Dual Individual Distributed Genetic Algorithm for Minimizing the Energy of Protein Tertiary Structure

Tomoyuki HIROYASU[1], Mitsunori MIKI[1],
Takashi IWAHASHI[2], Yuko OKAMOTO[3,4]

1 Department of Engineering, Doshisha University, 1-3, Tataramiyakodani, Kyoutanabe, Kyoto, 610-0394,Japan
2 Graduate School of Engineering, Doshisha University, 1-3, Tataramiyakodani, Kyoutanabe, Kyoto, 610-0394, Japan
3 Institute for Molecular Science, Okazaki National Research Institues, 38, Nishigonaka, Myoudaiji, Okazaki, Aichi,444-8585, Japan
4 Department of Functional Molecular Science,The Graduate University for Advanced Studies , 38, Nishigonaka, Myoudaiji, Okazaki, Aichi, 444-8585, Japan

Abstract: This paper describes Genetic Algorithm (GA) for minimizing the energy of protein tertiary structure. In the conventional study, Simulated Annealing (SA) is used to be applied for this problem. In the previous studies, it is also reported that it is difficult to find the optimum solutions by GAs. Dual individual Distributed Genetic Algorithm (Dual DGA) is one of DGAs and is good at global search. The Dual DGA also maintains the diversity of the solutions. Therefore, it can be supposed that they can get a good solution in energy minimization of protein tertiary structure. In this study, Dual DGA is applied to protein tertiary structure. The target protein in this paper is Met-enkephalin that consists of 5 amino acids sequences. The results show that Dual DGA has the higher searching capability than SA.

Keywords: Protein, Genetic Algorithm, Simulated Annealing

1. Introduction

Protein is a substance directly connected to the biological phenomena and its biological functions are said to be derived from its tertiary structure[1)]. Thus, clarification of its tertiary structure leads to an explanation of the biological phenomena process. Tertiary structure of the protein is believed to correspond to the conformation with the lowest residual potential energy[2)]. Therefore, as one of the methods to predict the tertiary structure of proteins, minimization of the energy function has been studied. Simulated Annealing (SA) has often been employed as the optimization method to predict the tertiary structure of proteins by the minimization of the energy[3)].

Genetic Algorithm (GA) is one of the powerful emergent optimization algorithms[4)]. GA is expected to be an alternative method for finding protein tertiary structure that has the minimum energy. However, it is reported that it is very difficult to find the optimum solution only by GAs[5)]. Because there are many sub optimum points in the landscape of the protein energy function, the search is converged in the early stage.

In this paper, Dual Individual Distributed Genetic Algorithm (Dual DGA) is used for finding the structure. The Dual DGA is extended version of Distributed GAs (DGAs)[6)]. Usually, there is only one population in GAs. On the other hand, in DGA, the total population is divided into sub populations. Each sub population is often called "island". In each sub population, normal genetic operations are performed for several generations. After the certain generation, some of the individuals are chosen and are moved to the other island. This operation is called "migration". Because the population size in each island is small, the early convergence may happen in each island. However, the migration operation prevents the early convergence and maintains the diversity of the solutions during the search. Therefore, it was reported that DGA has a higher searching capability than conventional GA [7)]. In Dual DGA, there are only two individuals in each island. Dual DGA can maintain the diversity of the solutions in the early stage of the search and has the high searching capability. DGA has many parameters that users should determine before the simulation. In Dual DGA, since some parameters, such as the number of population, migration rate, and crossover rate are determined automatically, users can use Dual DGA easily.

In this paper, Dual DGA is applied for minimizing the energy of protein structure. The target protein is Met-enkephalin that consists of five amino acids. Through the simulation, it can be said that Dual DGA find the minimum energy of protein structure and GAs can predict the protein structure.

2. Energy Minimization of Protein Tertiary Structure

2.1 Related Works

It is said that the real protein tertiary structure has the minimum energy of the structure among the pos-

sible states. Therefore, one of the strategies to predict the protein tertiary structure is minimizing the energy by changing the structure. In the several former studies, the energy of protein tertiary structure is tried to minimize using GAs. In those studies, two types of the protein models are used; those are the grid type and the real molecular type. In the grid model, protein structure is simplified into grid structure. In this model, since combinatorial numbers of the possible states are reduced, the calculation cost becomes small[8]. However, the precise prediction of the structure cannot be expected. Using this model, Unger and Moult applied the GA[9] and Krasnogor applied multimeme GA[10]. On the other hand, in the real molecular model, the energy is determined by considering all the molecular effects in protein. Therefore, the precise prediction can be expected[2]. However, the calculation cost is very huge. Kobayashi et al. used the real coded GA[11] and Okamoto developed Simulated Annealing (SA) approach[2].

