correspond necessarily to a viable conformation.

SEARCHING OF FLEXIBLE MOLECULES

Given a database of bounded distance matrices, a search for a pharmacophoric pattern can be effected using a multistage retrieval algorithm that is a natural extension of the techniques that are used to search a database of rigid 3D structures: in what follows, we shall refer to such a conventional pharmacophoric pattern search as a *rigid search* and a pharmacophoric pattern search of a database of conformationally flexible 3D molecules as a *flexible search*. A rigid search involves an initial screening search, which is based on distance ranges. The molecules that match the query at the screening level then undergo the geometric search, which utilizes a subgraph isomorphism algorithm. Molecules that match the query at the geometric level are then output to the user, e.g., for input to a modeling package. The first two stages are equally appropriate in the context of a flexible search, since it is possible to use a bounded distance matrix for the screen and geometric searches in much the same way as one uses a conventional interatomic distance matrix. However, flexible searching also requires an additional conformational search, as discussed below.

Matching criteria

The matching criteria that are required for a flexible search may best be illustrated by first considering a rigid search. Assume that the distance D_{IJ}, between two atoms I and J in a database structure is to be compared with the distance D_{KL} between two atoms K and L in the query pharmacophore. The condition for a match between these two distances in the geometric search is then

$$D_{IJ} = D_{KL}$$

In fact, query distances are normally expressed as a range $D_{KL} \pm \epsilon$, where ϵ is a tolerance value, e.g., 0.5 Å. The query distance thus can be expressed as a range between a lower bound, L_{KL}, and an upper bound, U_{KL}, so that the matching condition is now

$$L_{KL} \leq D_{IJ} \leq U_{KL}$$

Let $F(D)$ be a function that takes a distance (or distance range) and returns the corresponding bit (or set of bits) from a screenset. Then the condition for a match in the screen search is

$$F(D_{IJ}) \cap F(L_{KL}:U_{KL}) \neq \emptyset$$

where $L_{KL}:U_{KL}$ denotes the distance range between L_{KL} and U_{KL} and where $\emptyset$ denotes a null bit string.

Consider now the case of a flexible search, where the database structure distance is also represented by a range, $L_{IJ}:U_{IJ}$. In this case, the two distance ranges can be considered as being equivalent in the geometric search only if there is a nonzero overlap between them. This condition may be expressed formally as

$$max\{L_{KL}, L_{IJ}\} \leq min\{U_{KL}, U_{IJ}\} \qquad (1)$$

There are four ways in which this inequality can be satisfied:

(i) The structure distance range overlaps the upper end of the query distance range, i.e.,

$$U_{IJ} \geq U_{KL} \geq L_{IJ}$$

(ii) The structure distance range overlaps the lower end of the query distance range, i.e.,

$$U_{IJ} \geq L_{KL} \geq L_{IJ}$$

(iii) The structure distance range subsumes the query distance range, i.e.,

$$U_{IJ} \geq U_{KL} \quad \text{and} \quad L_{IJ} \leq L_{KL}$$

(iv) The query distance range subsumes the structure distance range, i.e.,

$$U_{KL} \geq U_{IJ} \quad \text{and} \quad L_{KL} \leq L_{IJ}$$

These four types of matching criteria are illustrated in Figure 1.

In the case of the screen search, the matching condition is now

$$F(L_{IJ}:U_{IJ}) \cap F(L_{KL}:U_{KL}) \neq \emptyset \qquad (2)$$

Conditions (1) and (2) need to be checked for each and every query distance in the geometric search and in the screen search, respectively, when checking for the presence of a query pharmacophore in a database structure.

Screen search

The screen search requires the characterization of each of the molecules in the database by a bit string. This is created using a generalization of the procedure described by Murrall and Davies,[15] where individual sets of bits are assigned for each of some number of low-energy conformations.

Assume that a set of distance range screens has been obtained and that these are to be assigned to a database structure, just as with a rigid searching system except that here we are seeking to represent a set of interatomic distance bounds rather than a set of interatomic distances. Consider an upper and a lower bound, U_{IJ} and L_{IJ}, in B. The quantities U_{IJ} and L_{IJ} are each compared with the screenset to identify the screens, X and Y, that contain the distances U_{IJ} and L_{IJ},

