Similarity Searching in Databases of Three-Dimensional Molecules and Macromolecules

Peter J. Artymiuk, Peter A. Bath, Helen M. Grindley, Catherine A. Pepperrell, Andrew R. Poirrette, David W. Rice, David A. Thorner, David J. Wild, and Peter Willett*

Departments of Information Studies and of Molecular Biology and Biotechnology, Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, England

Frank H. Allen

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

Robin Taylor

ICI Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY, England

Received June 15, 1992

This paper discusses algorithmic techniques for measuring the degree of similarity between pairs of three-dimensional (3-D) chemical molecules represented by interatomic distance matrices. A comparison of four methods for the calculation of 3-D structural similarity suggests that the most effective one is a procedure that identifies pairs of atoms, one from each of the molecules that are being compared, that lie at the center of geometrically-related volumes of 3-D space. This atom mapping method enables the calculation of a wide range of types of intermolecular similarity coefficient, including measures that are based on physicochemical data. Massively-parallel implementations of the method are discussed, using the AMT Distributed Array Processor, that achieve a substantial increase in performance when compared with a sequential implementation on a UNIX workstation. Current work involves the use of angular information and the extension of the method to field-based similarity searching. Similarity searching in 3-D macromolecules is effected by the use of a maximal common subgraph (MCS) isomorphism algorithm with a novel, graph-based representation of the tertiary structures of proteins. This algorithm is being used to identify similarities between the 3-D structures of proteins in the Brookhaven Protein Data Bank; its use is exemplified by searches involving the NAD-binding fold motif.

1. INTRODUCTION

Most current computer-based chemical information systems use two-dimensional (2-D) connection tables as the primary means of representing chemical structure information.[1] One of the most important facilities provided by such systems is similarity searching, in which those molecules are retrieved that are most similar to an input query molecule.[2,3] Such systems are of particular importance for the rational design of new drugs and pesticides, since molecules that are structurally similar to each other may well exhibit similar activity characteristics.

There are very many ways in which the similarity between a pair of molecules might be defined.[4] Current similarity-searching systems generally use a computationally-efficient approach in which the similarity between a pair of molecules is based on the fragment substructures that they have in common; specifically, a similarity search is implemented by matching the fragment bit strings that are used for screening purposes in substructure searches. A similarity search involves the user inputting a *target* molecule, e.g., a lead compound that had been shown to exhibit some beneficial chemical or biological property. The similarity is calculated between the target and each compound in the database by comparing the corresponding fragment bit strings to identify the bits (and hence substructural fragments) that they have in common; this information is then used to calculate a similarity coefficient for each structure in the file, which is finally sorted into order of decreasing similarity with the target. Interesting compounds from near the top of this ranking can then be used as the basis for subsequent searches.[3]

Recent developments in molecular modeling[5] mean that it is now relatively easy to generate the 3-D atomic coordinates for a small molecule from its 2-D connection table, thus allowing the creation of in-house databases of 3-D structures from the corresponding databases of 2-D structures. Several software systems have subsequently been developed for substructure searching in these 3-D databases, as reviewed by Martin et al.[6] and by Willett.[7] In this paper, we present an overview of several studies that are in progress at the University of Sheffield to develop methods for determining the similarity between pairs of 3-D structures that could complement these substructure-searching systems; an alternative approach to 3-D similarity searching is being developed by van Geerestein et al.[8] We also describe work on a related project that is developing methods for similarity searching in 3-D macromolecules, specifically the protein structures in the Brookhaven Protein Data Bank.

2. ATOM MAPPING METHOD

2.1. Comparison of Similarity Methods Using Activity Data. The most important characteristic of a similarity-searching system is the effectiveness of the measure that is used to quantify the degree of structural resemblance between a pair of molecules. Our work on 3-D similarity searching in databases of small molecules is aimed at supporting drug and pesticide discovery. An obvious measure of the effectiveness of a similarity measure is thus the extent to which the calculated similarities in structure mirror similarities in activity: if an active molecule is searched against a database containing both

* To whom all correspondence should be addressed at the Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

active and inactive structures, a good similarity measure will be one in which the similarity of the target compound with the actives will tend to be greater than its similarity with the inactives, resulting in a clustering of the active molecules at the top of the ranking. Different similarity measures can then be compared by the extent to which this clustering occurs.

A 3-D molecule is commonly represented by an interatomic distance matrix, in which the *ij*th element denotes the distance between the *i*th and *j*th atoms. Pepperrell and Willett[9] have reported a detailed comparison of four different ways of calculating the similarity between pairs of such distance matrices. Their comparison used 10 small datasets from the medicinal chemistry literature; these datasets covered a wide range of structural types, including both structurally homogeneous and structurally heterogeneous sets of molecules, and contained between 112 and 209 structures.[9] The 3-D atomic coordinates for these sets of molecules were obtained using the CONCORD program.[10] The sets of coordinates were used to calculate a distance matrix for each molecule in each dataset, and these matrices acted as the input to the four similarity procedures that were tested.

Similarity searching presumes the availability of at least one active molecule[3,11] that can be used as the target to identify other, potentially active molecules. An active molecule in a dataset was selected, and its similarity was calculated with each of the other molecules in that dataset, using one of the similarity methods that were being tested. The molecules were then ranked in decreasing order of similarity, and cutoffs were applied to retrieve some fixed number of the top-ranked compounds. These molecules were then checked to determine whether they were active or inactive in the particular biological test system associated with that dataset. The overall utility of a similarity method was calculated by taking the mean numbers of active molecules at the cutoff position when averaged over all of the active compounds that had been used to generate a ranking. Thus, if there were ACT actives in a dataset, and if the use of the *i*th active as the target identified $N(i)$ actives above the chosen cutoff, then the overall effectiveness of the current similarity method was given by

$$\frac{1}{\mathrm{ACT}}\sum_{i=1}^{\mathrm{ACT}} N(i)$$

The comparison suggested that the most cost-effective similarity measure of those tested was the *atom mapping* method, which gave the best overall level of predictive performance at a computational cost that was not such as to render it totally infeasible for the searching of large databases of 3-D structures.[9] The workings of the method are described in the next section.

It should be emphasized that the comparison of similarity methods has been restricted to an evaluation of their abilities at predicting biological activities, this focus reflecting the intended use to support drug and pesticide research. A similar comparison could also be carried out using molecular properties. Such data was used in our earlier studies of 2-D similarity measures,[3] and we hope to carry out an analogous comparison of 3-D similarity measures in the near future.

2.2. The Method. The atom mapping method provides a quantitative measure of the similarity between a pair of 3-D molecules, A and B, that are represented by their interatomic distance matrices. The measure is calculated in two stages.

1. The geometric environment of each atom in one molecule, A, is compared with the corresponding environment of each atom in the other molecule, B, to determine the similarity between each possible pair of atoms.
2. The interatomic similarities resulting from step 1 are used to identify pairs of similar atoms, and these equivalences allow the calculation of the overall, intermolecular similarity.

Assume that the molecule A contains $N(A)$ non-hydrogen atoms. The distance matrix for A, DA, is then an $N(A) \times N(A)$ matrix such that the *ij*th element, $DA(i,j)$, contains the distance between the *i*th and *j*th atoms in A, and similarly for the matrix DB representing the $N(B)$-atom molecule B; in the following, we shall assume that $N(A) \leq N(B)$. The first stage of the procedure compares each atom from A with each atom from B to identify the extent to which each pair of atoms lies at the center of comparable volumes of 3-D space, as defined by the geometric arrangement of the atoms that are contained within these volumes.

Consider the *i*th atom in A, $A(i)$: then the *i*th row of the distance matrix DA contains the distances from i to all of the other atoms in A. This set of distances is compared with each of the rows, j, from the distance matrix DB to identify the matching distances, where a matching distance is one that is the same in the two molecules to within some user-defined tolerance, e.g., ±0.5 Å, and that separates atom pairs that are of the same elemental type(s) in A and B. Let there be $C(i,j)$ such distance matches when $A(i)$ is compared with $B(j)$; then the similarity, $S(i,j)$, between $A(i)$ and $B(j)$ is given by the Tanimoto coefficient:[3]

$$S(i,j) = \frac{C(i,j)}{N(A) + N(B) - C(i,j)}$$

$S(i,j)$ is the *ij*th element of an $N(A) \times N(B)$ matrix, S, that contains the similarities between all pairs of atoms, $A(i)$ and $B(j)$, from A and B. This matrix is referred to subsequently as the *atom match matrix*.

Once the atom match matrix has been created, the interatomic similarities contained within it are used to establish a set of equivalences of the form $A(i) \leftrightarrow B(j)$. A greedy algorithm is used to establish the set of equivalences, as follows:

1. Sort the elements of the atom match matrix into order of decreasing similarity.
2. Scan the atom match matrix to find the remaining pair of atoms, one from A and one from B, that have the largest calculated value for $S(i,j)$.
3. Store the resulting equivalences as a tuple of the form $\{A(i) \leftrightarrow B(j); S(i,j)\}$.
4. Remove $A(i)$ and $B(j)$ from further consideration.
5. Return to step 2 if it is possible to map further atoms in A to atoms in B.

The overall degree of similarity between A and B is then calculated as the mean of the similarities over all of the atoms in A, using the information in the tuples that is stored in step 3 of the algorithm above. Thus, the intermolecular similarity is given by

$$\frac{1}{N(A)}\sum_{i=1}^{N(A)} S(i,j)$$

Ideally, we would like to map the atoms from A onto those from B so as to give the largest possible value for the intermolecular similarity coefficient; however, as discussed by Pepperrell and Willett,[9] the identification of the maximal value for the coefficient involves a combinatorial procedure

Table I. Mean Number of Actives in Top-Ranked Molecules Using Atom Mapping Method and Using Random Selection[a]

	no. of top-ranked molecules					
dataset	5		10		20	
A	4.27	3.85	8.22	7.41	16.15	14.52
B	4.12	3.46	7.60	6.54	14.66	12.69
C	4.43	4.08	8.40	7.94	16.25	15.65
D	4.33	3.77	7.82	7.24	14.87	14.16
E	3.77	2.87	7.09	5.22	12.34	9.90
F	4.11	3.30	7.39	6.18	13.58	11.93
G	4.51	3.30	8.57	6.17	15.73	11.91
H	4.50	3.87	8.63	7.45	16.83	14.61
I	4.35	3.43	8.44	6.46	16.24	12.54
J	2.68	1.99	4.19	3.22	6.81	5.68

[a] The first and second figure in each element of the table are the mean numbers of actives identified using the atom mapping method and using random selection, respectively.

that is far too demanding of computational resources for a practicable similarity-searching system. The greedy algorithm used here provides an efficient and effective heuristic that usually, but not invariably, identifies the most similar pairs of atoms.

