Thesis
Speciation in Digital Organisms

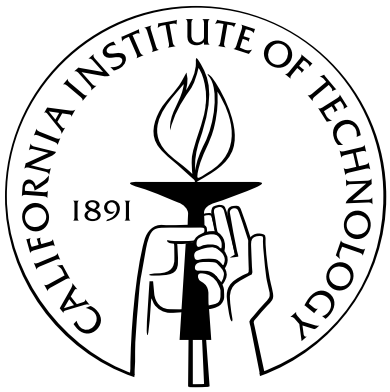
Thesis by

Stephanie S. Chow

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy



California Institute of Technology

Pasadena, California

2005

(Defended May 18, 2005)

Acknowledgements

I would like to thank my advisor, Dr. Christoph Adami, for his advice and guidance on all aspects of my graduate career in his lab. He helped make this thesis a reality. I would also like to thank Dr. Claus Wilke for his close collaboration and patience with my many questions, which taught me a lot on how to properly conduct research. My thesis committee asked me some very interesting questions: Dr. Michael Cross, Dr. Steven Quartz and Dr. Barbara Wold. My fellow graduate students in the lab, past and present, contributed many useful suggestions: Alan Hampton, Allan Drummond, Evan Dorn, Jesse Bloom and Robert Forster.

In less direct ways, my friends made this thesis possible. Without their friendship, support, and advice on many aspects of my work and life, the stresses of graduate school would have been impossible to deal with, and life would be a lot less fun.

Finally, I would like to thank my parents and my sister. Their love, support, and high expectations, not to mention the examples that they set in their own careers made a Ph.D. seem like an achievable goal.

Abstract

Current estimates of the number of species on Earth range from four to forty million total species. Why are there so many species? The answer must include both ecology and evolution. Ecology looks at the interactions between coexisting species, while evolution tracks them through time. Both are required to understand aspects of environments which promote speciation, and which promote species persistence in time.

The explanation for this biodiversity is still not well understood. I argue that resource limitations are a major factor in the evolutionary origin of complex ecosystems with interacting and persistent species. Through experiments with digital organisms in environment with multiple limited resources, I show that thes conditions alone can be sufficient to induce differentiation in a population. Moreover, the observed pattern of species number distributions match patterns observed in nature. I develop a simple metric for phenotypic distance for digital organisms, which permits quantitative analysis of similarities within, and differences between species. This enables a clear species concept for digital organisms that may also be applied to biological organisms, thus helping to clarify the biological species concept. Finally, I will use this measurement methodology to predict species and ecosystem stability.

Contents

List of Figures

Chapter 1

Introduction

Life is found everywhere on Earth, and its variety is impressive. Microbes have been found in Lake Vostok, which lies beneath the Antarctic ice sheet and is subject to high pressure (350 atmospheres), low temperatures (-3^o) and permanent darkness [88]. At another extreme, evidence of microbes has also been found in oceanic volcanic glass [34]. The total number of species may lie somewhere between 3 and 30 million [62], while some estimate that there may be more than a billion species of bacteria alone [27].

What is the source of this range and variety? Charles Darwin addressed these issues in his famous 1859 book “On the Origin of Species by Means of Natural Selection” [13]. Nearly one hundred and fifty years later, an understanding of the origin of species, or speciation, remains incomplete. Although there is an outline of a theory of speciation, the details of the causes and processes by which diversity and complexity arise and are maintained remains to be uncovered.

1.1 Speciation

The most important cause of speciation is geography. If a population of organisms from a single species experiences restricted movement due to geographic barriers such as dry land (for aquatic species), or mountains, the free flow of genetic materials within the species is impeded. For the subpopulations divided by the barrier, isolation leads to independent processes of natural selection and drift. Under different ecological conditions, the subpopulations can diverge significantly [84]. After sufficient time, they may become different enough to be called two species. Sexually reproducing organisms may diverge enough to lose the ability to recombine genetic material, after which the two subpopulations will have split permanently [63]. This effect is called allopatric speciation.

Australia's unique flora and fauna are good examples of allopatry. Approximately 45 million years ago, Australia became a separate continent, isolating its species and allowing many of them to evolve independently of those on other continents. The kangaroo is a grazer, eating grasses and the shoots and leaves of plants. Other animals with similar dietary habits, but native to other continents, are the deer and its relatives such as gazelles and antelope. In some ways, kangaroos and deer occupy the same *niche* on different continents, an ecological term which can be defined as "a way of making a living".

Less easy to explain is speciation in the absence of a geographic barrier, or *sympatric speciation*. The finches of the Galapagos Islands, located 600 miles west of Ecuador, have had a few million years to evolve. Thirteen species of finches that Darwin observed there are found nowhere else. This increase in diversity of a population that descended from one

Figure 1.1: The kangaroo, native only to Australia, eats grasses and the shoots and leaves of plants [106].

or a few founders is called adaptive radiation. As the birds evolve, they fill different niches. Some species eat seeds, others eat insects, and one drills for insects like a woodpecker. Some live on the ground, some live in trees. Allopatric speciation has occurred, since interbreeding is limited by the distance between the islands, but distinct species are also found on the same island.

Much recent work in speciation has focused on models of sympatric speciation in sexual organisms. Proposed explanations include assortative mating [86, 96], often due to ecological interactions and resource constraints [18, 20, 21], sexual conflict [38], disruptive

Figure 1.2: The deer, and its relatives such as gazelles and antelope, are native to most continents, but not Australia. It eats grasses and the shoots and leaves of plants [107].

selection [50], and offspring dispersal distance [49]. Ecological constraints have also been cited as a cause of speciation in biological organisms [42, 84].

As can be seen from the diversity of finches in the Galapagos, ecological constraints are a major source of evolutionary pressure. As the finches of the Galapagos increased in numbers, they experienced increased competition for food. The population diversified, adapting to different resources, as the challenge of finding nourishment provided an incentive to speciate in sympatry. Nonetheless, studies of evolution and ecology tend to be separate, thus leaving the emergence of populations of differentiated, interacting organisms infrequently examined [60].

This thesis will address the effect of environment on the emergence and maintenance of diversity in asexual populations, as well as the limitations imposed on diversity by ecological constraints. Sex and other forms of genetic recombination complicate the study of evolution. In bacteria, for instance, the extent of recombination in bacteria varies widely [33], and its effect is not necessarily beneficial [31, 51].

1.2 Single niches and competitive exclusion

The simplest type of environment is a homogeneous one with a single source of food. In this environment, the *genotype*, or specific genetic makeup, that produces the best *phenotype*, or set of physical traits, to consume the food and to power its reproduction is expected to dominate the population, and to drive all other genotypes to extinction. When the frequency of a gene has reached 100% of the population, this is called *fixation*. Only a single genotype will be the best competitor for the single source of food, and its genes will tend to go to fixation. The given set of conditions of the environment provides only one niche, which in turn supports only one species. More niches support more species. There is "one niche, one species", a common way of phrasing the fundamental principle of ecology known as the *competitive exclusion principle*.

The competitive exclusion principle states that in competition between species which seek the same ecological niche, one species will outcompete the others. With a single limiting resource, there may be more than one species co-existing, but one will be on its way in, and the other will be on its way out.

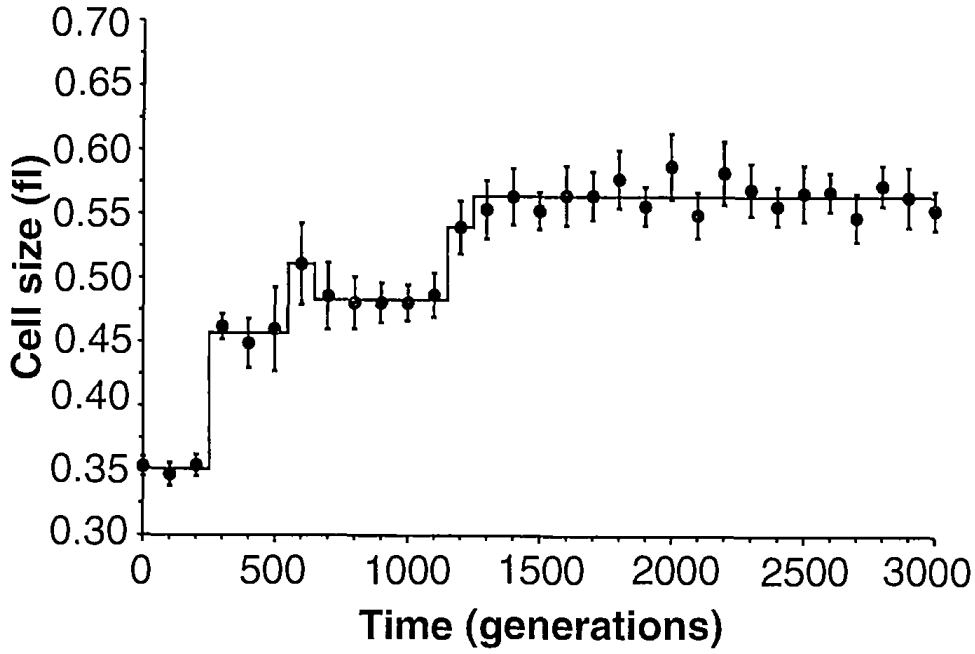


Figure 1.3: Changes in average cell size in a population of *E. coli* during 3000 generations of experimental evolution. The solid line shows a best fit of a step function model. From [30].

In a single niche environment, evolution occurs, but with a distinctive pattern. As mutations occur, they may be either beneficial or deleterious. Although difficult to quantify, studies of mutation rates in *Caenorhabditis elegans* [14] and *Drosophila melanogaster* [87] indicate that mutations reduce fitness on average, so natural selection will tend to limit the number of mutants. Rarely, there will be a beneficial mutation. Barring bad luck, this fitter mutant will outcompete and out-reproduce its comrades; that is, the mutation will go to fixation. Phenotypic changes often show a stepwise pattern, as seen in figure 1.3. These periods of stasis alternating with brief periods of rapid change are an example of punctuated equilibria [41]. The rate of beneficial mutations must be low to see this pattern [30]. At higher mutation rates, in addition to punctuation it is more likely to see an effect called

clonal interference, as beneficial mutations compete with each other [39, 73].

1.3 Multiple niches and adaptive radiation

Recall that in allopatric speciation, geographic separation can induce differentiation. Another way that geography can promote differentiation in a population is through *parapatric* speciation. An environmental gradient or change can result in adaptation to local conditions in the absence of a explicit barrier to gene flow. In both cases, heterogeneity provides niches and opportunities for mutants.

Adaptive radiation is the development of a variety of phenotypes from a single ancestral form. As the founders reproduce and mutate, the resulting population may rapidly fill many ecological niches [85]. Figure 1.4 shows three morphologically distinct types, or morphotypes, arising from adaptive radiation in the bacterium *Pseudomonas fluorescens* propagated in a heterogeneous environment. Cultures of *Pseudomonas fluorescens* can rapidly diversify their genotypes and phenotypes as they specialize into and adapt to different environments in the unstirred vessel containing the growth medium: the surface, liquid phase and the bottom wall [78].

Another determinant of diversity or species richness in an ecosystem is productivity. One definition of productivity is the rate of production of biomass in an ecosystem; the mostly widely used is probably “the rate at which energy flows through an ecosystem” [81]. Species richness tends to increase with increasing ecosystem productivity, but is sometimes observed to decline at high productivity levels [36, 46, 47, 66, 80, 92, 99].

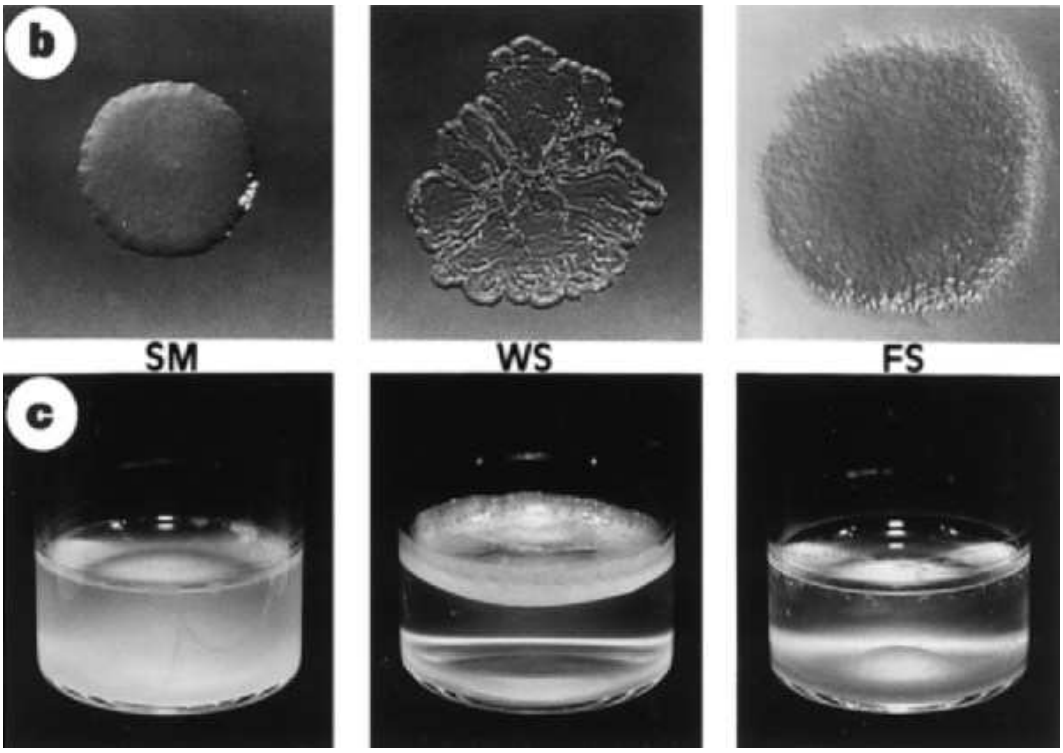


Figure 1.4: Three basic types of *Pseudomonas fluorescens* morphs arising as a result of adaptive radiation in an unshaken medium. The 'smooth morph' (SM) occupies the broth phase, 'wrinkly spreader' (WS) occupies the air-broth interface at the surface, and 'fuzzy spreader' (FS) occupies the bottom of the culture vessel. From [78].

Two types of proposed diversity-productivity relationships are monotonically increasing curves and unimodal ("hump-shaped") curves, where diversity eventually decreases. Data and theory mostly support unimodal curves [2,81,92] in heterogeneous environments, although there is also support for monotonically increasing curves as well [1].

When cultures of *Pseudomonas fluorescens* are grown in heterogeneous environments in a range of productivity levels, they show a clear unimodal diversity-productivity curve [47], but show much less diversification in a homogeneous medium [47] [78], as seen in figure 1.5.

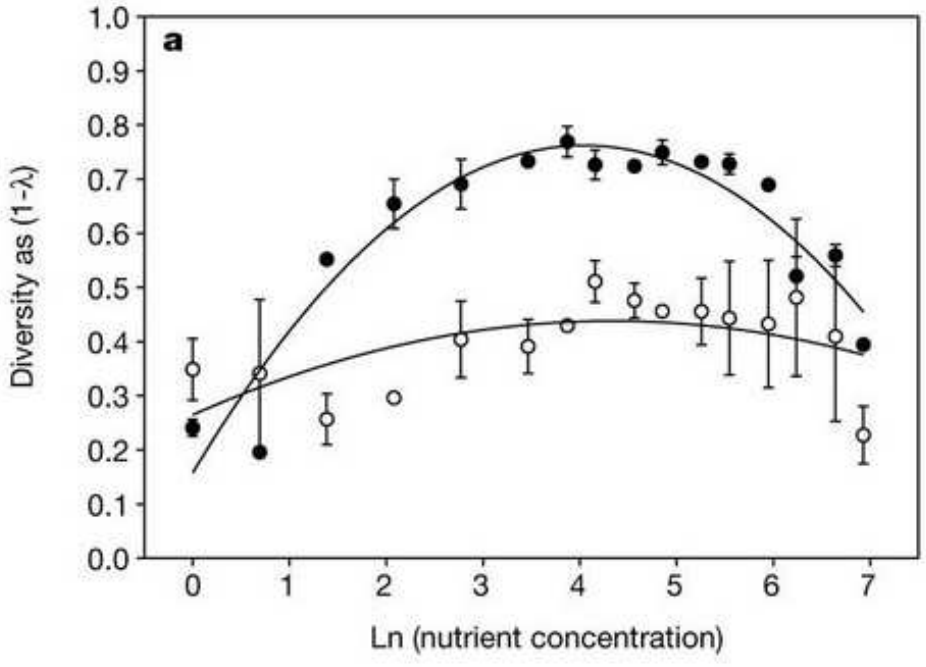


Figure 1.5: Diversity as a function of nutrient concentration in one replicate with *Pseudomonas fluorescens*. Black dots represent heterogeneous (undisturbed) environments, while white dots represent homogeneous (shaken) environments. P-values for the quadratic effect are < 0.0005 in the heterogeneous case, and 0.06 in the homogeneous case. Diversity is given as Simpson's Index of Diversity $1 - \lambda$. From [47].

Heterogeneity in the environment provides an opportunity for the emergence of stable polymorphisms, and perhaps eventual speciation through allopatric or parapatric means. Another way populations increase diversity is through resource partitioning, where subpopulations minimize competition by focusing on different resources [61, 77], although spatial structure can be a factor in the extent of differentiation. *Escherichia coli* grown in a homogeneous medium with a single limiting sugar may undergo adaptive radiation via cross-feeding, in which metabolites produced by certain phenotypes are consumed by other phenotypes [43, 82, 83, 95, 97], resulting in stable polymorphisms.

1.4 Motivation

That environmental heterogeneity or multiple resources may induce differentiation and adaptive radiation within a population has been established in experiments with *Pseudomonas fluorescens* and *Escherichia coli*. The question remains as to the necessity of the first, and the limits of the second. The principles underlying the emergence and maintenance of a diverse population of organisms, as well as factors limiting its diversity, in a resource-limited homogeneous environment are not yet well-understood.

The goal of my thesis research is to investigate speciation in well-stirred environments with multiple limited resources through experiments with digital organisms. First, I will address the issue of speciation, establishing a definition of species in digital organisms, and examining the conditions for and dynamics of adaptive radiation and sympatric speciation [6, 103]. I will also look at the dynamics of adaptation of individual species in the population [103]. Second, I will show that species richness peaks at intermediate productivity [6], a pattern which matches some experimental observations such as the ones in figure 1.5. Furthermore, this pattern is observed without relying on environmental heterogeneity as in many models and experiments. Factors which support long-term coexistence and stability in evolved ecosystems are also elucidated and tested. Third, I will show that the coevolved species show distinctive phenotypic characteristics relative to each other. I will continue on to develop a measure of phenotypic difference, and use it to show that this measure verifies a fundamental ecological principle with respect to the definition of a species. Finally, I will use the measure to generate and test predictions of the outcome of

invasion, when a novel species is introduced into an established ecosystem.

Chapter 2

Avida

While the ecology and evolution literature may cover the full range of life on earth, experimentalists often limit themselves to a much smaller selection. Bacteria are a common experimental organism in evolution, as are viruses and yeast [32], since they satisfy to various degrees the following list of desirable properties for organisms that can be used for experimental evolution research:

- Organisms are abundant and easy to breed.
- Organisms have a short generation time.
- The experimental environment that organisms live in is simple and controllable.
- It is easy to make measurements of genotypic and phenotypic characteristics.
- It is easy to store historical lineages.

Other organisms less commonly used for experimental evolution include multicellular organisms such as *Caenorhabditis elegans* and *Drosophila melanogaster*.

