Journal of Computer-Aided Molecular Design, 9 (1995) 532–549
ESCOM

J-CAMD 307

A genetic algorithm for flexible molecular overlay and pharmacophore elucidation

Gareth Jones[a,*], Peter Willett[a] and Robert C. Glen[b,**]

[a]*Department of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.*
[b]*Department of Physical Sciences, Wellcome Research Laboratories, Beckenham, Kent BR3 3BS, U.K.*

Received 28 February 1995
Accepted 14 July 1995

Keywords: Genetic algorithm; Flexible conformational search; Superimposition

Summary

A genetic algorithm (GA) has been developed for the superimposition of sets of flexible molecules. Molecules are represented by a chromosome that encodes angles of rotation about flexible bonds and mappings between hydrogen-bond donor proton, acceptor lone pair and ring centre features in pairs of molecules. The molecule with the smallest number of features in the data set is used as a template, onto which the remaining molecules are fitted with the objective of maximising structural equivalences. The fitness function of the GA is a weighted combination of: (i) the number and the similarity of the features that have been overlaid in this way; (ii) the volume integral of the overlay; and (iii) the van der Waals energy of the molecular conformations defined by the torsion angles encoded in the chromosomes. The algorithm has been applied to a number of pharmacophore elucidation problems, i.e., angiotensin II receptor antagonists, Leu-enkephalin and a hybrid morphine molecule, 5-HT_{1D} agonists, benzodiazepine receptor ligands, 5-HT_3 antagonists, dopamine D_2 antagonists, dopamine reuptake blockers and FKBP12 ligands. The resulting pharmacophores are generated rapidly and are in good agreement with those derived from alternative means.

Introduction

One approach to drug design is to seek to develop new ligands on the basis of structural information about the biological receptor site at which the molecules are expected to show the activity of interest [1]. However, often there is little or no information about the 3D structure of the receptor. In such cases, alternative methods such as the active analogue approach [2] must be invoked that seek to rationalise the ligand–receptor interaction on the basis of structural characteristics of those active molecules that have been identified thus far. One such approach involves aligning the active molecules to identify common structural features (however defined), with the aim of elucidating the pharmacophore that is responsible for the observed activity. There are many ways in which the alignment can be carried out, as reviewed recently by Klebe [3]. However, the procedures that have been described thus far typically have one or more of the following limitations: they require intervention to specify at least some points of commonality, thus biasing the resulting overlays; they encompass conformational flexibility by considering some number of low-energy conformations, rather than the full conformational space of the molecules that are to be overlaid; or they are extremely time-consuming in operation. In this paper, we describe the use of a genetic algorithm (GA) for the overlay of sets of molecules that seeks to overcome these limitations.

The first report on the use of a GA for the superimposition of flexible molecules, and the starting point for the work reported here, was a paper by Payne and Glen [4]. However, this GA was controlled by fitting to constraints or by minimising the distance between known pharmacophore points in the two molecules that were being compared. The GA presented here encodes not only conformational information but also intermolecular map-

*To whom correspondence should be addressed.
**Present address: Tripos Associates Inc., St. Louis, MO 63144, U.S.A.

1. **A set of reproduction operators (crossover, mutation etc.) is chosen. Each operator is assigned a weight.**
2. **An initial population is randomly created and the fitnesses of its members are determined.**
3. **An operator is chosen using roulette-wheel selection based on operator weights.**
4. **The parents required by the operator are chosen using roulette-wheel selection based on scaled fitness.**
5. **The operator is applied and child chromosomes are produced. Their fitness is evaluated.**
6. **The children replace the least fit members of the population.**
7. **If an acceptable solution has been found, stop; otherwise goto 3.**

Fig. 1. Summary of an operator-based GA.

pings between important structural features (such as lone pairs, hydrogen-bond donor protons and aromatic rings) that may be required for activity; in addition, the algorithm does not require any prior knowledge regarding either the constraints or the nature of the pharmacophoric pattern. Indeed, one of the main applications of the procedure described here is the identification of such patterns, since these can then be used to search a corporate or public database of 3D structures for additional, potentially active molecules.

Methods

Genetic algorithms

A GA is a computer program that mimicks the process of evolution by manipulating a collection of data structures called *chromosomes*. A steady-state-with-no-duplicates GA [5–7] was used in the experiments reported here, as summarised in Fig. 1. Starting from an initial randomly generated population of chromosomes, the GA repeatedly applies two genetic operators, *crossover* and *mutation*, which result in chromosomes that replace the least fit members of the population. Crossover combines chromosomes, while mutation introduces random perturbations. Both operators require parent chromosomes that are randomly selected from the existing population with a bias towards the fittest, thus introducing an evolutionary pressure into the algorithm. This selection is known as *roulette-wheel selection*, as the procedure is analogous to spinning a roulette wheel, with each member of the population having a slice of the wheel that is proportional to its fitness. This emphasis on the survival of the fittest ensures that, over time, the population should move towards the optimum solution, e.g., to the correct structural overlay of a series of active molecules presumed to bind to a biological receptor in a similar fashion.

Given a set of active molecules, the GA selects one of them as a *base molecule*, to which the other molecules are fitted. A chromosome in the GA encodes a range of information that is necessary to ensure an appropriate overlay of a molecule onto the base molecule. Specifically, each chromosome contains binary strings that encode angles of rotation about the rotatable bonds in all of the molecules, and integer strings that map hydrogen-bond donor protons, acceptor lone pairs and ring centres in the base molecule to corresponding sites in each of the other molecules. A least-squares fitting process is used to overlay molecules onto the base molecule in such a way that as many as possible of the structural equivalences suggested by the mapping are formed. The fitness of a decoded chromosome is then a combination of the number and similarity of overlaid features, the volume integral of the overlay and the van der Waals energy of the molecular conformations. This GA exploits methods that were developed in the flexible docking algorithm described by Jones et al. [8].

In the following, a detailed account is given of the various components of the program, i.e., the routines that are used to initialise each molecule in the overlay, the chromosome representation that is used to characterise molecular conformations and the features that are to be overlaid, the fitness function that is used to evaluate these chromosomes, and the genetic operators that are applied to the chromosomes.

Initialisation of input structures

In the absence of refined crystallographic co-ordinates, an input structure was normally created using the SYBYL BUILD module [9] and hydrogen atoms were added to all atoms with free valences. Groups within the input structure were ionised if this was appropriate at physiological pH (e.g., alkyl amine, carboxylic acid) and specific atoms were protonated if this was indicated by pK_a or NMR data. A low-energy conformation was generated using molecular mechanics (the SYBYL MAXIMIN energy minimiser with Gasteiger–Marsilli charges [9]). Following this procedure, each input structure was written out from SYBYL as a MOL2 file. All of the rings in each structure

TABLE I
ALLOWED DONORS AND ACCEPTORS BASED ON SYBYL ATOM TYPES

SYBYL atom types	Donor	Acceptor
N3, N2, O3	Y	Y
N1, NAR, O2, OCO2, F, BR, CL	N	Y
NAM, NPL3, N4	Y	N

were identified using a smallest-set-of-smallest-rings (SSSR) algorithm [10], and each structure was then analysed to determine the features that were present, where a feature is a hydrogen-bond donor proton, a lone pair or a ring. The base molecule was defined to be the one with the smallest number of features.

Hydrogen-bond donor and acceptor atoms were identified in each of the input structures using the SYBYL atom-type characterisation in Table 1 (with the restriction that each donor must be bonded to at least one hydrogen). Donor hydrogens could then be identified, and lone pairs were added to acceptors at a distance of 1.0 Å from the acceptor. All freely rotatable acyclic single bonds that were not connected to terminating atoms were selected as being rotatable. Additionally, single cyclic bonds could also rotate if the technique of bond breaking and ring closure was used (see the overlay of two FKBP12 ligands in the Results). Prior to superimposition, a random translation was applied to each input structure (including the base molecule) and random rotations were applied to all rotatable bonds.

The chromosome representation

A chromosome of $2N-1$ strings was used to encode a molecular overlay involving N molecules. This contained N Gray-coded [5,7] binary strings, each encoding conformational information for one structure with each byte encoding an angle of rotation about a rotatable bond, and $N-1$ integer strings, each encoding a mapping between features in a molecule (other than the base molecule) to features, of the same type, in the base molecule. For example, a lone pair in one molecule could be mapped to a particular lone pair in the base molecule, under the implicit assumption that the lone pairs in both molecules interacted with the same hydrogen-bond donor in the receptor.

On decoding a chromosome, the fitness function of the GA would attempt to satisfy the specified mapping by using a least-squares fitting technique. In order to make the mapping chemically sensible, the mapping was one-to-one between similar features. For example, it would not make sense if a lone pair was mapped to two different lone pairs in the base molecule, or if it was mapped onto a hydrogen-bond donor proton. Each integer string had a length L, where L is the number of features in the base molecule. Because the mappings were one-to-one, the integer string was constrained to have no duplicate values. Each feature in every molecule was assigned a unique label. The labels of the base-molecule features were then arranged in a list of length L. If V was the integer value at position P on the integer string and B the Pth element in the list of base-molecule labels, then the feature with label V was mapped onto the base-molecule feature with label B. By associating features in each molecule to base-molecule features, these mappings suggested possible pharmacophoric points. On decoding the chromosome, the GA used a least-squares routine to attempt to form as many points as possible.

It was possible for the base molecule to have a larger count of a particular feature than another molecule. This problem was solved by assigning dummy labels, so that each molecule had as many feature labels of each particular type as the base molecule. The dummy labels were negative integers, and identical dummy labels were not permitted in the same chromosome (that is, each individual mapping had its own dummy labels). Dummy labels were ignored by the fitness function, but treated as normal values by the genetic operators, which are discussed below.

Extra dummy labels were assigned to rings. Consider the example in Fig. 2, in which it is assumed that molecule A is to be superimposed onto molecule B. In order to overlay the sulphur and amide functional groups, the best superimposition is to map ring-1 in A to ring-1 in B, and to have ring-2 unmapped. However, if ring-1 in A is mapped to ring-1 in B then ring-2 in A must be mapped to ring-2 in B, unless extra dummy ring labels are provided. A number of extra dummy ring labels were thus made available, the number being equal to the difference between the number of rings in the base molecule (as defined by the SSSR routine mentioned above) and the number of cyclic regions in the base molecule. Thus, for the example in Fig. 2 there is one extra dummy ring label and ring-2 in A can be unmapped, i.e., mapped to the dummy label.

The fitness function

The fitness function was evaluated in six stages, as follows:

(1) A separate conformation was generated for each molecule by applying the bond rotations encoded in the appropriate binary string.

(2) Each molecule was superimposed on top of the base molecule using a transformation obtained from a least-squares procedure that fitted to the mapping encoded in the appropriate integer string.

