Pharmacophoric pattern matching in files of three-dimensional chemical structures: Use of smoothed bounded distances for incompletely specified query patterns

David E. Clark and Peter Willett

Department of Information Studies, University of Sheffield, Western Bank, Sheffield, UK

Peter W. Kenny

ICI Pharmaceuticals, Mereside, Alderley Park, Macclesfield, Cheshire, UK

This paper describes a technique for increasing the screenout of pharmacophoric pattern searches in the databases of three-dimensional chemical structures when only some of the interatomic distances in the query pattern are specified. The technique involves the application of a distance bounds-smoothing procedure to the query distances; this smoothing allows the calculation of upper and lower bounds for the unspecified distances. The bounded distances can then be used to set screens additional to those that are set to describe the distances that have been specified by the searcher. Evidence is presented to suggest that use of the technique can lead to increases in the efficiency of substructure searches for partially specified query patterns.

Keywords: Bounds smoothing, distance geometry, geometric search, pharmacophoric pattern matching, screen search, screenout, smoothed bounded distances

INTRODUCTION

Computerized information systems for the storage and retrieval of information about chemical structures are extensively used in chemical research and development.[1] One of the most important facilities in chemical information systems is that of *substructure searching*, which involves the identification of all molecules in a database that contain a user-defined partial structure. Systems for two-dimensional (2D) substructure searching have been available for over two decades: The last few years have seen an explosion of interest in the development of substructure searching techniques for databases of three-dimensional (3D) chemical structures, where atomic coordinate information is available from X-ray or molecular mechanics studies. Following early work by Gund and coworkers,[2] studies at the University of Sheffield[3–5] and at Abbott Laboratories[6] have demonstrated that effective 3D substructure searches can be carried out using techniques suitably adapted from those used for 2D substructure searching. These studies have resulted in a range of both in-house and commercial 3D substructure searching systems.[7]

Substructure searching is usually implemented by means of a two-stage retrieval algorithm in which an initial *screening* search is used to eliminate from further consideration large numbers of molecules that cannot possibly contain the query substructure; only those molecules that match the query substructure at the screen level then undergo a detailed and time-consuming subgraph isomorphism search. The screens that are used for 3D substructure searching are often based on interatomic distance ranges,[3,8] with each screen being represented by setting a bit in a bit-string that describes each molecule in the database. Then, when the query pattern is searched against a database, all of the bits that are set in the query bit-string must also be present in the string describing an individual molecule if that molecule is to be passed onto the subgraph isomorphism search. This second stage is far more demanding of computational resources than is the screening search. The overall efficiency of a substructure searching system is thus crucially dependent on the *screenout*, i.e., the fraction of the database that is eliminated from the subgraph isomorphism search by the screening search, and there accordingly has been considerable interest in the development of techniques that will ensure a

Address reprint requests to Dr. Willett at the Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.
Received 26 February 1991; accepted 12 March 1991

high level of screenout in searches for query substructures.[3,8,9]

It is often the case that only some of the interatomic distances within a pharmacophoric pattern are known with any precision, and thus the screenout will be less than if all of these distances were known. In this short communication, we discuss the use of a simple technique, derived from *distance geometry*,[10,11] to improve the screenout that can be obtained when only a small number of pattern distances have been specified.

USE OF SMOOTHED BOUNDED DISTANCES

Distance geometry is being used increasingly to calculate the 3D structures of both small molecules and macromolecules. The basic approach involves the use of some known distances within a 3D molecule to deduce upper and lower bounds on those distances that are unknown. This may be illustrated by considering a structure that contains three atoms, X, Y, and Z. We shall assume in what follows that all three atoms are of the same elemental type (although the analysis is not dependent on this assumption), that the distance $D(X, Z)$ is unknown, and that the other two distances $D(X, Y)$ and $D(Y, Z)$ are known; e.g., if X is bonded to Y and Y is bonded to Z, then the corresponding distances might come from lists of standard bond lengths. The use of the triangle inequality allows one to calculate upper and lower bounds for the unknown distance since, necessarily,

$$D(X, Y) - D(Y, Z) \leq D(X, Z) \leq D(X, Y) + D(Y, Z)$$

For an N-atom molecule, an $N \times N$ *bounded distance matrix* is created, the upper and lower elements of which contain the upper bounds and the lower bounds, respectively, for each of the distances. If the distance is known precisely, then the two bounds are the same, while all of the diagonal elements are zero-valued. The geometric inconsistencies in this matrix are eliminated by *triangle smoothing*, an iterative process that takes each possible set of three atoms in turn and repeatedly applies the triangle inequality to produce a *smoothed, bounded distance matrix*. The full distance geometry algorithm then proceeds to calculate the refined 3D coordinates for the molecular structure so that they satisfy the distance constraints contained in this smoothed matrix.[11] In the present context, however, we are interested in just the smoothed, bounded distance matrix.

