J-CAMD 071

Automated site-directed drug design: Approaches to the formation of 3D molecular graphs

Richard A. Lewis

Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.

Received 30 June 1989
Accepted 7 August 1989

This paper presents two methods for the automated generation of 3D molecular graphs. The objective is to obtain molecular graphs that span the binding site and incorporate predicted ligand points at their vertices. The steric surface at the site forms the mould into which a developing ligand has to fit. Patterns of molecular forces are found at the receptor's accessible surface that are important in controlling drug binding. In many cases of drug–receptor interaction, key atoms termed site points are located at the site surface. These give rise to loci in which putative ligand atoms may be non-covalently attached to the site. The drug designer has to link up the ligand points by a network of bonds to form a molecular graph for the ligand. Structure generation is a combinatorial problem that can only be solved in exponential time by brute-force methods, so heuristics must be used to obtain answers. The site-point ligand-point model can be used to indicate putative ligand points and provide spatial constraints for the evolving ligand. This model has been successfully used when filling a binding site with a minimal 2D molecular graph [1]. The goal of this work is to extend these methods into 3D through the generation of connecting chains or graphs. Once a graph has been established, appropriate atoms and bonds will have to be placed at the vertices and along the edges of the graph; if the ligand is to be recognised by the site, then the pattern of molecular forces generated by the ligand must match a complementary pattern of forces presented by the site. An automatic method of 3D structure generation would represent an important addition to the armoury of the drug designer.

A drug designer would like to create a structure that can occupy empty ligand points in the binding site and join them to other ligand points, or to seed atoms, in an existing ligand. This objective can be achieved by finding the minimal molecular graph that incorporates these specified points. The graph should (i) not cut across the receptor-accessible surface, thereby causing repulsive steric interactions; (ii) match the local patterns of electrostatic and hydrophobic forces in the receptor site; (iii) be able to meet these requirements without having to adopt a high-energy conformation. The resulting subgraph should then be near an enthalpic minimum: any entropic penalty can be reduced by later addition of rings which reduce the torsional freedom of the chain. An algorithm is presented that generates molecular graphs to fulfill the steric and torsional requirements. An alternative database strategy is also discussed.

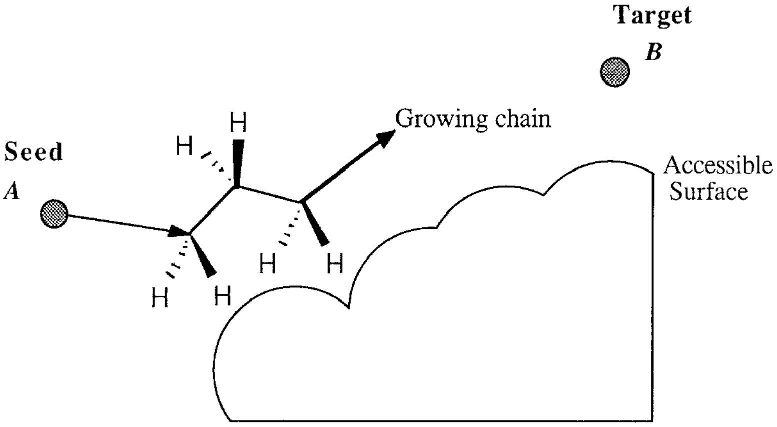

Fig. 1. The new chain is 'grown' along the surface of the receptor. The spare valencies are filled with hydrogen atoms.

The simplest case is the design of a graph that bridges a gap between a point *A* (a seed atom or ligand point) and a ligand point *B* (Fig. 1). It would be possible to search a large database of small chemical building blocks, or fragments, for a suitable subgraph to connect *A* and *B*. Each fragment in the database would have well-defined geometric and molecular properties. The geometry will determine if the gap between *A* and *B* can be bridged without incurring steric penalties. The molecular properties of all the geometrically feasible fragments could then be compared to those of the nearby site atoms so that a good complementary match is obtained. This strategy will only work if the geometries and properties of the fragments show little dependence on the environment created by the adjoining atoms and bonds in the growing molecular template. This approach will be especially useful for placing ligand groups in regions of the receptor with marked hydrophobic or electrostatic properties.

