JMB

Development and Validation of a Genetic Algorithm for Flexible Docking

Gareth Jones[1]*, Peter Willett[1], Robert C. Glen[2], Andrew R. Leach[3] and Robin Taylor[4]

[1]*Department of Information Studies and Krebs Institute for Biomolecular Research, University of Sheffield, Western Bank, Sheffield S10 2TN, UK*

[2]*Department of Physical Sciences, Wellcome Foundation Research Laboratories Beckenham, Kent BR3 3BS UK*

[3]*Glaxo Wellcome Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, UK*

[4]*Cambridge Crystallographic Data Centre, 12 Union Road Cambridge CB2 1EZ, UK*

**Corresponding author*

Prediction of small molecule binding modes to macromolecules of known three-dimensional structure is a problem of paramount importance in rational drug design (the ''docking'' problem). We report the development and validation of the program GOLD (Genetic Optimisation for Ligand Docking). GOLD is an automated ligand docking program that uses a genetic algorithm to explore the full range of ligand conformational flexibility with partial flexibility of the protein, and satisfies the fundamental requirement that the ligand must displace loosely bound water on binding. Numerous enhancements and modifications have been applied to the original technique resulting in a substantial increase in the reliability and the applicability of the algorithm. The advanced algorithm has been tested on a dataset of 100 complexes extracted from the Brookhaven Protein DataBank. When used to dock the ligand back into the binding site, GOLD achieved a 71% success rate in identifying the experimental binding mode.

Keywords: docking problem; genetic algorithm; molecular recognition; protein ligand docking

Introduction

Protein binding sites exhibit highly selective recognition of small organic molecules, in that evolution has equipped them with a complex three-dimensional ''lock'' into which only specific ''keys'' will fit. This has been exploited by medicinal chemists in the design of molecules selectively to augment or retard biochemical pathways and so exhibit a clinical effect. X-ray crystallography has revealed the structure of a significant number of these binding sites. It would be advantageous in attempting the computer-aided design of therapeutic molecules to be able to predict and to explain the binding mode of novel chemical entities (the ''docking'' problem) when the active site geometry is known.

Any solution to the docking problem requires both a powerful search technique to explore the conformation space available to the protein and the ligand and a good understanding of the process of molecular recognition to devise scoring functions that can reliably predict binding modes. Furthermore, since many putative dockings will require evaluation before elucidating the binding mode, any scoring function must be rapid in operation.

There are currently many different approaches to solving the docking problem (Blaney & Dixon, 1993; Jones & Willett, 1995). Early approaches to ligand docking consider both protein and ligand to be rigid, as typified by the DOCK program (Kuntz *et al*., 1982). Since the bioactive conformation of a bound ligand rarely corresponds to the isolated ligand X-ray structure (Nicklaus *et al*., 1995), recent techniques have dealt with the issue of conformational flexibility. Deterministic approaches include the FLOG system of Miller *et al*. (1994) and FLEXX of Rarey *et al*. (1996). The latter algorithm is very efficient and has been verified on 19 protein-ligand complexes. Alternative, stochastic sampling techniques include genetic algorithms (Jones *et al*., 1995a; Judson *et al*., 1994; Oshiro *et al*., 1995), simulated annealing (Goodsell & Olsen, 1990) and evolutionary programming (Gehlhaar *et al*., 1995).

Inspection of the X-ray crystallographic structures of proteins with associated high-affinity

Present address: R. C. Glen, Tripos Inc., 1699 South Hanley Road, St Louis, MO 63144, USA

Abbreviations used: GA, genetic algorithm; PDB, Protein Data Bank; DHFR, dihydrofolate reductase; CSD, Cambridge Structural Database; IMPT, intermolecular perturbation theory.

0022–2836/97/130727–22 $25.00/0/mb960897

ligands reveals that the ligands appear to conform closely to the shape of the binding cavity, maximising the hydrophobic contribution to binding, and to interact at a number of hydrogen bonding sites. The optimal binding mode may thus involve the ligand forming hydrogen bonds at key hydrogen-bonding sites, accompanied by hydrophobic surface area burial. The most significant contributions to apolar surface area burial are likely to be dispersive interactions between protein and ligand atoms together with an entropic contribution from the displacement of ordered water from the active site into the solvent. Sufficiently accurate simulation of many of these interactions may be enough to predict the binding mode of the majority of high-affinity ligands.

We have reported the use of a genetic algorithm, hereinafter a GA (Davis 1991; Goldberg, 1989; Holland, 1992) to perform protein docking (Jones *et al*., 1995a), where an evolutionary strategy is employed to explore the conformational variability of a flexible ligand while simultaneously sampling available binding modes into a partially flexible protein active site. The GA provides a search paradigm that enables the rapid identification of good, though not necessarily optimal, solutions to combinatorial optimisation problems. Of particular interest is the use of GAs in performing conformational analysis of both small molecules (Jones at al., 1995b; Brodmeirer & Pretsch, 1994; Clark *et al*., 1994) and macromolecules (Dandekar & Argos, 1996; Sun 1993).

Here, we describe a docking program called GOLD (Genetic Optimisation for Ligand Docking) that is based on the algorithm described by (Jones *et al*., 1995a). GOLD performs automated docking with full acyclic ligand flexibility, partial cyclic ligand flexibility and partial protein flexibility in the neighbourhood of the protein active site. In order to search the space of available binding modes efficiently, hydrogen bond motifs have been directly encoded into the GA. A simple scoring function was used to rank generated binding modes. This comprised a term for hydrogen bonding (which took account of the fundamental requirement that water must be displaced from both donor and acceptor before a bond is formed); a pairwise dispersion potential that was able to describe a significant contribution to the hydrophobic energy of binding; and a molecular mechanics term for the internal energy of the ligand. The original algorithm has now been substantially enhanced, as detailed in Materials and Methods. The resulting algorithm has been tested on a number of complex ligands and the result of docking NADPH into dihydrofolate reductase (DHFR) is reported here as an example of the power of this technique. In order to probe the strengths and weaknesses of GOLD in a more rigorous manner, 100 protein-ligand complexes were selected from the Protein Data Bank (PDB: Bernstein *et al*., 1977). These complexes were selected on the basis of pharmacological interest and whether or not the ligands involved were ''drug like''. The result was a varied and demanding test set of complexes. We report here the results obtained by using GOLD to predict the binding modes for these test complexes and compare these predictions against the crystallographically observed binding modes.

Results

The GA described in Materials and Methods required as input the approximate size and location of the active site, together with coordinates of the protein and a ligand conformation. As GOLD used a cavity detection procedure to further define the active site, the size and location input by the user was not critical. Although the determination of the active site is not currently automated, there are techniques available that are capable of predicting the location of the active site with considerable accuracy (Peters *et al*., 1996). The output was the ligand and protein conformations associated with the fittest chromosome in the population when the GA run terminated. Since GAs are non-deterministic algorithms, 20 GA runs were performed in each docking experiment in order to ensure that most high-affinity binding modes would be explored. In most cases similar results were obtained and, in general, the algorithm does not need to be run 20 times to elucidate a binding mode. In the examples that follow, the best solution refers to the predicted binding mode (i.e. the result from the 20 runs that had the highest GA fitness score) and not to the solution that was closest to that observed in the crystal structure. Depending on the complexity of the docking problem, each GA run would take anything between three and 35 minutes on a Silicon Graphics R4400 Indigo II workstation. Due to the rapid and sustained increase in CPU performance, computing time can be expected to reduce substantially in the future.

In order to illustrate the effectiveness of the technique we first describe the docking of NADPH into DHFR. Following this introductory example, the results of testing GOLD on a dataset of 100 complexes are presented in detail.

A number of results are illustrated in the colour Figures, where the experimental result is shown coloured by atom type and the GA prediction is shown in red. For clarity, only a few protein residues are displayed.

NADPH in dihydrofolate reductase

Dihydrofolate reductase (DHFR) catalyses the NADPH-linked reduction of folate to dihydrofolate then onto tetrahydrofolate. The three-dimensional structure of the ternary complex of DHFR with NADPH and the anti-cancer drug methotrexate has been determined by X-ray crystallography to a resolution of 1.7 Å (Bolin *et al*., 1982) and the coordinates have been deposited in the PDB (entry 3DFR). The NADPH cofactor is a large and fairly complex molecule and thus provided a demanding

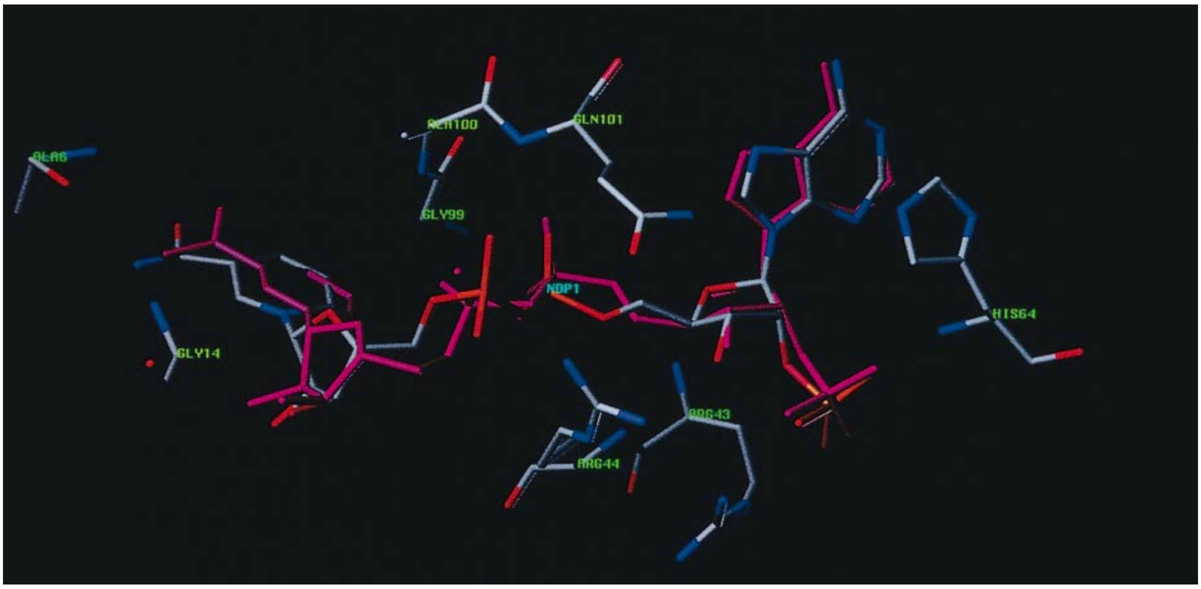


Figure 1. Docking of NADPH into dihydrofolate reductase.

test system for GOLD. Methotrexate, NADPH and all water molecules were removed from the ternary complex. Using GOLD, NADPH was docked back into DHFR and the algorithm's predictions were compared with the experimentally observed binding mode.

This is an extremely complicated problem: acyclic flexibility of NADPH was described using 17 rotatable bonds, while cyclic flexibility was accounted for by ten free corners. The active site was determined by flood-filling to a radius of 12 Å. After accounting for the probe radius, atom size and the final flood-fill from the detected cavity (see Materials and Methods for more details), this roughly corresponded to solvent-accessible protein cavity atoms in a sphere of radius 15 Å around the binding cleft. Of the 20 dockings, the solution with the highest fitness score is shown in Figure 1. Given the difficulty of the problem, this is an extremely powerful prediction, having a root-mean-square (r.m.s.) deviation between predicted and experimental heavy-atom coordinates of 1.2 Å. In the remaining 19 GOLD solutions there were four solutions with a r.m.s. deviation of under 1.5 Å and the average r.m.s. deviation over all 20 solutions was 2.6 Å. The solution that was closest to the crystal structure was ranked fourth and had an r.m.s. deviation of 1.1 Å.

Because of the many degrees of freedom in the NADPH-DHFR system this calculation was relatively time-consuming with the average GA run taking 26 minutes and 44 seconds on an R4400 CPU.

Experiments on the dataset of 100 PDB complexes

In order to achieve insight into the strengths and deficiencies of GOLD a dataset of 100 protein ligand complexes was selected from the PDB for the purposes of evaluation. This selection was done by one of us (R.T.) who was not involved in the development of the algorithm. These complexes were selected on the basis of pharmacological interest, with preference being given to "drug-like" molecules and to ligands that formed interesting or unusual interactions with the protein. The test set was highly varied: the number of heavy atoms in the ligand varied between six and 55 while the number of rotatable bonds in the ligand varied between zero and 30; and there were many different types of protein, including a number of metalloenzymes. We believe that the dataset is an extremely demanding testset for any docking technique, and it is also much larger than the datasets used in previously reported research. Those docking programs that have been tested on a number of complexes (but still fewer than the results presented here) include the GA for flexible docking used by Judson *et al.* (1995) which has been tested on ten complexes and the FLEXX algorithm that has been tested on 19 complexes (Rarey *et al.*, 1996).

The PDB codes of 99 of the complexes are listed in the third column of Table 1. The final complex was 1ACL (Harel *et al.*, 1993). The ligand in this complex (decamethonium) has no polar group and GOLD was thus unable to make a prediction for the binding mode of this ligand, since the principle driving force of our algorithm is the identification of hydrogen-bonding interactions between the ligand and the protein.

Preparation of the dataset

During preparation of the input structures (as described in Materials and Methods) considerable care was taken over the correct assignment of protonation and tautomeric states of both the protein and the ligand. Of particular importance in pro-

Table 1. Results of docking predictions on dataset of 100 complexes.