(3) A van der Waals energy was obtained for the internal steric energy of each molecule.

(4) A volume integral was obtained for the common volume between each molecule and the base molecule.

(5) A similarity score was generated by determining which features were common to all molecules in the current overlay.

(6) A final fitness score was generated by performing a weighted sum on the terms calculated in steps 3, 4 and 5.

Each Gray-coded byte in the binary string was decoded to give an integer value between 0 and 255. This integer value was linearly rescaled to give a real number between 0 and 2π, which was used as an angle of rotation, in radians, for the appropriate rotatable bond. The randomised 3D co-ordinates for the molecule were used as a starting configuration. Bond rotations were successively applied around the rotatable bonds to generate a new set of co-ordinates for the molecule. The resulting conformations were then passed on to the least-squares fitting procedure.

A virtual point, representing a donor or acceptor atom in the receptor with which the molecules interact, was created for each hydrogen-bond donor proton and acceptor lone pair in a molecule at a distance of 2.9 Å from the donor or acceptor, in the direction of the hydrogen or lone pair. A virtual point was created at the centre of each ring. Consider the superimposition of one molecule, A, onto the base molecule. Let N be the number of base-molecule features (minus any dummy labels that are required by molecule A), so that decoding a chromosome would give rise to N pairs containing a virtual point in the base molecule and a virtual point in A. Then, a Procrustes Rotation [11], with a correction to remove inversion, yielded a geometric transformation that, when applied to all the virtual points from molecule A in the N pairs, minimised the least-squares distance between all of the virtual points from molecule A and the corresponding base-molecule virtual points. As not all possible features in the base molecule will necessarily be included in a pharmacophore, a second least-squares fit was applied to minimise the distance between those pairs of points that were less than 3 Å apart.

The least-squares fit failed if the second pass contained fewer than three pairs of points, in which case the fitness function returned an error and the chromosome was excluded from the population. This limitation often caused severe problems during the initialisation of the GA population, where the values of all elements in each chromosome were assigned randomly, and the fitting procedure thus needed to be modified to accommodate this problem. If, during initial population generation, a failure in fitting occurred, the fitness module returned the identity of the molecule that failed the fitting process. The portion of the chromosome which mapped this molecule to the base molecule was then regenerated randomly and returned to the fitness module for refitting. Up to 10 regenerations were permitted per chromosome before the chromosome was finally discarded.

Calculation of the van der Waals energy

The internal steric energy for each molecule was calculated using a Lennard-Jones 6–12 potential [12]. This potential was of the following form:

$$E_{ij} = k_{ij}(1.0/a_{ij}^{12} - 2.0/a_{ij}^{6}) \quad (1)$$

where i and j are a pair of atoms, a_{ij} is the distance between i and j divided by the sum of their van der Waals radii and k_{ij} is the arithmetic approximation to the geometric mean of k_i and k_j, where the k values are dependent on the atom type and are parameters of the Lennard-Jones potential. The values for the radii and k values for each atom type were taken from the SYBYL general-purpose Tripos force field (v. 5.2) [13]. In order to efficiently calculate the interaction, an energy lookup table and a cutoff distance of $a_{ij} = 1.0$ were employed.

The steric energy of a molecular conformation was expressed as the difference between the 6–12 energy of this conformation and the 6–12 energy of the original input molecular conformation, prior to the randomisation of molecular coordinates (recall that the input structures are normally minimised and thus are low-energy conformations). If this difference was negative (indicating that the conformation was of lower energy than the input structure), then the steric energy for that molecule was set to zero. This was done to prevent the GA optimising on van der Waals energy rather than similarity. In order for the van der Waals energy term in the final fitness score to be independent of the number of molecules in the overlay, the mean 6–12 energy per molecule was determined. We call this energy *vdW_energy*.

The steric energy term is intended to ensure that structures generated by the GA are low-energy conformers. For reasons of efficiency, no full molecular mechanics potential was implemented. Thus, a final optimisation of the structures generated by the algorithm may be appropriate. However, given that only single bonds with no significant energy barrier to rotation are manipulated by the GA, a van der Waals energy should be sufficient to generate low-energy structures. This assumption is supported by the fact that few large changes in geometry are observed when minimising overlays generated by the GA.

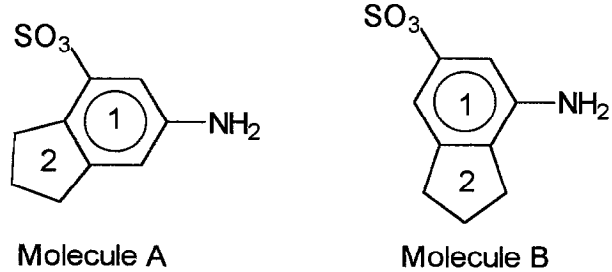


Fig. 2. Ring mapping problem. Mapping rings 1 and 2 in A to rings 1 and 2 in B results in the two molecules being incorrectly overlaid.

Calculation of the volume integral

In order to predict which portions of the actives are in contact with the active site, the algorithm should determine common molecular surface areas as a measure of similarity. However, such a calculation is extremely time-consuming, and pairwise common volumes were thus determined between the base molecule and each of the other molecules. In order to speed up the determination of common molecular volume, the calculation was approximated by treating atoms as spheres and summing the overlay between spheres in the two different molecules.

To determine the volume integral between two molecules A and B, each atom i in A was compared with every atom j in B. Let d_{ij} be the interatomic distance between the two atoms, R_i the van der Waals radius of the atom in A and R_j the van der Waals radius of the atom in B. The values of the van der Waals radius were dependent on atom type and were the same as those used in the SYBYL general-purpose Tripos force field (v. 5.2) [13]. Four separate cases could result when determining the volume integral between the two atoms:

(1) The distance d_{ij} is greater than either R_i or R_j. Here, the common volume between the two atoms is 0.

(2) $R_i > R_j$ and $d_{ij} < R_i - R_j$. Here, atom j is completely enclosed in atom i, and the common volume is given by $4/3\ \pi R_j^3$.

(3) $R_j > R_i$ and $d_{ij} < R_j - R_i$. Here, atom i is completely enclosed in atom j, and the common volume is given by $4/3\ \pi R_i^3$.

(4) Otherwise the two atomic spheres partially overlap. Here, the common volume is defined by the equation for the intersection of two spheres. Let

$$x = (R_i^2 - R_j^2 - d_{ij}^2) / 2d_{ij} \tag{2}$$

then the common volume is given by

$$\pi/3\,(2R_i^2 + 2R_j^2 + d_{ij}^2) - \pi\,(xR_i^2 + (R_j^2 + x\,d_{ij})\,(d_{ij} - x)) \tag{3}$$

The total volume integral between molecules A and B was determined by summing all the individual terms from atomic common volumes. In order for the volume integral term in the final fitness score to be independent of the number of molecules in the overlay, the mean volume integral per molecule with the base molecule was determined. We call this term *volume_integral*.

Calculation of the similarity score

A similarity score was determined for the overlaid molecules. This score, *similarity_score*, is the sum of three terms. The first term is a score for the degree of similarity in position, orientation and type between hydrogen-bond donors in the base molecule and in the other molecules; the second term is a score derived from comparing hydrogen-bond acceptors; and the third term is a score that results from comparing the position and orientation of aromatic rings. Thus:

$$\text{similarity\_score} = \text{donor\_score} + \text{acceptor\_score} + \text{ring\_score} \tag{4}$$

Similarity index for donor and acceptor atom types In order to assign similarity scores, the GA required the use of a function that determines how similar two hydrogen-bond donor or acceptor types are. Let type_sim[a, b] be a weight between 0 and 1 that is a measure of the similarity between hydrogen-bond types a and b, where a and b are either both donor types or both acceptor types. Donor and acceptor types were labelled by the fragment types used in Ref. 8 to model their bonding strength (the labels NPLCG and NPLCA refer to the guanidine and arginine donor types). Let ACCEPTORS be the set of hydrogen-bond acceptor atom types and DONORS the set of hydrogen-bond donor atom types:

$$\begin{aligned}\text{DONORS} &= \{\text{N4, NPL3, N3DA, NAM, O3DA, N2DA, NPLCG, NPLCA}\} \\ &= \{d_1, d_2, \ldots\ldots, d_8\} \\ \text{ACCEPTORS} &= \{\text{BR, N2DA, O2, OCO2, CL, N1, N3A, O3A, F, N2A, N3DA, O3DA, NACID}\} \\ &= \{a_1, a_2, \ldots\ldots, a_{13}\}\end{aligned}$$

Let bond_strength[d_i, a_j] be the strength of the hydrogen bond between donor d_i and acceptor a_j. The experimental determination of the solvated hydrogen-bonding energies for all pairs of donors and acceptors has been detailed by Jones et al. [8]. These bonding energies account for the displacement of water during bond formation.

For donors, type_sim[d1, d2] was defined as follows:

$$\text{ratio\_sim[d1, d2]} = \sqrt[13]{\prod_{i=1}^{13} \frac{\text{bond\_strength[d1, a}_i\text{]}}{\text{bond\_strength[d2, a}_i\text{]}}}$$

$$\begin{aligned}&\text{if} && \text{ratio\_sim[d1, d2]} > 1 \\ &\text{then} && \text{type\_sim[d1, d2]} = \frac{1}{\text{ratio\_sim[d1, d2]}} \\ &\text{else} && \text{type\_sim[d1, d2]} = \text{ratio\_sim[d1, d2]}\end{aligned}$$

For acceptors, type_sim[a1, a2] was defined as follows:

$$\text{ratio\_sim[a1, a2]} = \sqrt[8]{\prod_{i=1}^{8} \frac{\text{bond\_strength[d}_i\text{, a1]}}{\text{bond\_strength[d}_i\text{, a2]}}}$$

$$\begin{aligned}&\text{if} && \text{ratio\_sim[a1, a2]} > 1 \\ &\text{then} && \text{type\_sim[a1, a2]} = \frac{1}{\text{ratio\_sim[a1, a2]}} \\ &\text{else} && \text{type\_sim[a1, a2]} = \text{ratio\_sim[a1, a2]}\end{aligned}$$

TABLE 2
DONOR ATOM-TYPE SIMILARITIES (ratio_sim[d1, d2])

d1	d2							
	N4	NPL3	N3DA	NAM	O3DA	N2DA	NPLCG	NPLCA
N4	1.00	4.05	3.20	7.68	2.88	4.17	1.93	2.84
NPL3	0.25	1.00	2.27	0.88	0.44	1.84	0.39	0.48
N3DA	0.31	0.44	1.00	0.55	0.80	0.00	0.68	1.13
NAM	0.13	1.13	1.80	1.00	0.36	2.22	0.24	0.33
O3DA	0.35	2.25	1.25	2.77	1.00	4.02	0.66	0.96
N2DA	0.24	0.54	0.00	0.45	0.25	1.00	0.33	0.38
NPLCG	0.52	2.60	1.47	4.25	1.52	3.00	1.00	1.48
NPLCA	0.35	2.09	0.89	3.04	1.04	2.61	0.68	1.00

These atom types have been defined by Jones et al. [8] in their program for docking flexible ligands.

