Journal of Computer-Aided Molecular Design, 8 (1994) 283–298
ESCOM

J-CAMD 241

Automated molecular design: A new fragment-joining algorithm

Andrew R. Leach* and Simon R. Kilvington

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

Received 1 September 1993
Accepted 26 October 1993

Key words: Computer-aided molecular design; Structure-based design; Random tweak algorithm; Drug design; Conformational analysis

SUMMARY

A popular first step in the problem of structure-based, 'de novo' molecule design is to identify regions where specific functional groups or chemical entities would be expected to interact strongly. When the three-dimensional structure of the receptor is not available, it may be possible to derive a pharmacophore giving the three-dimensional relationships between such chemical groups. The task then is to design synthetically feasible molecules which not only contain the required groups, but which can also position them in the desired relative orientation. One way to do this is to first link the groups using an acyclic chain. We have investigated the application of the 'tweak' algorithm [Shenkin, P.S. et al., Biopolymers, 26 (1987) 2053] for generating families of acyclic linkers. These linking structures can subsequently be 'braced' using a ring-joining algorithm [Leach, A.R. and Lewis, R.A., J. Comput. Chem., 15 (1994) 233], giving rise to an even wider variety of molecular skeletons for further studies.

INTRODUCTION

The number of biologically active macromolecules whose 3D structures have been determined using X-ray crystallography or NMR is increasing at an exponential rate [1]. The detailed structural information provided by such studies is extremely useful, particularly in the design of novel inhibitors [2]. Two general approaches to the problem of automated molecular design can be distinguished. One can try to identify a previously synthesised molecule which would be expected to interact with the receptor. This approach is usually implemented by searching a 'three-dimensional' database [3]. Alternatively, one can attempt to design entirely new molecules 'from scratch'. Details of a number of programs that fall into this category have been published, including (in alphabetical order) BUILDER [4], CONCEPTS [5], GENSTAR [6], GROUP-

*To whom correspondence should be addressed.

BUILD [7], GROW [8], LEGEND [9], LUDI [10], SPLICE [11], SPROUT [12], TORSION [13] and the programs of Lewis and Dean [14,15] and Lewis [16].

When a detailed 3D structure of the site is available, a popular first step in such de novo design is to identify regions where specific functional groups or chemical entities would be expected to show a strong interaction. This method was pioneered by Goodford and is implemented in his GRID program [17]. A more recent variant is MCSS [18]. If no high-resolution structure of the receptor is available, then it may be possible by exploring the conformational space available to known inhibitors to deduce which functional groups should be present in a candidate molecule and to deduce constraints on the distances between these functional groups [19]. In either case, the starting point of the design process is a number of functional groups distributed in Cartesian space. These must then be combined in some way to give a chemically reasonable molecule. A powerful first step is to link the fragments with acyclic chains. These then constitute basic molecular skeletons for subsequent stages in the design process. We have recently developed an algorithm which can add rings to such chains, to give even more molecular skeletons which have a smaller entropic penalty and may be synthetically more attractive [20].

The problem of generating acyclic chains which link together molecular fragments is an example of a more general challenge in molecular modelling: to determine in which conformation(s) an acyclic chain can satisfy one or more imposed geometric constraints. This problem arises in a number of situations, such as the modelling of loops in protein homology modelling, the conformational analysis of cyclic molecules, and when exploring '3D' databases to identify possible lead compounds. In most of these applications the constraints to be satisfied are expressed as interatomic distances or distance ranges, though other types of constraint may also be imposed. For example, when modelling protein loops any proposed conformation should not clash with the remainder of the protein. In some of these problems the chemical constitution of the chain is predetermined; for molecular design, however, we should consider a wide range of chemically different linkers. Lewis and co-workers have previously described a number of algorithms for use in site-directed ligand design. Lewis and Dean initially proposed the use of 2D 'spacer skeletons' [14,15]. Lewis subsequently extended the method into three dimensions using diamond lattices [16]. As recognised at the time, however, few 'real' molecules fit exactly on regular lattices. Lewis [13] has also employed the Gō–Scheraga ring closure algorithm [21] and the Bruccoleri–Karplus variant [22], which give analytical expressions for the dihedral values that will bring the ends of a chain some desired distance apart. When the chain contains more than six rotatable bonds, the Gō–Scheraga algorithm cannot provide a solution; in such cases it is necessary to explore the additional degrees of freedom in some other way (e.g., using a systematic search). Lewis, Roe, Kuntz and co-workers have also developed an approach [4] which explores an irregular lattice derived by docking a large number of molecules into the site. This algorithm requires that the structure of the receptor is available. However, many molecules of biological interest have resisted the efforts of X-ray crystallographers and NMR spectroscopists and thus frequently a pharmacophore is the only information available.

METHODS

In this paper we describe an alternative approach to the problem of linking molecular fragments, which is applicable both when the receptor is available and when it is not and which can

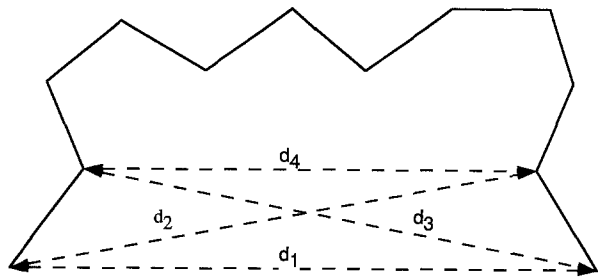

Fig. 1. Four interatomic distance constraints, derived between the two pairs of atoms at the ends of an acyclic chain.

