J-CAMD 299

The atom assignment problem in automated de novo drug design. 3. Algorithms for optimization of fragment placement onto 3D molecular graphs

M.T. Barakat and P.M. Dean*

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

Received 1 February 1995
Accepted 23 May 1995

Keywords: De novo drug design; Simulated annealing; Molecular graphs; Molecular electrostatic potential; Molecular complementarity; Molecular similarity

Summary

Atom assignment onto 3D molecular graphs is a combinatoric problem in discrete space. If atoms are to be placed efficiently on molecular graphs produced in drug binding sites, the assignment must be optimized. An algorithm, based on simulated annealing, is presented for efficient optimization of fragment placement. Extensive tests of the method have been performed on five ligands taken from the Protein Data Bank. The algorithm is presented with the ligand graph and the electrostatic potential as input. Self placement of molecular fragments was monitored as an objective test. A hydrogen-bond option was also included, to enable the user to highlight specific needs. The algorithm performed well in the optimization, with successful replications. In some cases, a modification was necessary to reduce the tendency to give multiple halogenated structures. This optimization procedure should prove useful for automated de novo drug design.

Introduction

In the first paper of this series on the atom assignment problem in automated de novo drug design, we presented a general scheme for optimizing the assignment of atoms onto molecular skeletons [1]. Atom assignment is a combinatorial problem that has to be reduced by optimization techniques if it is to be solved effectively. Use has been made of the fact that a number of properties of interest to drug designers are transferable if the atoms are collected together into fragments. This is especially so with electronic properties [2]; consequently, the electrostatic potential can be approximately reconstructed from molecular fragments, without the necessity of re-computing the residual charges each time a new atomic assignment is tried. Other properties, such as hydrogen-bonding capability and hydrophobicity, can also be handled by a fragment procedure.

In the previous paper [3] a method was presented for perception of molecular graphs and molecular fragments in terms of their connectivities at each node. A canonicalization of atom numbering, together with a hashing procedure for efficient storage and accessing of fragments from a fragment library, was presented. Combinatorial optimization problems can be tackled by at least three well-studied strategies: genetic algorithms, neural networks and simulated annealing. Few comparisons between these different approaches have been made, although Ingber and Rosen [4] have compared genetic algorithms with simulated annealing on a variety of mathematical test problems. From their work it would appear that simulated annealing is more efficient. However, this may not be so in all cases. Genetic algorithms may have the advantage of ease of use if the objective function can be expressed in a form of string notation; this would allow ready mutation and crossover of portions of the structure. The relative merits of neural networks compared with these two optimization techniques are not known.

In the current paper, we describe an optimization algorithm for efficient reduction of the combinatorial atom assignment problem. The algorithm employs simulated annealing. The number of combinations for fragment

*To whom correspondence should be addressed.

placement, N, that are possible on a molecular graph containing j nodes is given by:

$$N = \prod_{i=1}^{j} \sum_{f=1}^{g_i} r_f \qquad (1)$$

where the node i can be assigned a fragment f from a set containing g_i members; there are r_f orientations of the fragment f at that node. Of course this estimate of the size of the combinatorial problem is for atom placement only; it is not the number of *actual* molecular structures that can be made. Bond order violations between adjacent atoms drastically reduce the number of molecular possibilities.

In contrast to the parallel computations of neural networks, simulated annealing is essentially a serial optimization procedure which is thermodynamically motivated. The method reduces an exponential increase in computing time to only small powers of a measure of the size of the problem. The optimization heuristic relies on the Monte Carlo condition as modified by Metropolis et al. [5]. The ensemble starts with a large value for the control parameter, T (thermodynamically equivalent to temperature), and the objective function for the system (thermodynamically equivalent to energy) is calculated (E_1). A random perturbation is then applied to the system, and the new value E_2 is found. The difference, ΔE, is calculated, where $\Delta E = E_2 - E_1$. If ΔE is negative, the perturbation is accepted unconditionally, and the new state E_2 is maintained. The acceptance of a perturbation with a positive ΔE, on the other hand, depends on the Metropolis condition, Eq. 2:

$$P(s'|s) = e^{-\Delta E/T} \qquad (2)$$

where $P(s'|s)$ is the probability of accepting a state s' which has an energy greater than the energy of the existing state, s, by an amount ΔE. If a random number generated between [0,1) is less than $P(s'|s)$, then the new state s' is accepted. After several perturbations and when equilibrium is reached, T is reduced, and the cycle is repeated. In this way the system is given the opportunity to escape from a local minimum at high values of T. Optimization is ergodic. This approach has been applied successfully to many combinatorial problems of molecular matching [6–10].

In de novo drug design, our aim is to take a molecular graph generated within the site and optimize the assignment of atoms on that graph by a fragment placement procedure. The fragments have precalculated properties; thus only a small amount of computation should be necessary to monitor the perturbations produced by exchanging fragments on the molecular graph. The property of principal interest is the electrostatic potential, but the method is not restricted to this property; other properties may be included as penalty terms within the objective function to be optimized. Moreover, it is possible to style the objective function to consider de novo drug design within the framework of molecular complementarity to a site *or* with respect to molecular similarity from a set of 'similar' molecules [11,12].

Theory and Methods

The theory and methods for fragment placement will be considered in three separate sections: the annealing parameters peculiar to the algorithm, the rules for fragment placement and an outline of the algorithm.

The annealing parameters

Objective function

For atom assignment onto molecular graphs, the drug designer may want to optimize the placement according to electrostatic potential (E_e) for certain structures, corresponding to hydrophobic potential (E_h) for others, or apportioned by a combination of the two potentials. The algorithm therefore needs user-specified input of the required relative contributions of both electrostatic and hydrophobic potentials in the objective function. The electrostatic potential, V (in kJ mol^{-1}), from a fragment-placed molecular graph to a point j on its surface is calculated from Eq. 3:

$$V_j = \sum_{i=1}^{n} \frac{q_i}{r_{ij}} c \qquad (3)$$

where n is the number of atoms in the graph, q_i is the residual charge on atom i in atomic units, r_{ij} is the distance (Å) between atom i and point j, and $c = 2626.57$ kJ mol^{-1} au^{-1}. The method chosen to calculate the hydrophobic potential was that of Fauchère, Quarendon and Kaetterer [13], which has been discussed in paper 1 and has an exponential decline of potential with distance. The value of the absolute temperature used in the calculation is 293 K, and the hydrophobic potential is in kJ mol^{-1}.

Hydrogen-bonding information does not participate in the actual objective function; it is accounted for at the choosing stage in an all-or-none fashion: if an acceptor is required at a graph vertex, only a fragment with an acceptor atom is chosen, and similarly with donors.

Once the character of the objective function (electrostatic and/or hydrophobic) has been chosen, the drug designer specifies whether the assessment method of the sets of potential should be according to Pearson's correlation coefficient r, Spearman's rank r_s, pair sums (optimization of complementarity), or pair differences *eo* (optimization of similarity). Each method of assessment gives a different emphasis to the final atom assignment. For example, minimization according to Spearman's rank will produce a placement which primarily has the desired pattern of potential, whereas Pearson's coefficient will produce a placement which takes into account actual values of potential.