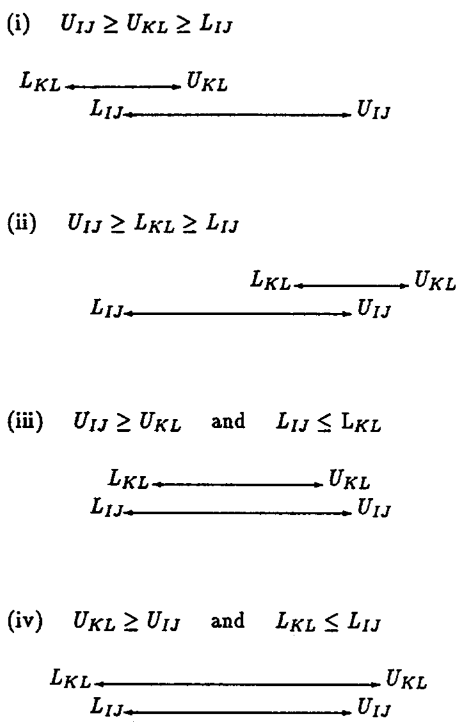
(i) $U_{IJ} \geq U_{KL} \geq L_{IJ}$
L_{KL}
U_{KL}
L_{IJ}
U_{IJ}
(ii) $U_{IJ} \geq L_{KL} \geq L_{IJ}$
L_{KL}
U_{KL}
L_{IJ}
U_{IJ}
(iii) $U_{IJ} \geq U_{KL}$ and $L_{IJ} \leq L_{KL}$
L_{KL}
U_{KL}
L_{IJ}
U_{IJ}
(iv) $U_{KL} \geq U_{IJ}$ and $L_{KL} \leq L_{IJ}$
L_{KL}
U_{KL}
L_{IJ}
U_{IJ}

Figure 1. Possible overlap of a distance range in a query pharmacophore, L_{KL}:U_{KL}, with a bounded distance, L_{IJ}:U_{IJ}, for a database structure.

respectively. The molecule, as represented by B, is then assigned the screens X and Y, as well as all of the screens that lie between X and Y; i.e., the function $F(D)$ yields the bit positions of all screens that cover distances in the range from L_{IJ} to U_{IJ}. This procedure is repeated for each distinct pair of bounded distances in B. Thus, whereas each interatomic distance in a rigid molecule will result in the assignment of a single screen, each pair of bounded distances may result in the assignment of several screens, the precise number being determined by the magnitude of $U_{IJ} - L_{IJ}$ and by the size of the distance ranges denoted by each of the screens.

In this way, bit string representations of database structures and of query pharmacophores are obtained that can be processed using Condition (2), in a manner that is analogous to the screening component of a rigid searching system.

Geometric search

Molecules matching a query pharmacophore in the screening search are then passed on for the geometric search, which utilizes the subgraph isomorphism algorithm due to Ullmann.[32] Work in Sheffield has demonstrated the applicability of this algorithm to the searching of 2D and 3D small molecules and 3-D macromolecules,[6,21] and it is now used in operational 3D searching systems.[13,15]

Ullmann's algorithm operates by means of a backtracking tree search in which database atoms are tentatively assigned to query atoms and the match is extended in a depth-first manner until a complete match is obtained or a mismatch is detected. In the latter case, the search backtracks to the previous assignment and an alternative match is considered. The amount of backtracking is minimized in Ullmann's algorithm by the use of an heuristic, the *refinement procedure*, that limits the number of levels of the search tree that have to be investigated before a mismatch is identified. Specifically, when the algorithm is used for rigid 3D searching, it makes use of the fact that if one query atom, $Q(K)$, is a specific distance from another query atom, $Q(L)$, and if a database atom, $S(J)$, matches $Q(L)$, then there must be database atom, $S(I)$, at the appropriate distance from $S(J)$, that matches $Q(K)$; this condition is checked for all of the neighbors of each query atom. The results of these checks are used to update a matrix, the *M matrix*, which contains all of the possible equivalences between query atoms and database atoms. A subgraph isomorphism has been detected when each of the query atoms has been mapped to a single, distinct database atom; a mismatch is identified if there are no mappings possible for one or more of the query atoms.

In a flexible geometric search, the algorithm must allow the matching of atoms on the basis of distance ranges, rather than on the basis of exact distances. Thus, if query atoms $Q(K)$ and $Q(L)$ are identical to the database atoms $S(I)$ and $S(J)$, respectively, then a match is obtained if there exists a nonzero overlap between the distance ranges characterizing the separations $Q(K) - Q(L)$ and $S(I) - S(J)$ (as discussed above). It is relatively simple to modify the refinement procedure in Ullmann's algorithm to allow the identification of such overlaps, and hence to allow the mapping of query and database atoms via the M matrix.

Conformational search

Once the geometric search has identified the possible subgraph isomorphisms, the final stage of the processing involves a conformational searching routine. It should be noted that it is possible for a structure to pass the geometric search and still not be a true hit for the query. Although some conformation of the structure will match each of the query distance ranges, this does not imply that any single conformation will match all of the ranges.[33] Thus, a conformational search is required to identify specific conformations that match the query, subject to the atomic equivalences and distance constraints resulting from each single isomorphism. There are several techniques available for conformational searching, as discussed in the review by Howard and Kollman.[34] In the work reported here, we have used a distance geometry embedding routine, as embodied in the DGEOM program,[35] which randomly selects a set of coordinates from the space defined by the database structure's distance bounds.[26] A match is obtained if the generated coordinates satisfy the query-to-structure constraints resulting from the geometric search to within the DGEOM default tolerances of ±0.5 Å. Several embeddings may need to be tested before a match is confirmed.

It must be emphasized that this is only one way of processing the output from the geometric search. Other conformational searching procedures additionally allow an energy calculation to be carried out for each of the conformations as they are produced. This information could be used to rank the isomorphisms for a given structure, or the structures themselves, in order of increasing energy; the utility of such a procedure is the subject of current study in our laboratory.

