An Algorithm To Directly Identify a Molecule's "Most Different" Conformations

Andrew R. Leach

Department of Chemistry, University of Southampton, Southampton, Hampshire SO9 5NH, U.K.

Received September 1, 1993[®]

Many molecules have a large number of minimum energy conformations. It is frequently desired to select from these conformers a smaller, representative sample for subsequent modeling studies or storage in a structural database. An obvious approach to this problem is to generate all conformations and then use cluster analysis to divide the structures into families from which the representative sample can be derived. Both the conformational search and the cluster analysis can be computationally demanding. Here we describe an alternative yet equivalent approach which can directly generate a molecule's "most different" conformations. The method combines the improved leader cluster algorithm and the A* search method. In this direct clustering search, the lowest energy conformation is first determined using the A* algorithm, as we have described before (Leach, A. R.; Prout, K. *J. Comput. Chem.* **1990**, *11*, 1193–1205). The second conformation to be generated is the one which is "most different" from this structure. Subsequent structures span the conformational space in accordance with the improved leader clustering method. We also describe a variant on this algorithm that can directly identify the two conformations corresponding to the maximum and minimum distances between a specific pair of atoms. Such an algorithm has potential applications in the derivation of distance screens for searching "three-dimensional" databases.

INTRODUCTION

The properties of a molecule are often intimately linked to the three-dimensional structures it can adopt. Most molecules of interest to organic, bioorganic, and medicinal chemists can adopt more than one conformation. The purpose of a "conformational search" is to determine a molecule's thermally accessible minimum energy conformations. A wide variety of methods for searching conformational space have been described in recent years, including systematic and random algorithms, distance geometry, and molecular dynamics.[1,2] Although a molecule may have a large number of minimum energy conformations, a high fraction of these conformations may be very similar, differing in extreme cases only by the rotation of a hydrogen atom (e.g. in a hydroxyl group). It is important to consider the conformational flexibility of molecules, but an upper limit must often be imposed on the number of conformations that can be stored and processed. Under such circumstances it is clearly desirable that the structures that are chosen provide a good representation of the total set. Such a set of representative conformations can be identified using cluster analysis.[3] All clustering algorithms require some means of determining "how different" any pair of objects are. This is usually calculated as a distance in the multidimensional space of the properties which are used to define each object. There are a number of ways in which the distance between two conformations can be calculated. Perhaps the most commonly used method is the root-mean-square distance (RMSD), which is the square root of the minimum mean square distance between corresponding atoms of one conformation relative to the other. A number of alternatives to the RMSD have been employed, such as the sum of the differences in the interatomic distance matrices of the two conformations. A third measure which we shall employ in the work described here is given by the root mean sum of the squares of the differences in the torsion angles of the two structures:

$$d_{\mathrm{ij}} = \left(\frac{\sum_{k=1}^{N}(\tau_{\mathrm{ik}} - \tau_{\mathrm{jk}})^2}{N}\right)^{1/2} \tag{1}$$

Here, d_{ij} is the torsional "distance" between the two conformations i and j, and $(\tau_{\mathrm{ik}} - \tau_{\mathrm{jk}})$ is the smallest difference between the values of the torsion angle k in the two structures, taking into account the 2π periodicity of torsion angles.

Both the conformational search and the cluster analysis can be computationally demanding. The processing demands are particularly important when considering very flexible molecules with many minima or when examining large numbers of molecules (e.g. to create a structural database). In this paper we describe a conformational search algorithm that is able to combine the two steps of conformational search and cluster analysis. The algorithm is thereby able to generate, in sequence, a molecule's "most different" conformations. The algorithm has been implemented in the COBRA program,[4] which is based in part on the approach taken in the WIZARD-II program.[5] These programs can rapidly and automatically identify the conformations of a molecule at or near minima on the energy surface. First we will describe those features of the approach that are pertinent to the current discussion and then consider the problem of directly generating a representative sample of conformations.

Conformations of a molecule are constructed by joining together three-dimensional structures (referred to as *templates*) of molecular fragments (called conformational *units*). The program takes each unit contained in the knowledge base and performs a substructure search[6] to determine whether it is present in the molecule. To illustrate the division of a molecule into units, we show in Figure 1 the units used to construct ochratoxin A. Each unit is permitted to adopt one or more discrete conformations which are stored as separate template files. The templates are obtained from the analysis of experimental data (e.g. crystal structures) and from theoretical studies and represent the conformations that the unit is observed to adopt in minimum energy conformations of molecules. For example, the cyclohexane unit has chair,

[®] Abstract published in *Advance ACS Abstracts*, March 15, 1994.

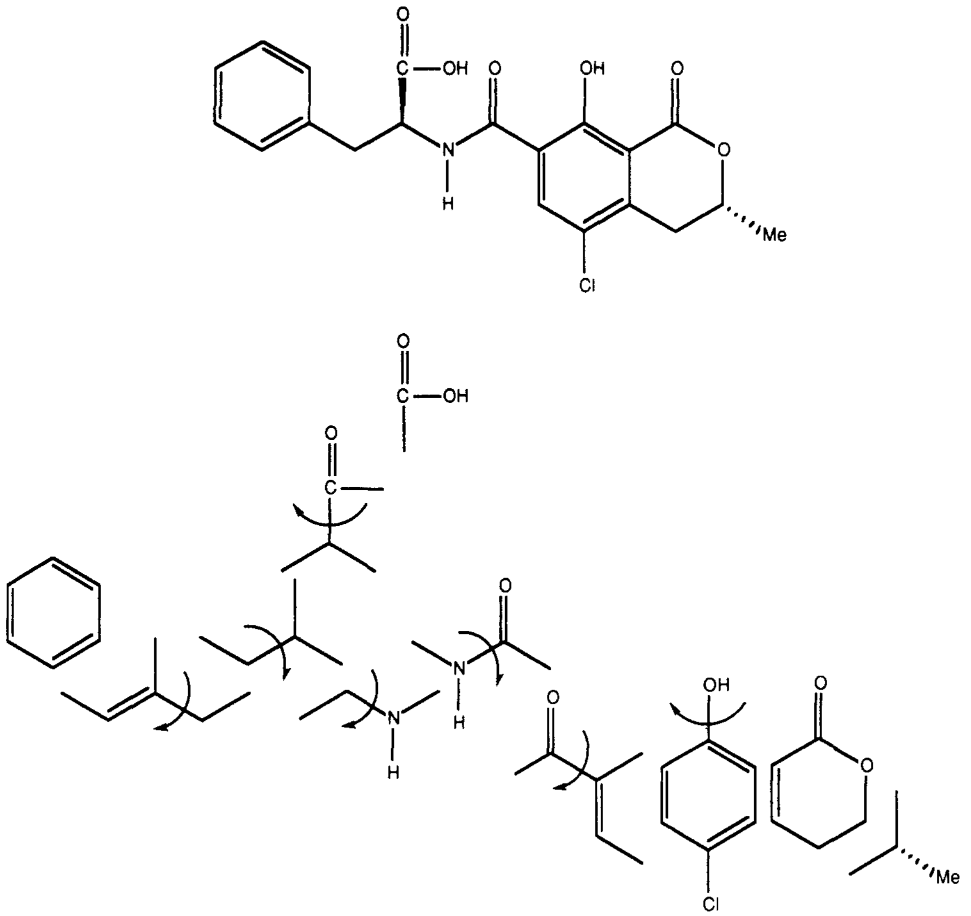


Figure 1. Conformational units used to construct ochratoxin A. For this molecule, each unit contains a single ring, a single rotatable bond (as indicated), or a single functional group.

twist-boat, and boat templates. It is also necessary to consider the orientations of each template; for example, the chair conformation of cyclohexane has two unique orientations. A specific template/orientation will be referred to as a *subconformation* of the unit. A conformation of the molecule is constructed by selecting one template for each unit, by assigning it one of its permitted orientations, and then joining[7] the templates together in a stepwise fashion. Different conformations of the molecule are obtained by joining different combinations of unit subconformations. Combinations in which the templates do not fit together properly or in which there are high-energy steric interactions are rejected as soon as they are identified.

