Laboratory excercise written by Andrew J. Crawford <andrew@dna.ac>
with the support of CIES Fulbright Program and Fulbright Colombia.
http://filogeografia.dna.ac
Enjoy!

Filogeografía: *genética evolutiva espacial.*

BIOL 4211, Universidad de los Andes, Bogotá. Coord.: -074.0657, 04.6012
25 de enero a 01 de abril 2006

Lab 9b
Divergence time estimation using *PAUP** 4b10, *MEGA* 3.1, and *TreeEdit* 1.0a10

25 marzo 2006

NB: Always download the latest version of all software
and read carefully the latest accompanying documentation.

The goal for lab 9b is to make some rough calculations of divergence times with confidence intervals for your data set. For this task, you will have to choose a few nodes of particular interest to focus on. You will apply external rate calibrations to the data and attempt to account for various sources of error in divergence time estimation.

The basic exercise of estimating divergence times implies that we know the rate of evolution of DNA sequence in our taxon of study. However, rates can vary among as well as within lineages. Therefore, ideally we would like to have several fossils of known ancestry and age, and whose direct descendents are included in the analysis. We could then use this temporal information to calibrate rates of evolution and infer the ages of all the other nodes in our phylogeny. Rarely in phylogeography do we have such fossil information, and when we do, often the ages are very uncertain. For example, with luck a fossil may be identifiable as a member of, or direct ancestor of, an extant lineage. This fossil, however, might be dated as "Pleistocene" in age, *i.e.*, between 1.8 million and 10,000 years, a level of uncertainty encompassing 2 orders of magnitude. Our second option is to use biogeographic events of known (assumed) age that we also assume caused the splitting event that resulted one or more nodes in our phylogeny. However, such inferred events are often unavailable, or sometimes we prefer to estimate rather than to assume its age.

Without fossils or reasonable biogeographic events to calibrate rates of evolution for our group of interest, we are forced to assume that our DNA sequences diverged at the same evolutionary rate estimated for similar genes from a related group of organisms. Recall that we need to apply a substitution (or divergence) rate, not a mutation rate as we used in the *IM* lab. Time will be calculated as (genetic divergence) ÷ (rate). We should account for at least four sources of error or uncertainty in our divergence time estimation:

I. Stochastic error or sampling error in estimating divergence from DNA sequence data.
We will calculate a standard error around our point estimate of genetic divergence.

II. Polymorphism within populations may inflate estimated divergence between populations.
We will calculate net divergence as $\pi_B = \pi_D - (\pi_{S2} + \pi_{S1})/2$ (Nei and Li, 1979).

III. Uncertainty in rates of evolution.
We will apply a fast and a slow rate calibration.

IV. Rate calibration and our estimates of genetic divergence should involve same type of distance.
We will calculate "uncorrected" genetic distances to match rate calibrations from the old literature. Thus, the larger the true genetic distance, the more our divergence times will be underestimates.

If we lack a calibration point and our data reject molecular clock hypothesis (Lab 7), we are "on thin ice." Continue with the lab today, but keep in mind that without a calibration point or a molecular clock, our divergence time estimates are very uncertain. Recall that our other option (which we are <u>not</u> doing today) is to add sequences from congeners, other genera, families, etc., that might have an associated, relevant fossil or biogeographic-based calibration point. Students interested in more sophisticated divergence time analyses should consider investigating other software packages, such as ***multidivtime*** (`http://statgen.ncsu.edu/thorne/multidivtime.html`), ***BEAST*** (`http://evolve.zoo.ox.ac.uk/beast/`), or ***r8s*** (`http://ginger.ucdavis.edu/r8s/`). All of these software

packages implement "relaxed clock" models of evolution for the estimation of divergence times from DNA sequence data.

Software: availability and assistance with analyses

Download the latest version of *MEGA* here:
`http://www.megasoftware.net/`
On-line help pages for *MEGA:*
`http://www.megasoftware.net/WebHelp/helpfile.htm`
Download the latest version of *TreeEdit* here:
`http://evolve.zoo.ox.ac.uk/software.html?id=treeedit`
An excellent forum for keeping up with the latest developments in phylogenetics:
`http://www.yphy.org/phycom/`
O'Meara, Brian (2006) guide to preparing NEXUS batch files and running the basic analyses using *PAUP**.
`http://www.brianomeara.info/phylogenetics.html`

Citations for software

Kumar S, Tamura K, Nei M (2004) *MEGA3*: Integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics* **5**, 150-163.
Rambaut A, Charleston M (2001) *TreeEdit* version 1.0 alpha 10. Distributed by the authors.
Swofford DL (1998) *PAUP*. Phylogenetic Analysis Using Parsimony *and Other Methods*, Version 4b10. Sinauer Associates, Sunderland, Massachusetts.