deal with any length of chain. Our approach uses the 'tweak' algorithm of Shenkin et al. [23,24]. This algorithm provides the changes in dihedral angles that will enable an acyclic chain to satisfy distance constraints between pairs of atoms along its length. Shenkin et al. used the tweak algorithm to generate a wide variety of backbone conformations of hypervariable loops for modelling antibody structures. The essence of their approach is as follows. The starting point is a conformation of the peptide backbone in which the backbone dihedral angles ϕ and ψ have been assigned random values. This random conformation will invariably fail to satisfy the imposed geometric constraints (i.e., that the end points of the loop should coincide with the appropriate atoms in the fixed part of the antibody, Fig. 1). The 'tweak' algorithm is then used to iteratively change the dihedral angles, so that the loop structure does meet the requirements. The tweaking algorithm uses a linearised Lagrange multiplier method which keeps the changes in the dihedrals to a minimum, whilst still meeting the constraints. Given an axis characterised by a unit vector $\hat{\boldsymbol{\theta}}$ and a point not on the axis whose distance to a second fixed point is d, then the variation of d with rotation about the axis is given by:

$$\frac{dd}{d\theta} = \left(\mathbf{r} \times \hat{\boldsymbol{\theta}}\right)\hat{\mathbf{d}}$$

Here, $\mathbf{r}$ is the radius vector of the moving point from a point on the rotation axis, θ is the magnitude of the rotation and $\hat{\mathbf{d}}$ is the unit vector from the moving point to the fixed point (Fig. 2). It is then possible to derive the following expression for the changes $\Delta\theta_i$ in each rotatable bond i which will enable the distance constraints to be satisfied:

$$\Delta\theta_i = \mathbf{D}\left(\mathbf{D}^T\mathbf{D}\right)^{-1}\Delta d_j$$

where

$$\mathbf{D}_j^i = \frac{\partial d_j}{\partial \theta_i}$$

$\mathbf{D}$ is an $m \times n$ matrix, where m is the number of distance constraints, $[\Delta d_j]$ is the matrix giving the errors between the current and required distances and n is the number of rotatable bonds. For the full derivation we refer the reader to the original papers [23,24]. The matrix to be in-

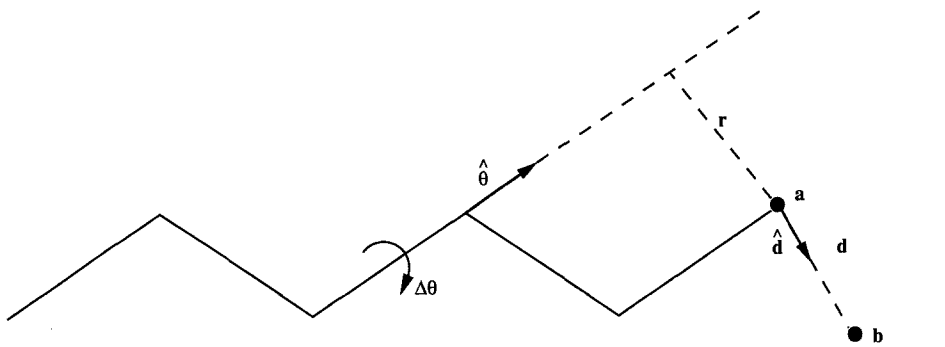

Fig. 2. Graphical illustration of the tweak algorithm: the aim is to move the end of the chain (point a) to the fixed point b by rotating about the bond shown.

verted, $\mathbf{D}^T\mathbf{D}$, is of size $m \times m$ and so the tweak algorithm alone approximately scales with the number of constraints to be satisfied rather than with the number of rotatable bonds. For the antibody modelling studies, the chain was of a predetermined constitution (equivalent to the peptide backbone of the hypervariable loop) and four sets of interatomic distance constraints were used for each chain, corresponding to all possible pairs of distances between the two pairs of atoms at the ends of the chain (Fig. 1). The conformations were checked after tweaking to ensure there were no high-energy atomic overlaps, either internally or with the rest of the protein model. The cyclic nature of the angular degrees of freedom is lost when deriving the constraint equations and so values of $\Delta\theta$ greater than 360° can sometimes be specified. The results above also rely on the approximation that the $\Delta\theta_i$ are small. Shenkin et al. therefore limited the maximum change in any $\Delta\theta_i$ to 10° and repeatedly applied the tweak algorithm until the chain satisfied the constraints. In the studies reported in this paper we also restrict the change in any dihedral for any iteration to this same maximum value.

The tweak algorithm applied to general acyclic chains

When linking molecular fragments, we may need to construct more than one chain because there may be more than two fragments. We may want to generate chains which contain a variety of atom types and bond orders, so as to introduce greater diversity into the molecules eventually suggested. We may wish to impose constraints on the torsion angles, so that the resulting structures will be of low energy. We will wish to avoid unfavourable interactions with the receptor. In this paper we discuss how the basic tweak algorithm can be incorporated into a general approach to tackle these varied problems. One problem faced by all approaches to de novo design is that the set of possible molecules is essentially infinite and it would require infinite computation to find all solutions. It is therefore necessary to have some means of directing the search towards the more 'reasonable' solutions. This introduces an element of subjectivity into the problem. Our objective is to provide a set of tools that can be used in both 'automated' and 'interactive' modes. The software therefore contains default modes of operation, as will be discussed below, but these defaults can be extended or overridden at any stage when used interactively.

