# Molecular Binding: a Case Study of the Population-based Annealing Genetic Algorithms

Leuo-hong Wang Cheng-yan Kao Ming Ouh-young Wen-chin Chen
Department of Computer Science and Information Engineering
National Taiwan University, Taipei, Taiwan, R.O.C.
d1506006@csie.ntu.edu.tw cykao@csie.ntu.edu.tw ming@csie.ntu.edu.tw wcc@csie.ntu.edu.tw

*ABSTRACT*

A class of population-based annealing genetic algorithms(PAGs) has been defined. One of the special designs which incorporate an annealing scheme with the normal probability density function as the neighbor generation method is proposed. We use the algorithm to solve a real world application: computer-aided drug design. Using a dihydrofolate reductase enzyme with the anti-cancer drug methotrexate and two analogs of antibacterial drug trimethoprim in our model, PAGs can find a drug structure within a couple of hours, and the experimental results indicate that the solutions are reasonable.

## 1. Introduction

In the past two decades, computational pharmacology has been receiving broader attention from research institutions and the pharmaceutical industry. The main idea of this field is the integration of numerical algorithms with computer graphics techniques to simulate models of theoretical chemistry. One of the research topics in computational pharmacology, computer-aided drug design, rationalizes the simulation model by associating computational results of known drugs with experimental observations of the drug's activities and molecular structure features. When applying the model to new classes of drugs, biological activities of new drugs can be predicted. The most important step in this simulation, called molecular binding, is actually an energy minimization process. However, a basic problem of the molecular binding is that the energy minimization must be done on a complex hypersurface with many local minima. In the past, many researchers have addressed this problem. Nevertheless, no matter what methods are applied, although the solutions may be significant in certain cases, none of them can claim to solve the problem completely.

In the last few years, genetic algorithms(GAs) have been applied to the optimization problems in computational chemistry. The results indicated that GAs worked well for these applications to some extent. Hence, it is intuitive to introduce GAs to solve the molecular binding problem. But, as we know, the framework of simple GAs conceals a problem called genetic drift. As a result, the best parameter setting of GAs, especially the population size, is difficult to determine. Generally speaking, increasing the population size can be used to reduce the influence of genetic drift. But the efficiency of GAs will be restricted by the large population size. To enhance the power of GAs and reduce the influence of genetic drift, we adopt the hybrid approach of GA design which incorporates GA with the simulated annealing method and apply this framework to solve the molecular binding problem.

The rest of this paper is organized as follows. Section 2 shows the model of the population-based annealing genetic algorithm. Section 3 gives the details of the molecular binding problem. Experimental results of the algorithm are listed in section 4. Some interesting phenomena observed in our experiments are included in this section too. The final section contains our conclusion and future work.

## 2. Population-based annealing genetic algorithm

### 2.1. Genetic Algorithm(GA)/Simulated Annealing(SA) Combination

The research incorporating simulated annealing with genetic algorithms can be roughly spilt into two complementary categories, one using the genetic approach to design parallel simulated annealing[7, 10] and the other considering the simulated annealing method as a neighborhood operator of genetic algorithms[2, 3, 4, 9, 15]. The categories are not clear in some cases, since these cases can be explained in different ways from different perspectives.

Observing the GA/SA hybrids, there is a com-

0-7803-2759-4/95/$4.00  

mon characteristic. No matter what the combination between GA and SA, a partial sequence of SA is performed on each individual of the population. From the view point of GA, this partial sequence of SA improves each individual in GA. We can treat the SA as a neighborhood operator which searches the neighbors of an individual guided by the Metropolis condition. Motivated by the observation mentioned above, we fix the number of steps of SA performed on each individual and call this type of mechanism a population-based simulated annealing. We define a class of population-based SAs as neighborhood operators of GAs. These operators have the property of diversity maintenance.

**Definition 1** *A population-based SA(PSA) is an operator with three tuples: $PSA = \{K, \mathcal{N}_g, \mathcal{R}_A\}$, where $K$ is the number of steps processing by PSA, $K \geq 1$, $\mathcal{N}_g$ is a neighbor generation method, and $\mathcal{R}_A$ is the acceptance criterion applied to the current point and its neighbors.*

According to Definition 1, incorporating PSA operators with GAs will form a class of GA/SA hybrids, hereafter called population-based annealing genetic algorithms(PAGs).

