Linkage Learning, Overlapping Building Blocks, and Systematic Strategy for Scalable Recombination

Tian-Li Yu
Illinois Genetic Algorithms
Laboratory
University of Illinois at
Urbana-Champaign
104 S. Mathews Ave., Urbana,
IL 61801, USA
tianliyu@illigal.ge.uiuc.edu

Kumara Sastry
Illinois Genetic Algorithms
Laboratory
University of Illinois at
Urbana-Champaign
104 S. Mathews Ave., Urbana,
IL 61801, USA
kumara@illigal.ge.uiuc.edu

David E. Goldberg
Illinois Genetic Algorithms
Laboratory
University of Illinois at
Urbana-Champaign
104 S. Mathews Ave., Urbana,
IL 61801, USA
deg@illigal.ge.uiuc.edu

ABSTRACT

This paper aims at an important, but poorly studied area in genetic algorithm (GA) field: How to design the crossover operator for problems with overlapping building blocks (BBs). To investigate this issue systematically, the relationship between an inaccurate linkage model and the convergence time of GA is studied. Specifically, the effect of the error of so-called *false linkage* is analogized to a lower exchange probability of uniform crossover. The derived qualitative convergence-time model is used to develop a scalable recombination strategy for problems with overlapping BBs. A set of problems with circularly overlapping BBs exemplify the recombination strategy.

Categories & Subject Descriptors

G.1.6 [Mathematics of Computing]: Global Optimization–Analyze.

General Terms

Algorithms, Design, Theory.

Keywords

Genetic Algorithms, Linkage Learning, Model Building, Overlapping Building Blocks.

1. INTRODUCTION

In 1975, Holland [6] suggested that operators learning linkage information to recombine alleles might be necessary for genetic algorithm (GA) success. More recently, a GA design theory proposed by Goldberg, Deb, and Clark [2] indicated that proper problem decomposition is one of the keys to ensure the effectiveness of GA. Since then, many

such GA designs [1] that respect linkage have been developed. The linkage model can be implicit (*e.g.* linkage learning GA [5]) or explicit (*e.g.* linkage identification by nonlinearity check procedure [9]), probabilistic (probabilistic model building GAs [10], or estimation of distribution algorithms [7]) or deterministic (*e.g.* dependency structure matrix GA [18]). Some of these linkage learning methods should be capable of identifying overlapping building blocks (BB [1]), and DSMGA is known to have such ability.

A question that needs to be answered is: Given overlapping BBs, how do we design a scalable recombination operator to maximize BB mixing and minimize BB disruptions? Consider a problem where BB_1 and BB_2 are overlapped. If the crossover operator exchanges only BB_1, the information carried by BB_2 is disrupted. The same thing happens if crossover operator exchanges only BB_2. However, if BB_1 and BB_2 are exchanged together, there is no chance to recombine the information carried by BB_1 and BB_2. Here we are facing a decision-making problem between higher BB disruption or lower BB mixing, and we need a recombination strategy to help us make the decision.

To systematically design a recombination strategy for problems with overlapping BBs, we need to understand the effect of BB disruption and BB mixing on the GA convergence time, and the former has been studied [17]. This paper focuses on the latter issue, namely, how the mistaken concatenation of BBs (called *false linkage* in this paper) reduces the mixing rate, and how the lower mixing rate affects GA convergence.

This paper first shows the effects of several different crossover operators on an illustrative problem with overlapping BBs. Then it defines several commonly used terms and briefly recaps the results in [17]. The effect of false linkage is analogized to uniform crossover with a lower exchange probability. A number of experiments are done to verify the analogy. Given the qualitative convergence-time model by considering both BB disruptions and BB mixing, a recombination strategy is proposed for problems with overlapping BBs. The convergence time model can take care of the tradeoff, and more details can be found in section 6. Finally, discussions and possible future work conclude this paper.
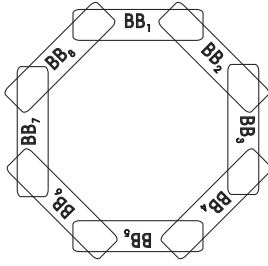
Figure 1: A problem with 8 overlapping BBs. The fitness of each BB is calculated by a trap₅ function. The overlapping length is 2. If we exchange BB_1 only, both BB_2 and BB_8 are disrupted.

2. EFFECTS OF DIFFERENT CROSSOVER OPERATORS ON A PROBLEM WITH OVERLAPPING BUILDING BLOCKS

Consider the following problem composed of overlapping trap₅ functions, where the trap₅ function is defined as:

$$trap_5(x_1 x_2 x_3 x_4 x_5) = \begin{cases} \frac{4-u}{5}, & u = 0, 1, 2, 3, 4 \\ 1, & u = 5 \end{cases}, \quad (1)$$

where x_i are binary values, and u denotes the number of ones among x_i's.

The overlapping length is 2, and the overlapping scheme is circular (figure 1). For example, for a problem with 3 BBs, the fitness is $trap_5(y_1 y_2 y_3 y_4 y_5) + trap_5(y_4 y_5 y_6 y_7 y_8) + trap_5(y_7 y_8 y_9 y_1 y_2)$, where y_i is a random permutation of genes x_i. Note that the arrangement of genes is randomly shuffled so the problem in general is not of tight linkage.

