Tópicos

Depto. Ciencias Biológicas, UniAndes
Profesor Andrew J. Crawford <andrew@dna.ac>
Semestre 2009 -II

Lab – Coalescent simulation using *SIMCOAL*
17 septiembre 2009

Coalescent theory provides a powerful model for analyzing population genetic data. In phylogeography, we can use coalescent simulations to compare observed data with theoretical expectations under a given model. By simulating datasets under different models we can ask which model or models are compatible with our data and which models we may be able to reject. Today's lab provides us with the opportunity to observe the stochastic variation in gene trees that is inherent in the coalescent process. Stochasticity comes from two ancestral processes (1) the random joining or coalescences among lineages as we look back in time, and, more importantly, (2) the time interval between coalescent events (with exponentially longer waiting times as the number of lineages drops). We will observe that data simulated under even a simple population model reveals a surprising amount of variation in the topology and branch lengths of gene trees. This variation among data sets under a single model illustrates the danger in "over-interpretation" of a single inferred gene tree. We will also explore the process of incomplete lineage sorting and compare it with migration as a source of paraphyly and polyphyly.

Today's lab will provide instructions and tips on how to run *SIMCOAL*. The major difference between *SIMCOAL* and *SIMCOAL2* is that the latter allows recombination in the simulations, however the former appears much easier to run. Since we will not concern ourselves with recombination in today's lab we will stick to the older version (*SIMCOAL* version 1). This program only works with the Windows operating system. To visualize simulated gene trees, we will use the program, *FigTree*.

Software

SIMCOAL (for Windows 2000, Windows XP, and Linux)
`http://cmpg.unibe.ch/software/simcoal/`
FigTree (for Windows OS, Mac or Linux)
`http://tree.bio.ed.ac.uk/software/figtree/`

Citation

Schneider S, Roessli D and Excoffier L, 2000. Arlequin: a software for population genetics data analysis. User manual ver 2.000. Ver. 2.000. Geneva: Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva. *(I believe this is the official citation!)*

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.

Further information

Throughout this lab, recall that *SIMCOAL* has a help page online:
(http://cmpg.unibe.ch/software/simcoal/#Introduction)

GOALS:

Learn the basic options for coalescent simulation.
Observe stochastic variation in gene trees, and understand its cause.
Understand the difference between the true genealogy of samples and a gene tree estimated from the data.
Understand better the process of lineage sorting.
Understand the causes of incomplete lineage sorting.
Investigate the similarities and differences in the affects of incomplete lineage sorting versus migration on genealogies.
Invent potential uses of coalescent simulations in testing historical models.

Today we will run a basic coalescent simulation using *SIMCOAL*. We will first run simulations under a simple population model: the idealized Wright-Fisher population model (Fisher 1922, 1930; Wright 1931). In all models we will look at today, we assume that there is no natural selection affecting the molecular markers of interest. We further assume that all individuals are interchangable and all have an equal probability of leaving offspring. Note, while all individuals have an equal chance at reproducing themselves, not all individuals will. By random chance, some will produce more some fewer, some none at all. This variance in reproductive success is the single source of random genetic drift in a Wright-Fisher population (Charlesworth 2009). Sexual selection, for example, would violate this assumption, and would reduce the effective size, ***Ne***, relative to the demographic size, ***N***. In an ideal Wright-Fisher population ***Ne*** is equal to ***N***, by definition.

For our first model, we will look at the simplest case by assuming that there is no geographic population structure and population size is constant. We can add population structure or population size changes later. In all of today's simulations, we will assume that we are working with DNA sequence data and that we have a single gene sequence with no intragenic recombination. We will further assume that our sample size, ***n***, is small compared to the population size, ***N***. Under this assumption, we are safe in assuming that usually zero, sometime 1, but never 2 coalescent events happen in a single (ancestral) generation. This assumption simplifies the mathematics behind the model (Wakeley 2009). We won't look at recombination today, but those interested can try running *SIMCOAL2*, later on their own time (the features we need today happen to work much easier in the original, *SIMCOAL*, as far as I can tell).