### 2.2. The algorithm

In order to realize the influence of the number of steps K, the neighbor generation methods, $\mathcal{N}_g$, and the acceptance criterion rules, $\mathcal{R}_A$, we first designed an algorithm proposed in[15]. In [15], we proposed a PSA operator using the traditional SA as the neighbor generation method. The experimental results indicated that the performance was better than a simple GA or SA. In this paper, we use a new PSA operator which applies the normal probability density function(normal P.D.F.) on the whole search space as the neighbor generation method. The new algorithm is listing in the following:

A. Initialize the parameters, i.e., population_size, $T_0$, and decreasing factor $\alpha(0 < \alpha < 1)$
B. Randomly generate the initial population.
C. Repeatedly generate the new population as follows :
   1. **For each individual do**
      Best_point=Current_point=Current_individual ;
      Do K times :
         a. Generate Next_point using the normal P.D.F. applied on the whole solution space.
         b. Accept Next_point as Current_point by:
            $Prob = \frac{exp(-C_{next}/T_k)}{exp(-C_{current}/T_k)+exp(-C_{next}/T_k)}$ ;
         c. **if** $(C_{current} < C_{best})$ **then**
            Best_point := Current_point ;
      Pick Best_point into the transient population.
   2. **Genetic stage.**
      Apply genetic operators to the transient population;
   3. **if** (it is the first stage) **then**
         determine the initial temperature $T_1$
         $T_1 = \frac{Max_i(C^i_{max}-C^i_{min})}{population\_size/2}$
      **else**
         $T_{k+1} = \alpha \cdot T_k$ ;

There is something worth noting in the initialization of the system temperature. It is recognized that the execution time of SA depends on both the initial temperature and the decreasing factor of the temperature. Since PAGs include SA as an operator, the efficiency of PAGs also depend on the initial temperature. We provide a strategy defined as follows to determine a reasonable range of initial temperature. For the sake of efficiency, we define the acceptance probability of a detrimental move in the first generation to be 0.6. From the Metropolis criterion $Prob(\Delta C) = \exp(-\Delta C/T)$, we obtain $T = \frac{-\Delta C}{ln 0.6} \cong 2 \cdot \Delta C$, where $\Delta C$ is determined by largest possible detrimental move of the current generation. Since PAGs generate piecewise Markov chains from each individual of the current population, these individuals may be located at very different hills. We have to consider all chains to determine the largest possible detrimental move. Therefore, we calculate the initial temperature as :

$$T_{init} = \frac{Max_i(C^i_{max} - C^i_{min})}{population\_size/2} \quad (1)$$

where $C^i_{max}$ ,$C^i_{min}$ are the largest and lowest cost of the ith sequence of the Markov chain generated by the ith individual of the first generation. We take the maximal difference of individuals as $\Delta C$ to determine the initial temperature.

In addition, at the genetic stage, we use a ranking algorithm[16] as the selection mechanism. Crossover and mutation operators are performed according to the following steps.

1. At first, two parents are selected from the population randomly. The crossover operator is applied with a predefined crossover rate. After that, two offspring are produced.
2. The offspring survive only when the costs of these two offspring are both less than the average cost of the previous generation. Otherwise, give up the offspring and continue to apply the mutation operator to parents. The mutation operator is an annealing-like operator which borrows the exploration capability of SA to explore the neighborhood of the parents.
3. Finally, the offspring or the mutated parents are copied into the new population.

## 3. Problem description

In the past two decades, it has been well recognized that the biological activity of a drug happens at a specific receptor site, such as a protein. That is, the procedure of computer-aided drug design predicts a new three-dimensional(3D) molecular structure which can bind well with a

specific receptor site. More concretely, given an organic molecule(receptor), the work of computer-aided drug design generates 3D structures from a number of possibilities, measuring the fitness of them with the given molecule and deciding which one is best. Finally, the best one can be a kind of new drug dependent on its feasibility of synthesis and success of clinical experiments. However, as all molecules are deformable, many degrees of freedom exist. This results in the combinatorial explosion of possible structures of a drug molecule. Finding a good structure by brute force is computationally intractable.

Many researchers have devoted themselves to computer-aided drug design. Various structure generation methods have been described[1, 5, 6, 13]. In our research, we apply the PAG algorithm to generate favorable drug structures.

### 3.1. Structure encoding and scoring function

The general description of the structure optimization process is given as follows.

Given two molecules which consist of a number of atoms defined by their three dimensional coordinates, one defines the drug molecule and the other defines the receptor molecule. Intuitively, the spatial location of the drug and its three rotational angles relative to 3 axes are all adjustable. Moreover, the molecules have a number of deformable single bonds. Each single bond is a degree of freedom. Based on the description, the energy minimization between these two molecules becomes:

1. Fix the location of the receptor molecule. Initialize the structure of the drug molecule. Evaluate the interaction energy based on the scoring function.
2. Repeatedly adjust different degrees of freedom, including translating and rotating the drug molecule and twisting single bonds inside the drug, to fit the receptor. Evaluate the interaction energy of each new configuration.
3. Find the best configuration with the lowest interaction energy from these configurations.

