Genetics and Molecular Biology, 27, 4, 605-610 (2004)
Copyright by the Brazilian Society of Genetics. Printed in Brazil
www.sbg.org.br

Research Article

A genetic algorithm for the ligand-protein docking problem

Camila S. de Magalhães[1], Hélio J.C. Barbosa[1] and Laurent E. Dardenne[2]

[1]Laboratório Nacional de Computação Científica, Departamento de Matemática Aplicada e Computacional, Petrópolis, RJ, Brazil.
[2]Laboratório Nacional de Computação Científica, Departamento de Mecânica Computacional, Petrópolis, RJ, Brazil.

Abstract

We analyzed the performance of a real coded “steady-state” genetic algorithm (SSGA) using a grid-based methodology in docking five HIV-1 protease-ligand complexes having known three-dimensional structures. All ligands tested are highly flexible, having more than 10 conformational degrees of freedom. The SSGA was tested for the rigid and flexible ligand docking cases. The implemented genetic algorithm was able to dock successfully rigid and flexible ligand molecules, but with a decreasing performance when the number of ligand conformational degrees of freedom increased. The docked lowest-energy structures have root mean square deviation (RMSD) with respect to the corresponding experimental crystallographic structure ranging from 0.037 Å to 0.090 Å in the rigid docking, and 0.420 Å to 1.943 Å in the flexible docking. We found that not only the number of ligand conformational degrees of freedom is an important aspect to the algorithm performance, but also that the more internal dihedral angles are critical. Furthermore, our results showed that the initial population distribution can be relevant for the algorithm performance.

Key words: ligand-protein docking, flexible docking, genetic algorithms.

Received: September 22, 2003; Accepted: May 12, 2004.

Introduction

With the increasing amount of molecular biological structures available, docking approaches have been very important and useful tools in structure-based rational drug discovery and design (Gane and Dean, 2000). For a protein/receptor with known three-dimensional structure, the ligand-protein docking problem basically consists in predicting the bound conformation of a ligand molecule within the protein active site. The docking problem is a difficult optimization problem involving many degrees of freedom, and the development of efficient docking algorithms and methodologies would be of enormous benefit in the design of new drugs (Marrone *et al.*, 1997). One of the major problems in molecular docking is how to treat the protein and the ligand flexibility, taking into account hundreds of thousands of degrees of freedom in the two molecules. In the last few years several docking programs have been developed (Diller and Verlinde, 1999; McConkey *et al.*, 2002). Some docking programs treat the receptor and the ligand as rigid body molecules considering only the ligand translational and orientational degrees of freedom (Ewing and Kuntz, 1997). Other docking algorithms also include the ligand flexibility and account for the ligand conformational degrees of freedom (Jones *et al.*, 1997; Rarey *et al.*, 1996). In the two docking classes above, the protein structure is fixed in the position of the experimental crystallographic structure. Docking large, highly flexible ligands is still a challenge for even the most sophisticated current docking algorithms (Wang *et al.*, 1999), and adding the receptor flexibility remains a major challenge (Carlson and McCammon, 2000).

Genetic Algorithms are inspired in Darwin’s theory of evolution by natural selection and are powerful tools in difficult search and optimization problems (Holland, 1975; Goldberg, 1989).

They have been shown to be a promising search algorithm for the ligand-protein docking problems (Morris *et al.*, 1998). The GA works with a population of individuals where each individual represents a possible solution for the problem to be solved and, in ligand-protein docking problem, it is the position of the ligand with respect to the protein. Therefore, a ligand conformation is represented by a chromosome constituted by real valued genes representing

Send correspondence to Camila Silva de Magalhães. Laboratório Nacional de Computação Científica, Departamento de Matemática Aplicada e Computacional, Av. Getúlio Vargas 333, sala 1A-24, 25651-075 Quitandinha, Petrópolis, RJ, Brazil. E-mail: camilasm@lncc.br.

