Journal of Biomolecular NMR, 5 (1995) 41–47
ESCOM

J-Bio NMR 279

A genetic algorithm-based protocol for docking ensembles of small ligands using experimental restraints

Robert P. Meadows* and Philip J. Hajduk

Pharmaceutical Products Division, Abbott Laboratories, Abbott Park, IL 60064, U.S.A.

Received 14 December 1994
Accepted 17 March 1995

Keywords: Genetic algorithm; Protein–ligand interactions; Ensemble averaging

Summary

A genetic algorithm (GA) based method for docking ensembles of small, flexible ligands to receptor proteins using NMR-derived constraints is described. In this method, three translations and rotations of the ligand and the dihedral angles of the ligand are represented by binary strings and evolve under the genetic operators of cross-over, mutation, migration and selection. The fitness function for the selection process includes distance and dihedral restraints and a repulsive van der Waals term. The GA was applied to a three-atom model system as well as to the streptavidin–biotin complex using simulated intermolecular distance restraints. In both systems, the GA was able to obtain low-energy conformations when only a single binding site was simulated. Calculations were also performed using distance restraints from two distinct binding sites. In these simulations, the GA was able to obtain low-energy conformations corresponding to ligand molecules in each of the two sites. The inclusion of additional ligands in the ensemble did not result in an energetic benefit, confirming that only two ligand conformations were necessary to fulfill the distance restraints. This method allows for a direct investigation of the minimum number of ligand orientations necessary to fulfill experimental distance restraints, and simultaneously yields detailed structural information about each site.

Introduction

Several problems are encountered in the structure determination of protein complexes with small or weakly bound ligands using NMR-derived restraints. First, because of imprecise distance bounds, multiple conformations may satisfy the distance restraints. Second, no single conformation may satisfy all of the restraints, indicating multiple and distinct binding sites and/or orientations. Thus, in the structural determination of such systems, techniques must be employed which are able to search vast regions of conformational space, but also incorporate the possibility of conformational averaging. A variety of searching algorithms are available which at least partially fulfill these criteria. Monte Carlo and Distance Geometry (Havel and Wüthrich, 1984) based techniques can effectively search large regions of space, but the problem of conformational averaging cannot typically be included. Molecular Dynamics (Nilges et al., 1988) with the use of time-averaged restraints (Torda et al., 1989; Schmitz et al., 1993; Mierke et al., 1994) can incorporate multiple conformations, but conformations which do not correspond to the 'true' minima may be sampled in the averaging process, and the results are sensitive to the time constant (τ) used in the simulation, the length of the simulation, and the initial configuration of the system.

This paper presents an alternative method for NMR-based docking of a ligand to its receptor using the computer-based paradigm of genetic algorithms (GAs). Borrowing from biological principles of adaptation and evolution, computer-based genetic algorithms are becoming an increasingly popular method for optimization in a variety of complex nonlinear systems (Goldberg, 1989; Holland, 1992; Forrest, 1993). The feature of GAs which makes them distinct from other stochastic or deterministic searching algorithms is that they use a population of individual solutions which are able to share information with each other through the genetic operators. In a 'survival of the fittest' scheme, better solutions will reproduce and pass on their genetic information more frequently,

*To whom correspondence should be addressed.

while less fit individuals will die off. Recently, GAs have been successfully used for conformational analysis of small peptides and organic molecules (Judson et al., 1993; McGarrah and Judson, 1993) and in an investigation of the folding of small proteins (Dandekar and Argos, 1994) by evolving a set of dihedral angles, $[\Phi_1,...,\Phi_N]$, to describe particular conformations. Here we have encoded a GA which evolves not only a binary representation of the ligand dihedral angles, but also three translations (x,y,z) and three rotations of the ligand coordinates. In order to investigate whether multiple binding orientations are involved, ensembles of ligand molecules were generated by concatenation of the gene strings for each individual conformation and evolved under the influence of the GA. These ensembles were averaged before fitness evaluation. The utility of this technique is demonstrated on a model three-spin system and on the streptavidin–biotin complex.

