Similarity Searching in Files of Three-Dimensional Chemical Structures: Flexible Field-Based Searching of Molecular Electrostatic Potentials

David A. Thorner, David J. Wild, Peter Willett,* and P. Matthew Wright

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield,
Western Bank, Sheffield S10 2TN, UK

Received December 26, 1996[⊗]

Similarity measures based on the molecular electrostatic potential (MEP) have been used to quantify the degree of resemblance between pairs of rigid three-dimensional (3-D) molecules. This paper discusses the effect of including molecular flexibility on the similarities that are calculated using such measures in searches of large 3-D databases. Experiments with a graph-theoretic approach to MEP-based similarity searching demonstrate the limitations of using multiple conformations to represent the variations in an MEP that arise from torsional flexibility. Better results have been obtained with a genetic algorithm (GA) that has been developed for calculating the similarity between the MEPs of pairs of molecules, one or both of which is flexible. Experiments with two small files of flexible 3-D structures show that this GA identifies significantly better alignments than does a comparable GA for searching rigid 3-D molecules.

INTRODUCTION

The development of effective and efficient programs for generating three-dimensional (3-D) structures from two-dimensional (2-D) structure diagrams[1,2] has led to the development of a range of sophisticated systems for 3-D substructure searching.[3] Early systems considered just rigid searching, with little or no account being taken of the torsional rotations that are possible in many small molecules,[4,5] but these approaches were soon supplanted by more sophisticated flexible searching systems, which permit the retrieval of all molecules that can adopt an energetically-feasible conformation that contains a query pharmacophoric pattern.[6−12]

The development of effective and efficient techniques for 3-D substructure searching has spurred the development of analogous techniques for 3-D similarity searching, where the aim is to identify those molecules in a database that are most similar to a user-defined target structure, using some quantitative measure of intermolecular structural similarity.[13,14] Several similarity measures for 3-D similarity searching have already been discussed in the literature,[15−20] and a current project in this department is evaluating the use of similarity measures based on molecular fields.[21−23] Following previous studies by Richards and his co-workers (see, *e.g.*, refs 24−26), the principal focus of our work thus far has been the development of methods for searching databases in which molecules are represented by their molecular electrostatic potentials (or MEPs). Specifically, we have described search methods that use either a genetic algorithm (hereafter a GA)[22] or a graph-theoretic algorithm[23] to match a target structure against each of the structures in a database. Thus far, we have considered only rigid 3-D molecules: in this paper, we discuss the extension of these methods to encompass conformational flexibility. The next section briefly reviews the two algorithms we have developed previously for rigid 3-D similarity searching, and we then discuss a range of experiments that demonstrate the limitations of the graph-theoretic approach when conformationally flexible molecules need to be searched. The fourth section describes the use of the GA approach for flexible searching, demonstrating the improvements in the molecular alignments that can be achieved when compared with a search of a database of rigid structures. Our main conclusions are summarized in the final section. Full details of all of the experiments reported here are presented by Wright.[27]

FIELD-BASED SIMILARITY SEARCHING

The MEP around a molecule can be represented by a 3-D grid, in which the *ijk*th element is the real-number value of the MEP at the location (*i*, *j*, *k*). The similarity between a target structure and a database structure is obtained in two stages: the corresponding grids are aligned so as to maximize the degree of overlap, and the similarity corresponding to this alignment is then calculated using a measure such as the cosine coefficient.[21] The similarity calculation normally involves the systematic comparison of all corresponding pairs of grid-point values, which is extremely time-consuming unless the grid step-size is quite large. However, Good *et al*. have shown that the speed of the similarity calculation can be substantially increased by means of a Gaussian approximation procedure.[25] This yields similarities that are generally only slightly different from those obtained when the full element-by-element comparison is carried out, and we have thus used this approximation in all of the work reported here. The method of Good *et al*. does, however, still requires the initial specification of an appropriate alignment of the MEPs of the molecules that are being compared. The efficient generation of appropriate alignments is, we believe, the most important problem that needs to be faced before field-based similarity searching can be carried out on databases of non-trivial size, and our efforts to date have thus focused on techniques for the generation of such alignments.

Our initial studies of field-based similarity searching adopted a graph-theoretic approach to the representation and searching of MEPs.[23] The basic idea underlying this

* To whom all correspondence should be addressed: Email P.WILLETT@SHEFFIELD.AC.UK.

[⊗] Abstract published in *Advance ACS Abstracts,* April 15, 1996.

approach is that it is possible to summarize the most important parts of an MEP by a much smaller number of points. The resulting set of points is represented by a *field-graph*, in which the points are the nodes of the graph and the interpoint distances are the edges. The field-graphs are generated in two stages. In the first stage, a threshold potential is applied to the 3-D grid representing an MEP to identify those grid-elements that have the largest magnitudes (either positive or negative). In the second stage, a clustering-like procedure is applied to the resulting subset of the original grid-elements to find the connected components that are present. The center of each such component is taken to be one node in the field-graph. The distance is calculated between the centers of all of the components that have been identified, and the resulting set of inter-center distances then forms the edges of the field-graph. Thorner *et al*. describe a variety of ways in which field-graphs can be generated[23] and all of the experiments reported in the next section of this paper use their recommended procedure.

A database is created by applying the graph-generation procedure to all of the constituent structures, and a similarity search is effected by comparing the field-graph representing the target structure with the field-graphs of each of the molecules in the database. Each such comparison is done by means of a maximal common subgraph (MCS) isomorphism algorithm, which identifies the largest subgraph common to the pair of field-graphs. The algorithm used here is a modification of the Bron-Kerbosch clique-detection procedure, which previous studies have shown to be well suited to the comparison of 3-D chemical graphs.[28,29] The MCS resulting from the application of this algorithm to a target-structure field-graph and a database-structure field-graph specifies an alignment of the corresponding MEPs. This alignment enables the calculation of the intermolecular similarity, which is done using the fast Gaussian approximation procedure mentioned previously. The graph-matching can result in the identification of more than one MCS, in which case the similarity calculation is carried out for each possible alignment and the intermolecular similarity is taken to be the largest of the calculated values.

