JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 7, 2016

Pp. 585–596
DOI: 10.1089/cmb.2015.0190

EDGA: A Population Evolution Direction-Guided Genetic Algorithm for Protein–Ligand Docking

BOXIN GUAN, CHANGSHENG ZHANG, and JIAXU NING

ABSTRACT

Protein–ligand docking can be formulated as a search algorithm associated with an accurate scoring function. However, most current search algorithms cannot show good performance in docking problems, especially for highly flexible docking. To overcome this drawback, this article presents a novel and robust optimization algorithm (EDGA) based on the Lamarckian genetic algorithm (LGA) for solving flexible protein–ligand docking problems. This method applies a population evolution direction-guided model of genetics, in which search direction evolves to the optimum solution. The method is more efficient to find the lowest energy of protein–ligand docking. We consider four search methods—a tradition genetic algorithm, LGA, SODOCK, and EDGA—and compare their performance in docking of six protein–ligand docking problems. The results show that EDGA is the most stable, reliable, and successful.

Key words: automated docking, drug design, evolutionary direction, genetic algorithm, protein–ligand docking.

1. INTRODUCTION

PROTEIN–LIGAND DOCKING is a typical problem for computer-aided drug discovery and drug design (Brooijmans and Kuntz, 2003; Moitessier et al., 2008; Huang and Zou, 2010; Jug et al., 2015). The aim of the problem is to identify the best ligand conformation and orientation relative to the active site of a target protein with the lowest energy. An efficient docking consists of a good scoring function and an efficient search algorithm.

The scoring function is a free energy of binding interaction between protein and ligands. Scoring function can help a docking to efficiently explore the binding space of a ligand. It is also responsible for evaluating the binding affinity once the correct binding pose is identified (Bharatham et al., 2014; Li et al., 2015).

The search algorithm for solving the docking problem of flexible docking aims to identify the docked conformation with the lowest energy. Some metaheuristics (Blum et al., 2011; López-Camacho et al., 2015) have been successfully applied in the docking problem, and many researchers have tried improving the optimization algorithm of protein–ligand docking. For instance, simulated annealing (SA) (Goodsell and Olson, 1990) is a generic probabilistic metaheuristic for the global optimization problem of locating a good

College of Information Science & Engineering, Northeastern University, Shenyang, People's Republic of China.

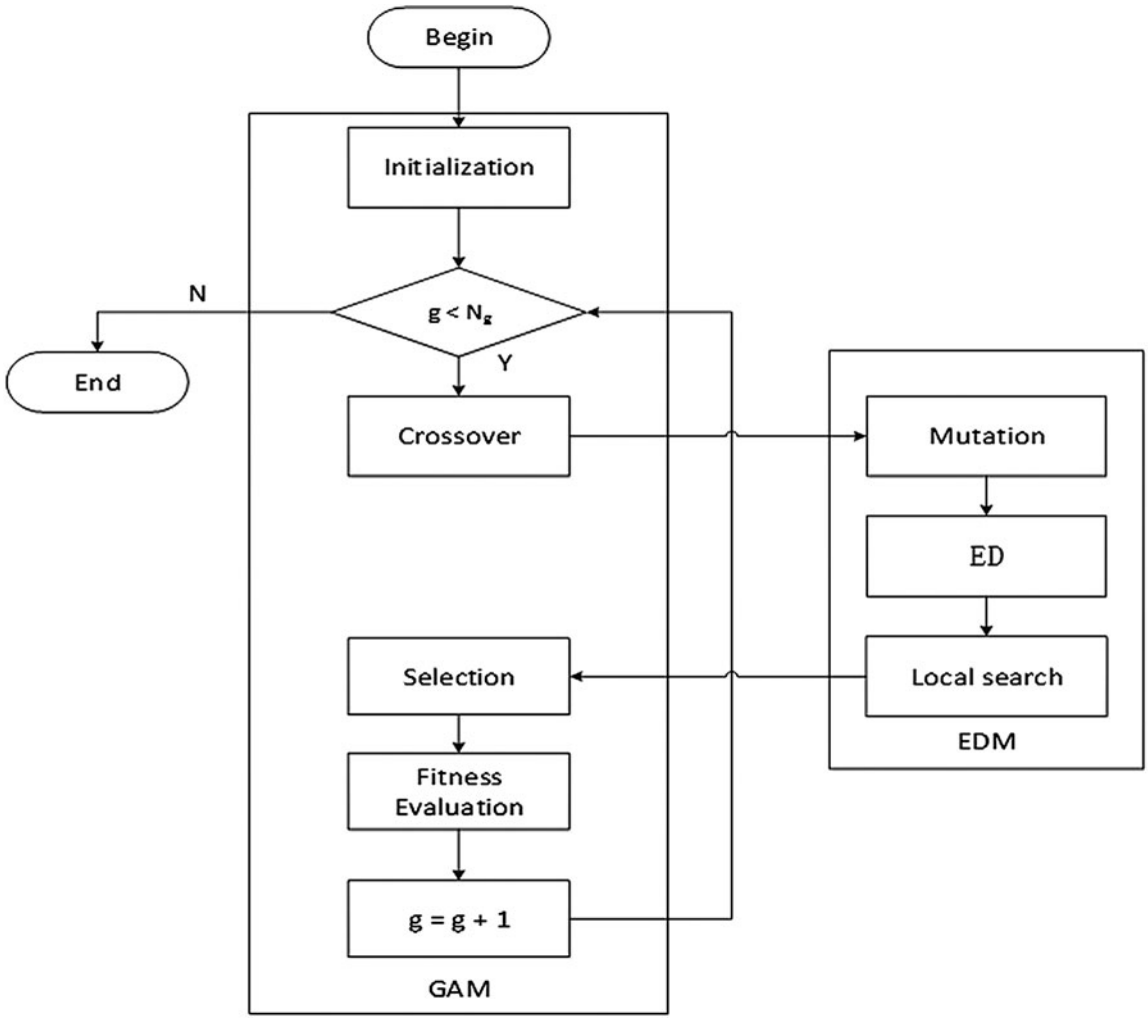

FIG. 1. Block diagram of EDGA.

