Chapter 2

The Model

A model is a tool for theoretical exploration as well as a guide for experimental work (Wimsatt 1987). It is an active form of a theory about how nature works. I argued in Section 1.4.2 that the application of configuration models to biological questions requires us to first specify the components of a model that are required in order to examine a hypothesis. Then we may construct a model so as to include only those components that are necessary to implement the hypothesis. This is the principle of parsimony as applied to modeling. We may argue later about the importance of missing complexities in the model, in light of the results. And, depending on the outcome of those arguments, we may want to elaborate the model. However, for now, I will try to keep the model as simple as possible.

2.1 Requirements

The purpose of the model is to examine the macroevolutionary dynamics of diversification through the microevolutionary dynamics of organisms. What aspects of biology must be represented in the model?

2.1.1 Microevolution

Requirement 1 *Organisms.* The components of evolution on an organismic and genic level are well understood. For natural selection to work, you need something to select. In this case, we will focus on organisms.

Requirement 2 *Reproduction.* Natural selection is based on differential reproduction. Some organisms must reproduce more than others. This may be caused by differential survival, but it need not be. Natural selection may still occur if no organisms die, as long as some produce more offspring than others.

Requirement 3 *Heredity.* If there is no relationship between a parent's differential reproduction and its offspring's reproduction, if parents cannot pass on the secrets of their success to their children, then natural selection will not be able to *select* anything.

In biology, the correlation between a parent's success and its offspring's success is mediated through the gift of genes. These genes must have some influence over the differences between organisms' reproductive success. Organisms (or any unit of selection), differential reproduction, and heredity are all that are really needed for natural selection to kick in. It is kind of surprising that "death" has not yet made the list. Modern biologists tend to define evolution as the change in proportions of alleles, or different genes, in a population. Given a potentially infinite population, differential reproduction is sufficient to produce changes in the proportions of alleles in the population.

Requirement 4 *Mutation.* Besides altering the proportion of alleles in a population, mutation is the engine of change in evolution. It is the ultimate source of novelty, and as such, it is a necessary ingredient for testing hypotheses that focus on evolutionary innovation as the driving force behind diversification (Benton 1990).

All the previous requirements establish a framework simply to allow evolution to play out in the model. However, a number of the hypotheses for diversification make explicit reference to ecological interactions (Vermeij 1987; Van Valen 1984).

Requirement 5 *Competition.* Competition is the darling of most diversity arguments (Raup 1991; Flessa et al. 1986). It is thought to be the causal agent behind most extinctions at high diversities (Ricklefs & Schluter 1993; Cornell 1993). Competition also defines the boundaries of viable niches that colonizers may hope to fill. Competition is a sort of "horizontal" interaction within a single trophic level of a food web.

Requirement 6 *Predation.* Predation characterizes most interactions between different trophic levels of a food web. Predation is probably also important in shaping diversity trends (Vermeij 1987). Most discussions of diversification revolves around the concept of niche space (Benton 1990; Valentine & Walker 1986; Walker 1985, see Chapter 6 for a full discussion). If we adopt a simplified view of niches as the set of resources upon which a species depends, then the inclusion of predation, the consumption of resources, introduces niches to the model. However, the boundaries of a niche, the constraints on predation, are central to most of these discussions.

Requirement 7 *Specialization.* Some organisms specialize on a small number of food sources. Koala bears (*Phascolarctos cinerus*) are notorious for subsisting on only the leaves of Eucalyptus trees (genus *Eucalyptus*). Other organisms are generalists and can eat a wide variety of foods. Humans (*Homo sapiens sapiens*) are notorious for being willing and able to eat just about anything. After evolutionary innovations that open up new niches for colonization, Benton (1990) argues that specialization has been the second most important factor in diversification.

2.1.2 Macroevolution

Requirement 8 *Species.* We are fundamentally interested in macroevolutionary dynamics. The dynamics of species, speciation, and extinction. This requires that we be able to group organisms into species. If we are to use the reproductive concept of species, then two further requirements follow.

Requirement 9 *Sexual reproduction.* The reproductive species concept only applies to organisms that reproduce sexually.

Requirement 10 *Reproductive barriers.* Furthermore, the characteristics that prevent individuals from successfully mating must themselves evolve. These are called reproductive barriers. Putting this together with the microevolutionary dynamics, the evolution of reproductive barriers implies that there must be some genes that control aspects of those barriers.

Requirement 11 *Speciation.* We are not simply interested in species. We are interested in the creation of new species by the splitting of previous ones. So there must be some representation of speciation.

Requirement 12 *Death.* If we assume that parents and their offspring will be similar enough that they will not be separated by the reproductive barriers, then it is impossible for a gene pool to split apart without death.

It is also nice (as a model designer) to include death in a model because computers have difficulty representing infinite populations. Once a finite population capacity is filled, all evolutionary change will stop without death. This is not to say that death is adaptive, or evolved for the good of the population. It is simply to say that a model would not be very interesting if it froze once the population had hit the carrying capacity. These design issues aside, death is necessary to split apart gene pools and so allow speciation.