A genetic algorithm, or GA, is a computational problem-solving method that mimics some of the principal characteristics of biological evolution and genetic reproduction.[30−32] GAs have become increasingly popular over the last few years for providing good approximate solutions to a wide range of combinatorial optimization problems. Applications in computational chemistry have already included conformational analysis, substructure searching, *de novo* ligand design, multiple regression, and ligand docking, *inter alia*,[33] and we have recently described a GA for aligning a pair of MEPs.[22] A chromosome in this GA encodes a set of translations and rotations that, when applied to the 3-D coordinates of one molecule, will align its MEP with the MEP of the other molecule, which is considered to be fixed in space. The fitness function for the GA is the similarity value resulting from a Gaussian similarity calculation and the GA hence seeks to identify that alignment which will maximize the value of this calculation.

The experiments we have carried out thus far suggest that the graph-theoretic and GA approaches provide a comparable level of performance in terms of both effectiveness (as denoted by the magnitudes of the similarity coefficients for the alignments generated by the two approaches) and efficiency (with both of them requiring 1−2 CPU s on a medium-level Unix workstation to match a pair of MEPs).[22] The experiments have, however, been limited in that they have assumed that the molecules that are being processed are completely rigid in nature; specifically, our work has used the structures that are produced by the well-known CONCORD program.[34] However, most organic molecules contain one or more rotatable bonds, thus allowing the molecule to exist in some, or very many, different conformations, and it is hence necessary to consider how torsional flexibility will affect MEP-based similarity searching.

The MEP at any point in 3-D space is a function of the partial atomic charges and the distances of each atom from that point. A full treatment of the effect of conformational flexibility on MEP-based similarity searching would take account of the changes in both the atom-to-point distances and the partial atomic charges that can occur as a molecule flexes.[35] The calculation of the partial atomic charges that are required for the generation of an MEP is time-consuming, even if a semiempirical method is used (as was the case in all of our experiments, which involved the MNDO routines in MOPAC[36]), and totally infeasible in the context of a program for flexible database searching. Accordingly, we focus here on the changes in the atom-to-point distances and make the assumption that the partial charges do not vary with the conformation adopted by a flexible molecule.

The rationale for our work is the expectation that improved intermolecular similarity relationships will be identified in a database search if the MEP-alignment procedure is able to take account of flexibility in a target structure and/or in a database structure. In seeking to find an appropriate alignment procedure, we have been guided by the extensive studies that have been carried out into techniques for flexible 3-D substructure searching.[6−12] This involves the retrieval of all of those molecules from a database that contain a user-defined query pharmacophore, which typically consists of a set of atoms (or, more generally, pharmacophore points such as ring centroids, hydrogen donors, and hydrogen acceptors) and the corresponding interatomic distances. Two general approaches to flexible substructure searching have been described. In the first approach, a flexible molecule is characterized by a small number of conformations that are checked to ascertain whether any of them contain a query pharmacophore.[6,37,38] The alternative approach involves a torsional optimization procedure that permits an exploration of the full conformational space of a flexible molecule at search time, seeking to determine whether it can adopt a conformation that contains the pharmacophore.[8−10] The application of field-graphs to the multiconformation approach and of GAs to the torsional optimization approach are considered in the remainder of this paper.

SEARCHING FLEXIBLE MOLECULES USING FIELD-GRAPHS

The successful use of the field-graph approach for searching rigid molecules has been described previously.[23] The successful use of the approach for flexible searching will be determined by the extent to which torsional rotations will affect the similarities that result from matching the field-graphs that represent a target structure and a database structure. Specifically, our initial experiments have focused on the numbers of field-graphs that are required to delineate

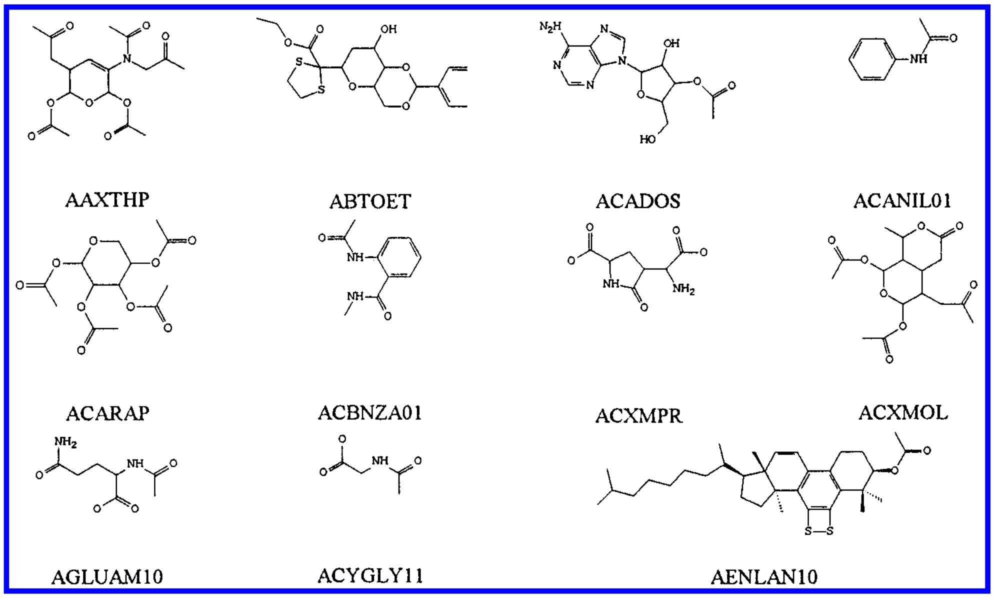

Figure 1. The 11 molecules that were used for the field-graph experiments, each with its CSD identifier.

fully the variations in MEP that result from variations in conformation, since the field-graph approach will only be applicable to searching databases of nontrivial size if these numbers are not large.