Since the optimization procedure is set to be a minimization, the objective function (r or r_s) is made *negative* in certain situations. These situations include optimization of similarity in envelope-directed atom assignment [14], and comparison of hydrophobic potentials by the objective function [15]. This change in sign of the correlation coefficients occurs only during the optimization, and on output, the signs of the coefficients revert to their true value.

No scaling of the objective function itself is performed; in the case of both Pearson's and Spearman's coefficients, the range is from −1 to +1. However, the next section describes how the *change* in objective function is scaled in the Metropolis condition.

Scaling of the change in objective function

Since the behaviour of the annealing algorithm is dependent on the Metropolis condition, the value of the change in objective function, ΔE, in the equation must be scaled so that it becomes independent of the size of the molecular graph, the number of points and the type of objective function used. A positive ΔE is only accepted if the statement defined in Eq. 4 is satisfied:

$$\text{random } [0,1) < e^{-\Delta E_{scaled}/T} \tag{4}$$

Assuming that the distribution of ΔE values is normal, a scaling is required such that the initial acceptance of the maximum ΔE is the same using different objective functions. For a normal distribution, 99.7% of the population have values smaller than three times the standard deviation, σ. ΔE is therefore scaled according to Eqs. 5–7.

$$\Delta E_{scaled} = f_e\, C\left(\frac{\Delta E_{unscaled}}{3\sigma_{\Delta E} + \langle \Delta E\rangle}\right)_e + f_h\, C\left(\frac{\Delta E_{unscaled}}{3\sigma_{\Delta E} + \langle \Delta E\rangle}\right)_h \tag{5}$$

$$f_e + f_h = 1 \tag{6}$$

$$\langle \Delta E\rangle = 0 \tag{7}$$

where subscripts e and h refer to electrostatic and hydrophobic potential changes, respectively, f_e and f_h are the fractional contributions of electrostatic and hydrophobic potential to the objective function, C is a constant, and $<\Delta E>$ is set to zero since, initially, there is a similar probability of generating a certain negative or positive ΔE value.

The maximum acceptance for a positive ΔE occurs when ΔE tends to zero, and the acceptance probability is 1. Consider just electrostatic changes (i.e. $f_h = 0$). For a positive $\Delta E = 3\sigma_{\Delta E}$, a constant C is needed such that the acceptance probability at the initial value for T is sufficient to prevent trapping in unfavourable local minima. We suggest a value of $C = 2$ or 4. In the algorithm, the value for C is given as an input option (default is 2.0). The same starting value for T can be used for different objective functions (each has its own $\sigma_{\Delta E}$), for different molecular graphs, and for different sets of surface points.

Transition mechanism

A state is defined as a particular arrangement of potential values at the surface of the molecular graph. Initially, a random placement of fragments is made. Then a node is chosen at random, and a fragment is chosen at random from the allowed number of fragments that could fit the chosen node. The partial charges (or atomic hydrophobic parameters) of this new fragment then replace the underlying atomic values. The new potential generated is calculated at the surface of the graph; this new set of potential values forms the new state. If the objective function improves, the state is accepted unconditionally; otherwise the Metropolis condition is applied. This is repeated several times before T is reduced. Since the value of the potential at point j is a summation over all atoms, considerable savings in computation time can be made by simply adding the difference in the contribution to the potential created by the exchange of atoms in the proposed transition.

Initial value for T

The initial value for T was chosen so that the initial acceptance probability for positive changes in the objective function was between 0.7 and 0.8 for 68.3% of the population of states. An initial T of 2.0 satisfied this condition.

Length of Markov chains

In simulated annealing, the Markov chain length represents the number of trials proposed at a given T. The cooling rate is dependent on both the length of the Markov chain, T, and the constant C, Eq. 8:

$$rate \propto \frac{l_m C}{\Delta T} \tag{8}$$

The length, l_m, must be sufficient to ensure that equilibrium has been reached at each value of T. In the atom assignment algorithm, the maximum length of the Markov chain was fixed to the sum of the total number of fragment rotations possible at all the nodes in the molecular graph. As an example, this length was between 100 and 200 for a 40-vertex molecular graph. If the number of accepted trials at a particular value of T equalled half the maximum length of the Markov chain, equilibrium was assumed to have been reached and the chain was ended and a new one started at a lower value of T.

Decrement of T

The decrement in T influences the quality of the convergence of the algorithm; large decrements result in entrapment in poor local minima, while small decrements take too much CPU time. The method employed in the atom assignment algorithm uses a dynamically varying fraction, $f_{dynamic}$, which depends on the current performance of the algorithm in the annealing process. The calculation of $f_{dynamic}$ is based on the algorithm of Aarts, Korst and Van Laarhoven [16], Eq. 9:

$$f_{dynamic} = 0.95\left[1+\left(\frac{T\langle E\rangle_T \ln(1+\delta)}{3\sigma_E}\right)\right]^{-1} \tag{9}$$

where the parameter $\delta = 0.02$ (the smaller δ, the slower the cooling), and $3\sigma_E$ is the standard deviation of the objective function (not ΔE) at a particular T. With Pearson's and Spearman's correlation coefficients, the equation is modified so that negative values of E do not result in a value for $f_{dynamic}$ greater than 1 (Eq. 10).

$$f_{dynamic} = 0.95\left[1+\left(\frac{T(1+\langle E\rangle_T)\ln(1+\delta)}{2}\right)\right]^{-1} \tag{10}$$

The advantage of this method is that the algorithm is able to monitor the progress of the optimization: as the system starts to 'crystallize' at a certain 'critical' value of T, the value for E declines rapidly, signalling the stage when the slowest cooling is required, and $f_{dynamic}$ is accordingly increased.

Stop criteria

Several stop criteria have been suggested by different authors for terminating the algorithm. The one most frequently used is based on the ratio of the number of accepted trials over the number of generated trials in a Markov chain; this is known as the acceptance ratio. If this acceptance ratio falls below some minimum value (0.08 has been found to be suitable for fragment placement), the system is assumed to have found a good local, if not the global, minimum. An alternative stop criterion, suggested by Huang, Romeo and Sangiovanni-Vincentelli [17], and previously used by us in molecular matching routines, was not used here because it tended to stop the algorithm prematurely. If the acceptance ratio failed to fall below 0.08, the algorithm was terminated after reaching the maximum number of Markov chains (set at 500 000 so that it was not limiting), or exceeding a default time limit of 2 h of CPU time.

Application of graph perception and simulated annealing to atom assignment

The second paper in this series [3] described how a molecular graph can be perceived, and how the same rules can be applied to the perception of fragments in the fragment library. The canonicalized connectivity patterns can be used for direct accessing of the fragment library after conversion of each pattern to a fragment bin address via a hash function. Searching, however, is only one of the computationally costly processes involved. The large number of possible placements of fragments onto a molecular graph requires optimization. This section combines the fast searching procedure of hashing with simulated annealing in an algorithm for optimizing fragment placement onto molecular graphs.

Designer options

A drug designer would like to have some flexibility in the system. To cater for various needs, the following designer options have been incorporated into the fragment placement program.

