Species Adaptation Genetic Algorithms: A Basis for a Continuing SAGA

Inman Harvey

CSRP 221, 1992

Cognitive Science Research Paper

Serial No. CSRP 221

The University of Sussex
School of Cognitive and Computing Sciences
Falmer
BRIGHTON BN1 9QH
England, U.K.

This paper appears in:
"Toward a Practice of Autonomous Systems",
Proceedings of First European Conference on Artificial Life,
F.J. Varela & P. Bourgine (eds.), MIT Press/Bradford Books, 1992

Species Adaptation Genetic Algorithms: A Basis for a Continuing SAGA.*

Inman Harvey
School of Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QH, England
email: inmanh@cogs.susx.ac.uk

Abstract

For Artificial Life applications it is useful to extend Genetic Algorithms from a finite search space with fixed-length genotypes to open-ended evolution with variable-length genotypes. A new theoretical analysis is required, as Holland's Schema Theorem only applies to fixed lengths. It will be argued, using concepts of epistasis and fitness landscapes drawn from theoretical biology, that in the long run a population must have genotypes of nearly equal length, and this length can only increase slowly. As the length increases, the population will be nearly converged, and hence evolving as a species.

1 Introduction

Genetic algorithms (GAs) are a form of search technique, primarily used for function optimization, modelled on Darwinian evolution. Some basic knowledge of GAs, is assumed for the purposes of this paper; the best introduction is (Goldberg 1989). Holland's Schema Theorem has provided the theoretical underpinning for GAs (Holland 1975, Goldberg 1989); this Schema Theorem assumes that all the genotypes in a population are the same length, and remain so through successive generations. In the messier world of natural evolution these assumptions do not hold, which prompts questions such as:

- Could some more generalized version of this theorem be extended to include variable length genotypes?
- Are there circumstances in which they might be of use in GAs?

In speaking of variable length genotypes I will be making some assumptions, spelt out later, about how those extra parts on long genotypes, not present on the shorter ones, contribute to their fitness. But the answers to these two questions will be, firstly: no, there is no such immediate generalization, but rather a very different process is at work as genotypes change length, which must be analysed independently. And secondly: for traditional function optimization problems they are unlikely to be of use, but they will be in Artificial Life.

Manipulation of schemata in the conventional analysis of GAs can be interpreted in terms of intersections of hyperplanes in the predefined search-space — for instance in the case of binary genotypes of length l, the search-space is a hypercube of l dimensions. If this length is variable, it is not easy to extend this notion of a search space satisfactorily. An alternative characterization of a genotype search space, perhaps less familiar to the GA community, is borrowed from theoretical biology; this lends itself more easily to variation in length of genotype.

It will be argued that for progress through such a space to be feasible, it only makes sense for genotypic variation in length to be relatively gentle. It follows that instead of attempting a generalization of the Schema Theorem to genotypes of any length, the analysis of the convergence of a population of nearly uniform length can and should be decoupled from the analysis of changes in length. A general trend towards increase in length is associated with the evolution of a *species* rather than global search. The word *species* I am using to refer a fit population of relative genotypic homogeneity.[1]

As to the question of under what circumstances variable lengths might be of use in GAs, it would seem that for such traditional GA concerns as function optimization in a pre-defined domain, one would do best to stick to fixed lengths. In the context of Artificial Life, however, where an animat is evolving in an environment with unknown complexity, then variability in genotype length becomes relevant. A genotype space can be open-

*Appears in "Toward a Practice of Autonomous Systems", Proceedings of First European Conference on Artificial Life, F.J. Varela & P. Bourgine (eds.), MIT Press/Bradford Books, 1992

[1] It will follow from this that crosses between members of the same species have a good chance of being another fit member of the same species; whereas crosses between different species will almost certainly be unfit.

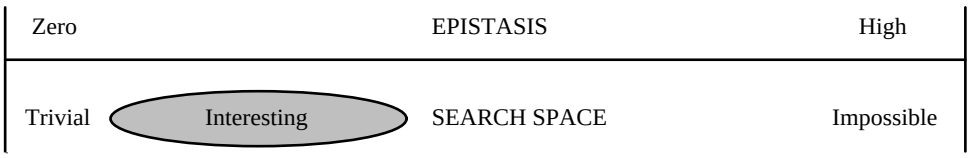

Figure 1: *Low, but non-zero, epistasis is associated with a search space that is possible, but non-trivial.*

ended if the environment itself alters over time, perhaps in response to the evolution of the animat itself. The classic case is the *Red Queen* (or *Arms Race*) phenomenon of coevolution of different species interacting with each other, where one can expect over time both the phenotype complexity and the genotype length to increase.