In calculating the similarities, the geometric mean of bond strength ratios was preferred to the arithmetic mean, since this resulted in the property type_sim[a, b] = type_sim[b, a].

A correction was made to the above calculation for those cases where the bonding energy was positive, i.e., the bond was unattractive. In these instances the calculated bond energy is largely meaningless, and it is likely that, following optimisation, the model fragments would not be in a hydrogen-bonding position. For this reason, the geometric mean ratio_sim was only determined over those ratios whose bonding energies were less than -0.25 kcal mol^{-1}. The calculated values of ratio_sim are shown in Tables 2 and 3. The geometric means are shown rather than the atom-type similarities, so that it is clear which of the two donors or acceptors in the ratio is the stronger.

The rationale behind this approach of measuring similarity is that the GA should overlay donor or acceptor groups of similar strength. While the method described here is intuitively acceptable, the main justification for the adoption of this similarity index is the success of the algorithm in superimposing sets of known actives.

Donor similarity score Each virtual point corresponding to a hydrogen-bond donor proton in the base molecule was used to define a hydrogen-bonding centre with the potential to interact with acceptors within a receptor molecule. A virtual point from every other molecule, corresponding to the hydrogen-bond donor proton that was geometrically closest to the base-molecule virtual point, was added to each of these hydrogen-bonding centres. A score, vec_wt × sim_wt, was then assigned to the hydrogen-bonding centre, where vec_wt is a measure of closeness of virtual point positions and hydrogen-bond vectors in the hydrogen-bond centres and where sim_wt is a measure of the similarity of the donors involved in the hydrogen-bond centre.

The centroid of the virtual point positions was determined in order to estimate the similarity of virtual point positions within the hydrogen-bond centre. Let vp_distance_wt be a measure of how close the centroid is to the base-molecule virtual point in the centre and let vp_d be the distance between this virtual point and the centroid of the virtual point positions. If vp_d was less than 0.5 Å, then vp_distance_wt was 1, or if vp_d was greater than 1.75 Å, then vp_distance_wt was 0. Otherwise, vp_d lay in the interval [0.5, 1.75] and was linearly rescaled to the interval [1, 0] and squared to give vp_distance_wt.

TABLE 3
ACCEPTOR ATOM-TYPE SIMILARITIES (ratio_sim[a1, a2])

a1	a2											
	BR	N2DA	O2	OCO2	CL	N1	N3A	O3A	F	N2A	N3DA	O3DA	NACID
BR	1.00	0.18	0.29	0.12	0.68	0.37	0.19	0.95	2.11	0.20	0.61	0.38	0.31
N2DA	5.58	1.00	1.21	0.22	2.41	2.05	1.30	1.26	0.00	1.12	0.98	1.17	0.39
O2	3.47	0.83	1.00	0.27	1.99	0.95	0.51	2.76	6.94	0.71	0.58	1.11	0.65
OCO2	8.11	4.54	3.72	1.00	8.73	9.77	6.25	10.3	13.1	7.85	5.63	4.55	2.35
CL	1.46	0.42	0.50	0.11	1.00	0.85	0.45	0.52	1.74	0.47	0.89	0.44	0.19
N1	2.73	0.49	1.05	0.10	1.18	1.00	0.54	1.25	0.00	0.74	0.61	0.74	0.18
N3A	5.14	0.77	1.96	0.16	2.22	1.87	1.00	2.34	0.00	1.39	0.90	1.16	0.30
O3A	1.05	0.79	0.36	0.10	1.91	0.80	0.43	1.00	1.44	0.60	0.49	0.40	0.23
F	0.47	0.00	0.14	0.08	0.57	0.00	0.00	0.70	1.00	0.00	1.00	0.22	0.22
N2A	4.96	0.89	1.41	0.13	2.14	1.34	0.72	1.68	0.00	1.00	0.82	0.99	0.25
N3DA	1.64	1.02	1.71	0.18	1.12	1.63	1.11	2.04	1.00	1.22	1.00	1.29	0.33
O3DA	2.65	0.86	0.90	0.22	2.29	1.35	0.86	2.49	4.61	1.01	0.77	1.00	0.51
NACID	3.20	2.56	1.54	0.43	5.40	5.45	3.39	4.26	4.47	3.29	3.05	1.95	1.00

These atom types have been defined by Jones et al. [8] in their program for docking flexible ligands.

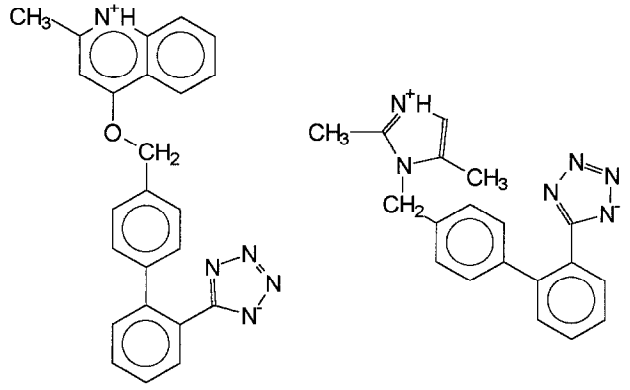


Fig. 3. Angiotensin antagonists.

A correction to this method was applied to ensure that the possible points of interaction with the receptor, represented by the hydrogen-bonding centre, did not lie within the van der Waals volume of the overlaid molecules. A hydrogen-bond centre was only formed for a base-molecule hydrogen-bond donor proton if the virtual point associated with that hydrogen lay outside the van der Waals volume of the base molecule. When choosing a virtual point from another molecule, M, to add to the hydrogen-bond centre, only those virtual points that did not lie inside the van der Waals volume of M were considered. If no virtual point was available from M to add to the hydrogen-bond centre, then a vec_wt of zero was assigned.

An indication of the similarity of hydrogen-bonding vectors can be obtained by considering the closeness in positions of the hydrogen-bond donors associated with the virtual points in the hydrogen-bonding centre. Let such a measure be donor_distance_wt and let donor_d be the distance between the donor associated with the base-molecule virtual point and the centroid of all hydrogen-bond donors that were connected to the virtual points in the hydrogen-bonding centre. As before, if donor_d was less than 0.5 Å, then donor_distance_wt was 1, or if d was greater than 1.75 Å, then donor_distance_wt was 0. Otherwise, donor_d lay in the interval [0.5, 1.75] and was linearly rescaled to the interval [1, 0] and squared to give donor_distance_wt.

The value of vec_wt could now be determined. If either donor_distance_wt or vp_distance_wt was 0, then vec_wt was set to 0. Otherwise, if donor_distance_wt was less than vp_distance_wt, then

$$\text{vec\_wt} = (\text{donor\_distance\_wt} + \text{vp\_distance\_wt}) / 2.0$$

and else

$$\text{vec\_wt} = \text{vp\_distance\_wt} / 2.0$$

This combination was chosen so that the GA would not superimpose donors in preference to virtual points, since it is the virtual points that represent the points of interaction with the receptor.

The donor similarity term, sim_wt, was then determined for the hydrogen-bond centre. The similarity index type_sim (defined above) determined the similarity between two donor types. However, if the overlay was of three or more molecules, a similarity term that determined the similarity between many donor types was required. Let d_b be the type of the donor atom in the base molecule that was connected to the virtual point at p_b. Let d_m be the donor type of a donor (in another molecule) that was associated with another virtual point in the hydrogen-bond centre and let mol_sim_wt = type_sim [d_b, d_m]. Different values of mol_sim_wt were determined for all d_m donor types in the hydrogen-bond centre, and sim_wt was then set to the smallest value thus found.

Once vec_wt and sim_wt had been determined for a given hydrogen-bonding centre, the contribution vec_wt × sim_wt was determined, and donor_score was then the sum of all such contributions from all hydrogen-bonding centres containing donor hydrogens.

One possible problem encountered with this similarity measure was that it was weighted towards matching those donors that contained a larger number of donor hydrogens. For example, an NH_2 group in the base molecule may be able to contribute a score of 2.0 to donor_score, whereas an OH group can at most contribute only 1.0. In order to redress this problem, a correction was applied to the contribution that each hydrogen-bonding centre made to donor_score. Each of the donors connected to a hydrogen-bond donor proton in the hydrogen-bonding centre was examined in turn. A count was made of the number of donor hydrogens and lone pairs that were connected to

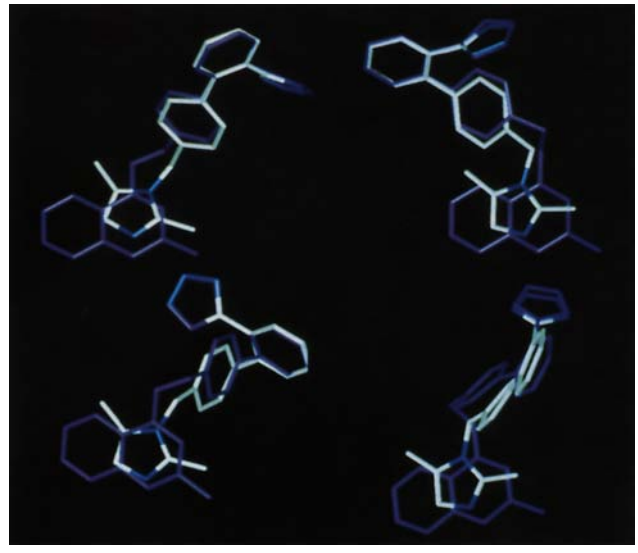


Fig. 4. Overlays of two angiotensin II receptor antagonists. Structure **23** is shown coloured by atom type and structure **5f** is purple.

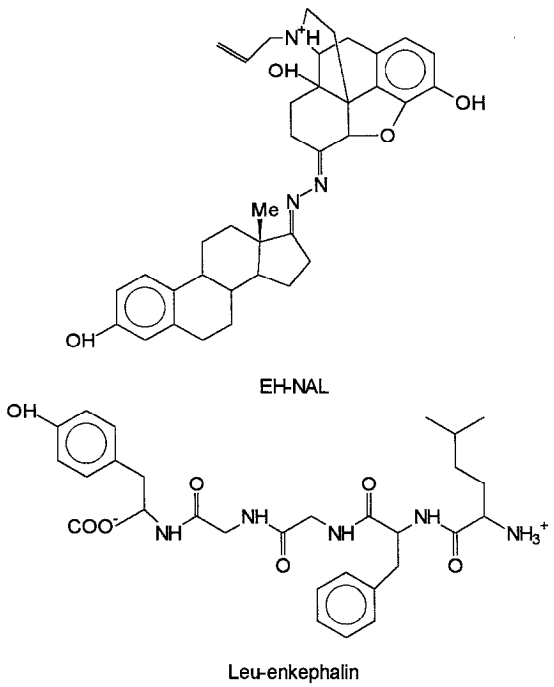


Fig. 5. Hybrid morphine and Leu-enkephalin.

that donor. Let the smallest such count, belonging to donor D, be min_points. If all donor hydrogens (and lone pairs, if any) attached to D were superimposed, then the maximum contribution to donor_score would be min_points, and the possible contribution from D, score, was hence normalised by dividing by min_points.