Subjective result	Number	PDB identification code									
Good	41	1ABE	1ACM	1ACO	1CBX	1COY	1CPS	1DBB	1DBJ	1FKG	1FKI
		1HDY	1HEF	1HYT	1LST	1MDR	1MRK	1PBD	1PHD	1POC	1SRJ
		1STP	1TPP	1ULB	1XIE	2ADA	2CGR	2CHT	2CTC	2PHH	2SIM
		3AAH	3PTB	3TPI	4DFR	4PHV	7TIM	8GCH	1AEC	1AHA	1ASE
		1HSL									
Close	30	1BLH	1DIE	1DR1	1DWD	1EPB	1GHB	1GLQ	1IDA	1IVE	1LDM
		1PHA	1PHG	1RNE	1SLT	1TKA	1TMN	1XID	2DBL	2PK4	2YHX
		3CPA	3GCH	3HVT	4CTS	5P2P	6ABP	6RNT	1APT	1AZM	4EST
Errors	9	1BAF	1EAP	1ETR	1HDC	1LIC	1RDS	1ROB	6RSA	1ACK	
Wrong	19	1AAQ	1ACJ	1DID	1EED	1ETA	1HRI	1ICN	1IGJ	1MCR	1MUP
		2R07	1NIS	1TDB	2AK3	2MTH	2PLV	3CLA	4FAB	2MCP	

teins are the ionisation and tautomeric states of histidine residues, which can be positively charged, neutral with a proton on N^{δ}, neutral with a proton on N^{ε} or even negatively charged and deprotonated. Also aspartic and glutamic acid, though generally ionised, can occasionally be protonated on either oxygen atom. In the case of aspartic proteases a proton was added to one of the two proximal, catalytic Asp residues, in the position where it seemed to form the best possible hydrogen bond to the neighbouring Asp. In other cases, the oxygen atom that was more solvent-exposed was normally protonated. When preparing the ligand, thought was given as to the charge state of basic nitrogen, acidic oxygen and acidic nitrogen atoms. If necessary, literature cited in the PDB file was consulted to help determine ionisation states. Atom typing of the ligand was particularly awkward in the case of 3HVT (Wang *et al.*, 1994). Here, a normally trigonal cyclic ligand nitrogen atom was stressed to a non-planar conformation and it was not clear if this atom could act as a nitrogen acceptor. However, the analysis by Allen *et al.* (1995) indicated that the nitrogen geometry was not sufficiently distorted for the ligand to accept hydrogen bonds. The ligands in complexes 1EED (Cooper *et al.*, 1992) and 1IGJ (Jeffrey *et al.*, 1993) were not complete and, after consulting the literature, the remainder of the ligands was created using the SYBYL BUILD module (Clark *et al.*, 1989) prior to docking.

When preparing the protein, all water molecules were normally removed. However, the following exceptions were made for some metal ion complexes: 2CTC (Teplyakov *et al.*, 1994) where the ligand is hydrogen-bonded to a water molecule co-ordinated to a zinc ion; 1MDR (Landro *et al.*, 1994) where a water molecule was tightly bound to a Mg ion; and 1NIS (Lauble *et al.*, 1994), where a water molecule is co-ordinated by a Fe ion. In the case of 1HEF (Murthy *et al.*, 1992) the whole of the active site was created by applying crystallographic symmetry operators.

A check was made of each ligand structure to see if conformational flexibility of ligand rings was likely to be observed. If this was the case then GOLD was run with corner flipping turned on, as in the following test systems: 1FKI, 1IGJ, 1MRK, 1RDS, 1TDB, 2ADA, 2AK3 and 6RSA. In 1FKI (Holt *et al.*, 1993) the ligand, FK506, has a large macrocycle for which corner flipping is inadequate (corner flipping flips one atom while holding neighbouring cyclic atoms fixed and hence will not significantly deform large macrocycles): in this test system the docking was therefore only partially flexible. This system is further complicated by the fact that the amide bond in FK506 appears to flip between planar-*cis* and planar-*trans* on binding to the protein (Van Duyne *et al.*, 1993). While this change would not be predicted by GOLD, it is not surprising given the rotamase activity of the FKBP protein.

The algorithm requires that the user specify the rough size and location of the active site. On average, flood-fill and cavity detection was performed to a radius of 10.2 Å (roughly corresponding to all solvent-accessible cavity atoms within a radius of 13 Å of the active site centre), though the radius of flood-fill varied between 5 Å and 15.5 Å, depending on the size of the active site. In all cases the specified size was sufficient to encompass the whole of the active site. Indeed, in many cases where the algorithm failed, GOLD had placed the ligand in a different pocket or in a completely different area of the active site.

Results

For each test system the GA was run 20 times. The solution with the highest GA fitness score was then compared with the crystallographically observed binding mode. Depending on how close the predicted binding mode was to the crystallographically observed binding mode the result was assigned to one of four subjective categories. The first, *good*, was for those predictions where the binding mode, all hydrogen-bonding and metal co-ordination interactions and other close contacts between the protein and the ligand were reproduced correctly. If an acceptable result was generated with the ligand binding mode reproduced, but with some displacement of ligand groups from the experimental result, the GA prediction would be assigned to the *close* category. A third category, *errors*, was used for those predictions that were partially correct but contained significant errors. Finally, the fourth category, *wrong*, was used for

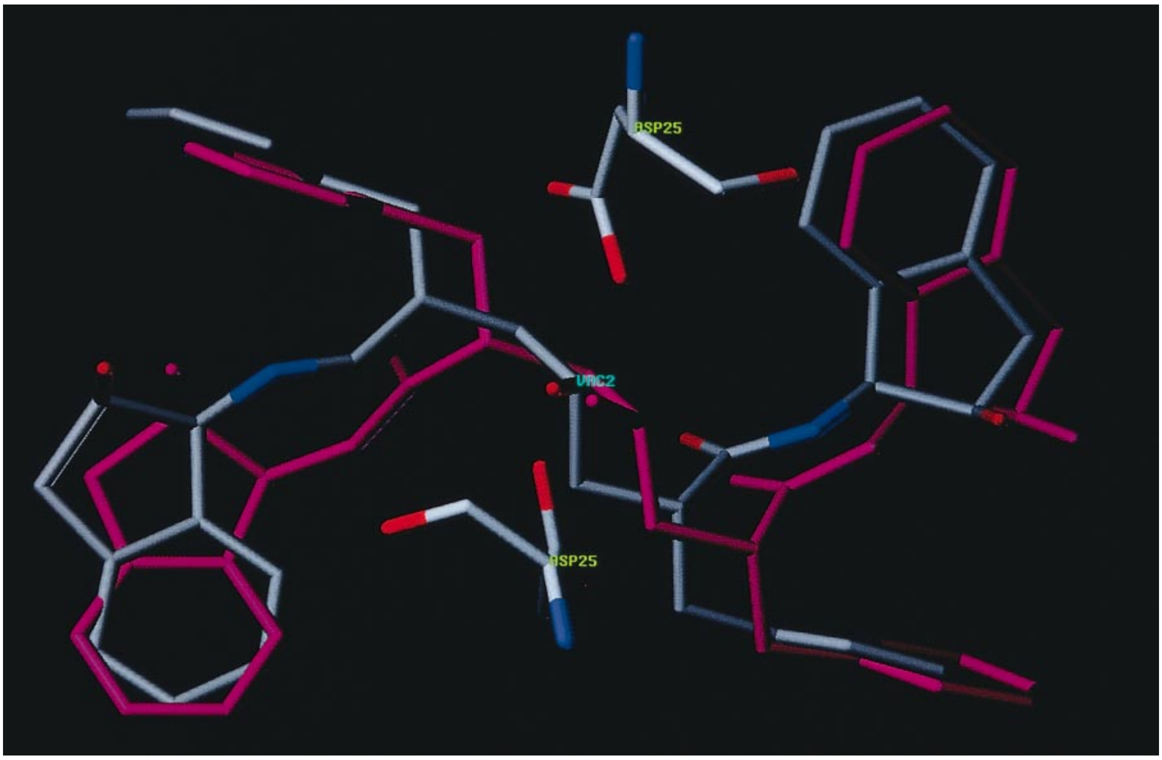


Figure 2. 4PHV. Example of *good*. A peptide-like ligand docked into HIV protease.

completely incorrect predictions. These categories are illustrated in Figures 2 to 5.

Figure 2 shows an example of good, with all ligand groups correctly positioned. An example of the *close* category is shown in Figure 3, where the ligand binding mode is clearly reproduced, though the nitrophenyl moiety and acid groups are displaced from the experimental result. Figure 4 shows an example of the *errors* category. While the benzene ring and phosphonate group are correctly positioned, GOLD has identified a different interaction with the acid group and there is some displacement of the alkyl side-chain. A *wrong* prediction is shown in Figure 5, where GOLD has completely failed to predict the binding mode of oleate within a fatty acid binding protein. When assigning solutions to each category an allowance was made for solvated, and thus mobile, ligand groups (for example in 4DFR, the ligand methotrexate, has a solvent- exposed acid group that is

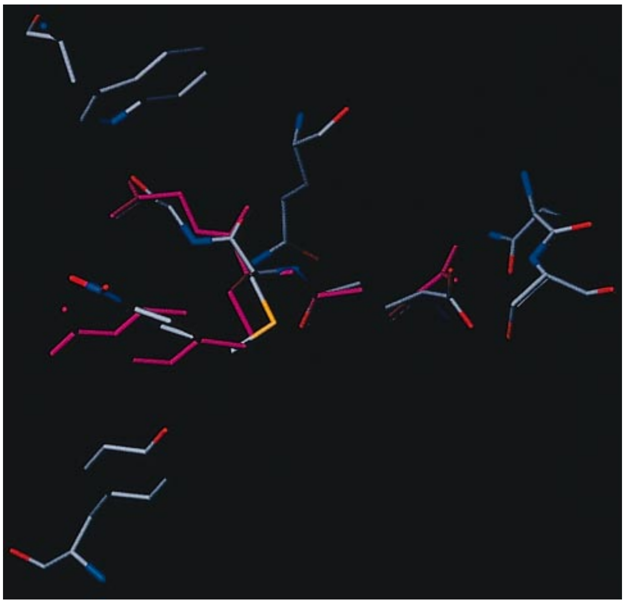


Figure 3. 1GLQ. Example of close. A nitrophenyl-substituted peptide ligand docked into glutathione S transferase.

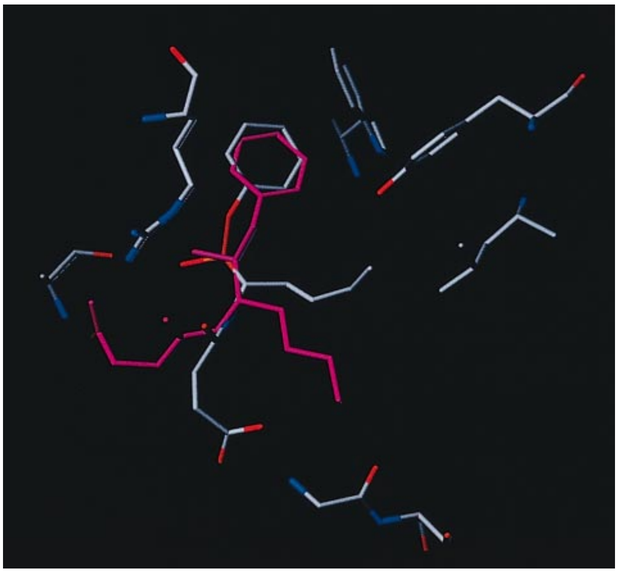


Figure 4. 1EAP. Example of *errors*. A succinylamino phosphonate ligand docked into an antibody.

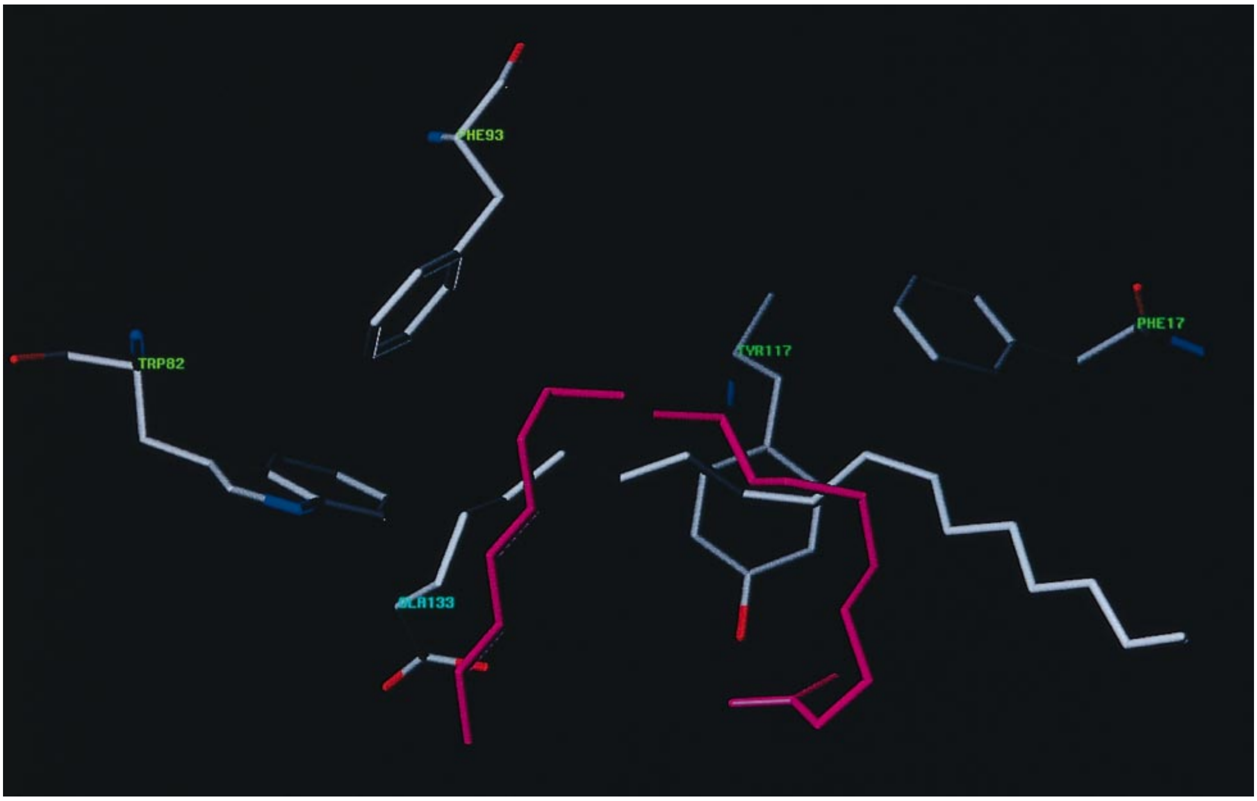


Figure 5. 1ICN Example of *wrong*. Oleate docked into a fatty acid binding protein.

highly mobile in the GOLD solutions) and for those cases where the experimentally observed geometry of the ligand appeared unreasonable (see the descriptions for 3GCH and 1IVE below).

Table 1 shows the complexes assigned to each category. If we consider acceptable results to be contained in the *good* and *close* categories, then GOLD achieved a 71% prediction rate. On average, each GA run (of which 20 were performed for each of the 100 complexes) took 12 minutes and ten seconds. The longest run took 34 minutes and three seconds, while the shortest run was two minutes and 52 seconds. The length of each GA run was primarily dependent on the size and flexibility of the ligand.

Table 1 shows that the best (i.e. top scoring) answer of 20 GA runs was essentially correct (i.e. *good* or *close*) for 71 of the test structures. The question arises of whether the correct answer can usually be found in less than 20 runs. Table 2 shows how many of the 71 complexes are predicted correctly after the first two, the first five and the first ten runs. It can be seen that as many as 49 correct answers are obtained after only two runs (therefore, 1/10 of the CPU time) and 63 complexes have had their binding modes predicted correctly after five runs. Thus, in general, GOLD does not require 20 GA runs to elucidate a binding mode. Note that 4EST was correctly predicted after five GA runs but incorrectly predicted after ten runs.