Although bacteria are easily propagated, have fairly short generation times, and are easy to grow, evolution experiments are still very time-consuming. One generation of *E. coli* takes approximately 20 minutes under optimal conditions. In practice, experimental conditions rarely allow reproduction to proceed at the maximum possible rate. Experiments with *E. coli* may run for 10000 generations [58, 76], and long-term studies have gone for 20000 generations or more [83]. Moreover, the observation of cross-feeding [43, 82, 83, 95, 97], where some bacteria consume the waste products of others, shows that aspects of environment can be difficult to control precisely. Finally, sufficient measurements of fitness, genomics and other data for high statistical accuracy may be impractical. One solution to these problems is to use digital organisms, which share many properties of biological organisms [102], but exceed them in others.

Digital organisms are self-replicating computer programs that have the three necessary and sufficient ingredients for Darwinian evolution. These are mutation (variation), replication (inheritance) and differential fitness (selection). They live in, and adapt to, an environment controlled by the experimenter. The complexity of the organism conferred by their computational genomic basis leads to complexity in their behaviours, complexity in their interactions, and evolutionary population dynamics among others. Recent studies with digital organisms have addressed issues of genome complexity and genetic interactions [56], robustness to mutations [104] and the evolution of complex features [57].

Many other prior computational approaches to evolution have been simulations rather than experiments [4, 15, 16, 49, 64].

2.1 Overview of Avida

The software platform for my experiments in computational evolutionary biology is called Avida. It is free and open-source (downloadable at www.sourceforge.net). All experiments were performed with version 1.99 of the Avida platform, compiled for linux with `gcc-2.95`.

Avida is a computer software and genetic system in which a population of asexual digital organisms evolves by natural selection. An organism is defined by its genome, a self-reproducing string whose elements are commands from a Turing-complete instruction set. Each genome executes on a virtual central processing unit (CPU). Mutations in an organism's genome correspond to changes in its instruction string. These can be deleterious, resulting at worst in an inability to reproduce, but may also be neutral or beneficial, for instance increasing replication efficiency or adding functionality. Simple Darwinian selection will then act upon these new organisms as they compete with the rest of the population. By correctly performing simple binary operations on up to three arbitrary 32 bit numbers, an organism increases its "merit", which increases the speed of its virtual CPU relative to other organisms. This, in turn, speeds up its reproduction.

The software has three main modules. The first is the *Avida core*, which maintains a population of digital organisms (genome or code, and virtual hardware), an environment with reaction rules (artificial chemistry) and resources, a scheduler to allocate CPU cycles to the organisms, and data collection objects. The second is the *graphical user interface* (GUI) with which the researcher can interact with the software. The last is a collection of

analysis and statistics tools to collect data and perform analyses of organism properties, lines of descent and many other features.

2.2 Organisms in Avida

Each organism in Avida is a self-contained automaton that can construct new automata. The organism, following the instructions in its genome, uses its virtual hardware to interact with the Avida environment as well as to attempt to make a perfect copy of its genome. The new genome is then passed to the Avida world, which gives the (possibly imperfect) copy its own virtual hardware, and places it somewhere in the population. If the environment has a grid structure, then the new offspring may be put in a location adjacent to the parent. If the environment is modelled on a chemostat and therefore has no spatial structure, the offspring may be be put in any random location. In either case, the Avida world has a fixed population size, so placing an offspring in a location means first killing and removing the previous occupant.

2.3 Virtual hardware

The basic structure of an Avida machine or organism is shown in figure 2.1. At the core is the CPU, which executes the instructions in the genome, and modifies the states of its components accordingly. There are three registers to store and manipulate data in the form of a 32-bit number: `AX`, `BX` and `CX`. There are also two stacks to store data.

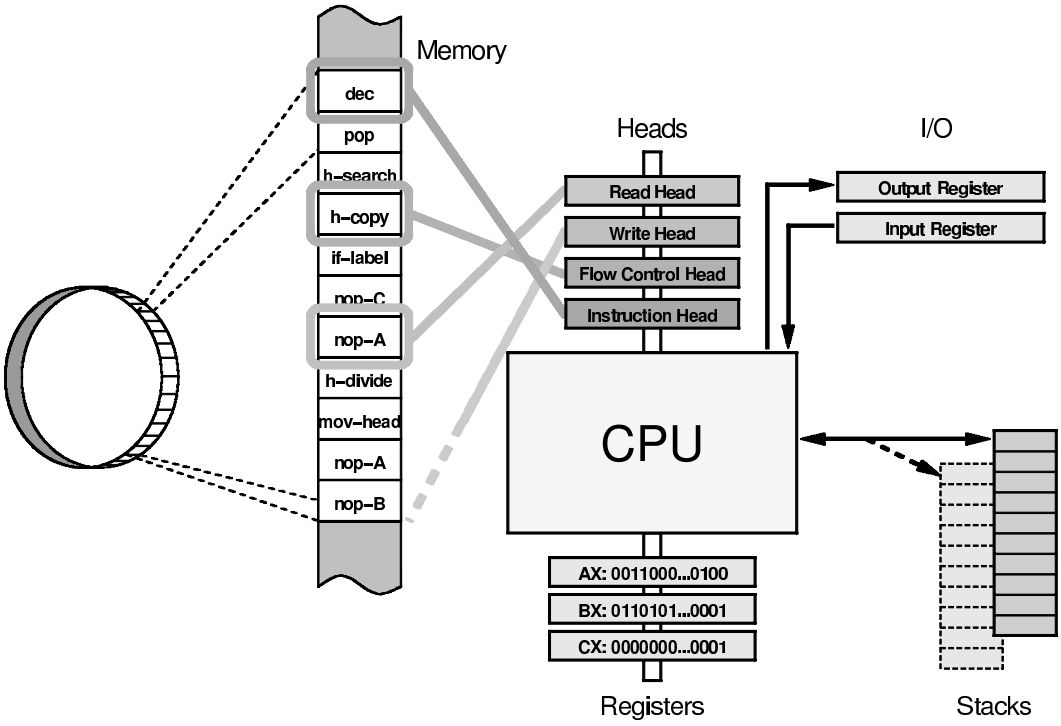


Figure 2.1: The basic hardware architecture of an Avida organism: CPU, registers, stacks, IO buffers. From [72].

At the initialization of an organism, the genome is loaded into memory. Execution starts with the first instruction and proceeds sequentially, unless an instruction (such as a jump) explicitly commands otherwise. When the last instruction is reached, the program loops back to the beginning. In effect, the genome is circular, just as in most types of bacteria.

The CPU structure supports four *heads*, which are essentially pointers to locations in memory: the instruction head, which points to the location of the instruction being executed; the read head and write head, which together allow the organism to read an instruction at one location and write it to another; and the flow-control head used for jumps and loops.

Finally, there is an input buffer and an output buffer, which allow the organism to interact with the environment. The machine can read in one or more inputs, perform computations, and write the results to the output buffer.

2.4 Genomes and Instruction Sets

As in biological organisms, it is the order and combination of the basic units of genetic information that define the genome. The elements of the assembly-like language are part of the full range of instructions that can be executed by the Avida CPU. Certain subsets of the instructions, called “instruction sets”, form logical units, each of which can be considered a programming language. A genome is a sequence of instructions from the instruction set, in the way DNA is a sequence of bases from the set of four nucleotides.

The instruction set used for the experiments in this thesis is the default instruction set, consisting of the following 26 instructions:

```
nop-A
nop-B
nop-C
if-n-equ
if-less
pop
push
swap-stk
swap
shift-r
shift-l
inc
dec
add
sub
```

```
nand
IO
h-alloc
h-divide
h-copy
h-search
mov-head
jmp-head
get-head
if-label
set-flow
```

The first three instructions are the `nop` ("no-operation") instructions, which do nothing on their own, as their name suggests. They are used for template matching, like labels. For example, positions in the genome can be tagged, allowing jumps and function calls to go to the right place even if the genome is subject to change, especially insertion and deletion mutations that shift instruction positions. Nops have circular complementarity: the complement of `nop-A` is `nop-B`, the complement of `nop-B` is `nop-C`, and the complement of `nop-C` is `nop-A`. For example, `h-search nop-A nop-B mov-head ... nop-B nop-C` puts the flow-control head after the complement to `nop-A nopB`, that is, after `nop-B nop-C`. Then `mov-head` moves the instruction head to the position of the flow-control head, and execution of the genome continues from the new position. Nops can also be used as arguments to the preceeding instruction. For example, `inc nop-A` increments the `AX` register.

To replicate, `h-alloc` allocates memory at the end of the genome for a copy, `h-copy` works with other structures in the hardware to create the copy, and `h-divide` splits the doubled genome in two, passing the copy to Avida which uses it to create a new organism in the world.

The sequence of instructions below is a simple, handwritten (as opposed to evolved)

example of a genome, which can copy itself and nothing more.

```
h-alloc    # Allocate space for child
h-search   # Locate the end of the organism
nop-C      #
nop-A      #
mov-head   # Place write-head at beginning of offspring.
nop-C      # Insert as required to achieve desired genome length
h-search   # Mark the beginning of the copy loop
h-copy     # Do the copy
if-label   # If we're done copying...
nop-C      #
nop-A      #
h-divide   # ...divide!
mov-head   # Otherwise, go back to the beginning of the copy loop.
nop-A      # End label.
nop-B      #
```

In summary, a self-replicating genome typically allocates space for the offspring, puts the write head at the beginning of the allocated space, iteratively copies each instruction to the child segment of the expanded genome, and then divides. The seed organism used in many of the experiments in this thesis has an almost identical genome, but with many more `nop-C`'s as placeholders to extend the genome length for reasons detailed later.

2.5 Mutations

The variability in the population that is required for selection is provided by mutations in the genome. The main type is the copy mutation. With a probability set by the experimenter, the instruction `h-copy` does not copy the instruction pointed to by the read head, but instead writes a random instruction to the location pointed to by the write head. Other possible types of mutations, again with probabilities set by the experimenter, are inser-

tion and deletion mutations, which may insert or delete a random instruction in the child organism; and cosmic-ray or point mutations, in which random changes in the genome occur during its execution at a set rate. A final type is the implicit mutation, caused not by explicitly set mutation parameters, but by an incorrect copy algorithm. Offspring with these mutations can be automatically discarded by setting the `FAIL_IMPLICIT` option in the genesis configuration file.

2.6 The Avida world

The Avida world has a fixed number N of cells or positions, each of which can hold at most one organism. The maximum population size is therefore also N. There are two possible topologies: a 2D grid with Moore neighbourhoods (each cell has eight neighbours), and a fully connected or *well-stirred* topology, where every cell is a neighbour to every other cell, and there is in effect no spatial structure. When an organism reproduces, depending on the parameter set by the experimenter, the offspring is placed in a random cell in the neighbourhood, or in the oldest cell (with a preference for empty cells) in the neighbourhood. The experiments in this thesis use the well-stirred structure with random replacement.

2.7 Resources and the rate of reproduction

All organisms have a virtual CPU, and each CPU can run at a different speed. In the simplest case, often used to start an experiment, all CPUs run at the same speed (instructions

executed per unit time). To simulate CPUs running in parallel, Avida provides a scheduler which executes one instruction on one machine, then one on another, and so on. A unit of time in Avida is an *update*, a period in which the average organism has executed k instructions ($k = 30$ by default).

If speeds differ between organisms, then the scheduler allocates CPU cycles proportional to the organism's *merit*, a unitless value which has meaning only when compared with the merits of other organisms. This allocation can be either *perfectly integrated*, or *probabilistic*. The perfectly integrated is the default, and the one used for experiments in this thesis.

All organisms start with an initial *merit* depending on their sequence length, since the length affects the time required to reproduce. This gives them an initial CPU speed. An Avida environment can contain resources that an organism can consume to change their merit. To absorb a resource, the organism must carry out the corresponding computation, or *task* as explained below.

The environment consists of a set of resources and a corresponding set of reactions required to interact with them. A reaction is defined by a computation that must be performed, a resource that is consumed as a result, the effect on merit (which may be proportional to the amount available in the environment), and a possible by-product. A resource has an initial level, an inflow rate in units per update, and an outflow rate as a fraction of the amount in the environment per update. If inflows are finite, then the resources are depletable, so that performance of the task corresponding to a resource results in a reduction in the amount of

that resource. “Infinite” inflows are possible by making resources non-depletable. Reactions can have conditions to control when they will be successful and therefore when they can change merit. These conditions could be a limit on the number of times a reaction can be performed, or be a requirement that another reaction must have been triggered first.

For natural selection to occur in Avida, there must be variability in the rate of reproduction, and some limit on population size. In a given organism, this is determined by a combination of CPU speed and *gestation time*, the number of instructions it needs to execute to produce an offspring.

To summarize, there are three types of resources in Avida. First, there is the basal resource, which is available to all organisms and allows them to replicate even if they cannot compute any logic functions. Second, there are computational resources which can be associated with the logic functions. There are plans to remove the basal resource in future versions of Avida, so that organisms can survive only when they can utilize computational resources, but currently, the basal resource is always available to every organism, and cannot be depleted. Finally, there is space. The memory of the virtual computer in which the Avida organisms live can hold only a finite number of organisms, which are replaced at random when new organisms are born.

It is important to note that in Avida, rewards do not depend on how the computation is performed. Avida presents the organism with inputs, and looks for a an output corresponding to a rewarded computation on the input(s).

2.8 Experimental setup

For the experiments in this thesis, the environment has nine resources corresponding to nine simple one and two input logical functions: Not, Nand, And, OrNot, Or, AndNot, Nor, Xor, and Equals. Resource levels start at zero, flow in at a constant rate, and flow out at a rate of one percent. The following are other basic features of the experiments:

- Population size of 3000
- Fixed genome length of 100
- Experiment duration of 400000 updates
- Seeded with an unevolved ancestor, giving a clonal population
- Same inflow and outflow rates for all resources, which give the same reward per unit consumed
- Inflow rates varied over six orders of magnitude, from 0.1 units per update to 100000 units per update
- Copy (point) mutations only, with a per site probability of 0.005 ($100 \times 0.005 = 0.5$ mutations per generation)
- No recombination

A population size of 3000 individuals is enough to allow maximum diversification (measured as number of species) without wasting time and computing power. A fixed

genome length simplifies analysis, since genomes are easily compared to each other. The length is maintained by having only point mutations during the copy loop and disregarding offspring with other lengths (which can happen if the copy loop is faulty and splits the offspring off at the wrong time). A length of 100 and an experiment duration of 400000 updates is more than sufficient for organisms to acquire all nine rewarded tasks if conditions are favourable. The range of inflow rates is guided by the range in productivity required to generate the hump-shaped diversity pattern in *Pseudomonas fluorescens* [78] as seen earlier in figure 1.5.

In biological organisms, mutation and sometimes recombination provide the variability in populations upon which natural selection acts. A genomic mutation rate (number of mutations per generation) 0.5 allows evolution to occur quickly without adding too much noise to the system. By comparison, the approximate genomic mutation rate in RNA viruses is 1, and in DNA-based microbes, it is much lower [24]. Since biological genomes are significantly larger than the digital ones in these experiments, the genomic mutation rates correspond to much lower per site and per gene mutation rates. Although recombination or gene transfer is a factor in rates of bacterial evolution [37, 70], its extent in natural populations varies between bacterial species [33] and its overall impact is not clear [31]. For these reasons, as well as ease of tracing lines of descent, there is no recombination in these experiments.

Chapter 3

Species in Avida

The species is a unit of evolution and of classification, and its definition is important to the study of speciation. And yet, Mallet notes that “we still do not all accept a common definition of what a species is” [60]. The number of species is a widely used measure of biodiversity, and will be used to measure diversity in digital organisms as well.

3.1 Species concepts

According to Darwin, the “undiscovered and undiscoverable essence of the term species” meant that “we shall have to treat species in the same manner as those naturalists treat genera, who admit that genera are merely artificial combinations made for convenience” [13].

Since that time, there have been many more concrete species concepts – Coyne [10] lists some recent ones: biological [19,63], evolutionary [101], phylogenetic [12,17], recognition [59], cohesion [91], ecological [98] and internodal [52].

The biological species concept has endured for more than half a century and remains

probably the best known. Species are “groups of interbreeding populations in nature, unable to exchange genes with other such groups living in the same area” [63]. Although there are already many difficulties with the concept with respect to sexual reproduction [23, 60, 89], it is even harder to apply to asexual reproduction due to its reference to the concepts of recombination or gene transfer.

It is well-known that some bacteria which differ significantly both genotypically and phenotypically are still able to exchange genetic material [26]. Lateral (or horizontal) gene transfer is a common phenomenon in bacteria and is an important factor in their evolution [37, 54, 70] although its extent varies significantly [33] and its effect is not necessarily beneficial [31, 51]. Similarly, viral recombination is also common [35, 100]. Furthermore, the exchange of genetic material has been shown to extend to gene transfers between kingdoms – between prokaryotes, eukaryotes and archaea [105], clearly not members of the same species. Recombination is not a good basis for a species definition in asexual organisms.

3.2 The digital species concept

Although consensus is lacking on a definition of species even when restricted to bacteria, a solution is to omit rates of recombination or lateral gene transfer in the genome, or at least treat it differently. In bacteria, definitions of species may rely on phylogeny or descent in gene trees [28], which trace the lineage of a particular segment of DNA rather than the entire genome; another possibility is to apply the concept of an “ecotype” – a population of

cells in the same ecological niche [8]; a third is to combine taxonomy (classification based on physical characteristics) and phylogeny [68].

Organisms in nature have been observed to form tight clusters [11, 40] rather than a continuum, although successive species sometimes coexist for a time [41]. In a single niche, organisms form a single cluster, implied by the stepwise pattern seen earlier in figure 1.3. A species can be considered to be a collection of genotypes which share many common genotypic and phenotypic features with each other, but differ considerably from members of other species [60], a definition which has been used for bacterial species [7, 77].

I will base the concept of a digital species on phylogeny, an important tool in evolution [74]. Organisms which are closely related, i.e., have recent common ancestors, are more likely to be of the same species than more distantly related ones. Recall that there is no recombination in my experiments, so relatedness is easy to measure. Underlying the use of phylogenetic information is the hypothesis, supported by results reported in this thesis, that closely related organisms also share common genotypic and phenotypic features. By the phenotype of a digital organism, I am referring to the ecological niche, or pattern of resources consumed.

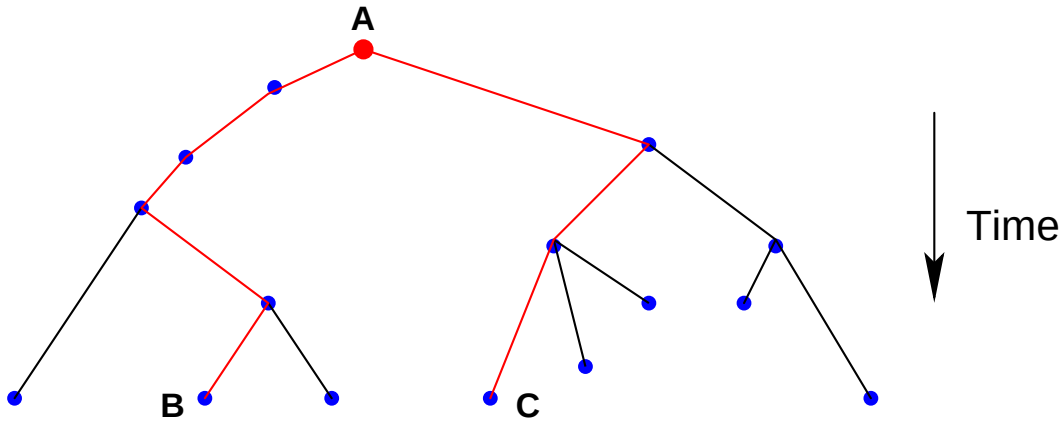


Figure 3.1: A simple example of a phylogenetic tree, showing how to measure the phylogenetic distance between any two organisms or genotypes who share a common ancestor. Each dot or node represents a genotype. Each edge connects a parent genotype with a child genotype which is at least one mutation away from its parent. In this example, *A* is the ancestor of all the genotypes in the tree. The distance between any two nodes can be found by counting the number of edges between them. The distance between *B* and its ancestor *A* is 5, and the distance between *B* and *C* is 8.