However, it was reported that the minimization of protein tertiary structure is very difficult by GAs[5, 9].

2.2 Energy function of protein structure

Protein consists of 20 types of amino acids. Between the atoms, energies, such as an interaction and a hydrogen bond, exist. The summation of these energies becomes the total energy of protein structure. In this paper, the energy function of protein structure that was introduced by Okamoto et al.[2] is also applied. The following equation is the energy function.

$$E_{tot} = E_P + E_S \tag{1}$$

In the equation (1), the total energy function $E_{tot}(kcal/mol)$consists of molecular' structure energy E_P and solvent free energyE_S. E_S is the term that is determined by several solvent contributions. E_P consists of electrostatic term E_C, Lennard-Jones term E_{LJ}, hydrogen bond term E_{HB}, and torsion energy term E_{tor}. This expression is shown in Equation2.

$$\begin{aligned} E_P &= E_C + E_{LJ} + E_{HB} + E_{tor} \\ E_C &= \sum_{(i,j)} \frac{332\, q_i\, q_j}{\epsilon\, r_{ij}} \\ E_{LJ} &= \sum_{(i,j)} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right) \\ E_{HB} &= \sum_{(i,j)} \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) \\ E_{tor} &= \sum_{i} U_i \left(1 \pm cos(n_i\, \chi^i) \right) \end{aligned} \tag{2}$$

In this equation, r_{ij} expresses the distance between ith atom and jth atom. ϵ is dielectric constant and χ^i is the dihedral angle of ith bonding.

Parameters and geometric information of ECEPP/2[12, 13, 14] are used for the energy function. The simulation is supposed in gas-phase case. The dielectric constant ϵ is 2.0 and E_S is 0.0.

3. Dual Individual Distributed Genetic Algorithm

DGAs are powerful algorithms that can derive better solutions with lower computation costs than Canonical GAs. Therefore, many researchers were studied on DGAs[15, 16, 17]. However, DGAs have the disadvantage that they require careful selection of several parameters, such as the migration rate and migration intervals, that affect the quality of the solutions.

Dual DGA[18] is an extended version of DGA. The Dual DGA has only two individuals in each island. The concept is shown in Figure 1.

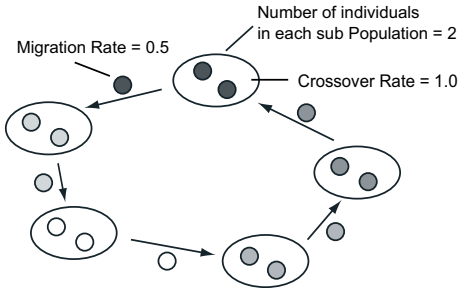

Figure 1: Fixed parameter of Dual DGA

In the Dual DGA, the following operations are performed.

The operation of Dual DGA is summarized in Fig.2.

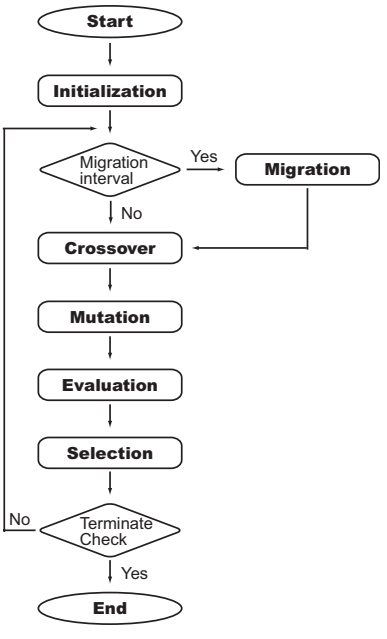

Figure 2: Flowchart of Dual DGA

- There are only two individuals in each island.
- Migration method: The individual who will migrate is chosen in random. The chosen individual is copied to the island. In this island, the migrated individual is substituted for the worst individual.
- Migration topology: The stepping stone method is performed and the direction of the migration is determined randomly at every migration event.
- Crossover: In this paper, one point crossover is used. From two parent individuals produce two children. These four individuals are remained until the selection is performed.
- Mutation: One of the bits of the individual is chosen in random and the bit is flipped. The chosen bit for each individual in an island should be different.
- Selection: The individual who has the best fitness value among the four individuals is chosen to remain. The individual who has the best fitness value in the previous generation is also chosen to remain.