Requirement 13 *Geography.* The best supported theory for speciation suggests that geographic isolation is an essential factor in the process. We must therefore represent geography in order to examine allopatric speciation.

Requirement 14 *Migration.* Geography allows small populations to be physically isolated from other organisms. This isolation typically begins with the migration of a small number of colonists to a new location. Migration has the effect of connecting geographically separated gene pools and so plays an important role in speciation.

Requirement 15 *Habitat.* Most of the abiotic hypotheses for diversification are based on fragmentation of populations by the geography as well as differential selective pressures on the subpopulations due to their exposure to different climates or habitats (Vermeij 1987; Cracraft 1985; Benton 1990). The only missing ingredient left, then, is habitat.

2.2 An Instantiation

There are many potential instantiations of the requirements to model diversification. Furthermore, the details of an instantiation may have dramatic effects upon the results of a model. The results of a single model cannot be claimed to be the final judgment on the validity of a hypothesis. Rather, robust results from models must come through the agreement of parallel and independent efforts with alternative implementations of the critical details. The implementation that follows, dubbed "a model of diversification" or MoD, may not be the cleanest or the best. However, it draws on the common techniques in artificial life for representing individuals, their interactions, and evolution (Ray 1992; Holland 1992;

Maley 1997). All italicized terms below are parameters to the model. Their experimental settings are generally parenthesized.

Both time and space are discrete in the model. Time moves in "time steps." During each time step, all the organisms have a chance to eat, die, and possibly mate. The details follow.

2.2.1 Microevolution

Microevolution rests on the requirements of implementing organisms, reproduction, heredity and mutation. An organism is represented by: a set of 4 chromosomes, a tally of the number of other organisms it has managed to eat since the last time it reproduced, and a tally of the number of predators that have attempted to eat it during the current time step. All of these aspects have an ecological function which leads to the differential reproduction of the organisms, described in Section 2.2.2. Each "chromosome" is represented as a string of 32 bits, or "loci" in the language of genetics. The loci of a chromosome[1] represent the set of characters that influence a particular aspect of the life history of the organism. Rather than representing a number, these bits are just treated as a pattern of 1's and 0's. The organisms are haploid, in that they only have one copy of each chromosome, not two, like most animals.

Once an organism has managed to collect a number of meals equal to the *energy-conversion* (3) parameter, it garners a chance to reproduce[2]. Setting *energy-conversion* to 3 enforces a 3 to 1 conversion of prey matter into predator matter. After locating a mate[3] in the organism's patch, the chromosomes for their offspring are constructed by recombination from the chromosomes of the parents. I have used two-point crossover to implement recombination, as shown in Figure 2-1. For each pair of chromosomes from the two parents, copies of the bit patterns are lined up and then two points are randomly chosen along the sequence of bits. The subsequence between the two points is exchanged between the two parental chromosomes, and then one of the two new "mosaic" results is randomly chosen to be the chromosome for the offspring. This is repeated for each of the four chromosomes.

Choosing to cut the chromosomes at two points avoids a bias inherent in one-point

[1]Most artificial life models would call these chromosomes "genes."
[2]Eating another organism provides the diner with one meal or energy unit.
[3]The mate need not have collected 3 meals or energy units.

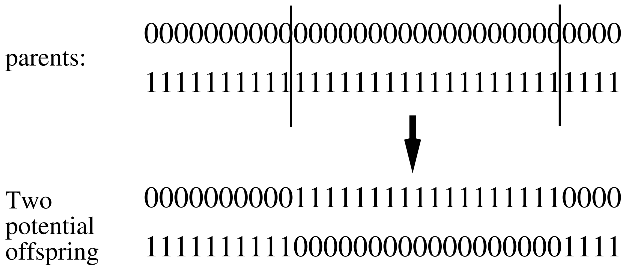

Figure 2-1: Two-point crossover.

crossover. Under one-point crossover, when the chromosomes are only cut once, the bits at either end of the chromosome are almost certainly going to be separated by recombination. However, two-point crossover, when implemented correctly, is equivalent to forming the bit strings into circles and then exchanging equivalent sections of the circles between parental chromosomes, as depicted in Figure 2-2. The chance that two bits will be separated by recombination is proportional to their distance from one another on the circle. All bit positions are thus treated equivalently.

