Journal of Computer-Aided Molecular Design, 11 (1997) 163–174
ESCOM

J-CAMD 387

Similarity searching in files of three-dimensional chemical structures: Representation and searching of molecular electrostatic potentials using field-graphs

David A. Thorner[a], Peter Willett[a,*], P. Matthew Wright[a] and Robin Taylor[b,**]

[a]*Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.*
[b]*Zeneca Agrochemicals, Jealott's Hill Research Station, Bracknell, Berkshire RG12 6EY, U.K.*

Received 1 August 1996
Accepted 23 November 1996

Keywords: Clique-detection algorithm; Database searching; Field-graph; Molecular electrostatic potential; Similarity searching

Summary

This paper reports a method for the identification of those molecules in a database of rigid 3D structures with molecular electrostatic potential (MEP) grids that are most similar to that of a user-defined target molecule. The most important features of an MEP grid are encoded in *field-graphs*, and a target molecule is matched against a database molecule by a comparison of the corresponding field-graphs. The matching is effected using a maximal common subgraph isomorphism algorithm, which provides an alignment of the target molecule's field-graph with those of each of the database molecules in turn. These alignments are used in the second stage of the search algorithm to calculate the intermolecular MEP similarities. Several different ways of generating field-graphs are evaluated, in terms of the effectiveness of the resulting similarity measures and of the associated computational costs. The most appropriate procedure has been implemented in an operational system that searches a corporate database, containing ca. 173 000 3D structures.

Introduction

The last few years have seen substantial interest in the development of systems for the storage and retrieval of three-dimensional (3D) chemical structures. These systems are designed to identify all of those structures in a 3D database that contain a user-defined pharmacophoric pattern, or pharmacophore [1–3]. The first such systems focused on the representation and searching of pharmacophores that consisted of patterns of atoms and interatomic distances, but the techniques have now been extended to encompass pharmacophores that are defined in terms of a much wider range of types of structural feature. Current systems, such as ChemDBS-3D, MACCS-3D or UNITY, allow the specification of ring centroids, lone pairs, lines or planes through user-defined sets of atoms, included and excluded volumes, and the use of angular, as well as distance, constraints. The ability to define a query pharmacophore in great detail and then to carry out a rapid search of a large database has aroused much interest, and there have been several reports of the use of such facilities in drug-discovery programmes [4–7].

Current 3D database systems characterise a molecule by means of its constituent atoms (or aggregates thereof), with the implicit assumption that these also characterise the electrostatic, steric and hydrophobic fields that most strongly affect the ability of a ligand to bind at a biological receptor site. Such an assumption is an approximation at best, and it is thus likely that very different types of structure would be retrieved from a database system in which these crucial determinants of biological activity were described in an explicit manner. The retrieval of such novel structures hence requires the development of methods for the representation and searching of the fields around molecules.

The importance of molecular fields in correlating structure and biological activity is evidenced by the rapid and widespread take-up of Comparative Molecular Field Anal-

*To whom correspondence should be addressed.
**Present address: Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

ysis (CoMFA) [8] and related approaches to the study of structure–activity relationships. However, few of these techniques are directly applicable to large files of disparate structures, such as might be found in a corporate database. In this paper, we describe algorithms and data structures that allow a chemist to input a *target* molecule, e.g. one that has been shown to exhibit activity in a biological test system, and then to retrieve those molecules in a large database that have patterns of surrounding field values that are most similar to the pattern surrounding the target molecule. The work described here considers the molecular electrostatic potential (MEP), but the techniques are applicable, in principle at least, to any field-like attribute that can be represented by real values in a 3D grid surrounding a molecule.

Similarity calculations using molecular electrostatic potentials

Carbo et al. [9] were the first to suggest that the similarity between a pair of molecules could be estimated by a similarity coefficient based on the overlap of the molecules' electron charge clouds. This idea has been taken up by several workers, using either electron densities or, more commonly, MEPs (see e.g. Refs. 10–16). A molecule is positioned at the centre of a 3D grid and the electrostatic potential is calculated at each point in the grid. The similarity between a pair of molecules is estimated by aligning them in some way so that similar features are superimposed, taking the product of the two molecular potentials at each point and then summing over the entire grid, with a suitable normalising factor being used to bring the similarities into the range −1.0 to +1.0. This numerical approach necessarily involves the matching of very large numbers of grid points and is thus extremely demanding of computational resources; however, Good et al. [17] have reported an alternative approach in which the potential distribution is approximated by a series of Gaussian functions that can be processed analytically, with a substantial increase in the speed of the similarity calculation and with only a minimal effect on its accuracy. This elegant idea removes one of the main limitations of field-based approaches to 3D similarity searching, but still requires the alignment of the two molecules that are being compared prior to the calculation of the similarity.

Field-based similarity searching in large databases will be feasible only with the identification of an appropriate alignment procedure. Exhaustive searching for alignments is computationally infeasible, and even stochastic searching procedures can run for extended periods if one wishes to conduct a detailed examination of the alignment space [18]. The absence of an appropriate alignment procedure has thus meant that field-based processing of large databases is not practicable, unless an initial screening procedure is used to eliminate the great majority of the structures from the time-consuming calculation. Examples of the latter approach have been suggested by Good et al. [15], who consider only the output of a conventional 3D pharmacophoric-pattern search, and by Van Geerestein et al. [19], who consider only the top-ranked molecules resulting from an initial similarity search based on distance information. Unfortunately, such approaches are limited in that the screening procedures use angular and distance information, rather than field information; accordingly, it is possible for a molecule to be eliminated from the second stage of the search even though it has a high degree of similarity with the target structure at the field level. Such retrieval failures are inevitable unless the screening stage also uses field information, and in the remainder of this section we discuss the use of graph theory to provide such a field-based screening procedure.

Chemical database systems use methods of representation and search that were first developed for two-dimensional (2D) chemical substructure searching, where the atoms and bonds of a chemical compound are denoted by the nodes and edges of a labelled graph and where searching is effected by the application of a subgraph isomorphism algorithm to such graph representations [20]. Analogous techniques are used for 3D pharmacophore searching, where the nodes of a graph again denote the atoms of a chemical compound but where the edges denote the interatomic distances (or distance ranges) in a rigid (or flexible) 3D molecule [1,2,21,22]. Graph-based similarity searching methods have also been described, in which a maximal common subgraph (MCS) isomorphism algorithm (rather than a subgraph isomorphism algorithm) is used to find molecules that are similar to a target molecule [23–26].

