J-CAMD 287

Flexible ligand docking using a genetic algorithm

C.M. Oshiro[a], I.D. Kuntz[a,*] and J. Scott Dixon[b,*]

[a]*Department of Pharmaceutical Chemistry, School of Pharmacy, University of California,
San Francisco, CA 94143-0446, U.S.A.*
[b]*Department of Physical and Structural Chemistry, SmithKline Beecham Pharmaceuticals, P.O. Box 1539,
King of Prussia, PA 19406-0939, U.S.A.*

Received 29 August 1994
Accepted 21 November 1994

Keywords: Genetic algorithm; DOCK; Flexible conformational search; Protein–ligand interactions; Dihydrofolate reductase; Thymidylate synthase; HIV protease

Summary

Two computational techniques have been developed to explore the orientational and conformational space of a flexible ligand within an enzyme. Both methods use the Genetic Algorithm (GA) to generate conformationally flexible ligands in conjunction with algorithms from the DOCK suite of programs to characterize the receptor site. The methods are applied to three enzyme–ligand complexes: dihydrofolate reductase–methotrexate, thymidylate synthase–phenolpthalein and HIV protease–thioketal haloperidol. Conformations and orientations close to the crystallographically determined structures are obtained, as well as alternative structures with low energy. The potential for the GA method to screen a database of compounds is also examined. A collection of ligands is evaluated simultaneously, rather than docking the ligands individually into the enzyme.

Introduction

Many drugs in clinical use today have been discovered by random screening of either natural products or corporate collections of compounds. In recent years, alternative methods for identification of lead compounds have been explored, including mechanism-based inhibitors and structure-based drug design [1,2]. This last method has been of increasing interest because of the availability of high-resolution structures of enzymes of critical metabolic pathways. With these structures, computer-based methods can be used to identify or design ligands that possess good structural and chemical complementarity to various sites on the enzyme. Structure-based design covers many strategies, including both manual and automated approaches. While progress has been made [1], a core problem is searching the very large conformational space of typical ligand–enzyme complexes.

Many different methods have been developed to help solve the conformational search problem. These methods can be divided roughly into two classes: the search for new molecules with properties similar to a known ligand, and the search for new molecules with properties complementary to the enzyme. That is, one approach finds ligands satisfying intramolecular constraints; the other finds ligands satisfying intermolecular constraints.

The first approach concentrates on finding ligands which contain certain pharmacophore properties or low-energy conformations. Numerous methods exist. For example, various AI methodologies, such as the programs WIZARD [3] and COBRA [4], have been used to search for low-energy conformations. Clark et al. recently compared several different methods of searching the conformational space of small to medium-sized molecules [5], including the systematic search method of Dammkoehler [6], the directed tweak approach described by Hurst [7], and the genetic algorithm [8–10]. They concluded that the last two methods are very fast and show the most promise for such conformational searching.

The second approach, which we consider in this paper,

*To whom correspondence should be addressed.
Abbreviations: GA, genetic algorithm; dhfr, dihydrofolate reductase; mtx, methotrexate; ts, thymidylate synthase; fen, phenolpthalein; HIV, human immune deficiency virus; hivp, HIV protease; thk, thioketal haloperidol.

concentrates on finding ligands with structural and chemical complementarity to the enzyme. One strategy is to examine a very limited set of favorable conformations, one at a time. In this treatment, the enzyme and ligand are taken as rigid objects and the search is reduced to finding energetically or geometrically favorable configurations of the ligand within the active site of the enzyme. Many of these methods have been applied primarily to docking of macromolecules. Examples of these rigid-body search methods include the work of Kuntz et al. [11], Bacon and Moult [12], Lawrence and Davis [13], Nussinov et al. [14,15], and Kasinos et al. [16]. Even with the rigid-body restriction, the number of possible ligand orientations is enormous and the computational problem belongs to the class known as 'nondeterministic polynomial time NP complete' problems [17].

Another strategy, often referred to as de novo drug design, builds molecules within the rigid enzyme. Fragments of molecules are docked separately into the enzyme, then linked together. Alternatively, molecules are grown in a stepwise fashion or in a random manner. Examples of these different de novo design methods include the work of Lewis [18], the programs LUDI [19,20] and GroupBuild [21,22], and the genetic programming methodology [23]. Structure generation is a computationally intensive task [24] and it is often difficult to properly orient fragments for tight binding [21].

A third strategy is to dock conformationally flexible ligands. This is an extension of the first approach and has the potential to identify a greater number and variety of known ligands. Monte Carlo methods have been used [25–29] as well as a combination of the program DOCK and systematic search [30], or a restricted search with AMBER [31,32] refinement [33]. These efforts are computationally expensive.

In this paper, we evaluate the usefulness of the genetic algorithm (GA) to dock conformationally flexible ligands into a rigid enzyme. The long-term objective of this work is to develop a method to computationally screen and evaluate conformationally flexible ligands as potential lead compounds. We are seeking a method that is both rapid and reasonably accurate. However, unlike most applications of GA thus far which explored intramolecular constraints [8–10], we search the conformational and configurational search space of known small molecules within the active site of an enzyme.

The GA method has been used to solve problems involving large search spaces, where traditional optimization methods are less efficient [34–36]. It is based on ideas borrowed from genetics and natural selection. A population of 'chromosomes' encodes solutions – good or bad – to the problem at hand. The population 'evolves' through a process loosely analogous to biological evolution, including simplified implementations of natural selection and genetic crossover. Chromosomes encoding good partial solutions survive, reproduce and combine to produce chromosomes hopefully encoding good global solutions in succeeding generations. The strength of GA lies in its ability to handle a large and diverse set of variables. For example, it has been used to optimize a function without the use of gradients [34,36] and in the conformational analysis of DNA structures [34]. In designing GA to solve a specific problem, however, it can be difficult to find an appropriate representation of the variables and a measure of fitness. Some classes of problems have been found to be 'GA deceptive' [35,36], where good partial solutions lead to poor global solutions and are difficult to solve using GAs.

In our application of GA, each chromosome represents a ligand in a particular orientation and conformation; the orientational and conformational space available within the active site of a known enzyme structure is sampled by a diverse population of chromosomes. The 'fitness' of a chromosome is taken to be the ligand–enzyme binding energy, approximated as the sum of electrostatic and van der Waals (vdW) interaction energies. GA allows ligands with improved orientation and conformation to be built, over time, by combining information from previous generations.

GA can be applied in different ways to the conformational search and structure-based design problem. Here, only two tasks are considered: (i) the reproduction of the ligand–protein structure, to test the accuracy of the method; and (ii) an initial study of screening a compound database for ligands complementary to the active site of an enzyme.

We first illustrate the potential of the GA method for flexible docking by using it to generate different orientations of a rigid ligand in a receptor. Specifically, the GA method replaces the sphere–atom matching step in the DOCK algorithm (which docks rigid conformations). This step is computationally difficult and is currently done in DOCK by variants on clique matching techniques [11,37].

We then evaluate the performance of the GA method in flexible docking by attempting to (re)generate the known ligand crystal conformation and orientation of three enzyme–inhibitor complexes: dihydrofolate reductase–methotrexate (dhfr–mtx), thymidylate synthase–phenolpthalein (ts–fen), and HIV-1 aspartic protease–thioketal haloperidol (hivp–thk). Two different variations of the method are used. One variant samples widely within the orientational and conformational space of the enzyme–inhibitor complexes; the other explores in detail the conformational space within a more restricted orientational space. We compare the conformations and orientations found by GA to the true solution, using a root-mean-squared deviation per atom (rmsd). We compare the two GA variants, and we also compare the GA solutions to the structures provided by various other existing conformational search methods.

We next examine the potential of one of the GA methods for screening large databases of compounds. Note that the primary objective of a computational database screen is to rapidly identify lead compounds for further analysis and not to find a unique binding conformation of the ligand (although that would be highly desirable). Since our scoring function approximates the enzyme–ligand binding energy, a low energy indicates that we may have identified a potential ligand. Hence, in the second part of this work, we examine the extent to which low-energy structures are generated, without considering the actual binding orientation of the ligand.

We first apply the GA method to the same three test systems, but using smaller population sizes and shorter computational times. Next, GA is used to evaluate a collection of different ligands of known binding affinities. Rather than docking ligands separately and individually into the known structure of an enzyme, the collection of ligands is evaluated simultaneously. This (computational) idea is inspired by the (experimental) combinatorial synthesis-screening methods now used in drug discovery [38–41]. In combinatorial chemistry, a large variety of compounds are created simultaneously and tested for activity; only those compounds that bind to an enzyme are kept and identified [38–41]. In our GA application, the different ligand 'species' are combined into one large population. Competition and cross-breeding in successive GA generations can then efficiently generate low-energy structures.

Methods and Procedures

This section is divided into three parts: descriptions of (i) the computational procedures; (ii) the test systems used in this work; and finally (iii) the computational experiments and parameters themselves.

Computational procedures

Overview

We first give an outline of the procedures, then describe the various techniques in more detail below. The starting point of all our calculations is the crystal structure of an enzyme from an enzyme–inhibitor complex. The basic ligand structure is also presumed to be known (i.e., the atomic connectivities), but the actual binding conformation of the ligand, described by a set of rotatable bonds, is taken as unknown. Both the binding conformation and the orientation of the binding conformation of the ligand within the active site of the enzyme are to be determined.

We explore two major variations in the use of the GA method for flexible docking. Both variants have certain features in common. In both cases, GA is used to generate and optimize different conformations and orientations of the ligand. Both variants evaluate the different conformations and orientations with a molecular mechanics interaction energy [42], using a scoring grid.

The primary difference between the two variants is the manner in which the orientation of the ligand is represented and varied. We call the first approach *sphere-based*, because GA is used to match ligand atoms with spheres (characterizing the shape of the receptor site) to implicitly orient the ligand. This approach explores both the conformational and orientational space within the docking site; it provides broad sampling of a complex system.

We call the second approach *explicit-orientation-based*, since it directly specifies the orientation of the ligand within the receptor site with a translation vector and three Euler angles. This approach consists of two main parts: restricting the orientation of the ligand within the active site and GA optimization of the ligand orientation and conformation within the site. It provides detailed ligand conformational sampling within a restricted portion of the active site.