As an alternative to subjective analysis, r.m.s. deviations between heavy atoms in the prediction and experimental result may be used. The subjec-

Table 2. Complexes correctly predicted after 20 GA runs, that are incorrectly predicted after 2, 5 or 10 runs

Number of GA runs	Number of complexes correctly predicted	Complexes incorrectly predicted				
2	49	1DBB	1DBJ	1FKG	1HDY	1XIE
		2ADA	4DFR	4PHV	8GCH	1DWD
		1EPB	1GLQ	1IDA	1PHG	1RNE
		1TMN	1XID	2DBL	5P2P	6RNT
		1APT	4EST			
5	63	1DWD	1EPB	1GHB	1GLQ	1RNE
		5P2P	6RNT	1APT		
10	65	1DWD	1EPB	1GLQ	1RNE	1APT
		4EST				

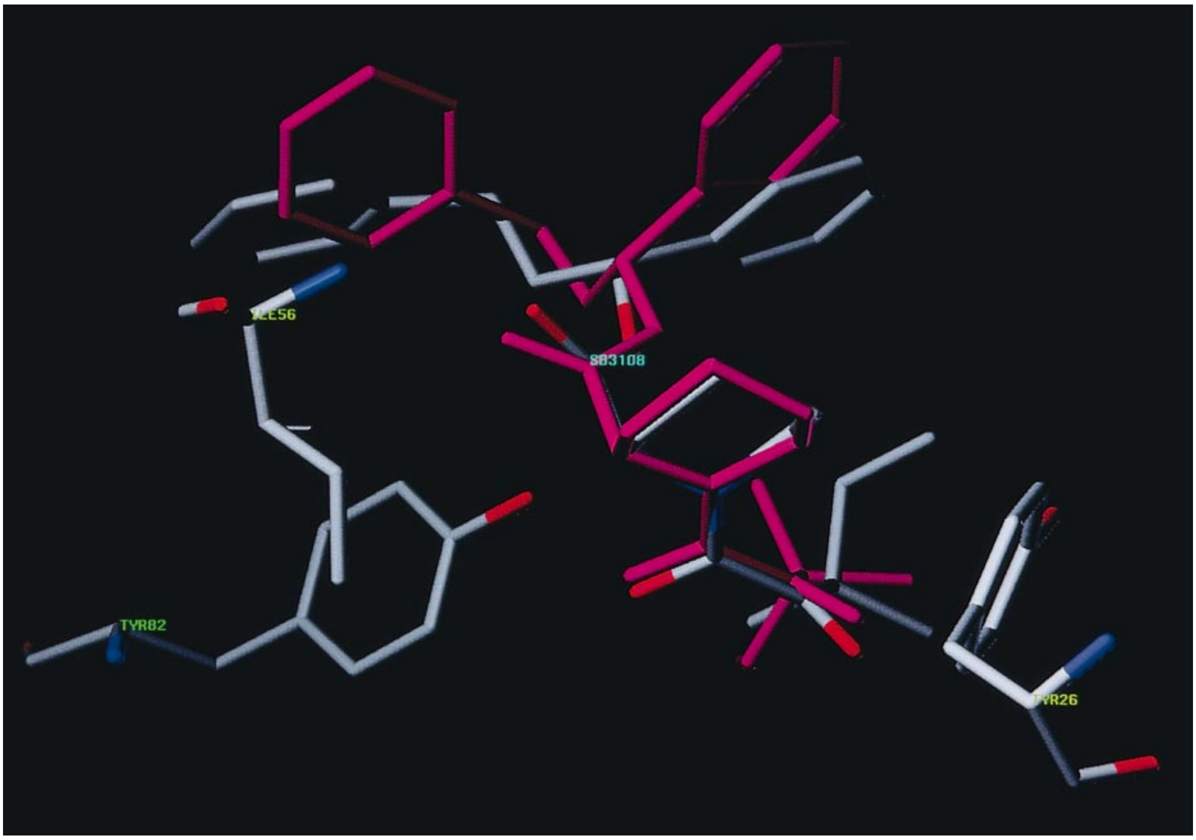


Figure 6. 1FKG. Docking of a rotamase inhibitor to FK506BP.

tive characterisation was preferred, since a small r.m.s. result may mask a significant error, while a reasonable result may have a large r.m.s. score if near-symmetrical groups are flipped. For example the GOLD prediction for 1FKG had a relatively high r.m.s. deviation of 1.81 Å, while the result, illustrated in Figure 6, was clearly excellent. The high r.m.s. deviation resulted from unimportant differences in solvent-exposed groups and GOLD rotating a bond so that a methyl and an ethyl group were exchanged, relative to the experimental binding mode. A summary of r.m.s. values is given in Table 3, where it can be seen that 66 complexes had r.m.s. deviations of 2.0 Å or less, while 71 had r.m.s. deviations of 3.0 Å or less. The one *close* result with an r.m.s. deviation greater than 3.0 Å is for 1IDA (Tong *et al.*, 1995), where a peptide-like ligand bound in HIV protease was predicted to bind in the same conformation as in the experimental result, but rotated by 180° around the protease 2-fold axis. Given the 2-fold symmetry of the active site this was felt to be an acceptable result, although the alternative binding mode predicted by GOLD is not the observed binding mode (the two alternate binding modes are shown in Figure 4 of Tong *et al.*, 1995). In 1HEF, another HIV protease complex, the ligand is observed in two distinct binding modes (Murthy *et al.*, 1992); the r.m.s. value calculated here was for the ligand orientation closer to the GOLD prediction.

Analysis of ligand composition

In order to try and predict when the algorithm is likely to succeed or fail, the performance of the algorithm was analysed in terms of ligand flexibility, size and polarity. The algorithm was considered to have succeeded in the case of a *good* or *close* prediction, and to have failed if an *errors* or *wrong* prediction was obtained. Table 4 shows how success or failure varies with the minimum, maximum and average values for the number of ligand heavy atoms, the percentage of polar atoms and the number of ligand torsions and free corners. Inspection of the Table shows that the GA is more likely to fail if the ligand is large or highly flexible. This is not surprising, as both these qualities are indicators of the complexity of the problem. However, it is clear from the Table that the algorithm is capable of predicting the binding mode of both large and highly flexible ligands. Lastly, an analysis of the proportion of polar atoms in the ligand shows that the GA is more likely to succeed if the ligand is polar. Given the design of the algorithm and the fitness function used this is not at all surprising. In fact, GOLD is most likely to fail given a large, flexible, hydrophobic ligand. There are three fatty-acid

Table 3. Summary of r.m.s. deviations between predictions and experimental results.

r.m.s.	*Total*	*Good*	*Close*	*Errors*	*Wrong*
$\leqslant 0.5$	8	8	0	0	0
$>0.5, \leqslant 1.0$	27	24	3	0	0
$>1.0, \leqslant 1.5$	20	7	13	0	0
$>1.5, \leqslant 2.0$	11	2	9	0	0
$>2.0, \leqslant 2.5$	2	0	2	0	0
$>2.5, \leqslant 3.0$	3	0	2	1	0
>3.0	28	0	1	8	19

like ligands in the dataset (1LIC, 1ICN and 2PLV) and GOLD fails in all cases. In order to estimate the statistical significance of these indicators, a χ^2 test was applied to the three variables in Table 4. For the number of ligand heavy atoms $\chi^2 = 0.33$ ($p = 0.85$, $\nu = 1$), for the number of ligand rotatable bonds and free corners $\chi^2 = 1.84$ ($p = 0.20$, $\nu = 1$) and for the proportion of polar atoms in the ligand $\chi^2 = 3.98$ ($p < 0.05$, $\nu = 1$). Thus, the statistical evidence would indicate that the effectiveness of GOLD is not too related to ligand size or flexibility but that performance of the algorithm is highly dependent on ligand hydrophobicity.

Problems in protein structure

In any docking experiment it is required that the co-ordinates of the active site be known to reasonable accuracy. Intuitively, it would not be surprising if the algorithm started to fail on poor-quality protein structures. There are two main reasons for this. Firstly, poor resolution means that the experimental result is less precise, i.e. the ligand might be wrongly positioned by the crystallographer. Secondly, one of the reasons for a poor resolution might be high thermal motion or even disorder, in which case the ligand might be highly mobile, with no clearly defined binding mode. The complexes in the dataset varied considerably in resolution. Table 5 shows how the subjective result obtained varied with resolution of the protein structure. An analysis of the Table shows that if the protein structure had a resolution of 2.5 Å or better then GOLD achieved a prediction rate of 77%. However, if the resolution was poorer than 2.5 Å then the prediction rate dropped to 52%. A χ^2 test on these data indicated that there is less than a 5% probability of this distribution arising through chance ($\chi^2 = 4.91$, $p < 0.05$, $\nu = 1$).

Another problem frequently encountered in protein structures containing small molecules is the poorly determined geometry of ligand groups. One example is 1APT (resolution 1.8 Å, James *et al.*, 1983) where a carbon atom in an ester group has tetrahedral rather than planar geometry. In 1TDB (resolution 2.65 Å, Perry *et al.*, 1993) a normally tetrahedral carbon has planar geometry. In both of these cases the minimised structure docked by GOLD was significantly different from the bound structure. In spite of this, GOLD did fairly well in the case of 1APT but failed to predict the binding mode of the ligand in 1TDB. The ligand in complex 1HEF (resolution 2.2 Å, Murthy *et al.*, 1992) had a bad bump involving two methyl groups and an unusual ester conformation.

In complex 1IVE (resolution 2.4 Å, Jedrzejas *et al.*, 1995) the benzoic acid ligand, 4-(acetylamino)-3-aminobenzoic acid, is unusual in that the ligand has a non-planar (almost 90°) amide group and a very short contact is seen between the terminal O of Asp151 and the ligand benzene ring. The minimised structure docked by GOLD clearly had a *trans* amide bond. Despite this, GOLD did fairly well, with the plane of the benzene ring, the acid group and the amide substituent in the correct position. However, in the GOLD prediction the benzene ring was flipped (moving the amino group), in order to avoid a bad internal steric clash with the amide group (this does not happen in the crystal structure because of the 90° amide bond). This was felt to be an acceptable result and the prediction was assigned to the *close* category, because of the unusual problems in docking this ligand.

In 3GCH (resolution 1.9 Å, Stoddard *et al.*, 1990) the bond angle between the ligand and the protein, through the terminal oxygen atom of Ser195, is only 82°. Since GOLD has an angle-bending potential for covalently bound ligands it will not reproduce this geometry. However, given the circumstances, GOLD does a reasonable job and positions the benzene ring of the cinnamic ester between two portions of the backbone, as seen in the crystal structure.

Table 4. Maximum, average and minimum values for the number of ligand heavy atoms, number of ligand torsions and free corners and percentage of ligand polar atoms

	Heavy atoms			Torsions and free corners			% H-bonding		
Result	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min
Good + Close	52	20.4	6	28	7.9	0	66.7	31.9	8.7
Errors + Wrong	55	24.3	9	40	11.4	0	53.9	25.1	4.8

Table 5. Protein structure resolution and subjective results obtained

Resolution	Total	*Good* + *Close*	*Errors* + *Wrong*
>1.0, ⩽1.5	2	2	0
>1.5, ⩽2.0	44	34	10
>2.0, ⩽2.5	32	24	8
>2.5, ⩽3.0	20	11	9
>3.0	1	0	1

In 1BLH (resolution 2.3 Å) a phosphonate inhibitor is covalently bound to β-lactamase (Chen *et al.*, 1993). In Figure 7 the crystal structure is shown coloured by atom type and the GOLD prediction is shown in red. The significant difference between the GA prediction and the experimental result is in the ester torsion angle of the carbamate group. In the crystal structure the C-O-C=O torsion angle is *trans* (172°), whereas in the prediction it is *cis* (1.4°). A search of the Cambridge Structural Database (CSD: Allen *et al.*, 1991) reveals 418 small molecule crystals containing 491 acyclic torsions of this type. In all cases the torsion is *cis*. It is worth noting that the torsional potential used in GOLD will not distinguish between these two conformations and in the 20 solutions generated, some of the low scoring predictions were identical with the crystal structure. It is notable that the crystal structure contains a torsion angle not observed in small molecule crystallography, while GOLD, with no bias against that torsion, predicts a binding mode that is consistent with the small molecule observations. It is interesting to speculate if the GOLD binding mode would be consistent with the experimental structure factors. Unfortunately, the phenyl group in the GOLD solution makes a close contact with a crystallographic water molecule (two carbon atoms are 2.4 Å from the water molecule, which was removed from the active site prior to docking). Although the water molecule is not hydrogen-bonded to the protein, it's presence may argue against the GOLD prediction.

Analysis of wrong *solutions*

In order to gain an understanding of why GOLD fails, all the solutions in the *errors* or *wrong* categories were examined in detail to see if the cause for failure could be identified.

