Study of Crossover in One Max Problem by Linkage Analysis

Hiroshi Furutani
Department of Information Science, Kyoto University of Education
Fushimi-ku, Kyoto, 612-8522 JAPAN

Abstract

The paper describes an investigation of crossover in genetic algorithms (GAs) by means of linkage analysis. We study the rate of change in population fitness using Fisher's fundamental theorem of natural selection as a guiding principle. We consider the GA applied to One Max problem whose fitness function has a very simple structure for the analysis. The important concept used in the investigation is that of linkage disequilibrium which represents the genetic correlation between loci. It is found that the effect of crossover can be estimated by calculating the degree of linkage disequilibrium. It is shown that the theory can explain features of numerical experiments in GAs on One Max problem with and without crossover.

1 INTRODUCTION

It has been a long-standing problem how crossover works in genetic algorithms (GAs), and by what mechanism it improves the performance of GAs [1, 2]. This paper describes an approach to give one possible solution to such a problem.

We focus on the variance of fitness, which will be shown to have a close relationship with the rate of change in the average fitness. Fisher already pointed out the importance of the variance for evolution in his famous fundamental theorem of natural selection [3]. We use Fisher's theorem as a guiding principle for investigating the effect of crossover in GAs. The variance of fitness in GAs was discussed by many authors from various points of view; for example, Goldberg and Rudnick [4], Davidor [5] and Aizawa [6].

Another important quantity is linkage disequilibrium, which also plays an essential role in this study [7, 8]. It will be shown that the variance of fitness can be decomposed into two parts. The first part comes from the variance at each bit, on which crossover does not have any effect. The second part is proportional to the sum of coefficients of linkage disequilibrium. Since crossover reduces the absolute values of these coefficients, it changes the variance of fitness and then accelerates the speed of evolution. The importance of linkage disequilibrium in the analysis of GAs was already noted by Holland [1], and several studies from the viewpoint of linkage have been reported [9].

2 METHOD

The GA studied in this paper is the one similar to Simple Genetic Algorithm defined in the book of Goldberg [2]. The mathematical details are described in [10, 11].

2.1 Mathematical Description

A binary string B_i of length ℓ represents the ith genotype, and is given by

$$B_i = < i_\ell, i_{\ell-1}, \cdots, i_2, i_1 >,$$

where i_k $(k = 1, \ldots, \ell)$ stands for the kth bit of unsigned integer i in binary representation. The number of genotypes is $n = 2^\ell$. We use the notation of

$$|i| = \sum_{k=1}^{\ell} i_k$$

to show the number of bit ones in B_i.

The relative frequencies of genotypes are used for representing the distribution of a population

$$x_i(t) = \frac{N_i}{N} \quad (0 \le i \le n-1), \tag{1}$$

where N_i and N stand of the number of the ith genotype and size of the population, respectively. Since $N = N_0 + N_1 + \ldots + N_{n-1}$, the frequencies satisfy the normalization condition

$$\sum_{i=0}^{n-1} x_i(t) \; = \; 1. \tag{2}$$

In the following, we only consider populations of infinite size, and neglect effects of random sampling.

2.2 Walsh Transformation

Here we briefly describe Walsh functions [12] and Walsh transform of the relative frequencies. The jth Walsh function for the ℓ bit string i is defined as

$$W_{ij} \equiv \prod_{k=1}^{\ell} (-1)^{i_k \cdot j_k}, \tag{3}$$

where we use the following representation of integers i and j

$$i =< i_\ell, i_{\ell-1}, \ldots, i_1 >, j =< j_\ell, j_{\ell-1}, \ldots, j_1 > .$$

Walsh transform $\tilde{x}_i(t)$ of the relative frequency $x_i(t)$ is given by

$$\tilde{x}_i(t) \equiv \sum_{j=0}^{n-1} W_{ij} x_j(t), \tag{4}$$

and its inverse transform is

$$x_i(t) = \frac{1}{n} \sum_{j=0}^{n-1} W_{ij} \tilde{x}_j(t). \tag{5}$$

Hereafter we call Walsh transform $\tilde{x}_i$ as Walsh coefficient.