The conformational space explored by the program consists of all possible combinations of unit subconformations and is conveniently represented as a search tree. If each unit i has S_i subconformations, then the total number of possible conformations of the molecule is $\Pi_{i=1}^{N_u} S_i$, where N_u is the number of units in the molecule. This therefore is the total number of terminal nodes in the search tree. A *goal node* corresponds to an acceptable conformation of the molecule, one in which the templates fit together satisfactorily and in which there are no unacceptably close interatomic contacts. For this reason the number of goal nodes in the search tree may be considerably fewer than the number of terminal nodes.

A variety of algorithms can be used to explore search trees such as this to find acceptable solutions. Searching algorithms involve the *expansion* of nodes in which the successor nodes of the current node are identified. Two commonly used search algorithms are the depth-first search and the breadth-first search.[8] The basic algorithm used in COBRA is the depth-first search, with backtracking to find all solutions. We have also investigated the application of alternative algorithms, including the A* search method.[4] The A* algorithm[9] is a method for finding the optimal or *least-cost* path to a goal node. This requires that each edge in the search tree has associated with it a cost. For example, in the classic traveling salesman problem an appropriate cost function would be the distance between each pair of cities. An obvious cost function to use in conformational analysis is the internal energy of the conformation, as might be calculated using molecular mechanics. The least-cost goal node then corresponds to the global minimum energy conformation. In our previous work the algorithm was extended to generate, in order of increasing energy, a user-specified number of the molecule's lowest energy conformations.[4] As we show below, the cost associated with each edge in the search tree can also be calculated using functions other than the internal energy. First, however, we provide the basic details of the A* algorithm and show how it can be used to identify the global minimum energy conformation.

THE A* ALGORITHM IN CONFORMATIONAL ANALYSIS

Central to the operation of the A* algorithm is an evaluation function termed f^*. The value of this function is calculated for each node visited by the algorithm. The value of f^* for any node n is given by

$$f^* = g^* + h^* \qquad (2)$$

g^* is the cost of reaching n from the root node, and h^* is an estimate of the additional cost to reach a goal node. An

Table 1. Energies of All Partial Conformations Containing Two Units and All Fully Constructed Conformations Containing Three Units (Arbitrary Units)[a]

A_1B_1	6	$A_1B_1C_1$	20
A_1B_2	10	$A_1B_1C_2$	18
A_2B_2	11	$A_1B_2C_1$	15
A_2B_2	16	$A_1B_2C_2$	22
A_3B_1	16	$A_2B_1C_1$	16
A_3B_2	17	$A_2B_1C_2$	30
		$A_2B_2C_1$	50
		$A_2B_2C_2$	27
		$A_3B_1C_1$	25
		$A_3B_1C_2$	23
		$A_3B_2C_1$	22
		$A_3B_2C_2$	24

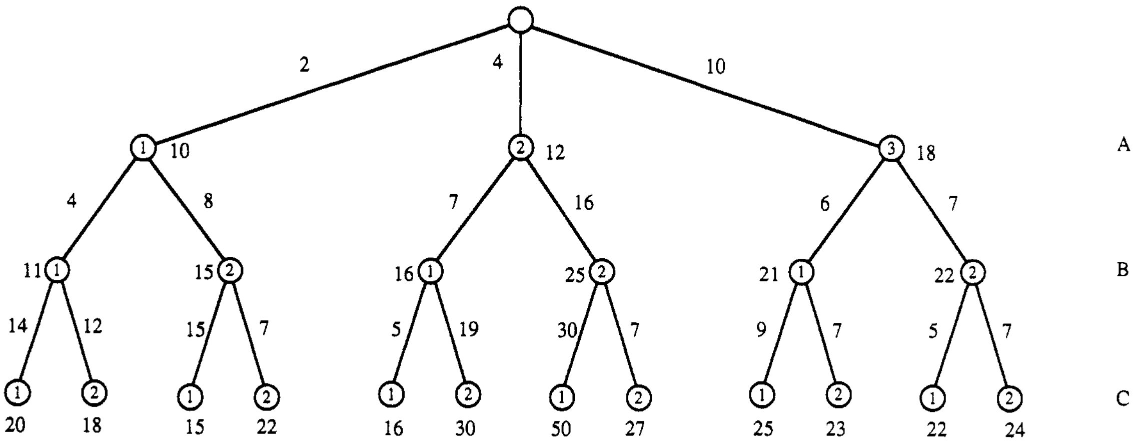
[a] The search tree for this problem is given in Figure 2.

important requirement for correct operation of the A* algorithm is that h^* should not overestimate the actual cost. It is in the use of the h^* estimates that the A* algorithm differs from a branch-and-bound search, which only uses the distance from the root node to decide which node to expand next. To implement the A* algorithm, a list is maintained of the nodes visited, ordered according to their values of f^* (lowest first). At each iteration of the procedure the node at the head of the list is taken. If it is a goal node, then the corresponding conformation is output. Nodes that are not goal nodes correspond to partially constructed conformations of the molecule: if there are m edges between such a node and the root node, then the corresponding partial conformation contains m units. To expand the node, all templates of the $(m+1)$th unit are taken and joined to this partial conformation in turn. Each of the resulting structures (now containing $m + 1$ units) is assessed for ill-fitting templates, steric problems, etc. If acceptable, the intramolecular energy of the partially constructed molecule is calculated using the COSMIC molecular mechanics force field.[10] This energy corresponds to the g^* value of the respective node in the tree. h^* is an estimate of the additional energy required to complete the conformation. This is determined by examining the units still to be added, identifying for each of these units the minimum energy template (also calculated using COSMIC), and adding together the energies of these minimum energy templates.

To illustrate the application and implementation of the algorithm, let us consider a molecule that is constructed from three units (A, B, and C) such that unit A has three subconformations and units B and C have two subconformations each. There are thus 12 terminal nodes in the search tree (=3 × 2 × 2). We assume that the units are joined in the order A, B, C. The internal energies (in arbitrary units) of the isolated templates are as follows: $A_1(2)$, $A_2(4)$, $A_3(10)$, $B_1(3)$, $B_2(5)$, $C_1(5)$, $C_2(7)$. In Table 1 we show the energies of all possible two-unit partial conformations (containing units A and B) and three-unit molecular conformations. The search commences by expanding the root node to give the three subconformations of unit A. The g^* value for each of these three nodes is equal to the internal energy of the template. The h^* values are obtained by finding the minimum energy subconformations of the two remaining units. These are B_1, which has an energy of 3, and C_1, which has an energy of 5, giving a total h^* of 8. The node with the lowest f^* value is thus A_1, which when expanded gives two partial conformations (A_1B_1 and A_1B_2), each containing two units. These have g^* values of 6 and 10, respectively, and h^* values of 5. We show in Figure 2 the complete search tree. The actual costs are indicated by each edge, and the f^* values are shown next to each node. The list of nodes varies as indicated; note how the lowest energy conformation obtained ($A_1B_2C_1$) is not the first total conformation of the molecule that is constructed ($A_1B_1C_2$), but it is the first goal node to reach the head of the list. We also show in Figure 2 how further processing of the list of nodes provides additional conformations, in increasing order of energy.