approximation to the global optimum of a given function in a large search space. Genetic algorithm (GA) (Jones et al., 1997; Morris et al., 1998; Thomsen, 2003; Cao and Li, 2004) is a method to search the optimal solution by simulating the natural evolution process. Lamarckian genetic algorithm (LGA) (Fuhrmann et al., 2010) is a hybrid of GA and the local search, and it is more successful than SA and GA for the docking problem. SODOCK (Chen et al., 2007; Jason et al., 2008) based on particle swarm optimization (PSO) integrates with a local search, and it is designed for flexible docking.

There are two criteria used to verify the performance of different optimization algorithms: fitness accuracy (energy based) and pose accuracy (root-mean-square deviation [RMSD] based) (Guo et al., 2014). For fitness accuracy, the lower binding energy is the greater binding activity that can also provide better

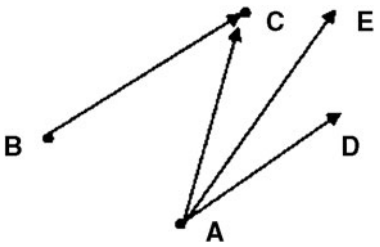

FIG. 2. Schematic diagram of GE.

Table 1. Results of GA Energy (Kal Mol^{-1}) and rmsd (Å)

PDB code	*Lowest energy*	*Lowest energy*	*Mean energy*	*Mean rmsd*	*Number of energy evaluations*
3ptb	−11.26	2.86	−8.81	7.57	1.50×10^6
2mcp	−7.76	1.46	−4.36	5.50	1.50×10^6
1stp	−11.03	2.84	−7.33	5.09	1.50×10^6
1hvr	−31.28	4.28	−18.72	4.28	1.50×10^6
4hmg	−8.44	1.69	−6.63	2.96	1.50×10^6
4dfr	−10.27	3.49	−5.98	4.79	1.50×10^6

PDB, Protein Data Bank; rmsd, root-mean-square positional deviation.

Table 2. Results of LGA Energy (Kal Mol^{-1}) and rmsd (Å)

PDB code	*Lowest energy*	*Lowest rmsd*	*Mean energy*	*Mean rmsd*	*Number of energy evaluations*
3ptb	−11.56	2.02	−10.81	3.90	1.50×10^6
2mcp	−8.22	1.33	−8.05	1.31	1.50×10^6
1stp	−13.41	2.55	−12.66	2.14	1.50×10^6
1hvr	−30.85	0.62	−16.64	4.94	1.50×10^6
4hmg	−10.09	1.70	−8.94	2.97	1.50×10^6
4dfr	−11.50	4.79	−9.66	4.77	1.50×10^6

drug efficiency. RMSD is used to determine if two docked conformations are similar enough to be included in the same cluster. A docked conformation with a smaller RMSD is considered as a more accurate solution to the docking problem. Based on these two standards, the existing algorithms are proven to have obvious shortcomings. Therefore, an efficient optimization algorithm that can find lower docking energy and RMSD is desirable.

The novel algorithm is improved on the basis of LGA. LGA is proven to be efficient, but its search direction is blind random. Because function evaluation of solving the lowest energy is the most computational-intensive process in the search algorithms, such randomness is clearly a waste of computational resources. Furthermore, LGA has no effective use of feedback information, and so its search speed is slow and it cannot obtain a more accurate solution. To improve LGA, a modified LGA that has evolutionary direction is presented in this article to solve the aforementioned docking problem.

The implementation of EDGA adopts the environment and scoring function of AutoDock 4.2.6 (Morris et al., 1996, 2009; Kitchen et al., 2004). AutoDock is the most widely used automated docking program. To

Table 3. Results of EDGA Energy (Kal Mol^{-1}) and rmsd (Å)

PDB code	*Lowest energy*	*Lowest energy*	*Mean energy*	*Mean rmsd*	*Number of energy evaluations*
3ptb	−12.18	1.95	−11.80	2.12	1.50×10^6
2mcp	−9.25	1.23	−8.65	1.33	1.50×10^6
1stp	−13.67	1.25	−13.21	1.93	1.50×10^6
1hvr	−28.10	0.75	−16.50	3.89	1.50×10^6
4hmg	−10.47	4.58	−9.15	3.32	1.50×10^6
4dfr	−12.74	5.91	−10.01	4.97	1.50×10^6

Table 4. Results of SODOCK Energy (Kal Mol^{-1}) and rmsd (Å)

PDB code	*Lowest energy*	*Lowest rmsd*	*Mean energy*	*Mean rmsd*	*Number of energy evaluations*
3ptb	−11.57	2.00	−10.74	3.95	1.50×10^6
2mcp	−7.72	1.42	−5.98	3.30	1.50×10^6
1stp	−13.52	1.00	−11.41	2.45	1.50×10^6
1hvr	−30.01	0.68	−24.59	2.42	1.50×10^6
4hmg	−10.08	1.36	−8.85	3.02	1.50×10^6
4dfr	−11.50	3.60	−10.14	5.25	1.50×10^6