Genetic algorithm: General description

The genetic algorithm is used to generate and optimize low-energy conformations and orientations of the ligand. The procedure is often used to search for global minima in a large parameter space. GA is characterized by the unique manner in which the parameter set is represented and altered. GAs have been used increasingly in a wide range of optimization problems [34–36]. The general method is described in detail elsewhere (Ref. 36 and references cited therein). We give an outline of the method here, then tailor the description to our particular implementation later.

In GA, the variables to be optimized are encoded as a sequence of bits in a bit string or a string of numbers in a vector. Each variable is referred to as a gene and the string containing all genes as a chromosome or genome. An initial population of chromosomes (strings) is generated, each chromosome in the population representing a different set of values for the variables. Each chromosome in the population is evaluated using an appropriate scoring scheme to measure how well it solves the problem under consideration. Chromosomes with good scores are selected for propagation into a new generation. Chromosomes with poor scores perish. The best performing chromosome from one generation is frequently saved and survives unchanged into the next generation (the elitist strategy). A portion of the better chromosomes are copied directly into the next generation. Another portion of the better chromosomes also 'mate' by a procedure called 'crossover' to produce offspring, with a mixture of genes from both parents. In this procedure, a pair of chromosomes is lined up, a 'crossover' point along the chromosome is selected, and portions of the bit string beyond the crossover point in the chromosome are swapped. Occasionally, mutations are also produced, by a random alteration of single bits within the bit string. The population size of each generation remains constant. With succeeding generations, better scoring chromosomes are formed. A sample of the best scoring chromosomes from all generations is stored for further analysis. The ability of GAs to form improved solutions is a result of their ability to combine strings containing partial solutions (for a detailed discussion of this aspect, see Goldberg [36] and Holland [43]).

The principal parameters of the method are the number of runs (experiments), population size, number of generations of evolution, crossover rate, mutation rate, and survival rate. Large numbers of runs and large population sizes increase the likelihood of good solutions. The generation number determines, in part, the extent to which the population has converged to a particular solution or range of solutions. The crossover, mutation and survival rates all affect the rate of convergence to good solutions. With too high a crossover or mutation rate the combination of good scoring strings is disrupted. In the limit of high crossover and mutation rates, GA becomes a random search. With too low a survival rate, diversity in the population is easily lost and the system can converge prematurely to poor solutions. With too low a crossover, mutation, or survival rate, the search space is sampled inefficiently, and the time needed to reach good solutions increases.

In our application of GA, each chromosome represents a ligand in a particular conformation and orientation. Each chromosome generally has two different parts, one representing the conformation of the ligand and another representing the orientation of the ligand. The conformation of the ligand is represented by a set of torsion angles about its rotatable bonds. The orientation of the ligand is represented in different ways, depending upon the GA variant used. A detailed description of the two different variations of GA is given in the next section. In both cases, the survival 'fitness' of each chromosome – ligand orientation and conformation – is primarily an interaction energy, and is described in detail in the Scoring section.

Sphere-based GA method

The sphere-based method orients the ligand by matching ligand atoms with spheres which are generated within the active site. The basic method of identifying and characterizing the active site by a set of overlapping spheres (program SPHGEN) has been described in detail elsewhere [11]. The packed spheres form a negative image of the enzyme active site and serve as a template for possible ligands. The molecular surface of the enzyme, as defined by Richards [44], is first generated using the Connolly MS program [45]. Spheres of varying radius are generated analytically to fill the different invaginations and clefts on the enzyme surface. Spheres touch the molecular surface at two points; they do not intercept any enzyme atom, and their centers lie along the surface normal of one of the surface points. Spheres are generated over the entire surface and then separated into clusters on the basis of radial overlap. The largest cluster is typically the ligand binding site.

In the sphere-based GA method, the spheres are used as targets at which ligand atoms are placed. Sets of ligand atoms match sets of sphere centers. This matching implicitly determines the transformation matrix of the ligand. The orientation of the ligand is computed by the rotation and translation matrix that best overlaps matched pairs of atoms and spheres, in much the same manner as with the DOCK algorithm. Unlike the DOCK algorithm, however, the pairing and selection is done by the GA method.

The orientation of the ligand is encoded by a series of pairs of integers. Each pair represents a matching of a sphere with a ligand atom. In this work, 12 pairs of atoms and spheres are used for the orientation. Each pair is encoded in the gene as a pair of numbers, one representing the atom number and the other representing the corresponding sphere number. The conformational degrees of freedom of the ligand are encoded as integer representations of the torsion angles. Consistent with the second method described in this paper, the torsion angles can vary in increments of 0.1 rad, or 5.6°.

For the sphere-based method, originally the program GAucsd [46] was used, but the experiments described here were done with a specially written GA program which uses a string of integers as the genome representation. Crossover was done at two points (frequency 90%) and mutation was accomplished by randomly increasing or decreasing a selected integer, with each integer having a 0.01 chance of being changed in each generation. This mutation method corresponds to the creep operator described by Davis [34].

Explicit-orientation-based method

The explicit-orientation-based method has two sequential parts: restricting the orientation of the ligand within the active site, and GA optimization of the ligand orientation and conformation within the site. We first describe the manner in which the limits for the orientation of the ligand within the active site are determined, then the GA optimization itself.

Determining limits of ligand orientation To orient a ligand, six parameters are needed, i.e., three coordinates to translate the ligand and three Euler angles to rotate it. The translation vector is initially bounded by a box enclosing the spheres characterizing the active site; the Euler angles are unbounded. To determine feasible starting orientations within the active site, more restrictive limits on these values are determined by either (i) DOCKing a rigid section of the ligand or (ii) running GA in a preliminary mode.

In the first case, a rigid section of the ligand is identified and oriented into the active site of the enzyme using the program DOCK [11]. Briefly, atoms in the ligand are matched to sphere centers in the active site. These sphere–atom pairs are then used to determine the orientation of the ligand within the active site, as described in the previous section. Structures are evaluated using the scoring grid described below. A set of the best scoring orientations are kept. These low-energy configurations of the rigid portion of the molecule are then subjected to a rigid-body minimization [47]. The maximum and minimum values of the parameters (translation vector and Euler angles) describing the set of low-energy orientations are then used as the limits on the orientation of the ligand for the GA runs.

In the second case, GA is run in a preliminary mode, to dock the entire (flexible) ligand. The center of mass of the ligand is limited to the sphere centers; the Euler angles are not limited. The low-energy configurations are kept and the maximum and minimum values of the translation vector and Euler angles are used as new limits on the orientational parameters for the final GA runs.

GA optimization of low-energy conformations and orientations Once the limits on the orientation of the ligand have been determined, GA is then used to generate and optimize low-energy orientations and conformations of the ligand. The variables used to fix the orientation of the ligand, as well as its conformation, are explicitly determined by the GA chromosomes. The translation vector and Euler angles which orient the ligand within the enzyme and torsion angles which specify the ligand conformation are all represented as floating point numbers with six bits of precision. Gray coding is used, so that a small change in bits corresponds to a small change in the corresponding floating point number [36]. The incremental difference in orientation represented by one bit depended upon the previously determined limits. Hence, each of the Cartesian components of the translation vector was varied by different increments for each of the test systems, ranging from 0.016 to 0.08 Å. Similarly, the Euler angles were varied by increments between 0.01 and 0.067 rad. The torsion angles were varied by increments of 0.1 rad, or 5.6°. The program GAucsd [46], based on the program Genesis [48], is used to generate, select and cross populations of strings. In each generation, approximately 10–20% of the population perished. (The GAucsd parameter sigma was set between 1 and 2.) Of the surviving population, between 60–70% participated in two-point crossover. The probability that a single bit would mutate was generally around 0.0065. The population size varied, depending upon the test system and the manner in which GA was used, and is described in a later section.

Evaluation of ligand orientations and conformations: Scoring

The different orientations and conformations of the ligand are evaluated primarily with a molecular mechanics interaction energy, based on the AMBER potential function [31,32], which approximates the ligand–enzyme binding energy. The interaction energy is calculated for each ligand atom using a scoring grid (generated by the program CHEMGRID [42]). The same scoring grid is used when DOCKing a rigid portion of the molecule, when evaluating the GA orientations and conformations and when performing the rigid-body minimization (see below). It is generated within a box enclosing the spheres which are packed into the active site.

To generate the interaction energies on the scoring grid, the enzyme factors of the interaction energy are separated from the ligand factors. At each grid point, three potentials due to nearby enzyme atoms are calculated: the vdW attractive potential, the vdW dispersive potential, and the electrostatic Coulomb potential. The comparable ligand terms of the interaction energy are stored separately with each ligand atom. For each orientation and conformation of the ligand, the ligand atom factors are combined with the enzyme factors of the grid point closest to the ligand atom. The total interaction energy is the sum of the interaction energies over all ligand atoms.

For all our test systems, parameters for the vdW and electrostatic energy of the enzyme atoms were taken from the united-atom parameter set of the AMBER package [31,32]. AMBER vdW parameters were also used for the ligand atoms [31,32]; here, the all-atom parameter set was used. Electrostatic charges of the atoms in the ligand were determined by the Gasteiger–Marsili method [49,50] using the program SYBYL from TRIPOS [51]. A distance-dependent dielectric was assumed. Receptor atoms within 10 to 20 Å of a grid point were used to calculate the receptor atoms 'potential'. The grid spacing was 0.3 Å.

The score for each conformation and orientation of the ligand is generally taken just as the interaction energy. However, when the interaction energy is particularly low, a crude conformational energy term can be added to the score. This term is based solely on a check of the interatomic distances in the ligand against the sum of their vdW radii. When the distance between two atoms in a ligand is less than the sum of their vdW radii, the intramolecular vdW energy is added to the score. This is a vast simplification of the intramolecular energies usually calculated, and is used primarily to prevent two different atoms in the ligand from occupying the same space.

Rigid-body minimization

As a final step to optimize the structure, after the GA runs, the low-scoring GA solutions which were stored for further analysis are subjected to a rigid-body minimization using either a quasi-Newton or a simplex optimizer [52]. Details have been described previously [47]. In this step, the conformation of the ligand is held fixed, and the rigid-body orientation of the ligand is varied slightly. The energy of the ligand is calculated on the same scoring grid used for the GA runs, but for each ligand atom, the enzyme terms from several nearby grid points are averaged to calculate a more accurate interaction energy. Intramolecular energy penalties are not calculated in this final step.