Three different crossover operators are tested: (1) allele-wise two-point crossover, (2) BB-wise uniform crossover, and (3) least-disruptive crossover. The allele-wise two-point crossover does not exploit the information of BBs. The BB-wise uniform crossover is similar to an ordinary allele-wise uniform crossover, but it performs on BBs instead of genes. The least-disruptive crossover performs like a two-point crossover on BBs. It randomly chooses two BB boundaries as cross sites on the circle, and swap the two partitions. The least-disruptive crossover is so named because it disrupts only 2 BBs during the recombination. It is not difficult to see that the least-disruptive crossover yields the least number of BB disruptions among any non-trivial recombination.

The testing problem contains 20 circularly overlapping BBs. The results shown in figure 2 are averaged out of 100 independent runs. Both the allele-wise two-point crossover and the BB-wise uniform crossover failed to find all correct BBs. On the contrary, the GA with the least-disruptive crossover successfully converged to the optimum. Note that we terminated the GA when 19 out of 20 BBs is correctly converged.

The least-disruptive crossover achieves the lowest BB disruption at the price of a lower mixing rate than the BB-wise uniform crossover because it concatenates many BBs as one big chunk during recombination. To fully understand why it outperforms other crossover operators on this type of problems, the effects of BB disruptions and BB concatenation need to be studied. A brief recap of the study of the relation-
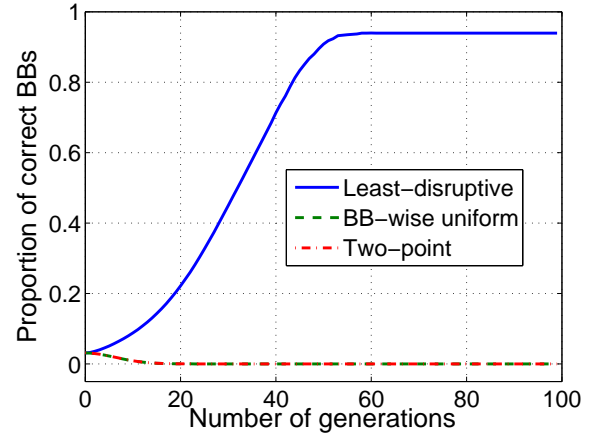

Figure 2: The convergence time for the circularly overlapping problem. Three different crossover operators are used. Both allele-wise two-point crossover and BB-wise uniform crossover fail, and the lines are overlapped. Only the least-disruptive crossover which respects both BBs and overlapping topology succeeds.

ship between BB disruptions and convergence time in [17] can be found in the next section.

3. ASSUMPTIONS AND THE ERRORS OF A LINKAGE MODEL

In this paper, several assumptions are made to simplify derivations:

- **Selectorecombinative GAs:** In this paper, crossover probability is always 1, and mutation operator is not taken into consideration. We focus only on the mixing behavior of the crossover operator.

- **Fixed-length χ-ary encoding:** The chromosome length is assumed to be fixed. The results can be applied to non-binary encoding, but the size of alphabets should be fixed.

- **Infinite population size:** This is a necessary assumption to use the convergence-time model derived in [8, 16]. To mimic the asymptotic behavior, we use population size 10 times larger than what the gambler's ruin model [4] predicts during all experiments. For finite population, readers are referred to [12].

- **BB-wise uniform crossover:** The BB-wise uniform crossover is similar to allele-wise uniform crossover, but now it acts on BBs not genes. The BB information is retrieved by some linkage learning algorithm, and it is not necessarily accurate. The least-disruptive crossover can be thought as a BB-wise uniform crossover with a linkage model which concatenates BBs into two chunks.

Next, we define some terms used in this paper. Some of them are popular in GA research field, but their meanings are different in different papers. The remainder of this section tries to clarify the meaning of these terms when they are used in this paper.

The term "linkage" is widely used in GA field, but defining linkage is not an easy task. In this paper, the linkage is roughly defined as follows [17]. *If linkage exists between two genes, recombination results in low-fitness offspring with high probability if those two genes are not transferred together from parents to offspring.* A group of highly linked genes forms a linkage group, or a building block (BB) in [1].

A linkage model is a model telling which genes form linkage groups. For instance, the boolean flags in LEGO [14], the genetic ordering in LLGA [4], the clustering model in eCGA [3], and the DSM clustering arrangement in DSMGA [18] are all linkage models. Two different types of errors can happen when a linkage model is adopted to describe the genetic linkage. One is that the linkage model does not link those genes which are linked in reality, which is called *detection failure*. The other is that the linkage model links those genes which are not linked in reality, which is called *false linkage*. Yu and Goldberg [17] has investigated the error of detection failure, and this paper primarily focuses on the error of false linkage.

The quality of a linkage model can be quantified by the number of errors it makes. For example, consider a problem with four BBs, where $\{BB_1, BB_2, BB_3, BB_4\}=\{\{1,2,3\}, \{4,5,6\} , \{7,8,9\}, \{10,11,12\}\}$, and a linkage model $\{BB_1', BB_2', BB_3', BB_4', BB_5'\}=\{\{1,2\}, \{3,4,5,6\}, \{7,8\}, \{9,10,11\}, \{12\}\}$. The linkage model has 3 detection failures and 1 false linkage.