In order for GOLD to work properly it is normally required that the ligand be hydrogen bonded to the binding site. However, it was found that this was not always the case. It is obvious that this assumption was misplaced in the case of 1ACL,

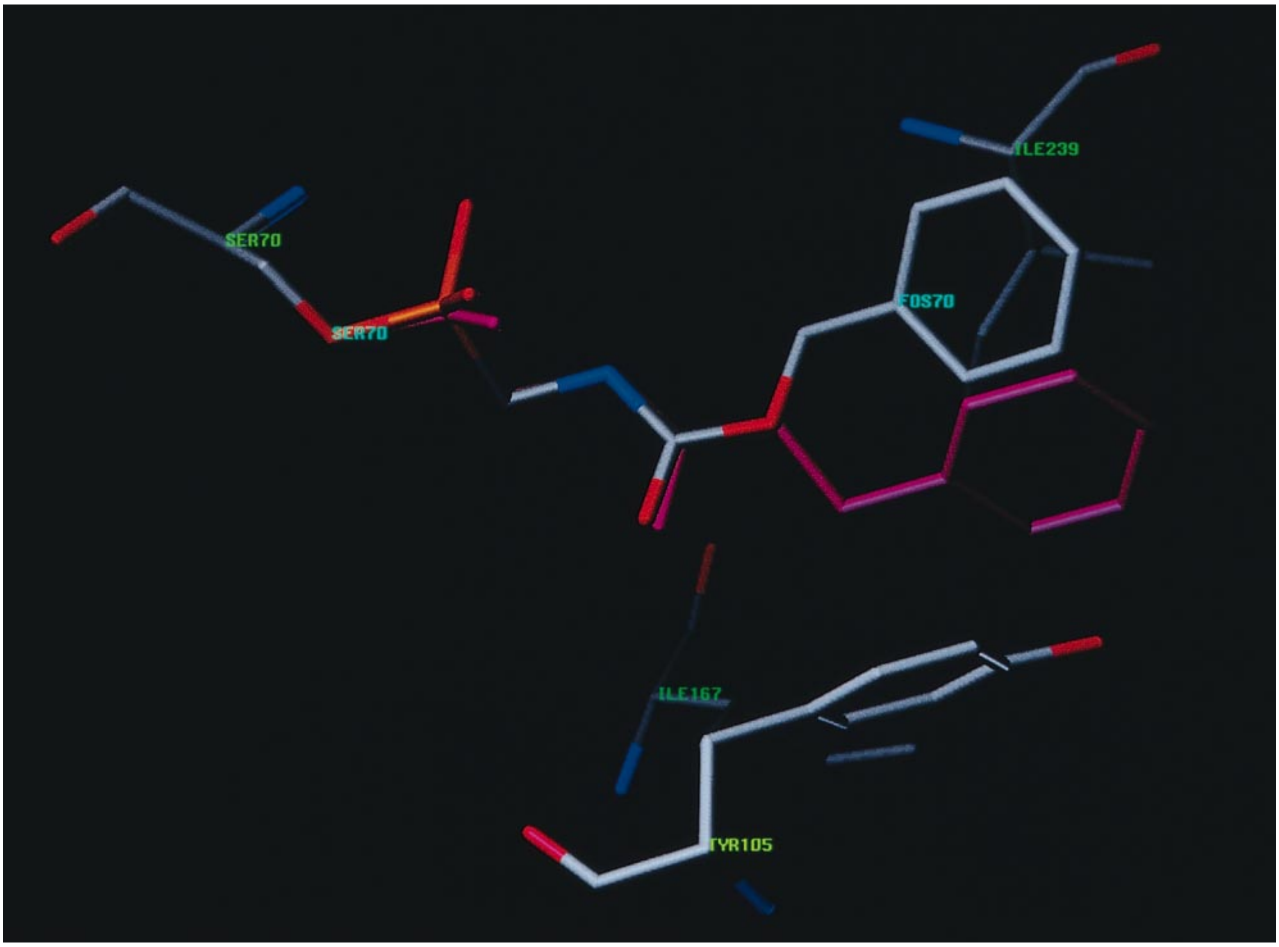


Figure 7. 1BLH. Docking of a covalently bound phosphonate inhibitor to β-lactamase.

where GOLD failed to give an answer. Additionally, the failure to predict the binding mode of the complexes 1HRI (Zhang *et al.*, 1993), 1IGJ (Jeffrey *et al.*, 1993), 1MUP (Böcskei *et al.*, 1992), 3CLA (Leslie, 1990) and 2R07 (Badger *et al.*, 1989) was due to the ligand not forming any hydrogen bonds to the protein. It is also the case that the ligand in 3HVT does not form any hydrogen bond to the protein. However, GOLD is able to predict the ligand binding mode correctly as a protein donor is relatively close to a ligand acceptor (though they are certainly not hydrogen bonded). For the same reason one of the 20 solutions for 3CLA was relatively close to the observed binding mode (r.m.s. 2.2 Å).

The design of GOLD favours the docking of hydrophilic ligands. There are two reasons for this. Firstly, the chromosome encoding in GOLD means that the GA samples binding modes by searching patterns of hydrogen-bonding motifs. Thus the algorithm is directed to find hydrogen-bond networks, whereas it is not guided to find hydrophobic interactions. Secondly, the fitness function contains a term for dispersive interactions but does not have a term for desolvation. With a significant part of the hydrophobic contribution to binding missing from the fitness score, the algorithm is likely to underestimate the contribution to binding from hydrophobic interactions. The complexes 1ICN (Eads *et al.*, 1993), 1LIC (LaLonde *et al.*, 1994), and 2PLV (Filman *et al.*, 1989) all contain ligands with a hydrocarbon chain and a polar head group hydrogen-bonded to the protein. However, in all three cases GOLD failed to dock these highly flexible and hydrophilic ligands, presumably for the two reasons outlined above. The prediction for 1LIC was better than the predictions for 1ICN and 2PLV with the hydrophobic tail in the correct place, but the ligand was orientated back to front. 1ICN was a particularly difficult problem as the acid head group in the oleate ligand was disordered in the crystal structure.

The problem of underestimating the hydrophobic contribution to binding is illustrated in the results obtained for 1ACJ (Harel *et al.*, 1993). Here, the predicted solution is clearly wrong but ranked solutions 2 to 6 are correct predictions with little difference in fitness scores from the highest ranking solution. A similar situation is observed in 4FAB (Herron *et al.*, 1989), where the crystallographically observed binding mode is present in the set of GOLD solutions but the top scoring prediction has solvent-exposed hydrophobic groups and strong hydrogen bonds. Analysis of the solutions shows that the predicted solution has exposed the hydrophobic part of the ligand to the solvent in order to form strong hydrogen bonds to groups on the protein surface. The failure to predict the binding mode of the ligands in 1ACK (Harel *et al.*, 1993) and 6RSA (Borah *et al.*, 1985) also appeared to be due to the same effect, though the experimental binding mode was not present in any of the GOLD solutions.

In some cases GOLD failed because of interactions between the protein and the ligand that were not expressed in the GA fitness function. For example, the algorithm does not properly represent the interactions between electron-rich and electron-deficient groups. In 1ACK (Harel *et al.*, 1993) an electron-deficient ligand choline group is not properly stacked on top of a Trp residue. In 1MCR (Edmundson *et al.*, 1993) there appears to be a NH..pi interaction contributing to binding. The inability of GOLD to recognise this bond may explain the failure to predict the binding mode. In 1ETA (Hamilton *et al.*, 1993) the binding of the ligand would appear to be partially driven by a short I..O contact. Such contacts have been observed in small molecule crystals (Lommerse *et al.*, 1996) and, again, this interaction is not recognised by GOLD.

It is not clear why GOLD failed in the cases of 1NIS (Lauble *et al.*, 1994), 1ROB (Lisgarten *et al.*, 1993), 1TDB (Perry *et al.*, 1993), 2AK3 (Diederichs & Schultz, 1991), 2MCP (Padlan *et al.*, 1985) and 2MTH (Smith & Dodson, 1992). In 1TDB the predictions seemed very unstable, with hardly any consistency or clustering in the 20 binding modes produced by GOLD. The predictions were particularly bad for 2MCP and 2MTH, where the ligand was predicted to bind to a completely different region, in both cases approximately 9 Å from the experimentally observed binding mode. 1TDB contained a ligand with very poor geometry (see above) and 2MTH had a crystal structure resolution of 3.1 Å. These factors may have contributed to failure. Some systems seemed to fail because of their complexity: 1AAQ (Dreyer *et al.*, 1992) and 1EED (Cooper *et al.*, 1992) both contained highly complex ligands (26 and 30 rotatable bonds, respectively) and it was not possible to identify any other cause of failure (although the number of rotatable bounds was not found to be a statistically significant cause of failure for this dataset, the algorithm will undoubtedly start to fail as ligand flexibility increases). In 1HDC (Ghosh *et al.*, 1994) the steroid ligand shows large differences in the location of the hydrophobic skeleton between the GA prediction and the experimental result, while the crystallographically observed binding mode is also present in the set of GOLD solutions. However, in this case a visual examination of the results did not show the GOLD prediction to be unreasonable in that the hydrophobic skeleton made good contact with the protein. In 1DID the observed binding mode was not ranked highly because the empirical parameters for Mn underestimated the coordination energy between N.3 and Mn.

Some of the complexes in the *errors* category were quite good solutions, with the binding mode of the majority of the ligand correctly predicted. This was the case for the complexes 1BAF (Brünger *et al.*, 1991), 1ETR (Brandsetter *et al.*), 1EAP (Zhou *et al.*, 1994) and 1RDS (Nonaka *et al.*, 1993). The prediction for 1EAP is shown in Figure 4. A subjective analysis of the observed and predicted binding modes did not suggest any reason for preference of

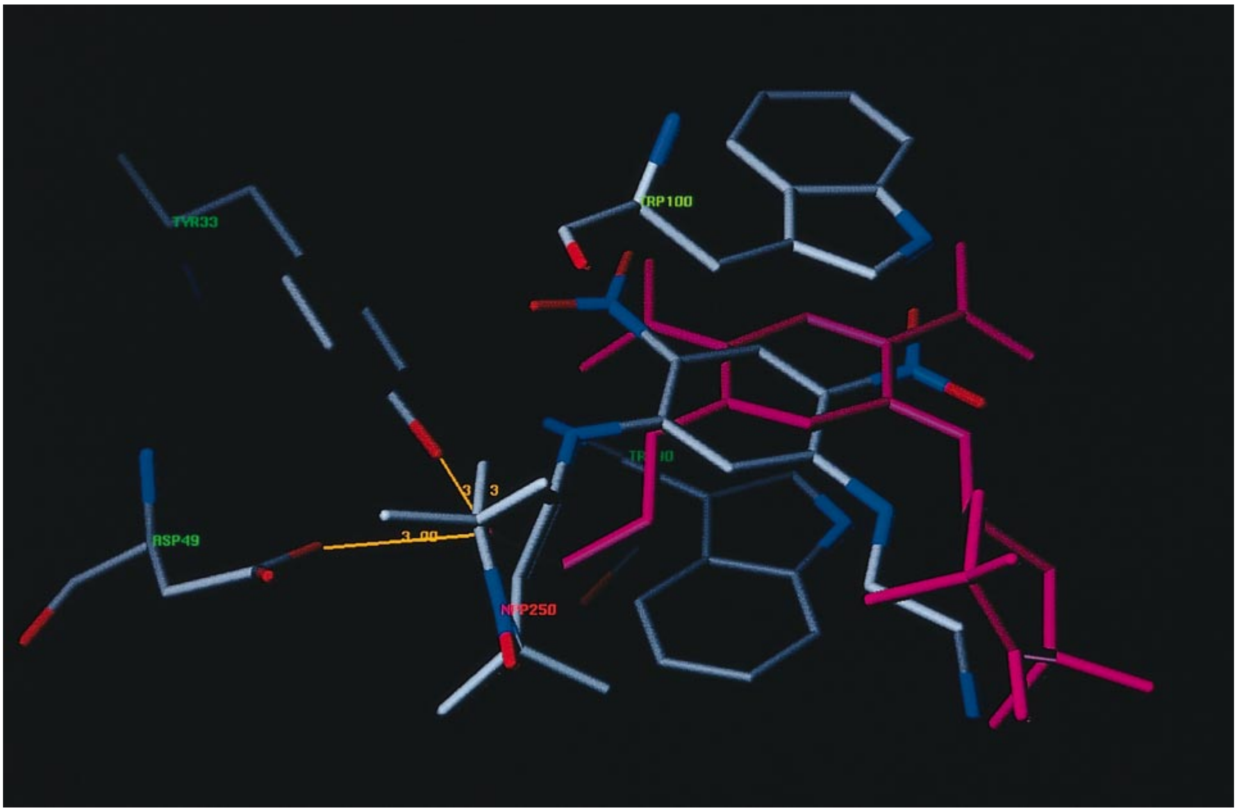


Figure 8. 1BAF. Docking of dinotrophenyl piperidine *N*-oxide ligand to an antibody.

the crystallographic binding mode. Although the prediction for 1RDS was incorrect, performing a further 20 GA runs produced 40 GA solutions, of which the highest scoring was close to the experimental result. In 1ETR the ligand contains a nitrogen bicycle that is highly mobile in the GOLD set of solutions. It is the position of this group that is wrong in the GOLD prediction and it is not easy to explain the preference for the crystallographic binding mode. However, this group is linked to the rest of the ligand through a sulphonamide group. The difference may be due to the Tripos forcefield torsional potential used by GOLD lacking parameters for sulphonamide torsions (Clark *et al.*, 1989). A search of the CSD for acyclic torsions of the form C-NH-SO_2-benzene retrieved 93 hits, only one of which was consistent with the GA prediction. However, the experimental result was consistent with the largest peak in the histogram of torsional frequencies obtained from the search.

The case of 1BAF is worth further comment. Here an antibody is complexed with a dinitrophenyl piperidine *N*-oxide ligand (Brünger *et al.*, 1991) and the crystal structure resolution is 2.9 Å. Figure 8 shows the ligand and some important residues from the crystal structure coloured by atom type while the GOLD prediction is shown in red. Ligand binding is dominated by the sandwiching of the dinitrophenyl ring between two tryptophan residues. GOLD correctly predicts the position of this ring. However, closer inspection shows that the predicted binding mode has the ligand flipped and the two side-chains exchanged, relative to the experimental result. GOLD has formed strong hydrogen bonds between the ligand protonated nitrogen atom and Tyr33 and Asp49. In the crystal structure these residues are partly desolvated by hydrophobic ligand groups. Given the low resolution of the crystal structure and the unusual experimental binding mode it is again interesting to speculate, as with 1BLH, if the structure factors are consistent with the predicted binding mode.

In the complexes 1EAP, 1ETR, 1HDC, 1ACJ, 1DID, 1ROB and 4FAB, GOLD made an incorrect prediction (in that the top-ranked solution did not correspond to the crystallographically observed binding mode) yet an acceptable solution was found in one of the 19 other solutions. In the complexes 1EED, 2AK3, 2MTH, 3CLA and 1RDS a much better (though still partially incorrect) solution was found within the set of solutions. In some senses this is good, as an experienced modeller examining the predicted binding modes generated by GOLD may be able to identify the solution closest to the crystallographically observed binding mode as the most likely. However, the fact that solutions closest to the observed binding mode score lower than incorrect solutions may indicate a failure in the discriminatory ability of the fitness function.

Docking using unbound protein structures

In the docking problems described above, we have docked the ligand back into the bound conformation of the active site. However, for several of the test systems, the structure of the unbound protein has been solved. In order to investigate the effectiveness of the algorithm when using the uncomplexed protein, docking into the unbound active site was performed for three of the test systems: 1BLH, 1DR1 and 4DFR. In the case of 1DR1 the crystal structure of uncomplexed chicken liver DHFR (PDB entry 8DFR: Matthews *et al*., 1985) was used for the protein and a good result was obtained. In fact, this result was superior to the docking obtained using the bound protein. For 4DFR the crystal structure of unbound *Escherichia coli* DHFR (PDB entry 5DFR: Bystroff & Kraut, 1991) was used for the protein and a close result was obtained. In this case the prediction for the ligand, methotrexate, was poorer than was obtained using the complexed protein, though all important protein ligand contacts were observed. In both of these test systems the conformation of the bound and unbound protein active sites were sufficiently similar to enable successful experiments to be performed on the unbound protein.