ligand translational, orientational and conformational degrees of freedom. The individuals are evaluated by a fitness function, that is, the total interaction energy between the protein and the ligand molecule and the intramolecular ligand energy. Individuals in the population are selected for reproduction in accordance with their fitness, and undergo mutation and crossover reproduction operators, to generate new individuals. In this paper, a non-generational also referred to as steady-state GA (Whitley, 1995) is adopted. In a steady-state GA (SSGA) there is no separation between consecutive generations of the population. Instead, each offspring is created and immediately tested for insertion in the population. In the following, the term generation will be associated with the creation of a single offspring (candidate solution) and its evaluation. The variable maxgen will thus denote the maximum number of objective function evaluations (which is equal to the total number of offspring generated). A pseudo-code for the steady-state GA used here is displayed as follows:

```
Begin
   Initialize the population $P$
   Evaluate individuals in $P$
   Sort $P$ according to the fitness value
   Repeat
    select genetic operator
    select individual(s) for reproduction
    apply genetic operator
    evaluate offspring
    select individual $\mathbf{x}^i$ to survive
    if $\mathbf{x}^i$ is better than worst individual in $P$ then
         remove worst individual from $P$
         insert $\mathbf{x}^i$ in $P$ according to its rank
    endif
   until stopping criteria are met
End
```

The SSGA differs from traditional GA basically by applying only one operator and replacing only one individual in each generation. In this work, we are interested in testing the use of a SSGA using a grid-based methodology in the rigid and flexible ligand docking cases. The algorithm performance is tested in five HIV-1 protease-ligand complexes with known three-dimensional structures. In all five tested complexes the receptor structure is assumed to be rigid. All ligands tested are highly flexible, having more than 10 conformational degrees of freedom.

Methods

In the implemented SSGA the individual chromosome has three genes representing the ligand translation, four genes representing the ligand orientation and the other genes representing the ligand conformation. The translational genes are the X, Y, Z reference atom coordinates (usually the closest atom to the ligand center of mass), the orientational genes are a quaternion (Maillot, 1990) constituted by a unit vector and one orientational angle. The conformational genes are the ligand dihedral angles (one gene to each dihedral angle). The ligand-protein energy function used is the GROMOS96 (van Gunsteren and Berendsen, 1987; Smith *et al.*, 1995) classical force field implemented in the THOR (Pascutti *et al.*, 1999) program of molecular mechanics/dynamics. The force field parameters are adjusted to reproduce experimental results (*e.g.*, structural and thermodynamic properties) or higher level *ab initio* quantum calculations (Brooks III *et al.*, 1988). The GROMOS force field is given by:

$$\sum_{\text{Protein}}\sum_{\text{Ligand}}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}+\frac{q_iq_j}{Dr_{ij}}\right)+\sum_{\text{Ligand}}\sum_{\text{Ligand}}\left(\frac{A_{ij}}{r_{ij}^{12}}-\frac{B_{ij}}{r_{ij}^{6}}+\frac{q_iq_j}{Dr_{ij}}\right)+$$
$$\sum_{\substack{\text{Dihedral}\\\text{Angles}}}\gamma_k(1+\cos(\varpi_k\theta_k-\theta_{0k}))$$

where r_{ij} is the distance between the atoms i and j; A_{ij} and B_{ij} are the Lennard-Jones parameters; q_i and q_j are atomic charges and D is a sigmoidal distance-dependent dielectric constant (Arora and Jayaram, 1997).

The first term of the equation corresponds to van der Waals interaction and electrostatic interaction between the protein and the ligand molecule, and the last two terms correspond to the ligand internal energy interaction, which also have one term for van der Waals interaction and one term for electrostatic interaction. The ligand-protein docking problem involves millions of energy evaluations, and the computational cost of each energy evaluation increases with the number of the atoms of the complex ligand-protein which has thousands of atoms. To reduce the computational cost, we implemented a grid-based methodology where the protein active site is embedded in a 3D rectangular grid and on each point of the grid the electrostatic interaction energy and the van der Waals terms for each ligand atom type are pre-computed and stored, taking into account all the protein atoms. In this way the protein contribution at a given point is obtained by tri-linear interpolation in each grid cell. A random initial population of individuals is generated inside the grid. For translational genes, random values between the maximum and minimum grid sizes are generated. For flexible docking, we also generated the initial population using a Cauchy distribution. The individual translational genes are generated by adding a random perturbation (drawn from a Cauchy distribution) to the grid center coordinates. In this way individuals are generated with higher probability near the grid center, while still permitting that individuals be generated far from the center. The Cauchy distribution is given by:

$$C(\alpha,\beta,x)=\frac{\beta}{\pi(\beta^2+(x-\alpha^2))}$$
$$\alpha\geq 0,\ \beta>0,\ -\infty<x<\infty$$

where α and β are Cauchy distribution parameters. In this work we used $\alpha = 0$ and $\beta = 0.75$. For genes corresponding to angles (dihedrals and/or orientationals), random values ranging from 0° to 360° are generated. Finally, for the genes corresponding to the orientational unit vector, random values between -1 and 1 are used. The individuals are evaluated, and then are selected to suffer recombination or mutation. A rank-based selection scheme (Whitley, 1995) was used. A new individual is inserted in the population if its fitness is better than the fitness of the worst individual in the population. The algorithm evolves until the maximum number of the energy evaluations is reached. The reproduction operators used are classical two-point crossover and non-uniform mutation operators (Michalewicz, 1992). The non-uniform mutation operator, when applied to an individual i at generation ngen, mutates a randomly chosen variable c_i according to the following:

$$c_i^{new} = \begin{matrix} c_i + \Delta(ngen, b_i - c_i), \text{ if } \tau = 0 \\ c_i + \Delta(ngen, c_i - a_i), \text{ if } \tau = 1 \end{matrix}$$

$$c_i \in (a_i, b_i),\ \Delta(ngen, y) = y(1 - r^{\left(1-\frac{ngen}{max\ gen}\right)^b})$$

where a_i and b_i are respectively the lower and upper bounds for the variable c_i, τ is randomly chosen as 0 or 1, r is randomly chosen in [0,1] and the parameter b set to 5. In the flexible docking, initially one randomly decides if a conformational gene will be mutated or not. Then a gene in the chosen group (conformational or not) is randomly selected for mutation. In this way, the seven translational/orientational genes have the same probability of being mutated as the conformational ones.

Results

We tested the algorithm with five HIV-1 protease-ligand complexes where the structures were obtained from the Protein Data Bank (PDB ID 1bve, 1hsg, 1ohr, 1hxw, 1hxb). The ligands tested are shown in Figure 1. The ligands tested have conformational degrees of freedom ranging from 12 to 20 dihedral angles. The DMP323 ligand in the HIV-1 protease active site is shown in Figure 2. The grid was centered in the protein active site and we used a grid dimension of 23 Å in each direction and a grid spacing of 0.25 Å. The algorithm success is measured by the RMSD (root mean square deviation) between the crystallographic conformation (from the corresponding PDB file) and the conformation found by the algorithm. A structure with a RMSD less than 2 Å is classified as docked and it is consid-

