Species Adaptation Genetic Algorithms:

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

Inman Harvey
School of Cognitive and

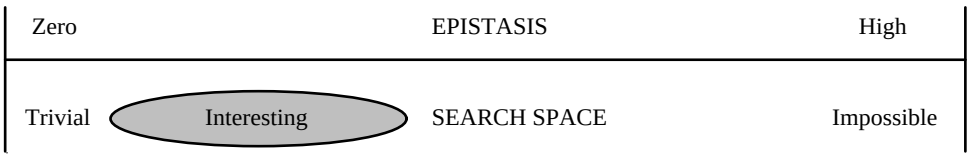


Figure 1: *Low, but non-zero, epistasis is associated with a search space that is possible, but non-trivial.*

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.

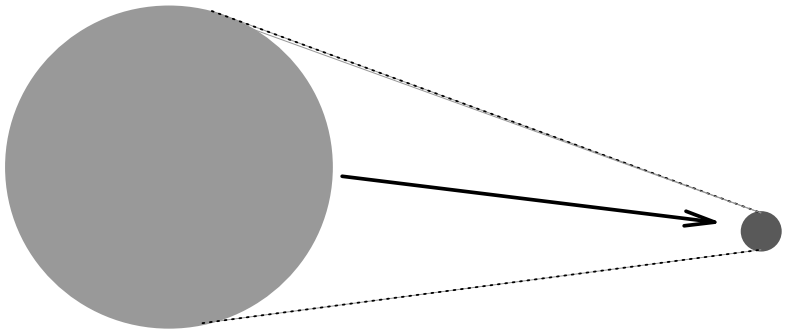
The obvious way to extend the crossover operator from fixed-length to variable-length genotypes is by randomly choosing different crossover positions for each of the two parents; an offspring may then inherit two short portions, or two long portions, and in general will have a genotype of significantly 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 ew59.5199(o-0Td(and)T

[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.



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 physicala alandscaph

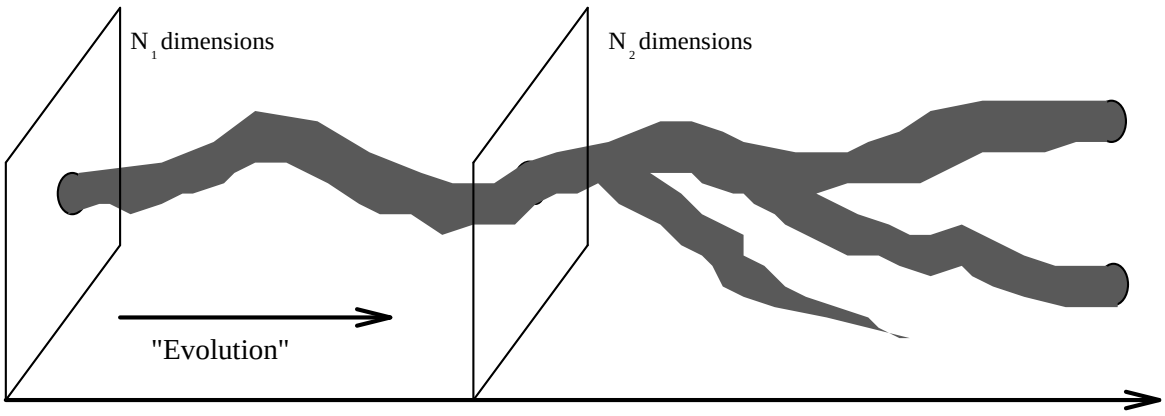


Figure 3: *The progress of the always compact course of a species; the* z *axis indicates both time and the (loosely correlated) number of dimensions of the current search space. The* x *and* y *axes represent just two of the current number of dimensions.*
The possibility of splitting into separate species, 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

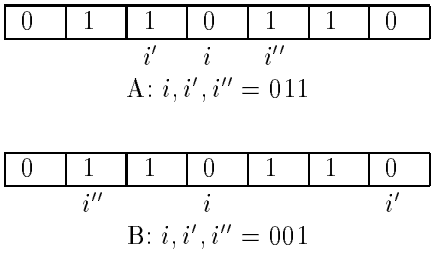


Figure 4: *At the top, gene i is linked to neighbours i', i''. The values 011 point into a fitness look-up table for i. Below, i' and i'' are no longer immediate neighbours.*

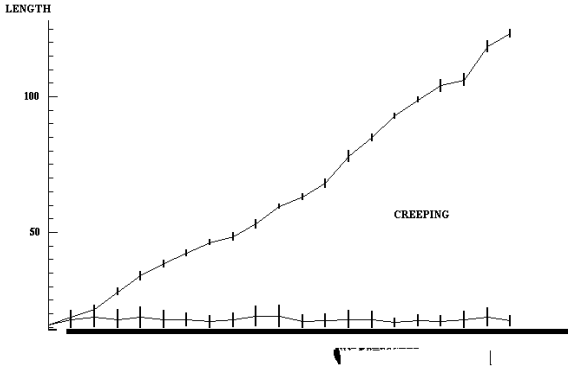
	000	0.141
value of B	001	0.592
	010	0.653
value of A	011	0.589
	100	0.793
	101	0.233
	110	0.842
	111	0.916

Figure 5: *Fitness table*

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.


LENGTH
100
50
CREEPING

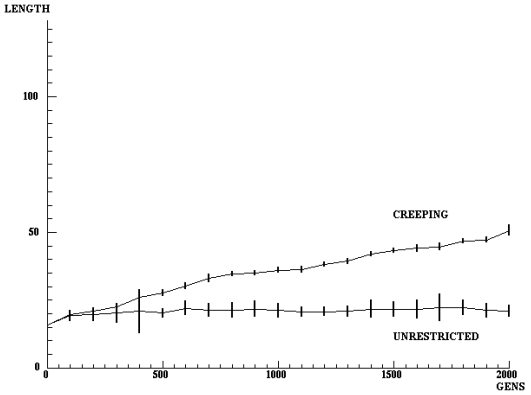

LENGTH
100
50
CREEPING
UNRESTRICTED
0
500
1000
1500
2000
GENS

Figure 6:

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, pfe69692(lar,i40T5i)-1920f(839ssnt0.7ethe) therees11999.3ssion

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