EXPERIMENTAL DETAILS

Datasets

To test the utility of B for the representation of conformational flexibility, a database of flexible structures is required. In the work reported here, we have used a database of rigid molecules as input to a distance geometry algorithm that produces a comparable database of bounded distance matrices. The structures used were a subset of the POMONA89 2D database, as supplied by Daylight Chemical Information Systems:[36] this subset contained no ionic or disconnected structures, nor any molecules that contained less than two heteroatoms and one carbon atom. The selected molecules were converted to 3D atomic coordinates by means of the CONCORD program.[12,37] The resulting sets of coordinates were then submitted to the PREP and SMOOTH modules of Smellie's distance geometry package[30] to produce the appropriate bounded distance matrices. The final search file contained 1538 such matrices.

A set of eight published pharmacophoric patterns was chosen from those used by Jakes et al.:[10] these literature patterns are illustrated in Figure 2. Further "random" query patterns were generated from the database by selecting every hundredth structure and, from each of these, extracting the required number of query atoms. The 3D coordinates of the randomly selected atoms were used to calculate the interatomic distances in the query pattern, to which the desired tolerance was added. In this manner, 15 3-atom and 15 5-atom queries were produced with tolerances of 0.1 Å and 0.5 Å. The maximum numbers of carbon atoms permitted in the sets of 3-atom and 5-atom query patterns were limited to 1 and 2, respectively, this reflecting the fact that pharmacophoric patterns typically contain a preponderance of heteroatoms. Since the interatomic distances in the queries, together with their tolerances, are easily expressed as upper and lower bounded distances, it was simple to assign screens to the queries in a similar manner to that for the database molecules.

Generation of screensets

The 3D coordinates for 200 of the POMONA89 structures were taken as the basis for the generation of the screen sets, which was carried out using the screen set generation algorithm described by Cringean et al.[8] This algorithm involves an analysis of a file of structures, typical of those that are to be screened, to generate all of the (non-hydrogen) interatomic distances in the file. This results in a list of fragments of the form $A_IA_JD_{IJ}$, where A_I and A_J ($A_I \le A_J$) are the elemental types of the pair of atoms that are being considered, and D_{IJ} is the distance between them. This file of interatomic distance descriptors is sorted into increasing alphanumeric order and cumulated, so that each interatomic distance is stored together with its frequency of occurrence. The file is then divided into P partitions, where P is the number of screens required, so that each of the partitions corresponds to one of the screens that are available for assignment to database structures or to query substructures. This subdivision is carried out in such a way that each of the resulting partitions contains approximately the same number of interatomic distance occurrences. The experiments reported below used a screen set containing 1024 screens.

Once the screens had been generated, a binary search algorithm was used to assign the relevant distance range screens to the full set of molecules in the database, using the procedure discussed above. The assignment of a screen to a particular structure was denoted by setting the corresponding bit in a bit map that formed the basis for the screening search: for the 1538 structures and 1024 screens considered here, the bit map was thus of dimensions 1538 × 1024. Screens were assigned to the queries in the same way. The database was searched for molecules matching a given query at the screening level by appropriate combination of logical OR and AND operations on the bit strings representing queries and database structures.

The screensets that form the basis for the results in the following section were generated as described, using 200 of the original interatomic distance matrices as the source of the fragment occurrence data. However, it may be felt that there is a logical inconsistency here, since we have used a set of rigid molecules to obtain frequency statistics for the selection of screens that will subsequently be used to characterize flexible molecules. Accordingly, we also have generated screens that are based on frequency statistics for flexible molecules. Specifically, a set of 1024 screens was generated using 200 of the bounded distance matrices. For a distance range L_{IJ}:U_{IJ}, fragments of the form $A_IA_JD_{IJ}$ were generated for each value of D_{IJ} in steps of 0.1 Å in the range

$$L_{IJ} \le D_{IJ} \le U_{IJ}$$

For example, a pair of atoms with associated lower and upper bounds of 5.6 and 6.9 Å, respectively, would produce the following fragments: $A_IA_J5.6$, $A_IA_J5.7$, $A_IA_J5.8$, . . . , $A_IA_J6.7$, $A_IA_J6.8$, and $A_IA_J6.9$ (rather just a single fragment if a rigid molecule was being used). The voluminous body of fragment occurrence data was sorted, cumulated, and input to the screen set generation algorithm as described previously. The resulting screen sets contain different sets of distance ranges from those in the screen sets resulting from the set of 200 rigid molecules; however, the two different types of screen set have comparable levels of screenout in pharmacophoric pattern searches. We thus conclude that it is sufficient to use files of rigid molecules to generate screen sets for the characterization of files of flexible molecules.