The experiments used a dataset of 11 compounds taken from a larger dataset selected by Ghose *et al*. to evaluate conformational searching methods.[39] Each of the molecules, which are shown in Figure 1, had a high-resolution crystal structure, with an *R* factor below 0.05, available in the Cambridge Structural Database (CSD)[40,41] and between 3 and 15 rotatable bonds. The SYBYL SEARCH module was used to generate a number of conformations (between 48 and 1589) for each of the 11 molecules by systematic increments of their rotatable bonds. A field-graph was generated from each of the resulting conformations, using the graph-generation procedures described by Thorner *et al*.[23] The set of field-graphs for the set of conformations for each molecule was next converted into a database that could be searched using our existing routines for MEP-based 3-D similarity searching, with the target structure in each case being the field-graph that was generated from that molecule's CSD crystal structure.

The similarity was calculated after aligning the field-graphs representing the target structure and each of the different conformations. The results were summarized as shown in Figure 2, which is a sorted list of the similarities calculated for the 301 conformations that were generated for the molecule AGLUAM10. It will be seen that there are a few conformations with similarities in excess of 0.80 but that the great majority of the conformations have much smaller similarities: for this dataset, the mean similarity is 0.601 with a standard deviation of 0.096. Similar results are obtained with all of the other molecules, as shown in Table 1. If all of the conformations for a particular molecule gave comparable MEPs, and hence comparable field-graphs, then most of the similarities with the target structure would be near to 1.0. In fact, most of them are very much smaller than this, with one of the conformations for ACARAP yielding a similarity as low as 0.202. It is thus clear that torsional rotations can bring about substantial changes in the similarity between two different conformations of the same molecule. Moreover, the smoothness of the curve shown in Figure 2 suggests that there is no clustering of the similarity values, such as might have been expected if some small number of conformations was sufficient to describe the range of field-graphs arising from torsional rotations.

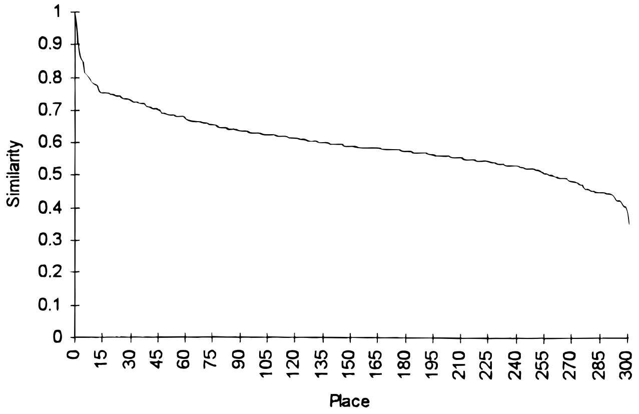

Figure 2. Similarity values for the molecule AGLUAM10 after sorting into descending order.

Table 1. Mean and Standard Deviations for the Similarity between a CSD Target Structure and Its Conformers

molecule	conformations	mean (SD)
AAXTHP	965	0.540 (0.092)
ABTOET	625	0.887 (0.058)
ACADOS	478	0.640 (0.067)
ACANIL01	48	0.334 (0.080)
ACARAP	1589	0.494 (0.016)
ACBNZA01	85	0.478 (0.084)
AENLAN10	356	0.759 (0.013)
ACXMOL	428	0.719 (0.013)
ACXMPR	232	0.718 (0.095)
AGLUAM10	301	0.601 (0.096)
ACYGLY11	627	0.308 (0.071)

The relationship of the torsional similarities between pairs of conformations to the corresponding MEP similarities was investigated in a manner analogous to that suggested by Leach.[42] Let τ_{ti} and τ_{ci} be the *i*th torsion angle ($1 \leq i \leq n$, where n is the number of rotatable bonds) in the target structure and in a particular conformation, respectively. The torsional similarity between the target structure and that conformation is then defined by

$$1 - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\tau_{ti} - \tau_{ci}}{180}\right)^2$$

with a value for 1.0 corresponding to a pair of structures that have identical sets of torsion angles. The relationship between the torsional similarity and the MEP similarity between a conformation and the target structure is illustrated in the scatter plot shown in Figure 3, which is for AGLUAM10. It will be seen that the relationship between the torsional and MEP similarities is not a strong one, the product-moment correlation coefficient here being 0.334. Comparable plots were obtained with the other molecules in our test dataset.

The figure shows that while conformers that are torsionally similar to the target structure tend to have high MEP similarities, the relationship is far too inexact for predictive purposes. In extreme cases, indeed, just a small change in a single torsion angle can bring about a change of up to 40% in the value of the MEP similarity between the target structure and the resulting conformation. There are two reasons for such drastic changes in the similarity. The first, and most obvious, is a change in the MEP itself arising from the rotation of a bond near the center of a molecule. This can result in large-scale alterations in the overall geometry of the molecule and hence in the atom-to-point distances that are used in the calculation of the potential at each point in the 3-D grid surrounding a molecule. This problem is exacerbated by the second reason, which is the lack of robustness in the routine that is used to generate a field-graph from the set of point potentials. We have found that the form of the field-graph resulting from this routine can be overly sensitive to the precise grid-point potential values, with only small changes in these values sometimes leading to changes in the number of nodes in the field-graph representing that set of values and thus to substantial changes in the resulting intermolecular similarities.

