AN INFORMATION BASED GENETIC ALGORITHM APPROACH TO FAST PEPTIDE DOCKING

Chen-Wei Yeh[1], Ji-Zheng Chu[2], Shi-Shang Jang[1]* and David Sun-Hill Wong[1]
1 Department of Chemical Engineering, National Tsing-Hua University
101 Guang Fu Road, Sec. 2, Hsin Chu, Taiwan 30043
2 Department of Automation, Beijing University of Chemical Technology
Beijing 100029

Abstract

In genetic algorithm (GA), there are two main methods of determining trial candidates: crossover and mutation. While crossover directs search between fit candidates, mutation plays a role on jumping out local optimal. In molecular docking calculations, it is desirable to chart as much unexplored search space as possible. Therefore it is desirable that mutation results in a uniform distribution of sampling points in solution space. In this work an information entropy based mutation procedure is developed. Instead of random mutation, mutation is directed to parts of the solution space that is least populated. Such a procedure is implemented in AUTODOCK and used to study the docking of Peroxisome Proliferator-Activated Receptors gamma (PPAR-γ). Results show that the proposed information-based genetic algorithm is superior to conventional GA for docking both in speed and precision.

Keywords

Information based genetic algorithm, Peptide Docking, PPAR-γ, AutoDock

Introduction

With rapid development in genomics, proteomics, combinatory chemistry, the technology for fast measurement of high-resolution structures of proteins, other biological macromolecules and their complexes, quantitative structure-activity relationship (QSAR) for drug-like molecules, and computing technology of both hardware and software, rational drug design (or structure-based drug design, or drug discovery) has come into reality in the last decade. With the success that a number of new drugs under clinical tests have been initiated, designed and modified with the help of computer, rational drug design gives us a new perspective for efficient and analytical discovery of drugs. In a broader sense, rational drug design also provides a platform to integrate knowledge.

From the viewpoint of molecular pathology and pharmacology, a drug is such a molecule which binds to a specific spot called the active site, of a target protein or a peptide or other biological macromolecule responsible for a disease, to abrupt its functionality causing the disease. Before a long list of crucial traits such as ADME/T (absorption, distribution, metabolism, and excretion/toxicity) is verified for a molecule to be a practical drug, a molecule as a possible lead is selected or designed chemically according to its geometric and chemical complementation with the receptor macromolecule. The infinite possibility of molecules, their conformations and complementary geometry and the highly nonlinear characteristics of chemical affinity make drug discovery an intelligent and extremely complex activity, which may exceed the intuition of all the chemists and thus be a task of labor and chance. It is in this aspect that a computer exhibits its extraordinary capability.

As for the design of a lead drug molecule, there are two basic iterative jobs: propose a structure and evaluate its fitness. While the former is primarily a problem of combinatory chemistry, the latter is mathematically a question of optimization known as docking. Since

* Correspondence author, whose email is: ssjang@che.nthu.edu.tw

Goodford (1985) proposed the seminal method of GRID for fast evaluation of affinity energy, a dozen of software systems have been developed to address the discovery of lead drug molecules (see reviews by Gane and Dean, 2000; Neamati and Barchi, Jr., 2002;). The objective of docking is to find the most suitable position at which and the most suitable conformation with which the ligand exists with respect to the macromolecule. The suitability or the so-called fitness is evaluated based the binding energy of the ligand with the macromolecule. Docking constitutes a formidable problem of calculation if detailed structures of both ligands and macromolecules are considered, not even to mention using rigorous models for binding energy. The current practice of docking calculations is using rigid-body macromolecules and allowing some flexibility in ligand molecules. In a partially flexible ligand, all the bond lengths and angles are fiexed, but possible rotations around bond are allowed. For such a ligand, the degrees of freedom include (Joseph-McCarthy, 1999; Yang, 2001):

	Degree of freedoms
Mass center of the ligand.	.3
Axial orientation	2
Axial rotation	1
Bond torsion	*n*
Total	*n*+6

Even under such simplifications, the solution space is still huge enough to challenge nowadays most efficient optimization procedures.