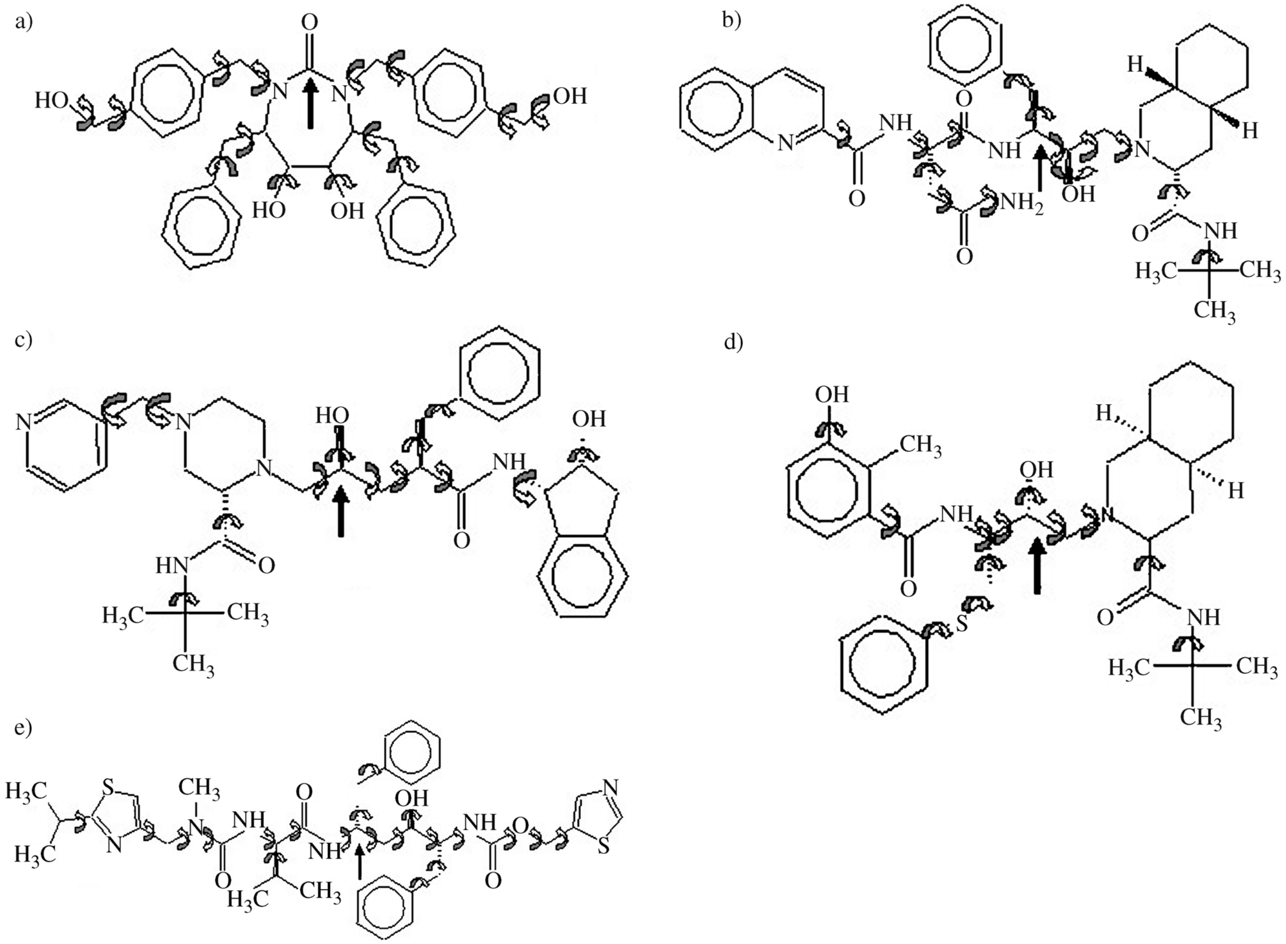

Figure 1 - HIV-1 protease ligands: (a) DMP323, (b) Saquinavir, (c) Indinavir, (d) Nelfinavir and (e) Ritonavir. The ligands' dihedral angles are shown by curved arrows. The right arrows show the ligands' reference atom. The more internal dihedral angles are the neighbors' angles to the reference atom.

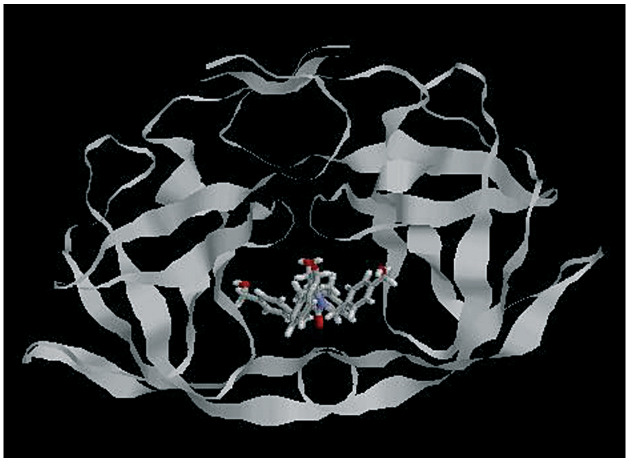

Figure 2 - The DMP323 ligand in the HIV-1 protease active site.

ered a very good result. A structure with a RMSD less than 3 Å is classified as partially docked. The success rate is the number of conformations found with RMSD less than 2 Å in 10 runs.