Yu and Goldberg [17] indicated that if the linkage model has e_d detection failures, the convergence time t_{conv} elongates by a factor of $\left(1 - \frac{e_d}{I\sqrt{m}}\right)^{-1}$, where m is the number of BBs, and I is the selection intensity. When e_d is greater than $2I\sqrt{m}$, the GA will fail to find all correct BBs.

4. FALSE LINKAGE AND EFFECTIVE EXCHANGE LENGTH

The positions where false linkages occur can be either fixed or random. For example, in the four-BB problem from the previous section, if the model builder always links BB_1 and BB_2, that is, the linkage model is $\{1,2,3,4,5,6\}$, $\{7,8,9\}$, $\{10,11,12\}$, then the false linkage occurs at a fixed position. Since two BBs of order k are linked together, the population size required to ensure the presence of at least one copy of the correct BB_1 and BB_2 in the initial population is $O(2^{2k})$. Note that whereas there is no false linkage, the population size dictated by BB supply is $O(2^k)$. Even if we rely on GA operators to generate the correct allele values in BB_1 and BB_2, the number of mutations needed is still $O(2^{2k})$, and crossover does not help because of the inaccurate linkage model. The false linkage at fixed position occurs when the linkage-learning method is applied off-line, or the linkage model is biased to some particular positions for some reason.

On the other hand, if the linkage-learning method is applied on-line (*e.g.* every generation) and is non-biased to particular positions, the false linkages occur at random positions. In other words, the linkage model randomly links BBs by mistake when false linkages happen. This paper fo-

cuses on errors due to false linkage at random positions. It should be noted that when false linkage occur in random positions then population size dictated by BB supply does not change and is $O(2^k)$. If the linkage model mistakenly links two BBs together in the current generation, there is a non-zero probability that the linkage model would not link these two BBs in the next generation so that crossover operator will mix them. The remainder of this section will focus on how false linkage affects BB mixing.

Consider the following two scenarios: (1) the linkage model links BB_1 and BB_2, and (2) the linkage model identifies BB_1 and BB_2 correctly (as two separate BBs), and the BB-wise uniform crossover by chance transfers BB_1 and BB_2 together. These two scenarios produce the same crossover result, but with different probabilities. In the first scenario, the information of BB_1 and BB_2 are transferred with a probability 1; while the scenario does the same thing with a probability 0.5. The difference changes the *effective exchange length* (EEL) during crossover. EEL is defined as the effective number of BBs of which the information is exchanged during crossover. Note that in a problem with m BBs, the minimal EEL is 0 and the maximal EEL is $\frac{m}{2}$ because exchanging m' BBs is the same as exchanging $(m-m')$ BBs, where $m' > \frac{m}{2}$.

It is not difficult to calculate EEL for uniform crossover with perfect BB information. Suppose that the problem has m BBs. There is a probability of $\frac{C_0^m + C_m^m}{2^m}$ that the crossover exchanges 0 BB (and m BBs), where C_b^a is the binomial coefficient of the order-b term in an order-a binomial or called "a choose b". For simplicity, we assume m is even. The EEL can be then expressed as follows.

$$EEL_{unif} = 2^{-m}\left[\frac{m}{2} \cdot C_{\frac{m}{2}}^m + \Sigma_{i=0}^{\frac{m}{2}-1} i \cdot (C_i^m + C_{m-i}^m)\right]. \quad (2)$$

EEL_{unif} can be reduced by the following three arithmetic relations: (1) $C_b^a = C_{a-b}^a$, (2) $b \cdot C_b^a = a \cdot C_{b-1}^{a-1}$, where $a, b > 0$, and (3) $\Sigma_{i=0}^a C_i^a = 2^i$.

$$
\begin{aligned}
EEL_{unif} &= 2^{-m}\left[\left(2\Sigma_{i=1}^{\frac{m}{2}} i \cdot C_i^m\right) - \frac{m}{2} \cdot C_{\frac{m}{2}}^m\right] \\
&= 2^{-m}\left[\left(2\Sigma_{i=1}^{\frac{m}{2}} m \cdot C_{i-1}^{m-1}\right) - \frac{m}{2} \cdot C_{\frac{m}{2}}^m\right] \\
&= 2^{-m}\left[\left(m\Sigma_{i=0}^{m-1} C_i^{m-1}\right) - \frac{m}{2} \cdot C_{\frac{m}{2}}^m\right]. \\
&= \frac{m2^{m-1} - \frac{m}{2} C_{\frac{m}{2}}^m}{2^m} \\
&= \frac{m}{2}\left(1 - \frac{C_{\frac{m}{2}}^m}{2^m}\right). \quad (3)
\end{aligned}
$$

Consider the Stirling approximation [15], $m! \simeq \sqrt{2\pi m} \cdot m^m e^{-m}$. $C_{\frac{m}{2}}^m$ can be approximated as $2^m \sqrt{\frac{2}{\pi m}}$. Therefore, the EEL of uniform crossover can be approximated as

$$EEL_{unif} \simeq \frac{m}{2}\left(1 - \sqrt{\frac{2}{\pi m}}\right). \quad (4)$$

Unsurprisingly, for a large m, uniform crossover on average exchange $\frac{m}{2}$ BBs, which is the maximal information exchange.

