Using an Annealing Genetic Algorithm to Solve Global Energy Minimization Problem in Molecular Binding

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.

Abstract

Molecular binding, important in drug design, explores the accurate binding structures between molecules. This exploration can be formulated as an global optimization problem. However, the problem in molecular binding is that the search space is very large and the computational cost increases tremendously with the growth of the degrees of freedom. In this paper, we utilize a new algorithm called the annealing genetic algorithm to solve the global optimization problem in molecular binding. Using an protein with three anti-cancer drugs in our model, our algorithm can find a binding structure with a complicated energy computation within a couple of hours and the experimental results indicate that the solutions are reasonable.

1 Introduction

Owing to the rapid evolution of the pharmaceutical industry in recent years, thousands of kinds of drugs against diverse diseases have been discovered. Traditionally, the discovery of these drugs depended largely on skilled choice and empiricism. Recently, the developmental procedures of drug design have evolved systematically. Research indicates that by exploiting certain physicochemical properties of compounds and associating these properties with the biological activities of binding between compounds at a molecular level, the design of new drugs can be guided. One of the associations is that the basis of biological activity of many drug molecules is the interaction of the drug molecules with a specific receptor site on a protein molecule. Based upon that observation, the procedure of drug design can be viewed as whether the drug molecule can bind with its receptor. Such an approach that binds the drug molecule into a receptor site on a macromolecule, a protein for instance, is called the macromolecular-fitting approach [1].

With respect to the biological activity of interaction between molecules, the corresponding physicochemical property can be described as an energy minimization problem which actually minimizes the binding energy between molecules. However, the interaction between molecules is so complicated that the minimization problem is very difficult to solve. In the past, many researchers have developed various optimization techniques to solve the problem [2][4]. In [2] the Ellipsoid algorithm is used to solve the optimization problem. In [4], a stochastic optimization technique called simulated annealing is used. No matter what algorithms are applied, they are either too time consuming or can only solve a small scale problem. Instead of applying algorithms to solve the energy minimization problem, other researchers designed systems to allow experienced users to interactively search for the binding position of drug molecule on its receptor [3][15]. These systems focused on improving the perception of human users via the assistance of the visualization of 3D molecular structure and used simplified energy calculation to simulate the binding process dynamically. Through the aids of visual cues, with the force field display and force feedback provided by these systems, human users had new understanding about how and why the drug molecule binds with its receptor [3]. However, these systems need real experts such as biochemists designing drugs to guide the search, and even so, binding an unknown drug molecule into the receptor site is not a trivial job.

As mentioned above, no matter what approach is applied, the correctness of the solution is either obtainable only in small problems or heavily dependent on the domain knowledge of the human users. Therefore, in our research, we try to both extend the dimension

1063-6730/94 $4.00 © 1994 IEEE

of the problem to be solved and avoid the man-in-the-loop situation as much as possible. After a lot of experimentation, we believe that a new framework called the annealing genetic algorithm can be used to meet our objectives if parameters of the framework are tuned properly.

The rest of this paper is organized as following. Section 2 gives details of the definition of molecular binding and describes the scale of the problem that we will solve. Section 3 shows the spirit of our framework, the annealing genetic algorithm, and the outlines of the tuning process. 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 the directions of our future work.

2 Problem definition

In essence, the macromolecular-fitting approach applies the techniques of computer graphics to display the detailed conformations of the drug and its receptor molecules and then uses some numerical optimization methods, such as the steepest descent algorithm, to determine the best position and orientation of binding between these molecules. When we say that the position and orientation of a drug molecule with respect to receptor molecules is better fit than before, it means that the extent of the fitness (geometric fit, electrostatic fit, hydrogen-bond fit, etc.) between molecules is more compact. In physics, the extent of the fitness can be described approximately by the potential energy between binding molecule pairs. The lower the potential energy, the better the binding configuration. Hence , when the potential energy is lower than some predefined threshold, the macromolecular-fitting approach will be able to predict whether the drug molecule possesses desired properties with a great probability. These properties are related to a drug being able to inhibit a specific disease that we want to cure.

