Journal of Computer-Aided Molecular Design, 11 (1997) 557–569.
KLUWER/ESCOM

J-CAMD 419

Comparison of protein surfaces using a genetic algorithm

Andrew R. Poirrette[a,b,*], Peter J. Artymiuk[b], David W. Rice[b] and Peter Willett[a]

Krebs Institute for Biomolecular Research, Departments of [a]Information Studies and [b]Molecular Biology and Biotechnology, University of Sheffield, Firth Court, Western Bank, Sheffield S10 2TN, U.K.

Received 16 May 1997
Accepted 10 August 1997

Keywords: Protein structure; Molecular surfaces; Structure comparison; Binding sites; Evolutionary computation

Summary

A genetic algorithm (GA) is described which is used to compare the solvent-accessible surfaces of two proteins or fragments of proteins, represented by a dot surface calculated using the Connolly algorithm. The GA is used to move one surface relative to the other to locate the most similar surface region between the two. The matching process is enhanced by the use of the surface normals and shape terms provided by the Connolly program and also by a simple hydrogen-bonding descriptor and an additional shape descriptor. The algorithm has been tested in applications ranging from the comparison of small surface patches to the comparison of whole protein surfaces, and it has performed correctly in all cases. Examples of the matches are given and a quantitative analysis of the quality of the matches is performed. A number of possible future enhancements to the program are described which would allow the GA to be used for more complex surface comparisons.

Introduction

The development of automated methods for the comparison of molecular surfaces, be they of small or large molecules, is important in areas of drug discovery, in the understanding of structure–function relationships in biological macromolecules and in exploring the interactions that are crucial in molecular recognition.

Many studies have been performed on the comparison of small molecules to identify how they may be optimally superimposed, with the aim, for example, of identifying new lead compounds in drug discovery. These studies have involved the comparison of many different features, including interatomic distances [1], electrostatic fields [2], pharmacophores [3] and functional groups [4]. A number of these methods have involved comparison of the shapes of molecules, whether represented by their solvent-accessible surface or by some other means and these have been of varying success in the examples studied.

For example, Chau and Dean [5] used a technique called gnomonic projection to provide an overall, even sampling of points on the surface of small molecules and thus aid their comparison. Masek et al. [6] represented the surfaces as a fixed thickness skin, where the first skin surface was formed from the solvent-accessible or van der Waals surface of the molecule and a second surface was created at a fixed distance from the first, to allow comparison of the shapes of small molecules by determination of the 'skin' intersection volume. In a similar approach to that of Masek, Perkins et al. [7] also used the idea of a surface volume intersection, in this case using a grid-based matching system and simulated annealing to optimise the overlay between small molecules.

Macromolecular superposition is a computationally far more demanding problem than small molecule superposition and the methods used for small molecule superposition have generally not been applied to the analysis of macromolecular surface similarity. Comparisons between proteins to determine their degree of similarity in three dimensions (as opposed to the many one-dimensional sequence alignment techniques) have largely concentrated on the superposition of atoms from the protein main chain or of representations of secondary structure elements. These techniques have enabled the identification of proteins with similar folds, and hence possibly similar function, where there is no homology at the primary

*To whom correspondence should be addressed. Present address: Zeneca Pharmaceuticals, Mereside, Alderley Park, Macclesfield, Cheshire SK10 4TG, U.K.

sequence level and have provided evidence for establishing the functional and evolutionary relationships between similar proteins [8–10].

However, there has been recent interest in the use of shape-based methods to compare protein exteriors. Thus, Leicester and co-workers have developed a technique that makes use of Fourier series descriptors to describe the global surface shape of proteins [11,12]. A series of proteins were represented in this manner, and then compared and clustered based on their shape descriptors. The technique was moderately successful, but it was found that the descriptor suffered from a lack of sensitivity as the lower order shape descriptors tended to dominate so that the method could only be applied for global shape comparison, and not for finding localised similarities. In another approach to this problem, Lewis and Rees [13] and Zachmann and Brickmann [14] have made use of the concept of fractals to characterise the roughness of protein surfaces. In this way, Zachmann and co-workers have described a parameter based on fractals called the Hausdorff mass. A number of globular proteins were represented by their Connolly surfaces and the Hausdorff masses were calculated both at the local and global level. The results showed that it was possible to perform a comparison of proteins, as represented by their Hausdorff mass, at the global level but that, once again, the use of measures of fractality was not applicable at a more local level, e.g., for the identification of active site areas.

The two approaches described above have allowed a numerical measure of shape similarity but have not involved any attempt to perform a superposition of the proteins based on the surface descriptor. Alternative representations for protein surfaces have been used for protein–protein docking studies that might also be applicable to the comparison and possible superposition of proteins. The representations described have included (i) the use of B-splines to smooth the Connolly surface and remove small surface bumps whilst leaving major features alone [15]; (ii) the use of sparse critical points to characterise the major features of the Connolly surface, thus reducing the number of points that represent the surface, whilst maintaining detail, in order to make computation quicker [16]; (iii) the calculation of local and global curvature profiles [17], again based on the Connolly surface; and (iv) the use of solid angles to determine surface concavity and convexity [18].

Nussinov and co-workers [19] have developed a program based on algorithms adapted from research into computer vision. They have described a method for protein comparison at a number of different levels, both global and local and using 'surface' descriptors or backbone atoms. An active site surface motif was generated for each site to be compared by identifying all residues that lay within a certain distance of the site. The comparison and superposition of the site surface motifs was then based on the atomic coordinates of the active site residue atoms. The program was shown to be successful in comparing the active site surface motifs of haemoglobin and myoglobin, and of alcohol dehydrogenase and lactate dehydrogenase [19].

Recently there has been great interest in the use of genetic algorithms (GA) for a variety of applications in structural molecular biology. A GA is a non-deterministic algorithm which attempts to mimic the processes of Darwinian evolution to find good, but not necessarily optimal, solutions to computationally demanding optimisation problems that cannot be handled satisfactorily by deterministic algorithms. GAs were originally developed by Holland [20] and have been widely implemented in many different areas. In the chemical and molecular biological fields, applications of GAs have included the investigation of processes of protein folding [21,22], ligand binding to protein receptor sites [23] and the determination of small molecule similarity [4]. In an application that is more similar to the work described here, May and Johnson [24,25] describe the use of a GA for protein structural comparison where the transformations for shifting one protein relative to the other are encoded in the chromosome population (see below) and the proteins are then compared at the α-carbon level using an evaluation function based upon a dynamic programming routine.

In this paper, we present an approach to the comparison of proteins that is based on superposition of their solvent-accessible surfaces using a GA. The steric component of molecular recognition is of prime importance in the recognition process, since proteins that bind to a common receptor or have a common ligand are likely to have similarly shaped binding surfaces. The aim of the work presented in this paper has been to develop an algorithm to identify areas of maximum surface similarity between two proteins and thus allow the identification of areas on their surfaces that may bind a common ligand or bind to a common receptor. The molecular surfaces are generated using Connolly's MS program [26] and are represented as a series of surface points, or 'dots'. We demonstrate that this approach provides an effective tool for the detection of similarity in molecular surfaces in a variety of applications ranging from fragmentary surfaces representing localised binding sites to complete protein surfaces.

Computational Methods

Surface representation

The solvent-accessible surface of each protein molecule was represented by a series of 3D coordinates or 'dots', using Connolly's surface program [26,27]. The latter program is easy to use, relatively fast in its operation and it allows the generation of a surface at any level of resol-

ution, based on a user-specified probe size. In addition, the program also calculates additional information associated with each surface point, as discussed below, and this has proved very useful for the matching process described in this paper. The Connolly surface is generated by rolling a probe over the van der Waals surface of the protein and is based on the molecular surface definition of Richards [28], which consists of two components, the contact surface and the reentrant surface. The contact surface is defined by the locations where the probe touches the surface at one point, whilst the reentrant surface is defined by the locations where the inward facing face of the probe is in contact with more than one atom.