The algorithms to be described below will be primarily illustrated using a simple molecule: *n*-heptane. This molecule is constructed from six conformational units, each of which corresponds to a single carbon–carbon bond (Figure 3). Units 5 and 6 are terminal methyl groups and have just a single template; units 1, 2, 3, and 4 have three templates, corresponding to the anti, gauche(+), and gauche(−) conformations of butane. We first consider the application of the energy-based A* algorithm to heptane. For the purposes of illustration only, we will assume a simple energy model in which each gauche template for units 1–4 contributes an energy ϵ to the intramolecular energy of the molecule; the trans/anti templates of these units contribute zero energy. Each of the terminal methyl units is also assumed to contribute zero energy to the intramolecular energy. We assume that the C–C–C–C torsion angles in the gauche templates are exactly ±60° (the angle in the templates used by the program is slightly greater than 60° due to vdw (van der Waals) interactions between the terminal carbon atoms). The units are joined in the order 1, 2, 5, 3, 4, 6. This order is determined by finding the most highly connected unit and then adding units in successive shells. The search tree thus has a depth of six with 81 terminal nodes. Only 41 of these terminal nodes correspond to acceptable conformations of the molecule; there are 40 conformations which contain a gauche(+)/gauche(−) combination of units 1 and 2, 2 and 3, or 3 and 4, giving rise to high-energy hydrogen–hydrogen steric interaction. As each conformation of heptane is constructed, the coordinates of various atoms in the molecule are defined. A partial conformation containing a single unit (unit 1) has coordinate positions defined for atoms 1, 2, 3, and 4, together with the hydrogens bonded to atoms 2 and 3. When unit 2 is added, then the coordinates of atom 5 are defined, together with those of the hydrogen atoms bonded to atom 4. Addition of unit 5 defines the coordinates of the hydrogen atoms bonded to atom 1; addition of unit 3 defines the coordinates of atom 6 and the hydrogens bonded to atom 5, and so on.

When the root node is expanded by the A* algorithm, three nodes are placed on the list, corresponding to the trans, gauche(+), and gauche(−) conformations of unit 1. These have energies of 0, ϵ, and ϵ, respectively. To determine their h^* values (the estimates of the energy required to complete the conformation), we take the lowest energies of the remaining units still to be added. The minimum template energies for these units are 0 in all cases, resulting in the following list of nodes: [a(0), g+(ϵ), g−(ϵ)]. [In this illustrative example nodes with the same f^* values are ordered such that the node with the largest number of units is closer to the head of the list. The values in brackets are the energy f^* values. "a" refers to an anti/trans conformation of the butane unit, "g+" to a gauche(+) conformation, and "g−" to a gauche(−) conformation. "m" is used to indicate the single conformation adopted by a terminal methyl unit.] In the second iteration the node at the head of the list is taken and expanded. This gives the three partial conformations aa, ag+, and ag− with energies (g^* values) of 0, ϵ, and ϵ. The h^* values are 0, as before. The list of nodes is now [aa(0), ag+(ϵ), ag−(ϵ), g+(ϵ), g−(ϵ)]. Unit 5 is next added to the aa partial conformation to give the partial structure aam(0). During the first 10



1. A_1(10) A_2(12) A_3(20186)
2. A_1B_1(11) A_2(12) A_1B_2(15) A_3(18)
3. A_2(12) A_1B_2(15) A_3(18) $A_1B_1C_2$(18) $A_1B_1C_1$(20)
4. A_1B_2(15) A_2B_1(16) A3(18) $A_1B_1C_2$(18) $A_1B_1C_1$(20) A_2B_2(25)
5. **$A_1B_2C_1$(15)** A_2B_1(16) A_3(18) $A_1B_1C_1$(20) $A_1B_2C_2$(22) A_2B_2(25)
6. **$A_2B_1C_1$(16)** A_3(18) $A_1B_1C_1$(20) $A_1B_2C_2$(22) A_2B_2(25) $A_2B_1C_2$(30)
7. A_3(18) $A_1B_1C_1$(20) $A_1B_2C_2$(22) A_2B_2(25) $A_2B_1C_2$(30)

Figure 2. Search tree for a hypothetical molecule that is constructed from three units: A, B, and C. Unit A has three allowed subconformations; units B and C each have two. The numbers beside each edge in the tree give the actual cost (energy); numbers beside each node indicate the f^* values as used in the A* search. The list of nodes varies as shown in the figure. Complete conformations of the molecule are shown in bold.

iterations the list of nodes varies as given in Chart 1 (complete conformations in bold).

Chart 1

[a(0), g+(ε), g–(ε)]
[aa(0), ag+(ε), ag–(ε), g+(ε), g–(ε)]
[aam(0), ag+(ε), ag–(ε), g+(ε), g–(ε)]
[aama(0), aamg+(ε), aamg–(ε), ag+(ε), ag–(ε), g+(ε), g–(ε)]
[aamaa(0), aamag+(ε), aamag–(ε), aamg+(ε), aamg–(ε),
ag+(ε), ag–(ε), g+(ε), g–(ε)]
[**aamaam(0)**, aamag+(ε), aamag–(ε), aamg+(ε), aamg–(ε),
ag+(ε), ag–(ε), g+(ε), g–(ε)]
[aamag+(ε), aamag–(ε), aamg+(ε), aamg–(ε), ag+(ε),
ag–(ε), g+(ε), g–(ε)]
[**aamag+m**(ε), aamag–(ε), aamg+(ε), aamg–(ε), ag+(ε),
ag–(ε), g+(ε), g–(ε)]
[aamag–(ε), aamg+(ε), aamg–(ε), ag+(ε), ag–(ε), g+(ε), g–(ε)]
[**aamag–m**(ε), aamg+(ε), aamg–(ε), ag+(ε), ag–(ε), g+(ε), g–(ε)]

The g^* energies and h^* estimates are in fact calculated using a molecular mechanics force field which contains contributions from bond stretching, angle bending, rotation about bonds, and van der Waals nonbonded interactions. We have found that in some cases the requirement that h^* is less than the actual path cost can be violated. This can occur when there are sizeable attractive (i.e. negative energy) interactions between atoms in different units in parts of the molecule not yet constructed. Such interactions cannot be predicted from the internal energies of individual templates but only arise when the templates are joined together. This can sometimes mean that the first conformation to be generated is not the global minimum energy structure. However, we have found that the first conformations produced by the energy-based A* algorithm are always among the very lowest energy structures of the molecule. More details and a discussion can be found elsewhere.[4]

A DIRECTED CLUSTERING CONFORMATIONAL SEARCH ALGORITHM

We now consider how the A* algorithm can be used to find a molecule's "most different" conformations, rather than its lowest energy conformations. We refer to such a method as a *directed clustering search algorithm*. Our directed clustering search is based upon the improved leader clustering algorithm.[3] When using this algorithm to cluster a group of conformations, a "central" conformation is first determined. This can be chosen in several ways; for example, it may be the conformation closest to the average structure. The conformation most dissimilar to this central conformation, according to the currently operable difference measure, is then found. These structures define two "leaders". Next, the entire set of conformations is sorted into two clusters according to which of the two leaders they are most similar to. During sorting the conformation most dissimilar from its leader is found; this then becomes the third leader. In the next step the conformations are sorted into three clusters, and so on. After $N-1$ passes the conformations are thus divided into N clusters. This algorithm has the advantage of speed over methods which require a similarity matrix of order N to be precomputed, without some of the drawbacks associated with other quick-partition algorithms. However, it can be dependent on the selection of the initial leader. When applying this algorithm to the clustering of conformations, we choose the first structure found by the energy-based A* algorithm (the global minimum energy structure) to be the first leader. This alleviates any order dependency problem.