However, how can we describe the fitness quantitatively? In [5], a complicated model is presented. The potential energy consists at least of four kinds of forces, including van der Waals forces, electrostatic forces and hydrogen-bond forces, molecular dynamics(bond stretching, bond angle bending, and torsion angle twisting), and hydrophobic forces(in aqueous solvent). Based upon the above model, the fitness of binding can be formulated quantitatively. But, experience indicates that in applying this empirical model to describe the problem, it becomes too complicated to solve directly. Therefore, we first try to solve a simplified problem. This simplified model considers the first two kind of forces mentioned above. When we consider that the binding energy approximately consists of electrostatic and van der Waals forces, it can be defined by the Lennard-Jones 6-12 potential function [15]:

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

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, $\varepsilon, A_{rd}, B_{rd}$ are the dielectric and non-bond constants. In this function, the first summation simulates the electrostatic interaction between each pair of atoms, the second and third summation simulate the repulsive and attractive term in van der Waals interaction energy. Using this function, we can consider the molecular binding problem to be an energy minimization problem.

Unfortunately, when we try to solve the minimization problem formulated above, the true global minimum is very hard to find. Since the receptor protein molecules have hundreds to thousands of atoms and the drug molecules have at least 6 degrees of freedom for translation and rotation and dozens of single bonds that can be twisted, the local minima abound.

The more formal description of this energy minimization problem 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, the other defines the receptor molecule. Essentially, each molecule viewed as a rigid body can be maintained as a graph in which the vertex are the atoms and the edges are chemical bonds between two atoms. Differing from a pure rigid body, the molecule has certain deformable single bonds. We mark each such edge in the graph explicitly. Based on these elementary data sets, the energy minimization process between these two molecules becomes:

1. Fix the location of the receptor molecule. Search for optimal solutions of the drug molecule in the binding. Intuitively, the search space is around and inside the receptor molecule.

2. Repeatedly adjust different configurations, including translating and rotating the drug

molecule and twisting single bonds inside the drug molecule, to fit the receptor.

3. Find the best configuration with the lowest binding energy from these configurations.

Intuitively, the adjustment of the configurations actually selects a feasible combination from the possible degrees of freedom about the molecules. In the case described above, the degrees of freedom include x, y, z translation and rotation about the drug molecule and single bond twisting inside the drug molecule because these operations will alter the relative position of molecules and change the binding energy. Obviously, it is a combinatorial optimization problem. There are over sixteen variables (6 basic degrees of freedom about rotation and translation and ten single bonds inside the drug molecule). Even while the given problem is just a simplified case, the global optimal solution is still very difficult to obtain.

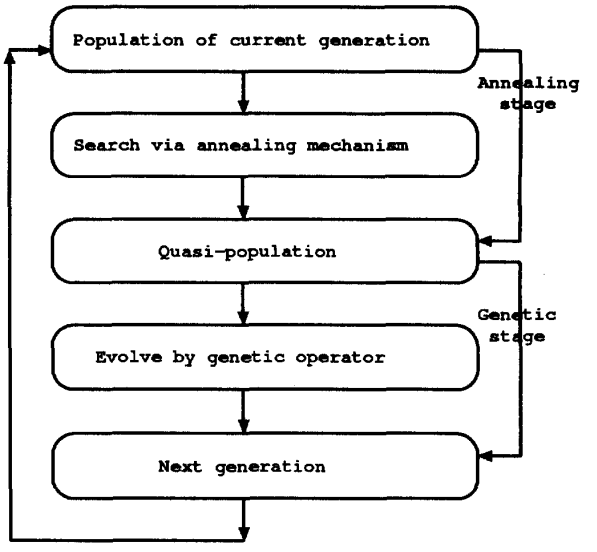

Figure 1: *The concept of the annealing genetic algorithm.*

3 The algorithm

Because of the complexity of the problem, optimization algorithms such as gradient-based techniques are virtually impossible for the molecular binding problem. In the literature, there are two kinds of probabilistic techniques, simulated annealing [9][11]and genetic algorithms [7][8], which can efficiently approximate the global minimum of the combinatorial optimization problems.