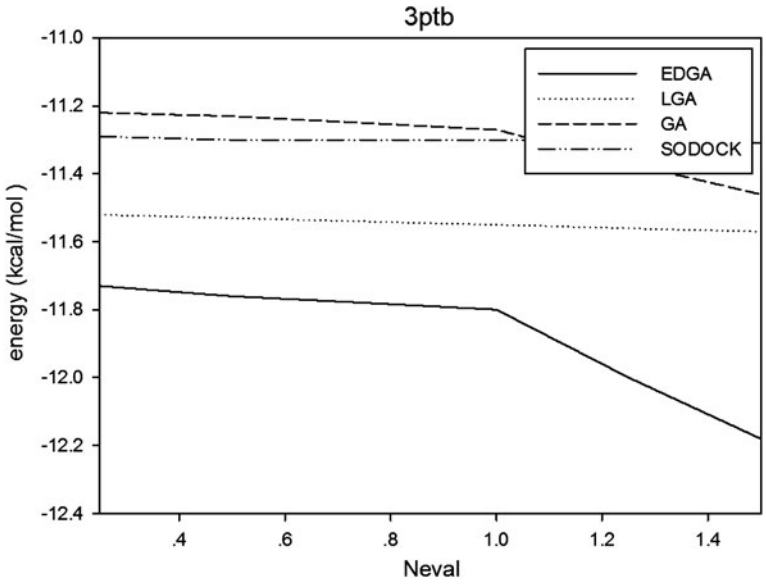

FIG. 3. Convergence diagram of 3ptb.

evaluate the method, we perform six protein–ligand docking problems from the Brookhaven Protein Data Bank (PDB) (Berman et al., 2000, 2002). In this article, we compared the performance SA, GA, SODOCK, and EDGA. Computer simulation results reveal that EDGA is superior to the other methods in terms of convergence performance, robustness, and obtained energy, especially for highly flexible ligands. Simulation results also reveal that EDGA can yield more accurate results than the other methods in terms of RMSD.

2. SCORING FUNCTION

AutoDock 4.2.6 uses a semiempirical free-energy force field to evaluate a docked conformation. The force field includes six pair-wise evaluations (V) and an estimate of the conformational entropy lost upon binding (ΔS_{conf}):

$$\Delta \boldsymbol{G} = \left(\boldsymbol{V}^{L\_L}_{bound} - \boldsymbol{V}^{L\_L}_{unbound}\right) + \left(\boldsymbol{V}^{P-P}_{bound} - \boldsymbol{V}^{P-P}_{unbound}\right) + \left(\boldsymbol{V}^{P\_L}_{bound} - \boldsymbol{V}^{P\_L}_{unbound} + \Delta \boldsymbol{S}_{conf}\right)$$

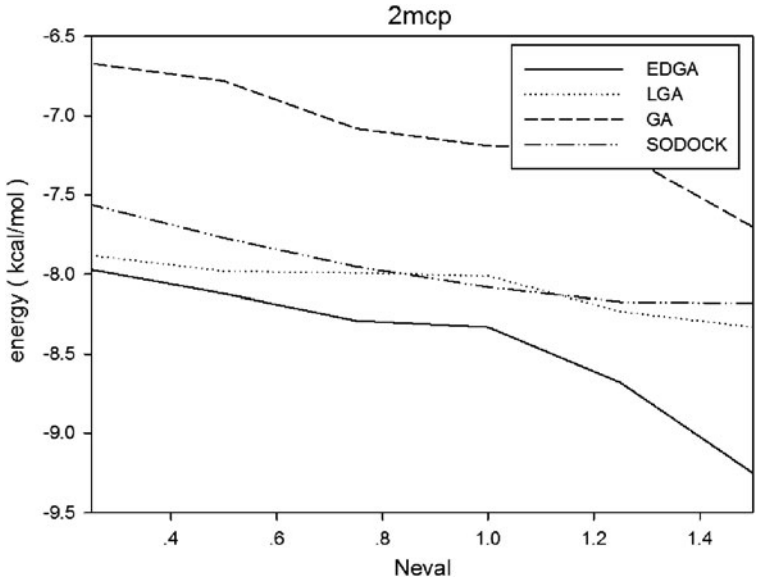

FIG. 4. Convergence diagram of 2mcp.

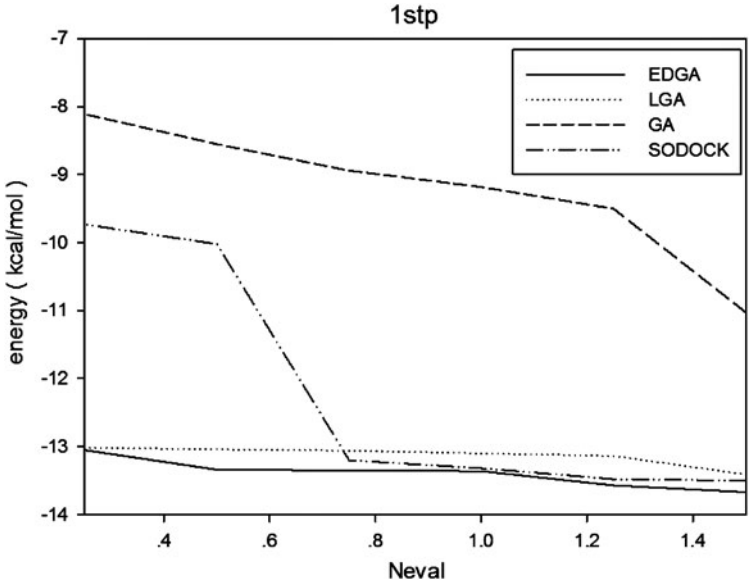

FIG. 5. Convergence diagram of 1stp.