It is simple to develop analogous graph methods for the processing of grid-based representations of the molecular fields discussed above, since a molecular grid can be represented by a graph in which the nodes correspond to each of the grid points and in which the edges correspond to the interpoint distances. The overlap between two sets of molecular fields, and hence the similarity of the two corresponding molecules, will then be given by the overlap between the two graphs, this being estimated most obviously by their MCS. Unfortunately, such a procedure is totally infeasible owing to the sizes of the graphs that need to be compared. The graph representing a molecule centred in a cubic grid with sides of length 25 Å and with a step-size of 1 Å will contain over 15 000 nodes, whereas the NP-complete nature of MCS detection [27,28] means that current MCS algorithms are capable of processing graphs containing only a few tens of nodes at most. Accordingly, the use of such an algorithm requires a reduction in the size of a grid-based graph by several orders of magnitude. Moreover, the few nodes remaining from such a reduction procedure must still encompass the

main features of the underlying fields if the resulting intermolecular similarities are to be chemically (and hence biologically) meaningful. We do not believe that it is possible to achieve such a drastic reduction whilst maintaining the effectiveness of the calculated similarities. Instead, the procedures we have developed use an MCS procedure for the alignment stage of an MEP calculation, this alignment being carried out using only small numbers of nodes that, hopefully, encompass the main features of the MEP around a molecule. Once the relative positions of the two molecules have been fixed in this way, the full similarity calculation is carried out using the Gaussian approximation procedure described previously. Our procedures hence provide an approach to the generation of molecular alignments that is an alternative, and complementary, to those generated by conventional, atom-based MCS procedures.

The two-stage procedure suggested here requires a mechanism for reducing a full grid, containing a very large number of points, to a simple graph containing (at most) a few tens of nodes that can then be used in the alignment stage. The mechanism that we have developed for the creation of these simple *field-graphs* is presented in the next section.

Generation and use of field-graphs

A field-graph is generated from a set of grid points by identifying a subset of them that have potential values meeting some criterion (the criteria that we have tested are discussed below) and then grouping points that meet the chosen criterion and that are 'close' to each other (in some sense). There are many ways in which a field-graph can be generated from a set of potential values, depending on the criterion that is used and the definition of 'close' that is employed to determine whether two grid points are to be considered as belonging to the same group. There are no obvious guidelines, a priori, as to how field-graphs should be created, and we have accordingly evaluated a range of different procedures for creating nodes. In fact, we have simplified the problem by considering only a single measure of closeness, which is that two grid points, P and Q, are considered as being contained within the same graph node if P and Q are vertically, horizontally or diagonally adjacent to each other. This may be illustrated by considering the set of grid points in Fig. 1 (which is in just two dimensions for simplicity). If the criterion was that a grid-point value was to be at least 10 kcal/mol, then the four boxed elements would be selected as forming a node in the field-graph.

We have evaluated eight criteria that can be applied to the grid-point values to determine whether they are significant and whether the corresponding grid points should, accordingly, contribute to one of the nodes in the field-graph. In all cases, the resulting subset of the original grid points is then grouped using the adjacency criterion described above. In what follows, the reader should note that the methods are described in terms of applying a positive threshold and then reducing it to identify the important positive parts of the MEP; in each case, the inverse procedure was followed to identify the most important negative parts of the MEP.

7	8	9	5
8	11	13	9
4	10	9	7
1	6	10	5
3	2	1	0

Fig. 1. Application of a threshold of +10 to a set of grid-point values to generate a field-graph. In this simple, planar grid, the four values surrounded by bold lines are taken as forming a single field-graph node since they all meet the threshold criterion and each is adjacent to at least one other grid point that meets the threshold criterion. The resulting node is shaded.

(1) The first, and simplest, approach involves thresholding the grid-point values using a fixed threshold potential of 15 kcal/mol. The application of this threshold identifies all of the positive grid points with values greater than or equal to the threshold value.

(2) As criterion (1), but with a threshold of 10 kcal/mol.

(3) Rather than using a fixed threshold value, the third approach starts by setting the threshold level to 30 kcal/mol and then reducing it by 1 kcal/mol towards a value of 3 kcal/mol until at least three nodes have been identified, one of which must be of the opposite sign to the other two.

(4) As criterion (3) but using 20 kcal/mol and 2 kcal/mol as the threshold values.

(5) The fifth approach seeks to maximise the number of nodes in the field-graph representing a structure. The threshold level is progressively lowered from a value of 30 kcal/mol towards 3 kcal/mol and that level is selected for which the number of nodes identified is a maximum. If two (or more) different threshold levels give the same maximum number of nodes, then the set with the higher (or highest) threshold level is chosen. The positive and negative grid-point values are considered separately, so that the threshold level needed to maximise the number of positive nodes is generally different from the level needed to maximise the number of negative nodes.

(6) This approach seeks to maximise the number of nodes in the field-graph that contain at least some minimal number of points. An extensive series of experiments was carried out to determine how this might best be achieved, using a range of relationships between the threshold level, and the number and size of the resulting

nodes. The best results were obtained by initially setting the threshold level at 30 kcal/mol and then progressively reducing it by 1 kcal/mol towards 3 kcal/mol until at least two nodes are present that contain at least $10 + 1.075^{30-CL}$ grid points, where CL is the current threshold level. The minimally acceptable number of points in each node thus increases, the smaller the threshold level that is applied to the grid; this is done to allow for the fact that there are very large numbers of grid points with low MEP values whereas it is the higher values that are of greater importance in determining the degree of similarity between a pair of molecules.

(7) As criterion (6) but with a threshold of 8 kcal/mol.

(8) As criterion (6) but with a step-size of 2 kcal/mol.