3.3 Species determination

3.3.1 Phylogenetic distance

To measure relatedness of two genotypes, I use phylogenetic distance. Avida can output a complete record of every genotype in the population as well as every genotype in the lines of descent all the way back to the seeding population. Each record includes the genome, useful data such as the current and past numbers of organisms with this genotype, as well as an identifier of the parent genotype. With the lineage information, it is simple to recreate a phylogenetic tree (or trees for non-clonal seed populations) which contains the entire ancestry of every genotype in the current update. Each edge of the tree joins a parent genotype with its mutated child genotype. The tree does not distinguish between back

mutations (a second mutation which changes the genome back to the original), multiple mutations, and the more common single point mutations. Hamming distance, the number of point to point differences between two strings of symbols, between an ancestral genotype and its descendants increases with time, so back mutations are rare. Intuitively, since a mutation picks a replacement with equal probability from the instruction set, and the default instruction set has 26 possible instructions, back mutations are unlikely. Multiple mutations are also very rare, given reasonable mutation rates.

The phylogenetic distance between any two genotypes in a tree is obtained by counting the number of edges in a path (no repeated nodes) connecting the two genotypes. A simple example is illustrated in figure 3.1. Start at either one of the genotypes, count up (i.e., back in time) the edges between child and parent genotypes in the tree to the most recent common ancestor (MRCA) of both. Then continue counting down (i.e., forward in time) the tree from the MRCA to the second genotype. If the MRCA is not known, start at one genotype and flag its entire line of descent back to a known common ancestor. Then start at the other genotype and trace its ancestry until you hit a flagged genotype. The flagged genotype is the MRCA. The phylogenetic depth of a genotype to an ancestral genotype is the number of mutants or edges in the tree between the two. The phylogenetic distance between two genotypes is therefore equivalent to the sum of the phylogenetic depths of the two organisms relative to their MRCA.

Other measures of distance between two genotypes include Hamming distance and time (rather than number of genotypes) to and from the MRCA. When applied to relatedness,

they were both found to be inadequate for different reasons. Details are in the appendix.

Intuitively, a smaller phylogenetic distance between two genotypes implies increased relatedness and a correspondingly increased likelihood of being members of the same species. Similarly, for any subpopulation or cluster of organisms, a score based on the pairwise distances between genotypes in the cluster, such as a sum, will indicate the relatedness within the cluster.

3.3.2 Clustering and calibration

The partitioning of the population into species is achieved by a clustering algorithm designed by Charles Ofria expressly for species determination in digital organisms in Avida. I coded the algorithm in C++ and it is available online [5].

In the clustering algorithm, all genotypes are initially in a single cluster, and this cluster is given a score based on phylogenetic distance. The score is compared to a cutoff, and cluster is repeatedly divided until the score falls below the cutoff. The final number of clusters or species is not known in advance. The detailed steps are as follows:

1. Put all genotypes into a single, initial cluster.

2. Find the genotype that minimizes the sum of the distances from it to all the other genotypes in the cluster; this is the centroid of the cluster.

3. Sum the distances; this is the score for the cluster.

4. Compare the cutoff to the total. If the cutoff is greater than the total, stop.

5. Otherwise, search through the remaining non-centroid genotypes and find a new centroid – the genotype which minimizes the sum of distances of all genotypes to the nearest centroid.

6. The score of the new cluster is the sum of the distances.

7. Compare the cutoff to the new score. If the cutoff is greater than the total, stop.

8. Repeat the previous three steps as needed.

Recall that a single niche environment usually contains only one species, although successive species sometimes coexist for a time [41]. In Avida, a single niche environment can be created by making resources undepletable. The sole limit to growth is population size, and the evolved population will usually have a single, dominant species.

If the single niche population is considered to be a single cluster, the resulting score will be low since all genotypes are closely related. I will use the distribution of single niche scores to determine a cutoff for the clustering algorithm.

The setup for the calibration experiments matches the basic procedure in section 2.8 except that resources are undepletable. There are 50 replicates, each with the same virtual environment seeded with a population of 3000 clones of a simple, unevolved organism. When I examined the resulting evolved populations, there were strong phenotypic and genotypic similarities, suggesting that organisms were closely related. At 100,000, 200,000, 300,000, and 400,000 updates, I calculated the score of a second cluster in the evolved replicate population. At 100,000 updates, the second cluster scores are lower than at later updates, likely

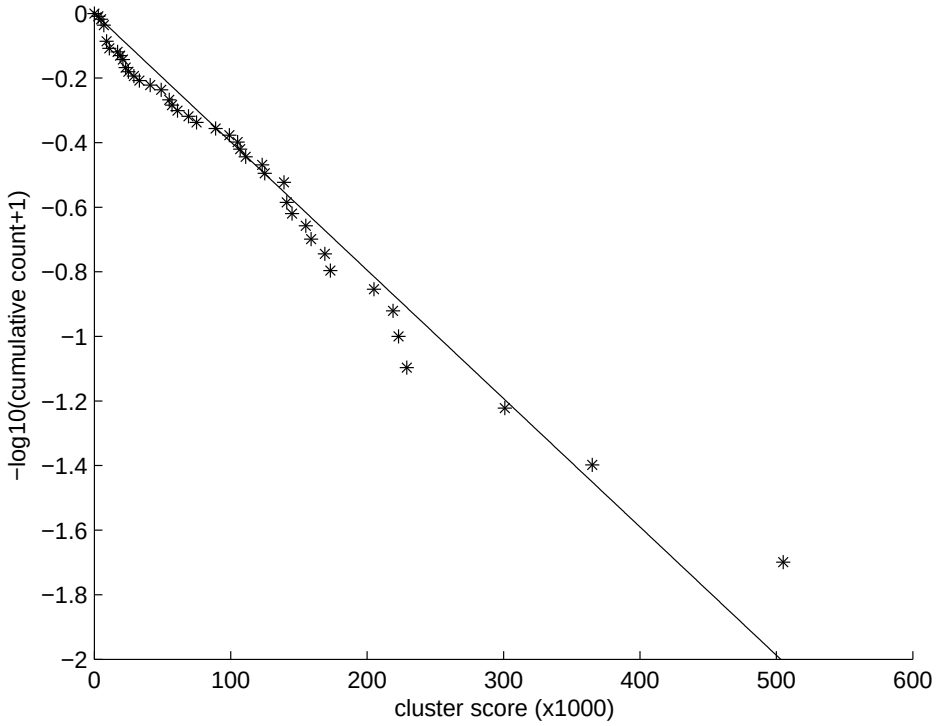


Figure 3.2: Adjusted log cumulative proportion of replicates versus the second cluster score in infinite (non-depletable) inflow runs. The values in the ordinate have been transformed by adding one before taking the base 10 logarithm. The solid line is the best linear fit through the origin. A cluster score threshold of 151,467 classifies 75% of the infinite inflow replicates as having a single species.

since the population has not had much time to differentiate from its clonal beginning. At updates 200,000, 300,00 and 400,000, the distributions of scores are quite similar. All of the score distributions are well described by an exponential. I transformed the scores from update 400,000 with $log(1 + \frac{cumulative\ count}{total\ count})$, and calculated the linear best fit through the origin in figure 3.2.

With the fitted function, I calculated cutoff scores at a variety of probabilities, where probability refers to the chance of counting the population as a single species rather than a

pair. A probability of 75% of classifying the single niche population as one species works well for sorting well-differentiated populations generated at intermediate inflow rates into distinct genotypic and phenotypic clusters. The corresponding cutoff value is 151,467.

Chapter 4

Adaptive radiation and emergence of species

The simplest ecosystem has a single niche. As discussed earlier in section 1.2, evolution takes place via successive sweeps of increasingly fit genotypes. A plot of fitness in *E. coli* in figure 4.1 shows a punctuated pattern, similar to the plot of cell size in figure 1.3.

4.1 A single niche in Avida

When digital organisms face a single niche environment, as in the threshold calibration experiments detailed in section 3.3.2, their fitnesses in figure 4.2 are indistinguishable from those of figure 4.1. Since resources are unlimited, the environment has no effect on growth rates and thus the speed of reproduction is the only determinant of fitness. As fitter digital organisms emerge through random mutation, selection causes the familiar pattern of long periods of stasis followed by brief periods of rapid change.

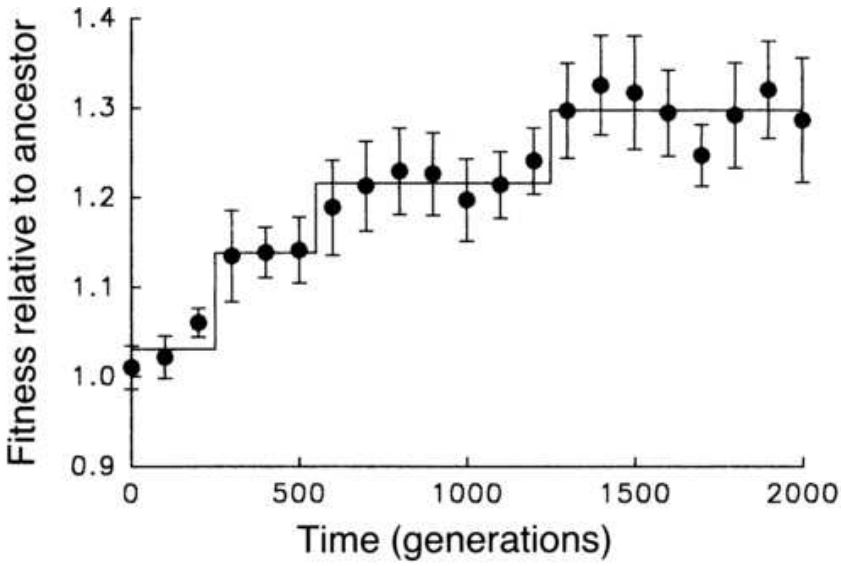


Figure 4.1: Fitness of *E. coli* in a glucose-limited environment. Fitness is measured by calculating the reproductive rate of the population relative to that of the founder, and displays punctuated increases as fitter mutants sweep through the population. (data from R. Lenski, MSU)

4.2 Multiple niches in Avida

Adaptive radiation of a clonal population of bacteria can occur as a result of different conditions, including seasonal environments, resource partitioning and spatial heterogeneity [78, 94]. The resulting evolved outcomes can be sensitive to small changes in environmental conditions [58, 93, 94]. Bacteria propagated under identical environments sometimes achieve similar fitnesses [93], but other experiments have found that replicate populations can diverge significantly from one another in morphology and mean fitness [58], suggesting that random events play an important role in evolution.

When a well-stirred digital environment has limited supplies of multiple resources, adaptive radiation can also result, even in the absence of seasonal environments and spatial

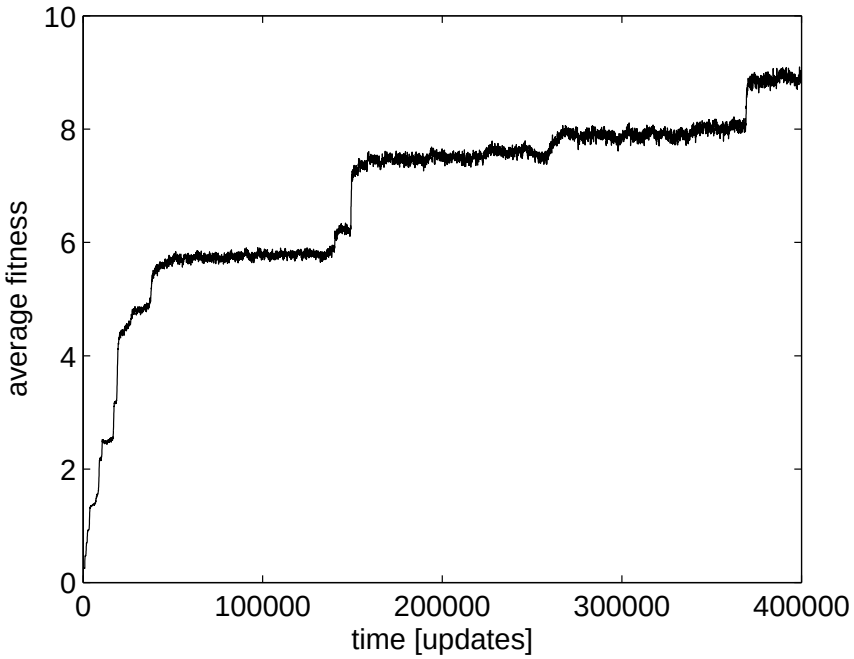


Figure 4.2: Fitness of digital organisms in Avida in a single niche environment. Fitness is measured by the reproductive rate, and displays punctuated increases as fitter mutants sweep through the population.

heterogeneity. The multiple resources provide an opportunity for organisms to minimize competition by differentiating from each other, specializing into different niches. Nonetheless, a given environment may produce a wide range in the number of species. As seen in bacteria [58], replicate populations propagated under identical conditions can diverge significantly from one another.

4.3 Patterns of differentiation and speciation

When a lineage evolves, speciation can occur in one of two ways: the lineage may change so much that it is justifiably called a new species, an event called anagenesis, or the lin-

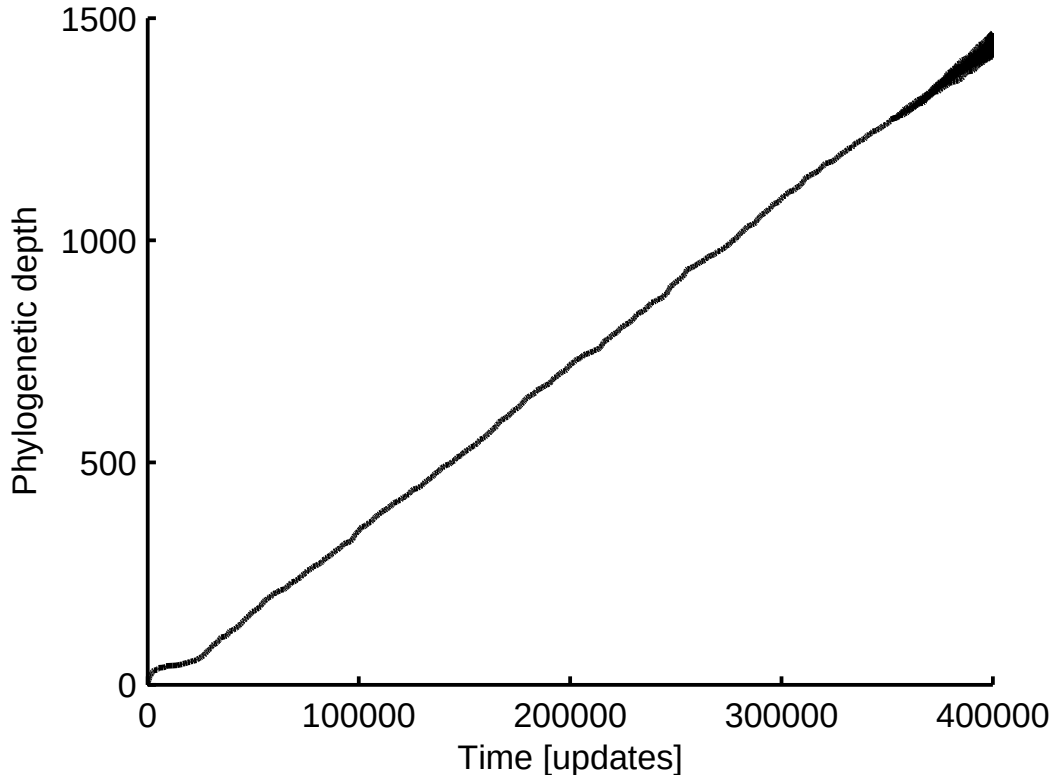


Figure 4.3: Phylogenetic depth versus time in an environment that does not promote speciation. Each line traces an organism in the final population back to the original unevolved ancestor. From [103].

eage may branch into new forms, called cladogenesis. The earlier figures 1.3 and 4.1 with their punctuated pattern, suggest anagenesis, while the morphotypes of *Pseudomonas fluorescens* in figure 1.4 suggest cladogenesis.

Multiple niche environments can exhibit both forms of speciation.

4.3.1 Low diversity populations

Figure 4.3 shows a typical phylogenetic depth pattern for a community of digital organisms that has not speciated. Following the final community backwards in time, all lines of

descent quickly coalesce. The most recent common ancestor (MRCA) of the community lived at around update 350,000, approximately 50,000 updates before the end of the experiment. From the MRCA, a single line of descent leads back to the founding ancestor. The small phylogenetic distance from an organism in the final population to the MRCA reflects the strong genotypic and phenotypic similarities within the final population. Although in this example the depletable resources were very abundant, this type of ecosystem evolved at least once out of 50 replicates at every inflow rate.

4.3.2 High diversity populations

By contrast, figure 4.4 shows a typical phylogenetic depth pattern for a community of digital organisms that has speciated. As before, each line traces an organism in the final community backwards in time. This time, the lines of descent do not quickly coalesce. There are four deep, clear branches that separate early, at around update 100,000. Near the end of the experiment, at around updates 300,000 to 350,000, these four branches fan out in the same way as the single branch in figure 4.4. The interpretation of this observation is simple: In a community that has speciated, all organisms within a single species share a fairly recent common ancestor, but the most recent common ancestor of all species lies in the distant past.

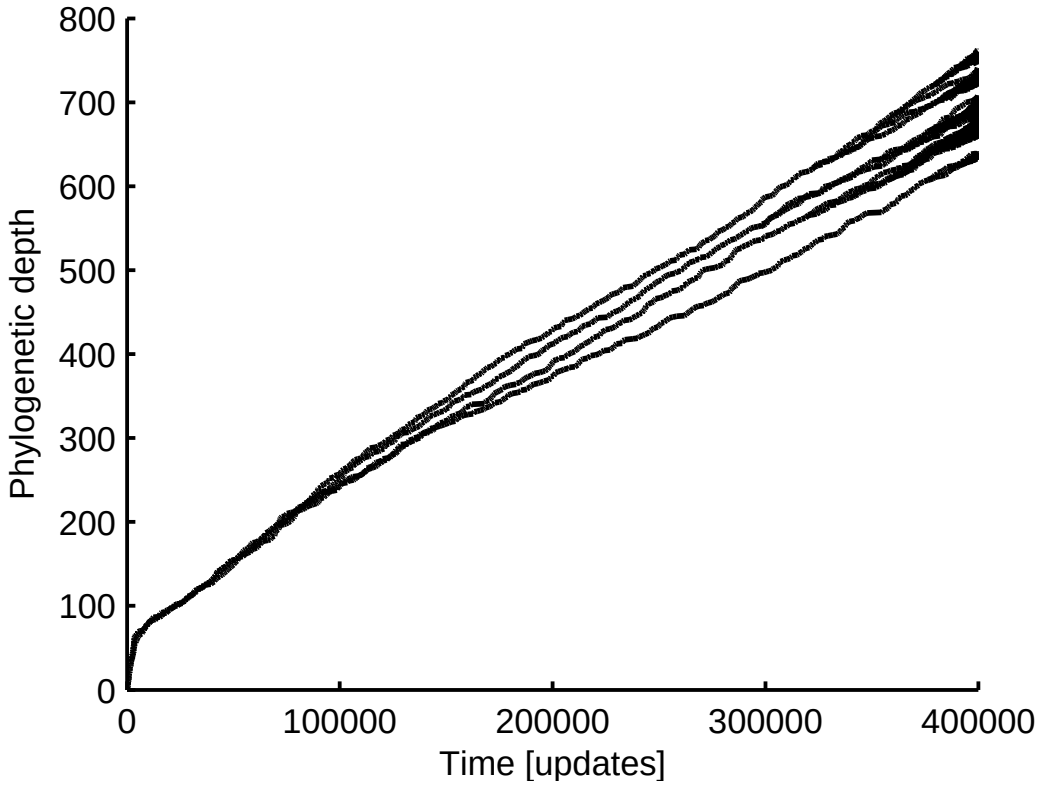


Figure 4.4: Phylogenetic depth versus time in an environment that does promote speciation. Each line traces an organism in the final population back to the original unevolved ancestor. From [103].

