Using Holey Fitness Landscapes to Counteract Premature Convergence in Evolutionary Algorithms

Gregory Paperin
Monash University, Faculty of Information Technology
Clayton Campus, Building 63, Wellington Road, Clayton, 3800 Victoria, Australia
gpaperin@infotech.monash.edu.au

ABSTRACT

Premature convergence is a persisting problem in evolutionary optimisation, in particular – genetic algorithms. While a number of methods exist to approach this issue, they usually require problem specific calibration or only partially resolve the issue, at best by delaying the premature convergence of an evolving population. Analytical models in biology show that resiliently diverse populations evolve on high-dimensional fitness landscapes with "holey" rather than "rugged" topographies, but the implications for artificial evolutionary systems remain largely unexplored. Here I show how *holey fitness landscapes* (HFLs) can be incorporated in an evolutionary algorithm and use this approach to investigate the ability of HFLs to maintain genetic diversity in an evolving population. The results indicate that an underlying HFL can counteract premature genetic convergence and sustain diversity. They also suggest that HFL may provide a flexible mechanism for dynamic creation and maintenance of subpopulations that concentrate their evolutionary search in different regions of the solution space. Finally, I discuss on-going work on using the HFL model in optimisation problems.

Categories and subject descriptors: G.1.6 [numerical analysis] Optimisation – *global optimisation*, s*tochastic programming*.

General terms: algorithms, performance, theory.

Keywords: evolutionary algorithm, genetic algorithm, holey fitness landscape, gene flow, premature convergence, reproductive isolation.

1. INTRODUCTION

Premature convergence is a common problem in genetic and other evolutionary algorithms. A number of approaches have been introduced to counter against it, however, solutions are often case-specific and the general mechanisms that maintain diversity in natural populations are not well understood. In nature, populations often maintain a resilient genetic diversity under strong selection pressures, and various generic approaches to achieving this effect in evolutionary optimisation have been based on mechanisms thought to facilitate the maintenance of diversity in nature. Most of these methods are variants of the so-called *niching* approach [1] (e.g. crowding [2], sharing [3], island model [4]). Other approaches attempt to vary environmental factors such as maximum population size [5], or the evolutionary goal [6], however the latter processes are not always present in diverse natural populations.

In general, niching aims to introduce a degree of reproductive isolation (RI) between groups of candidate solutions in order to concentrate the evolutionary search on different regions of the solution space. A recurring difficulty in applying niching-based algorithms is that the optimal degree of RI and the number of reproductively isolated groups (RI groups) are usually problem-specific and must be artificially tuned or set arbitrary. In addition, while often successfully delaying or slowing premature convergence, niching algorithms are rarely successful at preventing it completely [1]. This difficulty corresponds to a number of results from theoretical biology: In biological terms, niching introduces *prezygotic* RI, i.e. RI caused by not mating with members of other groups rather than by offspring inviability. However, the maintenance of sustained prezygotic RI presents a theoretical challenge for biologists: Prezygotic RI based on ecological divergence or physical barriers is often transient, collapsing when selection pressures change. In addition, even moderate migration between populations leads to high gene flow making the extinction or merging of RI groups likely [7]. On the contrary, RI is likely to be sustained once *postzygotic* reproductive barriers have evolved and genetic incompatibilities make hybrid viability unlikely [7].

Recent advances in theoretical biology suggest that assumptions about the relative fitness of individuals have profound implications for our understanding of the above problems [8]. In particular, Gavrilets and Gravner [9] showed that when fitness landscapes have high dimensionality (as is likely for real organisms as well as for many computational problems), the topology of the landscape changes from "rugged" to "holey". However, integrating such *holey fitness landscapes* into computational models remains a challenge.

In this study, I briefly examine the notion of the holey fitness landscape (HFL) and its implications for genetic diversity. I outline a method for integrating HFL genetics into an evolutionary algorithm. Using this method, I explore conditions for maintenance of genetic variation and reproductive isolation in an artificial evolution. Finally, I discuss on-going work [7, 10, 11] on potential applications to concrete optimisation problems.

