To the Graduate Council:

I am submitting herewith a thesis written by Hai Li entitled "Modeling Speciation on Holey Adaptive Landscapes." I have examined the final copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Science, with a major in Computer Science.

Michael D. Vose, Major Professor

We have read this thesis
and recommend its acceptance:

Brad Vander Zanden

S Gavrilets

Accepted for the Council:

Associate Vice Chancellor
and Dean of The Graduate School

# MODELING SPECIATION ON HOLEY ADAPTIVE LANDSCAPES

A Thesis

Presented for the

Master of Science

Degree

The University of Tennessee, Knoxville

Hai Li

May 1998

# ACKNOWLEDGEMENTS

There are many people who have contributed in various ways toward the completion of this thesis. First and foremost, I would like to thank my committee members, Dr. Vose, Dr. Gavrilets, and Dr. Vanderzanden for their guidance and patience in the development and preparation of this thesis.

I would also like to thank all of my friends for their support during my years in graduate school. They were always there when I needed help.

Last but not least, my very special thanks to all of my family members for their faith in me. Without their encouragement, this thesis would not have been possible.

# ABSTRACT

Sewall Wright's powerful metaphor of "rugged adaptive landscapes" (1932) has been the basis for studying the dynamics of evolution and speciation for more than sixty years. However, the validity of this model has been debated and has come under doubt. An alternative model, "holey adaptive landscapes" has been described by Gavrilets (1997) for the study of dynamics of evolution and speciation of macro-evolution. The work presented in this thesis is a implementation of a base simulation model for both rugged and holey adaptive landscapes. It is the first ever implementation that closely models real organisms in terms of the number of loci supported (in hundreds). By observing the dynamics of the "digital organisms" in simulation, our intuition can be trained and important characteristics and parameters can be identified in simulated evolution and speciation. Results from simulation studies with the base model on holey adaptive landscapes are also presented.

# TABLE OF CONTENTS

CHAPTER PAGE

# LIST OF FIGURES

# LIST OF TABLES

**TABLE** **PAGE**

# CHAPTER 1

# Introduction

## 1.1 Overview

Evolutionary biology is the study of how the present diversity of plants, animals arrived to their current stage of form from the earliest and most primitive organisms. It is a science that covers a large range of topics, from the study of molecular genetics to animal behaviors in the jungles [12]. The work presented here is in the area of population genetics, one of the major research areas of evolutionary biology. Population genetics is the study of of the distribution and dynamics of inherited variation among a group of organisms of the same species [11].

In 1932, Sewall Wright published a paper that used the metaphor of multi-dimensional rugged adaptive fitness model to discuss evolution and speciation [14]. The emphasis of this model is on the adaptive peaks and valleys where speciation is perceived to be the transition across adaptive valleys to adaptive peaks. This particular framework has been the basis of studying the dynamics of evolution and speciation for more than sixty years.

An alternative to the rugged adaptive landscape model, holey adaptive landscapes has been described by Gavrilets (1997). Instead of focusing on adaptive

peaks and valleys, emphasis is placed on the existence of "ridges" of highly fit genotypes. Speciation occurs when they come to the opposite ends of "hole" in a holey adaptive landscape. A theoretical study of this model is presented in paper by Gavrilets and Gravner [6]. The holey adaptive landscapes model is more suitable for the study of macro-evolution as opposed of micro-evolution with rugged adaptive landscapes.

The goal of this thesis is to implemented a simulation model that is capable of simulating evolution with both rugged and holey adaptive landscapes with organisms that have large number of loci. As will be shown in Chapter 4, evolution of organisms with close to 400 loci have been simulated. This is a vast improvement considering previous simulation models have only been able to simulate organisms with less than 30 loci [3]. This is a significant distinctive feature of the model implementation. In comparison, recent biology researches have shown that the number of loci in the *Escherichia coli* (*E-coli*) is about 4288 [1].

The significance of this study is that, by simulating evolution of organisms with large number of loci, we are more closely modeling the evolution of real organisms. Being able to simulate models to study realistic complex system is important because of difficulties with experimenting with the real system. This is true in Cosmology, Seismology, Nuclear Physics as well as Biology. Evolution has taken place over millions of years and it is not possible to experiment with this complex system over similar time scales. In a simulation we use a computer to evaluate a model *numerically*, and data are gathered in order to estimate the desired true

characteristics of the model [8]. Results learned from numerical simulations can be performed much more quickly and then used to to train our intuition about the order of the real system and identify important parameters and components.

In implementing the model, every attempt have been made to be efficient without sacrificing functionality. A linear random number generator from Vose [13] is used to generate necessary random numbers for the model. The implementation of one of the fitness function used is obtained from Lemke and Campbell [9]. Additional algorithms for generating fitness values, calculating distance distributions and mating functions are developed with guidance from Vose and Gavrilets.

## 1.2 Thesis Outline

In the next chapter, background concepts on evolutionary biology as well as population genetics, definitions of biology terms that are necessary for the understanding of this paper, and literature reviews of both rugged and holey adaptive landscapes will be presented. Chapter 3 will describe the code implementation and discuss the design of some of the essential components of the model. In the last chapter, analysis of the model, some results that were obtained as well as potential future enhancement to the model will be presented.

As this paper progresses, the reader will notice many similarities between the model implementation with typical implementation of Genetic Algorithms (GA) [7]. This is true. In implementing the model, coding techniques based on

implementations of genetic algorithms have been used. An example is the implementation of organisms as bit strings and the implementation of the mating and crossover function. However, it is important to point out here that the fundamental difference between the two is that, this paper focus on the studies of gene pool, or the population behavior, rather than finding the fittest individual (or solution, in Genetic Algorithms). In addition, terms used in this paper come from the field of biology and do not necessarily reflect their use in the field of genetic algorithms.

# CHAPTER 2

# Concepts and Literature Review

## 2.1 Overview

This chapter describes some background concepts as well as defines terms that are used in this paper in describing the model implementation. Please note that these definitions are provided only for the benefits of understanding the model presented here and therefore should not be treated as precise scientific definitions. For stricter and more complete definitions of these terms, one can consult reference [11] as listed in the bibliography. In section 2.3, the definition for adaptive landscapes is provided as well as two interpretations of this definition. In the last section, the articles mentioned in the first chapter on rugged and holey adaptive landscapes are reviewed.

## 2.2 Definitions

Evolution itself means change. The subjects of evolutionary biology is evolution and speciation. There are some well recognized facts about species in nature and how they evolve. It is recognized that species are not fixed, but evolve to form new species. This process is referred to as *speciation.* It is also recognized

that natural selection plays a significant role in the evolutionary process which is also influenced by external environmental changes and random genetic innovation (*mutation*). When multiple species can not mate because of geological, seasonal, or behavioral properties, we say they are *reproductively isolated* from each other. Speciation is often described as the process by which a population of a group of inter-breeding populations becomes reproductively isolated from another population, or group of inter-breeding populations [2]. In the biological evolution process, species will sometime acquire properties that will enable them to survive and reproduce. This process is referred to as *adaptation.* In studying evolution, we use the word "*fitness*" to indicate the measure of viability or fertility of an organism which determines its probability of contributing to next generation.

A *population* consists of a group of organisms or individuals, where an *organism* is a individual living system that is capable of reproduction, growth and self maintenance. The genetic information of organisms are encoded in *chromosomes*, which consist of *genes.* In population genetics, gene is sometime used interchangeably with *allele*, which has a biological definition indicating alternative forms of genes. A gene is the fundamental unit of the living system that determines a particular characteristic of the organism. An *haploid* organism is an organism with a single set of unpaired chromosome. An *diploid* organism is an organism with twice the haploid number of chromosomes. In general, bacteria, algae and fungi are haploid organisms, where higher plants, animals are diploid [10]. The *locus* is the position of a gene in an organism, the plural form of which is *loci.* The

*genotype* of an organism is its genetic composition. The collection of all genotypes is referred to as *genotype space.* The study of the relationship of genotype and its associated fitness is a major focus of this research because this relationship "is one of the most important factors in determining the evolutionary dynamics of populations" [4]. The tendency of organisms to be more likely to mate with others who resemble them is referred to as *assortive mating.*

## 2.3 Adaptive Landscapes

*Adaptive landscape* refers to a multivariate function (defined on the genotype space) that gives a genotype's fitness. It is usually represented by a graph of a surface. In the research of population genetics, there are two typical interpretations of adaptive landscapes used in the discussion of evolution and speciation.