Flexible 3-D substructure searches make use of extensive screening strategies to minimize the number of molecules for which a detailed search needs to be carried out.[6,7,10] One strategy that has been found to be particularly effective is to associate a *distance range* with each pair of atoms in a flexible molecule, where the lower-bounds and upper-bounds of the range correspond to the minimum and maximum separations of the two atoms as the molecule flexes. The set of distance ranges for a molecule will contain all of the geometrically-feasible conformations which that molecule can adopt. This representation allows the use of graph-based screening procedures that are analogous to those used for 2-D and rigid 3-D substructure searching, with only those molecules that match the query at the graph level proceeding to the final, detailed conformational search.[7]

Some preliminary experiments were carried out to investigate the extent to which it might be possible to associate a *potential-range* with each point in the 3-D grid surrounding a molecule, in a manner analogous to the distance-ranges that have been mentioned previously. Specifically, let the *i*th point in the grid be characterized by not one but two values, max_i and min_i, which are the maximum and the minimum values of the potential that are observed at that point as the molecule flexes. Consideration was given to defining a version of the cosine coefficient that used potential-ranges, rather than individual potential-values as is normally the case, with the aim of screening out molecules that had a low degree of similarity to a target structure. Such a procedure is only likely to yield meaningful measures of resemblance if the potential-ranges, $|max_i - min_i|$, are not large; unfortunately, our experiments have demonstrated that this is not the case, with large values of $|max_i - min_i|$ being observed at many of the grid-points that have non-zero values of the potential.[27] Accordingly, it seems unlikely that potential-ranges of the sort studied here can serve any useful screening function.

The results we have discussed thus far suggest that it will be difficult for a field-graph representation derived from a single conformation to provide an adequate description of the variations in MEP that can occur as a result of torsional rotations. Recent work by Smellie *et al*. suggests that, if chosen appropriately, 50−100 conformations are needed to describe the conformational space of a flexible molecule for the purposes of 3-D substructure searching,[38] and the results reported above would suggest that comparable numbers (at least) will be required for 3-D similarity searching, where one is matching conformations from the target and/or database structures (rather than conformations from the same molecule as in the experiments reported here). We are hence forced to conclude that a multiconformation approach to field-based searching would be very slow in operation. Our current algorithms for matching pairs of rigid-molecule MEPs[22,23] take 1−2 CPU s on conventional workstation equipment, and thus a multiple-conformation MEP search of a rigid target structure against a flexible database structure would be expected to require 50−200 CPU s for each match, even if no account is taken of the further substantial increase in the execution time that would (presumably) be required also to encompass flexibility in the target structure.

There are at least three further problems associated with the field-graph approach. First, the generation of each graph requires as input a single, fixed MEP, and this generation procedure would hence have to be repeated very many times to create a database for flexible searching (with consequent storage and processing costs). Secondly, the experiments reported above have demonstrated that this procedure is not very robust, in that small conformational changes can result in large changes in the resulting field-graph. Finally, the criteria that are used to identify the field-graph nodes can result (in about 6% of cases for the molecules we have used) in a field-graph containing less nodes than are required to generate a unique alignment when two molecules are compared using the MCS algorithm. For these reasons, we believe that the GA-based approach to similarity searching may be more appropriate, especially as GAs have been shown previously to be well suited to the processing of flexible molecules.[33] Accordingly, the remainder of this paper describes the design and testing of a GA for flexible similarity searching.

SEARCHING FLEXIBLE MOLECULES USING GENETIC ALGORITHMS

The Algorithm. The GA that we have developed is an extension of that described by Wild and Willett for rigid similarity searching.[22] The algorithm can be used in two ways. In the first, which has formed the basis for most of our experiments thus far, a rigid target structure is assumed but each of the database structures is allowed to be flexible. Alternatively, the target structure can also be allowed to be flexible.

The GA is designed to identify a set of geometric transformations (these including rotations, translations and torsional rotations) that results in the maximal overlap of a database-structure's MEP with that of the target structure. These transformations are encoded in a chromosome that contains five, 1-byte components plus an extra one-byte component for each rotatable bond in the database structure, so that, *e.g.*, each of the chromosomes describing a molecule with four rotatable bonds would be 9 bytes long. A single byte encodes 256 possible rotations, either of the entire molecule or of individual rotatable bonds, thus giving a step-size of about 1.4 degrees. The translations carried out along the X, Y, and Z axes are each encoded in a single byte, with a step size for the translation along a particular axis being obtained by measuring the size of the molecule along that axis, adding 10% and then dividing by 256.

The matching of the target structure and a database structure is initiated by positioning their centers of mass at the origin, *i.e.*, at the point (0,0,0), so that a large amount of time is not wasted bringing the two molecules into the same general area of 3-D space, as can occur if the database structure is initially positioned at random. The initial population for the GA is created by generating random values for each of the components of each of the chromosomes. The rotations and translations encoded in an individual chromosome are used to define a conformation and a location for the database structure and thus to define its alignment with respect to that of the target structure. A simple van der Waals radius bumpcheck procedure is used to ensure that the torsion angles encoded in the chromosome do not represent a high-energy conformation: if a contact is detected then that chromosome is allocated a smaller fitness than all of the other members of the current population and processing continues with the next chromosome. If the bumpcheck is successful, then the alignment acts as the input to a routine that calculates the overlap of the two MEPs, and hence the intermolecular similarity, using the Gaussian approximation method.[25] The resulting similarity value is the raw fitness value for that chromosome, and these values form the input to the roulette-wheel selection procedure that is used to select parent chromosomes for the next generation of the GA.

Roulette-wheel selection selects individual chromosomes from the current population with a probability that is directly proportional to their fitnesses. Three different approaches were studied here. The first, and simplest, was to use the fitness values resulting from the Gaussian similarity calculation. The second approach used a windowed fitness measure, which scaled each raw fitness value to a predefined range and then added a fixed starting increment to each resulting fitness. Finally, the chromosomes were ranked in decreasing order of raw fitness from the size of the population down to one (which is the least-fit chromosome in the current population) with their position in the ranked list being taken as the fitness. A starting value was also added to each chromosome to provide a fitness window. This last approach was found to give the best results, with a value of 1.5 being used for the selection pressure, *i.e.*, the ratio of the fitness of the fittest chromosome to the mean fitness of the population. This value is higher than is commonly used in GAs, with the consequent possibility of premature convergence, but was felt to be necessary in view of the need to process the large numbers of molecules that are involved in database searching.