2. HOLEY FITNESS LANDSCAPES

The notion of a holey fitness landscape (HFL) was introduced by Gavrilets [8, 9, 12]. Generally, a HFL is "an adaptive landscape where relatively infrequent high-fitness genotypes form a contiguous set that expands throughout the genotype space" [12].

To build some intuition for this model, first recall a few results from percolation theory which play an important role in the analytical treatment of HFLs. Consider a 2-dimensional lattice of cells which can assume one of two states: "black" or "white" (figure 1). Let every cell be black with some probability p independently of all other cells, or white with probability $1 - p$. If p is small, the lattice will contain a few black cells, which may be grouped in a number of small, isolated clusters. As p increases, these clusters grow and merge. Once p crosses a certain threshold p_c, most of the black cells merge together into a single giant cluster that percolates the whole lattice (figure 1). For a 2-dimensional square lattice this

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-131-6/08/07.

percolation threshold is known to be $p_c \approx 0.5927$ [13]. However, for lattices of higher dimensions the percolation threshold lies around the reciprocal of the lattice dimension [14], meaning that for a high dimension lattice a small proportion of black cells is sufficient for the emergence of a giant percolating cluster of connected black cells.

For the HFL model, a genotype is assumed to be viable with probability p independent of all other genotypes, and inviable with probability $1 - p$. For the purpose of this discussion, the exact fitness of a genotype is irrelevant, thus let the fitness of all viable and inviable genotypes be 1 and 0 respectively. Consider all possible (haploid) genotypes with L loci and A alleles at each locus ordered in an abstract genotype space, in which the distance between the genotypes describes the ease of transformation from one genotype to another. The dimensionality of this genotype space is $D = L \times (A - 1)$, and the corresponding percolation threshold is $p_c = 1/D$. Even for short genotypes (on biological scales) a relatively small value of p will result in an extensive network of high-fitness ridges extending through the genotype space (e.g. for $L = 10^5$ and $A = 5$, $p_c \approx 20 \times 10^{-7}$). The traditional picture of rugged highly-dimensional fitness landscapes is therefore misleading, as these landscapes are characterised by the existence of percolating nearly neutral networks. These high fitness networks are important as adaptive walks along such networks can proceed far without any substantial loss to fitness.

There are a number of analytic models of HFLs (e.g. see [8, part 1]), however the application of this concept to simulation and computational scenarios is largely unexplored. One of the reasons for this gap can be attributed to difficulties in the implementation of HFL-models. The difficulties arise because the space and the time complexity of computing an appropriate set of viable genotypes is in the order of A^L. In [7] and [10] my colleagues and I discuss this issue in detail and outline an algorithm that allows creating an HFL for large L using a desktop computer within a few minutes. In short, a set $\mathbb{V}' \in \mathbb{G}$ is created, where $\mathbb{G}$ is the set of all genotypes of length L and $\mathbb{V}'$ is a connected percolating subset that is uniformly distributed in $\mathbb{G}$. The diallelic genotypes are represented as bit-strings and stored in a manner that allows an efficient implementation of a function *viable(G)* that takes an arbitrary bit-string and returns true iff $G \in \mathbb{V}'$.

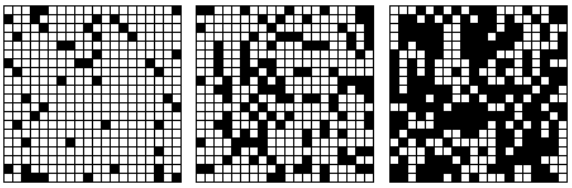

Figure 1. Percolation on a square lattice. The cells are black with probability $p = 0.1$ (left), $p = 0.3$ (middle) and $p = 0.6$ (right).