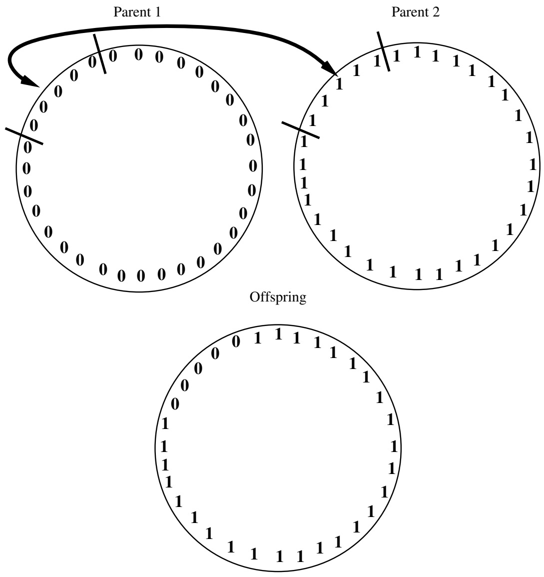

Figure 2-2: Two-point crossover on strings is equivalent to recombination between strings formed into circles. This makes it clear that there are no biases in the crossover based on the position of a bit in the string.

After we have determined the chromosomes of the new organism through recombination of the parental chromosomes, the new chromosomes are exposed to point mutations. A point mutation is simply the flipping of a single bit in a chromosome. The *mutation-rate* parameter (0.02) sets the average rate of bits flipped per chromosome per reproductive event. The number of bits to be flipped was determined by a Poisson process[4] Then a particular bit was chosen to be flipped by generating a random number[5]. between 1 and 32 with equal probability.

2.2.2 Ecology

The required ecological interactions include competition, predation, and the elaboration of predation by specialization or generalism. In other words, the interactions necessary to set up a basic food web.

The ecological interactions were implemented through 3 of the 4 "chromosomes." "Chromosomes" are metaphors here for phenotypic characters that are perfectly heritable. Thus, there is no distinction between the genotype and phenotype of an organism. This avoids the complexity, and our vast ignorance, of the process of development by which an organisms genes and the cell contents of the zygote create the full blown, multicellular organism. While this is a blatant simplification of the processes in evolution, the hypotheses for diversification do not make particular reference to the genotype-phenotype distinction. So, until the time when validation of the results may require us to include it, we are obligated to elide[6] the process of development. I will alternatively refer to a "chromosome" or a set of "characters" as equivalent aspects of the model.

Organisms have chromosomes for "predation resistance" characters, a preferred "prey template," and "generalism." These correspond to the three genes that Lindgren and Nordahl (1994) use to set up their ecosystem. An organism's prey template bit pattern encodes the bit pattern of the predation resistance chromosome in the prey organisms that it prefers to eat. On top of this, the 1's in an organism's generalism chromosome specify which bit

[4]Poisson processes are a form of stochastic processes that are typically used to model the random occurrence of discrete events over time. If the average rate of events is λ per time period, then the probability of observing k events over t time periods of a Poisson process is

$$p(k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!} \tag{2.1}$$

We can generate a random number with a poisson distribution by first generating a random number (the probability) between 0 and 1 with uniform probability. Then we repeatedly subtract $p(k)$ for $k = 0, 1, 2, \ldots$ until the result is negative. The value of k when the result first becomes negative is the poisson random number.

[5]I have used a modified version of Knuth's (1981, pp. 171–172) subtractive method for generating pseudo-random numbers coded by Terry Jones.

[6]In computer science slang, elide = "punt," as in, "I will punt on the issue of biological development."

positions do not matter in its prey template chromosome when it tries to consume prey. In other words, the generalism chromosome creates "wild-card" positions in the organism's prey template chromosome. I will refer to the predator's prey template chromosome, modified by wild-cards from the predator's generalism chromosome, as a "search pattern." The predation resistance chromosome represents characters relevant to the sufferance of predation. This might include both behavioral and structural characteristics of the organism that help it to avoid predation. Figure 2-3 diagrams an 8-bit example (the real chromosomes are 32 bits) of a match between a predator and a prey organism as well as a mismatch that prevents the predator from capturing and consuming the prey.

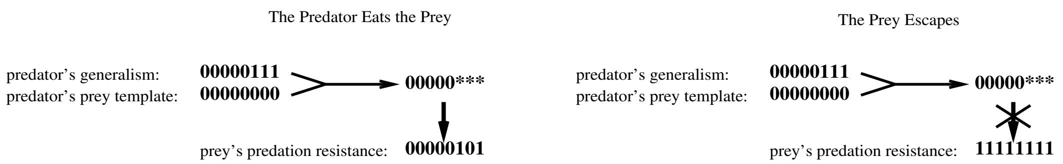

Figure 2-3: Determination of a predator-prey interaction. The predator's generalism and prey template chromosomes are combined to form a search pattern. This search pattern, with wild-card positions, is matched against prey organisms' predation resistance chromosomes. If all of the non-wild-card positions in the search pattern match the prey's resistance chromosome, then the prey may be consumed. Otherwise the prey may escape.

Notice that an organism with all 1's, all wild-cards, in its generalism chromosome, is an extreme generalist. It can eat any organism. At the other extreme, an organism with all 0's in it generalism chromosome is a Koala bear-style specialist. It can only eat prey that exactly match its prey template.