Consider a problem with only 4 BBs. If the linkage model is perfect, the probability that crossover operator exchanges $\{0, 1, 2, 3, 4\}$ BBs is $\frac{1}{2^4}\{C_0^4, C_1^4, C_2^4, C_3^4, C_4^4\} = \frac{1}{16}\{1, 4, 6, 4, 1\}$ respectively (note that exchanging 3 BBs is
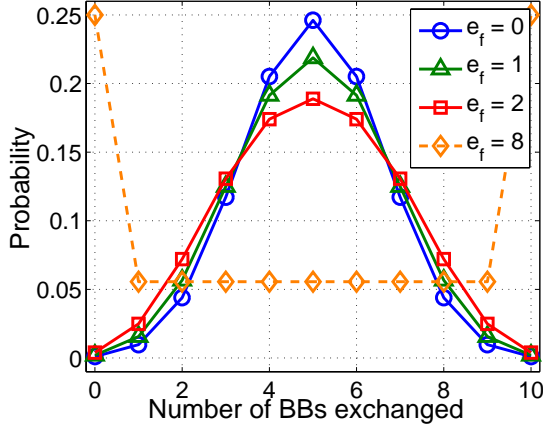
Figure 3: Probabilities of a certain number of exchanged BBs for a problem with 10 BBs. Parameter e is the number of false linkages in the linkage model. EEL for $e_f = 0$ is roughly $10/2{=}5$, and that for $e_f = 8$ is roughly $10/8{=}1.25$.

Figure 4: Relationship between EEL and false linkage count e_f. The problem is composed of 20 BBs.

the same as exchanging 1 BB). The EEL is $0 \cdot \frac{2}{16} + 1 \cdot \frac{8}{16} + 2 \cdot \frac{6}{16} = 1.25$. Now suppose that the linkage model mistakenly links BB_1 and BB_2. If BB_1 and BB_2 are not exchanged, the probability that $\{0, 1, 2, 3, 4\}$ BBs are exchanged is $\frac{1}{4}\{1, 2, 1, 0, 0\}$; if BB_1 and BB_2 are exchanged, the probability that $\{0, 1, 2, 3, 4\}$ BBs are exchanged is $\frac{1}{4}\{0, 0, 1, 2, 1\}$. Therefore, to sum up, the probability that $\{0, 1, 2, 3, 4\}$ BBs are exchanged given 1 false linkage is $\frac{1}{8}\{1, 2, 2, 2, 1\}$ respectively. The EEL is reduced from 1.25 (perfect linkage model) to 1.00 (1 false linkage).

When there are 2 false linkages, the calculation becomes somewhat more complicated, because the linkage model might have two linked chunks of two BBs, or might have one big chunk of three BBs. The effect of false linkage for a problem with 10 BBs is illustrated in Figure 3. The dashed line represents $e = 8$. Note that $(m - 2)$ is the maximal number of false linkages which still yields information exchange. The m BBs become only one chunk when there are $(m - 1)$ concatenations, and hence no information exchange is possible during crossover. When there are $(m - 2)$ false linkages for a problem with m BBs, the linkage model has only two big chunks of BBs. With a probability 0.5, these two chunks might be exchanged together (m BBs) or no exchange at all (0 BB). With a probability 0.5, exactly one chunk is exchanged, and the effect is similar to two-point crossover. The EEL is then roughly $\frac{1}{2}\frac{m}{4} = \frac{m}{8}$, where the $\frac{m}{4}$ comes from the EEL of two-point crossover.

The relationship between EEL and false linkage for a problem with 20 BBs is shown in figure 4. The EEL drops from roughly $\frac{m}{2} = 10$ to $\frac{m}{8} = 2.5$ for no false linkage ($e_f = 0$) to maximal false linkage ($e_f = (m - 2) = 18$).

To sum up, a perfect linkage model with uniform crossover has an EEL of $\frac{m}{2}$; and a nearly worst linkage model which contains $(m - 2)$ false linkages has an EEL of $\frac{m}{8}$. The worst linkage model with $(m - 1)$ false linkages has only one big chunk, and the GA does not work unless with a $O(2^\ell)$ population size (enumeration).
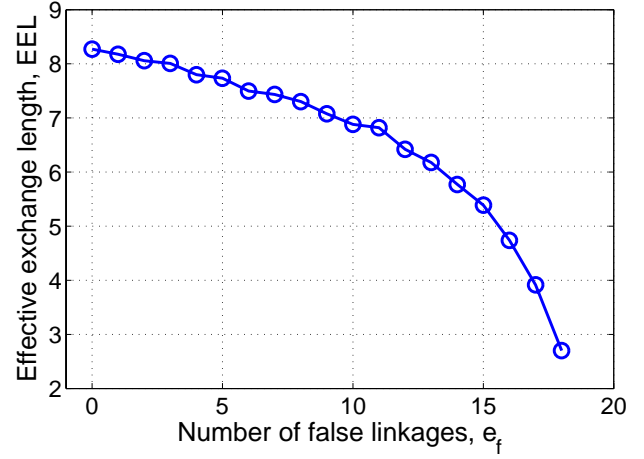
5. EFFECTIVE EXCHANGE LENGTH AND CONVERGENCE TIME