Test systems

For the first part of the work, in which we attempt to regenerate the crystal structure, three different test systems were selected: dihydrofolate reductase–methotrexate (dhfr–mtx), thymidylate synthase–phenolpthalein (ts–fen), and HIV protease–thioketal haloperidol (hivp–thk). These enzyme systems were selected partly because they have been targets for inhibitor design; each plays a crucial role in a particular metabolic pathway [53–56]. They also represent a wide range of enzyme–inhibitor interactions. mtx fits tightly into the dhfr binding site, which is charged, and has an IC_{50} on the order of 1 nM [54]. Phenolpthalein, on the other hand, is small relative to the ts binding site and does not fully fill the cavity; it forms hydrogen bonds to a water molecule, and has an IC_{50} of 15 μM [57]. The hivp–thk system represents an intermediate case; it fits loosely in the active site and has an IC_{50} of 15 μM [58].

For the dhfr–mtx system, the enzyme structure was taken from entry 4dfr [59] from the Brookhaven Protein Data Bank [60]. All crystallographic waters were removed. Hydrogens were added to dhfr. The mtx structure was also taken from 4dfr. In addition, another mtx structure was used; this was taken from the Fine Chemicals Directory [61] and was generated by the program Concord from TRIPOS [51]. We refer to the former mtx structure as 'crystal-generated mtx' and to the latter as 'Concord-generated mtx'. Hydrogens were added to mtx, and eight bonds were allowed to rotate in this structure. For the explicit-orientation-based GA, the rigid part of mtx was taken to be the pteridine ring.

For the dhfr–mtx system, two different sphere sets were generated, reflecting the manner in which they were used. For the sphere-based GA method, dhfr molecule 'A' from the 4dfr unit cell was used to generate spheres, whereas for the explicit-orientation-based system, molecule 'B' from the unit cell was used to generate spheres. The former set of spheres covered the entire mtx binding pocket well and was used to dock the entire ligand in the sphere-based GA method. The latter set of spheres defined primarily the pteridine binding area and was used in the explicit-orientation-based method to dock the pteridine ring. Scoring grids were calculated using the appropriate molecule from the unit cell, although in all cases the mtx from molecule 'B' was used (little difference was found in the results when the other mtx was used).

For the ts–fen system, the crystal structure of the enzyme–ligand complex [57] was obtained from K. Perry (Biochemistry Department, UCSF). The crystallographic waters within the active site were retained; all other waters were removed. In addition, the phosphate counterion was included. Hydrogens were added to the ts and fen. Four bonds on phenolpthalein were allowed to rotate. However, two of these bonds only affected the hydrogens on hydroxyl groups, which were not used in any rmsd calculations.

For the hivp–thk system, the crystal structure of the HIV protease enzyme–ligand complex [58] was obtained from R. Keenan (Biochemistry Department, UCSF). The enzyme was the Q7K mutant of the protease. All waters were removed, but the chloride ion bound to the flaps of the enzyme was kept. Hydrogens were added to the protease, including one between the catalytic aspartic oxygens, and also to thk from the crystal structure. For the sphere-based GA, seven bonds were allowed to rotate. For the explicit-orientation-based GA, the rigid part of thk was taken to be the chloro-phenyl piperidinol rings and five bonds were allowed to rotate. All ligands and their rotatable bonds are illustrated in Fig. 1.

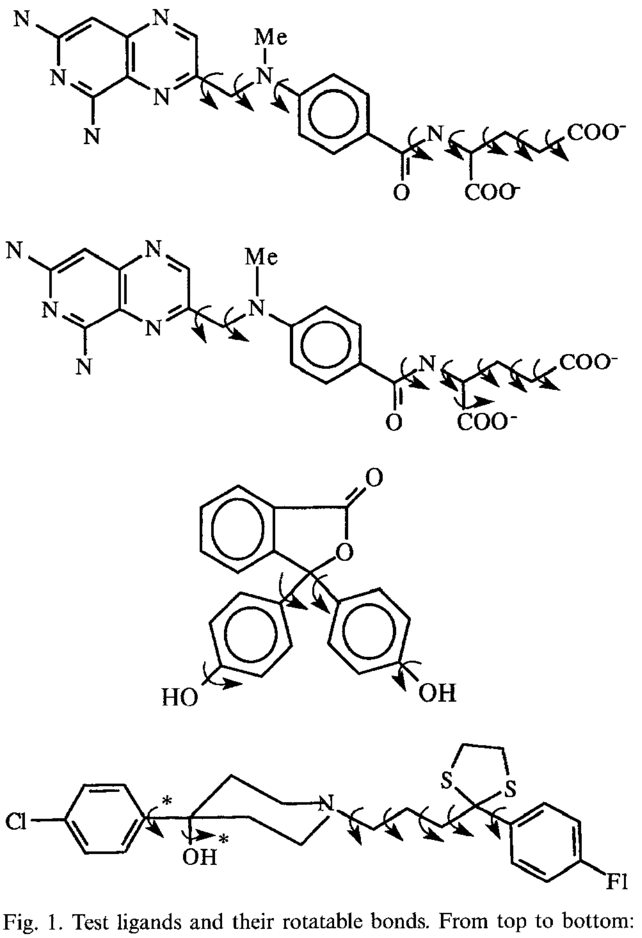

Fig. 1. Test ligands and their rotatable bonds. From top to bottom: crystal-generated methotrexate, Concord-generated methotrexate, phenolpthalein and thioketal haloperidol.

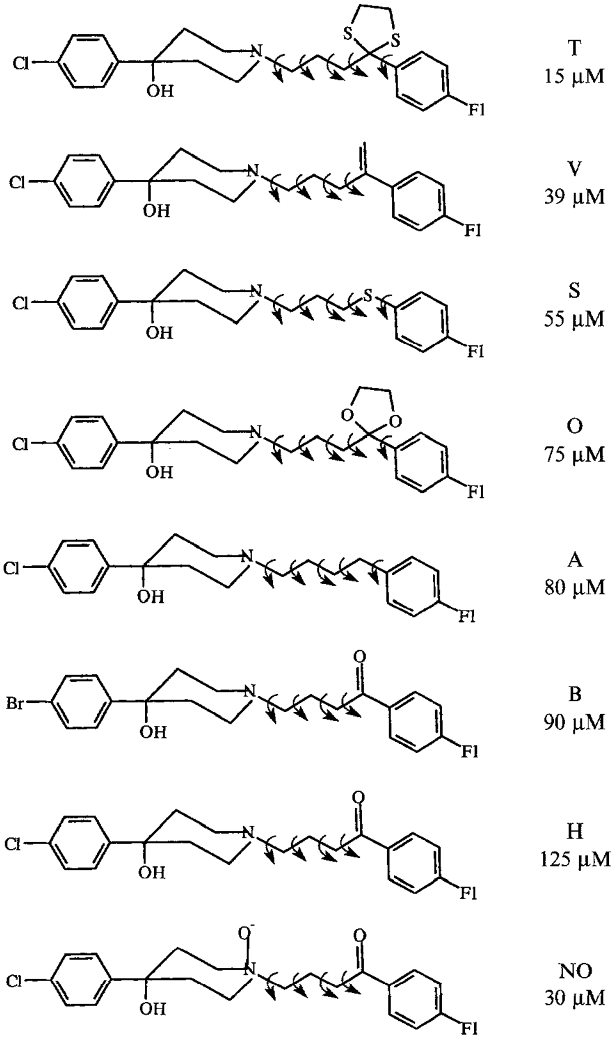

Fig. 2. Derivatives and analogues of haloperidol. Rotatable bonds are indicated. Ligand acronyms and IC_{50} values are listed on the right (Ortiz de Montellano, P.R. et al., personal communication).

In the second part of the work, GA was examined as it would be applied to screening a database of compounds. The same three test systems were used. In addition, a mixture of ligands was tested against the HIV protease structure, as described above. Test ligands were all derivatives or analogues of haloperidol. The ligands, listed descriptively, were: thioketal haloperidol (T), a vinyl derivative of haloperidol (V), a thio analogue of haloperidol (S), a ketal haloperidol (O), an alkyl derivative of haloperidol (A), bromoperidol (B), haloperidol (H) itself, and an N-oxide-substituted haloperidol (NO). The latter molecules, with their rotatable bonds, are illustrated in Fig. 2. Although only the crystal structure of the HIV protease–thioketal complex has been determined, the IC_{50} values of these ligands are known (Ortiz de Montellano, P.R. et al., personal communication), and are also listed in Fig. 2.

Computational experiments

In the first part of this work, the accuracy of the GA method is examined. First, to illustrate the potential of the method, the sphere-based GA is used to generate different orientations of a rigid ligand in a receptor. In this case, only the dhfr–mtx test system is used and the chromosome represents strictly the orientation of the rigid ligand. Next, both GA methods are applied to the three test systems to test the ability of the method to reproduce the crystal structure and to generate low-energy conformations and configurations of flexible ligands. The population size, number of generations and total number of evaluations for both flexible docking methods are given in Table 1.

In the second part of this work, we focus on the explicit-orientation-based GA as it would be applied for screening a large database of compounds. The objective is to examine whether or not potential ligands can be rapidly identified. Since our scoring function approximates the enzyme–ligand binding energy, in the second part of this work we investigate the extent to which low-energy structures – not necessarily close to the crystal structure – are generated. Unlike the first part of this work, where the run parameters were selected to provide sufficient sampling so that GA solutions close to the crystal structure would always be found (for the explicit-orientation-based method), now a different set of GA parameters was selected, more appropriate for the screening of a database of compounds. We first examine the same three test systems described above, but with more limited population sizes and computational times. The different population sizes, as well as the number of runs, are given below in the Results section. Each run comprised approximately 65 generations.

Finally, as part of our application of the explicit-orientation-based GA to screening a database of compounds, we also study the behavior of a GA population containing a mixture of ligands. We apply GA in a manner inspired by combinatorial chemistry [38–41]. Various de-

TABLE 1
GA PARAMETERS FOR THE TEST SYSTEMS

System	Population	Generations	Total evaluations
Sphere-based			
dhfr–mtx	800	50	940 332
ts–fen	50	200	240 282
hivp–thk	50	200	256 370
Explicit-orientation-based			
dhfr–mtx	12 167	39	376 843
ts–fen	279 841	21	4 520 297
hivp–thk	761	102	51 420

Listed are the population sizes, number of generations and total number of evaluations used for the different test systems for the first part of the work (see text).

rivatives and analogues of haloperidol are simultaneously docked into HIV protease and are allowed to interbreed and compete. The enzyme is held fixed while a mixture of ligands of different conformations and configurations is generated and evaluated. In this case, each GA chromosome has a ligand identification tag, in addition to the conformation and orientation information. Survival of each ligand orientation and conformation is determined relative to its own (sub)population of strings and the best of each class survives. The lowest energy GA solution of each haloperidol derivative and analogue from the mixture is determined. Each of these haloperidol derivatives and analogues is also evaluated separately, to verify that low-energy conformations were found in the computational mixture. Three sets of GA runs were performed: (i) a population of size 761, containing a mixture of eight ligands; (ii) a population of size 761, containing a single ligand; and (iii) a population of size 112, containing a single ligand.