Most similarity measures are *global* in character in that they return a number that describes the overall degree of similarity between a pair of objects. This is nearly always the case with nonchemical applications of similarity and also applies to the fragment-based measures that are used in 2-D similarity-searching systems, where the calculated similarities do not provide any information as to which particular parts of the molecules that are being compared are responsible for the observed degree of similarity. Global measures are not inappropriate when we have little or no information about the particular parts of a molecule that are responsible for the biological activity of interest, e.g., when an active lead has been identified and when there is a need to identify analogous structural types. Such measures may be less appropriate in the context of 3-D similarity searching, since biological activity at a receptor site is determined by the presence of a particular set of atoms in a particular geometrical arrangement, i.e., the pharmacophore. The atom mapping method is global in nature since it provides a single, real-valued number that measures the overall degree of similarity between the target and each of the database structures. However, it can also be regarded as a *local* similarity measure since the atom match matrix provides information about the structural equivalences that apply to individual pairs of atoms and about the contribution of each of these pairs to the overall, global similarity. The importance of providing local information has been emphasized by van Geerestein et al.[8] when describing their program, called SPERM, which provides a global measure for 3-D similarity searching.

2.3. Validation of Method. It is obviously important that the ranking of a database that is obtained from the use of a similarity measure is significantly different from that which would be obtained from a randomly-generated ranking of that database. The significances of the rankings that result from the use of the atom mapping method have been investigated by comparing the activities of the top-ranking molecules that are obtained when the atom mapping method is used with the activities of randomly-selected molecules. These experiments used the 10 small datasets mentioned in section 2.1 (denoted here by the letters A–J), with the aim of determining whether the nearest neighbors that were obtained for each active molecule using the atom mapping method contained a significantly greater number of actives than would have been obtained if the same number of nearest neighbors had been selected at random (the latter is easily calculated given the known numbers of active and inactive molecules in each of the datasets). The results of these experiments are detailed in Table I, where it will be seen that the atom mapping method always gives a greater number of actives at the top of the rankings than does the random selection of the same numbers of compounds. A similar result was demonstrated by a more detailed, simulation study of the rankings resulting from the atom mapping method,[13] thus validating the use of the method to identify those molecules in a database that are structurally most similar to an input target structure.

Further support for the effectiveness of the atom mapping method comes from an investigation of the extent to which variations in the various components of the procedure affect the rankings that are produced. Experiments were carried out in which:

(a) The Tanimoto coefficient was replaced by the Overlap coefficient.[3]

(b) The mapping of atoms based on their elemental types was replaced by the mapping of atoms based on augmented atoms (topological fragments that are extensively used for screening purposes in 2-D substructure searching systems[1]).

(c) The mapping of atoms based on the interatomic distances in which they are involved was replaced by the mapping of atoms based on the valence angles in which they are involved.

These variations were found to have little effect on the utility of the rankings that were produced.[13] This suggests that the effectiveness of the atom mapping method is not crucially dependent on the precise way in which it is implemented and, accordingly, that it provides a robust and generally-applicable approach to the calculation of molecular similarity.

2.4. Computational Requirements. A complexity analysis of the atom mapping method shows that the time requirement for the comparison of two molecules containing $N(A)$ and $N(B)$ atom is of the order $O(N(A)^3)$ if $N(A) \approx N(B)$ (with the computation being dominated by the calculation of the atom match matrix).[12] It is thus clear that extremely efficient algorithms are required if the atom mapping method is to be used to search large databases of 3-D structures. Accordingly, we have investigated several upperbound strategies that can reduce the numbers of molecules that need to be compared with the target structure during a similarity search, while still ensuring the retrieval of the true *nearest neighbors*, i.e., those molecules that are structurally most similar to the target compound. Upperbound algorithms have been used previously for 2-D similarity searching[3] and for 3-D maximal common substructure searching.[14]

In the present context, the user who has input the target structure is asked to specify how many of the top-ranked molecules are to be displayed as the output from the similarity search. Let this number of molecules be M, for which typical values might be 20 or 50. Two upperbound strategies are used to increase the efficiency of the search; a detailed account of these strategies is presented by Pepperrell et al.[12] The first strategy uses the molecular formulas of the target molecule and each of the molecules in the search file to calculate an upperbound to the similarity that would be obtained if the atom mapping similarity measure was calculated; the atom mapping procedure is invoked if, and only if, this upperbound is greater than the similarity for the Mth nearest neighbor identified up to that point in the course of the search. If the atom mapping procedure does need to be invoked for a

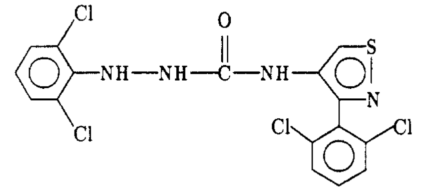

Figure 1. Target structure for an atom mapping search.

particular database structure, then a second upperbound is calculated after the calculation of the interatomic similarities but prior to the mapping stage that results in the calculation of the overall intermolecular similarity; the latter stage is executed if, and only if, this second upperbound is greater than the similarity for the *M*th nearest neighbor identified up to that point in the course of the search. Both of these strategies result in a reduction in the computational requirements of a similarity search, when compared with a simple sequential scan of the whole database. In fact, further increases in performance are achieved if this scan is carried out in decreasing order of the molecular formula upperbounds mentioned previously, since this permits the elimination of further structures from the atom mapping stage. The overall reduction in computation depends on the size of the file, on the number of nearest neighbors required, and on the number of atoms in the target structure: our experiments would suggest that it is possible to reduce the run time by between 60 and 80% of the time required for a sequential scan when none of these upperbound strategies are used.[12]

A prototype search system has been implemented at ICI Agrochemicals, using the upperbound procedures on a file of 4500 structures. A typical search of this file for the 50 nearest neighbors of an input target structure requires about 60 s, using a Fortran 77 program running on an Evans and Sutherland ESV 3 workstation. Thus, similarity searches on a typical corporate database containing a quarter of a million structures would be expected to take about 1 h using comparable equipment, the precise time depending on the target structure and the search parameters that are used. Pepperrell et al. discuss the results of searches using this system and demonstrate the substantial differences that exist between 2-D and 3-D similarity searches that use the same target molecules.[12] An example of a 3-D similarity search is shown in Figure 2, which illustrates the five structures that the atom mapping method judged to be most similar to the target structure shown in Figure 1.

3. USE OF PHYSICOCHEMICAL DATA

The description of the atom mapping procedure that has been presented in section 2.2 assumes that some atom, $A(i)$, in the target structure is considered for mapping to some atom, $B(j)$, in a database structure if, and only if, both $A(i)$ and $B(j)$ are of the same elemental type. However, there is no a priori reason why this should be so, and it is thus possible, if the user so wishes, for the program to match molecules on the basis of the arrangement of some specific atomic property (or properties) in 3-D space. Specifically, atoms can be described in terms of their hydrogen-bonding or their partial-charge characteristics, with each atom in a molecule being assigned an integer class number signifying the particular value of the chosen characteristic that is associated with that atom; the use of property values for characterizing atoms in 3-D substructure searching has been described by Guner et al.[15]

The hydrogen-bonding characteristics of the atoms are defined in terms of the following five classes: (i) neither an H-bond donor nor acceptor, but electronegative; (ii) neither an H-bond donor nor acceptor, not electronegative; (iii) H-bond donor only; (iv) H-bond acceptor only; and (v) both an H-bond donor and acceptor. The rules by which these classes are defined have been described by Pepperrell et al.;[13] for simplicity, only nitrogen and oxygen are considered to be capable of forming H-bonds, but the basic procedure can be readily extended to include other elemental types.

The atomic partial charges are calculated by the MOPAC program using the MNDO semiempirical molecular orbital method.[16] Five classes are defined in terms of the charge range, as follows: (i) <–0.4; (ii) –0.4 to –0.1; (iii) –0.1 to 0.1; (iv) 0.1 to 0.4; and (v) >0.4. The charges are calculated for each molecule in the database or for a target structure, and each atom (including hydrogen in this case, since the hydrogen atoms in a molecule can make a considerable contribution to its charge properties) is allocated the class integer corresponding to its charge; these integers are then used as the atomic descriptors in the atom mapping program. An entirely comparable approach is used for the assignment of the hydrogen-bonding class integers.

The options that have been described so far all characterize each atom in a molecule in terms of a single parameter, i.e., its elemental class, its hydrogen-bonding class, or its partial-charge class. Molecular recognition is determined by several factors, and thus the final option involves a multivariate characterization of the atoms (including hydrogen) in the molecules that are being searched. Three atomic properties are used, these being the hydrogen-bonding class, the partial-charge class, and the van der Waals radius (though the approach can be extended to include any type of property that is thought to be important in the context of the activity of interest). The first two properties are obtained as described previously: the number of charge types is reduced to three; while just two hydrogen-bonding categories are identified, these being atoms that can be involved in hydrogen bonds and all other atom types. The van der Waals radii are obtained from information provided within the SYBYL molecular-modeling system,[17] and these radii are then used to define three size classes: small, medium, and large. Each atom can hence be characterized by a variable XYZ, where X is the charge class (1, 2, or 3), Y is the size class (1, 2, or 3), and Z is the hydrogen-bonding class (1 or 2). Thus, an atom of class 222 would have a neutral charge, be in the medium-size category, and be incapable of taking part in hydrogen bonds. There is a total of 18 different classes (111, 112, 121, 122, ... 332) as shown in Table II.