We'll start with a single population with the following simulation parameters:

Number of "demes":	1 population
Population size, ***N***:	10,000 (or 5,000 diploid individuals)
Sample size, ***n***:	20 haplotypes
Population growth rate:	0 (constant population sizes)
Migration:	0
Historical events:	0 (e.g., population splitting, dispersal, expansion, etc.)
Type of molecular marker:	DNA
DNA sequence length:	1000 base pairs
Mutation rate per generation per gene:	0.0001 (i.e., 10E-7 per site)
Transition:Transversion rate:	0.66 (i.e., transitions are twice as likely)
Mutation rate heterogeneity among sites:	0.5 (4 rate categories)

To implement these conditions in a coalescent simulation exercise in *SIMCOAL*, the infile format is as follows:

```
//Parameters for the coalescence simulation program : simcoal.exe
1 samples to simulate
//Population effective sizes (number of genes)
10000
//Samples sizes
20
//Growth rates : negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration
matrix index
0 historical events
//Mutation rate per generation for the whole sequence
0.0001
//Number of nucleotides to simulate
1000
//data type either DNA, RFLP, or MICROSAT : If DNA, we need a second term for the transition
bias
DNA 0.81
//Gamma parameter (if 0: even mutation rates, if >0 :shape parameter of the Gamma distribution
0.50 4  // Second parameter is the number of discrete rate categories to simulate: if zero:
continuous distribution
```

Text and comments to the right of the // are free to vary. To prepare an infile, the safest and quickest method may be to open the example file, "testdna1.par," that was distributed with the program, save under a new and more descriptive name, modify your new **infile** appropriately, and save again. Be sure you save the file with the ending ".par"

To keep your infiles and outfiles well organized, I recommend you create a new directory (*carpeta*) with the same name as your infile, followed by the text: "_OUTFILES". Place your new infile in this directory, along with a copy of the program, *simcoal.exe*.

For a statistical test using simulation, you should run perhaps 10,000 simulations to estimate a null distribution. Today, however, we will just be using visual inspection of gene trees estimated from simulation, so 10 – 30 replicate simulations would be plenty. To run the program, double click on the executable (*simcoal.exe*), which will open the command prompt:

1) Type the name of your infile but without the ".par" suffix (ending).
2) Type the number of simulations to run. Try: **10**

If the program finished correctly (takes only a few seconds), *SIMCOAL* will have created 15 new files for you:

1) A numbered ".arp" file for each simulation (10 in this case, numbered 0-9)
2) A batch file ".arb" usable with the software, *Arlequin*
3) A NEXUS file ".paup" hopefully readable by PAUP or other phylogeny programs.
4) Summary of pairwise distances and age of genealogies among simulations in a ".gen" file.
5) Tree file containing true genealogy obtained from each simulation, ending with "_true_trees.trees"
6) Tree file containing genealogy inferred from mutations added to genealogy according to our specified mutation model. File ends with "_mut_trees.trees"

I. Stochastic variation among gene trees

Identical conditions can produce very different gene genealogies. The coalescent process includes an element of chance, or stochasticity. Lineages coalesce at random. The time between coalescent events

follows an exponential distribution, resulting in a sizable variance around the expected coalesence times. To observe this variation, select the outfile that ends with "_true_trees.trees" and open it with the application, *FigTree*. We simulated 20 samples (10 times), which are labeled on the right. The ".1" refers to the population number (we simulated only 1 population). Notice the scale bar at the bottom of the first tree figure, and observe how it changes among trees. Scroll through the 10 genealogies (of 20 samples each) we simulated by using the arrows in the upper menu bar, above "Prev/Next".

Recall from lecture (hopefully) that the expected time to the final coalescence (2 lineages coalescing into 1) equals half the total expected depth of the genealogy. How often do you observe this expectation? In other words, how often (among simulations) does the more ancestral half of the tree (the left half in *FigTree*) include more than two lineages? Alternatively, how often to you see two ancestral lineages occupying **more** than half the total depth of the tree?

In a balanced tree, we would observe 50% of samples descending from each of the two ancestral lineages. How often to you observe a balanced tree? What's the minimum number of samples you observe descending from one of the two ancestral nodes? Does the basal splitting (coalescent) event ever separate (join) 1 sample from (to) the other 19 samples? Would you say this one sample is "basal" or "ancestral" to the other 19? Why or why not?

Coalescent theory predicts that the rate of coalescent events slows as the number of lineages declines. Do you notice a faster rate of coalescences at the tip of the trees? Are there exceptions? Do you notice any polytomies in any genealogies? Should we expect to observe polytomies? Why or why not? Notice the lengths of the tip (external) branches. Would you expect to be able to detect even the shortest branches if you attempted to infer this "true" genealogy from DNA sequence data?

