Molecular recognition using a binary genetic search algorithm

A.W.R. Payne* and R.C. Glen†

Management Services Division and Department of Physical Sciences,† Wellcome Research Laboratories, Beckenham, Kent, UK*

A genetic algorithm has been devised and applied to the problems of molecular similarity, pharmacophore elucidation, and determination of molecular conformation. The algorithm is based on a binary representation of molecular position and conformation. Using the genetic operators, crossover, mutation, and selection near optimum conformations and orientations of molecules may be determined which best-fit defined constraints. The constraints may be any useful function for example, intermolecular or intramolecular distances, electrostatic potential on a surface, or volume overlap. Problems with up to 30 degrees of freedom have been tackled successfully.

Keywords: genetic algorithm, evolution, molecular similarity, pharmacophore recognition, conformational analysis

INTRODUCTION

One of the challenges in formulating useful structure–activity relationships is in deciding which regions of molecules in a series spatially overlap in their common molecule-receptor interactions. The solution of this problem is of fundamental importance in any drug design strategy. Incorrect assignment of overlapping or similar regions of a molecular series will invariably lead to inaccurate predictions of the activities of new molecules whose structural modifications are based on these molecular similarities.

There are many examples of structurally dissimilar molecules that are thought to act at a common receptor, for example, the endogenous hormone enkephalin, a pentapeptide, and the nonpeptide analogue naloxone, an opiate receptor antagonist.[1] Other examples are adenosine, a natural central nervous system depressant, and caffeine, an antagonist at the adenosine receptor.[2]

Dissimilar molecules may be compared using molecular modeling techniques to generate conformations and overlapping spatial arrangements that accommodate perceived molecular properties. More numerical methods include comparison of electron densities, atom charge distributions, dipole moments, volume overlaps, rms fitting of similar functional groups, and electrostatic and lipophilicity potentials.[3–9] The active analogue approach[8] of Marshall et al. (which is probably one of the most successful computer-aided methods currently applied to the drug design problem), relies heavily on the identification of similar regions in a series of molecules to select as overlapping regions or common functional groups.

The related problem, when the structure of an active site is known from X-ray crystallographic analysis, is how to dock a putative ligand into the active site in the most energetically favorable position. This is a nontrivial problem, as subtle differences in the active site can force a completely different binding mode upon the same molecule in, for example, two versions of the same enzyme from different species. An example is DHFR (dihydrofolate reductase) inhibited by TMP (trimethoprim). TMP is bound in two different modes, which differ by a rotation of almost 180° between the bacterial and avian enzymes.[10,11] This example highlights a major problem in comparing molecules in isolation of the receptor.

Some current methods of overlaying molecules in their most similar conformations and orientations in a more automated way include least-squares optimization of the overlay of common functional groups[12] using constrained molecular dynamics or simulated annealing; overlap of molecular fields (CoMFA);[13] molecular rotation followed by cluster analysis of molecular rotational space;[14] or simplex optimization of molecular rotational space (ASP).[15]

There are a number of major problems that must be addressed for successful orientation and overlaying of molecular series. The first step is to generate a suitable target function to optimize, e.g., a binding energy or a similarity index that is accurate, yet simple enough to evaluate in a reasonable time (reasonable equates to how fast a computer is available). At one extreme, *ab initio* molecular orbital methods generate accurate wavefunctions from which similarity indices between molecules may be calculated.[9] However, at present these calculations are impractical due to the size of molecules of interest and the repetitive nature of the fitting algorithm. Here we have adopted descriptors, which in this algorithm are evaluated initially and remain unchanged (e.g., atom-centered charges) or can be quickly reevaluated (e.g., molecular volume) during the course of the run.

The second step is to solve the multiple-minima problem.

Color Plates for this article are on pages 121–123.

Address reprint requests to Dr. R.C. Glen at the Department of Physical Sciences, Wellcome Research Laboratories, Langley Court, Beckenham, Kent, BR3 3BS, UK.

Received 28 April 1992; revised 28 May 1992; accepted 2 June 1992

Since most molecules of interest contain rotatable bonds and flexible rings, this results in a combinatorial explosion of possible solutions. To address this problem, we have adopted a genetic algorithm[16] (GA) that attempts to mimic some of the optimization qualities of natural selection. Darwin[17] propounded the theory that organisms tend to produce offspring varying slightly from their parents, that the process of natural selection tends to favor the survival of those best adapted to their environment, and that by the operation of these factors, new species may arise widely differing from each other and from their common ancestors. Genetic algorithms are an attempt to use the power of natural selection in finding optimum solutions to complex functions. We wish to mimic this process to optimize the orientation and conformation of molecules by quickly and efficiently fitting them to constraints. Molecules in particular conformations and orientations that best fit the constraints are analogous to species, and these are allowed to breed more offspring similar to themselves. The constraints applied to the molecules are analogous to the environment.

Third, not all parts of molecules in a series need to overlap at the same time. Some parts of molecules interact with the receptor, and some do not (only one side of a molecule, for example, may be interacting with a receptor; the other side may be exposed to solvent). Indeed, since most interactions of this type are of a non-covalent nature, there is an ensemble of binding modes between the receptor and ligand determined by the energetics of interaction in different binding modes, as well as resolvation into the medium.

These problems impose limitations on what can be expected from any similarity exercise in the absence of a full description of the receptor and its environment. However, in many cases the high degree of similarity required of receptor selective analogues is probably enough to provide sufficient information and allow the determination of a reasonably likely overlap.

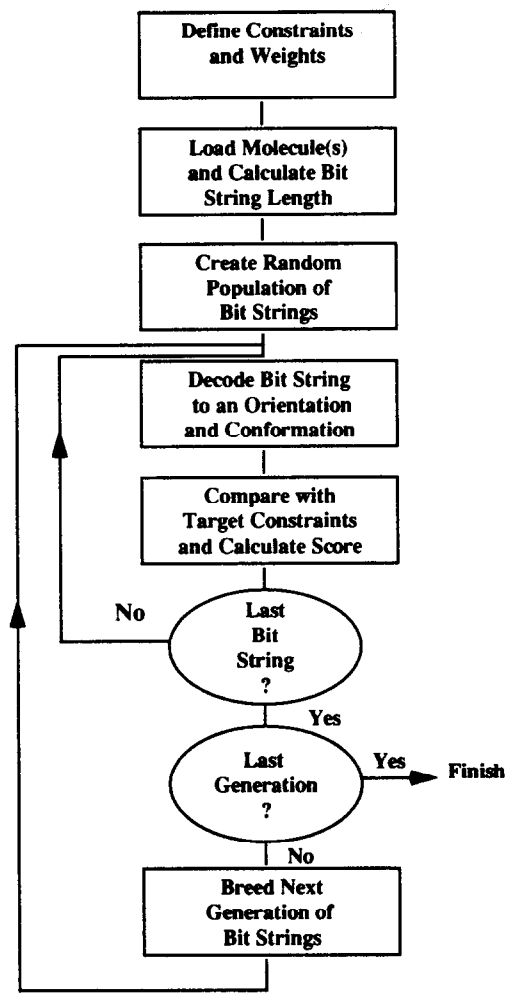


Figure 1. Flowchart showing GA optimization of orientation and conformation.

METHOD

A genetic algorithm[16] (GA) has been devised and implemented that attempts to optimize the fit of flexible molecules to a set of constraints. The genetic algorithm uses a blind search strategy, requiring no knowledge of the properties of the function that it is optimizing, so enabling the algorithm to be applied to a variety of molecular fitting problems. The constraints may be any useful function, e.g., intramolecular distance constraints, shape similarity, or charge distributions.

The algorithm uses a sequence of binary digits (a bit string) to describe the orientation and conformation of a given molecule. Bit strings are analogous to DNA, and contain information that is translated into an orientation and conformation of a molecule.

When the genetic operators, crossover, selection, and mutation (which may also be applied to DNA sequences), are applied to a population of these binary sequences, strings that code for near optimum orientations and conformations arise after a number of generations (iterations).

Figure 1 contains a flowchart showing the sequence of steps involved in a GA optimization of the orientation and conformation of a molecule.

Definition of constraints

The first step is to define a set of constraints within which a molecule will be fitted or compared. In the experiments involving the optimum fit of one molecule onto another, or fitting to a pharmacophore, constraints may be simple distances between atoms or functional groups. In this case, intramolecular or intermolecular distances are defined, and each is associated with a target value (or range) and a weight. The weight term allows different degrees of importance to be attached to each constraint. For example, a carbonyl and an amine in a molecule may be constrained to be 7.5 Å apart with a weight of 1.5.

In addition, the shape and charge distribution may be defined by mapping these molecular properties onto a suitable surface. The template molecule (or, if there is no static molecule, the first) is centered on the origin and is surrounded by a surface of points.[18,19] The surface may be configured to be spherical with a defined radius and point density (which is used here in similarity searching) or more complex, for example, a solvent-accessible surface at an extra radius. The shape of the template or target molecule is described by a set of distances; each distance is associated with one of the points

on the surface. Here, two different measures of shape are investigated. The first (Shape 1) calculates the distance from a surface point to the nearest atom in the molecule minus the van der Waals radius of that atom. The second (Shape 2) measures the distance along a radial vector from the origin to the van der Waals surface of the molecule; each radial vector starts at the origin and points in the direction of a calculated surface point.

The charge distribution of the molecule may be calculated using any preferred method, for example *ab initio*,[20,21] semi-empirical[22] (Gaussian-80 UCSF and Mopac Version 5.0 are used here) molecular orbital methods, or partial equalization of orbital electronegativity (PEOE)[23–26] (only the sigma contributions to atom charges have been used here).

Charge distributions may be calculated from an initial comformation, and remain unchanged during the course of the fitting procedure. Ideally, these would be recalculated upon conformational change; however, consistent charges may be good enough[27] to allow reasonable comparison. Electrostatic potential or distributed multipole-derived charges are preferred, but are time consuming to calculate.[21] GA methods are currently being developed to evolve novel molecules within constraints that require a method of charge calculation that is very fast. This will be reported elsewhere.

Two methods for expressing the charge distribution are evaluated: the first method calculates the electrostatic potential[28–30] at each of the points on the surface using the atom-centered charges and atomic screening constants. This calculation method has deficiencies, particularly for aromatic systems, but is very fast. The second method maps the partial charge of the nearest atom to the nearest point of the surface. The coordinates of the points on the surface and the associated shape and charge descriptions of the template molecule are then available for comparison with other molecules.

As well as these distance and surface constraints, molecular volume and overlap volume[31] between molecules may be calculated and used to optimize molecular similarity. For example, we may wish to optimize molecular overlap, and hence minimize the increase in volume in overlaying a series of molecules. Van der Waals contacts (used to prevent atomic overlap) are also calculated to eliminate impossible conformations.