First, we wished to investigate the effectiveness of the tweak algorithm when applied to acyclic chains, typical of those that might be used to link fragments of organic molecules. Our approach

was to systematically generate all possible acyclic chains containing between four and nine bonds. The bond orders allowed were single, double, triple and aromatic. Our initial investigations were limited to chains containing just carbon atoms, with the bond lengths and angles being assigned standard values [25]. Chemically unsatisfactory chains were rejected; no consecutive unsaturated bonds were permitted, except for a sequence of aromatic bonds (i.e., no C=C=C, C=C≡C etc.). Chains symmetrically equivalent to one previously generated (e.g., C–C=C–C–C and C–C–C=C–C) were also discarded at this stage. We chose to use four constraints between the four end atoms (i.e., between the following pairs of atoms: 1,N; 2,N; 1,(N–1); 2,(N–1)) of the chain as in Fig. 1, rather than just a single constraint between the two end atoms (1,N). This ensures that the geometry of the fragment will be better preserved. For example, when joining to a benzene ring, the position of the next atom in the chain can be determined from the coordinates of the adjoining atoms in the ring. The distances between the four end atoms are unchanged by rotation about the two bonds at the ends of the chain and so are largely unaffected by the nature of these two end bonds (between atoms 1,2 and N,(N–1)). The two end bonds were thus assigned an order of one in the investigations reported in this section. Thus, for a chain with b bonds we considered a maximum of $(b-2)^4$ possible chains, given the four different bond orders. The rotatable bonds were identified as the single and triple bonds in the chain, which was initially set to an extended conformation. A second conformation of the chain was then generated by changing the torsion angle of each rotatable bond in the chain by a random amount. The four sets of interatomic distances between the two pairs of atoms at the ends of the chain in the new conformation were determined. The tweak algorithm was then used to try and modify the original (extended) conformation to satisfy the constraints within some tolerance. Subsequently, the second conformation was randomly changed to give a third conformation and the tweak algorithm applied to the second conformation to try and satisfy the constraints imposed. This procedure was repeated 100 times for each chain.

In our investigations, the acceptable tolerance for each of the four interatomic distance constraints was permitted to be 0.1 or 0.25 Å; the maximum torsional deviation for each rotatable bond was $\pm 180°$, $\pm 90°$ or $\pm 30°$; the maximum number of tweak iterations was set to 10, 50 or 100; and the constraints were applied simultaneously as well as in a sequential manner. When applying the constraints sequentially, the constraint between the two end atoms (1,N) was required to be satisfied first, then the 2,N, 1,(N–1) and finally 2,(N–1). If any single constraint failed, the trial was abandoned. No van der Waals (vdW) checks were performed on the chains, as we only wished to investigate the ability of the algorithm to satisfy the constraints, rather than necessarily generating a 'chemically' reasonable structure. The results are presented in Tables 1 and 2 as the proportion of 'tweaks' which successfully generated a conformation that satisfied the constraints, given as an average over the 100 trials for each distinct set of bond orders. We also confirmed that the results were unaffected by randomly translating and rotating the chain before 'tweaking'. As can be seen from the tables, the success rate of the tweak procedure is enhanced by a small maximum angular deviation, a large constraint tolerance, a large maximum number of iterations, and by applying the constraints sequentially rather than all at once.

We measured the average time taken to tweak chains with varying numbers of bonds and with various maximum numbers of iterations of the tweak loop. This showed that the time required depends linearly on the number of bonds. This is because most time in our implementation is spent twisting bonds in the chain, rather than performing the tweak calculation. The time also

TABLE 1
THE EFFECT OF VARIOUS PARAMETERS ON THE FAILURE RATE OF THE TWEAK ALGORITHM: ALL CONSTRAINTS TO BE SATISFIED AT ONCE

Chain length	Total number of chains	Failure rate (%) at various constraint tolerances (Å)					
		0.1[a]	0.25[a]	0.1[b]	0.25[b]	0.1[c]	0.25[c]
Maximum deviation = 30°							
4	4	23	3	19	1	22	1
5	15	18	1	21	2	18	2
6	29	28	6	29	5	30	5
7	77	37	9	36	9	35	9
8	169	44	13	44	14	45	14
9	430	48	19	48	18	47	18
Average		33	9	33	8	33	8
Maximum deviation = 90°							
4	4	50	30	48	19	52	24
5	15	54	25	56	26	57	25
6	29	66	38	68	38	66	38
7	77	74	42	74	47	74	46
8	169	82	56	80	54	80	54
9	430	84	62	82	59	81	58
Average		68	42	68	41	68	41
Maximum deviation = 180°							
4	4	64	46	63	48	62	45
5	15	70	49	70	49	70	49
6	29	80	62	80	60	78	60
7	77	87	70	86	68	86	68
8	169	92	78	91	75	90	75
9	430	93	81	91	78	90	76
Average		81	64	80	63	79	62

[a] Maximum number of iterations is 10.
[b] Maximum number of iterations is 50.
[c] Maximum number of iterations is 100.

depends in a linear fashion on the maximum number of iterations, as chains that cannot be tweaked to satisfy the constraints will be subjected to the maximum number of tweak iterations before being rejected.

