Protein Docking Using a Genetic Algorithm

Eleanor J. Gardiner,[1]* Peter Willett,[2] and Peter J. Artymiuk[3]
[1]*Department of Information Studies and Department of Molecular Biology and Biotechnology, Krebs Institute, Sheffield University, Sheffield, United Kingdom*
[2]*Department of Information Studies, Krebs Institute, Sheffield University, Sheffield, United Kingdom*
[3]*Department of Molecular Biology and Biotechnology, Krebs Institute, Sheffield University, Sheffield, United Kingdom*

ABSTRACT **A genetic algorithm (GA) for protein–protein docking is described, in which the proteins are represented by dot surfaces calculated using the Connolly program. The GA is used to move the surface of one protein relative to the other to locate the area of greatest surface complementarity between the two. Surface dots are deemed complementary if their normals are opposed, their Connolly shape type is complementary, and their hydrogen bonding or hydrophobic potential is fulfilled. Overlap of the protein interiors is penalized. The GA is tested on 34 large protein–protein complexes where one or both proteins has been crystallized separately. Parameters are established for which 30 of the complexes have at least one near-native solution ranked in the top 100. We have also successfully reassembled a 1,400-residue heptamer based on the top-ranking GA solution obtained when docking two bound subunits. Proteins 2001;44:44–56.**

Key words: shape complementarity; macromolecular docking; protein–protein interaction; molecular recognition; protein–protein docking; genetic algorithm

INTRODUCTION

Protein–protein recognition represents a fundamental aspect of biological function. However, although protein structures are now routinely determined by methods such as X-ray crystallography, it is much more difficult to ascertain the structure of protein complexes, and only about 100 of these are currently deposited in the Protein Data Bank (PDB).[1] Protein docking studies are therefore important as an aid to our understanding of the ways in which proteins associate.[2] Methods for protein ligand docking, such as DOCK,[3] FLOG,[4] FLEXX,[5] and GOLD,[6] are widely used in drug-discovery programs. However, because of the complexity of the problem, protein–protein docking is still largely at the theoretical stage. Many solutions have been proposed but, although very promising results have been obtained in some cases,[7] there is still considerable scope for the development of methodology and exploration of alternative or ancillary approaches.

Protein docking methods are generally based on the idea of complementarity between the interacting molecules. This complementarity may be geometric or electrostatic, or both. Recently there have been two investigations into the validity of the complementary shape hypothesis. Betts and Sternberg[8]investigated the conformational changes that occurred on complex formation for a total of 30 complexes. These investigators concluded that, for the enzyme–inhibitor complexes that comprised most of their data, recognition could be "treated as lock and key to a first approximation." In a study of specific interprotein interactions (e.g., nonpolar buried surface area, polar buried surface area, number of hydrogen bonds), Norel et al.[9] found no correlation between any of these and the likelihood that a putative docking was correct. Instead, their finding confirmed the "long held notion of the importance of molecular shape complementarity in the binding and hence in docking." Both proteins are usually treated as rigid bodies, since to allow flexibility may be computationally intractable. Many shape-based docking algorithms have been proposed.[10–15] Some include other information (e.g., hydrophobicity, electrostatics) either in conjunction with shape matching[16] or as a subsequent filter.[17] Several geometric docking programs use the Connolly molecular surface representation.[18] Three are of particular interest, as they have some similarity to the method we describe below.

Connolly[19] described a method for selecting "critical" points of the Connolly surface. The idea of the critical points is to preserve the important features (local maxima and minima) of the protein surface, while reducing the computational intractability associated with processing very large numbers of surface points. These critical points (termed "knobs" and "holes") represent points that would be expected to match (knobs on one protein with holes on another and vice versa) in a successful docking. Connolly required four pairs of critical points to match in order to define a rigid body translation and this proved too restrictive in one of the two complexes he attempted to dock. However, Norel et al.[20] showed that pairs of two critical points plus the normals to the surface at these points are sufficient to dock the example at which Connolly's algorithm failed. In further work, the same group docked 26

Grant sponsor: Engineering and Physical Sciences Research Council; Grant sponsor: Biotechnology and Biological Sciences Research Council.

*Correspondence to: Eleanor Gardiner, Department of Information Studies, Sheffield University, Western Bank, Sheffield S10 2TN, United Kingdom. E-mail: e.gardiner@sheffield.ac.uk

Received 10 November 2000; Accepted 23 February 2001

"large" protein–protein complexes from the PDB using the whole of the surface of both proteins and without using any chemical information such as hydrogen bonding potential.[21] The dockings usually took only a few minutes of CPU time and generated a manageable number of solutions with the correct solution ranked high (in 22 cases, ranked 1). In further work, they applied the same algorithm to 19 complexes using data taken from the native protein structures.[9] Recently Hou et al.[22] used a genetic algorithm in combination with a tabu search to match dots of the Connolly surface in the initial global search stage of their docking. They also required the surface normals to match; the fitness score was then calculated according to the amount of buried surface area, weighted by a penalty for steric clash, proportional to the number of overlapping atoms. This is similar in form to our fitness function (detailed under Materials and Methods), except that we take account of matching dots rather than buried surface area and calculate overlap in a different manner. They followed this by performing local energy calculations for the top ranked solutions. Most of the computation in the global search is done by the tabu search element of their hybrid algorithm. The genetic element is used to diversify the search. The local search is then also performed by a GA; it attempts to minimize the interaction energy (van der Waals energy, electrostatic energy, and hydrogen-bond energy) between the participating molecules. They report good results on a test set of seven complexes (five bound and two native).

Most early attempts at protein–protein docking involved only reassembling complexes using coordinates taken from the complex. More recently there have been several studies using native proteins.[9,16,23–25] There have also been two protein–protein docking challenges.[7,26].

Most docking studies concentrate on enzyme–inhibitor complexes (most commonly serine protease–inhibitor complexes) and antibody–antigen complexes. A recent comparison of the protein–protein interactions in these two different types of complex concluded that, in general, a protease–inhibitor interface was more static and hence more easily predicted than an antibody–antigen interface.[27] This result is generally borne out by docking studies which have used unbound proteins.[23]

In our previous work on docking, we used techniques from graph theory to predict maximum sets of possible hydrogen bonds and used these to superpose the proteins to form a complex.[28] While this method was able to reassemble complexes, it proved less successful when docking native proteins, and here we have considered the use of a genetic algorithm (GA) for this purpose.

GAs have been shown to be of use in many areas of bioinformatics. A GA is a search algorithm based on accepted theories of biological evolution and natural selection. It operates on a population of solutions, applying the principle of "survival of the fittest." After a number of generations, a good, although not necessarily optimal, solution is obtained. Applications of GAs to structural biology include investigation of protein folding,[29] ligand binding to receptor sites,[30] protein structure and sequence comparison,[31,32] protein secondary structure prediction,[30] protein docking,[33] nuclear magnetic resonance (NMR), and X-ray crystallographic data analysis.[34] At Sheffield, we have used GAs successfully for many different applications, including database searching,[35] protein folding,[36] protein ligand docking and database screening,[37] identifying features common to sets of ligands,[38] selection of compounds for combinatorial libraries,[39] and protein surface comparison,[40] inter alia. In this article, we present a genetic algorithm for protein docking based on the surface comparison method of Poirrette et al.,[40] but which we have modified in order to capture the regions of greatest complementarity between the two proteins that are being docked. We demonstrate that, for a large number and variety of complexes, using mainly native data, our GA is able to produce highly ranked putative complexes that resemble the crystallographically determined complex.

MATERIALS AND METHODS

The basic method is based on that of Poirrette et al.,[40] who used a GA for surface comparison. A GA was used to generate rotations of the smaller (query) protein relative to the larger (target) protein surface, which was held static. Both proteins were treated as rigid bodies. We provide brief details, paying particular attention to the modifications made for protein docking.