There are therefore two distinct parts to the directed clustering search. First, the energy-based A* algorithm is used to locate the lowest energy structure. We next wish to find the conformation which is most different from this first "leader". We employ the torsional measure given in eq 1 to calculate differences between conformations. The torsion angles which are included in the summation are determined as follows. Each bond between two non hydrogen atoms in the molecule is examined. If the two atoms which comprise the bond are in turn bonded to at least one other non hydrogen atom, then the first non hydrogen quadruple is counted toward the torsional sum. Thus 4 torsion angles would be included for n-heptane ($\tau_{1,2,3,4}$, $\tau_{2,3,4,5}$, $\tau_{3,4,5,6}$, and $\tau_{4,5,6,7}$; Figure 3) and

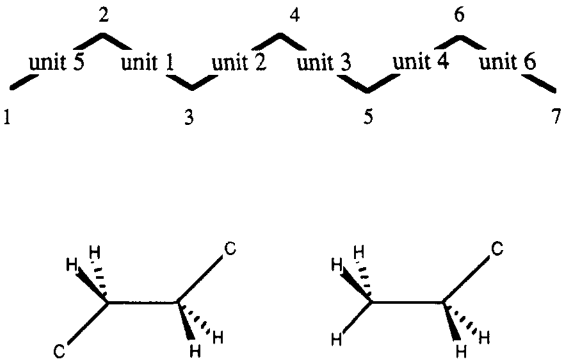


Figure 3. Atom numbering and conformational units used to construct *n*-heptane. Each of the terminal methyl units is permitted to adopt a single conformation; the butane units adopt three conformations, corresponding to the trans, gauche(+), and gauche(–) conformations.

23 torsion angles for ochratoxin A. In the clustering search we first wish to find the conformation which is most different from (i.e. the furthest distance from) the first structure generated. We therefore now seek a *maximum-cost path.* Just as a minimization algorithm can be used to maximize a function, so the A* algorithm can be easily modified to find maximum-cost paths. The h^* values now need to be *overestimates* of the actual path costs, and the list of nodes must be ordered in terms of *decreasing* f^* values. First, f^* values for the nodes in the list are calculated according to their torsional distances from the first structure generated. As before, each f^* value has two contributions, corresponding to actual and estimated torsional distances. For each partially or fully constructed conformation in the list, g^* is given by the sum of the squares of the differences in torsion angles for those angles where all four atoms have been assigned coordinates. The h^* values are determined by first identifying which of the torsion angles that contribute to the summation in eq 1 are undefined in the current structure. The maximum possible differences between the values of these torsions in the first conformation and those in the units still to be added give h^*. f^* is then calculated by dividing the sum of g^* and h^* by the number of qualifying torsion angles and taking the square root (eq 1). The second conformation to be obtained from the search is the one which is most different to the first structure produced (subject as always to the constraint that no conformation should contain any unacceptable high-energy steric interactions). If desired, an energy cutoff can be specified to eliminate any nodes for which the energy f^* value exceeds the energy of the global minimum by more than the cutoff. The third conformation that we wish to obtain is the structure furthest from its leader. To do this, we must compute new torsional f^* values for the nodes in the list, to take account of the fact that a partially or fully constructed conformation on the list may now be closer to the second structure rather than the first. The new torsional f^* values are determined by comparing each node on the list to the two completed structures, h^* again being given by the maximum possible torsional differences for those parts of the molecule not yet defined. In this way the third conformation is obtained, and so on. As each completed structure is obtained, new torsional f^* values are computed for the nodes in the list. The search proceeds until a specified number of conformations have been generated, until the f^* value of the node at the head of the list falls below a given value (which implies that all clusters are of a certain maximum size), or until there are no nodes left on the list.

To illustrate the operation of the directed clustering search algorithm, let us consider its application to heptane, using the simple model described above. The first conformation to be obtained from the energy-based A* search is the all-trans conformation. When this structure is generated, the list of nodes is [aamag+(ε), aamag–(ε), aamg+(ε), aamg–(ε), ag+(ε), ag–(ε), g+(ε), g–(ε)]; the first two nodes correspond to partial conformations with five units, the second pair of nodes correspond to partial conformations containing four units, and so on. The list is now reorganized according to the torsional difference between each of these nodes and the all-trans conformation. The first two partial conformations in the list differ by the conformation of only one unit from the conformation of the all-trans structure. Although there remains the terminal methyl unit 6 to be added, all of the atoms in the four torsion angles $\tau_{1,2,3,4}$, $\tau_{2,3,4,5}$, $\tau_{3,4,5,6}$ and $\tau_{4,5,6,7}$ have been assigned coordinates in these partial structures. The g^* values for these two nodes are thus 14 400 ($0^2 + 0^2 + 0^2 + 120^2$). As all torsion angles are defined (even though the molecule is not quite complete), the h^* values are 0 and f^* is thus 60 ($=[(0 + 14400)/4]^{1/2}$). For the second pair of nodes on the list three torsion angles are defined ($\tau_{1,2,3,4}$, $\tau_{2,3,4,5}$, and $\tau_{3,4,5,6}$), with values of 180°, 180°, and ±60°, respectively. The g^* values are thus 14 400 again. The h^* values for these two nodes, however, are 14 400 ($=120^2$), this being the maximum possible deviation of the three conformations of unit 4 from the trans conformation adopted in the lowest energy conformation. The total torsional f^* distance is thus 85 ($[(14400 + 14400)/4]^{1/2}$). For the final two nodes on the list (g+ and g–), the g^* values are 14 400 and the h^* values are 43 200 ($=120^2 + 120^2 + 120^2$), with f^* therefore being 120. The new list is thus [g+(120), g–(120), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]. The brackets now contain the torsional RMS value. The next node to be expanded is g+(120), which gives the g+g+, g+a, and g+g– combinations of units 1 and 2. These have torsional f^* values of 120, 104, and 120, respectively. The list of nodes varies as given in Chart 2.