It is often the case that only some of the interatomic distances in a query pharmacophore are known when a query pattern is to be searched against a database of 3D structures. This is likely to be especially true in the initial stages of an investigation, e.g., when only a few active structures have been identified. Consider the three-atom pattern described previously, where the distance $D(X, Z)$ is unspecified. Bits would be set in the query bit-string corresponding to the screens appropriate to the two known distances $D(X, Y)$ and $D(Y, Z)$, but it would not be possible to set any bits to describe $D(X, Z)$, with a consequent lowering of the screenout that could be obtained when compared to the situation when this distance was known. However, if the query pattern is submitted to a distance geometry routine, we can calculate the smoothed upper and lower bounds for $D(X, Z)$, $U(X, Z)$, and $L(X, Z)$, respectively. The screen dictionary is then consulted and bits are set for all screens that have an associated distance d such that

$$L(X, Z) \leq d \leq U(X, Z)$$

Thus, if a database structure is to match at the screen level, it must not only have bits set corresponding to the known distances but must also have at least one bit corresponding to a distance in the range $L(X, Z) - U(X, Z)$. The screenout clearly cannot be *less* than that obtained when bits are set only for the known distances in a query pattern: In the remainder of this paper, we describe experiments which suggest that the screenout can, in fact, be considerably *more* in some circumstances.

EXPERIMENTAL DETAILS

Datasets

The sample file that was used for the searches derived from 1538 structures from the Pomona College database. This file did not contain any ionic or disconnected structures or molecules containing less than two heteroatoms and one carbon atom. The atomic coordinates were obtained by use of the CONCORD program, which was developed by Pearlman and his associates at the University of Texas at Austin and which is now distributed by Tripos Associates.

It was necessary to have sets of query pharmacophoric patterns for searching this database that would allow us complete control over the number of interatomic distances that could be specified (and hence over the number of interatomic distances that would need to be approximated by the smoothing routine). Accordingly, the query patterns were generated by taking every hundredth structure from the data set and then extracting some fixed number of randomly selected atoms from each of the 15 resulting structures. Patterns containing 3, 5, and 7 atoms were used, with the maximum number of carbons permitted in these patterns being limited to 1, 2, and 4 atoms, respectively. The 3D coordinates of the selected atoms were then used to calculate the complete set of interatomic distances for each pattern.

A pattern containing Q atoms will have $Q(Q - 1)/2$ distinct distances. The effect of query specification was investigated by randomly selecting q of the distances, where $q \leq Q(Q - 1)/2$. For each such selection, the smoothed upper and lower bounds for the remaining $(Q(Q - 1)/2) - q$ distances, i.e., those that were being assumed to be unspecified, were calculated using the PREP and SMOOTH modules of the suite of distance geometry programs described by Smellie.[12] In this way, sets of patterns were obtained that contained fixed numbers of atoms and fixed numbers of both specified and unspecified distances. These patterns were then searched against the file of 1538 sets of coordinates.

Screening

The database structures and query patterns were represented for search by bit-string screen records as described in the introduction. The screens that were used consisted of a pair of atoms with an associated interatomic distance range, and were generated using the screen-set selection algorithm de-

scribed by Cringean et al.[9] This algorithm involves an analysis of a file of structures, typical of those that are to be screened, to generate all of the (nonhydrogen) interatomic distances present in the file. This results in a list of fragments of the form A_1A_2D, where A_1 and $A_2 (A_1 \leq A_2)$ are the atomic types of the pair of atoms that is being considered and where D is the distance (in angstroms) between them. This file of interatomic distance descriptors is then sorted into increasing alphanumeric order and cumulated, so that each interatomic distance is stored with its frequency of occurrence. The file is then divided into S partitions, where S 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. The subdivision is carried out in such a way that each of the resulting partitions contains approximately the same number of interatomic distance occurrences. In the work reported here, two sets of screens were used, with S equal to 512 and to 1024.

The screen sets were used to produce bit-string representations of each of the 1538 molecules; these bit-strings were then stored as a bitmap[4] for rapid matching against the screen records for each of the query patterns. When a query came to be searched against the database, the specified distances (using a tolerance of ± 0.1 Å) were used to set screens directly in the query bit-string. For the unspecified distances, all bits were set that corresponded to screens included in the range between the lower and upper bounds on the distance.

