Appendix A

Open Problems in Reproductive Graphs

We would like to know the expected number of components (species) in an evolving reproductive graph. These are graphs with nodes representing the different reproductive genotypes in the current population of the model. The graph starts with one node for each initial species. Over time, neighbor nodes appear due to mutations in the reproductive genes of the live organisms. At the same time, nodes disappear due to the death of all the organisms that carry the reproductive genotype represented by that node. These graphs are constrained to a 64-dimensional binary hypercube in the model. However, generalization to n-dimensional quadrary hypercubes may have more relevance to biological species, where each dimension corresponds to the four possible nucleic acids at a locus (G,A,T, and C).

One way to think about this is to consider the space of all possible graphs on the hypercube. If the binary hypercube has n dimensions, then there are 2^n possible nodes and so the power set of these nodes represents all 2^{2^n} possible graphs. However, it is not true that in one time step the model could transition to any other possible graph. The evolution of these graphs are constrained by the fact that only neighbors of current nodes may appear in the next time step. Thus, we might construct a Markov process with 2^{2^n} states along with the probabilities of making transitions between all these states.

The determination of these probabilities is non-trivial. Even if we ignore the bias to the probabilities introduced by predation (e.g., some organisms are more likely to die in a time step than others), it is still true that the probability of the loss of a node is a function of the number of organisms represented by that node. Furthermore, the probability of the appearance of a new node is a function of the number of organisms represented by neighboring nodes. In fact, the story is even more complex. Let us ignore the effects of the limits on searching for a mate, as well as the spatial partitioning of organisms into patches. Given these simplifications, the probability of a new node appearing is actually a function of the number of organisms in the neighboring nodes as well as the number of organisms in the neighbors of those neighbors. To see this, consider a two node graph with one populous node and a new mutant node. Now consider all the potential neighboring nodes to the mutant that are not neighbors of the populous node. The chance of these nodes appearing in the next time step is not simply a function of the number of organisms (1) in the mutant node. If the "populous" node also only has 1 organism, then there will be at most 2 new organisms produced in the next time step. Those new organisms might be mutants with reproductive genes that are neighbors of the parental mutant node. But since there are at most 2 new organisms, the chance of this is low. However, if the populous node has many organisms, then the 1 organism from the mutant node might engage in many mating events and so the probability of producing neighboring nodes would be higher.

Problem 1 *Given a mutation rate per mating m, a death rate per organism per time step d, a constant population size N, and starting with a homogeneous population in a single node, what are the expected number of components in the graph at equilibrium?*

The addition of the number of organisms represented by each node adds another dimension to the states of the Markov process. Each node in the graph could have from 1 to N organisms. This leads to a $N2^{2^n}$ state Markov process. If we assume a carrying capacity of the environment, then we may assume that N is constant. In the model, N is typically around $20,000$.

There are a number of critical elaborations to Problem 1:

Problem 2 *What are the expected number of components given different definitions of adjacency in the hypercube? What if nodes are connected when they have $\leq k$ bits different, can the hypercube results be generalized to these more complex topologies?*

Problem 3 *What are the expected number of components given positive and negative assortative mating? That is, given the tendency to mate with organisms in the same node (positive) or in neighboring nodes (negative).*

Problem 4 *What are the expected number of components given spatial structure with its concomitant limitations on the mating pool and migration between mating pools?*

A simplified version of these problems may be more tractable. Let us elide the underlying population dynamics and just assign birth and death rates to the nodes.

Problem 5 *Given a birth rate m for the probability that a new node will appear next to an extant neighbor, a death rate per node per time step d, and starting with a single node, what are the expected number of components in the graph at equilibrium?*

This can be explicitly computed for low n where n is the number of bits in the reproductive gene. Let us assume, in accordance with the model, that an extant node can give birth to neighbors and die in the same time step. However, a node that is born in a time step may not die or create new neighbors in that time step. Then the probability a reproductive graph in state s_1 makes a transition to state s_2 in a time step can be decomposed into the product of the individual dynamics for all the nodes and potential nodes in the hypercube. There are four possible cases for each node:

1. **A node is born:** The chance that a node appears in a time step is a function of the number of living neighbor nodes x:

$$Pr[\text{birth}] = 1 - (1 - b)^x \tag{A.1}$$

2. **A node is unborn:** The chance that a dead, or non-existent, node remains that way is just 1 minus the chance that it is born:

$$Pr[\text{not born}] = (1 - b)^x \tag{A.2}$$

3. **A node dies:** The chance that a node dies is just the node's probability of disappearing, d, times the probability that it is not replaced by a mutation from a neighboring node. That is d times the probability of no birth:

$$Pr[\text{death}] = d(1 - b)^x \tag{A.3}$$

4. **A node survives:** The chance that a living node survives into the next time step

is the chance that it either does not die or that it dies but is born again due to a mutation from a neighboring node:

$$Pr[\text{survival}] = (1 - d) + d\,(1 - (1 - b)^x) \tag{A.4}$$

The product of these probabilities over all 2^n nodes in the binary hypercube gives the state transition probability for the Markov process.