(1) The relative contributions of electrostatic and hydrophobic potential in the objective function are user-defined. There is fractional weighting (f_e and f_h) of the changes in electrostatic and hydrophobic potential (Eqs. 5–7) in the Metropolis condition. If only electrostatic potential needs to be optimized, f_e is set to 1.0 and f_h is set to 0.0. Furthermore, if there is some hydrophobic contribution in the optimization, then one can specify, for example, $f_e = 0.75$ and $f_h = 0.25$. Similarly, one could also request optimization with $f_e = 0.0$ and $f_h = 1.0$ if there were no electrostatic contribution at all.

(2) Another option concerns specifying if the procedure required, in the case of electrostatic potential optimization, is for complementarity placement (when site information is given) or for similarity placement (when there is an existing lead) to the electrostatic surface potential given. This option does not concern hydrophobic optimization, since the optimization is always a similarity placement to the hydrophobic surface potential given.

(3) The designer can specify that some atom types should be fixed. This is a useful option when an existing lead is present, and only minor modifications are needed to improve the structure.

(4) Hydrogen-bonding properties at certain vertices of the molecular graph can be specified. A vertex can be a hydrogen-bond acceptor, donor, or neither.

(5) The method of assessment of complementarity/similarity may be specified. For example, Pearson's correlation coefficient may be preferable to Spearman's rank for a particular molecular graph.

(6) An option is present for allowing only neutral fragments in the placement procedure.

(7) Finally, the designer may influence the rate of cooling of the annealing process by specifying the value of the constant C. The default value is 2.0; the greater the value, the faster the cooling rate.

Fragment placement rules

In the annealing process, fragments are chosen at ran-

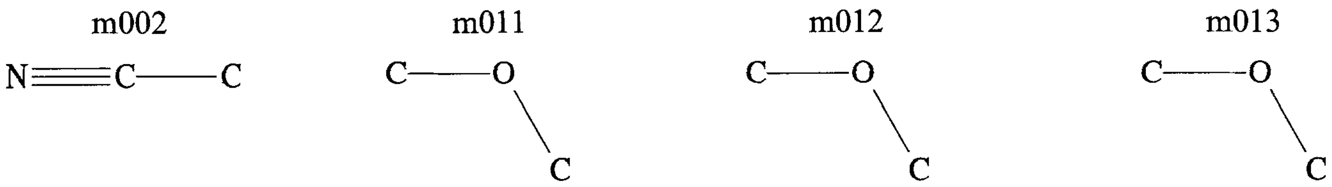

Fig. 1. Some fragments which are treated in a special way. Fragment m011 has two carbon atoms with sp^2 hybridization, fragment m012 has carbon atoms with sp^2 and sp^3 hybridization, and fragment m013 has two carbon atoms with sp^3 hybridization.

dom from the specified key address of a particular node (also chosen at random). Each time a fragment is placed on the graph, the atomic residual charges and atomic hydrophobic parameters of the fragment are transferred to the vertices of the graph node. The new surface potential (either electrostatic or hydrophobic) is then recalculated, and the new objective function (some measure of complementarity or similarity) is compared with the previous value.

At the start of the annealing process, there is no consideration of the inter-fragment bond order; otherwise, the placement would very soon become locked in an attempt at placing fragments which would perfect the inter-fragment connections rather than the objective function. It is only near the end of the annealing procedure that, firstly, connections between fragments are considered, and secondly, an average is taken of the atomic properties (charges or hydrophobic parameters) when there are overlapping fragments.

There are, however, a number of placements which are forbidden all through the running of the program:

(1) Bond angle violations. Only those three-atom fragments with the required bond angle can be placed on a particular three-vertex node. It is pointless placing a cyano-group fragment (m002, see Fig. 1) onto a three-vertex node in which the central bond angle is 109.6°.

(2) Hydrogen-bond-type violations. If the drug designer has specified that a particular vertex in the graph has to be a hydrogen-bond acceptor (or donor, or neither), then no fragment can be placed there unless it will impart the desired hydrogen-bond property.

(3) Hydrophobic parameter violations. The hydrophobicity of some fragments is not known (particularly that of the charged fragments), and when the objective function has a hydrophobic contribution, no fragment can be used unless its hydrophobicity is known. For example, the choice of a phosphate group is rejected for a five-vertex node.

(4) When placing a single aromatic ring fragment over a duplex cyclical node, any rotation of the single ring fragment is rejected if it results in the placement of a nitrogen atom at the intersection between the two subcycles of the duplex node.

(5) The aromatic fragments in the library cannot accommodate a branch connection via a double bond from an adjoining aliphatic fragment. As a result, the placement of any aliphatic fragment in which there is a double bond to the aromatic ring is prohibited.

(6) The hybridization of some aliphatic fragment carbon atoms is specified (e.g., fragments m011, m012 and m013, see Fig. 1), whereas the remaining aliphatic fragment carbon atoms have wild cards, and so can be treated as sp, sp^2 or sp^3 atoms. When it comes to placing an aliphatic fragment branch to an aromatic ring, fragments in which the C atom overlapping the ring has a wild card are rejected, and the placement is restricted to fragments in which an sp^2 carbon has been previously defined.

(7) The placement of an aliphatic fragment nitrogen atom over an aromatic fragment ring atom is rejected, except when the aromatic atom is an existing nitrogen atom. Similarly, the placement of an aliphatic fragment carbon atom over an aromatic fragment ring atom is rejected, except when the aromatic atom is an existing carbon atom.

(8) A drug designer may specify some atom types to be fixed. In that circumstance, the placement of an atom type different from the fixed type is prevented.

(9) If the option of using only neutral fragments has been invoked, the placement of a charged fragment is rejected.

(10) In some cases, the maximum number of halogen groups in the developing structure is set to two. This condition is applied here for testing.

Algorithm

The fragment placement program has been written in FORTRAN 77. The random number generator used was taken from Press et al. [18]. The following sections are a schematic representation of the program. These cover firstly the pre-annealing stage, secondly the annealing stage, and lastly the post-annealing stage of the fragment placement procedure.

Pre-annealing stage

The pseudocode for the pre-annealing stage is presented in this section. This stage involves reading in the user-specified options, the molecular graph, the surface points and potential, and the fragment library. The fragments are perceived and canonicalization is applied to every fragment. Fragment rotations are determined and the fragments are binned according to their canonicalized connectivity patterns (after transformation through a hash function). The same perception rules are applied to the noncyclical and cyclical nodes of the graph. The connectivity patterns of these nodes are canonicalized and each node is allocated a base 10 code by applying a hash function to the processed connectivity patterns (as explained in paper 2 of this series [3]). Each graph code refers to the fragment address location of suitable fragments in the fragment library which can fit the graph node in terms of connectivity.

A random set of fragments is placed on the molecular graph, and an initial value for the objective function is obtained. The annealing parameters, such as the initial value for T, the length and number of Markov chains, and the standard deviation of ΔE, are calculated.