The form of a field-graph will also depend on the mechanism that is used to define the geometric relationships (i.e. the edges of the graphs that are to be processed by the MCS algorithm) between pairs of nodes once they have been identified. Here, we have considered only a single, and extremely simple, way of defining these relationships. The location of a node is defined by the geometric coordinates of the centre of the corresponding cluster of points, with an associated label defining the threshold level (either positive or negative) at which it was formed and the size of the node, i.e. the number of grid points that it represents; the distance between two nodes is then simply the distance between the two centres. Thus, a field-graph containing N nodes (i.e. one that contains N clusters of grid points after the application of the clustering procedures described above) will contain $N(N-1)/2$ distinct distances (since the internode distance matrix is completely symmetric), and a database of 3D structures can be represented for search by the corresponding set of field-graphs.

Whichever generation procedure is chosen to characterise a database molecule, the field-graph nodes resulting from its use are input to an appropriate implementation of the Bron–Kerbosch clique-detection algorithm [29] to identify the MCS with the corresponding field-graph representation of the target structure. Given two field-graphs A and B, a pair of nodes a_i and a_j from A are regarded as matching a pair of nodes b_k and b_l , respectively, from B if the distances $D(a_i,a_j)$ and $D(b_k,b_l)$ are the same to within a user-defined tolerance (±1.0 Å in the experiments reported here) and if the matching pairs of nodes $\{a_i,b_k\}$ and $\{a_j,b_l\}$ are of the same type, i.e. positive or negative. Note that it is possible to adopt a more rigorous matching criterion that takes account of the magnitude of the local potential as well as its sign, or of the size of the node, i.e. the number of constituent grid points, or of some combination of the two; however, all of the experiments reported in this paper employed just the simple matching criterion described above. The algorithm identifies the MCS based on such node-to-node equivalences, and the constituent pairs of matching nodes are input to a least-squares fitting routine. This routine aligns the database structure with the target structure so as to minimise the squared distances between the centroids of the field-graph nodes from the database structure with the centroids of the field-graph nodes from the target structure to which they have been mapped. The resulting alignment is then used for the final grid-based similarity calculation.

The Bron–Kerbosch algorithm is designed, and normally used, to identify the largest subgraph common to a pair of graphs. In the present context, however, we have used it to generate all of the subgraphs common to a pair of field-graphs, and not just the largest such subgraph(s). This has been done to increase the number of possible alignments that are considered in the precise similarity calculation, and thus to ensure that the minimum possible number of close neighbours to the target structure are overlooked because of the approximations involved in the generation of the field-graphs. The use of such a procedure will maximise the recall of a database search, but it does mean that some database structures can give rise to a very large number of alignments (and a consequent number of similarity calculations) since it is known that pairs of 3D chemical graphs can have extremely large numbers of small, submaximal subgraphs in common [28]. The computational implications of this point are considered further when we discuss the operational search system that has been implemented at Zeneca Agrochemicals; for the moment, we note merely that the similarity between an individual database structure and the target structure is taken to be the largest of the calculated similarities over all of the possible alignments resulting from the first-level, clique-detection search.

Comparison of field-graph generation methods

Measurement of effectiveness

An important characteristic of research into similarity-based retrieval is the need for some quantitative means of evaluating the effectiveness of the similarity measures that are being tested. Previous studies, both in our laboratories and elsewhere, have made extensive use of the *similar-property principle* of Johnson and Maggiora [30]. Here, simulated property-prediction experiments are carried out using data sets for which both structural and property data are available, so as to ascertain which methods (e.g. which similarity coefficients) result in measures of structural similarity that are most closely correlated with measures of property similarity. We have adopted a rather different approach in the work reported here. The extensive studies of Richards and co-workers (see e.g. Refs. 10, 11, 15 and 17) have shown clearly that there is a strong correlation between biological activity and the similarities that result from grid-based MEP calculations. Accordingly, we can estimate the effectiveness of a par-

TABLE 1
PERFORMANCE OF THE EIGHT DIFFERENT CRITERIA THAT WERE TESTED FOR THE GENERATION OF FIELD-GRAPH NODES

Criterion	E_5	E_{10}	E_{20}
1	0.74	0.71	0.66
2	0.74	0.71	0.68
3	0.70	0.66	0.60
4	0.74	0.71	0.68
5	0.77	0.75	0.73
6	0.74	0.72	0.69
7	0.70	0.67	0.69
8	0.74	0.71	0.68

The figures quoted are mean E_n values averaged over a set of 62 target structures.

ticular field-graph (and hence of the node-identification algorithm that was used to create it) by the magnitude of the similarities resulting from the grid-based calculations, since a high second-stage similarity will be achieved if, and only if, an appropriate alignment (and hence an appropriate graph) has been used in the initial graph-matching stage.

The main performance measure we have used for evaluating the effectiveness of a graph-creation procedure is hence the MEP similarities that are calculated after a target field-graph has been mapped onto each of the field-graphs in the database. Specifically, let S(I) be the grid-based MEP similarity for the Ith most similar molecule to the mth target molecule; then the performance measure E_{mn} is defined by

$$E_{mn} = \frac{1}{n}\sum_{l=1}^{n} S(I) \qquad (1)$$

where typical values for n are 5, 10 or 20, i.e. E_{mn} is the mean MEP similarity for the n nearest-neighbour structures of the mth target structure. If there are a total of T target molecules, then the overall effectiveness of the set of searches for the n nearest neighbours of each target structure is given by

$$E_n = \frac{1}{T}\sum_{m=1}^{T} E_{mn} \qquad (2)$$

The results of using this evaluation criterion with the eight different methods for the creation of field-graphs are discussed in the remainder of this section.

Data sets

The molecules in all of the experiments reported in this section were drawn from the Fine Chemicals Database. The 3D structures of these molecules were generated using CONCORD [31] and then the atomic charges were calculated using the MNDO routines in MOPAC v. 5.00 [32].

The following procedure was adopted to generate the field-graph for a molecule from a CONCORD structure. The structure is read in from backing storage and a grid is constructed around it. This grid extends for 5.0 Å beyond the maximum and minimum extents of the molecule in each plane (xy, xz and yz), as determined by the centres of the atoms on the molecular surface, and has a step-size that is defined by the user (0.5, 1.0 or 2.0 Å in our experiments). A probe is placed at each point on the grid and the MEP is calculated at each such point from the equation

$$332.17\sum_{i=1}^{n}\frac{q_1 q_i}{\varepsilon d_i} \qquad (3)$$

where q_1 is the charge on the probe atom (set to 1.0 here), q_i is the charge on the *i*th of the n atoms in the molecule, ε is the dielectric constant, d_i is the distance between the probe and the *i*th atom and the constant of 332.17 is to give a value in kcal/mol. A grid point is ignored if it falls within the van der Waals radius of any of the atoms in the molecule. The resulting grid is then input to the chosen field-graph generation procedure.