RESULTS AND DISCUSSION

Assume that a query pattern containing Q atoms is to be searched. Some user-defined number q of the $Q(Q - 1)/2$ distances is selected at random and the remaining distances are calculated as described previously. Searches are then carried out in which the query pattern is represented, first, only by the screens for the q specified distances and, second, not only by these screens but also by the bounded distance screens for the unspecified distances. The efficiency of a substructure search is determined primarily by the number of molecules that match the query at the screen level and that need to undergo the time-consuming subgraph isomorphism search. Thus, the efficiency of the technique suggested in this paper may be determined by noting this number when only the user-specified distances are used in a substructure search and when all of the distances were used, i.e., when the user-defined distances are augmented by those derived from the smoothing procedure. These two numbers of structures, which are referred to subsequently as X and Y, respectively, were noted for each substructure search that was carried out, and the results were then averaged over the 15 query patterns for each size of Q atoms. The results are detailed in Table 1, which lists the mean values for X and Y that were obtained for patterns with $Q = 3$, 5, or 7, respectively, and with varying values for q. The tables also include the percentage improvement P obtained from the smoothing procedure, where P is defined to be

$$P = \frac{100 \times (X - Y)}{X}$$

Table 1. Effect of query specification on screenout with (a) three-atom, (b) five-atom, and (c) seven-atom patterns: q is the number of distances specified in the original query pattern; X and Y the mean numbers of molecules, averaged over all of the queries of size q-atoms, that match the query at the screen level when just the specified distances are used and when both specified and calculated distances are used; and P is the percentage improvement in performance. The first and second columns in each element of the tables represent the results obtained with sets of 512 and 1024 screens, respectively

q	X		Y		P	
1	761	732	557	524	26.8	28.4
2	349	300	309	254	11.5	15.3

(a)

q	X		Y		P	
1	1080	1067	464	416	57.0	61.0
3	300	259	216	172	28.0	33.6
5	156	114	124	84	20.5	26.3
7	87	58	83	54	4.6	6.9
9	71	45	69	43	2.8	4.4

(b)

q	X		Y		P	
1	1080	1067	449	418	58.4	60.8
2	888	857	407	363	54.1	57.6
3	627	601	328	297	47.7	50.6
4	439	403	249	229	43.3	43.2
5	210	179	147	118	30.0	34.1
6	126	95	95	70	24.6	26.3
9	88	60	73	49	17.0	18.3
12	60	38	52	33	13.3	13.2
15	44	29	41	27	6.8	6.9
18	36	24	34	22	5.6	8.3

(c)

An inspection of the results justifies the approach suggested in this paper because it will be seen that Y is consistently less than X, i.e., that fewer molecules need to undergo the subgraph isomorphism search when the smoothing procedure is used than would otherwise be the case. The magnitude of P, which measures the percentage improvement in performance, increases as q decreases, so that the efficiency of the technique is greatest for very poorly defined query patterns (although the largest values of P listed in the table are rather unrealistic in that one would expect that at least some nontrivial percentage of the query distances would be specified in normal circumstances). When the pattern is well defined, the improvements in performance are small but still significant, typically being on the order

of a 5–10 % reduction in the number of molecules that pass the screening search.

The augmentation of a query pattern by distance bounds-smoothing means that the efficiency of 3D substructure searching will always be at least as good as when just the specified distances are used. Our results show that the performance is, in fact, rather better, thus suggesting the use of distance bounds-smoothing as a simple but effective method for query optimization in 3D substructure searching systems.

ACKNOWLEDGEMENTS

We thank the Department of Education and Science for the award of an Information Science Advance Course Studentship, and Catherine Pepperrell and Andrew Poirrette for programming support.

REFERENCES

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

2 Gund, P., Wipke, W.T. and Langridge, R. Computer searching of a molecular structure for pharmacophoric patterns. In: *Proceedings of the International Conference on Computers in Chemical Research and Education*. Ljubljana, 1973, 33–38

3 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

4 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

5 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

6 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 searching of three-dimensional molecular structures. *J. Comp.-Aided Mol. Design* (1989) **3**, 225–251

7 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, 213–263

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

9 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

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

11 Crippen, G.M. and Havel, T.F. *Distance Geometry and Molecular Conformation*. Research Studies Press, Letchworth, 1988, 245–253

12 Smellie, A.S. *Distance Geometry: New Methods and Applications*. DPhil thesis, Univ. of Oxford, Oxford, 1989