It is also possible to specify weights when the elemental, hydrogen-bonding, or partial-charge classes are used. This option allows the user to designate certain atoms as being of greater importance than others; thus, a greater degree of importance might be assigned to hydrogen-bonding classes iii, iv, and v (i.e., to those atoms that can form H-bonds) than to the other two classes. If very high weights are assigned to some atoms, it is possible to obtain an output in which the top-ranked structures will contain all (or most) of the highly-weighted atoms, so that one can execute a form of ranked 3-D substructure search.

The inclusion of physical-property data and/or weighting involves simple modifications to the similarity calculation in the atom mapping algorithm described in section 2.2 and to the upperbound calculations described in section 2.4. The availability of a range of search options allows the user to focus the output of a similarity search for a target molecule on those structural features that are thought to be of most

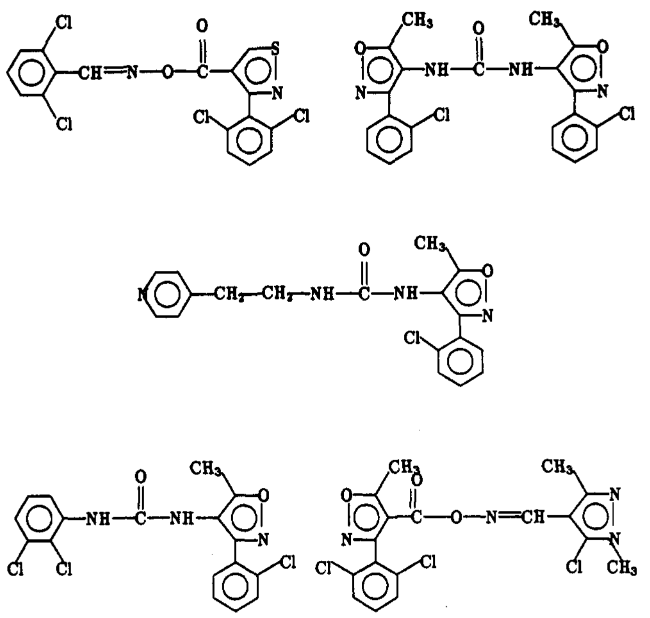

Figure 2. Five nearest neighbors for the target structure shown in Figure 1 in an atom mapping search using elemental types.

Table II. Parameter Classes Used for Multivariate Characterizations of Atoms in Atom Mapping Method

property	class	range
charge	1	<-0.1
	2	-0.1 to 0.1
	3	>0.1
van der Waals radius	1	<1.3
	2	1.3-1.6
	3	>1.6
hydrogen-bonding ability	1	yes
	2	no

importance in determining the similarity between the available target molecule (or molecules) and each of the structures in the search file. It is thus possible to retrieve very different sets of nearest neighbors for a single target molecule, depending upon the parameters that are specified at the start of the search. This behavior is exemplified by the structures shown in Figure 3. These are the nearest neighbors for the target structure shown in Figure 1 when a 3-D similarity search is carried out using the combined physicochemical properties (rather than using the elemental types as in Figure 2). It will be seen that the most similar structure, i.e., the first nearest neighbor, is the same in both cases; however, the other structures are different from each other. Further examples of atom-based and property-based 3-D similarity searches are discussed by Pepperrell et al.[12]

4. PARALLEL IMPLEMENTATION OF ATOM MAPPING METHOD

4.1. Distributed Array Processor. While it is reasonably fast in operation (as discussed in section 2.4), our present implementation of the atom mapping method is too slow for interactive searching. This suggests the use of parallel computer hardware to increase searching rates, and we are currently evaluating the use of the Distributed Array Processor (DAP) for this application. The DAP is a massively-parallel array processor, which is manufactured by Active Memory Technology and which has been the subject of extensive previous study in this department for the searching and clustering of both textual and chemical databases.[18]

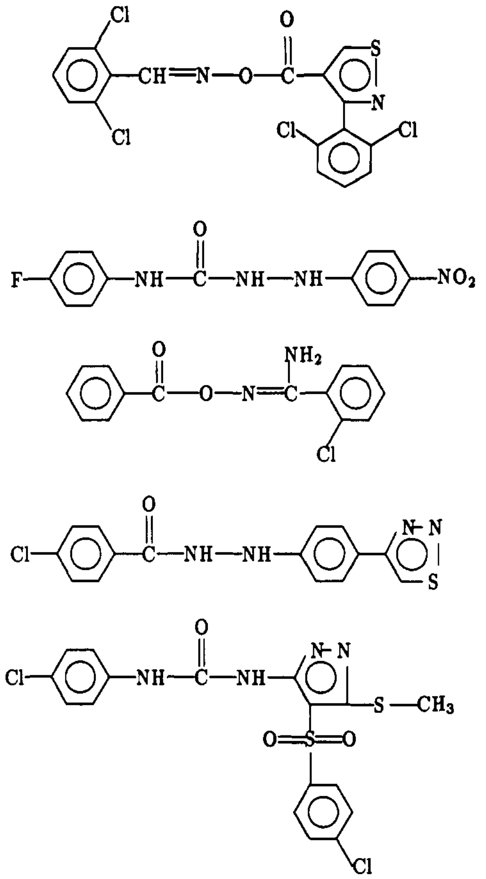

Figure 3. Five nearest neighbors for the target structure shown in Figure 1 in an atom mapping search using multivariate physicochemical property data.

The DAP is an example of an *array processor*, a type of computer in which there is a 2-D synchronous array of simple processing elements, or PEs, under the supervision of a single master control unit, or MCU, which broadcasts instructions for execution in the processor array. This type of machine architecture is commonly referred to as a *single instruction stream, multiple data stream* or SIMD computer[19] with all of the bit-serial PEs executing the same sequence of instructions in parallel. Our experiments have used a DAP 610 at the Centre for Parallel Computing, Queen Mary and Westfield College, University of London; this machine is hosted by a VAX workstation and contains a 64 × 64 array of PEs, each of which has 8 KB of local storage, giving 32 MB overall.

There are two main ways in which an algorithm can be implemented on an array processor. The simpler of the two approaches is to exploit what is variously referred to as *database parallelism* or *outer-loop parallelism*. This involves the application of the basic sequential algorithm to different database records at the same time, i.e., the database is distributed across the available PEs. A search is then effected by loading the query record into the array processor's MCU and broadcasting the attributes that have been specified in the query for comparison with the sets of attributes characterizing the records stored in each of the PEs. The alternative approach, which is referred to as *algorithm parallelism* or *inner-loop parallelism*, involves the identification of any inherent parallelism in the algorithm that is to be implemented, so that different parts of it can be executed concurrently. On an SIMD machine, this involves applying the same operations to (subsets of) the data structures involved in the algorithm; as we shall see, we have identified two different types of inner-loop parallelism that can be exploited when the atom mapping method is implemented on the DAP (or any massively-parallel processor).

The inner-loop and outer-loop approaches to parallel database searching are well illustrated by a recent study in which we have implemented a 2-D substructure searching algorithm on the DAP,[20] and we are now evaluating their relative merits for the implementation of the atom mapping method. The calculation of the interatomic similarities is the most time-consuming part of the atom mapping method, and our attempts at parallelization have thus focused on this part of the method (although we have also been able to make some improvements in the sorting procedures that are used to identify the matching pairs of atoms in the second part of the atom mapping method). It is possible to visualize three levels at which the similarity calculation can be parallelized:

1. *Distributed Molecules.* This is a simple outer-loop implementation in which each PE is assigned one of the database structures, and the similarity calculation is done sequentially on large numbers of database structures in parallel. The remaining two approaches are examples of inner-loop implementations, in which the calculation of the Tanimoto coefficient is carried out in parallel.
2. *Distributed Atoms.* Here, each PE is assigned one atom, so that each database structure is distributed over several PEs. The similarity is calculated for each atom in the target structure in turn, with each such atom being matched against all of the atoms in a database structure in parallel.
3. *Distributed Distances.* Here, each PE is assigned one distance, so that the distances associated with each atom are distributed over several PEs (and thus a single database structure over many PEs). The similarity is calculated for each distance in the target structure in turn, with it being matched against all of the database distances in parallel.

It must be emphasized that these three levels of parallelism are not completely distinct, since it is possible to exploit molecular parallelism at the distributed atoms level. This is achieved by processing the atoms from several database structures together (since there are generally significantly more PEs available than atoms in a single database structure). In a similar way, we can exploit both molecular and atomic parallelism at the distributed distances level, with the distances for several (or all) of the atoms in one molecule (and the distances of further molecules) being processed in parallel.

4.2. Implementation of Methods. 4.2.1. Distributed Distances. On a serial processor, the distances associated with each atom are kept in sorted order, so that the matching distances for a pair of atoms (one in the target structure and one in a database structure) can be identified using an efficient sort–merge algorithm.[12] This is done to ensure that the maximum number of common distances is identified. To see why this is so, consider the following unsorted list of distances associated with an atom from a four-atom target structure:

0.00, 2.73, 2.90, 3.10

where the distances are in angstroms, where the first distance is that from the atom to itself, and where (for simplicity) we have assumed that all of the atoms are of the same elemental type. Consider also the corresponding list from a five-atom database structure:

0.00, 2.79, 2.33, 3.05, 4.60

Assume that a match has been obtained for the second target distance, 2.73 Å, with the second database distance, 2.79 Å, using a tolerance of ±0.5 Å: this would mean that the latter distance could not be matched with the third target distance, 2.90 Å. However, if the second target distance were to be mapped to the third database distance, 2.33 Å (which is also within the allowed tolerance of ±0.5 Å), this would allow the third target distance to be mapped to the second database distance, thereby giving a larger number of matching distances. The maximum number of matching distances is obtained only if ordered lists are used.