where L represents the ligand and P represents the protein. Each of the pair-wise energetic terms is expressed as the sum of dispersion/repulsion in which the parameters are based on the Amber force field, hydrogen bonding, electrostatics, and desolvation.

$$V = W_{vdw}\sum_{i,j}\left(\frac{A_{ij}}{r_{ij}^{12}}\frac{B_{ij}}{r_{ij}^{6}}\right) + W_{hbond}\sum_{i,j}E(t)\left(\frac{C_{ij}}{r_{ij}^{12}}\frac{D_{ij}}{r_{ij}^{10}}\right) + W_{elec}\sum_{i,j}\frac{q_i q_j}{e(r_{ij})r_{ij}} + W_{sol}\sum_{i,j}\left(S_i V_j + S_j V_i\right)e^{\left(-r_{ij}^2/2\sigma^2\right)}$$

3. METHODS

3.1. Hybrid search of EDGA

In this article, we propose a novel algorithm to high protein–ligand docking. The new algorithm based on LGA integrates with a guided evolutionary direction mechanism so that solutions are in a better direction. The algorithm will be abbreviated as EDGA, which stands for a Population Evolution Direction-Guided Genetic Algorithm. EDGA is considered as a genetic algorithm module (GAM) cooperating with the

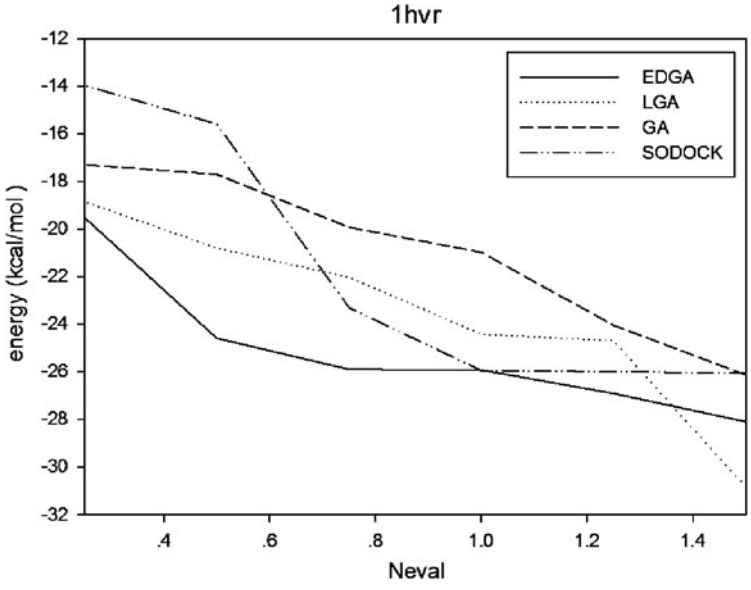

FIG. 6. Convergence diagram of 1hvr.

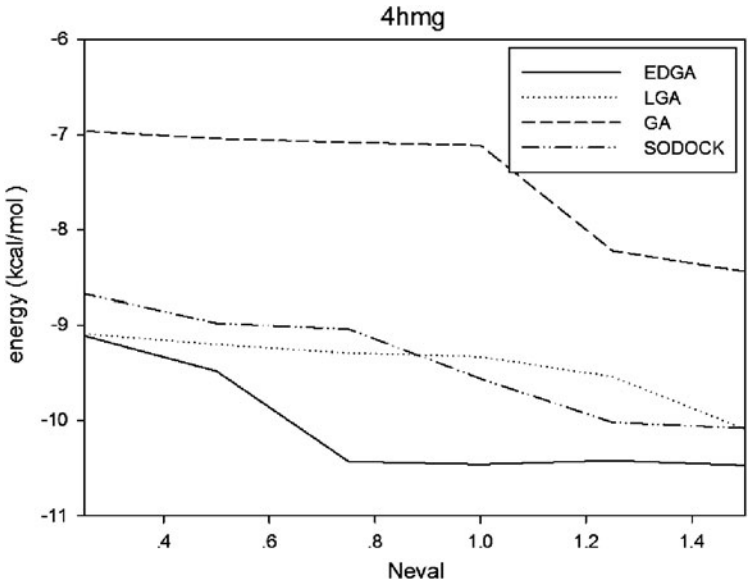

FIG. 7. Convergence diagram of 4hmg.

standard GA operators such as population initialization, crossover operator, and selection operator, while evolution direction module (EDM) consists of the mutation, the guided evolutionary direction, and applying the local search. Figure 1 shows the block diagram of EDGA.

EDGA starts with initializing the population. Afterward, the offspring population is generated by genetic operators such as crossover and mutation. The new population is then selected from the parent population and the offspring population. The reproduction process and the selection process are repeated until the number of iterations exceeds a predetermined value.