Results

The results are divided into two parts, each with two sections, depending upon the manner in which GA was used. Part I concentrates on single-ligand experiments, attempting to evaluate the accuracy of the method. Both variants of GA are used. The results from each method are reported separately for clarity. We report the deviations from the crystal structure, describe the families of low-energy GA solutions, and provide possible reasons why these alternative low-energy solutions – which deviate from the crystal structure – are generated. Part II concentrates on using GA to screen a database of compounds. In this case only the explicit-orientation-based GA method is considered. All calculations were performed on a 33 MHz SGI Indigo.

I. Single-ligand experiments: Comparison with crystal structure

Rigid docking

To illustrate the power of the GA method, we first compare the sphere-based GA method in a rigid docking mode to DOCK. For the dhfr–mtx system, DOCK found 273 orientations of mtx which scored below −40 kcal mol^{-1} (using 854 CPU seconds). The best scoring orientation was −58.3 kcal mol^{-1} and it differed from the crystal structure by an rmsd of 0.72 Å. Ten runs of the sphere-based GA method found several good orientations, including a best score of −70.4 kcal mol^{-1} (rmsd 0.37). The total run time for all 10 runs was 317 s. Further refinement of the 273 DOCK hits by rigid-body minimization took 155 s and gave a best scoring orientation of −69.8 kcal mol^{-1} (rmsd 0.39), while rigid-body minimization of the GA hits did not appreciably change the best score (−70.6 kcal mol^{-1}). This comparison illustrates that the sphere-based GA method is a fast and powerful way to perform rigid docking as an alternative to the matching techniques incorporated into DOCK. In general, we have found the GA method to be faster and to yield better scoring hits (Dixon, J.S., unpublished results). While the rigid-body minimizer is able to refine the DOCK orientations until they are essentially as good as the GA ones, it takes extra computer time.

Flexible ligand docking

Our primary interest is the use of the GA method to allow conformational freedom of the ligand. The accuracy of the GA method incorporating flexible ligand docking is examined and summarized for the three test systems in Table 2. The deviation from the crystal structure and the interaction energy of the best scoring and most accurate conformations are listed in each case. The sphere-based and explicit-orientation-based methods seemed to explore different regions of the solution space; this is probably partly due to the construction of each method. With the sphere-based method, several orientations and conformations with good energies were found for all test systems, but they were not necessarily close to the crystal structure. With the explicit-orientation-based method, energies comparable to the crystal structure were obtained for all test systems, with errors on the order of 1.0 Å; torsion angle errors based on the crystal structure averaged less than 60° for the best cases, positioning the dihedral angles

TABLE 2
ACCURACY OF GA SOLUTIONS

Test system	Sphere-based GA				Explicit-orientation-based GA			
	Energy		rmsd (heavy atoms)		Energy		rmsd (heavy atoms)	
	Best scoring	Lowest rmsd	Best scoring	Lowest rmsd	Best scoring	Lowest rmsd	Best scoring	Lowest rmsd
dhfr–crystal-generateded mtx[a]	−71	−66	3.2	2.8	−75	−65	1.2	0.6
dhfr–Concord-generated mtx[b]	−65	−65	2.1	2.1	−68	−64	1.4	1.0
ts–fen[c]	−35	−29	5.0	2.4	−36	−34	1.2	0.5
hivp–thk[d]	−52	−46	2.7	1.8	−50	−50	3.3	1.0

[a] Crystal energy = −70 kcal mol^{-1}.
[b] Crystal energy = −70 kcal mol^{-1}.
[c] Crystal energy = −33 kcal mol^{-1}.
[d] Crystal energy = −48 kcal mol^{-1}.

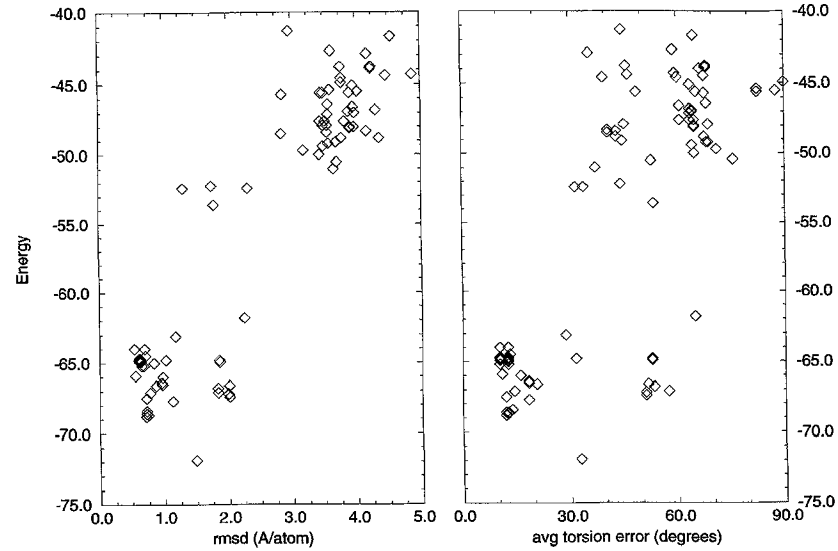

Fig. 3. Comparison of a crystal structure and GA structures: rmsd (left) and average torsion angle error (right) for the dhfr–crystal-generated mtx test case.

in the proper conformational minima (Figs. 3–6). Table 3 lists GA computational times used for these systems. Table 4 gives the size of the orientational space of the explicit-orientation-based method, both before and after the restriction of the space.

dhfr-mtx: Sphere-based GA Preliminary results [62] have shown that the sphere-based GA method can find good orientations for mtx using a different set of spheres. However, the spheres generated from molecule 'B' of the dhfr–mtx structure did not prove suitable for the flexible sphere-based method. They do not describe the mtx binding site well, with the exception of the pteridine binding pocket. In contrast, the spheres generated from molecule 'A' more uniformly described the whole mtx binding site and so were used in these calculations. For the crystal-generated mtx, the best scoring solution had an energy of −71 kcal mol^{-1}, a value lower than the crystal structure energy. However, as indicated in Table 2, the rmsd was 3.2 Å. This illustrates the inadequacy of the force field-based scoring function, since the structure is oriented very close to the crystal structure, except for the pteridine ring system; the latter system is extended down into the cofactor binding pocket. Thus, the rmsd is high, although the score is identical to that for the observed binding orientation. Since much of the binding energy comes from appropriate positioning of the glutamyl side chain to maximize electrostatic interactions, the orientation of the pteridine rings is not well reproduced.

dhfr–mtx: Explicit-orientation-based GA For the dhfr–mtx test system, GA produced two groups of low-energy structures for both the crystal-generated mtx and the Concord-generated mtx: one differing by 1 Å/atom or less from the crystal structure and one differing 1 Å/atom or more from the crystal structure. The latter contained

TABLE 3
COMPUTATIONAL TIMES FOR VARIOUS GA PROCEDURES

System	CPU time docking (min)	CPU time GA (min)	CPU time rgb (min)
Sphere-based GA			
dhfr–mtx		85.5	0.125[a]
ts–fen		15.7	0.133[a]
hivp–thk		18.9	0.063[a]
Explicit-orientation-based GA			
dhfr–mtx	1.3[b]	66	2[d]
ts–fen	6[c]	198	2.2[d]
hivp–thk	3[b]	6.7	3[d]

All calculations were performed on a 33 MHz SGI Indigo.
[a] Simplex rigid-body minimization time for 30 structures.
[b] DOCK run times for rigid structure.
[c] GA preliminary run times.
[d] Quasi-Newton rigid-body minimization time for 100 structures.

TABLE 4
EXPLICIT-ORIENTATION-BASED GA: LIMITS OF ORIENTATION

System	Cartesian volume (Å)	Eulerian volume (rad)
Original[a]		
dhfr–mtx	24 × 10 × 26	2π × 2π × π
ts–fen	23 × 14 × 24	2π × 2π × π
hivp–thk	12 × 9.5 × 19.5	2π × 2π × π
Reduced[b]		
dhfr–mtx	1.8 × 3.3 × 1.0	0.6 × 0.8 × 1.2
ts–fen	2.0 × 3.0 × 3.0	π × π × π
hivp–thk	5.0 × 1.5 × 4.0	1.6 × 4.3 × 1.2

[a] The original active site volume and Euler angle range.
[b] The reduced active site volume and Euler angle range after the site was restricted.

the lowest energy structure (−73 kcal mol^{-1} for crystal-generated mtx, −69 kcal mol^{-1} for Concord-generated mtx), while the structures close to the crystal structure had a higher energy (−69 and −63 kcal mol^{-1}, respectively). The energy of the crystal structure was −70 kcal mol^{-1} by our scoring method.

As can be seen from Fig. 3, crystal-generated mtx structures with low energy generally have smaller deviations from the crystal structure, as measured by rmsd and average torsion angle error, than do higher energy structures. In this case, GA solutions which deviate from the crystal structure by approximately 1 Å have an average torsion angle error of less than 30°. For the Concord-generated mtx, low-energy (and low-rmsd) GA solutions generally had an average torsion angle error between 30 and 60° (data not shown).

In general, the most variable angles are outside the pteridine-binding site, especially at the terminal carboxyl group. We illustrate this by examining in more detail the torsion angle error of the Concord-generated structure (Fig. 4). We use the method of parallel coordinates [63,64] to display the individual torsion angle errors of GA conformations. With this method, the x-axis represents the torsion angle number and the y-axes are the errors for each torsion angle. (Torsion angle numbers start from the left in Fig. 4 and increase from left to right.) The errors are connected with lines to aid in the visualization.