We first generated the proteins' Connolly dot surfaces using the program MS,[18,41] with a probe radius of 1.5 Å and a dot density of 1 dot/Å^2. In addition to three-dimensional (3D) coordinates for each dot, this program calculates the direction of the normal to the surface and also the type of surface (1 = convex, 2 = saddle, 3 = concave). We modified the program slightly so that each dot was also labeled with the hydrogen bonding potential of the nearest atom. The classification we adopted was as follows: -H-donor, H-acceptor, H-donor/acceptor, and non-H-bonding. All oxygen atoms were labeled as H-acceptor with the exception of the OH of tyrosine, the OG of serine and the OG1 of threonine, all of which were labeled as H-donor/acceptors. For simplicity, all nitrogen atoms were labeled as H-donors, including the histidine nitrogens ND1 and NE2 whose individual protonation states are difficult to assign. All other atoms were classified as non-H-bonding.[40]

Both proteins were translated so that their centers of gravity were positioned at the origin of space. This origin was also the center of a 3D grid placed around the target protein; it was sufficiently large that any query dot could contact any target dot, and the entire query molecule still remain within the grid. Each grid box was a 2-Å cube. Dots contained within the same grid box were given the same grid coordinate, which was that of the corner of the grid box nearest to the origin. This "coarsening" of the surface representation allowed for some tolerance in the matching of the protein surfaces, which was necessary to accommodate movements of the proteins from the native to the bound conformation, and also it enabled the implementation of a very fast routine for dot matching, thus increasing both the effectiveness and the efficiency of the program.

Each chromosome, corresponding to a potential docking, consisted of six integer elements, corresponding to the six degrees of freedom necessary to define the movement of one rigid body relative to another. The first three elements corresponded to rotations of 0–359° about the x, y, and z axes, respectively, and the remaining three to translations (measured in tenths of an Ångstrom) along the original x, y, and z axes. An initial population of chromosomes was generated at random, each was applied to the query protein dots (the target remained fixed), and the fitness of the resulting complex was calculated (as described below).

The GA was of a type known as steady-state-with-no-duplicates, meaning that a fixed proportion (in our case 5%) of the chromosome population was replaced after each generation and that duplicate chromosomes were not allowed.[42,43] The chromosomes replaced were those that were least fit in the population. New chromosomes were generated by applying the genetic operators mutation and/or crossover to parents chosen by roulette wheel selection. Mutation took one parent chromosome and produced one child. In standard mutation, a position was chosen at random along the parent and randomly mutated to another value (keeping within the range 0–359 if a rotation was being mutated). Small-creep mutation operated in the same way except that the replacement value was close to the original (within 5° for a rotation and within 2 Å for a translation). When mutation was the chosen genetic operator, standard or small-creep mutations were chosen at random (with equal likelihood). Crossover took two parents and produced two children, in one of two ways. In single-point crossover a position along the chromosome was chosen at random and all elements 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 element was swapped between the two chromosomes. When crossover was the chosen genetic operator, single-point or exchange-crossover were again chosen at random (with equal likelihood). After a predetermined number of generations (1,000 for all the tests reported here) the GA was presumed to have converged and the fittest population member was saved as a solution.

We have also applied a technique called niche restriction,[44] which we use to force the GA to explore different regions of solution space. After the first run through, an area (niche) around the fittest solution was removed from the solution space. The area removed was a sphere of radius 2 Å centered on the center of gravity of the fittest solution. The GA was then restarted as before, except that no chromosomes were allowed to place the center of gravity of the query protein within the restricted niche. After the second runthrough, a second niche was defined and so on. In our case, we obtained 5 niches per GA run, each niche corresponding to a potential solution.

Fitness Function

The GA thus described is similar to the surface matching algorithm of Poirrette et al.[40] The critical difference in the new method lies in the fitness function since for protein docking we are looking for complementary rather than similar surfaces, with minimal overlap of the protein interiors. However, during initial testing we found that too stringent a complementarity requirement was unsuccessful in that "fit" solutions did not correspond to correct orientations of the query protein. Thus we have adopted a principle of "not similar" matching where necessary. So, for a pair of target and query dots, a match was declared if (1) their grid coordinates were the same (thus, they occupy the same region of space); (2) their surface normals were nearly opposite (the angle between them should be close to 180°—the actual value was a parameter); (3) their Connolly shape type was not the same, unless they were type 2. (i.e., 1 matched 2 or 3, 3 matched 1, or 2 and 2 matched anything); or (4) their hydrogen bonding potential was satisfied, i.e., H-donor matched H-acceptor or H-donor/acceptor, H-acceptor matched H-donor or H-donor/acceptor, H-donor/acceptor matched H-donor, H-acceptor or H-donor/acceptor, non-H-bonding matched non-H-bonding.

We also wished to apply a penalty for overlap of the protein interiors whilst allowing for slight surface clashes. We did this by defining a set of "interior" points of the target protein. These were grid points that were near to a target protein atom and that were further than a user defined "thickness" parameter from any surface dot. In all the tests described in this report, a thickness of 2 Å was used. The coordinates of the interior points were hashed to integer values and stored in a sorted list. As their creation took 5–30 min on a Silicon Graphics 270-MHz O2 workstation with R12000 processor (depending on the size of the target), a set of interior points for a particular thickness was only created once and was stored on file for subsequent GA runs. For a possible orientation of the query with respect to the target, we counted the number of matching dots and the number of clashes between query dots and target interior points. The penalty was then given by

$$\text{Penalty} = \begin{matrix} J^{*}\ \text{number of clashes} & \text{if any dots matched} \\ 100{,}000 & \text{if no dots matched} \end{matrix}$$

where J, the penalty multiplier, was an input parameter.

The fitness of the chromosome was then given by

$$\text{Fitness} = \text{number of matches} - \text{penalty}$$

We recognize that this is not an even-handed approach. For example, a long side-chain may protrude from the target. A single side-chain is "thin" (i.e., by our definition, it has no interior points) and thus, if it penetrates the query, this will not be penalized. However, for the query protein, such a side-chain's penetration of the target would attract a penalty, as it would clash with the target interior points. We have taken this pragmatic approach to minimize computation (and hence time) required for docking. Calculating the new positions of the query points is one of the most time-consuming parts of the GA. If interior points were to be defined for the query, their position would also have to be recalculated for each alteration of each chromosome, which would be a further time burden. Some results of this "lopsidedness" is discussed later.

The GA may be used with entire protein surfaces or one or both proteins may be just a site. If a site was used, surface dots were generated for only that part of the protein surface. However, if that protein was the larger of

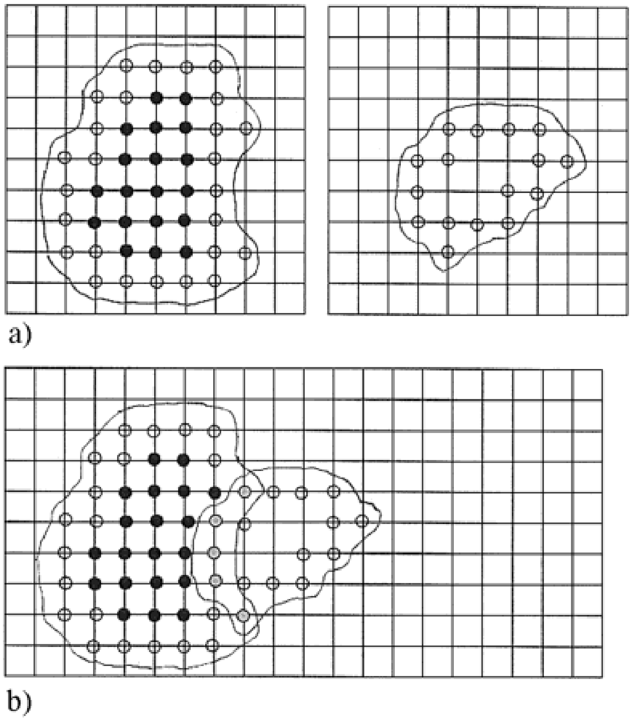

Fig. 1. **a:** Target and query proteins have their surfaces represented by grid point dots (○). The target also has interior points (●). **b:** Surface grid dots are matched (gray filled) to produce a putative complex. (N.B.: The proteins are positioned too closely together.)

the two, it could still be considered as the target. In that case, interior points were generated for the entire protein, not for the site alone. Figure 1 illustrates surface representation and matching.

During preliminary tests, we observed that, irrespective of whether a predictive complex resembled a crystallized complex, the GA tended to position the two proteins infeasibly closely. This is due in part to the protein representation we use. As surface points are mapped to grid points that are slightly nearer the origin, when grid points from two proteins are superposed, this causes the proteins to be positioned too closely together (Fig. 1). We therefore developed a separation program that we applied to each fit solution proposed by the GA. This program used a least squares superposition to find the best alignment of the pairs of normals which belong to a set of matching dots. We then moved the query protein 2 Å away from the target in the direction of this alignment.

Based on extensive testing we have determined that the parameters settings given in Table I caused the GA to converge and, as demonstrated below, appeared to be appropriate for this application domain.