3.2. Evolutionary direction

In the original algorithm, the mutation is random and the search direction is blind random. In the early stage, the randomness plays a very good guide effect for the global search. The direction of the optimal solution is not known. With the development of the algorithm, the search experience is accumulated, the direction is gradually clear, the search space begins to converge, and the difference between each solution is

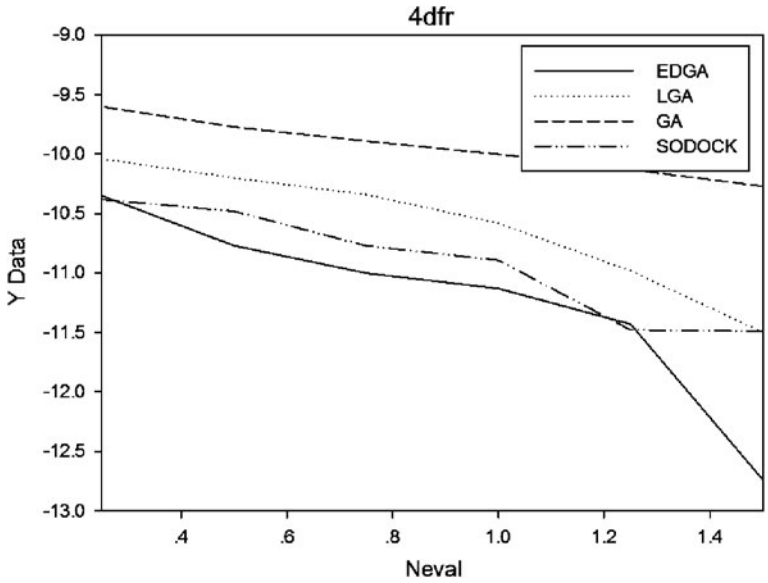

FIG. 8. Convergence diagram of 4dfr.

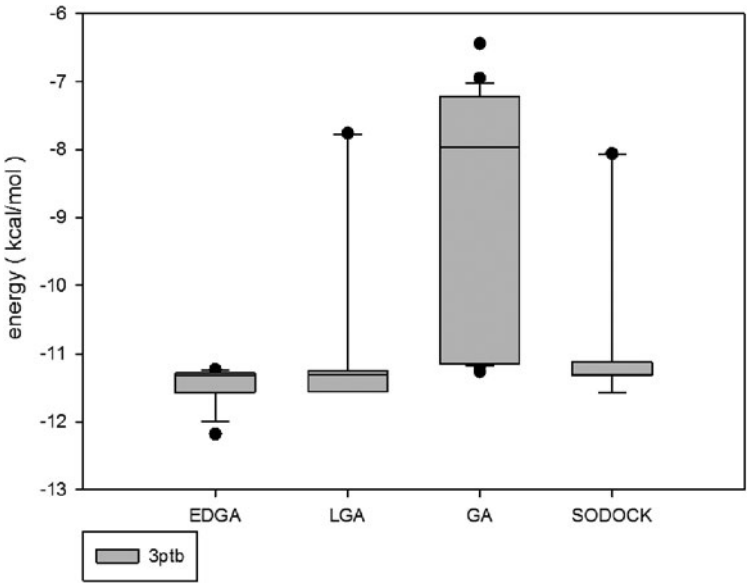

FIG. 9. Box plot of 3ptb.

getting smaller, which means inevitably that the abandoned solution is the current best—even *the* best. When the algorithm search space reduces to a very small range, if you continue to take the random pattern of the original algorithm, then it is possible to lead away from the search target. At this time, the search strategy needs to be adjusted. Therefore, a new mutation method is introduced.

In the method, an equilibrium factor β is introduced. When the random factor $\phi < \beta$, the random mutation is used; otherwise, the ''mixed mutation'' is used. This new mutation combines the information provided by the optimal solution and the suboptimal solution of the history.

In the improved algorithm, the mutation operator will be generated according to the following formula:

$$x_{ij} = \begin{cases} x_{\min} + \phi(x_{\max} - x_{\min}) & \text{if} |\phi| < \beta \\ x_{ij} + \phi\left(x_{optimum} - x_{sub}\right) + \delta\left(x_{optimum} - x_{ij}\right) & \text{otherwise} \end{cases}$$

where x_{ij} is the solution vector of the current value, ϕ is a random number between 0 and 1, β is a particular adjustable parameter (the range is 0 to 1), $x_{optimum}$ represents the solution vector of the historic optimal solution, and x_{sub} represents the solution vector of the historic suboptimal solution.

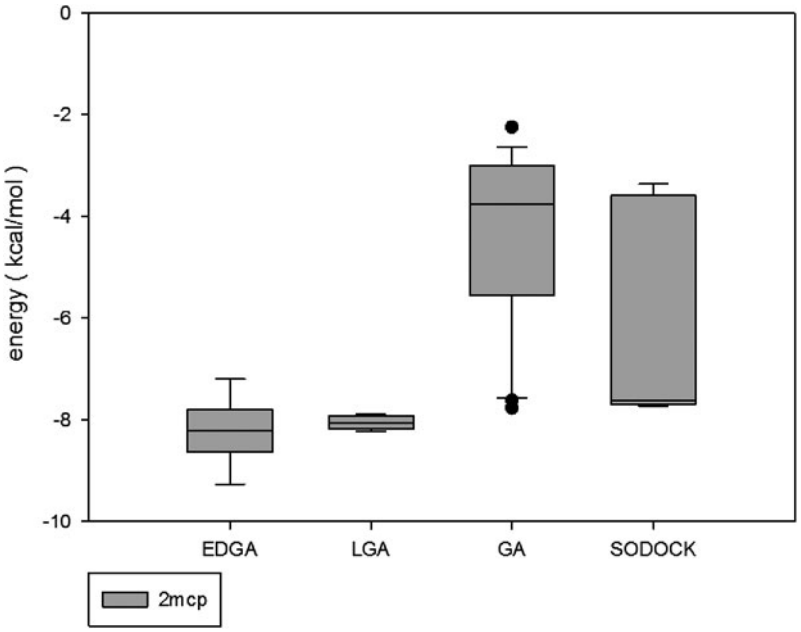

FIG. 10. Box plot of 2mcp.