I have computed the $n = 2$ case of the Markov process merely to check if the abstraction is reasonable. A sign of a bug in the abstraction would be a high probability for an equilibrium at a single node graph. We know from running the model that the number of reproductive genotypes at any given time is about two orders of magnitude greater than the number of species.

The results for computing the Markov process with $n = 2$ depend on the relation of the birth probability b and the death probability d for the nodes. If the birth rate is relatively high, then the system is most likely to settle into the full 4-node graph. The 3-node graphs are the next most likely, and so on, except that there is a fair chance that the process gets stuck in the 0-node graph. The 0-node graph is an absorbing boundary on the process. If the death rate is relatively high, then, predictably, the process tends to get stuck in the 0-node state. In between these extremes, there tends to be a bimodal distribution in the probability for the different states. The 0 and 4-node states tend to collect most of the probability, with the other graphs garnering a smear of the remaining probabilities in between. In other words, the results for $n = 2$ show reasonable behavior.

Unfortunately, since the problem scales super-exponentially, even moderate values of n are not computable.

Appendix B

Derivations

B.1 Bit Differences in a Heterogeneous Environment

To generate a heterogeneous environment, I flipped 8 bits in the habitat bit pattern of every patch. A bit was selected by generating a random number with uniform probability from 1 to 32. The selected bit was flipped in the current patch. This was repeated 7 more times. The same bit could be flipped and then flipped back to its original value. Each patch was treated independently. The question then arises, what is the expected number of bits that differ between any two patches. This is important because it determines the expected disadvantage a plant would suffer if it was perfectly adapted to the habitat of its parent's patch but migrated to a new patch.

The best way to think of this is question is to model it with a birth-death Markov process (Drake 1967). First off, the difference between two patches, each with 8 bits flipped, is equivalent to the difference between the original state of a patch and its state after 16 bits have been flipped. If we start with all bits set to 0, as we did in the model, then we can represent the possible states of the patch's habitat by states in the Markov process labeled with the number of 1's that have been set. If the habitat pattern has i 1's in it, then the chance that the next bit flip adds another 1 is $(32 - i)/32$. And similarly, the chance that the bit chosen to be flipped is already a 1, and so is changed to a 0, is $i/32$. One end of this Markov chain is shown in Figure B-1.

To determine the probability of being in state i after n bit flips, we first construct a matrix M where the (i, j)-th entry indicates the probability of moving from state i to state j in one step. These are the values labeling the arrows in Figure B-1. Raising this matrix

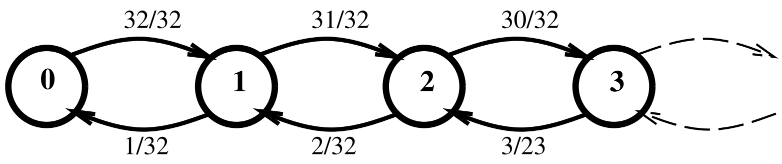

Figure B-1: A simple birth-death Markov process to model the number of bits set to 1 in a 32 bit habitat pattern for a patch. Only the first 4 out of 33 states are shown.

to the n-th power results in a matrix where the (i, j)-th entry is the probability of ending in state j if you started in state i and took n random steps. Since we know that we start in state 0, the zeroth row of the matrix gives the probabilities of being in each of the 33 possible states. Then to find the expected state we use the formula for the expectation of a probabilistic event:

$$E[\text{state}] = \sum_{j=0}^{32} jM(0, j) \tag{B.1}$$

For this problem $E[\text{state}] = 10.3$. On average, patches will differ in 10.3 bits under the heterogeneous environment condition. The progression of the number of bits set to 1, starting from all 0's, is shown in Figure B-2. There is an asymptote at 16 because a 32-bit string, when fully randomized by an infinite number of bit flippings, has on average half of its bits set to 1 and the other half to 0.

