LAMARC exercise: Module 21, Summer Institute 2014

Getting the programs

If you cannot get the programs from the SISG web site, they can be downloaded as follows:

- `Lamarc` and `lam_conv`, the lamarc file conversion utility, both from http://evolution.genetics.washington.edu/lamarc/lamarc.html (pick the executable appropriate to your machine)
- `Tracer`, found at http://tree.bio.ed.ac.uk/software/tracer/ (pick the appropriate source for your machine)

`Lamarc` and `lam_conv` executables will load directly from the web site. `Tracer` is a bit more complex. On a Linux machine, download the compressed file and double click on the tracer.jar file, as per the instructions. But you then need to change to the bin directory and turn on the execution flag of `Tracer` (chmod +x tracer) before it will run. After that you can either run `Tracer` from the command line or double click on it from a window, whichever you prefer. On a Mac, using the Finder, click on the downloaded files (.dmg extensions) to extract the files into a mounted folder. Then copy the mounted folder to a known location (for example, the Desktop). When you run these programs, you will double-click on the icon in your own folder, or the icon on the Desktop.

Getting data files

From the WWW, navigate to http://evolution.gs.washington.edu/lamarc/sisg-2014/demo/ and download all files from the **demo** link:

File	Description	Data Type	Pops	Format
duck_onepop.phy	Single population duck OD7 sequences	DNA	1	Phylip
duck_onepop.xml	Single population duck OD7 sequences	DNA	1	Lamarc
duck_short.xml	Two-population duck OD7 sequences	DNA	2	Lamarc
lpl_short.xml	reduced lipoprotein lipase	SNP	3	Lamarc
lpl.xml	lipoprotein lipase	SNP	3	Lamarc
lpl.mig	lipoprotein lipase	SNP	3	Migrate
convert_lpl.xml	utility file for LPL data	–	–	Lamarc

Alternatively, using SCP or another secure copy program:

host: `griffiths.gs.washington.edu`
user: `summerinst`
password: `summer`
folder: `demo`

These resources will only exist for the length of this module. If you would like copies of these files to do the exercises on your own, make copies now.

Data set one: single population data

We will examine two South American duck species collected by Dr. Kevin McCracken: the lowland Silver Teal (*Anas versicolor*) and the highland Puna Teal (*Anas puna*). For most of our runs we will use only the lowland data. The sequences are from an intron of the nuclear gene OD7.

Kevin's original data set contains 154 sequences, far too many for quick results. I randomly cut the data set down to 30 sequences from each species. Please note that such removals MUST be random. It is tempting to remove duplicate sequences as they are "less interesting" but doing so creates a severe bias.

File conversion

NOTE: On some Linux systems the file converter (lam_conv) executable needs a library that it can't find: the result is an error message mentioning "libexpat". This may be fixable by typing the following (and providing a password when requested):

sudo ln -s /usr/lib/libexpat.so /usr/lib/libexpat.so.0

If this does not work, please skip the converter and use the pre-converted file duck_onepop.xml.

Input files for the `lamarc` executable are produced using `lam_conv`, our file converter utility. We will create an input file using the `duck_onepop.phy` sample file.

- Start up `lam_conv`
- Use the mouse[1] to invoke the `File > Read Data File` menu commands and get a file browser.
- Navigate to the location where you downloaded your data files, select the file `duck_onepop.phy` and click `Open`.
- Find the box labeled `contiguous segment` and note that the datatype is represented as three question marks, "???". This is because the file converter is unable to tell if this file contains DNA or SNP data.
- Double click in the box labeled `contiguous segment`. A new window should appear. Click the check box labeled `DNA` within the `Data Type` pane and then click `Apply`.
- To save your converted datafile, choose `File > Write Lamarc File`, navigate to the folder/directory where you'd like to place your file, choose a name for your file,[2] and click `Save`.
- You're done with `lam_conv` for the moment so go ahead and close it.