Adjusting the value of each degree of freedom generates a new configuration . The whole search space is the combination of possible values of all degrees of freedom. Therefore we encode all degrees of freedom as a chromosome:

$$(t_x, t_y, t_z, r_x, r_y, r_z, s_1, ..., s_m)$$

where $t_x$ , $t_y$ and $t_z$ represent the position of the drug molecule relative to the centroid of the receptor, $r_x$ , $r_y$ and $r_z$ are the rotational angles of the drug and $s_i's$ are the twisting angles of single bonds inside the drug molecule. For convenience, we use a real-coded scheme to encode each parameter. That is, all parameters are real numbers.

In addition, we use the scoring function:

$$\mathcal{V} = V_\phi + V_{nb} + Ve \tag{2}$$

where $V_\phi$ is the bond torsion force field, $V_{nb}$ is the non-bonded interaction force field, and $V_e$ is the electrostatic interaction energy. In our scoring function, we use the Lennard-Jones 6-12 potential function to represent non-bonded interaction $V_{nb}$ and electrostatic interaction $V_e$. The Lennard-Jones equation is as follows[14]:

$$V_{tot}(r,d) = \sum_{r,d} \frac{332 q_d q_r}{\varepsilon |(\vec{R_r} - \vec{R_d})|} + \sum_{r,d} \frac{A_{rd}}{|(\vec{R_r} - \vec{R_d})|^{12}} - \sum_{r,d} \frac{B_{rd}}{|(\vec{R_r} - \vec{R_d})|^{6}} \tag{3}$$

where $V_{tot}$ is the total energy of binding, $q_r$ and $q_d$ are the charges of the atoms in the receptor and the drug respectively, $|\vec{R_r} - \vec{R_d}|$ is the distance between the receptor and the drug, and $\varepsilon, A_{rd}, B_{rd}$ are the dielectric and non-bond constants. In this function, the first summation simulates the electrostatic interaction between each pair of atoms, and the second and third summation simulate the repulsive and attractive terms of the van der Waals interaction energy.

Moreover, following the approach of [14], the bond torsion term $V_\phi$ also depends on Equation 3.

## 4. Molecular binding experiments

In order to verify that PAGs work well for generating a structure with near optimal binding energy, we tried to bind a real receptor molecule, dihydrofolate reductase enzyme(DHFR) with three drug molecules, methotrexate(MTX), and two analogs(inhibitor 91 and inhibitor 309) of trimethoprim. Methotrexate is an anti-cancer drug which is used clinically to cure patients, and trimethoprim is an anti-bacterial drug. There has been much research to analyze the binding structure of DHFR with methotrexate molecule [11, 12] or trimethoprim[8].

We have implemented the PAG algorithm on the Sun SparcStation 10. Each drug molecule is evaluated using the algorithm listed above. The results are given in Table 1. In this table, PAG executes 5 times for each case. The decreasing factor of temperature is 0.9. The processing steps of the PSA operator K is equal to the degrees of freedom. We use one-point crossover and blend crossover as the crossover operators. The probabilities of

| Drug name | | SB | Times (sec.) | Gen | Evaluations | Energy Kcal/mol |
|---|---|---|---|---|---|---|
| MTX | 30 | 10 | 42807 | 156 | 140226 | -90.64 |
| | 50 | | 52803 | 120 | 175478 | -91.20 |
| | 100 | | 168099 | 191 | 558604 | -93.83 |
| 91 | 30 | 10 | 31868 | 124 | 109922 | -60.03 |
| | 50 | | 70030 | 158 | 233436 | -59.78 |
| | 100 | | 144494 | 163 | 481647 | -63.65 |
| 309 | 30 | 6 | 14721 | 101 | 65010 | -55.53 |
| | 50 | | 32484 | 132 | 144666 | -61.57 |
| | 100 | | 57643 | 133 | 286996 | -59.63 |

Table: 1: *The results obtained from the PAG with the neighbor generation method which applies normal P.D.F on the whole search space. The numbers appear in the right side of the first column are the population sizes. Moreover, there are 557 atoms in DHFR(just considering the active site of DHFR). The drug molecules have about 50 atoms.*

crossover and blend crossover are both 0.5. Based on the design of the genetic stage explained in the previous section, the mutation rate is dependent on the failure of the crossover operators. Since we adopted the real-coded scheme to encode the chromosome and each degree of freedom is represented by a floating point with 32 bits, there are $2^{16 \cdot 32} = 2^{512}$ points totally in the search space for the case of 16 degrees of freedom.