The notion of a search space is a metaphor which is usually a useful one. It does, however, imply a space of pre-defined extent, with a pre-defined or recognizable goal. In the natural world, tempting though it may be for any one species to think of evolution as a 4 billion year search for a goal of something very like them, it is evident that any such notion of a goal can only be *a posteriori*. So in order to distinguish the space of possibilities that a species can move in from that of a conventional search space, I shall use the term *SAGA space*[2]. This corresponds to the acronym for Species Adaptation Genetic Algorithms, the altered and extended version of GAs necessary to deal with such a space.

2 Variable lengths in GAs

Variable length genotypes have been used in GAs in, for instance, Messy GAs (Goldberg *et al.* 1990), LS-1 classifiers (Smith 1980), Koza's genetic programming (Koza 1990). The first of these in fact uses an underlying fixed-length representation. The analyses offered in the other two examples do not satisfactorily extend the notion of a schema such that schemata are preserved by the genetic operators.

For instance, Koza's genetic programming (Koza 1990) uses populations of programs which are given in the form of LISP S-expressions; these can be depicted as rooted point-labeled trees with ordered branches. The primary genetic operator of crossover, or recombination, swaps complete sub-trees between the parents, and if these sub-trees are of different size then the offspring will have genotypes of different lengths from their parents.

Koza suggests that the equivalent of a schema in the search space of such programs can be specified initially by any one specific sub-tree. Since the set of all potential programs containing that sub-tree is infinite, Koza finds it necessary to partition it into finite subsets indexed by the length of the program, and it is these subsets that are considered as schemata. The number of occurrences in the reproductive pool of examples of a particular schema which, as sampled in the parental pool, shows above-average fitness, will indeed tend to increase. But this does not cater for the fact that the crossover operator will in general turn the offspring into programs of different lengths, and hence disrupt the schema which has been defined by program length. A possible way to minimize this disruption would be to restrict the possible variations in length to only minimal changes, and indeed this will be echoed in the conclusions reached further on in this paper.

The obvious way to extend the crossover operator from fixed-length to variable-length genotypes is by randomly choosing different crossover positions for each of the two parents; an offspring may then inherit two short portions, or two long portions, and in general will have a genotype of significantly different length. It will be shown that this approach is flawed.

3 Epistasis

A gene is the unit of analysis in determining the phenotype, and hence its fitness, from the genotype; it is coded for by a small subsection of the genotype. The term epistasis refers to the linkage between genes on the genotype, such that the expression of one gene modifies or over-rules the expression of another gene.

If there is no epistasis, in other words if the fitness contribution of each element on the genotype is unaffected by the values of any of the others, then optimization can be carried out independently on each element; simple hill-climbing is adequate. At the other end of the epistatic scale, where there are many dependencies between the elements, the only useful building blocks that a GA tries to manipulate are too long, and easily disrupted by genetic operators. Indeed in the limit of maximum epistasis only random search is feasible. The appropriate region on the epistatic scale suitable for GA type search is between these two extremes, and GA representations need to be chosen with this in mind (Davidor 1990).

4 Uncorrelated Landscapes

A model of a genotype search space which allows explicit setting of low or high degrees of epistasis is based

[2] "**Saga** ... story of heroic achievement or adventure; series of connected books giving the history of a family etc. [Old Norse = narrative]." Concise Oxford Dictionary.

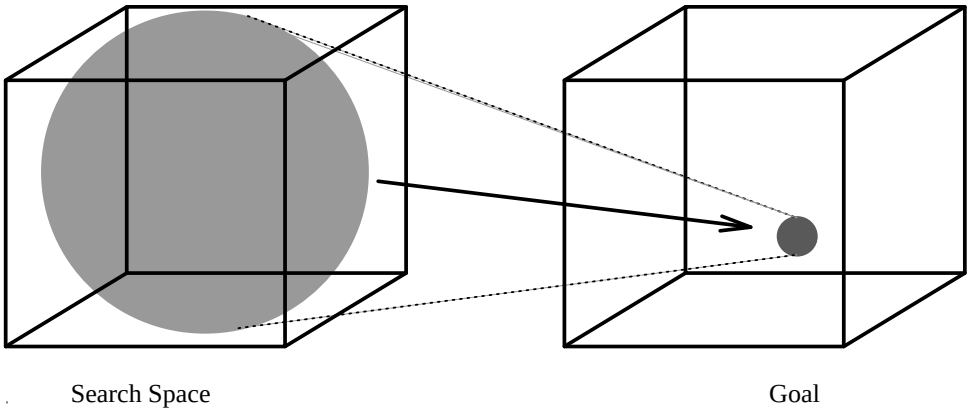

Figure 2: *The evolution of a standard GA in a fixed-dimensional search space; the population initially spans the whole space, and in the end focusses on the optimum.*