Simulated annealing is based on thermodynamics and can be viewed as an algorithm that generates a sequence of Markov chains controlled by gradually decreasing temperature of the system. However, as control of the parameter called system temperature is not a trivial job, the efficient annealing schedule is hard to design.

Genetic algorithms(GAs) which are based on natural selection try to inherit the genes with good fitness from generation to generation. Using reproduction plans to exhibit the selection pressure and applying genetic operators to populations to explore diverse search space, the genetic algorithm borrows the power of the natural selection to solve optimization problems.

Since these two techniques suffer from some drawbacks, Lin et. al [12], try to combine the concepts of the SA and GA to produce a new stochastic approach called annealing genetic algorithm(AG). The annealing genetic algorithm incorporates the genetic algorithm with simulated annealing. Empirical studies on several combinatorial optimization problems which are in the class of NP indicate that the performance of the AG algorithm is promising. The concept of the annealing genetic algorithm is shown as follow:

In Figure 1, the AG consists of a two-stage cycle, the annealing stage and the genetic stage. The populations of the current generation first search for better candidates for further evolution via the Markov chain generation process of the annealing stage. These better candidates served as quasi-populations providing the sources of evolution to which genetic operators can be applied. After applying the genetic operators, individuals of the next generation are produced. This new generation will become a feedback to the first stage until the whole population of the generation converges to some extent. In summary, AG can be viewed as either a simulated annealing algorithm with population-based state transition or a genetic algorithm with the Boltzmann-type selection operator.

Our algorithm essentially follows the spirit of the AG proposed in [12]. Problems studied in [12] are all represented by the traditional binary string encoding scheme used in genetic algorithms. Based on the characteristics of our problem, however, using real numbers to encode each parameter is a natural way to represent the solution space. Therefore, we apply the concept of the real-coded genetic algorithm [6][18] into AG. Moreover, we adopt a set of operators that have been used in real-coded GA to improve

the performance, like linear crossover [18] and blending crossover(Blx-0.5) [6]. These operators work well when the coding scheme uses real number and indeed serve as an important role in our algorithm. We use the simulated annealing algorithm as a mutation operator in our algorithm. Under the control of the system temperature, simulated annealing has a good feature in that the diversity of exploration is guaranteed when the temperature is high enough, and the guidance of better solutions is provided as the temperature decreases gradually. A similar concept appeared in [16]. Differing from previous approach, we control the temperature more carefully and we elitism between generations. In the following, we will summarize the different features compared with the original AG and our AG algorithm.

3.1 Different features in AG

During the implementation of AG, we have found alternatives that will be able to make the AG more efficient. There are two parts: the generation of quasi-population and the reasonable initial temperature. These issues are described as follow:

1. *The generation of quasi-population :* The quasi-population plays an important role in the evolutionary stage of the AG. They are the sources of further evolution in the search for better solutions. In the original AG, individuals of the quasi-population are selected from several piecewise Markov chains, which are generated by the move generation strategy that satisfies the Metropolis criterion from each individual of the current generation. Since the system temperature is so high that the Metropolis criterion is satisfied easily in the first few generations, the quasi-population is dominated by the Markov chains generated by only a few individuals of these generations. This will decrease the diversity of population and result in premature convergence. In order to avoid this situation, we modify the generation of quasi-population so that we choose the best point in every Markov chain generated by each individual of current generation to become the candidates of the quasi-population. In this case, the population diversity is preserved hopefully. Experiments show that our modification produces the expected effect and largely decreases the chance of premature convergence.