Noting $W_{0j} = 1$ for all j, and from the normalization condition(2), we have

$$\tilde{x}_0(t) = \sum_{i=0}^{n-1} x_i(t) = 1. \tag{6}$$

We call $k = |i|$ of $\tilde{x}_i$ as Walsh coefficient of the degree k, and it will be demonstrated that the degree of $\tilde{x}_i$ plays a very important role in the analysis of GAs. When we want to show the degree of $\tilde{x}_i$ explicitly, we use the following notation

$$\tilde{x}_i \; \equiv \; \tilde{x}^{(k)}[b_1, b_2 \ldots b_k],$$

where the degree $k = |i|$, and $b_1 < b_2 < \ldots < b_k$ stand for the positions of bit ones in the binary representation of i. Using this notation, we have an expression

$$\tilde{x}^{(k)}[b_1, b_2 \ldots b_k] = \sum_{j=0}^{n-1} \prod_{m=1}^{k} (-1)^{j(b_m)} x_j, \tag{7}$$

where $j(b_m)$ stands for the bit of j at the position b_m.

We can calculate the probability of bit zero at each position m $(1 \leq m \leq \ell)$ in a population by noting the following relation

$$\begin{aligned} \frac{1}{2}(1 \, + \, \tilde{x}^{(1)}[m]) &= \frac{1}{2}(\tilde{x}_0 \, + \, \tilde{x}^{(1)}[m]) \\ &= \frac{1}{2} \sum_j (1 + (-1)^{j_m}) \, x_j. \end{aligned}$$

We define an operator to extract a string j having bit zero at position m

$$\delta(j_m) = \frac{1 + (-1)^{j_m}}{2}.$$

Using this operator, we have

$$\frac{1}{2}(1 \, + \, \tilde{x}^{(1)}[m]) = \sum_j \delta(j_m) \, x_j.$$

In the same manner, we have an operator to extract strings having bit one at position m

$$\delta(1 - j_m) = \frac{1 - (-1)^{j_m}}{2},$$

and therefore

$$\begin{aligned} p_m^{(0)} &= (1 + \tilde{x}^{(1)}[m])/2 \\ p_m^{(1)} &= (1 - \tilde{x}^{(1)}[m])/2, \end{aligned} \tag{8}$$

where $p_m^{(0)}$ and $p_m^{(1)}$ stand for the relative frequencies of bit zero and one at position m, respectively.

We can also define the probability that a string has both bit ones at positions m and m' as $p_{m,m'}^{(11)}$. This probability can be represented in terms of Walsh coefficients

$$p_{m,m'}^{(11)} = \frac{1}{4}\{1 - \tilde{x}^{(1)}[m] - \tilde{x}^{(1)}[m'] + \tilde{x}^{(2)}[m, m']\}. \tag{9}$$

2.3 Rate of Change in Population Fitness

We consider here the rate of increase in the average fitness $\bar{f}(t)$ under selection. The relative frequencies at $t+1$ are given by those at t

$$x_i(t+1) = \frac{f_i}{\bar{f}(t)} x_i(t) \quad (i = 0, \ldots, n-1). \tag{10}$$

Here f_i stands for the fitness of genotype B_i, and $\bar{f}(t)$ stands for the average fitness of the population at t

$$\bar{f}(t) = \sum_{i=0}^{n-1} f_i x_i(t). \tag{11}$$

It is easy to obtain the solution of these equations [13]

$$x_i(t) = \frac{f_i^t x_i(0)}{\sum_j f_j^t x_j(0)} \quad (i = 0, \ldots, n-1). \qquad (12)$$

Then we derive a very important relation from the evolution equation including only selection (10) [3] . Multiplying the both sides of equation (10) by f_i, and summing over i, we have

$$\bar{f}(t+1) \; = \; \frac{1}{\bar{f}(t)} \sum_i f_i^2 x_i(t).$$

Then the rate of change in the average fitness

$$\Delta \bar{f}(t) = \bar{f}(t+1) - \bar{f}(t),$$

can be given by

$$\begin{aligned} \Delta \bar{f}(t) &= \frac{1}{\bar{f}(t)} \mathrm{VAR}(f), & (13) \\ \mathrm{VAR}(f) &= \sum_i f_i^2 x_i(t) - \bar{f}(t)^2. \end{aligned}$$

This is a version of Fisher's fundamental theorem of natural selection [3]. This theorem states that the increase in the average fitness at each generation is proportional to the ratio of the variance of fitness VAR(f) to its average. It therefore means that diversity of a population is a crucial factor for the performance of GAs.