EXPERIMENTAL RESULTS

Both flexible and rigid screen searches were carried out; the results of these searches (using the 1024-member screen set) are presented in Table 1. This table lists the average numbers of structures matching the query after the completion of the screen search so that the smaller the number of structures, the greater the screenout. The general trends in these figures are as expected; i.e., screenout increases with an increasing number of atoms in the query, and decreases with an increase in the query tolerance. Of much greater importance in the context of the present study is the finding that the flexible searches result in a substantial level of screenout, albeit, and unsurprisingly, much less than that in the rigid

Table 1. Average numbers of molecules matching a query pharmacophoric pattern in flexible and rigid screen searches

Source of query pattern	Tolerance (Å)	Flexible		Rigid	
		Mean	Median	Mean	Median
Literature	See Figure 2	484	374	93	70
3-atom	±0.1	373	492	189	144
random	±0.5	448	588	348	359
5-atom	±0.1	180	160	37	15
random	±0.5	243	259	119	116

searches. The method of screen assignment means that more bits are set in the bit string describing a flexible molecule than in the bit string describing a rigid molecule, since bits are set for all of the screens that apply to a particular interatomic distance range. The difference between the bit strings characterizing the rigid and the flexible versions of a molecule will increase with the degree of flexibility in that molecule. For the 1024-member screensets used here, the mean density of the bit strings for the rigid and flexible molecules were 10.2% and 58.7%, respectively; i.e., over five times as many bits were set to describe a flexible molecule as were set to describe a rigid molecule. Despite this, the figures in Table 1 demonstrate that the bounded distance matrix provides a representation of the conformational space of a flexible molecule that is sufficiently precise to allow a fair degree of discrimination in a screen search of a database of flexible 3D molecules.

The structures matching the query in the screening stages of the rigid and flexible searches were processed by the Ullmann algorithm, as detailed above. The results of these experiments are summarized in Table 2, which lists the average numbers of structures matching the query after the completion of both the screen search and the geometric search, and thus shows the overall effectiveness of the bounded distance matrices for searching. It will be seen that the number of hits for flexible searches is again very much greater than for the rigid searches. This is easily understandable in the light of the fact that while the rigid representation of a molecule may not contain the query pattern, it may yet be contained in the totality of the conformational space for the molecule in question.

The search results for the eight literature queries are detailed in Table 3, which gives the numbers of structures matching the query pharmacophore after the completion of the screen, geometric, and conformational searches. Once a structure has been identified as a hit in the geometric search, it must undergo the final, conformational search. More precisely, each of the subgraph isomorphisms identified in the geometric search must be tested, and there may be several, or many, such isomorphisms for each hit structure. Table 4 gives the total number of isomorphisms identified in each of the eight sets of geometric searches, together with the mean and maximum numbers of isomorphisms per hit structure (the minimum number is one if a structure is to be regarded as a hit after the geometric search). The numbers of hits listed in the righthand columns of Table 3 are the numbers obtained when just the first isomorphism for each matching structure was tested; the results are for the number of hits identified after a single embedding and after a maximum of ten embeddings. It will be seen that the number of hits rises with an increase in the number of attempts that are made to embed the query pharmacophore in the database structure. In fact, after ten attempts, it is possible to embed the majority of the isomorphisms, with the sole exception of the third query: thus, even if only a single isomorphism is tested, the majority of the structures resulting from the geometric search can be shown to contain the pharmacophore (although many of the matching conformers will have rather high energies arising from geometric distortions or close nonbonded contacts[38]). This finding provides further support for the use of a bounded distance matrix to represent a conformationally flexible molecule.

Taking Tables 3 and 4 together, we note that the screen search and the geometric search can eliminate about 90% of the structures in a flexible search of a database, but that each of the remaining 10% of the structures then spawns an average of nine possible matches for the query that must

Table 2. Average numbers of molecules matching a query pharmacophoric pattern in flexible and rigid geometric searches

Source of query pattern	Tolerance (Å)	Flexible		Rigid	
		Mean	Median	Mean	Median
Literature	See Figure 2	147	151	16	1
3-atom	±0.1	249	252	77	17
random	±0.5	344	407	202	132
5-atom	±0.1	33	12	11	2
random	±0.5	80	60	19	12

Table 3. Numbers of matching structures after the screen, geometric and conformational searches, for eight literature pharmacophoric patterns. The conformational search results are those obtained after testing just the first of the isomorphisms for each of the hits resulting from the geometric searches

Query (See Figure 2)	Hits after screen search	Hits after geometric search	Hits after conformational search	
			1 Embed	≤10 Embeds
1	498	145	63	83
2	312	186	120	151
3	707	125	46	58
4	364	73	42	53
5	290	190	99	136
6	384	216	156	185
7	234	86	41	59
8	1084	157	114	133

be evaluated in the conformational search. Thus, for the datasets considered here, the total number of pharmacophoric pattern matches is only slightly smaller than the total number of structures in the database that is being searched. This finding exacerbates a problem that has already been encountered in rigid 3D searching systems, viz., the very large numbers of hits that are obtained when a user specifies a query pharmacophoric pattern. We believe that novel approaches to the processing of search output will be required if full use is to be made of the enhanced retrieval capabilities offered by flexible searching.