It is clear that, although any hydrogen-bond donor proton from the base molecule can appear in only one hydrogen-bonding centre, it is nevertheless possible that a hydrogen-bond donor proton from another molecule could appear in more than one hydrogen-bonding centre. The following procedure was adopted, in order to correct for this anomaly. Let F be a feature that occurs in two hydrogen-bonding centres, d1 the distance between F and the base-molecule feature in the first hydrogen-bonding centre, and d2 the distance between F and the base-molecule feature in the second centre. Suppose (without loss of generality) that d1 is less than d2, then the contribution from donor_score from the second centre was set to 0. This mechanism prevented the same feature from contributing twice to the donor similarity score.

Acceptor similarity score Each acceptor lone pair in the base molecule was used to define a hydrogen-bonding centre with the potential to interact with donors within a receptor macromolecule. The process used to generate the score acceptor_score was entirely analogous to that used when determining donor_score.

Aromatic ring similarity score The third term in the similarity score, ring_score, required the use of normals to aromatic rings. Given an aromatic ring (determined using the atom types of its constituent atoms) of n atoms with position vectors p_1, p_2, ..., p_n and a centre

$$c = \frac{1}{n}\sum_{i=1}^{n} p_n \qquad (5)$$

a mean normal direction was determined:

$$m = (p_n - c) \otimes (p_1 - c) + \sum_{i=1}^{n-1} (p_i - c) \otimes (p_{i+1} - c) \qquad (6)$$

The mean normal m was scaled to a magnitude of size 2.9 Å, so that similarity scores calculated on ring normals would be comparable to those calculated on virtual points.

There are clearly two normal directions, m and –m, to any aromatic ring. The deduction of the most similar normals in two rings would thus involve four comparisons. In order to reduce this overhead, only one normal was used for each ring. The directional components (along the x-, y- and z-axis) of m were examined and the largest component was identified. If this component was positive then m was chosen, otherwise –m was chosen as being representative of the ring.

A third type of hydrogen-bonding centre was created for aromatic rings. Each aromatic ring in the base molecule was used to define a hydrogen-bonding acceptor centre. The ring centre normal that was geometrically closest to the base-molecule ring centre normal was added to the list.

In order to estimate the similarity of ring positions within a hydrogen-bond centre, the centroid of ring centres was determined. Let d be the distance between the centroid and the base-molecule ring centre. As before, if d was less than 0.5 Å then ring_distance_wt was 1, and if d was greater than 1.75 Å then ring_distance_wt was 0. Otherwise, d lay in the interval [0.5, 1.75] and was linearly rescaled to the interval [1, 0] and squared to give ring_distance_wt.

In order to measure the similarity of ring orientation, the centroid of the normals to the rings that comprised the hydrogen-bonding centre was computed. Let normal_distance_wt be a measure of how close the base-molecule

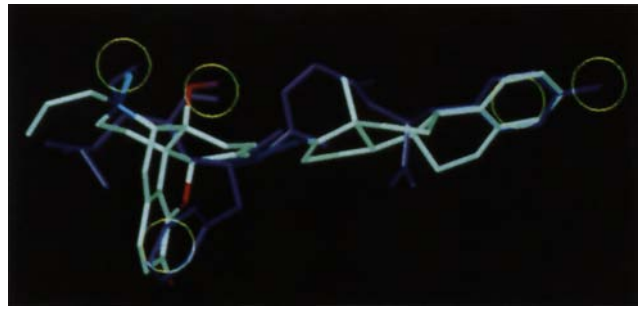


Fig. 6. Overlay of Leu-enkephalin and hybrid morphine. The hybrid morphine, EH-NAL, is shown coloured by atom type and Leu-enkephalin is shown coloured purple. The elucidated pharmacophore is indicated by yellow circles.

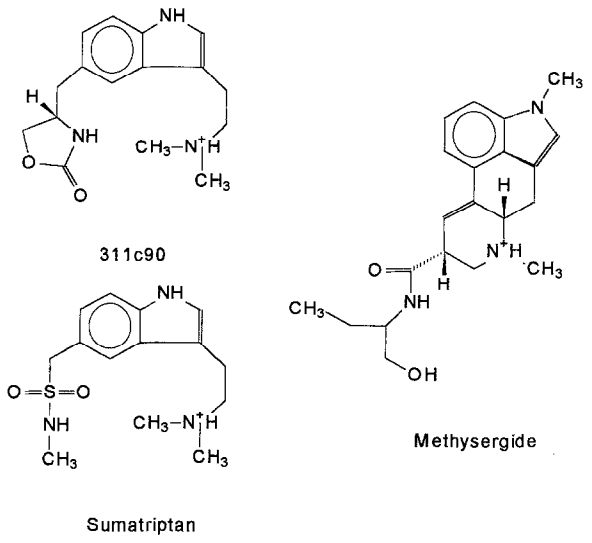


Fig. 7. 5-HT_{1D} agonists.

normal is to the centroid, where normal_distance_wt is determined using the same method described in the previous paragraph. If either normal_distance_wt or ring_distance_wt were 0 then ring_score was set to 0. Otherwise,

$$\text{ring\_score} = \frac{(\text{normal\_distance\_wt} + \text{ring\_distance\_wt})}{2.0} \quad (7)$$

A correction (as described above for the donor similarity score) was made if any aromatic ring normal appeared in two hydrogen-bonding centres.

The final fitness score

The final fitness score was determined by a weighted sum of the common volume, similarity score and steric energy. The fitness score is given by:

$$\text{volume\_integral} + 750 \times \text{similarity\_score} - 0.05 \times \text{vdW\_energy} \quad (8)$$

The weights of 750 and 0.05 were determined by empirical adjustment to give reasonable overlays (where the algorithm is driven to generate good pharmacophores without producing high-energy structures) over a wide range of examples. The selection of ideal weights is a complex process and it is an area of current investigation.

Genetic operators

The GA employed two genetic operators: mutation and crossover. The crossover operator required two parents and produced two children. The mutation operator required one parent and produced one child [5–7]. The operator weight for mutation was set equal to the operator weight for crossover. Parents were selected using the technique of roulette-wheel selection on linear normalised fitness values [6]. The GA was parameterised to give a selection pressure of 1.1, where the selective pressure represents the relative probability that the best individual will be chosen as a parent compared to the average individual. This low selection pressure reduced the likelihood of the GA converging to suboptimal solutions.

The crossover operator performed two-point crossover on integer strings and one-point crossover on binary strings. The one-point crossover was the traditional GA recombination operator [5], while the two-point crossover was the PMX crossover operator (including the duplicate removal stage) that has been described by Brown et al. [14] and by Goldberg [5]. As noted previously, each chromosome consisted of $2N-1$ strings (N binary strings and $N-1$ integer strings, where N is the number of molecules in the overlay). Crossover proceeded as follows. A random number, r, between 1 and $2N-1$ was generated. Crossover was then applied to the rth string in the parent chromosomes (using integer-string two-point crossover if this string was a mapping and one-point binary crossover otherwise). The remainder of the parent chromosomes were then copied to the children unchanged.

The mutation operator performed binary-string mutation on binary strings and integer-string mutation on integer strings. The binary-string mutation was identical to that described by Davis [6]. Each bit in the binary string had a probability of mutation equal to 1/L, where L is the length of the binary string. If the binary string remained unchanged after one application of the mutation operator, the operator was repeatedly applied until the string was mutated. The integer-string mutation was identical to that described by Brown et al. [14]. A position was randomly chosen on the integer string and the value at that position was mutated to a (different) new value that was randomly chosen from the set of allowed integer values. If this new value occurred elsewhere on the integer string, it was replaced by the original value at the mutated position. The operator proceeded as follows. A random number, r, between 1 and $2N-1$ was generated. Mutation was then applied to the rth string in the parent chromosome (using integer-string mutation if this string was a mapping and binary mutation otherwise). The remainder of the parent chromosome was then copied to the child unchanged.

The island model

It should be possible to make the algorithm very efficient by utilizing the fact that GAs are well suited to implementation in a distributed environment by means of the so-called *island* model. This involves separate subpopulations and the migration of individual chromosomes between the subpopulations [15,16]. The island model has attracted growing interest, not only because it represents a practical and efficient method of parallelising the GA,

and thus reducing the observed run time, but also because it has been observed that the resulting distributed GA with several small subpopulations often outperforms a GA with a single large population equal in size to the sum of the distributed subpopulations.

A simple island model was implemented using a serial algorithm. N_ISLANDS subpopulations were created by the GA and arranged in a ring, such that each island had two neighbours. Genetic operators were then applied to each subpopulation in turn, with parents being selected from that subpopulation and children being inserted into that same subpopulation. Let MAXOPS be the maximum number of operations applied by the GA. Thus (if the termination conditions were not satisfied), MAXOPS/N_ISLANDS operations would be applied to each subpopulation over the course of a GA run.

The migration operator required one parent and produced one child. The child was an exact copy of the parent. Let p be the subpopulation to which the migration operation was applied and let n be a subpopulation randomly selected from the two neighbours of p. Roulette-wheel parent selection was performed on n to produce the parent and the child was then inserted into p. An operator weight was used to determine how many migrations were performed, relative to mutation and crossover, and it was found that a 5% migration rate gave good results.

Initial experiments showed that there was no perceptible difference in performance between using five subpopulations of size 100 and a single population of size 500, although the island model showed slightly shorter run times. However, on implementation of a parallel version on five processors (either on a multiprocessor machine or a workstation LAN), up to five times speed-up could be achieved. This is currently under evaluation and will be reported elsewhere.

Results

General

This section describes the application of the GA to a number of diverse overlay and pharmacophore-generation problems; further examples of the application of this algorithm have been presented by Jones [17]. The parameters MAXOPS, OPS_INC and FITNESS_INC were required by the GA in order to set the termination conditions. The GA would terminate if the number of operations exceeded MAXOPS. Otherwise, the GA would terminate if the fitness of the best individual in the population had not increased by the value of FITNESS_INC in the past OPS_INC operations. In the following, except where noted otherwise, the following default parameter values were used: N_ISLANDS = 5; POPSIZE = 100; OPS_INC = 6500; FITNESS_INC = 0.01; and MAXOPS = 50 000. The GA was run 10 times for each of the problems, and the resulting fittest solutions from each run were ranked in order of decreasing fitness. In the following, references to the 'best solution', the 'worst solution', etc. correspond to the position of the solution in the ranked fitness list of final solutions. All CPU times are for a Silicon Graphics R4000 Indigo II workstation. The experimental results are illustrated in the colour figures (Figs. 4, 6, 8, 10, 12, 14, 16, 18 and 21), where the base molecules are coloured by atom type; hydrogens and lone pairs are not generally displayed, unless they are of particular significance in the overlay.