Existing methods for searching databases [2, 3] may not be appropriate here if they do not allow for ligand flexibility or steric occlusion of the site. The number of possible entries increases exponentially with the number of allowed atoms in the fragment. Good initial filters, such as upper and lower distance screens, should be used to avoid expensive optimisation calculations. More general database screening programs that do allow for flexibility and steric violations, such as the Shape–Search program [4], are computationally intensive and the quoted limit of database entries is of the order of 5000. Further partitioning of the data into clusters based on charge or hydrophobicity may be possible so that poor potential matches can be discarded very quickly. Ideally, application of these heuristics should produce <20 candidate fragments for rigorous fitting and matching.

For a fragment size of 7 or more heavy atoms, there are many thousands of different fragments; the database strategy is unworkable. Attempts to design a chain manually proved a very difficult process and highlighted the complex energetic play-off between occupation of the ligand points and avoidance of intersections with the receptor surface. A better method for joining points *A* and *B* is to generate an acyclic chain between them. A spacer skeleton is defined as the union of a set of subgraphs [5]: the union of alkyl chains is the diamond lattice. Each atom in the diamond lattice is sp^3-hybridised and all the torsion angles in the lattice are staggered. This conformation may not be the one finally adopted by the chain but it should be a good approximation to it, if the conformation energy is to be favourable. The chain is now generated by the following algorithm, BEELINE:

(i) Start with an arbitrarily large cubic diamond lattice and place one of the lattice atoms, *L0*, at point *A*. Point *B* will now fall somewhere within the boundaries of the lattice. More specifically, *B* will lie inside an adamantane subunit of the lattice (Fig. 2). The distance d_{AB} is calculated together with the sorted vector, *V*, of distances $d_{L0\,Li}$ between lattice atoms *L0* and *Li*.

(ii) Search *V* to find all the lattice atoms which lie within a distance *R* ($R < 2.0$ Å) of *B*. There will be about 10 such neighbours. Find the nearest neighbour, *Ln*, such that $d_{L0\,Ln}-d_{AB}$ is a positive minimum. This implies that the generated chain will always extend past, or 'overshoot', point *B*. The extra length of chain produced allows a greater freedom of movement in the chain during the subsequent removal of small steric violations during optimisation.

(iii) The initial orientation of the lattice with respect to the vector ***AB*** was arbitrary so that there will be some rotation about the point *A* which decreases the distance between *Ln* and *B* to a minimum. This transformation is calculated and stored in the matrix ***M***.

(iv) Perform a directed depth-first-search along the lattice starting at *L0* and heading towards *Ln*. Every time a new node is reached, the coordinates of its unvisited neighbours are transformed by ***M*** and the new points are checked for exclusion by the accessible surface of the binding site. Small violations are allowed.

(v) From the new trio of prospective nodes, select the node that (a) has no excluded neighbours, and (b) is the nearest to *Ln*. If all of the prospective nodes are forbidden, backtrack up the chain and try the next allowed node. Otherwise repeat steps (iv) and (v) until *Ln* is reached.

(vi) Set *Ln* to be the next nearest neighbour and repeat steps (iii) to (v). This will generate a new chain of different geometry.

The overshoot strategy employed in step (ii) gives a small positive safety margin of error in the positioning of *Ln*, to allow for subsequent flexing of the chain during optimisation. If there were any violations of the accessible surface, the chain could be iteratively adjusted until they are removed. The small cost in torsional energy inherent in any movement of the atoms will be compensated for by the decrease in repulsive steric overlap. The process of adjustment will 'shorten' the distance between the chain ends but will not pull *Ln* far away from *B* because of the overshoot that had been built in during generation.

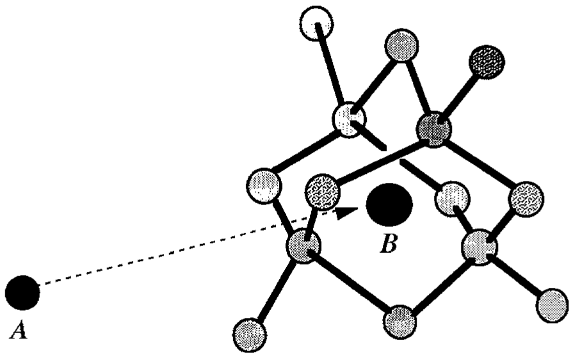

Fig. 2. The diamond lattice is rotated so that point *B* falls within an adamantane cage. The atoms of the cage will be used in turn as the target atom *Ln* in the directed search.