So far, I have described predation implying that only a perfect match (0 mismatching bits) between a predator's search pattern and a prey's predation resistance chromosome allows the predator to consume the prey. The reality of the algorithm is not so simple. It is not sufficient to model predation as a deterministic function of a perfect match between predator and prey. First of all, this "all or nothing" type of interaction creates sharp cliffs in fitness space, such that a single bit mutation is likely to change a well adapted lineage into a lineage with a search pattern that fails to match any of the potential prey organisms. When a fitness landscape has sharp drops, or is "rugged" in Kauffman's (1993) terminology, mutation with natural selection cannot easily herd a population up a fitness peak. Evolution is much more effective on smoother fitness landscapes. But more to the point, the model does not behave reasonably with such a simplistic predation function (see Sections 3.1 and

3.2).

Predation in the model is a probabilistic function. The chance that a predator captures and consumes a prey organism is a function of the number of bits that do not match between the predator's search pattern and the prey's predation resistance chromosome. It is also a function of the *prey-location* probability parameter (0.15), the number of times the prey organism has been pursued by predators during the current time step, and a *predation-distribution* parameter (2.0). The intuition and motivation for including these additional factors is discussed in Sections 3.1 and 3.2. In sum, the probability that a predator successfully captures a prey is

$$Pr[\text{prey capture}] = le^{-(m+nd)} \tag{2.2}$$

where l is the *prey-location* probability, m is the number of mismatching bits between the predator's search pattern and the prey's predation resistance chromosome, n is the number of times that prey organism has escaped predation this time step, and d is the *predation-distribution* parameter. The idea here is that the chance of a predator capturing the prey should decay with the increasing mismatch between the predator's search pattern and the prey's predator resistance chromosome. However, a probability of 1 when there is a perfect match is too big. So we scale it by l, the *prey-location* probability. Finally, as we will see in Section 3.2, competitive interference had to be added to MoD in order to elicit realistic predation dynamics. This is added through nd.

A random number between 0 and 1 is generated, and if it is below the probability of Equation 2.2 then the predator kills the prey organism and increments its record of the number of prey it has eaten. If this is the third prey it has eaten since its last mating attempt, the predator is granted a new mating attempt. However, even if it fails to reproduce, it will have to capture another three prey organisms before it is allowed to mate again. I have tied mating to survival for efficiency reasons. There is a general trade-off between modeling the complexity of an individual's life and modeling long time scales (with population size as a third constraint). By keeping the behavior of the organisms simple, we are able to run the model for more generations and so paint a better picture of evolution.

The introduction of predation would seem to establish a food web. However, as presented so far, the food web lacks grounding. What prey do the organisms on the lowest level of the food web eat? Every food web needs a foundation of organisms that can translate non-biological matter into biological matter. These are the autotrophs. The autotrophs with which we are most familiar are the plants. For simplicity I will refer to autotrophs as plants and the heterotrophs (organisms that eat other organisms) as animals. Every organism carries an additional flag indicating if it is a plant or an animal. These flags do not mutate. Only plants are allowed to mate as well as compete for space with other plants. The same is true for animals. Instead of matching their search pattern against a prey's predation resistance chromosome, a plant matches against a bit pattern in their environment that represents the habitat. The chance that a plant absorbs an energy unit, or "meal," from the environment is

$$Pr[\text{absorbing an energy unit}] = h \times \textit{energy input} \tag{2.3}$$

where h is the fraction of bits that the plant's search pattern matches with its habitat and *energy input* is a parameter (1.0) of the model so that I may pour more or less energy into the ecosystem. If this probability is greater than 1, the plant is granted energy units equal to the integer part of that value plus another unit with probability equal to the fraction. So if $h\times$ *energy input* $= 1.5$, then the plant would have a 50% chance of garnering 2 energy units and a 50% chance of absorbing only 1 energy unit.

Given this implementation of predation, one might expect all lineages to evolve towards greater generalism. All of the previous artificial ecosystems that implement generalism have used a trade-off between generalism and metabolism, or energy consumption (Lindgren & Nordahl 1994; Saruwatari et al. 1994; Herráiz et al. 1997). The fact that we see specialists in biology seems to imply that there is some trade-off such that specialists might be favored over generalists by natural selection under some conditions. In fact, there is very little evidence pointing to such a trade off. A competing hypothesis postulates that specialists evolve from generalists through random drift. Genes for utilizing resources that aren't currently present are not under selective pressure for maintenance and so may be destroyed by mutations. In other words, specialists may have no advantage over generalists. There is some evidence for this (Thompson 1994). The real story is bound to be a lot more complicated. Thompson (1994) takes a middle of the road approach to the problem arguing that any given lineage will have a collection of three different types of genes: (1) Genes that help utilize a resource but are independent of genes for utilizing other resources. (2) Genes that help to utilize more than one resource. (3) Genes that help utilize a resource at the expense of interfering with the utilization of other resources. Only the final form of gene results in a tradeoff between specialism and generalism. Since the jury is still out on this issue, I have made the simplest assumption, that there is no tradeoff between specialism and generalism.