It is difficult to produce an equivalent parallel algorithm for the distributed distances approach for two reasons:

1. A simple 'one-step' parallel operation cannot be performed as each target distance may need to be tested against more than one database distance. This is overcome by carrying out a total of N operations (for a database structure containing N atoms), shifting the list of database distances by one place each time.
2. A target distance is compared with all of the database distances in parallel; the ordering of the database distances is thus lost, and it is accordingly not possible to ensure that the maximum number of matching distances is obtained.

The second problem is a serious one, since it means that the distributed distances method is not guaranteed to give the same results as our serial programs; in practice, several heuristics are used to minimize this problem, and the rankings produced by the serial and parallel implementations are very similar to each other.

The distances were arranged on the DAP's array of PEs, so that each of the 64 rows of PEs contained the distances for one atom (with dummy values stored for molecules that contained less than 64 atoms). Although an upper limit of 64 atoms per molecule is assumed, it is easy for the algorithm to be adapted for a greater number of atoms, as the DAP allows the implicit use of *virtual PEs*, where a single PE carries out processing that would normally require some, or many, PEs. The arrangement of data on the DAP can be represented by a parallel array in which the ijth PE contains the ith distance for the jth atom; an analogous array was used to store the target distances.

4.2.2. Distributed Atoms. The distributed atoms method involves allocating one of the 64 rows of PEs to each molecule; within each such row, the interatomic distances associated with a particular atom are stored in the corresponding PE. The distribution of data can thus be visualized as a three-dimensional array in which the ijkth data location contains the kth distance for the jth atom in the ith molecule.

Some types of a massively-parallel processor would allow a sort–merge-type algorithm to be used, by keeping track of the current 'positions' in the target and database lists of distances and by keeping track of the numbers of mappings that had been identified at each point. However, the indexing facilities required for such an implementation are not available on the DAP; the need for such facilities in database-searching applications is discussed in detail by Willett and Rasmussen.[18] Instead, an exhaustive comparison had to be used in which every target distance was compared with every database distance, with all of the database atoms being processed in parallel. This approach avoids the need for the expensive shifting operations used in the distributed distances method and ensures that the same set of mappings is obtained as with the serial algorithm.

4.2.3. Distributed Molecules. This is a simple outer-loop implementation in which the basic serial algorithm is executed on 4096 structures in parallel.

The matching of the distances is carried out as in the distributed atoms method, except that it must be carried out for each atom of a database structure in turn (since all of the atoms of each such structure are stored on the same PE, instead of being distributed across the PEs). The allocation of data here may be visualized as a four-dimensional array, in which the *ijkl*th location contains the *l*th distance for the *k*th atom of the *ij*th molecule (where the first two dimensions correspond to the 64 × 64 array of PEs). Thus all of the 4096 database structures represented by these first two dimensions should be processed in parallel, while the third and fourth dimensions should be processed sequentially.

Unfortunately, it was not possible to implement the distributed molecules algorithm. This was because the PEs on the DAP that was available to us were too small to store the interatomic distance information associated with each of the 4096 molecules that were to be processed in parallel; there were also problems associated with the packing of this information to allow it to be transferred into the DAP from the VAX workstation host.

4.3. Comparison of Algorithms. The distributed distances and distributed atoms algorithms were run on a file of structures provided by ICI Agrochemicals, with several different target molecules containing between 6 and 31 non-hydrogen atoms.

We have adopted the common policy of measuring the performance of a massively-parallel algorithm by the *speed-up*, *S*, which is given by

$$S = \frac{T_{\text{Serial}}}{T_{\text{Parallel}}}$$

where T_{Serial} and T_{Parallel} are the execution times of the most efficient sequential algorithm implemented on a sequential processor and of the algorithm under investigation on the chosen parallel processor, respectively.[21] The T_{Serial} times here were those obtained for the calculation of the similarities on an ESV 3 UNIX workstation, using the upperbound sequential algorithm that has been described in section 2.4.

The mean *S* values obtained in our initial experiments for the distributed distances and distributed atoms algorithms were 7.4 and 39.6, respectively. The former algorithm is thus noticeably slower; moreover, as noted previously, it is not guaranteed to give the same results as the sequential algorithm. The results for the distributed atoms algorithm is comparable with the best that have been obtained in our previous studies of the use of the DAP for database processing applications.[18] However, it must be emphasized that, thus far, this work has considered only the calculation of the interatomic similarities and has assumed that all of the necessary data is stored within the array of PEs. An operational implementation would require the rapid refilling of the DAP with new data as each subset of a database is matched against the target structure; unfortunately, while such high-speed disk units are available for the DAP, the machine that we have been using has only limited communication links to backing storage and to the host processor, making it impossible for us to evaluate a full system at present. Even so, our initial results suggest that the DAP provides an appropriate architecture for the implementation of the atom mapping method, and we are now hoping to test this conclusion on other types of massively-parallel processor.

5. DEVELOPMENT OF ATOM MAPPING METHOD

In addition to the work on parallel implementations of 3-D similarity searching, we are currently studying two developments of the basic method. These are the development of bit string representations of 3-D structure that could be searched far more efficiently than the interatomic distance matrices that are required for the atom mapping method and of methods for searching for patterns of field values rather than for patterns of atoms as in all of our previous work on 3-D searching.

5.1. Use of Bit String Representations. Bit strings denoting the presence or absence of fragment substructures provide an effective means for screening both 2-D and 3-D substructure searches, and they are also used in 2-D similarity-searching systems. However, Pepperrell and Willett's comparison[9] of different 3-D similarity methods showed that a 3-D equivalent of this, in which the similarity between a pair of molecules was determined by the number of interatomic distances that they had in common, gave far less effective rankings than did the atom mapping method.

The reason for the poor performance is that a molecule containing $N(A)$ non-hydrogen atoms will have order $O(N(A))$ screens assigned to it in a 2-D substructure-searching system; however, the same molecule will have $O(N(A)^2)$ distance screens associated with it in a 3-D substructure-searching system. Accordingly, the fact that two molecules have an interatomic distance in common in a 3-D similarity-searching system says much less about their overall similarity than does their sharing of a common topological fragment in a 2-D similarity-searching system. Useful measures of similarity are only obtained in the 3-D case when a more complex matching operation is used that takes account of the environments in which these common fragments occur, so that the atom mapping method is based on sets of distances associated with individual atoms rather than the individual distances. Another limitation of the use of individual distances is that they are known to exhibit markedly nonequifrequent frequency distributions. These are generally characterized by a broad maximum, which then tails away to insignificance with increasing distances, with only a few distances greater than 20 Å. In addition to the broad maximum, there are usually one or more large, and fairly sharply defined, peak(s) at relatively short interatomic distances; these sometimes being accompanied by other, less well-marked peaks at longer distances. Examples of such distributions are given by Jakes and Willett.[22] The irregular nature of the distributions means that matches on some distances count for much more than matches on others, i.e., that a form of implicit, frequency-based weighting has been used in the similarity calculation.

These reasons have led us to investigate bit string representations of 3-D structure that

1. Involve the setting of far less bits than in the case of interatomic distance screens
2. Take more account of the overall geometry of the molecule than in a representation based solely on individual distances
3. Are based on geometric characteristics with less disparate frequency distributions than interatomic distances.

We are now studying combined angle and distance descriptors that are derived from the work of Bartlett et al. on the CAVEAT program[23] and of Poirrette et al.[24,25] on the selection of screens for angle-based substructure searching. Given a set of four atoms A, B, C and D, in which A is bonded to B and C is bonded to D, but in which B is not bonded to C, then the fragments we have chosen to use comprise three components:

1. The torsion angle between the vectors **A–B** and **C–D**, as viewed along the vector **B–C**

2. The valence angles ABC and BCD, these being combined into a single value
3. The distance between the atoms B and C, subject to a minimum-distance constraint that avoids the overspecification of very commonly-occurring substructural moieties, such as six-membered rings and that also reduces the total numbers of bits that are set[25]

This three-part representation of the structure of a 3-D molecule meets all of the criteria listed above. The three values, when rounded, are used as indices to a three-dimensional bit–matrix that characterizes the overall geometry of a molecule. The similarity between a pair of molecules is then measured by a Tanimoto coefficient that is calculated from the numbers of bits that the corresponding bit matrices have in common. We are now evaluating the effectiveness of the measure using the structure–activity datasets that were used for the comparison of similarity methods reported in section 2.

5.2. Field-Based Similarity Searching. The work described in section 3 has shown that it is possible to characterize atoms by property values rather than by elemental types, for 3-D similarity searching. However, the development of property-prediction methods such as comparative molecular field analysis[26] has demonstrated that biological activity is determined, *in* large part, by the distribution of property values across the elements of a 3-D grid around a molecule, rather than by the distribution of these values across that molecule's constituent atoms. There is thus a need for *field searching*, where we wish to identify molecules that are similar to a target structure in the way that some property (e.g., electrostatic potential or hydrophobicity) is distributed in 3-D space.

There have already been several studies of field-based searching,[27–29] but the approaches that have been described to date are all extremely demanding of computational resources and inappropriate for a database-searching system that must be able to match a target structure against many tens, or hundreds, of thousands of structures. We are thus studying techniques that will provide an approximate, but rapid, mapping of one 3-D structure onto another. A simple modification of the atom mapping method described in this paper provides an obvious way of implementing such a system, since it could be used to map the points of one grid onto another, with the coarseness of the grid determining the precision of the mapping that was made. The mappings that resulted from this procedure could then be used as the input to one of the more rigorous procedures referenced above.

6. SIMILARITY SEARCHING IN DATABASES OF 3-D MACROMOLECULES

6.1. Representation of 3-D Proteins. The very rapid developments in 3-D structure handling that have taken place over the last few years show clearly the great potential of graph-theoretical methods for the representation and searching of 3-D molecules. An ongoing project at the University of Sheffield is seeking to develop comparable methods for the representation and searching of the 3-D structures in the Brookhaven Protein Data Bank.[30,31] Although some of our work has considered searching for patterns of C$^\alpha$ atoms,[32] we have focused primarily on the geometric arrangement of the *secondary structure* elements of proteins, specifically the α-helix and β-strand elements.