Pseudocode for pre-annealing

```
Read in options:
    relative contribution of electrostatic potential, $f_e$; set
        $f_h = 1 - f_e$
    similarity or complementarity?
    are there any fixed atom types?
    are any H-bond properties desired?
    assess with Pearson's, Spearman's or pair sums/pair
        differences?
    are only neutral fragments required?
    which value for the acceptance constant $C$ is required?
Read in molecular graph coordinates
Read in surface points (with a tessellation frequency of 1)
Read in potential (electrostatic and/or hydrophobic) at
    each point
If similarity placement, set potential = – potential
                    ! to maintain a minimization
Read in fragment library
    apply canonicalized numbering scheme to the fragments
    find fragment rotations
    calculate connectivity keys of fragments
    get hash function (key address) of each fragment
    place each fragment in the bin whose number is the
        key address
Perceive all nodes in skeleton
    perceive all noncyclical and cyclical nodes
    apply canonicalized numbering scheme to the nodes
    calculate connectivity keys of nodes
    apply hash functions to get fragment bin addresses by
        conversion to base 10
For each node, allocate a fragment randomly chosen
    from the bin corresponding to the node
Calculate the initial electrostatic objective function, $E_e$
    and/or the initial hydrophobic objective function, $E_h$
Set up the annealing parameters:
    set $T_{initial} = 2$
                    ! initial annealing value for T
    set $n_m = 500\,000$
                    ! maximum number of Markov chains
    set $l_m$ = sum of total number of fragment rotations at
        all nodes
                    ! maximum length of Markov chains
Calculate the standard deviation of $\Delta E_e$ and $\Delta E_h$ for $l_m$
    trials, accepting none
```

Annealing stage

The simulated annealing algorithm pseudocode is now described. At a particular value for T, different fragments are randomly chosen for placement onto the molecular graph. Each time a new fragment is introduced, the Metropolis condition is applied and only accepted changes to the objective function are retained. This is repeated until equilibrium is reached in a Markov chain, and the value for T is reduced. The whole process continues until a stop criterion comes into effect.

Inter-fragment bond orders are then checked, and if found to be inappropriate, the offending fragments are reannealed until the correct bond order is generated. Adjacent aliphatic fragments can only be connected via single bonds. The final arrangement of fragments is processed to adjust the total charge to an integer. Also, the bond lengths are corrected using standard covalent bond radii [19] rather than fragment values, since there may be some fixed atom types (which are not included in the fragment library). The corrected structure is then repositioned onto the original using the superpositioning algorithm of McLachlan [20], which employs a least-squares procedure. The new objective function is calculated and stored. After this first annealing step, a second annealing is performed for the 25% of atoms whose surface potential is poorest.

First annealing pseudocode

Step 1, placement

```
do $j = 1, n_m$
    do $i = 1, l_m$
        if ((a random number [0,1) < 0.5) or ($i = 1$)), then
            choose a node at random
        else
            set node = current node
        endif
        choose a fragment at random from the bin of the
            node
        set nrot = number of rotations of the chosen fragment
        do $k = 1$, nrot
            reject fragment rotation if it results in a prohibited placement
            transfer the properties of the new fragment to the
                graph
            calculate new $E_e'$ and/or $E_h'$
            calculate $\Delta E_e$ and/or $\Delta E_h$
            apply Metropolis condition
            if (accepted change) then
                retain new fragment having rotation $k$
                set $E_e = E_e'$
                set $E_h = E_h'$
                if (number of acceptances = $l_m/2$) then end chain
                if (number of acceptances < 0.08) end annealing
            else
                reject fragment rotation $k$
            endif
        end do
    end do                              ! end of Markov chain
    calculate $f_{dynamic}$              ! dynamic T factor
    set $T = T * f_{dynamic}$            ! reduce T
cnd do
```

Step 2, elimination of molecular inconsistencies

Check inter-fragment bond orders of best structure: adjacent aliphatic fragments can only be connected via single bonds

Reanneal fragments which have unsuitable inter-fragment bond orders

Step 3, structure refinement

Reassign charges, taking averages at regions of overlapping fragments

Adjust all atomic charges so that the overall formal charge is an integer

Correct the bond lengths (from standard covalent bond radii [19])

Reposition corrected structure onto original molecular graph [20]

Calculate new E_e and/or E_h

End of first annealing step

Second annealing step

The 25% of atoms whose surface potential gives rise to the poorest part of the objective function are found, and the first annealing step is repeated for only these atoms. In practice, more than 25% of the atoms are used since neighbouring fragments are included to remove possible bond violations.

Post-annealing stage

The results of the first and second anneals are compared, and the best of the two, representing the global (or good local) minimum, is chosen. The coordinates and atom types of this best structure are output.

Methods summary

The drug designer is allowed certain options to tailor the results according to particular requirements. There are a number of forbidden placements, and these prevent unnecessary applications of the computationally costly Metropolis condition. The placement procedure consists of three stages. The pre-annealing stage involves reading in and perceiving the fragments in the fragment library. The noncyclical and cyclical nodes of the molecular graph are perceived according to the same rules used for the fragments. The connectivity pattern of each node of the graph is canonicalized and converted by a hash function to the library location address of the fragments which can fit the node. The pre-annealing stage ends with the calculation of the annealing parameters.

The next stage is divided into two parts. A first annealing step is carried out, in which a run of fragment placements is allowed until the system reaches equilibrium. The best placement is adjusted to have standard bond lengths, and stored. A second annealing is made of the 25% of atoms corresponding to the worst surface potential. In the post-annealing stage, the corrected best placement of this second annealing step is compared with the stored result of the first anneal, and the most favourable placement is output.

Computer times are quoted for FORTRAN 77 programs running on a Sun SPARCstation IPX.

Results

The operation of the algorithm will be demonstrated on five molecular structures taken from ligand–protein co-crystal complexes from the Brookhaven Protein Data Bank. With these examples, a self-test protocol is applied. The objective is to examine whether an electrostatic potential surrounding a molecular graph can be used to generate a molecular structure which is identical to that used originally to create the electrostatic potential. One example was used to test the hybrid objective function composed of the electrostatic potential and the hydrophobic potential. Where hydrogen-bonding points are available on the molecular graph, the results are analysed by using the specified hydrogen-bond option. The electrostatic potential was computed from the crystallographic coordinates and CNDO charges, since the fragments database contained charges obtained by that method [2]; there is no restriction on the method employed for calculating the atomic residual charges.

The tables of data used for comparison of the results all have the same format. Column 1 identifies the ligand. Column 2 specifies the method of assessment of the objective function. The classes are: statistical correlation by Pearson's product moment correlation coefficient, *r*; rank correlation coefficient, r_s; and difference in electrostatic potential values, *eo*. The results are replicated up to 10 times. Column 3 gives the mean value of the final objective function calculated by fragment placement, together with its standard deviation. Column 4 gives the average time taken for the placement by that method. Column 5 gives the fraction of replicates which exactly recreated the original structure. Column 6 indicates the mean number

TABLE 1
ANNEALING RESULTS USING THE ELECTROSTATIC POTENTIAL FROM AMP AND ITS MOLECULAR GRAPH

Method of calculation	Method of assessment	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.988 ± 0.000	49.0 ± 4.9	10/10	0
	r_s	0.987 ± 0.000	63.8 ± 10.9	10/10	0
	eo	12 350 ± 25 320	31.8 ± 6.8	9/10	4
Electrostatic potential with hydrogen bonds specified	r	0.988 ± 0.000	39.2 ± 3.9	10/10	0
	r_s	0.987 ± 0.000	33.6 ± 7.6	10/10	0
	eo	4608 ± 88.52	28.7 ± 7.6	0/10	1.0 ± 0.0

The values in each row represent means over 10 runs for a given set of conditions. The fifth column represents the fraction of runs in which AMP itself was reproduced, while the last column represents the average number of atoms which differ from AMP in the remaining runs.

of atoms which were incorrectly placed on the inexact recreation(s) of the original structure.