3. SIMULATION MODEL

My objective is to investigate the extent to which HFL can sustain existing RI (and therefore, diversity) between spatially isolated sub-populations under different levels of migration. For this I created a simulation model (previously introduced in [7]) in which individuals are located on a homogeneous landscape consisting of a cell-grid. Individuals, whose fitness (viability) is defined by the HFL, mate with other individuals within the same cell and then migrate to a neighbouring cell with a certain probability. The lifecycle of the individuals is reproduction – selection – migration and generations are non-overlapping. As common in biological models (e.g. [15]), I use a number of neutral loci to measure the level of gene flow between the populations in different cells for different migration rates.

Here I give the parameter values used in the experiments. For motivation of particular values and sensitivity analysis see [7].

Representation: Individuals are represented by their genotype, which consists of a *coding* section and a *neutral* section. The coding section consists of $L_C = 26$ diallelic loci that code for vital traits. The coding section of a genotype is used as a parameter to the *viable* function of the HFL described at the end of the previous section. This function is used to determine whether an individual is viable. The neutral genotype section consists of $L_N = 5$ loci with $A_N = 128$ different alleles possible at each locus. The neutral loci do not affect the fitness (viability) of an individual and are used to measure the genetic divergence.

Reproduction: Each individual is selected once as a mother and a partner is selected randomly from the same cell. The offspring genotype is determined through free recombination. Mutation is applied with probability $p_M = 10^{-4}$ per locus. If a coding locus is mutated, its binary value is flipped. The neutral loci are subject to a circular stepwise mutation model [16].

Selection: All individuals within a single cell of the spatial landscape compete to reach the age of reproduction. A landscape cell provides enough resources for the survival of $C_{mc} = 250$ mature individuals. If a cell is inhabited by more than C_{mc} individuals, C_{mc} viable individuals are selected with equal probability and the rest are discarded (as in this HFL model a particular individual is either fit or inviable).

Dispersal: To avoid edge artefacts the landscape is represented as a torus. Initially, migration between the cells is disabled and the model is iterated for 100,000 generations in order to allow the allele distribution to reach equilibrium. Then, migration is permitted at a specific rate (see results section) and the model is iterated for 300,000 further generations.

A quantity of prime interest in this model is the number of reproductively isolated groups (RI groups) present in the model at any one time as well as various attributes of such groups. The aim is to detect groups of genotypes that could mate successfully, not groups of individuals who actually do so. In order to cluster the genotypes of a population into RI groups I employ the Markov Clustering algorithm (MCL) [17]. The details of this approach and an analysis of the applicability of the results is given in [7]. On the basis of the RI groups, the average genetic divergence in neutral loci between the groups is measured using the fixation index F_{st} [18]. This measure is close to 1 when the RI groups in the population exhibit a strong genetic divergence at neutral loci and close to 0 when no significant divergence is present [7]. For each of the scenarios discussed below 10 independent model runs were performed and the results were averaged.

4. SIMULATION RESULTS

Consider first the 2×2 grid layout. As a basis for comparison a set of runs with a migration rate of 0% was performed. As expected, the

number of RI groups corresponds to the number of cells and the divergence at neutral loci grows ($F_{st} \approx 1$) (figure 2)[1].

In the next scenario the migration rate was increased to 1% after the first 100,000 generations. This lead to a slight increase in the number of distinct coding sections in the population which is due to viable hybrids resulting from breeding with immigrants. Some of these hybrids spontaneously form RI groups, however such groups cannot persist due to low population numbers in comparison to native populations. These viable hybrids facilitate a limited gene flow between the populations: after 300,000 generations F_{st} has decreased to ca. 0.8 (figure 2.D).

In the next scenario the migration rate was set to 5%. Qualitatively, the results are similar to the 1% scenario. Quantitatively, the gene flow between the populations is higher (F_{st} falls to ca. 0.7, not shown). The higher migration rate leads to an increased probability for formation of RI hybrid groups (figure 2.A). Genetic drift within a larger number of RI groups as well as hybridisation between more diverse individuals lead to a larger number of coding genotype sections in the population (figure 2.B) and to a higher rate of discovering new viable adaptations (figure 2.C). Further rises in the migration rate to 15% and 20% (figure 2) increase the strength of the above effects.