For each surface point the Connolly program outputs its 3D coordinates, an associated normal vector and a numerical indicator of the shape of the surface at that point, classified as convex, concave or saddle-shaped. The program was adapted slightly so that each point was additionally labelled with a status indicator of hydrogen-bonding potentiality, based on the type of protein atom nearest to the surface point. In this implementation four values were possible: H-donor, H-acceptor, H-donor/acceptor, and non-H-bonding. All oxygen atoms were labelled as H-acceptors with the exception of the O^{η} of tyrosine and the O^{γ} of serine and threonine, both of which were labelled as H-donor/acceptors. For simplicity all nitrogen atoms were labelled as H-bond donors, including the histidine nitrogens, N^{δ} and N^{ε}, whose individual protonation states are difficult to assign. All other atoms were labelled as non-H-bonding. Clearly, it would be quite simple to adopt alternative schemes for the characterisation of surface points in future investigations.

As a starting point for the comparison process, the centres of gravity of both target and query protein coordinates were translated to the origin of space. A 3D grid was placed around the target surface, with the relative dimensions of the grid being set sufficiently large such that any surface point from the query molecule could be positioned on the target surface while the query molecule surface was contained within the grid boundaries. The size of the individual grid elements was a user-definable option and was set at 2 Å for all the comparisons described here. In a similar method to that employed by Jiang and Kim in their protein–protein docking work [29], surface points contained within the same grid element were given the same coordinates, which in this case were set to be equal to the corner of the grid nearest to the origin. This procedure had the effect of coarsening the definition of the surface and thus allowed surfaces to be considered as matching when they might otherwise have been missed. In all the comparisons described here, the larger of the two molecules being compared (as defined by the number of surface points) was used as the target. The GA-encoded transformations were hence applied to the molecule with fewer surface points, thus improving the efficiency of the program. A rotation interval of 3° and a translational step size of 0.1 Å were defined for enabling the transformations.

The genetic algorithm

The GA was used to generate translations and rotations of the query protein surface relative to the target protein surface which was held static. The rotation and translation was applied and then a fitness function (see below) was used to determine the degree of overlap and similarity between the rotated query surface and the target surface. In this implementation a steady-state-with-no-duplicates GA has been designed [30,31]. The query protein surface was free to be moved to any position within the bounds of the previously defined grid and both proteins were considered to be entirely rigid.

In a GA, potential solutions to the problem being studied are represented as data structures called chromosomes. For this particular problem an integer-string chromosome was used. Each chromosome in the population contained six elements corresponding to the six degrees of freedom necessary to completely move one rigid body relative to another. The first three elements encoded rotations about the x, y and z axes, while the remaining three elements encoded translations along the original x, y and z axes. An initial population of chromosomes was generated at random and for each chromosome the encoded rotations and translations were applied to the query protein surface, which was first translated within the grid surrounding the target and then rotated about axes centred on its centre of gravity. The fitness function of the GA was then used to evaluate the efficacy of the individual chromosomes and each chromosome was given a fitness score, those that produced better solutions being given a higher fitness and vice versa.

The fitness function evaluated the degree of similarity between the target and query surfaces, after the transformations encoded by a chromosome had been applied. The underlying assessment of similarity was a simple count of the number of surface points in common to within a user-defined distance tolerance, as determined by the grid element size, but this was qualified in a number of different ways. Two points were considered to be matching if they occupied the same region of space and, optionally, had the same chemical and/or shape characteristics, i.e., they had to have the same Connolly shape characteristic, the same H-donor/acceptor characteristic and surface normals that pointed in the same direction, to within some predefined tolerance. At present, all points that contribute to a match are given an equal weight and one future enhancement of the program may be the introduction of features which allow two potential solutions with an identical number of surface points in common to be distinguished, perhaps in terms of the degree of clustering of matched grid points. This has not yet been done,

as the program has performed well using the present simple fitness function.

Having evaluated the initial population, the GA then ran in a series of cycles and at each iteration a subset of members of the existing population was replaced with an equal number of new members and the fitness of these new members was evaluated, the aim being to increase the average fitness of the population and hence move towards the optimal solution to the problem. The replacement of population members was controlled by the use of three genetic operators, i.e., parental selection, crossover and mutation.

Roulette-wheel parental selection [31] was used to choose parents for the production of new members for the next generation. Conceptually, roulette-wheel selection works by giving each member of the population a slice of a roulette wheel, the size of the slice being proportional to the fitness of the solution. In this way, when the wheel is spun, there is a chance that any member from the original population may be chosen as a parent but there is an increased likelihood that a fitter member will be chosen over a less fit member. In order to avoid giving too much weight to the fittest individuals in a population, and hence causing premature convergence of the GA on a local minimum, the fitness values for all of the population members were subject to a linear normalisation routine prior to roulette-wheel selection. The normalisation was carried out with a selection pressure of 1.1. The selection pressure represents the probability of choosing the fittest individual relative to an average individual [4].

The genetic operators crossover and mutation were applied to the chosen parents to create new population members. New chromosomes which by chance were duplicates of pre-existing members of the population were discarded and new members were chosen as parents. Crossover was applied in three possible ways. In all cases, two parents were chosen. In single-point crossover a position along the chromosome was chosen at random and all points subsequent to the chosen point were then swapped over between the two chromosomes. In exchange-crossover, a point was again chosen at random and just that position was swapped between the two chromosomes. The final crossover operator, called average crossover, took two parents and produced only one child, the elements of which were the arithmetic mean of the equivalent parental positions. Mutation was applied in two possible ways, both requiring only one parent and producing one child chromosome. In standard mutation, a position was chosen at random along the parent chromosome and this was randomly mutated to a new value. Random-creep mutation operated in the same way, except that in this case the chosen position was randomly altered within a small predefined range of ±5° for rotations and ±2 Å for translations, the ranges having been found to be effective in preliminary testing of the program.

The algorithm was halted after some predefined number of iterations, or when a certain fitness value had been reached and had remained unchanged for a number of iterations. The fittest member of this final population was considered to be the best solution to the problem. Because of the nature of the GA, different runs with identical parameter settings will often result in different solutions and it is the usual practice to run the GA a number of times and to keep the overall fittest solution from all of the runs.

After the initial tests of the GA, two additional optional procedures were incorporated to enhance the basic matching procedure described above. A further shape matching condition was introduced, involving the location of surface invaginations [32] prior to running the GA. The shape matching condition then allowed the sub-classification of the surface points into two types, those lying within invaginations and those considered to be outside invaginations. Matching could then be further limited by the criterion that all points matching in space and fulfilling all the previous criteria must also match in their invagination designation.

The GA was also modified to include a technique termed niche restriction [33]. Standard GAs, similar in design to that discussed above, are designed to converge to a single optimal solution. The niche restriction technique has been developed to enable GAs to provide solutions to multimodal problems, i.e., where more than one solution exists. In this case, however, the technique was applied to enable the identification of sub-optimal solutions and remove them from the search space. Our algorithm used a modification of the sequential niche technique of Beasley et al. [33]. The GA was run N times, where N is the number of niches that will ultimately be identified. After the first run-through, the fittest solution was identified and an area surrounding it was removed from the search space. The actual area removed was a sphere of user-definable size centred on the coordinates of the centre of gravity of the fittest solution. The value must be kept sufficiently small so that the optimum solution is not inadvertently excluded from the search space by virtue of the GA finding a local minimum nearby, i.e., within the niche sphere diameter setting, whilst keeping it sufficiently large to enable the GA to perform a better overall search of the space available. The GA was then re-run as before, but during this run no members of the population were allowed to be in the area defined by the niche. After this second run a second niche area was defined and the process was repeated until N runs had been completed. The fittest solution from the N runs was then identified as the overall best solution.