3 VARIANCE AND LINKAGE DISEQUILIBRIUM

3.1 Linkage Disequilibrium

The definition of the coefficient of linkage disequilibrium is given by [7]

$$D[m, m'] = p_{m,m'}^{(11)} - p_m^{(1)} p_{m'}^{(1)}. \qquad (14)$$

Substituting equations (8) and (9) into the definition, we have

$$D[m, m'] = \frac{1}{4}\{\tilde{x}^{(2)}[m, m'] - \tilde{x}^{(1)}[m]\tilde{x}^{(1)}[m']\}. \qquad (15)$$

The effects of crossover in the Walsh representation are given by Vose [14, 15]. Therefore we have an expression for the effects of crossover on the coefficient of linkage disequilibrium.

Crossover does not change the value of the first order Walsh coefficient [15]

$$\widehat{C}\tilde{x}^{(1)}[m] = \tilde{x}^{(1)}[m],$$

where $\widehat{C}$ stands for the action of crossover with a crossover rate $\chi = 1$.

The effect of crossover on the second order Walsh coefficient depends on the type of operator. For example, if the operator is that of uniform crossover, we have [16]

$$\widehat{C}\tilde{x}^{(2)}[m, m'] = \frac{1}{2}\{\tilde{x}^{(2)}[m, m'] + \tilde{x}^{(1)}[m]\tilde{x}^{(1)}[m']\}.$$

Thus, for uniform crossover, we obtain

$$D'[m, m'] \; = \; \frac{1}{2} D[m, m'],$$

where D' denotes the coefficient of linkage disequilibrium after crossover. For one point crossover, we have

$$D'[m, m'] \; = \; (1 - \frac{m' - m}{\ell - 1}) D[m, m'].$$

In any case, the absolute value of D decreases when crossover is included in a GA.

3.2 Variance of Fitness

The fitness function used in One Max problem is given by

$$f_i \; = \sum_{k=1}^{\ell} i_k \; = |i|.$$

Let us consider two statistical quantities of a population in the GA on One Max problem. One is the average fitness, and another is the variance of fitness. The average and variance of fitness at generation t are given by

$$\bar{f}(t) = E\{|i|\} = \sum_{i=0}^{n-1} |i| \, x_i(t)$$

and

$$\mathrm{VAR}(f) = V\{|i|\} = \sum_{i=0}^{n-1} |i|^2 \, x_i(t) \; - \; E\{|i|\}^2,$$

respectively .

To calculate these quantities, we define a generating function $G(s, t)$ [16]

$$G(s, t) \; = \sum_{i=0}^{n-1} \exp\{|i| s\} \, x_i(t).$$

It is convenient to calculate $G(s, t)$ by expanding $x_i(t)$ in terms of Walsh coefficient $\tilde{x}_j(t)$ defined in equation (5)

$$G(s, t) \; = \; \frac{1}{n} \sum_{i,j} \exp\{|i| s\} \, W_{ij} \, \tilde{x}_j(t).$$