Overlay of two angiotensin II receptor antagonists

Bradbury et al. have reported a study of 2-alkyl-4-(biphenylmethoxy)quinoline derivatives as angiotensin II receptor antagonists [18]. Two of their structures, numbered **5f** and **23** in their paper, are shown in Fig. 3, with structure **23** as the base molecule to which structure **5f** was fit. For this relatively simple problem, the default value for MAXOPS was lowered to 15 000 and N_ISLANDS was set to 1, with the average time for each GA run being 1 min and 30 s.

The eight best solutions, when ranked in order of decreasing fitness, produced overlays where every feature was successfully mapped to an equivalent feature in structure **5f**, with the remaining two runs failing to match one of these features. However, the GA solutions showed considerable variation in molecular conformation. Figure 4 shows four different overlay conformations that were representative of the eight successful GA solutions. Structure **23** is shown coloured by atom type and structure **5f** is shown in purple. One of these conformations (the top left overlay) is similar to a stereoview overlay shown in Ref. 18. The generation of different possible overlay conformations is a useful side-effect of the nondeterministic nature of the algorithm.

Masek et al. have described a method for the comparison of molecular shapes and exemplified it by comparing two angiotensin II receptor antagonists that were virtually identical to the antagonists shown in Fig. 3 [19]. Starting with a set of low-energy conformations of both molecules, the method involves a set of exhaustive volume comparisons to determine the optimal structural overlay. Using these two structures and the parameterisation described above, a series of 10 GA runs were performed. Near-optimal overlays were observed in eight of the runs, with the other two runs failing to match one feature. The average execution time of the GA was 1 min and 40 s. Stereoviews of the results obtained by Masek et al. [19] appear to be in accordance with those found by the GA; however, their procedure is far more time-consuming, requiring 96 h of CPU time on a Silicon Graphics R4000 workstation. The long run time of this algorithm is presumably due to the overhead required to determine exact intersection volumes.

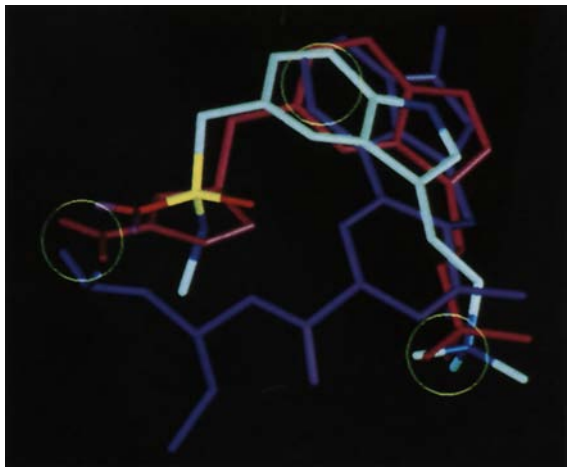


Fig. 8. Overlay of 5-HT_{1D} agonists. The base molecule, sumatriptan, is coloured by atom type. Methysergide is shown in purple and 311c90 in red. The elucidated pharmacophore is indicated by yellow circles.

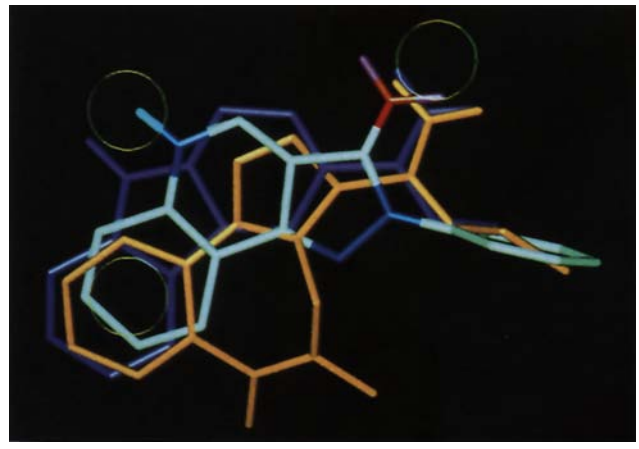


Fig. 10. Overlay of benzodiazepine receptor ligands. CGS-8216 is coloured by atom type, Rol15-1788 is in orange and methyl-β-carboline-3-carboxylate is in purple. Points of interest are indicated by yellow circles (see the text for discussion).

Overlay of Leu-enkephalin on a hybrid morphine

The second superimposition problem to be considered here involved two very different structures, specifically, the hybrid morphine molecule EH-NAL, a mixed azide between estrone and naloxone [20], and Leu-enkephalin, as shown in Fig. 5. Although only two molecules are involved, this is an extremely demanding problem, as Leu-enkephalin is highly flexible, containing 20 rotatable bonds. In contrast, EH-NAL has just six rotatable bonds in side chains. The GA was run 10 times to generate 10 possible overlays. The mean run time was 9 min and 13 s.

Figure 6 shows the best solution (ranked by GA fitness score) that was obtained. The base molecule, EH-NAL, is shown coloured by atom type, while Leu-enkephalin is shown in purple. The pharmacophore identified by the GA contains five features, indicated by yellow circles, i.e., two aromatic rings, one phenol group, the protonated nitrogen (for which the connected hydrogens are shown to illustrate their common directionality), and the GA has also overlaid an sp^3 oxygen in EH-NAL with an sp^2 oxygen in Leu-enkephalin, such that a lone pair from each (displayed in Fig. 6) is aligned in the same direction. Six of the 10 runs (including the five best runs) identified the first four of these five features, with just the fittest identifying the oxygen overlap.

The bound conformations of these molecules are not known, so it is not possible to judge the accuracy of the GA. It is, however, interesting to note that Kolb [20] has obtained a very similar fit (superimposing both rings, the phenol group and the protonated nitrogen) using molecular dynamics with simulated annealing. However, this approach is not fully automated, unlike the GA, and would appear to be much more time-consuming in operation (although exact times are not available).

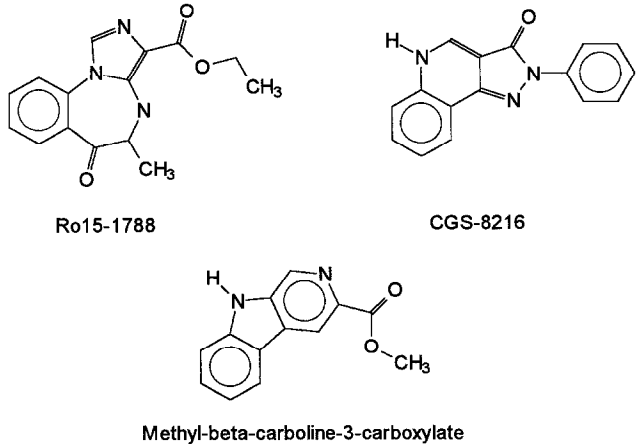


Fig. 9. Benzodiazepine receptor ligands.

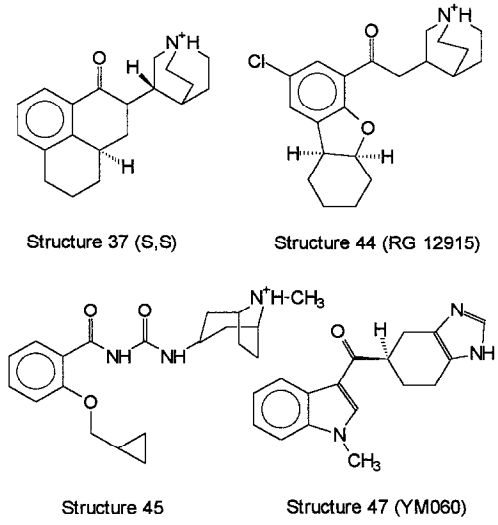


Fig. 11. 5-HT_3 receptor agonists (the structure numbers used are from Ref. 23).

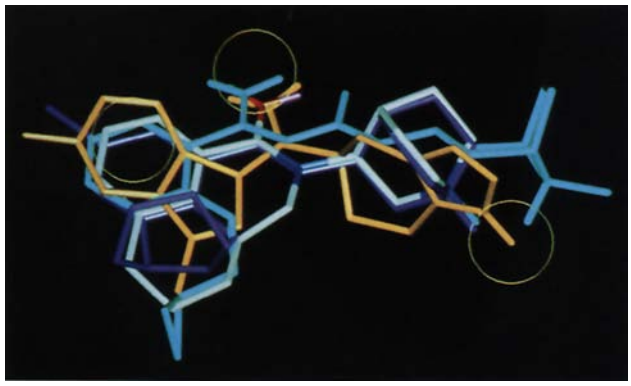


Fig. 12. Overlay of 5-HT_3 antagonists. Structure **37** is coloured by atom type, **44** is in purple, **45** is cyan and **47** is orange. The elucidated pharmacophore is indicated by yellow circles.

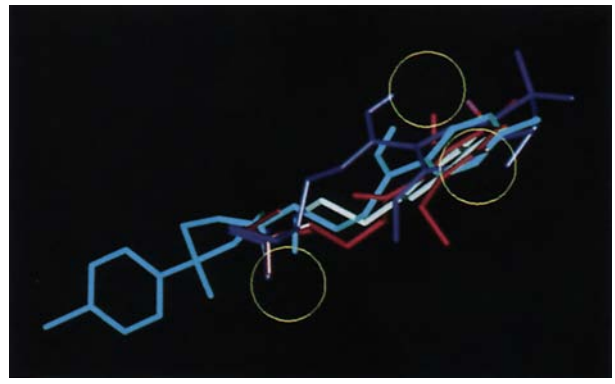


Fig. 14. Overlay of dopamine ligands. Dopamine is shown coloured by atom type, amisulphride is purple, haloperidol is cyan and raclopride is red. The elucidated pharmacophore is indicated by yellow circles.

Overlay of three 5-HT_{1D} agonists

Potent 5-HT_{1D} agonists have been tested as migraine therapeutic agents [21], and an overlay was attempted of three such structures, as shown in Fig. 7. It was found that the GA often took a long time to converge in this problem, and the parameter MAXOPS was accordingly set to 60 000. Even so, the run time of the GA was only 7 min and 44 s when averaged over the 10 runs that were carried out.