One way in which a flexible search might be implemented would be to allow a user to specify the number of hits required. A conventional rigid search would then be carried out; if insufficient hits were obtained, a second search would be executed using some small number of low-energy conformations (as in the ChemDBS-3D molecule produced by Chemical Design Limited), with the full flexible search being invoked only if the first two searches failed to provide sufficient structures. In fact, the user would still have some control over the number of structures retrieved, even in this third and final stage, by specifying the number of attempts that should be made to embed the matches resulting from the flexible geometric search. Alternative criteria would be appropriate if different types of conformational search were to be used to implement the third stage of the search. For example, one could vary the torsion angle increments in a systematic searching algorithm, such as that described by Dammkoehler et al.[39] There might also be merit in an initial, prescreening of the search file using a flexibility index, such as those described by Kier[40] and by Fisanick et al.[41] This would serve to remove structures that are unlikely to bind strongly to the receptor at an early stage and thus save much computational expense in conformational searching. Alternatively, if it is not wished to discard structures entirely, the index could be used to prioritize the structures passing the geometric search—the more rigid molecules being submitted preferentially for the conformational search.

Table 4. Numbers of isomorphisms resulting from the geometric search, for eight literature pharmacophoric patterns

Query (See Figure 2)	Number of isomorphisms		
	Total	Mean	Maximum
1	497	3.2	30
2	1136	6.1	96
3	1620	12.5	104
4	928	12.7	114
5	1138	6.0	80
6	3210	14.9	300
7	705	8.2	48
8	1296	8.6	36

The performance of the screen and geometric searches could be increased by a further tightening of the distance bounds in the database structures; this could be effected by the use of the tetrangle inequality, in addition to the triangle inequality, in the bound-smoothing process. Previous experiments in this area have suggested that useful improvements can be obtained in this way.[28,42] The application of a more stringent geometric constraint means that the bounded distance matrix produced is a better approximation to the molecule's conformational space and it is thus anticipated that improvements in screenout would be produced. By tightening the distance ranges in the molecule, the probability of overlap with the pattern ranges would be diminished, and thus an increase in the efficiency of the refinement procedure in the Ullmann algorithm might also be reasonably expected. Since an algorithm for tetrangle inequality bound smoothing is both complex to implement and computationally expensive,[28,43] simulation experiments were carried out to investigate the benefits that might be expected to accrue from its implementation. The magnitude of the distance bounds contained in each of the bounded distance matrices was uniformly reduced by $R\%$ and then a new bitmap was produced and screen and geometric searches performed, using various values for R. The results, in terms of the percentage decrease in the number of molecules matching the query at the screen and geometric levels for the literature pharmacophores, are shown in Table 5. The

Table 5. Percentage decrease in the numbers of matching structures in the screen and geometric searches resulting from a reduction of *R*% in the extent of the distance bounds for each of the database structures

	Percentage decrease	
R	Screen search	Geometric search
10	5.0	8.5
20	10.9	23.4
40	27.5	53.6
60	41.6	78.4
80	54.4	94.7

figures show that tightening the distance bounds has much less effect on the number of matches in the screen search than in the geometric search.

It should be noted that these experiments are extremely artificial in nature, since each distance bound in a structure is reduced by an identical amount, a situation that is quite different from that which would pertain were an actual tetrangle inequality bound-smoothing algorithm to be applied. Nonetheless, we believe that the implementation of this algorithm, complex as it is, may prove useful at some future stage. Alternatively, both Fisanick et al.[41] and Bradshaw and Maliski[44] have recently described database-analytic procedures that can be used to set upperbounds to interatomic distances that are tighter than those resulting from the bound-smoothing procedure that has been used in our experiments.

COMPUTATIONAL REQUIREMENTS

The discussion so far has focused on the retrieval performance of the various search algorithms, without consideration of their efficiencies. The following discussion will show that flexible searching is extremely demanding of computational resources.

Let $N(FS)$ and $N(RS)$ denote the numbers of structures matching a query pharmacophore after the flexible screen search and in the rigid screen search, respectively; let $N(FG)$ and $N(RG)$ denote the corresponding numbers of matching structures after the two types of geometric search. Let $T(FG)$, $T(RG)$, and $T(FC)$ denote the run times for a single flexible geometric search, for a single rigid geometric search, and for a single conformational search (which only applies in the flexible context). Then, the time for the flexible geometric search is given by

$$N(FS) \times T(FG)$$

and that for the rigid geometric search by

$$N(RS) \times T(RG)$$

Let I denote the average number of isomorphisms identified in each hit from the flexible geometric search; then the overall time for the conformational search is given by

$$I \times N(FG) \times T(FC)$$

Thus, the ratio, α, of the run times for the flexible and rigid searches is given by

$$\frac{I \times N(FG) \times T(FC) + N(FS) \times T(FG)}{N(RS) \times T(RG)} \tag{3}$$

It will be noted that this analysis ignores the times, $T(FS)$ and $T(RS)$, for the flexible and rigid screening searches. This is because these times are negligible in comparison with the times required by the subsequent stages of the two searches.