RESULTS AND DISCUSSION

Docking Bound Complexes

Our aim in docking is correctly to predict the structure of the complex formed when two native proteins bind. First, however, we tested the method and determined which new parameter values were successful, using the well-known (and often docked) complexes chymotrypsin–inhibitor complex (PDB code 1cho), subtilisin–inhibitor complex (PDB code 2sni), and the antibody Fab–antigen complex HyHel-5–lysozyme (PDB code 3hfl). We separated each complex into two PDB files and randomly reoriented the smaller protein and moved it away from the larger one. For the antibody–antigen docking, we used only those residues of the Fab within 8 Å of the known antibody-binding site. The parameter settings we wished to establish were the angle between the surface normals and the penalty multiplier. We varied the normal angle between 180° ± 10° and 180° ± 30° in 5° steps, and the penalty multiplier J between 0.5 and 3.5 in steps of 0.5. For each parameter combination we obtained 100 possible solution chromosomes by running the GA 20 times, each run yielding 5 niches. These were then ranked in decreasing order of fitness.

TABLE I. Parameter Settings Used by the GA

Parameter	Value
Population size	400
No. of runs	1,000
Percentage of population replaced each iteration (%)	5
Mutation/crossover rate (%)	50
Selection pressure	1.1
Niche sphere diameter (Å)	2
No. of niches	5

The best results for these three complexes are summarized in Table II. The root-mean-square deviations (RMSD) quoted for these bound complexes were calculated by superposing the enzymes of the predicted and crystallized complexes and then comparing the positions of the *C*α atoms of the respective inhibitors (after the 2-Å separation described under Materials and Methods). Encouragingly, for several different combinations of parameter settings, the fittest (top ranked) solution was within 3 Å of the crystallographic solution and, in all cases but two, there were several correct hits ranked in the top 10. The parameter combinations shown in columns 2 and 3 were the most consistent in yielding good hits; these were the values chosen for subsequent docking of native proteins. Of these four, the parameter combination of normal angle ±15° and $J = 1.0$ does seem to give slightly better results than the others, with the top-ranked hit resembling the crystallized complex in all three cases. However, when docking using native coordinates, we expected that looser constraints might be necessary in order for surface dots to match and therefore we decided to retain all four parameter combinations to increase the chance of obtaining correct dockings. The results for the antibody–antigen complex were not as good as those of the enzyme–inhibitor complexes. This is in agreement with the prediction of Jackson.[27]

Docking Native Complexes

Each of the above are separated crystallized complexes. We then turned our attention to predicting the structure of complexes when starting from native component proteins. A recent study of protein–protein complexes gave a list of 31 complexes (18 enzyme–inhibitor, 7 antibody–antigen and 6 “other” complexes) whose structures have been

TABLE II. Docking Bound Complexes*

Complex (PDB Code)	Normal angle[a]	J[b]	Top rank[c]	Top 10 hits[d]	Total hits[e]	RMSD[f]	Best RMSD[g]
1cho	±30°	3.0	6	2	4	1.47	1.47
	±20°	1.5	4	1	5	0.77	0.77
	±20°	3.0	2	4	5	1.05	0.93
	±15°	1.0	1	7	9	1.69	0.70
2sni	±30°	3.0	5	1	1	1.63	1.63
	±20°	1.5	1	5	8	1.60	1.32
	±20°	3.0	2	3	6	2.99	0.76
	±15°	1.0	1	7	10	1.40	1.36
3hfl	±30°	3.0	0				
	±20°	1.5	25	0	2	1.63	1.63
	±20°	3.0	34	0	4	0.94	0.94
	±15°	1.0	1	3	5	2.00	2.00

*Each row corresponds to 20 runs of the GA, each producing 5 niches giving 100 solutions per row.

[a] 180° normal angle is the angle allowed between the surface normals for a permitted match.

[b] Penalty multiplier.

[c] Rank (by fitness score, of 100) of the highest ranked solution which resembled the complex.

[d] Number of complex-resembling hits in the 10 top ranked solutions.

[e] Number of complex-resembling hits (of 100).

[f] Root-mean-square deviation of the inhibitor $C\alpha$ atoms of the most highly ranked predicted complex when its enzyme is superposed on the Protein Data Bank (PDB) complex enzyme.

[g] Predicted complex closest to the actual complex. Predicted complexes were deemed to resemble the complex if their RMSD was <3 Å.

solved by X-ray crystallography to a resolution of ≥ 2.8 Å, and each of which has at least one of its component proteins crystallized separately to a similarly high resolution.[8] In order to test the GA on a wide variety of complexes, we have attempted to dock each of these complexes using whatever native data was available. We also included the well-known barnase–barstar complex (PDB code 1brs), another antibody–antigen complex (PDB code 3hfm) and the β-lactamase–inhibitor complex, which was the subject of a recent docking challenge.[25] To our knowledge, this is the largest docking study of native proteins to date. Table IIIA–C contains details of these 34 test complexes.

For our first attempt at these native dockings, we used the entire surfaces of both proteins for the enzyme–inhibitor complexes. For the antibody–antigen complexes, we used sites of the target antibody as these proteins are very large and the binding sites of Fabs are well known to be in the complementarity determining regions (CDR). As in the bound test, for the antibody–antigen dockings we used only those residues of the target antibody protein which were within 8 Å of the known antigen-binding site. This gave target sites that were still big enough to give the GA plenty of scope for finding incorrect orientations. We also take the view that protein docking should be a pragmatic exercise—all available biochemical information can be used if necessary.

For the “other” complexes, we first attempted to dock using the whole of both proteins. Given the mixed results of this strategy (see below), we later chose the active site for the larger protein of each complex and repeated the GA runs with these sites. These sites were chosen by looking at the known structure of the crystallized complex and selecting any residue of the target protein, which had an atom within 8 Å of a query protein atom. The exceptions were 3hhr (human growth hormone–receptor complex) and 2btf (β-actin–profilin complex), in which all residues within 6 Å of a query atom were chosen (the shape of these targets indicating that most target atoms are within 8 Å of a query atom). Clearly, this procedure would be impossible when predictively docking two proteins, the structure of whose complex was unknown. However, when two proteins are known to associate, there is frequently some biochemical evidence (e.g., from mutagenesis studies or prediction studies) to indicate the likely position of a binding site residue, so it is reasonable to assume some knowledge of the general region in which binding occurs.

The results of these native dockings are given in Table IVA–D. As suggested by Gabb et al.,[16] we have assessed the quality of these dockings by calculating the RMSD of the $C\alpha$ interface atoms after the $C\alpha$ atoms of both proteins of the predicted complex are superposed on those of the crystal complex.[16] (Any atom within 10 Å of the opposite protein is considered an interface atom.) This value is a good measure of the correctness of the predicted complex as it removes the distorting effect that small errors at the interface may have on the position of atoms distant from that interface. For each complex, we also created a “reference structure” by separately superposing the $C\alpha$ atoms of the unbound proteins onto their bound counterparts. We then calculated the RMSD between the $C\alpha$ interface atoms of the reference structure and the crystallized complex. This base RMSD value is a good approximation to the best that could be expected of a rigid-body docking process. We consider a ”correct“ docking to be one in which the $C\alpha$ interface RMSD minus base RMSD is <4 Å, and dockings are included in Table IV only if this criterion was satisfied. All timings are in CPU minutes for a Silicon Graphics 270-MHz O2 workstation with R12000 processor.