The torsion angle errors from GA solutions with small deviations from the crystal structure are displayed at the bottom of Fig. 4; errors from alternate GA solutions are displayed at the top for comparison. The torsion angle errors for the low-rmsd structures were concentrated along the part of the ligand away from the pteridine ring. That is, the angles responsible for the 'bend' in mtx had very low errors (this is true in general also for the crystal-generated mtx case). However, GA solutions which differed from the crystal structure had much higher errors here. As can be seen from the plot at the top of Fig. 4, the first three torsion angle errors display a 'crankshaft' motion. This allows the ligand to bend in the active site, albeit in a manner slightly different from that in the crystal structure.

Alternate low-energy GA solutions The fact that GA structures are generated with energies lower than the crystal structure can easily be understood after examination of the enzyme structure and our force field. The primary difference between the two families of GA solutions was the orientation of the terminal carboxylic acid group. (This is consistent with the torsion angle errors mentioned earlier.) The location of this charged functional group on the surface of the enzyme – which has many different charged groups – made alternative electrostatic contacts possible. For example, for the Concord-generated mtx, the best scoring solution has the oxygen close to a lysine side chain, rather than the arginine side chain (as in the crystal structure). Furthermore, the oxygen is slightly closer to the lysine charge group than it is to the arginine group in the crystal structure. The closer proximity of two oppositely charged groups resulted in a lower energy and accounts for the majority of the energy difference between the two structures.

ts-fen: Sphere-based GA Using the sphere-based method, several families of low-scoring orientations could be located. One family had an rmsd of about 2 Å, one about 5 Å and one about 6 Å; all had energies comparable to

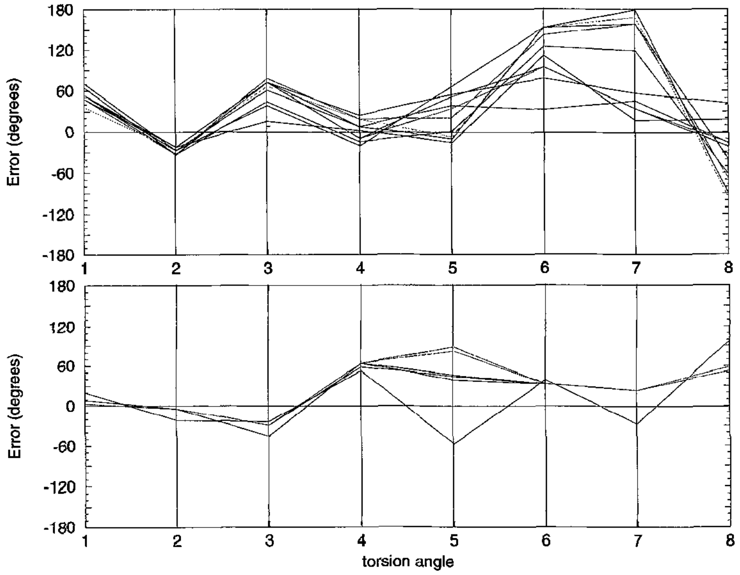

Fig. 4. Torsion angle errors of GA structures for the Concord-generated mtx test case. GA structures with small deviations from the crystal structure are given in the lower part; those with large deviations are given in the upper part. For a definition of the torsion angle numbers, see Fig. 1. Errors from each structure are connected together with lines.

the crystal structure. The 2 Å and 5 Å families were oriented in the active site in roughly the same position as the crystal structure. In the 2 Å family, the pterin and phenol roughly overlapped with the corresponding rings from the crystal structure. In the 5 Å family, the pterine ring had swapped with one of the phenol rings, but the placement of all three rings overlapped roughly with that of the crystal structure. GA solutions from both families made similar contacts with the receptor; hence, similar energies were obtained. The last family of structures was located roughly 4.5 Å away from the crystal structure position. In this case, one of the hydroxyl groups on the ligand made good electrostatic contact with the phosphate ion, rather than the water molecule, in the active site.

Several families of GA solutions were generated probably because there was no requirement in our scoring function that the oxygens be oriented in a particular manner. Hydrogen bonding, however, appears to play an important role in orienting phenolpthalein in the active site [57]. Since our simplified force field lacks an explicit hydrogen bonding term, the oxygen atoms on phenolpthalein were free to be oriented in any direction, as long as they were within a reasonable (electrostatic) distance of the enzyme. In addition, the large number of conformations of GA in the 2 Å family is probably also due to the lack of an internal torsion angle term. Hence, these alternative families of GA solutions are all allowed.

ts–fen: Explicit-orientation-based GA Although there were many conformations and orientations with very similar energies, the low-energy structures fell into only two general families: solutions which were within an rmsd of 2.5 Å from the crystal structure and solutions which differed from the crystal structure by more than 4.2 Å. This can be seen in Fig. 5. The structure which most closely resembled the crystal structure had an rmsd of 0.6 Å/atom. The average torsion angle error for the 2.5 Å family varied substantially, from 0 to π, in large part because this system is the most sterically hindered that we examined and it is the most poorly determined by our limited evaluation of conformational energy.

The large variety of conformations in the 2 Å family is partly the result of the lack of an internal torsional angle energy term in our simplified force field. The primary difference between the many low-rmsd structures and the crystal structure is the orientation of the phenol rings: when the deviation is relatively high, both rings are oriented perpendicular to the orientation in the crystal structure. Inclusion of an internal torsional energy term would probably help distinguish the different conformations for this sterically hindered system. In addition, as mentioned in the previous section, the inclusion of an explicit hydrogen bonding term could help to place the oxygens on phenolpthalein in a particular position.

Most of the GA solutions from the 4.2 Å family had an interaction energy dominated by electrostatic interactions; 95% of its energy comes from the close proximity of the two hydroxyl hydrogens to the enzyme or water oxygens. Although its score is on the same order as the crystal structure, there are very few good vdW interactions, which is unreasonable. Consequently, this structure can be eliminated from further consideration.

hivp–thk: Sphere-based GA In the sphere-based GA method, there were roughly four families of orientations/conformations, all with energies less than or equal to the crystal structure energy. We refer to the different families by their deviation from the crystal structure. Since it was not necessary with this method to use a rigid portion of the ligand, all seven bonds were allowed to rotate. Perhaps for this reason, the lowest scoring structures were slightly lower in energy.

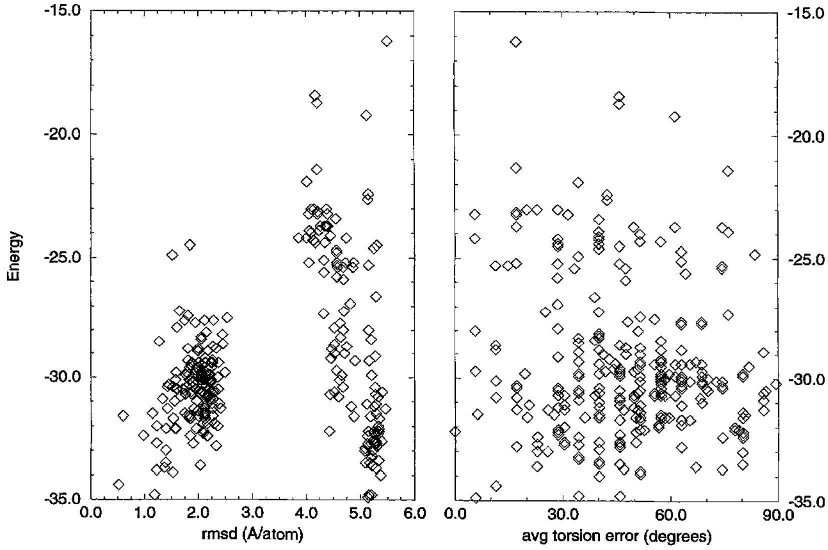

Fig. 5. Comparison of a crystal structure and GA structures: rmsd (left) and average torsion angle error (right) for the ts–fen test case.

The GA solution with the smallest deviation from the crystal structure (1.8 Å/atom) had an energy of –46 kcal mol^{-1}. This family of solutions, with an rmsd of 2 Å or less and the second lowest energy, was very similar to the crystal structure. Both the 3 Å family and the 5 Å family had GA solutions with lower energies. In the latter two families, the Cl-phenyl ring was rotated relative to the crystal structure. Due to this rotation, the Cl-phenyl ring made better vdW contacts with the enzyme than did the same ring in the crystal structure, resulting in a lower interaction energy. In addition to the rotation of the Cl-phenyl ring, in the 5 Å family the positions of the fluorophenyl and thiol rings were exchanged compared to the crystal structure. This portion of thk extends outside the active site; neither ring makes any particular favorable contacts with the enzyme. As a result, this family of solutions and the 3 Å family of solutions have very similar energies.

In general, for the first three families of GA solutions, the piperidinol ring remained close to the crystal structure position. On the other hand, the last family of GA solutions differed greatly from the crystal structure: it was flipped end-to-end relative to the crystal structure. The piperidinol group was not in the same position as the crystal structure. In this case, the hydroxyl group on the piperidinol group made good contact with a backbone carbonyl oxygen close to the aspartyl active site. Good vdW and electrostatic contacts were maintained.

hivp–thk: Explicit-orientation-based GA The low-energy GA solutions fell into two families: those which differed from the crystal structure by less than 2 Å rmsd/atom and those which differed by more than 3.2 Å/atom (see Fig. 6). The lowest interaction energy in each family was the same, when only intermolecular interaction energies were considered. The average torsion angle error for the low-energy structures was on the order of 50°, with the most variable angles in the fluorophenyl/thioketal region of the molecule.

The generation of alternate GA structures with energies comparable to the crystal structure can easily be understood after examination of the enzyme–ligand complex. As mentioned earlier, the primary difference between the crystal structure and solutions of both the 2 Å and 3.2 Å families lies in the fluorophenyl/thioketal region of the molecule. The large variety of conformations in the 2 Å family was largely due to a rotation of the fluorophenyl ring. In the 3.2 Å family, the positions of the fluorophenyl ring and the thioketal ring are exchanged. This portion of thk lies on the surface of the enzyme and protrudes out of the active site; with our force field scoring, neither ring has any particularly favorable interaction with the enzyme in the crystal structure. There is no driving force to generate the crystal conformation.

Summary

For all test systems examined, GA conformations and orientations were found with interaction energies comparable to or lower than those of the crystal structure. For the explicit-orientation-based method, a low-scoring conformation was always found with a deviation of 1 Å/atom or less; these structures had most of their torsional angles in the crystallographically observed minima. Alternate low-energy conformations were also found, which is probably the result of our simplified force field; we discuss this below.