4.3.3 Time course of speciation

In this section, I investigate the ways in which a community evolves. The alterations may involve changes to the resources used, as well as phenotypic changes at branching points of the lineage and at other times.

Figure 4.5 shows resources used as a function of time in a replicate resulting in four species seeded with the basic unevolved ancestor. Species are numbered in order of branching. At the beginning of the experiment, rapid evolutionary change occurs, as organisms explore genotype and phenotype space. After approximately 70,000 updates, the founder

Figure 4.5: Resource use pattern as a function of time in a community of four species in an experiment seeded with an unevolved ancestor. Black areas indicate that a given resource is used at a particular point in time. Each subplot of the four corresponds to the line of descent from a representative genotype (the most numerous one) of one of the four species in the evolved population. Species subplots are sorted in order of their time of branching from the main line of descent, with the top one branching first, and so on. The experiment had the standard duration of 400,000 updates, but only the first 200,000 are shown since there are no subsequent changes to the figure.

of species 1 branches off from the rest of the population, and after branching, the lineage changes its phenotype and then settles on resources 2 and 3. At update 80,000, the ancestor of species 2, 3 and 4 has chosen other resources than species 1, consuming resources 1, 4 and 7. At update 110,000, species 3 and 4 branch off nearly simultaneously. By update 130,000, each species has settled on a different subset of resources. Some resources are shared by more than one species, but no two species use the exact same set. There are no changes to the four sets of resources used by the four species after update 130,000.

4.3.4 Methods of adaptation

The replicate in figure 4.5 above was seeded with a basic, unevolved genotype, so perhaps it is not surprising that rapid changes in phenotype occurred at the beginning of the experiment. Do rapid changes occur when a well-evolved seed genotype is used instead? Figure 4.6 shows resources used as a function of time in another replicate resulting in four species, this one seeded with a generalist who uses all nine resources [1]. The same basic environment with the same basic resource inflows was used. Note that figure 4.6 shows 400,000 rather than the 200,000 updates in figure 4.5, since the resource usage continued to change. To contrast with the unevolved ancestor, the generalist seed was the most numerous (and therefore successful) genotype in a population which had evolved for 400,000 updates.

In the beginning, rapid evolutionary change occurs, but not as rapidly as observed with the unevolved ancestor. On the other hand, evolutionary change in resource usage continues

[1] Generalists are difficult to evolve unless the environment is specifically structured to reward the use of many resources. This is done by limiting the reward an organism can absorb from performing any particular calculation, and scaling the rewards so that those which often take longer to evolve are more highly rewarded.

for much longer. The first species emerges at approximately update 30,000, the second at update 100,000, and the third and fourth split at update 160,000. The first two species to emerge show few changes in resource usage after speciation, while the last two continue to adapt. By the end of the experiment, perhaps thanks to the longer period of adaptation, all four species have settled on non-overlapping resources: species 1 on resources 1, 2, 4 and 5; species 2 on 7; species 3 on 6, and species 4 on 8.

One method of adaptation is through *gain of function*, easily seen in the early updates of figure 4.5. The various descendants of a genotype using no resources have acquired the ability to use each one of the resources at some point in time.

Another less obvious method of adaptation is through *loss of function*, also observed in bacteria [65, 71]. Loss of function can be seen in the early updates of figure 4.6, as the descendants of the generalist adapt to new environment by differentiating, this time by specializing on different resources to minimize competition.

The third important method of adaptation can be seen in both figures, more easily in the middle to later updates, namely the gain by a lineage of one resource often coincides with the loss of another, easily seen at update 90000 in species 2, 3 and 4 of figure 4.5, as well as in the switching of species 3 in 4.6 between resources 3 and 6 between updates 250,000 to 310,000. In the genomes of this section of the lineage, the code to use resource 3 is is derived from the code to use resource 6, and vice versa. This adaptation of existing genetic code to new uses has been observed in bacteria [9], and is called *antagonistic pleiotropy*.

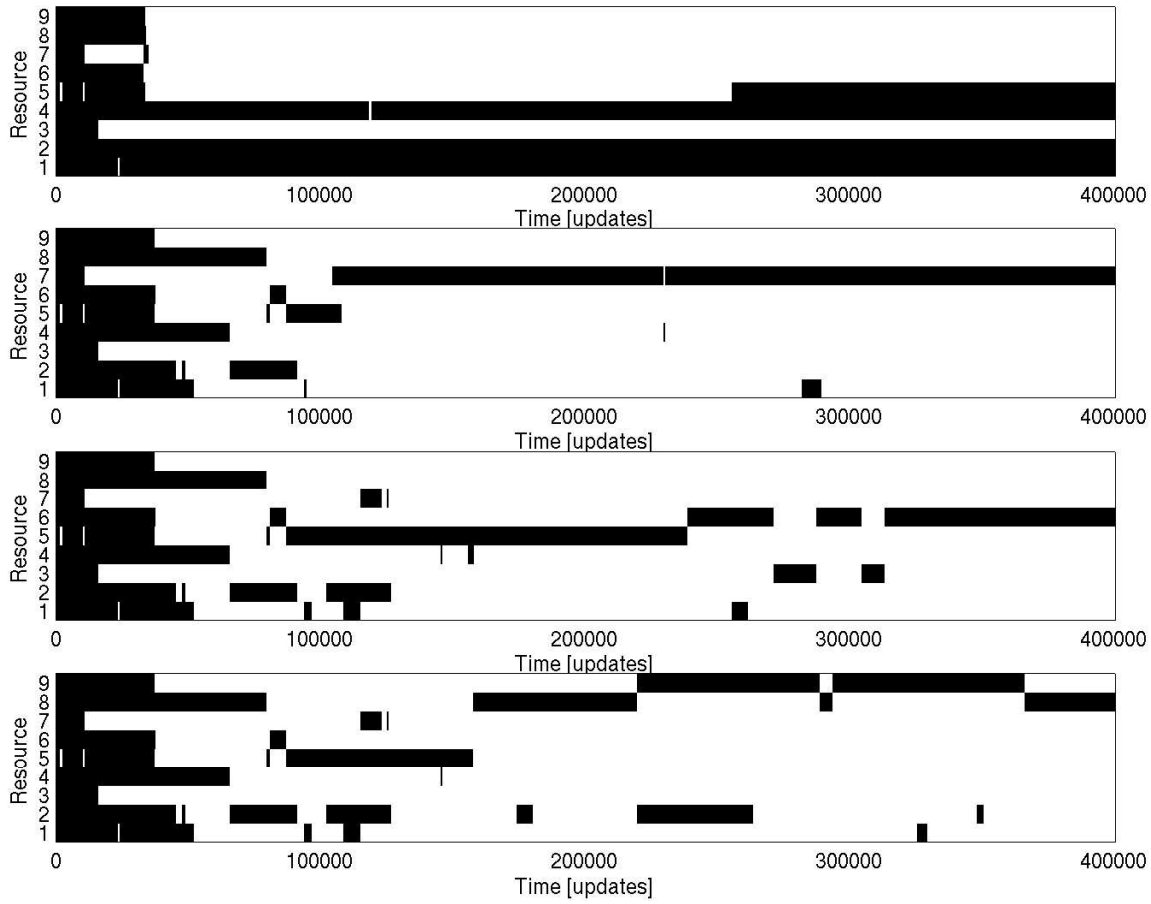


Figure 4.6: Resource usage pattern as a function of time in a community of four species in an experiment seeded with an evolved generalist who consumes all nine resources. Black areas indicate that a given particular point in time. Each subplot corresponds to the line of descent from a representative genotype (the most numerous one) of one of the four species in the evolved population. Species subplots are sorted in order of their time of branching from the main line of descent, with the top one branching first, and so on. The full duration of the experiment is shown.

Chapter 5

Diversity and productivity

Since digital environments can produce differentiated populations under conditions of multiple, but limited, resources, as can the biological ones described in section 1.3, will the resulting pattern of diversity in response to resource supply, i.e. inflow or productivity, resemble those found in biological organisms such as seen in figure 1.5? The answer is yes, but depends on how diversity is measured.

5.1 Measures of diversity

Kassen *et al.* [47] used Simpson's Index of Diversity $1-\lambda$ with stirred and unstirred cultures of *Pseudomonas fluorescens*. λ is given by

$$\lambda = \sum_{species} \frac{{n_s}^2}{N}$$

where n_s is the number of organisms in species s, and N is the total number of organisms. While the index commonly used is $1-\lambda$, occasionally $\frac{1}{\lambda}$ is found. The quantity λ sums the

squares of relative population sizes so a population consisting of 90% of one species and 10% of another has a λ of 0.82, while another population with 50% each of two species has λ of 0.5. Figure 1.5 showed a very weak peak in diversity when measured with Simpson's Index, but the experimenters found many different colony morphs at intermediate productivity values, although these were rare.

In later experiments, again with *Pseudomonas fluorescens*, Kassen *et al.* [48] further investigated diversity, this time expressing it with both forms of Simpson's Index as well as the number of morphotypes, i.e., different forms. This is shown in figure 5.1. In the daily disturbance regime, there is a small peak in diversity when the measure is the number of morphotypes, but it disappears when Simpson's Index is used.

When number of species is the measure of diversity, the relationships between the number of rodent species and the log of rainfall levels in two types of habitats are well fitted by second order polynomials [2]. Since the number of species (or morphotypes) is a simple and sufficient diversity measure, I will use it to measure diversity in evolved populations of digital organisms.

5.2 Productivity and species in Avida

The peak in diversity observed in natural populations at intermediate productivity also appears in populations in Avida. Figure 5.2 summarizes the pattern of diversity plotted by resource inflow rate (A) and time (A and B). In A, the number of species as determined by the clustering algorithm in section 3.3.2 is shown at four time points through the repli-

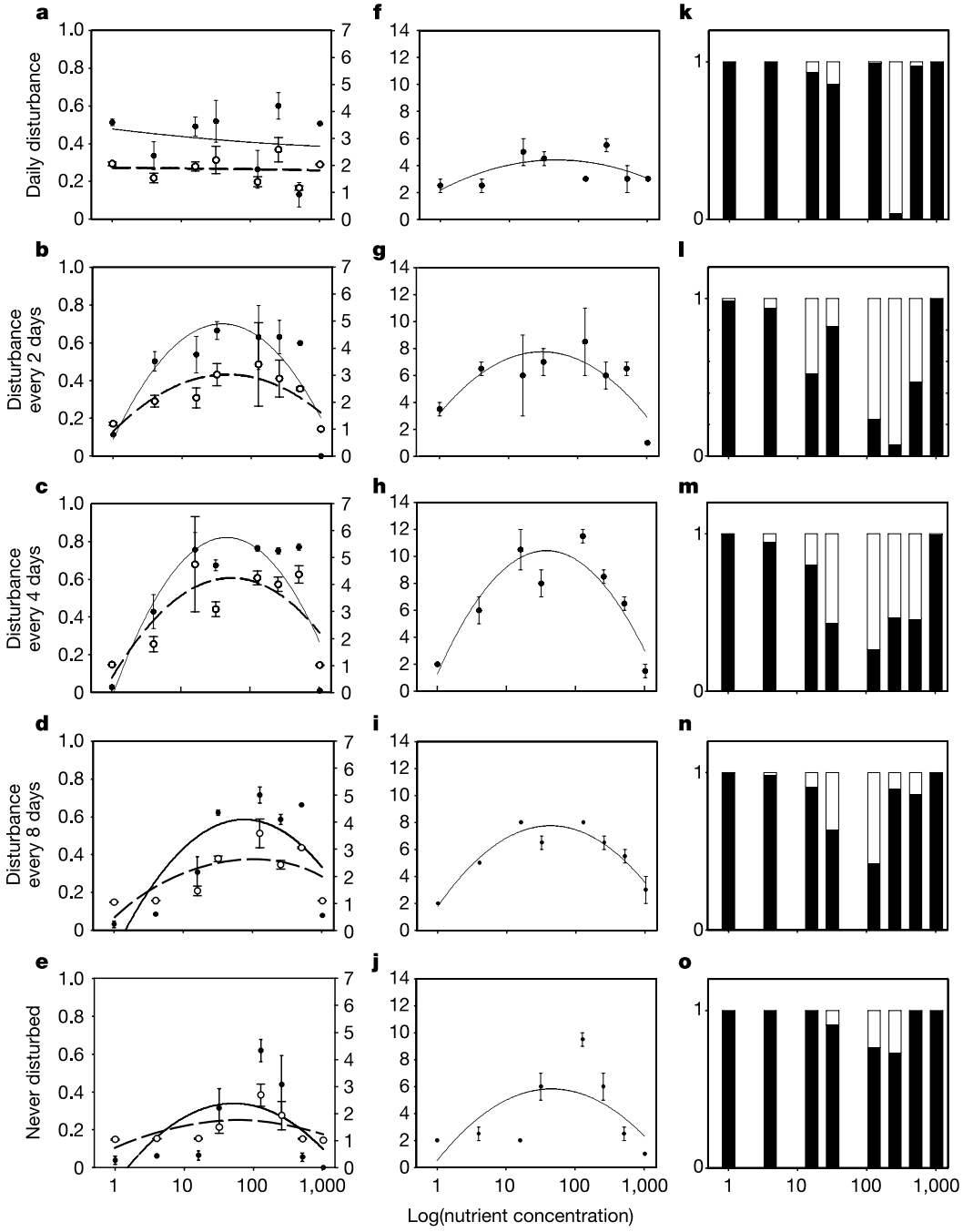


Figure 5.1: Diversity in relation to productivity as a result of adaptive radiation in *Pseudomonas fluorescens*, by disturbance regime. More frequent disturbances produce more homogeneous environments. Rows correspond to the disturbance regime, and columns correspond to different diversity measures. In the first column, diversity is expressed as $1 - \lambda$ (solid line) and $\frac{1}{\lambda}$ (dashed line); in the second column, as the number of distinct colony morphs; in the third column, as the relative frequency of the smooth morphotypes (filled sections) with respect to other types. From [48].

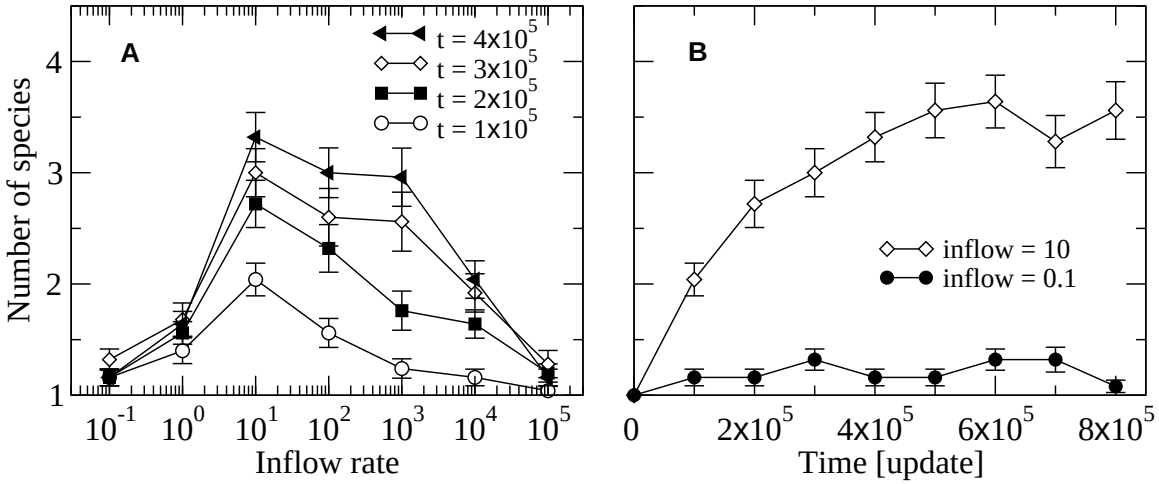


Figure 5.2: Mean number of species as a function of inflow rate (A) and time (B). Error bars indicate standard error over 25 replicates. All runs are seeded with an unevolved ancestor unable to use any resources. From [6].

cate: 100,000, 200,000, 300,000, and 400,000 updates. Every replicate starts off with a clonal population, i.e., every organism is identical. The hump shape becomes apparent at 100,000 updates. As time passes, the average number of species at intermediate inflows heads upwards. By update 400,000, the highest average of about 3.5 species is achieved by the replicates that receive an intermediate inflow rate into the environment of 10 units of each resource per update, followed closely by experiments with inflow rates of 100 and 1000 that show an average of three species.

The upward trend in diversity with respect to time in part A of figure 5.2 suggests the possibility that maximum diversity has not yet been achieved. To check this, I allowed all 25 replicates from inflow 10 (which had many species) and inflow 0.1 (which had very few) to continue for double the usual 400,000 updates, to see if diversity would continue to increase. The results can be seen in figure 5.2B. The average species count levels off at

around 3.5 in the inflow 10 case, and stays near one in the inflow 0.1 case. The other inflow rates confirm that diversity does not significantly increase past 400,000 updates.

Interestingly, none of the replicates ever reached the maximum number of species allowed by the principle of competitive exclusion. Nine species, each specializing on a single resource, would minimize interspecies competition, and maximize diversity.

5.3 Limits on diversity

The question of why we do not observe populations that achieve the maximum number of possible species can be answered in part by analyzing resource usage patterns in detail. In section 4.3.3, I showed that species often use more than one resource, and often share one or more of these resources with other species. This suggests that genes vary in their degree of difficulty to acquire and use. If, for instance, modifying a genome to add the `NAND` task takes few mutations and does not add appreciably to gestation time, then it may be cost-effective to add it, even if `NAND` is already popular. The use of multiple resources shrinks the number of unexploited niches which a mutant might occupy. Biological species sometimes also continue to consume widely preferred resources without greatly compromising their specializations on less exploited resources [79].