This section investigates the relationship between EEL and convergence time of GA. As shown in the previous section, uniform crossover with exchange probability 0.5 on average exchanges $0.5m$ BBs. This statement can be generalized for other exchange probabilities by examining the nature of binomial distribution. The mean of a binomial distribution is Ap, where A is the number of Bernoulli trials and p is the probability that the Bernoulli trial gives a true value. Although the calculation is somewhat different (we treat exchanging $(m - m')$ is the same as exchanging m' BBs), it is not difficult to show that for large m, the EEL of uniform crossover with exchange probability p is approximately mp by the similar derivations in the previous section. Therefore, the effect of crossover with EEL η (call it EEL-crossover) should be similar to uniform crossover with exchange probability $\frac{\eta}{m}$. The EEL-crossover simply randomly chooses η BBs to exchange. Figure 5 illustrates the similarity of these crossover operators. We can see that EEL-crossover with crossover length η is similar to uniform crossover with exchange probability η/m.

By assuming perfect mixing and infinite population size, the following convergence time model has been developed [8, 16].

$$t_{conv} = \frac{c_c \sqrt{\ell}}{I},\qquad(5)$$

where t_{conv} is the convergence time, ℓ is the problem size, I is the selection intensity, and c_c is a constant. Assuming the order of BBs k is fixed, the problem size ℓ is proportional to the number of BBs m. In the derivations of Eq 5, binomial distribution is assumed, *i.e.*, if current population has a proportion p of correct BBs, the variance of correct BBs for each individual is $\sqrt{mp(1 - p)}$. Rabani, Rabinovich, and Sinclair [11] derived bounds of the relaxation time that uniform crossover needs to make the population exactly random as follows.

$$\frac{\ln n}{2 \ln q^{-1}} \le \tau_{unif} \le \frac{2 \ln n}{\ln q^{-1}},\qquad(6)$$

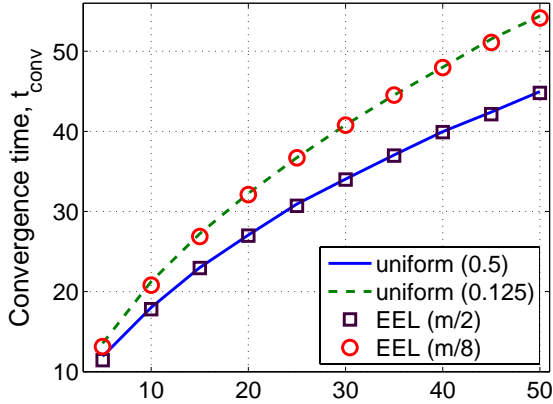where q is the probability that two positions are not sep-

Figure 5: Convergence time for different crossover operators. EEL-crossover with crossover length η has the roughly the same convergence time as uniform crossover with exchange probability η/m.

Figure 6: The relationship between convergence time and false linkage. The testing problem contains 20 BBs.

arated. In the case of uniform crossover with exchange probability p, $q = p^2 + (1-p)^2$. For the exchange probability reduced from $\frac{1}{2}$ to $\frac{1}{8}$, the relaxation time increases roughly by 2.8 times. Therefore, the distribution of correct BBs can no longer be modeled as a binomial distribution. The variance should be smaller than $\sqrt{mp(1-p)}$, and that makes GA converges more slowly. Followed by the notion of facetwise modeling in [1], and empirical findings (figure 5), Eq 5 still yields good approximation, but with a different constant $c'_c = r \cdot c_c$ when uniform crossover probability is not 0.5. We call parameter r the *elongation factor*. For m varying from 10 to 50, the elongation factor varies only from roughly 1.18 to 1.21. Therefore, the elongation factor is not tightly related to m, and treating r as invariant to m should give us a good-enough model to use.

From the previous section, we know that given an inaccurate linkage model with e_f errors of false linkage, the EEL η is reduced ($\eta < \frac{m}{2}$). The effect of the uniform crossover with a reduced EEL η is similar to the effect of the uniform crossover with exchange probability $p = \frac{\eta}{m} < \frac{1}{2}$. Then finally, it elongates the convergence time $t_{conv} = \frac{r \cdot c_c \sqrt{m}}{I}$. Figure 6 shows the relationship between e_f and t_{conv}. It is easily seen that η is a function of e_f, and r is a function of $p \simeq \frac{\eta}{m}$. Both relationships are non-linear and difficult to model quantitatively. However, the qualitative model provides a way to compare the impacts of the error of detection failure versus the error of false linkage.