Methods

The basic GA

As shown in Fig. 1, the first step of the GA approach requires that an appropriate representation scheme be defined. This scheme must encode all of the values of the evolved parameters into a format suitable for schema processing (Holland, 1992) (typically binary encoding), and the components of the representation are generally referred to as 'genes' and 'chromosomes'. A single gene contains the value of one parameter, whereas a chromosome contains multiple genes. Next, an initial population is generated which consists of random bit strings. The third step of the GA protocol calls for phenotypic evaluation. In this stage, the gene for each member of the population is 'decoded' and the resulting parameters (phenotype) are submitted to the fitness function. After a fitness has been associated with each gene in the population, individual members of the population are selected for mating. Mating probabilities assigned to each member are generally based on their contribution to the total fitness of the population, and the least fit members very often have mating probabilities close to zero. At this stage the actual selection of information occurs, since the best fit individuals will have a higher probability of producing 'offspring' for the next generation. The penultimate step in the GA is analogous to the evolutionary processes of inheritance, where the offspring inherit portions of their parents' genetic material. The genetic operator for the inheritance of such information is called cross-over (shown schematically in Fig. 1). Another commonly used genetic operator is mutation (also shown in Fig. 1), which causes random changes in the genes. Finally, the offspring replace the parents and the process returns to the fitness evaluation stage. The evolution continues until some energy criterion is met or a maximum number of generations is exceeded.

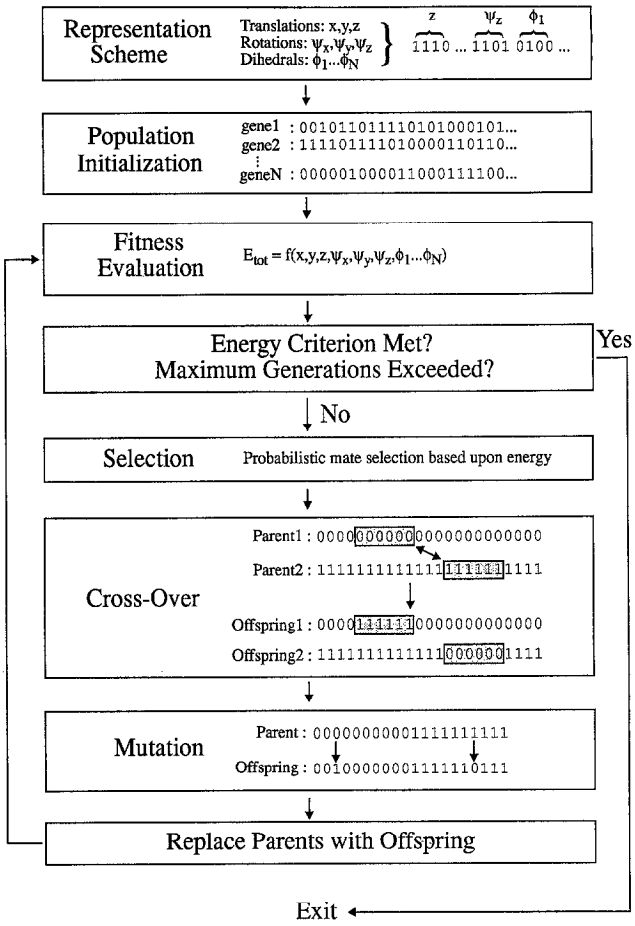


Fig. 1. Flow chart for the genetic algorithm.

Assuming constant bond lengths and bond angles, the conformation of any molecule can be defined by a set of dihedral angles, and this representation scheme has been successfully used in several applications (Judson et al., 1993; McGarrah and Judson, 1993; Dandekar and Argos, 1994). We have incorporated this representation scheme into our GA and have also included x-, y- and z-coordinate translations and rotations of the ligand molecule for the purpose of docking the ligand to the receptor. These parameters were encoded in the GA as follows: each member of the population was represented by a binary string (the chromosome) which contained multiple genes, and each gene encoded the value of an evolved parameter. For example, a typical gene would consist of eight bits and can have genomic values ranging from 0 to 255. The chromosome (11001001)(01000111)(11111111) contains three genes with values of 201, 71 and 255, respectively. These values can then be scaled to yield an appropriate value for the geometric parameter of interest: genes which represented dihedral angles or coordinate rotations were scaled to range between 0–360°, while coordinate translations were scaled to place the ligand within a specified region of Cartesian space (i.e., –20 to 20 Å).