Another limit on diversity is the number of resources that can be consumed cost-effectively. While an unexploited resource may be relatively abundant in the environment, the cost of acquiring or keeping the gene may still outweigh its benefit. Figure 5.3 shows that except at intermediate inflows whose environments have the highest diversity, many

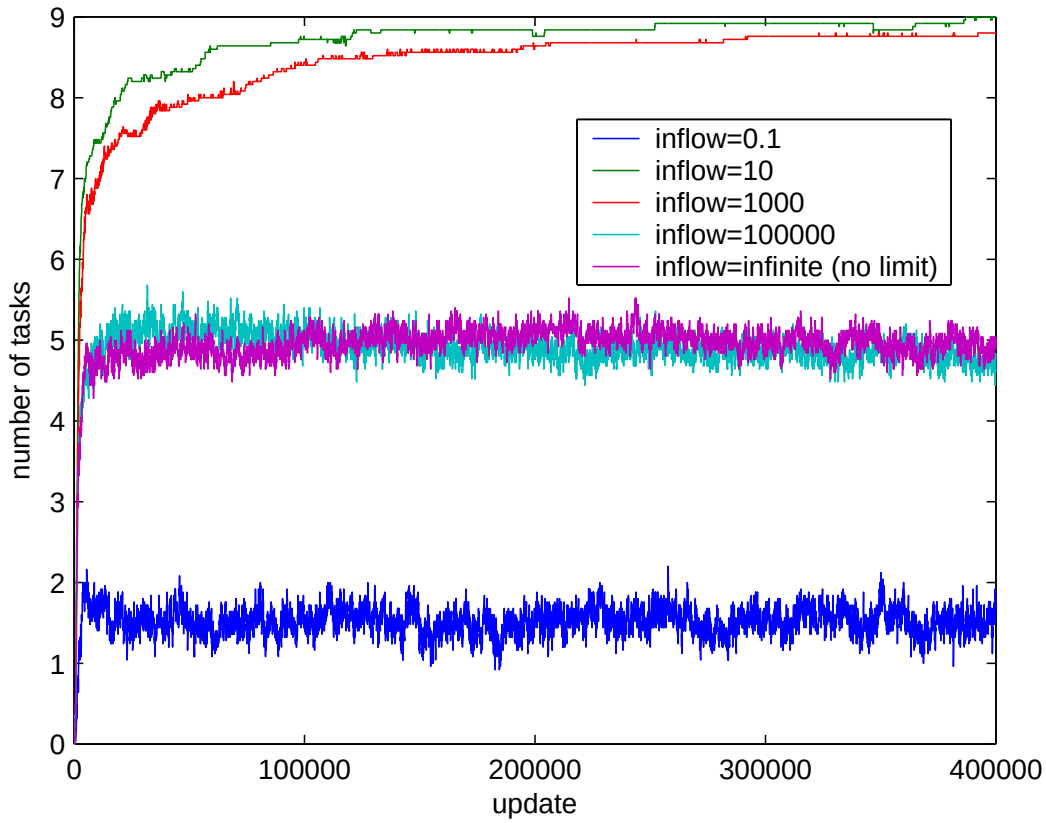


Figure 5.3: Number of resources consumed by the population, by inflow rate and time. At each time point, number of resources consumed by at least one member of the population in a replicate is counted. For clarity, only every second inflow rate, plus the infinite (non-depletable) case, is plotted. At a very low inflow rate, few resources are exploited. At intermediate inflows, all or nearly all nine are consumed, and at extremely high and infinite inflow rates, an intermediate number are used.

resources remain untapped by the population.

"Not" istypically one of the first to arise. "Equals", on the other hand, is much more difficult to evolve, and typically only arises if the reward for it is very high compared to other tasks. In order to study the influence of task complexity on the time necessary to evolve a task, I measured the mean time until the first consumption of the nine task-associated resources, shown in Figure 5.4.

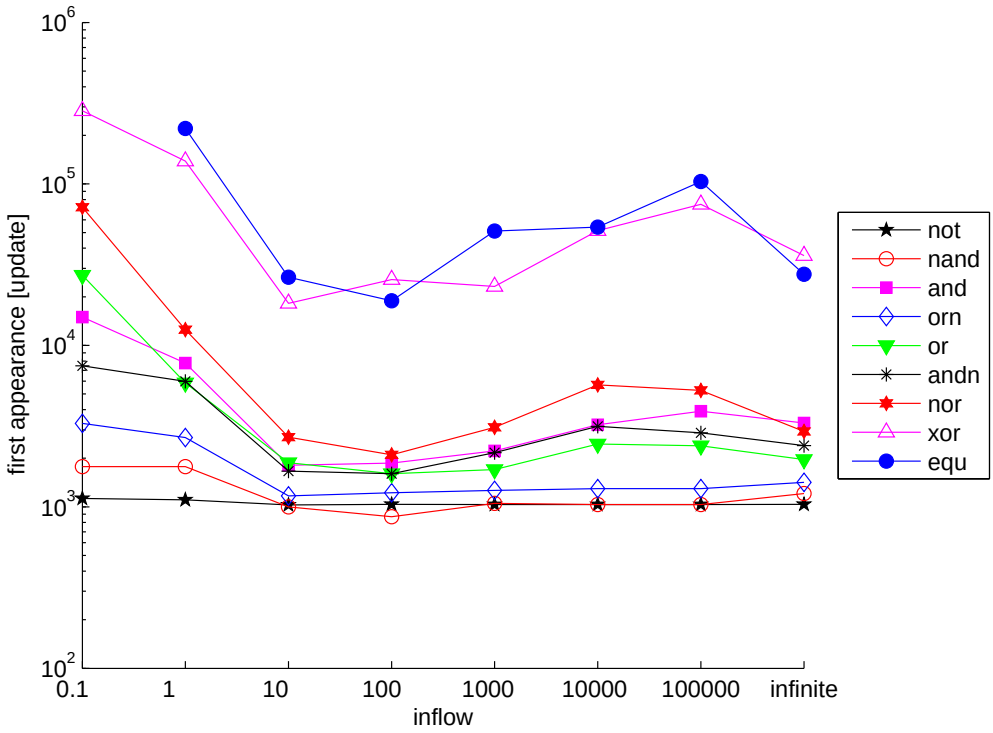


Figure 5.4: Mean time until first consumption of a resource, in replicates where the resource was used. A higher time value implies a more difficult gene to acquire. Note: populations were sampled every 10000 updates, so transitory genes may escape detection. If there is no data point, the gene was not found in any replicate.

A gene for a task can appear *de novo*, as it does when an unevolved organism acquires it for the first time. Another method to acquire a new gene is to modify an existing one to perform a new function. In Avida, the transformation of one gene into another can be inferred from the coincidence of the addition of one computation and the deletion of another, observed in the resource usage patterns in figures 4.5 and 4.6. This phenomenon resembles antagonistic pleiotropy in biological organisms, such as *E. coli* [9], where a gene which is no longer useful in the current environment is adapted to a new function.

A simple cost-benefit argument can explain the unimodal diversity-productivity pattern. At low inflow rates, the cost of adding a gene for a computation is not likely to be repaid by the limited supply of an untapped resource. At high inflow rates, even levels of exploited resources may no longer be limiting, so it may be easier to improve the efficiency of the current genome than to add another gene [1].

5.4 Frequency-dependent selection and stability

In evolution, mutation provides variability, and natural selection acts to limit that variability by weeding out all but the fittest. Nonetheless, the existence of multiple co-evolved species with distant common ancestors implies that there must be a force acting to maintain diversity in some environments.

The aspect of the experimental environments which maintains species after they have arisen is the presence of depletable resources, whose levels vary inversely with the number of organisms consuming them. A stabilizing force called negative frequency-dependent selection (FDS), which I explain below, can act to maintain diversity.

In frequency-dependent selection, the fitness of a phenotype is a function of its proportion in the population. I show a simple example of both positive and negative FDS for a population consisting of two competing phenotype in figure 5.5.

In positive FDS, the fitness of a phenotype increases with its frequency. Any equilibrium where both phenotypes have equal fitness is unstable. If phenotype 1 is subject to

[1] In all experimental Avida environments, there is maximum quantity of any resource which can be absorbed, since unlimited growth rates are biologically implausible.

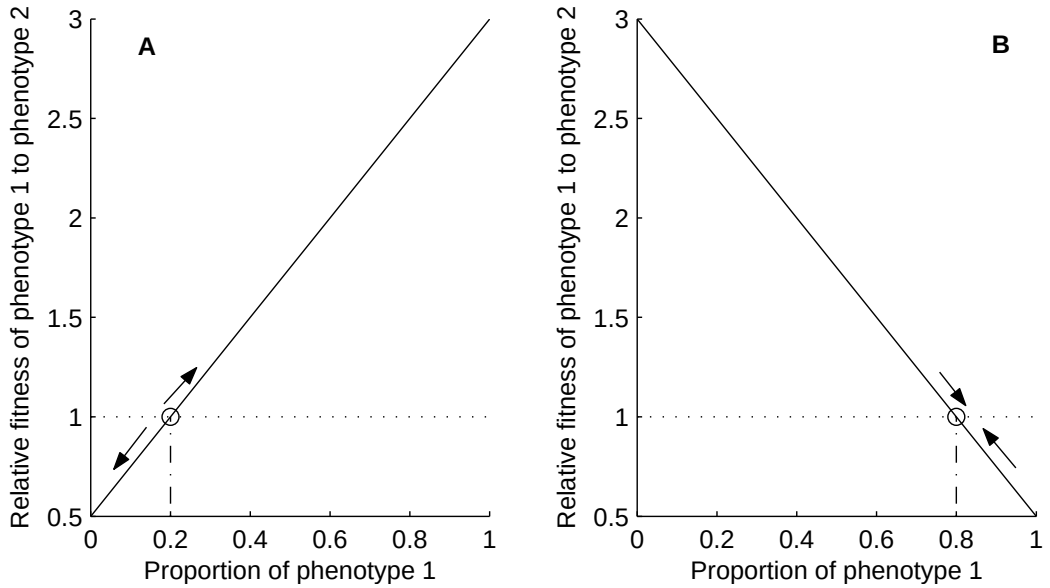


Figure 5.5: Relative fitness of a phenotype in a two-phenotype population as a function of its proportion in the population, illustrating positive (A) and negative (B) frequency-dependent selection. In A and B, a population consisting of 20% phenotype 1 and 80% phenotype 2 has a relative fitness ratio of 1, indicating that the two phenotypes are in equilibrium. In **A**, the equilibrium is unstable. If phenotype 1 goes over 20% of the population, then it will be fitter than phenotype 2 and increase its proportion in the population until it takes over. Conversely, if phenotype 1 is below 20%, it will be less fit than phenotype 2 and go to extinction. In B, the equilibrium stable. If phenotype 1 is at proportion greater than its equilibrium, it has a lower fitness than its competitor phenotype 2, and will decrease in frequency. Conversely, if phenotype 1 is at a lower proportion than equilibrium, it will be fitter than phenotype 2, and increase in frequency.

positive FDS, then a disturbance which increases its proportion past equilibrium will send it to fixation, while one which decreases its proportion will drive it to extinction. In biology, this can happen when there are gains to cooperation, such as for bacteria that gain resistance to antibiotics when they aggregate to form biofilms [25, 90].

In negative FDS, the fitness of a phenotype decreases with its frequency. An equilibrium where both phenotypes have equal fitness is stable. A disturbance that increases its

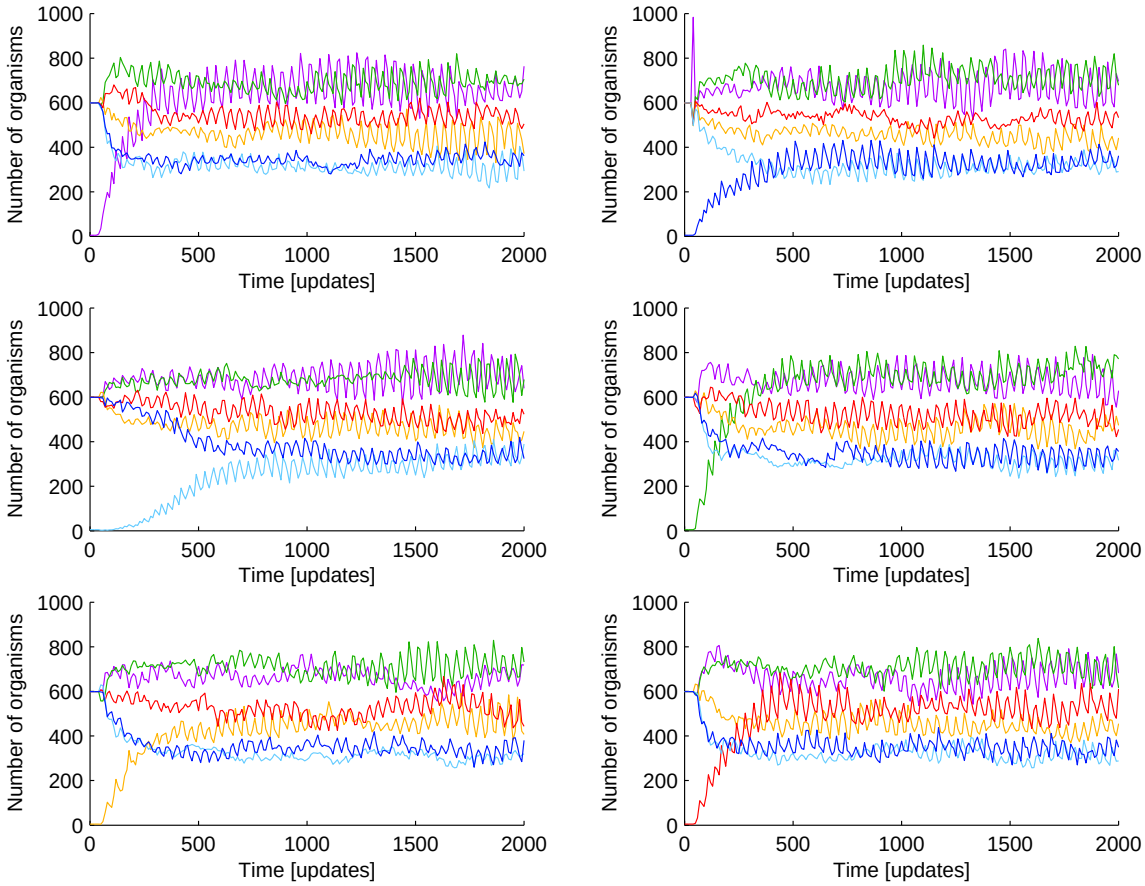


Figure 5.6: Species invasion when rare in a replicate with six species. Each species in the replicate is able to invade a population of the remaining five, where each species is represented by its most numerous genotype. The effect of negative frequency dependent selection is clear, as species adjust their numbers until an equilibrium is reached. From [6].

proportion past equilibrium will result in reduced fitness, returning the population to equilibrium, and vice versa. In resource-limited environments, an overabundance of a species or phenotype will result in a shortage of the resources it prefers, reducing that phenotype's fitness and in turn its abundance. Conversely, uncommon species or phenotypes may find their preferred niches relatively uncrowded, enabling them to invade when rare.

The co-evolved species in Avida ecosystems are subject to resource limitations, and

therefore negative FDS may explain why a phylogenetic tree can exhibit long, persistent branches, i.e., species. I conducted a series of invasion experiments on evolved ecosystems with more than one species. I created a new population from the old ecosystem by adding low numbers of one species to high numbers of the other species. Each species was represented by its most numerous genotype, and I turned off mutations to avoid any changes to species. In each of the invasion experiments in figure 5.6, the rare species was able to increase its numbers until the population reached equilibrium. Moreover, the final equilibrium frequencies in all six invasion experiments were the same. At equilibrium, all organisms are equally fit.

5.4.1 Phenotypes of co-evolved species

A closer look at the ecosystem depicted in figure 5.6 allows us to identify the origin of the FDS dynamics. Figure 5.7 shows the mean number of times that each rewarded computation is performed in one life cycle by a species. While the mean number of times a resource is consumed is proportional to the overall frequency of consumption of the resource in the population, the actual consumption also depends on its abundance in the environment.

Species 1 is the biggest consumer of resources 5 and 9. Species 2 likes resource 7, species 3 prefers resource 1, species 4 focuses on resource 2, species 5 dominates in resource 3, and species 6 consumes resource 6. Many species consume resources 1 and 4. The evolved species specialize on different resources, although some overlap in consumption remains. Contrary to some analyses in biological organisms [29], both generalists

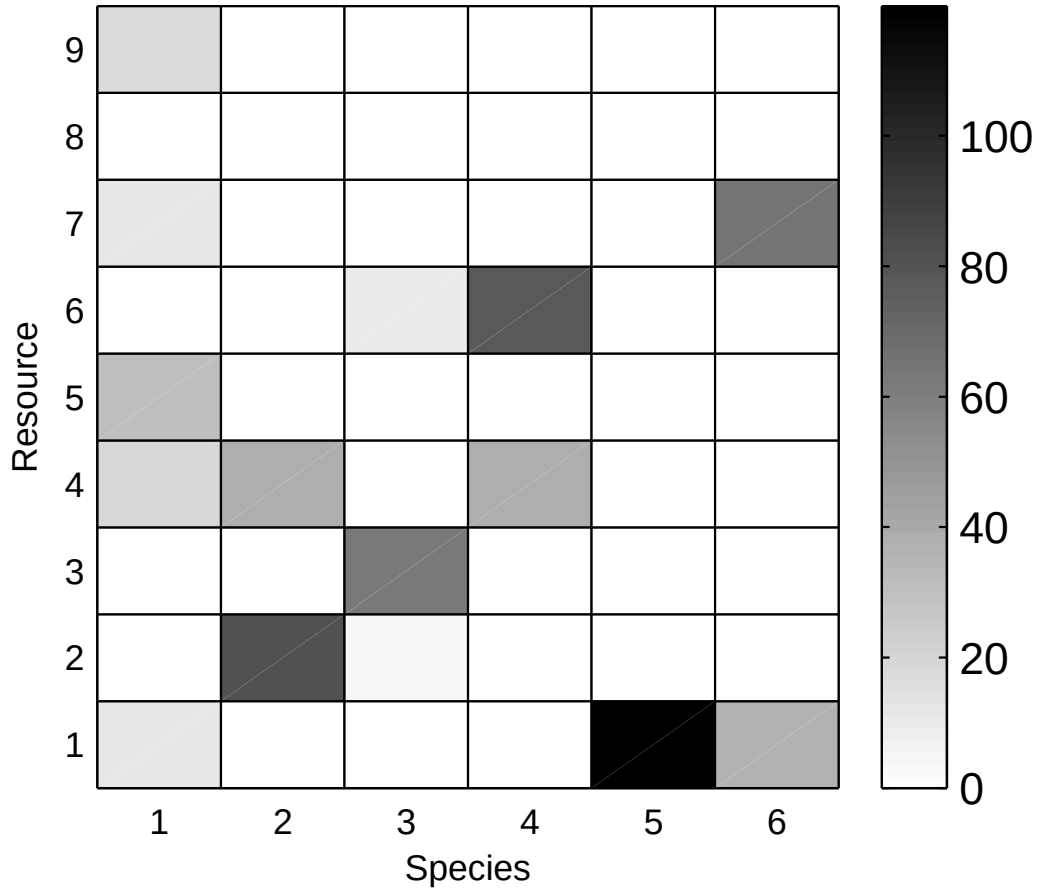


Figure 5.7: Matrix of resource use by the species depicted in figure 5.6. Each rectangle is shaded according to the number of times the particular resource is consumed in the life cycle of an average member of the species. Although there is some overlap in resource use, each species dominates in at least one.

(species 1) and specialists (species 5) can evolve in the same environment.

Figure 5.7 also confirms phylogeny as a basis for a definition of species. The species in the figure exhibit clear within-cluster similarities and between cluster differences in phenotype.

5.5 The founder effect

One possible concern is that the evolutionary outcomes reported above are dependent on the choice of seeding genotype for an experiment. Since the basic genome is hand-written and unoptimized, it could be argued that the genome is easy to evolve since there are many ways to make it fitter. The adaptive radiation arises in part through gains in function, as the mutant descendants of the founder develop genes *de novo*.

To address the possibility that the evolutionary trajectory of the population may be strongly influenced by the characteristics of the seed genotype, resulting in a *founder effect*, I generated another 25 replicates for each of the seven inflow rates, this time seeding them with five different evolved genomes. Each genome coded for a perfect generalist, that is, one with the ability to consume all nine resources. The generalists were evolved in a special infinite-inflow environment with a reward structure that encouraged generalization to as many resources as possible. In order to ensure that the seed organisms are fit, the generalists were also required to be the most numerous genotype in their replicate population. Each generalist seeded five replicates for each inflow. Otherwise, the experimental setup matches that of the basic replicates detailed in section 2.8.

Earlier in section 4.3.3 and in figure 4.6, I showed that diversity could arise from an evolved generalist as it could from the unevolved ancestor. The overall pattern of diversity produced by generalist ancestors is also similar.

