From Sequences to Population Parameters: Using LAMARC for Data-Mining

Lucian Smith, Mary Kuhner
University of Washington, Department of Genome Sciences
Poster available at: http://evolution.gs.washington.edu/lamarc/Tutorial_poster.pdf

Figuring out the right Question

Issues to Consider

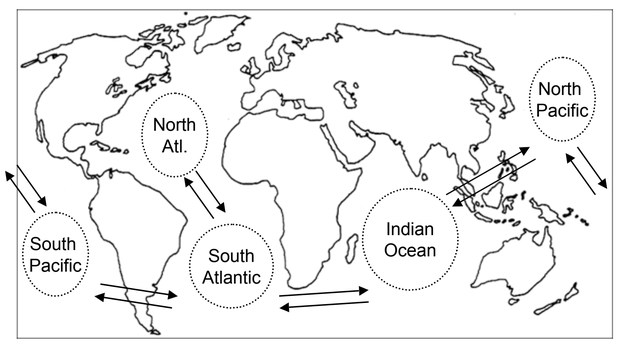
LAMARC can estimate:

Population size/mutation rates (θ) per population,
Growth rates per population,
Migration rates between populations, and
Recombination rate of your organism as a whole.

All of these parameters can be estimated at once, but figuring out which parameters you're interested in, and which parameters are merely nuisance parameters can help guide your collection of data.

The Hypothetical Experiences of Walter Wilsford, Whale-Watcher

Walter is studying a newly-classified species of surface-dwelling whales, *Balaenoptera obscurificus*. He is about to go on a world-wide expedition, and will be collecting samples from whales from five different areas. He suspects the populations have been shrinking, so wants to measure growth rate in all five areas. He's not as interested in the migration rates between the populations, nor does he wish to estimate recombination rates.



Collecting Data

Certain types of data contain more information than others. The amount of new information you collect as you sample more individuals, for example, goes down with each new individual. The number of possible genealogical topologies increases much faster than exponentially with each new individual, meaning that you have to spend more and more time searching for less and less added information with each new sample--a reasonable upper limit is 25 samples per population, or 12 individuals for a diploid species.

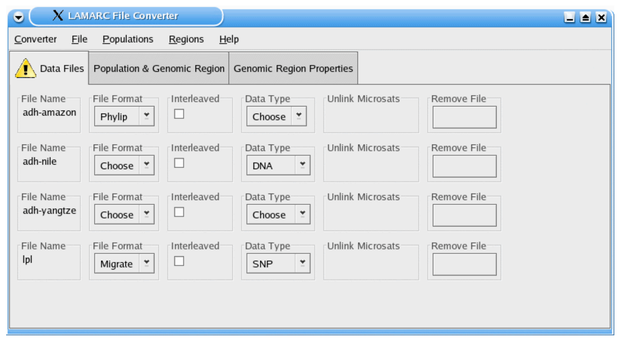
The information present at a new genetic locus, however, is huge--it reflects a completely independent sample from the history of the species. As such, it is almost always better to spend your effort collecting data from new loci than new individuals.

If you're studying recombination, a new locus might help, but what will give you the most added information is extending your sequence, to better pick up locations of unique recombination events.

Walter sends his two graduate students on expeditions, and between them they manage to collect samples from seven to ten whales from each population they wished to study. Back in the lab, Walter then develops primers for two short stretches of DNA from different chromosomes, and begins sequencing. Since growth is a particularly difficult parameter to estimate, Walter knows two loci will not be nearly enough, but fortunately for him, his friend and colleague who has studied the Minke whale in the past offers to give him primers for the eight microsatellite loci she studied in her research, so he adds those to his study.

Converting Data to LAMARC input

Once you have your data, you can convert it into LAMARC-ready input by using our file converter (shown to the right). Once you have done this, you can run LAMARC and use its menu system to give it more information about your data--what evolutionary data model you wish to use, whether you want to constrain any nuisance parameters to be constant (so as to not waste time estimating them), etc. You can also decide at this point whether to perform a Bayesian or a Likelihood analysis (or both!)



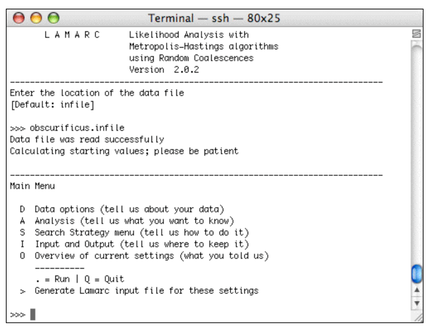
Walter now plugs his data into the converter, and ends up with a LAMARC input file. When he opens this file in LAMARC, his first order of business is to constrain the migrations. He knows that it is highly unlikely that whales from the North Atlantic will migrate directly to the North Pacific, for example, so he constrains those migration rates to be 'invalid'. He does the same for a variety of other migrations, until he is left only with those shown on the above map. He also decides he will run both a Bayesian and a Likelihood analysis, so he can compare the results.