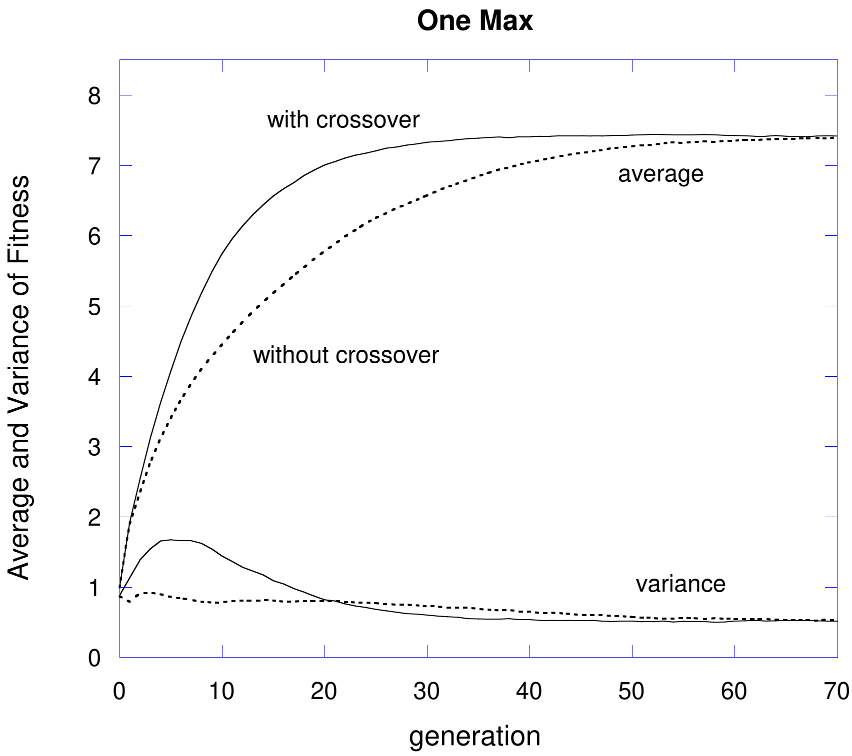

Figure 1: Average and variance of fitness in One Max problem . $\ell = 8$, $p = 0.01$, and $N = 200$. Solid lines show the results with crossover, and dotted lines without crossover.

Then, $\sum_i \exp\{|i|s\} W_{ij}$ is given by

$$\begin{aligned}\sum_{i=0}^{n-1} \exp\{|i|s\} W_{ij} &= \sum_i \prod_{k=1}^{\ell} e^{i_k s} (-1)^{i_k j_k} \\ &= (1+e^s)^{\ell-|j|}(1-e^s)^{|j|},\end{aligned}$$

where the identity

$$\sum_{i=0}^{n-1} = \sum_{i_1=0}^{1} \sum_{i_2=0}^{1} \cdots \sum_{i_\ell=0}^{1}$$

is used. From this, the generating function G is given by

$$G = \sum_{j=0}^{n-1} \left(\frac{1+e^s}{2}\right)^{\ell-|j|} \left(\frac{1-e^s}{2}\right)^{|j|} \tilde{x}_j(t). \qquad (16)$$

The average fitness $E\{|i|\}$ can be given by the partial derivative of this generating function. Noting

$$E\{|i|\} = \frac{\partial}{\partial s} G(s,t)\big|_{s=0},$$

and $1 - e^s = 0$ when $s = 0$, we obtain

$$E\{|i|\} = \sum_{m=1}^{\ell} p_m^{(1)} = \sum_{m=1}^{\ell} \frac{1-\tilde{x}^{(1)}[m](t)}{2}, \qquad (17)$$

Since $p_m^{(1)}$ is the probability of $j(m) = 1$, this result is quite reasonable.

Similarly, we can obtain an expression for the variance of fitness

$$V\{|i|\} = \frac{\partial^2}{\partial s^2} G(s,t)\big|_{s=0} - E\{|i|\}^2.$$

The variance $V\{|i|\}$ can be described as the sum of the additive variance V_A and epistatic variance V_I [7]. The definition of the epistatic variance is different from that given by Davidor [5].