After 400,000 updates, we can see in figure 5.8 the species count that emerges with generalist seed organisms. Qualitatively, the species count closely resembles that in figure

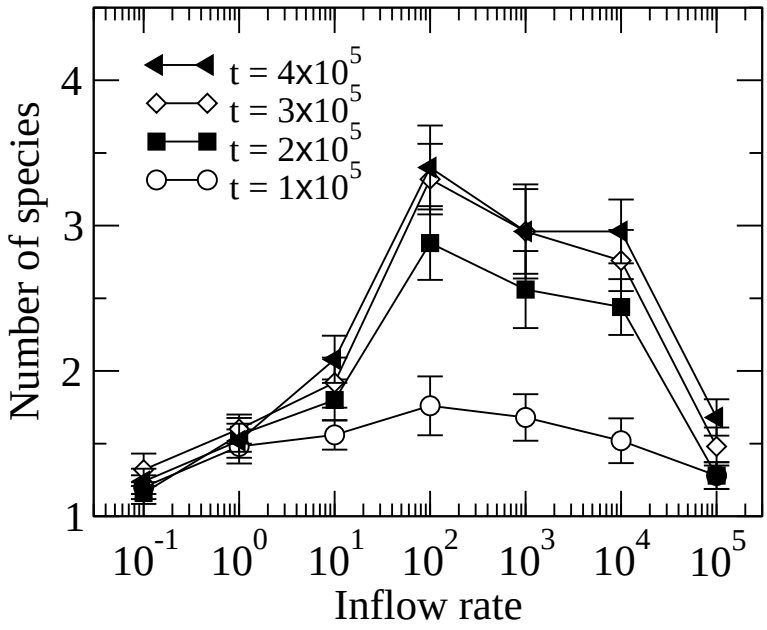


Figure 5.8: Mean number of species as a function of inflow rate where runs are seeded with evolved generalist organisms. Error bars indicate standard error over 25 replicates. All runs are seeded with a clonal population based on one of five generalists who use all nine resources.

5.2, which were obtained with a simple, unevolved founder. The hump shape in species count is still there, but with a slightly lower number of species at inflow 10.

As the descendants of generalists adapt to their environments, they specialize on different resources. As can be seen in figure 5.9, the six species in a replicate seeded with a generalist exhibit distinctly different phenotypic profiles, as did the six species in figure 5.7. After sufficient time, the diversity-productivity relationship appears to be independent of the nature of its founders.

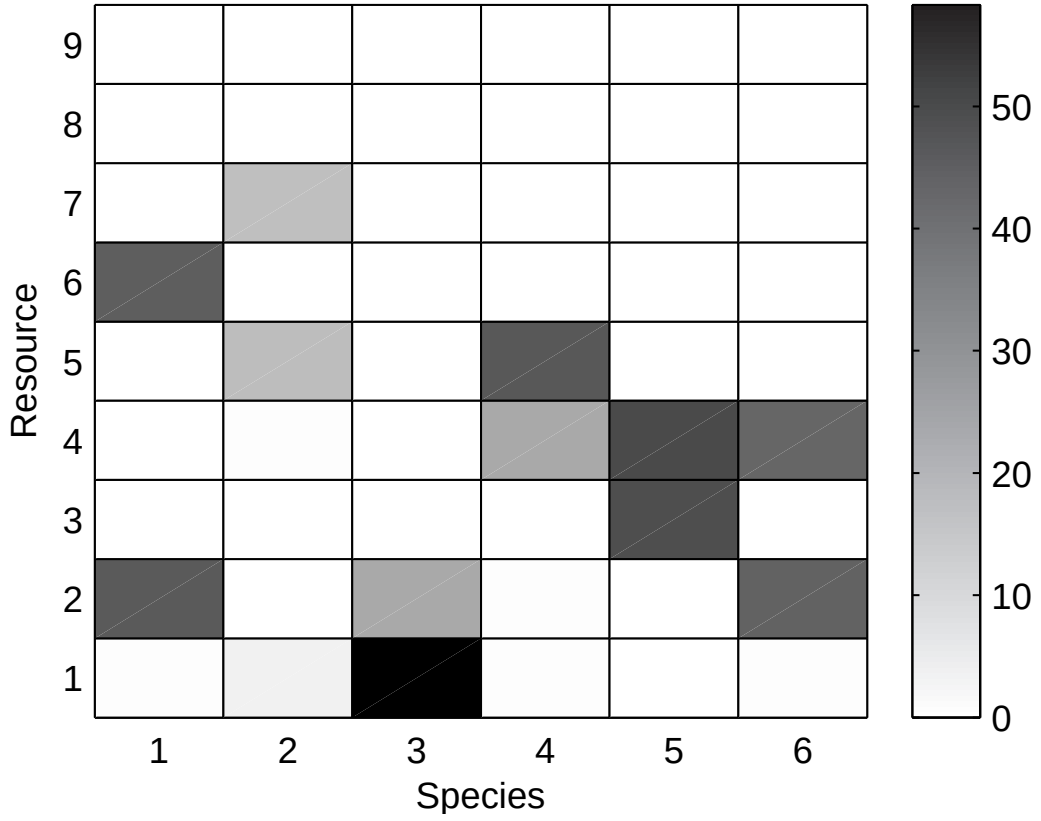


Figure 5.9: Matrix of resource use by in a replicate with six species with a generalist founder. Each rectangle is shaded according to the number of times it is used in the life cycle of an average member of the species. Although there is some overlap in resource use, each species dominates in at least one.

5.6 Population size limitations

The well-stirred environment occupied by digital organisms is subject not only to resource limitations, but also to population size limitations. In this section, I analyze the effect of population size on the diversity-productivity relation. In the Avida software, the population size is fixed. As one organism is born, it is placed in a cell and any previous occupant is killed and removed. It is possible in principle that this limitation may change the experimental outcomes, since population sizes are unable to vary dynamically and therefore

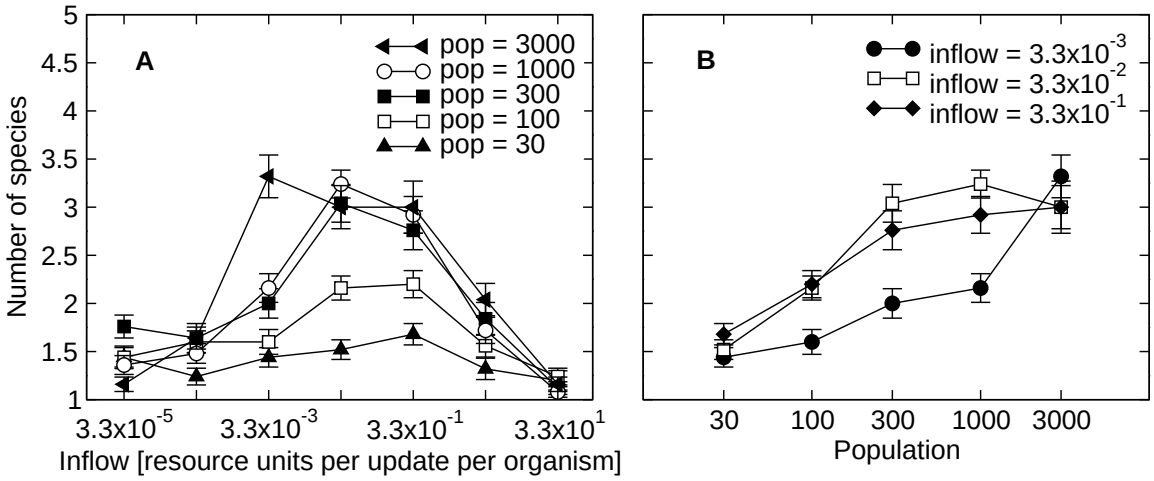


Figure 5.10: Mean number of species as a function of inflow rate per organism (A) and population (B). Error bars indicate standard error over 25 replicates. All runs are seeded with an unevolved ancestor unable to use any resources.

population size effects on resource levels can be muted. Even though version 1.99 of the Avida software does not allow experiments with variable population sizes, we can nonetheless study the effect of pure population size.

In the experiments, a general result is that population size does not affect species richness as long as it is not too small. At very small population sizes, random drift may play a disproportionately large role, frequently eliminating nascent species before they can establish themselves. As can be seen in figure 5.10, there is an upward trend in species count with respect to population which starts to level off at a population size of 300 in some environments with intermediate inflow rates. The trend past a population size of 3000 is not established. Although the computational resources available to me consists of a large computer cluster with approximately 200 CPUs, they are not sufficient to address 25 replicates at population size of 10,000 at this time. A single replicate with a population size of 3000

may take a day or more on a single CPU.

5.7 Cutoff dependence of the clustering algorithm

For the experiments with digital organisms reported in this thesis, diversity, as given by the number of species, is a critical measure. Starting with a single cluster containing all genotypes, the clustering algorithm iteratively divides the population, calculates a score for the current state based on phylogenetic distance, and then compares the score to a cutoff value. If the score is above the cutoff, then the population is divided and the algorithm continues. When the score falls below the cutoff, the algorithm stops, and each genotype is assigned to a species.

The cutoff is determined from an empirical distribution over cluster scores from 50 replicates of a single-niche environment. The single niche environment is created by making resources non-depletable. The score is calculated for a two-cluster case according to the algorithm described in section 3.3.2. From a linear fit through the log-transformed empirical distribution of scores, I can estimate the cumulative proportion of replicates whose scores fall below a given cutoff. A score below the cutoff indicates a replicate that is classified as having one species, while a score above the cutoff corresponds to a replicate that is classified as having two or more species. A cutoff value determines a cutoff proportion, an estimate of the percentage of single-niche replicates evaluated to have one species. The number of species as a function of the cutoff proportion can be seen in figure 5.11.

In choosing a threshold, a balance must be struck between erring on the side of too many

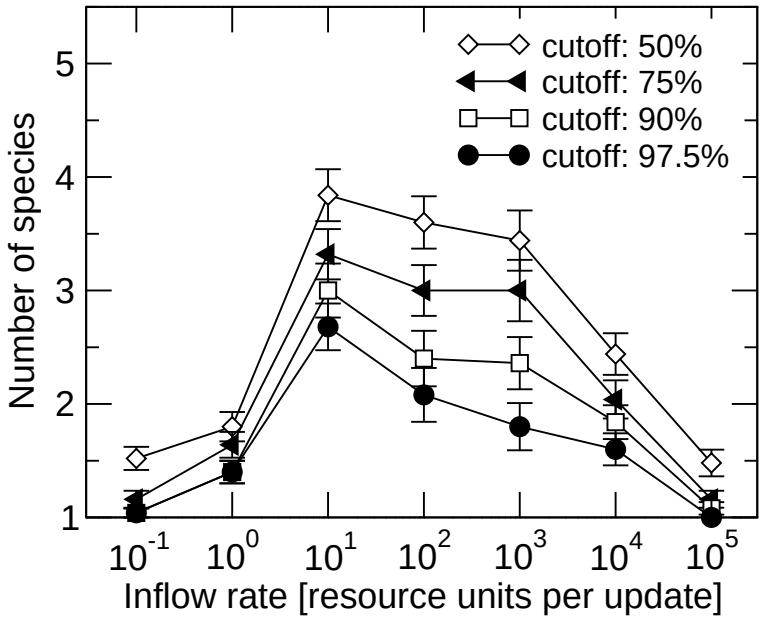


Figure 5.11: Mean number of species as a function of the cutoff percentage of the species clustering algorithm. Error bars indicate standard error over 25 replicates. The cutoff proportion is the estimated probability that an infinite (non-depletable) inflow replicate is determined to have one species rather than two. The estimate is derived from a linear fit through the origin of the log-transformed cluster scores.

versus not enough species. The species as determined by the algorithm should be both phenotypically and genotypically distinct from each other, but exhibit strong within-species similarities. A cutoff proportion of 75% corresponds well to phenotypic sorting of well-differentiated populations generated at intermediate inflow rates, although small numbers of organisms within a cluster are occasionally phenotypically distinct from other members of the cluster. At very high or very low inflow rates, a cutoff proportion of 75% occasionally assigns phenotypically similar organisms to different species. Nonetheless, we can see from figure 5.11 that the hump-shaped pattern of diversity with respect to productivity is robust to the exact value of the cutoff.

Chapter 6

On the orthogonal nature of species

The principle of competitive exclusion states that in competition between species that seek the same ecological niche, one species survives while the other does not under a given set of environmental conditions. This is also phrased as “one niche, one species”. It follows that stably co-existing species in an environment should be phenotypically distinct from one another, forming tight clusters [11, 40].

There have been few attempts to quantify phenotypic differences between and within species. One complicated approach has been to consider speciation to be a self-organizing process of correlations in species properties [53]. Below I propose a simpler measure which has an easy geometric interpretation, but is nonetheless sufficient to support the digital species concept in section 3.2, allows me to measure phenotypic properties of stable ecosystems, and therefore lets me make and test predictions about the outcome of the invasion of novel species into established ecosystems.

Not	Nand	And	OrNot	Or	AndNot	Nor	Xor	Equals
96	0	0	0	96	0	95	0	0

Figure 6.1: Computation profile of an Avida organism. In vector form, the profile is (96, 0, 0, 0, 96, 0, 95, 0, 0). The values are taken from the most numerous genotype in the replicate of figures 5.6 and 5.7.

6.1 A representation of phenotype in Avida

A simple representation of phenotype can be summarized as a list of the number of times that each (rewarded) computatation is performed by an organism, genotype or species in one life cycle: a resource use profile. Figure 5.7 showed the resource use profiles of the species in an ecosystem in matrix form. Similarly, figure 6.1 shows a resource use profile of an individual digital organism. If two phenotypic profiles resemble each other, their organisms would be in direct competition over the same resources. If these organisms belonged to different species, then competitive exclusion would suggest that only one of the species would survive in the long run.

An ordered array of numbers such as a phenotypic profile can be represented as a vector. Two phenotypic profiles automatically give two vectors. If the genotypes specialize on completely non-overlapping resources, their vectors will be orthogonal to one another. If they overlap on many resources, they are in strong competition, and the vectors will tend to point in similar directions. A measure that captures this difference in resource use is the angle between the two resource use vectors, and the angle can be calculated with the cosine law:

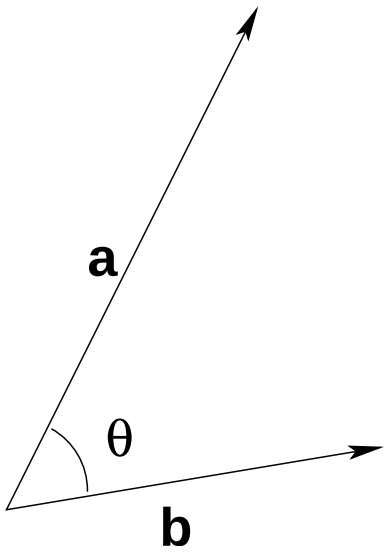


Figure 6.2: The angle θ between two vectors **a** and **b**.

$$cos\theta = \frac{\underline{a} \cdot \underline{b}}{\|\underline{a}\| \cdot \|\underline{b}\|}$$

Because resources are consumed and there are no byproducts by the rules of the environment, all components of a vector are non-negative. A phenotypic profile vector in the current experiments must therefore lie in the non-negative orthant. The maximum angle between two non-zero vectors (at least one component is non-zero) is therefore 90 degrees.

A digital organism does not need to consume any resources to live, because as described in section 2.7, Avida provides a "basal" resource to all organisms even if they cannot perform any rewarded computations. Such an organism would be assigned a phenotype vector of all zeros. Now the angle between a zero vector and any other is not defined. For the purposes of phenotypic profiles I will assign an angle between a zero vector and another as follows: if both vectors are zero vectors, the two genotypes concerned are direct com-

petitors even they do not compete for limited resources, so the angle between them should also be zero. If one of the vectors is not a zero vector, then the organisms are effectively specialized on completely non-overlapping resources (those associated with computations, not the basal resource), so I will assign an angle of 90 degrees.

6.2 Species orthogonality

Earlier we saw in figure 5.7 that species of digital organisms exhibit distinct and well-differentiated phenotypes as do their biological counterparts. This suggests that the angles between a pair of phenotype vectors from two species in an ecocystem should be strictly greater than zero. Moreover, if species form tight clusters [11, 40], then the angles between species phenotype profiles should be larger than the angles between genotypes in the same species. Using phenotype data from the original set of replicates, we can see in figure 6.3 that the mean angles between co-evolved species are significantly greater than the angles within species.

While between-species angles lie pretty close to 90 degrees, they are not precisely so. Among the non-extreme (1 to 10,000 units per update) inflow replicates, a few have one or more small (i.e., less than 45 degrees) pairwise angles. One possible explanation for this non-orthogonality is that evolution has not stopped. Indeed, species co-evolve, but they do not co-evolve in lock step. There will be times when they overlap in resource use, even if none of the resources are necessary for survival. Recall that this was the case in figures 4.5 and 4.6, where species often shared resources.

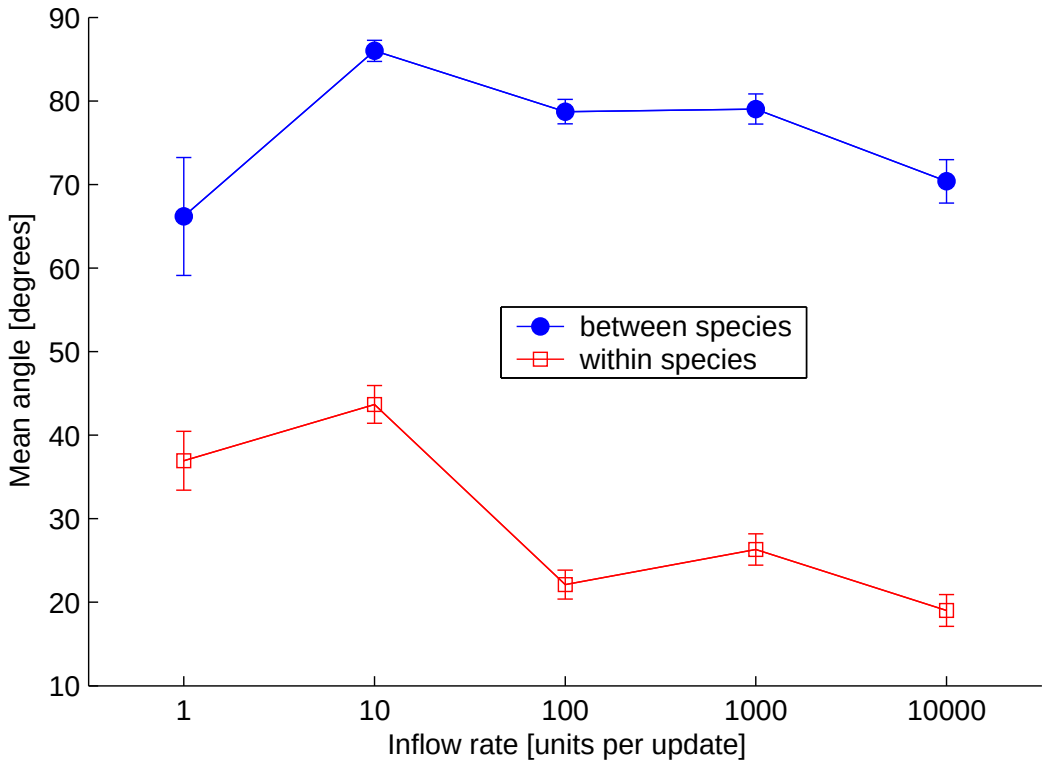


Figure 6.3: Pairwise between-species and within-species task profile angles by inflow. Between-species values are calculated over all pairwise angles between co-evolved species, where a species is represented by its most numerous genotype at 400000 updates. The within-species values are calculated over all the pairwise angles within a species. Error bars indicate standard error in the 17, 111, 90, 97, and 35 pairs of species at inflows of 1, 10, 100, 1000, and 10000 respectively. Inflows rates of 0.1 and 100000 are omitted because there were only 4 pairs each.