One of the advantages of Dual DGA is that users are free from setting some of the parameters. By limiting the population to two individuals on each island, the Dual DGA model enables the following parameters to be determined automatically:

- crossover rate: 1.0
- number of islands: total population size/2
- migration rate: 0.5

4. Experiments

To discuss the searching ability of the Dual DGA for minimizing the energy of protein tertiary structures, the small protein is targeted and the results of the Dual DGA are compared with those of the DGA.

4.1 Met-enkephalin

The target protein in this chapter is Met-enkephalin. Met-enkephalin is the protein that consists of five amino acids. The sequence of amino acids of Met-enkephalin is illustrated in Fig.3. It has been already derived that the energy of this protein is less than $E-11kcal/mol$ [19] that is derived by ECEPP/2 energy function [12, 13, 14]. In this paper, it is also considered that the optimum structure is derived when the structure has the energy less than $E-11kcal/mol$.

In the experiments, we choose 19 dihedral angles of Met-enkephalin as design variables. Ten of them are $\phi_1, \psi_1, ..., \phi_5, \psi_5$ in the main chain and nine of them are $\chi_1^1 ... \chi_5^4$ in the side chain. The range of these design variables is from $-180°$ to $180°$.

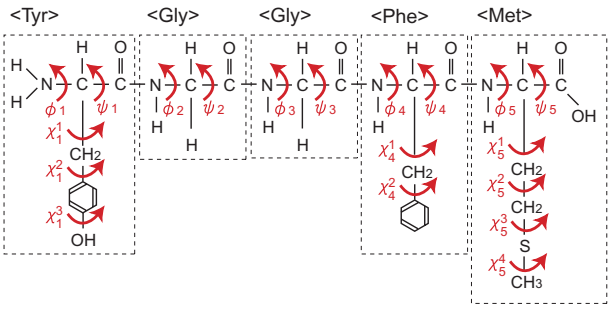

Figure 3: Met-enkephalin molecule

4.2 Parameters

The energy of Met-enkephalin is minimized by Dual DGA and DGA. There are several parameters in Dual DGA and DGA. The used parameters are summarized in Table 1.

Table 1: Used Parameters

model	DGA	Dual DGA
Total Population size	800, 1600, 3200, 6400	
Sub Population Size	16,8,4	2
Number of Design Variables	19	
Chromosome Length	171 (= 19 × 9 Design Variable)	
Selection	Tournament	-
Tournament Size	2	-
Crossover Rate	1.0	
Crossover	1pt. crossover	
Mutation Rate	0.006 (= 1 / 171)	
Migration Interval	1,2,3,4,5,6,7,8,9,10	
Migration Rate	0.25	0.5
Number of Elites	1	-
Terminal Criterion	1,900,000 evaluations	

In these parameters, several types of parameters for the population size and migration interval are prepared. The rest parameters are fixed. There are 4, 8, 16 islands in DGA. When DGA has n islands, it is expressed "DGA(n)". Simulation will be terminated when the protein energy evaluation function is called 1,900,000 times. This terminal condition is same as Okamoto's experiment [19]. In the following sections, all the results are derived from 30 trials.

4.3 Success Rate

In Fig.4, the success rates of each DGA and Dual DGA case are shown. The success ratio is the rate of the number of trials that derive the optimum solutions with the number of all trials.

Dual DGA with 6400 population size and three migration interval derives the best success rate ($0.93 = 29/30$). In the most cases, when the population size becomes bigger, the success rate becomes better. Huge migration interval also affects the better solutions. However, when the population size is 6400, there is the optimum point for the migration interval. These results

indicate the following things; the population size 6400 is adequate for this terminal condition and the optimum migration interval is existed.

4.4 Best, average, and worst energy values of each method

In Fig.5, the average energy values of each method are shown. In the same way, the best values are shown in Fig.6 and the worst values are shown in Fig.7.