A given chain may be able to satisfy the distance constraints in more than one possible conformation. By performing multiple iterations of the tweak algorithm, using different starting chain conformations, we aim to cover the space of possible conformations. Therefore, when a large number of iterations are used, it is commonly observed that many of the resulting conformations are very similar. Under such circumstances it would clearly be inefficient to use every identical chain in the subsequent stages of the design procedure. The conformations are therefore subjected to a clustering analysis to identify a representative family of structures. We currently apply the 'unweighted pair-group method using arithmetic averages' (UPGMA) algorithm [26] to do so, using as the measure of difference between two conformations the following root-mean-square torsional measure:

TABLE 2
THE EFFECT OF VARIOUS PARAMETERS ON THE FAILURE RATE OF THE TWEAK ALGORITHM: CONSTRAINTS APPLIED SEQUENTIALLY

Chain length	Total number of chains	Failure rate (%) at various constraint tolerances (Å)					
		0.1[a]	0.25[a]	0.1[b]	0.25[b]	0.1[c]	0.25[c]
Maximum deviation = 30°							
4	4	5	2	5	2	1	1
5	15	9	1	5	1	7	1
6	29	9	2	7	2	6	2
7	77	10	4	8	2	6	1
8	169	11	4	7	2	6	2
9	430	11	9	7	2	6	2
Average		9	4	7	2	5	2
Maximum deviation = 90°							
4	4	29	11	10	5	10	4
5	15	37	16	23	6	22	7
6	29	42	21	27	9	24	9
7	77	27	27	26	8	23	7
8	169	54	34	29	11	25	9
9	430	56	37	30	12	26	9
Average		44	24	24	9	22	8
Maximum deviation = 180°							
4	4	36	21	11	6	12	5
5	15	47	28	30	10	31	8
6	29	54	36	34	14	32	13
7	77	60	41	34	13	32	10
8	169	69	50	38	17	34	13
9	430	71	53	39	18	34	13
Average		56	38	31	13	29	10

[a] Maximum number of iterations is 10.
[b] Maximum number of iterations is 50.
[c] Maximum number of iterations is 100.

$$d_{ij} = \sqrt{\sum_{k=1}^{N} (\tau_{ik} - \tau_{jk})^2 / N}$$

Here, d_{ij} is the distance between the two conformations i and j, and $(\tau_{ik} - \tau_{jk})$ 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. The UPGMA algorithm first calculates the difference between all conformations. The closest two structures are then merged into a single cluster. The distance from all the other clusters (which only contain a single structure) to this cluster is then recalculated as the average of the distances to its two members. The next closest pair of clusters is identified and interatomic distances recomputed. The distance between any pair of clusters i and j is in general given by:

$$d_{ij} = \frac{\sum_{k=1}^{N_i} \sum_{l=1}^{N_j} d_{k,i} d_{l,j}}{N_i N_j}$$

where N_i and N_j are the numbers of conformations in clusters i and j. The clustering proceeds until the size of any cluster (as measured by the largest difference between any pair of structures in a cluster) exceeds a predetermined value and/or the number of clusters falls below a specified maximum value. A representative structure must then be chosen from each cluster. We currently choose the structure that is closest to the 'average' conformation in each cluster, where the torsion angles in the 'average' conformation are the average values in the structures contained in the cluster.

Generating the initial chains

Next we consider how to generate the initial chains which will subsequently be 'tweaked'. For the purposes of this section we assume that we are trying to join just two fragments and that the four atoms (two from each fragment) which will be at the ends of the chain have already been identified. The implementation of these two steps is described below. The generation of each initial chain requires two distinct operations. Firstly, it is necessary to determine what length of chain to use (i.e., how many bonds) and what the bond orders and atom types should be (which determines the bond lengths and angles). Secondly, we must generate an initial conformation for the chain, which will then be modified to satisfy the constraints.

For a given distance to span there will clearly be a minimum chain length, which can be derived by considering the end-to-end distances of extended chains. There is no limit to the number of bonds that could be present in the chain, giving conformations from 'almost extended' to 'very bent' (Fig. 3). The default is to use chains of the smallest length (i.e., as near to the extended form as possible). The smallest length chain should have the smallest deviation from the constraints and so the changes in the torsion angles should be relatively small. As we showed above, the tweak algorithm is most effective when the maximum tweak angle is small, which will be the case when the initial structure is close to the final structure. If chains are to be generated systematically, then the minimum-length chain will have the fewest atoms and so the space will be searched most rapidly. If a random search is used, then fewer trials will be needed to produce a given coverage of the expected space. However, a curved chain might be needed to avoid interactions with the receptor. We describe below how we tackle this particular problem. If all chains of the minimum length are not able to join the fragments (i.e., such that the constraints are satisfied within the specified tolerances) or if all of the chains initially suggested are deemed unacceptable later in the design process, then longer chains can be manually specified by the user.

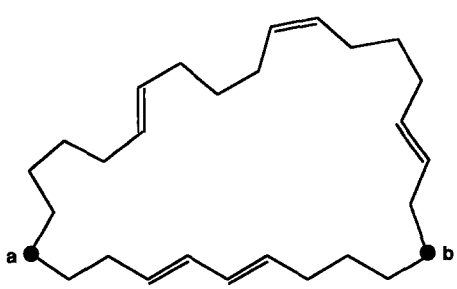

Fig. 3. Fragments (labelled a and b) can be linked using chains ranging from 'almost straight' to 'very bent'.

Deciding which fragment atoms to connect, and how to join more than two fragments

Many fragments contain more than one atom. For a given pair of fragments, it is thus necessary to decide which atoms should be joined together. The default approach is to use the pair of atoms, one from each fragment, which are closest in space. When we want to join together more than two fragments, there may be a number of ways to join the fragments together. This is very similar to the problem faced in cluster analysis, where groups of objects are merged until eventually a single cluster is obtained [26,27]. Just as there are a number of algorithms that can be used to form the clusters in such an analysis, there are also a number of ways in which the fragments could be joined together. The simplest nontrivial case contains three fragments. The fragments may be approximately linear, in which case perhaps the most obvious and sensible option would be to join them in a linear fashion (Fig. 4). The fragments may also lie approximately at the corners of an equilateral triangle. In this case we could join the fragments along the edges of the triangle. We could first join two of the fragments and then connect the third fragment, possibly to part of the linking chain. We could also identify a point near the centre of