When the migration rate is set to 25% or more RI can no longer be sustained. A large number of reproduction events that lead to inviable offspring implies a high chance of extinction for any native cell population. As seen in figure 2.A only one RI group remains under 25% migration. Sporadically small RI groups arise due to drift, but do not persist long enough to achieve a significant divergence at neutral loci (figure 2.D). The main population evolves as a single RI group and the number of distinct coding sections in the population is small (figures 2.B & 2.C).

Next, the above experiments were repeated on a 1×2 grid. In large, the model behaviour is similar, however the migration rate has a larger impact on the smaller landscape. Readily a migration rate of 1% causes F_{st} to decrease to ca. 0.5 after 300,000 generations of migration (not shown here, but see figure 5.C in [7]). A migration rate of 10% causes the generic divergence of the two RI groups to decrease to insignificant levels within 50,000 generations of migration. However, RI can be sustained at up to 15% migration – the number of RI groups stays around 2 which shows that the significant gene flow is not sufficient to break RI. At 20% migration, RI collapses rapidly and the entire model population evolves as a single reproductive group (figure 5 in [7]). Next, the experiments were repeated on a 3×3 grid (not shown here). As expected, a larger grid makes it possible to sustain RI at higher migration rates. At 30% migration RI is sustained and the number of RI groups lies above 40. At 35% migration, RI collapses in a way similar to the previous scenarios.

As a basis for comparison, all of the above experiments were repeated without the HFL. In these runs all individuals are viable and selection is thus random. The detailed results of these control runs are presented in [7]. In summary, for a migration rate of 0%, gene divergence is clearly measurable and grows with time. However, for all grid sizes, a migration rate of 1% is sufficient to cause gene divergence to rapidly drop to a value around zero.

[1] The graphs in this paper were created and processed using the LiveGraph exploratory data analysis and visualisation framework [19].