From these figures, the following three tendencies can be derived. Firstly, in the most cases, the results of Dual DGA are better than those of DGA. Secondly, when the population size is bigger, the results are getting better. This tendency was also derived in the former section. Finally, when the migration interval becomes longer, the results become better in the most cases. However, since the population size 6400 is adequate with this terminal condition, the optimum point of migration interval is also existed in this case.

4.5 Search transition

In Fig.8, the search transition is illustrated. This is the case when the population size is 6400 and migration interval is 3. The results are the average of 30 trials. From

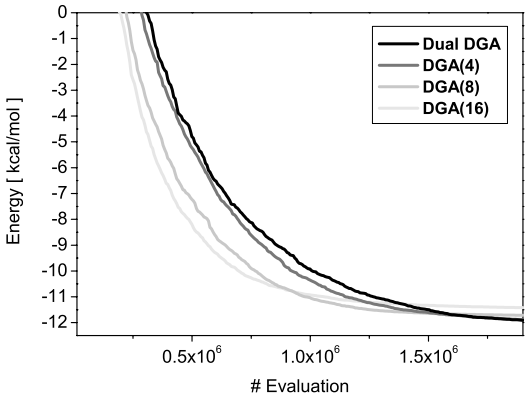

Figure 8: Energy transition of Met-enkephalin (Populaiton size:6400, Migration Interval:3)

this figure, it is found the following two tendencies. In the former part of the search transition, the model who has the smaller number of the islands finds the better solutions. On the other hand, in the latter part of the search transition, the model who has the bigger number of the islands finds the better solutions. From these results, it can be said that, in the first stage of the search, the model who has the bigger number of the islands, especially Dual DGA, maintains the diversity of the solutions and it is searching in the global area. Therefore, the best values of this model are worth than that of the model who has the smaller number of island. However, in the second stage, Dual DGA may change to start to search around the solution that has the best value. Therefore, Dual DGA can derive the better solution at the end of the simulation. From these discussions, it can be concluded that the model who has the bigger number of islands and Dual DGA can maintain the diversity of the solutions during search and they can find the better solutions.

4.6 Comparison with the other methods

In this section, the success rates of Dual DGA and DGA are compared with other methods; those are Okamoto's SA[20, 21] and Parallel Simulated Annealing using Genetic Crossover (PSA/GAc)[21]. The results are also the average of 30 trials.

In Fig.9, the success rates of Dual DGA and DGA are better than that of SA. At the same time, it can be found that Dual DGA and DGA are the same as PSA/GAc. Since PSA/GAc is the very good algorithm for minimizing the protein energy, it can be said that Dual DGA and DGA derived the very good result in finding the minimum energy of Met-enkephalin.

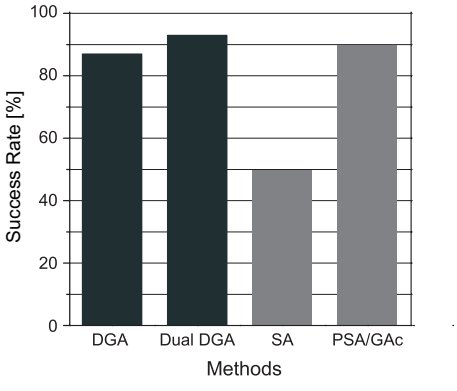

Figure 9: Success Rate

5. Conclusion

In this study, the energy of Met-enkephalin that consists of five amino acids is minimized by changing its structure. The minimum energy state can be equal to the real structure. The optimization methods are DGA and Dual DGA. In the former studies, the optimum solutions could not be derived by GA. On the other hand, in this paper, it is found that DGA and Dual DGA can successfully derived the optimum solution of Met-enkephalin.

In the future research, Dual DGA is tried to apply for deriving the minimum energy of the other huge proteins. At the same time, the other types of GAs, such as real coded GAs should be discussed.