TABLE III. Enzyme–Inhibitor, Antibody–Antigen, and Other Complexes

Complexed proteins			Uncomplexed proteins					
PDB code	Res[a] (Å)	Description	PDB code	Res[a] (Å)	No. of residues	PDB code	Res[a] (Å)	No. of residues
A. Enzyme–Inhibitor Complexes								
1brb	2.1	Trypsin/pancreatic trypsin inhibitor	1bra	2.2	223	1bpi	1.1	51
1cgi	2.3	α-Chymotrypsinogen/pancreatic trypsin inhibitor	1chg	2.5	245	1hpt	2.3	56
2kai	2.5	Kallikrein A/pancreatic trypsin inhibitor	2pka	2.1	232	1bpi	1.1	57
2ptc	1.9	β-Trypsin/pancreatic trypsin inhibitor	3ptn	1.7	223	4pti	1.5	58
2sic	1.8	Subtilisin/streptomyces inhibitor	1sup	1.6	275	2ssi	2.6	107
2sni	2.1	Subtilisin/chymotrypsin inhibitor	2sbt	2.8	275	2ci2	2.0	64
1acb	2.0	α-Chymotrypsin/Eglin C	4cha	1.7	245	1acb[b]		63
1brc	2.5	Trypsin/APPI	1bra	2.2	223	1brc[b]		56
1cho	1.8	α-Chymotrypsin/Ovomucoid	4cha	1.7	245	1cho[b]		53
1cse	1.2	Subtilisin Carlsberg/Egin C	1scd	2.3	274	1cse[b]		63
1ppe	2.0	Trypsin/CMT-I	3ptn	1.7	223	1ppe[b]		29
1sbn	2.1	Subtilisin Novo/Eglin C	1sup	1.6	275	1sbn[b]		63
1stf	2.4	Papain/Stefin B	1ppn	1.6	212	1stf[b]		98
1tab	2.3	Trypsin/BBI	3ptn	1.7	223	1tab[b]		36
1tgs	1.8	Trypsinogen/pancreatic trypsin inhibitor	1tgt	1.5	225	1tgs[b]		56
2tec	2.0	Thermitase/Eglin C	1thm	1.4	279	2tec[b]		63
4htc	2.3	α-Thrombin/Hirudin	2hnt	2.5	290	4htc[b]		61
1udi	2.7	Uracil-DNA glycosylase/inhibitor	1udh	1.8	227	1udi[b]		83
1brs	2.0	Barnase/Barstar	1a2p	1.5	107	1a19	2.8	89
B. Antibody–Antigen Complexes								
1mlc	2.1	Fab D44.1 (a,b-chains)/lysozyme	1mlb	2.1	432	1lza	1.6	129
1vfb	1.8	Fv D1.3 (a,b-chains)/lysozyme	1vfa	1.8	223	1lza	1.6	129
1nca	2.5	Fab NC41 (1,h-chains)/Neuraminidase	1nca[b]		435	7nn9	2.0	389
1nmb	2.5	Fab NC10 (1,h-chains)/Neuraminidase	1nmb[b]		388	7nn9	2.0	389
1igc	2.6	Fab (1,h-chains)/*Streptomyces* protein G	1igc[b]		435	1igd	1.1	58
1jel	2.8	Fab JE142 (1,h-chains)/HPR	1jel[b]		432	1poh	2.0	85
3hfl	2.7	Fab HyHel-5(1,h-chains)/lysozyme	3hfl[b]		427	1lza	1.6	129
3hfm	3.0	Fab HyHel-10(1,h-chains)/lysozyme	3hfm[b]		429	1lza	1.6	129
C. Other Complexes								
1atn	2.8	Actin/deoxyribonuclease I	1atn[b]		258	3dni	2.0	373
1gla	2.6	Glycerol kinase/GSF III	1gla[b]		489	1f3g	2.1	161
1spb	2.0	Subtilisin/subtilisin prosegment	1sup	1.6	264	1spb[b]		71
2btf	2.6	β-Actin/profilin	2btf[b]		375	1pne	2.0	160
3hhr	2.8	Human growth hormone/receptor	3hhr[b]		393	1hgu	2.5	185
1mda	2.5	Methylamine dehydrogenase/Amicyanin	2bbk		489	1aan	2.0	103
temblip[c]	1.7	β-Lactamase/β-lactamase inhibitory protein	TEM1[c]	1.7	288	BLIP[c]		165

[a] Res is the resolution in Ångstroms of the crystallized structure.
[b] Data taken from complex.
[c] The coordinates of TEM-1, BLIP, and the temblip complex were obtained from Professor M.N.G. James.

Enzyme–Inhibitor Complexes

Table IVA shows that the GA performed extremely well on these complexes. In every case, for at least one parameter combination, a complex that was close to the PDB complex was found. This means that, although there are literally millions of ways in which one protein may be oriented with respect to the other, the GA produced a list of 400 such orientations, which always contained at least one (and in most cases many) approximately correct prediction. For 13 of the 19 complexes, correct hits were found for every parameter combination. For 11 of the complexes, at least one good hit was found ranked in the top 5 for at least one set of parameters. Figure 2a shows the $C\alpha$ trace of the best predicted docking of the Kallikrein A (PDB code 2pka)–pancreatic trypsin inhibitor (PDB code 1bpi) superposed onto the actual complex (PDB code 2kai). The predicted complex is shown in lime green (enzyme) and red (inhibitor) and the actual complex in blue (enzyme) and turquoise (inhibitor).

The two complexes, 1cho and 2sni, which gave similarly high-ranking hits when docked using bound coordinates gave quite different results when docked using native coordinates. The worst top rank for the native 1cho docking was 5, whilst for 2sni the best native docking rank

TABLE IV. Results of Enzyme–Inhibitor, Antibody–Antigen, Other Complex, and Other Complex (Site)

				Normal angle ±20° *J* 1.5			Normal angle ±15° *J* 1.0			Normal angle ±20° *J* 3.0			Normal angle ±30° *J* 3.0			
Complex[a]	Tdots[b]	Qdots[c]	Base[d]	Rank[e]	RMSD[f]	Nhits[g]	Rank[e]	RMSD[f]	Nhits[g]	Rank[e]	RMSD[f]	Nhits[g]	Rank[e]	RMSD[f]	Nhits[g]	Total time[h]
A. Enzyme/Inhibitor Results																
1brb	8008	2768	0.38	4	1.40	5	13	3.52	7	3	2.76	9	5	4.08	11	380
1cgi	8941	2705	1.48	83	5.14	2	None			7	4.24	4	34	5.04	8	360
2kai	9095	2768	1.00	6	1.23	13	9	3.78	7	48	4.26	5	4	4.39	8	380
2ptc	8184	2670	0.74	13	3.15	4	17	4.70	4	2	3.63	3	1	4.38	6	360
2sic	8348	4547	0.38	70	4.00	1	62	2.87	1	2	2.65	9	25	3.91	5	560
2sni	8694	2976	0.85	50	2.91	3	25	3.89	2	14	4.08	5	51	2.05	2	400
1acb	8995	2855	0.59	None			None			None			2	2.32	1	380
1brc	8008	2287	0.39	1	2.41	13	1	2.41	20	1	2.41	14	28	2.63	5	280
1cho	8995	2573	0.33	1	1.87	13	2	2.06	7	5	2.60	2	1	1.79	9	300
1cse	8961	2912	0.29	37	2.28	2	25	3.89	2	11	2.04	5	36	2.40	8	360
1ppe	8184	1581	0.69	1	0.97	13	36	0.77	7	1	1.00	13	2	0.96	12	200
1sbn	8348	2903	0.29	4	3.58	12	1	4.25	9	1	3.80	16	4	2.49	14	380
1stf	8083	4592	0.27	56	2.01	2	None			18	2.80	3	37	3.00	1	600
1tab	8184	2043	0.69	13	4.36	1	None			None			None			260
1tgs	8547	2729	0.67	None			56	1.89	1	31	1.97	1	None			360
2tec	8403	2965	0.19	None			12	0.79	1	38	1.58	2	18	1.27	4	360
4htc	11179	2916	0.76	33	2.82	1	7	2.55	2	15	3.13	2	2	1.91	2	380
1udi	9649	3715	0.37	6	3.12	3	29	1.68	1	9	2.98	2	13	2.23	3	460
1brs	4681	4061	0.47	36	3.94	2	16	3.25	1	28	3.05	1	1	3.14	2	520
B. Antibody–Antigen Results																
1mlc	3382	5032	0.88	74	3.19	1	None			18	3.42	1	55	3.17	2	660
1vfb	3974	5032	1.12	None			None			None			None			660
1nca	4829	14207	0.38	None			None			None			74	1.19	1	2000
1nmb	4443	14207	0.22	None			None			None			None			1920
1igc	6110	2887	0.74	19	4.28	4	44	3.45	1	31	2.98	1	20	4.60	4	400
1jel	4193	3609	0.27	None			None			None			100	3.58	1	500
3hfl	3172	5032	0.42	76	2.64	2	None			70	2.76	2	83	2.50	1	660
3hfm	3917	5032	0.44	67	3.63	1	None			None			50	3.38	1	660
C. Entire Other Complex Results																
1atn	14931	9620	0.33	None			None			None			None			1340
1gla	17669	5527	0.49	None			None			None			None			600
1spb	8348	3495	0.35	None			None			None			43	3.00	1	420
2btf	15173	5450	0.29	None			None			None			None			800
3hhr	16045	10170	2.92	None			None			None			91	6.78	1	1800
1mda	20445	4127	2.69	None			None			None			None			640
temblip	9803	6630	0.88	66	3.10	1	None			None			None			780
D. Other Complex (Site) Results																
1atn	3747	9620	0.33	50	3.97	1	66	2.06	2	48	1.35	4	90	3.98	1	1340
1gla	1672	5527	0.49	None			None			None			40	4.35	2	600
1spb	3495	3913	0.35	None			None			40	1.12	1	73	1.42	1	440
2btf	3446	5450	0.29	20	4.26	1	33	2.85	2	None			16	1.90	2	580
3hhr	6687	10170	2.92	None			35	5.80	1	None			56	0.90	1	1800
1mda	5796	4127	2.69	12	5.00	6	21	6.44	4	10	4.95	6	1	5.05	18	640
temblip	6630	6469	0.88	10	2.16	3	53	3.08	1	8	3.51	6	8	2.81	4	780