All but three of the GA solutions (specifically, the worst, second worst and fourth worst ranked solutions) identified a pharmacophore comprising an aromatic ring, a protonated nitrogen and an oxygen acceptor group. The solution obtained by the best run is shown in Fig. 8, with the functional groups comprising the pharmacophore indicated by yellow circles. For clarity, the lone pairs

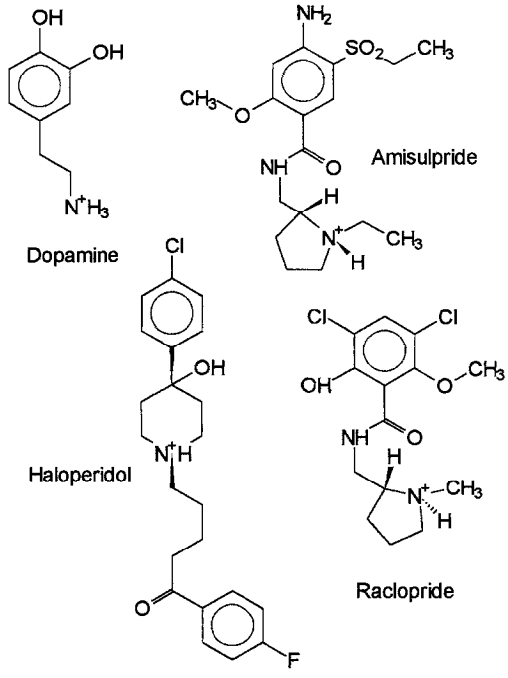


Fig. 13. Flexible dopamine agonists.

connected to the oxygen acceptor group and the proton connected to the charged nitrogen are displayed. The overlay of oxygen acceptor atoms comprises sp^2 oxygens from 311c90 and sumatriptan and an sp^3 donor–acceptor oxygen from methysergide (sp^2 and sp^3 oxygen acceptors have a high acceptor atom-type similarity of 0.9). The second best solution overlaid sp^2 oxygens from all molecules, but the superimposition of the virtual points associated with the lone pairs was not as good. Nevertheless, the pharmacophore suggested by this second best solution is the same as that identified by Glen et al. [21]. In both the top two overlays, methysergide follows a different path compared to the other two molecules to reach the hydrogen-bond acceptor side. This is not unreasonable, in that experiments have indicated that a portion of the indole group in 311c90 occupies a selective volume that reduces the affinity of the drug for 5-HT_2 receptors [21]. Methysergide, however, shows no selectivity for 5-HT_1 (over 5-HT_2) receptors.

Overlay of three benzodiazepine ligands

Codding and Muir [22] have produced an overlay of ligands that bind to the benzodiazepine receptor, using structure–activity studies and functional group similarities. Three of the ligands used in their analysis are shown in Fig. 9; two of these molecules are inverse agonists (which promote convulsions on binding), while Ro15-1788 is an antagonist that has no convulsant effects. With the exception of MAXOPS, which was set to 30 000, the default parameters were used. The mean execution time of the GA was 2 min and 10 s.

The best solution obtained from the 10 runs is shown in Fig. 10, where the base molecule, CGS-8216, is shown coloured by atom type, Ro15-1788 is in orange and methyl-β-carboline-3-carboxylate is shown in purple. The pharmacophore elucidated by the GA comprised a benzene ring and an sp^2 oxygen acceptor; these features are indicated in the figure by yellow circles and the lone pairs

connected to the sp^2 oxygens are also displayed. Also of interest is the fact that the donor nitrogens in the two inverse agonists are closely positioned (this is again indicated by a circle and the donor hydrogens are also displayed). It has been suggested that an absence of this donor is required for antagonism [22]. All of the other GA runs produced this overlay, although the worst run produced a very untidy fit.

The GA solution is in fair agreement with Codding and Muir's structure–activity studies, which predicted a binding site that recognises four features: an aromatic ring, an sp^2 oxygen, a hydrophobic side chain and an N-H donor group, although this last feature is not present in Ro15-1788. The GA does not identify hydrophobic regions, but the volume overlay term ensured that the side chains comprising this feature were correctly overlaid.

Payne and Glen obtained poor results when fitting Ro15-1788 onto the β-carboline, using electrostatic and shape constraints [4]. The GA described here has a simpler fitness function, which drives it to elucidate pharmacophores explicitly and has yielded much better results on this data set.

Overlay of four 5-HT_3 antagonists

Clark et al. have synthesised several series of *N*-(quinuclidin-3-yl)aryl and heteroaryl-fused pyridones and tested them for 5-HT_3 receptor affinity [23]. An overlay was attempted of the four antagonists shown in Fig. 11. The default parameters were used, with the mean run time being 6 min and 9 s.

The GA elucidated a pharmacophore consisting of a nitrogen donor, an sp^2 oxygen acceptor and an aromatic ring in all but the least fit run. Figure 12 shows the superimposition obtained by the GA run that generated the highest fitness score. The base molecule, structure **37**, is shown coloured by atom type, **44** is coloured purple, **45** is cyan and **47** is orange. The yellow circles in Fig. 12 indicate the three pharmacophore points, i.e., the normals of the aromatic ring, the lone pairs of the sp^2 oxygen and the donor hydrogens bonded to the nitrogens. Although the nitrogens are not overlaid, their donor hydrogens are clearly in a position to interact with the same point in the receptor. This pharmacophore is the same as that identified by Clark et al. [23], although the centres of the aromatic rings in their overlay do not appear to be as close as in the GA solutions.

Using NMR and X-ray studies, Bradley et al. have shown that structure **45** probably binds with the central amide bond in a cis configuration [24]. This conformation is stabilised by internal hydrogen bonds and gives a better overlay of basic nitrogens. Unfortunately, the GA does not currently include a term for internal hydrogen bonding. Furthermore, as there is a large barrier to rotation in an amide bond, this bond is not considered as being flexible by the GA. In the experiment described above, the overlays elucidated by the GA had the central amide bond in a trans configuration, as this was the configuration of the input structure. The experiment was repeated using structure **45** with the amide bonds in the configuration described in Ref. 24 and the GA elucidated the same pharmacophore as before, with little of the expected improvement in the grouping of the protonated nitrogens. It would be a simple matter to extend the chromosome encoding for those cases when it is not known whether or not a bond is in a trans or cis conformation. For example, a bit in the binary bit string that encodes molecular conformation could be set to 1 to encode the fact that a given bond should be in the trans configuration and it could be set to 0 if the bond should be cis. Thus, in the initial randomly generated population half the overlays would contain structures with the bond in a cis configuration and half would have the bond in a trans configuration. The same mechanism could be used to encode molecules of unknown chirality.

Overlay of four dopamine agonists

The algorithm was next applied to four flexible dopaminergic ligands that have been described by Höberg and Norinder [25]. These molecules (the dopamine receptor agonist, dopamine, and three antipsychotic dopamine D_2 receptor antagonists) are shown in Fig. 13. Using the default parameters, the mean run time for the GA was 6 min and 51 s.

This problem was particularly deceptive for the GA, as most runs generated pharmacophores comprising an aromatic ring and oxygen sp^3 groups. However, the best solution generated a pharmacophore comprising an aromatic ring and the protonated nitrogen, and an overlay of sp^3 oxygen donors was also observed. This overlay is shown in Fig. 14, where the base molecule dopamine is coloured by atom type, amisulpride is purple, haloperidol is cyan and raclopride is red. The pharmacophore points are indicated by yellow circles and the hydrogen-bond donor proton attached to the protonated nitrogen is displayed for clarity, as is the lone pair attached to the sp^3 oxygen.

The fact that the best overlay appeared only once indicates a failure by the GA to sample the whole search space. The runs were hence repeated with POPSIZE set to 800 and MAXOPS to 80 000, with the selection pressure being maintained at 1.1. These changes resulted in an increase in the mean run time (which was still just 10 min and 14 s) and in not only the best but also the two next-best runs managing to overlay the aromatic ring and the protonated nitrogen (together with sp^3 oxygen groups). Thus, the larger population size and the greater number of operations appeared to increase the reliability of the GA in this case, without an unacceptable increase in the computational requirements.

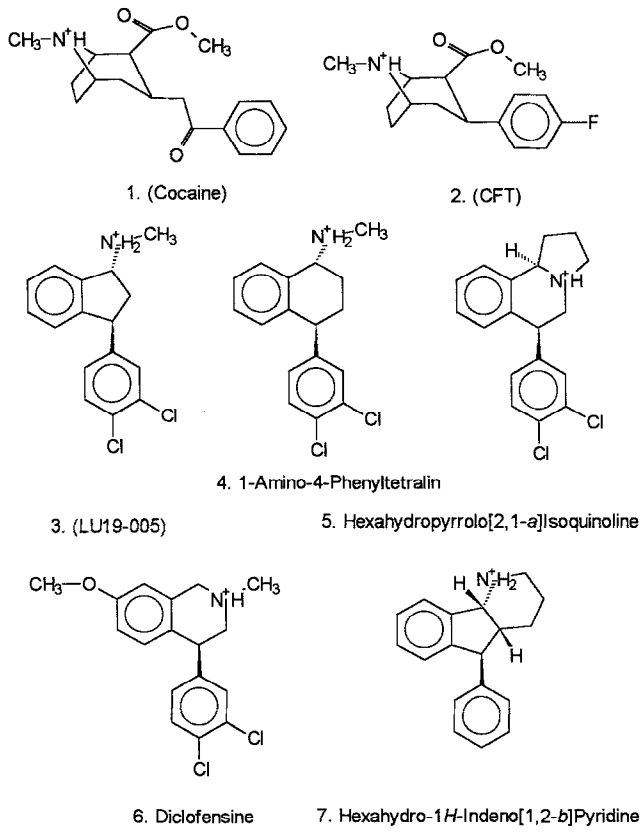


Fig. 15. Dopamine reuptake blockers (from Ref. 26).

Overlay of seven dopamine reuptake inhibitors

Froimowitz has performed conformational analysis on a set of seven dopamine reuptake inhibitors that probably have a common mode of action [26]. These ligands are shown in Fig. 15. The GA was run with the default parameters, and the mean run time was 5 min and 50 s.

All 10 runs yielded a pharmacophore containing an aromatic ring and a protonated nitrogen. However, the fitter runs also succeeded in overlaying the second aromatic ring that occurs in structures **3**–**7**. Figure 16 shows the overlay generated by the best GA run. This superimposition was exceptional, in that the volume overlay was much better than that observed in any of the other runs. Structure **7**, the base molecule, is shown coloured by atom type, **1** is purple, **2** is orange, **3** is cyan, **4** is green, **5** is red and **6** is yellow. Structures **7** and **3** are overlaid exactly. For clarity, the donor hydrogens connected to the protonated nitrogen are displayed and the pharmacophore points are indicated by yellow circles.