In rigid docking tests, we fixed the ligand dihedral angles in the position of the crystallographic structure for all ligands, and only translational and orientational movements are applied to the molecule. The individual chromosome has only the translational and orientational genes, and the last two terms are not evaluated for the energy function. We use a population of 500 individuals, 200,000 energy evaluations, and probability of 0.3 for two-point crossover and 0.7 for non-uniform mutation. The results are shown in Table 1.

In flexible docking tests, all terms of the energy are considered. We use a population of 1,000 individuals, 1,000,000 energy evaluations, and probability of 0.3 for two-point crossover and 0.7 for non-uniform mutation. We first tested flexible docking for DMP323 ligand with 10 and then with 14 dihedral angles (Table 2). The results for DMP323 flexible docking with and without the Cauchy distribution are shown in Table 2. For all other ligands, we used the same parameters together with the Cauchy distribution. The results are shown in Table 3. We also fixed two (three for the Ritonavir ligand) more internal dihedral angles (Figure 1). The results are shown in Table 4.

Discussion

In the rigid docking analyses, satisfactory results were found. For all ligands tested the mean RMSD ranged from 0.046 Å to 0.099 Å. This is considered a very good result in docking problems. The SSGA was able to find the corresponding crystallographic conformation in all 10 runs for all ligands tested, with a success rate of 100%.

In the DMP323 flexible docking analyses, we can see that the inclusion of only four additional dihedral angles (Table 2) can interfere directly in the algorithm performance, decreasing the success rate from 100% to 30%, and increasing the mean RMSD from 0.373 Å to 6.812 Å. However, with the use of the Cauchy distribution in the initial population the success rate returned to 100% and with a

Table 1 - Rigid docking results.

Ligands	DMP323	Nelfinavir	Ritonavir	Indinavir	Saquinavir
Lowest rmsd	0.038	0.066	0.090	0.037	0.058
Mean rmsd[3]	0.046 (0.012)	0.068 (0.002)	0.099 (0.004)	0.053 (0.009)	0.077 (0.011)
Energy of lowest rmsd	-58.37	-82.18	-100.18	-87.08	-86.78
Mean energy[3]	-58.36 (0.028)	-82.17 (0.005)	-100.20 (0.011)	-87.08 (0.002)	-86.78 (0.027)
Success ratio[4] (%)	100	100	100	100	100

[1]Energy (kcal/mol) and rmsd (Å); [2]The parameters used were 10 runs, 500 individuals, 200,000 energy evaluations, two-point crossover (prob. = 0.3) and non-uniform mutation (prob. = 0.7); [3]Mean in 10 runs; [4]Percent of conformations found by the algorithm with rmsd < 2 Å. Standard deviations are given in parentheses.

Table 2 - DMP323 flexible docking results.

Initial population distribution	without Cauchy		with Cauchy
Dihedral angles considered	10	14	14
Lowest rmsd	0.290	0.619	0.420
Energy of lowest rmsd	-31.77	-32.91	-33.08
Mean rmsd[3]	0.373 (0.117)	6.812 (4.072)	0.596 (0.268)
Mean energy[3]	-31.72 (0.111)	16.71 (81.974)	-32.77 (0.907)
Success ratio[4] (%)	100	30	100

[1]Energy (kcal/mol) and rmsd (Å); [2]10 runs, 1,000 individuals, 1.0 x 10^6 energy evaluations, two-point crossover (prob. = 0.3) and non-uniform mutation (prob. = 0.7); [3]Mean in 10 runs; [4]Percent of conformations found by the algorithm with rmsd < 2 Å. Standard deviations are given in parentheses.

