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 2 – Alignment of DNA sequences, visualization and export
04 February 2006

By now you have found your data set to work with for the course. At a minimum, your data include one mitochondrial gene sequenced from multiple individuals from multiple populations within a single species. Your data should also include an outgroup. Your data may also include additional species and additional genes. Today's lab will provide instructions and tips on how to create an alignment using *ClustalX* and manipulate this alignment in *MEGA* (Windows OS only) and *Mequite*. To visualize preliminary trees, we will use *TreeEdit*. These programs should already be installed on your machines. Otherwise, the programs are available here:

Software

ClustalX (for all major computer operating systems):
`ftp://ftp.ebi.ac.uk/pub/software/unix/clustalx/`

MEGA version 3.1 (for Windows or with Windows emulator software):
`http://www.megasoftware.net/`

Mequite:
`http://mesquiteproject.org/`

Additional information

Using ClustalX:
`http://bips.u-strasbg.fr/fr/Documentation/ClustalX/`
Use the 'Help' menu in *ClustalX*, or look at the above URL for detailed information on *ClustalX* and its various options.

TreeEdit:
`http://evolve.zoo.ox.ac.uk/software.html`
A free program for Macintosh for visualizing and manipulating trees quickly and easily.

TreeView:
`http://taxonomy.zoology.gla.ac.uk/rod/treeview.html`
Like *TreeEdit*, but fewer features.

MacClade:
Is not free, but it is useful. I believe it is already on the Mac computers in the lab. Feel free to use this program as well. Not available for Windows.

Text editors:
For preparing data files, I recommend a free, flexible text editor such as *TextWrangler*.
On a Macintosh OSX machine, you can also use a unix-based text editor such as *emacs*.

On a PC machine, you might use *WordPad*, but under View > Options > Text (or Rich Text) > Word wrap, select "No wrap."

GOALS:

Organize your sequences into one file.
Gain familiarity with *ClustalX*.
Change multiple sequence alignment parameters and note any resulting changes in alignment.
Save one or more final alignments for use in the next 8 weeks.
For protein-coding sequences, confirm that the data do not contain premature stop codons.

I. Gathering you data together.

To begin the lab, for each gene fragment or multigene contiguous DNA sequence you should have one file. In passing data from GenBank to ClustalX, the easiest format to use would appear to be FASTA, in which each block of DNA sequences is preceded by one line of text containing the name of the sequence. This line starts with the "greater than" symbol (>). For example:

```
>gi|66270413|gb|AY968888.1| Myioborus albifacies isolate ALB_190_1 NADH
dehydrogenase subunit 2 (ND2) gene, partial cds; mitochondrial
ATAAACCCCCAAGCAAATCTAATTTTCATCATCAGCCTAATCCTAGGAACTACCATTACCATCTCAAGCA
ACCATTGAGTCATAGCCTGAACCGGCCTTGAAATCAACACGCTTGCCATCCTCCCACTAATCTCAAAATC
CCATCACCCCCGAGC
>gi|66270414 etc etc.
```

All your sequences of any one gene can be in one file, cut and paste from one file into a new, third file if necessary (or you can load the different files one at a time directly into ClustalX). Before proceeding with the next step, you may want to shorten the name of each sequence, leaving only the minimum necessary information, such as GenBank accession number, species name, specimen number, and the population it came from! (The name of the gene can be part of the name of your new file.) However, try to keep the number of characters in the name limited to 26 or fewer, and try not to use any symbols or characters besides English letters and numbers. Use the underscore symbol (_) instead of spaces in the sequence name. If you have multiple genes, make your sequence names such that if you were to alphabetize the sequences in both genes, the individuals would come out in the same order! This will help you a lot later on! You might think to create names for your sequences such that if you were to automatically put them in alphabetical order (such as by using *Mequite*) the samples would be grouped by population.

II. Multiple sequence alignment using *ClustalX*

Launch *ClustalX*
Select File > Load sequence and navigate to your data file and select it.
In *ClustalX*, verify that you have your expected number of sequences.
Choose Alignment > Multiple Alignment > Multiple Alignment Parameters
Check the following parameters and note their default values:

Gap Opening: Too high = few gaps but also fewer matching bases.
Too low = lots of gaps but with more matches.
Gap Extention [Extension]: Penalty for additional gaps in a row.
Delay Divergent Sequences: Align "weird" sequences last.
DNA Transition Weight: Optionally you can 'downweight' transitions
1 → no weighting; 0 → transitions = mismatch.