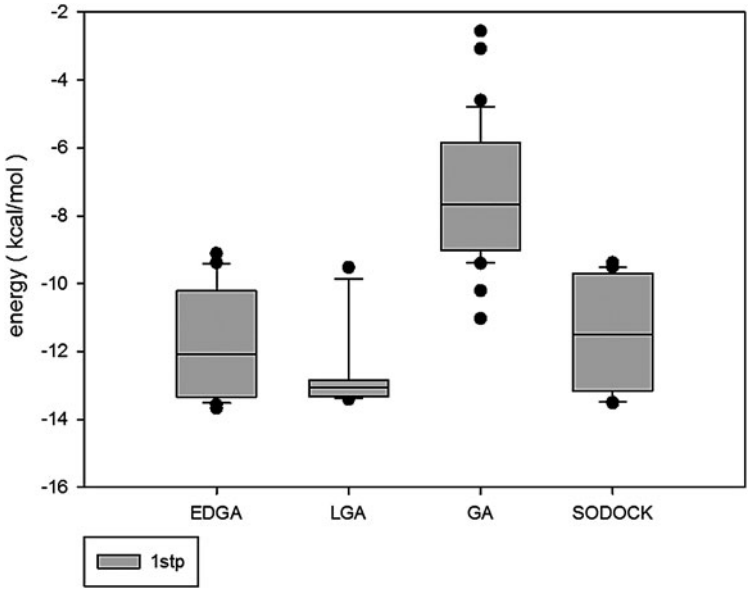

FIG. 11. Box plot of 1stp.

In the formula, the purpose of introducing the current solution and the second best solution is to keep the search direction to the optimal solution as far as possible. In principle, as shown in Figure 2, A is the current optimal solution, B is the historic suboptimal solution, and C is the historic optimal solution. Thus, ϕ (C − A) + ϕ (C − B) = ϕ AD + AC = AE, where ϕ is 0 to 1. As a result, the search direction will be closer to the AE direction because the historic optimum solution is in the vicinity of the direction, and so the possibility of finding the global optimal solution becomes larger. On the other hand, if the current optimal solution is the historic optimal solution, that is, when C and A coincide in the schematic diagram, the AE direction is coincident with the BC direction. The actual moving direction of the search is changed into AD. In this case, the optimal solution of the group is abandoned, but it can also guarantee the correctness of the direction.

4. EXPERIMENTS AND DISCUSSION

Six protein–ligand complexes (Hu et al., 2004) were chosen from the Brookhaven PDB (Berman et al., 2000, 2002) to compare the performance of the docking techniques. The six docking problems are summarized in the following:

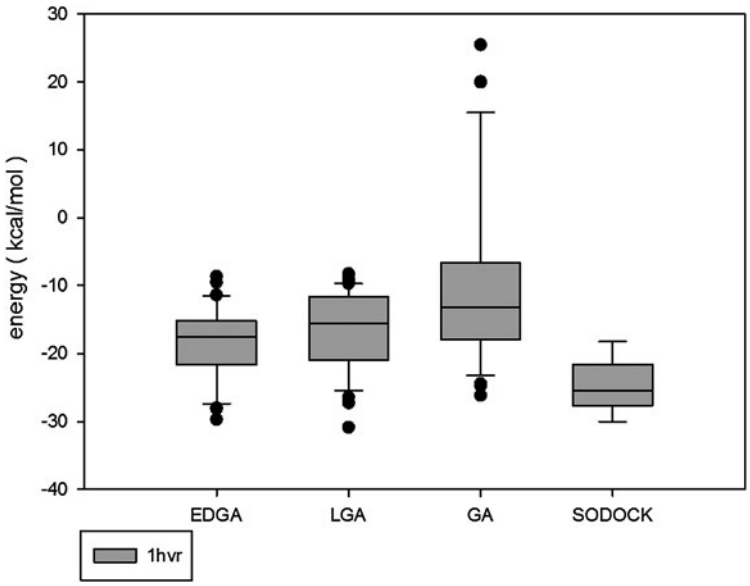

FIG. 12. Box plot of 1hvr.

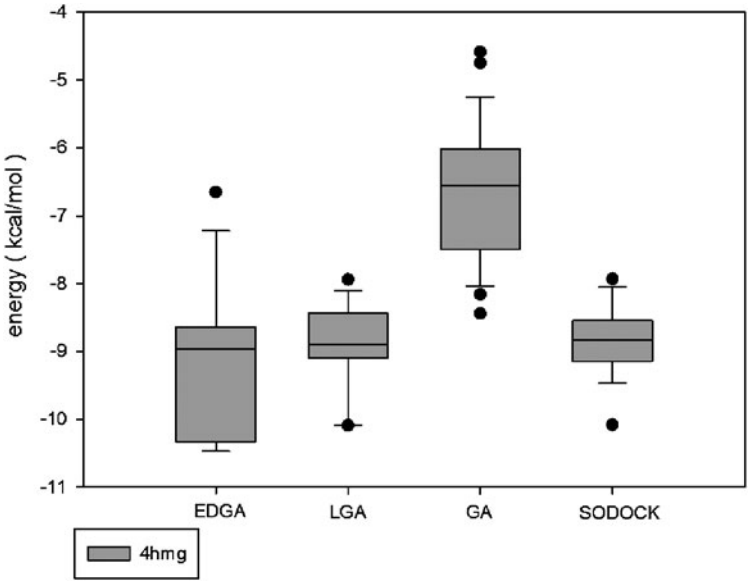

FIG. 13. Box plot of 4hmg.