[a] Complex is the Protein Data Bank (PDB) code of the complex.
[b] Number of surface dots generated for the target protein.
[c] Number of surface dots generated for the query protein.
[d] Root-mean-square deviation (in Å) obtained by superposing the $C\alpha$ atoms of each unbound component onto the corresponding bound protein and then calculating the RMSD between interface atoms of this reference structure and the crystallized complex.
[e] Rank (by fitness score, of 100) of the highest ranked solution that resembled the complex.
[f] Root-mean-square deviation of interface $C\alpha$ atoms.
[g] Total number (out of 100) of complex-resembling hits found.
[h] Total time (CPU mins) is the time taken for 20 runs giving 100 solutions.

was 14. However, this could be explained by the fact that the inhibitor for the native 1cho docking was taken from the complex, whereas for 2sni, both proteins were native. The majority of the incorrect orientations predicted by the GA, and certainly most of the high-ranking incorrect predictions, did position the inhibitor in the active site of the enzyme, with the error being in the orientation of the inhibitor within the site. This reflects the more deeply concave nature of the enzyme active sites compared with the remainder of the enzyme surface.

The run times given are the total times for 20 runs of the GA, producing 100 solutions. The time needed to generate

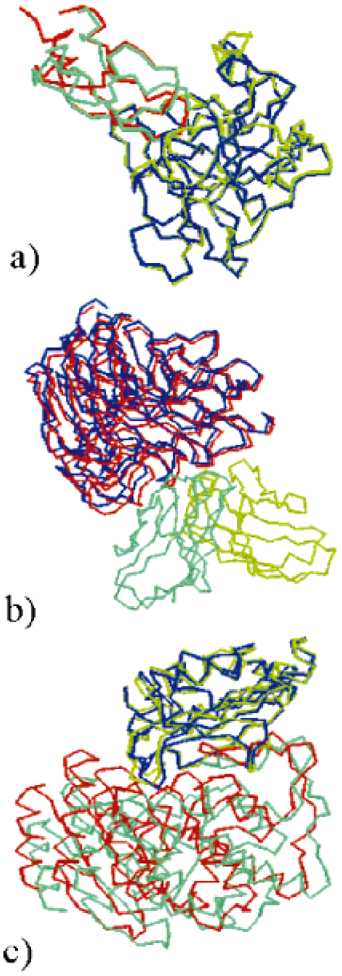

Fig. 2. Docking native complexes. The $C\alpha$ traces of the predicted complex with the lowest RMSD is shown superposed onto the crystallized complex. **a:** 2kai. The predicted enzyme and inhibitor are shown in lime green and red with the crystallized enzyme and inhibitor in blue and turquoise respectively. **b**:1nca. Only part of the Fab is shown, with lime green for the light chain and turquoise for the heavy chain. This is the same for both crystallized and predicted complex as the docked Fab is taken from the crystallized complex. The predicted neuraminidase is shown in red and the actual in blue. **c:** temblip. The predicted enzyme and inhibitor are shown in lime green and red, with the crystallized enzyme and inhibitor in blue and turquoise, respectively.

400 solutions varied between approximately 17 h (for the smallest complex, 1tab) and 40 h (for 1stf, the largest complex). However, as the fitness score produced by the GA was also a ranking of the solution, no subsequent ranking or screening was necessary. The run times increased, particularly with the size of the query protein. This was to be expected, as this was the protein upon which most computation was carried out. Thus docking the barnase–barstar complex (1brs) required a similar time to that required by papain–stefin B (1stf) because the query proteins had a similar number of surface dots, although papain had nearly twice as many dots as barnase.

Antibody–Antigen Complexes

For each of these complexes, the query protein was docked against a site within the CDR of the antibody

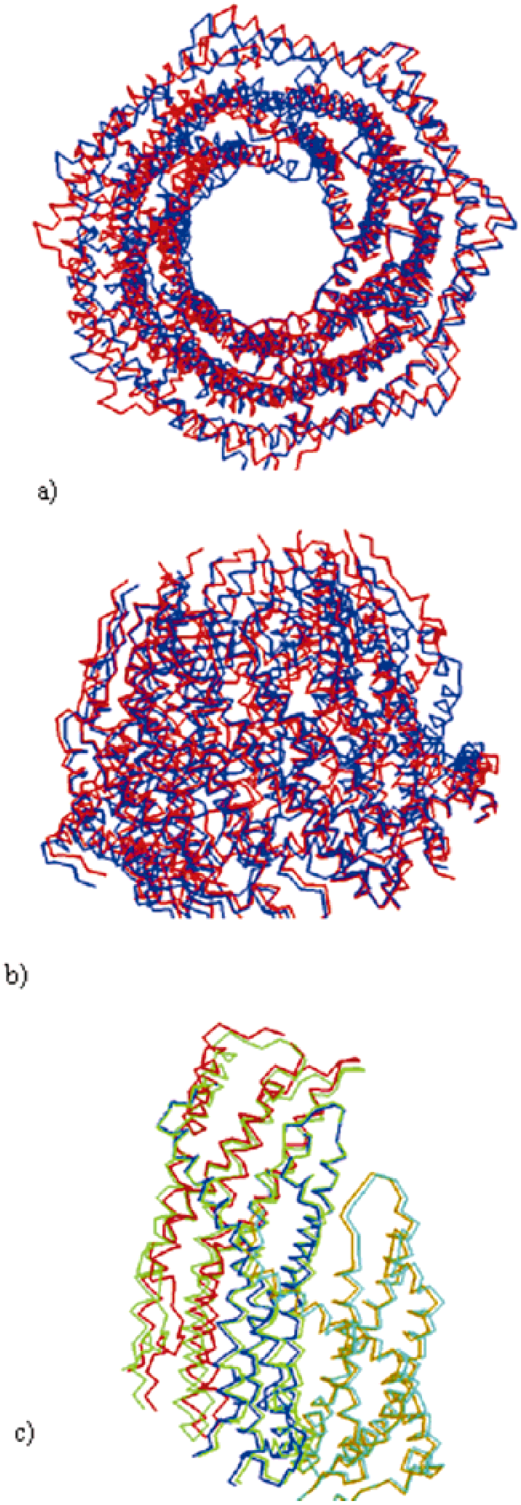

Fig. 3. Reassembling a heptamer. Protein Data Bank complex 1avo reassembled by repeatedly adding a docked subunit. **a:** End-on view of the $C\alpha$ chains of the helical assembly, actual complex in blue with predicted superposed in red. **b:** Side-on view of a. **c:** Small errors in docking the first pair of subunits mean that subsequent subunits become further out of step. This is just the first three subunits; the yellow chain is the first target, the blue is the second, and the red would be the target for the fourth subunit (not shown). The actual complex is in turquoise.

(selected as described above). Even so, the dockings still generally involved more surface dots than for the enzyme–inhibitor cases described above and thus provided a more severe test, the results of which are given in Table IVB. The bound docking tests suggested that the GA did not perform quite as well on antibody–antigen complexes, and the native test confirmed this. Nevertheless, in 6 of 8 cases, the GA did find at least one hit in the top 400, the two exceptions being the Fab–lysozyme complex 1vfb and the Fab–neuraminidase complex 1nmb. In fact, in the case of 1nmb, which (with 1nca) has the largest query protein of any complex considered (neuraminidase has 389 residues), hits were found that somewhat resembled the crystallized complex but whose RMSD (of the order of 7 Å) were too great too allow them to be considered correct.