In the first interpretation, an adaptive landscape is the graph of the mean fitness of the population as a function of gene frequencies. In this representation, each point on the surface represents an individual population. The difficulty with this type of representation is that because neither population size nor genetic variability are represented in the picture, it is hard to model or detect the forming of new species.

In the second interpretation, an adaptive landscape is the graph of the individual fitness as a function defined on the genotype space. In this representation, each population is represented by a cloud of points on the graph. This represen-

tation of adaptive landscapes is the more fundamental representation of the two and is used in the discussion of adaptive landscapes in this paper [4].

## 2.4 Literature Review

In this section, we will review two papers that are closely related to the topics of the study. The first is the paper by Sewall Wright, titled *"The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution"* [14]. The second paper is by Sergey Gavrilets, titled *"Evolution and Speciation on Holey Adaptive Landscapes"* [4]. Both of these papers are on the subject of adaptive landscapes.

### 2.4.1 Sewall Wright

The paper by Wright (1932) is the first paper that used the concept of contoured graph in order to explain the adaptiveness of species (or combinations of genes). This contoured graph representation is shown in Figure 2.1. The corresponding rugged landscape in three dimension is shown in Figure 2.2 [4]. The rugged adaptive landscape contains many local "adaptive peaks" and "adaptive valleys", where species are represented by adaptive peaks and unfit hybrids are represented by adaptive valleys. With this analogy, evolution is viewed as sort of "hill-climbing" [10] where populations, through natural selection, mating, and mutation, move to "nearby" adaptive peaks to form new species. The multiple scenarios of such movement is described in the paper. As mentioned in the first chapter, this particular model has been the basis of research in evolution and spe-

ciation for more than sixty years. The emphasis of this in particular is on adaptive peaks and valleys where evolution takes place via crossing valleys. As such, it is a more suitable model for the study of micro-evolution.

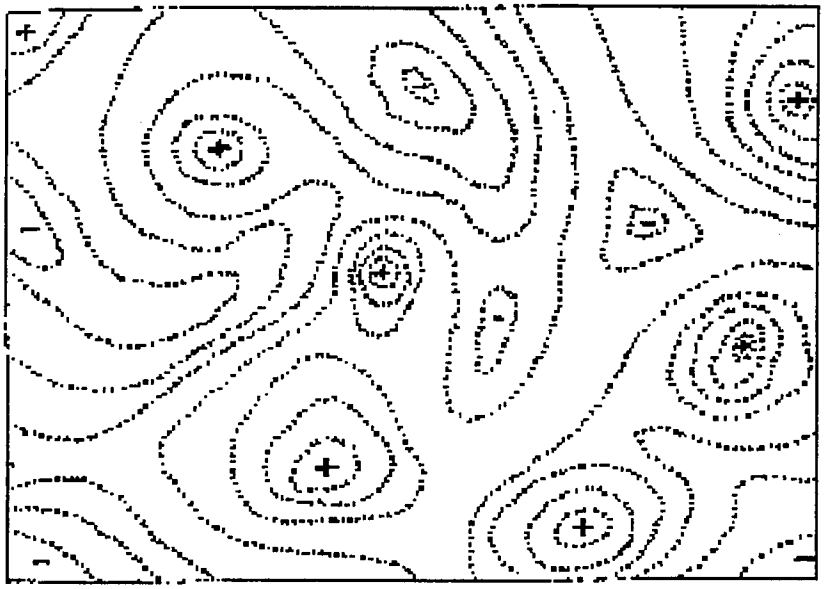

Figure 2.1: Contour Graph - Peaks and Valleys

As pointed out by recent papers, the implicit assumptions made by this particular model are questionable [6]. First of all, the assumption that different species have different fitness may not be valid in the study of speciation [12]. Another fundamental question is "how can a population evolve from one local peak to another across an adaptive valley when selection opposes any changes away from the current adaptive peak?" The shifting-balance theory offered by Wright to answer this question can work in theory, but it may not be feasible for a comprehensive analytical and numerical study [10]. In addition, conditions of this theory are too strict to be satisfactory [4].

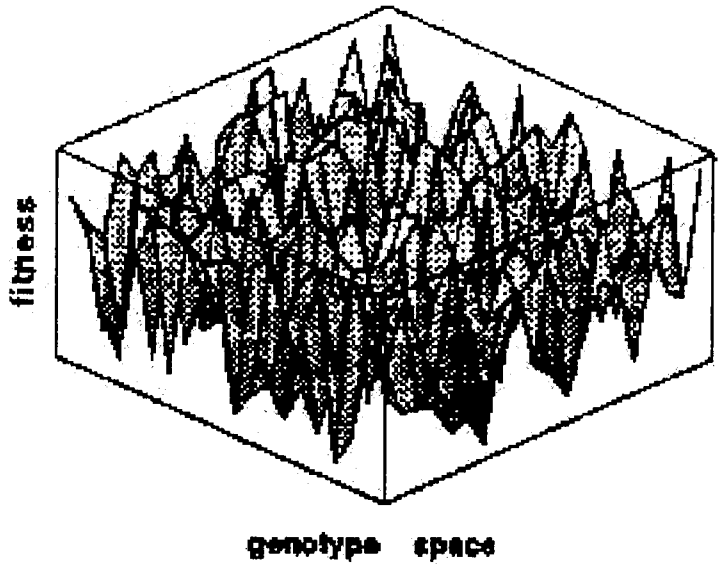


Figure 2.2: Rugged Adaptive Landscapes

One of the study that is based on the concept of rugged adaptive landscapes is the paper by Finjord [3]. This paper is not relevant to the work presented here other than to compare the simulation implementation between Finjord's model and the model implemented here. In Finjord's implementation, studies were performed on population sizes of around 100 and genome size of about 20. As will be shown in the later sections, the model implemented here is capable of handling problems of at least one magnitude larger.

### 2.4.2 Sergey Gavrilets

The paper by Gavrilets on holey adaptive landscapes is an alternative to the rugged adaptive landscapes model. Instead of viewing evolution as crossing adaptive valleys to adaptive peaks, it is viewed as movement of species along the

"ridges" of highly fit genotypes that extend throughout the genotype space. The model defines individuals as having only two different fitnesses, 1 (viable) or 0 (not viable). In doing so, it creates adaptive landscapes which in three dimensions are just surfaces with many holes (see Figure 2.3). Another basic assumption made by this model is that fitness values are generated randomly as well as independently. The study of correlated fitness landscapes is a direction for future studies. The author explored the question of whether rapid speciation can occur under strong selection, many loci and finite population. The holey adaptive landscapes is a step toward bridging the gap between micro-evolution and macro-evolution. Both adaptive landscapes can be studied by simulating evolution with holey adaptive landscapes as implemented here.

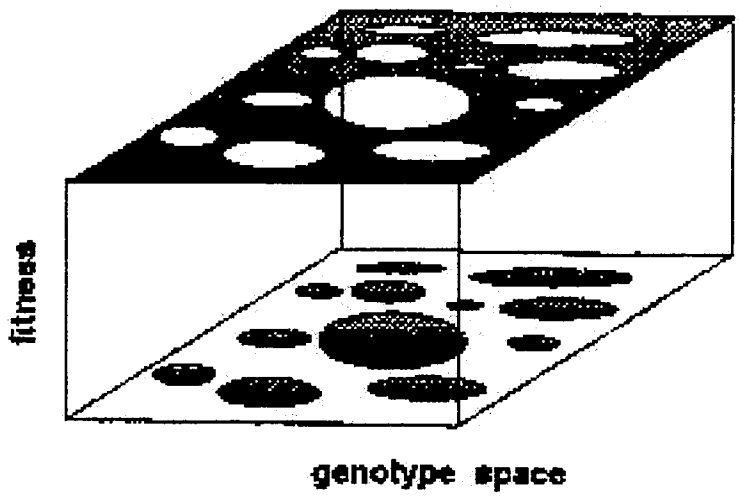


Figure 2.3: Holey Adaptive Landscapes

# CHAPTER 3

# Code Implementation

## 3.1 Overview

The following is a list of criteria used in guiding model implementation:

- it must be handle large problem sizes
- it must be able to handle simulations with both rugged and holey adaptive landscapes
- it must be flexible where key variables are parameterized for execution without recompiling
- it must be efficient (i.e. fast)

The problem size of the model is defined to be a combination of the population size, the number of loci in the chromosome and the number of evolution cycles. Simulations done previously on rugged adaptive landscapes have only handled a population size of around 100, and a chromosome length around 20. It is our goal to be able to handle these parameters at least one magnitude larger. To minimize the execution time of the model, a linear time random number generator [13] is used. Genetics of the population are stored in binary strings, and mating is accomplished with binary bit-wise operators. The mutation operation is accomplished by flipping bits from 0 to 1 or 1 to 0 depending on a specified probability.