Recall that $t_{conv}(e_d) = \frac{1}{1 - \frac{e_d}{2I\sqrt{m}}} \frac{\pi\sqrt{m}}{I}$ for the detection failure error, where the elongation factor r_d is $\frac{1}{1 - \frac{e_d}{2I\sqrt{m}}}$. The quality of a linkage model can be defined as the probability that the model produces an error (either detection failure or false linkage). We can calculate when one type of error dominates the other as follows.

$$
\begin{aligned}
& r_f < r_d \\
\Rightarrow \quad & r_f < \frac{1}{1 - \frac{e_d}{2I\sqrt{m}}} \\
\Rightarrow \quad & m < \frac{1}{\left[2I\left(1 - \frac{1}{r_f}\right)\right]^2} e_d{}^2.
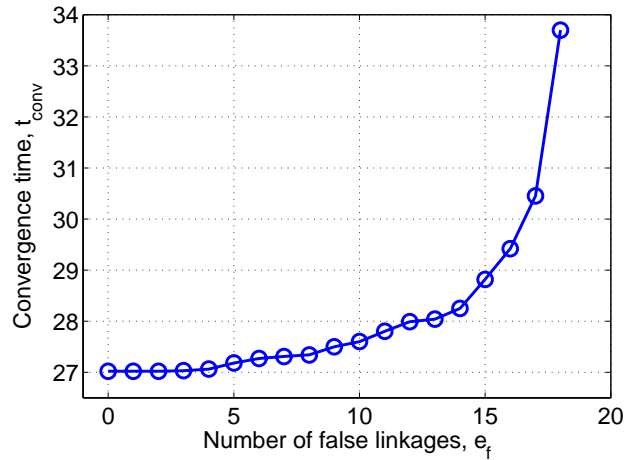\end{aligned}
\tag{7}
$$

Recall that r_f varies only from 1 to 1.2 for e_f from 0 to $(m-2)$. The term $\frac{1}{\left[2I\left(1 - \frac{1}{r_f}\right)\right]^2}$ can be treated like a constant over wide range of m. Call the constant c_s. In addition, assume that the linkage model has equal probabilities p to produce detection failure or false linkage errors. The quality Q of the linkage model can be then defined as $(1-p)$. Given that the number of detection failures ranges from 0 to m, Q can be approximated as $1 - p \simeq 1 - \frac{e_d}{m}$. The above inequality relation can be then re-expressed as

$$
m > \frac{1}{c_s(1-Q)^2}. \tag{8}
$$

Given the empirical findings that selection intensity $I \simeq 0.5$ for problems composed of $trap_5$ functions, a control map can be drawn (figure 7). In most problems, the detection failure error affects the convergence time more severely than false linkage. The error of false linkage dominates only when the problem size is small and the linkage model is very accurate.

To sum up, two important things can be learned from the qualitative model. First, the error of false linkage results in a longer convergence time, and for a specific e_f, the elongation factor r_f is a constant independent of the problem size. Second, for many problems, the linkage learning algorithm should put more efforts on eliminating detection failure than false linkage. This argument is used to develop the a recombination strategy for problems with overlapping BBs in the next section.

6. RECOMBINATION STRATEGY FOR PROBLEMS WITH OVERLAPPING BBS

Given the argument that the detection failure error elongates the convergence time much more severe than false linkage, the recombination strategy is proposed as follows: Treat the whole problem as two big BB chunks. If we want to increase the mixing rate and hence increase the number of cross sites, the number of BB disruptions would grow very fast. Therefore, the recombination strategy when dealing with a group of overlapping BBs is as follows:
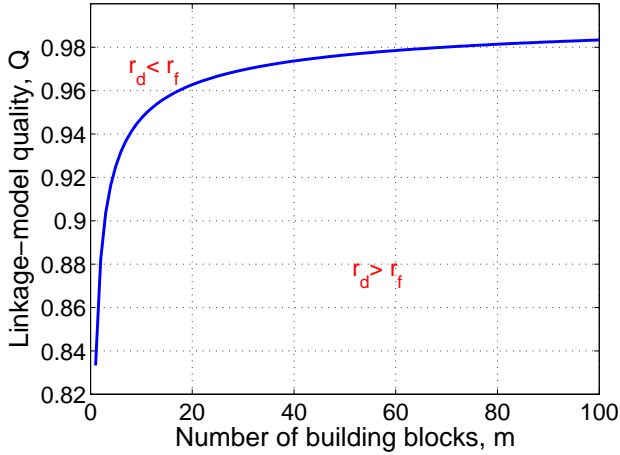
Figure 7: **The control map of the two types of errors, where r_d is the elongation factor of detection failure, and r_f is that of false linkage. In most problems, the detection failure error affects the convergence time more severe than false linkage. The error of false linkage dominates only when the problem size is small and the linkage model is very accurate.**

1. Perform BB identifying algorithm to capture the overlapping topology.

2. Construct a graph $G = (V, E)$ where the nodes are BBs, and the edges are overlapping relations between BBs. There is an edge between two BBs if and only if the two BBs overlap.

3. Randomly choose two nodes n_1 and n_2. Then partition the graph G into two subgraphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ which satisfy the conditions: $n_1 \in V_1$, $n_1 \in V_1$, and $|E| - |E_1| - |E_2|$ is minimal.

In other words, the two chunks need to be chosen at random, and the choice disrupts minimal number of overlapping BBs.