The progress of the GA is driven by the fitness of the individuals within the population. Very fit individuals

should mate more often and produce more offspring. Therefore, the choice of the fitness function is critical to the success of the GA and must meet several criteria. First, the fitness function must determine the optimum values of the evolved parameters. Second, the differences between fit and unfit individuals must be appreciable and appropriate. Finally, the individuals with highest fitness must be reasonable solutions to the problem. The first criterion is fairly obvious and simply states that the parameters which are evolved must actually affect the fitness of the individual, else the evolution is undirected. The second criterion states that there must be large enough differences in fitness between fit and unfit individuals. If such significant differences are not present, then unfit individuals will have essentially the same mating probability as the fit individuals, and no evolution will occur. However, if these differences are too large and the fit individuals dominate the mating pool at the early stages of the simulation, 'premature evolution' may occur (Goldberg, 1989) and only a local solution will be found. The third criterion states that the solutions must be reasonable. For example, given a distance restraint list for some N-atom system, solutions may be found which satisfy the restraints but have many van der Waals (vdW) violations. These solutions are not reasonable, and the vdW terms must be included in the fitness function for proper solutions to be obtained.

Given these criteria, we have chosen a fitness function defined by the total energy of the conformation, E_{tot}, given by:

$$E_{tot} = \sum_{i=1}^{L} F_{NOE}(r_{i,error})^2 + \sum_{i=1}^{M} F_{vdW}(r_i)^{-4} \qquad (1)$$

where F_{NOE} and F_{vdW} are the force constants for the NOE and vdW interactions, respectively, L and M are the number of NOE and vdW interactions, and r_i is the interatomic distance. For the NOE interactions, the error in the internuclear distance, $r_{i,error}$, is given by:

$$r_{i,error} = \begin{cases} \langle r_i^{-3} \rangle^{-1/3} - r_{i,upper} & \text{for } \langle r_i^{-3} \rangle^{-1/3} > r_{i,upper} \\ \langle r_i^{-3} \rangle^{-1/3} - r_{i,lower} & \text{for } \langle r_i^{-3} \rangle^{-1/3} < r_{i,lower} \end{cases} \qquad (2)$$

where $r_{i,upper}$ and $r_{i,lower}$ are the upper and lower bounds of a square-well potential, and

$$\langle r_i^{-3} \rangle^{-1/3} = \left(\frac{1}{N} \sum_{k=1}^{N} r_i^{-3}(k) \right)^{-1/3} \qquad (3)$$

where $r_i(k)$ is the interatomic distance for the kth member in the ensemble, and N is the number of members in each ensemble. Dihedral angles were not restrained in these calculations, but this can easily be incorporated.

In this application we have included several modifications to the basic GA, all of which were developed to maintain high diversity between individuals within the population while decreasing the amount of time spent searching nonproductive regions of conformational space. A distributed GA with migration between subpopulations (Schaffer, 1989) was used with a migration probability of 2.0% per subpopulation per generation, and up to 30% of the size of the subpopulation was allowed to migrate. Typically, four subpopulations were used. The members who migrated were randomly chosen from the 'visiting' population and replaced members in the 'host' population. A proportional selection scheme was used to select mates. In this process, each gene was assigned a mating probability which was proportional to its contribution to the total fitness of the population. For example, a population of four genes with fitnesses of 20, 50, 30 and 100 (arbitrary units) would have mating probabilities of 10, 25, 15 and 50%, respectively. Mates were then randomly chosen according to these probabilities until the number of offspring equaled the size of the population. A maximum of two two-point cross-overs were used with a cross-over probability of 95% per mated pair and a maximum splice length of 30% of the length of the entire chromosome. Exponentially weighted mutation rates (Schaffer, 1989) were used with base mutation probabilities of 0.001% for the most significant bit and 1.0% for the least significant bit. 'Sigma' fitness scaling (Goldberg, 1989; Brodmeier and Pretsch, 1994) was used with a scale factor of 1.0. Elitism (Goldberg, 1989) was also incorporated, where the single best gene from each generation was maintained intact in the next generation. Finally, to enhance diversity, if identical genes survived the mating process, one of the genes was kept and all others were replaced with random bit strings. The GA progressed