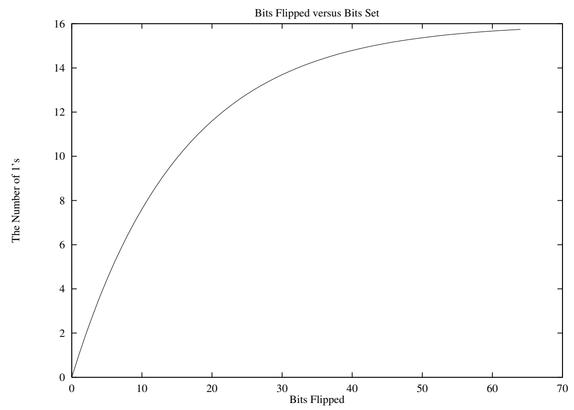

Figure B-2: The expected number of bits set to 1 after flipping a given number of bits. We start with all bits "cleared," set to 0.

B.2 Migration Rates and the Fixation Index

First let us derive a rough estimate of the number of migrants that might be exchanged between neighboring patches with 0.9 migratory barriers, in one generation. Once we have an estimate for the number of migrants, we can derive the fixation index. This will be a function of the migration rate. And finally, we will be able to calculate the fixation index by plugging in the rate of migration that we will have calculated.

B.2.1 Number of Migrants

What is the probability that a plant seed will land in a particular neighboring patch? First of all, the seed takes a random walk of 0 to 2 steps on the grid of patches. So 1/3 of the seeds will take exactly one step. With 0.9 migratory barriers, only 1 in 10 of these seeds will make it over the barrier. To be conservative, we will consider the worst case, when the patch of origin is not on the edge of the environment, and so it has 9 neighbors (including itself). In general, the plant populations were limited by the carrying capacity of their patches, so a new seed only had a 1 in 2 chance of replacing an endemic plant in the destination patch. The product of all these probabilities is:

$$\textit{origin} \overset{1/3}{\rightarrow} \textit{one step} \overset{1/10}{\rightarrow} \textit{over barrier} \overset{1/9}{\rightarrow} \textit{destination} \overset{1/2}{\rightarrow} \textit{take root} = \frac{1}{540} \tag{B.2}$$

Since we expect 1000 plants to reproduce in a generation, this results in about 2 migrants that take one step into a particular neighboring patch. Added to this are the seeds that attempt to take 2 steps and end up in that same neighboring patch. There are two cases for these 2-steppers. First, a seed might be blocked on the first step with probability 9/10, and then proceed in the next step, exactly as the one-steppers. Second, a seed might make it into the neighboring patch on the first step, but get trapped there on the second step with probability 9/10. Any two-stepper who first moves into a different patch and then moves into the destination patch would have to cross the migratory barriers twice. The chance of doing that is so small that they can be safely ignored. The total of the one-stepper and two-stepper migrants is:

$$1000\left(\frac{1}{540} + \frac{9}{10}\left(\frac{1}{540}\right) + \frac{9}{10}\left(\frac{1}{540}\right)\right) \approx 5 \tag{B.3}$$

B.2.2 The Fixation Index in a Haploid Population

The fixation index is generally calculated for a diploid population (Hartl & Clark 1997, p.194). That is, a population that has two copies of each chromosome, like humans. However, the formulas can easily be derived for a haploid population, with half the alleles of the analogous diploid population.

In a haploid population with N individuals, there are N alleles. We let m be the proportion of new alleles introduced into the population through migration. Thus, if we choose two alleles randomly from this population, we can ask what is the chance, F_t, that the two alleles in generation t are copies of the same allele in some previous generation? we assume that if either allele is from a migrant, then the two alleles cannot be identical by descent. There are two possible cases for identity by descent. In one case, neither allele is one of the new migrant alleles, $(1-m)^2$, but they both are copies from an allele in the previous generation. This is equivalent to randomly choosing the same allele twice from the previous generation. The probability that the second allele chosen comes from the same one as the first allele chose is $1/N$. In the second case, the two alleles are not migrant alleles and they don't come from the same allele in the previous generation, $(1-m)^2(1-1/N)$, but the two ancestral alleles from the previous generation are themselves identical by descent, with probability F_{t-1}. This leads to the recurrence relation:

$$F_t = (1-m)^2 \frac{1}{N} + (1-m)^2 \left(1 - \frac{1}{N}\right) F_{t-1} \tag{B.4}$$

To solve this we look for the equilibrium value of $\hat{F}$ when $F_t = F_{t-1}$. We can get an elegant and reasonable approximation if we throw out all terms with m^2 and m/N. These terms are too small to matter.

$$\hat{F} \approx \frac{1}{N} + \left(1 - 2m - \frac{1}{N}\right) \hat{F} \tag{B.5}$$

$$\hat{F} \approx \frac{1}{1 + 2Nm} \tag{B.6}$$

Since Nm is exactly the number of new migrants in the population we can plug in the value $Nm = 5$ as calculated above to get $F \approx 0.09$.