mean RMSD of 0.596 Å, with only 1,000,000 energy evaluations. This is a very good result considering that all 14 dihedral angles are being considered, and that current docking programs use about 1,500,000 energy evaluations even in ligands with less conformational degrees of freedom (Morris *et al.*, 1998). For all ligands tested the SSGA was able to find the corresponding crystallographic structure with RMSD less than 2 Å at least once in 10 runs. We obtained a mean RMSD ranging from 3.585 Å to 5.755 Å and a success rate ranging from 10% to 30% in finding docked structures, and 10% to 60% in finding partially docked structures (Table 3). When we fixed two (three for the Ritonavir ligand) more internal dihedral angles (Figure 1) we found better results (Table 4). We obtained a mean RMSD ranging from 1.449 Å to 3.733 Å and a success rate ranging from 20% to 90% in docked structures, and 50% to 90% in partially docked structures, with 10 to 17 ligand dihedral angles. The superior performance of DMP323, when compared to the others ligands, may be due to a minor dependence among its dihedral angles and to the fact that its correct conformation is placed in the center of the protein active site; that is, privileged by using a Cauchy distribution to generate the initial population. The other ligands have a more "open" geometry with larger arms and consequently there is a major dependence among the dihedral angles. In this sense we observed (see Table 4) that the more internal dihedral angles are critical. This seems to be due to the fact that small variations in internal dihedral angles may cause larger motions in the molecule than variations in the other more external dihedral angles.

The results obtained show the difficulty in dealing with highly flexible ligands, *i.e.*, containing many conformational degrees of freedom. Moreover, the enclosed active site of the HIV-1 protease is a considerable challenge for a docking program (Gehlhaar *et al.*, 1995). The EPDOCK program had a success rate of 34% in finding the corresponding crystallographic structure of the AG-1343 HIV-1 protease ligand, with nine conformational degrees of freedom (Gehlhaar *et al.*, 1995). Current docking programs present a decreasing performance with the increasing number of conformational degrees of freedom considered (McConkey *et al.*, 2002). The implemented SSGA demonstrated a good performance in docking rigid ligand molecules to molecular targets in a few minutes (using a Pentium III 800 MHZ), and may be used for screening compounds in large databases. The flexible docking methodology needs to be improved. This may be done by designing new problem-specific operators that take into

Table 3 - Flexible docking results using the Cauchy distribution.

Ligands	DMP323	Nelfinavir	Ritonavir	Indinavir	Saquinavir
Dihedral angles	14	12	20	14	15
Lowest rmsd	0.420	0.267	1.848	1.943	0.147
Energy of lowest rmsd	-33.08	-57.23	-78.09	2.61	-65.77
Mean rmsd[3]	0.596 (0.268)	4.185 (3.260)	4.237 (2.620)	5.755 (3.110)	3.585 (1.391)
Mean energy[3]	-32.77 (0.907)	-26.24 (21.648)	-41.43 (21.293)	36.60 (41.370)	-19.51 (22.493)
Success ratio[4] (%)	100	30	10	10	10
Success ratio (partially docked structures)[5] (%)	100	50	10	10	60

[1]Energy (kcal/mol) and rmsd (Å); [2]10 runs, 1,000 individuals, 1.0 x 10^6 energy evaluations, two-point crossover (prob. = 0.3) and non-uniform mutation (prob. = 0.7); [3]Mean in 10 runs; [4]Percent of conformations found by the algorithm with rmsd < 2 Å; [5]Percent of conformations found by the algorithm with rmsd < 3 Å. Standard deviations are given in parentheses.

Table 4 - Flexible docking results using the Cauchy distribution without the more internal dihedral angles.

Ligands	DMP323	Nelfinavir	Ritonavir	Indinavir	Saquinavir
Dihedral angles considered	14	10	17	12	13
Lowest rmsd	0.420	0.056	0.924	0.659	0.341
Energy of lowest rmsd	-33.08	-58.05	-101.46	-60.50	-64.11
Mean rmsd[3]	0.596 (0.268)	1.449 (1.752)	3.733 (2.309)	3.118 (1.036)	3.106 (1.419)
Mean energy[3]	-32.77 (0.907)	-53.48 (11.012)	-70.49 (15.008)	-17.22 (26.407)	-25.63 (17.992)
Success ratio[4] (%)	100	90	30	20	20
Success ratio (partially docked structures)[5] (%)	100	90	60	60	50