AMP

The ligand was taken from the crystal structure complex of adenylate kinase (1AK3) [21]. Graph perception revealed 35 vertices, and a total of 11 placeable nodes in the AMP skeleton, giving rise to around 400 million ways of arranging fragments; this value would rise dramatically to 10^{21} if the aromatic ring were to be overlaid with aliphatic fragments. Therefore, this option was prohibited to prevent poor transferability of fragment charges. The number of surface points used was 136. Results are shown in Table 1.

Placement using electrostatic potential with either correlation coefficient yielded excellent reproducibility. The placement results were not quite so good if absolute differences *eo* were used as the objective function; with the hydrogen-bond option the algorithm always placed one atom, chlorine, incorrectly. When the hydrogen-bond option was switched off, the absolute difference method worked well in all cases but one. Placement with the hydrogen-bond option was a little quicker.

cAMP

The structure of the *Escherichia coli* catabolite gene activator protein (3GAP) has been crystallographically refined at a resolution of 2.5 Å [22]. Each of the subunits (A and B) is bound to one molecule of the allosteric activator, cyclic AMP (cAMP_a and cAMP_b). The formal charge on cAMP was −1 (see Fig. 2). Hydrogen-bonding acceptors in cAMP were: O2, O3, O4, O7, O9, O11, N15, N19 and N21, while the hydrogen-bond donors were H28, H32 and H33.

Cyclic AMP had 33 vertices and 10 placeable nodes, giving rise to 10^8 possible fragment placements, as compared to 10^{21} if aliphatic fragments were allowed to overlay planar rings. Using a tessellation frequency of 1, 114 surface points were present for the cAMP_a conformation. The annealing results for the molecular graph of cAMP_a are shown in Table 2.

The self-placement results showed that cAMP_a was reproduced in all runs except one (the pair-sum method with no hydrogen-bonding information), which failed to converge. The fragment placement method was successful for this graph; the correlation coefficients were close to 1.0. The results for cAMP_b were very similar and are not given here.

Folate

Folate was taken from recombinant human dihydrofolate reductase (1DHF) [23]. The structure was found to have 49 vertices, and a total of 19 placeable nodes in the folate skeleton, giving rise to 10^{11} ways of arranging fragments (compared to 10^{31} if the aromatic ring were to be overlaid with aliphatic fragments). The number of surface points used was 181. The results of the annealing runs are shown in Table 3.

Folate itself was not reproduced in the self-placement runs, except in one case where Spearman's rank correlation coefficient was the method of assessment with no hydrogen-bonding information. Correlation coefficients were close to 1.0. The new structures created were mostly halogenated aromatic ring variants of folate, and may be

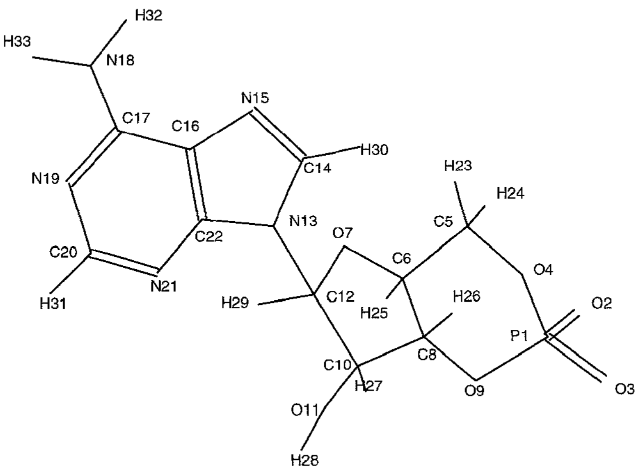

Fig. 2. The structure and numbering scheme of cAMP_a. Note that the numbering scheme for cAMP_b is identical.

TABLE 2
ANNEALING RESULTS USING cAMP_a AS A MOLECULAR GRAPH AND VARYING ANNEALING CONDITIONS

Method of calculation	Method of assessment	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.958±0.001	70.1±19.9	10/10	0
	r_s	0.967±0.002	87.5±20.4	10/10	0
	eo	4372±62.85	48.0±8.3	9/9	0
Electrostatic potential with hydrogen bonds specified	r	0.958±0.001	82.2±8.8	10/10	0
	r_s	0.968±0.000	111.7±23.0	10/10	0
	eo	4352±28.20	85.1±22.2	10/10	0

See the footnote to Table 1 for an explanation of the parameters.

related to a transferability problem with some of the fragment placements, leading to entrapment of the algorithm in nonoptimal minima.

Restricting the number of halogenated groups

Careful scrutiny of the results with the folate molecular graph and a few others revealed some placements which had unfavourable multiple halogenation. This phenomenon was also observed in site-directed placements [14]. Two possible explanations could be given for the placement of multiple halogens on the skeleton. First, the annealing procedure could fail to reach equilibrium owing to the increased number of possible fragments to choose from in the fragment bins: given a particular four- or five-vertex connectivity pattern in which at least two termini are present, there are several halogenated fragments as compared with only one 'correct' fragment. The effect is to create a difficult landscape with many local minima that prematurely trap the algorithm before the global minimum is found. This phenomenon has been described elsewhere for an analogous problem [7]. The second explanation for multiple halogenation, in the case of site-directed atom assignment, could be that the placement is actually better than the original. The runs using folate molecular graphs were repeated using the limit of two halogen groups; this restriction was included as an option to be used in the atom assignment method. The results of the modified annealing procedure are given in Table 4.

This restriction on the number of halogen atoms to be placed clearly improved the reproducibility of folate from its molecular electrostatic potential. This was observed with either hydrogen-bond option and with all methods for evaluating the objective function. Use of the option did not appreciably affect the CPU time.

Dihydroxy benzoate

The DHB structure was taken from the crystal of *p*-hydroxybenzoate hydroxylase (1PHH) [24] and had 16 vertices and 62 surface points. The number of placeable nodes was 7, giving a total number of placements of 2000. When aliphatic placements are considered on the ring, the number of nodes becomes 13, with around 10^{11} possible placements; this aliphatic placement is forbidden to prevent poor transferability of fragment charges. The results of the annealing runs using the DHB molecular graph are shown in Table 5.

As with the AMP molecular graph, the annealing self-placement results with DHB displayed a perfect performance using either Pearson's correlation coefficient or Spearman's rank, while the pair-sum method gave marginally poorer results. Placement was extremely fast due to the small number of nodes and fragments which could be placed.