on the concept of a protein space, originally introduced in (Maynard Smith 1970). This space has a point for each possible example of a genotype, and a neighbourhood metric which gives all those other points which can be reached by a single mutation from a given point. Compared with the traditional GA analysis, which takes a global view of the whole search space, and considers how a population of points in this search space can effectively range across it by use of recombination, we now have a very different perspective. Here mutation is the only genetic operator, instead of just a background one to prevent irremediable loss of an allele.

Kauffman (Kauffman and Levin 1987, Kauffman 1989) has extended this model to produce a general theory of adaptive walks on rugged fitness landscapes — where the distribution of fitness values across the space is visualised as a landscape with fitness represented by the height. It should be noted that the fitness values are ascribed to points on a lattice rather than a continuum. Nevertheless the landscape can be imagined as a mountain range, where ruggedness implies a relative lack of correlation of heights of nearby points, which in turn is associated with high epistasis on the genotype.

Gillespie's assumptions (Gillespie 1984), that the mutation rate is slow compared to the assimilation of any fitter mutation by the population as a whole, are being used. The population as a whole is considered to be at a single point in the space, with mutations of single members sampling the immediately neighbouring points. Any less fit mutations die out rapidly, whereas any fitter one causes, by this assumption, the whole population to move to that point.[3]

If the fitnesses of neighbouring points on this landscape are completely uncorrelated, then it is maximally rugged. In an adaptive walk on such a landscape if the first step upwards, from the bottom rank of fitness, takes one unit of time, then the next step upwards, where only half the neighbours are fitter, takes on average 2 units, then 4, 8, ..., doubling each time (Kauffman and Levin 1987).

5 Correlated landscapes

The above discussion is for a completely uncorrelated landscape — which can be considered equivalent to maximum epistasis between the genes on the genotype. In most fitness landscapes there is, however some local correlation, in that neighbours will tend to have similar fitness values, and certainly this is true of any search space in which GAs are to be of use. Let length be defined in this space using the distance metric of how many point mutations are necessary to move from one genotype to another. A long jump is defined to be the equivalent of several *simultaneous* mutations, long enough to jump beyond the correlation lengths in the landscape. Moves via such long jumps will in general display important similarities with the characteristics of uncorrelated landscapes (except that in the limit of long jumps all points are accessible, and hence the notion of a local optimum becomes meaningless). In particular the above result still holds: that the waiting time until finding a fitter variant by such long jumps doubles after each such improvement.

Kauffman further considers a different assumption from that used above; suppose that instead of a single mutant being sampled at each unit of time, there

[3]This is a more restricted assumption than that in (Eigen and Schuster 1979), where a population under the influence of selection and a low mutation rate in general moves to form a *quasi-species*, with a probability distribution centred about a point. Eigen and Schuster show that for a given selective pressure, the maximum length of genotype that can be reliably held in a tight distribution at an optimum is of the order of magnitude of the reciprocal of the mutation rate. Mutation rates of 5×10^{-4} per base by single-stranded RNA replication is adequate for a phage with length 4500. The lower mutation rates of order 10^{-9} associated with DNA replication and recombination in eukaryotes allow for the genotypes of length order 10^9 that humans have.

is a large population of fixed size simultaneously sampling different mutants, and the population then moves as a whole to the fittest of any improved variant encountered. It is shown that the above result on waiting times remains almost unchanged.

This search process is of course very different from that analysed in conventional GAs, where a population of points effectively spans the search space, and recombination allows effective moves to predominate. The distinction between these two types of search process must be kept in mind when we turn to looking at variable length genotypes.

6 Variable length genotypes

Let us spell out some assumptions about a genetic system with variation in the length of genotypes, within which many different types of representation, or mapping from genotype to phenotype to fitness, could be allowed.

- Firstly, it is assumed that the genotype can be analysed in terms of a number of small building blocks, or genes, that are coded for individually on it; possibly by a single symbol, or a sequence of symbols. These genes can be uniquely identified, either by their position by reference to an identified end of the genotype, as in conventional GAs; or by an attached tag or template, such as those used in messy GAs (Goldberg *et al.* 1990). Longer genotypes will code for genes that are not present at all on shorter ones.