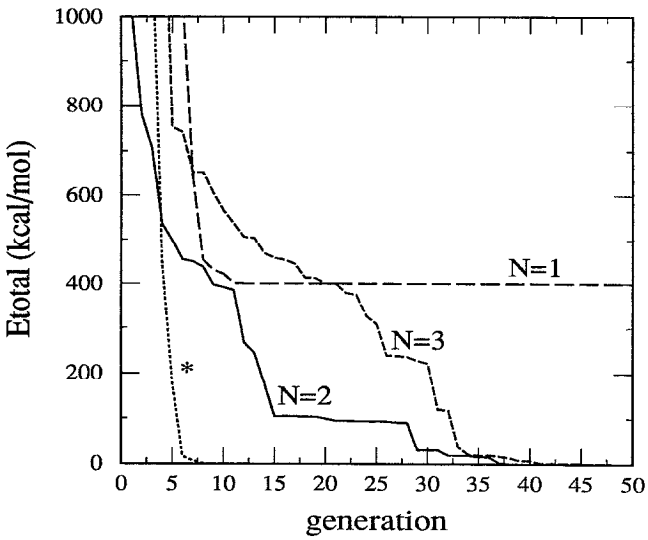


Fig. 2. Energy profiles showing the best total energies for several calculations on the three-atom model system. The dotted line (denoted by the asterisk) represents the best total energy for the single binding-site case as described in the text (only one member was included in the ensemble). For the two-site simulations, the results of evolving N = 1, 2 and 3 conformations are shown.

until either an energy criterion was fulfilled or a maximum number of steps was exceeded.

Evolving ensembles

The procedure for evolving ensembles of ligand molecules is essentially identical to that described above, save that chromosomes for N individuals were concatenated to form a 'superchromosome'. For example, (10101010)(11111111) and (00000000)(01010101) are two chromosomes with two genes of eight bits. Concatenation of these two chromosomes produces the superchromosome (10101010)(11111111)(00000000)(01010101) which encodes parameters for two distinct molecules. For example, genes 1 and 2 can describe x- and y-translations for the first conformation while genes 3 and 4 describe x- and y-translations for the second conformation. During the evolution, the superchromosomes are evaluated, and N distinct conformations are 'docked' onto the protein. vdW violations were calculated and summed for each of these conformations, but the interatomic distances were $\langle r^{-3} \rangle^{-1/3}$ averaged before submission to the NOE evaluation step; see Eq. 3. Importantly, it was the fitness of the ensemble that determined the superchromosome's probability for mating. It should be noted that the population size is independent of the ensemble number. Increasing the population size increases the *number* of chromosomes (or superchromosomes), while increasing the ensemble number increases the size of each chromosome.

Model systems

The GA was first applied to a model system which consisted of two atoms fixed in two-dimensional Cartesian space at (–3,0) and (0,3) (in Å) and a third atom whose x- and y-coordinates were evolved. The third atom was allowed to move within ranges of –6.0 to 6.0 Å and –2.0 to 2.0 Å in the x- and y-dimensions, respectively. Simulations were performed to mimic one- and two-site interactions. For the one-site system, the third atom was required to be within 1.0 Å of the atom at (–3,0). For the two-site system, the third atom was required to be within 1.0 Å of both fixed atoms. A force constant of 50 $Å^{-2}$ was used with no vdW restraints. A single population of 100 members was evolved in these calculations.

The method was also applied to the streptavidin–biotin complex. Coordinates of the crystal structure were obtained (Weber et al., 1989), protons were added (using InsightII, Biosym Technologies, Inc.) and ^1H-^1H intermolecular distance restraints were estimated between the nonexchanging protons of biotin and streptavidin. A total of 32 distances were observed between 2.5 and 4.0 Å and 15 of these were randomly chosen as NOE distance restraints for the simulations. For the purpose of testing the efficacy of ensemble averaging, an alternate ligand position was obtained by manually docking the ligand on the protein surface (see Fig. 5B). No energetic or shape criteria were used, except to avoid vdW violations. For this ligand position (between 2.5 and 4.0 Å), 16 ^1H-^1H distances were observed and 11 of these were used as additional distance restraints. A lower bound of 1.8 Å was used on all restraints, and upper bounds of 3.3 and 4.0 Å were used appropriately. This resulted in six and five restraints between 1.8 and 3.3 Å for the crystal structure site and the manually docked site, respectively, the remaining nine and six restraints being between 1.8 and 4.0 Å. Force constants of 50 $Å^{-2}$ and 50 Å (Torda et al., 1989) were used for the NOE and vdW interactions. The biotin molecule was allowed to move within a 40 $Å^3$ box centered about the streptavidin center-of-mass. As shown in Fig. 5A, the initial population consisted of ligands within a box around streptavidin; no vdW violations between the ligands or between the ligand and the protein were allowed. Three translations and rotations of the biotin molecule were evolved, as well as five dihedral angles of the biotin 'tail', for a total of 11 parameters. Four subpopulations of 100 members (a total population of 400) were evolved, and the minimum-energy conformations were typically found in less than 500 generations. Each generation took less than 5 s on a Silicon Graphics Indy workstation. All software was written in-house.