The results of the GA using the above recombination strategy for the circularly overlapping problem is shown in figure 8. The solid line is the GA convergence theory. The dashed line considers only the effect of false linkage, and the dotted line considers only the effect of detection failure. The elongation factors r_d and r_f are empirically measured and plotted in figure 9. Since $e_d = 2$ is a constant, r_d tends to be 1 when m goes to infinity, and r_f remains constant.

According to Yu and Goldberg [17], the elongation factor r_d caused by detection failure can be approximated as $\left(1 - cm^{-0.5}\right)^{-1}$, where c is a constant decided by the selection intensity, the number of misidentified BBs. We can verify the above approximation by the following derivations.

$$t_{conv} = \frac{\pi\sqrt{m}}{I} / \left(1 - cm^{-0.5}\right) \tag{9}$$

$$\Rightarrow \quad cm^{-0.5} = 1 - \frac{\pi\sqrt{m}}{I t_{conv}} \tag{10}$$

$$\Rightarrow \quad c_1 - 0.5\log(m) = \log(1 - \frac{\pi\sqrt{m}}{I t_{conv}}). \tag{11}$$

$$\tag{12}$$

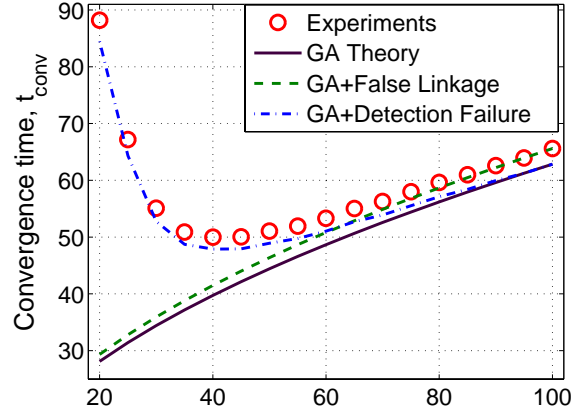If we plot $\log(m)$ versus $\log(1 - \frac{\pi\sqrt{m}}{I t_{conv}})$, we should get a

Figure 8: **The convergence time using the proposed recombination strategy for the circularly overlapping problem with different problem size. The solid line is the GA convergence theory. The dashed line considers only the effect of false linkage, and the dotted line considers only the effect of detection failure.**
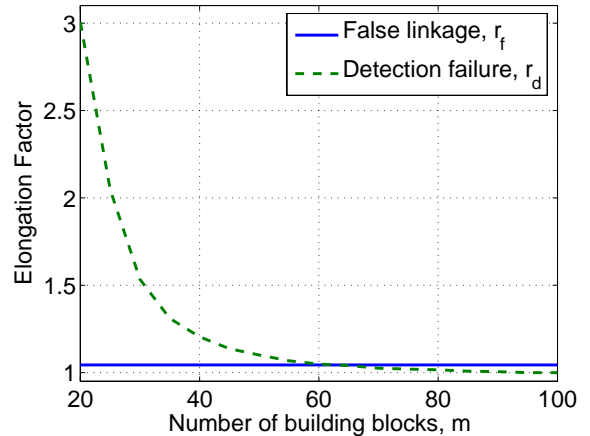

Figure 9: **The elongation factor r_d and r_f for detection failure and false linkage respectively. r_d decreased as m increases, and r_f remains constant.**
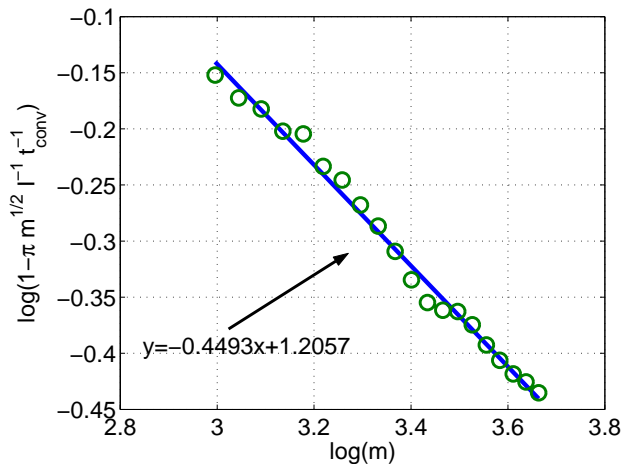
Figure 10: The slope of the trend line is roughly -0.45 empirically. In other words, the elongation factor r_d can be approximated as $(1 - cm^{-0.5})^{-1}$.

straight line with slope -0.5. Empirically, the slope is found be roughly -0.45 (Figure 10), which validates the approximation.

Note that the above strategy seems similar to an ordinary two-point crossover, but it has a significant difference. The ordinary two-point crossover does not respect the concept of linkage. It will succeed only for problems with tight linkage. For problems with overlapping BBs and random linkage, (1) if no BB-identification method is applied, the GA will fail, and (2) if BBs are correctly identified but careless crossover (which does not respect the overlapping topology, *e.g.*, a blind BB-wise uniform crossover), the GA will fail, too. When dealing with overlapping BBs, crossover operators that respect both BBs and overlapping topology are necessary for GA success. Finally, the strategy of splitting the group of overlapping BBs into two chunks is only applied to those overlapping BBs. If a problem contains overlapping BBs and non-overlapping BBs, BB-wise uniform crossover should be applied to non-overlapping BBs to achieve maximal mixing.