II. Database screening

Single-ligand experiments

When screening a database of compounds for a poten-

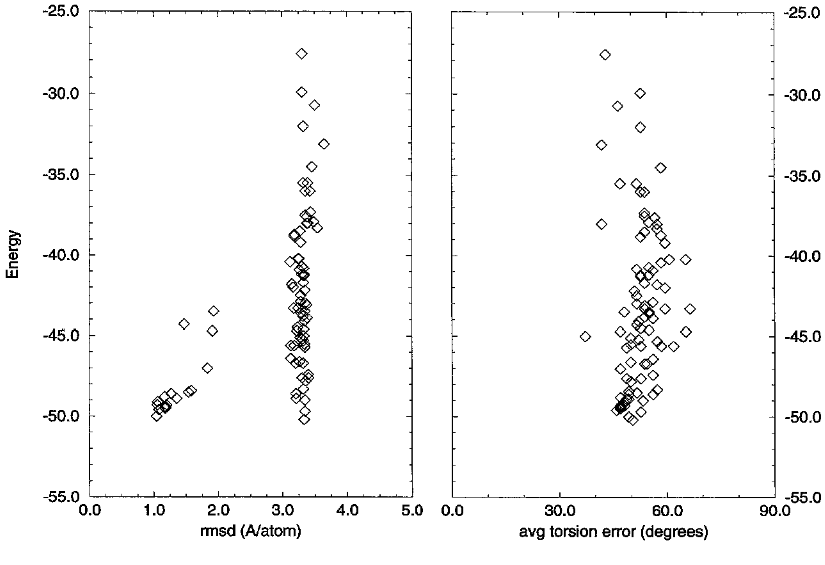

Fig. 6. Comparison of a crystal structure and GA structures: rmsd (left) and average torsion angle error (right) for the hivp–thk test case.

tial inhibitor of an enzyme, the actual binding orientation and conformation of the ligand is not known. The objective is not, as it was in the previous section, to regenerate a known binding structure, but rather, to search for ligands with low interaction energies. Here, we explore the population size and computational time needed to reach low-energy GA solutions, ignoring the deviations of these solutions from the crystal structure. We focus on the explicit-orientation-based GA method here and in the next section. In general, multiple runs with population sizes considerably smaller than those used in the previous section generated low-energy solutions. Most notably, a large decrease in CPU time was seen in the ts–fen system, as reported below.

Table 5 summarizes the results for all test systems for the different population sizes. The number of runs, number of instances where energies were within 20% of the crystal energy, and the computational time for each run are given. As can be seen in Table 5, for the hivp–thk system the low-energy conformation was generated in 5 min of CPU time. (Note that, although the population size for the hivp–thk system was the same as in the previous section, fewer generations were used to generate low-energy conformations; this slightly decreased the CPU time compared to the previous section.) For the ts–fen system, the low-energy conformation was generated in 2 min of CPU time. Both these energies were comparable to the crystal structure energies. For the dhfr–mtx system, several GA runs were necessary to generate structures with energies as low as that of the crystal structure. Each run, however, took only 6 min of CPU time.

TABLE 5
SUMMARY OF RESULTS, PART II: SINGLE-LIGAND EXPERIMENTS

Population	No. of runs	Energy within 20% of crystal energy[a]	CPU time per run (min)
dhfr–mtx			
12 167	5	4	66
761	8	3	5.7
112	10	0	1.1
ts–fen			
761	5	5	1.8
112	8	5	0.6
23	8	6	0.3
hivp–thk			
761	5	5	4.8
381	5	2	3.3
112	10	2	0.8

[a] The number of instances that the energy of a GA structure was within 20% of the crystal energy, as a function of population size. dhfr–mtx: crystal energy = –70 kcal mol^{-1}; energy < –53 kcal mol^{-1}. ts–fen: crystal energy = –33 kcal mol^{-1}; energy < –25 kcal mol^{-1}. hivp–thk: crystal energy = –48 kcal mol^{-1}; energy < –38 kcal mol^{-1}.

Plots of lowest energy vs. generation number for different population sizes of the hivp–thk system are given in Fig. 7. The behavior of the curves is similar to that observed for the other test systems, although the actual population size used for each test system varied.

As can be seen in Fig. 7, with a larger population, any one GA run is far more likely to converge to a low-energy solution. As the population size decreases, the envelope of low-energy solutions produced by multiple runs with small populations often overlaps with the (smaller) envelope of low-energy solutions from the GA runs with larger populations.

HIV protease screen, mixture of ligands

As part of our application of GA to screening a database of compounds, we also studied the behavior of a GA population containing a mixture of ligands. Again, we focus simply on the explicit-orientation-based method. GA, which makes use of diversity within a population, naturally lends itself to such a screening. In this last set of experiments, we study a GA population containing a mixture of ligands which were all derivatives or analogues of haloperidol. All ligands were docked to HIV protease.

The rigid portion of all the ligands was presumed to bind in the same manner as in the hivp–thk crystal structure. The limits on the translation vector and Euler angles are as described above. Three sets of GA runs were performed: (i) a population of size 761, containing a mixture of eight ligands; (ii) a population of size 761, containing a single ligand; and (iii) a population of size 112, containing a single ligand. Several runs were performed for each population size.

The lowest energy obtained for each ligand in the mixture of ligands from each run was saved. The range of this lowest energy from each run is given in Table 6. In addition, the table lists the range of the lowest energy obtained for each ligand from individual GA runs with large and small populations.

As can be seen by a comparison of the energy values in the table, low-energy conformations and orientations are always obtained for seven out of eight ligands when they are contained in a mixture. These low energies are comparable to those obtained in single GA runs with the large population. With the small population, the lowest energy solution varies considerably with multiple runs, as was seen in the previous section.

Discussion

As mentioned in the Introduction, the long-term objective of this work is to develop a computational method to screen databases of conformationally flexible ligands with properties complementary to an enzyme active site. Such methods must achieve a balance between accuracy and processing time. In this section, we first consider the

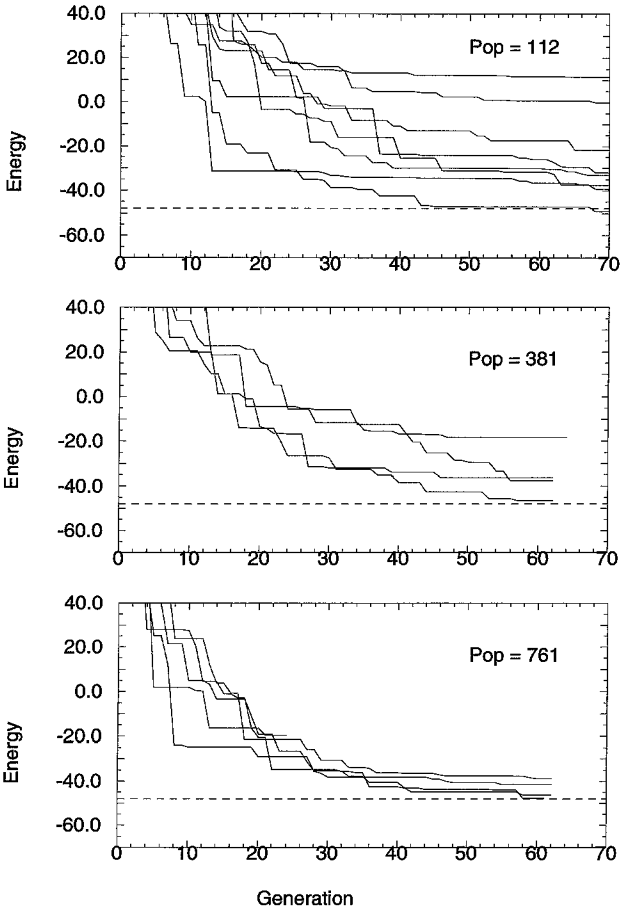

Fig. 7. Plot of the lowest energy as a function of generation number for the thioketal test system with different population sizes. Different lines in each figure refer to different runs.

accuracy of the method and compare it to that of other published docking methods. We then compare the two different GA methods. Next, we examine the processing time required for screening a database of compounds. We then discuss the results of applying our method to a mixture of ligands. Finally, we suggest variants on our method which may also be useful for flexible docking.

Comparison with crystal structure

Both GA methods always found low-energy solutions. In addition, however, the explicit-orientation-based GA with a force field fitness function, followed by rigid-body minimization, did reasonably well at reproducing the crystal structure. For all test systems examined, a low-scoring conformation was always found with a deviation of 1 Å/atom or less. This is especially gratifying for the Concord-generated mtx–dhfr case, where the starting mtx structure is generated independently, with standard bond lengths and bond angles.

GA (or any other optimization method) will do no better than the accuracy of its scoring method. The best we could hope to achieve would be deviations of 0.2 Å/atom for mtx, 0.3 Å/atom for thk and 0.5 Å/atom for fen, since with our scoring method, the lowest energy orientation of the crystal conformation differs from the true crystal orientation by these amounts.

The scoring function determined not only the limits of the accuracy of the method, but also allowed for the generation of conformations that were different from the crystal structure, but with energies comparable to or lower than that of the crystal structure. The force field used to evaluate each conformation and configuration was, by construction, simple; we did not calculate the total energy of the enzyme–ligand complex. For example, our force field lacked an explicit hydrogen bonding term, a hydrophobicity term, a solvation/desolvation term, and intramolecular energy terms. We selected such a force field because it is both rapid and reasonably accurate for screening a database of compounds.

For each of our test systems, the fact that alternate low-energy solutions are generated can easily be understood when taking into consideration our simplified scoring scheme (vide supra). That is, alternate structures are primarily a result of our method of scoring, rather than of the GA method itself.

Finally, it is worthwhile noting that in these structures of ligand–enzyme complexes, the ligand portion and conformation can be considerably less well determined than the structures as a whole, as seen, for example, in the reported thermal factors [57–59]. Thus, we must be cautious in using the agreement with the observed structure as the only criterion of merit.

