Archipelago

Jeet Sukumaran

December 13, 2009

Introduction

Archipelago is a forward-time simulation of macro-evolutionary diversification processes in a spatially-explicit framework. The motivation for this simulator is to provide for null distributions of diversity, i.e., numbers of species or lineages found in different regions, under the model being simulated.

The Simulation Model

The Diversification Process

The birth-death process has two parameters. The birth rate, λ, is the probability that each species in the system speciates, or splits into two daughter species. The death rate, μ, is the probability that each species in the system goes extincts. A special case of the birth-death process is when the death rate, μ, is 0, in which case we have a pure-birth process, which is also referred to as the *Yule* model.

In the pure birth process, the expected number of lineages, $E(n)$ at time t is given by:

$$E(n) = e^{\lambda t},$$

corresponding to population growth equations of the same form.

With extinction, the expected number of lineages, $E(n)$ at time t is:

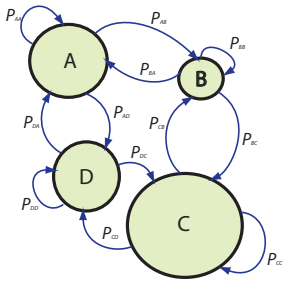
$$E(n) = e^{(\lambda - \mu)t}.$$

The Geographic Template

The spatial aspect of the simulation model is represented by the *geographical template*, which the defines the fundamental atomic spatial units of the simulation, *regions*, and the connectivity between these units, which determines the rate of dispersal of lineages from one region to another.

The geographical template is specified as a $M \times M$ matrix, where M is the number of regions in the system and the value of cell $M_{i,j}$ specifies the probability of dispersal from the i^{th} region to the j^{th} region. The rows of the matrix are required to sum to 1, and cell $M_{i,i}$ represents the probability of no dispersal.

For example, the following schematic representation of the geographical template:

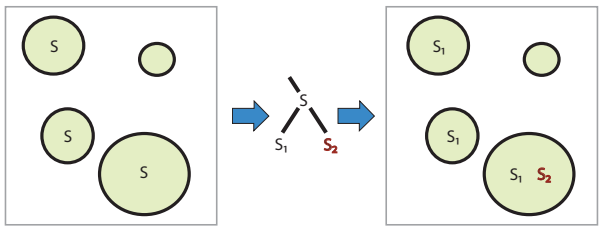


would be represented by the following matrix:

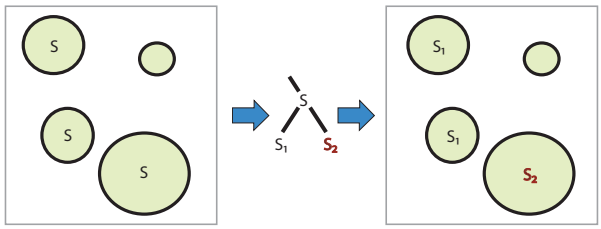
	A	B	C	D
A	P_{AA}	P_{AB}	P_{AC}	P_{AD}
B	P_{BA}	P_{BB}	P_{BC}	P_{BD}
C	P_{CA}	P_{CB}	P_{CC}	P_{CD}
D	P_{DA}	P_{DB}	P_{DC}	P_{DD}

Speciation Modes

Two types of speciation modes are supported: *sympatric* and *allopatric* speciation. In sympatric speciation, a speciation event results in one region receiving both daughter lineages while the remaining regions receive just one of the two (the same one for all the remaining regions):



In allopatric speciation, one region region receives one of the daughters, while all the others receive the other:



The Simulation Procedure

The simulation is set up by specifying the birth rate, λ, the death rate, μ, and the geographical template as a matrix of dispersal probabilities, as well as a random seed to use. The birth rate, death rate and random seed are specified as command line arguments to the program, while the dispersal probability matrix is given a text file provided to the program.

The termination condition for the simulation also needs to be specified. Two types of termination conditions are supported. The first, *target diversity*, specifies the number of species or lineages that need to be generated. The second, *generation limit*, specifies the number of generations that need to be run, irrespective of the number of species or lineages that are generated. The "target diversity" termination condition is ideally suited for generating distributions of diversity that match some observed total diversity, to statistically compare the distributions of the same total diversity under random diversification and dispersal against that observed, to see whether additional factors might be operating in the observed system.

Once the simulation is set up with the appropriate parameters and termination condition, a particular region is *seeded* with an initial species or lineage. The main simulation cycle then runs until the specified termination condition is reached.