In addition to the huge solution space, the landscape of affinity energy is also very rugged with many local minima. Simulated annealing (SA) (Goodsell et al., 1996; Morris et al., 1996), genetic algorithm (GA) (Welch et al., 1996), and various variants of them such as family competition evolutionary approach (FCEA) (Yang, 2001) etc. have been used to search for best conformation of a ligand and its spatial position relative to a receptor. In the open-source code software AutoDock (Version 3.0) (Morris et al., 1998), Morris et al. proposed the Lamarckian genetic algorithm (LGA) in which environmental adaptations of an individual's phenotype are reverse transcribed into its genotype and become inheritable traits, with the environmental adaptations produced by a local search. Morris et al. also claimed that LGA can handle ligands with more degrees of freedom than SA and is more efficient and reliable than both SA and a traditional GA. LGA together with an empirical free energy correlation calibrated with known binding constants makes AutoDock a successful toolbox for docking.

Our experience with AutoDock shows that it is liable to producing local minima. In GA, there are two main methods of determining trial candidates: crossover, and mutation. While crossover directs search between fit candidates, mutation plays an important role on jumping out local optimal. To overcome this shortcoming, one can follow the suggestion of Morris et al. (1998), namely, have parallel runs or change the Cauchy distribution parameters to have a mutation more biased toward large deviates to cover broader solution space. In this study, however, an information entropy based mutation procedure is developed. Instead of random mutation, mutation is directed to parts of the solution space that is least populated. Such a procedure is implemented in AutoDock and used to study the docking of Peroxisome Proliferator-Activated Receptors gamma (PPAR-γ) .

Problem Description

For a partially flexible ligand (*n* rotatory bonds, fixed bond lengths and angles) and a rigid receptor, docking can be formulated formally as:

$$\min_{x,q,\theta} f(x,q,\theta) \qquad (1)$$

under constraints of

$$x_{i,\min} \le x_i \le x_{i,\max} \quad (i = 1, 2, 3) \qquad (2)$$

$$\sum_{i=1}^{4} q_i = 1 \qquad (3)$$

where, x=coordinates of the mass center of the ligand, $x \in R^3$; q=quaternion for the orientation of the ligand, $q \in R^4$; θ=rotation angles of bonds, $\theta \in R^n$; f = fitness function; $x_{i,min}$=lower bound on x_i; $x_{i,max}$=upper bound on x_i.

The constraint in (3) is set by structure of the active site of the receptor macromolecule, whereas that in (4) is for a unit vector of the quaternion. The fitness function f is the same as the Equation (2) of Morris et al (1998), which stands for the best development in correlating binding energy in molecular docking and accounts for the free energy change caused by dispersion/repulsion, directional hydrogen bonding, Coulombic electrostatic potential, unfavorable entropy due to conformational restriction, and desolvation. In AutoDock, the fitness (or scoring) function is an approximate free energy:

$$\Delta G = \Delta G_{VDW} + \Delta G_{Hbond} + \Delta G_{elec} + \Delta G_{sol} \qquad (4)$$

where, ΔG_{VDW} is Van der Waals potential energy, ΔG_{Hbond} is energy of hydrogen bond, ΔG_{elec} is electrostatic potential, ΔG_{tor} free energy change of torsion, and ΔG_{sol} is desolvation free energy.

Information Based Genetic Algorithm

The information-based genetic algorithm proposed in this paper is the same as that used in AutoDock (Morris et al., 1998) except that the mutation stage is guided by information entropy of the existing samples relative to the whole solution space, instead of the Cauchy distribution

used by Morris et al. in AutoDock. Suppose that there already exist (*N*-1) samples $V' \in \Omega$ and

$$V' = \{(f_i; x_i, q_i, \theta_i) | i = 1, 2, \cdots, N-1\} \quad (5)$$

Our problem is to find the optimum position $\{x_N, q_N, \theta_N\}$ at which a new sample will be taken, and as a result that Ω will be covered to the largest degree of uniform by the addition of this new sample. The diversity of variables will be compared through calculating their information entropy. The higher information entropy of the variable means the distribution of the variable is more divers. While one variable in GA operation with higher diversity means it behaves well in larger solution space, it will be more suitable for changed while the solution space with many constrains. On the other hand, the less discovered space of the variable will be tried. The behind idea is to have more diversity of samples in the solution space through mutation. To find the position of $\{x_N, q_N, \theta_N\}$, we can maximize the information entropy:

$$\max_{x_N, q_N, \theta_N} S \quad (6)$$

with

$$S = \sum_{i=1}^{M} p_i \ln(p_i) \quad (7)$$

$$p_i = \frac{s_i}{M} \quad (8)$$

where *S* is the entropy, *p* is the probability, *M* is the total number of subspaces, s_i is the number of samples in subspace *i*. It is seen that the above entropy is related to the existing samples and the partition of the whole solution space.

