Recombination Strategy Adaptation via Evolution of Gene Linkage

Jim Smith  
Faculty of Computer Studies & Mathematics  
University of the West of England  
Bristol, BS16 1QY  
U.K.  
jim@ics.uwe.ac.uk

T.C. Fogarty  
Faculty of Computer Studies & Mathematics  
University of the West of England  
Bristol, BS16 1QY  
U.K.  
tcf@ics.uwe.ac.uk

Abstract

This paper analyses recombination strategies resulting from evolving gene linkage on a variety of fitness landscapes with known properties. The evolution of “blocks” of linked genes within a gene pool makes it possible for the recombination strategy to vary in both the amount of genetic material exchanged during the formation of a new individual, and in the number of parents contributing genetic material. The strategies evolved are examined in the light of the known properties of the landscapes.

The results obtained explain the previously recorded good performance of the algorithm as a self-adaptive mechanism for function optimisation, as the amount and type of recombination is able to adapt to suit the landscape being searched unlike conventional “fixed” operators.

1: Introduction

In recent years the interaction between the Genetic Algorithm (GA) and Evolutionary Strategies communities has led to an increasing interest in the use of adaptive operators within genetic algorithms e.g. [1, 2, 3], as it has been recognised that the optimal rate of application of a given operator will depend both on the nature of the landscape being searched and on the current state of the search.

The aim of this work is to develop a representation and operators which allow for self-adaptation of the genetic search to the nature of the landscapes. This is intended to provide a robust function optimisation method which is not prone to the poor performance that “fixed” operators can show on certain types of landscape.

This paper examines a model within which the evolution of heritable “links” between neighbouring genes provides the basic means for the self-adaptation of recombination strategies. Measurement of the amount of gene linkage also provides a natural means of analysing the recombination behaviour of the algorithm as evolutionary time progresses.

This algorithm has already been shown to be competitive against a variety of other recombination strategies and across a range of mutation rates when tested in the setting of a generational GA [4]. Subsequent experiments have shown that this potentially disruptive mechanism works better in a “steady state” setting, a result which was also found in work on multi-parent recombination techniques [5]. This paper therefore concentrates on analysing how the nature of the search adapts to different types of landscapes.

Kauffman’s NK model is used to provide a means of creating and characterising the landscapes searched in terms of the amount of interaction between loci (epistasis). This is briefly described in Section 3.

2: The LEGO Model for Adaptive Evolution of Gene Linkage

The encoding and recombination used for this work, and described below, is the “LEGO” model (Linkage Evolving Genetic Operator) [4]. The recombination operator works by considering the GA’s population as a gene pool comprised of “blocks” of genes defined over certain loci, with these blocks potentially varying in size from a single gene to an entire chromosome. In order to achieve this each gene has associated with it two boolean flags determining whether it will link to the genes to its left and right, and two adjacent genes are said to be linked if the appropriate flags are both set to “true”. Blocks are thus comprised of chains of linked genes. This representation is similar to Schaffer and Morishima’s use of “punctuation marks” to encode for crossover points [1], however in this case the recombination is not restricted to two parents.

A new individual is created by holding a series of competitions to fill its loci, starting at the left hand edge. When a given position is filled a choice is made from all the blocks eligible to fill that space (i.e. whose left-most defined position is at the locus being filled) based on their fitnesses. The entire block (with its links) will be copied. A new competition is then held to fill the next locus after the end of the block. Since each block comes from a complete parent, it will always be possible to find at least one suitable candidate block for each competition. This is shown schematically in Figure 1.

If all of the blocks are of unit length (no links) then this reduces to considering the problem as a set of separate populations of alleles (one for each locus) which is the basis of Bit Simulated Crossover [6]. At the other extreme, if a chromo-
some is fully linked (one block extends over its whole length) then selection of this block for reproduction will result in an asexual strategy i.e. mutation only evolutionary search.

The population is initialised so as to provide a variety of different block sizes at the outset, and the linkage flags are also subject to mutation, allowing a variety of possible recombination strategies to be explored.