The complexity of the docking was also reflected in the longer run times. The run time is generally dependent on the size of the query protein. Hence, all the runs with lysozyme (which has 129 residues) as the query took about half an h for one run and thus to obtain 400 solutions took about 40 h. Neuraminidase is a very large query protein (389 residues) and the two neuraminidase complexes took about 120 h for 400 solutions. Figure 2b shows the $C\alpha$ trace of the predicted Fab–neuraminidase (PDB codes 1nca, 7nn9) complex, with the lowest RMSD, superposed on the actual 1nca complex. For clarity, only the part of the Fab that contacts the neuraminidase is shown, with its light chain in lime green and its heavy chain in turquoise. This is the same for the predicted and actual complexes, since the Fab coordinates used in docking were taken from the complex. The predicted neuraminidase is shown in red and the complexed in blue.

Other Complexes

These complexes provided a very stiff test indeed for the GA. To the best of our knowledge, most have not previously been used in theoretical docking studies. Exceptions are the TEM-1/BLIP complex, which was the subject of a docking challenge,[7] and the human growth hormone–growth hormone receptor complex (3hhr), which has recently been studied.[45] Although human growth hormone is itself a complex, consisting of two chains that are identical in sequence, but different in conformation, we have followed Betts and Sternberg[8] in treating the hormone as a single preassembled unit; this we have attempted to dock with its receptor. All the complexes are very large, and the dockings involved many surface dots. Another particular problem was that, in some cases, the conformation of the native protein had some significant differences to that of the complex. For example, the RMSD between native methylamine dehydrogenase (2bbk) and that in the 1mda complex with amicyanin is 6 Å—considering just those residues 4 Å from the interface gives an RMSD of 2.8 Å. This clearly poses a problem for a rigid body docking procedure.

In view of these difficulties, the GA performed better than might have been expected, even when docking entire proteins (Table IVC). For three of the complexes—3hhr, 1spb (subtilisin–subtilisin prosegment complex), and TEM1-BLIP—it found a correct docking, although in each case it was only 1 from 400 possible solutions. However, when attention was restricted to docking a site of one protein against the whole of the other protein, its performance was, as was hoped, much improved (Table IVD). The GA found at least one hit for each complex. For four of the complexes—1atn, 1gla, 1mda, and TEM1-BLIP—it found several hits for most of the parameter combinations, while for two complexes—1gla and 1spb—only one hit was found. It was pleasing to find that the GA was able to find solutions ranked in the top 10 for the TEM-1/BLIP complex, as our previous attempts (using clique-detection methods) had completely failed for this complex.[28] Figure 2c shows the predicted TEM-1/BLI P complex with the lowest RMSD, superposed on the actual complex. A direct comparison with the docking challenge result is difficult as we are docking with the benefit of hindsight, whereas the challenge participants did not know the answer. However, although all the participants ranked near-native complexes top in their submissions, some did use their judgment to "overrule" the ranking of their scoring procedure,[25] so the fact that our best hit was ranked 8 by the GA does not mean that we would have failed the docking challenge. However, in this example, it is notable that the group of Abagyan and Totrov and that of Sternberg and Jackson scored the correct solution top in their respective ranking systems.[7]

Performance of the GA

The GA was run with four combinations of the normal angle/penalty multiplier parameters. Although there was no consistently best performer for the enzyme–inhibitor complexes, overall it did seem that the 180°± 30° with penalty multiplier $J = 3.0$ combination was the most satisfactory. Table V, which summarizes the results for all complexes for this parameter combination, shows that for 30 of the 34 complexes, the GA found at least one solution ranked in the top 100, when these parameters were applied. For 21 of these cases, more than one good solution was found, and for 15 cases, a good solution was found ranked in the top 20. In 6 cases, this parameter combination produced a docking that resembled the complex which was ranked one or two. However, using all four parameter combinations, at least one good solution among 400 was found for all but two of the test complexes, although this was at the expense of having more false hits.

We have been quite generous in our definition of what constituted a hit as we wished to assess the utility of the GA in docking as wide a range of complexes as possible. The interface RMSD of our "hits" ranged from 0.77 Å (enzyme–inhibitor complex, 2tec) to 6.78 Å (human growth hormone–receptor complex, 3hhr), although in the case of 3hhr, the base interface RMSD was 2.92 Å. There are, however, different "grades" of nonhit. For example, the fittest docking of the enzyme–inhibitor complex 2sic with an RMSD of <4 Å is ranked 25 (Table V). In fact, 8 of the top 10 fittest solutions had RMSD values of <5 Å. Because we chose a cutoff RMSD value of 4 Å (relaxed by the base RMSD), none of these were classed as hits although they clearly resembled the crystallized complex more closely than a prediction with RMSD of 20 Å.

TABLE V. Docking All Complexes Using Normal ±30° and J 3.0

Complex	Rank	RMSD	Nhits
1brb	5	4.08	11
1cgi	34	5.04	8
2kai	4	4.39	8
2ptc	1	4.38	6
2sic	25	3.91	5
2sni	51	2.05	2
1acb	2	2.32	1
1brc	28	2.63	5
1cho	1	1.79	9
1cse	36	2.40	8
1ppe	2	0.96	12
1sbn	4	2.49	14
1stf	37	3.00	1
1tab	None		
1tgs	None		
2tec	18	1.27	4
4htc	2	1.91	2
1udi	13	2.23	3
1brs	1	3.14	2
1mlc[a]	55	3.17	2
1vfb[a]	None		
1nca[a]	74	1.19	1
1nmb[a]	None		
1igc[a]	20	4.60	4
1jel[a]	100	3.58	1
3hfl[a]	83	2.50	1
3hfm[a]	50	3.38	1
1atn[a]	90	3.98	1
1gla[a]	40	4.35	2
1spb[a]	73	1.42	1
2btf[a]	16	1.90	2
3hhr	91	6.78	1
1mda[a]	1	5.05	18
temblip[a]	8	2.81	4

[a] Indicates that only part of the target protein was used.

The GA performed well in comparison with two other recent large docking studies that have used unbound proteins. In their study, Ritchie and Kemp[24] docked 18 enzyme–inhibitor and antibody–antigen complexes and found solutions ranking in the top 100 in only 4 cases, when their dockings were constrained to search only the receptor-binding site. They required some knowledge of the ligand-binding site to generate more highly ranked solutions. However, they successfully docked the Fab fragment–antigen complex 1vfb, which the GA was unable to dock using any of our parameter combinations. Our results are also comparable to those of Norel et al., [9] who docked 19 unbound complexes. In 15 cases, they found a solution in the top 100, and 5 of these were in the top 10. However, their test cases only involved 8 different complexes—the multiple results came from using the same proteins crystallized to different resolution or in different crystal forms. Thus, although large, their test set was not as extensive as ours. In general, the GA is somewhat slower than comparable docking methods, with one hundred orientations taking 3–33 h. For example, the methods of Norel et al.,[9] and Ritchie and Kemp[24] took only a few minutes and about 2 h, respectively. The "correct" hits produced by the GA also tend to have slightly worse RMSD than those of these other groups.

Docking Subunits

We are also interested in docking protein subunits to form oligomers, as some proteins, such as bacterial toxins, are monomeric in solution but associate to form oligomeric pores (e.g., aerolysin, hemolysin E). The intersubunit interactions that occur when dimers (or other such complexes) form are thought to be largely driven by hydrophobic forces.[46] As our GA modeled these interactions (requiring non H-bonding atoms to match with similar), we wished to see how well it performed in this context. A difficulty with this kind of docking is that subunits are not often able to be crystallized separately from the complete protein. Thus, docking tests are usually performed using bound coordinates. If the GA were ever to be used to predict the structure of an oligomer, the minimum test it must pass is to reassemble crystallized oligomers. This problem has been studied by Cummings et al.,[47] who reassembled diubiquitin from the subunits of the crystallized dimer and also by Hendrix et al.,[45] who, as part of a three-protein docking of human growth hormone and its receptor (which is composed of two monomers, each of which interacts with the hormone in a different fashion), reassembled the entire complex using the bound coordinates. We successfully assembled this complex (PDB code 3hhr, Table IVC,D) by treating the receptor as a single protein (a dimer). To the best of our knowledge, the reassembly of a unit larger than a trimer has not yet been attempted.