- Secondly, it is assumed that each gene makes a separate additive contribution to the fitness of the whole; but that the contribution of any one gene can be modified by epistatic interactions with a number K of the other genes. This number K is less than the total number of genes available, otherwise the fitness landscape would be uncorrelated.

- Thirdly it is assumed that the total of all these additive contributions is then normalized in some way such that the final fitness remains within some predefined bound regardless of how many genes there are.

This last condition reflects the fact that any fitness function is only relevant in so far as it affects the selection process. On average in the long term each member of a viable population will be replaced by just one offspring. Less than one and the population is heading for extinction, more than one implies exponential growth. But there are always finite physical resource limitations which prevent such unlimited growth, and this has to be taken account of in the fitness function.

All these assumptions allow a standard GA to operate when lengths are fixed. In addition to the normal genetic operators of mutation and crossover, we assume that there are further operators, perhaps *cut* and *splice*, or *increase-length*, which allow offspring to have their length changed by arbitrary amounts, although still retaining at least some genetic material from their parent(s).

Suppose that there are a total of G different genes represented in the population, some perhaps represented in all genotypes and some in only a few. Then by adding an extra allele for each gene, to indicate whether it is 'absent' in a particular genotype, a new representation of the population can be formed in which every gene is represented in every member. Genetic operators which do not introduce a completely new gene into the population allow this to be analysed as a normal GA.

Suppose now that the genetic operators allow, by lengthening a genotype, the creation of a *single* new gene, giving a new total of $G + 1$. By the second and third assumptions made explicit above, the epistasis of this new gene is similar to that of the previous ones. The new population can now be considered as being spread across a new $(G + 1)$-dimensional search space, except that all bar one member is confined within the previous G-dimensional sub-space. This can still be analysed as a normal GA with an initially skewed population. If a single advantageous new gene appears in the population, it can become widespread through crossovers.

In contrast to this, an alternative possibility is that the genetic operators allow the creation in one generation of a *large number* g of new genes on one genotype. In the new $(G + g)$-dimensional search space, the old population is based entirely inside the original (in relative terms, very small) G-dimensional subspace, with just the one new point exploring elsewhere. This is obviously a 'long jump' and the fitness will be uncorrelated with that of any of the previous generation. If such a long jump is successful, in the sense that the new genes are retained in the population, with a resulting general increase in the fitness of the population, then the chances of a successful further long jump will be significantly less. Any such long jump adaptation will suffer from the problem of the doubling of waiting time after each jump.

The picture now emerges of two very different processes going on at independent timescales in this SAGA space. Given a genetic operator which allows unrestricted changes in length of genotypes, we can expect the following sequence of events in a locally correlated landscape:

- An early population could fluctuate in length through 'long jump' adaptation which effectively acts in an uncorrelated landscape; but as average fitness increases the doubling of waiting times will slow down this process drastically.

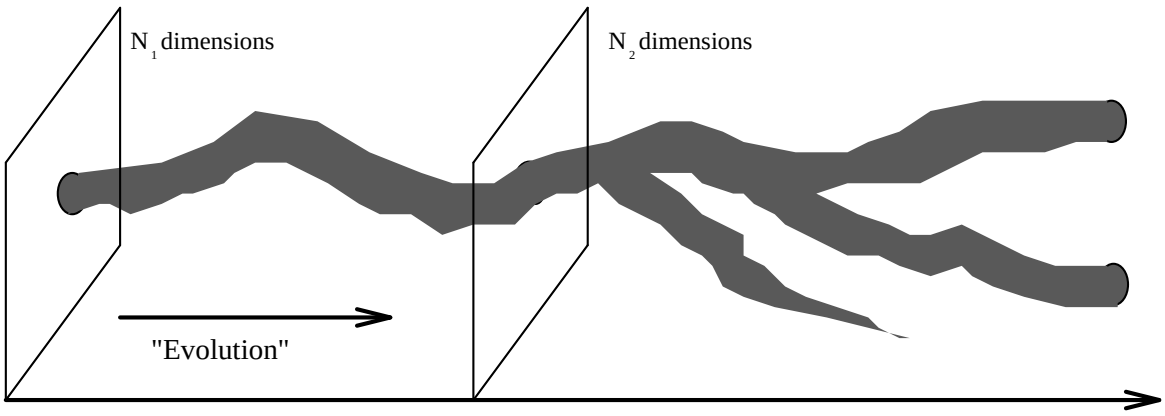

Figure 3: *The progress of the always compact course of a species; the* z *axis indicates both time and the (loosely correlated) number of dimensions of the current search space. The* x *and* y *axes represent just two of the current number of dimensions.*
The possibility of splitting into separate species, and of extinction, are indicated in the sketch, although not here discussed.