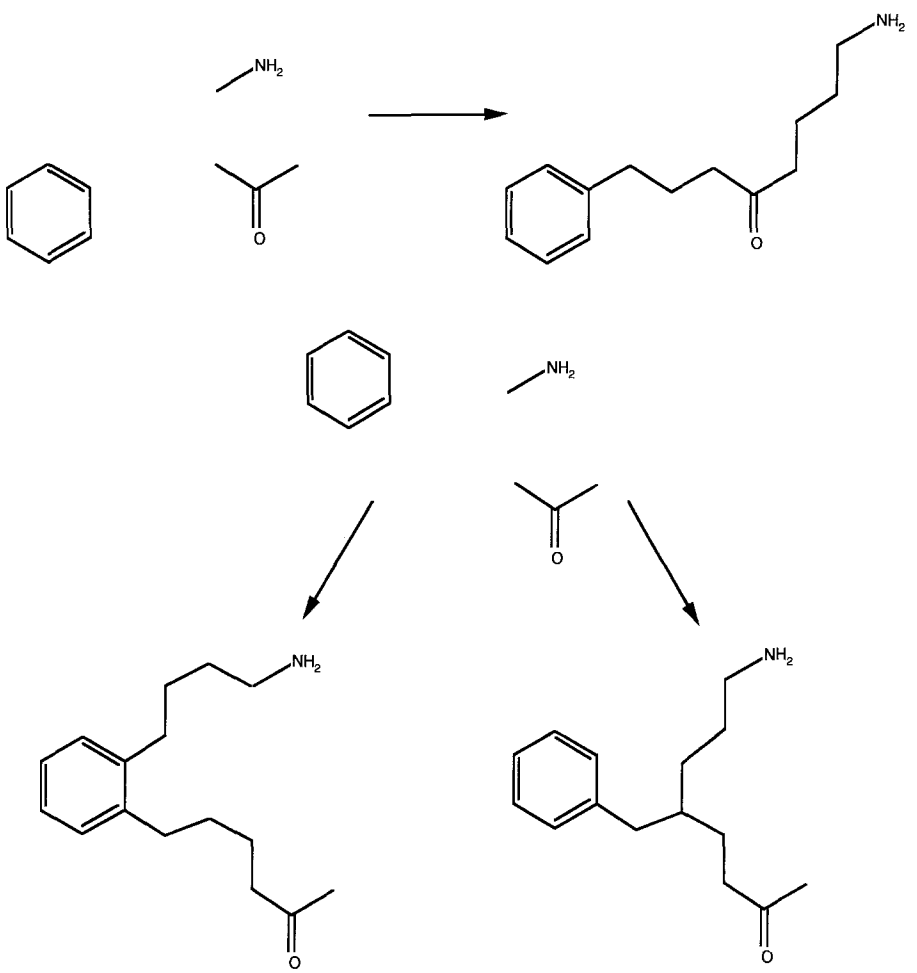

Fig. 4. Several ways in which three fragments can be joined together, depending on their relative orientations.

the triangle and then construct chains to each of the fragments (Fig. 4). There is no 'best' solution. Our default approach is analogous to the single-linkage cluster algorithm. First, the closest pair of fragments, i.e., the two fragments with the smallest atom–atom distance, is identified. These two fragments are then joined to give a larger composite fragment. If it proves impossible to join the two atoms then the next closest pair (not necessarily from the same pair of fragments) is considered. The new set of fragments (now containing one composite fragment) is then analysed to identify the closest pair of fragments, which are then joined, and so on. At each stage more than one chain may be possible, giving rise to a variety of solutions. As with some of the other steps described above, there is always the possibility of choosing to join different pairs of atoms in the fragments and of joining the fragments in a different order, should the initial suggestions be subsequently deemed unacceptable.

Taking account of interactions with the receptor site

A linking chain should not interact unfavourably with the receptor site. In some cases a near-extended chain will not satisfactorily join together two fragments, due to interactions with the receptor. Under such circumstances a number of approaches are possible. We could generate longer chains and hope that the tweak algorithm would produce a conformation that fits around the protrusion. However, this might require a considerable computational effort until a satisfactory chain is obtained. We have therefore investigated how the algorithms described above might be used to construct linking chains which are composed of a number of smaller chains. The latter would themselves link together points in the receptor site, chosen so that straight lines drawn between successive points do not intersect the accessible surface of the receptor (Fig. 5). The problem then is to determine which additional points to use. We have invented a recursive procedure to do this. First, the algorithm determines whether a straight line connecting the two fragments intersects the accessible surface of the receptor. If the line does not intersect the surface, then the two fragments are connected as above and the resulting chains are checked for nonbonded interactions with the site. However, if the line does intersect the accessible surface, then the points where the line intersects the surface are determined. These points are taken in pairs along the line. For each pair of points the mid-point is calculated. A disc passing through the mid-point and perpendicular to the line is swept out. The point on the accessible surface is identified where the distance from the line to the surface is minimal (Fig. 6). This new point is projected out along the normal by the length of a carbon–carbon bond from the surface to give

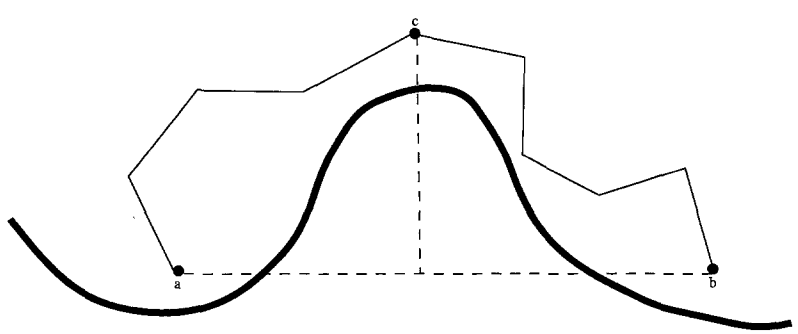

Fig. 5. When unfavourable interactions with the receptor are present, the chain is constructed from a network of smaller fragments which skirt around the receptor surface.