Chromosomes are selected for processing by the two conventional genetic operators: crossover and mutation. Three different crossover methods were tested, these being conventional one-point crossover, two-point crossover, and uniform crossover. Of these, two-point crossover was found to give by far the best results. Here, two positions within a chromosome string are selected at random, and the two parent chromosomes then exchange all of their genetic material between the two selected positions. The mutation operator works by checking each individual bit of the chromosome in turn and then flipping it (*i.e.*, changing it from one to zero or *vice versa*) if a randomly-generated number in the range 0−100 is less than the user-defined mutation rate. The choice of which operation to perform in each generation is made by generating a random number in the range 0−100: if this number is less than the crossover rate then crossover is performed, otherwise mutation. The chromosome(s) to which the chosen operation is to be applied (two chromosomes for crossover and one for mutation) are identified by roulette-wheel selection. Once the appropriate operator has been applied, the population for the next generation is created by means of a steady-state-without-duplicates replacement strategy in which a single new chromosome replaces the least-fit member of the current population (unless it is identical to an existing member of that population in which case a further new chromosome is generated). Initial tests showed that this was more effective than an alternative generational replacement strategy.

Choice of Parameter Values. It will be clear that there are, as with most GAs, very many parameters for which appropriate values must be obtained if the algorithm is to perform effectively. These values were obtained by taking seven target molecules and then calculating the similarity between each of them and each of a database of 451 molecules drawn from the Fine Chemicals Database, with both the target and database molecules being processed using the CONCORD and MOPAC procedures described previously. A GA is inherently nondeterministic in nature and each run was hence repeated ten times and the similarity coefficients noted in each case. One combination of parameter settings was assumed to be superior to another such combination if the mean similarity of the former, when averaged over all 7 × 451 × 10 similarities, was greater than the mean similarity of the latter. Wright presents the results of the extremely detailed parameterization tests that were carried out.[27] These tests led to the selection of the following parameter values for use in our main experiments (as detailed below), with the range of values that were tested included in brackets: a population of 150 chromosomes (varied in the range 5−500); a selection pressure of 1.5 (varied in the range 1.05−1.90); 1250 generations (varied in the range 50−5000); a crossover rate of 35% (varied in the range 2−90%); and a mutation rate of 7% (varied in the range 0.1−90%).

The parameter values we have used cannot be guaranteed to be optimal for producing the best alignments and hence the largest possible similarities. This is because consideration also needed to be given to the length of time taken for a run, since some of the runs involving large populations and/or large numbers of generations would have been far too

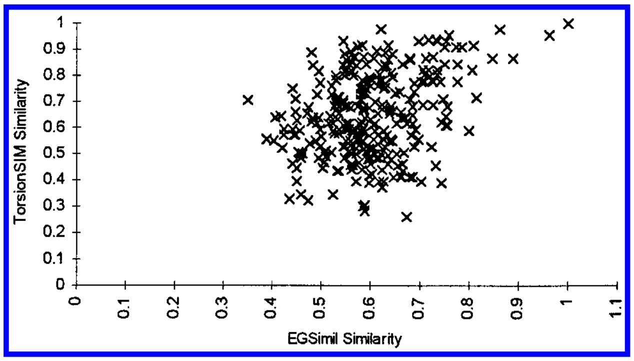

Figure 3. Scatter plot of torsional similarity against MEP similarity for the molecule AGLUAM10.

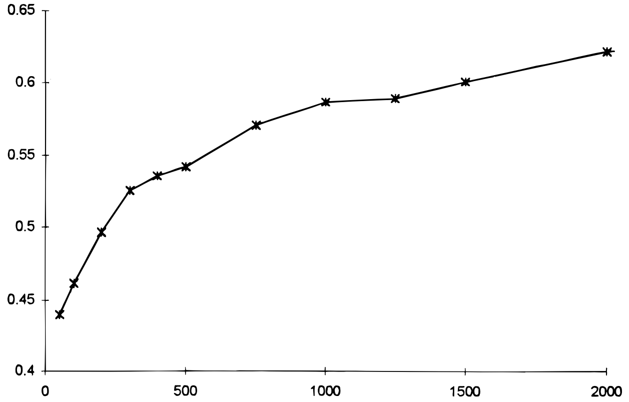

Figure 4. Effect of increasing the number of generations on the calculated similarity values. The Y axis of this figure is the mean similarity when averaged over all 10 runs for matching each of seven target structures against 451 database structures. The results were obtained with a population containing 150 chromosomes.

slow if applied to a large database, and to the consistency over the set of ten runs, as manifested in the range of similarity values that were obtained for a given combination of parameter values. Figure 4 shows the increases in the observed mean similarities that result from increases in the number of generations, and it will be seen that the GA has not converged with the parameter values we have chosen to use. Similar comments apply to the decrease in the range of values observed in a set of ten runs as the number of generations is increased, as exemplified in Figure 5. The shapes of these figures, which were obtained with the default population size of 150 chromosomes, are in line with those resulting from the use of other population sizes.

Figures 4 and 5 suggest that the GA could usefully be run for many more than the 1250 generations used here and also repeated more than 10 times; however, these modifications would result in the GA being too slow for searching databases of substantial size. As it is, the listed parameter values result in run times of about 3.5 CPU s for calculating the similarity between a rigid target structure and a flexible database structure, using an implementation of the algorithm in the C programming language on a Silicon Graphics R4000 workstation. The fitness calculation, which has to be carried out each time that a new chromosome is created, dominates the computational requirements of the algorithm, even with the use of precalculated look-up tables for the exponents that are needed for the Gaussian similarity function.