If you are unable to convert the file, a pre-converted input file is available as `duck_onepop.xml`.

[1] Keyboard shortcuts and navigation do not yet work seamlessly on all dialogs and platforms. Please feel free to experiment with them, but use the mouse when in doubt. Shortcuts are displayed on the right hand side of menu items. Keystroke navagation is enabled on Windows and Linux via the `<Alt>` key.

[2] You may place the file in whatever directory you like, just remember the name and location! On the Mac, the easiest place is the folder in which the program resides.

Exercise 1: Estimating Theta alone

To use the Lamarc menus, enter the single character in the leftmost column (case does not matter) and hit Enter or Return. Be aware that the meaning of a command is context sensitive, so `S` may mean different things on different screens: look at the context to be sure. Also, watch out for letter l versus number 1, and letter O versus number 0. One other wrinkle is that when you choose a toggle (for example `X` on the Growth Rate menu) nothing obvious may happen. Look closely and you'll see that the value associated with `X` has switched to the other option.

If you land on a menu you don't expect, you can use the `<enter>` / `<return>` key to back out.

Run `lamarc` and when asked for the location of a data file, give it the full name of your Lamarc input file. Be careful to include extensions such as "xml" or "doc" at the end of a filename.[3] After successfully entering the file name you should see a text–based menu with the heading `Main Menu`.

The default run length is generally too short for publication, but too long for a demo! We'll adjust the run length using the following steps:

- From the `Main Menu`, choose `S` to get the `Search Strategy` menu.
- Under the `Search Strategy` submenu, choose `S` again to get the `Sampling Strategy` submenu.
- Use menu item `4` to change the number of samples to discard from initial chains to `100`.
- Use menu item `8` to change the number of samples to discard from final chains to `100`.
- Use menu item `6` to change the number of recorded genealogies in the final chain to `1000`.

Since we know this is mammalian nuclear DNA, the default TTRatio value should work. For mtDNA we would want to increase it substantially. To check this:

- Press the `<Return>` key twice to go back to the main menu.
- Select submenu `Data options`.
- Select submenu `edit data model(s) for Region 1` (this is the "one" digit).
- Use option `T` to change the `TT Ratio` value. We might also want to look at the base frequencies. They are calculated from the data, which is usually fine but might be questionable if the sequences are very short. [4]
- Run the program! (enter `.` the period) Progress reports should appear, including a guess as to how long the analysis will take.

At the end of this section are some tips for deciding whether your run is satisfactory. You may want to consider these as you watch the progress reports; you will certainly want to consult them when the run has finished. Your output will be in `outfile.txt` unless you told the program to rename this file. (On a Mac, the file will be in the same folder as your lamarc.app subfolder.)

[3] To see file extensions under the Windows OS, open any folder then choose "Folder Options" from the "Tools" drop down menu, then open the "View" tab and uncheck the box for "Hide extensions for known file types".

[4] PAUP* is an ideal tool for finding an optimal data model for your data; I strongly encourage its use.

Questions for the Θ-only Silver Teal data run

- What is the Maximum Likelihood Estimate (MLE) of Theta?
- What are the 95% confidence interval boundaries for this estimate?

Validation

TRACER is not very useful for a likelihood-based run such as this one, so we will look at LAMARC's own diagnostics.

Examine the output file from your run, looking for the following symptoms (you won't find them all in this data set, this is a general check list):

- Parameter estimates are following a trend throughout the run (increasing or decreasing) rather than settling around a value. *Run is too short. Try more chains.*
- The entry "Data lnL" in the run reports, which shows the log likelihood of the best tree found during that chain, follows a trend throughout the run rather than settling around a value. *Run is too short. Try more or longer chains.*
- The entry "Posterior lnL" in the run reports, which shows the degree to which the trees in that chain improve on their starting values, is greater than 2x the number of parameters being estimated even in the final chain. *Run is too short. Try more chains.*
- Results of two runs with the same data and settings, but different random number seeds, are not similar. (Compare the between-run difference with the confidence intervals.) *Run is too short, or data are uninformative. Try more or longer chains; if this doesn't work, reconsider data.*
- Estimates fluctuate wildly from one chain to the next. (Compare the between-chain difference with the confidence intervals.) *Run is too short, or data are uninformative. Try longer chains; if this doesn't work, reconsider data.*
- Acceptance rate is above 50% or below 5%. *A high acceptance rate suggests that the data are mutationally saturated. In practice, this is often an alignment failure. A low acceptance rate may be an unavoidable feature of some data sets, but it suggests the need for particularly long runs.*