Comparison of the two GA methods

Comparison of the results obtained by the explicit-orientation-based method and the sphere-based method are interesting. While both methods seem successful at finding ligand orientations which score at least as well as the crystal structure, the explicit-orientation-based method was also able to find GA solutions close to the crystal structure. This might be due to the manner in which each method was constructed: the sphere-based method provided broad orientational and conformational sampling in the active site of the enzyme, while the explicit-orientation-based method explored detailed conformational sampling of a restricted portion of the active site. In some of our tests, the two methods seem to have explored different regions of the solution space. Clearly, the sphere-based method depends critically on the placement of the spheres. When SPHGEN spheres, used for docking the pteridine ring, were employed to DOCK mtx into dhfr, they proved to be problematical; using our normal parameters, it was difficult to find any docked orientations of mtx close to the X-ray structure. Similarly, these spheres could not be used for the sphere-based flexible docking. We are exploring alternative ways to generate spheres or other target points for flexible docking. In docking a large database of molecules, the explicit-orientation-based method and the sphere-based method may provide complementary results. This issue remains to be investigated.

Comparison to other docking procedures

One of our GA methods (explicit-orientation-based GA + rigid-body minimization + force field scoring) appears to reproduce the crystal structure as well as other methods. Here, we compare the accuracy of the results using this method with the results of other workers.

dhfr–mtx has often been used as a test system, although not specifically the 4dfr system. For our dhfr–mtx test system, our low-scoring, low-rmsd GA solution is 0.8 Å/atom. Using the crystal conformation and rigid-body docking methods, Bacon and Moult reported a value of 0.6 Å/atom [12]; Hart and Read found 0.7 Å/atom [27]. Using flexible docking methods, Leach and Kuntz reported a value of 0.8 Å/atom [30]. Yamada and Itai, however, found a much lower value of 0.3 Å/atom [33]. Their method searched for reasonable mtx structures by matching hydrogen bonding points in mtx and dhfr, then minimized the structures using the AMBER force field [31,32]. This is a more accurate method than ours, calculating more terms for the energy. Furthermore, although much of the hydrogen bonding interaction is in the electrostatic term in our force field, we do not calculate an explicit hydrogen bonding term, and hence we lose some of the directional requirements for hydrogen bonding. It is not very surprising, then, that Itai's reported rmsd is much lower than ours.

In other test systems, flexible docking methods have generally produced structure deviations on the order of 1 Å/atom. For example, Goodsell and Olson, using a Monte Carlo flexible docking method, reported an rmsd of 1 Å/atom for McPC603–phosphocholine [25]. Stoddard and Koshland, using the same method, found a value of 1 Å/atom for the aspartate receptor–aspartate system [29]. Judson et al. reported an rmsd value of 1.1 Å/atom for a thermolysin–peptide complex using a GA method similar to ours [65]. All our test systems have similar deviations.

Computational processing time

In the previous section, we discussed the accuracy of the method. In this section, the computational processing time required for the method is considered. In particular, for the ts–fen system, the large difference in CPU time required to generate an accurate GA structure vs. the CPU time required to simply generate a low-energy structure is rationalized. We note, however, that although some consideration has been given to minimize the CPU time, much of the code has not been optimized. The programs are prototypes and not production programs. Nevertheless, we feel that it is important to get a sense of the time required to carry out these kinds of searches; therefore we discuss it here. Consistent with our long-term goals, the discussion is limited to the explicit-orientation-based GA method, with emphasis on the CPU times when screening a database of compounds.

The total processing time required is a function of the population size, the number of generations and the number of GA runs needed to reach a low energy. The bulk of the CPU time for GA is spent evaluating each chromosome, generating the actual ligand structure and scoring the structure. Depending upon the test system, GA evaluates between 95 structures (for the dhfr–mtx system) and 335 structures (for the ts–fen system) per CPU second on the 33 MHz SGI. As expected, the number of evaluations per second decreases with the number of torsion angles and the number of atoms in the test ligand. For the dhfr–mtx and hivp–thk systems, between 25–35% of the total CPU time is spent in the intramolecular energy calculation. Note that, unlike other optimization schemes, GA does not require the calculation and storage of the derivatives of the energy with respect to the variables being optimized. The 'direction' of the search is included in its fitness value. Hence, such calculations are not part of the processing time.

An appropriate selection of population size and number of experiments is important, since these affect total computational time. However, from our work it is difficult to determine the minimum population size and num-

TABLE 6
SUMMARY OF RESULTS, PART II: MIXTURE OF LIGANDS

Ligand	Range of low-energy structures		
	Mixture population 761	Single population 761	Single population 112
Thioketal	−50 to −46	−48 to −45	−41 to −8
Vinyl	−58	−58 to −51	−36 to +1
Thio	−55 to −52	−53 to −49	−51 to −22
Ketal	−54 to −52	−54 to −55	−38 to −10
Alkyl	−55 to −50	−51 to −50	−43 to −22
Bromo	−54 to −55	−55	−55 to −42
Haloperidol	−57 to −55	−56 to −53	−50 to −28
N-oxide	−5 to −1	−32 to −27	−25 to +12

Range of lowest energy scores for eight ligands, either in a mixture or evaluated separately (labeled 'single' above), using different population sizes with multiple runs. Ligand acronyms are defined in the text; the structures are given in Fig. 2.

ber of runs required to generate the low-energy conformation. For every test system, different values seem appropriate. In general, there seems to be a population size, different for each system, where low-energy solutions (within 10–20% of the crystal energy) are likely to occur. Decreasing the population size decreases the likelihood of generating such low-energy structures with any one run and additional runs are needed to achieve a low-energy solution. However, with a small population (112) and a limited number of runs (10), the lowest energy GA solution was 20% higher than the crystal structure energy for the dhfr–mtx system. It is difficult to decide the balance between population size and number of computational experiments to efficiently sample the space for all systems.

The ts–fen test system required a large population and a long CPU time to accurately reproduce the crystal structure. On the other hand, GA solutions with energies comparable to the energy of the crystal structure, but differing in structure, were generated with small populations in significantly less time. This is probably due to the occurrence of many GA solutions with very similar energies. Note that the fen IC_{50} value is relatively high, indicating that K_i is probably relatively low. Correspondingly, the energy of the crystal structure is relatively high (-33 kcal mol^{-1}), so there can be many orientations and conformations of fen with very similar energies. In this case, GA does not preferentially converge to the crystal structure, which does not score significantly better than the other structures. A large GA population was needed to provide sufficient sampling to ensure that the crystal structure was produced. On the other hand, a small population and short amount of time were needed to reach the (relatively high) minimum energy.

Conformational search space

For our systems, low-energy GA solutions – not necessarily the crystal structure – are generated by evaluating only a small fraction of the total number of possible orientations and conformations, similar to other GA problems [36]. This can be seen in Table 7, which lists estimates for the number of possible orientations and conformations for each of the test systems, the product of the two and the (approximate) number of evaluations needed to reach a low-energy conformation. For our test systems, with the explicit-orientation-based method, the number of evaluations was on the order of the cube root of the number of possibilities.

For the explicit-orientation-based method, we estimate the size of the conformational and orientational search space used in Table 7 in the following manner. We assume that all unique conformations of the ligand can be represented by the six bits used to represent the precision of these angles, i.e., unique conformations which differ by approximately 5° increments. The number of conformations is then the number of possible values for each torsion angle, raised to the power of the number of rotatable bonds. In our case, it is 64 raised to the power 2, 5, or 8. To estimate the number of orientations, we assume that unique orientations differ by 0.3 Å, which is equal to our grid spacing. The number of unique translational values is then the bounds for each Cartesian direction divided by 0.3. To estimate the number of unique Euler angles, we further assume that the rigid portion of the ligand has a radius of approximately 5 Å. A rotation of approximately 0.06 rad would then move the end of the rigid part of the ligand by 0.3 Å. The number of unique orientation angles is thus given by the bounds for each Euler angle divided by 0.06.

For the sphere-based method, we estimate the size of the conformational and orientational search space in the following manner. Again, the number of conformations is the number of possible values for each torsion angle, raised to the power of the number of rotatable bonds. The orientational space is calculated to be the number of possible k-pairs of N spheres and M ligand atoms. Each sphere can be paired with more than one ligand atom. Hence, the size of the orientational space is on the order of $N^k * (M!/(M-k)!\ k!)$. Note that the solution space of the sphere-based method can be different from that of the explicit-orientation-based method, since different combi-

TABLE 7
CONFORMATIONAL SEARCH SPACE

Test system	No. of orientations	No. of conformers	Total possible	No. of GA evaluations
Sphere-based GA				
dhfr–mtx	1.7×10^{34}	2.8×10^{14}	4.8×10^{48}	9.4×10^{5}
ts–fen	4.4×10^{36}	4.1×10^{3}	7.5×10^{43}	2.4×10^{5}
hivp–thk	8.7×10^{34}	4.4×10^{12}	3.8×10^{47}	2.6×10^{5}
Explicit-orientation-based GA				
dhfr–mtx	4.1×10^{5}	2.8×10^{14}	1.1×10^{20}	5.4×10^{5}
ts–fen	7.5×10^{7}	4.1×10^{3}	3.1×10^{11}	1.0×10^{3}
hivp–thk	4.1×10^{7}	1.1×10^{9}	4.5×10^{16}	3.4×10^{4}

The table contains an estimate of the number of possible orientations and conformations and the number of GA evaluations needed to generate low-energy solutions.

nations of sphere–atom pairs can lead to the same solution.

Mixture of ligands

GA runs with a mixture of ligands produced low-energy structures for most of the ligands. Separate GA runs with individual ligands also generated low-energy solutions, provided that the same large population size was used. However, when the population size for these separate runs was decreased to the number of individual ligands present in the mixture, the low-energy solutions were not reliably generated. A larger population allows one to better sample the conformation space, so it is not surprising that the runs with the smaller population did not always generate the low-energy conformations. However, it is somewhat surprising that most of the ligands would reach low-energy conformations and orientations in the mixture. There must be a similarity in the orientation and conformation of all the ligands which is 'transferable'.

When the rigid portion of all ligands is the same, our tests show that GA can be used to evaluate a mixture of similar ligands simultaneously. This could result in a faster computational evaluation process. In this way, GA inherently offers an added advantage over other sequential screening methodologies.

The one ligand which did not reach a low energy in the mixture was the one containing the N-oxide group. This group is attached to the rigid portion of the molecule, and the ligand is the one ligand in the mixture which has a substituent on the rigid section. Because of this negatively charged group, the low-energy orientation of the rigid section for this ligand is slightly different from that of the other ligands. In the HIV protease system used by us, a chloride ion is present [58]. The presence of a negatively charged group causes a slight rotation and translation of the rigid portion of the ligand away from the chloride ion in the separate GA run. When this ligand is part of the mixture, it picks up the 'consensus' orientation of the other ligands. This places two similarly charged groups close together, resulting in a higher interaction energy in the mixture.