Search Performance. Our main experiments used a further set of 491 structures from the Fine Chemicals Database, 39 of which were used in turn as the rigid target structure for a search using the parameter values given previously. For comparison, we have carried out an exactly comparable set of runs in which the database structures were kept entirely rigid so as to determine the increase in performance, if any, resulting from the inclusion of torsional flexibility in the matching process. Exactly the same algorithm was used as for the flexible runs, with the exception that a population of size 15 was used with 750 generations; while these values are less than in the flexible runs, they are both larger than the standard values that were derived from our earlier studies of rigid MEP-based similarity searching[22] and thus provide a reasonable basis for comparison.

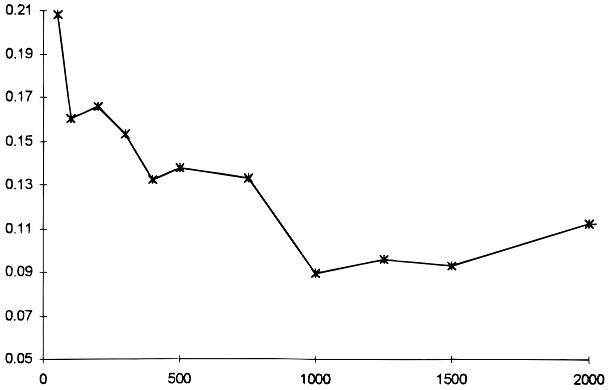

Figure 5. Effect of increasing the number of generations on the range of the calculated similarity values. Let $AVSIM_{max}$ and $AVSIM_{min}$ be the largest and the smallest average similarities calculated in a set of ten runs using one of seven target structures. The Y axis of this figure is the value of $AVSIM_{max} - AVSIM_{min}$, when averaged over all 451 similarity calculations and all seven of the target structures. The results were obtained with a population containing 150 chromosomes.

The mean similarity of each target structure to each database structure, when averaged over all 39 target structures, all 491 database structures in each case, and all ten sets of runs, was 0.465 for the searches of the rigid database structures and 0.556 for the searches of the flexible database structures. Rigid searching is a limiting case of flexible searching, and we thus expect that the similarities obtained from a flexible search will be at least as good as those obtained from a rigid search: our results support the contention that the inclusion of conformational flexibility enables the identification of better MEP overlaps than if only rigid molecules are considered. The distribution of the calculated similarities is shown in Figure 6, which illustrates the marked shift to higher similarities that results from the inclusion of database-structure flexibility in the matching algorithm.

Let $S_{flexible}$ (or S_{rigid}) be the similarity between a target structure and a particular database structure when the latter is (or is not) allowed to flex. In general, $S_{flexible} > S_{rigid}$, as would be expected (and as is demonstrated by Figure 6); however, there are many exceptions, with 14.4% of the database structures on average for each target structure having $S_{rigid} > S_{flexible}$. We do not believe that these cases should be regarded as being failures of our algorithm; rather, they reflect the inherently nondeterministic nature of any GA and the need to suggest appropriate overall parameter values that may well be suboptimal for a specific pair of structures.

An analysis of the values of $S_{flexible}-S_{rigid}$ shows that the difference is positively correlated with the number of

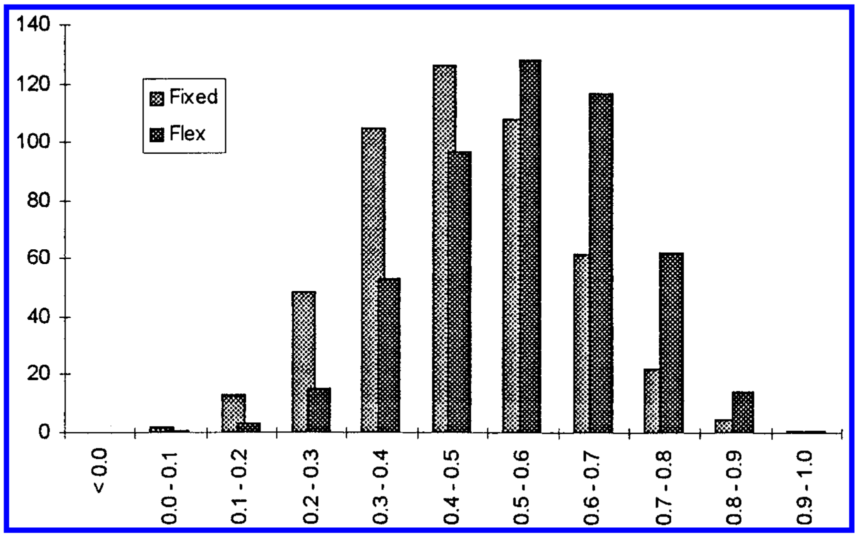

Figure 6. Distribution of observed similarities for rigid and flexible 3-D similarity searching. Each column denotes the mean number of similarities in the indicated range, when averaged over the 39 target molecules that were used in these experiments.