Froimowitz carried out superimposition studies on structures **3**–**7** to define a pharmacophore based on the aromatic rings and the protonated nitrogen, which was in excellent agreement with the pattern suggested by the GA overlay. It was also discovered that the superimposition of **2** on the pharmacophore placed the ester methoxy group on the second phenyl ring of structures **3**–**7** [26]. Froimowitz's analysis of **1** and **2** showed that the preferred conformers contained an electrostatic interaction between the ammonium hydrogen and carbonyl oxygen of the ester group. This effect may highlight the need to include a term for internal hydrogen bonding in the GA fitness function (as was observed above).

Overlay of six angiotensin II receptor antagonists

Perkins and Dean have described a novel strategy for the superimposition of a set of flexible molecules, using a combination of simulated annealing and cluster analysis [27]. The conformational space of each molecule is searched using simulated annealing. Significantly different low-energy conformations are extracted from the conformational analysis history using cluster analysis. For each pair of molecules, every possible combination of conformations found by cluster analysis is matched by simulated annealing, using the difference distance matrix as the objective function. The molecules are then superimposed using the match statistics, either by reference to a base molecule or by a consensus method. The algorithm was tested on the six angiotensin II antagonists shown in Fig. 17, and we have also used these structures to evaluate the GA. The butyl side chains in five of the six structures were replaced by methyl groups, as this simplification was also performed in Ref. 27. As the GA proved slow to converge, the parameter MAXOPS was set to 60 000. The average run time of the GA was 7 min and 56 s.

Six of the 10 overlays (including the three fittest solutions) generated a pharmacophore comprising an aromatic ring and a protonated nitrogen. Inspection of the structures in Fig. 10 shows that these are the only features common to all compounds. The best overlay is shown in Fig. 18. The base molecule, L-158809, is shown coloured by atom type, GLAXO is magenta, SEARLE is orange, SKB 108566 is cyan, TAK is green and DuP 753 is yellow. The yellow circles indicate the two pharmacophore points and the hydrogen-bond donor proton bonded to the protonated nitrogen is also displayed. It was hoped that the GA might have been able to overlay an acidic group from each structure. However, this was not the case, although it was able to overlay acid groups from all structures except SKB 108566 (which is the most dissimilar structure from L-158809). This overlay is indicated by a yellow circle and includes all the tetrazol groups. The fact that this overlay of five acidic groups is present in the final solution suggests that there may have been chromosomes within the population that encoded an overlay of acidic groups from all structures.

The superposition obtained by Perkins and Dean [27] has some similarity with the GA result, in that the imidazole group and benzene rings are also successfully superimposed (due to their structural similarity). However, their procedure is far more time-consuming. Conformational analysis took about 8 min (for SKB 108566), while the pairwise matching process took about 6 h for

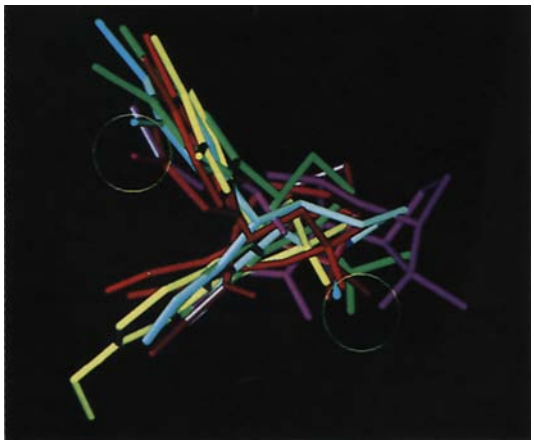


Fig. 16. Overlay of dopamine reuptake blockers. Structure **7** is coloured by atom type, **1** is purple, **2** is orange, **3** is cyan, **4** is green, **5** is red and **6** is yellow. The elucidated pharmacophore is indicated by yellow circles.

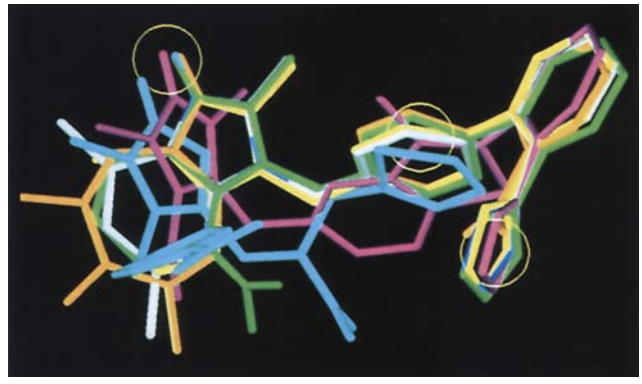


Fig. 18. Overlay of six angiotensin II receptor antagonists. The base molecule L-158809 is shown coloured by atom type, GLAXO is coloured magenta, SEARLE is orange, SKB 108566 is cyan, TAK is green and DuP is yellow. Points of interest are indicated by yellow circles (see the text for discussion).

each pair of molecules (based on the time used for matching all conformers of DuP 753 and SKB 108566). These times are for a Sun SPARCstation IPX workstation (a CPU that is comparable to that used for the GA): a superposition of the six structures, by reference to a base molecule, should thus take about 31 h of CPU time, with a consensus superposition taking considerably longer.

Overlay of two FKBP12 ligands

The immunosuppressant FK506 is a microbial product that blocks T-cell activation and proliferation and binds strongly to the human immunophilin FKBP12. The three-dimensional structure of FK506 has been determined by X-ray crystallography [28] and the co-ordinates of the bound complex have been deposited in the Brookhaven Protein Data Bank (PDB) [29]. Holt et al. have described the design, synthesis and evaluation of a number of high-affinity FKBP12 ligands [30]. The atomic structures of three FKBP12–ligand complexes were determined by X-ray crystallography and deposited in the PDB. The structures of FK506 and compound **9** from Ref. 30 are shown in Fig. 19.

The crystal structures of FK506 and compound **9** complexed with FKBP12 were extracted from the PDB. The structure of FKBP12 was found to be highly similar on binding to each ligand (using SYBYL [9], the two proteins were superimposed and the mean rms deviation in their atomic co-ordinates was found to be 1.1 Å). Both ligands were extracted from the crystal structures, hydro-

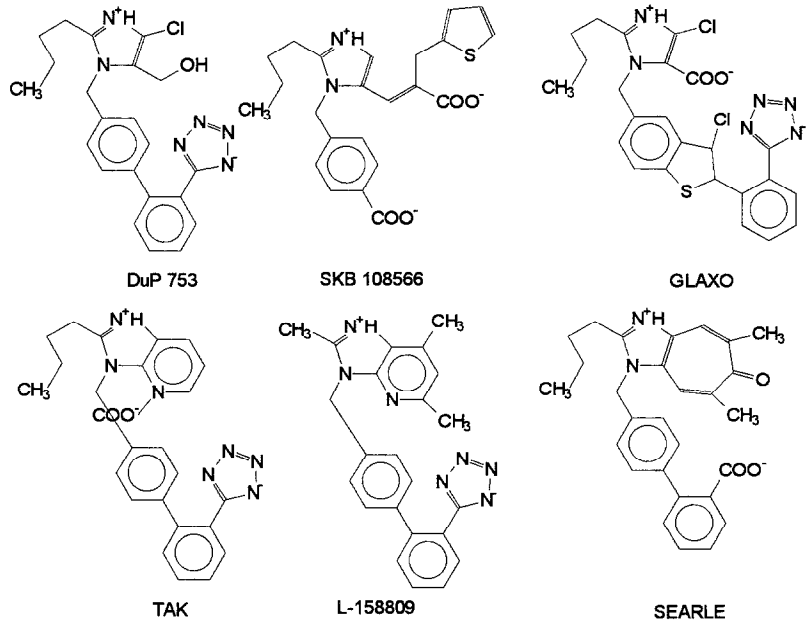
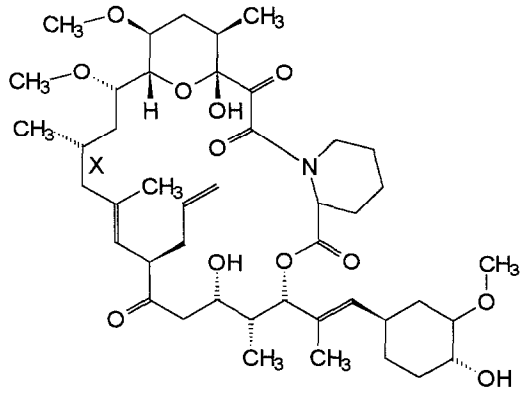


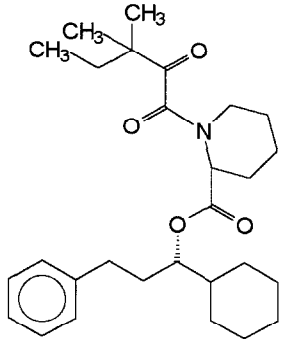
Fig. 17. Six angiotensin II antagonists (structures from Ref. 27).

gens were added and the structures were minimised using SYBYL molecular mechanics [9]. The GA was then used to superimpose the ligands and to elucidate a pharmacophore.

Inspection of Fig. 19 shows that FK506 contains a large cyclic region. The technique of bond breaking and ring closure to allow GAs to perform conformational analysis of cyclic regions has been reported previously [31]. In order to take account of the flexibility of FK506, the bond labelled B in Fig. 19 was broken and single bonds within the previously cyclic region were allowed to rotate. In order to ensure closure of the ring, constraints were applied. Figure 20 illustrates the use of these constraints. Using the initial input structure, the bond between A and B was deleted. Two dummy atoms, D1 and D2, were then created in such a way that D1 was positioned on B and bonded to A and D2 was positioned on A and bonded to B. Selection of rotatable bonds then proceeded as described in Methods. On decoding a chromosome, the distances d1 (the distance between D2 and A) and d2 (the distance between D1 and B) were determined. The fitness of the chromosome was then reduced



FK506



Compound 9

Fig. 19. Two ligands of the human immunophilin FKBP-12 receptor. FK506 is a natural substrate and compound **9** (from Ref. 30) is a high-affinity ligand. The bond labelled X in FK506 was broken during superposition (see text for discussion).

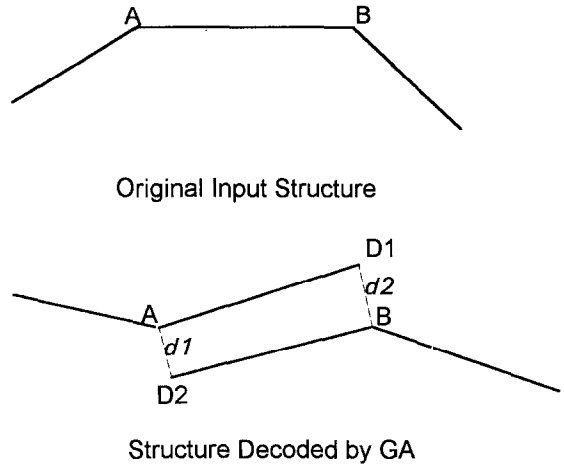


Fig. 20. Replacement of a single bond (between A and B) by two dummy atoms (D1 and D2) and two constraints (d1 and d2) for the conformational analysis of large cyclic regions.

by $100\times((d1)^2+(d2)^2)$ on the basis of empirical parameterisation. This correction to the fitness score had the effect of reducing d1 and d2 and thus closing the ring correctly.