Similarity measure

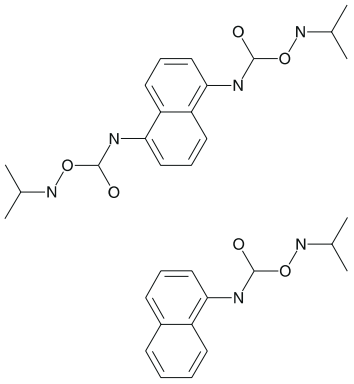
There is extensive literature relating to the coefficients that can be used to measure the degree of similarity between a pair of MEPs [11,33–35]. In the work reported here, we have used the so-called Carbo index, which is in fact identical to the cosine coefficient, an association coefficient that has been used in cluster analysis for many years (see e.g. Ref. 36) and that we have found performs as well as other measures when used for the calculation of field-based similarities [37]. If two molecules, A and B, have been aligned in some way and if P_A and P_B are the MEPs at some aligned point, then the Carbo index is given by

$$\frac{\sum P_A P_B}{\sqrt{\sum P_A^2 \times \sum P_B^2}} \qquad (4)$$

where the summations are over all the points in the grid. In fact, as noted previously, we have used the technique of Good et al. [17] to replace the summations by Gaussian functions, with a huge increase in the speed with which the similarity can be calculated. Good et al. have discussed the use of both two-term and three-term Gaussian functions for this purpose: in our work, we have used the former to minimise the calculation time.

Experimental results

A large number of experiments were carried out to determine the effectiveness of the two-term Gaussian similarities, when compared with the full grid-summation similarities. In one such set of experiments, 25 molecules



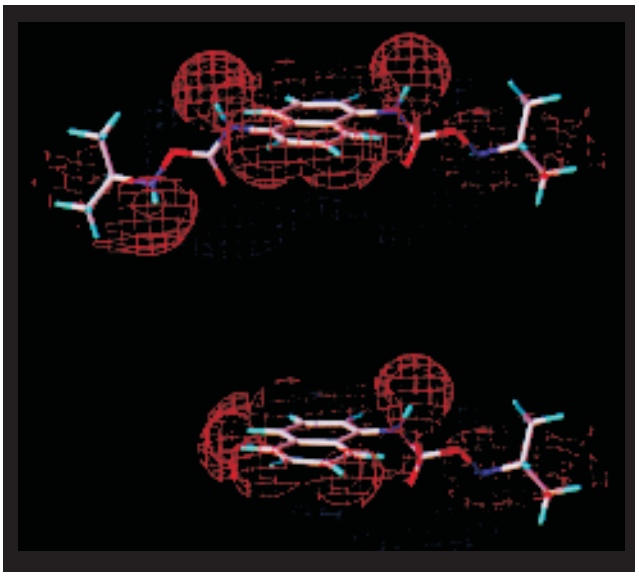
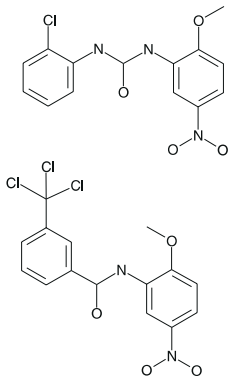


Fig. 2. Target structure and nearest-neighbour structure in a search of 1000 molecules from the Fine Chemicals Database, and the MEP of the corresponding, aligned structures.

were used as the targets for similarity searches of files of about 100 heterogeneous structures, giving a total of 2554 similarities that were calculated using the two approaches. No less than 55.5% of the similarities were within ±0.1 of each other and 83.8% of them were within ±0.2 of each other. Several of these, and other, searches were timed using the two approaches: the Gaussian searches were between 290 and 4058 times as fast as the grid-summation searches. Thus, while a few of the Gaussian similarities were found to be seriously in error, the results are sufficiently good (in terms of both effectiveness and efficiency) for us to have used the two-term Gaussian calculation as the basis for the full grid-based searches in all our work.

The main comparative study used a larger file containing 1000 structures. The field-graphs were generated for these structures using the eight criteria described previously and then searches were carried out using a set of 62 target structures selected at random from the file. Equation 2 defines the performance measure, E_n, that we have used to compare the eight criteria, and Table 1 details the E_n values that were obtained in the searches. It will be seen that the best level of performance, i.e. the largest E_n values, is given using criterion (5), which seeks to maximise the number of nodes in the field-graphs and thus to provide the most discriminating description of an MEP. The magnitudes of the E_n values listed in the table demonstrate that this has, indeed, been achieved. However, the large numbers of nodes that result from this procedure mean that searches based on it are extremely time-consuming, running for 5–6 times longer than searches based on any of the other criteria that we have tested and taking too long for consideration in an operational environment. Of the remaining approaches, the best results are given by criteria (2) and (6); further searches with these two criteria revealed the general superiority of the latter, which involves the identification of the maximum number of nodes that contain at least some minimal number of grid points. Criterion (6) was thus chosen for



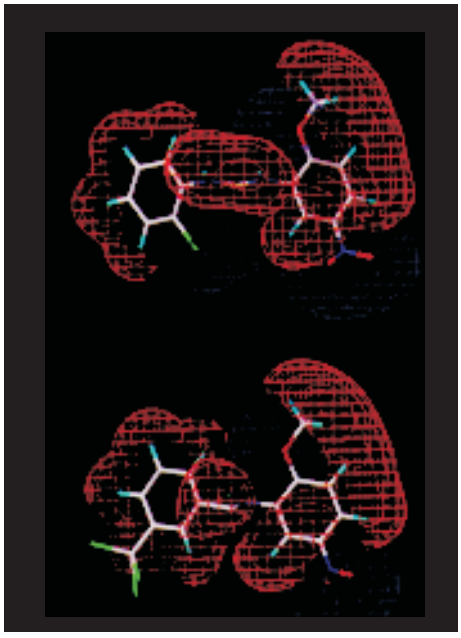
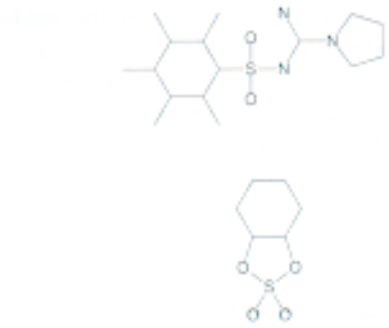


Fig. 3. Target structure and nearest-neighbour structure in a search of 1000 molecules from the Fine Chemicals Database, and the MEP of the corresponding, aligned structures.