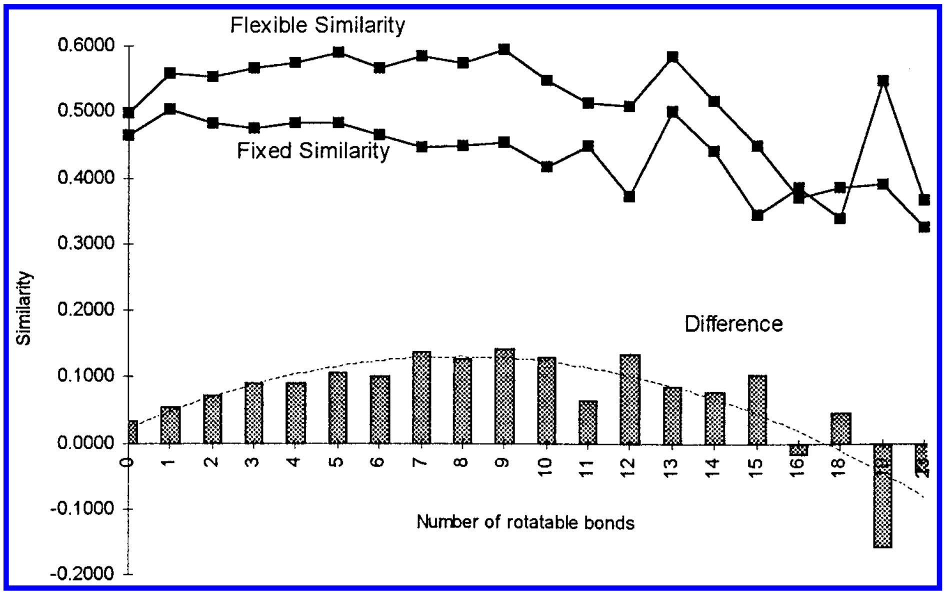

Figure 7. Effect of the number of rotatable bonds in a database structure on the calculated similarities. The upper part of the figure plots the mean similarity, averaged over all of the database structures with a given number of rotatable bonds, against that number of rotatable bonds for the rigid and for the flexible searches, while the lower part plots the difference of these two similarities.

rotatable bonds but then becomes increasingly negative as more highly flexible molecules are considered. This behavior is illustrated in Figure 7, which plots the mean similarities, $S_{flexible}$ and S_{rigid}, and the mean differences, $S_{flexible}-S_{rigid}$, the averages being taken over all of the database structures having the same numbers of rotatable bonds. It will be seen that the difference grows steadily until about eight rotatable bonds, at which point it levels off and then starts to decrease. The merits of including flexibility are thus increasingly apparent up to a certain number of rotatable bonds; thereafter, there would seem to be an increasing lack of convergence in the GA as the search space grows in size.

Thus far, we have compared the rigid and flexible GAs by the extent to which they are able to identify effective alignments, as reflected in the resulting similarities, of a target structure with a database structure. It is necessary also to consider the efficiencies of the two algorithms, as reflected in the computational times necessary for the generation of an alignment. The mean run-time for the rigid and flexible GAs in the comparative experiments were 4.12 and 3.55 CPU s, respectively, when the algorithms were implemented in C on an R4000 Unix workstation. We have noted previously that the GA parameters were chosen so as to given broadly comparable run-times to those necessary for the flexible search, and it is thus rather surprising that the flexible GA should be faster. The reason for this is that while it is parameterized to run for 1250 generations, some of these will represent conformations that fail the bumpcheck and these are assigned a very small fitness value without the actual fitness being calculated. The similarity calculations that are needed for the fitness function are by far the most time-consuming part of the GA and the elimination of some of these calculations thus yields nontrivial reductions in the observed run-times. Alternatively, we could have run the GA for the full set of 1250 generations, neglecting any that did not need the full fitness calculation.

CONCLUSIONS

This is the fourth paper describing work in our laboratory on the applicability of field-type information for similarity searching in databases of 3-D structures. In it, we have extended our previous studies[21−23] to consider the inclusion of conformational flexibility in the matching of the MEPs representing a target structure and each of the structures in a database.

The first set of experiments here investigated the effect of variations in the conformation of a molecule on its MEP, as reflected in its field-graph.[23] These experiments demonstrate that it is possible to use the field-graph approach for flexible MEP searching only if each molecule is represented by large numbers of field-graphs, with consequent, and substantial, processing costs. We have hence investigated an alternative approach, based on the GA we have described previously for rigid MEP searching,[22] and demonstrated that this algorithm can be used for flexible searching with only minor modifications. Experiments using this flexible GA demonstrate clearly that it is more effective than a comparable GA for searching rigid database structures and is also at least as efficient in operation. In the future, it may be possible to achieve further improvements in the effectiveness of searching without any increase in the response time. This is because changes in the number of generations had the largest effect on the calculated similarities from amongst all of the variants of our GA that were tested systematically (crossover rate, mutation rate, population size, selection pressure, replacement strategy, and crossover operator).[27] Better alignments, and correspondingly larger intermolecular similarities, may thus be obtained simply by increasing the number of generations that are carried out; accordingly, improvements in hardware or the use of multiple machines may be expected to translate directly into improvements in the performance of our algorithm.

We are currently studying two developments of the work reported here, both of which will be reported shortly. Firstly, we have considered thus far only the flexibility of the database structures, while keeping the target structure rigid. It is, however, simple to extend the algorithm to include any inherent flexibility in the target structure. All that is required is to extend the chromosome by a further byte for each rotatable bond in the target structure and to treat these bytes in just the same way as is currently done for the corresponding bytes in a description of a database structure. Preliminary experiments suggest that the inclusion of flexibility within the target structure will enable further improvements in the alignments that can be identified, albeit at a substantially increased computational cost. Secondly, the principal focus of our work to date has been the development of appropriate algorithms and data structures for MEP-based searching. Previous studies with small QSAR datasets have demonstrated the relationships that exist between MEP-based similarities and biological activity data.[24−26] We believe that the approach described here provides a firm basis for carrying out such correlation studies on a much larger scale than previously, using public databases that contain both structural and activity data.

ACKNOWLEDGMENT

We thank the James Black Foundation, the Biotechnology and Biological Sciences Research Council, the Engineering and Physical Sciences Research Council and Zeneca Agrochemicals for funding; Tripos Inc. for software support; Val Gillet for comments on an earlier draft of this paper; and Jonathan Davies, Caroline Low, and J. Andrew Vinter for helpful advice. The Krebs Institute for Biomolecular Research is a designated Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Co-ordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(2) Ricketts, E. M.; Bradshaw, J.; Hann, M.; Hayes, F.; Tanna, N.; Ricketts, D. M. Comparison of Conformations of Small Molecules from the Protein Data Bank with Those Generated by Concord, Cobra, ChemDBS-3D and Converter, and Those Extracted from the Cambridge Structural Database. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 905−925.

(3) Bures, M. G.; Martin, Y. C.; Willett, P. Searching Techniques for Databases of Three-Dimensional Chemical Structures. *Topics Stereochem.* **1994**, *21*, 467−511.