The only remaining issue in implementing the microevolutionary requirements of MoD is the realization of competition. A plant with a search pattern that is a better match to the habitat will collect more energy than its competitors with poor matches to the habitat. It will thus reproduce more quickly, passing on those "genes" for the match to the habitat, and eventually come to dominate the population. Similarly, plants with predation resistance chromosomes that do not match any of the search templates of the predators, along with predators whose search templates match with many of the available prey, will be more likely to reproduce than the poor sods who succumb to predation or starvation.

2.2.3 Macroevolution

The purpose of this model is to examine macroevolutionary dynamics through the modeling of the microevolutionary interactions. It is thus appropriate that the heart of MoD has already been described in the previous section on microevolution. There are only two remaining phenomena that have not been explained: regional level issues, like geography, and species, with which we are primarily concerned.

Geography and Migration

As was mentioned before, space is discrete. However, MoD does not have the fine grained spatial scale of cellular automata. Rather, space is organized into "patches." There is little spatial structure within a patch. Each patch can hold up to 2K (2048) organisms[7]. Plants are limited to only fill half of a patch, otherwise they might take over the entire patch and so compete with animals for space. The 16 patches are organized into a 4 by 4 grid. The ecological interactions are constrained to organisms within a single patch. An organism may

[7] The organisms are stored in an array. In addition, free spaces in the array are wired up into a linked list stack so that finding a free space is an $O(1)$ time operation. This means that initially, new organisms are placed into the patch in consecutive positions of the array. However, as the pattern of deaths in the array becomes random, the placement of new organisms will also become random.

only mate[8] with and eat[9] other organisms within its patch. During each time step every organism in a patch is given a chance to eat[10]. Failing this, it starves to death. However, if it succeeds in finding food, and it has eaten thrice since its last mating attempt, it will try to find a viable mate. An organism may be eaten before it gets its chance to mate. However, new organisms produced in a time step are held in a separate data structure until the end of the time step. This prevents them from being born and acting or being eaten in the same time step. In this way, predation occurs before prey competition for space, and so reduces the competitive pressures amongst the prey (Crawley 1992).

Migration, or movement between patches only happens at the beginning of an organism's life. When a new organism is born, it takes a random walk across the environment to locate the patch in which it will spend the remainder of its life. The random walk is 0–2 steps long[11]. For each step of the walk a random number between 1 and 9 is generated. These numbers correspond to the patch where the organism is currently located along with the eight potentially neighboring patches. If the random number dictates a step off the edge of the 4 by 4 environment, the organism is placed in the border patch nearest to its intended destination. These random walks roughly model the dispersal of real organisms. The chance of finding an organism at a distance d from its point of origin is approximately d^{-2}.

The theory of allopatric speciation does not just require spatial structure, it requires "geography." This includes barriers to migration. The canonical example is a mountain range that separates two populations by making migration between them exceedingly rare. This is modeled by modifying the random walk algorithm. Every patch in the model is associated with a "migratory barrier" between 0 and 1. For every step of the random walk in an organism's migration, a random number between 0 and 1 is generated. If that number is below the migratory barrier of the patch that the organism is trying to enter, it fails to enter that patch and must remain in its current patch for that step of the random walk. If there are more steps to follow in the walk, it may end up trying again, and perhaps even

[8] Potential mates are found by starting at a random position in the array and scanning the entire population until either a compatible mate is found or the population in that patch has been exhausted. When the end of the array is hit, the search continues from the beginning of the array.

[9] Similar to mating, a prey organism is located by starting at a random position in the array and then scanning the next *search-radius* (512) positions. This continues until either a prey organism is consumed, or 512 positions in the array have been checked.

[10] The live organisms are also kept in a doubly linked list so as to not waste time processing empty positions in the array of organisms. The double linkage is useful since an organism in the middle of the list may die and have to splice itself out of the list. New organisms are inserted at the front of the list.

[11] The length of the random walk is determined randomly with uniform probability.

succeeding the next time. Thus, if all the patches have 0.9 migratory barriers, then there is only a 10% chance that an organism will be able to enter a neighboring patch on a step of its random walk. As was mentioned above, habitats are modeled by associating a bit pattern with each patch representing the relevant climate or habitat characteristics for the plants.

Species and Speciation

A species is a reproductively isolated gene pool. If two organisms can mate and produce viable offspring, then they are, by our definition, part of the same species. This is a non-transitive relationship. Organism A may be able to reproduce with organism B, and organism B may be able to reproduce with organism C, but it sometimes happens that organism A cannot reproduce with organism C. Yet all three are part of the same species. It is still reasonable to say that organisms A and C are part of the same species, despite their inability to mate, because their genes can still come together in a grandchild or later descendant. They are part of the same gene pool.