Thought questions

- These are data from only the lowland population. Does this make our estimate of Θ questionable?
- We are assuming no population growth and no recombination. Are these assumptions defensible? How could they influence our results?
- These estimates of Θ are generally higher than estimates for human (around 0.002). How could this be explained? Are Silver Teal more abundant than humans? What other factors might be involved?

Exercise 2: Estimating Theta and growth rate

We're going to repeat the analysis, but this time estimating population growth as well as Theta.

You may want to use the "menusettings_infile.xml" produced by the previous run as your starting point, as it will already have the desired ttratio and search strategy.

- Save your previous results by either copying/moving/renaming the `outfile.txt` just created, or by telling `lamarc` to save its output to a file with a different name (change this in the `Input and Output` top level menu).
- Read in `menusettings_infile.xml` from the previous run. (Did you save a copy first? You might wish to as this run will overwrite it.)
- From the `Main Menu` choose the `A Analysis` and then `G Growth rate` menus.
- Turn on estimation of growth using the `X` option. This is a toggle so no data entry is required. The display will reflect the change.
- Run the program.

Questions for the Silver Teal run with growth

- What is the MLE for Theta?
- What are the 95% confidence interval boundaries for this estimate?
- What is the MLE for the growth rate g? Is this growth or shrinkage?
- What are the 95% confidence interval boundaries for this estimate? Can we reject either growth or shrinkage?
- Is there a lot of correlation between Theta and g?
- How much does this estimate of Theta differ from one made without taking growth into account?
- Could the apparent signature for growth be due to something else? What dangerous assumptions are we making?

Validation

Consider the points raised for Exercise 1 above. An additional potential problem:

- The estimates of both Theta and g are enormous and the upper bounds of the confidence intervals are practically infinite. *If the population has grown too rapidly, the tree becomes star-like and not enough information is present to co-estimate Theta and g. Additional unlinked loci might help. Otherwise, you can estimate Theta while holding g constant, or g while holding Theta constant, but not both at once. If multiple time points are available, consider using* `BEAST` *instead.*

Exercise 3: Estimating subpopulation Thetas and immigration rates

To estimate subpopulation parameters you will need data with more than one population. We provide a file `lpl.xml` with SNP data from the lipoprotein lipase (LPL) locus in three human "populations:" European Finns from North Karelia; assorted Europeans from Rochester, Minnesota; and African-Americans from Jackson, Mississippi. A shorter version, better for class exercises, is in `lpl_short.xml`. This is a complete `Lamarc` infile to perform a Bayesian analysis with one final chain sampling 7000 parameter values split randomly among all of the available parameters.

If you'd like to experiment with some of the more complex features of the data converter, we provide the raw migrate format file containing the full dataset in `lpl.mig`, plus the xml command file for use with the converter in `convert_lpl.xml`. Depending on your platform, you may need to place both these files in the same directory/folder as your executable.

- Start up `lamarc` and read in file `lpl_short.xml`.
- Navigate through the `Search Strategy` and `Sampling Strategy` menus and verify this is a Bayesian run.
- Run the program.

An important point to remember is that LAMARC estimates rates of **immigration**. (Random emigration has no substantial effect on a population's genetic makeup, whereas immigration brings in new haplotypes and can have a huge effect.)