The model implementation is fairly straightforward. The following is a high-level overview:

```
begin
  parameters (static and command line)
  initialization
  repeat
    premating selection
    mate
    postmating selection
  until N generations
  write results
end
```

Each of the steps as well as some key components not listed (such as the fitness function) will be described in detail in the next few sections.

### 3.2 Parameters

Table 3.1 and table 3.2 are lists of both static and command line parameters used in the modeling code. By explaining the use of each parameter, it is hoped that the reader will gain some insights into the features available in the model which would otherwise be too lengthy to describe.

The static definitions are located in the global header file global.h. They are the most basic parameters that control the data structure and memory space for code execution. *HAPLOID/DIPLOID* is used to specify the type of organism to simulate. Both type of organisms are implemented in the code with matching operations appropriate for each. *CHROMOSOMES* is used to specify the number

Table 3.1: Static Parameters (global.h)

| HAPLOID | DIPLOID |
|---|---|
| CHROMOSOMES | CHROMOSOMELENGTH |
| RANDOM_MATING | PREMATING_RI |
| POSTMATING_RI | LOCATIONS |
| SELECTION_ON_TRAIT | RANDOM_FITNESS |
| HAMMING_DISTANCE | |

Table 3.2: Command-line Parameters

| Seed | Popsize | Locsize |
|---|---|---|
| Max_Mating | Iterations | Max_Difference |
| Min_Fitness | Mrate | Crate |
| Loc_prob | Drift_Save | Av_Hist_Save |
| Slide_Save | | |

of number of loci and *CHROSOMELENGTH* is used to specify the number bits to represent genes at each loci. If *RANDOM_MATING* is defined, all selection and mating are accomplished in a random fashion. This gives a base scenario where no control is placed on either the selection of individuals for mating or the fitness of the individual as a result of mating. *PREMATING_RI* and *POSTMATING_RI* determine whether either pre-mating or post-mating reproductive isolation should be applied. *LOCATIONS* specifies the number of locations in the population. This static variable is typically used in combination with *PREMATING_RI* for location-based reproductive isolation. *SELECTION_ON_TRAIT* controls whether the default fitness function or trait-based fitness should be used (explained in section 3.6.1). *RANDOM_FITNESS* causes fitness values to be assigned randomly without regard to the genetic composition (bit pattern) of the

organism. *HAMMING_DISTANCE* determines whether hamming distance or genetic distance function should be used (by default, genetic distance is used).

The values for each of the following variables can be supplied on the command line at code execution time. *Seed* is used to initialize the random number generator. *Popsize* stores the overall size of the population (including all locations). *Locsize* specifies the size of the subpopulation at each location. *Iterations* specifies the number of evolutions to simulate. *Max_Difference* specifies the maximum percentage of difference (bitwise) that is allowed when mating two individuals. This is used in pre-mating reproductive isolation. *Min_Fitness* specifies the minimum fitness of the new individuals as result of mating. When this variable is greater than 0.0, it serves the purpose of post-mating reproductive isolation where either a individual is viable, or not. *Max_Mating* is the maximum number of mating attempts that will be made in order to produce a viable new individual for the next population (more detail will be provided in the mating section). *Crate* and *Mrate* specify the crossover and mutation rates. Mutation rate specifies the probability of a bit change in an organism. Crossover rate specifies the probability of crossover at each loci. *Loc_prob* specifies the probability of mating between locations. The effect of this variable can be overrode by specifying a inter-location mating probability matrix where these probabilities are non-uniform between locations. *Drift_Save*, *Av_Hist_Save* and *Slide_Save* are control parameters that determine if genetic drift of the populations, average histogram of the populations, and a slide file should be saved during the evolution process.

## 3.3 Initialization

The initialization process is divided into three parts. In the first part, command line variables as listed in section 3.2 are assigned either the default, or user specified values. The random number generator is initialized by either the specified seed, or the current time (obtained with a system call). The fitness lookup tables, crossover masks, mutation masks that conform to the specified distribution are pre-generated. The fitness lookup table will be discussed in more detail in the postmating reproductive isolation section.

The crossover masks are generated based on a user specified crossover probability. A crossover mask is a series of 0s and 1s where an 1 bit indicates a crossover point (where the swapping of bits occur between two individuals). Since organisms are stored in binary strings of varying length, the current implementation handles two different cases. In the first case, the length of each individual organism is less than 16 bits, all possible crossover masks for that length are generated and assigned appropriate probabilities. In the case where the length of the organism is greater than 16, it is required by the model that the length of the organism then be a multiple of 16. All possible 16-bit crossover masks are generated and assigned appropriate probabilities. This particular implementation of crossover is uniform, where the probability of a crossover at each bit position is the same. The mutation masks are generated much in the same way as the crossover masks. An 1 bit in the mask indicates that bit position should be flipped.

In the second section, the initial population, referred to as *Eden*, is generated randomly. All individuals at each location are the same at start up, but generated randomly. The reason behind this is so that genetic variability can be tracked between evolutions by comparing each new population with the original.

In the third part, the X display is initialized. A default window size of 800 by 500 pixels is used to graph the hamming distance histogram on a per location, per iteration basis. Figure 3.3 shows an example of the X-window used in simulation. At the upper left hand corner are three numbers which are used to indicate the location number, the size of the population at the location, and the generation number. Although in the initial phase of the experimentation, population size does not change, it is displayed in anticipation of future enhancements allowing the size of the population to change depending on the evolution cycle and other parameters.

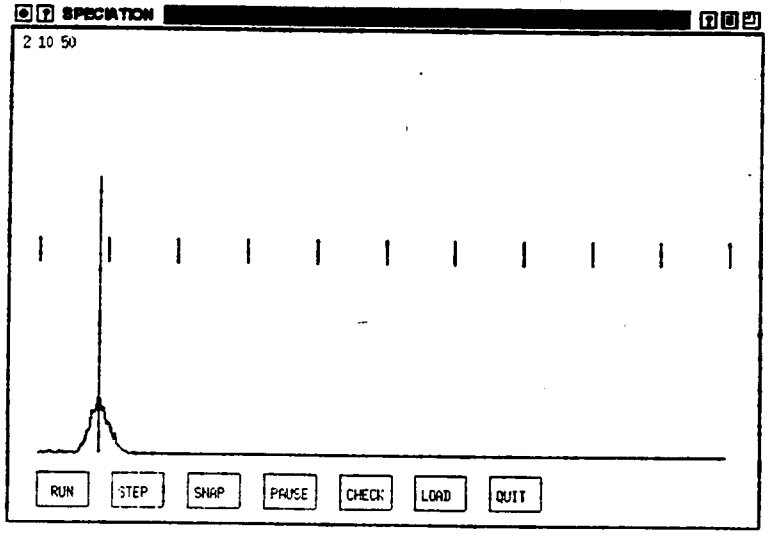


Figure 3.1: X Display for Simulation

There are seven buttons available in the window for controlling the course of the simulation. Through the use of these buttons, one can pause the simulation, step through each evolution cycle, get a snapshot, save the evolution state, retrieve previously saved evolution state, or quit in the middle of the simulation.

## 3.4 Premating Selection

Premating selection is the process of selecting two individuals from a original or intermediate populations to form the new individual for the next generation. The selection of individuals is a key component of the mating process. The implementation allows for a couple of methods of selecting the pair of individuals.

When there is only one location in the population, selection of the first individual is made randomly. The selection of the second individual is based on either the hamming distance distribution or the genetic distance distribution of all other individuals with respect to the first individual.