II. True genealogies vs. inferred gene trees

Previously we were looking at simulated "true" genealogies. We can never know the true genealogy of a sample of DNA sequences. We can only hope to estimate it from the variation among DNA sequences (or other molecular markers). The amount of mutations in our simulated DNA sequence datasets is determined by the mutation rate we selected, and by the population size. Bigger populations should have deeper genealogies, and deeper trees have more total history (length) where mutations might be observed.

To observed the genealogies as estimated from the simulated mutational process, close the "._true_trees.trees" file and open the file that ends with "_mut_trees.trees" again in *FigTree*. Flip through the 10 trees, briefly. You might see from 0 to 13 mutations in the genealogy. How can you count mutations? Note the shortest non-zero branch length among trees. This length should equal 1 inferred mutational difference. If you like, try looking at unrooted trees using the 3rd button under "Layout" at the top of the vertical menu on the left of the screen. Which view do you prefer?

Notice there is not much "phylogenetic information" for estimating the full genealogy. Many samples have identical haplotypes. This situation is quite normal for a sample of haplotypes from a single population. Our simulation was somewhat accurate for mtDNA in vertebrates, with a mutation rate of 10E-7 per site in a population of 2N=10,000. The expected average pairwise distance among samples is $\boldsymbol{\pi} = 4\boldsymbol{N\mu} = 4(5000)(0.0000001) = 0.002$ per site, or 2 mutations among 1000 base pairs. How does

this expectation compare with the simulation? Look in your file ending with ".gen" and look at the mean and S.D. of number of pairwise differences. In my set of 10 simulations I observed a mean of 1.28 (SD=0.716), which is reasonably close to expectations.

III. Lineage sorting

Now we will simulate a pair of diverged populations. Thus, not all individuals will be equivalent. *SIMCOAL* will indicate each simulated individual by a sample number followed by a ".1" for population 1 and ".2" for population 2, of course. Models where not all individuals are equivalent are referred to as "structured coalescent" models. We will simulate a simple case of a single ancestral population dividing into two daughter populations at some time in the past. Time we will measure in numbers of generations. Later we'll add migration between our two daughter populations, but not yet.

We can use a structured coalescent model to help us visualize the process of lineage sorting and the parameters that affect this process. Lineage sorting is one of the most important processes in phylogeography. The process is fundamentally very simple, yet at the same time difficult to understand or conceptualize. *SIMCOAL* asks use to set time in generations, but coalescent theory allows us to count the number of generations in units of ***N***, the population size (since the rate of coalescence is tied to genetic drift which is determined by ***N***). We can set up 2 simulations, the first will assume a time of separation ***N*** generations in the past, and the second simulation will assume an older pair of populations separated 3***N*** generations ago.

Prepare 2 new directories, appropriately labeled (perhaps "2pop_young_OUTFILES" and "2pop_old_OUTFILES"). We might raise the mutation rate a bit and/or increase the length of DNA sequences, just to obtain empirical gene trees (mutation trees) with more potential phylogenetic information. The first infile might look like this:

```
//Parameters for the coalescence simulation program : simcoal.exe
2 samples to simulate
//Population effective sizes (number of genes)
10000
10000
//Samples sizes
10
10
//Growth rates : negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration
matrix index
1 historical event
10000 0 1 1 1 0 0
//Mutation rate per generation for the whole sequence
0.0001
//Number of nucleotides to simulate
10000
//data type either DNA, RFLP, or MICROSAT : If DNA, we need a second term for the transition
bias
DNA 0.8
//Gamma parameter (if 0: even mutation rates, if >0 :shape parameter of the Gamma distribution
0.5 4  // Second parameter is the number of discrete rate categories to simulate: if zero:
continuous distribution
```

Place a copy of the executable (program) in your folder (*carpeta*) and run this infile as explained above, but perhaps do 10-30 simulations. If you haven't done so already, close/quit *FigTree* before

using it to open your new trees. In your new "true" trees, notice the .1 or .2 following the sample number, these designate the population of origin of each sample. How many of your ~20 simulated trees show reciprocal monophyly of the two populations? How many show paraphyly of one population (with the other population monophyletic)? How many show polyphyly? Enter your results in the third column (I did 30 simulations):

Topology	Div time = N AJC's results	Div time = N Your results	Div time = 3N AJC's results	Div time = 3N Your results
Reciprocal monophyly	**7**		**23**	
Paraphyly of 1 population	**14**		**6**	
Polyphyly	**9**		**1**	

Now look at the trees inferred from the mutational history. Can you determine the proportion of monophyletic, paraphyletic and polyphyletic trees? Are these results different from the "true" genealogy file? Why or why not?