The molecules in conventional chemical information systems are represented by labeled graphs, in the form of 2-D or 3-D connection tables, and we have developed an analogous graph-theoretical representation of the tertiary structure of a protein. The representation makes use of the fact that the α-helix and β-strand are both approximately linear repeating structures, which can hence be described by a vector drawn along their major axes. The set of vectors corresponding to the secondary structure elements in a protein can then be used to describe the structure of that protein in 3-D space, with the secondary structure elements and the inter-secondary structure element angles and distances corresponding to the nodes and to the edges, respectively, of a graph.[33,34] In fact, each edge in such a labeled graph is a three-part data element that contains the angle between a pair of lines, the distance of closest approach, and the distance between their midpoints. Mitchell et al.[34] provide a detailed account of the methods that are used to create the set of graphs that represent the proteins in the Protein Data Bank; an alternative approach to the graph-theoretical description of protein structures has been developed recently by Kaden et al.[35] These graphs can then be processed by graph-matching algorithms that are entirely analogous to those that are used for the processing of small 2-D and 3-D molecules. Our initial studies of protein searching resulted in a program, called POSSUM, that allows searches to be carried out for secondary-structure motifs, i.e., patterns of α-helices and/or β-strands in 3-D space, using a subgraph isomorphism algorithm.[33,34,36,37] We now describe a more recent program, called PROTEP (the PROtein Topographic Exploration Program), that allows similarity searches to be carried out in the Protein Data Bank, using a maximal common subgraph isomorphism algorithm.[37,38]

6.2. Use of Bron–Kerbosch Algorithm. The maximal common subgraph procedure that we have used in this work is based on the clique-detection algorithm of Bron and Kerbosch,[39] where a clique is a subgraph of a graph in which every node is connected to every other node and which is not contained in any larger subgraph with this property.

Given a pair of graphs A and B, a *correspondence graph*, C, can be formed by the following process:

1. Create the set of all pairs of nodes, one from each of the two graphs, such that the nodes of each pair are of the same type.
2. Form the graph C whose nodes are the pairs from step 1. Two nodes $[A(i),B(x)]$, $[A(j),B(y)]$ are marked as being connected in C if the values of the arcs from $A(i)$ to $A(j)$ and $B(x)$ to $B(y)$ are the same.
3. Maximal common subgraphs then correspond to the cliques of the correspondence graph. The relationship between clique detection and maximal common subgraph isomorphism seems to have been first noted by Barrow and Burstall.[40]

The identification of the MCS for a pair of 3-D chemical structures is hence equivalent to the identification of the largest clique in the correspondence graph linking together the two structures; this is, in its turn, equivalent to identifying the largest possible overlap of one structure on the other, given the user-defined geometric tolerances. The maximal common subgraph isomorphism thus defines a local similarity measure, as discussed in section 2.2, since it is derived from the secondary structure elements that are common to the two proteins that are being compared.

Clique detection was first studied in Sheffield in the context of finding the maximal substructure common to pairs of small 3-D molecules. An evaluation of the clique-detection approach[41] showed that it was far faster in operation than an alternative, breadth-first search procedure that had been described previously for this application by Crandell and Smith,[42] and a comparison of different clique-detection algorithms showed that the most efficient in this application domain was the algorithm of Bron and Kerbosch.[39] A subsequent paper discussed the use of this algorithm for the implementation of similarity searching in files of 3-D small molecules,[14] and it has also been used in ligand-binding studies by Crippen and co-workers.[43,44]

The Bron–Kerbosch algorithm operates by means of a back-tracking tree search. At each level, D, of the tree search there are two sets, $N(D)$ and $C(D)$, of nodes of the graph which are connected to every node in the set $M(D)$, which consists of the D nodes under consideration for inclusion in the next potential clique. $N(D)$ contains the nodes which have already been tried in the attempt to enlarge $M(D)$, and $C(D)$ contains those candidate nodes which have yet to be tried. The algorithm moves to the next level of the tree search by moving a candidate node from $C(D)$ to the trial set $M(D)$, which then becomes $M(D + 1)$. The sets $N(D + 1)$ and $C(D + 1)$ are then calculated by removing from $C(D)$ and $N(D)$ those nodes not connected to the candidate node. When back-tracking occurs, the node most recently added to $M(D + 1)$ is added to $N(D)$ and removed from $C(D)$, and the level of the search becomes D, from its previous value of $D + 1$. A clique is found when both $C(D)$ and $N(D)$ are empty; if only $C(D)$ is empty, then $M(D)$ is a subset of a clique which has already been output. A search-ordering heuristic is used to increase the likelihood that the largest cliques will be identified as rapidly as possible during the tree search. This involves selecting a 'best' candidate node from $C(D)$, at each level D, that is connected to the greatest number of other nodes in $C(D)$. Subsequently at each level D, further candidate nodes are selected that are not connected to the initial 'best' candidate, which has, by now, been transferred to $N(D)$. Brint and Willett[41] give a worked example of the use of this algorithm when it is used for clique detection in 3-D small molecules.

PROTEP uses the Bron–Kerbosch algorithm to identify patterns of α-helices and β-strands in 3-D space that are common to a pair of proteins. The algorithm can be used to identify just the largest common pattern (and hence the term 'maximal common subgraph isomorphism'); however, this is identified only as the result of an exhaustive tree search that identifies all structural patterns that are common to the two graphs that are being compared. PROTEP can thus identify all of the common patterns that are larger than some user-defined threshold size. This restriction is imposed to ensure that the user is not overwhelmed with a very large output consisting primarily of small common substructures that are of little structural significance: this effect has been noted previously in studies of the use of MCS algorithms with small 3-D molecules.[41] The precision of the matches that are identified is determined by the tolerances that are used. The angular tolerance is specified in terms of numbers of degrees, but the distance tolerances are specified either in angstroms or as a percentage of the distance in the target structure. Three types of distance match are possible: closest approach distances (CAD), midpoint distances (MD), or both types of distance (BD). In the experiments reported here, runs were carried out in which the angular tolerance was varied between 10° and 50° in 10° intervals, and the distance tolerance was varied between 10% and 50% in 10% intervals and also between 1 and 4 Å in 1-Å intervals. The user thus has a very large degree of control over the number and the quality of the matches that are identified by the program.

PROTEP has been extensively tested using sets of structurally-related proteins and has been shown to be capable of identifying correctly the structural features that they have in common. A wide range of searches is discussed by Grindley et al.;[38] in the next section, we discuss another example, this being a series of searches involving proteins that contain the dinucleotide-binding fold. These searches were carried out on the 371 labeled graphs that resulted from the use of our database-creation programs on the April 1989 release of the Protein Data Bank.

7. PROTEP SEARCHES FOR NAD-BINDING FOLD MOTIF

7.1. Specification of Motif. The dinucleotide-binding fold is an important and well-defined structural and functional motif within a group of proteins known generally as the oxidoreductases, which include dehydrogenases, reductases, hydroxylases, etc.[45] The April 1989 release of the Protein Data Bank contained four NAD-linked dehydrogenase structures that contained the characteristic Rossmann fold for their NAD-binding domains: lactate dehydrogenase (LDH), *s*-malate dehydrogenase (MDH), D-glyceraldehyde-3-phosphate dehydrogenase (GAPDH), and liver alcohol dehydrogenase (ADH). In addition, work in Sheffield had produced a preliminary structure for the NAD-linked glutamtae dehydrogenase, which possesses a related nucleotide-binding domain. The dinucleotide-binding fold consists of a central core of parallel β-sheet whose strands are interconnected invariably by α-helices, and is made up of two distinct structural moieties: the adenine-binding part and the nicotinamide-binding part. Both of these moieties are composed of the sequence of secondary structures β–α–β–x–β, which is recognizable as a form of the classic Rossmann fold; in general, the x structure may be one or more α-helices. The two moieties are joined such that if the six strands are labeled a–f, then the folding sequence of strands in the sheet is cbadef. A connecting α-helix (helix B) between strands a and b is also spatially conserved in all of the five NAD-dependent proteins mentioned above.[45] Baker et al.[46] noted that a second helix (helix X) also appears to be common to all five proteins. This helix is positioned and oriented very similarly in space in all five cases, but may come from a noncontiguous part of the polypeptide chain. In addition to the NAD-dependent proteins, there are also three well-documented examples of NADP-dependent oxidoreductases: dihydrofolate reductase (DHFR), glutathione reductase (GTHR), and *p*-hydroxybenzoate hydroxylase (PHBH). These contain one or more dinucleotide-binding domains including NADP- and FAD-binding domains.

In all, the search file used in our work contained 13 examples of NAD-dependent dehydrogenases: lactate dehydrogenase (1LDX, 3LDH, 4LDH, 5LDH), liver alcohol dehydrogenase (4ADH, 5ADH, 6ADH, 7ADH), *s*-malate dehydrogenase (2MDH), and D-glyceraldehyde-3-phosphate dehydrogenase (1GD1, 1GPD, 2GPD, 3GPD). There are also two examples each of the NADP-dependent dehydrogenases dihydrofolate reductase (2DFR, 3DFR) and glutathione reductase (2GRS, 3GRS) and one of *p*-hydroxybenzoate hydroxylase (1PHH). The target motif that was used in this work was taken from the dinucleotide-binding fold in lactate dehydrogenase (4LDH). The motif consists of the five β-strands a–e and the two helices B and X described previously. The motif is illustrated in Figure 4. Searches were carried out with the angular and distance tolerances varying as described previously and with a minimum clique size of six secondary structure elements. Two such sets of runs were done; in the first case, runs were specified to be sequence ordered, and in the second case sequence order was specified as not important. Searches for this motif using POSSUM, our subgraph-isomorphism program, have been reported previously by Mitchell et al.[34]

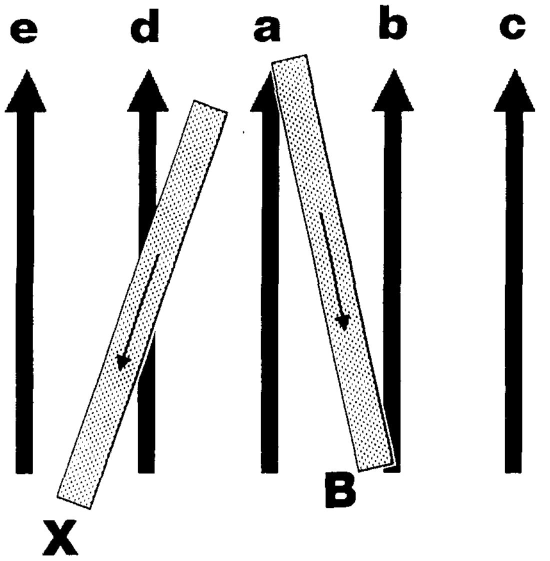

Figure 4. Diagrammatic representation of the dinucleotide-binding motif from lactate dehydrogenase (4LDH) that was used as the target for the PROTEP searches. The five parallel β-strands are labeled a–e, and the two α-helices are labeled B and X.