7. CONCLUSIONS

This paper systematically designs a recombination strategy for problems with overlapping BBs. To achieve the goal, the effect of false linkage on convergence time is investigated. A linkage model with more false linkages causes a lower effective exchange length (EEL) during the crossover. The reduced EEL has a similar effect as uniform crossover with a lower exchange probability. With the results of the earlier work [17], a qualitative convergence-time model for both detection failure and false linkage is given. The elongation factor of false linkage is constant in terms of the problem size. The control map of the two types of error is given. For most of the problems, the impact of detection failure dominates that of false linkage. The error of false linkage dominates only when the problem size is small and the linkage model is very accurate. Finally, the convergence-time model is then used to develop a recombination strategy for problems with overlapping BBs.

In this paper, only a qualitative model of the convergence

time is derived. To be able to derive a quantitative model, a better understanding of the relationship between mixing rate and convergence time is needed, which is one of the highlighted issues in GA research area. Once such a model of mixing rate versus convergence time is done, it can be plugged into this framework, and all the arguments should still be valid.

Combined with the earlier work [17], this paper investigates how an inaccurate linkage model affects GA convergence. Nevertheless, it is known that BB disruptions and BB mixing also affect the GA population sizing [13]. To be able to calculate the total number of function evaluations, we need to investigate the relationship between GA population sizing and the two types of error: detection failure and false linkage.

8. REFERENCES

[1] D. E. Goldberg. *The design of innovation: Lessons from and for competent genetic algorithms.* Kluwer Academic Publishers, Boston, MA, 2002.

[2] D. E. Goldberg, K. Deb, and J. H. Clark. Genetic algorithms, noise, and the sizing of populations. *Complex Systems*, 6:333–362, 1992.

[3] G. R. Harik. Linkage learning via probabilistic modeling in the ECGA. IlliGAL Report No. 99010, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL, 1999.

[4] G. R. Harik, E. Cantú-Paz, D. E. Goldberg, and B. L. Miller. The gambler's ruin problem, genetic algorithms, and the sizing of populations. *Proceedings of 1997 IEEE International Conference on Evolutionary Computation*, pages 7–12, 1997.

[5] G. R. Harik and D. E. Goldberg. Learning linkage. *Foundations of Genetic Algorithms 4*, pages 247–262, 1996.

[6] J. H. Holland. *Adaptation in natural and artificial systems.* University of Michigan Press, Ann Arbor, MI, 1975.

[7] P. Larrañaga and J. A. Lozano, editors. *Estimation of Distribution Algorithms.* Kluwer Academic Publishers, 2002.

[8] H. Mühlenbein and D. Schlierkamp-Voosen. Predictive models for the breeder genetic algorithm: I. Continuous parameter optimization. *Evolutionary Computation*, 1(1):25–49, 1993.

[9] M. Munetomo and D. E. Goldberg. Identifying linkage groups by nonlinearity/non-monotonicity detection. *Proceedings of the Genetic and Evolutionary Computation Conference 1999: Volume 1*, pages 433–440, 1999.

[10] M. Pelikan, D. E. Goldberg, and F. G. Lobo. A survey of optimization by building and using probabilistic models. IlliGAL Report No. 99018, University of Illinois at Urbana-Champaign, Illinois Genetic Algorithms Laboratory, Urbana, IL, 1999.

[11] Y. Rabani, Y. Rabinovich, and A. Sinclair. A computational view of population genetics. *Random Structures and Algorithms*, 12:314–330, 1998.

[12] M. Rattray and J. L. Shapiro. Noisy fitness evaluations in genetic algorithms and the dynamics of learning. *Foundations of Genetic Algorithms*, 4:117–139, 1997.

[13] K. Sastry and D. E. Goldberg. How well does a single-point crossover mix building blocks with tight linkage? *Proceedings of the International Symposium on Computer and Information Science*, 2002.

[14] J. Smith and T. C. Fogarty. Recombination strategy adaptation via evolution of gene linkage. *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pages 826–831, 1996.

[15] J. Stirling. Methodus differentialis. 1730.

[16] D. Thierens and D. E. Goldberg. Convergence models of genetic algorithm selection schemes. In *Parallel Problem Solving fron Nature, PPSN III*, pages 119–129, 1994.

[17] T.-L. Yu and D. E. Goldberg. Toward an understanding of the quality and efficiency of model building for genetic algorithms. *Proceedings of the Genetic and Evolutionary Computation Conference 2004*, pages 367–378, 2004.

[18] T.-L. Yu, D. E. Goldberg, A. Yassine, and Y.-p. Chen. Genetic algorithm design inspired by organizational theory: Pilot study of a dependency structure matrix driven genetic algorithm. *Proceedings of Artificial Neural Networks in Engineering 2003 (ANNIE 2003)*, pages 327–332, 2003. (Also IlliGAL Report No. 2003007).