Retinol

The structure of retinol was taken from the crystal of human retinol binding protein (1RBP) [25]. Retinol con-

TABLE 3
ANNEALING RESULTS USING FOLATE AS A MOLECULAR GRAPH AND VARYING ANNEALING CONDITIONS

Method of calculation	Method of assessment	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.968±0.009	110.3±9.6	0/10	5.0±1.1
	r_s	0.959±0.007	141.2±16.7	1/10	2.9±1.5
	eo	19 680±1915	60.9±11.6	0/10	4.5±0.8
Electrostatic potential with hydrogen bonds specified	r	0.974±0.001	133.3±13.0	0/10	3.1±0.8
	r_s	0.957±0.008	119.4±16.8	0/10	3.6±1.9
	eo	18 990±1915	73.3±11.6	0/10	4.4±0.9

See the footnote to Table 1 for an explanation of the parameters.

TABLE 4
ANNEALING RESULTS USING FOLATE AS A MOLECULAR GRAPH UNDER VARYING ANNEALING CONDITIONS

Method of calculation	Method of assessment	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.982±0.016	91.6±14.3	7/10	3.0±1.7
	r_s	0.957±0.010	108.3±17.7	7/10	2.0±0.0
	eo	9826±2772	52.8±5.6	7/10	2.7±1.2
Electrostatic potential with hydrogen bonds specified	r	0.989±0.005	125.5±12.4	8/10	2.0±0.0
	r_s	0.958±0.010	166.1±54.8	6/10	2.0±0.8
	eo	11 880±3379	68.9±8.1	4/10	1.7±0.5

See the footnote to Table 1 for an explanation of the parameters. In this annealing run, the number of halogens to be added was limited to a maximum of two.

sisted of 51 vertices and had 208 surface points. The number of placeable nodes was 21, giving 10^{21} possible arrangements of fragments. Two properties have been considered for this ligand, i.e., its electrostatic potential and its hydrophobic potential (Φ). The annealing results for self-placement on the retinol graph are shown in Table 6.

The program had considerable difficulty with recreating retinol itself, given its surface potential (electrostatic or hydrophobic). The electrostatic potential similarity was much weaker than that of the hydrophobic potential. Some very bad placements occurred, with one condition making about 16 atom misplacements. Most of the self-placement runs gave chlorinated derivatives of retinol. One possible explanation for this may be related to the configurational landscape created by a structure which is essentially made up of repeat building blocks. This leads to a degenerate system, where several arrangements of atoms give roughly the same surface potential, thus producing a terrain with numerous local minima. The best annealing conditions were found when the objective function was 50% electrostatic and 50% hydrophobic, hydrogen bonds were specified, and Pearson's correlation coefficient was the method of assessment. These conditions reproduced retinol in five out of 10 runs. Reproducibility problems again appeared to originate from too many halogen atom placements.

The same self-placement conditions as those just specified were used for the retinol molecular graph, with additional restrictions on the number of halogen groups (two or less) and no charged fragments. The results are listed in Table 7.

The fourth row of Table 7 illustrates the difficulty presented to the optimization algorithm for a pure electrostatic placement. The final value of the objective function is 0.441 with correct placements; for maximum similarity in electrostatic potential, this value should approach 1.0. This extremely low value would suggest that the atomic residual charges on the fragments give a poor representation of the electrostatic potential compared with charges calculated explicitly for the whole molecule. In this example the hydrophobic potentials, together with the hydrogen-bond option, score well for the objective function; they consistently reproduce retinol and give high values for the correlation coefficients. Moreover, the penalty function values used were equally effective. As with the folate results, the new restrictions gave improved performance in the self-placement runs, with considerably more perfect placements than in the runs performed in Table 6, especially when hydrogen-bonding information was given. However, these better atomic assignments have been achieved at the expense of poorer values for the objective function with electrostatic potentials.

Discussion

This paper has described a simulated annealing algorithm for the optimization of atom assignment onto molecular graphs by placement of small fragments with trans-

TABLE 5
ANNEALING RESULTS USING DHB AS A MOLECULAR GRAPH AND VARYING ANNEALING CONDITIONS

Method of calculation	Method of assessment	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.997±0.000	4.0±0.6	10/10	0
	r_s	0.996±0.000	4.6±1.0	10/10	0
	eo	1992±761.9	3.2	7/10	1.7±1.2
Electrostatic potential with hydrogen bonds specified	r	0.997±0.000	3.3±0.4	10/10	0
	r_s	0.996±0.000	3.8±0.5	10/10	0
	eo	1631±324	2.9±0.4	9/10	1

See the footnote to Table 1 for an explanation of the parameters.

TABLE 6
SELF-PLACEMENT ANNEALING RESULTS USING RETINOL AS A MOLECULAR GRAPH

Method of calculation	Method of assessment	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.343±0.080	227.5±17.4	0/9	8.2±2.3
	r_s	0.606±0.064	296.2±26.4	0/10	16.6±2.5
	eo	10 840±3934	254.0		7.5±1.9
Electrostatic potential with hydrogen bonds specified	r	0.424±0.037	117.5±49.1	1/10	2.3±1.2
	r_s	0.728±0.066	185.8±15.0	0/10	5.3±1.3
	eo	6990±2392	177.4±8.3	0/10	2.7±1.1
Hydrophobic potential	r	0.749±0.077	184.6±10.9	0/10	6.5±2.9
	r_s	0.836±0.042	337.6±31.2	1/10	4.7±2.2
	eo	60.70±18.15	145.1±16.0	0/10	4.2±1.5
Hydrophobic potential with hydrogen bonds specified	r	0.779±0.148	174.6±10.0	3/10	3.0±1.3
	r_s	0.817±0.029	297.4±24.4	1/10	2.6±0.8
	eo	47.79±13.04	144.3±9.5	1/10	1.9±1.1
Electrostatic and hydrophobic potential with hydrogen bonds specified	electrostatic 0.75 with hydrophobic 0.25; r	0.414±0.051	207.1±263.0	1/10	2.4±1.8
	electrostatic 0.5 with hydrophobic 0.5; r	0.853±0.093	121.8±10.8	5/10	2.2±0.8
	electrostatic 0.25 with hydrophobic 0.75; r	0.769±0.066	136.8±13.3	0/10	1.5±0.5

The surface potential used in the optimization of fragment placement here is that of retinol (electrostatic and/or hydrophobic). See the footnote to Table 1 for an explanation of the other parameters.

ferable molecular properties [1]. The notion of transferable molecular properties has been pioneered by Bader [26] and was recently summarised for drug design [27]. If automated de novo drug design is divided into two processes [12,28], then: firstly, crude structures can be created with good shape complementarity, and the dummy atoms can be stripped away to leave a 3D molecular graph (effectively a shape graph); secondly, atoms can then be assigned to the graph so that the resultant molecule has an optimized complementarity to the site with respect to various properties, as well as retaining the initial general shape complementarity. The currently implemented properties that can be handled by the routine are electrostatic potential complementarity and hydrophobic potential similarity. Moreover, the same algorithm can be used for a radically different approach to drug design from studies of molecular similarity [11].