Structure manipulation and optimization against constraints

The genetic algorithm is initialized by loading a list of the constraints to be used in the calculation. These are constraints for use in the optimization (for example, distance constraints), and a weighting factor to apply to each class of constraint. Also loaded, if required, are the surface points and the properties associated with each point, previously calculated from the template molecule (or pharmacophore).

The molecule to be fitted is loaded, and the rotatable bonds and free corners of rings are identified. A rotatable bond is defined here as a single bond (of bond order 1) not in a ring between two nonterminal atoms. A penalty may be introduced for amide bonds in the cis conformation, or they may be allowed to rotate freely. A free corner is defined as a ring atom that can be flipped across the plane of the ring without straining the ring in which it lies, distorting a neighboring

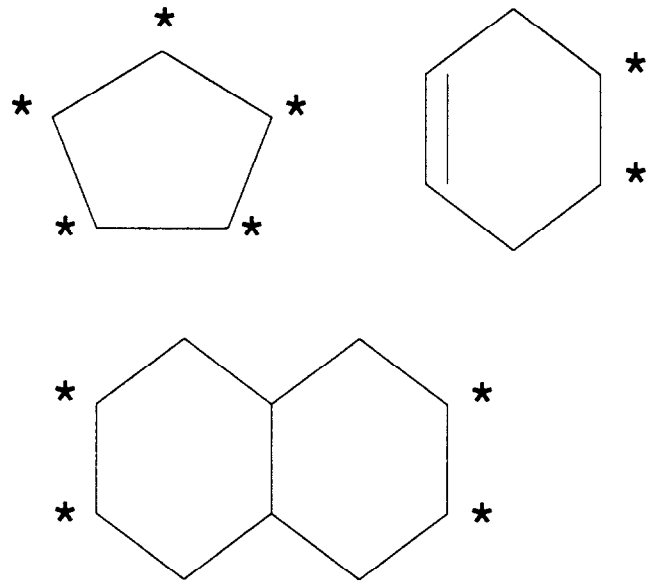


*Figure 2. Free ring corners (marked with a *) of three simple ring systems.*

ring, or causing a rotation about a double bond. Flipping free corners allows partial flexibility of rings without the algorithmic complexity of full conformational flexibility.[32] There are, of course, severe limitations on this algorithm; for example, a free ring corner in the plane of the ring will not be moved unless it is distorted initially from the plane.

To find the free corners in a molecule, the rings are analyzed with a spanning tree algorithm to identify the smallest set of smallest rings (SSSR).[33] Next, the bonds associated with the rings in the SSSR are identified, and the number of appearances of each bond in the SSSR is counted. If an atom is a free corner, then the bonds attached to the atom and the adjacent bonds must be single bonds, and in addition have occurred only once in the SSSR. Figure 2 shows the free ring corners of three simple ring systems.

The conformation and orientation of a molecule to be fitted to the constraints is described by a binary sequence. The binary sequence can be broken down into four regions: the first region describes the translation of the molecule along the three axes, the second describes the rotation of the whole molecule around the axes, the third describes rotations around each of the rotatable bonds in the molecule, and the fourth is the conformation of the rings within the molecule. The transformations are always applied to original molecules; hence favorable transformations accumulate that cause the molecules to fit the constraints.

The rotational, translational, and torsional degrees of freedom are each described by 8 bits, while each ring flip uses only 1 bit. The length of the binary sequence used to describe the conformation and orientation of the molecule can then be calculated as:

$$\text{sequence\_length} = 24 + 24 + (8 \times \text{rotatable bonds}) + \text{free ring corners}$$

The binary sequence shown in Figure 3 describes a molecule with four rotatable bonds and five free ring corners.

The first 24 bits represent translations of the molecule

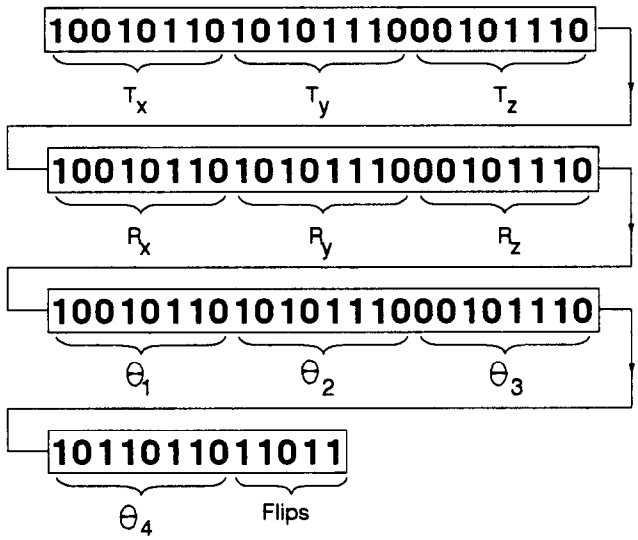


Figure 3. Coding of orientation and conformation with a binary string.

along each of the axes. The 8-bit code for each axis at present maps to a displacement of the molecule along the axis in the range ± 3 Å. This distance range can, of course, be modified to suit the problem.

The second 24 bits represent the rotations around the three coordinate axes. The 8-bit code for each axis maps to a rotation angle of 0–360°. In the example shown, a further 32 bits describe the rotations around the four rotatable bonds, with each group of 8 bits mapping to a rotation angle of 0–360°. Finally, a further 5 bits are used to describe the position of the five free ring corners. The value 1 represents a flip, and 0 represents no flip.

To initiate the problem, a population of random bit strings is generated (satisfying the previously described bit string length), with each string coding for a different orientation and conformation of the molecule. Before the quality of the fit to the constraints is measured, each string is decoded into an orientation and conformation of the molecule.

The decoding process begins by converting the 8-bit codes used to describe the translations and rotations into real numbers in the defined ranges (the 8-bit binary code does not need to be converted to an integer, as the computer stores integers in binary).

$$T_x = (6.0 \times (8\text{ bit code})/255.0) - 3.0$$

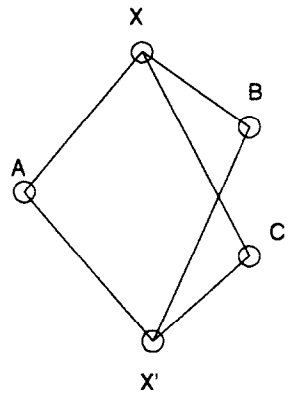
$$R_x = 360.0 \times (8\text{ bit code})/255.0$$

$$\Theta_i = 360.0 \times (8\text{ bit code})/255.0$$

where T_x is the translation in x (similarly for the y and z axes), R_x is the rotation about the x axis (similarly for y and z axes), and Θ_i is the rotation about a torsion angle in degrees clockwise.

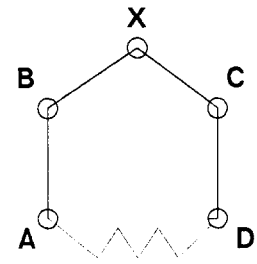
The molecule is then translated along the coordinate axes by the displacement defined by the bit string. Similarly, rotations about fixed coordinate axes are performed. The rotations around bonds involves multiplying the position vectors for all the atoms on one side of the rotatable bond by a homogeneous transformation matrix for rotation about an arbitrary axis[34] determined by the bond vector. The portion of the molecule having fewer atoms is moved during bond rotations, as this speeds up the transformation and assists optimization by minimizing the number of atoms moved.

If the bit associated with a free corner is set (to 1), the algorithm will attempt to flip the ring atom across the plane of the ring, adjusting the position of any atoms attached to the ring atom and its neighbors. To preserve the correct bond lengths and angles, the new position for the ring atom must be the same distance from its nearest neighbors and next nearest neighbors in the ring as in the original position. The new position can be found by calculating the points of intersection of three spheres centered on the two neighboring atoms and one of the next nearest neighbors in the ring;[35] one of the points of intersection must be the original position of the atom, and the other must be the position of the atom flipped across the plane formed by the three neighboring atoms (see Figures 4 and 5).



If the distance of an atom X from three stationary points A, B & C is known; then by finding the points of intersection of three spheres, the two positions for the atom can be calculated.

Figure 4. Three-point three-distance algorithm.



To find the alternative position for the ring atom X:

1/ Calculate the distances AX, BX, CX, DX.

2/ Calculate the positions for X' using the coordinates of A, B & C and the distances AX, BX and CX as described in diagram 4.

3/ Choose the position of X' which is different from the position of X.

4/ If the distance DX' is the same as DX then a new position for X has been found, which does not alter and ring bond lengths or ring angles.

5/ If DX' is different from DX then it is not possible to flip the atom.

Figure 5. Finding an alternative position for a ring atom using the three-point three-distance algorithm.

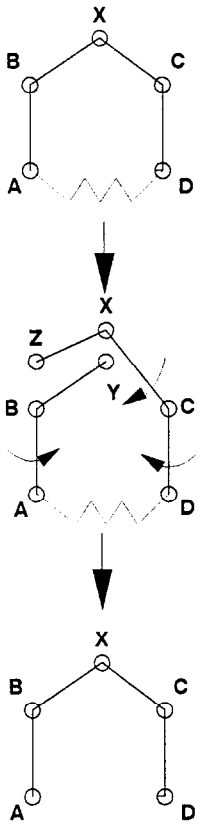


Figure 6. Flipping rings by sequential bond rotations.

The ring atom is rotated onto its new position by a series of three bond rotations; this ensures that atoms attached to the ring are repositioned correctly and that there is no alteration of chirality (see Figure 6).

Comparison with constraints and calculation of score

After decoding (analogous to the translation of RNA to protein), a score is generated by comparing the new molecule with the predefined constraints. Penalties are added if bad van der Waals contacts are found—the better the fit to the constraints, the lower the score. A score of zero indicates a perfect fit between the oriented conformation and the constraints.

The score generated for a conformation and orientation of a molecule is calculated by summing the weighted error terms for each of the constraints (some or all of these constraints may be used in a particular problem):

$$\text{Score} = W_1S_{\text{shp}} + W_2S_{\text{chg}} + W_3S_{\text{dist}} + W_4S_{\text{vdw}} + W_5S_{\text{vol}}$$

where S_{shp} is the error in the shape; S_{chg} is the error in the charge distribution; S_{dist} is the error in the distance constraints; S_{vdw} is a penalty term added for van der Waals contacts; S_{vol} is the error in the volume constraints; and W_{1-5} are the weights applied to each property.

The molecular shape and charge distribution terms (point charge distribution or electrostatic potential) are calculated for the fitted molecule at the points originally defined by the target molecule. The S_{shp} and S_{chg} values are the rms differences between each of the point properties of the target molecule and the oriented molecule:

$$S_{\text{shp}} = \left(\sum_{i=1}^{n} (S_{\text{fi}} - S_{\text{oi}})^2\right)^{1/2}$$

$$S_{\text{chg}} = \left(\sum_{i=1}^{n} (C_{\text{fi}} - C_{\text{oi}})^2\right)^{1/2}$$

where S_{fi} is the calculated value associated with the ith surface point for the fitted molecule; S_{oi} is the calculated value associated with the ith surface point for the original molecule; C_{fi} is the mapped charge or electrostatic potential associated with the ith surface point for the fitted molecule; C_{oi} is the mapped charge or electrostatic potential associated with the ith surface point for the original molecule; and n is the number of surface points.