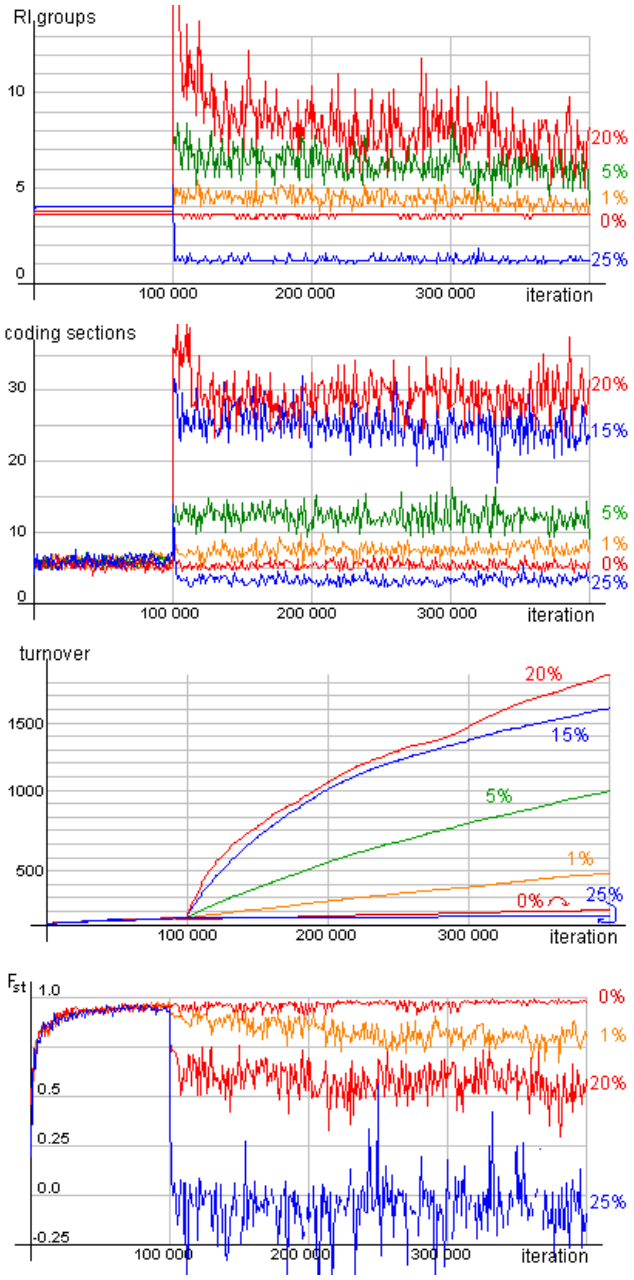

Figure 2. Evolution on a 2×2 grid for the migration rates 0% (red), 1% (orange), 5% (green), 15% (blue), 20% (red) and 25% (blue). Data averaged over 10 runs. Some values omitted for clarity.
A (top): The number of RI groups increases when the migration rate is higher. For very high migration rates the whole model population collapses into a single reproductive group. **B (2nd from top):** The number of distinct coding genotype sections in the population increases when the migration rate is high. As the population collapses to a single reproductive group at very high migration rates, the number of coding sequences falls. **C (3rd from top):** The rate of evolving new viable coding genotype sections increases when migration rate is higher due to drift in a larger number of IR groups and due to hybridisation between more RI groups. As the population collapses into a single reproductive group at very high migration rates, the turnover rate of coding sequences falls. **D (bottom):** Genetic divergence between RI groups measured using the fixation index. Higher migration rates lead to increased gene flow and thus lower genetic divergence.

5. DISCUSSION AND FUTURE WORK

The above results have important implications for the design of genetic and other evolutionary algorithms. Whereas in nature the "holey" structure of HFLs arises implicitly through gene incompatibilities, here the HFL was modelled explicitly. This explicit model may be used to support and, in some circumstances, to replace traditional niching algorithms. In particular, the island model [4] bears close resemblance to the current approach. In the present model the neutral loci were only used to measure genetic divergence, however, in a genetic algorithm they may be used to encode candidate solutions instead. The small effect of an increasing migration rate on the number of RI groups observed here implies that some hybrid populations exhibit real RI and are not simply fuelled by repeated hybridisation with immigrants. In population biology, hybrid populations that have strong genetic incompatibilities with the main population (such as those caused by HFL-genetics) are thought to be most stable [7]. In the current simulations such populations are short-lived because their small initial population size and the absence of prezygotic isolation make it unlikely that they successfully reproduce for a large number of consecutive generations. However, in the presence of a free niche, these hybrid groups can reproduce and persist. Such a niche may be given by under-explored areas of the solution space of an optimisation problem. Thus, the current model may be used as a mechanism for dynamic discovery and maintenance of multiple search directions in genetic and other evolutionary algorithms. Further experiments are necessary to verify how this approach performs for particular optimisation problems.

As mentioned above, the HFL model discussed here is explicit, while in nature, HFL arises implicitly because the majority of biochemically possible genotypes gives rise to inviable phenotypes. It is important to note that this is paralleled by many optimisation problems in which a valid solution encoding can result in an illegal or irrelevant solution. For instance, candidate solutions to the travelling salesman problem that contain incomplete loops or duplicate stops are inviable. Candidate solutions to multiobjective optimisation problems that conflict with one or more constrains can also be assumed inviable. Evolutionary search for problems for which viable genotypes (representations) build a connected cluster is carried out on a HFL. In such cases, the coding genes must not be considered either viable or inviable, as was the case here. Instead, the fitness of viable genotypes must be differentiated to express the goodness of a viable candidate solution. A better understanding of structure and dynamics of HFLs may provide new insights for the solution of such problems. A step in this direction has been undertaken in [10], where I provide a numerical analysis of a biological niche model with HFL genetics.

Natural populations evolving on HFLs have not evolved according to the aim of optimising the performance in some specific task. Similarly, it is not immediately clear that HFL-models will improve the performance of algorithms in the sense that better solutions may be found faster. Carefully engineered artificial methods can be expected to perform better for such measures. However, evolution on HFLs in nature leads to diverse populations that perform robustly under unexpected disturbances and are able to adapt to unforeseen circumstances. It is this type of tasks where engineering methods often fail and where HFL-based methods may yield a substantial benefit.