We have tested this use of the GA on three different proteins, HIV-1 protease (PDB code 4hpv), proteasome activator reg-alpha (PDB code 1avo) and diubuquitin (PDB code 1aar). HIV-1 protease is a dimer with two identical chains, A and B, each of 99 residues. We separated the PDB file into two files and docked the two chains, using A as the target. The proteasome activator is a heptamer with 14 chains, 7 pairs that form a circular helical assembly with a central hole. Each subunit consists of two chains of 59 and 139 residues, respectively. In this case, we docked two identical copies of the A,B-chain subunit. This is a slightly more stringent test than that imposed on 4hpv, as the RMSD between the A,B-chain subunit and the adjacent C,D-chain subunit is 0.2 Å. Diubiquitin is a dimer consisting of two distinct copies of the ubiquitin monomer; the $C\alpha$ RMSD between the two chains is 1.4 Å. We docked two identical copies of the A chain which provides a more severe test than docking the A chain against the B chain.

The results, which were very encouraging, are shown in Table VI. For the 4hpv dimer, the top 15 solutions for normal angle ±20° and $J = 1.5$ were all within 2.6 Å of the actual complex. The 1aar and 1avo results, whilst not as outstanding as 4hpv, were still very good with each parameter set providing multiple hits within the top ten ranked solutions. This time, taking all three complexes into consideration, the parameter combination of normal angle ±20° and $J = 1.5$ did seem significantly better than the other three since a correct hit was ranked top for every

TABLE VI. Docking Bound Subunits*

Complex (PDB code)	Normal angle[a]	J[b]	Top Rank[c]	Top 10 hits[d]	Total hits[e]	RMSD[f]	Best RMSD[g]
4hpv	±30°	3.0	1	10	18	1.99	1.52
	±20°	1.5	1	10	20	2.00	0.98
	±20°	3.0	1	10	13	1.86	1.72
	±15°	1.0	1	10	27	2.79	1.38
1avo	±30°	3.0	4	4	7	2.31	1.27
	±20°	1.5	1	4	6	1.42	1.42
	±20°	3.0	2	2	3	3.53	3.53
	±15°	1.0	4	2	6	1.81	1.81
1aar	±30°	3.0	4	3	5	2.66	2.47
	±20°	1.5	1	5	8	2.33	2.17
	±20°	3.0	1	3	4	2.32	2.32
	±15°	1.0	2	6	7	1.92	1.87

*Each row corresponds to 20 runs of the GA, each producing 5 niches giving 100 solutions per row.

[a] 180° normal angle is the angle allowed between the surface normals for a permitted match.

[b] Penalty multiplier.

[c] Rank (by fitness score, of 100) of the highest ranked solution which resembled the complex.

[d] Number of complex-resembling hits in the 10 top ranked solutions.

[e] Number of complex-resembling hits (of 100).

[f] Root-mean-square deviation of the query $C\alpha$ atoms of the predicted complex when the target monomer is superposed upon the monomer of the crystallized dimer.

[g] Predicted complex closest to the actual complex. Predicted complexes were deemed to resemble the complex if this RMSD was <4 Å.

complex. However, the docking was done using bound coordinates and we anticipate that native docking might again require looser angle matching constraints.

When attempting to predict the structure of an oligomer, one needs to assemble multiple copies of a subunit. We have done this for the proteasome activator reg-alpha, 1avo. Figure 3a,b shows views of a heptamer, based on the predicted 1avo dimer with the lowest RMSD to the actual dimer, superposed onto the actual heptamer. Naturally, the predicted assembly becomes slightly further out of step with the crystallized complex with each additional subunit. Figure 3c shows just the first three docked subunits. Even so the overall RMSD between all 1,400 $C\alpha$ atoms of our assembly and the crystallized complex is only 4.2 Å.

The case of 1avo illustrates the lack of even-handedness of the treatment of the query and target proteins. As both are the same, we anticipated that the GA would produce a similar number of dockings with the query bound "before" as "after" the target (Fig. 4a). In practice, however, most correct solutions (and all the correct solutions which rank highly) have the query after the target (in position 1, see Fig. 4a). The reason for this seemed to be that there was a helix (residues 1–31), which stuck out from the monomer (Fig. 4b). This helix was "thin" (had few interior points defined) in the target and so when the GA approached position 1 this did not cause a clash with query dots because the query has no interior points. However, when the GA approached position 2 and overlap of the query helix with the target interior occurred, this caused a penalized clash (Fig. 4c). This type of situation may be a reason for the failure of the GA in the unbound cases where few or no solutions were found.

The 1avo assembly also illustrates a situation in which the GA may be of immediate practical use. When a circular assembly (e.g., a trans-membrane pore) is expected to be composed of a number of subunits, a further constraint is imposed on the docking by the fact that when a number (which may be known) of the subunits are docked together, the last must also dock with the first. From a list of 100 dockings produced by the GA, only a few, if any, will satisfy this condition.

Diubiquitin is unusual, in that the structure of the monomer (ubiquitin) has also been determined alone. As a final test, we also attempted to form diubiquitin by docking two copies of ubiquitin. This was also attempted by Cummings et al.,[47] who reported success when they used only part of the monomer as their docking target. They also found it necessary to delete the C-terminal residue Gly 76 and to truncate the flexible residue Arg 42 by deleting all atoms beyond the CB. They give biochemical justifications for all their modifications. Our results proved very similar to theirs. The GA gave no correct hits when the entire surface of the monomer was used for both query and target. However, when the target protein was restricted to only those residues which are within 8 Å of the other monomer (in the crystallized dimer) and the same deletion of Gly 76 and alteration of Arg 42 performed, the GA found 6 correct hits using the parameter combination of normal angle ±30° and $J = 3$. The fittest correct hit (with an interface RMSD of 3.9 Å to the crystal dimer) was ranked.[4]

CONCLUSIONS

We have developed a GA for protein–protein docking, based on matching complementary Connolly surfaces. We have tested the GA first on three complexes using bound coordinates and then on 34 complexes (some very large) using unbound coordinates and it has performed very well.

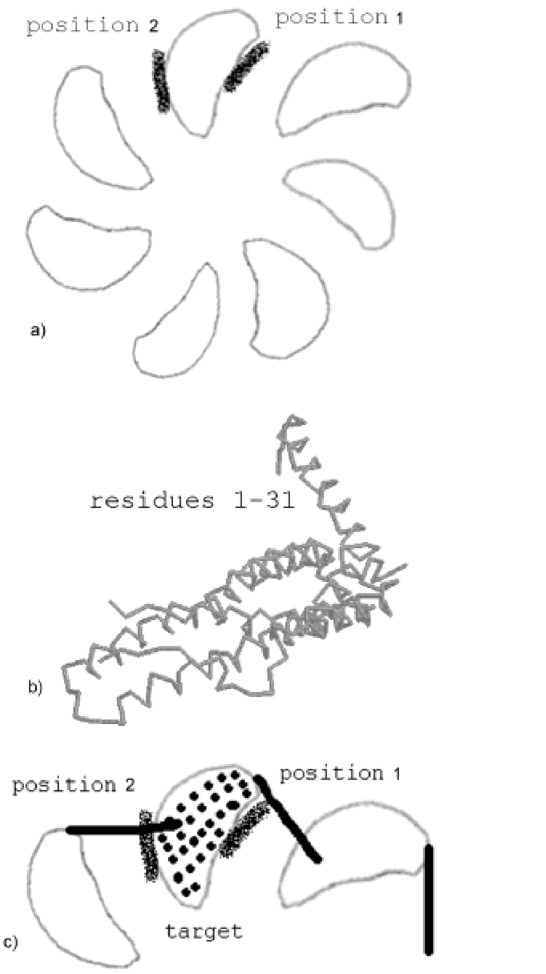

Fig. 4. The GA prefers one of two alternate binding sites for 1avo The two binding sites are shown by chalk lines at positions 1 and 2 of the target monomer. **a:** As 1avo is a heptamer, the query protein will bind in position 1 and position 2. However, almost all correct dockings found by the GA were in position 1. **b:** 1avo monomer showing residues 1–31 protruding. **c:** The target protein has interior points defined. If the query helix (black line) penetrates the target (position 2), a penalized clash with interior points occurs. If the target helix penetrates the query (position 1) no such clash occurs. Although no such clash occurs in the actual binding of the two monomers, this clash drives the GA away from the solution at position 2.

When proteins dock, many factors have an influence. We have modeled only two, the surface shape and the "chemical shape." Even so, these have been sufficient, in most cases, to find orientations near to the crystallized complex. We have established a set of parameters for which the GA found solutions ranked in the top 100 for 30 of the 34 complexes. The GA has also been used to dock subunits, including successfully reassembling a heptamer from bound components.