Questions for LPL without recombination

- What is the Most Probable Estimate (MPE) of Theta for each population? (Note that this is a Bayesian run so reports MPE and not MLE.)
- What are the 95% support interval boundaries for these estimates?
- Draw a migration diagram and note the size of each immigration rate M.
- Draw a second diagram, but instead of recording M, record M times the Theta of the recipient population (this is 4Nm, an intuitive measure of migration–it measures the number of individuals migrating in per generation). How do your impressions of the results change with this rescaling?

Validation

All of the validation tests from earlier exercises are relevant here. Since this was a Bayesian run, we can additionally look at the results using a helpful tool, `Tracer`, written by Drummond and Rambaut. `Lamarc` writes files with "tracefile" in the name that are ready to be read by `Tracer`.

Open the program `Tracer`. (If this doesn't work, make sure its executable flag is turned on.) Use the `File > Import Trace File` menu item to read `tracefile_LPL_1.txt`. There is a lot of data on this screen, it helps to focus on the subsections:

- View the trace and estimates for `Ln(Data Likelihood)`.
- Pick some of the other data statistics (middle left column) to see how their traces and estimates vary.
- Notice there are a lot of useful summary statistics in the top area above the Estimates plot.

[Note here for R/Matlab users: The curvefile_???.csv and profile_???.csv generated by Lamarc are easily analyzed and plotted in R and will give you different insights into the data. That is outside the bounds of an introductory class but can be very useful. If you use neither R nor Matlab, you may want to look into R as it is free and provides a wealth of statistical methods useful in all areas of biology.]

Potential problems:

- Parameter traces show a trend across the whole run. *Run the program longer.*
- Histograms (on Estimates tab) of parameters show multiple peaks. *Run the program longer. This can also, rarely, be a real feature of the data. Some data sets support two different migration structures; often, this takes the form of support for strong migration from A to B, or from B to A, but not both.*
- Values are piled up against one side of the histogram. *Unless it is a natural boundary such as zero, your prior may be poorly chosen: either much too wide, or not including all of the plausible values.*
- ESS (Effective Sample Size) below 200 for some parameters. *Run the program longer. Also consider whether you are trying to estimate too many parameters, or a parameter for which the data are uninformative. For example, it is useless to try to estimate recombination rates from very short sequences.*

It is a sad truth that while bad TRACER results nearly guarantee that the run was unsuccessful, good TRACER results do not prove success. Still, TRACER is a powerful validation tool; it catches a significant proportion of bad runs, and journal reviewers are beginning to ask for ESS as a measure of reliability for MCMC runs.

Exercise 4: Estimating Theta and recombination rate

If you have time, try estimating recombination rate from these data. We suggest using the previous run's `menusettings_infile.xml` as your new input file.

- Start up `lamarc` and read in file `menusettings_infile.xml`. (Did you make a copy? This one will get overwritten when the run starts.)
- Navigate to the `Analysis` and then `Recombination` menus.
- Turn on recombination.
- Use option `B` to view the Bayesian prior on recombination rate.
- The default prior is probably too wide for human data. Use option `1` to get to the `Bayesian Priors Menu for RecRate` and set the upper and lower bounds to more reasonable values, say an upper bound of 5.0 and a lower bound fairly close to zero, like 0.0001. (You can experiment with a linear prior here if you like; then the lower bound can actually be zero.)
- Run the program

Questions for LPL with recombination

- What is the MPE of Theta?
- What are the 95% support interval boundaries for this estimate?
- What is the MPE of the recombination rate r?
- What are the 95% support interval boundaries for this estimate?
- How much does this estimate of Theta differ from one made without taking recombination into account?
- (If you have time) How does changing the prior for r change these results?

Validation

This run should also be tested with `Tracer`. Runs with recombination generally need to be much longer than runs without, as the space of possible ancestral recombination graphs is vast.