If angle is a good indicator of species relatedness, we would expect within-species angles to be close to zero. Instead, we observe values which are significantly above zero, reaching nearly 45 degrees at an inflow rate of 10 units per update. It is possible that such excessive angles are due to the high mutation rate used in this study. In all of the experiments, I set the genomic mutation rate to 0.5, that is, 0.5 mutations on average per genome per generation. With a fixed genome length of 100, the per site mutation rate

is 0.005, which is high when compared with many biological organisms. RNA viruses, known for their high mutation rates, have mutation rates in the range of 10^{-3} to 10^{-5} per nucleotide site per round of copying [22]. I chose this mutation rate for practical reasons: it is high enough for evolution to proceed at a decent pace without adding too much noise to each generation. A side effect is that the "cloud" of mutants that form a species is relatively large, leading to large intra-species angles.

Another contributing facter to high within-species angles is possibly the threshold for the clustering algorithm. The same threshold is used for all inflow rates. In some cases at intermediate inflows, a small subcluster within what has been classified as one species could be phenotypically distinct, leading to higher within-species angles. Conversely, at low inflows, the same threshold occasionally leads the algorithm to split a population into two species, when both species in fact have identical or nearly identical phenotypes. The solution to this problem would be non-universal threshold where the value depended on the inflow. Nonetheless, the all-purpose threshold still does a good job overall in partitioning populations into species.

6.3 Invasion of novel species into stable environments

The introduction of a novel species into an established ecosystem has many possible consequences. If the invading species finds the environment unhospitable and goes extinct, then the disturbance is transient. On the other hand, the invader may find its new home congenial, and spread. The result for other species in the ecosystem may be increased

competition, including niche displacement, hybridization, introgression, predation and extinction [3, 67]. The particular outcome depends on the phenotypic characteristics of the invader and the ecosystem. There are many well-known cases of introduced species becoming pests, from rabbits in Australia to zebra mussels in North America. In this section, I propose to use phenotype vectors to predict the outcome of the the introduction of novel species into well-characterized environments.

If co-evolved and stably co-existing species are well-differentiated in phenotype, resulting in nearly orthogonal phenotype vectors, then perhaps near-orthogonality supports long-term co-existence and stability, as we saw in the example of frequency-dependent selection in figure 5.6.

The long-term co-existence of two species will be determined by the competition they experience, which should be predicted by the angle their phenotypes make with each other. As we saw in figure 6.3, nearly orthogonal phenotype vectors are a characteristic of co-evolved species, whose co-existence results from frequency-dependent selection. In the event of an invasion of a novel species to an established ecosystem, the invader will co-exist with the other species, or there will be an extinction, depending on the level of competition that results from the invasion. Less competition may enable co-existence, while more competition may result in the extinction of a species.

To see whether the phenotype angles between an invader and the established ecosystem species can predict outcome, I randomly chose 50 (ordered) pairs of ecosystems from the same resource inflow, for each of the seven inflow rates. Each species is represented by its

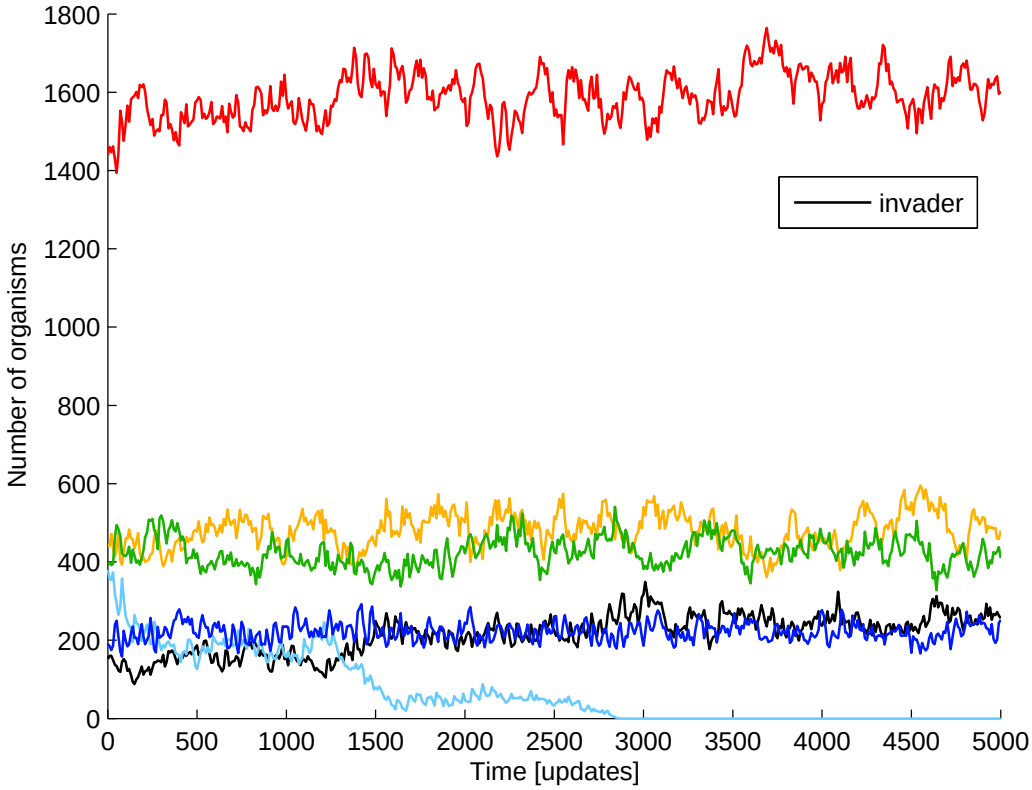


Figure 6.4: Species populations as an invader is introduced into a stable 5-species ecosystem. The invader is represented by a black line, while the various ecosystem species are represented by coloured lines. The invader replaces one of the ecosystem species (cyan).

most numerous genotype. In the first ecosystem, I introduced all species in equal numbers and let their numbers stabilize while mutations were turned off. In order to minimize the disruption to an otherwise stable ecosystem, I reduced the number of organisms in each species by 5%, and replaced the native species by invaders. The population remains at full size, and the native species are in equilibrium with each other.

Outcomes of the Avida invasion experiments included co-existence of the invader with the species in the ecosystem (that is, no extinction), extinction of a native species, or ex-

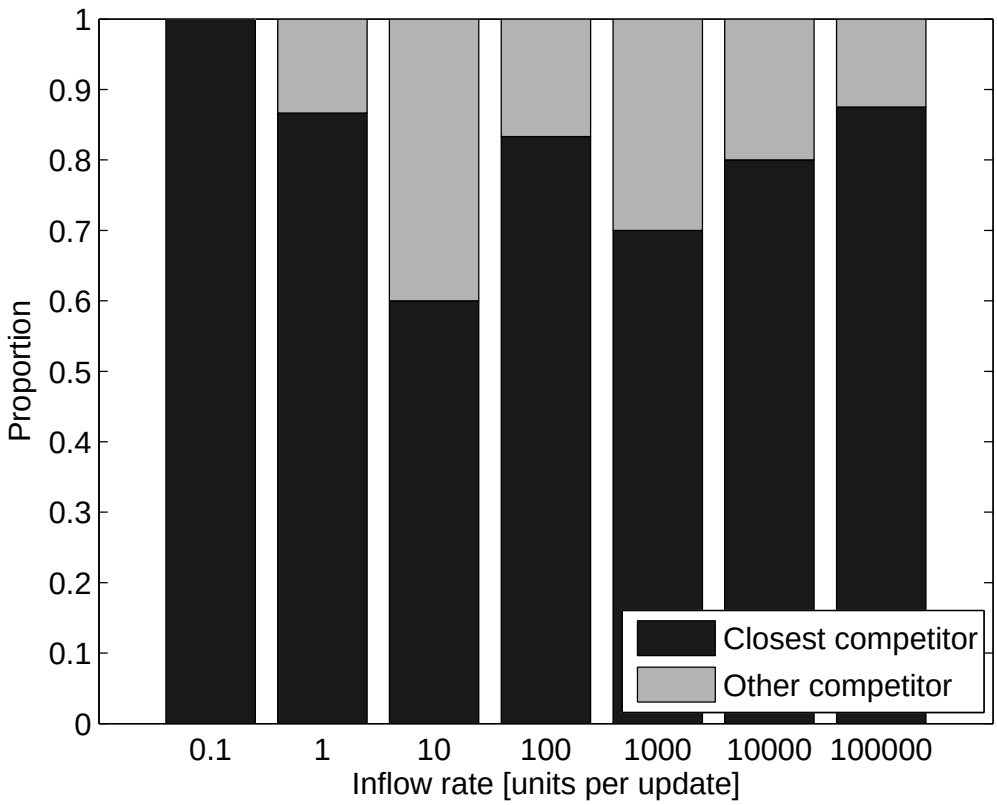


Figure 6.5: Proportion of extinctions following the invasion of a novel species that involve or do not involve the closest competitor species to an invader. The closest competitor of an invader is the species whose phenotype makes the minimum angle with the invader's phenotype.

tinction of the invader. In none of the 350 experiments did two or more species die out. An example of successful invasion and the concomitant extinction of a native species can be seen in figure 6.4.

Since the species in an ecosystem have co-evolved, then in the absence of a disturbance they will exhibit stable co-existence as a result of frequency dependent selection. When a novel species invades the ecosystem, an extinction will result if the invader competes too

strongly with a native species. I will call the ecosystem species whose phenotype makes the smallest angle with the invader phenotype, the *closest competitor* of the invader.

In the event of an unsuccessful invasion, it is difficult to claim that the extinction of the invader is due to its closest competitor and not any other. An invader can make acute angles with more than one species. On the other hand, if a native ecosystem species goes extinct, it will be due to the competition it experiences from the invader. The ecosystem species that makes the smallest angle to the invader should also be the one experiencing the most competition, and is therefore the most likely to die out.

We can see this in figure 6.5. A student's t-test of the null hypothesis that the closest competitor of an invader is not more likely to go extinct is rejected at the 5% level ($p = 7.1144 \times 10^{-4}$).

I plan to perform more experiments and refine my approach to improve the accuracy of prediction in the case of an unsuccessful invasion. A single pairwise angle does not capture all the competition that an invader may face. A predictor which is a function of all the angles that an invader faces, such as a weighted sum or a geometric mean, may capture the effect of the closest competitor while incorporating the information on less direct competition. Another alternative is to consider a measure which takes into account both the angle and magnitude of a vector, that is, projection. If two phenotype vectors are compared by examining the projection of one onto the second, then the longer vector would imply that its associated genotype had higher resource consumption which would in turn lead to increased fitness. A refinement may be to normalize vectors with gestation time,

since higher resource use may carry a cost of a longer life cycle.

Refinements and improvements notwithstanding, the simple prediction in figure 6.5 does a good job in predicting native species threatened by extinction in the event of an invasion.

Chapter 7

Conclusions and future research

7.1 Conclusions

There are approximately 1.75 million described species, and the total may be an order of magnitude more [44]. Species richness has been observed to vary among habitats, and the explanation for the differences has been called “perhaps the greatest unsolved ecological riddle” [69]. Geography, in the form of barriers to gene flow, is widely accepted to be the greatest contributor to biodiversity, reflected in the rapid speciation of allopatric populations in marine environments [75]. At smaller scales, spatial heterogeneity is required for adaptive radiation in *Pseudomonas fluorescens* in a laboratory microcosm, while *Escherichia coli* grown in a homogeneous medium exhibit adaptive radiation via cross feeding, in which metabolites produced by certain genotype serve as resources for others.

Another factor commonly thought to control diversity is productivity. Productivity is defined as “the rate at which energy flows through an ecosystem” [81]. Species richness typically grows with productivity, but sometimes declines at very high productivity levels

[47, 66, 92, 99]. Hypotheses to explain why diversity might be maximized at intermediate productivity depend on various ways on spatial heterogeneity or on predation [45, 69, 81].

In experiments with digital organisms, I have established that adaptive radiation giving rise to species can occur in the absence of spatial heterogeneity and predation. A variety of distinct phenotypes emerge when digital organisms are propagated in a homogeneous environment with multiple resources supplied in moderate abundance. Moreover, diversity in the population as a function of productivity, defined here by resource inflow to the system, reaches its maximum level at intermediate productivity and declines substantially at both lower and higher productivities. The hump-shaped pattern of species richness in digital organisms matches observations in biological organisms, which typically display an increase in diversity with productivity sometimes followed by a decrease at very high productivity levels [47, 66, 92, 99]. The decline in species richness at high productivity occurs because selection shifts from favouring exploitation of unused resources to favouring maximum replication when resources are no longer limiting.

Closer examination of the experimental data indicates that genotypes and phenotypes reflect phylogeny. A species can be considered to be a collection of genotypes which share many common genotypic and phenotypic features with each other, but differ considerably from members of other species [60]. Small phylogenetic distances correspond to small genotypic and phenotypic differences. Co-evolved digital species have distant common ancestors and exhibit profiles of resource usage which are clearly distinct. Additional experiments verified that negative frequency-dependent selection maintains digital species.

Competition favours rare types because their preferred resources are relatively unexploited.

The precise partitioning of resources varies from one replicate experiment to another, and varies in time when traced along a lineage. As species evolve, adaptation may occur via either acquisition or loss of functionality, and a new gene can come at the cost of another. Speciation is often initiated when a mutant acquires the ability to consume a previously untouched resource. Since resource usage comes at a cost, a single-species ecosystem will have unexploited resources and is therefore often not stable.

Finally, a difference measure for phenotype based on a vector representation quantifies species co-existence and predicts extinction in the event of an invasion. Phenotypes can be summarized as a list or vector of resource usage. Orthogonal vectors indicate specialization onto non-overlapping resources and adaptation to minimize competition. Co-evolved species exhibit nearly orthogonal phenotype vectors, while angles within a species are much lower. A small angle between two phenotype vectors indicates strong competition over shared resources, and is a good predictor of the extinction in an ecosystem following a successful invasion of a novel species.

7.2 Future research

7.2.1 Vector representation and speciation in biological organisms

Experiments in evolution with digital organisms give results which match many observations in biological organisms and provide evidence for the fundamental ecological princi-

ple of competitive exclusion, which is widely accepted but often difficult to test. Thus, as model organisms for experiments in evolution, digital organisms have proved their worth, not only in single niche environments [56,57,104], but in more complex ecological settings as well.

If digital organisms reflect patterns found in nature, does nature reflect patterns found in digital organisms? The vector representation of phenotype may lend itself to future experiments in biological organisms, since it provides clear predictions on the nature of co-evolved species. The well-studied plankton foraminifer *Globorotalia truncatulinoides* has a detailed fossil record and has undergone periods of rapid evolutionary change and speciation [55]. If the resource limitations of the different forms could be ascertained, then numerical phenotype representations can be extracted, and the angles between species can be compared to their long-run survival outcomes.

7.2.2 Natural extinction

Within Avida, the vector approach used successfully to predict ecosystem invasion outcomes may also lead to a better understanding of the related issue of natural extinction, where a lineage ends as a result of competition with other lineages and their effects on the environment.

I predict that lineages die out due to increased competition as measured by phenotype angle. A mutant in one lineage may affect not only the course of its own lineage, but also those of others as its phenotype changes. To study this in Avida will be similar to the inva-

sion experiments, although it will require a careful survey of more detailed historical data to find lineages that die out. When such lineage is found, the vector approach can be used to compare its phenotype to that of the other, surviving lineages in the time immediately preceding the extinction, with the outcome checked using stable co-existence tests with mutations turned off (as was done to verify frequency-dependent selection).

Chapter 8

Appendix

8.1 Alternate measures of distance between two genotypes

The clustering algorithm used to determine species bases its decisions on a measure of distance between two genotypes. For the algorithm's classifications to be useful, the measure of distance should correlate well to the relatedness of the two genotypes. Aside from the phylogenetic distance, there are some other natural distance measures.

One measure of distance between two strings, such as the string of instructions defining a genotype, is Hamming distance. Hamming distance counts the number of mismatches between a pair of strings of the same length, that is, the number of point mutations required to transform the first genotype into the second. When applied to digital organisms, Hamming distance tends to strongly underestimate the number of species when calibrated in the same way as phylogenetic distance. The size of change in genotype does not correspond well to the size of change in phenotype, so Hamming distance disregards the phenotypic effects of a mutation.

Levenshtein distance, a more sophisticated version of genetic distance, is not required since genomes were constrained to have identical lengths.

A third measure is a variation on phylogenetic distance. Instead of counting along mutants in the phylogenetic tree from on, the time to the most recent common ancestor (MRCA) relies on the fact that species are well-separated in time as well as genetic distance. If mutations are incorporated into the genome at a steady rate, then time to Unfortunately, time to MRCA tends to overestimate the the number of species. The assumption that mutations are incorporated at a constant rate, which would make time to MRCA equivalent to phylogenetic distance, is false. As we saw easily in the single niche environments of figures 4.1 and 4.2, periods of rapid evolution do not alternate with periods of stasis at a steady rate.

8.2 The default instruction set

The following descriptions are taken from the Avida documentation.

Type: The category of all subsequent keywords.

Keyword: The lookup term for the particular piece of information. Each keyword command starts a new entry

Desc: The description for this entry.

Alias: Other lookup terms that might be used for the most recently defined alias. (may have multiple alias lines)

Type: Instruction

Keyword: nop-instructions

Desc: The instructions nop-A, nop-B, and nop-C are no-operation instructions, and will not do anything when executed. They will, however, modifiy the behavior of the instruction preceeding it (by changing the CPU component that it affects; see also nop-register notation and nop-head notation) or act as part of a template to denote positions in the genome. Alias: nop-a nop-b nop-c no-operation

Keyword: IO

Desc: This is the input/output instruction. It takes the contents of the BX register and outputs it, checking it for any tasks that may have been performed. It will then place a new input into BX. Alias: io

Keyword: add

Desc: This instruction reads in the contents of the BX and CX registers and sums them together. The result of this operation is then placed in the BX register.

Keyword: dec

Desc: This instruction reads in the contents of the BX register and decrements it by one.

Keyword: h-alloc

Desc: This instruction allocates additional memory for the organism up to the maximum it is allowed to use for its offspring.

Keyword: h-copy

Desc: This instruction reads the contents of the organism's memory at the position of the read-head, and copy that to the position of the write-head. If a non-zero COPY_MUTATION_PROB

is set, a test will be made based on this probability to determine if a mutation occurs. If so, a random instruction (chosen from the full set with equal probability) will be placed at the write-head instead.

Keyword: h-divide

Desc: This instruction is used for an organism to divide off an finnished offspring. The original organism keeps the state of its memory up until the read-head. The offspring's memory is initialized to everything between the read-head and the write-head. All memory past the write-head is removed entirely.

Keyword: h-search

Desc: This instruction will read in the template the follows it, and find the location of a complement template in the code. The BX register will be set to the distance to the complement from the current position of the instruction-pointer, and the CX register will be set to the size of the template. The flow-head will also be placed at the beginning of the complement template. If no template follows, both BX and CX will be set to zero, and the flow-head will be placed on the instruction immediatly following the h-search.

Keyword: if-label

Desc: This instruction reads in the template that follows it, and tests if its complement template was the most recent series of instructions copied. If so, it executed the next instruction, otherwise it skips it. This instruction is commonly used for an organism to determine when it has finished producing its offspring.

Keyword: if-less

Desc: This instruction compares the BX register to its complement. If BX is the lesser of the pair, the next instruction (after a modifying no-operation instruction, if one is present) is executed. If it is greater or equal, then that next instruction is skipped.

Keyword: if-n-equ

Desc: This instruction compares the BX register to its complement. If they are not equal, the next instruction (after a modifying no-operation instruction, if one is present) is executed. If they are equal, that next instruction is skipped.