A great strength of the GA is that the solutions produced are already ranked which eliminates the need for a further screening/ranking step in the docking process. Whilst it is not true that the fittest solution was necessarily correct, a list of 100 putative orientations, of which at least one is probably close to the crystallographic complex, is a very good start, although there is still room for improvement.

The fitness function at present is very simple. We intend to try to enhance it by including some electrostatic elements and also by weighting some features more heavily than others. For example, in antibody–antigen interactions, tyrosine residues represent over a quarter of the total interaction energy donated by the antibody.[27] Therefore, we might choose to weight interactions involving tyrosines accordingly, just for antibody–antigen docking. Frequently, when docking, some feature of the binding site may be known. For example, mutagenesis studies may implicate a particular residue in binding. We intend to modify the GA so that such information may be included if available.

ACKNOWLEDGMENTS

The authors thank Professor M.G. James for supplying us with the coordinates of TEM-1 and BLIP and the Biotechnology and Biological Sciences Research Council and Tripos, Inc., for computing resources. The Krebs Institute is a BBSRC designated Biomolecular Sciences Centre and is a member of the BBSRC North of England Structural Biology Centre.

REFERENCES

1. Berman HM, Westbrook J, Feng Z, Gililand G, Bhat TN, Weissig, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
2. Hart TN, Read RJ. Multiple-start Monte Carlo docking of flexible ligands. In: Merz K Jr, Le Grand S, editors. The protein folding problem and tertiary structure prediction. Boston: Birkhauser; 1994. p 77–108.
3. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Perrin TE. A geometric approach to macromolecule-ligand interactions. J Mol Biol 1982;161:269–288.
4. Miller MD, Kearsley SK, Underwood DJ, Sheridan RP. Flog—a system to select quasi-flexible ligands complementary to a receptor of known 3-dimensional Structure. J Computer Aided Mol Design 1994;8:153–174.
5. Rarey M, Wefing S, Lengauer T. Placement of medium-sized molecular fragments into active sites of proteins. J Computer Aided Mol Design 1996;10:41–54.
6. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
7. Strynadka NCJ, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F, Olson A, Duncan B, Rao M, Jackson R, Sternberg M, James MNG. Molecular docking programs successfully predict the binding of a beta-lactamase inhibitory protein to TEM-1 beta-lactamase. Nature Struct Biol 1996;3:233–239.
8. Betts MJ, Sternberg MJE. An analysis of conformational changes on protein–protein association: implications for predictive docking. Protein Eng 1999;12:271–283.
9. Norel R, Petrey D, Wolfson HJ, Nussinov R. Examination of shape complementarity in docking of unbound proteins. Proteins 1999;36: 307–317.
10. Helmer-Citterich M, Tramontano A. Puzzle: a new method for automated protein docking based on surface shape complementarity. J Mol Biol 1994;235:1021–1031.
11. Jiang F, Kim SH. Soft docking—matching of molecular-surface cubes. J Mol Biol 1991;219:79–102.
12. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular-surface recognition—determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 1992;89:2195–2199.
13. Lenhof HP. New contact measures for the protein docking problem. Proceedings of the 1st Annual Conference on Computational Molecular Biology, RECOMB 1997; 182–191.
14. Meyer M, Wilson P, Schomberg D. Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking. J Mol Biol 1996;264:199–210.
15. Shoichet BK, Kuntz ID. Protein docking and complementarity. J Mol Biol 1991;221:327–346.

16. Gabb HA, Jackson RM, Sternberg MJE. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 1997;272:106–120.
17. Jackson RM, Sternberg MJE. A continuum model for protein–protein interactions—application to the docking problem. J Mol Biol 1995;250:258–275.
18. Connolly ML. Solvent-accessible surfaces of proteins and nucleic acids. Science 1983;221:708–713.
19. Connolly ML. Shape complementarity at the hemoglobin a1b1 subunit interface. Biopolymers 1986;25:1229–1247.
20. Norel R, Lin SL, Wolfson HJ, Nussinov R. Shape complementarity at protein–protein interfaces. Biopolymers 1994;34:933–940.
21. Norel R, Lin SL, Wolfson HJ, Nussinov R. Molecular-surface complementarity at protein–protein interfaces—the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 1995;252:263–273.
22. Hou TJ, Wang JM, Chen LR, Xu XJ. Automated docking of peptides and proteins by using a genetic algorithm combined with a tabu search. Protein Eng 1999;12:639–647.
23. Jackson RM, Gabb HA, Sternberg MJE. Rapid refinement of protein interfaces incorporating solvation: application to the docking problem. J Mol Biol 1998;276:265–285.
24. Ritchie DW, Kemp GJL. Protein docking using spherical polar Fourier correlations. Proteins 2000;39:178–194.
25. Palma PN, Krippahl L, Wampler JE, Moura JJG. BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 2000;39:372–384.
26. Dixon JS. Evaluation of the CASP2 docking section. Proteins 1997; 1(suppl):198–204.
27. Jackson RM. Comparison of protein–protein interactions in serine protease–inhibitor and antibody–antigen complexes: implications for the protein docking problem. Protein Sci 1999;8:603–613.
28. Gardiner EJ, Willett P, Artymiuk PJ. Graph-theoretic techniques for macromolecular docking. J Chem Inform Comput Sci 2000;40: 273–279.
29. Pedersen JT, Moult J. Protein folding simulations with genetic algorithms and a detailed molecular description. J Mol Biol 1997;269:240–259.
30. Vivarelli F, Giusti G, Villani M., Campanini R, Fariselli P, Compiani M, Casadio R. Lgann—a parallel system combining a local genetic algorithm and neural networks for the prediction of secondary structure of proteins. Computer Applications in the Biosciences 1995;11:253–260.
31. May ACW, Johnson MS. Improved genetic algorithm-based protein structure comparisons: pairwise and multiple superpositions. Protein Eng 1995;8:873–882.
32. Notredame C, Higgins DG. SAGA: sequence alignment by genetic algorithm. Nucleic Acids Res 1996;24:1515–1524.
33. Verkhivker GM, Rejto PA, Gehlhaar DK, Freer ST. Exploring the energy landscapes of molecular recognition by a genetic algorithm: analysis of the requirements for robust docking of HIV-1 protease and FKRP-12 complexes. Proteins 1996;25:342–353.
34. Pearlman DA. Automated detection of problem restraints in NMR data sets using the FINGAR genetic algorithm method. J Biomol NMR 1999;13:325–335.
35. Wild DJ, Willett P. Similarity searching in files of three-dimensional chemical structures. Alignment of molecular electrostatic potential fields with a genetic algorithm. J Chem Inform Comput Sci 1996;36:159–167.
36. Bayley MJ, Jones G, Willett P, Williamson MP. GENFOLD: a genetic algorithm for folding protein structures using NMR restraints. Protein Sci 1998;7:491–499.
37. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Further development of a genetic algorithm for ligand docking and its application to screening combinatorial libraries. In: Parrill AL, Reddy MR editors. Rational drug design: novel methodology and practical applications. ACS Symposium Series. Washington, DC: American Chemical Society; 1999. Vol 719, p 271–291.
38. Holliday JD, Willett P. Using a genetic algorithm to identify common structural features in sets of ligands. J Mol Graph Model 1997;15:221–232.
39. Gillet VJ, Willett P, Bradshaw J, Green DVS. Selecting combinatorial libraries to optimize diversity and physical properties. J Chem Inform Comput Sci 1999;39:169–177.
40. Poirrette AR, Artymiuk PJ, Rice DW, Willett P. Comparison of protein surfaces using a genetic algorithm. J Comput-Aided Mol Design 1997;11:557–569.
41. Connolly ML. Analytical molecular-surface calculation. J Appl Crystallogr 1983;16:548–558.
42. Goldberg DE. Genetic algorithms in search, optimisation and machine learning. Reading, PA: Addison-Wesley; 1989. 421 p.
43. Davis L. Handbook of genetic algorithms. New York: Van Nostrand-Reinhold; 1991. 385p.
44. Beasley D, Bull DR, Martin RR. A sequential niche technique for multimodal function optimization. Evol Comput 1993;1:101–125.
45. Hendrix DK, Klien TE, Kuntz ID. Macromolecular docking of a three-body system: the recognition of human growth hormone by its receptor. Protein Sci 1999;8:1010–1022.
46. Janin J, Miller S, Chothia C. Surface, subunit interfaces and interior of oligomeric proteins. J Mol Biol 1988;204:155–164.
47. Cummings MD, Hart TN, Read RJ. Monte-Carlo docking with ubiquitin. Protein Sci 1995;4:885–899.