When the population is subdivided into multiple locations, an extra step is taken in the selection of the second individual. First, the location for choosing the second individual is selected according to user specified inter-location mating probabilities. Then, the hamming distance or the genetic distance (depending on model parameter) of all individuals at that location with respect to the first individual is calculated. The selection of second individual is based on this distance distribution where the more "similar" (smaller distance), the higher the

probability of being chosen.

The premating reproductive isolation operation is also applied at this stage of the evolution. A comparison is made of the distance of the two individuals with the maximum difference threshold. Should this threshold be exceeded, the second individual is re-picked from the population. After a specified number of failed attempts to locate a mate have been made, the first individual is re-selected and the process repeats.

## 3.5 Mate Function

The mating function, given two individuals, applies the crossover and mutation operations and produces a new individual. Because organisms are stored as binary strings, the crossover and mutation operations are implemented using binary bitwise AND and OR operators.

The crossover operation is carried out by applying crossover masks to each of the two individuals and then combining the resultant strings to form the new individual before the mutation masks is applied.

## 3.6 Postmating Reproductive Isolation

When postmating reproductive isolation is applied, the fitness value of the new individual is compared with the specified minimum fitness. Should the new individual's fitness be less than the threshold, the selection and mating process is

repeated again until a viable individual is produced. This comparison, in effect, is where the assumption of fitness values being either a 1 (viable) or 0 (not viable) is implemented.

As described earlier, in order to study the general properties of the holey adaptive landscapes, the fitness values of organisms are generated randomly as well as independently. Two methods of generating such fitness values have been implemented and are described in the next section.

### 3.6.1 Fitness Functions

Both implementations of the fitness functions as described below generate random, non-correlated fitness values where a single bit change in the organism will result in a completely different fitness value.

The first fitness function implemented is based on the "random walks in a closed space" algorithm as described in [9]. The basic idea behind this algorithm is that the entire bit string is evaluated in arriving at a random number (fitness value). Each individual bit contributes differently to the final fitness value. The algorithm is implemented by a table lookup method because values for 16-bit segments can be pre-calculated and an appropriate multiplier can be applied later, depending on the positions of the bits.

The second fitness function implemented is a trait based method. With this method, all possible 16-bit binary patterns are pre-generated. For each of the 16 bits, a randomly generated "weight" is assigned to that bit. The weight of each bit

pattern is calculated by summing up the weights associated with each of the "1" bit. The fitness of a the whole organism by summing up all the 16-bit segments.

The first method of generating fitness function is currently being used because it is more robust and has been proven to produce non-correlated random numbers as long as the values of some parameters stay within specification [9].

## 3.7 Outputs

Standard measures in simulations of evolution are output by the program. These include: genetic drifts within each location, average distance between locations, unique segments counts of each location and other useful information such as hamming distance distribution within each location.

An additional feature of the implementation is that each hamming distance distribution for each population or subpopulation can be saved to a file. Therefore, the simulation of evolution can be saved for later viewing and analysis. This is useful when batch jobs in simulation experiments.

## 3.8 Additional Codes

Some of the additional works associated with implementing the code are:

- a separate program (slideshow) can be used to replay simulations
- a separate program (eval_fit) was used to generate random fitnesses for analysis with *xmgr* application

- result files such as genetic drifts and count of unique organisms in the population are graphed with *xmgr* application

In the next chapter, some results from the empirical studies of holey adaptive landscapes obtained from this implementation is described.

# CHAPTER 4

# Analysis

## 4.1 Overview

In this chapter, we will review the code implementation as outlined in the previous chapter. Results from simulation studies of holey adaptive landscapes using the code is presented. These results and details on the setup of these simulations are from the paper "Rapid Speciation on Holey Adaptive Landscapes", authored by Gavrilets, Li and Vose. The article is included as Appendix A of this document. In the last section, potential future enhancements to the code implementation are discussed.

## 4.2 Analysis

The implementation of the base simulation model has met the objectives as set forth in section 3.1. Trial simulations of evolutions of organisms with close to 1000 loci have been done without difficulty. The maximum number of loci is limited by the amount of available memory. Both rugged and holey adaptive landscapes are supported. The code is divided into separate components where any enhancement or modification to the implementation can be easily accomplished

in a modular fashion. By using command line parameters, adjustment to mutation rate, crossover rate, population size and those parameters associated with reproductive isolations can be easily made without recompiling the executable. The execution time of the program is bounded by $O(n^2)$ due to the fact that, in the selection process, the second individual might be selected from the rest of the population based on distance distribution of all other individuals with respect to the first.

### 4.3 Simulation with Holey Adaptive Landscapes

Results presented in this section are based on simulations with holey adaptive landscapes experimenting with strong natural selection, finite populations and organisms with large number of loci. The number of loci in the simulation is about 400, compared to the the *E-Coli* bacteria which has about 4000 loci.

Three different characteristics of evolution and speciation with holey adaptive landscapes (with haploid organisms only) were observed and results are described in the next three sub-sections (see Appendix A for a complete description).

The symbol $K$ is the premating reproductive isolation parameter where two individuals can only mate if they are different at no more than $K$ loci. Note that when $K$ equals to the number of loci, then there is no premating reproductive isolation. The symbol $N$ represents the number of individuals in the population. The number of loci with all organisms used in each of the three simulations is 384.

Mutation rate is fixed at $10^{-4}$ and crossover (recombination) rate between adjacent loci is set to 0.005. Unless otherwise specified, these parameters stayed the same for each of the three studies. The mating probabilities between locations varied depending on the aspect of each study. For each one of the figures, average genetic distances (AGD) within and between each location are graphed with respect to the number of generations.

### 4.3.1 Effects of Premating Isolation

The first sets of simulations are on the studies of effects of premating isolation on evolution and speciation. Two subpopulations (locations) are considered in the study. The migration rate (probability of mating between locations) is set to 0.01. Figure 4.1 illustrates three kind of dynamics. The top lines in each of the figure is the AGD between locations where bottom two lines are the AGD within each location.

In Figure 4.1a, no genetic divergence and speciation took place within 3000 generations. With a reduced $K$ value (restriction on migration / strong selection), in Figure 4.1b, genetics of subpopulations diverged quickly and the subpopulations become reproductively isolated from each other. With the same restrictions on migration as in Figure 4.1a, the increase in $N$ sped up the genetic divergence between two subpopulations (as shown in Figure 4.1c).