Choice of GA parameter settings

Successful use of a GA requires extensive testing to identify appropriate default values for the various para-

meters of the algorithm. It was not possible to perform exhaustive testing of all combinations of parameter settings, as the stochastic nature of the GA means that different runs with identical parameters will give different results. However, the evaluations carried out allowed the identification of those parameter combinations that performed well most often and we were able to identify a set of parameters (see Table 1) which were used in all the examples given in this paper, with the minor exception of two parameters in test 6 (described below). These parameters gave a satisfactory performance in all cases. The choice of crossover or mutation as genetic operators within an actual run of the GA is controlled by the mutation/crossover parameter. A value of 50% was used here, meaning that there was a 50% chance that the operator chosen would be crossover and a 50% chance that it would be mutation. Within each of the operator choices, there were then further parameters controlling which of the individual mutation or crossover types to choose. These parameters were all set so that there was an equal chance of choosing either random or random creep mutation if mutation was chosen as the operator or choosing single-point, exchange or average crossover if crossover was chosen as the operator.

TABLE 1
THE PARAMETER SETTINGS USED BY DEFAULT IN THE GA

Parameter	Value
Population size	400
Number of runs	1000
Percentage of population replaced each iteration (%)	5
Mutation/crossover rate (%)	50
Selection pressure	1.1
Niche sphere diameter (Å)	2 (5)
Angle allowed between normals (°)	±60
Number of niches created	5 (10)
Creep range for rotation (°)	5
Creep range for translation (Å)	2

Figures in brackets for the niche sphere diameter and the number of niches created were alternative values used in one of the test cases (test 6; see text for details).

Results

For testing purposes, the GA has been used for the comparison of a number of different protein surfaces. Five of the comparisons represent searches for the surface of a binding site from the query protein in a target protein known to bind the same ligand. The first query pattern was the HIV-protease binding site surface from an HIV protease in complex with a drug candidate and this was searched against the surface defined for the uncomplexed HIV protease. In the second case, the methotrexate binding site from dihydrofolate reductase was searched against the apo form of dihydrofolate reductase. Three further examples of 'site-based' searches used different proteins for the definition of both query and target surfaces. These were: a search for the haem binding site surface from myoglobin against the surface of haemoglobin; a search for the sialic acid binding site surface from bacterial sialidase against the surface of a viral sialidase; and a search for the NAD binding site surface from alcohol dehydrogenase against the surface of lactate dehydrogenase. A further example search, using the surface of lysozyme as the query and a lysozyme–antibody complex surface as the target, represents a search using an entire protein surface as the query. In the final example, the GA has been applied to a slightly different problem, involving the comparison of the surfaces of two elastase inhibitors (a large 56-residue polypeptide and a much smaller ligand) that are both known to bind to the same receptor.

As stated above, the same set of GA parameters, detailed in Table 1, were used in all the examples given in this paper, with the minor exception of two parameters in test 6. In terms of the fitness function, all the test runs were performed with the inclusion of all possible labelling of the individual surface points, i.e., each point was labelled with its Connolly shape indicator, its H-bond formation indicator, its invagination indicator (except for test 7, where the query pattern was a small molecule and this was inappropriate), and its normal vector.

Evaluation of the program

In order to test the effectiveness of the method, seven test comparisons were conducted. These were chosen to cover a wide range of applications, ranging from comparison of small surface patches representing binding sites and ligand surfaces to comparison of complete protein surfaces. All of the tests have involved searching for surface similarities where it was known that the target protein binds the same ligand or, as in the case of the lysozyme search (test 6), contains the query protein itself.

In the general case, the program is intended for locating similar surface regions in completely different proteins which may nevertheless have similar binding sites. In such a case there would be no detectable similarities at the level of the overall fold or local C^{α} traces. However, in order to validate the program, we have chosen examples where such similarities of fold do exist so that we can retrospectively validate the final surface superposition. It is important to emphasise that the information regarding C^{α}s is not used in our GA comparison method. We are therefore able to independently check the correctness of our surface superpositions by subsequent C^{α} superposition. Thus, if a surface superposition is correct, then the transformation obtained should also give a good superposition of the underlying atomic structures. Therefore, choice of these examples enables us to obtain proof-of-concept of our method.

TABLE 2
SUMMARY RESULTS FROM 10 RUNS OF THE GA FOR EACH OF THE 7 SURFACE COMPARISONS

Test number	Number of runs with rmse ≤ 2 Å	Best fitness (points in common)	Rmse associated with best fitness (Å)	Best overall rmse (Å)	Fitness of best overall rmse (points in common)	Best possible rmse (Å)	Search time (CPU min)
1	6	243	1.6	1.6	243	0.8	9
2	10	408	2.0	2.0	408	1.9	22
3	7	440	2.3	1.5	438	1.1	24
4	10	191	1.1	1.1	191	0.5	16
5	3	194	1.0	1.0	194	0.4	17
6	6	1350	0.8	0.8	1350	0.5	122
7	6	135	1.1	0.8	129	0.3	4

Column 2 shows the number of runs where the GA predicted solutions with rmse values below 2 Å. Column 3 indicates the best fitness achieved over the 10 runs, together with the associated rmse (column 4). The best rmse value from the 10 runs is shown in column 5, together with its associated fitness score (column 6). Rmse values are to the nearest tenth of an Å. For comparison purposes the best possible rmse values for superposition are given in column 7. The final column in the table gives representative CPU times for the comparisons (all performed on a DEC Alpha 3000).

As the molecules concerned may well be in different conformations, a root mean square error (rmse) value for the 'best possible' fit between the atoms has also been calculated and the relevant values are given in Table 2. This represents a lower limit on attainable values that might be achieved from an overlap predicted by the surface superposition. However, in practice, we would only expect a somewhat higher rmse to be achieved. This follows from the empirical nature of the scoring function and inherent differences in the molecular details of the surfaces that are being compared. Three of the seven surface comparisons were selected because they have been used by other researchers to test the efficacy of their own programs: tests 3 and 5 are very similar to those used by Fischer et al. [19] when comparing the active sites of proteins using the geometric hashing paradigm and test 7 was used by Masek et al. [6] and subsequently by Perkins and Dean [7] for the evaluation of their surface similarity programs.

In all the surface comparisons performed, the query and target surfaces were based upon PDB deposited coordinate sets [34] and generated at a Connolly surface resolution of 1 surface point/Å^2, using a probe with a diameter of 1.5 Å. Table 3 summarises the comparisons performed and provides details of the relative sizes and the number of surface points of the target and query structures. The GA was run 10 times for each of the seven surface comparisons and a summary of the results for these tests is given in Table 2. The table details how many of the 10 runs provided a solution transformation that, when applied to the ligand/main-chain atoms of the query, gave rmse values for superposition onto the target of less than 2 Å, from which it is clear that in all but one of the searches the GA has successfully superposed the query surface onto the target surface in the majority of the runs performed.