Running Preliminary Analyses

It can take a bit of familiarity with both your data and with the LAMARC program before you get it to work well. Because of this, it's a good idea to make a few exploratory analyses that only take a few hours to complete before doing a final analysis that can take a week or even months. The default settings for chain lengths are a decent place to start, though if you have a lot of data even those might be a bit long. A final Bayesian analysis is best set up as one long chain, but for these preliminary runs, having multiple chains so you can keep track of the estimates is fine.

You will probably also want to turn off profiling for your exploratory likelihood analyses, since this can take a significant portion of your run-time.

Finally, you may wish to analyze each genomic region separately, so you can get a sense of how much information it contains, and how long it takes to get a good estimate.



Walter sets up three preliminary analyses--one for his eight microsatellite loci, and one for each of his DNA sequences. He isn't interested in recombination, but turns it on for his DNA runs to see if it is estimated to be sufficiently high to interfere with his other results (it isn't). He also experiments with turning off growth estimation, but leaving migration completely unconstrained, to make sure there's no significant migration where he thinks there should be none. He does a literature search to discover the relative mutation rates between his microsatellite and DNA data, and notes that they seem to be reasonably consistent with the estimated θ values from his different genomic regions. Finally, he sees how repeatable his results are with different random number seeds, to get a sense of how long his final analysis might have to take.

Troubleshooting

One of the reasons to do preliminary analyses is that if you start seeing oddities, you can address the issues without having spent a week's worth of computer time. Some possible issues include:

Your analyses are not consistent from one run to the next (solvable by longer runs, replication, and/or heating)

Your constraints are not appropriate, or can be expanded

Your results for one region are not consistent with the results for a different region (relative mutation rates may be required)

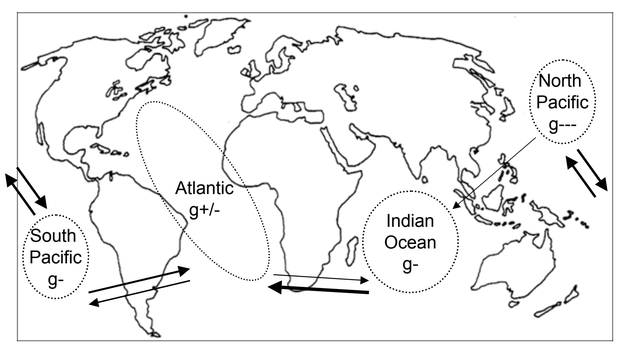
When Walter starts running his analyses, immediately he starts running into trouble. Even when constrained, some of his migration rates are sky-high, and when unconstrained, he has significant migration between populations that aren't supposed to be connected. Finally, he throws his data at 'Structure' (Pritchard, et al. 2000) and discovers that his two putative Atlantic populations are actually genetically a single population. He collapses his samples from the Atlantic, re-does the analysis, and finally things start making sense. Additionally, he discovers the repeatable result that there is little to no migration from the Indian ocean into the Pacific, though small levels of the reverse can be seen. At this point, his results are fairly consistent; enough so that he hopes a longer run will solve any lingering consistency problems.

The Final Analysis

Once you know how long a good run of LAMARC is going to take, you make sure your computer has enough memory,and that it's not going to crash any time soon, and start your final analysis. If you have extra computers and want to check reproducibility, you might start up multiple analyses with different random number seeds, and/or different runs with Bayesian or Likelihood analyses. It will probably behoove you to turn on percentile profiling as well for a comprehensive error analysis

Analyzing your output can be as simple as looking at the reported estimates and confidence intervals for your parameters. In a Bayesian run, you will also have 'curvefiles', which are pictures of the probability distributions of each parameter. In a Likelihood run, you don't have the pictures, but you can tell if your parameters are correlated--growth and theta values, for example, tend to be positively correlated with each other, while migration rates between two populations might be inversely correlated with each other.

After finally figuring out the appropriate constraints, Walter takes over his department's cluster for a few weeks, and runs three Likelihood and three Bayesian analyses of his data, each with different random number seeds. To his great relief, they all agree with one another to the limits of the estimated confidence intervals. The confidence intervals for his estimates of growth are larger than he would like, but he can at least tell that the population of his *Balaenoptera obscurificus* has clearly been shrinking in the North Pacific, has seen more moderate decreases in the Indian Ocean and South Pacific, and has been fairly stable in the Atlantic. He can also tell that the North and South Pacific populations exchange a lot of migrants, and that there is a definite influx of whales from the Indian Ocean into the Atlantic.



His Likelihood analysis shows only weak correlations between his parameters, so he can feel confident that the map on the left is an accurate representation of his data. He gets his graduate students to write up the conclusions, and publishes.