The directed search of step (iv) is much more efficient than a brute-force clipping of the lattice to the site: less than 10% of the lattice atoms are visited and tested during the generation of the chain. The testing of an atom against the site for steric violations is effectively like looking for a key in an array of unordered records; this O(*n*) process forms a significant part of the computational load. The directed search method means that the routine is cheap to use and so can be called many times as the ligand structure is evolving.

The criteria in step (v) used for selecting the next atom to be added to the chain and for finishing the search are both based on simple distance functions. There seems to be no real justification for a more complex function for atom selection at present. The termination function could be made more realistic by testing for inclusion of the current atom within a ligand point region, contoured at a suitable energy [6] or probability [7] level. This would prevent unnecessary elongation or rejection of chains due to a poor target point *B*.

This algorithm will derive from the diamond lattice all the simple molecular graphs which can join the two ligand points in a binding site. It will also build in steric complementarity into the chain. There is the added bonus of possible favourable hydrophobic interactions because of the way in which the chain is made to crawl over the accessible surface.

In future work, the chain atom positions will have to be refined through energy minimisation/ flexible fitting; this could be performed by three different methods. One approach would be to optimise a combined objective and penalty function. As the perturbations of the chain should be small in size, the objective function could be the sum of the conformational energy and the intermolecular interactions. The penalty function would be a sum of terms representing steric violations. The second method treats the local geometries of the start and end atoms in the chain as fixed (a reasonable approximation for hydrogen bonds and joins to an existing template). If there are 5 or less torsion angles in the chain, then each of the angles can be derived analytically from the given fixed geometries [Lewis R., unpublished results]. For 6 torsion angles, the root-finding algorithm of Gō and Scheraga [8] can be used. For 7 or more rotatable bonds, the system becomes underdefined and there is no analytical solution. The number of variables may be reduced to just 6 torsion angles by iterating through the angle space of the remaining (*n*-6) bonds and creating a new start triad explicitly from the (*n*-6) rotations. Test runs with random starts have shown that there are rarely more than 8 roots produced per given geometry for the 6-angle case. This means that the expensive optimisation of a multivariate energy function can be avoided by only examining 'ideal' conformations. The use of these analytical procedures will allow very quick tests to see whether a given chain is feasible and will save large amounts of computation time.

At present, the chain derived from the diamond lattice is composed of links 1.52 Å long to mimic C-C bonds. The lattice must be homogenous or the nature of the chain will depend very strongly on the identity of the start atom *L0*. If the initial set of generated chains fails to produce a satisfactory solution, small perturbations (in the form of changed bond lengths and angles) could be introduced and tested to produce a new set of chains; this is analogous to isostere replacement and would allow the generation of general alicyclic chains based on an alkyl template. The third procedure for flexible fitting would be to give the steric constraints infinite weight and to allow a looser fit of specified tolerance at either end of the chain. The docking could then be performed by a distance geometry method [9].

Initial studies have been performed on an isolated α-helical heptapeptide with a random sequence (Ser-Val-Leu-Ala-Ala-Gly-Asp) built using the SYBYL molecular modelling program.