(1) β-Trypsin/Benzamidine (3ptb)

Benzamidine is a reversible competitive inhibitor of trypsin, trypsin-like enzymes, and serine proteases. The recognition of benzamidine by β-trypsin is mainly because of the polar amidine moiety and the hydrophobic benzyl ring.

(2) McPC-603/Phosphocholine (2mcp)

Phosphocholine is an intermediate in the synthesis of phosphatidylcholine in tissues. The recognition of Phosphocholine by FabMcPC-603 is mainly because of the influence of ArgH52.

(3) Streptavidin/Biotin (1stp)

Biotin, also known as vitamin H or coenzyme R, is a water-soluble B vitamin. Streptavidin/biotin is one of the most tightly binding noncovalent complexes.

4) HIV-1 Protease/XK263 (1hvr)

The cyclic urea HIV-protease inhibitor, XK-263, has 10 rotatable bonds, excluding the cyclic urea's flexibility.

5) Influenza Hemagglutinin/Sialic Acid (4hmg)

The recognition of sialic acid by influenza hemagglutinin is chiefly because of hydrogen bonding.

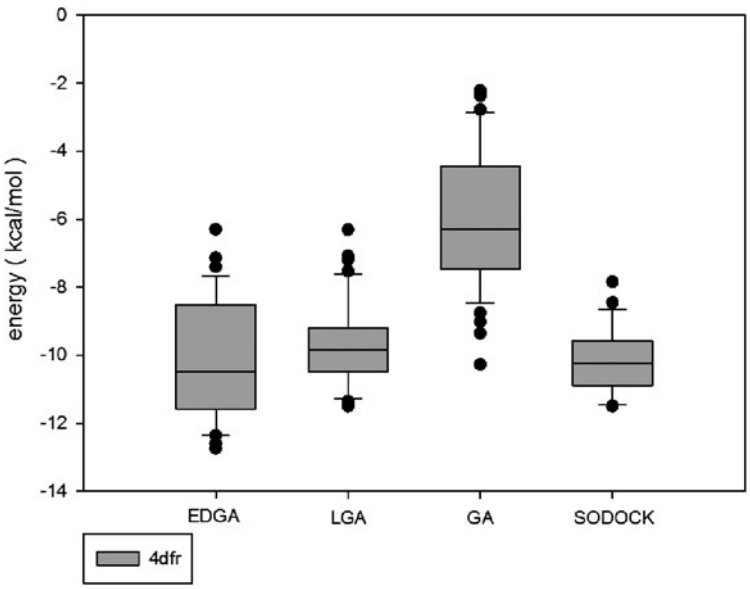

FIG. 14. Box plot of 4dfr.

Table 5. Hypothesis Test of 3ptb

	EDGA	LGA	GA	SODOCK
EDGA	—	0.991	0.996	0.953
LGA	0.009	—	0.997	0.500
GA	0.004	0.003	—	0.004
SODOCK	0.047	0.500	0.996	—

Table 6. Hypothesis Test of 2mcp

	EDGA	LGA	GA	SODOCK
EDGA	—	0.992	0.995	0.995
LGA	0.008	—	0.995	0.995
GA	0.005	0.005	—	0.232
SODOCK	0.005	0.005	0.768	—

Table 7. Hypothesis Test of 1stp

	EDGA	LGA	GA	SODOCK
EDGA	—	0.992	0.996	0.963
LGA	0.008	—	0.995	0.624
GA	0.004	0.005	—	0.004
SODOCK	0.037	0.376	0.996	—

Table 8. Hypothesis Test of 1hvr

	EDGA	LGA	GA	SODOCK
EDGA	—	0.962	0.995	0.992
LGA	0.038	—	0.913	0.377
GA	0.005	0.087	—	0.038
SODOCK	0.008	0.623	0.962	—

Table 9. Hypothesis Test of 4hmg

	EDGA	LGA	GA	SODOCK
EDGA	—	0.996	0.997	0.996
LGA	0.004	—	0.997	0.583
GA	0.003	0.003	—	0.003
SODOCK	0.004	0.417	0.997	—

Table 10. Hypothesis Test of 4dfr

	EDGA	LGA	GA	SODOCK
EDGA	—	0.995	0.995	0.996
LGA	0.005	—	0.995	0.664
GA	0.005	0.005	—	0.004
SODOCK	0.004	0.336	0.996	—

6) Dihydrofolate Reductase/Methotrexate (4dgr)

Methotrexate is an antimetabolite that attacks proliferating tissue and selectively induces remissions in certain acute leukemias.

We compared the performance of GA, LGA, SODOCK, and EDGA. The semiempirical free-energy force field presented above was used for energy evaluation in all cases. The main goal was to find the lowest energy in each docking problem. We also compared the root-mean-square positional deviation (rmsd) between the lowest energy docked structures. The rmsd tolerance was used to determine if two docked conformations were similar enough.

It is important to ensure that different search methods are treated equally. Therefore, in the three GAs, the population was 50, the number of generations was 27,000, and the energy evaluations was 1.5×10^6 in a docking. Therefore, the dockings were terminated by reaching the maximum number of generations. In SODOCK, the number of particle was 50, the number of immediate neighbors was 5, and the maximal number of function evaluation was 1.5×10^6. Five times were tested and each time consisted of 10 runs, and so there were 50 results.