Particular problems for this type of run:

- Sudden leaps in the estimate of a parameter, usually the recombination rate. *These indicate that the search has found a new area. They don't invalidate the analysis, but they should fall early in the run; if they were still happening late in the run, it should have been longer.*

- Multiple peaks in the estimate of a parameter, usually a migration rate. *These often indicate a too-short run, but they can also show a real feature of the data. Some data sets are genuinely ambiguous, supporting asymmetrical migration but not defining its direction. Multiple peaks in Theta, however, are almost certainly a too-short run.*
- Long "knees" on the lower side of the histogram for a parameter, usually the recombination rate. *If the prior puts a lot of weight on very low values of the recombination rate which the data cannot actually distinguish, you can get a "knee" which is due more to the prior than to the data. In general, a Bayesian posterior which looks like its prior is cause for concern! You may have an inappropriate prior, or little information in your data. For this particular data set, a flat prior on r may be better than a log prior.*

Thought questions for the LPL runs in general

- What dangerous assumptions are we making? In particular, are the data likely to violate our migrational model?
- If the runs are unsatisfactory, how could they be improved? (This will be explored a later session.)

LAMARC glossary

- Population parameters:
  - **Mutation rate** (μ) – neutral mutation rate per SITE per generation. Be careful when comparing with older studies which often use mutation rate per LOCUS per generation.
  - **Effective population size** (N_e) – population size of an idealized Wright-Fisher population with the same rate of genetic drift as the actual population. Usually less than the census size.
  - **Theta** (Θ) – effective population size times neutral mutation rate times a constant. In diploids, Theta=$4N_e\mu$. In haploids, $2N_e\mu$.
  - **Migration rate** (M) – Probability that a lineage migrates divided by probability that it mutates (m/μ). For analysis, it is often more useful to look at $4Nm$, the expected headcount of migrant chromosomes per generation. You can obtain $4Ne$ by multiplying `Lamarc's` estimate of M by its estimate of the recipient population Theta. Population genetics theory suggests that when $4Nm > 1$ the populations are homogenizing, and when $4Nm < 1$ they are diverging.
  - **Growth rate** (g) – Rate of exponential growth scaled by mutation rate. Positive g is growth, negative is shrinkage, but the magnitude can only be interpreted if μ is known: $\Theta_t = \Theta_0 e^{gt}$ (for t increasing into the future).
  - **Recombination rate** (r) – Rate of recombination (per site) scaled by mutation rate. `Lamarc` assumes uniform recombination across the sequence.
- Terms used to describe data:
  - **DNA** – full DNA sequence
  - **SNP** – single nucleotide polymorphism data. Lamarc currently assumes that every SNP was assayed; the next release will have a correction for ascertainment using a panel.
  - **Microsatellite** – variable-number sequence repeat, reported as number of repeats (not length). If only lengths are available, and the repeat size is known, it's okay to infer relative number of repeats.
  - **Region** – a stretch of data which is all linked; may include pieces which require different mutational models (for example, SNPs and microsatellites)
  - **Contiguous segment** – a stretch of data which is uninterrupted and all uses the same mutational model. (We could call this a "locus" but it may actually be multiple adjacent loci.)
- Terms used to describe the run:
  - **Chain** – series of tree rearrangements, after which we will report on results so far. In a likelihood run, multiple chains are used to improve the starting values. In a Bayesian run, this is not necessary and often only one lengthy chain is used.
  - **Heating** – running one or more "scout" searches on a flattened version of the likelihood surface in order to find distant peaks.
  - **Temperature** – in a "heated" run, controls how radically the scouts explore the terrain. The higher the temperature, the more randomly the scouts explore.
  - **MLE** – maximum likelihood estimate; the point estimate of a likelihood run
  - **MPE** – most probable estimate; the point estimate of a Bayesian run (the mode of the histogram representing sampled parameter values)

– **Prior** – in a Bayesian run, a distribution representing our prior expectations about the value of a parameter
– **Posterior** – the distribution resulting from multiplying the prior (our previous expectations about the situation) by the data-based likelihood (the contribution of our data). If the posterior looks just like the prior, your data are not contributing any information.