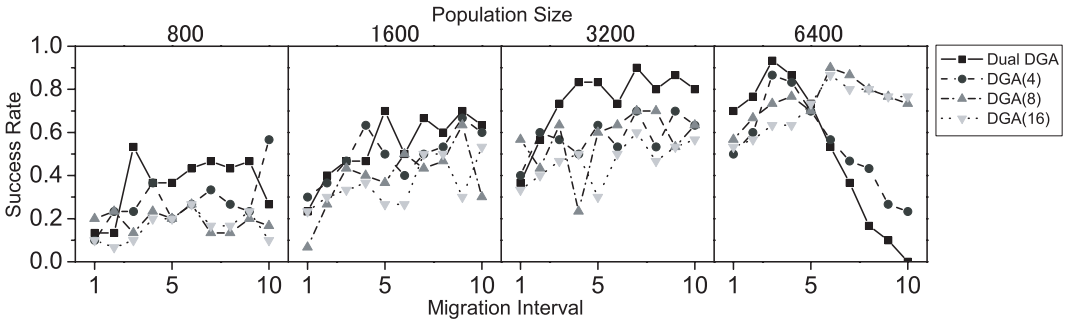

Figure 4: Success Rate

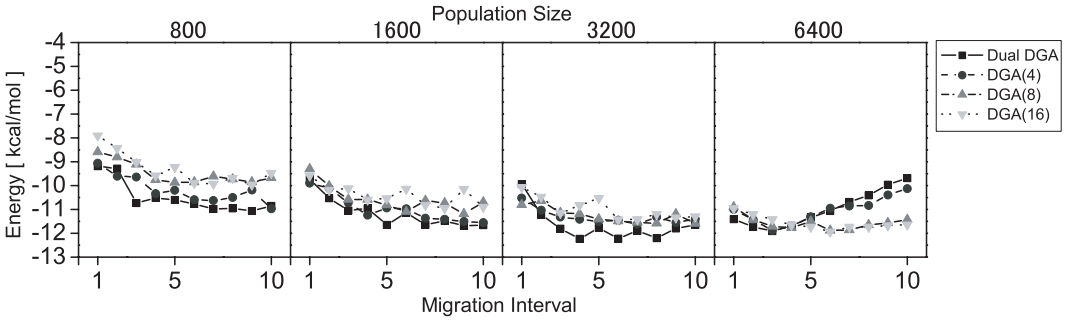

Figure 5: Average

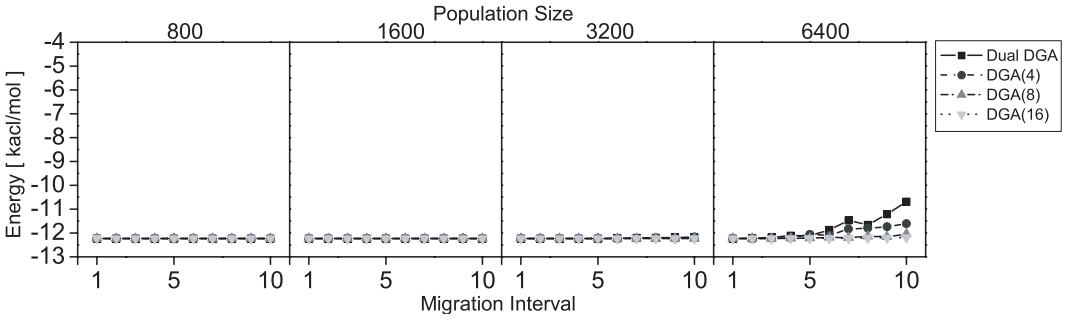

Figure 6: Best

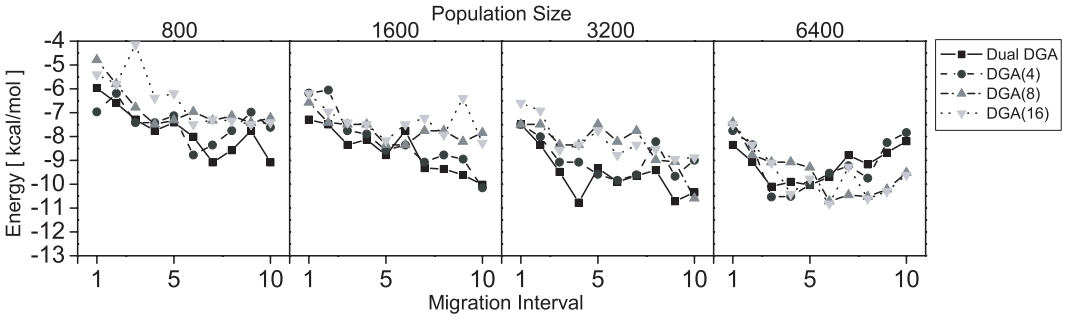

Figure 7: Worst

References

[1] Toshihiko Ikeuchi. *learning life Protein Science*. Ohmsha, 1999.