B.2.3 Migration Rates

What is the proportion of migrants in a generation of new organisms in a patch? Note that migrants do not disappear off the edge of the environment. Since the number of migrants is conserved by the model, the average proportion of immigrants into a patch should equal the average proportion of emigrants leaving the patch. It is slightly easier to analyze the amount of emigration.

Consider a migratory barrier b which is assumed to be high ($1 - b \ll 1$). This lets us ignore the rare cases where a migrant crosses a barrier on both steps of its random walk. We can split up the population of new organisms into three equal classes: the stay-at-home types, the 1-steppers, and the 2-steppers, corresponding to random walks of 0, 1, and 2 steps. The stay-at-home types contribute nothing to the population of emigrants. The 1-steppers contribute $1 - b$ of their number, and the 2-steppers contribute $1 - b^2$ of their number. So the total proportion of emigrants m in the population of new organisms (not the parental organisms) is:

$$m = \frac{1}{3}(1 - b) + \frac{1}{3}(1 - b^2) \tag{B.7}$$

If $b = 0.95$ then $m = 0.049$, and if $b = 0.9$ then $m = 0.097$. So, for high migratory barriers, the proportion of emigrants, which is also the proportion of immigrants, is very close to the proportion that can get over the migratory barrier in one step $(1 - b)$.

References

Alvarez, L. W., Alvarez, W., Asaro, F., & Michel, H. V. (1980). Extraterrestrial cause for the Cretaceous-Tertiary extinction. *Science*, **208**, 1095–1108.

Anderson, R. M. & May, R. M. (1982). Coevolution of hosts and parasites. *Parasitology*, **85**, 411–426.

Andersson, M. (1994). *Sexual Selection.* Princeton, NJ: Princeton University Press.

Ausich, W. I. & Bottjer, D. J. (1985). Phanerozoic tiering in suspension-feeding communities on soft substrata: Implications for diversity. In: *Phanerozoic Diversity Patterns: Profiles in Macroevolution* (Valentine, J. W., ed), pp. 255–274. Princeton, NJ: Princeton University Press.

Axelrod, R. (1984). *The Evolution of Cooperation.* New York, NY: Basic Books.

Bambach, R. K. (1985). Classes and adaptive variety: The ecology of diversification in marine faunas through the Phanerozoic. In: *Phanerozoic Diversity Patterns: Profiles in Macroevolution* (Valentine, J. W., ed), pp. 191–253. Princeton, NJ: Princeton University Press.

Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). *Genetic Programming.* San Francisco, CA: Morgan Kaufmann Publishers, Inc.

Barry, D. (1998). Bad science. *The Boston Globe Magazine*, pp. 8–9.

Baum, D. A. & Donoghue, M. J. (1995). Choosing among alternative "phylogenetic" species concepts. *Systematic Botany*, **20**, 560–573.

Bedau, M. A., Ronneburg, F., & Zwick, M. (1992). Dynamics of diversity in an evolving population. In: *Parallel Problem Solving from Nature, 2* (Männer, R. & Manderick, B., eds), pp. 95–104. Amsterdam, The Netherlands: Elsevier Science Publishers.

Begon, M., Harper, J. L., & Townsend, C. R. (1990). *Ecology.* London: Blackwell Scientific Publications.

Belding, T. C. (1995). The distributed genetic algorithm revisited. In: *Proceedings of the Sixth International Conference on Genetic Algorithms* (Eshelman, L. J., ed), pp. 114–121, San Francisco, CA: Morgan Kaufmann.

Benton, M. J. (1990). The causes of the diversification of life. In: *Major Evolutionary Radiations* (Taylor, P. D. & Larwood, G. P., eds), pp. 409–430. Oxford, UK: Clarendon Press.

Benton, M. J. (1995). Diversification and extinction in the history of life. *Science*, **268**, 52–58.

Bernstein, H., Byerly, H. C., & Michod, R. E. (1985). Sex and the emergence of species. *Journal of Theoretical Biology*, **117**, 665–690.

Booker, L. B. (1985). Improving the performance of genetic algorithms in classifier systems. In: *Proceedings of the First International Conference on Genetic Algorithms* (Grefenstette, J. J., ed), pp. 80–92, Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Booth, G. (1997). Gecko: A continuous 2-d world for ecological modeling. *Artificial Life*, **3**, 147–164.