A Test with Himmelblau Function

The proposed information based genetic algorithm is used to find the minimum of a benchmark problem, the modified Himmelblau function:

$$f = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 + x_1 + 3x_2 + 57 \quad (9)$$

The solution space is $x_1, x_2 \in [-5, 5]$. For comparison, conventional GA is used for another parallel run. In both the information based GA and conventional GA, the total population is 3000 and the mutant rate is 0.07. Every generation will have 10 populations. The selecting rule is based on Roulette-Wheel-Selection, and the top 5 fitness individual will be selected, the more fitness of the individual the more generation of which will be regenerated. Crossover is used to generate new individuals depend on their parents. We use two-point crossover to generate the next generations, which are calculated by linear interpolation from their parents:

$$x_3 = x_1 + \delta(x_2 - x_1) \quad (10a)$$

$$x_4 = x_2 - \delta(x_2 - x_1) \quad (10b)$$

where, $x_1$and x_2 are parent's gene (the two variables of the modified Himmelblau function), x_3 and x_4 are the children (the next generations), is a random number. Figure 1 shows the performance of both the information based GA and the conventional GA in searching for the minimum of Himmelblau function. It is clear that the information based GA performs much better than the conventional GA .

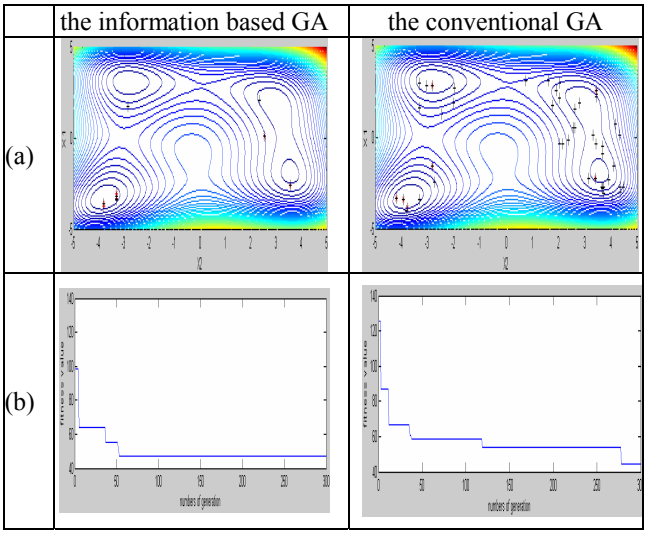

Figure 1. *Performance of the information based GA and the conventional GA in searching for the minimum of Himmelblau function: (a) distribution of samples in the background of fitness counter. (b) Fitness value changes with generation.*

Docking on PPAR-γ

Peroxisome Proliferator-Activated Receptors gamma (PPAR-γ) is important to metabolism of carbohydrate and the synthesis of lipid in human body, and is the receptor target of many projects for diabetes-drug discovery. In this study, the information based GA proposed in this study and the conventional GA are used to dock a ligand named as 544 in the active site (yellow rods at the left-bottom corner of Figure 2). Figure 3 compares binding energy of the best 40 conformations among 256 runs found by the two algorithms, from the same initial position and conformation. From this comparison, it is clear that the information based GA is much superior over the conventional GA, the former found much more low binding energy conformations than the later. The best conformation found by the information based GA is

shown in Figure 3. The reason for this may be that local search such as that suggested by Morris et al. is crucial to the conventional GA. However, the superiority of proposed approach is very clear as shown in Figure 3. One of the suboptimal conformation is compared with the natural conformation of the ligand(544) also as illustrated in Figure 2, the binding energy is -12.0 kcal/mol, the docking energy is -15.3 kcal/mol, and the root mean square distance (rmsd)of the conformation compared to natural ligand is 28.609Å