The complex 1BLH is described above. Here, the unbound conformation of β-lactamase was taken from PDB entry 3BLM (Herzberg, 1991). It was observed that residue Tyr105 had changed conformation upon binding to the phosphonate inhibitor (Chen *et al*., 1993). The GOLD prediction for binding to the unbound protein was different from the previous prediction obtained for binding to the bound protein as the earlier prediction made a close contact to this residue. A *close* result was obtained, with the ligand in a similar position to that observed in the crystal structure complex. As in the crystal structure, the carbamate C-O-C=O torsion in this prediction was *trans*.

Discussion

Here, we have described the development of GOLD, a GA for flexible ligand docking. The effectiveness of the approach has been illustrated by the docking of NADPH to DHFR. The method has been verified by testing the program on a set of 100 complexes selected from the PDB. During this process GOLD achieved a 71% success rate in reproducing the experimentally observed binding mode. While this was a very encouraging result, an analysis of the results was performed in order to determine common causes for failure and conditions for success. Inspection of ligand composition revealed that the algorithm was most likely to fail if the ligand was hydrophobic or if the protein active site was poorly resolved. Likewise, if the active site is poorly determined failure is clearly probable. The algorithm has been specifically designed to elucidate the binding mode of hydrophilic ligands and it runs into major problems when trying to dock hydrophobic ligands.

Presently GOLD has no effective mechanism for sampling hydrophobic interactions and the scoring function used in the GA does not include terms for desolvation. Future research will endeavour to make good these deficiencies. In the first instance hydrophobic groups in the ligand and points of hydrophobic interaction in the active site will be identified. These can be mapped to each other in the GA chromosome (in an analogous fashion to what is already done for donor hydrogen atoms and acceptors), thus directing the GA search to elucidate hydrophobic interactions. One possible method of taking account of desolvation might involve calculating the solvent-accessible surface area of the docked ligand (Connolly, 1983; Fraternali & Gunstren, 1996; Still *et al*. 1990).

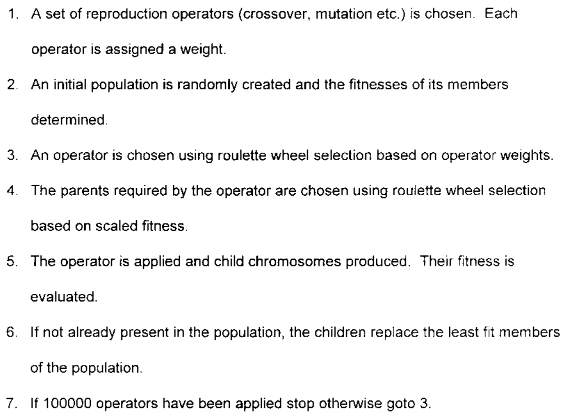
There remain some interactions that are not accounted for by the fitness function. Examples of these are: interactions between electron-rich and electron-deficient groups; unusual non-bonded contacts such as I...O; ring stacking; ring edge to face interactions and hydrogen bonds to pi systems. We hope to extend the algorithm to include these interactions.

Despite the algorithm's apparent success there are a number of problems with the methodology in general. Firstly, the algorithm is comparatively time-consuming. Two approaches are available for reducing the run-time of the procedure and these will be investigated in the future: increasing the effectiveness of the program so that each run will be more likely to identify a high scoring docking and thus fewer runs will be required to elucidate the binding mode; and increasing the efficiency of the program so that each run takes less time. One technique that has the potential to increase both efficiency and effectiveness of the algorithm is to search only those ligand torsional angles that are seen in small molecule crystals. The CSD could be used as a knowledge base for deriving torsional constraints (Allen *et. al*., 1991; Klebe, 1994).

Although the algorithm is almost completely automated, the user is required to input the approximate size and location of the active site together with the ionisation state of the protein and of the ligand. There are now techniques that can predict the location of the active site with considerable accuracy (Peters *et al*., 1996). Using a technique such as this would help fully automate the algorithm.

Frequently, one of the aims of docking experiments is to estimate the binding free energy of the ligand. Since the scoring function used by GOLD does not contain any entropic component it is unlikely that the program could predict binding free energies or rank actives correctly. However, given that GOLD can predict the binding mode effectively, there are a number of approaches that attempt to predict binding free energies accurately (Ajay & Murcko, 1995).

A more fundamental limitation of the algorithm is that (with the exception of a few terminal bonds) the binding site is essentially rigid. From our limited experiments on unbound complexes, it would appear that there are systems for which the conformational changes to the protein active site on binding are sufficiently small to enable docking to the uncomplexed protein. Furthermore, it is often a bound protein conformation that is used as target in structure-based drug design. We hope in the future to extend the flexibility of the protein to include highly mobile active site side-chains. Modelling large-scale conformational changes of the protein upon binding remains an intractable problem.



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 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 produced. Their fitness is evaluated.
6. If not already present in the population, the children replace the least fit members of the population.
7. If 100000 operators have been applied stop otherwise goto 3.

Figure 9. Steady state with no duplicates GA.

Materials and Methods

Genetic algorithms

A GA is a computer program that mimics the process of evolution by manipulating a collection of data structures called chromosomes. Each of these structures encodes a possible solution (i.e. a possible ligand orientation within the protein binding site) to the docking problem and may be assigned a fitness score based on the relative merit of that solution. A steady-state operator-based GA was used to explore conformation space and ligand binding modes (Davis, 1991). This GA is illustrated in Figure 9. The algorithm is similar to the software already described (Jones *et al.*, 1995a). However, there have been many significant improvements and developments.

Initialisation of the protein and of the ligand

The protein was prepared as described by Jones *et al.* (1995a,b), i.e., water molecules and ions were removed and hydrogen atoms were added at appropriate geometry, taking account of protonation states.

GOLD requires that the user indicates the approximate size and location of the ligand binding site: this is done using the user-definable parameters ORIGIN and RADIUS, where the binding site should lie within a sphere of radius RADIUS, around the point ORIGIN. A flood-fill algorithm (Ho & Marshall, 1990) was used to locate the solvent-accessible surface within distance RADIUS of the point ORIGIN. Following this a cavity-detection algorithm (Delaney, 1992) isolated concave solvent-accessible surfaces to which the ligand could bind. The routine was parameterised to locate all cavities less than 7.5 Å wide. A second pass of the algorithm removed isolated cavities that were less than 2.0 Å wide. The active site was defined by using a second flood-fill to find the solvent-accessible protein surface within 2.0 Å of the remaining cavities. All hydrogen-bond donor, donor hydrogen atoms and acceptors within this surface were identified using the SYBYL atom-type characterisation (Clark *et al.*, 1989) shown in Table 6 (a set of fragments illustrating SYBYL atom typing can be seen in Figure 9 of Jones *et al.* (1995a)). Lone pairs were added to acceptors at a distance of 1.0 Å.

If a terminal donor or acceptor was bound to the protein *via* a single bond, then that bond was selected as rotatable (where a terminating atom is the final heavy atom in an acyclic chain). This rule allowed terminating NH and OH groups to move into good positions for hydrogen bonding. However, any protonated nitrogen atom that could form a strong hydrogen bond to a nearby protein O.2 or O.co2 acceptor, was held fixed in that hydrogen-bonded conformation.

A check was made of all donors and acceptors (whose lone pairs or donor hydrogen atoms could not rotate) to determine if they would be accessible for binding to the ligand. A binding point was created for each donor-hydrogen and acceptor lone-pair at a distance of 2.9 Å, colinear with the bond to the donor or acceptor. Donor hydrogen atoms were considered available for ligand binding only if their binding point lay within the cavity determined above.

We assumed (Jones *et al.*, 1995a) that hydrogen bonds had a directional preference along the acceptor lone-pair. However, it appears that, while many acceptors exhibit this behaviour, some acceptors prefer to form bonds in

Table 6. Allowed donors and acceptors based on SYBYL atom types

SYBYL atom types	Donor	Acceptor	Acceptor geometry
N.1, N.ar, O.co2 (carboxylate acid), O.2 in NO2 nitro group	N	Y	Lone-pair
N,3, N.2, N.pl3 with only 2 connections (acidic N)	Y	Y	
O.2, O.2 in amide group	N	Y	Plane of lone-pairs
O.3[a]	Y	Y	
O.co2 or O.2 bonded to P or S, or O.co2 (singly charged oxygen)	N	Y	No lone-pair geometry
N.am, N.pl3, N.4	Y	N	N/A

N.B. Donors must have a hydrogen atom attached.

[a] O.3 can accept only if a donor, or if bonded to one C.3 atom and a C.2 or C.ar atom within a benzene ring, or if bonded to two C.3 atoms.

the plane of the lone-pairs, or have no preference in relation to the lone-pair positions (unpublished results). In Table 6, acceptors are characterised by their preferred hydrogen-bonding geometries, as determined by searches of the CSD. Acceptors may show a tendency to form hydrogen bonds along lone-pair directions (e.g. the oxygen atoms of nitro groups), or merely within the plane of the lone-pairs (e.g. ether oxygen atoms), or they may show no strong directional preferences at all (e.g. phosphate oxygen atoms). In actuality, phosphate oxygen atoms do show some lone pair preference for hydrogen bonding (unpublished results). However, this preference is not marked and GOLD was found to perform much better if it is assumed that phosphate oxygen atoms do not show hydrogen-bonding directionality, relative to the lone-pairs.

An acceptor was deemed available for binding if it lay in the detected cavity and showed no preference for hydrogen-bonding in the lone-pair directions or the lone-pair plane. If an acceptor preferred to form hydrogen bonds in the plane of its lone pairs and if the binding point of either lone-pair lay within the active site cavity then both lone-pairs were accessible for ligand binding. The acceptors that displayed hydrogen bond directionality along-lone pairs could bind to the ligand only through lone-pairs whose binding points lay in the active site cavity. If no such lone-pair was found then the acceptor was not available for ligand binding. In order to reproduce binding motifs observed in complexes, a correction was made for the two charged oxygen atoms in a carboxylic acid. If one of the *syn* lone-pairs on either of the two oxygen atoms had its binding point in the cavity then both *syn* lone-pairs were available for ligand binding.

Hydrogen atoms were added to the ligand and hydrogen bond donors, and acceptors within the ligand were identified. Prior to docking, the ligand was fully minimised using the MAXIMIN2 module, available within SYBYL (Clark *et al*., 1989). Lone-pairs were added at appropriate geometry. All acyclic single non-terminal bonds were marked as rotatable. In order to ensure that there was no bias for the original coordinates when docking, a random translation was applied to the ligand and random rotations were applied round ligand rotatable bonds.

The chromosome representation

A similar representation to that described by Jones *et al*. (1995a) was employed. Conformation information was encoded by two binary strings: one for the protein and one for the ligand, where each byte in the string encoded an angle of rotation about a rotatable bond. Thus each torsion was allowed to vary between -180° and 180° in step-sizes of 1.4°. Unlike the encoding used by Jones *et al*. (1995a), Gray-coding was not employed, as it was found to disrupt crossover. Two integer strings encoded mappings, suggesting possible hydrogen bonds between the protein and the ligand. The first of these strings encoded a mapping from acceptors in the ligand to donor hydrogen atoms in the protein, such that if V was the integer value at position P on the string, then the Pth acceptor in the ligand was mapped to the Vth donor hydrogen in the protein. V could also be a null value, indicating that the acceptor was not mapped to any protein hydrogen. In a similar manner, the second string encoded a mapping from donor hydrogen atoms in the ligand to acceptors within the ligand. On decoding a chromosome, GOLD utilised least-squares fitting to form as many of these hydrogen bonds as possible.

The fitness function

The fitness function was evaluated in six stages as follows. (1) A conformation of the ligand and protein active site was generated. (2) The ligand was placed within the active site using a least-squares fitting procedure. (3) A hydrogen bonding energy *H_Bond_Energy*, was obtained for the complex. (4) A pairwise energy, *Complex_Energy*, was obtained for the steric energy of interaction between the protein and the ligand. (5) Molecular mechanics expressions were used to generate the term *Internal_Energy* which was a measure of the internal energy of the ligand. (6) The energy terms were summed together to give a final fitness score.

With the exception that binary-coding was now preferred to Gray-coding, the methods described by Jones *et al*. (1995a) were used to generate the active site and ligand conformations.

For every donor hydrogen atom a virtual fitting point was created colinear with the bond at a distance of 2.9 Å from the donor. Fitting points were created at the centre of each acceptor. Although acceptors were now used in preference to lone-pairs as fitting points for generating hydrogen-bond motifs, the same two-pass least-squares fitting technique described by Jones *et al*. (1995a) was used to dock the ligand within the active site. When the second pass of least-squares fitting was applied, those pairs of points that were less than 1.5 Å apart were used. If three such pairs could not be found, the three closest pairs of points were used.

In order to determine the hydrogen-bonding energy of the complex, each possible combination of donor-hydrogen atom and acceptor was examined in turn to see whether or not a bond had been formed. The geometrical arrangement of donor hydrogen fitting point, acceptor and any lone-pairs was examined, and a weight between 0 and 1 assigned to the potential bond. The full bond energy between the donor and acceptor was then scaled by this weight (this full bond energy, E_{pair}, is described by Jones *et al*. (1995a) and discussed later). The weight is the product of two terms: a distance weight and an angle weight. Let *wt* be the weight of the hydrogen bond, such that $wt = distance\_wt \times angle\_wt$.

The distance weight is a function of d, the distance between the donor hydrogen fitting point and the acceptor. If d was less than 0.25 Å then *distance_wt* was 1 and if d was greater than *max_distance* then d was 0. Otherwise d lay in the interval (0.25, *max_distance*), in which case d was linearly rescaled to the interval (0, 1) and squared to give *distance_wt*. *max_distance* was 4.0 Å when the GA started, 1.5 Å after the application of 75,000 genetic operations and varied linearly between these bounds. The rationale behind varying *max_distance* over the course of a GA run was to allow long-range contacts between donors and acceptors to contribute to the fitness score at the beginning of a GA run (in the expectation that they would evolve into good hydrogen bonds), while ensuring that all hydrogen bonds that contributed to the fitness of the final solution were close contacts.

The hydrogen-bond directional preference of acceptors is listed in Table 6. If the acceptor had no lone-pair directional preference then *angle_wt* was 1.0. If the acceptor had hydrogen bond directionality in the plane of the lone-pairs then *angle_wt* was a function of Θ, the angle between the donor, donor hydrogen atom and the plane of the lone-pairs. If Θ was greater than 60 then *angle_wt* was set to 0 and if Θ was less than 20 then Θ was set to 1.0. Otherwise Θ lay in the interval (20, 60), in which case it was linearly rescaled to (1, 0) and squared to give *angle_wt*. Finally, if the acceptor had a directional preference along the lone-pair, then, for each acceptor lone-pair the hydrogen-bond angle between the two vectors donor, donor hydrogen atom and lone-pair, acceptor was determined. *angle_wt* was a function of Φ, the largest hydrogen-bond angle between the donor hydrogen atom and any of the acceptors lone pairs. If Φ was greater than 160 then *angle_wt* was 1.0 and if Φ was less than 60 then *angle_wt* was set to 0. Otherwise, Φlay in the interval (160, 60), in which case it was linearly rescaled to (1, 0) and squared to give *angle_wt*.