If we take the median results for the literature pharmacophoric patterns from Tables 1, 2 and 4, then $I = 8.4$, $N(FG) = 151$, $N(FS) = 374$, and $N(RS) = 70$. Substituting these values into Equation (3), we obtain

$$\alpha = \frac{8.4 \times 151 \times T(FC) + 374 \times T(FG)}{70 \times T(RG)} \tag{4}$$

For our FORTRAN 77 programs on an IBM 3083 BX processor, the values for $T(FG)$ and $T(RG)$ are about 0.3 and 0.03 seconds respectively, and on a VAX 8820, $T(FC)$ is approximately 10.0 CPU seconds. Thus, $T(FG)$ is about an order of magnitude larger than $T(RG)$. The reason for this difference is that there is a much greater probability that a query distance will match with a distance range (as in a flexible geometric search) than that it will match with a single distance (as in a rigid geometric search). Thus, the M matrix, which contains the possible matches of a database structure's atoms with the query atoms, will be much more densely populated in a flexible search than in a rigid search. This will result in a substantial reduction in the amount of search-tree pruning that can take place, with a consequent increase in the amount of backtracking that is required.

Substituting the values above into Equation (4), we obtain a value for α of about 6093. This should be regarded only as an order-of-magnitude estimate for several reasons (even if we ignore the possible effects of the hardware and software that is used to implement our algorithms):

- We have assumed a value for I of 8.4, as indicated by the results in Table 4. However, there are some applications which would not require all the isomorphisms to be checked. For example, consider the use of a 3D searching system to identify structures for biological screening. In this case, once a molecule has been identified as containing the query pharmacophore, it would go forward for testing and there would be no need to execute the conformational search for any further isomorphisms that had been identified in that molecule.
- The value for $T(FC)$ derives from the use of distance geometry, which is one of the slower methods for conformational search.
- The calculation has assumed that only a single embedding is required for each isomorphism; the run times will be correspondingly greater the more embeddings that need to be attempted before the molecule can be accepted or rejected as a match for the query pharmacophore.
- For simplicity, we also have assumed that the IBM and VAX machines run FORTRAN programs at a comparable speed; i.e., the value of $T(FC)$ obtained on the VAX 8820 has been used as if it were obtained on the IBM 3083. In reality, the more modern VAX is likely to be somewhat faster.
- We have considered queries that contain only inter-

atomic distance information, and the inclusion of other types of constraint might affect the relative search times for flexible and rigid searching.

The first two of the factors in this list could overestimate the true value of α; the third and fourth, conversely, would underestimate it; it is not possible to determine the effect of the fifth without further investigation of a wide range of types of query. It should also be emphasized that:

- Our experiments have involved only a very small database of structures, because of the extremely extended run times that are necessary for flexible searching.
- The quoted figures are median values, and individual patterns can involve very extended run times. For example, the large tolerances for two of the interatomic distances in the fourth query of Figure 2 resulted in an extremely slow geometric searching stage for this particular pattern, while the two carbons in the eighth query led to over 70% of the database having to undergo the geometric search, despite the fact that the tolerances for this query are very small.
- We have used only one size of screen set with one particular set of atomic descriptors. Different levels of screenout would be obtained if different types of screen set were to be used, e.g., one based on heteroatoms, ring centroids, etc.,[15] rather than on all non-hydrogen atoms as here.

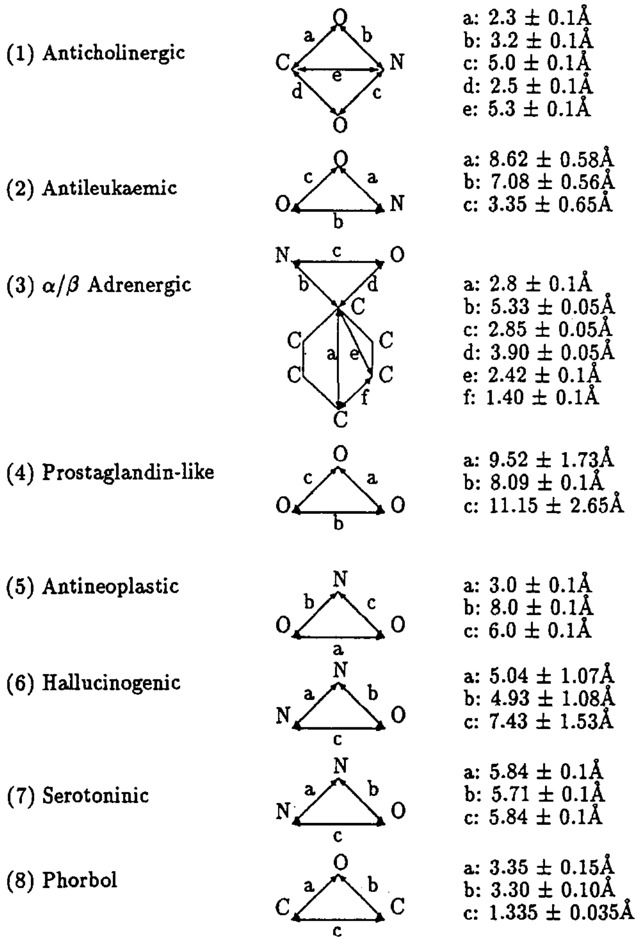

Figure 2. Query pharmacophoric patterns from the literature

Accordingly, no significance should be attached to the particular value of α noted above: a reasonable conclusion from our results would be that flexible 3D searching is at least two or three orders of magnitude slower than rigid 3D searching (which is, in its turn, very much slower than 2D searching).