It is necessary to use two flags to determine linkage, since this provides a means of building up longer blocks which would otherwise have a greater tendency to be destroyed by mutation. This merging occurs when two blocks (from separate parents) which have their appropriate flags set true are juxtaposed during the creation of a new individual.

These two mechanisms for changing the amount of material kept together during recombination provide enough variety for the evolution of successful recombination strategies without the need to keep statistics on the relative performance of offspring etc. as is commonly done in adaptive operator paradigms e.g. [7, 8, 9].

3: The NK Landscape Model

Kauffman’s NK model of fitness landscapes has been used by a number of researchers to investigate properties of the genetic algorithm e.g. [10, 11]. The abstract family of landscapes defined can be shown to be generic members of the AR(1) class of combinatorial optimisation landscapes. The reader is directed to [12] for a fuller description.

In brief the fitness of an individual is considered to be the average of the fitness contributions of its N loci. The fitness of each loci is dependent on K other loci, and this “epistasis” is considered to be so complex that is modelled by creating a look-up table of random fitnesses. This table has an entry for each of the $2^{k+1}$ possible combinations of values (assuming a binary representation) of bits in that locus and the K other loci on which it depends. Each landscape is thus defined by a lookup table of size $N \times 2^{k+1}$.

For K=0, there is no epistasis and only one optimum, but as K is increased to N-1 the landscape becomes more and more rugged with increasing numbers of lower optima and decreasing correlation between the fitness of neighbouring points in the landscape, as changes in the genotype value of one position affect the fitnesses of more and more other positions. At the limit of K = N -1 the landscape is entirely uncorrelated.

A further refinement can be made by considering the nature of the epistatic interactions. These can be adjacent (i.e. the K bits to the right of the locus are considered) or random (i.e. for each position K values are initially defined at random as being the epistatic loci for that position). Note that for the adjacent epistasis the chromosome is considered to be a circle for this purpose, and that when K = 0 or K = N-1, the two types of epistasis are the same.

Kauffman provides an empirical investigation of the structure of fitness landscapes by plotting the relative fitness of local optima against their Hamming distance from the fittest found optima for the value of N = 96. He summarises the results as “for K small - for instance, K = 2- the highest optima are nearest one another. Further, optima at successively greater Hamming distances from the highest optimum are successively less fit... our landscape here possesses a kind of Massif Central, or high region, of genotype space where all the good optima are located. As K increases this correlation falls away, more rapidly for K random than for K adjacent” ([12] pp61-62).

4: Experimental Conditions

In order to investigate the behaviour of the algorithm the population was scanned after every 20 evaluations and each individual assigned to one of the nominal classes (0-24%, 25-49%, 50-74% and 75+%) according to the proportion of its genes which were linked.

The algorithm was run for 10,000 evaluations using a “steady state” GA with both interaction models with a string length (N) of 16 and values for K of 0, 4, 8 and 15. The GA had a population of 100, and used a “delete-oldest” replacement strategy. Both problem representation and linkage genes
were subjected to (value flipping) mutation at a rate of 0.1% probability per position. Blocks were selected using Fitness Proportionate selection using the scaling

$$\text{expected} = 1.0 + (\text{fitness} - \text{avege}) / (\text{avege} - \text{worst})$$

For each value of K the averages of 20 runs were made, each run with a different landscape.

## 5: Population Analysis

Figure 2 shows the behaviour of the algorithm on landscapes with adjacent interactions for the four values of K. Only the first 2000 evaluations are shown as after this the populations remain stable before drifting under mutation towards 50% linkage in the converged populations.

In each case there is an initial “spike” in the number of highly linked chromosomes (the 75+% population), as a result of the “free” links joining together to create longer blocks. However, as can be seen from figure 2a, for the smooth unimodal problem (K =0), the population is taken over rapidly by individuals with very low linkage, i.e. the gene pool is dominated by small blocks as the problems of optimising the fitness at each loci are completely separable. The decline in the number of long blocks can be explained by the fact that once more than 50% of the bits are optimal, mutation will on average have a deleterious effect and so an asexual reproduction policy will yield diminishing returns.