2. *Reasonable initial temperature:* The execution time of the simulated annealing algorithm depends on the initial temperature and the decreasing factor of the temperature. Since the body of AG includes the simulated annealing, the efficiency of AG is also dependent upon the initial temperature. In the original AG, there is an equation used to determine the reasonable range of initial temperature,

$$T_{init} = \frac{(C_{max} - C_{min})}{population\_size/2} \qquad (1)$$

where C_{max} and C_{min} are largest and lowest cost respectively and T_{init} is the initial temperature. It will generate a very large initial temperature when applying the equation to our problem. This is because that the range of the function value in our problem spreads from over 10^{10} to 10^{-2}. Therefore we adopt an alternative strategy to determine the initial temperature. The strategy is defined as follow. The original AG algorithm defines the acceptance probability of a detrimental move to be 0.6. From Metropolis criterion $Prob(\Delta C) = \exp^{\frac{-\Delta C}{T}}$, we obtain $T = \frac{-\Delta C}{ln 0.6} = 2 * \Delta C$ where ΔC is determined by the largest detrimental move of current generation. Since our version of AG generates piecewise Markov chains from each individual of current generation which may locate on very different hills, we can not determine the largest detrimental move as the two points with highest and lowest costs. These two points belong to two hills with a great chance. The probability of moving from one of them to the other is much lower than 0.6. When this probability is much lower than 0.6, the equation (1) is inadequate to determine a reasonable initial temperature and the whole schedule of AG will be inefficient. Therefore, we modify the determination of the initial temperature as:

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

where C^i_{max} and C^i_{min} are the largest and lowest cost of ith sequence of Markov chain generated by the ith individual of the first generation. We take the maximal difference of the individuals as ΔC to determine the initial temperature. Experimental results demonstrate that by using equation (2), the efficiency of the AG is improved.

3.2 Outlines of the algorithm

Our version of the annealing genetic algorithm(AG) can be described as below:

1. First, we generate an initial population randomly.

2. Annealing process is applied to the current generation to generate the quasi-population. The initial temperature from the initial population is determined using Equation (2).

3. Genetic stage is applied to quasi-population. Here we use the ranking algorithm [7] to do reproduction because the range of the fitness value in our problem is so large that proportional reproduction dependent upon fitness value is unsuitable. As presented in [17], ranking can not only prevent GA from premature convergence but also provide a direct control on selective pressure that can affect the search speed. We adopt the dynamic ranking procedure which is adaptively changing the selective bias from generation to generation. The dynamic ranking procedure is more powerful than the static approach that fixes ranking bias during evolution. Because the population diversity is maintained during first few generations when the bias is low, the selective pressure increases at only last few generations. In our algorithm, we use 1.2 as initial bias and multiply the bias by 1.005 in each generation. This parameter settings find the best results. We use one-point-crossover at parameter level and blend crossover Blx-0.5 [18][6]. We use an annealing-like mutation operator to evolve the population from generation to generation. These genetic operators perform according to the following steps. At first, the parents are selected from quasi-population randomly. The one-point-crossover and crossover Blx-0.5 are applied randomly with the crossover rate of 0.5 respectively. After that, two offsprings are produced. The offsprings survive only when the costs of these two offsprings are both less than average cost of the old generation. Otherwise, they give up the offsprings and continue to apply the mutation operator to the parents. The mutation rate of the annealing-like mutation operator, which is applied to parents, is based on the current system temperature. Finally, the mutated parents are copied into the next generation.

4. We check whether the frozen condition is satisfied or not. If it is satisfied the AG is terminated, otherwise repeating step 3 and 4 until system is frozen. Following [12], the frozen condition is signaled when 80% of the population in a certain generation has an error rate less than 0.1% relative to the best point of the current generation.