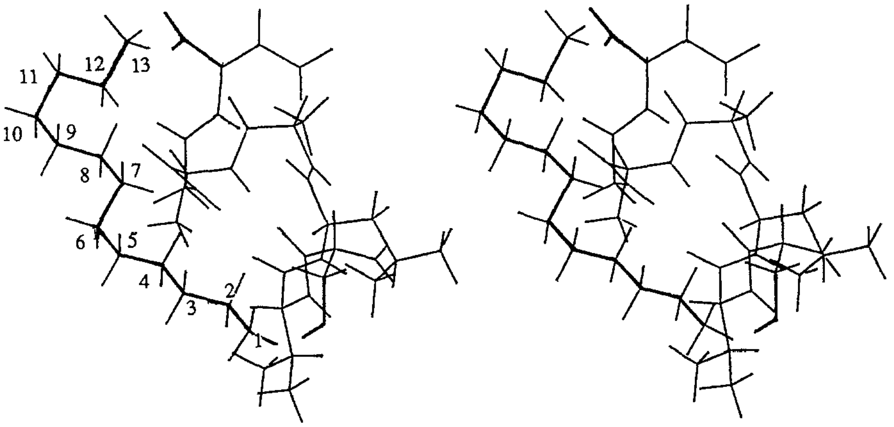

Fig. 3. A relaxed stereo plot of a chain (in bold) generated by BEELINE to join two hydrogen-bonding groups (bold) in a random heptapeptide helix. The generator proceeds along the chain by adding one atom at a time in the order shown.

Generation was started at the coordinate system defined by the hydrogen-bonding ligand point due to Ser-1.Oγ (Fig. 3, bottom centre) with the aim of joining this ligand point with the corresponding one due to Asp-7.OXT (Fig. 3, top centre); the distance between the points is 10.8 Å. The chain shown grows up the left hand surface of the helix towards the target, atom by atom, until it reaches a bump near to chain atom 8. A backtrack has clearly occurred as the chain doesn't grow in the 'best' direction (due to steric repulsion) but moves around the obstacle (atoms 9–11) before arriving at the target (atom 13). The best-fit chain of the 6 that were generated placed the final atom 0.01 Å away from the target ligand point. The steric interaction energy of 18.2 kJmol^{-1} (Buckingham potential) was due to steric violations by 3 atoms in the chain. Two of these atoms (accounting for $\sim$120 kJmol^{-1}) were H's in hydrogen-bonding regions so that this figure implies that the rest of the fit was energetically very good. The generation of each chain took 0.05 cpu s on an IBM 3084Q.

The chain-generation algorithm given above does not take into account the electrostatic properties of the site at the accessible surface: a hydrophobic chain may be generated in close proximity to polar regions of the surface. A solution to this difficulty might be to give an extra scaling factor to the van der Waals radius of a polar atom; this would effectively appear as an increased steric constraint in step (iv) of the generation algorithm and would force the chain away from the region. Furthermore, the scaling factor could be a (linear) function of the magnitude of the electrostatic potential and so could be made to favour neutral regions.

This investigation into methods for the formation of 3D molecular graphs has pointed out several new areas of research. The basic chain generation algorithm, BEELINE, has been shown to be capable of generating quite long acyclic graphs subject to steric constraints. Further tuning of the parameters needs to be performed to allow for electrostatic and hydrophobic effects. The future completion of this work would allow the extension of planar graphs generated by existing methods [1] into 3D or the joining of two ligand points, so filling up the given binding site. This would represent an important addition to the armoury of the drug designer.

ACKNOWLEDGEMENTS

R.A.L. would like to thank Philip Dean for many helpful comments and criticisms. This work was supported through the William Colton Research Fellowship at Queens' College, Cambridge.

REFERENCES

1 Lewis, R.A. and Dean, P.M., Proc. Roy. Soc., B236 (1989) 125.
2 Jakes, S.E., Watts, N., Willett, P., Bawden, D. and Fisher, J.D., J. Mol. Graph. 5 (1987) 41.
3 Van Drie, J.H., Weininger, D. and Martin, Y.C., J. Comput.- Aided Mol. Design, 3 (1989) 225.
4 DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1989) 722.
5 Lewis, R.A. and Dean, P.M., Proc. Roy. Soc., B236 (1988) 141.
6 Boobbyer, D.N.A., Goodford, P.J., McWhinnie, P.M. and Wade, R.C., J. Med. Chem., 5 (1989) 1083.
7 Danziger, D.J. and Dean, P.M., Proc. Roy. Soc., B236 (1989) 115.
8 Gō, N. and Scheraga, H.A., Macromolecules, 3 (1970) 178.
9 Blaney, J.M., Proceedings of the 3rd International School of Crystallography, Sicily, 1989.