Chart 2

[g+(120), g–(120), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]
[g+g+(120), g+g–(120), g–(120), g+a(104), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]
[g+g+m(120), g+g–(120), g–(120), g+a(104), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]
[g+g+mg+(120), g+g+mg–(120), g+g–(120), g–(120), g+g+ma(104), g+a(104), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]
[g+g+mg+g+(120), g+g+mg+g–(120), g+g+mg–(120), g+g–(120), g–(120), g+g+mg+a(104), g+g+ma(104), g+a(104), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]
[**g+g+mg+g+m(120)**, g+g+mg+g–(120), g+g+mg–(120), g+g–(120), g–(120), g+g+mg+a(104), g+g+ma(104), g+a(104), ag+(104), ag–(104), aamg+(85), aamg–(85), aamag+(60), aamag–(60)]
[g–(120), g+g–(104), g+a(104), ag+(104), ag–(104), g+g+ma(85), g+g+mg–(85), aamg+(85), aamg–(85), g+g+mg+a(60), g+g+mg+g–(60), aamag+(60), aamag–(60)]

As can be seen, the second conformation to be generated is the all-gauche(+) structure. At this point the f^* values of the nodes in the list are recomputed to take account of the fact that some of the structures may be closer to the second conformation than to the first. For example, the node g+g+mg+g– is closer in torsional space to the all-gauche(+) structure than to the all-trans structure and so is assigned a different (lower) torsional f^* value. The third conformation to be produced is g–g–mg–g–m, which has a torsional f^* value of 120° from the two structures generated.

AN ALGORITHM TO DIRECTLY IDENTIFY THE UPPER AND LOWER DISTANCE BOUND CONFORMATIONS

We now consider the problem of directly identifying the two conformations in which a specified interatomic distance achieves its minimum and maximum values. This is a problem of increasing importance due to the current interest in searching "three-dimensional" databases to identify molecules which might be able to satisfy a three-dimensional pharmacophore.[11] The efficiency of such searches can be greatly enhanced by the judicious application of appropriate screens. Thus, should the distance range permitted to a given pair of atoms in the target pharmacophore be outside the upper and lower bound distances of the corresponding atoms in a molecule, then that molecule can be eliminated from further consideration and a potentially time-consuming conformational search avoided. Our aim is to identify the conformation in which the distance between the two atoms adopts its maximum value and the conformation corresponding to the minimum value, without having to generate all conformations.

A number of approaches can be used to determine interatomic distance bounds, of which the best known is probably triangle smoothing. This procedure is an essential component of the distance geometry method for exploring conformational space.[12,13] However, the bounds matrix produced by triangle smoothing often does not represent the distance ranges that are possible in realistic, low-energy conformations. This is partly a consequence of the fact that triangle smoothing is typically applied only to the bounds matrix; for reasons of computational cost it is not usually applied to the distance matrix as well. In addition, it has been shown that three-dimensional objects must satisfy not only triangle inequalities but tetrangle, pentangle, and hexangle relationships as well;[13] an algorithm for performing tetrangle smoothing has been devised but is often not used due to the prohibitive computational effort required, particularly for large molecules.

In contrast to our original application of the A* algorithm,[4] which uses the intramolecular energy as the basis for calculating the f^* values, or the directed clustering algorithm, which uses a torsional measure of path cost, in the current problem we must express the f^* values in terms of the interatomic distance between a pair of atoms. The most direct sequence of bonds between the two atoms (the shortest path) is crucial to the calculation of these distance f^* values. In some cases more than one shortest path may be possible; we consider such situations below. Once again, we use heptane as our initial example. Let us consider how to generate the conformation in which the distance between the terminal carbon atoms is a maximum. The shortest bond path between the two atoms is determined; this consists of atoms 1, 2, 3, 4, 5, 6, 7. The units are joined in the order 1, 2, 5, 3, 4, 6, as before. When the root node in the search tree is expanded to give the three partial conformations corresponding to the three templates of unit 1, then the coordinates of the four carbon atoms 1, 2, 3, 4 are defined as are the coordinates of the hydrogen atoms bonded to carbon atoms 2 and 3. We need to estimate the maximum possible distance between atoms 1 and 7 subject to the constraints that are imposed by that part of the molecule already constructed. To do so, the algorithm first identifies those units which must be added in order to define the coordinates of the two atoms of interest. Units off the shortest path between the two atoms are ignored. Units 1, 2, 3, and 4 need to be present for the path between atoms 1 and 7 to be defined. "Reduced representations" of those units that must be added to define the coordinates of the two atoms are derived. Each reduced representation contains only those atoms that lie on the path, so for heptane, the reduced representations of units 2–4 contain just the carbon atoms (no hydrogens). A recursive procedure is used to perform a depth-first search over the different conformations available to these reduced representations. This enables the algorithm to calculate the distances that could be achieved between the two atoms of interest and thereby to determine the appropriate f^* value. Thus, for each template of unit 1 the algorithm examines the conformations that are available to atoms 5–7 using the reduced representations of units 2–4. This enables the possible distances between atoms 1 and 7 to be calculated and hence the maximum possible distance that could be achieved can be identified. In this simple case, the maximum distances are achieved with trans conformations of the remaining units. The f^* distance values for the trans, gauche(+), and gauche(−) templates of unit 1 are 7.6, 6.9, and 6.9 Å. The use of reduced representations ignores the possibility of unfavorable steric interactions that may arise when the actual units are joined together, but it considerably speeds up the calculation, particularly for larger units or for conformational entities (see below). Ultimately the all-trans conformation is produced. To determine the lower bound conformation, the list of nodes is ordered with the lowest f^* values at the head. In this case, conformations containing alternating g+g− combinations appear most promising (the two terminal carbon atoms in heptane can in fact be exactly superimposed using a g+g−g+g− sequence), but these are rejected due to high-energy steric interactions when the corresponding conformations are actually constructed. The first acceptable conformation is the g+amg+g+m structure, in which the distance between the end atoms is 4.8 Å.

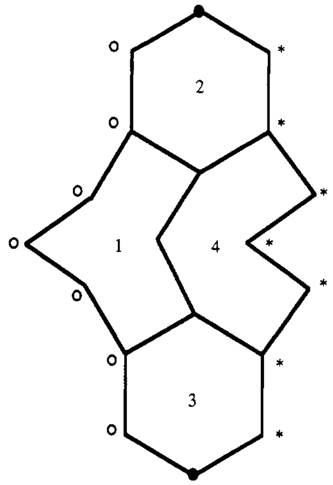


Figure 4. Four equivalent paths are possible between the two atoms (●), two of which are shown (* and O). Units 1 and 4 are cyclooctane units; units 2 and 3 are cyclohexane units.

We now consider an extension that naturally takes account of the potentially difficult situation when there is more than one shortest path between two atoms. This may arise when rings are present in the molecule. An example is shown in Figure 4. The difficulty is that during construction of the molecule some of the atoms on the chosen path may not be defined, yet an alternative path would be fully defined. For example, suppose (Figure 4) that the path indicated by "*" was chosen and that the molecule is constructed by joining the units in the order 1, 2, 3, 4. When units 1–3 have been joined together, the path "*" still contains undefined atoms, even though the path "o" would be completely defined. To alleviate these and other problems, which particularly arise for groups of fused and bridged rings, we employ a *conformational entity* description of the molecule. As originally defined,[14] a conformational entity contains one or more units, such that a group of fused and/or bridged cyclic units constitutes a single entity. Each acyclic unit and each isolated cyclic unit also constitute a single entity. Many advantages accrue from the use of conformational entities, such as the resolution of problems in strained conformations.[14] In the entity-based conformational search,[15] the conformational space of each individual entity is explored in isolation. Conformations of the whole molecule are then constructed by joining together the entity conformations. This is, in some respects, a "divide and conquer" problem-solving strategy, where the problem is broken down into a number of smaller problems whose solutions are then combined. One advantage provided by the entity-based search is that ill-fitting combinations of cyclic units are not repeatedly examined when exploring the search space, as would occur when a simple unit-based search is used. An obvious extension is to use entities which contain more than one connected acyclic unit and/or isolated cyclic unit, thereby enhancing the efficiency savings obtained from the divide-and-conquer principle. In our current version of the program, cyclic entities are identified as before. The remaining units are then collected into entities, subject to two requirements. First, an entity can only contain connected units. Secondly, a limit is imposed on the number of conformations permitted to each entity and the total number of combinations of unit templates for each entity is not allowed to exceed this limit. For example, if the maximum possible number of entity conformations is 20, then two entities would be defined for heptane: the first containing units 1, 2, and 5 (a total of nine possible combinations) and the second containing units 3, 4, and 6 (again, a total of nine possible combinations of unit conformations). Neither entity contains three units with three templates, as the number of possible conformations would then be 27 and so greater than the maximum permitted value of 20.