Now copy your infile to the other folder (e.g., "2pop_old_OUTFILES"), and save under a new name. Edit the age of the historical event (population splitting event in forward time, coalescent event in backwards time) from 10,000 to 30,000 (3***N***) generations. Review your resulting genealogies. Tabulate the frequencies of monophyletic, paraphyletic and polyphyletic genealogies. Can you explain WHY you (probably) observed more monophyletic trees when the splitting even happened longer ago in the past? If you have any non-reciprocally monophyletic trees, look at the ancestral lineages. Do you have >2 lineages extending close to the root of the tree? Is there any reason why you might expect more ancestral lineages extending farther back in the tree for those trees that are **not** reciprocally monophyletic?

IV. Migration vs. lineage sorting

Lack of monophyly between two sister populations could be due to incomplete lineage sorting, or it might also be due to migration. In this example, we will simulate 2 sister populations that diverged 5***N*** generations in the past. In the absence of migration, these populations are likely to be reciprocally monophyletic. However, in this simulation we will add a recent dispersal event from one population to the other at time = (***N***/10) generations ago. During the dispersal event, each member of one population has a 0.2 probability of migrating to the other population, but population sizes will remain the same. Modeling migration with population sizes that remain constant is referred to as "conservative migration" (e.g., Nagylaki 1998).

Create a new directory, labeled e.g., "2pop_MIG_OUTFILES"), and copy the executable plus the infile (".par" file) into it. Leave most of your simulation parameters the same, but change the historical events as follows:

```
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration
matrix index
2 historical event
1000 0 1 0.2 1 0 0
50000 0 1 1 1 0 0
```

Run the program and examine your "true" genealogies as before. How many of your simulated genealogies show reciprocal monophyly? Paraphyly? Can you detect any different in genealogies (in tree shape or branch lengths) between trees that were non-monophyletic due to the recent dispersal event versus trees that were non-monophyletic due to incomplete lineage sorting (in the above exercise)?

Can you think of a way of potentially distinguishing non-monophyly due to incomplete lineage sorting versus non-monophyly due to recent dispersal or introgression? Could you use *SIMCOAL* to do a power analysis – i.e., explore the question of how much data you would need to distinguish between certain historical scenarios or hypotheses? Most studies of incomplete lineage sorting versus introgression do not compare directly the two models; rather they assume a no-migration model as the null hypothesis and try to reject that. For example, one can estimate time and population size from the data, assume these values plus the null hypothesis of no introgression in conducting coalescent simulations, and then ask what is the chance of observing non-monophyly in the absence of migration among the replicate simulations? If non-monophyly appears to be highly unlikely in the absence of migration, under the simulated conditions, then migration is inferred (e.g., Buckley et al. 2006). Given the flexibility of statistical testing by simulation, however, the student has the opportunity to create novel tests of hypotheses.

References:

Buckley TR, Cordeiro M, Marshall D, Simon C. 2006. Differentiating between hypotheses of lineage sorting and introgression in New Zealand alpine cicadas (*Maoricicada dugdale*). *Systematic Biology* 55: 411-425.

Charlesworth, B. 2009. Fundamental concepts in genetics: Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics* 10: 195-205.

Fisher, RA (1922) On the dominance ratio. Proc. Roy. Soc. Edinb. 52: 312–341.

Fisher, RA (1930) The distribution of gene ratios for rare mutations. Proc. Roy. Soc. Edinb. 50: 205–220.

Nagylaki, T (1998) The expected number of heterozygous sites in a subdivided population. Genetics 149: 1599–1604.

Wakeley, J (2009) *Coalescent Theory: An Introduction.* Ben Roberts, Greenwood Village, Colorado.

Wright, S (1931) Evolution in Mendelian populations. *Genetics* 16: 97-159