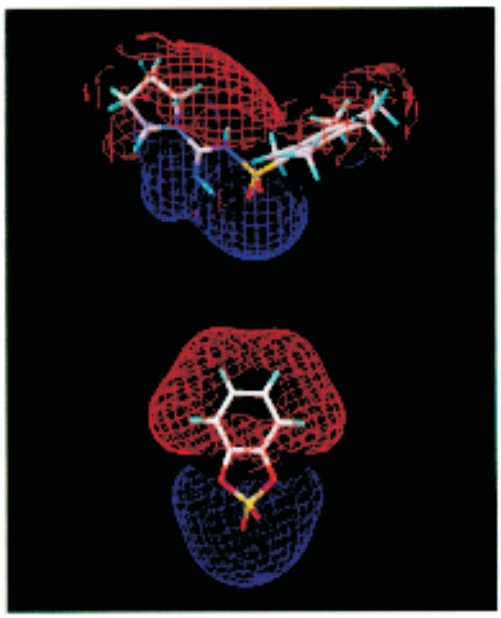
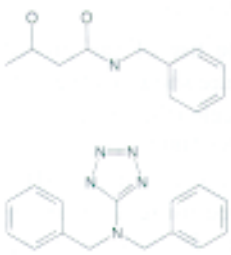


Fig. 4. Target structure and nearest-neighbour structure in a search of 1000 molecules from the Fine Chemicals Database, and the MEP of the corresponding, aligned structures.



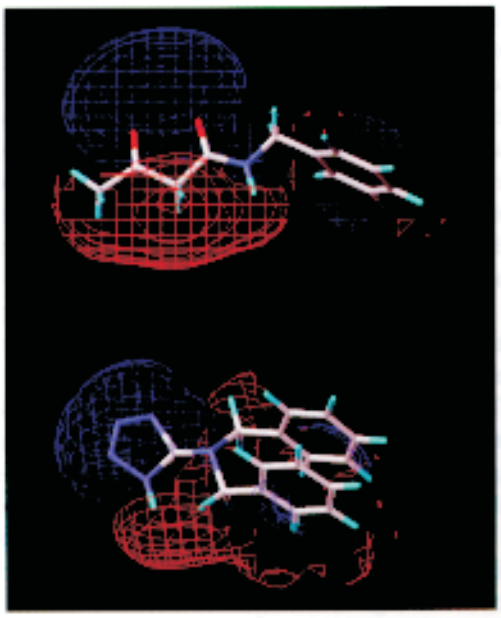
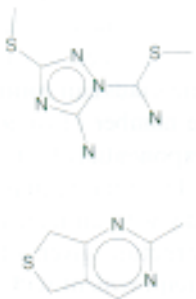


Fig. 5. Target structure and nearest-neighbour structure in a search of 1000 molecules from the Fine Chemicals Database, and the MEP of the corresponding, aligned structures.



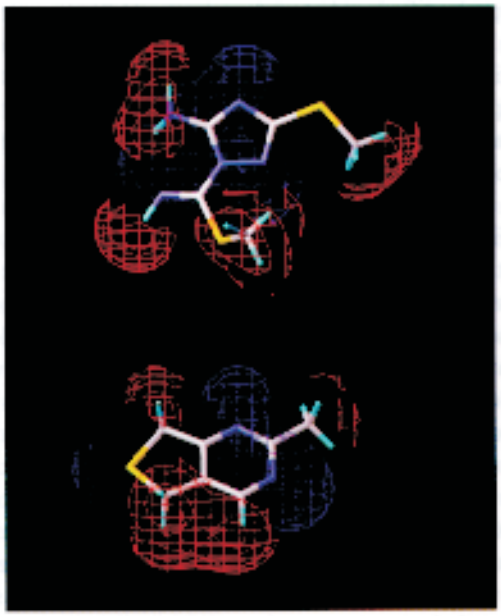


Fig. 6. Target structure and nearest-neighbour structure in a search of 1000 molecules from the Fine Chemicals Database, and the MEP of the corresponding, aligned structures.

the operational implementation that is described in the following section.

Examples of the use of the procedure are shown in Figs. 2–6. Each figure shows (i) the 2D structure diagrams of the target molecule (the upper molecule) and its nearest neighbour in the 1000-structure file mentioned previously; and (ii) the corresponding MEPs contoured at ±1 kcal/mol and aligned using the mapping that results from the MCS procedure. In looking at the MEPs, the reader should remember that the contoured displays may well represent levels that are very different from the variable thresholds that result from the application of criterion (6).

Figures 2 and 3 illustrate the fact that molecules that have very similar 2D structures may also have very similar MEPs, since the nearest neighbours found here would also stand a strong chance of being retrieved in a conventional, fragment-based 2D similarity search. In the first example, the nearest neighbour is a substructure of the target structure, with an excellent overlap of the common structural moiety being obtainable using the MEPs, while in the second example, the strong negative contribution from the CCl_3 group in the nearest neighbour is outweighed by the contributions from the rest of the molecule. The calculated similarities here were 0.72 and 0.84. The other pairs of molecules illustrate cases where the target and its nearest neighbour belong to different structural classes, thus providing a mechanism for the identification of possible bioisosteres that would not be obvious at the level of the 2D or 3D arrangement of the constituent atoms. In Fig. 4, the match of 0.83 is obtained because of the dominant contribution of the SO_2 grouping to the calculated value for the similarity coefficient, even though the remaining parts of the molecules' MEPs have little similarity to each other. The major region of MEP overlap identified in Fig. 5 (similarity of 0.77) is that between the β-diketone in the target structure and a tetrazole ring in the database structure, while the excellent overlap (similarity of 0.88) in Fig. 6 is based in part on the mapping of a triazole ring to a pyrimidine ring.