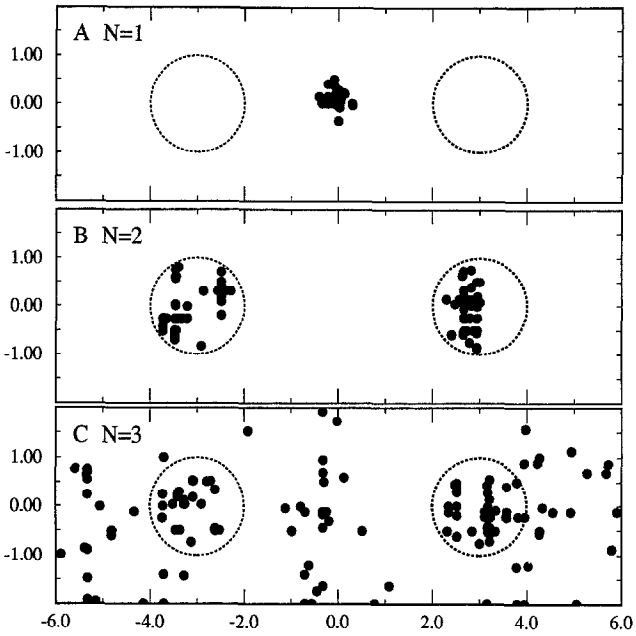


Fig. 3. Cartesian coordinate results for the two-site simulations on the three-atom model system using N = 1, 2 and 3 members in the ensemble. The third atom (represented by the filled circles) was required to be within 1.0 Å of coordinate pairs (–3,0) and (3,0). The regions of space defined by these restraints are enclosed by the dotted circles. The 50 lowest energy solutions are shown for each case.

Results

Three-atom system

Figures 2 and 3 show energy profiles and Cartesian coordinate solutions from several GA simulations on the

three-atom model system. When the third atom was constrained to be within 1.0 Å of the atom at (−3,0), the GA rapidly (< 10 generations) found low-energy conformations (dotted line in Fig. 2). In a second simulation, designed to mimic a two-site interaction, no low-energy conformations were found when only a single conformation was encoded by each chromosome (N = 1). However, low-energy solutions were obtained when two or three conformations were encoded by each chromosome (N = 2 and N = 3; see Fig. 2). For the case of N = 2, all solutions corresponded to one ensemble member in the vicinity of each of the two fixed atoms, as expected. In the case of N = 3, many conformations were obtained which were > 1.0 Å from either fixed atom. These solutions did fulfill the distance restraints (one member within 1.0 Å of each of the two fixed atoms), but the Cartesian coordinates of the third member were virtually unconfined.

Streptavidin–biotin complex

Figure 4 shows the energy profiles for several GA trials on the streptavidin–biotin complex. As an initial test, only restraints corresponding to the single conformation observed in the crystal structure were used. It is clear from Fig. 4A that the GA rapidly evolved to the optimum conformation ($E_{tot} < 20$ in less than 200 generations). In addition, no advantage was gained by ensemble averaging; a low-energy conformation was obtained with a single member. The GA was then applied using restraints from the two ligand binding sites. The progress of the GA with N = 1 (no averaging), and ensemble averaging over two and three conformations is shown in Fig. 4B. As in the three-atom model discussed above, no low-energy conformation was found with only a single member in the ensemble (N = 1). However, low energies were obtained with N = 2 or N = 3. A Cartesian coordinate superposition of the 10 lowest energy ensembles for the N = 2 simulation is shown in Fig. 5B.