Drug	Time	Iteration	Evaluation	Energy
MTX	20873	128	80928	-79.63
91	19566	118	74970	-35.70
309	10792	101	50333	-58.55

Table 1: *The results evaluated by AG. Inhibitor MTX and 91 both have 10 single bonds and 309 have just 6 single bonds.* Time *represents the total execution time in seconds.* Iteration *is the total generation evaluated by AG.* Evaluation *means the count of evaluation about Lennard-Jones Equation.* Energy *is the minimum found by AG.*

4 Results

In order to verify that the modified version of AG is viable for solving the molecular binding problem, we use a real receptor molecule, dihydrofolate reductase enzyme(DHFR), three drug molecules, methotrexate(MTX), and two analogues(inhibitor 91 and inhibitor 309) of trimethoprim in our simulation model. Methotrexate is an anti-cancer drug which is used clinically to cure patients, and trimethoprim is an antibacterial drug. There has much research which analyzes the binding structure of DHFR as containing the methotrexate molecule [13][14] or trimethoprim[10]. Techniques of the x-ray crystallography were used to obtain the three dimensional binding structure of the molecules. We have implemented the annealing genetic algorithm on the Sun SparcStation 10. The three drug molecules are evaluated respectively. The results are given in Table 1.

In Table 1, AG executes 9 times for each drug molecule. The population size of AG is 50 and the decreasing factor of temperature is 0.9. The precision of search space is 0.2 angstrom in translation and 5 degrees in rotation. The probability of crossover and blend crossover are both 0.5, the mutation rate is dependent on the success of the crossover operators. That means mutation will execute under the condition that crossover operators can not improve the individuals.

The data listed above are the average of the results of 9 runs except the minimal energy which is the result of the best run. According to the results listed in Table 1, we find that AG converges to a near optimal solution in about 100 generations for all cases. It indicates that AG is steady and powerful. The differences are that the execution time of inhibitor 309 is shorter and the binding energy of inhibitor 91 is much

Figure 2: *The binding structure of MTX with DHFR.*

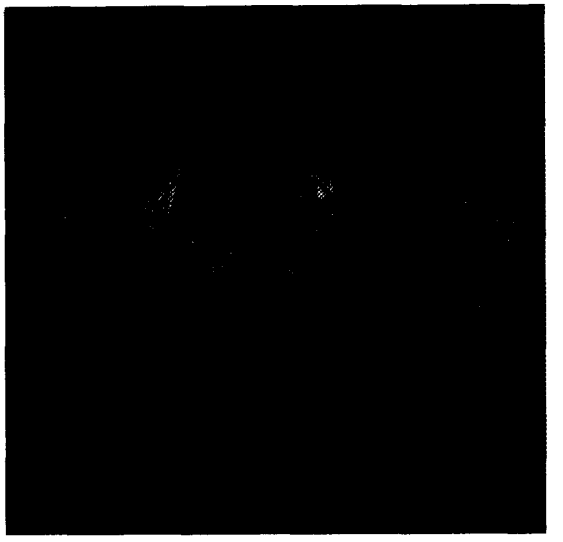

Figure 3: *The binding structure of inhibitor 91 with DHFR.*

higher than the others. The reason of the former is straightforward. It is because the degrees of freedom of inhibitor 309 is small. As to the later one, because there are two primary hydrogen-bonds formed at the same position with the others, the configuration is still reasonable. In general, the binding energy of a drug lower than -40Kcal/mol indicates that the drug is curative with a great probability. The results verify the medicative of these three drugs.

Figure 2 and Figure 3 show the schematic illustration of the binding structure of two drug molecules (MTX and 91) with DHFR, respectively. The brighter line segments constitute the drug molecule, the others belong to DHFR. In addition to binding energy, the existence of hydrogen-bonds is another criterion that determine the goodness of fitting between molecules. According to the results presented in [13] which have shown the binding structure of the MTX with DHFR, there is a pocket that exists in DHFR and when the interaction between MTX and DHFR tends to be stable, MTX will be buried deeply in the pocket. It is because of that there are two primary hydrogen-bonds formed in the pocket. Undoubtedly, in Figure 2, our results show that AG makes the MTX drug molecule bounded in this pocket since two hydrogen-bonds are found (The arrow points to the position of the primary hydrogen-bonds). In Figure 3, the inhibitor 91 has the same situation. After careful examination of these figures, one can see that the primary hydrogen-bonds are attached to the same atoms in DHFR. Based on the observation, we claim that the results obtained by AG method are very significant.