7.2. Sequence-Ordered Runs. For this particular target motif, the various search types, i.e., CAD, MD, and BD runs for percent and angstrom distance tolerance, all give similar numbers of cliques. The results of the runs are conveniently displayed by means of a 3-D bar chart, such as that shown in Figure 5a, where the total numbers of cliques found in the runs for each set of tolerances are indicated by the height of the relevant bar in the bar chart. This figure summarizes the results that were obtained in the CAD searches where the tolerances were expressed as a percentage of the distances specified in the target.

In general, on ascending from lowest to highest tolerances, the order in which members of groups of proteins first occur is dehydrogenases, ATCases, glutathione reductase, *p*-hydroxybenzoate hydroxylase, tyrosyl-tRNA synthetase/flavodoxin/galactose-binding protein, carboxypeptidases, kinases, and xylose isomerase/Ef Tu-binding protein. This ordering applies to all of the distance types used, so that hits are being found in the 'expected' oxidoreductase proteins at lower tolerances than in other proteins.

Cliques occur in nine of the 13 NAD-dependent dehydrogenase proteins at various tolerance levels throughout the whole set of runs. The correct fold cliques in these nine proteins were usually of size 7, but with some of size 6 (where a 'correct' fold implies the presence in a database protein of the correct five strands a–e and the two helices B and X or a subset of these if the clique contains less than seven secondary structure elements). The four dehydrogenases for which matches were not obtained are all members of the ADH protein group (i.e., 4ADH, 5ADH, 6ADH, and 7ADH). There are two rather different factors that account for the absence of these four structures, as described below.

Examination of the four structures reveals that the position along the polypeptide chain of the helix equivalent to helix X in the target motif is different. Instead of this helix being in the sequence after the stretch containing helix B and the five strands a–e, it is in a part of the sequence that is at the beginning of the polypeptide chain and thus prior to helix B and to the five strands a–e. Hence a match of all seven secondary structure elements including helix X will not be found in a sequence-ordered search: however, if this was the only factor affecting the searches, we would still expect to detect six-membered cliques for these proteins. That this is not so is due to what we have called the *crossover problem*, which arises from the way in which the intervector torsion angles are calculated for pairs of secondary structure elements that are nearly coplanar. The angles between helix B and strand a in the target motif and in 5ADH are −163° and 146°, respectively. Thus, while the absolute values of the two angles are very similar, the signs are different, which means that one of the elements in the target structure is (just) on the opposite side of the plane from that which it occupies in the database structure. When a search is carried out for the target motif, PROTEP will attempt to match secondary structure elements in the target motif and in a database structure by checking to see whether the associated angles and distances are equal to within the search tolerances. Angles in PROTEP (and also in POSSUM) are measured in the range −180° to 180°; the difference between the two angles is thus 360° − (146° − 163°), i.e., 51°. This is greater than even the largest tolerance used here, which was 50°, with the result that it was never possible to match these pairs of secondary structure elements. Checks have now been built into PROTEP (and also into POSSUM) to identify those occasions where sign reversal could lead to the nondetection of such structural equivalences.

Up to 22 correct cliques were found in some runs. These included all the size 6 and size 7 cliques in the NAD-dependent dehydrogenases (as mentioned in the previous paragraph) and also included correctly-located matches with the target motif in three of the NADP-dependent dehydrogenases, these being 2GRS, 3GRS, and 1PHH. These correct fold cliques were examined and were found to superimpose well with the target motif in terms of position and orientation of the matched secondary structure elements. 'Partially correct' cliques were found in the oxidoreductases. These are cliques in which a subset of the matched secondary structures in the target motif and database structure is in equivalent positions in the dinucleotide-binding fold. However, one or more of the strands, for instance, in the target motif may be matched to a neighboring strand in the β-sheet to that to which it would normally be adjacent; these cliques do not appear in anything like large numbers, except at the highest tolerances of 50° and 50% or 4Å.

The remainder of the cliques found in this set of runs occurred in the following non-oxidoreductase proteins: ATCases (4ATC, 5ATC), flavodoxin (1FX1), galactose-binding protein (1GBP), carboxypeptidases (1CPB, 3CPA, 4CPA, 5CPA), kinases (1HKG, 3PGK), tyrosyl-tRNA synthetase (1TS1), Ef TU-binding protein (1ETU), and xylose isomerase (1XIA). All of these proteins appeared as just one clique except for the carboxypeptidases (3CPA, 4CPA, 5CPA) and ATCase (4ATC), which had a total of two cliques each over the whole set of runs. Inspection of these non-oxidoreductase proteins showed that the cliques in both 4ATC and 5ATC of size 6 were good matches with the target in terms of position and orientation. Aspartate transcarbamylase contains regulatory subunits binding for CTP and ATP which are mononucleotides, thus suggesting a possible reason for the matches. The size 6 clique in 1FX1 contained four strands and two helices and was also a good match with the exception of a strand equivalent to strand c in the target motif; this finding is in line with previous work by Rao and Rossmann,[47] who have described finding a similarity between parts of lactate dehydrogenase and flavodoxin in studies of the nucleotide-binding fold. A clique of size 6 was located in 3PGK, which included the five strands of the sheet and one helix equivalent to helix B and superimposed well onto the corresponding target

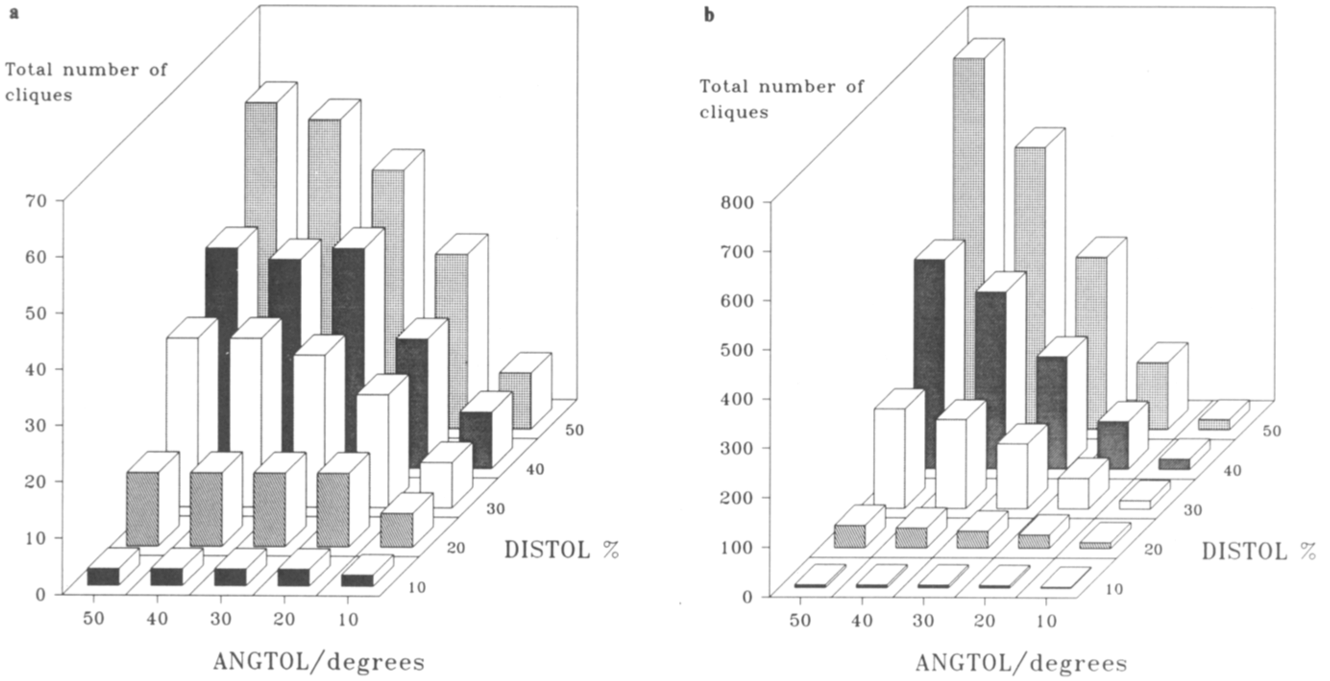

Figure 5. Total numbers of cliques containing at least six secondary structure elements found in PROTEP searches for the dinucleotide-binding fold. The results correspond to CAD runs where the distance tolerance is specified as a percentage and where (a) the sequence order is specified and (b) the sequence order is not specified. The numbers attached to each bar denote the number of cliques that were located using each combination of search parameters.

secondary structures. The five strands and one helix are known to form a dinucleotide-binding type fold,[48] and the substrate ATP for 3PGK binds at the mononucleotide-binding fold comprising the strands cab and helix B. The cliques in the carboxypeptidases are all of size 6, the common missing element being a match with strand c in the target motif. There is a strand at this point in the carboxypeptidases, but it is not considered to be equivalent as it is not in the correct sequence order as in the target motif: otherwise the orientation and positioning of this match is quite convincing.

