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

Filogeografía

BIOL 4211, Universidad de los Andes
25 de enero a 01 de abril 2006

Lab 1 – Looking for relevant DNA sequence datasets
28 January 2006

During this 10-week course each pair of students is going to be analyzing their own data sets. Finding these data sets is the responsibility of each pair of students. Today's lab will provide instructions and tips on how to locate potential data sets via the web and download data from GenBank.

The DNA sequence data contained in all published filogeographic, phylogenetic and population genetic studies should be available in GenBank or related databases. Therefore, we can either search GenBank directly or look for published studies of interest via a literature database. If we can obtain the papers, we can note the GenBank accession numbers for the published data then download the data from GenBank.

GOALS:

Find a data set based on your taxonomic group of interest.
Find a data set from Amazonia, Colombia, South America, Neotropics, if possible, please!
Confirm that the data include sufficient geographic and intraspecific sampling.
Get the original publication, look for locality data or museum accession numbers.
Look for additional sequences in GenBank of this species or close relatives.
Look for additional genes for these same individuals.

A great starting place for any search for data (and papers) is:
`http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi`

There you will find some very useful links:

PopSet: you can search for complete data sets by taxon, author, place names, subject etc. This is the most useful source of data sets. However, older studies are often missing from this data base. If you don't find data you want here, try one of the following.
Nucleotide: search all DNA sequences, regardless of when or how it was submitted.
PubMed: search for published papers.
PubMedCentral: search for published papers in open access journals. Yay!
Taxonomy: search by species, genus, or any other level in the taxonomic hierarchy.

Searching by publication.

Search literature databases via the library:
`http://biblioteca.uniandes.edu.co/`

Try google scholar:
`http://scholar.google.com/`

A list of Open Access journals can be found here:
`http://www.doaj.org/ljbs?cpid=68`

If you find a potentially useful data set or a provocative abstract, but you cannot access the original paper, ask the Profe to see if he might be able to obtain the PDF. Failing that, contact the authors directly.

When you find your preferred data set, you and your team mate will want to download the data in FASTA format. FASTA is pronounced "FAST-Aye", and stands for "FAST-All", because it works with any alphabet, an extension of "FAST-P" (protein) and "FAST-N" (nucleotide) alignment.

If you find a data set via Entrez PopSet, notice the dropdown menu labeled "Display." Select FASTA from this menu, then where it says "Send to" select File and save your data.

Do not forget to search GenBank for more sequences of the same gene/s from the same or related species, perhaps ones that were part of a different study. Also check for different genes from your same organisms.

If you have your data, please have the Profe check and approve it.

If your data set is approved, you can start visualizing and aligning your data using ClustalX, MEGA (PC only), Genedoc and Mequite, if these programs have been installed. These programs are free, and you should install them in your own computers, too.

ClustalX `http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html`
MEGA `http://www.megasoftware.net/`
Genedoc `http://www.psc.edu/biomed/genedoc/`
Mequite `http://www.mesquiteproject.org/mesquite/mesquite.html`

To look at a NJ from ClustalX, for example, use:

TreeEdit `http://evolve.zoo.ox.ac.uk/software.html?id=TreeEdit`
TreeView `http://taxonomy.zoology.gla.ac.uk/rod/treeview.html`
FigTree `http://tree.bio.ed.ac.uk/software/figtree/`

For geographic data, you may have to check the data for the actual specimens. Many natural history collections are on-line. You can also check new multi-institutional data bases such as:
Global Biodiversity Inventory Facility: `http://www.gbif.org/`

If you are having trouble finding a data set, ask the Profe for help or suggestions.

Good luck!!