The distance constraint term S_{dist} is calculated by summing the square of the differences between the desired constraint distances (from a rigid analogue, for example) and the corresponding interpoint distances in the oriented molecule:

$$S_{\text{dist}} = \left(\sum_{i=1}^{n} (Td_i - d_i)^2\right)^{1/2}$$

where n is the number of distance constraints; Td_i is the ith target distance; and d_i is the ith distance in the fitted molecule.

The van der Waals (vdw) contact term S_{vdw} is calculated by summing penalty terms for each bad atom–atom contact found in the conformation. For r_{ab} (the distance between two atoms) less than $vdw_1 + vdw_2$ (the van der Waals radii of the atoms[36]),

$$S_{\text{vdw}} = \sum_{i=1}^{n} \text{const}\,(vdw_1 + vdw_2) - r_{ab}$$

where n is the number of bad vdw contacts.

The constants used are 0.6 if the atoms are possible hydrogen bonds (i.e., one is a potential donor—e.g., oxygen of OH— and the other a potential acceptor—e.g., = O) and 0.8 for non-hydrogen bonded atoms.

After all the strings have been decoded and scored, a fitness value (F') for each string is calculated:

$$F' = 1/Score$$

The fitness values are normalized so that the sum of the fitness values (F) for a population is 1.0,

$$F = \frac{F'}{\sum_{i=1}^{ps} F_i'}$$

where ps is the population size.

Breeding the next population of bit strings

The genetic operators: selection, crossover, and mutation are employed to create a new population of strings (these are the "offspring"). Selection is the process of selecting the binary strings coding for the "best" molecules for breeding. Crossover is the mechanism by which information, from both

parents, is copied to the new offspring. Mutation is a random operator used to introduce diversity into the population of bit strings.

Figure 7 contains a flowchart showing how the crossover operator is used to generate a new population of bit strings.

Figure 8 shows how the mutation operator induces point mutations during the generation of the new strings.

Selection

Strings are selected from the mating pool of the old population using roulette wheel selection.[16] This is a method for selecting the parents most fit to breed the next generation. Each member of the population is allocated a slot in the roulette wheel; the size of the slot, expressed as the proportion of a full rotation of the wheel, represents the fitness of the individual. When the wheel is spun (by selecting a random number between 0 and 1), the probability that any individual is selected as a parent will depend on the slot size and therefore the fitness of the individual relative to the rest of the population.

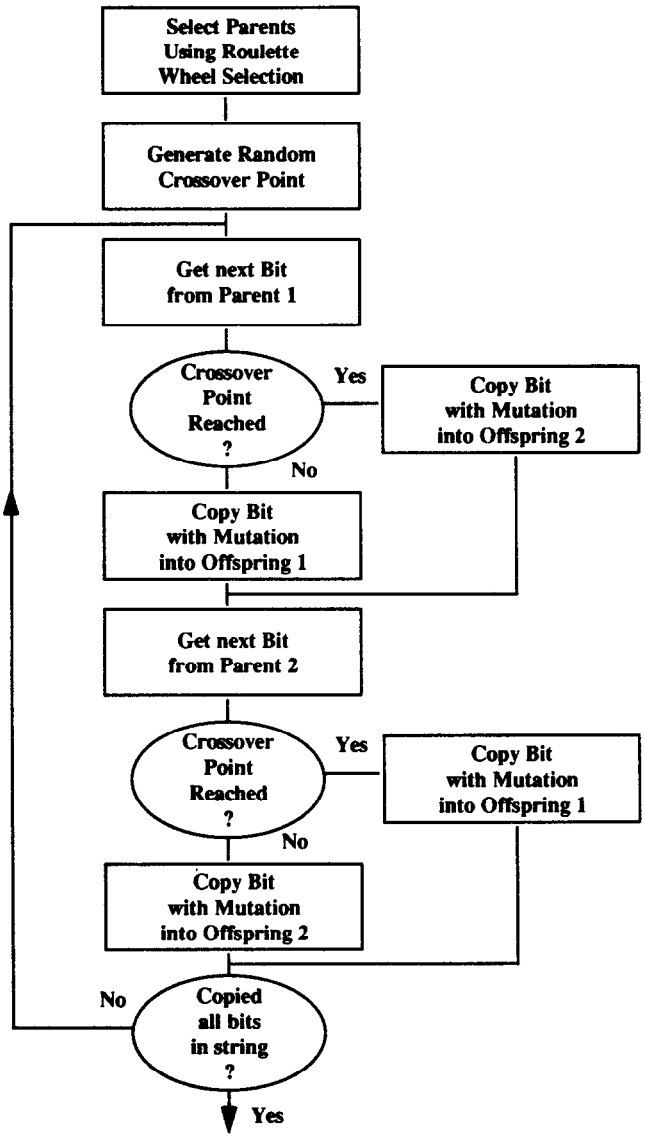


Figure 7. Flowchart showing the generation of new binary strings using the crossover operator.

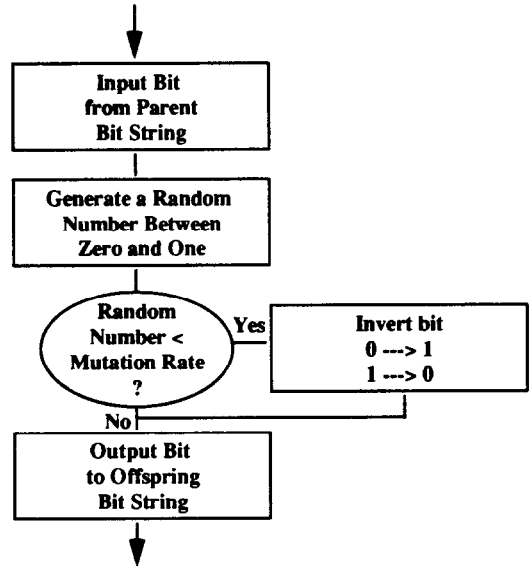


Figure 8. Flowchart showing how point mutations are introduced.

Crossover

The crossover operator is responsible for combining features of two individuals to produce offspring that may contain advantageous properties of each parent.

In crossover, the two bit strings are mated and produce two offspring, each containing parts of the bit strings from both parents. This is done by generating a random crossover point in the range of values from 1 to the length of the sequence. The bit string from the first parent is copied into the bit string for the first offspring until the crossover point is reached; at that point copying is switched to the second parent and proceeds until the end of the sequence is reached. The second offspring is produced in the same way, except that the second parent is copied until the crossover point is reached, after which the first parent is copied.

Mutation

The third genetic operator, mutation, is used to maintain the genetic diversity of the population, and decreases the risk of convergence to a local minima.

Mutations are introduced while copying the bit string from parents to offspring. As each bit is copied, a random number is generated in the range 0 to 1.0. If the random number is below a threshold value (a probability value for mutation) the bit is inverted. The threshold is set so that the probability of a mutation occurring during the creation of an offspring is less than a predetermined mutation rate. This rate is best determined by experiment on examples where the optimum result is known.

The three genetic operators, outlined above, are used again to generate a second population of bit strings. The bit strings in this second population (generation) will contain a predomi-

nance of the substrings from the best individuals of the first population (due to the roulette wheel selection).

The bit strings of the second population are decoded and scored as described above. Fitness values are calculated from the scores, and subsequent generations are produced. This process is repeated many times. As the optimization progresses, the substrings that improve the score of the individuals accumulate in the population. Eventually, the diversity in the population of bit strings decreases to a point where all the bit strings are nearly identical; at this point, the optimization (for this population) is complete, although the mutation operator will continue to introduce diversity at a low frequency.

The power of genetic algorithms applied to this problem arises from the way in which these high-quality substrings can be combined in many different ways, creating new strings that probe the conformational hyperspace in areas that are likely to contain optimum fits.

The optimization process of genetic algorithms has been analyzed in terms of *schemata*, a convenient description of the accumulation of small change leading to improvement in populations.[16]

RESULTS

Optimizing intramolecular distance constraints

There are many instances where distance constraints may be used to obtain realistic conformations for molecules of interest. The constraints may be derived, for example, from nuclear magnetic resonance data or from a pharmacophore hypotheses. In these instances, the objective is to fit a flexible series of molecules to a set of predefined distance constraints while maintaining reasonable low-energy conformations. A genetic algorithm similar to the implementation described here has previously been applied to fitting DNA sequences to nuclear magnetic resonance-derived constraints.[37] To test the applicability of the genetic algorithm to this type of problem, a number of test cases were generated.

Fitting a simple branched chain molecule

A simple branched chain molecule (Structure 1, see Figure 9) with ten rotatable bonds was generated using the SYBYL[38] molecular modeling program in a series of random low-energy conformations. The molecular geometry was optimized using the MAXIMIN[38] force field.

Three distance constraints were measured from one of the generated low-energy conformers (so that it was certain the molecule could achieve this conformation from a random starting point), and these are shown in Figure 9.

The experimental conditions used in the genetic algorithm are shown in Table 1. This problem has ten degrees of freedom, resulting in a binary string length of 80 bits. The angle resolution is 360/256 = 1.4°. The number of possible conformers is 256^{10} (1.2×10^{24}).

The results over a number of runs were consistent, and showed that all three distances could be rapidly optimized to within 0.05 Å (maximum observed error). A plot of the rate of convergence from a typical example (Figure 10) indicates that the algorithm has converged at about 30 generations.

Fifty generations on an IBM RS6000/320 using the present (unoptimized) code took about 60 seconds of cpu time (this varies slightly due to the random nature of the algorithm). Table 2 shows the goodness of fit of the distance constraints in two sample runs. The resulting conformers, due to the van der Waals overlap constraint, were of low steric energy. Clearly, the algorithm converges rapidly in this simple prob-

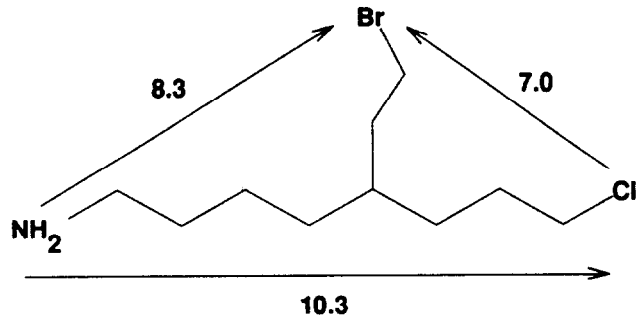


Figure 9. Simple branched chain molecule (Structure 1) with three intramolecular distance constraints in Angstroms.