For the value of K = 4, where the landscape is more rugged, the nature of the population has changed, and the snapshots of the population reveal far more individuals falling into the higher linkage classes. This represents a gene pool composed of longer blocks, which in turn means that the creation of a new individual will involve the recombination of “information” from fewer parents and will be more likely to preserve long schema.

For the very rugged problems with K = 8 and K = 15 this trend is even more pronounced, with the population being increasingly dominated by individuals composed of very long blocks (75+% linkage). Since many of these blocks will be incompatible with each other, this corresponds to a much reduced, or zero probability of recombination i.e. the search proceeds mainly via mutation. This is a result of the landscape containing so many optima in uncorrelated positions that the chances of the successful juxtaposition of two sub-solutions via recombination becomes very small.

The results described above show that as K increases the more disruptive recombination strategies (represented by individuals composed of series of small blocks) are increasingly less successful whereas the evolution of links between co-adapted genes is increasingly favoured. This can be understood in the light of the first part of Kauffman’s findings, namely that the concentration of peaks into a small region of space lessens as K is increased and the fitness correlation of the landscape decreases. Therefore combining information from two fit parents is less likely to be a successful policy.

The second part of the statement, that the correlation falls away more rapidly for random interactions than adjacent, might be expected to translate in this setting to a faster decline.
in the fortunes of the smaller blocks in the gene pool. If so for any given value of K between 2 and 15, there would be a noticeable difference between the results for the two different types of interactions. Figure 3 compares the behaviour of the algorithm with the two types of interaction for K = 4 and K = 8.

For the smoother landscape the random interactions (figure 3c) show populations with consistently higher linkage than adjacent interactions (figure 3a). Unlike the populations on the adjacent landscape, those on the random landscape show a virtually zero incidence of individuals exhibiting low gene linkage (0-24%). This shows that recombination has proved comparatively less successful on the “random” landscape as a result of the greater distance between peaks.

For the very rugged landscapes, both searches are initially dominated by highly linked individuals (long blocks). However after this initial phase the populations on the “random” landscapes show a markedly faster decline towards a gene pool with a wider mixture of block sizes than do the populations adapting on the “adjacent” landscapes. If Kauffman’s empirical observations are valid, there are two possible reasons for this apparent anomaly - one (and this is a potential criticism of the LEGO model) being that the system is unable to adapt when the nature of the epistasis does not match the linear pattern of the linkage. The second is that the effects of selection and convergence on the population have outweighed selection pressure towards the formation of longer blocks.

In order to investigate the behaviour of the algorithm on the two landscapes more fully, the experiments were repeated but this time the number of parents contributing blocks to the make-up of each new individual was measured as an average over each period of 20 evaluations. This provides an alternative measure of the evolving search strategies since there are two scenarios in which the degree of linkage can be misleading. The first is the case of a highly converged population (Schaffer & Morishima noted this effect as they found that the number of crossover points encoded in their populations increased although the number of “productive” crossover operations did not). The second is the opposite case where an individual may be reproduced whole due to a lack of competition at each of its block boundaries. In this case even a moderately linked population could preserve discrete niches through lack of inter-niche recombination. In figure 4, the absolute (total) and distinct (ignoring identical copies) numbers of parents are plotted for K values of 2, 4 and 8.