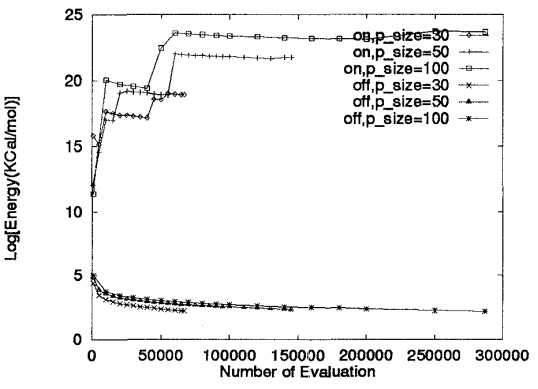


Fig. 1: *The on-line and off-line performance of the PAG. The on-line performance is the average of the solutions evaluated so far. The off-line performance is the average of the best solution found so far.*

In addition, we also plotted the off-line performance and on-line performance of the PAG for the cases of different population size(Figure 1).

According to the results , we find that the PAG has the following properties:

1. For the cases of different population size, the PAG converges to a solution before 200 generations. When the population size increases, the number of generations, total number of evaluation, and execution time all rise. The only exception is the case of MTX(p_size=30). It indicates that PAG is stable.
2. The best binding energies of the three drugs are distributed from -40Kcal/mol to -120Kcal/mol. The PAG obtains the minimal energy at the range from -40Kcal/mol to -100Kcal/mol. Since all solutions are not refined by a local minimizer, the results roughly verify PAG's problem-solving power. Another reason which confirms the power of PAGs is the binding structure generated by PAGs. This will be explained later.
3. The off-line performance of the PAG is also interesting. The case of p_size=30 converges quickly to a good solution. Compared to the results of Table 1, the solution is almost as good as the other cases with larger population sizes. As in the other version of PAG in[15], the quality of solutions generated are proportional to the population size. It indicates the new PAG listed above needs less population size to generate the same quality of solution. Moreover, this PAG never traps into a bad local minimum when the population size is larger than 30.
4. In Figure 1, the on-line performance of the PAG is greater for all cases. This shows the diversity maintenance property of the PAG. Because we use the PSA operator which generates the next point from the whole search space by applying the normal P.D.F, it explores the new points, including very bad solutions, from generation to generation. Even as individuals are getting better, the probabilities of generating bad points never decrease. The on-line performance shows the situation.

In addition to the binding energy, the existence of hydrogen-bonds is another criterion to determine the quality of the fit between molecules. According to the results presented in[11] which have shown the binding structure of the MTX with DHFR, there existed a pocket in DHFR. When the interaction between MTX and DHFR converged, MTX would be buried deeply in the pocket. Observing all the cases, every generated drug structure is bound to this pocket, because hydrogen bonds form in the position. That is, all drugs are attached to the same atoms in DHFR by the PAG. Figure 2 shows the illustration of the binding structure of drug molecule(Inhibitor 309) with DHFR. In this figure, we show the partial molecular surfaces of DHFR as dots and represent the drug molecule by solid surfaces. We find the drug molecule indeed burying into DHFR deeply. Based on this observation, we have more strong evidence to claim that the results obtained by the

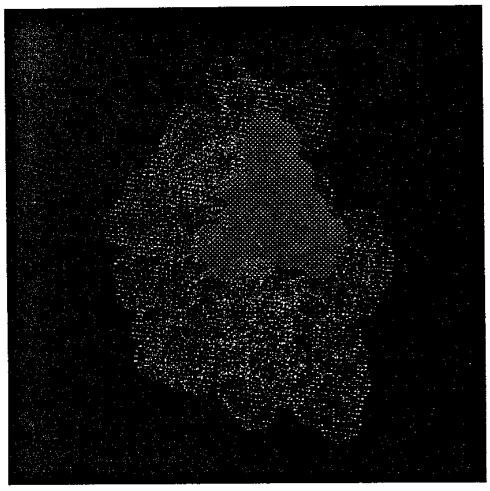

Fig. 2: *The binding structure of Inhibitor 309 with DHFR. The dots indicate the molecular surface of DHFR, The balls are the atoms of Inhibitor 309.*

PAG are very significant.

## 5. Conclusion and future work

In this paper, we proposed a class of neighborhood operators called population-based simulated annealing(PSA) for genetic algorithms. Incorporating different PSAs with genetic algorithms formed new GA/SA hybrids(PAGs). We designed a new PAG and applied it to solve the molecular binding problem. Our case studies not only indicate the power of PAGs but also the properties of the PSA operator when applying the normal P.D.F.