This problem proved to be very demanding, with a total of 46 rotatable bonds (33 in FK506 and 13 in compound **9**). Accordingly, the parameter MAXOPS was set to 100 000 and a single population of 800 individuals was used. The average run time of the GA was 30 min and 15 s; this long run time is a consequence of the large number of rotatable bonds.

The best overlay is shown in Fig. 21. The yellow circles indicate the elucidated pharmacophore, which comprised all the oxygen atoms from **9**. The two ring-closure constraint distances, d1 and d2, in FK506 were found to be 0.23 and 0.27 Å. In order to verify the effectiveness of the ring-closure procedure, SYBYL was used to recreate the bond and the resulting structure was minimised. The rms

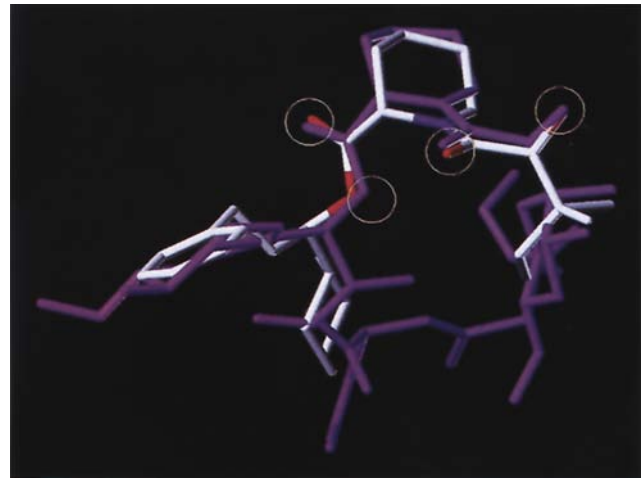


Fig. 21. Overlay of compound **9** on FK506. The base molecule, **9**, is shown coloured by atom type and FK506 is shown in purple.

deviation of heavy atoms in the minimised structure from the GA solution was found to be only 0.44 Å. Using only the oxygen pharmacophore atoms elucidated by the GA, the crystallographic co-ordinates of FK506 were fitted onto the GA solution. Following fitting, the rms deviation for the pharmacophore atoms in the GA solution from those in the crystal structure was only 0.38 Å, while the rms deviation for all atoms was found to be 3.35 Å. The procedure was repeated for compound **9**. Following fitting, the rms deviation in pharmacophore points was found to be only 0.34 Å and the rms deviation over all atoms was 2.67 Å.

This is a good result for the GA, as the pharmacophore observed by the GA is identical to that seen in the crystal structure. Considering all atoms in the ligands, the deviation from the crystal structure is not too bad, especially when one considers that large portions of each ligand are solvent exposed and thus fairly mobile. One interesting difference between the GA-predicted overlay and the crystallographically observed binding modes is the superposition of the benzene ring in compound **9** on the cyclohexane ring in FK506. In the GA superposition, the rings lie on top of each other; in the crystal structures the planes of the rings make an angle of 90°. However, this must be regarded as a limitation of the overlay problem, since it is hard to see how, in the absence of any knowledge of the receptor, any algorithm could predict this fit.

Conclusions

The design and implementation have been described of a GA for the superimposition of flexible molecules and the use of the resulting overlays to suggest possible pharmacophoric patterns. The experiments reported here demonstrate the effectiveness and versatility of the algorithm, in that it has been possible to superimpose flexible molecules on structurally diverse test systems with results that are both intuitively acceptable and often in agreement with overlays suggested by alternative means; this said, there are several additions and improvements that could be made to the program and these are currently being investigated. Inactive compounds that are similar to known actives are often incorporated in a structure–activity analysis. It is possible to extend the GA to incorporate inactives or to include biological activity. As the algorithm attempts to find pharmacophore points that are common to all input structures, a current limitation of the fitness function is that the GA will have difficulty incorporating molecules that do not fit the pharmacophore for some reason. Further simple improvements that could be made to the algorithm include encoding with the chromosome for molecules of unknown chirality and the addition of internal hydrogen bonding.

The use of the GA to elucidate possible pharmacophoric patterns has been emphasised in this paper. However, other applications of the approach are equally feasible. The overlays may be used as a starting point for investigation of a data set by 3D QSAR, which requires an initial alignment of the molecules that are to be analysed [32]. The potential of this approach is well illustrated in a recent paper by Calder et al. [33], who have used an overlay program to generate the alignments for a CoMFA analysis of six classes of compounds that block GABA receptors. Another possible application is similarity searching in 3D databases [34]. Current similarity searching systems are normally based on the use of fragment occurrence data or interatomic distances, since superimposition-based procedures are generally far too demanding of computational resources for use in a database context, even if attention is restricted to rigid 3D structures [35]. The speed of the GA when performing pairwise superimpositions, which is what is required to match a target structure against each of the structures in a database, is such as to suggest that it might be feasible to consider the use of a modified version of the algorithm for flexible 3D similarity searching. We hope to investigate these possibilities in the future.

Acknowledgements

We thank the referees for their comments on the initial draft of the paper, the Science and Engineering Research Council and The Wellcome Foundation for funding, and Tripos Inc. for hardware and software support. This paper is a contribution from the Krebs Institute for Biomolecular Research, which is a designated Centre for Biomolecular Sciences of the Biotechnology and Biological Sciences Research Council.

References

1 Greer, J., Erickson, J.W., Baldwin, J.J. and Varney, M.D., J. Med. Chem., 37 (1994) 1035.
2 Marshall, G.R., Barry, C.D., Bosshard, H.E., Dammkoehler, R.A. and Dunn, D.A., In Olson, E.C. and Christofferson, R.E. (Eds.) Computer Assisted Drug Design, American Chemical Society Symposium Series, Vol. 112, American Chemical Society, Washington, DC, 1979, pp. 205–226.
3 Klebe, G., In Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993, pp. 173–199.
4 Payne, A.W.R. and Glen, R.C., J. Mol. Graphics, 11 (1993) 74.
5 Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley, Wokingham, 1989.
6 Davis, L. (Ed.) Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, NY, 1991.
7 Forrest, S., Science, 261 (1993) 872.
8 Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245 (1995) 43.
9 SYBYL molecular modelling software, available from Tripos Associates Inc., St. Louis, MO.
10 Zamora, A., J. Chem. Inf. Comput. Sci., 16 (1976) 40.
11 Digby, P.G.N. and Kempton, R.A., Multivariate Analysis of Ecological Communities, Chapman and Hall, London, 1987, pp. 112–115.

12 Hirschfelder, J.O., Curtiss, C.F. and Bird, R.B., Molecular Theory of Gases and Liquids, Wiley, New York, NY, 1964.
13 Clark, M., Cramer III, R.D. and Van Opdenbosch, N., J. Comput. Chem., 10 (1989) 982.
14 Brown, R.D., Jones, G., Willett, P. and Glen, R.C., J. Chem. Inf. Comput. Sci., 34 (1994) 63.
15 Starkweather, T., Whitley, D. and Mathias, K., In Schwefel, H.P. and Manner, R. (Eds.) Parallel Problem Solving From Nature, Springer, Berlin, 1990, pp. 176–185.
16 Tanese, R., In Schaffer, D. (Ed.) Proceedings of the Third International Conference on Genetic Algorithms and their Applications, Morgan Kaufmann, San Mateo, CA, 1989, pp. 434–439.
17 Jones, G., Ph.D. Thesis, University of Sheffield, Western Bank, 1995.
18 Bradbury, R.H., Allott, C.P., Dennis, M., Fisher, E., Major, J.S., Masek, B.B., Oldham, A.A., Pearce, R.J., Rankine, N., Revill, J.M., Roberts, D.A. and Russell, S.T., J. Med. Chem., 35 (1992) 4227.
19 Masek, B.B., Merchant, A. and Matthew, J.B., J. Med. Chem., 36 (1993) 1230.
20 Kolb, V.M., Prog. Drug Res., 36 (1991) 49.
21 Glen, R.C., Hill, A.P., Martin, G.R. and Robertson, A.D., Headache, 34 (1994) 307.
22 Codding, P.W. and Muir, A.K.S., Mol. Pharmacol., 28 (1985) 178.
23 Clark, R.D., Miller, A.B., Berger, J., Repke, D.B., Weinhardt, K.K., Kowalczyk, B.A., Eglen, R.M., Bonhaus, D.W., Lee, C., Michel, A.D., Smith, W.L. and Wong, E.H.F., J. Med. Chem., 36 (1993) 2645.
24 Bradley, G., Ward, T.J., White, J.C., Coleman, J., Taylor, A. and Rhodes, K.F., J. Med. Chem., 25 (1992) 1515.
25 Höberg, T. and Norinder, U., In Krogsgaard-Larsen, P. and Bundgaard, H. (Eds.) A Textbook of Drug Design and Development, Harwood Academic Publishers, Reading, 1992, pp. 55–91.
26 Froimowitz, M., J. Comput. Chem., 14 (1993) 934.
27 Perkins, T.D.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 7 (1993) 155.
28 Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L. and Clardy, J., J. Mol. Biol., 229 (1993) 105.
29 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, F., Bryce, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
30 Holt, D.A., Luengo, J.I., Yamashita, D.S., Oh, H., Konialian, A.L., Yen, H., Rozamus, L.W., Brandt, M., Bossard, M.J., Levy, M.A., Eggleston, D.S., Liang, J., Schultz, L.W., Stout, T.J. and Clardy, J., J. Am. Chem. Soc., 115 (1993) 9925.
31 Sanderson, P.N., Glen, R.C., Payne, A.W.R., Hudson, B.D., Heide, C., Tranter, G.E., Doyle, P.M. and Harris, C.J., Int. J. Pept. Protein Res., 43 (1994) 588.
32 Kubinyi, H. (Ed.) 3D QSAR in Drug Design: Theory, Methods and Applications, ESCOM, Leiden, 1993.
33 Calder, J.A., Wyatt, J.A., Frenkel, D.A. and Casida, J.E., J. Comput.-Aided Mol. Design, 7 (1993) 45.
34 Dean, P.M. (Ed.) Molecular Similarity in Drug Design, Blackie, Glasgow, 1994.
35 Downs, G.M. and Willett, P., Rev. Comput. Chem., in press.