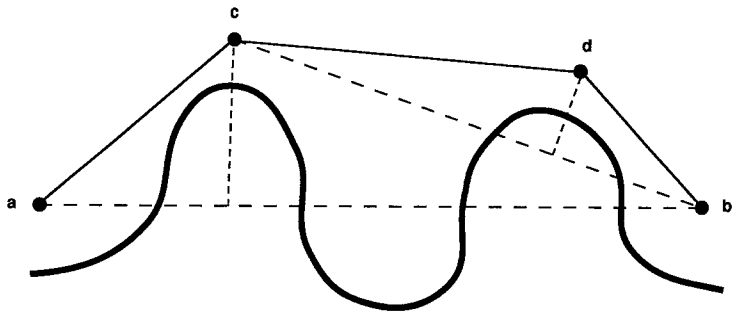

Fig. 6. The algorithm to determine intermediate points to avoid interactions with the receptor surface. First, point c is constructed, and then point d. The line a-c-d-b then skirts the protrusions in the receptor site. (– – –) = shortest path between original points; (- - - - -) = lines used to add additional points; (——) = final path used to connect original points.

a third point, intermediate between the original two. Lines are then drawn between this third point and the two original ones and the procedure is applied again in a recursive fashion. Ultimately a network of points is created such that no line connecting any pair of successive points intersects the accessible surface (Fig. 6). Whenever a second, third etc. point is required, the search is restricted so that the angle between the two planes, one containing the new point and the other containing its predecessor point, does not exceed a relatively small angular deviation (currently $\pm 60°$).

RESULTS

We describe here the application of the algorithms described above to the generation of linking chains in the active site of citrate synthase. The X-ray crystal structure of this enzyme, with coenzyme A and citrate bound, has been determined [28]. We chose the following four groups from the structure of the bound CoA: the adenine ring system, the phosphate group, the pyrophosphate group, and the terminal amide group (Fig. 7). These four fragments were then connected, using our joining algorithm.

Three chains are required to join four fragments together. Initially, we chose to connect the atoms which are themselves joined in the parent CoA molecule. We generated chains containing just carbon atoms, using at each step a systematic algorithm which produced all chemically sensible sequences of bond orders for a given chain length in an initially extended conformation. Each of these chains was then subjected to 1000 attempts of randomisation and tweaking.

Three chemically different linkers were able to provide satisfactory structures: an all-sp^3 chain, a chain with one double bond and a chain with one aromatic bond. In all three cases, structures similar in conformation to the actual link passing through the sugar were obtained. In addition, for the all-sp^3 chain an alternative family of structures was produced. For the link between the pyrophosphate group and the terminal amide a large number of solutions were found. The default link contains fewer bonds than are present in the actual molecule, so none of the generated chains exactly corresponded with the observed structure. If a link of the appropriate length was used, then chains corresponding to the crystal conformation were obtained. However, for the link between the phosphate and pyrophosphate groups solutions were found only for the all-sp^3 chain, none of which was particularly close to the structure observed in the conformation of

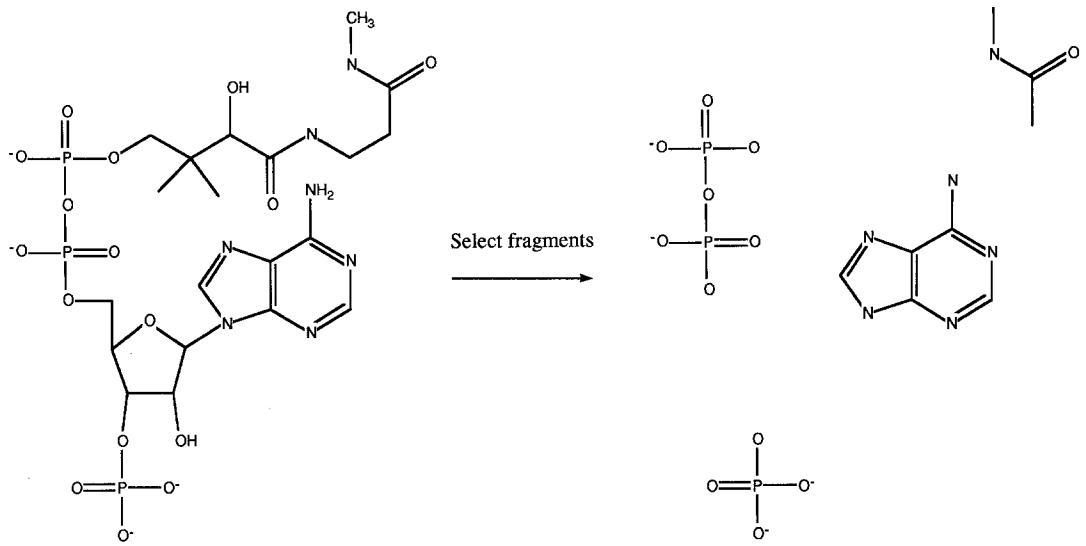

Fig. 7. The four functional groups extracted from CoA which provide a starting point for testing the connecting algorithm.

CoA. In this case the conformation(s) of each linking chain depends critically on its geometry. To verify this, we attempted to join the pyrophosphate and phosphate groups again, using the geometry in the crystal structure (in which the bond lengths and angles deviate from their ideal values) and we were able to obtain a conformation much closer to that observed.