The remaining non-oxidoreductase proteins did not superimpose as well in terms of orientation and position of secondary structures. However, the clique found in 1TS1 contained the secondary structure elements referred to Webster et al.,[49] which appear to form a Rossmann mononucleotide-binding fold, in common with the dinucleotide-binding oxidoreductases. Similarly, although the match in terms of orientation and position of secondary structure elements was not so good in 1ETU, LaCour et al.[50] describe a β-sheet core surrounded by interconnecting helices that make up the GDP/GTP dinucleotide-binding site in this protein. The cliques in both 1HKG and 1XIA did not superimpose at all well onto the set of matched secondary structure elements in the target.

7.3. Nonsequence-Ordered Runs. For comparison with the results above, nonsequence-ordered searches were carried out, with the threshold clique size again being set to six. Hardly surprisingly, this led to a much higher number of cliques being located, as can be seen from the heights of the bars of the example bar chart in Figure 5b. Table III lists the types of protein that were identified in the sequence-ordered and nonsequence-ordered searches. It will be seen that very similar types of proteins are being found as hits, with the addition of the subtilisins, some additional kinases, and a few 'other'[Hl] proteins (i.e., proteins that do not obviously belong to any of the other families of proteins that are listed in the table).

Cliques with the liver alcohol dehydrogenases (4ADH, 5ADH, 6ADH, 7ADH) and dihydrofolate reductases (3DFR, 4DFR) are located in this set of runs, since the sequence ordering of the secondary structure elements in these proteins differs from that of the target motif. The ADH proteins are found in all six types of run and first appear at very low tolerances. However, the DHFR proteins only occur in CAD/percent and angstrom runs and first appear at moderate tolerances; inspection of these cliques showed that the target was reasonably well-matched with DHFR, although the orientation of some of the strands at the edge of the sheet was poor.

A clique in adenylate kinase (2ADK, 3ADK) was found at quite low tolerances, and this appeared to superimpose well in terms of orientation and position of the matched secondary structures. Liang et al.[51] and Walker et al.[52] describe regions of secondary structure in adenylate kinase that form nucleotide-binding sites for ATP and AMP, and those described by Walker et al. are present in the cliques found in this set of runs. Walker et al. also describe comparable regions in phosphofructokinase, and matches with this protein (1PFK, 2PFK, 3PFK, 4PFK, 5PFK) were found as hits throughout the set of runs. Subtilisins were also identified, in line with the similarity between parts of lactate dehydrogenase and subtilisin that has been noted by Rao and Rossmann in their studies of nucleotide-binding folds.[47]

7.4. Search Times. It will be seen from the search results above that PROTEP provides a highly effective means of identifying structural resemblances between the tertiary structures of proteins. It is also efficient in operation, with the run times being little affected by the particular search parameters that are used. Taking the sequence-ordered CAD searches as an example, with the angular tolerance varied from 10° to 50° in steps of 10° and with the distance tolerance varied from 10% to 50% in steps of 10% (a total of 25 distinct searches), the run times were all in the range 517–536 CPU on a MicroVax 3600; the run times with modern Evans and Sutherland or Silicon Graphics workstation equipment are about one-fourth of this.

Table III. Comparison of Types of Proteins Found in Sequence-Ordered and Nonsequence-Ordered Runs of PROTEP Using Dinucleotide-Binding Motif

protein family	sequence ordered	nonsequence ordered
dehydrogenases	1GD1, 1GPD, 2GPD, 3GPD	1GD1, 1GPD, 2GPD, 3GPD
	1LDX, 3LDH, 4LDH, 5LDH	1LDX, 3LDH, 4LDH, 5LDH
	2MDH	2MDH
		4ADH, 5ADH, 6ADH, 7ADH
	1PHH	1PHH
	2GRS, 3GRS	2GRS, 3GRS
		3DFR, 4DFR
aspartate transcarbamylase	4ATC, 5ATC	4ATC, 5ATC
tyrosyl-tRNA synthetase	1TS1	1TS1
flavodoxins	1FX1	1FX1, 3FXN, 4FXN
carboxypeptidases	1CPB, 3CPA, 4CPA, 5CPA	1CPB, 3CPA, 4CPA, 5CPA
kinases	1HKG	1HKG
	3PGK	2PGK, 3PGK
		2ADK, 3ADK
		1PFK, 2PFK, 3PFK, 4PFK, 5PFK
		1PYK
subtilisins		1CSE, 1SBC, 1SBT, 1SEC, 1SIC, 1SNI, 2SBT
isomerases	1XIA	1XIA, 2XIA, 1PGI, 1TIM
periplasmic-binding proteins	1GBP	1GBP, 1ABP
others	1ETU	1ETU, 1EFM
		1AAT, 1ENL, 1SRX, 1WSY, 2PRK, 2TAA, 3PGM

8. CONCLUSIONS

In this paper, we have described the use of two algorithms, atom mapping and clique detection, for similarity searching in databases of small 3-D molecules and of 3-D macromolecules. Both of the similarity measures studied here are local measures in that they specify not just the numbers but also the locations and the identities of the substructural features common to a pair of substances. We believe that such measures are more appropriate to 3-D structural information than the global measures that are routinely used in 2-D similarity-searching systems.[3]

Our extensive experiments with files of 3-D structures for which activity data are available[9] have demonstrated the strong relationships that exist between the structural similarities resulting from use of the atom mapping method and the corresponding molecules' biological activities. Having demonstrated its effectiveness, we believe that the atom mapping method has several other characteristics that make it well-suited to similarity searching in corporate databases. It is robust in operation, giving results that are not crucially dependent on the precise way that the method is implemented; it is flexible, in that the atoms can be characterized in several different ways as dictated by the needs of an individual search; and it is sufficiently fast in operation to allow searches to be carried out on files of nontrivial size. In the longer term, we expect that atom-based searching procedures of the sort discussed in this paper will be complemented, or even supplanted, by procedures that involve the matching of molecular fields, and we are beginning to investigate this possibility. However, such methods can be expected to be highly demanding of computational resources, and there will thus be a continuing need for ways of increasing search speeds, e.g., by using parallel hardware or by using an initial, bit string screening search (as discussed in section 5.1).

Turning to macromolecular similarity searching, the nucleotide results, and others that have been reported elsewhere,[37,38] indicate that PROTEP correctly identifies the features common to sets of proteins that belong to the same families of proteins. Moreover, as is illustrated by some of the matches obtained in the NAD searches, PROTEP also identifies substructural equivalences between proteins that are not related. This is a particularly valuable characteristic of the program, and one that has already been used, for example, to demonstrate a striking structural resemblance between leucine aminopeptidase (1LAP) and carboxypeptidase A (5CP).[53] The matching substructure here contains no less than eight β-strands in a sheet and five of the 10 α-helices in the C-terminal domain of leucine aminopeptidase; this strong 3-D resemblance is quite unapparent if the two proteins are compared at the sequence lvel, where the degree of homology is only 7% even in the matched region. Another study in our laboratory[38] has confirmed the suggestion of Vriend and Sander[54] that there is a close resemblance between ubiquitin (1UBQ) and ferredoxin (3FXC). In fact, searches with 1UBQ as the target also identified a size-5 match with photosynthetic reaction center protein (1PRC), with four strands from the twisted, curved β-sheet and the single flanking α-helix in ubiquitin being equivalenced to a very similar fold in 1PRC; two of the strands in this match were linked by a short connecting loop that was positioned and oriented very similarly in the two proteins. At present, it is not possible to represent and to search for such loops, and it is clearly desirable to extend PROTEP (and our protein substructure searching program, POSSUM[34]) to encompass this type of structural feature. In addition, we hope to include other types of features such as turns, binding sites, and disulfide bridges and to integrate searching at the secondary-structure level with searching at the residue level, using either sequence data or the C$^{\alpha}$ coordinates. In the longer term, it may also be possible to integrate our graph-theoretical methods with some of the other approaches that have been suggested for the searching of 3-D protein structures.[55]

Similarity searching in databases of 3-D structures is still at a very early stage, but we believe that it will prove to be of great importance in areas such as the discovery of novel, biologically-active molecules and the investigation of protein structure–function relationships. In fact, its immediate potential is probably greater in the case of 3-D macromolecules, owing to our current lack of knowledge about the structural relationships that exist between proteins and owing to the fact that there are, as yet, only a few structural motifs that are sufficiently well-defined to allow substructure searches to be carried out; in the small-molecule case, conversely, the last few years have seen a rapid increase in the numbers of pharmacophoric patterns that are known, and new ones are being reported constantly, e.g., the pattern for nonpeptidic inhibition of HIV-1 protease that has been identified recently by Bures et al.[56]

ACKNOWLEDGMENT

We thank the Cambridge Crystallographic Data Centre, ICI Agrochemicals, the Science and Engineering Research Council, the Wellcome Trust, and Tripos Associates for funding and the Centre for Parallel Computing at Queen Mary and Westfield College, University of London, for the provision of DAP facilities. ICI Agrochemicals in the United Kingdom is part of Imperial Chemical Industries PLC. P.J.A. is a Royal Society University Research Fellow, and D.W.R. is a Lister Institute Research Fellow. The Krebs Institute for Biomolecular Research is a designated Centre for Molecular Recognition Studies under the Molecular Recognition Initiative of the Science and Engineering Research Council.

REFERENCES AND NOTES

(1) Ash, J. E.; Warr, W. A.; Willett, P., Eds. *Chemical Structure Systems*; Ellis Horwood: Chichester, 1991.

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

(3) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, U.K., 1987.

(4) Johnson, M. A. A Review and Examination of the Mathematical Spaces Underlying Molecular Similarity Analysis. *J. Math. Chem.* **1989**, *3*, 117–45.

(5) Cohen, N. C.; Blaney, J. M.; Humblet, C.; Gund, P.; Barry, D. C. Molecular Modeling Software and Methods for Medicinal Chemistry. *J. Med. Chem.* **1990**, *33*, 883–894.

(6) Martin, Y. C.; Bures, M. G.; Willett, P. Searching Databases of Three-Dimensional Structures. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1990; pp 213–260.

(7) Willett, P. *Three-Dimensional Chemical Structure Handling*; Research Studies Press: Taunton, 1991.