Test 1: HIV protease inhibitor site

The surface of the active site from HIV-1 protease in complex with a drug candidate [35; PDB code 1aaq] was

TABLE 3
THE COMPARISONS PERFORMED BY THE GA

Test number	Query pattern	Number of amino acids	Size (surface points)	Target protein	Number of amino acids	Size (surface points)
1	1aaq	46	1686	1hhp[a]	198	7810
2	4dfr	15	619	5dfr	159	6517
3	4mbn	32	1474	2hhb[b]	141	6759
4	1nnb	31	1173	2sim	381	13 651
5	2ohx	28	1070	9ldt[b]	332	14 497
6	3lym	129	4970	1fdl	561	21 485
7	DFKi	[c]	324	TOMI	56	2514

For each query pattern and target protein the number of amino acids involved in the creation of the surface and the number of surface points generated are detailed.

[a] The PDB deposition for this protein contained the coordinates of one monomer. The active protein is a dimer and this was created using the PDB cryst record supplied.

[b] The target patterns for this protein were created from only one of the monomers in the asymmetric unit.

[c] The DFKi elastase inhibitor contains three amino acid residues in addition to two other chemical moieties.

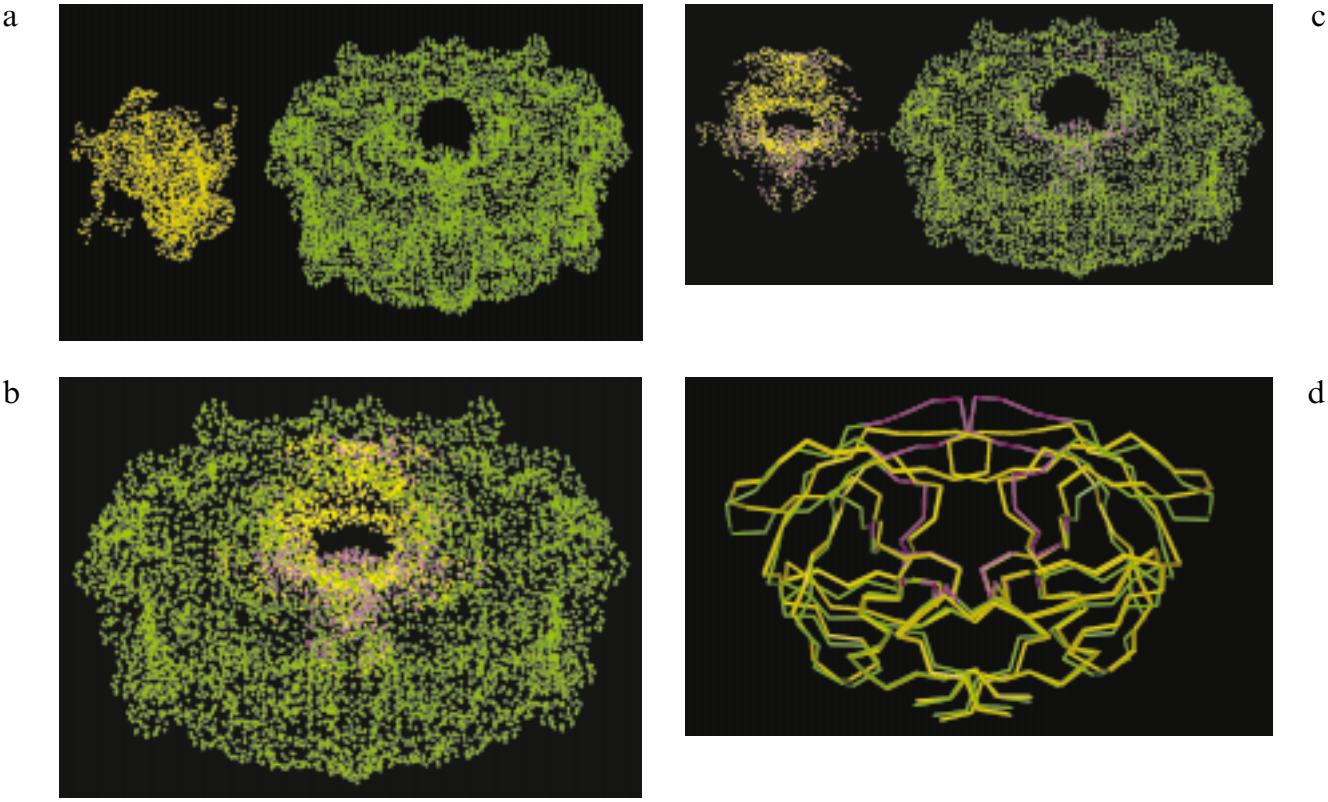


Fig. 1. (a) Connolly dot surfaces for the active site query from HIV-1 protease 1aaq (yellow) and the target HIV-1 protease 1hhp (green). (b) The best overlay of the active site query surface on the target surface predicted by the GA with the matching surface dots highlighted in magenta (other colours as for (a)). (c) As in (b), but with the matching query surface translated to the left of the target for clarity. (d) Overlays of the α-carbon trace of HIV-1 protease (1aaq) on the α-carbon trace of uncomplexed HIV-1 protease (1hhp), based on the transformations output by the GA for superposing the active site surface of 1aaq on the surface of 1hhp. The chain traces for the 10 overlays of 1aaq produced by the GA are virtually superimposable and are shown in yellow. The chain trace for 1hhp is shown in green, with the active site region highlighted in magenta to illustrate the large change in conformation between the complexed and uncomplexed forms of the enzyme.

searched against the surface of HIV-1 protease in uncomplexed form [36; PDB code 1hhp]. The ligand was removed from the structure and the query pattern surface was generated by selecting all those residues that lay within 12 Å of the centre of the active site. In all, a total of 46 residues were identified (23 from each monomer in this dimeric enzyme) and these were used to create a Connolly surface of 1686 surface points (shown in yellow in Fig. 1a). The target surface was generated based on the active dimeric form of the enzyme and consisted of 7810 surface points (shown in green in Fig. 1a). The active sites of the complexed and uncomplexed forms of the enzyme are quite different, as a major conformational change occurs in the region of the active site on ligand binding. This means there is a much greater difference in residue positions between the sites than is seen with the dihydrofolate reductase search discussed below in test 2. For example, the rmse for superposition of all atoms of the 46 residues identified for creation of the active site surface is 2.8 Å, including several atoms separated by distances in excess of 8 Å. Nevertheless, the search was performed very successfully. As can be seen from Table 2, the best possible superposition of main-chain atoms is 1.9 Å, and

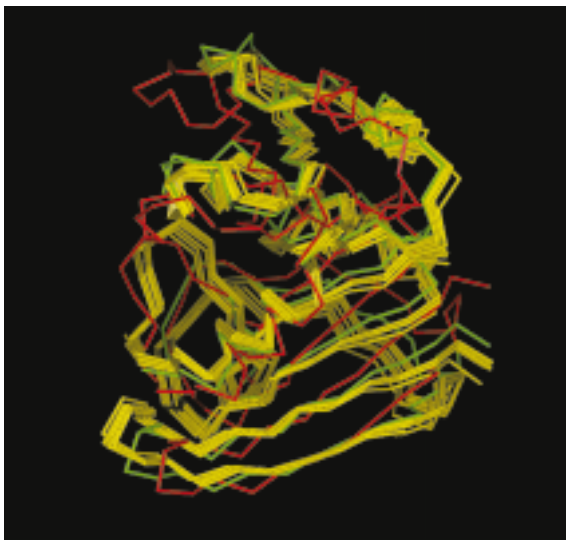


Fig. 2. Overlays of the α-carbon trace of dihydrofolate reductase (4dfr) on the α-carbon trace of apo-dihydrofolate reductase (5dfr), based on the transformations output by the GA for superposing the methotrexate binding site surface from 4dfr on the surface of 5dfr. The structure for apo-dihydrofolate reductase is in green, whilst the yellow traces represent the nine high-fitness solutions for 4dfr produced by the GA. The single poor solution for 4dfr is shown in red.

in 8 of the 10 runs of the GA a solution was found that mapped the main-chain atoms of the proteins with an rmse of 2.0 Å, whilst the two remaining runs provided solutions with rmse values of 2.1 Å. The superposition of surfaces corresponding to the best fitness is shown in Fig. 1b, where the surface dots that contributed to the match are shown highlighted in magenta. Figure 1c emphasises the matching dots and it is clear that, as would be expected, the majority are located in the region of the binding surface that undergoes the smallest conformational change on binding the ligand. In Fig. 1d the C^{α} trace for the uncomplexed target protein is shown in green with the active site region emphasized in magenta, whilst the 10 yellow traces show the 10 solutions predicted by the GA. The change in position of the main-chain atoms in the region of the site, that occurs on binding a ligand, can be clearly seen. It is clear that in this case the shape of the surface of the site was still sufficiently similar between the structures to allow the GA to successfully perform a match.