Discussion

Ensemble averaging within the GA was successful on both systems described in this paper, and the results highlight several advantages this approach may have over conventional techniques. First, the number of binding sites can be directly investigated. In both systems, two binding sites were used and reasonable energies were obtained only upon increasing the number of conformations in the ensemble from one to two. No decrease in final energy was observed upon increasing the ensemble number to three. This indicates that two conformations are the minimum necessary to fulfill the distance restraints. Second, detailed structural information can be obtained about each binding orientation when the correct number of conformations is used. Figures 3B and C show the dramatic differences in GA solutions for the three-atom system when two and three members are used in the ensemble, respectively. All conformations for the N = 2 case are within the allowed conformational space, while many conformations for the N = 3 case are significantly outside of this range. The same result was observed for the streptavidin–biotin complex (not shown). This indicates that detailed structural information about the ligand interaction site can be lost if the number of conformations in the ensemble exceeds the minimum. This is due to the fact that after the minimum number of conformations to adequately fulfill the constraints are employed, the $\langle r^{-3} \rangle^{-1/3}$ averaging of the interatomic distances allows the 'extra' members to adopt nonideal conformations with virtually no energetic penalty. In the biotin–streptavidin complex, two conformations were determined to be the minimum necessary to fulfill the restraints (as expected), and to enable structural information to be obtained about each site (see Fig. 5B).