We then subjected the four fragments to our linking algorithm, using the default approach described above. Thus, the algorithm first attempts to join the closest pair of fragments to give a composite fragment. After clustering the linkers, the process is repeated until all the fragments have been joined. We set the maximum number of tweak iterations to 10 and the constraint tolerance to 0.25 Å. We considered the effect of the clustering criteria by restricting the maximum number of clusters at each stage to between 1 and 20 chains. The calculations were performed on a Silicon Graphics R4000 Indigo computer. The total number of solutions obtained in each case, and the CPU time required, are given in Table 3. A wide variety of structures were obtained, due in part to the many different ways in which the fragments can be linked together. The apparent anomaly in Table 3, i.e., the fact that the number of structures does not increase when the maximum number of clusters is increased from six to eight, is due to the selection of different 'representative' conformations for each cluster. This is not very surprising; we would anticipate that selecting different representative conformations may ultimately give different results. We show in Fig. 8 two of these structures, to demonstrate the wide variety of structures that can be obtained using the approach.

DISCUSSION AND CONCLUSIONS

We have described in this paper a series of algorithms by which a set of molecular fragments distributed in space may be linked together. Our technique is an alternative to the methods

TABLE 3
THE NUMBER OF CHAINS OBTAINED, AND THE TOTAL CPU TIME REQUIRED, TO CONNECT THE FOUR FRAGMENTS OF CoA

Maximum number of clusters	Number of linking chains found	CPU time (s)
1	24	8
2	651	79
3	968	147
4	1032	158
5	1922	846
6	1935	912
7	1915	966
8	1934	948
9	2133	1408
10	2568	2507
20	3180	4866

Eleven experiments were performed, in which the maximum number of clusters permitted at each stage was increased from 1 to 20.

already proposed by Lewis and Dean [14,15], Lewis [16], and Lewis, Roe, Kuntz and co-workers [4]. If an appropriate set of parameters is used, the tweak algorithm provides an efficient method of adjusting an acyclic chain to satisfy distance constraints and it can therefore be used to connect molecular fragments. There is no restriction on the chain length, nor is the method dependent on having available an atomic structure of the receptor. The tweak algorithm itself scales according to the number of constraints rather than the length of the chain and is also able to provide a variety of solutions (providing they exist) if more than one starting conformation is used. The resulting structures span a wide range of molecular scaffolds for subsequent steps in the design of realistic molecules.

In this paper we have concentrated on the generation of all-carbon chains. However, we would not wish to restrict the atom types in any molecule ultimately suggested to just carbon atoms. It would be straightforward to include other atom types when suggesting possible linkers; the tweak algorithm is extremely fast and it is feasible to envisage a systematic search over both bond orders and atom types, particularly if some simple chemical rules were incorporated (such as the DENDRAL 'BADLIST' [29] which discards combinations such as OOO). An alternative is to

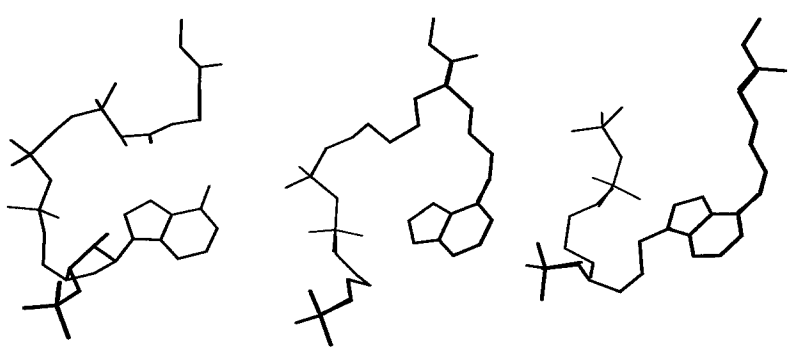

Fig. 8. Two structures which connect the four fragments from CoA in different ways, as generated by the connecting algorithm. CoA is shown on the left in a similar orientation.

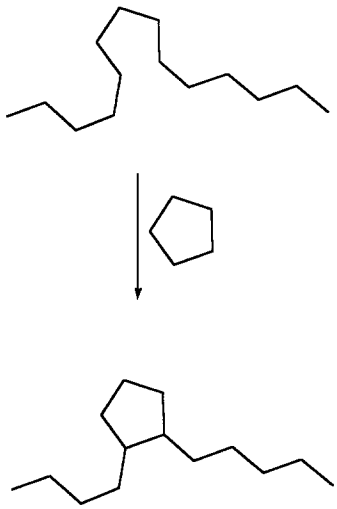

Fig. 9. An acyclic chain which is of high internal energy, but which can be braced with a ring to give a low-energy structure.

regard the all-carbon chains as generic structures in which the hybridisation is specified at each node, but not the precise atom type. This philosophy is similar to our use of generalised templates in the WIZARD-II [30] and COBRA [31] programs for exploring conformational space. When an atom type other than carbon is required, the bond lengths and sometimes the bond angles will be different. Changing the bond lengths and angles from their original all-carbon values will often mean that the new chain does not satisfy the constraints. We have found that one or two further iterations of the tweak algorithm are often sufficient to modify the initial generic structure to produce an acceptable conformation with the new atom types. This was done by generating 500 random conformations of all-carbon chains containing seven bonds and then randomly selecting an atom which was changed to N, O, P or S. The bond lengths (and bond angles where necessary) were changed to the values appropriate to the new chain and a maximum of 10 iterations of the tweak algorithm was performed to try and resatisfy the distance constraints between the ends of the chain. The proportions of chain which could satisfy the original constraints and which were 'similar' to the original chain (i.e., for which the rms difference was less than 0.25 Å) were 96% for N, 51% for O, 21% for P and 17% for S. It can thus be seen that, where large changes in bond lengths and/or angles are required (i.e. for P or S), significantly different conformations can be required to satisfy the constraints. Under these circumstances it might be preferable to generate a new family of skeletons from scratch. This was evident in our study of CoA, where it proved impossible to reproduce the skeleton which links the phosphate and pyrophosphate groups using all-carbon chains with ideal geometries.