Keyword: inc

Desc: This instruction reads in the contents of the BX register and increments it by one

Keyword: jmp-head

Desc: This instruction will read in the value of the CX register, and the move the instruction pointer (IP) by that fixed amount through the organism's memory.

Keyword: mov-head

Desc: This instruction will cause the IP to jump to the position in memory of the flow-head.

Keyword: nand

Desc: This instruction reads in the contents of the BX and CX registers (each of which are 32-bit numbers) and performs a bitwise nand operation on them. The result of this operation is placed in the BX register. Note that this is the only logic operation provided in the basic avida instruction set.

Keyword: pop

Desc: This instruction removes the top element from the active stack, and places it into the BX register.

Keyword: push

Desc: This instruction reads in the contents of the BX register, and places it as a new entry at the top of the active stack. The BX register itself remains unchanged.

Keyword: set-flow

Desc: This instruction moves the flow-head to the memory position denoted in the CX? register.

Keyword: shift-l

Desc: This instruction reads in the contents of the BX register, and shifts all of the bits in that register to the left by one, placing a zero as the new rightmost bit, and trunkating any bits beyond the 32 maximum. For values that require fewer than 32 bits, it effectively multiplies that value by two.

Keyword: shift-r

Desc: This instruction reads in the contents of the BX register, and shifts all of the bits in that register to the right by one. In effect, it divides the value stored in the register by two, rounding down.

Keyword: sub

Desc: This instruction reads in the contents of the BX and CX registers and subtracts CX from BX. The result of this operation is then placed in the BX register.

Keyword: swap-stk Desc: This instruction toggles the active stack.

Keyword: swap Desc: This instruction swaps the contents of the BX register with its complement.

Bibliography

[1] P.A. Abrams. Monotonic or unimodal diversity productivity gradients - what does competition theory predict. *Ecology*, 76(7):2019–2027, 1995.

[2] Z. Abramsky and M.L. Rosenzweig. Tilman predicted productivity diversity relationship shown by desert rodents. *Nature*, 309(5964):150–151, 1984.

[3] J.E. Byers. Competition between two estuarine snails: Implications for invasions of exotic species. *Ecology*, 81(5):1225–1239, 2000.

[4] G. Caldarelli, P.G. Higgs, and A.J. McKane. Modelling coevolution in multispecies communities. *J. Theor. Biol.*, 193(2):345–358, 1998.

[5] S.S. Chow, C.O. Wilke, C. Ofria, R.E. Lenski, and C. Adami. Supplementary material to [6]. Available at `http://dllab.caltech.edu/pubs/science04/`.

[6] S.S. Chow, C.O. Wilke, C. Ofria, R.E. Lenski, and C. Adami. Adaptive radiation from resource competition in digital organisms. *Science*, 305(5680):84–86, 2004.

[7] F.M. Cohan. Bacterial species and speciation. *Systematic Biology*, 50(4):513–524, 2001.

[8] F.M. Cohan. What are bacterial species? *Annu. Rev. Microbiology*, 56:457–487, 2002.

[9] V.S. Cooper and R.E. Lenski. The population genetics of ecological specialization in evolving *Escherichia coli* populations. *Nature*, 407(6805):736–739, 2000.

[10] J.A. Coyne. Ernst Mayr and the origin of species. *Evolution*, 48(1):19–30, 1994.

[11] J.A. Coyne and N.H. Barton. What do we know about speciation? *Nature*, 331:485–486, 1988.

[12] J. Cracraft. Speciation and its ontology: The empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In D. Otte and J.A. Endler, editors, *Speciation and Its Consequences*, pages 28–59. Sinauer, 1989.

[13] C. Darwin. *On the Origin of Species by Means of Natural Selection*. John Murray, 1859.

[14] E.K. Davies, A.D. Peters, and P.D. Keightley. High Frequency of Cryptic Deleterious Mutations in *Caenorhabditis elegans*. *Science*, 285(5434):1748–1751, 1999.

[15] T. Day. Competition and the effect of spatial resource heterogeneity on evolutionary diversification. *American Naturalist*, 155(6):790–803, 2000.

[16] T. de Meeus and J. Goudet. Adaptive diversity in heterogeneous environments for populations regulated by a mixture of soft and hard selection. *Evolutionary Ecology Research*, 2(8):981–995, 2000.

[17] K. de Queiroz and M.J. Donoghue. Phylogenetic systematics or Nelsons version of cladistics. *Cladistics*, 6(1):61–75, 1990.

[18] U. Dieckmann and M. Doebeli. On the origin of species by sympatric speciation. *Nature*, 400(6742):354–357, 1999.

[19] T. Dobzhansky. A criticque of the species concept in biology. *Philosophy of Science*, 2:344–355, 1935.

[20] M. Doebeli. A quantitative genetic competition model for sympatric speciation. *J. Evol. Biol.*, 9(6):893–909, 1996.

[21] M. Doebeli and U. Dieckmann. Evolutionary branching and sympatric speciation caused by different types of ecological interactions. *American Naturalist*, 156(S):S77–S101, 2000.

[22] E. Domingo and J.J. Holland. RNA virus mutations and fitness for survival. *Annu. Rev. Microbiology*, 51:151–178, 1997.

[23] M.J. Donoghue. A critique of the biological species concept and recommendations for a phylogenetic alternative. *Bryologist*, 88(3):172–181, 1985.

[24] J.W. Drake, B. Charlesworth, D. Charlesworth, and J.F. Crow. Rates of spontaneous mutation. *Genetics*, 148:1667–1686, 1998.

[25] E. Drenkard and F.M. Ausubel. *Pseudomonas* biofilm formation and antibiotic resistance are linked to phenotypic variation. *Nature*, 416(6882):740–743, 2002.

[26] K.E. Duncan, N. Ferguson, K. Kimura, X. Zhou, and C.A. Istock. Fine-scale genetic and phenotypic structure in natural-populations of *Bacillus-subtilis* and *Bacillus-licheniformis*: implications for bacterial evolution and speciation. *Evolution*, 48(6):2002–2025, 1994.

[27] D.E. Dykhuizen. Santa Rosalia revisited: Why are there so many species of bacteria? *Antonie van Leeuwenhoek International Journal of General and Molecular Microbiology*, 73(1):25–33, 1998.

[28] D.E. Dykhuizen and L. Green. Recombination in *Escherichia coli* and the definition of biological species. *J. Bacteriology*, 173(22):7257–7268, 1991.

[29] M. Egas, U. Dieckmann, and M.W. Sabelis. Evolution restricts the coexistence of specialists and generalists: The role of trade-off structure. *American Naturalist*, 163(4):518–531, 2004.

[30] S.F. Elena, V.S. Cooper, and R.E. Lenski. Punctuated evolution caused by selection of rare beneficial mutations. *Science*, 272(5269):1802–1804, 1996.

[31] S.F. Elena and R.E. Lenski. Test of synergistic interactions among deleterious mutations in bacteria. *Nature*, 390(6658):395–398, 1997.

[32] S.F. Elena and R.E. Lenski. Evolution experiments with microorganisms: The dynamics and genetic bases of adaptation. *Nat. Rev. Gen.*, 4(6):457–469, 2003.

[33] E.J. Feil, E.C. Holmes, D.E. Bessen, M.S. Chan, N.P.J. Day, M.C. Enright, R. Goldstein, D.W. Hood, A. Kalla, C.E. Moore, J.J. Zhou, and B.G. Spratt. Recombination within natural populations of pathogenic bacteria: Short-term empirical estimates and long-term phylogenetic consequences. *PNAS*, 98(1):182–187, 2001.

[34] M.R. Fisk, S.J. Giovannoni, and I.H. Thorseth. Alteration of oceanic volcanic glass: Textural evidence of microbial activity. *Science*, 281(5379):978–980, 1998.

[35] R. Froissart, D. Roze, M. Uzest, L. Galibert, S. Blanc, and Y. Michalakis. Recombination every day: Abundant recombination in a virus during a single multi-cellular host infection. *PLOS Biology*, 3(3):389–395, 2005.

[36] T. Fukami and P.J. Morin. Productivity-biodiversity relationships depend on the history of community assembly. *Nature*, 424(6947):423–426, 2003.

[37] S. Garcia-Vallve, A. Romeu, and J. Palau. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Research*, 10(11):1719–1725, 2000.

[38] S. Gavrilets. Rapid evolution of reproductive barriers driven by sexual conflict. *Nature*, 403(6772):886–889, 2000.

[39] P.J. Gerrish and R.E. Lenski. The fate of competing beneficial mutations in an asexual population. *Genetica*, 103:127–144, 1998.

[40] D.S. Glazier. Towards a predictive theory of speciation: The ecology of isolate selection. *J. Theor. Biol.*, 126:323–333, 1987.

[41] S.J. Gould and N. Eldredge. Punctuated equilibrium comes of age. *Nature*, 366:223–227, 1993.

[42] T. Hatfield and D. Schluter. Ecological speciation in sticklebacks: Environment-dependent hybrid fitness. *Evolution*, 53(3):866–873, 1999.

[43] R.B. Helling, C.N. Vargas, and J. Adams. Evolution of *Escherichia-coli* during growth in a constant environment. *Genetics*, 116(3):349–358, 1987.

[44] V.H. Heywood, editor. *Global Biodiversity Assessment*. United Nations Environment Programme. Cambridge Univ. Press, 1995.

[45] R.D. Holt, J. Grover, and D. Tilman. Simple rules for interspecific dominance in systems with exploitative and apparent competition. *American Naturalist*, 144(5):741–771, 1994.

[46] M.C. Horner-Devine, M.A. Leibold, V.H. Smith, and B.J.M. Bohannan. Bacterial diversity patterns along a gradient of primary productivity. *Ecology Letters*, 6(7):613–622, 2003.

[47] R. Kassen, A. Buckling, G. Bell, and P.B. Rainey. Diversity peaks at intermediate productivity in a laboratory microcosm. *Nature*, 406(6795):508–512, 2000.

[48] R. Kassen, M. Llewellyn, and P.B. Rainey. Ecological constraints on diversification in a model adaptive radiation. *Nature*, 431:984–988, 2004.

[49] M. Kawata. Invasion of vacant niches and subsequent sympatric speciation. *Proc. R. Soc. London B.*, 269(1486):55–63, 2002.

[50] A.S. Kondrashov and F.A. Kondrashov. Interactions among quantitative traits in the course of sympatric speciation. *Nature*, 400(6742):351–354, 1999.

[51] F.A. Kondrashov and A.S. Kondrashov. Multidimensional epistasis and the disadvantage of sex. *PNAS*, 98(21):12089–12092, 2001.

[52] D.J. Kornet. Permanent splits as speciation events - a formal reconstruction of the internodal species concept. *J. Theor. Bio.*, 164(4):407–435, 1993.

[53] P. Kral. Species orthogonalization. *J. Theor. Biol.*, 212(3):355–366, 2001.

[54] J.G. Lawrence. Catalyzing bacterial speciation: Correlating lateral transfer with genetic headroom. *Systematic Biology*, 50(4):479–496, 2001.

[55] D. Lazarus, H. Hilbrecht, C. Spencer-Cervato, and H. Thierstein. Sympatric speciation and phyletic change in *Globorotalia truncatulinoides*. *Paleobiology*, 21(1):28–51, 1995.

[56] R.E. Lenski, C. Ofria, T.C. Collier, and C. Adami. Genome complexity, robustness and genetic interactions in digital organisms. *Nature*, 400(6745):661–664, 1999.

[57] R.E. Lenski, C. Ofria, R.T. Pennock, and C. Adami. The evolutionary origin of complex features. *Nature*, 423(6936):139–144, 2003.

[58] R.E. Lenski and M. Travisano. Dynamics of adaptation and diversification – A 10,000-generation experiment with bacterial populations. *PNAS*, 91(15):6808–6814, 1994.

[59] M. Macnamara and H.E.H. Paterson. The recognition concept of species. *South African Journal of Science*, 80(7):312–318, 1984.

[60] J. Mallet. A species definition for the modern synthesis. *Trends in Ecology & Evolution*, 10(7):294–299, 1995.

[61] T.E. Martin. Are microhabitat preferences of coexisting species under selection and adaptive? *Ecology*, 79(2):656–670, 1998.

[62] R.M. May. How many species? *Phil. Trans. R. Soc. Lond. B*, 330:293–304, 1990.

[63] E. Mayr. *Systematics and the origin of species*. Columbia Univ. Press, 1942.

[64] J. McFadden and G. Knowles. Escape from evolutionary stasis by transposon- mediated deleterious mutations. *J. Theor. Biol.*, 186(4):441–447, 1997.

[65] A. Mira, H. Ochman, and N.A. Moran. Deletional bias and the evolution of bacterial genomes. *Trends in Genetics*, 17(10):589–596, 2001.

[66] G.G. Mittelbach, C.F. Steiner, S.M. Scheiner, K.L. Gross, H.L. Reynolds, R.B. Waide, M.R. Willig, S.I. Dodson, and L. Gough. What is the observed relationship between species richness and productivity? *Ecology*, 82(9):2381–2396, 2001.

[67] H.A. Mooney and E.E. Cleland. The evolutionary impact of invasive species. *PNAS*, 98(10):5446–5451, 2001.

[68] E. Moreno. In search of a bacterial species definition. *Revista de Biologia Tropical*, 45(2):753–771, 1997.

[69] P.J. Morin. Biodiversity's ups and downs. *Nature*, 406(6795):463–464, 2000.

[70] H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, 2000.

[71] H. Ochman and N.A. Moran. Genes lost and genes found: Evolution of bacterial pathogenesis and symbiosis. *Science*, 292(5519):1096–1098, 2001.

[72] C. Ofria and C.O. Wilke. Avida: A software platform for research in computational evolutionary biology. *Artificial Life*, 10(2):191–229, 2004.

[73] H.A. Orr. The rate of adaptation in asexuals. *Genetics*, 155(2):961–968, 2000.

[74] M. Pagel. Inferring the historical patterns of biological evolution. *Nature*, 401(6756):877–884, 1999.

[75] S.R. Palumbi. Genetic divergence, reproductive isolation, and marine speciation. *Annu. Rev. Ecol. and Systematics*, 25:547–572, 1994.

[76] D. Papadopoulos, D. Schneider, J. Meier-Eiss, W. Arber, R.E. Lenski, and M. Blot. Genomic evolution during a 10,000-generation experiment with bacteria. *PNAS*, 96(7):3807–3812, 1999.

[77] P.B. Rainey, A. Buckling, R. Kassen, and M. Travisano. The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends in Ecology & Evolution*, 15(6):243–247, 2000.

[78] P.B. Rainey and M. Travisano. Adaptive radiation in a heterogeneous environment. *Nature*, 394(6688):69–72, 1998.

[79] B.W. Robinson and D.S. Wilson. Optimal foraging, specialization, and a solution to Liem's paradox. *American Naturalist*, 151(3):223–235, 1998.

[80] M.L. Rosenzweig. *Species Diversity in Space and Time*. Cambridge Univ. Press, Cambridge, 1995.

[81] M.L.. Rosenzweig and Z. Abramsky. *Species Diversity in Ecological Communities*, pages 52–56. Univ. Chicago Press, 1993.

[82] R.F. Rosenzweig, R.R. Sharp, D.S. Treves, and J. Adams. Microbial evolution in a simple unstructured environment – genetic differentiation in *Escherichia-coli*. *Genetics*, 137(4):903–917, 1994.

[83] D.E. Rozen and R.E. Lenski. Long-term experimental evolution in escherichia coli. viii. dynamics of a balanced polymorphism. *American Naturalist*, 155(1):24–35, 2000.

[84] H.D. Rundle, L. Nagel, J.W. Boughman, and D. Schluter. Natural selection and parallel speciation in sympatric sticklebacks. *Science*, 287(5451):306–308, 2000.

[85] D. Schluter. Ecological causes of adaptive radiation. *American Naturalist*, 148(S):S40–S64, 1996.

[86] D. Schluter. Ecology and the origin of species. *Trends in Ecology & Evolution*, 16(7):372–380, 2001.

[87] S.A. Shabalina, L.Y. Yampolsky, and A.S. Kondrashov. Rapid decline of fitness in panmictic populations of *Drosophila melanogaster* maintained under relaxed naturalselection. *PNAS*, 94(24):13034–13039, 1997.

[88] M.J. Siegert, J.C. Ellis-Evans, M. Tranter, C. Mayer, J. Petit, A. Salamatin, and J.C. Priscu. Physical, chemical and biological processes in Lake Vostok and other Antartic subglacial lakes. *Nature*, 414(6864):603–609, 2001.

[89] R.R. Sokal and T.J. Crovello. Biological species concept - a critical evaluation. *American Naturalist*, 104(936):127, 1970.

[90] P.S. Stewart and J.W. Costerton. Antibiotic resistance of bacteria in biofilms. *Lancet*, 358(9726):135–138, 2001.

[91] A.R. Templeton. Using phylogeographic analyses of gene trees to test species status and processes. *Mol. Ecol.*, 10(3):779–791, 2001.

[92] D. Tilman and S. Pacala. *Species Diversity in Ecological Communities*, pages 13–25. Univ. Chicago Press, 1993.

[93] M. Travisano, J.A. Mongold, A.F. Bennett, and R.E. Lenski. Experimental tests of the roles of adaptation, chance, and history in evolution. *Science*, 267(5194):87–90, 1995.

[94] M. Travisano and P.B. Rainey. Studies of adaptive radiation using model microbial systems. *American Naturalist*, 156(Suppl. S):S35–S44, 2000.

[95] D.S. Treves, S. Manning, and J. Adams. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of escherichia coli. *Molecular Biology and Evolution*, 15(7):789–797, 1998.

[96] M. Turelli, N. Barton, and J. Coyne. Theory and speciation. *Trends in Ecology & Evolution*, 16(7):330–343, 2001.

[97] P.E. Turner, V. Souza, and R.E. Lenski. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology*, 77(7):2119–2129, 1996.

[98] L. Vanvalen. Ecological species, multispecies, and oaks. *Taxon*, 25(2–3):233–239, 1976.

[99] R.B. Waide, M.R. Willig, C.F. Steiner, G. Mittelbach, L. Gough, S.I. Dodson, G.P. Juday, and R. Parmenter. The relationship between productivity and species richness. *Annual Review of Ecology and Systematics*, 30:257–300, 1999.

[100] D.M. Walling, S.N. Edmiston, J.W. Sixbey, M. Abdelhamid, L. Resnick, and N. Raabtraub. Coinfection with multiple strains of the epstein-barr virus in human immunodeficiency virus-associated hairy leukoplakia. *PNAS*, 89(14):6560–6564, 1992.

[101] E.O. Wiley. Evolutionary species concept reconsidered. *Systematic Zoology*, 27(1):17–26, 1978.

[102] C.O. Wilke and C. Adami. The biology of digital organisms. *Trends in Ecology & Evolution*, 17(11):528–532, 2002.

[103] C.O. Wilke and S.S. Chow. Exploring the evolution of ecosystems with digital organisms. In M. Pascual and J. Dunne, editors, *Food Webs as Complex Adaptive Networks: Linking Structure to Dynamics*. Oxford Univ. Press, in press.

[104] C.O. Wilke, J.L. Wang, , C. Ofria, R.E. Lenski, and C. Adami. Evolution of digital organisms at high mutation rates leads to survival of the flattest. *Nature*, 412(6844):331–333, 2001.

[105] Y.I. Wolf, A.S. Kondrashov, and E.V. Koonin. Interkingdom gene fusions. *Genome Biology*, 1(6):research 0013.1–0013.13, 2000.

[106] Image: Eastern grey kangaroo. Available at `http://en.wikipedia.org/wiki/Eastern_Grey_Kangaroo`.

[107] Image: White-tailed deer. Available at `http://en.wikipedia.org/wiki/Image:White-tailed_deer.jpg`.