Variants of GA methodology

The GA methodology developed here considers primarily known binding ligands. Alternative protocols might be more appropriate when searching for novel leads. We consider variants of scoring and selection here.

Although the DOCK force field scoring [42] was used, other scoring methods could easily be adopted. For example, the DOCK shape scoring method [66] could be applied. A shape score might be more appropriate when the enzyme structure is poorly defined, as in the case where it is generated by homology modeling. In addition, different scoring methods could be used at different times, to try to capture conformations and orientations which simultaneously have a good fitness by both (or several) methods. Since gradients are not used with GA, this can be easily done.

Additionally, the selection of the new population could follow Boltzmann statistics. GA has, naturally, a population of 'particles'; in our case, each chromosome also has an interaction energy. To allow for a variety of structures, the 'temperature' can be set roughly equal to the average energy of the population; in this way, high-energy structures are possible. The 'temperature' of the system naturally lowers as better conformations and configurations are produced through succeeding generations. We have used this selection procedure in limited cases and found the GA populations to behave similarly to that of the standard GA selection procedures.

Conclusions

Two methods have been developed using the genetic algorithm to explore the orientational and conformation space of a flexible ligand within the active site of an enzyme. Both methods generate low-energy solutions comparable to the crystal structure binding mode. In addition, one method finds solutions with an rmsd accuracy of 1.0 Å/atom or less. With this method, the location and the orientation of the ligand must be restricted to a portion of the active site. This GA method can also be used to simultaneously evaluate several conformationally flexible similar compounds and identify potential ligands. This simultaneous evaluation of ligands can be far more rapid than sequential evaluation.

Acknowledgements

We gratefully acknowledge support from NIH Grants GM-31497 and GM-39552. We wish to thank R. Keenan and K. Perry for providing crystallographic coordinates, E. Meng for providing several CHEMGRID files, D. Gschwend for early release of the adapted RGDMIN program, N. Schraudolph for the GAucsd program, P. Turner for making the program xmgr available and the UCSF Computer Graphics Laboratory for the use of Midas. The GA library code used by the sphere-based method was written by Rich Humprey.

References

1 Kuntz, I.D., Science, 257 (1992) 1078.
2 Perun, T. and Propst, C.L. (Eds.) Computer-Aided Drug Design, Marcel Dekker, New York, NY, 1989.
3 Dolata, D.P., Leach, A.R. and Prout, K., J. Comput.-Aided Mol. Design, 1 (1987) 73.
4 Leach, A.R. and Prout, K., J. Comput. Chem., 11 (1990) 1193.
5 Clark, D.E., Jones, G., Willett, P., Kenny, P.W. and Glen, R.C., J. Chem. Inf. Comput. Sci., 34 (1994) 197.

6 Dammkoehler, R.A., Karasek, S.F., Shands, E.F. and Marshall, G.R., J. Comput.-Aided Mol. Design, 3 (1989) 3.

7 Hurst, T., J. Chem. Inf. Comput. Sci., 34 (1994) 190.

8 Judson, R.S., Jaeger, E.P., Treasurywala, A.M. and Peterson, M.L., J. Comput. Chem., 14 (1993) 1407.

9 McGarrah, D.B. and Judson, R.S., J. Comput. Chem., 14 (1993) 1385.

10 Payne, A.W.R. and Glen, R.C., J. Mol. Graph., 11 (1993) 74.

11 Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T., J. Mol. Biol., 161 (1982) 269.

12 Bacon, D.J. and Moult, J., J. Mol. Biol., 225 (1992) 849.

13 Lawrence, M. and Davis, P.C., Proteins, 12 (1992) 31.

14 Lin, S.L., Nussinov, R., Fischer, D. and Wolfson, H.J., Proteins, 18 (1994) 94.

15 Norel, R., Fischer, D., Wolfson, H.J. and Nussinov, R., Protein Eng., 7 (1994) 39.

16 Kasinos, N., Lilley, G.A., Subbarao, N. and Haneel, I., Proteins, 5 (1992) 69.

17 Kuhl, F.S., Crippen, G.M. and Friesen, D.K., J. Comput. Chem., 5 (1984) 24.

18 Lewis, R.A., Roe, D.C., Huang, C., Ferrin, T.E., Langridge, R. and Kuntz, I.D., J. Mol. Graph., 10 (1992) 66.

19 Böhm, H.J., J. Comput.-Aided Mol. Design, 6 (1992) 61.

20 Böhm, H.J., J. Comput.-Aided Mol. Design, 6 (1992) 593.

21 Rotstein, S.H. and Murcko, M.A., J. Med. Chem., 36 (1993) 1700.

22 Rotstein, S.H. and Murcko, M.A., J. Comput.-Aided Mol. Design, 7 (1993) 23.

23 Blaney, J., personal communication, 1993.

24 Gillet, V.J., Newell, W., Mata, P., Myatt, G., Sike, S., Zsoldos, Z. and Johnson, A.P., J. Chem. Inf. Comput. Sci., 34 (1994) 207.

25 Goodsell, D.S. and Olson, A.J., Proteins, 8 (1990) 195.

26 Goodsell, D.S., Lauble, H., Stout, C.D. and Olson, A.J., Proteins, 17 (1993) 1.

27 Hart, T.N. and Read, R.J., Proteins, 13 (1992) 206.

28 Stoddard, B.L. and Koshland Jr., D.E., Nature, 358 (1992) 774.

29 Stoddard, B.L. and Koshland Jr., D.E., Proc. Natl. Acad. Sci. USA, 90 (1993) 1146.

30 Leach, A.R. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 730.

31 Weiner, S.J., Kollman, P.A., Nguyen, D.T. and Case, D.A., J. Comput. Chem., 7 (1986) 230.

32 Weiner, S.J., Kollman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profeta Jr., S., and Weiner, P., J. Am. Chem. Soc., 106 (1984) 765.

33 Yamada, M. and Itai, A., Chem. Pharm. Bull., 41 (1993) 1203.

34 Davis, L., Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, NY, 1991.

35 Forrest, S., Science, 261 (1993) 872.

36 Goldberg, D., Genetic Algorithm in Search, Optimization and Machine Learning, Addison–Wesley, Reading, MA, 1989.

37 Miller, M.D., Kearsley, D.J., Underwood, D.J. and Sheridan, R.P., J. Comput.-Aided Mol. Design, 8 (1994) 153.

38 Geysen, H.M., Meloen, R.H. and Barteling, S.J., Proc. Natl. Acad. Sci. USA, 81 (1984) 3998.

39 Simon, R.J., Kania, R.S., Zuckermann, R.N., Huebner, V.D., Jewell, D.A., Banville, S., Ng, S., Wang, L., Rosenberg, S., Marlowe, C.K., Spellmeyer, D.C., Tan, R., Frankel, A.D., Santi, D.V., Cohen, F.E. and Bartlett, P.A., Proc. Natl. Acad. Sci. USA, 89 (1992) 9367.

40 Bunin, B.A. and Ellman, J.A., J. Am. Chem. Soc., 114 (1992) 109997.

41 Baum, R.M., Chem. Eng. News, 72 (1994) 20.

42 Meng, E.C., Shoichet, B.K. and Kuntz, I.D., J. Comput. Chem., 13 (1992) 505.

43 Holland, J.H., Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, MI, 1975.

44 Richards, F.M., Annu. Rev. Biophys. Bioeng., 6 (1977) 151.

45 Connolly, M.L., Science, 221 (1983) 709.

46 Schraudolph, N., GAucsd, Version 1.4, University of California, San Diego, CA, 1992.

47 Meng, E.C., Gschwend, D.A., Blaney, J.M. and Kuntz, I.D., Proteins, 17 (1993) 268.

48 Grefenstette, J., Genesis, Version 4.5, Naval Research Laboratory, Washington, DC, 1990.

49 Marsili, M. and Gasteiger, J., Croat. Chim. Acta, 52 (1980) 601.

50 Gasteiger, J. and Marsili, M., Tetrahedron, 36 (1980) 3210.

51 SYBYL, Version 6.02, Tripos Associates, St. Louis, MO, 1993.

52 Fletcher, R., Practical Methods of Optimization: Unconstrained Optimization, Vol. 1, Wiley, New York, NY, 1981.

53 Blakley, R.L., The Biochemistry of Folic Acid and Related Pteridines, Wiley, New York, NY, 1969.

54 Kraut, J. and Mathews, D.A., In Jurnak, F.A. and McPherson, A. (Eds.) Biological Macromolecules, Wiley, New York, NY, 1987, pp. 1–72.

55 Stryer, L., Biochemistry, 2nd ed., W.H. Freeman, San Francisco, CA, 1983, p. 527.

56 Wlodawer, A. and Erickson, J.W., Annu. Rev. Biochem., 62 (1993) 543.

57 Shoichet, B.K., Stroud, R.M., Santi, D.V., Kuntz, I.D. and Perry, K.M., Science, 259 (1993) 1445.

58 Rutenber, E., Fauman, E.B., Keenan, R.J., Fong, S., Furth, P.S., Ortiz de Montellano, P.R., Meng, E., Kuntz, I.D., DeCamp, D.L., Salto, R., Rose, J.R., Craik, C.S. and Stroud, R.M., J. Biol. Chem., 268 (1993) 15343.

59 Bolin, J.T., Filman, D.J., Mathews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1992) 13650.

60 Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.

61 Molecular Design Limited, Fine Chemicals Directory San Leandro, CA.

62 Dixon, J.S., In Wermuth, C.G. (Ed.) Trends in QSAR and Molecular Modelling 92 (Proceedings of the 9th European Symposium on Structure–Activity Relationships: QSAR and Molecular Modelling), ESCOM, Leiden, 1993, pp. 412–413.

63 Fiorini, P. and Inselberg, A., In Proceedings of the IEEE International Conference on Robotics and Automation, IEEE Computer Society Press, Scottsdale, AZ, 1989, pp. 1215–1220.

64 Inselberg, A. and Dimsdale, B., In Proceedings of the First IEEE Conference on Visualization, IEEE Computer Society Press, San Francisco, CA, 1990, pp. 361–378.

65 Judson, R.S., Jaeger, E.P. and Treasurywala, A.M., J. Mol. Struct., 114 (1994) 191.

66 DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 31 (1988) 722.