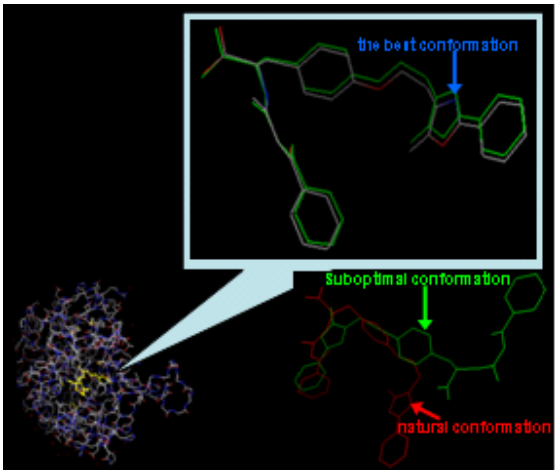

Figure 2*. PPAR-γ (the left bottom corner) and the best conformation of ligand 544 found by the information based GA, where green rods are the native conformation of 544, the colored rods are the conformation found by docking(the right top corner). The suboptimal conformation is compared with natural conformation on the right bottom corner.*

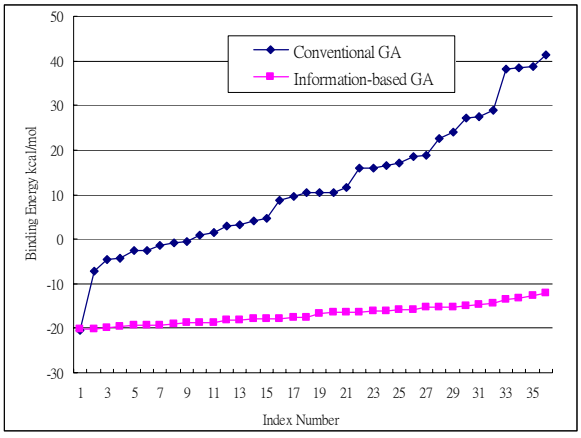

Figure 3. *The binding energy of the best 40 conformations found by docking using the information based GA and the conventional GA.*

Conclusion

An information based genetic algorithm (GA) has been proposed in this study and tested with a benchmark problem, the minimization of Himmelblau function and with docking a ligand onto Peroxisome Proliferator-Activated Receptors gamma (PPAR-γ). The novelty of the proposed algorithm is that the mutation is guided by maximizing the information entropy rather than random jumps, which facilitates the diversity of new solutions. Preliminary results from parallel runs show clear superiority of our information based GA over conventional GA both in speed and precision.

Acknowledgement

The Authors thank the financial support from National Science Council through the grant NSC92−2214−E007−008.

References

Gane, P.J. and P.M. Dean (2000). Recent Advances in Structure-Based Rational Drug Design. *Current Opinion in Structural Biolog*, 10, 401.

Goodford, P.J. (1985). A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.*, *28*, 849.

Goodsell, D.S., G.M. Morris and A.J. Olson (1996) Automated docking of flexible ligands: applications of AutoDock. *J Mol Recognit*, 9, 1.

Joseph-McCarthy, D. (1999). Computational approaches to structure-based ligand design. *Pharmacology & Therapeutics*, 84, 179.

Morris, G.M., D.S. Goodsell, R. Huey and A.J. Olson (1996) Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J Comput Aided Mol Des*, 10, 293.

Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K. and Olson, A. J. (1998), "Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function", *J. Computational Chemistry* , 19 : 1639.

Neamati, N. and J. Barchi, Jr. (2002). New Paradigms in Drug Design and Discovery. *Current Topics in Medicinal Chemistry*, 2, 211.

Welch, W., J. Ruppert and A.N. Jain (1996) Hammerhead: fast, fully automated docking of flexible ligands to protein binding sites. *Chem Biol.* , 3, 449.

Yang, J.-M. (2001). A family competition evolutionary approach of global optimization in neural networks, optical thin-film coatings, and structure-based drug design. *Doctorial thesis*, National Taiwan University.