- Thereupon the traditional GA operators of crossover and mutation will take over, and Holland's Schema Theorem will be applicable to this phase of the search.
- Those applications of the change-length operator which result in minimal changes of length will be moves on a correlated landscape, and therefore are feasible even if major changes are increasingly unlikely.
- If there are selectionary pressures which encourage the genotype lengths to increase, the population will become a nearly-converged 'species', with an almost uniform length that increase in small steps.[4]

7 SAGA and the Schema Theorem

A schema defines a subset of possible genotypes which share the same values at a specified number of genes. If there is no upper limit to the possible length of the genotype, these subsets will be infinite in size, and estimates of the 'average fitness' of a schema based on any finite sample become problematical.

We might be tempted to avoid this by saying, in this particular example we will restrict the space of possibilities to genotypes of length less than, e.g., 1000. But then 'nearly all' possible instances of a particular schema will refer to genotypes very close to this upper limit, and there may be no reason to expect the average fitness of this schema to bear any significant relationship to the fitness of the same schema restricted to genotypes of maximum length 100, 500 or even 950.

However, consider the case where all the population have the same G gene sites (though with variations in the values at each gene site); and we are considering the addition of one extra gene to one or more of the population. We can recast our analysis in terms of a population all of genotype length $G+1$, with the extra gene having one additional possible value of 'absent'. For any two schemata S_1 and S_2 that have the extra gene fixed as 'absent', let S_1' and S_2' be the corresponding schemata with the extra gene value allowed any value (including 'absent'). Given the assumptions of low epistasis, the relative fitnesses of schemata S_1 and S_2 will be closely correlated with the relative fitnesses of schemata S_1' and S_2'. This will still hold true if we allow an extra g genes rather than one, provided that g is small in relation to G and the assumption of low epistasis holds. It will not hold true when g is large, or epistasis is high.

Hence in the short term of small changes in genotype length in a population of nearly uniform genotype length, we can still apply the Schema Theorem.

8 Would variable lengths be useful?

Turning now to the second question posed in the introduction, under what circumstances might it be useful to have a genetic operator which allows an increase in the number of genes represented on the genotype? If the problem being tackled is basically a function optimization one, where there is a pre-defined search space with

[4] These ideas should be neutral in respect of the punctuated equilibria controversy. A succession of small steps may or may not be rapid in geological time — indeed there may well be good reasons why there should on occasion be such a cascade. What is being ruled out here is any single large step.

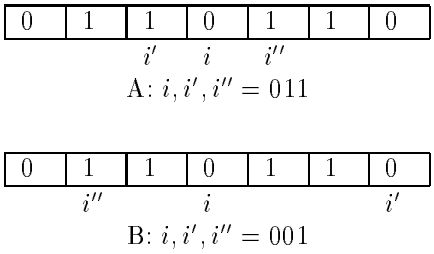

Figure 4: *At the top, gene i is linked to neighbours i', i''. The values 011 point into a fitness look-up table for i. Below, i' and i'' are no longer immediate neighbours.*

	000	0.141
value of B	001	0.592
	010	0.653
value of A	011	0.589
	100	0.793
	101	0.233
	110	0.842
	111	0.916

Figure 5: *Fitness table for gene i, filled with random numbers between 0 and 1. i, i' and i'' determine fitness contribution of gene i to fitness of the whole genotype.*

a fixed number of factors that can be coded for on the genotype, then it would be folly not to put them all in at the start, represented in such a way as to minimize the epistasis, and put one's trust in the Schema Theorem.

A major group of problems which cannot be specified in terms of a pre-defined search space involve coevolution of one population with another (or several) which in turn is affected by the first. Since one population is part of the environment for the other, the environment is continually changing (Hillis 1991, Husbands and Mill 1991). The same requirements of relatively few epistatic interactions between a gene and those aspects of the environment which it affects and is affected by, hold if an evolutionary process is going to be more than random search.