From these plots the following observations can be made: i) For both types of landscape the absolute amount of recombination decreases with increasing K - i.e. as K increases so longer linked blocks of co-adapted genes dominate the gene pool as noted above. ii) In all but one case (the exception being K = 2 Random Epistasis) the effect of converging (sub) populations is that the distinct number of parents is reduced to one i.e. as Hordijk & Manderick noted [10] the second phase of the search is mutation led. iii) For both types of landscapes the rate at which this happens...
increases as K is increased. These observations tie in with the first part of Kauffman’s findings on the structure of the landscape and observation (iii) ties in with the theoretical result that the mean walk length to a local optima decreases as K increases ([12] p57) since this implies that populations using an asexual strategy will be able to locate and explore local optima more quickly.

iv) For K= 2 the absolute number of parents involved in recombination is lower on the “random” landscapes (i.e. there is more gene linkage) but the distinct number is higher, falling after 2000 evaluations to 2 parents as opposed to 1 one the “adjacent” landscape. Despite this the “mean score” curves (not shown) are very similar. This suggests that on the “random” landscape the algorithm has maintained sub-populations on at least two optima of similar height. This ties in with Hordijk & Manderick’s findings that in fact the “central massif” for the case of K = 2 contains more peaks of a similar height close to the global optima in the random case than in the adjacent case.

v) For K = 4 the absolute number of parents involved in recombination is again consistently lower in the populations adapting on random landscapes, i.e. the amount of gene linkage is again higher. However this difference is much lessened.

vi) For K = 8 there is a different situation. The numbers of distinct parents involved are very similar in the two cases, falling rapidly to 2 and then declining slowly as the population converges. In both cases the absolute number of parents (inversely related to mean block size) falls rapidly, but then climbs once the population has converged. However for this value of K the absolute number of parents climbs to a higher value for “random” landscapes than for “adjacent” landscapes. This means that although the recombination mechanism evolved in each case is effectively sampling the same number of parents, the number of blocks picked is greater for random epistasis. This means that in terms of schema bias analysis [13] the recombination strategies which evolve via gene linkage show less positional bias for random epistasis than for adjacent epistasis.

Although space and clarity prohibits the display of the mean scores of the algorithm running on the two different types of landscapes, they are very similar, and the mean fitnesses after 2000 evaluations are not significantly different (using Student’s t-test at the 5% level).

These experiments have been repeated with a variety of selection mechanisms and very similar results have been obtained in each case. Comparisons on the model within a function optimisation setting showed it to be fairly insensitive to changes in mutation rates.

6: Conclusions & Discussion

The analysis of the behaviour of the algorithm suggests that the recombination strategies which evolve using this model correspond closely to what might be expected from theoretical and empirical investigation into the structure of the fitness landscapes concerned. In general it can be seen that as the degree of correlation of the fitness landscape decreases, so there is an increasing tendency for longer blocks of linked genes to take over the genepool. Since individuals created
from large blocks will necessarily involve fewer (absolute) parents than those created from more, smaller blocks, they will be less likely to combine information from a number of different peaks. Conversely if there is low epistasis and high fitness correlation then smaller blocks will take over the geneepool as individuals created from these are less likely to suffer from the deleterious effects of genetic “hitchhiking”.

However analysis of the strategies evolved over the course of the search also demonstrates that the most successful recombination strategy is not simply a function of the landscape being searched but of the distribution of the population over that landscape.

An obvious potential criticism of this model is that is geared around links between adjacent bits, and that on problem representations which do not follow this pattern it will perform poorly, and be unable to adapt to the structure of the landscape. However the results obtained with random interactions show that on landscapes where this would be expected to be a problem (i.e. high K) successful individuals are created without the extra overhead of a mechanism to allow links between any two loci. This is because with a nearly converged population the gene linkage is able to adapt according to the utility of positional bias in recombination.

This paper demonstrates an abstract class of problems for which the use of fixed recombination operators at fixed rates are strictly limited in their performance by the matching of operators and application rates to the nature of the landscape. For for most “real-world” problems (in which the structure of the landscape is not known) this will be very difficult. This suggests that despite the extra overhead of learning gene linkage, the recombination mechanism described above may provide a useful tool for self adaptive optimisation and studying new landscapes.

7: References:
14 Moran, Moreno, Merelo & Chacon (eds) 1995 “Advances in Artificial Life: Proceedings of Third European Conference on Artificial Life”, eds, Springer Verlag