The energy *H_Bond_Energy* was the sum of all individual bond energies found from all combinations of ligand donor hydrogen atom and protein acceptor and all combinations of ligand acceptor and protein donor hydrogen atom.

Following the placement of the ligand into the active site of the protein, a 4-8 potential with linear cut-off (similar in form to that used by Surles *et al*. (1994)) was used to determine the energy of interaction between the ligand and the protein. The 4-8 potential was of the form:

$$E_{ij} = \frac{A}{d_{ij}^{8}} - \frac{B}{d_{ij}^{4}}$$

where E_{ij} was the energy of interaction between two atoms and d_{ij} was the distance between them. The choice of the parameters A and B is described below. This form of pairwise interaction was chosen because it is much softer than the standard 6-12 potential. Adjustments were made to E_{ij} if the two atoms were involved in a hydrogen bond. E_{ij} was zero for the interaction between a donor hydrogen atom and an acceptor, while the distance between the donor and acceptor is scaled by a factor of 1.43. This was equivalent to reducing the van der Waals radii (Clark *et al*., 1989) of donor and acceptor by 70%. All pairwise interaction energies of ligand and protein atoms in close contact (a cut-off distance of 1.5 times the sum of the van der Waals radii of the two atoms was employed) were summed to give *Complex_Energy*.

Let $-k_{ij}$ be the minimum energy of interaction between two atoms i and j. For interaction energies, E_{ij}, greater than *scale* $\times$ k_{ij} the linear cut-off was applied. The gradient of the cut-off was such that E_{ij} was $1.5 \times$ *scale* $\times k_{ij}$ when d_{ij} was zero. This cut-off value was expressed in terms of k_{ij} not for any physical reason, but rather to enable the use of efficient lookup tables. The parameter *scale* varied throughout the GA run. After the application of 75,000 genetic operators it was set to 120.0, at the beginning of the GA run it was 1.0 and in-between it was varied on a logarithmic scale. This scaling was employed so that the GA was encouraged to form close interactions with the protein early in the course of a GA run, while ensuring that there was no steric clash when the algorithm terminated.

The 4-8 potential was parameterised to reproduce the minimum of the more usual 6-12 potential. The energy of association between two molecules can be represented using a Lennard-Jones 6-12 potential, where the second term accounts for the attractive dispersion energy between two molecules (Hirschfelder *et al*., 1964):

$$E_{ij} = \frac{C}{d_{ij}^{12}} - \frac{D}{d_{ij}^{6}} = \frac{C}{d_{ij}^{12}} - \frac{3}{2} \times 0.23 \times \frac{I_i I_j \alpha_i \alpha_j}{(I_i + I_j) d_{ij}^{6}}$$

where I_x is the ionisation potential of atom x (in eV) and α_x is the polarisability (in 10^{25}cm^3). Table 7 lists these parameters for the common SYBYL atom types. Ionisation potentials were taken from Hinze & Jaffe (1962) and atom-centred polarisabilities were determined using the method devised by Glen (1994). C was chosen so that E_{ij} was at a minimum when d_{ij} was equal to $r_i + r_j$, or the sum of the van der Waals radii of the two atoms. The parameters A and B were chosen so that the 4-8 potential used by GOLD had the same minimum as the Lennard-Jones 6-12 potential.

Table 7. Ionisation potentials and polarisabilities for SYBYL atom types

SYBYL atom types	Ionisation potential (eV)	Polarisability (10^{25}cm^3)
C.3	14.61	13.8
C.2 C.ar C.cat	15.62	13.8
C.1	17.47	13.8
N.3	18.93	8.4
N.2 N.ar	22.10	8.4
N.1	23.91	8.4
N.am N.pl3	19.72	8.4
N.4	33.29	8.4
O.3	24.39	5.4
O.2	26.65	5.4
O.co2	35.12	5.4
S.3 S.o S.o2	15.50	29.4
S.2	17.78	29.4
P.3	16.78	40.6
H	13.60	4.0
F	20.86	3.7
CL	15.03	21.8
BR	13.10	31.2
I	12.67	49.0

This potential proved to be particularly effective in reproducing experimental ligand binding modes. A 4-8 potential with a linear cut-off is much softer than the 6-12 potential that is traditionally used, allowing the GA to form close contacts with the protein more easily. Figure 10 shows the different forms of the following potentials: the 6-12 carbon-carbon (C.3-C.3) interaction used in the Tripos forcefield (Clark *et al*., 1989); the C.3-C.3 interaction used in GOLD; and the oxygen-oxygen (O.3-O.3) interaction used in GOLD. It can be seen that the C-C potential used in GOLD is both softer and deeper than that used in the Tripos forcefield. Additionally, the depth of the well for O.3-O.3 interactions is much less than for C.3-C.3 interactions and is shallower than in the Tripos forcefield (the minimum energy is 0.078 kcal mol^{-1} in GOLD and 0.12 in the Tripos forcefield). In general, the van der Waals potential used in GOLD will favour close contacts between hydrophobic groups. This is perhaps not surprising in view of the close relationship between polarisability and hydrophobicity.

The term *Internal_Energy* was a sum of the ligand steric and torsional energies. The steric energy was determined using a 6-12 potential of the form:

$$E_{ij} = \frac{C}{d_{ij}^{12}} - \frac{D}{d_{ij}^{6}}$$

where C and D were calculated as described above. The

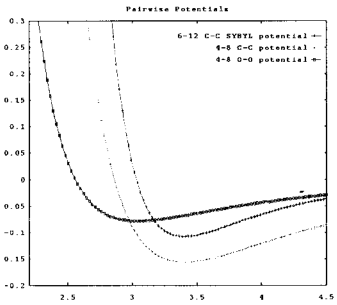


Figure 10. The form of the 4-8 potential used within GOLD.

Tripos forcefield torsional potential used was of the form:

$$E_{ijkl} = \frac{1}{2} V_{ijkl} \left[1 + \frac{n_{ijkl}}{|n_{ijkl}|} \cos(|n_{ijkl}| \cdot \omega_{ijkl}) \right]$$

where E_{ijkl} was the torsional energy associated with four consecutively bonded atoms i, j, k, l; ω was the torsional angle; n was the periodicity; and V the barrier to rotation (the last two parameters are taken from Clark *et al.* (1989)).

The final fitness score was determined by a sum of all the energy components. The fitness score was given by:

$$-H\_Bond\_Energy - (Internal\_Energy + Complex\_Energy)$$

The genetic operators

Rather than manipulate one large population of chromosomes a distributed environment known as the island model was employed. This involves several subpopulations and the migration of individual chromosomes between the subpopulations (Starkweather *et al.*, 1990; Tanese, 1989). The island model and migration operator were implemented as described in Jones *et al.* (1995b). As observed by Jones *et al.* (1995b) the island model did not improve the effectiveness of the GA, but an improvement was seen in its efficiency. Five subpopulations were used, each containing 100 individuals.

The GA made use of three genetic operators: crossover, mutation and migration. Crossover and mutation are described by Jones *et al.* (1995a). The migration operator (described by Jones *et al.*, 1995b) copied an individual from one island to a neighbouring island. Operators were chosen using roulette-wheel selection based on operator weights. These weights were chosen so that crossover and mutation were applied with equal probability and migration was applied 5% of the time. After the application of 100,000 genetic operations the algorithm terminated, outputting the highest scoring docking.

As Jones *et al.* (1995a), selection was based on linear-normalised fitness scores and, in order to prevent premature convergence, a low selection pressure of 1.1 was used (the selection pressure represents the relative probability that the best individual will be chosen as a parent compared with the average individual). The technique of nicheing (Goldberg & Deb, 1989) was used to further increase population diversity. When adding an individual to the population, nicheing involved comparing the individual against every member of the population to determine if any members inhabited the same niche as the new individual, which corresponded here to examining the docked ligands associated with two chromosomes to determine if they shared a niche. This was the case if the r.m.s. distance between all donors and acceptors in both ligand dockings was less than 1.0 Å. If more than one individual was found in the same niche as the new chromosome then the new chromosome replaced the least-fit chromosome in the niche, rather than the least-fit chromosome in the population.

Calculation of hydrogen-bond energies

The hydrogen-bond energy between a donor and an acceptor is an important component of the fitness function, since each hydrogen-bonding pair contributes to the overall energy of binding. As described by Jones *et al.* (1995a) hydrogen bond energies between donor and acceptor types were precalculated using model fragments and accounting for water-displacement. Initially the donor (d) and the acceptor (a) are in solution, but on coming together (da) water (w) is stripped off. Therefore to simulate the interaction energy, E_{pair} is composed of four terms:

$$E_{\text{pair}} = (E_{\text{da}} + E_{\text{ww}}) - (E_{\text{dw}} + E_{\text{aw}}).$$

Jones *et al.* (1995a) describe the generation of hydrogen-bonding energies for six donor and 12 acceptor types. However, searches of the CSD reveal that covalently bound halogens rarely accept hydrogen bonds. Thus Cl, Br and F are no longer counted as acceptors and the associated fragments CL, BR and F were removed from the set of fragments. In continuing the development of the algorithm we have added two metal ions (Mg and Zn) and three acceptor types. Within GOLD, metal ions that coordinate electronegative atoms are modelled in an analogous fashion to donors (see below). The new acceptor types are: amide oxygen (O2N), $N^{+}O_{2}^{-}$ nitro group (ONO2) and deprotonated nitrogen (NACID). The first two acceptor types are based around the SYBYL O.2 atom types while the NACID acceptor is represented by a N.pl3 atom singly bonded to only two neighbours. The methods described by Jones *et al.* (1995a) were used to calculate hydrogen-bond energies. The fragments used within the modelling experiments are illustrated in Figure 11, where the atoms coordinating the metal ions were similar to those observed in protein structures. It was assumed that the metal ions both had a formal charge of +2, giving the total fragment charge for MG as

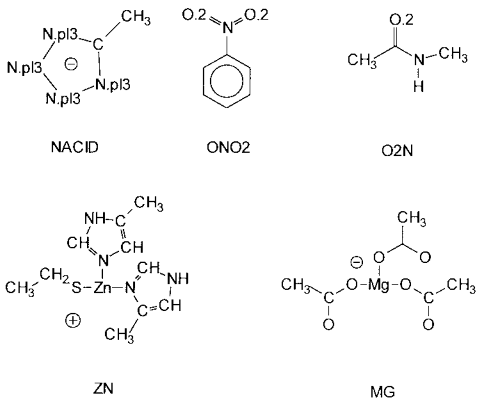


Figure 11. New donor and acceptor types and associated fragments used in hydrogen bonding energy determination.

Table 8. Coordination energies for metals and hydrogen-bonding energies for new acceptor types

		Energy (kcal mol^{-1})			
Donor or metal	Acceptor	Donor or metal	Acceptor	Complex	Bond
N4	O2N	−24.784	−7.955	−39.569	−8.732
NPL3	O2N	−4.427	−7.955	−11.024	−.544
N3DA	O2N	−1.147	−7.955	−5.206	1.994
NAM	O2N	−5.919	−7.955	−14.521	−2.549
O3DA	O2N	−33.236	−7.955	−42.200	−2.911
N2DA	O2N	−10.005	−7.955	−15.992	.066
N4	NACID	−24.784	3.122	−24.025	−4.265
NPL3	NACID	−4.427	3.122	−2.788	−3.385
N3DA	NACID	−1.147	3.122	6.984	3.107
NAM	NACID	−5.919	3.122	−5.696	−4.801
O3DA	NACID	−33.236	3.122	−34.187	−5.975
N2DA	NACID	−10.005	3.122	−6.442	−1.461
N4	ONO2	−24.784	11.604	−18.679	−7.401
NPL3	ONO2	−4.427	11.604	9.106	.027
N3DA	ONO2	−1.147	11.604	15.100	2.741
NAM	ONO2	−5.919	11.604	5.898	−1.689
O3DA	ONO2	−33.236	11.604	−21.913	−2.183
N2DA	ONO2	−10.005	11.604	4.635	1.134
MG	N2DA	−95.340	−10.356	−99.107	4.687
MG	O2	−95.340	−21.092	−106.459	8.071
MG	OCO2	−95.340	−22.926	−130.455	−14.091
MG	N1	−95.340	−3.408	−100.328	−3.482
MG	N3A	−95.340	−13.432	−93.791	13.079
MG	O3A	−95.340	−2.923	−79.361	17
MG	N2A	−95.340	−2.802	−102.587	−6.347
MG	N3DA	−95.340	−2.041	−92.878	2.601
MG	O3DA	−95.340	−34.695	−140.749	−12.616
MG	NACID	−95.340	3.122	−65.876	24.44
MG	O2N	−95.340	−7.955	−86.008	15.385
MG	ONO2	−95.340	11.604	−125.826	−43.992
ZN	N2DA	39.270	−10.356	36.181	5.365
ZN	O2	39.270	−21.092	8.993	−11.087
ZN	OCO2	39.270	−22.926	1.335	−16.911
ZN	N1	39.270	−3.408	38.178	.414
ZN	N3A	39.270	−13.432	8.619	−19.121
ZN	O3A	39.270	−2.923	26.565	−11.684
ZN	N2A	39.270	−2.802	29.903	−8.467
ZN	N3DA	39.270	−2.041	27.970	−11.161
ZN	O3DA	39.270	−34.695	−5.103	−11.58
ZN	NACID	39.270	3.122	28.300	−15.994
ZN	O2N	39.270	−7.955	23.996	−9.221
ZN	ONO2	39.270	11.604	44.939	−7.837

Dielectric constant = 1.0; water dimer energy = − 1.902.

−1 and ZN as +1. The values in Table 9 are energies in kcal mol^{-1} obtained using gas-phase molecular mechanics (Clark *et al*., 1989) with PM3 Mulliken (Stewart, 1992) charges and a dielectric constant of 1.0. A special fragment was not used for a deprotonated nitrogen atom that was also a donor. In the event that an acidic nitrogen atom was also a donor it used the hydrogen-bond energies obtained using the NPL3 donor fragment. The energy of interaction between the MG and ONO2 fragment was highly attractive, due to an electrostatic interaction between the nitrogen atom in the ONO2 fragment and one of the coordinating acid groups. Because this value seemed unreasonably attractive an empirical value of −10 kcal mol^{-1} was used in preference.