Reproductive barriers are implemented by giving the organisms a fourth chromosome, twice the length of the others, called the "reproduction" chromosome. The 64 bits (loci) of the reproduction chromosome[12] are assumed to represent the set of characters influencing reproduction such as mating season, reproductive morphology, courtship behavior, etc. A pair of organisms may only mate if their reproductive genotypes differ in at most one bit. That is, potential mates must have reproductive genotypes within Hamming distance one of their mate. As with earthworms, there are no male and female "sexes" per se. If there were only a few bits in the reproductive chromosome, then the probability of two lineages randomly evolving to within a single bit would be relatively high. Conventional wisdom in biology assumes that it is highly unusual for two species to coalesce into one.

Using a Hamming distance restriction on mating mimics the non-transitivity observed in nature. Sometimes individuals of a species may be well adapted to their environment across a gradient of resources in the environment. It occasionally happens that while individuals of the species may mate with neighboring individuals, individuals from more distant

[12]The chromosome is represented by two 32-bit integers. Mutation is applied to each of the 32-bit halves of the chromosome in turn, so the reproductive chromosome has an average of 0.04 bits flipped per new organism. The mutation rate was set so as to generate an expected total of one bit flip for every 10 new organisms (with 5, 32-bit chromosomes in each organism).

populations do not produce viable offspring (Oliver 1972; Fox 1951). As both Orr (1995) and Gavrilets (1997; 1997) point out, the accumulation of mating incompatibilities is necessarily a gradual process. If a dramatic mutation reproductively isolates and organism, it will likely have the same fate as Lonesome George[13]. With only one living member of an isolated gene pool, a species is doomed.

Orr (1995) has analyzed the dynamics of speciation due to genetic incompatibilities in the genomes of the parents. He found that speciation should progress at approximately the same rate regardless of whether the incompatibility is based on a pair of loci that have a dramatic effect or on many loci, each having a small effect on reproduction. There is thus reason to believe that the dynamics of the 1-bit restriction on mating should be similar to the dynamics of allowing mating with multiple differences in their reproductive chromosomes. However, allowing matings between organisms with multiple loci differences would probably require proportionally longer chromosomes to avoid the accidental coalescence of two species.

Notice that MoD includes both pre-mating and post-mating reproductive barriers. The reproductive genotypes encode the pre-mating reproductive barriers and any incompatibilities that prevent the production of viable offspring. However, a "viable" offspring is likely to have ecological chromosomes (the predation resistance, prey template and generalism chromosomes) that are different from both parents' ecological chromosomes, due to recombination. Depending on the result of this recombination, the offspring may acquire a prey search pattern that fails to match any of the organisms in its patch. It may also acquire a predation resistance chromosome that makes it extremely vulnerable to predation. In either case, even though the offspring is "viable" it will soon die because it is not well matched to the ecological context of its time. This sort of fitness reduction is not taken into consideration in our definition of species because of its fundamental dependence on context.

It may seem difficult to identify a gene pool amongst a large number of organisms with this non-transitive mating relationship. However, there is a surprisingly efficient algorithm ($O(n)$ where n is the number of different reproductive genotypes in the population) for identifying the reproductively isolated gene pools. Remember that, as long as there is an indirect connection between two organisms through which a series of matings might bring

[13]Lonesome George is a Galapagos Giant Tortoise who is the last survivor of his subspecies (*Galapagos elephantopus abingdoni*). All attempts to get him to mate with members of other subspecies have failed.

their genes together in a descendant, then those two organisms share the same gene pool. Consider a diagram, or "graph" in computer science lingo, of who can mate with whom. Each node in the graph represents a unique reproductive genotype in the population. Since many organisms can have the same reproductive chromosome, a node may represent many organisms. We draw a line, or "edge," between any two nodes whose organisms can mate. That is, we connect nodes that are only one bit different. If we imagine that the nodes are beads and the edges are threads connecting the beads, then a species is the dangling set of beads you would get if you picked up one bead. A group of beads connected by threads is called a "connected component" of the graph. A graph may have many connected components. An example of a reproductive graph with two species is shown in Figure 2-4.