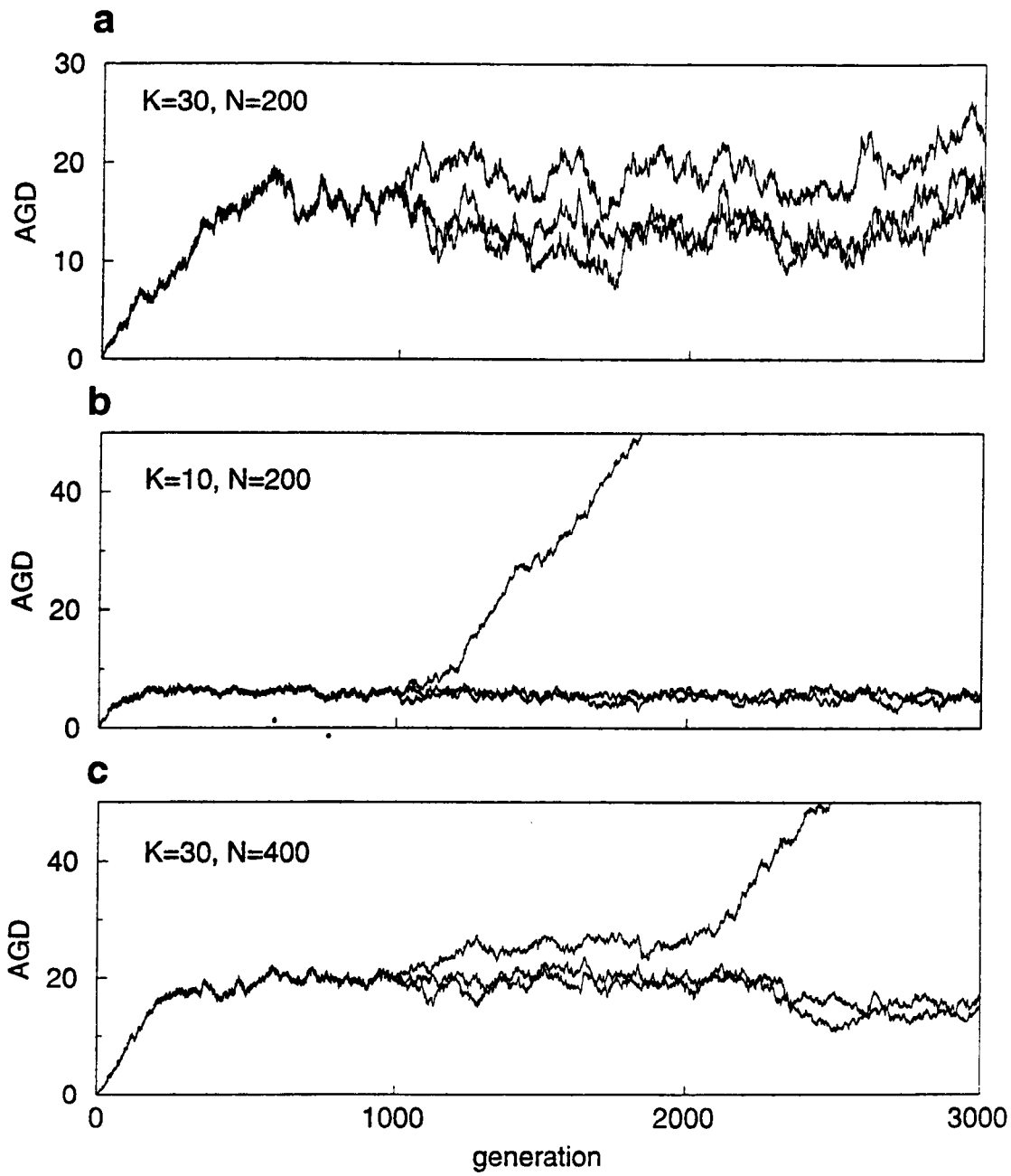


Figure 4.1: Effects of Premating Isolation

### 4.3.2 Effects of Population Size

The second sets of simulations are on the studies of effects of population size on speciation. The number of individuals $N$ was set to 200 and 400. Migration rate and other parameters (except $K$) are the same as in the previous section. Simulations were run for $K$ of 10 and 30 for each $N$. As shown in Figure 4.2(a,b), with strong selection ($K$ is small), the increase in population size actually delayed or prevented genetic divergence or speciation. However, as shown in Figure 4.2(c,d), the opposite is true when selection is weak (large $K$). Explanations for such dynamics can be found in Appendix A.

### 4.3.3 Effects of Population Subdivision

The third sets of simulations are on the studies of effects of population subdivision on evolution and speciation. For all diagrams in Figure 4.3, the population size $N$ was fixed at 800 and $K$ set to 20.

In Figure 4.3a, a single run of two subpopulations shows that no speciation took place after 3000 runs. However, by dividing population into four subpopulations, all AGD between locations increased after 1000 generations, indicating the emergence of 4 species (Figure 4.3b). Migration rate for both cases is kept at 0.01.

Figure 4.3(c,d) corresponds to a stepping-stone model (mating can only take place between neighboring locations) with migration rate $m/2 = 0.03$. In Figure 4.3c, the population is divided into 8 subpopulations. Only AGD between sub-

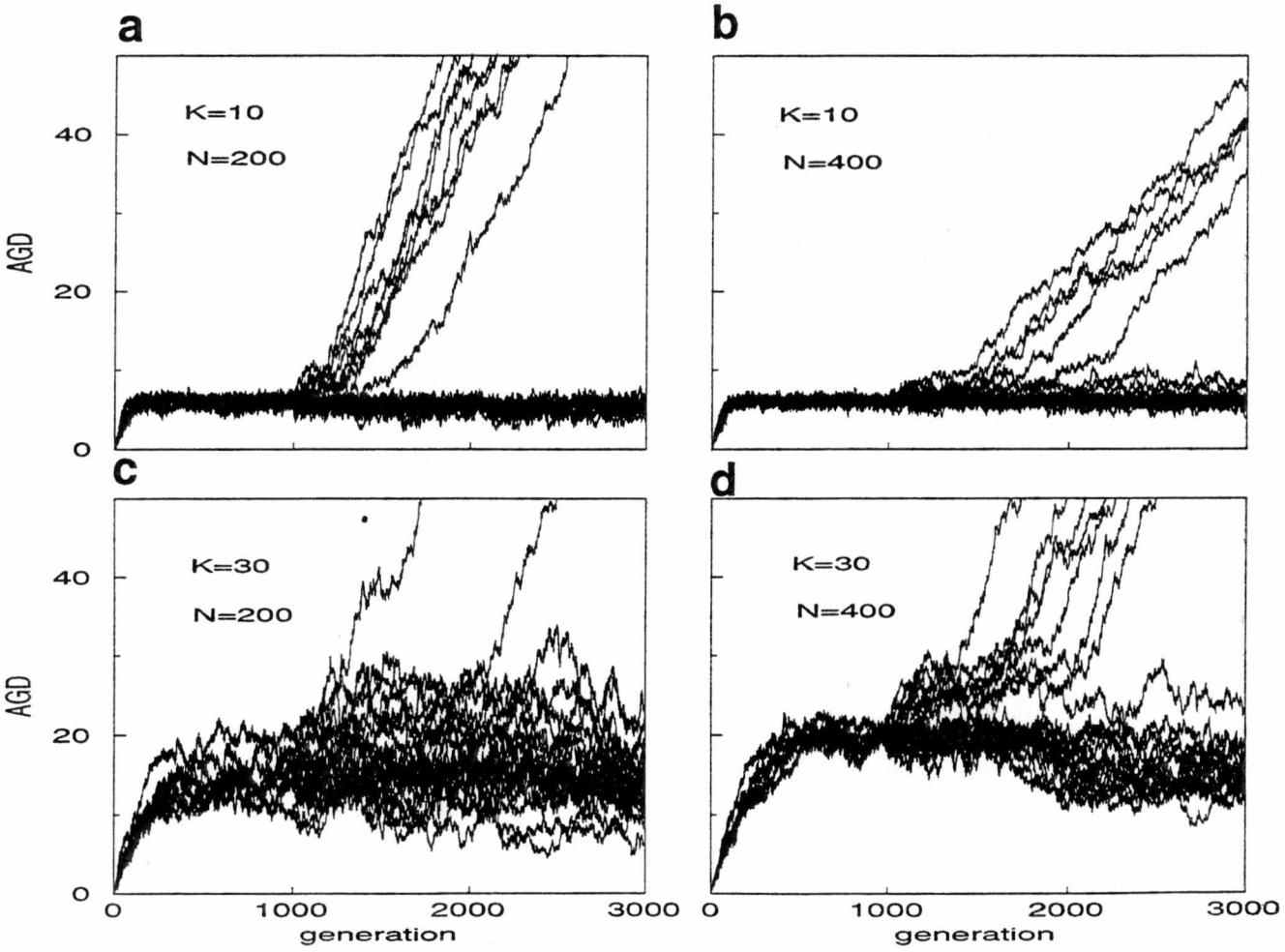


Figure 4.2: Effects of Population Size

populations are graphed. All seven distances remained below $K$ except between subpopulations 4 and 5 indicating the split into two species with 4 subpopulations each.

For Figure 4.3d, the population is divided into 4 subpopulations. Of the three curves in the figure, the bottom line is the AGD between neighboring subpopulations, the line above that is the AGD between subpopulations that are separated by one other subpopulations. The top line is the AGD between subpopulation 1 and 4. This corresponds to the intuition that genetic distance between subpopulations increases with the geographic distance.

### 4.3.4 Summary of Simulation Studies

What we learned from the simulation studies is that:

> Genetic divergence along the ridges in the adaptive landscapes can lead to rapid speciation (allopatric or parapatric). Increasing the number of considered loci has significantly increased the plausibility of speciation relative to that in earlier models. The plausibility of speciation is enhanced by population subdivision; simultaneous emergence of several new species from a subdivided population is highly probable.