For very similar sequences, use no weighting (weight = 1).
Select Alignment > Produce Guide Tree Only and follow instructions.
If you have *TreeEdit* or *TreeView* on your computer, use it to view your new tree.
By default his file will end with .dnd
Next select Alignment > Do Alignment from Guide Tree. By default, you new alignment file will end with .aln
On my Mac laptop, 50 sequences of 1000 bp each took about a half of a minute.
The other option, "Do Complete Alignment," took 5 minutes.

Notice the low horizontal window at the bottom of the screen. This graph indicates by position the relative quality of the alignment. You can change the relative scale via the menu: Quality > Column Score Parameters.
If you do not have any gaps inside any DNA sequences (as opposed to on either end of the sequence), then you can skip item **II.** below and proceed to part **III.**

III. Effect of alignment parameters.

If you have gaps in your alignment, you should check whether changing parameters affects your result.
First, try to make fewer gaps of larger size using values such as these:
Gap Opening: 50
Gap Extension: 1
Remember to save your new alignment under a new name. Look at the distribution of gaps before closing.
Next, try to make lots of gaps of small size:
Gap Opening: 5
Gap Extension: 5 (Do not bother setting the extension penalty higher than the gap penalty.

Did your alignment change among the three trials? If not, your alignment is robust to variation in alignment parameters. If your alignment changed a lot among trials, which alignment do you prefer and why? Note any regions of the alignment that changed a lot among the three trials.

Looking at your preferred alignment (you may have to quit & re-lauch *ClustalX*, and sometimes you have to run the alignment algorithm over again) and the Quality Score window below, note any regions of your alignment that have a low score. Such regions you may want to exclude from all future analyses.

IV. Saving and exporting

Your preferred alignment is already saved on the computer, along with zero or more other alignments that all end in .aln These files start with the following line of text:

```
CLUSTAL X (1.83) multiple sequence alignment
```

These files are readable directly by programs such as *MEGA*, *Mequite*, *MacClade*, but not by *PAUP**. It is always a good idea to save files in various formats, just to be sure that your data will be readable by the next software you may want to use.

In *ClustalX*, you can "export" your preferred alignment (or alignments) using File > Save Sequences As . . . Select "NEXUS" and "OK" The resulting file starts with lines of text such as:

```
#NEXUS
BEGIN DATA;
dimensions ntax=53 nchar=1041;
format missing=?
symbols="ABCDEFGHIKLMNPQRSTUVWXYZ"
interleave datatype=DNA gap= -;

matrix
```

This file should be readable by *PAUP**, *MacClade*, *Mesquite*, *MEGA*, etc.

However, for our purposes, "interleave" format is *not* preferred for single-gene data. (Later, if you combine multiple gene sequences in one NEXUS file, you will need the word "interleave" at that time.) For now, working with one gene at a time, you should open your data (either version) in one of the above programs and then export it again, making sure that the 'interleave' option is not selected. If you have protein-coding sequences, you can just do this re-exporting at the end of step V below.

V. Checking amino acid translation.

Using MEGA *(also* MacClade *does this, but* Mequite *does not, as far as I can tell.):*

Launch *MEGA*, and click on "Click me to activate a data file." Navigate to your file, and click "Open."

Utilities > Convert to MEGA Format, then in the "Select File and Format" window select the "Data Format" box and choose type of format your data *was* in (either ".aln (CLUSTAL)" or ".nexus (PAUP*, MacClade)". Note, .nexus and .nxs refer to the same format, as does .nex Finally, select "OK"

Close the new data window and the old data window and re-open your new .meg file.

Select "Nucleotide Sequences".

Reply "Yes" or "No" as to whether the sequence is protein coding. If yes, next select appropriate genetic code.

To translate the DNA sequence into the inferred amino acid sequence, click on the button "UUC→Phe". *MEGA* automatically tries to identify the correct reading frame by minimizing the number of stop codons.

Now check for stop codons! This is important!! I believe *MEGA* indicates a stop codon with an asterisk (*). For easier visualization, select Display > Color Cells.

If your protein-coding gene has stop codons anywhere besides the end, your sequence might be in the reverse complement state. *MEGA* I don't think makes this conversion, but *Mesquite* and *MacClade* do.

Option B, MacClade:

Of course you can use this program, but I prefer to use free software (although I'm still using PAUP*).

If you need to check the reverse complement of your DNA sequences to see if this way you see no pre-mature stop codons:

Launch the application, *Mesquite*, select File > Open File and choose either your .nxs file (NEXUS or PAUP* format) or your .aln file (CLUSTAL format), and then select the appropriate format to be read in, if necessary.