An explicit form of the variance of fitness is

$$V\{|i|\} = V_A + V_I, \qquad (18)$$

where

$$V_A = \sum_{m=1}^{\ell} p_m^{(1)} (1 - p_m^{(1)}), \qquad (19)$$

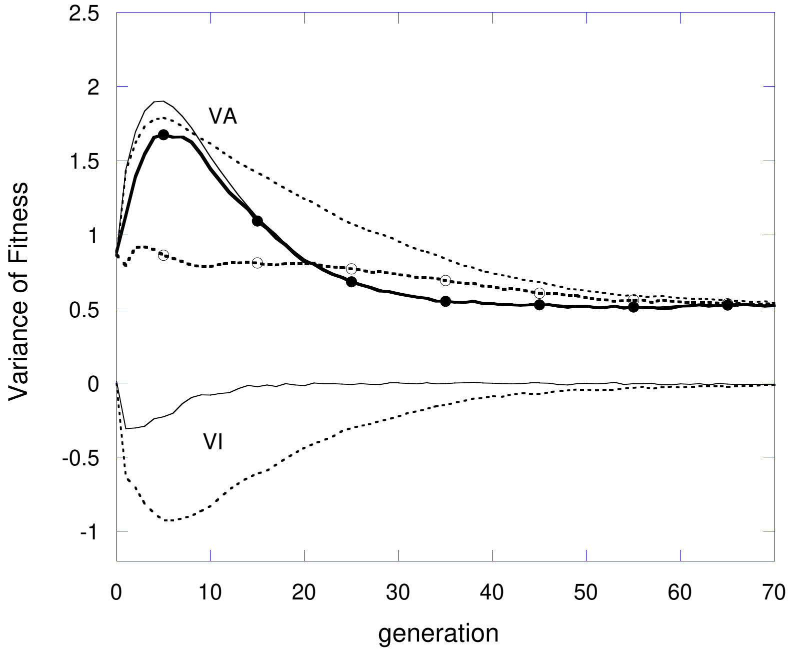

Figure 2: Variance of fitness V and its components V_A and V_I. Solid lines show the results with crossover and dotted lines those without crossover. Thick solid line with closed circles for the total variance V with crossover and thick dotted line with open circles for V without crossover.

and

$$V_I = \frac{1}{2}\sum_{m<m'}\{\tilde{x}^{(2)}[m,m'] - \tilde{x}^{(1)}[m]\tilde{x}^{(1)}[m']\} = 2\sum_{m<m'} D[m,m']. \tag{20}$$

The additive variance V_A is a function of terms $p_m^{(1)}$, and crossover does not have any effect on it. On the other hand, since the epistatic variance V_I is composed of the terms representing the interaction between loci, it may be changed drastically by crossover when some $D[m,m']$s are finite. Therefore, we expect that crossover changes the distribution of a population by reducing the epistatic variance.

We define the rate of change in the average fitness $\bar{f}(t) = E\{|i|\}$ over successive generations

$$v(t) \equiv \Delta\bar{f}(t) = \bar{f}(t+1) - \bar{f}(t). \tag{21}$$

4 NUMERICAL EXPERIMENTS

The numerical experiments for One Max problem were performed for comparing the results of the GAs with and without crossover. We used uniform crossover. Parameters of experiments are $\ell = 8$, $N = 200$ and a mutation rate $p = 0.01$. We carried out experiments with crossover rates $\chi = 0$ and 1. The GA calculations were performed repeatedly 100 times with different random numbers, and results were averaged over these runs. All members of the population were created with the probability of bit one $p_m^{(1)} = 1/\ell$ at the initial stage $t = 0$. Thus the population has an initial average fitness $\bar{f}(0) = 1$.

Figure 1 shows the average and variance of fitness $f(t)$ as functions of generation t with crossover probability $\chi = 1$ (solid line) and $\chi = 0$ (dotted line), respectively. This figure clearly demonstrates the advantage of crossover in the GA. The average fitness $\bar{f}(t)$ increases more rapidly in the GA with crossover than that without crossover. We also notice the large difference in the variance of fitness. The variance including the effect of crossover has a peak around $t = 7$, while the one without crossover is a slowly decreasing function of t.

Figure 2 demonstrates the components of the variances $V\{|i|\}$ shown in Fig.1. This figure includes the re-