(8) van Geerestein, V.; Perry, N. C.; Grootenhuis, P. D. J.; Haasnoot, C. A. G. 3D Database Searching on the Basis of Ligand Shape Using the SPERM Prototype Method. *Tetrahedron Comput. Methodol.* **1990**, *3*, 595–613.

(9) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 455–474.

(10) Rusinko, A.; Skell, J. M.; Balducci, R.; McGarity, C. M.; Pearlman, R. S. *CONCORD: a Program for the Rapid Generation of High Quality Approximate 3-Dimensional Molecular Structures*; University of Texas at Austin and Tripos Associates: St. Louis, MO, 1988.

(11) Christie, B. D.; Henry, D. R.; Guner, O. F.; Moock, T. E. MACCS-3D: a Tool for Three-Dimensional Drug Design. In *Proceedings of the 14th International Online Information Meeting*; Raitt, D., Ed.; Learned Information: Oxford, 1990; pp 137–161.

(12) Pepperrell, C. A.; Taylor, R.; Willett, P. Implementation and Use of an Atom Mapping Procedure for Similarity Searching in Databases of 3-D Chemical Structures. *Tetrahedron Comput. Method.* **1990**, *3*, 575–593.

(13) Pepperrell, C. A.; Poirrette, A. R.; Willett, P.; Taylor, R. Development of an Atom Mapping Procedure for Similarity Searching in Databases of Three-Dimensional Chemical Structures. *Pestic. Sci.* **1991**, *33*, 97–111.

(14) Brint, A. T.; Willett, P. Upperbound Procedures for the Identification of Similar Three-Dimensional Chemical Structures. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 311–320.

(15) Guner, O. F.; Hughes, D. W.; DuMont, L. M. An Integrated Approach to Three-Dimensional Information Management with MACCS-3D. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 408–414.

(16) Stewart, J. P. MOPAC: a Semiempirical Molecular Orbital Program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–105.

(17) SYBYL is produced by Tripos Associates, St Louis, MO.

(18) Willett, P.; Rasmussen, E. M. *Parallel Database Processing: Text Retrieval and Cluster Analysis Using the DAP*; Pitman: London, 1990.

(19) Hockney, R. W.; Jesshope, C. R. *Parallel Computers 2. Architecture, Programming and Algorithms*; Adam Hilger: Bristol, 1988.

(20) Willett, P.; Wilson, T.; Reddaway, S. F. Atom-by-Atom Searching Using Massive Parallelism. Implementation of the Ullmann Subgraph Isomorphism Algorithm on the Distributed Array Processor. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 225–233.

(21) Parkinson, D.; Liddell, H. M. The Measurement of Performance on a Highly Parallel System. *IEEE Trans. Comput.* **1983**, *C-32*, 32–37.

(22) Jakes, S. E.; Willett, P. Pharmacophoric Pattern Matching in Files of 3-D Chemical Structures: Selection of Inter-Atomic Distance Screens. *J. Mol. Graphics* **1986**, *4*, 12–20.

(23) Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Waterman, S. CAVEAT: a Program to Facilitate the Structure-Derived Design of Biologically Active Molecules. In *Molecular Recognition: Chemical and Biochemical Problems*; Roberts, S. M. Ed.; Royal Society of Chemistry: Cambridge, 1989; pp 182–196.

(24) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Characterization and Use of Generalized Valence Angle Screens. *J. Mol. Graphics* **1991**, *9*, 203–217.

(25) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Characterization and Use of Generalized Torsion Angle Screens. *J. Mol. Graphics*, in press.

(26) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959–5967.

(27) Burt, C.; Richards, W. G.; Huxley, P. The Application of Molecular Similarity Calculations. *J. Comput. Chem.* **1990**, *11*, 1139–1146.

(28) Hermann, R. B.; Herron, D. K. OVID and SUPER: Two Overlap Programs for Drug Design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 511–524.

(29) Manaut, F.; Sanz, F.; Jose, J.; Milesi, M. Automatic Search for Maximum Similarity between Molecular Electrostatic Potential Distributions. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 371–380.

(30) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. (1987) Protein Data Bank. In *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*; Allen, F. H., Bergeroff, G., Sievers, R., Eds.; Data Commision of the International Union of Crystallography: Cambridge, 1987; pp 107–132.

(31) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F., Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, M.; Tasumi, M. The Protein Data Bank: a Computer-Based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535–542.

(32) Brint, A. T.; Davies, H. M.; Mitchell, E. M.; Willett, P. Rapid Geometric Searching in Protein Structures. *J. Mol. Graphics* **1989**, *7*, 48–53.

(33) Artymiuk, P. J.; Mitchell, E. M.; Rice, D. W.; Willett, P. Searching Techniques for Databases of Protein Secondary Structures. *J. Inf. Sci.* **1989**, *15*, 287–298.

(34) Mitchell, E. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins. *J. Mol. Biol.* **1990**, *212*, 151–166.

(35) Kaden, F.; Koch, I.; Selbig, J. Knowledge-Based Prediction of Protein Structures. *J. Theor. Biol.* **1990**, *147*, 85–100.

(36) Artymiuk, P. J.; Rice, D. W.; Mitchell, E. M.; Willett, P. Structural Resemblance between the Families of Bacterial Signal Transduction Proteins and of G Proteins Revealed by Graph Theoretical Techniques. *Protein Eng.* **1990**, *4*, 39–43.

(37) Artymiuk, P. J.; Grindley, H. M.; Rice, D. W.; Ujah, E. C.; Willett, P. Searching Techniques for the Tertiary Structures of Proteins in the Protein Data Bank. In *Proceedings of the 1991 International Chemical Information Conference*; Collier, H., Ed.; Calne: Infonortics, 1991; pp 91–106.

(38) Grindley, H. M.; Artymiuk, P. J.; Rice, D. W.; Willett, P. Identification of Secondary-Structure Resemblances in Proteins Using a Maximal Common Subgraph Isomorphism Algorithm. *J. Mol. Biol.*, in press.

(39) Bron, C.; Kerbosch, J. Algorithm 457. Finding all Cliques of an Undirected Graph. *Commun. ACM* **1973**, *16*, 575–577.

(40) Barrow, H. G.; Burstall, R. M. Subgraph Isomorphism, Matching Relational Structures and Maximal Cliques. *Inf. Process. Lett.* **1976**, *4*, 83–84.

(41) Brint, A. T.; Willett, P. Algorithms for the Identification of Three-Dimensional Maximal Common Substructures. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 152–158.

(42) Crandell, C. W.; Smith, D. H. Computer-Assisted Examination of Compounds for Common Three-Dimensional Substructures. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 186–197.

(43) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A combinatorial algorithm for calculating ligand binding. *J. Comput. Chem.* **1984**, *5*, 24–34.

(44) Smellie, A. S.; Crippen, G. M.; Richards, W. G. Fast drug-receptor mapping by site-directed distances: a novel method of predicting new pharmacological leads. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 386–392.

(45) Birktoft, J. J.; Banaszaik, L. J. Structure–Function Relationship Among Nicotinamide-Adenine Dinucleotide Dependent Oxidoreductases. In *Peptide and Protein Reviews*; Hearn, M. T. W., Ed.; Dekker: New York, 1984; pp 61–101.

(46) Baker, P. J.; Farrants, G. W.; Rice, D. W.; Stillman, T. J. Recent Progress on the Structure and Function of Glutamate Dehydrogenase. *Biochem. Soc. Trans.* **1987**, *15*, 748–751.

(47) Rao, S. T.; Rossmann, M. G. Comparison of Super-Secondary Structures in Proteins. *J. Mol. Biol.* **1973**, *76*, 241–256.

(48) Blake, C. C. F.; Rice, D. W. Phosphoglycerate Kinase. *Philos. Trans. R. Soc. London* **1981**, *A293*, 93–104.

(49) Webster, T. A.; Lathrop, R. H.; Smith, T. F. Prediction of a Common Structural Domain in Aminoacyl-tRNA Synthetases through Use of a new Pattern-Directed Inference System. *Biochemistry* **1987**, *26*, 6950–6957.

(50) LaCour, T. F. M.; Nyborg, J.; Thirup, S.; Clark, B. F. C. Structural Details of the Binding of Guanosine Diphosphate to Elongation Factor TU from *Escherichia coli* as Studied by X-ray Crystallography. *EMBO J.* **1985**, *4*, 2385–2388.

(51) Liang, P.; Phillips, G. N.; Glaser, M. Assignment of the Nucleotide Binding Sites and the Mechanism of Substrate Inhibition of its *Escherichia coli* Adenylate Kinase. *Proteins: Struct., Funct. Genet.* **1991**, *9*, 28–36.

(52) Walker, J. E.; Saraste, M.; Runswick, M. J.; Gay, N. J. Distantly Related Sequences in the α- and β-Subunits of ATP Synthase, Myosin, Kinases and other ATP-Requiring Enzymes and a Common Nucleotide-Binding Fold. *EMBO J.* **1982**, *1*, 945–951.

(53) Artymiuk, P. J.; Grindley, H. M.; Park, J. E.; Rice, D. W.; Willett, P. Three-Dimensional Structural Resemblance between Leucine Aminopeptidase and Carboxypeptidase A Revealed by Graph-Theoretical Techniques. *FEBS Lett.* **1992**, *303*, 48–52.

(54) Vriend, G.; Sander, C. Detection of Common Three-Dimensional Substructures in Proteins. *Proteins: Struct., Funct. Genet.* **1991**, *11*, 52–58.

(55) Thornton, J. M.; Gardner, S. P. Protein Motifs and Database Searching. *Trends Biochem. Sci.* **1989**, *14*, 300–304.

(56) Bures, M. G.; Hutchins, C. W.; Maus, M.; Kohlbrenner, W.; Kadam, S.; Erickson, J. W. Using Three-Dimensional Substructure Searching to Identify Novel, Non-Peptidic Inhibitors of HIV-1 Protease. *Tetrahedron Comput. Methodol.*, in press.

Registry No. DH, 9035-82-9.