The main simulation cycle consists of two phases: the *dispersal* phase and the *diversification* phase.

In the dispersal phase, for each species or lineage in each region, a dispersal destination is selected based on the dispersal probabilities defined in the geographical template. This destination may the same as the source region, or it may be another region in the system. If the destination region is the same as the source region, or if the species or lineage already exists in the destination region, then no colonization of a new region occurs. Otherwise, the destination region is added to the range of the species.

In the diversification phase, a uniform random number, $u \sim U(0, 1)$ is generated. If $u < \lambda$ then a speciation event is modeled, while if $\lambda < u < \lambda + \mu$ then an extinction event is modeled. A speciation event is modeled by splitting the speciating lineage into two daughter species. Each daughter species inherits the range of their parent (i.e., the speciating lineage) in one of two modes. In the sympatric speciation mode, one daughter lineage inherits the entire range of its parent except for one region, while the other inherits

the entire range. In allopatric speciation mode, one daughter lineage inherits the entire range of its parent except for one region, while the other inherits the single region not inherited by its sister.

This cycle of dispersal and diversification repeats until the specified target diversity is reached or the specified maximum number of generations (cycles) has been run. When complete, the final output of the simulation is produced. The final output of the simulation consists of various kinds of summaries of the state of the system, including:

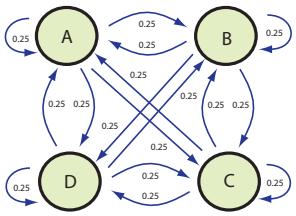
- A phylogenetic tree showing ancestor-descendent relationships of all the lineages in the system.
- An incidence matrix, showing presence-absence of the lineages in the regions of the system.
- A co-occurrence matrix, showing the overlap of ranges of lineages in the system.

This output provides the minimal data required to calculate various indices of beta and community diversity as well as phylogenetic diversity statistics. Future versions of the program will provide the facility to calculate and report these statistics natively, allowing for better pipelining workflows.

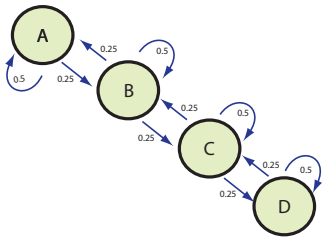
Application: Mid-Domain Effect

The mid-domain effect is an explanation for the observed latitudinal gradient in diversity, where tropical latitudes are seen to have higher numbers of species than temperate ones. The mid-domain effect has been justified analytically in terms of fitting a number of species with random range sizes into a finite-bounded space, resulting in the middle regions of the space accumulating more species than the marginal regions. Here I present the results of a study that demonstrates the mid-domain effect modeled purely in terms of random processes of dispersal and diversification under two different geographical geometries, without recourse to any specific pre-defined range size per species.

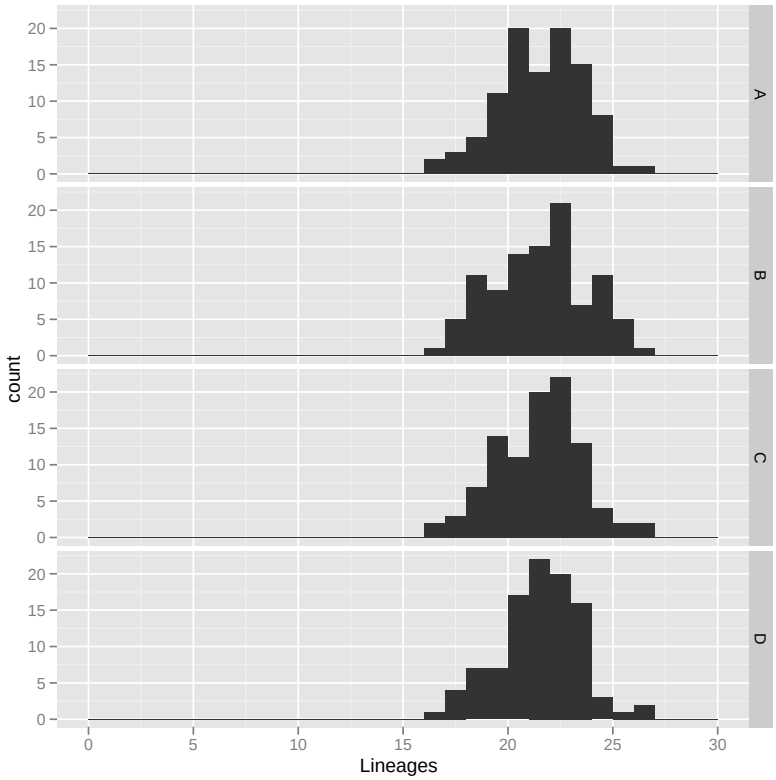
The two geographical geometries are "box" and "linear". In the "box" geometry, there is no distinct "mid-domain" as such: dispersal from any region to any other region is equal, and hence all regions are equidistant from each other in the Euclidean space induced by the dispersal probabilities:



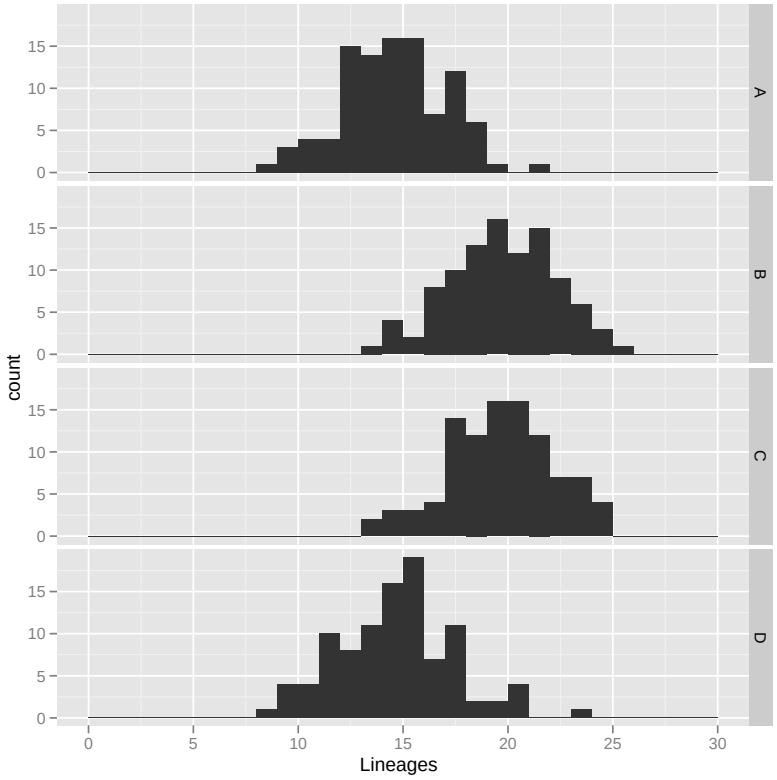
In the "linear" geometry, the dispersal probabilities are given such that there are distinct "satellite" (A and D) and "mid-domain" (B and C) regions:



A total of 100 replicates were run, with a birth rate, $\lambda = 0.1$, and a death rate, $\mu = 0.1$, and termination condition set to a target diversity of 30 species. The following plot shows the frequency distribution of the numbers of lineages in each region in order (A, B, C, and D) for the "box" configuration:



While the following shows the frequency distribution of the numbers of lineages in each region in order (A, B, C, and D) for the "linear" configuration:



While there is no clear statistical distinction between the two sets of distributions, it can be seen that there is a clear trend toward higher diversity in the mid-regions (i.e., regions B and C) in the linear configuration, whereas no such trend is evident in the box configuration.

These results are not particularly surprising given the model: the total flux into the marginal regions is exactly half that of the flux into the mid-domain regions, leading to an accumulation of higher total diversity. They do provide for an alternate justification of the mid-domain effect, however, in terms of random processes of diversification and dispersal.

Conclusions

By generating null distributions of diversity under the process of random (neutral) diversification and dispersal, and comparing these distributions using various summary statistics or indexes, such as the checkerboard score etc., to actual observed diversity, this simulator can be used to identify systems in which processes *other* than random diversification and dispersal are operating. There is a large scope for studies of this kind, especially in island archipelagos that are only recently beginning to be studied using modern phylogenetic

and ecological theory, such as the Philippine islands. The high diversity of the Philippine islands is unevenly distributed, with maximum diversity reached in different islands by different groups. As yet, there is no clear consensus on the process underlying this patter, or even if, indeed, there is any process apart from the historical contigency of stochastic colonization and radiation. By carefully calibrating the simulator to replicate conditions in the Philippines (i.e., diversity of groups of different sizes, dispersal/connectivity between the islands, etc.), and usage of suitable summary statistics (that should be able to index diversity both in terms of its richness as well as its distribution), it should be possible to determine whether one need investigate evolutionary or ecological processes as an explanation for the patterns.