Additional references

Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating divergence times from molecular data on phylogenetic and population genetic timescales. *Annual Review of Ecology and Systematics* **33**, 707-740.
Bromham L, Penny D (2003) The modern molecular clock. *Nature Reviews Genetics* **4**, 216-224.
Brown W Jr, George M Jr, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Sciences USA* **76**, 1967-1971. *(The classic 2% vertebrate mtDNA clock.)*
Fleischer RC, McIntosh CE, Tarr CL (1998) Evolution on a volcanic conveyor belt: using phylogeographic reconstructions and K-Ar-based ages of the Hawaiian Islands to estimate molecular evolutionary rates. *Molecular Ecology* **7**, 533-545, *(Bird cyt* b *rate of 0.016/Myr total divergence, K2P+Γ model.)*
Macey JR, Strasburg JL, Brisson JA, Vredenburg VT, Jennings M, Larson A (2001) Molecular phylogenetics of western North American frogs of the *Rana boylii* species group. *Molecular Phylogenetic and Evolution* **19**, 131-143. *(Lists a number of mtDNA calibrations for ectothermic vertebrates.)*
Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences USA* **76**, 5269-5273.
Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy. *Molecular Biology and Evolution* **14**, 1218-1231. *(The NPRS technique.)*
Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic Inference. In: Hillis DM, Moritz C, Mable BK (eds.), *Molecular Systematics,* pp 407-514. Sinauer Associates Inc., Sunderland, Massachusetts, USA.
Thorne JL, Kishino H (2002) Divergence time and evolutionary rate estimation with multilocus data. *Systematic Biology* **51**, 689-702.

Overview of today's lab:

1) Identify groups of samples (possibly but not necessarily sister clades) for which you want to estimate divergence times.
2) Import data from NEXUS file into *MEGA*, sort samples into groups of interest.
3) Calculate **net** between group divergence, or just (total) between group divergence if you are comparing two clades containing substantial divergence well beyond the level of intra-populational polymorphism.
4) Mean divergence **±** two (or one) standard errors yields a confidence interval for genetic divergence.

5) Alternately apply two rate calibrations to the minimum and maximum divergences.
6) Create a Non-parametric rate smoothed tree in *TreeEdit.*

I. Pick two rate calibrations

Decide first on what two rate calibrations you will apply to you data. Check your original paper for possible rates. Check the literature for your study group (*e.g.*, "bird rates" or "frog rates") or ask the Profe. When you find your rate calibrations, note the species, the genes, the rate, the age of the calibration used to obtain this rate, and the type of genetic distance (corrected? "Raw" distance?) assumed in making the calibration. The type of distances (*e.g.*, uncorrected, K2P, etc.) assumed in the original calibration will determine what type of genetic distance you will estimate below. If you have no specific literature references in mind, you should calculate uncorrected distances (p-distance), see below.

II. Using MEGA to obtain a matrix of net between-group distances with standard errors

Launch *MEGA*, and click on "Click me to activate a data file". Select your NEXUS file. Upon opening this file, *MEGA* will give you and error message. Do not worry. In this window, find the far right drop-down menu: Utilities > Convert to MEGA format. In the new window select your file name and format (NEXUS), which should already be the default choices. Visually inspect this new document to make sure it looks like the conversion was clean. *MEGA* can be fussier than *PAUP**. For example, *MEGA* will not accept taxon names that start with "_". If the conversion was successful, save the file, keeping the new default extension ".meg". (If on the other hand, the file was not converted correctly, try going back to your original NEXUS data file, executing this in *PAUP**, exporting the data into a new "clean" document, and then opening this new file in *MEGA*. This might help clean out any weird symbols, commands or comments that might be confusing to *MEGA*.) Close the new file, the old file, and the current window ("Sequence Data Explorer")

Open your new .meg file in *MEGA*. Answer the series of questions about your data. You will need to create two "taxa groups" for each node of interest. You can do this here or from the next window. If you are back at the very first *MEGA* window, select the button for "Select/Edit Taxa Groups." Click the + New Group button and name your new groups. With the new group name highlighted, select your sequences one at a time and after each one click the left-arrow button to send the chosen sample into the new group. When all your non-overlapping groups are created, click "√ Close".