The results of GA, LGA, EDGA, and SODOCK docking experiments are summarized in Tables 1–4, respectively. Through these tables, we concluded that EDGA found the lowest energy in five of the six protein–ligand docking problems. The number of the lowest energy corresponding to the lowest rmsd was 1, 1, 2, and 2 for GA, LGA, EDGA, and SODOCK, respectively. Thus, EDGA performed best in finding the lowest energy docked structure and the lowest rmsd. Furthermore, the number of the lowest mean energy was 0, 0, 4, and 2, using GA, LGA, EDGA, and SODOCK, respectively. However, the number of the lowest mean rmsd found by each of the four search methods was 1, 2, 2, and 1. In conclusion, considering their average performance, the best search method was EDGA.

Figures 3–8 are convergence diagrams, and these figures show that EDGA also had the best convergence performance among all methods. Figures 9–14 are box plots, and the median of EDGA was the lowest in five of the six docking problems and the data of EDGA were the most concentrated. Tables 5–10 are hypothesis tests. Through these tables, it can be seen that EDGA performed significantly better than LGA, GA, and SODOCK.

5. CONCLUSIONS

We have shown that, of the four search methods tested—GA, LGA, EDGA, and SODOCK—the most efficient, reliable, and successful is EDGA. We defined efficiency of search in terms of lowest energy found in a given number of energy evaluations, and reliability in terms of reproducibility of finding the lowest energy structure in independent dockings. The introduction of the EDGA search method extends the power and applicability of AutoDock to docking problems compared with the earlier versions of search methods.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation Program of China (61572116, 61572117, and 61502089), and the National Key Technology R&D Program of the Ministry of Science and Technology (2015BAH09F02).

AUTHOR DISCLOSURE STATEMENT

The authors declare that no competing financial interests exist.

REFERENCES

Berman, H., Battistuz, T., Bhat, T.N., et al. 2002. The protein data bank. *Acta Crystallogr. D Biol. Crystallogr.* 58, 899–907.

Berman, H.M., Westbrook, J., Feng, Z., et al. 2000. The protein data bank. *Nucleic Acids Res.* 28, 235–242.

Bharatham, N., Bharatham, K., and Shelat, A.A., et al. 2014. Ligand binding more prediction by docking: mdm2/mdmx inhibitors as a case study. *J. Chem. Inf. Model* 54, 648–659.

Blum, C., Puchinger, J., Raidl, G.R., et al. 2011. Hybrid metaheuristics in combinatorial optimization: A survey. *Appl. Soft. Comput.* 11, 4135–4151.

Brooijmans, N., and Kuntz, I.D. 2003. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* 32, 335–373.

Cao, T.C., and Li, T.H. 2004. A combination of numeric genetic algorithm and tabu search can be applied to molecular docking. *Comput. Biol. Chem.* 28, 303–312.

Chen, H.M., Liu, B.F., Hwang, S.F., et al. 2007. SODOCK: Swarm optimization for highly flexible protein-ligand docking. *J. Comput. Chem.* 28, 612–623.

Fuhrmann, J., Rurainski, A., Lenhof, H.P., et al. 2010. A new Lamarckian genetic algorithm for flexible ligand-receptor docking. *J. Comput. Chem.* 31, 1911–1918.

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

Guo, L.Y., Yan, Z.Q., Zheng, X.L., et al. 2014. A comparison of various optimization algorithms of protein-ligand docking programs by fitness accuracy. *J. Mol. Model.* 20, 2251.

Hu, X., Balaz, S., and Shelver, W.H. 2004. A practical approach to docking of zinc metalloproteinase inhibitors. *J. Mol. Graph. Model.* 22, 293–307.

Huang, S.Y., and Zou, X.Q. 2010. Advances and challenges in protein-ligand docking. *Int. J. Mol. Sci.* 11, 3016–3034.

Jason, S., Merkle, D., and Middendorf, M. 2008. Molecular docking with multi-objective particle swarm optimization. *Appl. Soft. Comput.* 8, 666–675.

Jones, G., Willett, P., Glen, R.C., et al. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267, 727–748.

Jug, G., Anderluh, M., and Tomašič, T. 2015. Comparative evaluation of several docking tools for docking small molecule ligands to DC-SIGN. *J. Mol. Model.* 21, 164–178.

Kitchen, D.B., Decomez, H., Furr, J.R., et al. 2004. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discov.* 3, 935–949.

Li, Z., Gu, J., Zhuang, H., et al. 2015. Adaptive molecular docking method based on information entropy genetic algorithm. *Appl. Soft. Comput.* 26, 299–302.

López-Camacho, E., Godoy, M.J., García-Nieto, J., et al. 2015. Solving molecular flexible docking problems with metaheuristics: A comparative study. *Appl. Soft Comput.* 28, 379–393.

Moitessier, N., Englebienne, P., Lee, D., et al. 2008. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *Br. J. Pharmacol.* 153, 7–26.

Morris, G.M., Goodsell, D.S., Halliday, R.S., et al. 1998. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* 19, 1639–1662.

Morris, G.M., Goodsell, D.S., Huey, R., et al. 1996. Distributed automated docking of flexible ligands to proteins: Parallel application of AutoDock 2.4. *J. Comput. Aid. Mol. Des.* 10, 293–304.

Morris, G.M., Huey, R., Lindstrom, W., et al. 2009. AutoDock4 and AutoDockTools 4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* 30, 2785–2791.

Thomsen, R. 2003. Flexible ligand docking using evolutionary algorithms: investigating the effects of variation operators and local search hybrids. *Biosystems* 72, 57–73.

Address correspondence to:
Prof. Changsheng Zhang
College of Information Science & Engineering
Northeastern University
No. 3, Lane 11, Wenhua Road, Heping District, Shenyang 110819
People's Republic of China

E-mail: zhangchangsheng@ise.neu.edu.cn