Test 2: Methotrexate binding site

The methotrexate binding site surface from *Escherichia coli* dihydrofolate reductase complexed with methotrexate [37; PDB code 4dfr] was searched against an apo-form of the same protein [38; PDB code 5dfr]. The PDB deposition of 4dfr contains two independent protein molecules in the asymmetric unit. For this search, the query pattern was created from the 15 residues identified in the PDB data set as site residues for subunit A. This was the smallest pattern tested, consisting of 619 surface points, and the only site-based pattern to be created by identifying the relevant residues from the PDB record, rather than by use of a sphere centred on the active site. Although 4dfr and 5dfr have practically identical sequences, 5dfr contains no ligand bound in the active site, and the site residues exist in a somewhat different conformation to that of the 4dfr structure. The rmse for superposition of the 15 active site residues for the two proteins is 1.0 Å, but several atoms are significantly displaced with respect to their counterparts in the other structure, e.g. the NH_2 atoms of residues Arg^{52} are 4.3 Å apart in the two structures, and the O^{γ} atoms of Ser^{49} are 2.2 Å apart. Some limited surface differences can therefore be expected.

To evaluate the effectiveness of the comparison, the surface superposition transformations were used to superpose the main-chain atoms of 4dfr on 5dfr (the 'best possible' rmse is 0.8 Å). The search was performed very successfully, with 6 of the 10 GA predicted solutions having rmse values below 2 Å and only one of the remaining solutions exceeding 2.2 Å. The superposition of the dihydrofolate reductase α-carbon traces based on the GA predicted transformations can be seen in Fig. 2; the trace for the target protein is shown in green, the fittest solutions are shown in yellow, whilst the poorest solution, with an rmse of 3.9 Å, can be seen in red.

Test 3: Haem binding pocket

The haem binding site surface from sperm whale myoglobin [39; PDB code 4mbn] was searched against the entire surface of one alpha subunit from human deoxyhaemoglobin [40; PDB code 2hhb]. The query pattern surface was generated by selecting all the amino acid residues containing atoms that lay within 10 Å of the iron atom, coordinated by the haem ligand. A total of 32 residues were identified as containing atoms within this distance and these were used to create a Connolly surface of 1474 surface points. The target surface was generated after the removal of the haem ligand and comprised 6759 surface points. In both proteins the haem site is a deep cleft on the surface and this example was chosen to demonstrate the efficacy of the program where the surface features to be identified are quite distinct from all other surface features. Fischer et al. [19] describe a wholly site-based search generated using the same two protein structures, although in their case the search was based on matching the residues of the two sites from the proteins rather than the surfaces. The results from the search for the haem binding site surface were also very good. In this case the coordinates of the haem moiety were used to evaluate the quality of the superposition, with a 'best possible' value of 1.1 Å. Figure 3 shows the superpositions of the haem ligand from haemoglobin (all atom colours) on the haem from myoglobin (green); all the predicted solutions are in the correct region of space and seven have an rmse below 2 Å, while the remaining three predicted solutions have rmse values below 2.6 Å.

Test 4: The active site of sialidase

The sialic acid binding site surface from influenza virus sialidase [41; PDB code 1nnb] was searched against the surface of a bacterial sialidase [42; PDB code 2sim]. The surface pattern for 1nnb was generated by selecting all those residues containing atoms that lay within 10 Å of the sialic acid ligand bound in the active site. A total of 31 residues were identified as containing atoms within the target distance and these were used to generate a surface of 1173 points. The target surface was generated after removal of the ligand and comprised 13 651 surface points. This was a much larger target surface than those in the previous tests and there was a large difference between the sizes of the target and query surfaces. Once again the search was performed very successfully, all 10 of the runs gave a predicted solution with an rmse of less than 2 Å, calculated on the atoms from the sialic acid ligand (best possible value = 0.5 Å.) The superpositions of the 2,3-dehydro-2-deoxy-*N*-acetyl neuraminic acid ligand from influenza virus sialidase (all atom colours) on that from bacterial sialidase (green) can be seen in Fig. 4.

Test 5: NAD binding site in dehydrogenases

The NAD binding site surface from equine alcohol

dehydrogenase [43; PDB code 2ohx] was searched against the surface of the A chain of porcine lactate dehydrogenase [44; PDB code 9ldt]. The surface pattern for 2ohx was generated by selecting all those residues that contained atoms that lay within 4 Å of any atom of the NADH cofactor bound in the active site. A total of 28 residues were identified as containing atoms within the target distance and these were used to generate a Connolly surface of 1070 surface points. In common with the previous test, this example also involved a large target surface, in this case consisting of 14 497 surface points. Fischer et al. [19] describe a search based on alternative depositions of the same structures, but their search was not global over the whole surface of the lactate dehydrogenase and was again based on matching the residues of the two sites rather than on matching surfaces. This was the least successful of all the searches, with the GA correctly predicting the location of the NAD binding site surface on lactate dehydrogenase in only 3 of the 10 runs. The 10 superpositions of the NADH cofactor (all atom colours) on the target NADH (green) can be seen in Fig. 5a. The three 'correct' solutions (rmse values of 1.0–1.6 Å compared with a best possible value of 0.4 Å) are clearly identified in Fig. 5b. However, the remaining seven solutions are very poor, with rmse values as high as 14.7 Å, indicating that the superposition is wrong. However, the fitness scores of these wrong solutions are appreciably lower (158–168) than the good solutions (172, 194 and 194), indicating that the GA had not reached a global minimum in these runs. This highlights the importance of conducting multiple runs (or longer runs) when using stochastic algorithms of this kind.

Test 6: Analysis of lysozyme/antibody complexes

The largest test performed by the GA involved searching the surface of hen egg-white lysozyme [45; PDB code 3lym] against a lysozyme–Fab complex [46; PDB code 1fdl]. The query protein, 3lym, contains 129 residues and these were used to generate a Connolly surface of 4970 points. The target surface was also the largest used and comprised 21 485 surface points. In this test there is a large surface common to the two proteins, i.e., the entire surface of the lysozyme molecule except where it is in contact with the antibody in 1fdl, and we would expect the lysozyme query surface to be able to locate the correct patch on the antibody complex. Ultimately, we hope to use the GA for the comparison of two whole protein surfaces where there is only a small part of the surface in common, e.g., a drug binding site; this example was chosen to give an indication of the CPU times required to perform such searches, in addition to providing some indication of their likely feasibility. Because in this case a whole protein surface was searched against a whole protein surface and the surfaces were therefore very large, the niche sphere diameter setting and the number of niches searched for were changed to allow for the larger search space (figures shown in brackets in Table 1).