The concept of the field-graph was developed to provide a compact representation of the principal features of an MEP. An inspection of the field-graph nodes resulting from the application of our node-generation procedure suggests that, in some cases, the nodes merely reflect the presence of hydrogen donors and acceptors in a molecule. We have hence carried out a series of searches in which the nodes that are input to the MCS algorithm are simply the donor and acceptor atoms in a structure, rather than those resulting from the field-graph procedure described above. Specifically, we have characterised the nitrogen and oxygens in a target structure or a database structure as donors, as acceptors, or as both donors and acceptors using the classification scheme developed previously in studies of the atom-mapping method for 3D similarity searching [38]. Searches were then carried out as described previously, but with the coordinates of the field-graph nodes replaced by the coordinates of the donors, acceptors and donors–acceptors. The searches involved a subset containing 47 of the set of 62 target structures used previously. The mean E_n values for these searches are listed in Table 2, where it will be seen that the donor–acceptor field-graphs are substantially inferior to those resulting from criterion (6).

An operational implementation

Data set

The field-graph generation procedure described above was applied to the structures in the Zeneca Agrochemicals corporate database. Specifically, the grid step-size was set to 0.5 Å, the grid-boundary limit size was set to 5.0 Å, and criterion (6) was used for the generation of the field-graph nodes. The processing took about 20 CPU days on an R4000 Silicon Graphics workstation, excluding the very extensive computation associated with the calculation of the atomic partial charges. The mean number of nodes in the resulting set of 173 197 field-graphs was 7.13 (standard deviation of 3.22), with the single, largest graph containing 41 nodes.

Optimisation of search speeds

Preliminary testing showed that the search times for a scan of the entire file were likely to be excessive unless a fair amount of program optimisation was to be carried out. Profiling revealed that the relative computational requirements of the MCS-detection and Gaussian-calculation stages for the matching of the target structure with a database structure depend on the number of possible alignments resulting from the application of the MCS algorithm. Specifically, if only a few possible alignments are identified then the MCS stage takes more time, while the Gaussian calculation takes more time if there are many alignments (since each one of them needs the similarity calculation to be carried out).

Most of the time requirement for the similarity calculation is occasioned by the need to calculate square roots and exponentials. The Gaussian routine was hence modified to minimise the number of square roots and to calculate most of the exponentials by means of a precalculated look-up table. The time requirement for the MCS algorithm rises rapidly with an increase in the number of nodes in the target structure, given a fixed set of database structures, and this requirement will hence be minimised if the target structure contains just a few nodes. A field-graph with only a few nodes will also generate only a small number of alignments, and hence reduce the time requirement of the final similarity calculations. A series of searches was hence carried out in which a threshold was applied to the sizes of the nodes in the field-graph representing the target structure. Specifically, a cutoff was

applied so that only those nodes containing more than n grid points, where n is defined by the user, were considered in the matching of a target structure with a database structure. This procedure will certainly increase the speed of searching, but may also mean that molecules that are, in fact, strongly similar to the target structure in the final similarity calculation will be overlooked since the appropriate alignments are not forthcoming from the MCS stage of the search. The extent of this problem was investigated by selecting 50 pairs of molecules for which the calculated Gaussian similarity was at least 0.70 when the full sets of graph nodes were used.

Field-graphs were created with n set to 1, 2, 3, 5 and 10 (with a value of 0 for n denoting the original field-graphs without the exclusion of any nodes), and the similarities were re-evaluated using the new field-graphs. The number of the similarities that were still ≥0.70 was recorded, as was the mean number of alignments per pair and the mean run-time for the calculation of the similarity for each pair (in CPU seconds for an unoptimised program on an R3000 Silicon Graphics workstation). The results of these experiments are detailed in Table 3, where it will be seen that the run-times and (to a lesser extent) the accuracy of the procedure (as measured by the percentage of the similarities that remain ≥0.70) fall rapidly as n increases. On the basis of these figures, we believe that the accuracy is too low for an acceptable retrieval system if n is greater than 3; accordingly, the alignment stage of the final search system considered only those target field-graph nodes that contained at least three grid points. The chosen value for n could, of course, be changed by a user if this was felt to be desirable in the context of a particular search.

Analogous timing experiments were carried out in which we varied the tolerance (in Å) that was acceptable for a match between two internode distances in the alignment; these variations were found to have less effect on the run-times and on the accuracies of the calculations. The operational system thus takes as its default the tolerance of 1.0 Å used in all of the other experiments reported in this paper.

The modifications to the basic procedure that have been described above, together with use of the maximum level of compiler optimisation for our Fortran 77 programs, resulted in the reduction of the run-time to an acceptable level for operational use of the program. Specifically, a search of a typical target structure against the entire database can be accomplished in about 16 h of elapsed time on an R4000 Silicon Graphics workstation, i.e. in a single overnight run, although large target structures can mean that a search will take a day or even more. It should be noted that some of the molecules in the database can lead to very large numbers of alignments, each of which must then be checked in the subsequent Gaussian calculation: searches with several target structures demonstrated that the search time could be reduced by ca. 40% by ignoring the top-ranked 5% of the database when the structures were ranked in decreasing order of the number of alignments that needed to be processed.

TABLE 2
PERFORMANCE OF FIELD-GRAPHS AND OF NITROGEN AND OXYGEN DONORS, ACCEPTORS AND DONOR–ACCEPTORS

Structural description	E_5	E_{10}	E_{20}
Criterion (6) field-graphs	0.73	0.71	0.68
Donors–acceptors	0.71	0.67	0.62

The figures quoted are mean E_n values averaged over a set of 47 target structures.

TABLE 3
EFFECT OF CHANGES IN THE MINIMAL SIZE OF THE FIELD-GRAPH NODES ON THE CALCULATED SIMILARITIES, ON THE MEAN NUMBERS OF ALIGNMENTS, AND ON THE MEAN ELAPSED TIME PER PAIR OF STRUCTURES

Minimal node size	Percentage of similarities ≥0.70	Mean number of alignments	Mean run-time
0	100	134	4.5
1	76	47	1.6
2	70	38	1.3
3	64	38	1.2
5	58	31	1.0
10	52	12	0.6

The comparison was restricted to pairs of molecules that had similarities ≥0.70 when the full field-graphs were used.