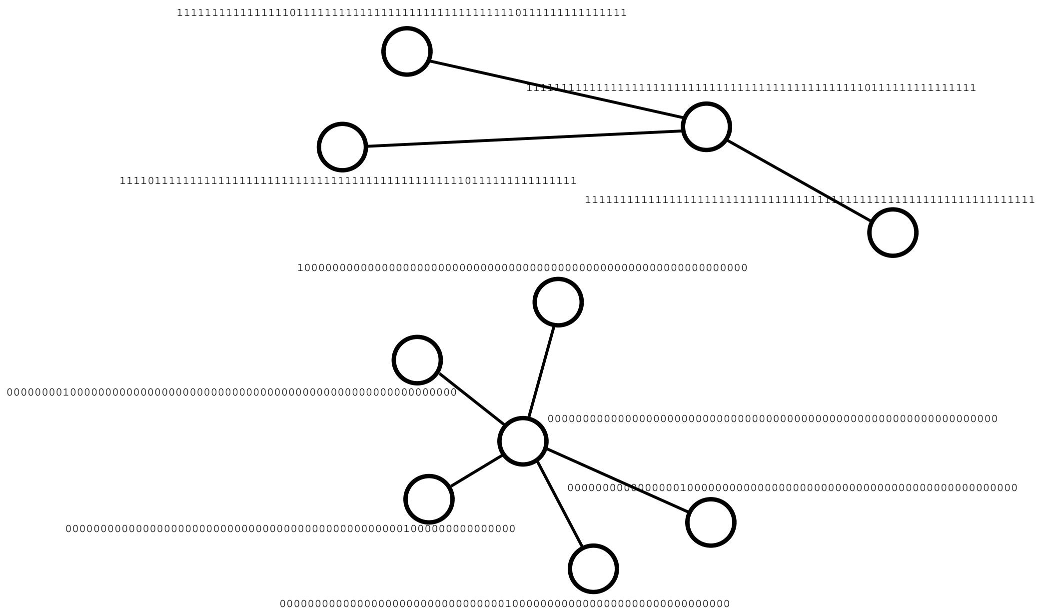

Figure 2-4: A graph representing the reproductive genotypes in a population and their relationships. The nodes represent the reproductive genotypes of the organisms currently alive in the model. An edge connects the nodes whose organisms are allowed to mate. These are the nodes that have only a single bit difference. There are two species in this graph, represented by the two connected components.

To identify a component we mark nodes with a depth-first search[14] of the graph, starting

[14] A depth-first search of a graph is a particular style of working through a graph. Starting at some arbitrary origin node, you mark it and then apply this same search recursively to each of its un-marked neighbors in turn, finishing with one neighbor before processing the next. If you think of the living members of a family tree as a graph, then this would correspond to starting with a grandmother, marking her, then moving to her first child, marking that child, then moving to that child's children, and so on. All the grandchildren of the first child will be marked before the second child of the grandmother is marked. This is why it is called "depth-first." You work as deeply as possible into the graph before popping back up levels

at any arbitrary node. All those nodes that get marked, and only those nodes that get marked are members of a single connected component, a single species. We repeat this until all the nodes have been marked and so all the species identified. Because this is an efficient algorithm, we can execute it every time step to track the species and identify exactly when a species splits apart.

A species splits when all of the organisms that carry a particular reproductive genotype die, and the loss of this node in the graph splits apart a connected component into more than one component. One strange byproduct of this implementation is that a species can split into more than two species. Notice that if either of the central nodes in Figure 2-4 disappear, many single node components will be created. We do not know if this happens in nature.

Death comes in three forms in MoD. An organism can be eaten by another organism. If an organism fails to find food in a time step, it will starve to death[15]. Finally, if a patch has been filled to its spatial capacity of 1K plants or 1K animals, then a new organism has a 50/50 chance of replacing an organism of its same kind (plant or animal). Thus, an organism can die in a sort of competition for space. Without this final form of death, evolution might grind to a halt. If there is no space left for new organisms, and no predators left to make space amongst the prey, then new mutations will never enter the population and nothing will change. As it turned out, plant population sizes were generally constrained by space, not herbivores. Because of the $3:1$ transformation of prey matter into predator matter, the animal populations rarely rose above a third of the plant population and so never neared the carrying capacity of the space in a patch.

The implementation of a species as a connected component of a reproductive graph bears a certain similarity to the "holey adaptive landscape" model of Gavrilets and Gravner (1997; 1997). Both systems are constrained to high dimensional hypercubes. On closer consideration, there are a few important aspects that distinguish the two approaches. First of all, Gavrilets and Gravner consider only monomorphic species. That is, there is no genetic variation within a species, and so they can use a distribution model to represent

to finish off the upper levels of the graph. See Cormen, Leiserson, and Rivest (1990, p.478) for an $O(n)$ depth-first search of a graph.

[15] I played with allowing organisms to survive for longer without eating but it only decoupled the population dynamics of the predators and prey and so destabilized the model thus increasing the probability that one of the two populations would go extinct.

its dynamics rather than a configuration model. Second, the reproduction chromosome of our model explicitly codes pre-mating reproductive barriers, while the post-mating barriers of inviability emerges out of the starvation and predation of an organism in the ecosystem dynamics. In contrast, the analysis of holey adaptive landscapes has not yet been extended to coevolutionary dynamics or pre-mating reproductive barriers (Gavrilets & Gravner 1997).

2.2.4 Parallelism

The remaining aspects of MoD concern how it was implemented on a parallel computer. This is only of interest to programmers, and so other readers may want to skip to the next section.