In addition to Zn and Mg, the metal ions Fe, Mn and Ca occur commonly in protein crystal structures. As semi-empirical methods such as MOPAC (Stewart, 1992) are not reliably parameterised for these metals we were unable to obtain good charges for the determination of metal coordination energies. The empirical energies listed in Table 9 were used by the GA for these metals.

In the algorithm described by Jones *et al*. (1995a) we treated the nitrogen atoms in all lysine residues as solvated and used a separate set of hydrogen-bonding energies for these donors. With the new pairwise energy term this rather arbitrary approach was found to be unnecessary, since if the ligand were to bind to a solvent-exposed lysine residue it could not then form close contacts with the active site. Thus, in the algorithm described here, lysine groups donated using the gas-phase energies calculated for N4 fragment.

In order to obtain more accurate hydrogen-bonding energies an attempt was made to recalculate values of E_{pair} using inter-molecular perturbation theory (IMPT: Hayes & Stone, 1984). Unfortunately this method produced bond energies that were far too attractive (J.P.M. Lommerse, personal communication). The implication must be that the bond energies produced for the algorithm, using molecular mechanics with PM3 Mulliken charges, are of questionable accuracy, since a much higher level of theory has produced results that are much less acceptable. In a separate study, Mitchell & Price (1991) used the IMPT method to estimate amide..amide

Table 9. Empirical coordination energies for Zn, Fe and Mn metal ions

Acceptor	Energy (kcal mol^{-1})
N2DA, N1, O3	0
N2, O2, N3, N3DA, O2N, ONO2	−10.0
OCO2, O3DA, NACID	−15.0

Table 10. Metal ion coordination distances

Metal Ion	Coordination distance
Mg	2.09
Zn	2.05
Mn	1.98
Fe	2.06
Ca	2.44

hydrogen-bond energies and emphasised that care must be taken in extrapolating results from small model systems to interactions with peptides and proteins. Thus, it would appear that the bond energies used within the program should not be taken literally, but should be regarded as empirical parameters that are effective in elucidating ligand binding modes. It is clear that the issue of solvation is very complicated, but the effect of the solvent may be better modelled using a larger bath of many water molecules, rather than the single water molecule in our simple dimer systems. However, this could be done only with significant computational costs.

Extensions to the algorithm

In order to be able to better predict binding modes for a wide variety of protein-ligand complexes several extensions were made to the basic algorithm described above.

Metal ions

A large number of observed protein-ligand interactions involve metal ion coordination. In order to predict these binding modes, common metal ion coordination geometries have been incorporated into the software.

As described above, energies for the coordination of electronegative atoms to positively charged metal ions have been added to the program, to enable it to dock ligands into proteins containing Mg, Mn, Zn, Fe or Ca ions. These ions are observed in protein crystal structures coordinating to electronegative atoms in predominantly tetrahedral or octahedral geometry (Glusker, 1991). The positively charged metal ions were modelled in a similar fashion to hydrogen bond donors. The coordination geometry (tetrahedral or octahedral) of a metal ion was determined as follows: coordinating atoms within the protein were located and coordinating positions identified, where a coordinating position could be either a coordinating atom or the mid-point between two coordinating electronegative atoms bonded to a common atom (for example a carboxylic acid group, or a coordinating water molecule). This enabled the program to recognise the bifurcated coordination of tetrahedral zinc and the EF hands that bind calcium (Glusker, 1991). Coordination angles between the metal and pairs of coordinating positions were measured: coordination angles of over 135° were assumed to indicate octahedral geometry and were counted as 90°; if the mean coordination angle was between 60° and 105° then octahedral geometry was assigned; while if the angle was between 105° and 135° then the metal ion showed tetrahedral geometry. Two exceptions were made to these rules: first Ca ions were assumed to be octahedral. Ca ions typically have coordination numbers of 6 to 10, so do not form tetrahedral geometry (Glusker, 1991). With the ability to recognise bifurcated metal liganding GOLD will correctly model most of these cases. Second, Fe ions coordinated by three or more sulphur atoms were assigned tetrahedral geometry (in the ferrodoxins, sulphur atoms coordinate Fe ions in a distorted tetrahedral geometry (Glusker, 1991)).

Table 10 shows metal ion coordination distances. These were determined from the CSD using mean oxygen-metal contact distances. A set of idealised coordination positions was fitted onto the metal ion, and those positions that were available for ligand binding determined. Within the chromosome string, ligand acceptors were allowed to map to metal coordination positions. When decoding such a chromosome, the 3D coordinates of the position were used as a virtual fitting point. A coordination energy was then determined, using the methods described above for hydrogen bonding.

Small ligand model

If a ligand has fewer than three donor hydrogen atoms and acceptors, then docking by least-squares fitting will not be possible. Additionally, docking may be unreliable if the ligand has a small number of polar groups, especially since some of these may be solvated in reality. Thus GOLD utilised an alternative encoding when docking ligands with fewer than five donors and acceptors.

One integer value and six bytes were added to the chromosome encoding. The integer value was constrained to lie between 1 and the number of donor hydrogen atoms and acceptors in the ligand. The chromosome was decoded as follows: the first three bytes were decoded to generate rotations between 0 and 360°, these being applied to the ligand around the x, y and z axes, respectively. Next, the integer was used to pick one of the mappings between the ligand and the protein. This mapping was then decoded by placing the appropriate acceptor on the donor hydrogen fitting point. The final three bytes were then decoded to produce relaxation distances between -0.5 Å and 0.5 Å, which were applied as translations to the ligand along the x, y and z axes. Following this process the ligand was docked in the active site. In the case that the ligand had only one acceptor or donor hydrogen atom, the additional integer value was not required. Unfortunately, if the ligand has no polar group GOLD is currently unable to perform a docking.

While this feature enabled the algorithm to produce answers for ligand with few polar groups, the additional encoding did not appear to be as effective in docking ligands as the least-squares fitting process. In fact, it is not clear that crossover can perform well on the additional binary encoding, whereas it clearly provides a highly effective mechanism for exchanging hydrogen-bond motifs.

Ligand cyclic flexibility

Free corner flipping was encoded into the algorithm (Goto & Osawa, 1989), in order to allow for limited cyclic flexibility of the ligand. This technique allows the corner atoms of rings to flip above and below their neighbours. For example, a ring in a boat conformation can flip a corner into a chair conformation. Each free corner was encoded as an additional bit in the binary string encoding ligand conformations. If a bit was set when decoding this string, then the corresponding free corner was flipped. The technique of free corner flipping within a binary string genetic algorithm is comprehensively described by Payne & Glen (1993).

Because bioactive conformations for small acyclic rings are often known with some reliability, and corner flipping can serve only to increase the complexity of the docking problem, corner flipping was turned off by default but was available as an option to the user. Those examples for which corner-flipping was used are listed in Results.

Covalently bound inhibitors

A modification to the algorithm was required to dock covalently bound inhibitors. Rather than attempt to elucidate covalent binding, the user was required to select covalent binding as an option and to inform the software of the protein and ligand atoms to be joined by a covalent bond. This was achieved by identifying the protein atom that would bind to the ligand (for example, a serine oxygen atom). This atom was then included in the ligand input file, appropriately bonded to the rest of the ligand. On docking, the least-squares fitting routine was modified to ensure that the protein atom in the ligand file was overlaid to its equivalent in the protein active site. Two energy terms were then added to *Complex_Energy*: the first was a torsional term to account for the torsional energy for the bond linking the protein to the ligand (this was identical with the SYBYL torsional term described above); the second was an angle bending term to ensure that the bond angle between the two bonds linking the ligand and the protein was correct. This second term was taken from the Tripos forcefield and was of the form:

$$E_{ijk} = k_{ijk} \times (\omega - \theta_{ijk})^2$$

where i,j and k are the three atoms forming the bond angle, $\omega$is the measured angle and k_{ijk} and θ_{ijk} are parameters of the forcefield (Clark *et al.*, 1989).

If the covalently bound inhibitor being docked had less than five donors and acceptors then mappings between the ligand and the protein were not included in the chromosome. Three bytes were added to the binary string encoding the protein conformation. The three bytes were decoded to give rotations about the x, y, and z axes. These were applied to the ligand, which was then docked by forming the covalent bond.

Acknowledgements

We thank J.C. Cole and J.P.M. Lommerse for useful discussions, the Biotechnology and Biological Sciences Research Council, the Cambridge Crystallographic Data Centre, the Department of Trade and Industry, Glaxo Wellcome Ltd. and the Medical Research Council for financial support and Tripos Inc. for the provision of software.

References

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). The development of versions 3 and 4 of the Cambridge Structural Database system. *J. Chem. Inform. Comput. Sci.* **31**, 187–204.

Allen, F. H., Bird, C. M. & Rowland, R. S. (1995). Correlation of the hydrogen-bond acceptor properties of nitrogen with the geometry of the N$sp^2 \rightarrow$ Nsp^3 transition in R_1(X=)C-NR_2R_3 substructures: reaction pathway for the protonation of nitrogen. *Acta Crystallog. sect. B,* **51**, 1068–1081.

Ajay & Murcko, M. A. (1995). Computational methods to predict binding free energy in ligand-receptor complexes. *J. Med. Chem.* **38**, 4953–4968.

Badger, J., Minor, I., Oliveira, M. A., Smith, T. J. & Rossmann, M. G. (1989). Structural analysis of antiviral agents that interact with the capsid of human rhinoviruses. *Proteins: Struct. Funct. Genet.* **6**, 1–19.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, F., Bryce, M. D., Rogers, J. R., Kennard, O., Shikanouchi, T. & Tasumi, M. (1977). The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Blaney, J. M. & Dixon, J. S. (1993). A good ligand is hard to find: automated docking methods. *Perspect. Drug Discov. Res.* **1**, 301–319.

Böcskei, Z., Groom, C. R., Flower, D. R., Wright, C. E., Phillips, S. E. V., Cavaggioni, A., Findlay, J. B. C. & North, A. C. T. (1992). Pheromone binding to two rodent urinary proteins revealed by X-ray crystallography. *Nature,* **360**, 186–188.

Bolin, J. T., Filman, D. J., Matthews, D. A., Hamlin, R. C. & Kraut, J. (1982). Crystal structures of *Escherichia coli* and *Lactobacillus casei* dihydrofolate reductase refined at 1.7 Å resolution. *J. Biol. Chem.* **257**, 13650–13662.

Borah, B., Chen, C. W., Egan, W., Miller, M., Wlodawer, A. & Cohen, J. S. (1985). Nuclear magnetic resonance and neutron-diffraction studies of the complex of ribonuclease-A with uridine vanadate, a transition-state analog. *Biochemistry,* **24**, 2058–2067.

Brandsetter, H., Turk, D., Hoeffken, H. E., Grosse, D., Sturzebecher, J., Martin, P. D., Edwards, B. F .P. & Bode, W. (1992). Refined 2.3 Angstrom X-ray crystal structure of bovine thrombin complexes formed with the benzamidine and arginine-based thrombin inhibitoes NAPDP 4-TAPAP and MQPA: a starting point for improving antithrombotics. *J. Mol. Biol.* **226**, 1085–1099.

Brodmeier, T. & Pretsch, E. (1994). Application of genetic algorithms in molecular modelling. *J. Comput. Chem.* **15**, 588–595.

Brünger, A. T., Leahy, D. J., Hynes, T. R. & Fox, R. O. (1991). 2.9 Å Resolution Structure of an anti-dinitrophenyl-spin-label monoclonal antibody fab fragment with bound hapten. *J. Mol. Biol.* **221**, 239–256.

Bystroff, C. & Kraut, J. (1991). Crystal structure of unliganded *Escherichia coli* dihydrofolate reductase-ligand-induced conformational changes and cooperativity in binding. *Biochemistry,* **30**, 2227–2239.

Chen, C. C. H., Rahil, J., Pratt, R. F. & Herzberg, O. (1993). Structure of a phosphonate inhibited β-lactamase. An analog of the tetrahedral transition state/intermediate of β-lactam hydrolysis. *J. Mol. Biol.* **234**, 165–178.

Clark, D. E., Jones, G., Willett, P., Kenny, P. W. & Glen, R. C. (1994). Pharmacophoric pattern matching in files of three-dimensional structures: comparison of conformational-searching algorithms for flexible searching. *J. Chem. Inform. Comput. Sci.* **34**, 197–206.

Clark, M., Cramer, R. D. & Van Opdenbosch, N. (1989). Validation of the general-purpose TRIPOS 5.2 force field. *J. Comput. Chem.* **10**, 982–1012.

Connolly, M. J. (1983). Analytical molecular surface calculation. *J. Appl. Crystallog.* **16**, 548–558.

Cooper, J. B., Quail, W., Frazao, C., Foundling, S. I. & Blundell, T. L. (1992). X-ray crystalographic analysis of inhibition of endothiapepsin by cyclohexyl renin inhibitors. *Biochemistry,* **31**, 8142–8150.

Dandekar, T. & Argos, P. (1996). Identifying the tertiary fold of small proteins with different topologies from sequence and secondary structure using the genetic algorithm and extended criteria specific for strand regions. *J. Mol. Biol.* **256**, 645–660.

Davis, L. D. (1991). Editor of *Handbook of Genetic Algorithms,* Van Nostrand Rheinhold, New York.

Delaney, J. S. (1992). Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph.* **10**, 174–177.

Diederichs, K. & Schultz, G. E. (1991). The refined structure of the complex between adenylate kynase from beef-heart mitochondrial matrix and its substrate AMP at 1.85 Å resolution. *J. Mol. Biol.* **217**, 541–549.

Dreyer, G. B., Lambert, D. M., Meek, T. D., Carr, T. J., Tomaszek, T. A., Fernandez, A. V., Bartus, H., Cacciavillani, E., Hassel, A. M., Minnich, M., Petteway, S. R. & Metcalf, B. W. (1992). Hydroxyethylene isostere inhibitors of human immunodeficiency virus-I protease: structure-activity analysis using enzyme kinetics, X-ray crystallography, and infected T-cell assays. *Biochemistry,* **31**, 6646–6639.

Eads, J., Sacchettini, J. C., Kromminga, A. & Gordon, J. I. (1993). *Escherichia coli*-derived rat intestinal fatty acid binding protein with bound myristate at 1.5 Å resolution and I-FABP$^{\mathrm{Arg106} \rightarrow \mathrm{Gln}}$ with bound oleate at 1.74 Å resolution. *J. Biol. Chem.* **268**, 26375–26385.

Edmundson, A. B., Harris, D. L., Fan, Z. C., Guddat, L. W., Schley, B. T., Hanson, B. L., Tribbick, G. & Geysen, H. M. (1993). Principles and pitfalls in designing site-directed peptide ligands. *Proteins: Struct. Funct. Genet.* **16**, 246–267.

Filman, D. J., Syed, R., Chow, M., Macadam, A. J., Minor, P. D. & Hogle, J. M. (1989). Structural factors that control transitions and serotype specificity in type 3 poliovirus. *EMBO J.* **8**, 1567–1579.