There are many coevolutionary worlds where an increase in complexity in one population stimulates an increase in complexity of the other, and so on, perhaps indefinitely. So in as much as length of genotype is associated with complexity of the phenotype, we can expect that there is selective pressure for long-term growth in their lengths. Lindgren (Lindgren 1990, Lindgren 1991) models a population of individuals competing with each other at the iterated Prisoner's Dilemma with noise — the population in practice breaks into sub-populations with different strategies. There is no recombination, the only genetic operators being mutation and gene doubling. The particular representation used treats a binary genotype of length 2^h as a look-up table; the history of the last h interactions between competing prisoners, coded in 0's and 1's and considered as a binary number, generates a pointer into this look-up table to determine the strategy. Application of the gene-doubling operator does not in itself generate new strategies, but allows later mutations to generate finer discriminations within that strategy. Hence his representation could be mapped into a different one where the length of the genotype only increases by one step at a time. His results show periods of stasis alternated by periods of unstable dynamics, with a long-term growth in the lengths of the successful sub-populations.

9 Simulation

The NK model (Kauffman and Levin 1987, Kauffman 1989) assumes a binary genotype of length N, where each position represents a gene which is affected by linkage with K others. This is an abstract model in which it is assumed that the fitness of the phenotype can be directly calculated from the values on the genotype.

The three assumptions itemized above hold for the fitness function. In the case of $K = 2$, the fitness contribution of gene i depends on the two others to which it is linked (which may be specified as immediate neighbours, or may be specified at random positions). The binary alleles of i and its 2 neighbours specify a 3-bit number which picks a fitness from an 8-place table of fitnesses associated with gene i — there are N such fitness tables prepared at the start of the simulation, with each place containing a fitness randomly chosen in the range 0 to 1.

The fitness of the genotype is then assessed by adding up the fitnesses thus determined for all N genes, and dividing by N. It can be seen that in this case of $K = 2$, the flipping of a single bit on the genotype will affect the fitness contributions of on average just 3 genes; the other $N-3$ being unaffected, this gives a reasonably correlated fitness landscape. In the limit of $K = N - 1$, where the fitness table associated with each gene would have 2^N places, the flipping of a single bit would alter everything, and the fitness landscape is totally uncorrelated.

This model can be extended to allow for changes in genotype length. The simplifying assumption is made that any new gene appears at the right-hand end of the genotype, and that the identity of the gene is uniquely determined by its position in the genotype. In the case of $K = 2$, if linkage is with immediate neighbours to left and right, the ends are assumed linked in a loop to avoid boundary conditions. A set of tables of random fitness values for each gene is set up for the minimal-length genotype. For each new gene added one new table is generated for it, and two further replacement tables for those genes which are neighbours of the newcomer. This can easily be generalized for $K > 2$, and also for choice of

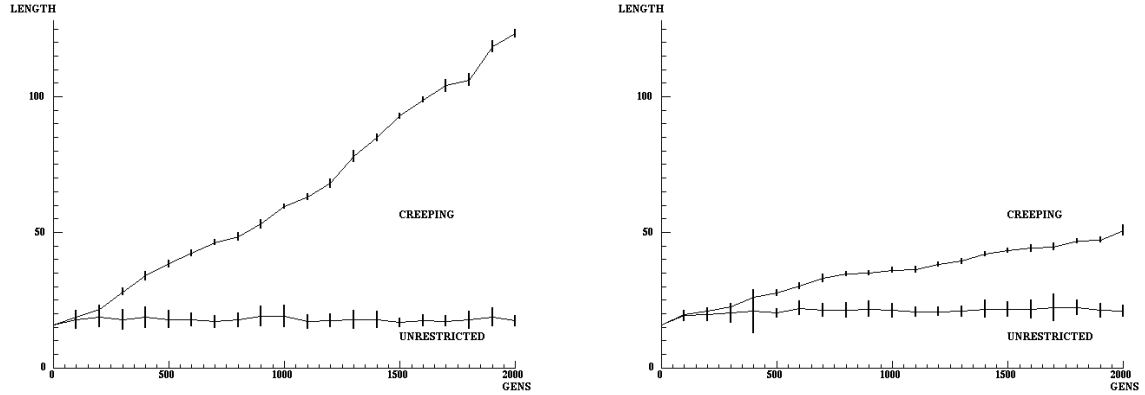

Figure 6: *Average genotype lengths against generations; vertical bars show standard deviations. Effects of 'creeping' and 'unrestricted' increase-length genetic operators on a population with the same fitness conditions, epistasis $K = 2$. Left graph, linkage with neighbouring genes. Right graph, random linkage.*

epistatic linkages to newcomers being randomly selected rather than restricted to neighbours.

This allows the setting up of models for simulations with genotypes of any length, with epistasis of any degree. A standard GA can then be run, with increase in the lengths of genotype allowed under specified conditions. Such simulations allow experimentation with variable lengths, in an abstract context, without the difficulties of choice of representation that normal problems give.