The GA performed very well for the search, with 6 of the 10 GA predicted solutions having rmse values below 2 Å (the 'best possible' value is 0.5 Å). In fact, only two of the runs had rmse values greater than 2.5 Å and the best superpositions of the lysozyme (yellow traces) on the lysozyme from the antibody complex (shown in green) can be seen in Fig. 6. The poorest solution, shown in magenta, can clearly be seen some way from the correct location, whilst the only other solution with an rmse exceeding 2.5 Å, shown in red, is actually centred in the correct region of the antibody complex surface but has not been rotated to match the surfaces correctly, presumably because there were insufficient generations in the GA run in this particular case. Here, as before, the best GA fitness scores (1350 grid points equivalenced) correspond to the best rmse (0.8 Å) when checked against the atomic coordinates and the worst score (338) corresponds to the worst rmse (25.5 Å). The correct solution could have been identified without prior knowledge, but this once again highlights the importance of conducting multiple runs of these non-deterministic algorithms.

Test 7: Comparison of elastase inhibitors

This search involved the comparison of the surfaces of two elastase inhibitors, difluoroketone inhibitor (DFKi) and turkey ovomucoid inhibitor (TOMI), a 56-residue peptide. For the purposes of this comparison the surface points associated with the fluorine atoms in DFKi were labelled as hydrogen-bond accepting. The 3D coordinates for these inhibitors were extracted from PDB files 4est [47], porcine pancreatic elastase, and 1ppf [48], human neutrophil elastase, respectively. This comparison was performed by Masek et al. [6] when determining the efficacy of their molecular skin representation for shape comparison of small molecules, and more recently by Perkins et al. [7] when performing a further small molecule comparison study. Masek et al. [6] identified 17 key residues (41, 57, 102, 189–195, 213–216 and 226–228) in the core of the active site region of the two elastases that could be superposed, based on main-chain atoms, with an rmse of 0.34 Å, and we have used rmse values based on the superposition of these atoms as a guide to the quality of the overlays produced between the two inhibitor surfaces. Figure 7a shows the 10 superpositions of the DFKi elastase inhibitor ('correct' solutions in all atom colours and poorer solutions in red) on the TOMI elastase inhibitor (green). The expected position of DFKi based on superposition of the elastase coordinates is shown in yellow. In 6 of the 10 runs the GA predicted solutions with an rmse of less than 2 Å and in a seventh run a solution with an rmse of 2.1 Å (note: these values are for the superposition of the 'core' main-chain atoms from the inhibited elastase structures, not for the superposition of

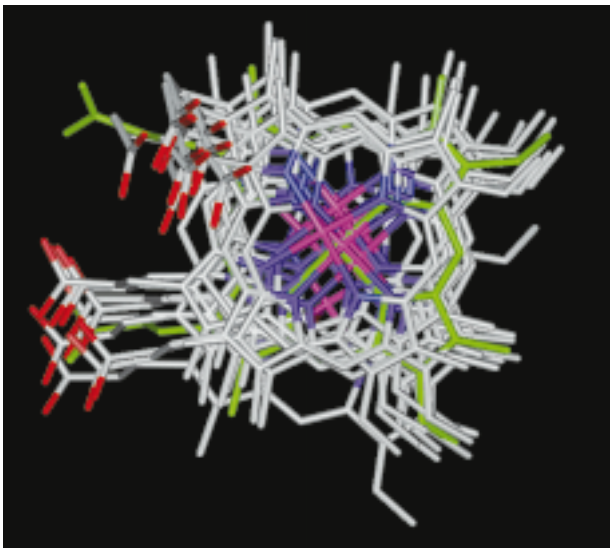


Fig. 3. Ten overlays of the haem from myoglobin (all atom colours) on the haem from haemoglobin (green), based on the surface transformations output by the GA.

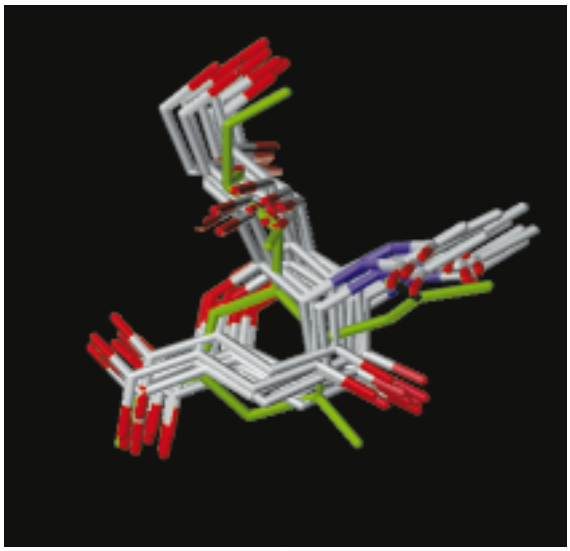


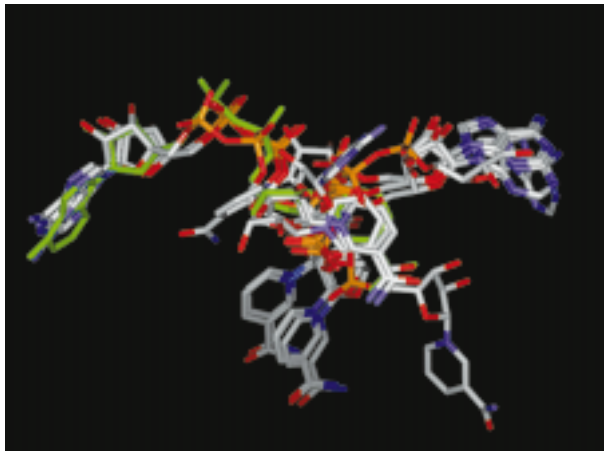
Fig. 4. Ten overlays of the neuraminic acid inhibitor from influenza virus sialidase (all atom colours) on that from bacterial sialidase (green), based on the transformations output by the GA for superposing the active site surface from the viral sialidase on the entire surface of the bacterial sialidase.

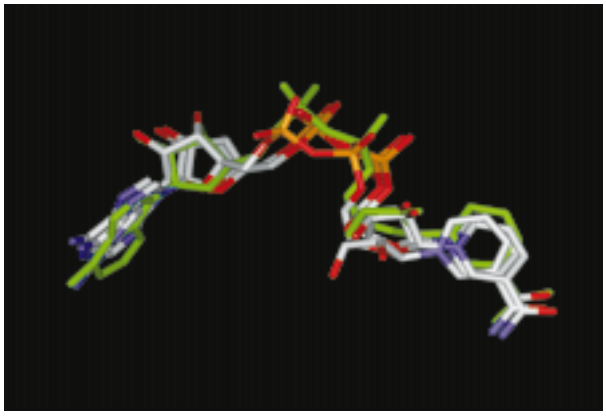
ligand atoms). Therefore these solutions were essentially correct. The 10 solutions can be seen more clearly in Fig. 7b, where the TOMI inhibitor has been removed for clarity. The 'correct' solutions can be clearly seen in atom colours and have fitnesses in the range 112–135 grid points equivalenced. The remaining three predicted solutions are shown in red and are poor, with fitnesses in the range 102–106, and are all clustered in the same region of space.

Performance of fitness function

The best results from each of the 10 runs based on the fitness score for the number of surface points in common are shown in Table 2. Although the program has been operating on the surface representations of the proteins, and not the atomic representations, one would nevertheless hope that the fittest solution predicted by the GA would also have the best rmse value for the ligand/main-chain atom superposition. In order to assess this, columns 5 and 6 in the table show the best overall rmse value for the 10 runs along with the fitness score that was associated with it. For five of the seven examples tested there is a good agreement between the fittest solution predicted by the GA and the best rmse for the overlay of the ligand/main-chain atoms, with the fittest solution predicted in each case providing transformations that produce the

a





b

Fig. 5. (a) Ten overlays of the NADH cofactor from alcohol dehydrogenase (2ohx; all atom colours) on the NADH cofactor from lactate dehydrogenase (9ldt; green), based on the transformations output by the GA for superposing the NADH binding site surface from 2ohx on the entire protein surface of 9ldt. (b) Close-up of the three overlays with the highest fitness.