Table 1. GA optimization conditions for distance constraints fitting problem

Population size	100
Number of generations	50
Mutation probability (per bit)	0.005
Weighting for van der Waals contacts	0.2
Weighting for distance constraints	1.0

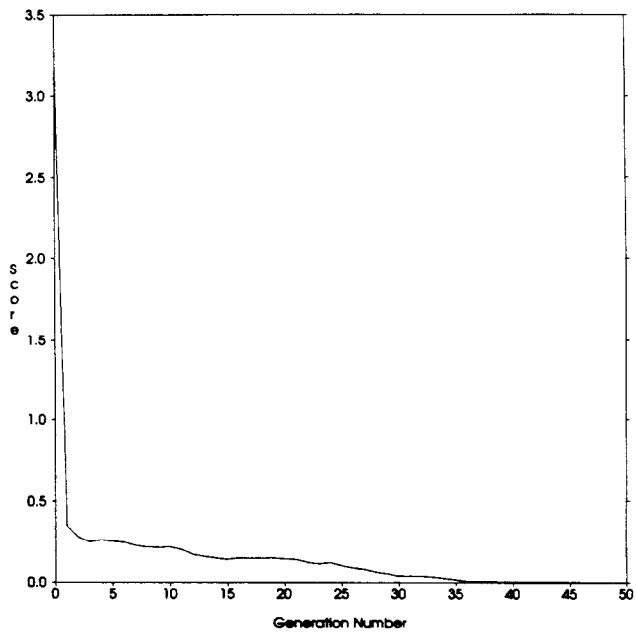


Figure 10. Convergence of the score, over 50 generations, during optimization of the conformation of the simple branched chain molecule (Structure 1) to the defined intramolecular distance constraints.

Table 2. Goodness of fit to distance constraints

Distance Constraint	Target Distance	Actual distance run 1	Actual distance run 2
N .. Cl	10.3	10.297	10.340
Br .. Cl	7.0	6.987	7.048
N .. Br	8.3	8.309	8.291

lem, and results in a very close fit to the constraints. A modified version of this GA may be useful in searching 3D databases for pharmacophoric matches.

Fitting molecules to a known pharmacophore

A common problem in computer-aided drug design is determining whether a molecule has an allowable conformation that matches a known or putative pharmacophore. The genetic algorithm was used to fit a series of N-methyl-D-aspartate antagonists to a putative NMDA pharmacophore (unpublished data) consisting of a small number of interatomic distances. These distances are defined for each of three test molecules (Structures 2, 3, and 4) in Figure 11.

In each case, the molecules were constructed as the neutral nonionized species to simplify computation (although, in solution at physiological pH these molecules would certainly be ionized). The molecules were built in an extended conformation using SYBYL,[38] and the geometry optimized using the MAXIMIN[38] force field. The conditions used in the genetic algorithm in the fitting of each molecule are shown in Table 3. The pharmocophore is composed of two distances, the first from the amine nitrogen to a phosphonate sp^2 oxygen, and the second from the carboxylic acid oxygen to the same phosphonate sp^2 oxygen. The nitrogen–carboxylic acid oxygen distance varies only slightly during bond rotation.

The molecules were generated from arbitrary starting points in rotational and conformational space. The two defined distances converged rapidly. Table 4 shows the closeness of fit of the functional groups in five example runs. It is clear that the algorithm achieves conformations that are reasonable (there are no van der Waals overlaps, and the structures have low strain energies), and have very close compliance with the desired constraints.

The scores calculated for each of the molecules in this problem are based solely on the distance constraints and atom–atom contacts, and do not depend on the orientation of the molecules. In addition, the molecules are not subject to any additional constraints, for example, a requirement to achieve maximum overlap between molecules.

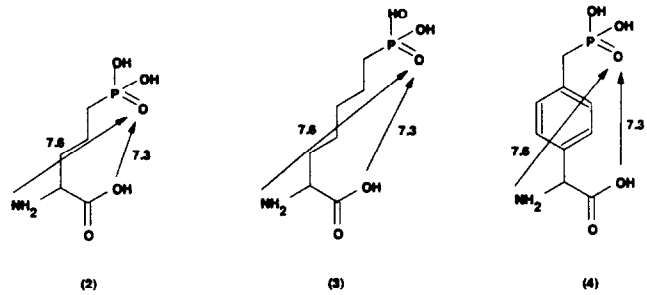


Figure 11. NMDA antagonists (Structures 2, 3, and 4) with distance constraints, in Angstroms, from a putative pharmacophore.

Table 3. GA optimization conditions used in fitting NMDA antagonists to distance constraints

Population size	100
Number of generations	50
Mutation probability (per bit)	0.005
Weighting for van der Waals contacts	0.2
Weighting for distance constraints	1.0

Elucidation of a pharmacophore

A more complex problem than fitting molecules to a known pharmacophore is in the determination of a putative pharmacophore when the only information available is a series of molecular structures, and possible key functional groups. In this case, the distances between the atoms of interest within any molecule are unknown. However, the distances between similar atoms or functional groups in different molecules

Table 4. Results of optimization against NMDA pharmacophore

Distance Constraint	Target	Run 1	Run 2	Run 3	Run 4	Run 5
OH in COOH to O in PO(OH)2, molecule 2	7.3	7.294	7.299	7.298	7.276	7.350
N in amine to O in PO(OH)2, molecule 2	7.6	7.573	7.587	7.599	7.451	7.592
OH in COOH to O in PO(OH)2, molecule 3	7.3	7.302	7.306	7.301	7.299	7.295
N in amine to O in PO(OH)2, molecule 3	7.6	7.607	7.595	7.599	7.599	7.593
OH in COOH to O in PO(OH)2, molecule 4	7.3	7.302	7.304	7.300	7.303	7.280
N in amine to O in PO(OH)2, molecule 4	7.6	7.590	7.588	7.600	7.616	7.588

(which may overlap in the receptor) can be set to zero between molecules, as these can be hypothesized to interact with common points in a receptor.

The genetic algorithm was adapted to allow the conformations and orientations of a set of molecules to be optimized simultaneously; this involved concatenating the bit strings coding for the rotation, translation, and conformation of each molecule and ignoring intermolecular van der Waals overlaps. The adapted genetic algorithm was applied to the problem of overlaying the three NMDA antagonists (Structures 2, 3, and 4) described above.

The conditions for this run are shown in the column labeled *Fit 1* in Table 5. The atoms to be fitted together in a common set of conformers for the three test molecules were the phosphonate sp^2 oxygens, the amine nitrogens, and the carboxylic acid sp^3 oxygens. Allowing for rotation, translation, and conformational freedom means that in this case there are 40 degrees of freedom (one molecule does not translate or rotate, to act as a reference point).

The overlayed molecules were monitored using computer graphics to assess the goodness of fit and the degree of diversity of the population at sampled generations (the diversity of molecular conformations and orientations decreases as the algorithm progresses and converges upon a solution). This problem ran for 10 days' cpu time on an IBM RS6000/320, at which time it was decided that the resulting overlay was reasonably good. This is shown in Color Plate 1. The interatomic distances between the designated atoms (which should be zero for perfect atomic overlap) are given in Table 6 in the rows labeled *Fit 1*.

Clearly this example, as defined, is having considerable difficulty in converging! The problem appears to be that there is no single target to optimize against. The stationary molecule, which acts as an origin, is in fact changing conformation, and hence interatom target distances are changing. The algorithm attempts to converge on a set of distances that are in continuous flux. Also, two of the molecules are translating and rotating during the fit, which increases the number of degrees of freedom.

To overcome this, a modification was introduced into the optimization. After each new set of molecule positions, rotations and conformations were generated, the specified functional group atoms were fitted by a least-squares fitting procedure[39,40] to minimize the interatom distances before calculation of the score. This dramatically simplified the problem by removing the degrees of freedom associated with the orientation of two of the molecules.

The conditions used for this run are shown in the column labeled *Fit 2* of Table 5, and the interatomic distances achieved between molecules are shown in the rows labeled *Fit 2* of Table 6. As before, zero distances describe a perfect fit.

An overlay of the resulting structures is shown in Color Plate 2. The elapsed time to achieve a better state of convergence than the previous 10-day run took 9 minutes on an IBM RS6000/320.

The resulting overlays, although displaying excellent fitting of the specified functional group atoms, did not, of course, maximize the overlap of the rest of the molecules. This would be desirable in a pharmacophore-fitting exercise.

In addition to the previous constraints, an atom-by-atom overlap integral calculation between atoms in different molecules was introduced.[31] This constraint is a measure of the degree of interpenetration between molecules, and is a maximum for best overlay.

The conditions used for this run are shown in the column labeled *Fit 3* of Table 5 and the interatomic distances are shown in the rows labeled *Fit 3* of Table 6.

The decrease in the total volume of the combined overlayed molecules, resulting from the additional overlap integral constraint, can be clearly seen in Table 7 and in Color Plate 3. The final combined volume[38] of the overlayed molecules was 34% larger than the volume (181 Å^3) of the largest individual molecule. The introduction of the volume criterion results in a more compact overlay of the molecules, which is probably more relevant in the elucidation of a pharmacophore.

Table 5. GA optimization conditions for the elucidation of a putative NMDA pharmacophore

Run time conditions	Fit 1	Fit 2	Fit 3
Population size	8000	500	2500
Number of generations	1299	93	500
Mutation probability (per bit)	0.003	0.003	0.002
Weighting for van der Waals contacts	1.0	1.0	1.0
Weighting for distance constraints	1.0	1.0	1.0
Weighting for volume overlap	N/A	N/A	5.0

Table 6. Interatomic distances for NMDA antagonists after molecular fitting

Atom	Run	2–3	2–4	3–4
OH in COOH	Fit 1	0.317	0.388	0.278
N in amine	Fit 1	0.274	0.311	0.341
O in PO(OH)2	Fit 1	0.276	0.599	0.382
OH in COOH	Fit 2	0.249	0.267	0.024
N in amine	Fit 2	0.183	0.209	0.060
O in PO(OH)2	Fit 2	0.088	0.128	0.046
OH in COOH	Fit 3	0.189	0.125	0.065
N in amine	Fit 3	0.105	0.165	0.070
O in PO(OH)2	Fit 3	0.283	0.022	0.285

Fitting a molecule to NOE distance constraints

We have used the genetic algorithm to search for conformations of a cyclic peptide, cyclo(-Arg-Gly-Asp-Ser-Lys)

Table 7. Total volume of combined molecules and elapsed time for run

Run	Total Volume (A^3)	Time (hours)
Fit 1	298.1	240
Fit 2	326.8	0.15
Fit 3	243.5	4.5

(Structure 5, see Figure 12), which fit a set of distance constraints derived from NMR experiments.[41]

To ensure that the whole of the allowed conformational space could be probed by the algorithm, the bond in the cyclic peptide between the Glycine C_α and the Glycine carbonyl carbon was removed. During the run, the conformation of the peptide is constrained by the distance constraints derived from NMR experiments (Table 8), and by three additional distance constraints that are used to ensure that the broken bond can be rejoined. The three additional distance constraints attempt to preserve not only the distance across the broken bond but also the bond angles at either end of the broken bond. To accentuate the importance of peptide recyclization, the distance constraints associated with the broken bond were weighted relative to the other distance constraints; this was achieved by simply adding these constraints more than once to the overall score.