Results

Searches with the system at Zeneca Agrochemicals demonstrate clearly that it often leads to the retrieval of structures with a high degree of novelty that would not be retrieved by conventional similarity-searching techniques (which are based on patterns of atoms, in either 2D or 3D). The system hence provides an effective way of suggesting novel bioisosteres for known active compounds. That said, the focus on just a single type of field means that while the top-ranked structures generally provide a reasonable level of MEP similarity, many of them are of little interest since their steric and/or lipophilic characteristics render them inappropriate for the biological system under investigation, even if the calculated MEP-based similarity is greater than 0.70 (which was found to be the threshold below which potentially interesting structures were unlikely to be identified). The principal value of the system is hence as an ideas-generator that can suggest previously unexplored chemical classes to the chemist requesting the search, albeit at the cost of (in some cases) a low level of search precision.

There are two other obvious limitations to the current system. The first is that suboptimal alignments are gener-

Molecule	Grid-point values
1	0 – 0 + 0 – 0 0
2	0 – 0 + 0 – 0 +
3	– – 0 + 0 – – 0
4	– – 0 + 0 – 0 0
5	0 – 0 + 0 – – 0

Fig. 7. Data set of five molecules, each surrounded by a grid containing just eight elements with values of + (strongly positive), – (strongly negative) or 0 (neither strongly positive nor strongly negative).

ated in some cases, with the result that some molecules that are similar to the target structure are ranked less highly than they should be. This is almost inevitable given the simplicity of the representation that is used. The second is that the similarities often seem to be dominated by a single strongly positive or strongly negative field-graph node, such as a nitro group. If such a node is present in the target structure, then it largely determines the ranking of the database structures that is produced, almost irrespective of the other features that are present (as is exemplified by the match shown in Fig. 4).

Use of average target molecules

Given a single known active molecule, a similarity search is normally carried out to identify other molecules that are structurally related and that are hence expected to exhibit the same activity. However, unless information is available from other sources, it is not possible to say which part of that molecule's field is responsible for the binding, i.e. we have only *global* information about the molecule's field and no *local* information. This is not the case when several molecules have been tested, since we can then compare these molecules to determine which features are responsible for activity or for inactivity (although we have not actually done the latter as yet). Examples of this approach are the use of substructural analysis [39] or CoMFA [8] methods to identify 2D substructural fragments or steric and electrostatic fields, respectively, that are most heavily correlated with activity. Related methods can be used for the generation of putative pharmacophores in both rigid and flexible 3D molecules [6,26,40,41], and we now describe how it is possible to utilise information from more than one active molecule to define a *pseudo-target* molecule that can be used as the basis for a field-based similarity search.

Consider a set of five molecules, each of which has been shown to be active in the biological test system of interest. Assume that their MEPs have been aligned and that each of the aligned grids contains just eight points; assume further that the potential in each such point is assigned one of the following three values, + (strongly positive), – (strongly negative) or 0 (neither strongly positive nor strongly negative). This situation is illustrated in Fig. 7, from which it will be seen that the electrostatic requirement for activity is that there is a strongly positive value in the fourth grid point, and strongly negative values in the second and sixth grid points (since these features occur in all of the molecules in this data set), i.e. the pseudo-target molecule is of the form

? – ? + ? – ? ?

where ? denotes a grid-point value that varies amongst the members of the active set. This pseudo-target provides a much more precise specification of the electrostatic requirements for activity, and may thus be expected to be more effective, i.e. to identify a larger proportion of actives in the output from a similarity search, than if just a single molecule is used as the target structure. In reality, of course, each grid point has an associated real value, and the grid-point values for a pseudo-target structure are obtained by taking the arithmetic mean of the grid-point values of its constituent structures after they have been appropriately aligned with each other. Thus, if the five molecules in Fig. 7 all had values of +10 (–10) in the strongly positive (strongly negative) elements and zero values in the remainder, the pseudo-target would have values

–4 –10 0 +10 0 –10 –4 +2

If a pseudo-target molecule is to be used in a search, then its constituent structures are aligned manually and a grid is constructed to extend 5.0 Å beyond the maximum extent of the largest molecule. The electrostatic potential for each molecule is calculated at each grid point, using the chosen step-size of 0.5 Å, and the corresponding values are summed and then divided by the number of molecules in the pseudo-target to give the average potential at that point. This pseudo-MEP is converted to a field-graph just as if it was an MEP for a single molecule, and the resulting graph then forms the input to the MCS procedure that aligns the pseudo-target with each database structure. Once the alignment has been obtained, the atoms of the constituent molecules are treated as if they belonged to a single molecule for the calculation of the Gaussian similarity. The main limitation of this approach is that there is a much larger number of atoms in a pseudo-target molecule than in a conventional target molecule, and the similarity calculation is thus much more time-consuming.

Conclusions

In this paper, we have presented algorithms and data structures that permit a user to submit a target molecule and then to identify those molecules in a database of single-conformation 3D structures that have the most similar MEPs to the MEP of the target structure. A two-

stage search procedure has been adopted in which an initial graph-matching search is used to align the target molecule with a database molecule prior to the calculation of the actual similarity. The graph matching is based on what we refer to as field-graphs, which summarise the main characteristics of the MEP field around a molecule, and the MEP calculation uses the Gaussian similarities first suggested by Good et al. [17]. We have evaluated several different approaches to the creation of such field-graphs and have implemented the most cost-effective of these in an operational similarity-searching system at Zeneca Agrochemicals.

It is important to emphasise the great degree of approximation that is involved in the search system that has been developed here. Firstly, the small number of nodes in a field-graph can provide only the most cursory description of the full MEP around a 3D structure. It is thus hardly surprising that the alignments resulting from their use are often inferior to those that could be obtained by graphics-based manual overlapping or by a more sophisticated automatic alignment procedure. An alternative graph representation of a molecule's electrostatic characteristics has been described recently by Meurice et al. [42]. Secondly, the similarities are calculated using two-term Gaussian functions, with the result that the similarities are not as accurate as those that would be obtained if a full-grid calculation was carried out. Finally, and most importantly, the work reported here has considered only rigid 3D molecules; the extension of these techniques to the representation and searching of conformationally flexible molecules is discussed by Thorner et al. [43]. Even with these limitations, the examples reported here suggest that our algorithms do permit the identification of molecules in a large database that are electrostatically similar to a user-defined target molecule, even if the structural similarity (as exemplified by 2D or 3D structural descriptors) is low. We thus believe that our techniques provide a novel, and potentially valuable, tool to support the identification of new lead compounds in drug- and pesticide-discovery programmes.