Fraternali, F. & van Gunsteren, W. F. (1996). An efficient mean solvation force model for use in molecular dynamics simulations of proteins in aqueous solution. *J. Mol. Biol.* **256**, 939–948.

Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J. & Freer, S. T. (1995). Molecular recognition of the inhibitor AG-1343 by HIV-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol.* **2**, 317–324.

Ghosh, D., Erman, M., Wawrzak, Z., Duax, W. L. & Pangborn, W. (1994). Mechanism of inhibition of 3α,20β-hydroxysteroid dehydrogenase by a licorice-derived steroidal inhibitor. *Structure,* **2**, 973–980.

Glen, R. C. (1994). A fast empirical method for the calculation of molecular polarizability. *J. Comput. Aid. Mol. Des.* **8**, 457–466.

Glusker, J. P. (1991). Structural Aspects of metal liganding to functional groups in proteins. *Advan. Protein Chem.* **42**, 1–76.

Goldberg, D. E. (1989). *Genetic Algorithms in Search Opimization and Machine Learning,* Addison-Wesley, Reading, MA.

Goldberg, D. E. & Deb, K. (1989). An investigation of niche and species formation in genetic function optimisation. In *Proceedings of the Third International Conference on Genetic Algorithms and their Applications* (Schaffer, D., ed.), pp. 42–50, Morgan Kaufmann, San Mateo, CA.

Goodsell, D. S. & Olson, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Struct. Funct. Genet.* **8**, 195–202.

Goto, H. & Osawa, E. (1989). Corner flapping: a simple and fast algorithm for exhaustive generation of ring conformations. *J. Am. Chem. Soc.* **111**, 8950–8951.

Hamilton, J. A., Steinrauf, L. K., Braden, B. C., Leipnieks, J., Benson, M. D., Holmgren, G., Sandgren, O. & Steen, L. (1993). The X-ray crystal structure refinements of normal human transthyretin and the amyloidogenic Val30 → Met variant to 1.7 Å resolution. *J. Biol. Chem.* **268**, 2416–2424.

Harel, M., Schalk, I., Ehret-Sabattier, L., Bouet, F., Goeldner, M., Hirth, C., Axelsen, P., Silman, I. & Sussman, J. (1993). Quaternary ligand binding to aromatic residues in the active site gorge of acetylcholinesterase. *Proc. Natl Acad. Sci., USA,* **90**, 9031–9035.

Hayes, I. C. & Stone, A. J. (1984). An intermolecular perturbation theory for the region of moderate overlap. *Mol. Phys.* **53**, 83–105.

Herron, J. N., He, X. M., Mason, M. L., Voss, E. W. & Edmundson, A. B. (1989). Three-dimensional structure of a fluorescein FAB complex crystallized in 2-methyl-2,4-pentanediol. *Proteins: Struct. Funct. Genet.* **5**, 271–280.

Herzberg, O. (1991). Refined crystal structure of β-lactamase from *Staphylococcus aureus* PC1 at 2.0 Å resolution. *J. Mol. Biol.* **217**, 701–719.

Hirschfelder, J. O., Curtiss, C. F. & Bird, R. B. (1964). *Molecular Theory of Gases and Liquids,* John Wiley, New York.

Hinze, J. & Jaffe, H. H. (1962). Electronegativity. I. Orbital electronegativity of neutral atoms. *J. Am. Chem. Soc.* **84**, 540–546.

Ho, C. M. W. & Marshall, G. R. (1990). Cavity search: an algorithm for the isolation and display of cavity like binding regions. *J. Comput. Aid. Mol. Des.* **4**, 337–354.

Holland, J. H. (1992). *Adaptation in Natural and Artificial Systems,* MIT Press, Cambridge, MA.

Holt, D. A., Luengo, J. I., Yamashita, D. S., Oh, H. J., Konialian, A. L., Yen, H. K., Rozamus, L. E., Brandt, M., Bossard, M. J., Levy, M. A., Eggleston, D. S., Liang, J., Schultz, L. W., Stout, T. J. & Clardy, J. (1993). Design, synthesis, and kenetic evaluation of high-affinity FKBP ligands and the x-ray crystal structures of their complexes with FKBP12. *J. Am. Chem. Soc.* **115**, 9925–9938.

James, N. M. G., Sielecki, A. R. & Moult, J. (1983). Crystallographic analysis of a pepstatin analogue binding to the aspartyl proteinase penicillopepsin at 1.8 Å resolution. In *Peptides: Structure and Function, Proceedings of the Eighth American Peptide Symposium*

(Hruby, V. J. & Rich, D. H., eds), pp. 521–531, Pierce Chemical Company, Rockford, IL.

Jedrzejas, M. J., Singh, S., Brouillette, W. J., Laver, W. G., Air, G. M. & Luo, M. (1995). Structures of aromatic inhibitors of influenza virus neuraminidase. *Biochemistry,* **34**, 3144–3151.

Jeffrey, P. D., Strong, R. K., Sieker, L. C., Chang, C. Y. Y., Campbell, R. L., Petsko, G. A., Haber, E., Margolies, M. N. & Sheriff, S. (1993). 26-10 FAB-digoxin complex: affinity and specificity due to surface complementarity. *Proc. Natl Acad. Sci., USA,* **90**, 10310–10314.

Jones, G. & Willett, P. (1995). Docking small molecule ligands into active sites. *Curr. Opin. Biotechnology,* **6**, 652–656.

Jones, G., Willett, P. & Glen, R. C. (1995a). Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **254**, 43–53.

Jones, G., Willett, P. & Glen, R. C. (1995b). A genetic algorithm for flexible molecular overlay and pharmacophore elucidation. *J. Comput. Aid. Mol. Des.* **9**, 532–549.

Judson, R. S., Jaeger, E. P. & Treasurywala, A. M. (1994). A genetic algorithm method for docking flexible molecules. *J. Mol. Struct.* **308**, 191–206.

Judson, R. S., Tan, Y. T., Mori, E., Melius, C., Jaeger, E. P., Treasurywala, A. M. & Mathiowetz, A. (1995). Docking flexible molecules: a case study of three proteins. *J. Comput. Chem.* **16**, 1405–1419.

Klebe, G. (1994). Mapping common molecular fragments in crystal structures to explore conformation and configuration space under the conditions of a molecular environment. *J. Mol. Struct.* **308**, 53–89.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R. & Ferrin, T. E. (1982). A geometric approach to macromolecule-ligand interactions. *J. Mol. Biol.* **161**, 269–288.

LaLonde, J. M., Bernlohr, D. A. & Banaszak, L. J. (1994). X-ray crystallographic studies of adipocyte lipid-binding protein complexed with palmitate and hexadecanesulfonic acid. Properties of cavity binding sites. *Biochemistry,* **33**, 4885–4895.

Landro, J. A., Gerlt, J. A. & Kozarich, J. W. (1994). The role of lysine 166 in the mechanism of mandelate racemase from *Pseudomonas putida*: mechanistic and crystallographic evidence for stereospecific alkylation by (*R*)-α-phenylglycidate. *Biochemisty,* **33**, 635–643.

Lauble, M., Kennedy, M. C., Beinert, H. & Stout, C. D. (1994). Crystal structures of aconitase with trans-aconitate and nitrocitrate bound. *J. Mol. Biol.* **237**, 437–451.

Leslie, A. G. W. (1990). Refined crystal structure of type III chloramphenicol acetyltransferase at 1.75 Å resolution. *J. Mol. Biol.* **213**, 167–186.

Lisgarten, J. N., Gupta, V., Maes, D., Wyns, L., Zegers, I., Palmer, R. A,, Dealwis, C. G., Aguilar, C. F. & Hemmings, A. M. (1993). Structure of the crystalline complex of cytidylic acid (2′-CMP) with ribonuclease at 1.6 Angstrom resolution-conservation of solvent sites in RNase-A high resolution structures. *Acta Crystallog. sect. D,* **49**, 541–547.

Lommerse, J. P. M., Stone, A. J., Taylor, R. & Allen, F. H. (1996). The nature and geometry of intermolecular interactions between halogens and oxygen or nitrogen. *J. Am. Chem. Soc.* **118**, 3108–3116.

Matthews, D. A., Bolin, J. T., Burridge, J. M., Filman, D. J., Volz, K. W., Kaufman, B. T., Beddell, C. R., Champness, J. N., Stammers, D. K. & Kraut, J. (1985). Refined crystal structures of *Escherichia coli* and chicken liver dihydrofolate reductase containing bound trimethoprim. *J. Biol. Chem.* **260**, 381–391.

Miller, M. D., Kearsley, S. K., Underwood, D. J. & Sheridan, R. P. (1994). FLOG: a system to select ′′quasi flexible′′ ligands complementary to a receptor of known three-dimensional structure. *J. Comput. Aided Mol. Des.* **8**, 153–174.

Mitchell, J. B. O. & Price, S. L. (1991). On the relative strengths of amide..amide and amide..water hydrogens bonds. *Chem. Phys. Letters,* **180**, 517–523.

Murthy, K. H. M., Winborne, E. L., Minnich, M. D., Culp, J. S. & Debouck, C. (1992). The crystal structures at 2.2- Å resolution of hydroxyethylene-based inhibitors bound to human immunodeficiency virus type 1 protease show that the inhibitors are present in two distinct orientations. *J. Biol. Chem.* **32**, 22770–22778.

Nicklaus, M. C., Wang, S. M., Driscoll, J. S. & Milne, G. W. A. (1995). Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **3**, 411–428.

Nonaka, T., Nakamura, T., Uesugi, S., Ikehara, M., Irie, M. & Mitsui, Y. (1993). Crystal structure of ribonuclease Ms (as a ribonuclease T_1 homologue) complexed with a guanyly-3′,5′-cytidine analogue. *Biochemistry,* **32**, 11825–11837.

Oshiro, C. M., Kuntz, I. D. & Dixon, J. S. (1995). Flexible ligand docking using a genetic algorithm. *J. Comput. Aided Mol. Des.* **9**, 113–130.

Payne, A. W. R. & Glen, R. C. (1993). Molecular recognition using a binary genetic search algorithm. *J. Mol. Graph.* **11**, 74–91.

Padlan, E. A., Cohen, G. H. & Davies, D. R. (1985). On the specificity of antibody 3-antigen interactions: phosphocholine binding to MCPC603 and the correlation of three-dimensional structure and sequence data. *Ann. Immunol. (Paris),* **C136**, 271–276.

Perry, K. M., Carreras, C. W., Chang, L. C., Santi, D. V. & Stroud, R. M. (1993). Structures of thymidylate synthase with a C-terminal deletion: role of the C-terminus in alignment of 2′-deoxyuridine 5′-monophosphate and 5,10-methylenetetrahydrofolate. *Biochemistry,* **32**, 7116–7125.

Peters, K. P., Fauck, J. & Frömmel, C. (1996). The automatic search for ligand binding sites in proteins of known three-dimensional structure using only geometric criteria. *J. Mol. Biol.* **256**, 201–213.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). Predicting receptor-ligand interactions by an incremental construction algorithm. *J. Mol. Biol.* **261**, 470–489.

Smith, G. D. & Dodson, G. G. (1992). The structure of a rhombohedral R6 insulin hexamer that binds phenol. *Biopolymers,* **32**, 441–445.

Starkweather, T., Whitley, D. & Mathias, K. (1990). Optimisation using distributed genetic algorithms. In *Parallel Problem Solving from Nature* (Schwefel, H. P. & Manner, R., eds), pp. 176–185, Springer-Verlag, Berlin.

Stewart, J. J. P. (1992). *MOPAC User Manual Version 6.0.* Quantum Chemical Program Exchange, Department of Chemistry, University of Indiana, Bloomington, IN.

Still, W. C., Tempczyk, A., Hawley, R. C. & Hendrickson, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129.

Stoddard, B. L., Bruhnke, J., Porter, N., Ringe, D. & Petsko, G. A. (1990). Structure and activity of two photoreversible cinnamates bound to chymotrypsin. *Biochemistry*, **29**, 4871–4879.

Sun, S. (1993). Reduced representation model of protein structure prediction: statistical potential and genetic algorithms. *Protein Sci.* **2**, 762–785.

Surles, M. C., Richardson, J. S., Richardson, D. C. & Brooks, F. P. (1994). Sculpting proteins interactively: continual energy minimization embedded in a graphical modeling system. *Protein Sci.* **3**, 198–210.

Tanese, R. (1989). Distributed genetic algorithms. In *Proceedings of the Third International Conference on Genetic Algorithms and their Applications* (Schaffer, D., ed.), pp. 434–439, Morgan Kaufmann, San Mateo, CA.

Teplyakov, A., Obolova, G., Wilson, K. S., Ishii, K., Kaji, H., Samejima, T. & Kuranova, I. (1994). Crystal-structure of inorganic pyrophosphatase from *Thermus thermophilus*. *Protein Sci.* **3**, 1098–1107.

Tong, L., Pav, S., Mui, S., Lamarre, D., Yoakim, C., Beaulieu, P. & Anderson, P. C. (1995). Crystal structures of HIV-2 protease in complex with inhibitors containing the hydroxyethylamine dipeptide isostere. *Structure*, **3**, 33–40.

Van Duyne, G. D., Standaert, R. F., Karplus, P. A., Schreiber, S. L. & Clardy, J. (1993). Atomic structures of human immunophilin FKBP-12 complexes with FK506 and rapamycin. *J. Mol. Biol.* **229**, 105–124.

Wang, J., Smerdon, S. J., Jäger, J., Kohlstaedt, L. A., Rice, P. A., Friedman, J. M. & Steitz, T. A. (1994). Structural basis of asymmetry in the human immunodeficiency virus type 1 reverse transcriptase heterodimer. *Proc. Natl Acad. Sci. USA*, **91**, 7242–7246.

Zhang, A., Nanni, R. G., Li, T., Arnold, G. F., Oren, D. A., Jacobo-Molina, A., Williams, R. L., Kamer, G., Rubenstein, D. A., Li, Y., Rozhon, E., Cox, S., Buontempo, P., O'Connell, J., Schwartz, J., Miller, G., Bauer, B., Versace, R., Pinto, P., Ganguly, A., Girijavallabhan, V. & Arnold, E. (1993). Structure determination of antiviral compound SCH 38057 complexed with human rhinovirus 14. *J. Mol. Biol.* **230**, 857–867.

Zhou, G. W., Gou, J., Huang, W., Fletterick, R. J. & Scanlan, T. S. (1994). Crystal structure of a catalytic antibody with a serine protease active site. *Science*, **265**, 1059–1064.

Edited by F. E. Cohen

(Received 22 August 1996; received in revised form 16 December 1996; accepted 20 December 1996)