Select Matrix > Alter/Transform > Other Choices . . . and select "Nucleotide Complement"

Repeat the above command but chose "Reverse Sequence" instead.

Now select File > Export . . . twice, choosing different formats each time, such as FASTA and "Old-Fashioned NEXUS", just to be safe, and perhaps adding "revcomp" to the end of the name. Save data file as well, under a new name.

Open one of these new files in *MEGA*, and follow the instructions above.

VI. Final data set.

Your final data set should be in NEXUS format and **not** interleaved.

If you have multiple genes, you should both save the genes in separate files, but also you will want to make a master NEXUS file.

To make a master file, you will want to put all your individuals in the same order in each data set. Do this by hand or by using the sort (alphabetizing) tool in *Mesquite*, the tool that looks like a little Mayan pyramid, then re-save or re-export your data. If some individuals are missing from some data sets, you will have to add them in and put a "?" for every nucleotide site (position).

Finally, you will past all your non-interleaved data sets into one giant text file in NEXUS format, one set of aligned sequences BELOW the other. Do not attempt to put all gene sequences of each individual on one line. You will have to change the "nchar=" parameter at the top of the NEXU file. The new nchar value will be the sum off all the nucleotide sites in all genes in this file. Also, you will have to add the word "interleave". Here's an example in which we are missing data from the first gene (COI) for the third individual, and the first individual has no data from the second gene (cyt *b*). Remember that we can put any comments we want to inside square brackets [like this] and they will be ignored by all programs that understand the NEXUS format (e.g., *PAUP*, *dnaSP*, *MrBayes*, *Mesquite*, *BioEdit*, etc.).

```
#NEXUS
BEGIN DATA ;
      DIMENSIONS  NTAX=5  NCHAR=59 [36 plus 23] ;
      FORMAT interleave DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
[                                          10        20        30
[                                          .         .         .
            [ *** THE COI DATA FIRST, NCHAR=36 *** ]
azueroen_KRL0680_noCytb              ??????????AGAAGTCTACATCCTAATCCTGCCGG
crass_AJC0085_Las_Cruces_CR          TTGGCCACCCGGAAGTATATATTCTTATTCTCCCAG
crass_AJC0189_Fortuna_noCOI          ????????????????????????????????????
crass_AJC0209_Nusagandi_PA           TTGGTCACCCGGAAGTCTACATTCTTATCCTCCCAG
crass_AJC0215_Campana_PA             ??????????GAAGTCTACATTCTTATCCTCCCAG

            [ *** THE cyt b DATA BELOW, NCHAR=23 *** ]
azueroen_KRL0680_noCytb                 ???????????????????????
crass_AJC0085_Las_Cruces_CR             CCTCAGGGCTATTTCTGGCCATA
crass_AJC0189_Fortuna_noCOI             TCTCAGGCCTGTTTCTCGCTATA
crass_AJC0209_Nusagandi_PA              ??????ACTATTTCTGGGCCATA
crass_AJC0215_Campana_PA                CCTCGGGAGTATTTCTCGCCATA
```

Good luck!!

(Lab continues on next page…)

REPORT:

Remember, the purpose of each lab report is so that by the end of the course each team will be able to prepare its final "manuscript" very easily. When I write a phylogeography manuscript, I might write the introduction while I'm still collecting the data. I definitely start writing the Methods and Results while I am doing the analyses. If you wait until you are "done" with all the analyses to write the Methods and Results, it is often difficult to find all the details of what you did. So, you can think of each lab report as making some progress on your final manuscript. Some basic information about the data will be repeated on various lab reports, as necessary.

Working with your lab partner:

Write a very brief summary of your alignment protocol, noting the parameter values you used in your multiple sequence alignment. Note briefly how or why you justify your choice of alignment. Did you check the amino acid translation for premature stop codons? How did you do this?

For your Methods section, be sure to always note the version of the software you are using. Also, note the citation. All software packages provide information on the proper way to cite them.

Your Results section should include the number of sequences from which genes of which species you have. Mention the total number of positions in your alignment, and mention how many gapped sites you have. Also mention now if you plan to EXCLUDE any sites from your analyses due to skepticism about the alignment in certain regions of the gene.

You should start making a Literature Cited section, including the original publication and any other papers associated with data on sequences, specimens, or localities. Also, include the citation for any software you used today.

For example, the citation for *ClustalX* is below. You can "cut n paste" this reference from the URL mentioned above:

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Research, 24: 4876-4882.

This brief report will be due on the 10th of February.