(4) Jakes, S. E.; Watts, N. J.; Willett, P.; Bawden, D.; Fisher, J. D. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Evaluation of Search Performance. *J. Mol. Graphics* **1987**, *5*, 41−48.

(5) Sheridan, R. P.; Nilakantan, R.; Rusinko, A.; Bauman, N.; Haraki, K. S.; Venkataraghavan, R. 3DSEARCH: a System for Three-Dimensional Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 255−260.

(6) Murrall, N. W.; Davies, E. K. Conformational Freedom in 3-D databases. 1. Techniques. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 312−316.

(7) Clark, D. E.; Willett, P.; Kenny, P. W. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Use of Smoothed-Bounded Distances Matrices for the Representation and Searching of Conformationally-Flexible Molecules. *J. Mol. Graphics* **1992**, *10*, 194−204.

(8) Clark, D. E.; Jones, G.; Willett, P.; Kenny, P. W.; Glen, R. C. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Comparison of Conformational-Searching Algorithms for Flexible Searching. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 197−206.

(9) Hurst, T. Flexible 3D Searching: the Directed Tweak Technique. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 190−196.

(10) Moock, T. E.; Henry, A. G.; Ozkabak, A. G.; Alamgir, M. Conformational Searching in ISIS/3D Databases. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 184−189.

(11) Greene, J.; Kahn, S.; Savoj, H.; Sprague, P.; Teig, S. Chemical Function Queries for 3D Database Search. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1297−1308.

(12) Milne, G. W. A.; Nicklaus, M. C.; Driscoll, J. S.; Wang, S. National Cancer Institute Drug Information System 3D Database. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1219−1224.

(13) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman and Hall: Glasgow, 1995.

(14) Downs, G. M.; Willett, P. Similarity Searching in Databases of Chemical Structures. *Rev. Comput. Chem.* **1995**, *7*, 1−66.

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

(16) Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching of CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664−674.

(17) Perry, N. C.; van Geerestein, V. J. Database Searching on the Basis of Three-Dimensional Molecular Similarity Using the SPERM program. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 607−616.

(18) Ho, C. M. W.; Marshall, G. R. FOUNDATION: a Program to Retrieve All Possible Structures Containing a User-Defined Minimum Number of Matching Query Elements from Three-Dimensional Databases. *J. Computer-Aided Mol. Design* **1993**, *7*, 3−22.

(19) Bath, P. A.; Poirrette, A. R.; Willett, P.; Allen, F. H. Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 141−147.

(20) Good, A. C.; Ewing, T. J. A.; Gschwend, D. A.; Kuntz, I. D. New Molecular Shape Descriptors: Application in Database Screening. *J. Computer-Aided Mol. Design* **1995**, *9*, 1−12.

(21) Turner, D. B.; Willett, P.; Ferguson, A. M.; Heritage, T. W. Similarity Searching in Files of Three-Dimensional Chemical Structures: Evaluation of Similarity Coefficients and Standardisation Methods for Field-Based Similarity Searching. *SAR QSAR Environmental Res.* **1995**, *3*, 101−130.

(22) Wild, D. J.; Willett, P. Similarity Searching in Files of Three-Dimensional Chemical Structures: Alignment of Molecular Electro-

static Potentials with a Genetic Algorithm. *J. Chem. Inf. Comput. Sci.* **1996**, in press.

(23) Thorner, D. A.; Willett, P.; Wright, P. M.; Taylor, R. Similarity Searching in Files of Three-Dimensional Chemical Structures: Representation and Searching of Molecular Electrostatic Potentials Using Field-Graphs. Manuscript in preparation.

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

(25) Good, A. C.; Hodgkin, E. E.; Richards, W. G. The Utilisation of Gaussian Functions for the Rapid Evaluation of Molecular Similarity. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 188−191.

(26) Good, A. C.; Peterson, S. J.; Richards, W. G. QSARS from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods. *J. Med. Chem.* **1993**, *36*, 2929−2937.

(27) Wright, P. M. Ph.D. Thesis, University of Sheffield, Manuscript in preparation.

(28) Kuhl, F. S.; Crippen, G. M.; Friesen, D. K. A Combinatorial Algorithm for Calculating Ligand Binding. *J. Comput. Chem.* **1984**, *5*, 24−34.

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

(30) Goldberg, D. E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison Wesley: New York, 1989.

(31) Forrest, S. Genetic Algorithms: Principles of Natural Selection Applied to Computation. *Science* **1993**, *261*, 872−878.

(32) Goldberg, D. E. Genetic and Evolutionary Algorithms Come of Age. *Commun. ACM* **1994**, *37*(3), 113−119.

(33) Willett, P. Genetic Algorithms in Molecular Recognition and Design. *Trends Biotechnol.* **1995**, *143*, 516−521.

(34) CONCORD is distributed by Tripos Associates, St. Louis, MO 63144, U.S.A.

(35) Reynolds, C. A.; Essex, J. W.; Richards, W. G. Atomic Charges for Variable Molecular Conformations. *J. Am. Chem. Soc.* **1992**, *114*, 9075−9079.

(36) Stewart, J. J. M. MOPAC: a Semi-Empirical Molecular Orbital Program. *J. Computer-Aided Mol. Design* **1990**, *4*, 1−105.

(37) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.

(38) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 2. Applications of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.

(39) Ghose, A. K.; Jaeger, E. P.; Kowalczyk, P. J.; Peterson, M. L.; Treasurywala, A. M. Conformational searching methods for small molecules. 1. Study of the SYBYL search method. *J. Comput. Chem.* **1993**, *14*, 1050−1065.

(40) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(41) Allen, F. H.; Kennard, O. 3D Search and Research Using the Cambridge Structural Database. *Chem. Design Automation News* **1993**, *8*(1), 31−37.

(42) Leach, A. R. An Algorithm to Directly Identify a Molecule's "Most Different" Conformations. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 661−670.

CI960002W