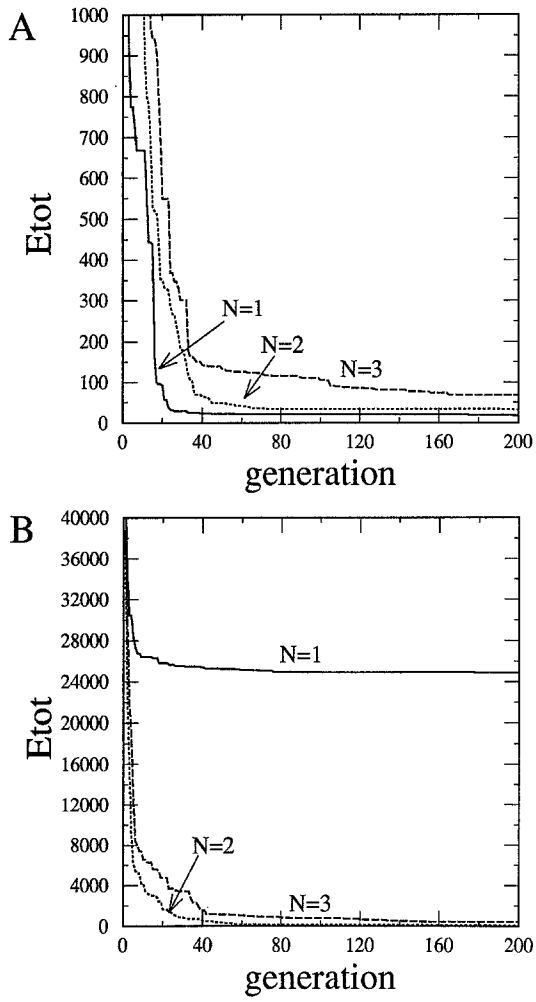
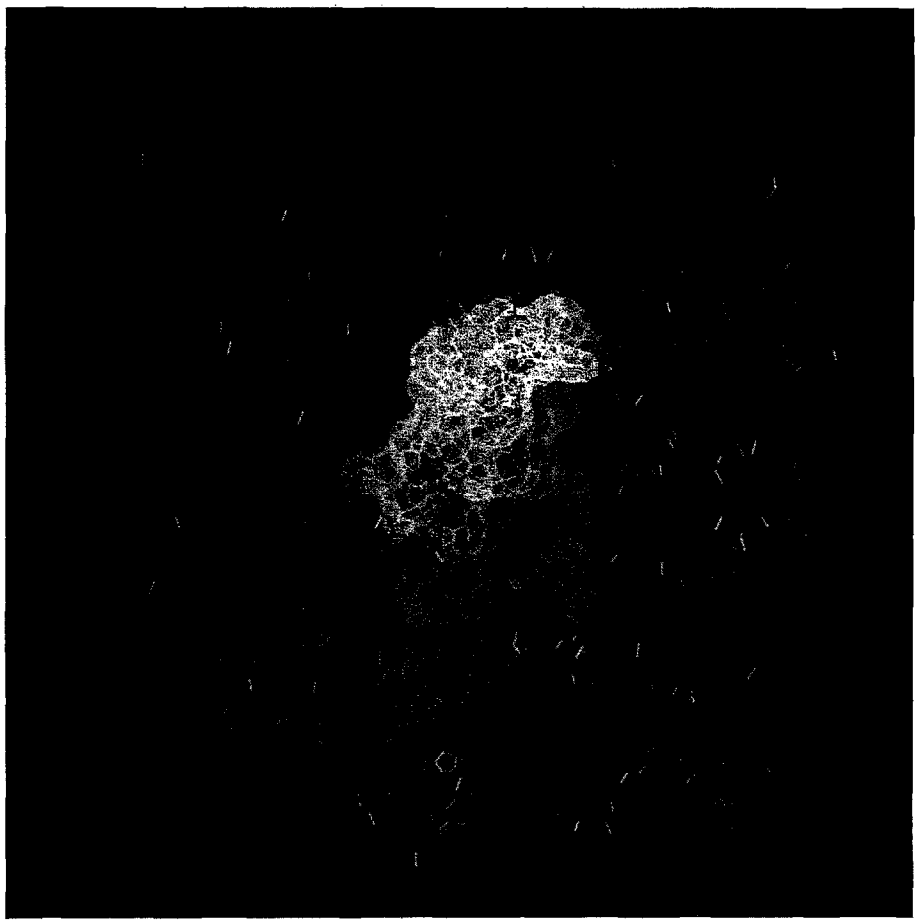
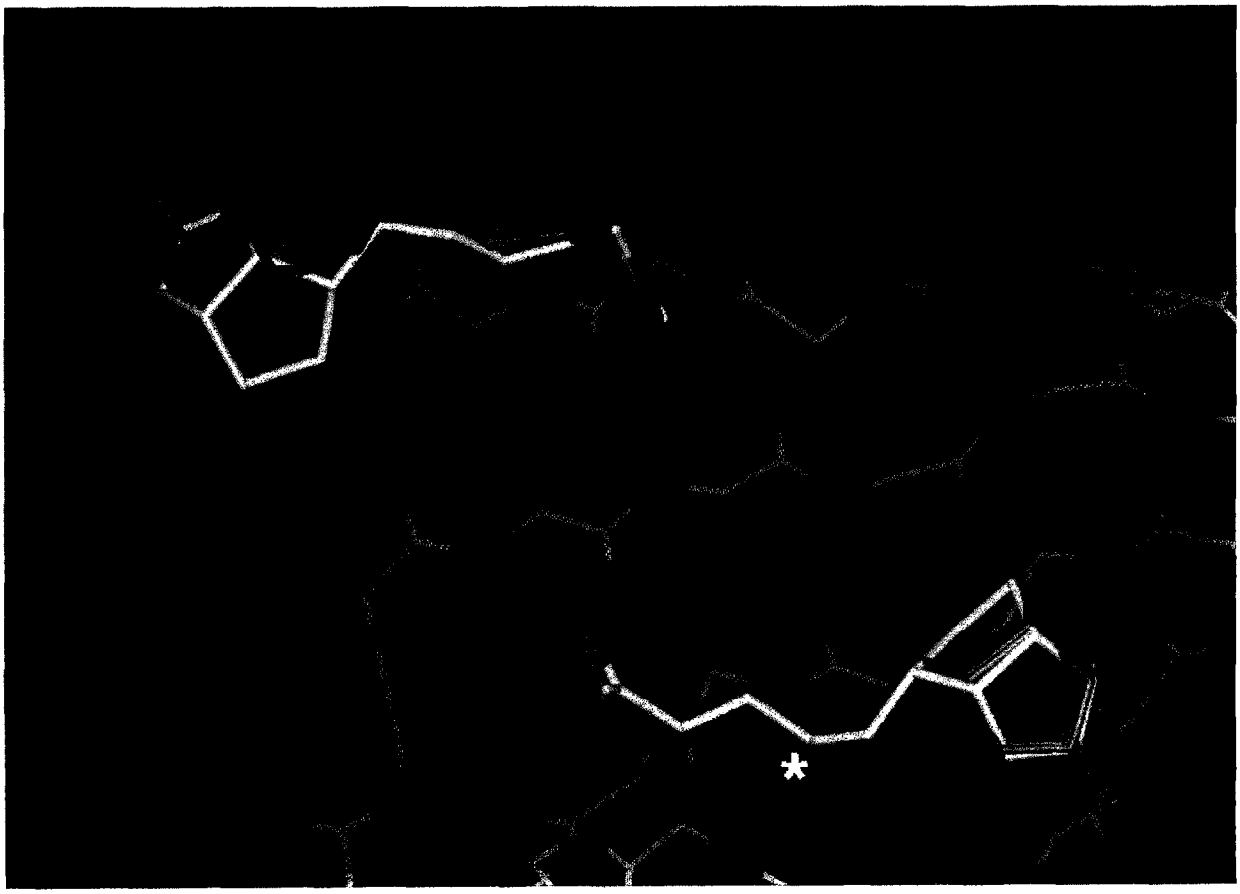