An important determinant of α is the choice of conformational searching procedure. Our experiments have used embedding, but there are several others with rather different computational requirements.[34] One approach, which we hope to evaluate in the near future, is the use of the conformational searching algorithm described by Dammkoehler et al.[39] This achieves efficiency in search by utilizing any geometric constraints that can be provided; in the database searching context, this would be the matching distance ranges that result from the geometric search. Since it is the conformational search that makes the greatest contribution to α, as defined by Equation (4), any increases in the efficiency of conformational searching will have a large effect on the overall run time of a flexible searching system. However, even if the time requirements of this final stage were to be very substantially reduced, flexible searching would be still far more demanding of computational resources than is rigid searching.

CONCLUSIONS

The current generation of 3D database searching systems provide efficient access to databases of rigid 3D structures, and it is also possible to carry out comparable searches on databases that use a small number of low-energy conformations to summarize the conformational space of a flexible 3D molecule. In this paper, we have described algorithmic techniques that allow the representation and searching of flexible 3D structures, without the need to select individual, representative conformations. Our results suggest that bounded distance matrices provide an effective means of representing a conformationally flexible molecule since we are able to answer one of the major unsolved problems in 3D database searching: "Can this molecule adopt a conformation so that it can match this query pharmacophoric pattern?" It must be emphasized that we can, at present, answer this question only in geometric terms, without taking explicit account of the energies of the conformations identified by our matching algorithms. We intend to evaluate the use of a range of energy calculation methods for processing the hits from the geometric search, with the aim of ranking the search output in order of increasing energy, as discussed above. Our results also suggest that flexible searching can yield extremely large numbers of hits in searches for typical pharmacophores.

While we have demonstrated that it is possible to carry out flexible searches, we have also shown that such searches are extremely demanding of computational resources, even without any subsequent energy calculation. We thus believe that there is very substantial scope for the development of improved software and hardware techniques for flexible searching: if such techniques are not forthcoming, then flexible searching may be restricted to small files of 3D structures or to file subsets that have been generated using alternative retrieval criteria.

ACKNOWLEDGMENTS

We thank ICI Pharmaceuticals and the Science and Engineering Research Council for the award of a CASE Studentship, and Michael Lynch for a careful reading of the manuscript. This work has profited from helpful discussions on the subject of conformational flexibility with David Leahy (ICI Pharmaceuticals), Tad Hurst (Tripos Associates), Robert Pearlman (University of Texas at Austin), and Yvonne Martin (Abbott Laboratories).

REFERENCES

1 Barnard, J.M. Recent developments in chemical structure handling. *Perspectives in Information Management* 1989, **1**, 133–168

2 Lipscombe, K.J., Lynch, M.F. and Willett, P. Chemical structure processing. *Ann. Rev. Inf. Sci. Technol.* 1989, **24**, 189–238

3 Cohen, N.C., Blaney, J.M., Humblet, C., Gund, P. and Barry D.C. Molecular modeling software and methods for medicinal chemistry. *J. Med. Chem.* 1990, **33**, 883–894

4 Gund, P., Wipke, W.T. and Langridge, R. Computer searching of a molecular structure for pharmacophoric patterns. *Proceedings of the International Conference on Computers in Chemical Research and Education, Ljubljana, July 12–17, 1973,* Elsevier, Amsterdam, pp. 33–38

5 Gund, P. Three-dimensional pharmacophoric pattern searching. *Progress in Molecular and Subcellular Biol.* 1977, **5**, 117–143

6 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

7 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

8 Cringean, J.K., Pepperrell, C.A., Poirrette, A.R. and Willett, P. Selection of screens for three-dimensional substructure searching. *Tetrahedron Comp. Method.* 1990, **3**, 37–46

9 Poirrette, A.R., Willett, P. and Allen, F.H., Pharmacophoric pattern matching in files of 3-D chemical structures: characterization and use of generalized valence angle screens, *J. Mol. Graphics* 1991, **9**, 203–217

10 Jakes, S.E., Watts, N.J., Willett, P., Bawden, D. and Fisher, J.D. Pharmacophoric pattern matching in files of three-dimensional chemical structures: evaluation of search performance. *J. Mol. Graphics* 1987, **5**, 41–48

11 Bawden, D., Devon, T.K., Faulkner, D.T., Fisher, J.D., Leach, J.M., Reeves, R.J. and Woodward, F.E. Development of the Pfizer integrated research data system SOCRATES. in *Chemical Structures: the International Language of Chemistry* (W.A. Warr, ed.) Springer-Verlag, Berlin, 1988 pp. 63–75

12 Rusinko III, A., Sheridan, R.P., Nilakantan, R., Haraki, K.S., Bauman, N. and Venkataraghavan, R. Using CONCORD to construct a large database of three-dimensional coordinates from connection tables. *J. Chem. Inf. Comp. Sci.* 1989, **29**, 251–255