[1]Energy (kcal/mol) and rmsd (Å); [2]10 runs, 1,000 individuals, 1.0 x 10^6 energy evaluations, two-point crossover (prob. = 0.3) and non-uniform mutation (prob. = 0.7); [3]Mean in 10 runs; [4]Percent of conformations found by the algorithm with rmsd < 2 Å; [5]Percent of conformations found by the algorithm with rmsd < 3 Å. Standard deviations are given in parentheses.

account critical factors of the problem such as the motion of more internal dihedral angles. The use of a Cauchy distribution in the initial population improved the algorithm performance in all cases, but only obtained a very good result with the DMP323 ligand. With the other ligands the improvement was not very significant, requiring the development of better docking strategies (Magalhães *et al.*, 2004).

Acknowledgments

This work was supported by CNPq (grant number 40.2003/2003.9) and FAPERJ (grant number E26/171.401/01). The authors would like to thank the Carcará Project at LNCC for computational resources.

References

Arora N and Jayaram B (1997) Strength of hydrogen bonds in σ helices. J Comp Chem18:1245-1252.

Brooks III CL, Karplus M and Pettitt BM (1988) Proteins: A theoretical perspective of dynamics, structure, and thermodynamics. Advances in Chemical Physics Vol LXXI. John Wiley & Sons, New York.

Carlson HA and McCammon JA (2000) Accommodating protein flexibility in computational drug design. Molecular Pharmacology 57:213-218.

Diller DJ and Verlinde CLMJ (1999) A critical evaluation of several global optimization algorithms for the purpose of molecular docking. J Comp Chem 20:1740-1751.

Ewing TJA and Kuntz ID (1997) Critical evaluation of search algorithms for automated molecular docking and database screening. J Comp Chem 18:1175-1189.

Gane PJ and Dean PM (2000) Recent advances in structure-based rational drug design. Current Opinion in Structural Biology 10:401-404.

Gehlhaar DK, Verkhivker G, Rejto PA, Fogel DB, Fogel LJ and Freer ST (1995) Docking conformationally flexible small molecules into a protein binding site through evolutionary programming. In: Proceedings of the Fourth Annual Conference on Evolutionary Programming, MIT Press, Cambridge.

Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, New York.

Holland JH (1975) Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI.

Jones G, Willett P, Glen RC, Leach AR and Taylor R (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267:727-748.

Magalhães CS, Barbosa HJC and Dardenne LE (2004) Selection-insertion schemes in genetic algorithms for the flexible ligand docking problem. Lecture Notes in Computer Science, Springer Verlag, Berlim, 3102:368-379.

Maillot PG (1990) In Graphics Gems. Glassner AS (ed), Academic Press, London, pp 498.

Marrone TJ, Briggs JM and McCammon (1997) Structure-based drug design. Annu Rev Pharmacol Toxicol 37:71-90.

McConkey BJ, Sobolev V and Edelman M (2002) The performance of current methods in ligand-protein docking. Current Science 83:845-855.

Michalewicz Z (1992) Genetic Algorithms + Data Structures = Evolution Programs. Springer-Verlag, New York, pp 87-88.

Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK and Olson AJ (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem 19:1639-1662.

Pascutti PG, Mundim KC, Ito AS and Bisch PM (1999) Polarization effects on peptide conformation at water-membrane interface by molecular dynamics simulation. J Comp Chem 20:971-982.

Rarey M, Kramer B, Lengauer T and Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470-489.

Smith LJ, Mark AE, Dobson CM and van Gunsteren WF (1995) Comparison of MD simulations and NMR experiments for hen lysozyme. Analysis of local fluctuations, cooperative motions, and global changes. Biochemistry 34:10918-10931.

Wang J, Kollman PA and Kuntz ID (1999) Flexible ligand docking: A multistep strategy approach. Proteins 36:1-19.

Whitley D, Beveridge R, Graves C and Mathias K (1995) Test driving three genetic algorithms: New test functions and geometric matching. Journal of Heuristics 1:77-104.

van Gunsteren WF and Berendsen HJC (1987) Groningen Molecular Simulation (GROMOS) Library Manual. Biomos, Groningen.