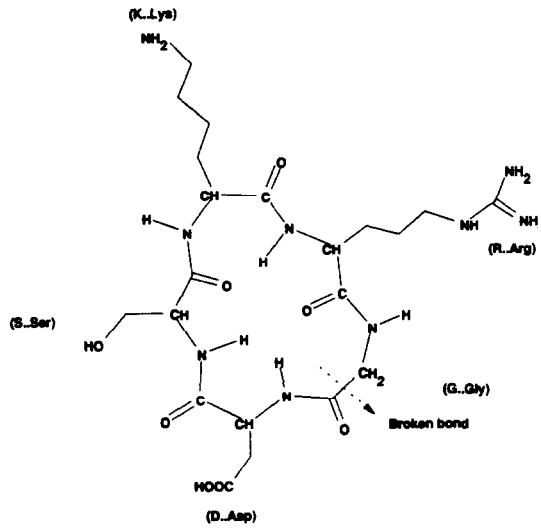


Figure 12. Cyclo [-Arg-Gly-Asp-Ser-Lys-] (Structure 5)

The conditions for the optimization are shown in Table 9. A large population of strings was used due to the complexity of the problem. Even though amide bonds were not allowed to rotate, the peptide contained 25 rotatable bonds, 9 of which occurred on the peptide backbone. The requirement for ring closure was therefore dependent on 9 rotatable bonds, and small changes in any one of these could prevent ring closure. This problem is particularly challenging for a genetic algorithm, because the probability is low that two strings coding for reasonable conformations of the cyclic peptide can be combined to provide a better conformation. The results of two test runs are shown in Table 8.

The algorithm produces conformations that are cyclic and that satisfy the distance constraints fairly well without severe van der Waals interactions. The hydrogen-bonding angles are, however, distorted (there are no angle constraints for hydrogen-bonds in these runs). The deviation of a trial structure from the distance constraints is recorded by Williamson et al.[41] as the square root of the sum of squares of the deviations from the desired distance constraints in angstroms. This was 2.7 Å for their best starting structure.

Using the genetic algorithm, the structures represented in

Table 9. GA optimization conditions for the conformational search on a cyclic peptide

Condition	Runs 1 and 2
No of Generations	1000
Population size	8000
Mutation Probability	0.003
vdw Contact Weight	0.2
Distance Weight	1.0

Table 8. Cyclic peptide target distance constraints and results

Distance constraint	Relative weight	Target distance	Actual distance run 1	Actual distance run 2	Energy minimized
Cyclization bond length	4	1.51	1.530	1.611	1.51
Cyclization distance 1	2	2.52	2.602	2.606	2.51
Cyclization distance 2	2	2.46	2.461	2.607	2.47
R_N–K_N	1	2.50	2.887	2.867	2.557
G_N–R_a	1	2.61	2.511	2.595	2.621
S_N–D_N	1	2.70	2.734	2.815	2.683
S_N–D_a	1	2.99	3.622	3.614	3.216
K_N–S_N	1	2.93	3.510	3.327	2.990
K_N–S_a	1	2.83	3.634	3.679	3.043
S_N–R_{CO}	1	2.50	3.086	3.079	2.549
R_N–S_{CO}	1	2.50	2.482	2.489	2.501

runs 1 and 2 had violations from the distance constraints of 1.4 Å and 1.3 Å.

These conformations are suitable for constrained minimization. The conformer resulting from run 1 on structure 5 was subject to constrained minimization using the SYBYL force field MAXIMIN[38] after remaking the omitted bond. The constraints were represented by Hooke's law, with constants of 200 Kcal/mol/Å^2. Also, by the addition of extra constraints, each hydrogen of the listed hydrogen bonds was constrained to lie about 2.1 Å from the H-acceptor to produce sensible donor-H-acceptor bond angles. The resulting structure is shown in Color Plate 4. The distance constraints and the calculated distances from the minimized structure are shown in Table 8. The calculated steric energy for the constrained structure was 81.7 Kcal/mol, of which the constraint energy was 12.1 Kcal/mol.

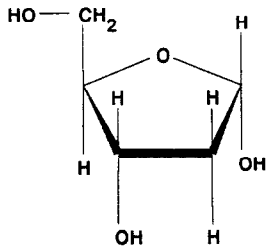
Fitting flexible molecules to shape and electronic constraints

In cases where there is no evidence for functional group fitting, we are faced with the more complex task of generating overlays based on combinations of all the possible functional group overlays (e.g., overlaying similar functional groups) or overlaying molecular properties. However, this may be a preferable method of comparison, as the receptor perceives the molecular structure and properties not of individual atomic or functional groups but as the result of their combination in the complete structure. Also, there is no bias introduced due to preconceptions about functional group correspondence between ligands.

The objective here is to find a region of conformational, translational, and rotational space close to the global minimum that corresponds to the best overlay of the computed properties near the molecular surface. To assess the ability of genetic algorithms to find optimal orientations and conformations of flexible molecules, we have performed a set of self-similarity experiments; only when fitting a molecule onto itself is it possible to compare the fit with a known global minimum. This also allows some experimentation and comparison with the GA parameters.

Fitting 2-deoxyribose to shape and charge parameters

To investigate the application of the genetic algorithm to the self-fitting problem and to investigate the run time conditions, the mono-saccharide 2-deoxyribose in the D-configuration (Structure 6) was fitted onto itself using varying constraints, population sizes, and mutation rates.



2-deoxyribose (6)

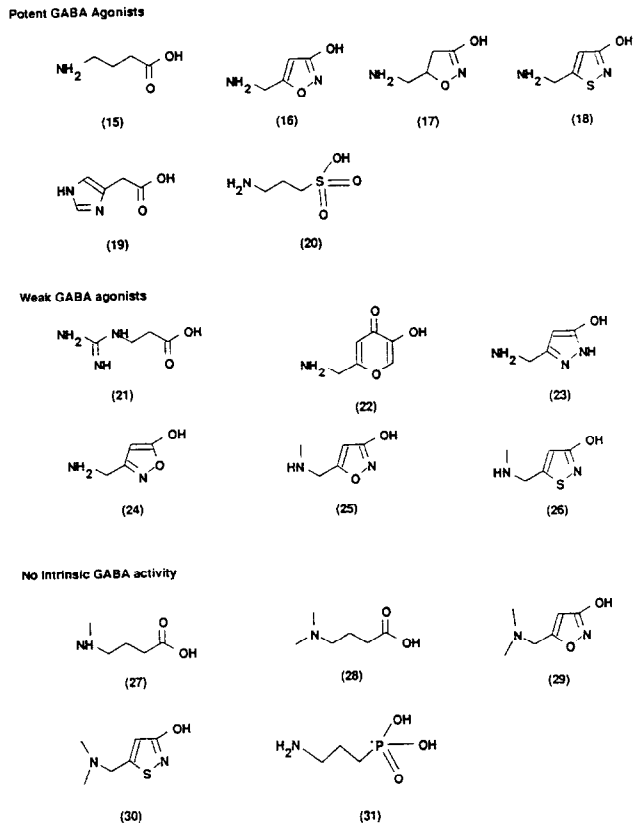


Figure 13. GABA analogues classified into three classes.

Table 10. GA optimization conditions with variable population size

Condition	Runs 1–5	Runs 6–10	Runs 11–15	Runs 16–20	Runs 21–25
No of generations	300	300	300	300	300
Population size	1000	500	250	125	50
Mutation probability	0.005	0.005	0.005	0.005	0.005
vdw contact weight	0.2	0.2	0.2	0.2	0.2
Shape weight	1.0	1.0	1.0	1.0	1.0
Electrostatic weight	1.0	1.0	1.0	1.0	1.0

The 2-deoxyribose molecule has four degrees of freedom associated with the four rotatable bonds, three translational degrees of freedom, and three rotational degrees of freedom. The ten degrees of freedom in conformational and orientational hyperspace can be represented by a binary string 80 bits long; the total number of possible conformations and orientations that can be searched with this representation is 2^{80}, or 1.21×10^{24}. To increase throughput, ring flexibility (pseudorotation of the 5-membered ring) was disabled for these runs. The 2-deoxyribose molecule was centered on the origin and surrounded by a 6-Å sphere of 129 equally spaced points.[18] A shape (shape-2) and electrostatic potential (from the PEOE charges, using nuclear screening constants[28,30])

Table 11. GA optimization conditions with variable mutation rate

Condition	Runs 26–30	Runs 31–35	Runs 36–40	Runs 6–10	Runs 41–45	Runs 46–50	Runs 51–55
No of generations	300	300	300	300	300	300	300
Population size	500	500	500	500	500	500	500
Mutation probability	0.05	0.02	0.01	0.005	0.002	0.001	0.000
vdw contact weight	0.2	0.2	0.2	0.2	0.2	0.2	0.2
Shape weight	1.0	1.0	1.0	1.0	1.0	1.0	1.0
Electrostatic weight	1.0	1.0	1.0	1.0	1.0	1.0	1.0

was calculated for each of the points on the sphere, and this was used as a set of target properties for the self-fitting problem.

The first set of experiments investigated the effect of varying population size on the quality of the fit to the constraints. The conditions used in a series of runs are shown in Table 10. Color Plate 5 shows the mean score of the best four individuals in the population averaged over five runs for each set of conditions.

The convergence of the algorithm improved as the size of the population increased; however, no further improvement was noted as the population grew beyond 500 individuals. The second set of experiments investigated the effect of varying the mutation probability. The conditions used in a series of runs are shown in Table 11. Color Plate 6 shows the mean score of the best four individuals in the population averaged over five runs for each set of mutation probabilities.

The runs using the highest mutation probabilities failed to converge after 300 generations; the high rate of point mutations precludes the accumulation of high-quality substrings required for convergence. The runs with the lowest mutation probabilities converged but failed to find conformations and orientations close to the global minimum; the low rate of point mutations causes a loss of genetic diversity. The third experiment examined the effects of using the alternative shape and charge properties. The GA conditions are shown in Table 12. In this experiment, the scores are not comparable, since they are calculated using different sets of constraints. However, the rms error in the positions of the atoms in the fitted molecule can be calculated and compared. Table 13 describes the use of the molecular shape descriptors, shape-1 or shape-2, combined with charge mapping or electrostatic potential (EP, calculated from PEOE atom centered charges with nuclear screening constants[28,30]), with the results averaged over five runs.