The major difficulty in any complex optimization problem is to define an adequate objective function. This difficulty is compounded if the user is not quite sure how to express the function. What is the best way of describing electrostatic complementarity, or similarity, projected onto a set of points? Absolute differences, or rms values, between the potentials could be used, but they have the disadvantage of comparability where the scales of potential are widely different. It is noteworthy that the use of absolute differences between potentials as an objective function for placement did not give consistency in self-placement tests of the algorithm. A correlation coefficient has the advantage of providing a uniform scale for comparison. We tried two methods, Pearson's correlation coefficient based on absolute values of potential, and rank correlation based on differences in rank; the latter method measures the complementarity in pattern of potential. The gradient of the regression line should be approximately −1 for complementarity and +1 for similarity; we did not investigate whether it would also be possible to use an amalgam of gradient and correlation coefficient in the objective function. If two or more parameters are to be compared simultaneously in the objective function, there is an added problem of how the parameters are to be relatively weighted. At this moment we do not have sufficient experience to offer a guideline for a composite of electrostatic and hydrophobic potentials.

Any optimization algorithm needs to be appropriately scaled, so that the trajectory is not trapped prematurely with differently sized problems. The scaling procedure used here is identical in concept to that investigated earlier for molecular matching [6]. The initial differences, ΔE, between transitions are scaled to give an appropriate acceptance probability at a given starting value of T. This enables the algorithm to be used efficiently for any problem size. Further efficiency is gained by allowing the algorithm to 'cool' faster where large changes are made to the objective function and more slowly when a solution is approached. However, the generator function in the transition mechanism has not been optimized; we have simply used random selection from a uniform distribution, both for the choice of the node for placement and for selection of a fragment from the bin. This procedure differs from other forms of simulated annealing, where the generator is based on a Boltzmann or Cauchy distribution [4,29]. An investigation of more elaborate generator functions would be an obvious area for future improvements.

An efficient method of graph perception for classifying fragments into bins, in a preprocessing step, is essential for the annealing algorithm. This procedure ensures that the connectivities are correct for all placements. We have found this to be a great advantage over an earlier, but abandoned, strategy for trying just to use the fragment to place a single atom on a node, without taking into account a more extended graph. A further improvement would be to have a larger set of fragments to select from the bins. For example, our library of rings is not yet large [30] and does not include the 5,6 duplex containing sulphur which is needed for an efficient placement of fragments onto JEDEUS [3]. However, the algorithm is able to cope with this problem by attempting an aliphatic placement of all atom types instead; nevertheless, it would be better to have an aromatic set of charges for all possible aromatic ring systems.

Graph perception and fragment placement are strongly connected in the annealing objective function. The characteristics of each node have been perceived and this information controls the choice of available fragments for placement at that node. Ridiculous placements have to be eliminated by rules, but the rules should not be too stringent to cause the algorithm to lock a placement and thereby impose a particular class of structure early in the trajectory. The algorithm must be free to try poor placements in order to generate an optimum in the objective function. Ordinary downhill gradient optimization would be disastrous; a hill-climbing procedure is obligatory. Our emphasis has been to allow sufficient leeway in placement by dividing the annealing of fragment placement into a

TABLE 7
SELF-PLACEMENT ANNEALING RESULTS USING RETINOL AS A MOLECULAR GRAPH UNDER VARYING ANNEALING CONDITIONS

Method of calculation	Potential type and assessing method	Mean value of the final objective function	Time (s)	Ratio perfect runs/total runs	New atoms
Electrostatic potential	r	0.317±0.115	101.2±14.0	2/10	3.2±1.8
	r_s	0.319±0.052	210.1±75.2	0/10	5.0±1.7
	eo	8090±3480	254.0±21.6	0/10	5.0±1.7
Electrostatic potential with hydrogen bonds specified	r	0.441±0.019	183.4±123.4	9/9	0
	r_s	0.259±0.084	187.8±115.3	5/8	2.0±0.0
	eo	4844±608.3	208.06±107.5	6/9	2.0±0.0
Hydrophobic potential	r	0.804±0.052	228.7±68.4	0/10	3.2±1.0
	r_s	0.846±0.031	341.4±38.4	0/10	2.4±0.8
	eo	54.27±10.21	158.4±50.1	0/10	2.4±0.8
Hydrophobic potential with hydrogen bonds specified	r	0.964±0.028	176.0±83.1	8/8	0
	r_s	0.972±0.025	224.1±28.2	7/7	0
	eo	15.24±3.310	208.0±92.0	8/8	0
Electrostatic and hydrophobic potential with hydrogen bonds specified	electrostatic 0.75 with hydrophobic 0.25; r	0.428±0.049	265.4±92.8	8/9	2.0±0.0
	electrostatic 0.5 with hydrophobic 0.5; r	0.973±0.016	280.2±116.9	8/8	0
	electrostatic 0.25 with hydrophobic 0.75; r	0.949±0.065	146.7±60.7	6/8	2.0±0.0

The surface potential used in the optimization of fragment placement here is that of retinol (electrostatic and/or hydrophobic). The number of halogen atoms to be added was limited to a maximum of two; no charged fragment placements were allowed. See the footnote to Table 1 for an explanation of the other parameters.

number of stages. The first annealing step has just a few rules, which are related to inconsistent geometry, input options such as placement of a particular hydrogen-bonding type at certain nodes, and disallowing double bonds to be connected directly to an aromatic fragment. We never consider monitoring any other bond order violations at this stage. If we were to attempt to get everything right during the annealing step, the algorithm would degenerate into a brute-force placement and be hopelessly slow. This first annealing step gives an optimum placement of fragments on the graph, but there will be inevitable bond order violations. These violations are checked in step 2. Adjacent aliphatic fragments can only be connected by single bonds; this does not mean that an ideal multiple bond is not placed between two nodes, but simply that the connection mechanism is shifted to the side of the multiple bond. A re-annealing is performed around the offending nodes to find a suitable local arrangement that does not cause a violation. The structure generated at this stage is a viable candidate assignment, but it needs refinement. Until this stage, the input geometry of the molecular graph has been retained irrespective of the atom types suggested by the fragment placement. The initial bond lengths are now replaced by appropriate bond lengths. It could be argued that this limited geometry optimization should be performed at each proposed transition; our guess is that it would not be a cost-effective strategy. The new graph, with corrected bond lengths, is repositioned over the old graph and the correlation coefficients are recalculated. In a final refinement step, the algorithm scans the atom assignments for atoms contributing poorly to the current optimal value of the objective function. The worst 25% are marked for re-annealing and the first annealing step is repeated on them. We have consistently found that this re-annealing step is of great value in removing the few bad assignments, due to the weak ergodicity imposed on the algorithm by time constraints. This improvement strategy has been used effectively before in molecular structure-matching problems [8]. It is quite possible that this general two-stage annealing strategy could be refined further, to reduce the CPU time, by doing a much faster anneal in the first stage. Detailed analysis of many annealing trajectories (Barakat, M.T. and Dean, P.M., unpublished observations) shows that the majority of acceptable transitions occur in early Markov chains; in later chains the algorithm makes few significant new exchanges. It may be possible to develop a first-stage stop criterion that is less stringent than the one used here.