The use of entities alleviates the potential problems arising from the presence of more than one path, for when the coordinates of the atoms in one path are defined, then so too will the coordinates of the atoms in all other equivalent paths. One final extension that enhances the efficiency of the search is to recompute the order in which the units are joined for each interatomic pair; the units which define the path between the two atoms are considered first and then the other units in the molecule. For example, if we wanted to determine the minimum and maximum distances between the carbonyl oxygen of the carboxylic group and the methyl carbon in ochratoxin A, then those units/entities that lie off the path between the two atoms would only be joined once the units on the path have been successfully connected.

APPLICATIONS

Our aim in this section is to examine the utility of the two searching algorithms in automated conformational analysis. It should be noted that exactly the same conformations can be obtained using the previously described algorithms[4,5,15] for exploring the conformational space (e.g. depth-first search). Our objective here is to investigate if, and by how much, the new searching techniques are more efficient as measured by the computer processing time required. We anticipate that the search techniques will be particularly useful for flexible molecule which have a very large number of minimum energy conformations. Methods to deal with very flexible molecules are important because such cases require a disproportionate amount of time to explore their conformational space. All calculations were performed on a Silicon Graphics 50MHz R4000 Indigo workstation.

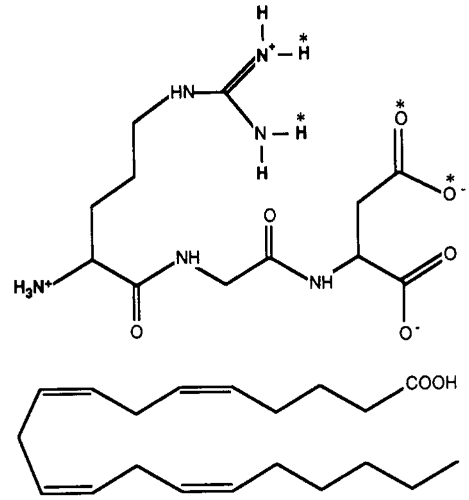


Figure 5. Arg-Gly-Asp tripeptide (top) and arachidonic acid.

As with other methods for exploring conformational space,[1,2] various user-selectable parameters determine how a conformational search is performed with COBRA and the criteria used to evaluate whether a conformation is "acceptable". Principal among these is the *pairwise interatomic close-contact ratio*: should the ratio of the sum of the vdw radii for any pair of atoms in a 1,*n* relationship ($n > 3$) to the interatomic separation of the atoms in a partial or full conformation exceed the threshold, then that conformation (or partial conformation) is rejected. This parameter was set to 2.0 in the current study. If the *template energy threshold* is activated, then higher energy templates (e.g. boat cyclohexane) are initially not used in the conformational search. Such higher energy forms are only considered if no acceptable conformations can be generated using their lower energy counterparts. In this study, however, no threshold was applied and all templates were available. The strain relief algorithms,[14] which are employed if no acceptable conformations can be produced using the default templates, are not at present applicable to the search methods based on the A* algorithm, and so molecules for which only strained conformations could be generated were not considered further. We have recently described the use of a distance geometry algorithm to generate conformations for parts of a molecule (e.g. large rings) that are not present in the fragment database and to construct conformations for highly strained ring systems which cannot be constructed by joining together rigid templates.[16] These algorithms were also not employed in the studies reported here.

We first consider the directed clustering search algorithm. To reiterate, this algorithm will produce a set of representative conformations for each molecule, and as such it combines conformational search and cluster analysis. We will consider three molecules as illustrative examples: ochratoxin A (Figure 1), arachidonic acid, and the tripeptide Arg-Gly-Asp, a sequence that is implicated in the binding of fibrinogen to activated platelets (Figure 5). We are not concerned with predicting the bioactive conformations of these molecules, but rather with characterizing the range of low-energy conformations possible. For each molecule a "normal" conformational search was performed using the default depth-first algorithm. A directed clustering search was also performed to find the five "most different" conformations according to the torsional criterion that we have described above. We summarize in Table 2 the results of these

Table 2. Summary of Directed Clustering Conformational Analyses of Ochratoxin A, Arachidonic Acid, and the Tripeptide Arg-Gly-Asp[a]

	depth-first search		Directed clustering search				
molecule	time	confs	time	diff 2	diff 3	diff 4	diff 5
ochratoxin A	25 s	216	18 s	75° (1)	51° (2)	50° (2)	50° (2)
arachidonic acid	102 min	>100 000	4 min	120 (1)	103° (1)	99° (1)	94° (1)
RGD	80 min	>100 000	22 min	132° (1)	113° (1)	103° (2)	100° (3)

[a] Column headed "confs" indicates how many conformations were obtained for the depth-first search. The search for arachidonic acid and RGD was terminated when 100 000 conformations had been generated. The improved leader clustering algorithm required approximately 8 s to cluster the 216 conformations of ochratoxin A. We did not attempt to cluster the 100 000 conformations of either arachidonic acid or the RGD tripeptide. Columns headed "diff 2"–"diff 5" indicate the distance to the closest conformation (given in parentheses) using the root-mean-square torsional measure (eq 1). See also Figures 6–8.

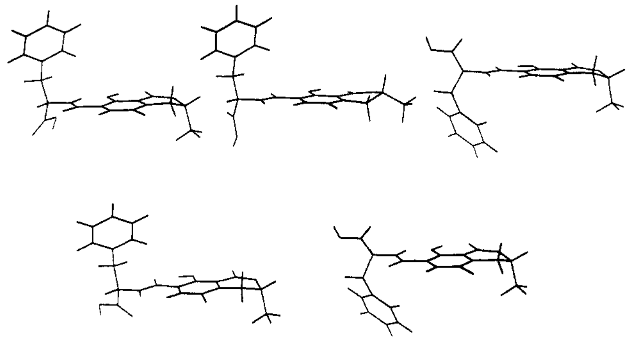


Figure 6. The five "most different" conformations found by the directed clustering search algorithm for ochratoxin A.