The tweak algorithm is a purely geometrical method and thus does not consider the energetics of the linking chains. We have, however, developed a variant in which the change in each dihedral is limited according to its torsional potential and in which the initial chain is at an energy minimum. Thus, for example, an amide bond would be limited to a relatively small change in dihedral angle, whereas a bond with a smaller barrier to rotation would be permitted a larger torsion value. There has been some debate on the importance of the conformational energy

surface of the isolated ligand when it is bound to its receptor. The global minimum-energy conformation of an isolated flexible molecule (as might be calculated using molecular mechanics, for example) will not necessarily be the conformation found in the intermolecular complex (indeed, it is rather unlikely that it will correspond to this conformation). However, there will certainly be an energy penalty incurred if the ligand has to bind in a high-energy conformation [32]. This should be considered in any approach to ligand design. It should, however, be recognised that it is the ultimately suggested conformation of the molecule that is important. For example, we show in Fig. 9 a very high energy conformation of an alkane chain. When the offending atoms are made part of a ring system, the resulting structure is of low energy.

ACKNOWLEDGEMENTS

A.R.L. thanks the SERC for support under the Advanced Fellowship scheme and for the provision of computing equipment. S.R.K. thanks the SERC and Zeneca Pharmaceuticals Ltd. for a CASE award. Molecular graphics images were produced using the MidasPlus software system [33].

REFERENCES

1 Protein Data Bank Q. Newslett., 60 (1992) 1.
2 Kuntz, I.D., Science, 257 (1992) 1078.
3 Martin, Y.C., J. Med. Chem., 35 (1992) 2145.
4 Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R. and Kuntz, I.D., J. Mol. Graphics, 10 (1992) 66.
5 Pearlman, D.A. and Murcko, M.A., J. Comput. Chem., 14 (1993) 1184.
6 Rotstein, S.H. and Murcko, M.A., J. Comput.-Aided Mol. Design, 7 (1993) 23.
7 Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.
8 Moon, J.B. and Howe, W.J., Protein Struct. Funct. Genet., 11 (1991) 314.
9 Nishibata, Y. and Itai, A., J. Med. Chem., 36 (1993) 2921.
10 Böhm, H.-J., J. Comput.-Aided Mol. Design, 6 (1992) 593.
11 Ho, C.M.W. and Marshall, G.R., J. Comput.-Aided Mol. Design, 7 (1993) 623.
12 Gillet, V., Johnson, A.P., Mata, P., Sike, S. and Williams, P., J. Comput.-Aided Mol. Design, 7 (1993) 127.
13 Lewis, R.A., J. Mol. Graphics, 10 (1992) 131.
14 Lewis, R.A. and Dean, P.M., Proc. R. Soc. London, Ser. B, 236 (1989) 125.
15 Lewis, R.A. and Dean, P.M., Proc. R. Soc. London, Ser. B, 236 (1989) 141.
16 Lewis, R.A., J. Comput.-Aided Mol. Design, 4 (1990) 205.
17 Goodford, P.J., J. Med. Chem., 28 (1985) 849.
18 Miranker, A. and Karplus, M., Protein Struct. Funct. Genet., 11 (1991) 29.
19 Dammkoehler, R.A., Karasek, S.F., Shands, E.F.B. and Marshall, G.R., J. Comput.-Aided Mol. Design, 3 (1989) 3.
20 Leach, A.R. and Lewis, R.A., J. Comput. Chem., 15 (1994) 233.
21 Gō, N. and Scheraga, H.A., Macromolecules, 3 (1970) 178.
22 Bruccoleri, R.E. and Karplus, M., Macromolecules, 18 (1985) 2767.
23 Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H. and Levinthal, C., Biopolymers, 26 (1987) 2053.
24 Fine, R.M., Wang, H., Shenkin, P.S., Yarmush, D.L. and Levinthal, C., Protein Struct. Funct. Genet., 1 (1986) 342.
25 Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, A.G. and Taylor, R., J. Chem. Soc., Perkin Trans. II, 12 (1987) S1.
26 Romesburg, H.C., Cluster Analysis for Researchers, Lifetime Learning Publications, Belmont, CA, 1984.
27 Hartigan, J.A., Clustering Algorithms, Wiley, New York, NY, 1975.
28 Remington, S., Wiegand, G. and Huber, R., J. Mol. Biol., 158 (1982) 111.

29 Lindsay, R.K., Buchanan, B.G., Feigenbaum, E.A. and Lederberg, J., Applications of Artificial Intelligence for Chemical Inference: The Dendral Project, McGraw-Hill, New York, NY, 1980.
30 Dolata, D.P., Leach, A.R. and Prout, K., J. Comput.-Aided Mol. Design, 1 (1987) 73.
31 Leach, A.R. and Prout, K., J. Comput. Chem., 11 (1990) 1193.
32 Williams, D.H., Cox, J.P.L., Doig, A.J., Gardner, M., Gerhard, U., Kaye, P.T., Lal, A.R., Nicolls, I.A., Salter, C.J. and Mitchell, R.C., J. Am. Chem. Soc., 113 (1991) 7020.
33 Ferrin, T.E., Huang, C.C., Jarvis, L.E. and Langridge, R., J. Mol. Graphics, 6 (1988) 13.