Fig. 4. Energy profile showing the best total energy for (A) a single orientation in the binding site; and (B) two distinct orientations. The results from evolving ensembles of 1, 2 and 3 conformations are shown in each panel. In (A), a restraint set of 15 randomly chosen intermolecular distances found in the streptavidin–biotin crystal structure was used (see text). In (B), 11 additional restraints were added which corresponded to a proximal site on the protein (see text and Fig. 5B).

A



B



←

Fig. 5. (A) The initial population for the biotin–streptavidin complex simulation. A total of 160 ligand coordinates were generated. (B) Cartesian coordinates results of evolving a two-membered ensemble for the two-orientation simulation (N = 2 in Fig. 4B). The lowest E_{tot} ensemble from 10 separate GA simulations is shown in red. The ensembles shown here consist of a single conformation in each of the two binding sites. Shown in white are the ligand positions for the crystal structure (denoted by the asterisk; 15 restraints used in the simulation) and the manually docked ligand (11 restraints used in the simulation). The alternate ligand position was chosen for clarity. No energetic or shape criteria were used, except to avoid vdW violations. Superpositions were based on protein backbone coordinates, which were held fixed during the simulations.

Although not utilized in these examples, the side chains of the protein can also be evolved. The inclusion of side-chain flexibility allows for small conformational changes of the protein in order to accommodate the ligand. This is especially important for side chains which are ill-defined by NMR data. However, the GA is not currently able to handle the evolution of *all* protein dihedrals to allow for large conformational changes upon binding. These searches become intractable without severely restricting the search space, and we have found that the GA is best suited for a small number of evolving parameters.

Conclusions

The results presented here support the viability of using the genetic algorithm for docking a ligand to its receptor protein using NMR-derived restraints. The conformations obtained in the single-site simulations of the biotin–streptavidin complex were free of any vdW violations and were essentially identical to that observed in the crystal structure. Conformational averaging is easily and predictably incorporated into the algorithm, allowing the experimentalist to determine not only the minimum number of binding conformations, but also detailed structural information about each site. All of these characteristics make the GA a valuable tool in the experimental investigation of protein–ligand complexes.

References

Brodmeier, T. and Pretsch, E. (1994) *J. Comput. Chem.*, **15**, 588–595.

Dandekar, T. and Argos, P. (1994) *J. Mol. Biol.*, **236**, 844–861.

Forrest, S. (1993) *Science*, **261**, 872–878.

Goldberg, D.E. (1989) *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, New York, NY.

Havel, T.F. and Wüthrich, K. (1984) *Bull. Math. Biol.*, **46**, 675–698.

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

Judson, R.S., Jaeger, E.P., Treasurywala, A.M. and Peterson, M.L. (1993) *J. Comput. Chem.*, **14**, 1407–1414.

McGarrah, D.B. and Judson, R.S. (1993) *J. Comput. Chem.*, **14**, 1385–1395.

Mierke, D.F., Kurz, M. and Kessler, H. (1994) *J. Am. Chem. Soc.*, **116**, 1042–1049.

Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.*, **239**, 129–136.

Schaffer, J.D. (Ed.) (1989) *Proceedings of the Third International Conference on Genetic Algorithms*, Morgan Kaufmann, San Mateo, CA.

Schmitz, U., Ulyanov, N.B., Kumar, A. and James, T.L. (1993) *J. Mol. Biol.*, **234**, 373–389.

Torda, A.E., Scheek, R.M. and Van Gunsteren, W.F. (1989) *Chem. Phys. Lett.*, **157**, 289–294.

Weber, P.C., Ohlendorf, D.H., Wendoloski, J.J. and Salamme, F.R. (1989) *Science*, **243**, 85–88.