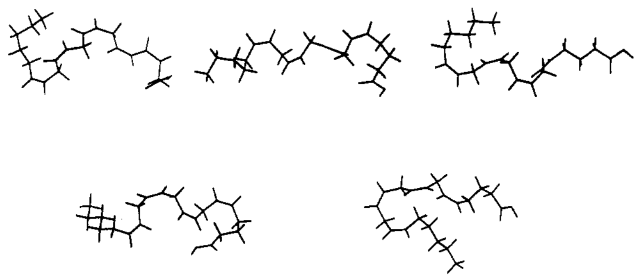


Figure 7. The five "most different" conformations found by the directed clustering search algorithm for arachidonic acid.

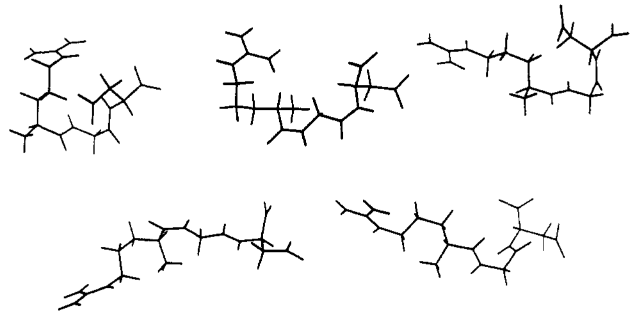


Figure 8. The five "most different" conformations found by the directed clustering search algorithm for the tripeptide Arg-Gly-Asp.

calculations. As can be seen, for ochratoxin A the depth-first computational search required slightly more cpu time than the directed clustering search, even without performing the cluster analysis. A very large number of conformations can be found for both arachidonic acid and for the RGD peptide. We did not attempt to cluster these structures; indeed, we estimate that to calculate the torsion angles in all the structures would require as much cpu time as to perform the directed clustering conformational search. It is clear that the directed clustering search is significantly more efficient than the alternative approach of identifying all conformations and then performing a separate cluster analysis in such cases. The torsional differences between the five conformations for each molecule are shown in Table 2, and the corresponding structures are drawn in Figures 6–8.

We now consider the distance bounds search. One potential application of this algorithm is in deriving the maximum and minimum distances that can be achieved between pairs of pharmacophoric groups for the molecules contained in a database. We searched the Cambridge Structural Database (CSD)[17] to find all entries that contained both a carboxylic acid bonded to a carbon atom (CCOOH) and an amide fragment (CONH or $CONH_2$). Compounds containing a metal (CSD class 4M) were ignored. Where more than one fragment was present in an entry, the largest molecular fragment was used. The crystallographic coordinates for the 423 molecules so identified were then converted into a form readable by COBRA; hydrogen atoms were added where necessary. For 95 molecules no coordinate information was present or was incomplete, giving a total of 328 molecules to analyze. We then applied our distance bounds searching algorithm to the problem of finding, for every possible pairing of a carboxylic acid fragment and an amide fragment in the 328 molecules, the two conformations corresponding to the maximum and minimum distances between the nitrogen of the amide fragment and the carbonyl oxygen atom of the carboxylic acid group. The stereochemistry of each molecule was taken to be that defined by the X-ray coordinates.

For 29 molecules the units contained in the database were insufficient to describe the molecule or only strained conformations could be generated. As noted above, COBRA contains algorithms specifically designed to deal with these situations, but these algorithms are not currently applicable to the A* bounds search. For six molecules the cpu time limit of 20 min was exceeded. The remaining 293 molecules contained a total of 443 different pairs of carboxylic acid and amide fragments. The average cpu time was 31 s, and the mean time was 4 s. By contrast, the average and mean times taken to perform a total search using the depth-first algorithm were 120 and 8 s. These figures demonstrate that, as anticipated, the bounds search is most useful for molecules with an extensive conformational space. The two graphs in Figure 9 show how the calculated maximum and minimum distances compare with the actual distances observed in the crystal structures. Figure 10 is a scatterplot of the maximum versus minimum distances for the 443 pairs of groups to indicate the range of distances present; the difference between maximum and minimum distances gives an indication of the degree of flexibility in a molecule.

For six molecules the crystallographic distance was smaller than the minimum or greater than the maximum by more than 0.5 Å. These discrepancies primarily arose for one of two reasons. The unit conformations sometimes did not correspond closely enough to the relevant conformations in the molecule. This often occurred when a precise unit required was not present in the unit library and the "adjusting" algorithms[16] were used to derive the conformations of the unit

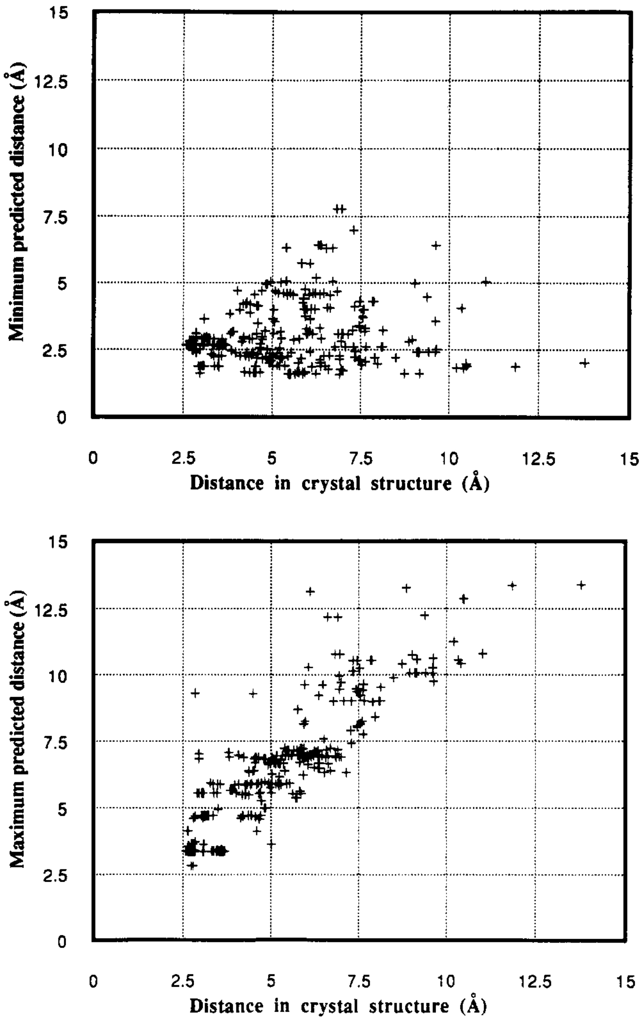


Figure 9. Scatterplots showing how the predicted minimum and maximum distances between the carbonyl oxygen of the carboxyl group and the nitrogen of the amide group compare with the actual distance in the crystal structure.

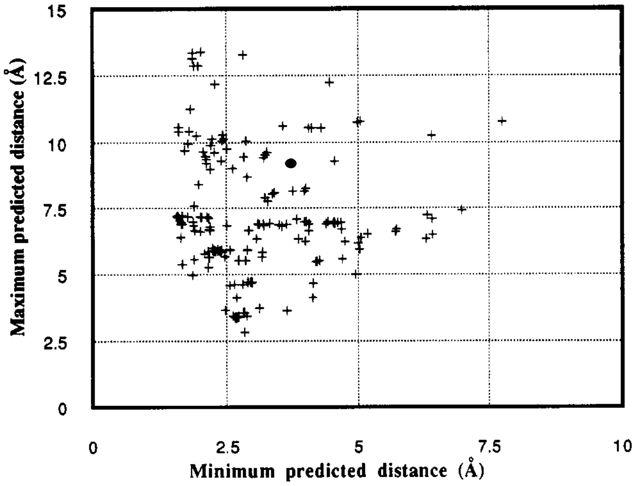


Figure 10. Comparison of the calculated minimum and maximum distances for 443 fragment pairs.

from those of a related unit. In other cases, the X-ray conformation did correspond to a combination of unit conformations, but the structure violated the close-contact ratio. No minimization is performed in our conformational search, and so it is sometimes possible that an initially high-energy structure might correspond to an accessible, low-energy conformation of the molecule. One obvious way to alleviate this is to employ a less stringent close-contact ratio.