5 Conclusion

In this paper, we introduced a new application of genetic algorithms to molecular binding problem and proposed a framework called the annealing genetic algorithm to solve the problem. As previously described, after tuning the framework carefully, we obtained near-optimal solutions to the molecular binding problem. The solutions can aid biochemists in the research on drug design. In the research, we verify that the genetic algorithms are suitable for solving the molecular binding problem. We believe these results are very useful for the biotechnology community. However, as we use only a simplified model to represent the interaction between molecules, the result is an approximation of the actual situation. Based upon the experience of this research , we plan to extend the model to include simulations of molecular dynamics into our model. The first problem encountered is the efficiency of the computation since the degrees of freedom increase tremendously. Nevertheless, there are a lot of methods which can be used to simplify the energy calculation (such as 3D tabulation [15]). The parallel genetic algorithms will also improve the efficiency. Previous experiences with AG have shown that

the empirical complexity of AG is $O(n^2)$ [12]. Therefore, the efficiency problem of applying AG to large problem seems quite controllable. We will simplify the energy calculation and design a parallel version of AG to conquer the complexity of the real world molecular binding problems.

References

[1] C. R. Beddel, "Designing Drugs to Fit a Macromolecular Receptor," *Chem. Soc. Rev.*, No. 13, pp. 279-319. 1984

[2] M. Billeter, A.E. Howard, I.D. Huntz, P.A. Kollman, "A New Technique to Calculate Low-Energy Conformation of Cyclic Molecules Utilizing the Ellipsoid Algorithm and Molecular Dynamics: Application to 18-Crown-6," *J. Am. Chem. Soc.*, No.110, pp. 8385-8391. 1988.

[3] F.P. Brooks,Jr., M. Ouh-Young, J.J. Batter, P. J. Kilpatrick, "Project GROPE - Haptic Displays for Scientific Visualization," *Computer Graphics*, Vol. 24, No. 4, Aug. 1990.

[4] A. T. Brunger, "Crystallographic Refinement by Simulated Annealing on Supercomputers," *Cray Channels*, pp. 16-19, Fall 1988

[5] P. M. Dean, "Molecular foundations of drug-receptor interactions," Chapter 3, Cambridge University Press, 1987.

[6] L. J. Eshelman, J. D. Schaffer, "Real-Coded Genetic Algorithms and Interval-Schemata," *Fundamentals of Genetic Algorithms*, pp. 187-202, 1993.

[7] D. E. Goldberg, "Genetic algorithms : In Search, Optimization and Machine Learning," *Reading*, Ma. : Addison-Wesley Publishing Company, 1989.

[8] J. H. Holland, "Adaptation in natural and artificial systems", *Reading*, The university of Michigan Press, 1975.

[9] S. Kirkpatrick, C. D. Gelatt Jr., M. P. Vecchi, "Optimization by simulated annealing," *Science*, Vol. 220, No. 4598, pp. 671-680, 1983.

[10] 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.*, pp. 1120-1122, No. 25, 1982.

[11] P. J. M. van Laarhoven, E. H. L. Aarts, "Simulated Annealing : theory and applications," *Reading*, Dordrecht, Holland : D. Reidel Publishing Company, 1987.

[12] F. T. Lin, C. Y. Kao, C. 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, Nov. 1993.

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

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

[15] 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.

[16] R. Unger, J. Moult, "A Genetic Algorithm for 3D Protein Folding Simulations," *Proceedings of the Fifth International Conference on Genetic Algorithms*, pp. 581-588, 1993.

[17] 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.

[18] A. Wright, "Genetic Algorithms for Real Parameter Optimization," *Fundamentals of Genetic Algorithms*, G. J. E. Rawlins (editor), Morgan Kaufmann, San Mateo, CA, pp. 205-218.