The current model was written in CILK, a parallel extension of C, and run on the Xolas cluster of Sun Ultra symmetric multiprocessors at MIT[16]. CILK is simply C with the addition of two new key words, `spawn` and `sync`. By using `spawn` in front of a function call, a new thread is started and the function is executed in parallel with the rest of the code. A `sync` command will freeze the processing in the current thread until all the threads that it has spawned finish their work. CILK guarantees load balancing amongst the processors without programmer control. These two commands provide an easily understandable abstraction of parallelism at the cost of some restrictions. Most strikingly, there is no inter-thread communication[17]. The language is best suited to problems where the work can be split up into independent pieces.

The execution of the model can be nicely divided into the different patches. Almost all of the work occurs within a patch, independently of the others. The only exception is migration between the patches. So every time step, a separate thread was spawned to handle the interactions between organisms within each patch. At the end of the time step, these threads were synchronized and a single thread handled the distribution of organisms through migration as well as the collection of any statistics about the state of MoD.

To maximize the amount of work that was done in parallel, newborns were put in a special array in their parents' patch. The random walk of a newborn was implemented in parallel with newborns in other patches except that, instead of placing the newborn into its

[16]See `http://xolas.lcs.mit.edu/` for information on the processors and links to the CILK distribution pages.

[17]There are a few additional control abstractions for aborting threads and protecting memory from collisions between threads, but their use is officially discouraged and turned out to be unneeded for this model.

destination patch, the location of that destination patch was noted in the data structure that held the newborns. So, once the threads finished their work in all the patches, migration was a simple matter of reading the destination location associated with each newborn and stuffing the newborn into its specified patch.

2.3 Caveats

Meeting the requirements dictated by the hypotheses necessitates a non-trivial degree of complexity. However, there are many dynamics and characteristics of real biological systems that have been ignored. There is even good reason to believe that some of the missing pieces have significant impacts on the evolution of diversity. Whether or not they are necessary to address diversification is an open question, and so, for the nonce, they have been left out.

As was mentioned in Section 2.2.2, MoD makes no distinction between genotypes and phenotypes. The mapping from genotype to phenotype has implications for the constraints on variation across a population as well as the evolutionary trajectory of a species. Unfortunately, just what those implications are remains an open question.

I have consciously avoided the encoding of the size of organisms into MoD. This seemed an unnecessary elaboration. However, the size of an organism has dramatic consequences on its ecological behavior (Haldane 1927; Charnov 1993). In particular, size is important to the success of parasites. There are no parasites in MoD. Unlike predation in the model, a host can provide ample resources for many parasites as well as the parasite's offspring. Parasites are thought to have played an important part in the evolution of life (Hamilton et al. 1990; Paterson & Gray 1997; Anderson & May 1982; Hurst 1991), but none of the hypotheses for the diversification of life made explicit mention of them.

It is highly unlikely in MoD that a new mutant organism will establish a new species by the fact of its mutation. For a start, this would require two bits being flipped in its reproductive chromosome, which happens with probability 0.0002. Then, lest the mutant suffer the fate of Lonesome George, it must survive long enough for a viable mate to enter its patch. In contrast, it is relatively easy for a mutation in a real plant to establish a new species. This is simply because most real plants can fertilize themselves, and so need not depend on the arrival of a compatible mate.

The self-fertilization of plants brings up another important omission of the model. I have adopted the reproductive species concept and the following chapters will use the elegance of this definition as leverage for many of their results. However, the reproductive species concept only applies to organisms with some degree of bi-parental reproduction. If an organism was completely self-fertilizing or was asexual, then the reproductive species concept would not be applicable. Historically, most of life has been asexual. The phenotypic similarity within asexual "species" probably does not stem so much from sharing a gene pool as it does from the constraints of the fitness landscape. The innovation of sexual reproduction is interesting in part because it binds together a population of organisms into a sort of super-organism. An innovation in one lineage can eventually be shared by all the lineages in a sexual species.

There is very little behavior allowed to the organisms. Perhaps most strikingly, organisms only migrate once during their life time. The spatial scale of a patch is supposed to be such that almost all movement of individuals can be considered to take place within a patch. Furthermore, many ecological models focus on nutrient cycling through the food web and environment. Besides the fact that the hypotheses do not make reference to this level of detail, I have consciously chosen to work at a higher level of abstraction for reasons of both parsimony and computational resources.

Finally, perhaps the most stringent limitation on MoD is the fact that the speed of computation restricts the model to a regional, not global, level of representation. With 16 patches, a population that hovered around 20K organisms, and 5000 time steps, one run of MoD took about 20 minutes on an 8 processor Sun Ultra. One advantage of this model is that, if patch sizes remain constant, processing time scales linearly with the population/spatial scale. So, to model a world with twice as many patches and twice as many organisms in the same amount of time, we would only require a doubling in processor speed. But for the moment, all the results of MoD are restricted to evolutionary dynamics on a regional scale.

Is there any reason to believe that the model is an adequate representation of even the regional dynamics? This poses the problem of invalidating the model, to which we will now turn.