Gingko: Forward-Time Simulation of Genealogical Trees Under Spatially and Environmentally Dynamic Biogeographical Histories

Jeet Sukumaran

1 Introduction and Overview

1.1 Objectives

GINGKO will simulate genealogical trees for multiple locii in populations of organisms evolving and interacting across a virtual landscape. The landscape will incorporate both spatial and environmental structuring, with the spatial relationships between discrete locations (*cells*) and the environmental characteristics of those locations, changing over time to simulate various biogeographical scenarios or histories. The spatial relationships between cells will effect the migration of organisms across the landscape, while the environmental characteristics of cells will effect the survival probabilities of organisms occupying those cells.

The trees generated by GINGKO, or DNA sequences simulated on those trees, will be used to answer the following questions:

- What is the effect of spatial structuring (barriers to migration) on divergence time and population size estimation methods, such as BEAST?
- What are the false positive and false negative error rates of Nested Clade Analysis given various phylogeographic histories?

1.2 Components

The full GINGKO pipeline will consist of three components:

1. A primary simulator program, which will carry out a single replicate of a forward-time simulation under a particular biogeographical history, and produce a file of ancestor identity histories for each neutral locus tracked in each species.
    - The GINGKO library will provide for landscape spatial and environmental data to be loaded dynamically from external ESRI ARC/INFO ASCII GRID files. A

schedule file will specify the suites of files to be loaded at particular generations to effect changes in spatial and/or environmental conditions.

- Organism classes in the GINGKO libraries will be specialized by client code specific to each simulation scenario, to control details such as mating and movement ecologies, fitness functions, etc..
- Thus, GINGKO will consist of multiple simulation programs, with each simulation program specialized in the particulars of the ecologies of its organismal agents, while landscapes and landscape histories are specified through different external input data files (landscape schedules and the ASCII GRID files).

2. A tree compilation script, which will take the ancestor identity history and compile a phylogenetic tree relating each allele.

3. A sequence generation script, which will generate DNA or protein character sequence data under various finite-state models on trees.

2 Simulation of Ancestor Identities Histories

- An ancestor identity history for alleles in a particular locus is simply a listing of ancestors for each allele in that locus in at a particular snapshot in time. Alleles are identified by an arbitrary number, guaranteed to be unique across all alleles within a particular locus across the entire lifespan of the simulation.