Experiments have been run with a population of 100 genotypes, all of initial length $N_{init} = 16$, with epistasis given by $K = 2$, linkages being with neighbouring genes on the genotype; the genetic operators will allow the lengths to increase. The initial fitness of any genotype, as defined by look-up tables of random numbers, was then adjusted by adding a factor proportional to $N/(N+\overline{N})$, where N is the length of the particular genotype and $\overline{N}$ is the current average genotype-length of the population. The constant of proportionality was chosen such that there was a selectionary pressure in favour of longer genotypes comparable to the selectionary pressures given by the initial fitnesses.

In the first trial the genetic operators were crossover, mutation, and an increase-length operator which in 10% of offspring allowed a genotype to increase in length by between 0 and 50%. In the second trial the 'creeping' increase-length operator only allowed an increase of genotype length in the offspring by exactly one. Despite this restriction, the average length increased steadily in the 'creeping' trial as compared to virtually no increase in the 'unrestricted' trial.

On separate trials with the epistatic linkages being with randomly placed genes instead of with neighbouring genes, the results were similar, although with the 'creeping' increase in length at a slower rate. In both sets of trials there was much more variation in lengths within the population in the 'unrestricted' case, compared to the 'creeping'.

10 SAGA and Development

By working with the evolution of a nearly converged species increasing in genotype length and in phenotype complexity over time, we have moved away from the usual GA notion of evolution as a search technique towards a notion of 'evolution as a tinkerer' (Jacob 1989), always adding to or altering something that is already viable.

The cumulative process of additions and alterations implies that a phenotype can be considered as being produced from a genotype by a developmental process. It will not be surprising if 'Ontogeny recapitulates Phylogeny', subject to the small but ever-present possibility of a later alteration bearing on a significantly earlier stage in the developmental process. The application of this approach to, for instance, the evolution of subsumption architectures for robots (Brooks 1991) would seem to correspond to the effective, tinkering, incremental approach that practical designers take.

One consequence will be that a species will only reach those parts of a SAGA space that are connected by a continuous chain of viable ancestors to the origin. Thus within the space of all possible genotypes of length G there may well be a host of fit and viable points or islands which, through isolation and lack of a viable pathway from the origin, are unattainable.

11 SAGA and Genetic Operators

In addition to the usual GA operators of mutation and/or crossover, an operator which allows change in genotype length is necessary. The example in the NK

model simulation above is the most trivial such operator, and depends on the identity of any gene being given by its position relative to one end of the genotype. Lindgren's (Lindgren 1991) doubling operator uses a representation which has this same dependency on position.

If the identity of a gene is given by a tag, or by template-matching as seems to happen in the real world of DNA, then absolute positions of genes on the genotype need not be maintained. This allows for duplication of a section of the genotype, after which mutations can differentiate the duplicated parts. The crossover operator can still be used in a fairly homogeneous population with slight variations in genotype length, although given any random crossover point in one parent, a 'sensible' corresponding crossover point in the other parent must be chosen. This can be uniquely defined as that point (or in some cases, any of a contiguous group of points) which maximises the longest common subsequences on both sides of the crossover. A version of the Needleman and Wunsch algorithm makes this computationally feasible (Needleman and Wunsch 1970, Sankoff 1972).

12 Conclusions

With fixed-length genotypes one can afford to think in terms of a fixed, pre-defined search space with a finite number of dimensions which, even if it is immense, is at least theoretically knowable by God or Laplace.

When one allows genotypes to vary in length the search space is potentially infinite and it stops making sense to think of it as predefined. Nevertheless, in the real world, evolution has taken place in such a fashion that we have very distant ancestors whose genotypes were much shorter than ours; the problems we face are not the problems they faced.

When looking at evolution, talking about 'problems being solved' can be very misleading. However, people using GAs are usually hoping to use lessons from evolution in order to find solutions to a problem that faces them. If they really do know the problem they have to solve, then they can define in finite terms the search space, and fixed length genotypes are appropriate. If, however, they are trying to evolve a structure with arbitrary and potentially unrestricted capabilities, then the problem space is not pre-defined, genotypes must be unrestricted in length, and a new approach is needed. Hence this discussion is probably more relevant to those looking at the evolution of animats or cognitive structures than it is to those looking at GAs as function optimizers.

One of the lessons demonstrated is that if genotypes can potentially increase indefinitely, they will in practice only do so on a slow timescale, so that within a population all genotypes will be very nearly the same length. Indeed, there will be a high degree of uniformity in the genotypes, and any significant variations, including changes in length, will spread through the whole population before the next variation occurs. This is in contrast to the relatively fast timescale on which the crossover operator, which is the power-house of standard GAs, very efficiently mixes and matches fitter schemata.