## 4.4 Future Research

There are several aspects of the base model implemented here that can be added or improved upon. There is no question that the code can be optimized further by the use of "table lookups". An example of such is the current method

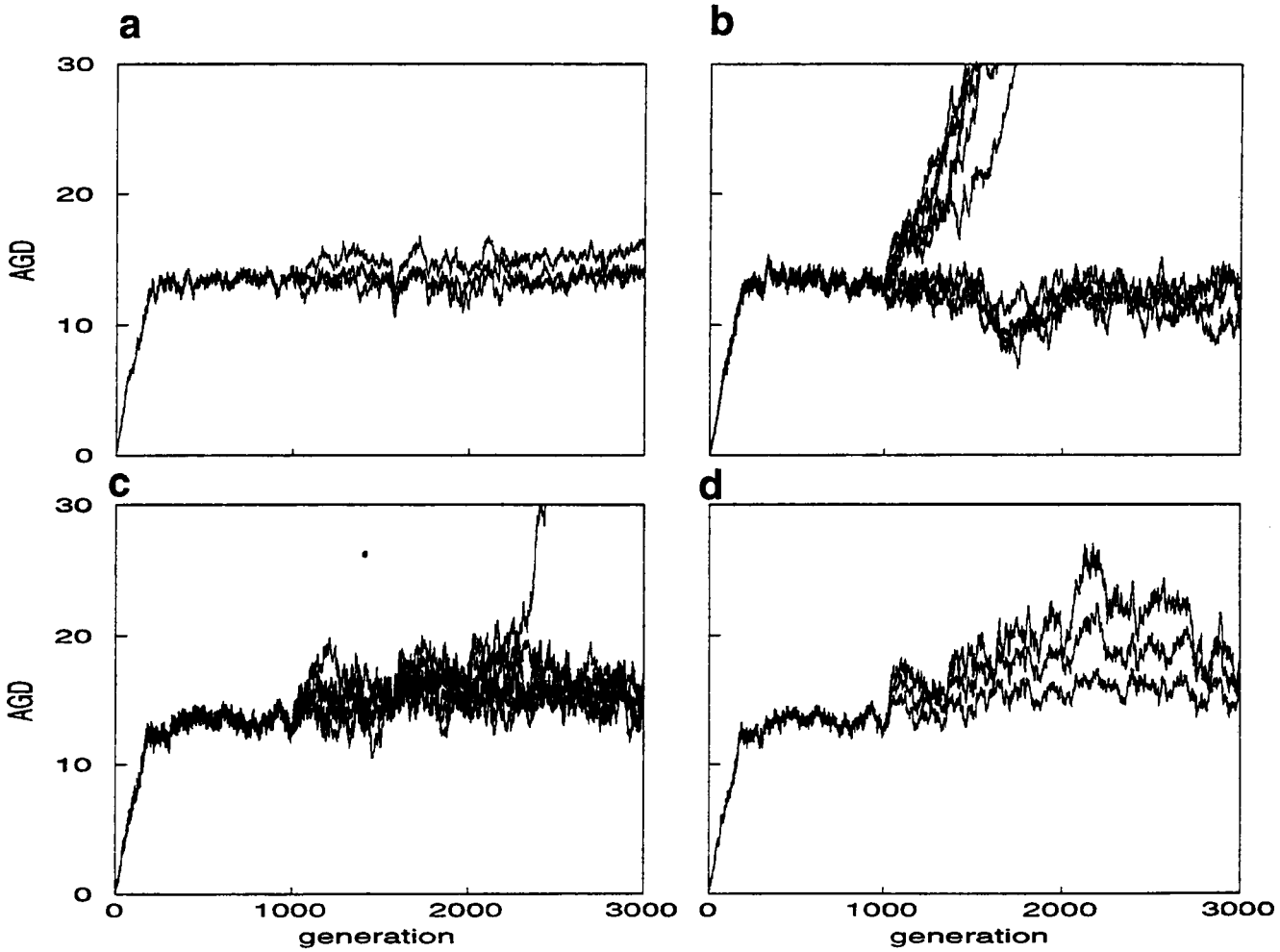


Figure 4.3: Effects of Population Subdivision

of calculating genetic distances between individuals in the selection process. All genetic distances are calculated between the first individual and all individuals at the chosen location for every choice of the first individual. Ideally, any given genetic distance should only be calculated once. Future encounters of distance calculation can then use the stored value from a previous calculation.

One aspect of modeling the dynamics of evolution and speciation is to be able to dynamically increase or decrease the size of the population during the course of evolution.

Finally, one could explore the idea of creating a parallelized version of the model implemented here. There is the potential to divide the workload associated with each evolution among multiple processors. This is especially true if the simulation of a large number of subpopulation is desired.

# BIBLIOGRAPHY

# BIBLIOGRAPHY


[1] Blattner, F. R., *et al.*, [1997]. "The complete genome sequence of Escherichia coli K-12", Science, vol. 277, September 5

[2] Dobzhansky, T. [1955]. "Evolution, genetics, and man", John Wiley & Sons, Inc.

[3] Finjord, J. [1996]. "Sex and self-organization on rugged landscapes", Journal of Modern Physics, vol. 7, no. 5, 705-715.

[4] Gavrilets, S. [1997]. "Evolution and speciation on holey adaptive landscapes", *TREE*, vol. 12, no. 8, 307-312.

[5] Gavrilets, S., Li, H., Vose, M. D. [1997] "Rapid speciation on holey adaptive landscapes", Submitted.

[6] Gavrilets, S., Gravner, J. [1997]. "Percolation of the fitness hypercube and the evolution of reproductive isolation", Journal of Theoretical Biology, vol. 189, 51-69

[7] Goldberg, D. [1989]. "Genetic Algorithms in search, optimization and machine learning", Addision-Wesley Publishing Company, Inc.

[8] Law, A. M., Kelton, W. D. [1991]. "Simulation modeling and alysis", McGraw-Hill, Inc.

[9] Lemke, N., Campbell, I. A. [1996]. "Random walks in a closed space", *PHYSICA*, vol. A 230, 554-562.

[10] Nagylaki, T. [1992]. "Introduction to theoretical population genetics", Springer-Verlag.

[11] Oxford, [1996]. "A dictionary of biology", Oxford University Press.

[12] Ridley, M. [1993]. "Evolution", Blackwell Scientific Publications, Inc.

[13] Vose, M. D. [1991]. "A linear algorithm for generating random numbers with a given distribution," *IEEE Transactions on Software Engineering.*

[14] Wright, S. [1932]. "The roles of mutation, inbreeding, crossbreeding and selection in evolution", *Proceedings of the Sixth International Congress of Genetics*, no. 1, 356-366

# APPENDICES

# APPENDIX A

## Rapid Speciation on Holey Adaptive Landscapes

# Rapid speciation on holey adaptive landscapes

Sergey Gavrilets*, Hai Li, Michael D. Vose

**In spite of significant advances in both theoretical and experimental studies of evolution, understanding speciation remains a major challenge (1). In recent years a view of speciation that can be traced to the classical Dobzhansky model (2) has become a point of renewed interest (3-9). Using the metaphor of "adaptive landscapes" (10), populations diverge genetically along the "ridges" of highly-fit genotypes and become reproductively isolated species when they come to be on opposite sides of a "hole" in a "holey" adaptive landscape (8,9). In the alternative view, which was first put forward by Sewall Wright (10,11) and has dominated most discussions of speciation, speciation involves transitions from one isolated "peak" to another across a "valley" of maladaptation in a "rugged" adaptive landscape (12,13). We consider here whether the mechanism implied by Dobzhansky-type models might result in rapid speciation. We simulated finite populations with hundreds of arbitrary linked diallelic loci. Our results demonstrate that parapatric (or allopatric) speciation on the time scale of hundreds of generations is plausible and is enhanced by increasing mutation rate, decreasing migration rate and increasing population subdivision. Simultaneous emergence of more than two new species from a subdivided population is highly probable.**

S. Gavrilets, Departments of Ecology and Evolutionary Biology and Mathematics, University of Tennessee, Knoxville, TN 37996, USA.

H. Li, M.D. Vose, Department of Computer Sciences, University of Tennessee, Knoxville, TN 37996 USA.

*To whom correspondence should be addressed. Email: gavrila@ecology.tiem.utk.edu).