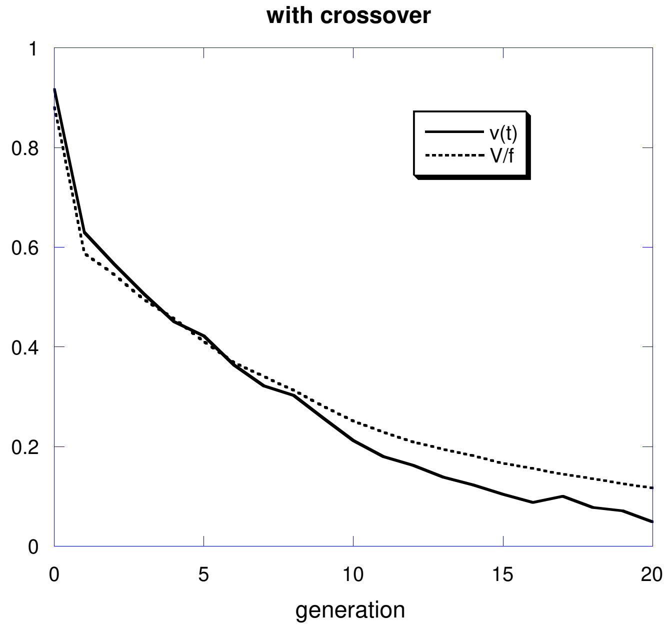

Figure 3: Examination of Fisher's theorem in the GA with crossover. Solid line for $v(t)$ and dotted line for $V/\bar{f}$.

sults for the total variance V, additive variance V_A and epistatic variance V_I. Solid lines represent the results with crossover, and dotted lines without crossover, respectively. The additive variances V_A in both calculations give large positive contribution to the total variances. The large difference is observed in the epistatic variance V_I. In both calculations, V_I is negative or zero at all generations. In the wide range of generations, two calculations show quite different behaviors. When the calculation without crossover starts, the variance V_I becomes negative and very small almost reaching to -1. Then the total variance V becomes very small as a result of the negative contribution of V_I. On the other hand, V_I in the calculation with crossover is almost zero except for the initial stage, and its contribution to the total variance V is very small. Therefore V is almost the same as the additive variance V_A.

As we have shown in equation (13), if mutation rate $p = 0$, the rate of change in average fitness $\Delta \bar{f}(t)$ is proportional to the ratio of the variance of fitness to the average fitness $V/\bar{f}$. In general, for finite mutation rate $p > 0$, this relation does not hold any more. However, when p is small, we expect that this relation still holds approximately.

Figures 3 and 4 give $v(t)$ and $V/\bar{f}$ as functions of generation t in calculations with and without crossover, respectively. These figures show the results in early stage of evolution $t \leq 20$. The solid line represents the speed of evolution $v(t) = \Delta \bar{f}(t)$, which is in good agreement with $V/\bar{f}$ given by dotted line. The speed of evolution $v(t)$ in Fig.4 is smaller than that in Fig.3, especially at the initial stage. As we have mentioned above, the effect of crossover appears in the total variance V through the negative contribution of the epistatic variance V_I. Since V without crossover is smaller than that with crossover, the speed of evolution becomes slow if crossover is absent.

5 SUMMARY

This paper was devoted to the analysis of crossover in the GA applied to One Max problem, in which the fitness function is linear. It has been shown that the analytical form of the variance can be obtained for linear fitness functions [4, 16]. In this case, the total variance V is decomposed into two parts, $V = V_A + V_I$. The first part V_A is the sum of the first order Walsh coefficients $\tilde{x}^{(1)}$, and is invariant under the action of crossover. On the other hand, the second part V_I includes the second order terms $\tilde{x}^{(2)}$, which may change if crossover is present. More precisely, the second part can be given in terms of the coefficients of linkage disequilibrium D. Since crossover makes the absolute value of D small, it reduces the magnitude of V_I.