The algorithm has been extensively tested and five representative examples are reported here. Various input options allow a wide variety of usage. Hydrogen bonds can be explicitly specified and the number of halogens can be limited to a certain number. Three methods for assessment of the objective function have been tested, i.e., two correlation coefficients and the summation of absolute differences. The latter method did not perform as well as the correlation coefficients. There was little to distinguish between the two correlation coefficients, but one can perhaps recommend Pearson's correlation coefficient as the method of choice. Comments that follow are confined to results using this method of correlation.

Assignment of atoms to the 3D graphs of AMP, cAMP or DHB provided correctly replicated answers with or without specifying hydrogen bonds. However, assignment onto folate was poor, with about three to five atoms constantly mis-assigned. These wrong placements were almost all chlorine/fluorine atoms. If the number of halogen atoms allowed was restricted to 2, the assignments improved significantly to give 70–80% perfect replications. This modification did not change the amount of CPU time needed for convergence of the algorithm. It would appear from this example that the halogen atoms are creating a landscape difficulty for the annealing routine by being electrostatically equivalent. However, this simple explanation of equivalence may not be the only one, since one would expect some halogens to be included when the restriction on their numbers is set to 2.

The retinol molecular graph provided an extremely difficult and complex test example. The electrostatic potential, derived from a correct fragment placement, was very badly correlated with the potential calculated explicitly. Thus, even with a poor representation of the electrostatic potential, the optimum similarity could be obtained by restricting the halogen atoms (compare Tables 6 and 7), whilst specifying the hydrogen bonds. This finding offers hope that the fragment placement procedure developed in this paper is not critically dependent on the method of calculating the residual partial atomic charges. The retinol molecule contains large hydrophobic segments. The use of a hydrophobic potential gives a reasonable correlation coefficient of about 0.96 for correctly placed fragments. A restriction on the number of halogen atoms improves the assignment with a hydrophobic objective function. A compound objective function of electrostatic and hydrophobic potential gives reproducible results only if there is a restriction on the number of halogen atoms. A penalty value in the range 0.25–0.75 for the electrostatic potential does not appear to affect the results appreciably.

It is quite clear from a comparison between Tables 3, 4 and 6, 7 that the restriction in the number of halogen atoms allowed is the most important factor affecting the reproducibility of the results. Where the number of halogen atoms, in an unrestricted placement, is close to 2, a restriction dramatically improves the atom assignment. However, if the number of halogen atoms is much greater than 2, a restriction of halogen atoms to 2 gives a minor improvement. This general finding has also been noted for complementarity-based atom placements [14].

Compound objective functions composed of different features present a number of theoretical problems that need to be considered. The simplest procedure is to weight the features in a penalty function, a notion that was adopted here. However, the difficulty with this is that the whole molecular surface is treated in the same way. Thus, the molecule is considered to be either electrostatic, hydrophobic, or a fractional representation of the sum of the potential. The fault with this strategy is that molecules usually exhibit regional differences in potential that are not reflected in that style of objective function. An objective function needs to be designed to take these regional differences into account. Thus, one would like to weight a particular region as strongly or weakly electrostatic, with a converse weight for the hydrophobic regions.

The atom assignment problem, although a combinatorial problem, can be effectively solved by an appropriate optimization method. The simulated annealing routine proposed here is adequate to reproduce the original potential given in the five self-placement tests. We have illustrated a selection of successes and difficulties encountered in the optimization and outlined how the difficulties can be surmounted. Multiple halogenation was the principal difficulty generated by the algorithm, although this can be overcome by a restricted but sensible placement of no more than two halogen atoms. The results presented here are encouraging for the optimization of automated de novo design problems. In two subsequent papers we extend our test system to the two important design strategies, design from complementarity with the site [14] and design from molecular similarity within a molecular surface envelope [15].

Acknowledgements

We wish to thank the Wellcome Trust for a Wellcome Prize Studentship (M.T.B.) and for a Principal Research Fellowship (P.M.D.). Part of this work was carried out in the Cambridge Centre for Molecular Recognition.

References

1 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 341.

2 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 407.

3 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 351.

4 Ingber, L. and Rosen, B., Math. Comput. Modelling, 16 (1992) 67.

5 Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A. and Teller, T., J. Chem. Phys., 21 (1953) 1087.

6 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 295.

7 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 4 (1990) 317.

8 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 107.

9 Papadopoulos, M.C. and Dean, P.M., J. Comput.-Aided Mol. Design, 5 (1991) 119.

10 Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 7 (1993) 155.

11 Dean, P.M., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic and Professional, London, 1994, pp. 1–23.

12 Dean, P.M., Barakat, M.T. and Todorov, N.P., In Dean, P.M., Jolles, G. and Newton, C.G. (Eds.) New Perspectives in Drug Design, Academic Press, London, 1995, pp. 155–180.

13 Fauchère, J.-L., Quarendon, P. and Kaetterer, L., J. Mol. Graphics, 6 (1988) 202.

14 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) in press.

15 Barakat, M.T. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) in press.

16 Aarts, E.H.L., Korst, J.H.M. and Van Laarhoven, P.J.M., J. Stat. Phys., 50 (1988) 187.

17 Huang, M.D., Romeo, F. and Sangiovanni-Vincentelli, A., In Proceedings of the IEEE International Conference on Computer-Aided Design, Santa Clara, 1986, pp. 381–384.

18 Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P., Numerical Recipes in FORTRAN, Cambridge University Press, Cambridge, 1992, pp. 273–274.

19 Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, G. and Taylor, R., J. Chem. Soc., Perkin Trans. II, (1987) S1.

20 McLachlan, A.D., Acta Crystallogr., A38 (1982) 871.

21 Diedrichs, K. and Schultz, G.E., J. Mol. Biol., 217 (1991) 541.

22 Weber, I.T. and Steitz, T.A., J. Mol. Biol., 198 (1987) 311.

23 Davies, J.F., Delcamp, T.L., Prendergast, M.J., Ashford, V.A., Freisheim, J.H. and Kraut, J., Biochemistry, 29 (1990) 9467.

24 Schreuder, H.A., Van der Laan, J.M., Hol, W.M.G. and Drenth, J., J. Mol. Biol., 199 (1988) 637.

25 Cowan, S.W., Newcomer, M.E. and Jones, T.A., Protein Struct. Funct. Genet., 8 (1990) 44.

26 Bader, R.W.F., Atoms in Molecules: A Quantum Theory, Oxford University Press, Oxford, 1990.

27 Popelier, P.L.A., In Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie Academic and Professional, London, 1994, pp. 217–244.

28 Barakat, M.T., Todorov, N.P. and Dean, P.M., In Sanz, F. (Ed.) Trends in QSAR and Molecular Modelling 94 (Proceedings of the 10th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling 94), Prous, Barcelona, 1995, in press.

29 Szu, H. and Hartley, R., Phys. Lett. Ser. A, 122 (1987) 157.

30 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 385.