To obtain a matrix of net between-group genetic distances plus standard errors, follow these steps starting from the first window of *MEGA*. Select Distances > Compute **Net** Between Group Means... Under Compute, change the setting Distances to read Distances & Std. Err. Under "Gaps/Missing Data", change Complete Deletion to Pairwise Deletion. Select the model of evolution that matches the model (or non-model) used in the calibrations you plan to apply. For uncorrected distances, select p-distance. In all cases, leave "Pattern among Lineages" as Same (Homogeneous). Under "Substitutions to Include", leave as d: Transitions + Transversions (*i.e.*, all sites. "Under Std. Err. computation by", select Bootstrap and change the number of replicates to 1000.

In the resulting output, the numbers below the diagonal are the mean between-group distances, while above the diagonal you see the corresponding standard error. Under File > Export/Print Distances, save this table to an appropriately named file.

III. Calculating divergence times with confidence interval.

From your output files obtained above, calculate a confidence interval around your genetic distances by both adding and subtracting **two** standard errors (some authors might use just **±** one standard error). Now simply divide each number by a given rate to get a confidence interval for divergence time. For example, with a mean net divergence of 0.054 between groups and a standard error of 0.005, I would calculate a genetic divergence interval of 0.044 to 0.064. Assuming a fast rate of 0.02 total divergence per million years (Myr), I would obtain a range of divergence times from 2.2 to 3.2 Myr. I could also apply a slower rate of 0.014 and obtain a divergence time interval of 3.1 to 4.6 Myr. To account for both the error in genetic divergence estimation and the uncertainty in rates of evolution, I would report the minimum age obtained under the fast clock and the maximum age obtained under the slow clock, or 2.2 to 4.6 Myr. Of

course both rates could be wrong, but in the absence of a direct calibration or temporal constraint on our estimate, this is perhaps the best we can do.

If your data conformed to the molecular clock hypothesis (Lab 7), you can be relatively happy with these results. If your data rejected the clock, then you also are faced with the uncertainty about which parts of your molecular phylogenetic tree may have been evolving faster or slower.

IV. Making a NPRS tree.

We can use *TreeEdit* to quickly construct an ultrametric tree in which node height is proportional to the estimated relative divergence time. Branch lengths are calculated using Non-parametric Rate Smoothing (Sanderson 1997). Simply locate your ML tree from Lab 7, the file ending with the `.tre` extension. Launch *TreeEdit* then open this tree file from with the program. If your tree is not already rooted, you need to root your tree now (using the rooting button in the *TreeEdit* tool bar). Now, select Trees > Transform branch lengths... , then click on Ultrametric: and select Non-parametric rate smoothing, leaving Model in its default state (in other words, I do not know what is the difference among the three choices of Model), and finally hit OK. Enjoy looking at your NPRS tree. Now save the results. First select File > Export... and export your NPRS tree as a nexus file. Then select File > Export Data and Stats and export useful data on this tree by selecting all of the boxes.

You can use the NPRS branch lengths instead of the original ML branch lengths to estimate your divergence times. In this case, you would simply look at the node depth. However, to get a confidence interval, you would have to bootstrap the data and obtain a 95% confidence interval on node depth based on at least 100 NPRS-corrected ML trees. For this exercise, you could either fix the topology, as we did in Lab 7, or hope your node of interest is well supported.

LAB REPORT:

You do not need to write a lab report. However, you do need to incorporate the methods and results of divergence time estimation, as well as provide a discussion of the results and their implications for the history of your species and your *a priori* hypotheses concerning environmental, ecological, or environmental processes and their relative timing. You also need to provide ample citations. You need to explain your rate calibrations, where they came from and what data the original calibrations were based on (*i.e.*, age of calibration, the nature of the assumed event or fossil, and the type of evolutionary model assumed, if any). Also, note what species and what genes were used. Also, check whether the calibration was based on complete DNA sequences or RFLP data. Your discussion should also include your comment on the reliability (or unreliability) of your divergence time estimate. In other words, what was good (or bad) about your divergence time estimation methodology?

In your final report, you are free to use or ignore the NPRS tree, as you wish.

Good luck!