The two shape descriptors gave similar results. However, the vector representation gives a biased description of shape in cases where the surface of the molecule is flat or oblate. In this case, most vectors cross the molecule surface near the center. Charge-mapping and electrostatic-mapping appear to give similar results, although charge-mapping is significantly faster than calculating electrostatic potential. Color Plate 7 shows a series of snapshots of 2-deoxyribose during self-fitting over the course of 170 generations using shape-2 and electrostatic potential constraints. The run was performed using a population size of 500 individuals and a mutation probability of 0.005. The time taken for 300 generations was 2.5 hours on an IBM RS6000/320. The initial molecular positions are shown at generation zero. As the run progressed, the conformation and fit of the best example from each generation was overlaid on the template molecule. Following through the series of pictures indicates a gradual convergence of the run until both the moving and static molecules are overlaid.

Table 12. GA optimization conditions for shape and charge calculation comparisons

Condition	
No of generations	300
Population size	500
Mutation probability	0.05
vdw contact weight	0.2
Shape weight	1.0
Electrostatic/charge Weight	1.0

Table 13. Results of different shape and electronic descriptions (rms errors in Angstroms)

	1	2	3	4	5
EP + shape-2	0.04	0.00	0.01	0.02	0.06
Charge map + shape-1	0.03	0.00	0.01	0.01	0.00
EP + shape-1	0.20	0.03	0.07	0.13	0.02
Charge map + shape-2	0.04	0.20	0.02	0.01	0.00

Self-fitting of Trimethoprim to shape and charge constraints

The drug molecule trimethoprim (TMP) (Structure 7) poses a more complex problem for the genetic algorithm.

The molecule contains two aromatic rings separated by a pair of rotatable bonds, and has an additional eight rotatable bonds on the periphery of the molecule. The conformational and orientational hyperspace required a binary string of 128 bits to represent the 16 degrees of freedom. The total number of possible conformations and orientations that could be searched with this representation is 2^{128}, or 3.4×10^{38}.

A conformation and geometry of TMP from an X-ray crystallographic structure[11] was centered at the origin and

surrounded by a 12-Å sphere of 331 equally spaced points. A shape (shape-2) and electrostatic potential property were calculated for each of the points on the sphere, and these were used as the set of target properties for self-fitting from a random starting point. The conditions for the run are shown in Table 14. Starting from a random conformation and orientation of TMP, the genetic algorithm found an overlay of the molecule that closely matched the target TMP. The rms distance between the target atom positions and fitted atom positions was 0.087 Å. The run took 15 hours on an IBM RS6000/320 using unoptimized code.

Color Plate 8 shows a series of snapshots of the best molecule of trimethoprim from selected generations, during self-fitting, over the course of the 661 generations. In this example, similar to the 2-deoxyribose case, the run appears to show gradual convergence to a good fit, correctly orientating and overlaying the moving molecule.

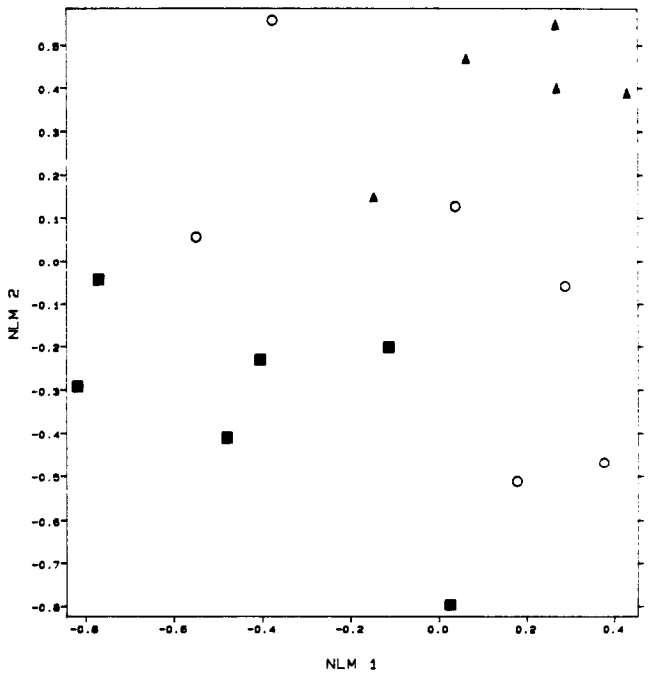
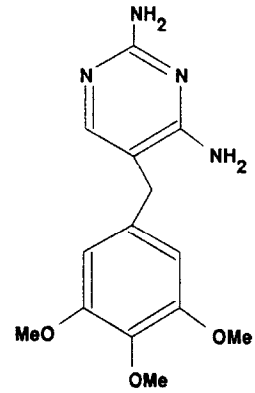


Figure 14. Nonlinear map of GABA analogues fitted to THIP using GA optimization: square, potent agonist; circle, weak agonist; triangle, inactive.

Table 14. GA optimization conditions for trimethoprim self-similarity run

Condition	Run 1
No of generations	661
Population size	1000
Mutation probability	0.005
vdw contact weight	0.2
Shape weight	1.0
Electrostatic weight	1.0



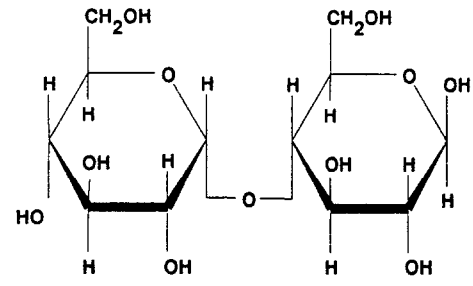
Trimethoprim (7)

Self-fitting of Maltose: Conformational variability in rings

Molecules with flexible ring systems pose considerable problems to current similarity methods. For example, some molecules have rings in which motion of a side chain induces cooperative motion in another side chain (e.g., sugars). This is more complex in multiple ring systems.

Molecular dynamics is one method of analyzing molecular motions. In comparing molecules, similar distances between key atoms or functional groups may be sought in conformations, which result in sensible overlays. Alternatively, the dynamics may be constrained to achieve a desired set of constraints. The process is somewhat random and does not often reach an optimum fit. Heating and cooling the system systematically is a great improvement (annealing), and in some ways is analogous to GAs in which each cycle results in a new parent that is subsequently cooled to a new series of conformers.

We have applied a simple ring flapping algorithm, useful only in cases where there are free corners, to fitting maltose (Structure 8) (a disaccharide) that has 12 rotatable bonds and 12 free corners, to itself from a random starting conformation, orientation, and position. Including rotation, translation, ring flip, and bond rotation, there are 30 degrees of freedom (bit string length = 156, made up from: 8 × 12 for the rotatable bonds, 1 × 12 for the ring flaps and 8 × 6 for the rotation and translation). The GA conditions applied to this problem are given in Table 15.



Maltose (8)

Color Plate 9 shows the self-fitting of maltose in chronological order from the random starting set to the best fit at generation 50. This run took 72 hours on an IBM RS6000/

Table 15. GA optimization conditions for self-fitting Maltose

Condition	
No of generations	60
Population size	8000
Mutation probability	0.005
vdw bump weight	0.2
Shape weight	1.0
Electrostatic weight	1.0

320, although it is apparent that the correct orientation and conformational preferences were detected early on in the run. The large population size was selected due to the complexity of the problem.

Despite the large population size, the quality of the fit obtained in this first run was not found to be generally reproducible. Subsequent runs had a tendency to optimize to alternative fits, not the best overlay. This therefore represents a limit on this version of a genetic algorithm and constraints as presently implemented.

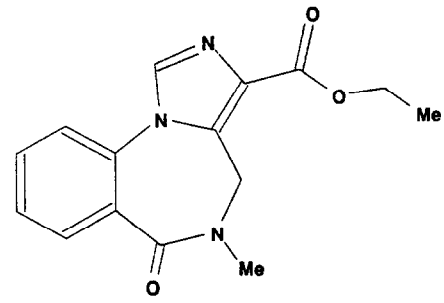
Fitting benzodiazepine receptor ligands to a β-carboline

In the previous examples of self fitting, it appears that a number of test molecules may be successfully fitted back onto themselves from random starting conformations and positions. We have, therefore, investigated the usefulness of these methods in the more relevant exercise of fitting very dissimilar molecules together. To do this, we need examples where there is strong evidence from structure–activity relationships (SAR) for a particular overlay pattern. One such example is in the overlay of dissimilar ligands at the benzodiazepine receptor, which is part of the GABAergic (GABA is γ-aminobutyric acid) receptor complex controlling chloride ion flux in neural tissue.[42] Compounds that enhance the action of GABA are agonists (anxiolytics), while those which lessen GABA binding are classified into antagonists (no biological effect) and inverse-agonists (convulsants). Codding et al.[43] proposed an overlay of antagonists (structures from X-ray crystallography) based on SAR and functional group similarities.

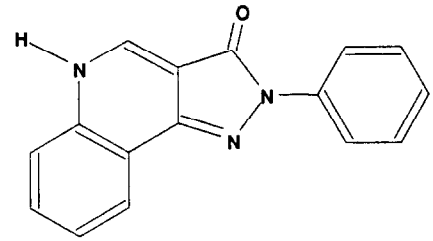
This overlay is shown in Color Plate 10 for Ro15-1788 (Structure 9), CGS-8216 (Structure 10), and methyl-β-carboline-3-carboxylate (Structure 11). Also shown in this Color Plate are the major receptor interactions identified from the molecular comparisons.

The conformation and geometry of Structure 11 used in the GA fitting process is from the available X-ray crystallographic structure.[44] The other two molecules were assigned random starting conformations, orientations, and translations from the original crystal structures.[43,45] The conditions used in this run are shown in Table 16.

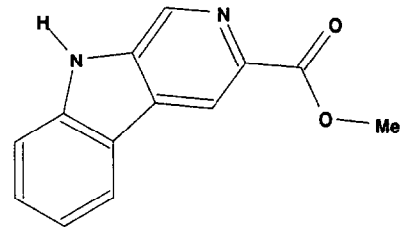
The target function was generated from the static molecule (Structure 11) and was composed of charge-mapping, shape (shape-2), and van der Waals contact constraints. The charge and shape were mapped, as described before, onto a sphere of radius 12 Å containing 331 equidistant points.



Ro15-1788 (9)



CGS-8216 (10)



methyl-beta-carboline-3-carboxylate (11)

After 100 generations, the fit of CGS8216 (Structure 10) onto methyl-β-carboline-3-carboxylate (Structure 11) (Color Plate 11) was very similar to that obtained by Codding et al. To further explore the goodness of fit, the fitted molecule (Structure 10) in the final position and conformation was used to generate constraints as before, and the previously static molecule (Structure 11) was fitted to these constraints starting from a random orientation and conformation. This is in effect running the problem in reverse. Structure 11 returned nearly to its original position (rms deviation of the final atom positions is 0.043 Å), which tends to justify the overlay obtained.