One factor to bear in mind here is that there is a relationship between mutation rate and the length of a genotype that can effectively evolve. Too little mutation, and there is not the variation to allow change; too much, and there is not sufficient stability to maintain fitness.

In contrast to the approach used in Holland's Schema Theorem, or the hyperplane analysis of schemata, where the population can effectively sample the whole search space, we must visualise a population in our new, infinitely though slowly expandable, search space as a localized cloud (with a high degree of consistency within the population) which can only sample 'nearby' points (those that can be moved to by one or a small number of applications of the genetic operators.) The question ceases being 'Where in this whole search space is the optimum?' and becomes instead 'From here, where can we move to that is better?'.

Acknowledgements

This work has been supported by a grant from the SERC, and I acknowledge helpful comments on an earlier draft from Phil Husbands, Jim Stone, Pedro de Oliveira and Shirley Kitts.

References

[Brooks 1991] R.A. Brooks. Intelligence without representation. *Artificial Intelligence*, 47:139–159, 1991.

[Davidor 1990] Yuval Davidor. Epistasis variance: A viewpoint on representations, ga hardness, and deception. *Complex Systems*, 4(4), 1990.

[Eigen and Schuster 1979] M. Eigen and P. Schuster. *The Hypercycle: A Principle of Natural Self-Organization.* Springer-Verlag, 1979.

[Gillespie 1984] J.H. Gillespie. Molecular evolution over the mutational landscape. *Evolution*, 38:1116, 1984.

[Goldberg *et al.* 1990] David E. Goldberg, K. Deb, and B. Korb. An investigation of messy genetic algorithms. Technical Report TCGA-90005, TCGA, The University of Alabama, 1990.

[Goldberg 1989] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley, Reading, Massachusetts, USA, 1989.

[Hillis 1991] W.D. Hillis. Co-evolving parasites improve simulated evolution as an optimization parameter. In C. G. Langton, J. D. Farmer, S. Rasmussen, and C. Taylor, editors, *Artificial Life II: Proceedings Volume of Santa Fe Conference Feb. 1990.* Addison Wesley: volume XI in the series of the Santa Fe Institute Studies in the Sciences of Complexity, 1991.

[Holland 1975] John Holland. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, USA, 1975.

[Husbands and Mill 1991] Philip Husbands and Frank Mill. Simulated co-evolution as the mechanism for emergent planning and scheduling. In R. K. Belew and L. B. Booker, editors, *Proceedings of the Fourth Intl. Conf. on Genetic Algorithms, ICGA-91*, pages 264–270. Morgan Kaufmann, 1991.

[Jacob 1989] François Jacob. *The Possible and the Actual.* Penguin, 1989.

[Kauffman and Levin 1987] Stuart Kauffman and Simon Levin. Towards a general theory of adaptive walks on rugged landscapes. *Journal of Theoretical Biology*, 128:11–45, 1987.

[Kauffman 1989] Stuart Kauffman. Adaptation on rugged fitness landscapes. In Daniel L. Stein, editor, *Lectures in the Sciences of Complexity*, pages 527–618. Addison Wesley: Santa Fe Institute Studies in the Sciences of Complexity, 1989.

[Koza 1990] John R. Koza. Genetic programming: A paradigm for genetically breeding populations of computer programs to solve problems. Technical Report STAN-CS-90-1314, Department of Computer Science, Stanford University, 1990.

[Lindgren 1990] K. Lindgren. Evolution in a population of mutating strategies. Technical report, Nordita, Copenhagen, 1990.

[Lindgren 1991] K. Lindgren. Evolutionary phenomena in simple dynamics. In C. G. Langton, J. D. Farmer, S. Rasmussen, and C. Taylor, editors, *Artificial Life II: Proceedings Volume of Santa Fe Conference Feb. 1990.* Addison Wesley: volume XI in the series of the Santa Fe Institute Studies in the Sciences of Complexity, 1991.

[Maynard Smith 1970] John Maynard Smith. Natural selection and the concept of a protein space. *Nature*, 225:563–564, 1970.

[Needleman and Wunsch 1970] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48:443–453, 1970.

[Sankoff 1972] David Sankoff. Matching sequences under deletion/insertion constraints. *Proceedings of the National Academy of Science, USA*, 69(1):4–6, 1972.

[Smith 1980] Stephen F. Smith. *A Learning System based on Genetic Adaptive Algorithms.* PhD thesis, Department of Computer Science, University of Pittsburgh, USA, 1980.