We also determined the triangle-smoothed bounds for each of the acid/amide pairs. In all cases the range of distances was larger with triangle smoothing than was predicted by our conformational search, as it should be. The difference between the lower bound suggested by triangle smoothing and the minimum distance predicted by COBRA was less than 1 Å in 67% of cases, between 1 and 2 Å for 15%, between 2 and 3 Å for 8%, between 3 and 4 Å for 6%, and greater than 4 Å for 3%. The corresponding figures were 95%, 4%, 1%, 0%, and 0% for the upper bound. A molecule that exemplifies the apparent discrepancy for the minimum distance is 4-acetamidobenzoic acid, in which the triangle-smoothed lower bound between the amide nitrogen and the carbonyl oxygen of the carboxylic acid group is equal to the sum of the vdw radii, compared to 6.4 Å in both the crystal structure and the COBRA conformations. However, as pointed out above, the triangle-smoothed bounds cannot necessarily be achieved in a low-energy three-dimensional conformation (as is obtained from the EMBED algorithm[12,13]); bounds from triangle smoothing are a necessary but not sufficient restriction. It should also be pointed out that the triangle-smoothing algorithm is usually substantially faster to execute than a conformational search, and as such it does represent a useful procedure for the development of distance screens for database searching.[18]

DISCUSSION AND CONCLUSIONS

In this paper we have described two new applications of the A* search algorithm in automated computational conformational analysis and search. The algorithms have been implemented and described within the context of the COBRA program but should also be applicable to other methods for exploring conformational space. The directed clustering algorithm has been considered herein on the basis of a torsional measure for determining the difference between two conformations; this has the advantage of being particularly amenable to the calculation of the estimates (h^*) required by the A* algorithm. Other measures could also be employed, if an appropriate method for deriving the estimates is available. For example, it would be relatively straightforward to base the f^* values on a specific interatomic distance and thereby generate a series of conformations that spanned that distance. This would provide a way to directly calculate distance "bins", without necessarily having to explore the full conformational space. Other measures of the difference between two conformations may be more difficult to incorporate, due to the problem of accurately estimating how close a partially constructed structure could come to one of the previously generated conformations.

Knowledge of the upper and lower bounds of the interatomic distance between potentially pharmacophoric atoms is now regarded as very valuable information when screening three-dimensional databases. The most efficient screens are those that eliminate as many molecules as possible, without of course incorrectly excluding compounds. A key feature of the method described above for finding the upper and lower bound conformations is that the algorithm identifies energetically reasonable conformations of the *whole* molecule. In this respect it differs from procedures such as triangle smoothing. The fragment of the molecule between the two atoms of interest may represent a much less demanding conformational search problem. However, there are many cases where the part of the molecule that lies off the path between the two groups or that lies beyond one or the other of the two groups has a significant effect on the interatomic distance. Two significant restrictions on our approach to exploring conformational space must be emphasized. First, only conformations corresponding to combinations of the units present in the template library can be constructed (except when special algorithms[14,16] are activated). Secondly, COBRA can sometimes reject some combinations of unit conformations which might, after energy minimization, have resulted in a low-energy structure. Subject to these limitations, our bounds search algorithm provides a means of identifying the conformations that correspond to the upper and lower bounds of any specified interatomic distance. Finally, although we have considered only interatomic distances, the algorithm can easily be extended to include other pharmacophoric elements such as ring centroids.

ACKNOWLEDGMENT

The SERC is acknowledged for personal support under the Advanced Fellowship scheme and for providing computing equipment. Molecular graphics images were produced using MidasPlus.[19]

REFERENCES AND NOTES

(1) Howard, A. E.; Kollman, P. A. An Analysis of Current Methodologies for Conformational Search of Complex Molecules. *J. Med. Chem.* **1988**, *31*, 1669–1675.

(2) Leach, A. R. A Survey of Methods for Searching the Conformational Space of Small and Medium-Sized Molecules. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH Publishers: New York, 1991; Vol II, pp 1–55.

(3) Hartigan, J. A. *Clustering algorithms*; Wiley: New York, 1975.

(4) Leach, A. R.; Prout, K. Automated Conformational Analysis: Directed Conformational Search Using the A* Algorithm. *J. Comput. Chem.* **1990**, *11*, 1193–1205.

(5) Dolata, D. P.; Leach, A. R.; Prout, K. WIZARD: AI in Conformational Analysis. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 73–85.

(6) Leach, A. R.; Dolata, D. P.; Prout, K. Automated Conformational Analysis and Structure Generation: Algorithms for Molecular Perception. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 316–324.

(7) Leach, A. R.; Dolata, D. P.; Prout, K. An Investigation Into The Construction Of Molecular Models By the Template Joining Method. *J. Comput.-Aided Mol. Des.* **1988**, *2*, 107–123.

(8) Winston, P. H. Artificial Intelligence, 3rd ed.; Addison-Wesley Publishing Co.: Reading, MA, 1992; pp 63–100.

(9) Hart, P. E.; Nilsson, N. J.; Raphael, B. A formal basis for the heuristic determination of minimum cost paths. *IEEE Trans. SSC* **1968**, *4*, 100–114.

(10) Vinter, J. G.; Davis, A.; Saunders, M. R. The COSMIC molecular mechanics force field. *J. Comput.-Aided Mol. Des.* **1987**, *1*, 31–51.

(11) Martin, Y. C. 3D Database Searching in Drug Design, *J. Med. Chem.* **1992**, *35*, 2145–2154.

(12) Crippen, G. M. Distance Geometry and Conformational Calculations. *Chemometrics Research Studies Series 1*; Wiley: New York, 1981.

(13) Crippen, G. M.; Havel, T. F. Distance Geometry and Molecular Conformation. *Chemometrics Research Studies Series 15*; Wiley: New York, 1988.

(14) Leach, A. R.; Prout, K.; Dolata, D. P. The Application of Artificial Intelligence to the Conformational Analysis of Strained Molecules. *J. Comput. Chem.* **1990**, *11*, 680–693.

(15) Leach, A. R.; Prout, K.; Dolata, D. P. Automated conformational analysis: Algorithms for the efficient construction of low-energy conformations. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 271–283.

(16) Leach, A. R.; Smellie, A. S. A Combined Model-Building and Distance Geometry Approach to Automated Conformational Analysis. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 379–385.

(17) Allen, F. H.; Bellard, S. A.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Rogers, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. The Cambridge Crystallographic Data Centre: Computer-Based Search, Retrieval, Analysis and Display of Information. *Acta Crystallogr.* **1979**, *B35*, 2331–2339.

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

(19) Ferrin, T. E.; Huang, C. C.; Jarvis, L. E.; Langridge, R. The MIDAS Display System. *J. Mol. Graphics* **1988**, *6*, 13–27.