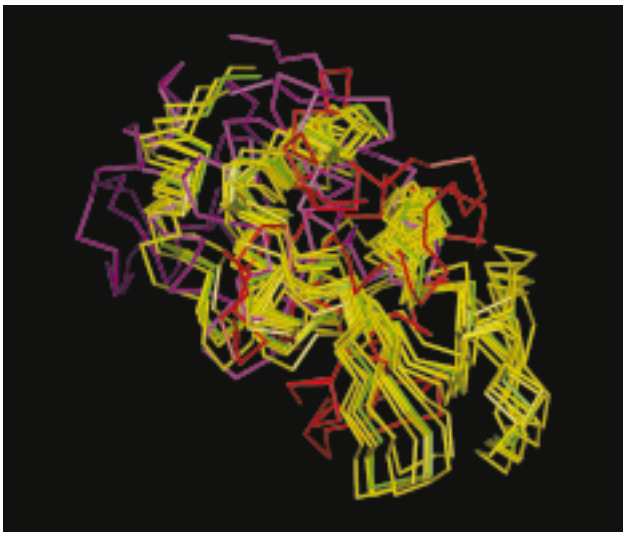


Fig. 6. Ten overlays of the α-carbon trace of lysozyme (3lym; yellow, red and magenta traces) on the α-carbon trace of lysozyme from the antibody–lysozyme complex 1fdl (green trace), based on the transformations output by the GA for superposing the surface of 3lym on the surface of 1fdl.

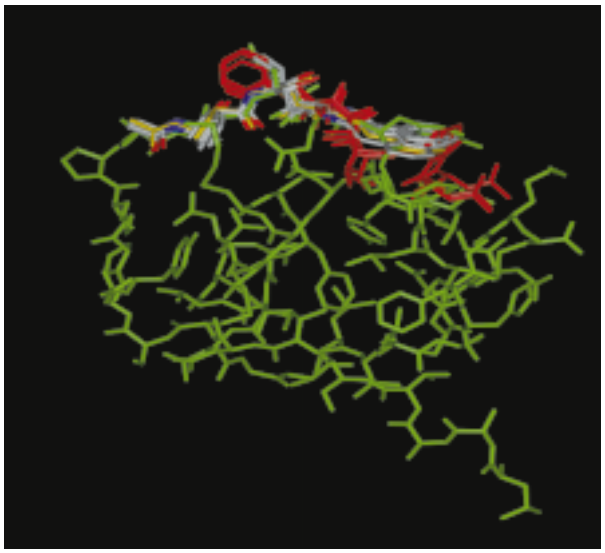
best rmse for the superposition of the ligand/main-chain atoms.

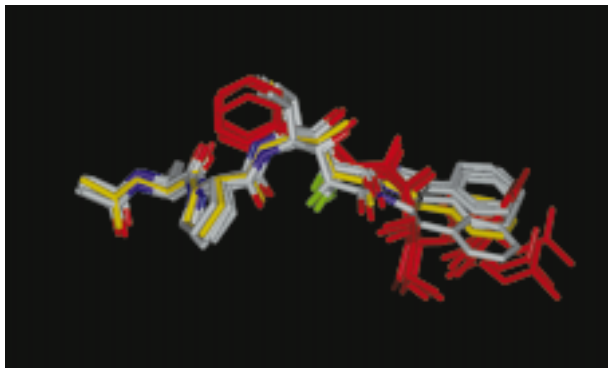
This is not the case for two of the searches, namely the haem binding pocket (test 3) and the elastase inhibitor example (test 7). The best rmse for superposition of the two haem groups was 1.5 Å, with an associated fitness score of 438, whilst the best fitness score achieved was 440 surface points in common with an associated rmse of 2.3 Å. Nevertheless, the latter solution is an essentially correct solution, and the small difference in fitness reflects the broad maxima associated with comparing fairly coarse surface representations. A similar result was also seen in the comparison of the two elastase inhibitors where the fittest solution predicted by the program (135 surface points in common) had a slightly higher rmse than the best overall rmse. In practice, however, these variations are of little consequence since the GA-predicted solutions are essentially correct in both cases.

Program timings

Details of representative timings (in CPU minutes on a DEC Alphastation 3000) for each of the searches performed are given in Table 2. Although at this stage optimising the program in terms of speed of operation has not been a priority, the program is relatively fast in operation, with all but one of the examples completing in under 25 CPU minutes per run. In addition, approximately 1 min was required to generate the initial Connolly surfaces in each example. Naturally there is an increase in the time taken to perform the searches as the number of surface points in the patterns increases. The fastest search, involving the surfaces of the two elastase inhibitors, took just 4 min, whilst the slowest search was the comparison of the lysozyme surface against the lysozyme antibody complex surface, which took 122 min. This latter search involved the largest patterns for both the query and the target and additionally was run with twice the number of niches, thus adding significantly to the overall run time. It can be seen from the search results that in order to be sure of identifying the best superposition between two surfaces it may be necessary to run the program more than once and this will obviously increase the time taken to provide a solution.

a





b

Fig. 7. (a) Ten overlays of the elastase inhibitor DFKi (best solutions in atom colours and worst solutions in red) on the inhibitor TOMI (green), based on the transformations output by the GA for superposing the surface of DFKi on the surface of TOMI. (b) Close-up of the overlays of the inhibitor DFKi shown in (a), but with TOMI coordinates removed for clarity; the expected position of DFKi based on superposition of the elastase coordinates is shown in yellow.

Discussion and Conclusions

The GA has performed very well in all of the examples discussed in this paper. In every case a solution was found that was able to superpose the atoms being used as guides to the quality of the match with very low rmse values. At present we have shown that the method will allow the location of a small prespecified surface patch on a whole protein surface (tests 1–5 and 7), and the location of a large surface region in a still larger one (test 6). One area of prior concern was the possibility that amino acid substitutions or side-chain flexibility between similar binding sites in different proteins would adversely affect the performance of the method. However, our tests described above have shown that the method works well – perhaps because the Connolly surfaces are quite coarsely sampled. This is borne out by the fact that in several of the examples chosen, the acid proteases (example 2), the sialidases (example 4) and the alcohol and lactate dehydrogenases (example 5), there are very appreciable differences in the position and type of binding site residues but the common binding site surfaces are correctly identified.

Ultimately we hope to make the program more flexible by enabling the user to define the kind of match that is required. Thus, in addition to the searches described above, it may under some circumstances be necessary to search one whole protein surface against another to locate only a small area of surface similarity between them. This would enable the location of binding surfaces of proteins known or believed to bind to the same receptor or ligand but for which no detailed structural data on the location of the binding site is available. Extension of the program to allow this will require extensive modifications, which we are currently exploring. The use of a variety of weighting and clustering measures is being investigated in order to make this possible. As an example, surface points that match and are located in invaginations could be scored with a higher weight than otherwise identical matching points that are not in invaginations. This would have the effect of directing the search to those matches that involved such regions and might be of benefit when looking for binding sites, which tend to be in cavity, or pocket, regions [49,50]. However, one of the advantages of the current representation is its simplicity and the ease with which the surface points can be labelled with additional information which can then be optionally included in the matching process. Thus, whilst the surface points are currently 'coloured' with a limited number of additional parameters, it may be possible to improve the specificity of the program by including more parameters to distinguish one point from another, for example, lipophilic and hydrophilic regions [51]. A further challenging application of the GA would be to investigate the protein–protein docking problem by modifying the program to look for surface complementarity rather than similarity. This more complex problem is a topic of very widespread interest [52]. Thus, the present program is not only valuable in its own right, but also forms a starting point for further investigations in this important area of research.