However, the fit of Ro15-1788 (Structure 9) onto (Structure 11) (Color Plate 12) was rotated by almost 180° relative to the predicted fit, which is not consistent with the SAR. After a number of repeat runs using the same constraints and alternatively with the constraints defined by charge-mapping and electrostatic potentials derived from *ab initio* 6-31G* potential-derived charges, the same result is obtained. The optimization process is probably finding a minimum of the target function; however, the constraints definition is not

Table 16. GA optimization conditions for the fitting of Structures (9) and (10) onto Structure (11)

Condition	Ro-1788 (9) and CGS-8216 (10)
No of generations	100
Population size	500
Mutation probability	0.005
vdw contact weight	0.2
Shape weight	1.0
Charge map weight	5.0

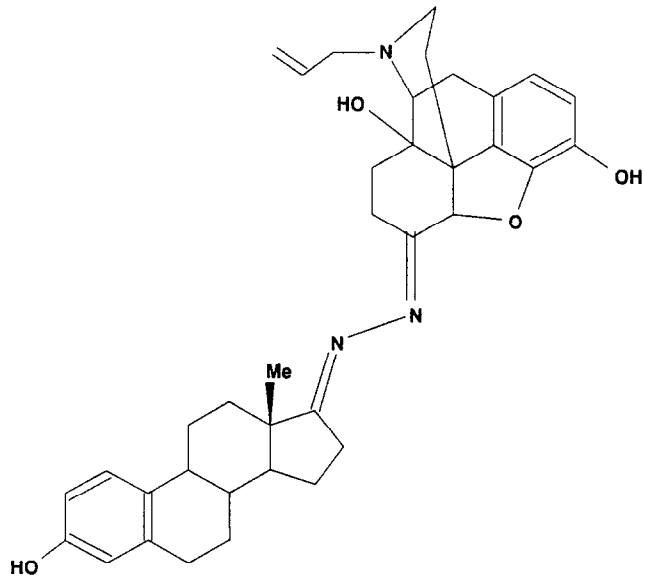
sufficient to completely define the problem. Since only a comparison between molecules is being attempted in the absence of the receptor, there is no knowledge in the model of the receptor topology. Therefore, extra volume of the molecule being fitted over and above the volume of the template molecule may be tolerated in the receptor, and indeed may contribute to binding, but will be classed in the simple constraints used here as a bad fit. This is seen in the comparison of Structures 9 and 11. In the predicted fit, there is extra volume caused by the presence of a methyl and phenyl group.

Attempts to encourage the predicted fit by increasing the weighting on electrostatics did not improve the situation. However, the allowed volume was increased by overlaying Structures 9 and 11 in the predicted fit before surface shape calculation. Thus, the extra volume was not preventing the predicted overlay. The correct orientation was then obtained. This appears to define a limitation on the blind fitting method.

Fitting leu-enkephalin to hybrid morphine

Another problem in which very different molecular structures probably interact at the same receptor site is seen in enkephalin and morphine analogues. Morphine analogues are highly constrained and are useful templates on which to fit very flexible enkephalin analogues.

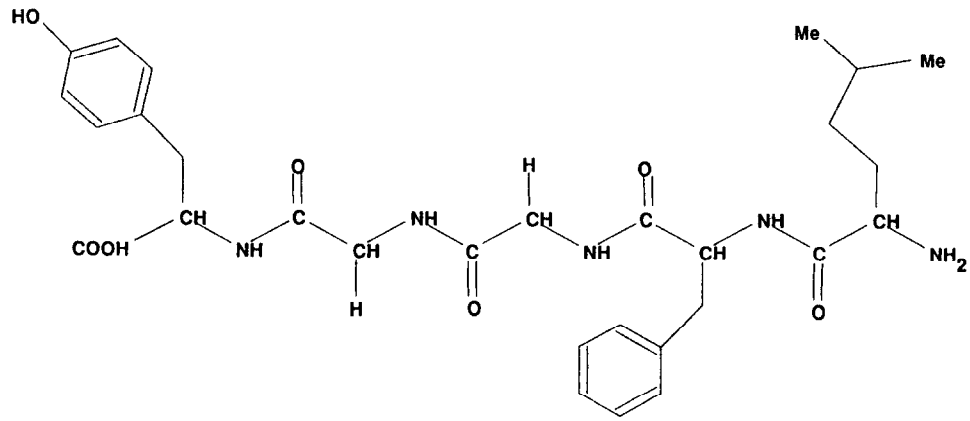
The semirigid hybrid morphine molecule EH-NAL (Structure 12), a mixed azine between estrone and naloxone,[46] was built using SYBYL and optimized in the MAXIMIN[38] force field to a low-energy geometry. Leu-enkephalin (Structure 13) was built as an extended structure, and the geometry similarly optimized. The carboxylic acid and amine groups in the structures were converted to their ionized states. This structure was surrounded by a 15-Å sphere of 552 equally spaced points.[18,19] The shape (shape-2) and charge properties (electrostatic potential using PEOE charges[28,30]) associated with each of the points were calculated and used as a set of constraints for fitting the flexible leu-enkephalin molecule (Structure 13).



EH-NAL (12)

Excluding the amide bonds leu-enkephalin has 21 rotatable bonds, and therefore the problem has 27 degrees of freedom in total. The conditions for the GA are given in Table 17.

Color Plate 13 shows leu-enkephalin (Structure 13) overlaid on the hybrid morphine (Structure 12). The overall shape of the two molecules is remarkably similar, and there is a good overlap between similar functional groups in the two molecules, particularly the protonated nitrogens, the aromatic regions, and the backbone atoms. There is, however, an overlay of a phenol and carboxylate region that does not appear optimal. There is presently no absolute way of deter-



Leu-enkephalin (13)

mining if this is the most similar fit or the receptor-bound conformation; however, the fit obtained appears to be remarkably similar to that obtained by Kolbe et al.[46] using molecular dynamics with simulated annealing, and could serve as a starting structure for optimization of molecular similarity.

To optimize the fit described by Kolbe et al., the overlay of four key functional groups was performed using the same GA conditions (Table 17). Starting from a random orientation and conformation, the target distance between each functional group pair was set to zero, while simultaneously optimizing the volume overlap. This gave the results shown in Color Plate 14. This is a remarkably compact overlay showing excellent correspondence between the identified functional groups.

Fitting GABA analogues to conformationally restricted THIP

There are examples of molecular fitting and the subsequent calculation of molecular properties that enable classification of biological activity. We wished to investigate the use of the GA in fitting a series of dissimilar molecules to a template consisting of a conformationally restricted molecule and to determine if the fit obtained was good enough to allow classification of their biological activities.

GABA (γ-aminobutyric acid) is an inhibitory neuro-transmitter concerned with the control of neuronal activity. A series of analogues of GABA (Figure 13) were classified into agonist, weak agonist, and inactives by Krogsgraad-Larsen, based on electrophysiological and binding studies.[42] Analysis of seven computer-generated molecular properties (surface area, dipole moment, and principal ellipsoid axes)[47] for each of the 17 GABA analogues showed that a plot of the first two principal components, or a two-dimensional nonlinear map clustered the molecules correctly into each class.

In this study, the 17 GABA analogues were built using the SYBYL[38] molecular modeling system in random conformations and rotations. These were conformationally and spatially restricted by fitting to THIP (4,5,6,7-tetrahydroisoxazolo[5,4-c]pyridin-3-ol) (Structure 14) using a genetic algorithm to match molecular properties generated from the test molecules and THIP on a molecular surface.

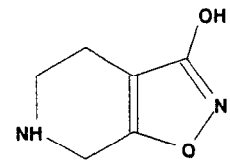
Molecular properties for the target function were generated from THIP, which was centered at the origin and surrounded by a 12.0-Å sphere of 331 equally spaced points.[18,19] The molecular shape (shape-1), electrostatic potential (calculated from atom-centered charges (PEOE) and incorporating nuclear screening constants[28,30]), and atom centered charge densities were calculated and mapped on to the appropriate points as previously described. The GA parameters used here are shown in Table 18. The resulting fitted structures are shown in Color Plate 15. The conformations and spatial orientations of the test structures resulting from the GA fit to THIP (Structure 14) were used to calculate key molecular properties (surface area, dipole, and principal ellipsoid axes) for input to pattern recognition. The dipole moment was calculated from CNDO/2 calculations[48] on the neutral species for the fitted molecules (CNDO/2 was used to maintain compatibility with the original study). It should of course be borne in mind that in this example the zwitterions would certainly be present at neutral pH. (However, it is a much more complex and expensive task to calculate the zwitterionic species with solvent and counter ions.)

Table 17. GA optimization conditions for the fit of leu-enkephalin (Structure 13) on EH-NAL (12)

Condition	
No of generations	300
Population size	1000
Mutation probability	0.002
vdw contact weight	0.2
Shape weight	1.0
Electrostatic weight	1.0

Table 18. GA optimization conditions used in fitting GABA analogues to THIP

Condition	
No of generations	300
Population size	1000
Mutation probability	0.001
vdw contact weight	10.0
Shape weight	1.0
Charge weight	1.0
Electrostatic weight	1.0



THIP (14)

Figure 14 shows a nonlinear map generated from the 7 molecular properties (surface area, dipole x, y, z, and principal ellipsoid axes x, y, z) generated for each structure.

The structures are classified in the map as agonist, weak agonist, and inactive. The degree of classification is excellent, and compares favorably to the map previously generated from manually overlaid structures[47] using rms fitting between similar functional group atoms.

Again, it is not known if this is the best fit for this molecular series; however, it is good enough to allow classification of the biological activity of the training set of molecules examined.

CONCLUSIONS

A binary genetic algorithm has been devised and applied to the problems of molecular similarity, pharmacophore elucidation, and conformational analysis. The genetic algorithm uses the operators crossover, mutation, and selection to optimize molecules within constraints. These constraints

have been devised to describe molecules in terms of distances, surface properties, and whole molecule properties.

In general terms, the optimization performs very well indeed on complex problems having many degrees of freedom, as shown by the self-fitting experiments. The main problem, as with most optimization methods, is one of early optimization to a false minimum, but this is to some extent improved by using larger populations. Fitness scaling in which the constraints are relaxed initially to generate more diverse populations, and parallel generation of populations to enable more diverse groups to develop, are areas for future investigation. This is particularly important, as GA methods are ideal for coarse grain parallelization, which would speed up the algorithm enormously on appropriate hardware.

The constraints are necessarily simple due to the iterative nature of the algorithm, and this imposes limitations on the quality of the results. Improvements in describing the molecules would be obtained by using fast semi-empirical quantum mechanical methods and by describing ring flexibility better, and these are areas for future investigations.

Of particular significance is the speed and efficiency of the algorithm in overlaying functional groups to elucidate a pharmacophore. Also, the ability to fit molecules to distance constraints quickly and efficiently may be beneficial in three-dimensional database searching.