Current renewal of interest in Dobzhansky-type models is coming from different sources. First, there is growing experimental support for the genetic architecture implied. These are data of genetic studies of reproductive isolation (4, 6), data on the relationships between primary and secondary structure of RNA (14) (which represent the only large-scale information about adaptive landscapes currently available), as well as data on selection responses, hybrid zones, and fossil records (9). On the theoretical side, there is a growing realization of serious problems with Sewall Wright's metaphor of "rugged" adaptive landscapes (9,15) and shifting balance theory (16). Both have for a long time have been a focus of theoretical studies of evolution. On the other hand, it recently has been demonstrated that the existence of "ridges" of highly-fit genotypes, which is postulated in Dobzhansky-type models, is actually a general property of multidimensional adaptive landscapes (8,17).

Earlier theoretical studies of speciation as a consequence of genetic divergence along "ridges" in the adaptive landscape involved various approximations such as small number of loci (3,4,7), linkage equilibrium (3), infinitely large number of highly mutable unlinked loci (18),

or absence of within-population variability during divergence (6). No attempts have been made to justify these assumptions empirically or to check the generality of the results obtained. Here we report results of individual based simulations that did not involve any approximations typical in studying multilocus systems. Using a linear algorithm for generating random numbers with a given distribution (19) has allowed us to model hundreds of linked diallelic loci potentially influencing fitness. The number of considered loci, which is only one order of magnitude smaller than that of bacteria (20), is a significant distinctive feature of our simulations.

We model finite subdivided populations of sexual haploid individuals. Evolutionary factors included were mutation, recombination, migration, genetic drift and selection. To decrease the number of parameters we used some symmetry assumptions. Mutation rates were equal across loci. Recombination rates between adjacent loci on the same chromosome were equal and recombination events took place independently of each other. The population was subdivided into subpopulations of equal size. To reflect the idea that reproductive isolation arises simultaneously with genetic divergence we posit that an encounter of two individuals results in fecund mating only if the individuals are different in no more than $K$ loci. Otherwise the individuals do not mate (premating reproductive isolation) or their offspring is inviable or sterile (postmating reproductive isolation). The neutral case (no reproductive isolation) corresponds to $K$ equal to the number of loci. Throughout our simulations individuals migrated before mating. To characterize the levels of genetic variability within subpopulations and genetic divergence between subpopulations we used the average genetic distances (AGD) defined as the number of loci at which two individuals (from the same subpopulation or from different subpopulations) are different. This is the standard Hamming distance.

Our simulations started with all $N$ individuals identical. During the first 1000 generations there were no restrictions on migration between subpopulations and the whole population evolved as a single randomly mating unit. One thousand generations was sufficient for the population to reach a state of stochastic equilibrium between factors affecting its dynamics. This was apparent from the behavior of AGD (see figures below) as well as from the behavior of the number of segregating loci and the number of different genotypes (data not presented). Starting with generation 1000, restrictions on migration were introduced. We used both the island model (11) and the one-dimensional stepping-stone model (21). In the former case, with probability $m$ an individual migrated to another subpopulation (with all subpopulations equally accessible). In the later case, the subpopulations were arranged along a line with migration only between neighboring subpopulations. The simulations were continued until generation 3000. We have chosen this relatively short time span for two reasons. First, here we are interested in the plausibility of *rapid* speciation. Second, only during relatively short time intervals may one assume the constancy of abiotic and biotic environment implied in our model. The number of runs for each parameter configuration was eight.

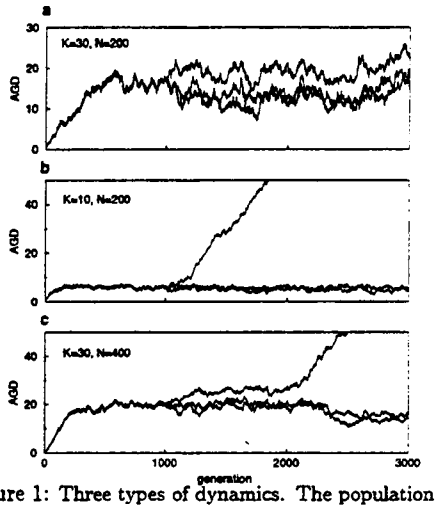


Figure 1: Three types of dynamics. The population is subdivided into two subpopulations. Of the three curves apparent after generation 1000, the upper curve gives the AGD between the subpopulations, whereas the two remaining curves give the AGD within each subpopulation. The population size $N$, and $K$ value, are given on the graphs. The rate of migration is $m = 0.01$, there are 384 loci per chromosome, the rate of recombination between adjacent loci is 0.005, and the mutation rate is $10^{-4}$.

Extensive simulations reveal the existence of three different dynamical regimes. In the first regime, the AGD both between and within subpopulations reach (stochastic) equilibrium values below $K$ (Fig.1a). Between-subpopulations matings are fecund. No further genetic divergence and speciation take place (at least until generation 3000). In the second regime, the AGD between subpopulations steadily increases after introducing restrictions on migration (Fig.1b). Subpopulations diverge genetically accumulating different mutations. After some time, the AGD between subpopulations significantly ex-

ceeds the AGD within subpopulations, subpopulations become reproductively isolated and evolve as independent units. According to any of the species concepts common in the literature (22) these two subpopulations can be considered as two different species. Here, speciation has occurred in the presence of gene flow and, thus, is parapatric. The third regime is a combination of the first two. After some transient time during which the AGD both within and between subpopulations have appeared to reach a new stochastic equilibrium, the latter starts suddenly increasing (Fig.1c). The system undergoes a stochastic transition resulting in speciation and further accumulation of genetic differences between subpopulations. After speciation has taken place, the genetic variability within (sub)populations decreases.

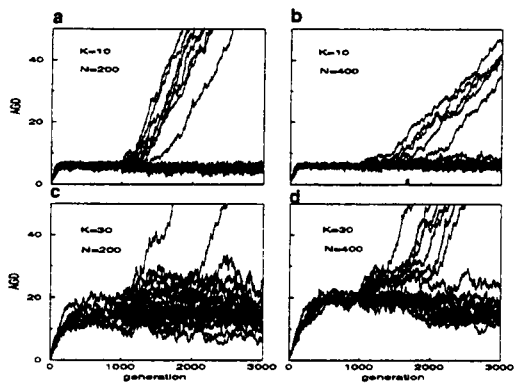


Figure 2: Effects of population size on evolutionary divergence and speciation. The population is subdivided into two subpopulations. AGD within and between subpopulations are shown (eight runs for each parameter configuration). a. With $K = 10$ and $N = 200$, subpopulations steadily diverged immediately after generation 1000 in all runs. b. However, with $N = 400$ the time until the initiation of divergence varied and divergence was not observed in 3 runs. Comparing the slopes of the curves representing AGD between subpopulations in a and b shows that increasing population size decreases the rate of divergence as well. c. With $K = 30$ and $N = 200$ speciation was observed in only two runs. d. With $N = 400$, speciation took place in 7 runs. Other parameters are the same as in Fig.1.

Dynamic behavior depends on the intensity of different factors operating in the system. As expected, reduced migration and increased mutation both increase the plausibility of speciation (data not presented). Population size plays a double role. With strong selection (small $K$), increased population size delays or even prevents speciation (see Fig.2a,b). With weak selection (large $K$), decreasing population size delays or prevents speciation (see Fig.2c,d). These effects can be explained in the following way. A necessary condition for genetic divergence and speciation is fixation of alternative alleles in different subpopulations. After the (sub)population becomes polymorphic at $K$ loci, new mutations are selected against because individuals carrying them have a reduced probability of fecund mating. Genetic drift operating in small populations overcomes the effect of selection and allows for genetic divergence. Genetic divergence of very large randomly mating subpopulations connected by migration will not occur. On the other hand, if (sub)populations are too small, they lack genetic variability necessary to initiate divergence. Genetic divergence of very small randomly mating subpopulations connected by migration will not occur either.