However, we used only a simplified model to represent the interaction between molecules, the results are an approximation of the actual situation. Based on the experience of this research , we plan to extend the model to include other force fields, such as bond stretching and bond angle bending, into our algorithm. Since the degrees of freedom will increase accordingly, the population size will be enlarged to reduce the chance of premature convergence. The efficiency of the algorithm will be influenced heavily, but the power of the PSA operator used by the PAG seems to be able to solve the problem by reducing the population size to a reasonable size. We will also try to apply PAGs to more complicated model of molecular binding.

## References


[1] R. Abagyan, M. Totrov, D. Kuznetsov, "ICM–A New Method for Protein Modeling and Design: Applications to Docking and Structure Prediction from the Distorted Native Conformation," *Journal of Computational Chemistry*, Vol. 15, No. 5, pp. 488–506, 1994.

[2] T. Boseniuk, W. Ebeling, "Boltzmann-, Darwin- and Haeckel-strategies in optimization problems," *Parallel Problem Solving from Nature* **496**, pp. 430–444, 1991.

[3] D. Brown, C. Huntley, A. Spillane, "A parallel genetic heuristic for the quadratic assignment problem," *Proceeding of the Third International Conference in Genetic Algorithms*, pp. 406–415, 1989.

[4] A. Eiben, E. Aarts, K. Van Hee, "Global convergence of genetic algorithms: a Markov chain analysis," *Parallel Problem Solving from Nature* **496**, pp. 4–12, 1991.

[5] D. Gehlhaar, K. Moerder, D. Zichi, C. Sherman, R. Ogden, S. Freer, " *De Novo* Design of Enzyme Inhibitors by Monte Carlo Ligand Generation," *Journal of Medicinal Chemistry*, Vol. 38, No. 3, pp. 466–472, 1995.

[6] V. Gillet, W. Newell, P. Mata, G. Myatt,S. Sike, Z. Zsoldos, A. Johnson, "SPROUT: Recent Development in the de Novo Design of Molecules," *Journal of Chem. Inf. Comput. Sci.*, Vol. 34, No. 1, pp. 207–217, 1994.

[7] D. Goldberg, "A Note on Boltzmann Tournament Selection for Genetic Algorithms and Population-Oriented Simulated Annealing," *Complex Systems* **4**, pp. 445–460, 1990.

[8] L. Kuyper, "Receptor-based Design of Dihydrofolate Reductase Inhibitors : Comparison of Crystallgraphically Determined Enzyme Binding with Enzyme Affinity in a Series of Carboxy-substituted Trimethoprim Analogues," *Journal of Med. Chem.*, No. 25, pp. 1120–1122, 1982.

[9] F. Lin, C. Kao, C. Hsu, "Applying the Genetic Approach to simulated Annealing in Solving Some NP-Hard Problems," *IEEE Transaction on System, Man, Cybernetics*, Vol. 23, No. 6, pp. 1752–1767, 1993.

[10] S. Mahfoud, D. Goldberg, "Parallel Recombinative Simulated Annealing: A Genetic Algorithm," *Parallel Computing*, Vol. 21, No. 1, pp. 1–28, 1995.

[11] D. Matthews, R. Alden, J. Bolin, S. Freer, "Dihydrofolate Reductase : X-ray Structure of the Binary Complex with Methotrexate," *Science* **197**, pp. 452–455, 1977.

[12] D. Matthews, et al., "Dihydrofolate Reductase from Lactobacillus casei : X-ray Structure of the Enzyme Methotrexate NADPH Complex," *Journal of Biological Chemistry*, Vol. 253, No. 19, Issue of October 10, pp. 6946–6954, 1978.

[13] A. Payne, R. Glen, "Molecular Recognition Using a Binary Genetic Search Algorithm,"

*Journal of Molecular Graphics*, Vol. 11, No. 2, pp. 74–91, 1993.

[14] N. Pattabiraman et. al., "Computer Graphics and Drug Design: Real Time Docking, Energy Calculation and Minimization," *Journal of Computational Chemistry*, Vol. 6, pp. 432–436, 1985.

[15] L. Wang, C. Kao, M. Ouh-young, W. Chen, "Using an Annealing Genetic Algorithm to solve Global Energy Minimization problem in Molecular Binding," *Proceedings of the Sixth International Conference on Tools with Artificial Intelligence*, pp. 404–410, 1994.

[16] D. Whitley, "The Gentitor Algorithm and Selection Pressure : Why Rank-Based Allocation of Reproductive Trials is Best," *Proceedings of the Third International Conference on Genetic Algorithms*, pp. 116–121, 1989.