One major problem in comparing molecules is that the receptor will usually have regions open to solvent or specific interactions with the ligand that cannot be incorporated into a simple similarity algorithm. This is seen in the benzodiazepine example. However it may be possible to incorporate solvent effects and specific ligand-receptor interactions; this is under investigation and will be reported elsewhere.

ACKNOWLEDGEMENTS

We would like to thank Dr. Brian Hudson for alerting us to the potential of genetic algorithms, Mrs. Valerie Rose, and Dr. John Wood for useful discussions.

REFERENCES

1 Cheney, B.V. Structure-Activity Relationships for Drugs Binding to the Agonist and Antagonist States of the Primary Morphine Receptor. *J. Med. Chem.* 1988, **31**, 521–531

2 Peet, N.P., Lentz, N.L., Meng, E.C., Dudley, M.W., Ogden, A.M.L., Demeter, D.A., Weintraub, H.J.R., and Bey, P. A Novel Synthesis of Xanthenes: Support for a New Binding Mode for Xanthenes with Respect to Adenosine at the Adenosine Receptor. *J. Med. Chem.* 1990, **33**, 3127–3130

3 Boudon, A., Szymoniak, J., and Chretien, J.R. A Molecular Electrostatic Potential Study of Phenothiazine Dopaminergic Antagonists. *Eur. J. Med. Chem.* 1988, **23**, 365–371

4 Waters, J.A., Spivak, C.E., Hermsmeier, M., Yadav, J.S., and Liang, R.F. Synthesis, Pharmacology, and Molecular Modeling Studies of Semirigid, Nicotinic Agonists. *J. Med. Chem.* 1988, **31**, 545–554

5 Arvidsson, L., Karlen, A., Norinder, U., Kenne, L., Sundell, S., and Hacksell, U.J. Structural Features of Importance for 5-Hydroxytryptaminergic Activity. Conformational Preferences and Electrostatic Potentials of 8-Hydroxy-2-(di-n-propylamino)tetralin (8-hydroxy-DPAT) and Some Related Agents. *J. Med. Chem.* 1988, **31**, 212–221

6 Tayar, N., Carrupt, P., van de Waterbeemd, H., and Testa, J. Modeling of β-Adrenoceptors Based on Molecular Electrostatic Potential Studies of Agonists and Antagonists. *J. Med. Chem.* 1988, **31**, 2072–2081

7 Croizet, F., Langlois, M.H., Dubost, J.P., Braquet, P., Audrey, E., Dallet, P.H., and Colleter, J.C. Lipophilicity Force field Profile: An Expressive Visualisation of the Lipophilicity Molecular Potential Gradient. *J. Mol. Graphics* 1990, **8**, 153–155

8 Marshall, G.R., Barry, C., Bosshard, H.E., Dammkoehler, R.A., and Dunn, D.A. The Conformational Parameter in Drug Design: The Active Analogue Approach. in *Computer-Assisted Drug Design.* (E.C. Olson and R.E. Christoffersen, Eds.) ACS symposium series 112, American Chemical Society, 1979

9 Carbo, R., Leyda, L., and Arnau, M., How similar is a molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quant. Chem.* 1980, **17**, 1185

10 Mathews, D.A., Bolin, J.T., Burridge, J.M., Filman, D.J., Volz, K.W., Kaufman, B.T., Beddell, C.R., Champness, J.N.C., Stammers, D.K., and Kraut, J. Refined Crystal Structures of *Escherichia Coli* and Chicken Liver Dihydrofolate Reductase Containing Bound Trimethoprim. *J. Biol. Chem.* 1985, **260** (1), 381–391

11 Champness, J.N.C., Kuyper, L.F., and Beddell, C.R. Interaction Between Dihydrofolate Reductase and Certain Inhibitors. *Molec. Graphics Drug Design.* Elsevier Science Publishers, B.V. (Biomedical Division). Ed. Burgen, A.S.V., Roberts, G.D.K., Tute, M.S., 1986

12 Kearsley, S.K. An Algorithm for the Simultaneous Superposition of a Structural Series. *J. Comp. Chem.* 1990, **11** (10), 1187–1192

13 Cramer, R.D., Patterson, D.E., and Bunce, J.D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* 1988, **110** (18), 5959–67

14 Dean, P.M., Callow, P., and Chau, P.L. Blind Searching for Regions of Strong Structural Match on the Surfaces of Two Dissimilar Molecules. *J. Mol. Graphics* 1988, **6** (1), 28–34

15 Burt, C., Richards, W.G., and Huxley, P. The Application of Molecular Similarity Calculations. *J. Comp. Chem.* 1990, **11** (10), 1139–1146

16 Goldberg, D.E. *Genetic Algorithms in Search, Optimisation and Machine Learning.* Addison-Wesley, Reading, 1989

17 Darwin, C. *The Origin of Species.* (Mathew C. Harrison, Ed.) published by Dent Gordon, 1973

18 Connolly, M. *Molecular Surface Program. QCPE 429.* Quantum Chemical Program Exchange, Dept.

of Chemistry, University of Indiana, Bloomington, IN, USA
19 Richards, F.M. Areas, Volumes, Packing and Protein Structure. *Ann. Rev. Biophys. Bioeng.* 1977, **6**, 151–176
20 Singh, U.C. and Kollman, P. *Gaussian 80-UCSF. QCPE 446.* Quantum Chemical Programme Exchange, University of Indiana, Bloomington, Indiana, USA
21 Amos, R.D. and Rice, J.E. *CADPAC: The Cambridge Analytic Derivatives Package, issue 4.0.* Cambridge, 1987
22 Stewart, J.J.P. *MOPAC Version 5.0.* Quantum Chemical Program Exchange, Department of Chemistry, University of Indiana, Bloomington, Indiana, USA
23 Gasteiger, J. and Marsili, M. Iterative Partial Equalisation of Orbital Electronegativity-Rapid Access to Atomic Charges. *Tetrahedron* 1980, **36**, 3219–3288
24 Heinz, J. and Jaffe, H.H. Electronegativity I. Orbital Electronegativity of Neutral Atoms. *J. Am. Chem. Soc.* 1962, **84**, 540–546
25 Heinz, J., Whitehead, M.A., and Jaffe, H.H. Electronegativity II. Bond and Orbital Electronegativities. *J. Am. Chem. Soc.* 1963, **85**, 148–154
26 Heinz, J. and Jaffe, H.H. Electronegativity IV. Orbital Electronegativities of the Neutral Atoms of the Periods Three A and Four A and of Positive Ions of Periods One and Two. *J. Am. Chem. Soc.* 1963, **67**, 1501–1505
27 Cieplak, P. and Kollman, P. On the use of electrostatic potential derived charges in molecular mechanics force fields. The relative solvation free energy of *cis-* and *trans*-N-methyl-acetamide. *J. Comput. Chem.* 1991, **12** (10), 1232–1236
28 Giessner-Prettre, C. and Pullman, A. Molecular Electrostatic Potentials: Comparison of *Ab-Initio* and CNDO results. *Theoret. Chim. Acta(Berlin)* 1972, **25**, 83–88
29 *Chemical Applications of Atomic and Molecular Electrostatic Potentials.* (D.G. Truhlar and P. Politzer, Eds.) Plenum Press, New York, 1981, 309–334
30 Giessner-Prettre, C. *QCPE 11.* Quantum Chemical Program Exchange, Department of Chemistry, University of Indiana, Bloomington, Indiana, USA, 1974
31 Hopfinger, A.J. *Conformational properties of macromolecules.* Academic Press, 1973, 74
32 Goto, H. and Osawa, E. Corner Flapping: A Simple and Fast Algorithm for Exhaustive Generation of Ring Conformations. *J. Am. Chem. Soc.* 1989, **111**, 8950–8951
33 Downs, G.M., Gillet, V.J., Holliday, J.D., and Lynch, M.F. Review of Ring Perception Algorithms for Chemical Graphs. *J. Chem. Inf. Comput. Sci.* 1989, **29**, 172–187
34 Newman, W.M. and Sproull, R.F. *Principles of Interactive Computer Graphics.* McGraw-Hill, 1981, 346–348
35 Senn, P. Determination of the position of an atom in space from three interatomic distances. *Computers & Chemistry* 1991, **15**(1), 93–94
36 *Handbook of Chemistry and Physics, 60th Ed.* CRC Press, 1979, van der Waals radii in Angstroms, D-194
37 Lucasius, C.B., Blommers, M.J.J., Buydens, L.M.C., and Kateman, G. A genetic algorithm for conformational analysis of DNA, in *A Handbook of Genetic Algorithms.* (Lawrence Davis, Ed.) van Nostrand Reinhold, 1991, Chapter 18
38 *SYBYL 5.3 molecular modeling package.* Tripos Associates, St. Louis, MO, USA, 1990
39 Digby, P.G.N. and Kempton, R.A. *Multivariate Analysis of Ecological Communities.* Chapman and Hall, 1987, 112–115
40 Press, W.H., Flannery, B.P., Teukolsky, S.A., and Vetterling, W.T. *Numerical Recipes.* Cambridge University Press, 1987, 52–64
41 Williamson, M.P., Davies, J.S., and Thomas, W.A. ^{1}H NMR studies on Cyclo[-Arg(Mtr)-Gly-Asp(Bu^t)-Ser(Bu^t)-Lys(Boc)-] and Cyclo(-Arg-Gly-Asp-Ser-Lys), Cyclic Analogues of the 'Adhesion' Domain of Fibronectin. *J. Chem. Soc. Perkin Trans 2,* 1991, **5**, 601–606
42 Krogsgaard-Larsen, P., Jacobsen, P., and Falch, E. Structure-Activity Requirements of the GABA Receptor. in *The GABA Receptors.* (S.J. Enna, Ed.) The Humana Press, Clifton, 1983, 149–176
43 Codding, P.W. and Muir, A.K.S. Molecular structure of Ro15-1788 and a model for the binding of benzodiazepine receptor ligands. *Molecular Pharmacology,* 1985, **28**, 178–184
44 Bertolasi, V., Ferretti, V., Gilli, G., and Borea, P.A. methyl-β-carboline-3-carboxylate. *Acta Crystallogr.* 1984, **C40**, 1981
45 Bertolasi, V., Ferretti, V., Gilli, G., and Borea, P.A., 2-phenyl-2,5-dihydropyrazolo(4,3-c)quinolin-3(3H)-one. *Acta Crystallogr.,* 1985, **C41**, 107
46 Kolbe, V.M. Opiate receptors: Search for new drugs. *Progress in Drug Research* 1991, **36**, 49–70
47 Glen, R.C. and Rose, V.S. Computer programme suite for the calculation, storage and manipulation of molecular property and activity descriptors. *J. Mol. Graph.* 1987, **5**(2), 79–86
48 *CNDO/2. QCPE 91.* Quantum Chemical Program Exchange, Department of Chemistry, University of Indiana, Bloomington, Indiana, USA