In the practical situation, we found that the second

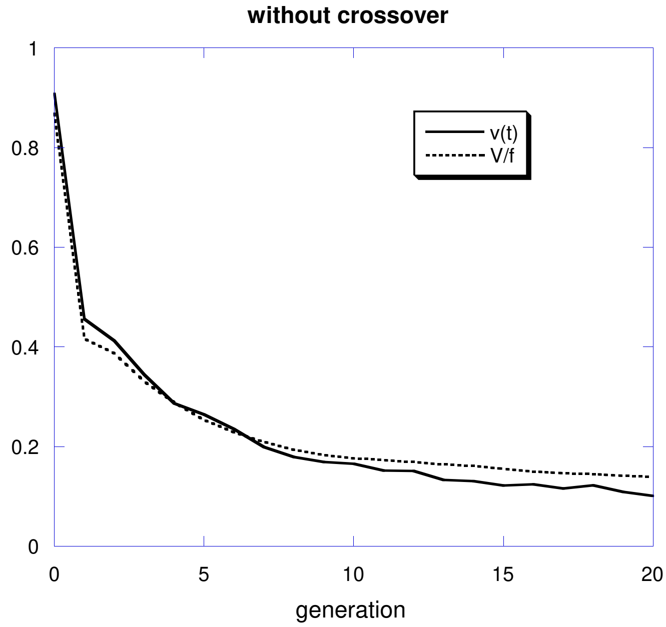

Figure 4: Examination of Fisher's theorem in the GA without crossover.

term V_I is negative in the calculations with and without crossover. In the case of the GA with crossover, the absolute value of V_I is small, almost zero except for regions of $t \leq 15$, while V_I is always negative in the GA without crossover. Therefore the total variance V with crossover is larger than that without crossover. This fact makes the large difference in the performance of the GA shown by the speed of evolution $v(t)$.

The essential point of this paper is that the epistatic variance V_I becomes negative in the GA without crossover. We have shown this result only experimentally, and cannot give any theoretical explanation for it. We are now trying to derive an analytical formulation for the mechanisms of generating linkage disequilibrium and inducing the negative V_I.

References

[1] Holland,J.H.: Adaptation in Natural and Artificial Systems, MIT Press, Massachusetts (1992).

[2] Goldberg,D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, Addison Wesley, Massachusetts (1989).

[3] Fisher,R.A.: The Genetical Theory of Natural selection, 2nd edition, Dover, New York (1958).

[4] Goldberg,D.E. and Rudnick,M.: Genetic Algorithms and the Variance of Fitness, Complex Systems, Vol.2, pp.265–278 (1988).

[5] Dvidor,Y.: Epistatic Variance: A Viewpoint on GA-hardness, Rawlins,G.J.E. (ed.) Foundations of Genetic Algorithms, pp.225–245 (1991).

[6] Aizawa,A.: Fitness Landscape Characterization by Variance of Decompositions, Belew,R.K. and Vose,M.D. (eds.) Foundations of Genetic Algorithms 4, pp.69–93 (1997).

[7] Maynard Smith,J.: Evolutionary Genetics, Oxford University Press, Oxford (1998).

[8] Weir,B.S.: Genetic Data Analysis II, Sinauer Associates, Massachusetts (1996).

[9] Munetomo,M. and Goldberg,D.E.: Linkage Identification by Non-monotonicity Detection for Overlapping Functions, Evolutionary Computation, Vol.7, pp.377–398 (1999).

[10] Furutani,H. : Application of Eigen's Evolution Model to Infinite Population Genetic Algorithms with Selection and Mutation, Complex Systems, Vol.10, pp.345–366 (1996).

[11] Furutani,H. : Analytical Solutions for Infinite Population Genetic Algorithms on Multiplicative

Landscape, Banzhaf,W., Daida,J., Eiben,A.E., Garzon,M.H., Honavar,M.H., Jakiela,M. and Smith,R.E. (eds.) Proceedings of the Genetic and Evolutionary Computation Conference, GECCO-99 Morgan Kaufmann, Vol.1, pp.204–211 (1999).

[12] Goldberg,D.E.: Genetic Algorithms and Walsh Functions: Part I, a Gentle Introduction, Complex Systems, Vol.3, pp.129–152 (1989).

[13] Goldberg,D.E. and Deb,K.: A Comparative Analysis of Selection Schemes Used in Genetic Algorithms, Rawlins,G.J.E. (ed.) Foundations of Genetic Algorithms, pp.69–93 (1991).

[14] Vose,M.D. and Liepins,G.E.: Punctuated Equilibria in Genetic search, Complex Systems, Vol.5, pp.31–44 (1991).

[15] Vose,M.D.: The Simple Genetic Algorithms, MIT Press, Massachusetts (1999).

[16] Furutani,H. : Study of Evolution in Genetic Algorithms by Eigen's Theory Including Crossover, Proceedings of 2000 IEEE International Conference on Industrial Electronics and Instrumentation, Nagoya, pp.2696–2703 (2000).