Increased population subdivision promotes speciation. For example, in a population with $N = 800$ subdivided into two subpopulations no speciation has been observed (Fig.3a). However, subdividing into four groups immediately initiates divergence and emergence of 4 species (Fig.3b). The migration rate in these simulations as well as in those presented above was small ($m = 0.01$). Speciation can also occur for higher migration rates with appropriate spatial subdivision. Fig.3c shows results for a single run of a stepping-stone model with 8 subpopulations and $m = 0.06$ (that is there are 6 migrants per subpopulation per generation). Here after some transient time the population splits into 2 independently evolving and reproductively isolated groups in proportions 4:4. Fig.3d illustrates the dynamics of the AGD between the subpopulations within one of these groups. As expected, genetic distance increases with the geographic distance (see Fig.3d). Note that most of the time after generation 1000 the AGD between the first and forth subpopulations is above $K$ meaning these two subpopulations are reproductively isolated. However, they evolve together and have a common evolutionary trajectory. In other seven runs the population was split into two pieces in proportions 4:4 and 6:2 once each and in proportion 3:5 twice. No split was observed in a single run. Twice the population was split into three pieces in proportion 2:3:3 resulting in the emergence of 3 species.

The dynamics of genetic divergence are not neutral. During the phases of stochastic equilibrium the AGD both within and between subpopulations are well below neutral expectations(23). After speciation has taken place, the rate of between-population divergence is smaller than the neutral expectation (which is twice

the expected number of mutations per generation).

Simulations described in Fig. 1-3 were performed with all 384 loci on a single chromosome. We also performed numerical simulations for the case of 6 chromosomes with 64 loci each. No apparent differences between these two cases have been observed.

Recently several cases of rapid speciation have been described (24- 26). The most spectacular one is probably the origin of hundreds of species of Lake Victoria cichlids in 12,000 years (26). The process of speciation via transitions between adaptive peaks across valleys of maladaptation is expected to be slow (13) and hardly can explain extremely rapid speciation observed. In contrast, the results presented here show that genetic divergence along the ridges in the adaptive landscape can lead to rapid speciation (allopatric or parapatric). Increasing the number of considered loci has significantly increased the plausibility of speciation relative to that in earlier models (3,4,7). The plausibility of speciation is enhanced by population subdivision; simultaneous emergence of several new species from a subdivided population is highly probable. The latter possibility is typically neglected in the contemporary methods for reconstructing phylogenies (27). The time scale in our simulations was short, meaning that restrictions on migration should not be long-lasting. Several hundreds generations might be sufficient for a significant divergence and evolution of reproductive isolation. Restoring migration after that to high levels will not return the system back to its initial state of free gene exchange between subpopulations. The model presented here assumes that the loci controlling reproductive isolation are different from those responsible for local adaptation. If this is not the case or if there is close linkage between the two types of loci, local adaptation might accelerate speciation (28).

## bf REFERENCES AND NOTES


1. J.A. Coyne, *Nature* **355**, 511 (1992).
2. T.H. Dobzhansky, *Genetics and the Origin of Species*. Columbia University Press (1937).
3. M. Nei, T.Maruyma, C-I. Wu, *Genetics* **103**, 557 (1983).
4. A. Wagner, G.P. Wagner, P. Similion, *Genetics* 138. 533 (1994).
5. H.A. Orr, *Genetics* **139**, 1803 (1995).
6. H.A. Orr, L.H. Orr, *Evolution* **50**, 1742 (1996).
7. S. Gavrilets, A. Hastings, *Am. Nat.* **147**, 466 (1996).
8. S. Gavrilets, J. Gravner, *J. Theor. Biol.* **184**, 51 (1997).
9. S. Gavrilets, *Trends Ecol. Evol.* **12**, 307 (1997).
10. S. Wright, *Proc. Sixth Int. Congr. Genet.* **1**, 356 (1932).
11. S. Wright, *Genetics* **16**, 97 (1931).
12. S.A. Kauffman, S. Levin, *J. Theor. Biol.* **128**, 11 (1987)
13. N.H. Barton, S. Rouhani, *Genetical Research* **61**, 57 (1993).
14. P. Schuster *et al.*, *Proc. R. Soc. Lond. Ser. B* **255**, 279 (1994); M.A. Huynen, P.F. Stadler, W. Fontana, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 397 (1996).
15. M.C. Whitlock *et al.*, *Annu. Rev. Ecol. Syst.* **26**, 601 (1995).
16. S. Gavrilets, *Evolution* **50**, 1034 (1996); J.A. Coyne, N.H. Barton, M. Turelli, *Evolution* **51**, 643 (1997).
17. C.M. Reidys, P.F. Stadler, P. Shuster, *Bull. Math. Biol.* **59**, 339 (1997).
18. P.G. Higgs, B. Derrida, *J. Mol. Evol.* **35**, 454 (1992); F. Manzo, L. Peliti, *J. Phys. A: Math. Gen.* **27**, 7079 (1994).
19. M.D. Vose, *IEEE Transactions on Software Engineering* **17**, 972 (1991).
20. F.R. Blattner *et al. Science* **277**, 1453 (1997).
21. M. Kimura,G.H. Weiss, *Genetics* **49**, 561 (1964).
22. J. Mallet, *Trends Ecol. Evol.* **10**, 294 (1995).
23. G.A. Watterson, *Theor. Popul. Biol.* **7**, 256-276; W.-H. Li, *Theor. Popul. Biol.* **10**, 303-308 (1976).
24. D. Schluter, J.D. McPhail, *Amer. Natur.* **140**, 85 (1992).
25. A.R. McCune, *Paleobiology* **22**, 34 (1996).
26. T.C. Johnson *et al.*, *Science* **273**, 1091 (1996).
27. G.A. Hoelzer, D.J. Melnick, *Trends in Ecology and Evolution* **9**, 104 (1994).
28. D. Schluter, *Amer. Natur.* **148**, S40 (1996).

29. We are grateful to Mark Camara, John Gittleman and Massimo Pigliucci for helpful comments. S.G. was partially supported by a Professional Development Award (University of Tennessee, Knoxville) and by grants from Université P. et M. Curie and École Normale Supérieur, Paris.

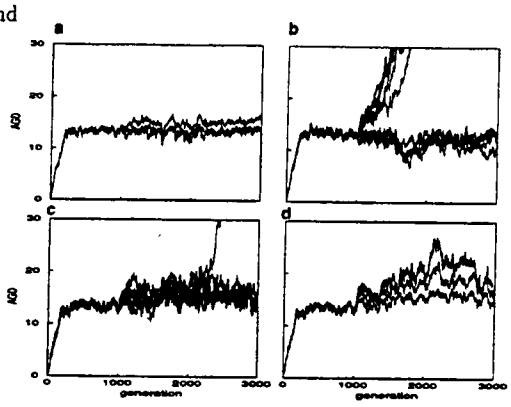


Figure 3: Effects of spatial subdivision ($N = 800, K = 20$; number of loci and the rates of recombination and mutation are the same as in Fig.1). Figures a and b show AGD within and between subpopulations for a single run. Results for the other seven runs are very similar. a, Population is subdivided into 2 equal subpopulations with migration rate $m = 0.01$. No speciation is observed (cf. Fig.1a). b, Population is subdivided into 4 equal subpopulations; island model with $m = 0.01$. All AGD between subpopulations steadily increase after generation 1000 indicating emergence of 4 species (cf. Fig.1b). Figures c and d correspond to a stepping-stone model with 8 local subpopulations with migration rate $m/2 = 0.03$ between neighboring subpopulations. Only results for a single run resulting in the emergence of two species are shown. c, AGD between neighboring subpopulations. All seven distances remain below $K$ (no reproductive isolation between neighboring subpopulations) except between subpopulations 4 and 5 which starts increasing around generation 2300 (cf. Fig.1c) indicating the split into two species with 4 subpopulations each. d, AGD between 4 subpopulations of the first species. Of the three curves apparent after generation 1000, the lower curve gives the average of the AGD between neighboring subpopulations, the middle curve gives the average of the AGD between subpopulations separated by one other subpopulations, and the upper curve gives the AGD between subpopulations 1 and 4 (separated by two other subpopulations).

# VITA

Hai Li is an international student from the People's Republic of China. He was born in Lanzhou, China on January 1, 1972 and first came to the on November of 1986. He enrolled the University of Tennessee in Spring of 1988. In Fall of 1991, he received a Bachelor of Science degree with major in Computer Science and a minor in Mathematics. He joined the graduate program in the University of Tennessee in the Spring of 1992 as a graduate teaching assistant. He joined the Network Service the subsequent year as a graduate assistant and went on to take a full-time position at the Telecommunication and Network Services at the University of Tennessee. He worked with Dr. Michael Vose and Dr. Sergey Gavrilets on modeling speciation on holey adaptive landscapes and graduated with a Master of Science degree in May, 1998.