Acknowledgements

We thank the Engineering and Physical Sciences Research Council, the James Black Foundation, the Science and Engineering Research Council and Zeneca Agrochemicals for funding, Tripos Inc. for hardware and software support, Fraser Williams (Scientific Systems) for the Fine Chemicals Database, Anne Mullaley and Harold Cox for assistance with the generation of the Zeneca database, and Anne Mullaley for first suggesting the idea of aligning pairs of molecules by means of the most important parts of the MEP. This paper is a contribution from the Krebs Institute for Biomolecular Research, which is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

References

1 Martin, Y.C., J. Med. Chem., 35 (1992) 2145.
2 Bures, M.G., Martin, Y.C. and Willett, P., Top. Stereochem., 21 (1994) 467.
3 Willett, P., J. Mol. Recog., 8 (1995) 290.
4 Bures, M.G., Hutchins, C.W., Maus, M., Kohlbrenner, W., Kadam, S. and Erikson, J.W., Tetrahedron Comput. Methodol., 3 (1990) 673.
5 Haraki, K.S., Sheridan, R.P., Venkataraghavan, R., Dunn, D.A. and McCulloch, D., Tetrahedron Comput. Methodol., 3 (1990) 565.
6 Bures, M.G., Black-Schaefer, C. and Gardner, G., J. Comput.-Aided Mol. Design, 5 (1991) 323.
7 Milne, G.W.A., Nicklaus, M.C., Driscoll, J.S. and Wang, S., J. Chem. Inf. Comput. Sci., 34 (1994) 1219.
8 Cramer, R.D., Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
9 Carbo, R., Leyda, L. and Arnau, M., Int. J. Quantum Chem., 17 (1980) 1185.
10 Hodgkin, E.E. and Richards, W.G., Int. J. Quantum Chem., Quantum Biol. Symp., 14 (1987) 105.
11 Burt, C., Richards, W.G. and Huxley, P., J. Comput. Chem., 11 (1990) 1139.
12 Hermann, R.B. and Herron, D.K., J. Comput.-Aided Mol. Design, 5 (1991) 511.
13 Manaut, F., Sanz, F., Jose, J. and Milesi, M., J. Comput.-Aided Mol. Design, 5 (1991) 371.
14 Richard, A.M., J. Comput. Chem., 12 (1991) 959.
15 Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Comput.-Aided Mol. Design, 6 (1992) 513.
16 Sanz, F., Manaut, F., Rodriguez, J., Lozoya, E. and Lopez-de-Brinas, E., J. Comput.-Aided Mol. Design, 7 (1993) 337.
17 Good, A.C., Hodgkin, E.E. and Richards, W.G., J. Chem. Inf. Comput. Sci., 32 (1992) 188.
18 Kearsley, S.K. and Smith, G.M., Tetrahedron Comput. Methodol., 3 (1990) 615.
19 Van Geerestein, V., Perry, N.C., Grootenhuis, P.G. and Haasnoot, C.A.G., Tetrahedron Comput. Methodol., 3 (1990) 595.
20 Ash, J.E., Warr, W.A. and Willett, P. (Eds.) Chemical Structure Systems, Ellis Horwood, Chichester, 1991.
21 Willett, P., Three-Dimensional Chemical Structure Handling, Research Studies Press, Taunton, 1991.
22 Clark, D.E., Willett, P. and Kenny, P.W., J. Mol. Graph., 10 (1992) 194.
23 Hagadone, T.R., J. Chem. Inf. Comput. Sci., 32 (1992) 515.
24 Brint, A.T. and Willett, P., J. Comput.-Aided Mol. Design, 2 (1988) 311.
25 Ho, C.M.W. and Marshall, G.R., J. Comput.-Aided Mol. Design, 7 (1993) 3.
26 Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1993) 83.
27 Levi, G., Calcolo, 9 (1972) 341.
28 Brint, A.T. and Willett, P., J. Chem. Inf. Comput. Sci., 27 (1987) 152.
29 Bron, C. and Kerbosch, J., Commun. Assoc. Comput. Machinery, 16 (1973) 575.

30 Johnson, M.A. and Maggiora, G.M. (Eds.) Concepts and Applications of Molecular Similarity, Wiley, New York, NY, U.S.A., 1990.
31 CONCORD is distributed by the University of Texas at Austin, TX, U.S.A. and Tripos Associates, St. Louis, MO, U.S.A.
32 Stewart, J.P., J. Comput.-Aided Mol. Design, 4 (1990) 1.
33 Reynolds, C.A., Burt, C. and Richards, W.G., Quant. Struct.–Act. Relatsh., 11 (1992) 34.
34 Good, A.C., J. Mol. Graph., 10 (1992) 144.
35 Petke, J.D., J. Comput. Chem., 14 (1993) 928.
36 Sneath, P.H.A. and Sokal, R.R., Numerical Taxonomy, Freeman, San Francisco, CA, U.S.A., 1973.
37 Turner, D.B., Willett, P., Ferguson, A.M. and Heritage, T.W., SAR QSAR Environ. Res., 3 (1995) 101.
38 Pepperrell, C.A., Poirrette, A.R., Willett, P. and Taylor, R., Pestic. Sci., 33 (1991) 97.
39 Cramer, R.D., Redl, G. and Berkoff, C.E., J. Med. Chem., 17 (1973) 533.
40 Marshall, G.R., ACS Symp. Ser., 112 (1979) 205.
41 Crandell, C.W. and Smith, D.H., J. Chem. Inf. Comput. Sci., 23 (1983) 186.
42 Meurice, N., Leherte, L., Vercauteren, D.P., Bourguignon, J.-J. and Wermuth, C.G., Helv. Chim. Acta, in press.
43 Thorner, D.A., Wild, D.J., Willett, P. and Wright, P.M., J. Chem. Inf. Comput. Sci., 36 (1996) 900.