The notion of holey fitness landscapes, while largely unchallenged in biology, has arguably received insufficient attention from computer scientists. The current model shows that simulating plausible fitness landscapes can considerably change predictions about the maintenance of diversity and the emergence of new adaptations (novel solutions). Representing fitness landscapes in a biologically plausible way may facilitate ongoing adaptive exploration and the continuous generation of novelty in evolving problem solutions. The approach described here may be useful in further exploring these issues and in developing more flexible and powerful genetic and other evolutionary algorithms.

6. REFERENCES

[1] S. W. Mahfoud (1995). Niching Methods for Genetic Algorithms. PhD thesis. Urbana Iniversity of Illinois.

[2] K. A. De Jong (1975). Analysis of the behaviour of a class of genetic adaptive systems. PhD thesis University of Michigan.

[3] D. E. Goldberg and J. Richardson (1987). Genetic algorithms with sharing for multimodal function optimization, *2nd Int. Conf. on GAs and their application*, pp. 41-49 Lawrence Erlbaum Associates Inc., Mahwah, NJ, USA.

[4] D. Whitley, S. Rana and R. B. Heckendorn (1999). The Island Model Genetic Algorithm: On Separability, Population Size and Convergence. *J. of Computing and IT*. 7 (1): pp. 33-47.

[5] J. Liu, Z. Cai and J. Liu (2000). Premature convergence in genetic algorithm: Analysis and prevention based on chaos operator, *Procs 3rd W. Congr. Intell. Ctrl and Automation*.

[6] N. Kashtan, E. Noor and U. Alon (2007). Varying environments can speed up evolution. *PNAS.* 104 (34).

[7] G. Paperin, S. Sadedin, D. G. Green and A. Dorin (2008). Holey Fitness Landscapes and the Maintenance of Evolutionary Diversity *Submitted to 11th International Conference on Simulation and Synthesis of Living Systems (ALife XI)*.

[8] S. Gavrilets (2004). *Fitness Landscapes and the Origin of Species*. Princeton University Press, Princeton / Oxford.

[9] S. Gavrilets and J. Gravner (1997). Percolation on the Fitness Hypercube and the Evolution of Reproductive Isolation. *J. of Th. Bio*. 184 (1): pp. 51-64.

[10] G. Paperin, D. G. Green and A. Dorin (2007). Fitness Landscapes in Individual-Based Simulation Models of Adaptive Radiation. In T. D. Pham and X. Zhou (eds.), *2007 Int. Symposium on Computational Models for Life Science* .

[11] G. Paperin, D. G. Green, S. Sadedin and T. G. Leishman (2007). A Dual Phase Evolution model of adaptive radiation in landscapes. In M. Randall, H. A. Abbass and J. Wiles (eds.), *The 3rd Australian Conference on Artificial Life (ACAL'07)*, Springer.

[12] S. Gavrilets (2003). Models of Speciation: What have we learned in 40 years? *Evolution*. 57 (10): pp. 2197-2215.

[13] M. E. J. Newman and R. M. Ziff (2000). Efficient Monte Carlo Algorithm and High-Precision Results for Percolation. *Physical Review Letters*. 85 (19): pp. 4104-4107.

[14] G. R. Grimmett (1999). *Percolation*. Springer.

[15] S. Gavrilets and A. Vose (2005). Dynamic patterns of adaptive radiation. *PNAS*. 102 (50): pp. 18040-18045.

[16] T. Ohta and M. Kimura (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a population. *Gen. Res*. 22 (2): pp. 201-204.

[17] S. Van Dongen (2000). Graph Clustering by Flow Simulation. PhD thesis University of Utrecht. Utrecht.

[18] R. R. Hudson, M. Slatkin and W. P. Maddison (1992). Estimation of Levels of Gene Flow From DNA Sequence Data. *Genetics*. 132 (2): pp. 583-589.

[19] LiveGraph - a framework for real-time data visualisation, analysis and logging. Retrieved on 01.03.2008 from: http://www.live-graph.org.