13 Sheridan, R.P., Nilakantan, R., Rusinko III, A., Bauman, N., Haraki, K.S. and Venkataraghavan, R., 3DSEARCH: a system for three-dimensional substructure searching. *J. Chem. Inf. Comp. Sci.* 1989, **29**, 255–260

14 Christie, B.D., Henry, D.R., Güner, O.F. and Moock, T.E. MACCS-3D: a tool for three-dimensional drug design. in *Online Information 90: 14th International Online Information Meeting Proceedings, London* (D.I. Raitt, ed.) Learned Information, Oxford, 1990, pp. 137–161

15 Murrall, N.W. and Davies, E.K. Conformational freedom in 3-D databases. 1. Techniques. *J. Chem. Inf. Comp. Sci.* 1990, **30**, 312–316

16 Davies, E.K. and Upton, R.M. Performance characteristics in 3-D database searching. in *Online Information 90: 14th International Online Information Meeting Proceedings, London* (D.I. Raitt, ed.) Learned Information, Oxford, 1990, pp. 129–136

17 van Drie, J.H., Weininger, D. and Martin, Y.C. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure search of 3D molecular structures. *J. Comp.-Aided Mol. Design* 1989, **3**, 225–251

18 Martin, Y.C. Computer design of potentially bioactive molecules by geometric searching with ALADDIN. *Tetrahedron Comp. Method.* 1990, **3**, 15–25

19 Bartlett, P.A., Shea, G.T., Telfer, S.J. and Waterman, S. CAVEAT: a program to facilitate the structure-derived design of biologically active molecules. in *Molecular Recognition: Chemical and Biochemical Problems* (S.M. Roberts, ed.) Royal Society of Chemistry, Cambridge, 1989, pp. 182–196

20 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

21 Willett, P. *Three-Dimensional Chemical Structure Handling*. Research Studies Press, Taunton, 1991

22 Perun, T.J. and Propst, C.L., eds. *Computer-Aided Drug Design.* Marcel Dekker, New York, 1989

23 Crippen, G.M. Distance geometry approach to rationalizing binding data. *J. Med. Chem.* 1979, **22**, 988–997

24 Behling, R.W., Yamane, T., Navon, G. and Jelinski, L.W. Conformation of acetylcholine bound to the nicotinic acetylcholine receptor. *Proc. Nat. Acad. Sci. USA* 1988, **85**, 6721–6725

25 Haraki, K.S., Sheridan, R.P., Rusinko, A., Venkataraghavan, R., Dunn, D.A. and McCulloch, D. Looking for pharmacophores in 3-D databases: does conformational searching improve the yield of actives? paper presented at the 4th Chemical Congress of North America, New York, 25–30 August 1991

26 Havel, T.F., Kuntz, I.D. and Crippen, G.M. The theory and practice of distance geometry. *Bull. Math. Biol.* 1983, **45**, 665–720

27 Ghose, A.K. and Crippen, G.M. Geometrically feasible

binding modes of a flexible ligand molecule at the receptor site. *J. Comp. Chem.* 1985, **6**, 350–359

28 Easthope, P.L. and Havel, T.F. Computational experience with an algorithm for tetrangle inequality bound smoothing. *Bull. Math. Biol.* 1989, **51**, 173–194

29 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

30 Smellie, A.S., *Distance Geometry: New Methods and Applications*. D. Phil. thesis, University of Oxford, 1989

31 Wenger, J.C. and Smith, D.H. Deriving 3D representations of molecular structure from connection tables augmented with configuration designations using distance geometry. *J. Chem. Inf. Comp. Sci.* 1982, **22**, 29–34

32 Ullmann, J.R. An algorithm for subgraph isomorphism. *J. Assoc. Computing Machinery* 1976, **23**, 31–42

33 Hurst, T., private communication

34 Howard, A.E. and Kollman, P.A. An analysis of current methodologies for conformational searching of complex molecules. *J. Med. Chem.* 1988, **31**, 1669–1675

35 Blaney, J.M., Crippen G.M., Dearing, A. and Dixon, J.S. DGEOM: distance geometry. Quantum Chemistry Program Exchange program number 590, Department of Chemistry, Indiana University, Bloomington, Indiana, USA

36 Daylight Chemical Information Systems, 3951 Claremont Street, Irvine, California, USA

37 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)

38 Kuntz, I.D., Thomason, J.F. and Oshiro, C.M. Distance geometry. *Meth. Enzymology* 1989, **177**, 159–204

39 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

40 Kier, L.B. An index of molecular flexibility from kappa shape attributes. *Quant. Struct.-Act. Relat.* 1989, **8**, 218–221

41 Fisanick, W., Cross, K.P. and Rusinko III, A. Characteristics of computer-generated 3D and related molecular property data for CAS Registry substances. paper presented at the 4th Chemical Congress of North America, New York, 25–30 August 1991

42 Crippen, G.M. A novel approach to the calculation of conformation: distance geometry. *J. Comp. Phys.* 1977, **24**, 96–107

43 Havel, T.F. An evaluation of 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

44 Bradshaw, J. and Maliski, E.G. Use of most restrictive paths in 3-D search strategy. Paper presented at the 4th Chemical Congress of North America, New York, 25–30 August 1991