Acknowledgements

We thank the Medical Research Council for funding this research, and Wellcome Trust, BBSRC and Tripos Inc. for computing resources. The Krebs Institute is a designated BBSRC Biomolecular Sciences Centre.

References

1 Pepperrell, C.A. and Willett, P., J. Comput.-Aided Mol. Design, 5 (1991) 455.
2 Hodgkin, E.E. and Richards, W.G., Int. J. Quant. Chem., Quant. Biol. Symp., 14 (1987) 105.
3 Martin, Y.C., Bures, M.G., Danaher, E.A., DeLazzer, J., Lico, I. and Pavlik, P.A., J. Comput.-Aided Mol. Design, 7 (1993) 83.
4 Jones, G., Willett, P. and Glen, R.C., J. Comput.-Aided Mol. Design, 9 (1995) 532.
5 Chau, P.L. and Dean, P.M., J. Mol. Graph., 5 (1987) 97.
6 Masek, B.B., Merchant, A. and Matthews, J.B., Proteins, 17 (1993) 193.
7 Perkins, T.D.J., Mills, J.E.J. and Dean, P.M., J. Comput.-Aided Mol. Design, 9 (1995) 479.
8 Mitchell, E.M., Artymiuk, P.J., Rice, D.W. and Willett, P., J. Mol. Biol., 212 (1989) 151.
9 Vriend, G. and Sander, S., Proteins, 11 (1991) 52.
10 Grindley, H.M., Artymiuk, P.J., Rice, D.R. and Willett, P., J. Mol. Biol., 229 (1993) 707.
11 Leicester, S., Finney, J. and Bywater, R., J. Math. Chem., 16 (1994) 315.
12 Leicester, S., Finney, J. and Bywater, R., J. Math. Chem., 16 (1994) 343.
13 Lewis, M. and Rees, D.C., Science, 230 (1985) 1163.
14 Zachmann, C.-D. and Brickmann, J., J. Chem. Inf. Comput. Sci., 32 (1992) 120.
15 Colloc'h, N. and Mornon, J.-P., J. Mol. Graph., 8 (1990) 133.
16 Lin, S.L., Nussinov, R., Fischer, D. and Wolfson, H.J., Proteins, 18 (1994) 94.
17 Zachmann, C.-D., Heiden, W., Schlenkrich, M. and Brickmann, J., J. Comput. Chem., 13 (1992) 76.
18 Connolly, M.L., J. Mol. Graph., 4 (1986) 3.
19 Fischer, D., Norel, R., Wolfson, H. and Nussinov, R., Proteins, 16 (1993) 278.
20 Holland, J.H., Adaptation in Natural and Artificial Systems, The University of Michigan Press, Ann Arbor, MI, U.S.A., 1975.
21 Dandekar, T. and Argos, P., Protein Eng., 5 (1992) 637.
22 Unger, R. and Moult, J., J. Mol. Biol., 231 (1993) 75.
23 Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245 (1995) 43.
24 May, A.C.W. and Johnson, M.S., Protein Eng., 7 (1994) 475.
25 May, A.C.W. and Johnson, M.S., Protein Eng., 8 (1995) 873.
26 Connolly, M.L., Science, 221 (1983) 709.
27 Connolly, M.L., J. Appl. Crystallogr., 16 (1983) 548.
28 Richards, F.M., Annu. Rev. Biophys. Bioeng., 6 (1977) 151.
29 Jiang, F. and Kim, S.H., J. Mol. Biol., 219 (1991) 79.
30 Goldberg, D.E., Genetic Algorithms in Search, Optimisation and Machine Learning, Addison Wesley, Reading, PA, U.S.A., 1989.

31 Davis, L., Handbook of Genetic Algorithms, Van Nostrand Reinhold, New York, NY, U.S.A., 1991.
32 Delaney, J.S., J. Mol. Graph., 10 (1992) 174.
33 Beasley, D., Bull, D.R. and Martin, R.R., Evol. Comput., 1 (1993) 101.
34 Bernstein, F.C., Koetzle, T.F., Williams, G.J.D., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanuchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
35 Dreyer, G.B., Lambert, D.M., Meek, T.D., Carr, T.J., Tomaszek Jr., T.A., Fernandez, A.V., Bartus, H., Cacciavillani, E., Hassell, A.M., Minnich, M., Petteway Jr., S.R. and Metcalf, B.W., Biochemistry, 31 (1992) 6646.
36 Spinelli, S., Liu, Q.Z., Alzari, P.M., Hirel, P.H. and Poljak, R.J., Biochimie, 73 (1991) 1391.
37 Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C. and Kraut, J., J. Biol. Chem., 257 (1982) 13650.
38 Bystroff, C. and Kraut, J., Biochemistry, 29 (1990) 3263.
39 Takano, T., J. Mol. Biol., 110 (1977) 569.
40 Fermi, G., Perutz, M.F., Shaanan, B. and Fourme, R., J. Mol. Biol., 175 (1984) 159.
41 Bossart-Whitaker, P., Carson, M., Babu, Y.S., Smith, C.D., Laver, W.G. and Air, G.M., J. Mol. Biol., 232 (1993) 1069.
42 Crennell, S.J., Garman, E.F., Laver, W.G., Vimr, V. and Taylor, G.L., J. Mol. Biol., 259 (1996) 264.
43 Al-Karadaghi, S., Cedergen-Zeppezauer, E.S., Petratos, K., Hovomoeller, S., Terry, H., Dauter, Z. and Wilson, K.S., Acta Crystallogr., D50 (1994) 793.
44 Dunn, C.R., Wilks, H.M., Halsall, D.J., Atkinson, T., Clarke, A.R., Muirhead, M. and Holbrook, J.J., Phil. Trans. R. Soc. London, 332 (1991) 177.
45 Kundrot, C.E. and Richards, F.M., J. Mol. Biol., 193 (1987) 157.
46 Fischmann, T.O., Bentley, G.A., Bhat, T.N., Boulot, G., Mariuzza, R.A., Phillips, S.E.V., Tello, D. and Poljak, R.J., J. Biol. Chem., 266 (1991) 12915.
47 Takahashi, L.H., Radhakrishnan, R., Rosenfield Jr., R.E., Meyer Jr., E.F. and Trainor, D.A., J. Am. Chem. Soc., 111 (1989) 3368.
48 Bode, W., Wei, A.Z., Huber, R., Meyer, E., Travis, J. and Neumann, S., EMBO J., 5 (1986) 2453.
49 Peters, K.P., Fauck, J. and Froemmel, C., J. Mol. Biol., 256 (1996) 201.
50 Laskowski, R.A., Luscombe, N.M., Swindells, M.B. and Thornton, J.M., Protein Sci., 5 (1996) 2438.
51 Heiden, W., Moeckel, G. and Brickmann, J., J. Comput.-Aided Mol. Design, 7 (1993) 503.
52 Strynadka, N.C.J., Eisenstein, M., Katchalski-Katzir, E., Shoichet, B.K., Kuntz, I.D., Abagyan, R., Totrov, M., Janin, J., Cherfils, J., Zimmerman, F., Olson, A., Duncan, B., Rao, M., Jackson, R., Sternberg, M. and James, M.N.G., Nat. Struct. Biol., 3 (1996) 233.