[2] Yuko Okamoto. Protein Folding Problem as Studied by Monte Carlo Simulated Annealing. *Physical Properties Research*, Vol. 70, No. 6, pp. 719–742, 1998.

[3] Hikaru Kawai, Takeshi Kikuchi, and Yuko Okamoto. A prediction of tertiary structures of peptide by the Monte Carlo simulated annealing method. *Protein Engineering*, Vol. 3, No. 2, pp. 85–94, 1989.

[4] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, 1989.

[5] Laurence D.Merkle, Gary B.Lamont, Jr. George H.Gates, and Ruth Pachter. Hybrid Genetic Algorithms for Minimization of a Polypeptide Specific Energy Model. *IEEE Conf on Evolutionary Computation*, pp. 396–400, 1996.

[6] Reiko Tanese. Distributed genetic algorithms. *Proc. 3rd International Conference on Genetic Algorithms*, pp. 434–439, 1989.

[7] Tomoyuki Hiroyasu, Mitsunori Miki, and Jiro Kamiura. A Presumption of Parameter Settings for Distributed Genetic Algorithms by Using Design of Experiments. *IPSJ JOURNAL*, Vol. 43, No. SIG 10(TOM 7), pp. 199–217, 2002.

[8] Ken A.Dill. Theory for the folding and stability of globular proteins. *Biochemistry*, Vol. 24, No. 1501, 1985.

[9] Jan T Pedersen and John Moult. Genetic algorithm for protein structure prediction. *Current Opinion in Structural Biology*, No. 6, pp. 227–231, 1996.

[10] N.Krasnogor, B.P.Blackburne, E.K.Burke, and J.D.Hirst. Multimeme Algorithms for Protein Structure Prediction. *PPSN VII*, pp. 769–778, 2002.

[11] Osamu Tomobe, Isao Ono, and Shigenobu Kobayashi. Experimental Study on Determination of Protein three dimensional Structure using Genetic Algorithm. *SICE 25rd Intelligent System Symposium*, pp. 35–40, 1998.

[12] F.A. Momany, F.A., R.F. McGuire, A.W. Burgess, and H.A. Scheraga. *J. Phys. Chem.*, Vol. 79, pp. 2361–2381, 1975.

[13] G. Nemethy, M.S. Pottle, and H.A. Scheraga. *J. Phys. Chem.*, Vol. 87, pp. 1883–1887, 1983.

[14] M.J. Sippl, G. Nemethy, and H.A. Scheraga. *J. Phys. Chem.*, Vol. 88, pp. 6231–6233, 1984.

[15] Enrique Alba and Jos M. Troya. A Survey of Parallel Distributed Genetic Algorithms. *Complexity*, Vol. 4, No. 4, 1999.

[16] Erick Cantú-Paz. *Effective and Accurate Parallel Genetic Algorithms*. 2000.

[17] M. Tomassini. *Parallel and distributed evolutionary algorithms : A Review*, pp. 113–133. J. Wiley and Sons, 1999.

[18] Tomoyuki Hiroyasu, Mitsunori Miki, Masaki Sano, Yusuke Tanimura, and Masahiro Hamasaki. Dual Individuals Distributed Genetic Algorithm. *SICE*, Vol. 38, No. 11, pp. 990–995, 2002.

[19] Yuko Okamoto, Takeshi Kikuchi, and Hikaru Kawai. Prediction of Low-Energy Structures of Met-Enkephalin by Monte Carlo Simulated Annealing. *CHEMISTRY LETTERS*, pp. 1275–1278, 1992.

[20] Ulrich H. E. Hansmann and Yuko Okamoto. Numerical Comparisons of Three Recently Proposed Algorithms in the Protein Folding Problem. *JOURNAL OF COMPUTATIONAL CHEMISTRY*, Vol. 18, No. 7, pp. 920– 933, 1997.

[21] Tomoyuki Hiroyasu, Mitunori Miki, Maki Ogura, and Yuko Okamoto. Examination of Parallel Simulated Annealing Using Genetic Crossover. *IPSJ JOURNAL*, Vol. 43, No. 7, pp. 70–79, 2002.