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 3 – DNA sequence polymorphism analysis using *dnaSP*
11 February 2006

Today we begin to explore the DNA sequence data set that you found two weeks ago on GenBank, and which you aligned last week. At a minimum, your data include one mitochondrial gene sequenced from multiple individuals from multiple populations within a single species. Your data might also include additional genes and additional species. Your data should also include one or more outgroup sequences.

Today's lab will provide instructions and tips on how to run *dnaSP.* We will be using version 4.10, dated October 21, 2005. This program only works with the Windows operating system, which is unfortunate. This program should already be installed on the non-Macintosh computers in the computer lab. The program is available from the Universitat de Barcelona website:

Software

dnaSP version 4.10.4 (for Windows OS only)
`http://www.ub.es/dnasp/`

Citation

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.
Maddison WP, Swofford DL, Maddison DR (1997) NEXUS: an extendible file format for systematic information. *Systematic Biology* **46**, 590-621.

Further information

Throughout this lab, recall that *dnaSP* has an excellent Help menu which may accessed via the Help menu on the far right of the menu bar. Select Help > Contents. You may have to help *dnaSP* find the help file! There you can find information on running the software as well as background information and citations concerning each type of analysis.

GOALS:

Define groups of taxa in your NEXUS file.
Gain familiarity with *dnaSP*, an "industry standard" in molecular population genetics.
Calculate summary statistics for your data.
Run statistical tests to see if your data conform to the Wright-Fisher standard neutral model.

I. Preparing your data for *dnaSP*.

You aligned DNA sequence should now be in one data file in the NEXUS format. If you have a multi-gene data set, all sequences will be in your master data file, but you should also still have data files with each gene alone. For this lab, you will analyze each gene (alignment) separately! I strongly suggest you put all your master data documents in one file on your computer, and then for each analysis you want to run, create a new file in which you will put a copy of your master data and in which you will store your output from each analysis. Sometimes, however, you will update your master files with additional blocks of information, such as assigning samples to different populations. The goal is that your master data files will have the most complete information available for your data.

Before beginning this lab, you should make some modifications to your master file that will also be useful for future analyses in future labs. Begin by defining any groups of sequences you might need to analyze together, such as a collection of sequences from one population, or a collection of sequences from one geographic region, or from one species. Use a text editor such as *WordPad*, *TextWrangler*, or *emacs* to modify your data file as follows, creating "blocks." Blocks contain commands and parameter settings that will be read by particular software packages every time you open your data in that particular software. Blocks will save you lots of time because you will not have to re-set all your parameters or re-write all your commands. Also, in your NEXUS file you can write notes to yourself by place any comments or alternative commands in `[ ]` square brackets. In NEXUS format, upper case vs. lower case is not important. In *dnaSP*, however, square brackets inside commands will cause problems, and some of these problems you might not even notice at first. Therefore, in *dnaSP*, `[]`-symbols should be used with caution. In *PAUP* the number of spaces, tabs and new lines does not matter, but in *dnaSP* these can be a problem! If *dnaSP* fails to accept your data file, try minimizing these extra white spaces.

In NEXUS format, different types of data and different types of commands are written in different "blocks." A block begins with the word "`begin`" followed by the name of the block. Possible blocks include `SETS`, `PAUP`, `TAXA`, `CHARACTER`, `DNASP`, `MrBayes`, etc. Each set contains one or more commands or settings. Each line of commands plus any options or settings will end with a ";" (but one command many occupy more than one line, in which case the command should not be interrupted by a ";"). Example:

```
begin SETS ;
     TaxSet California = 13-44 ;
END;
```

If your NEXUS file currently contains both a `TAXA` block (containing only names of your samples) and a `CHARACTER` block (containing the data matrix), you will need to modify your file for *dnaSP* such that it has only a `DATA` block. Put [ ] around your entire `TAXA` block so that *dnaSP* will ignore it. Then change the name of your `CHARACTER` block to read `DATA`, make sure it contains the correct values for `nchar` and `ntax`, and make sure there are no spaces on either side of the "=" signs. Your DATA block should follow this format:

```
#NEXUS
BEGIN DATA ;
      DIMENSIONS  NTAX=2  NCHAR=29 ;
      FORMAT DATATYPE=DNA  MISSING=? GAP=- ;
MATRIX
      G_mutisi        AATCCGACTACTGACACCGGGTATAAACC
      G_bogotana      AATCCGACCACTGACACCGGATATAGACC
;
END;
```

Your DNA sequence alignment (see example above) ends with a ";" and then "`END;`". Remember that every command (but not every line) in NEXUS format has to end with ";" Skip a line and add a "`SETS`" block which will contain the name of your populations or other groups of samples. The numbers refer to the samples in your aligned data, the first aligned sequence is "1", and on downward. Here is an example `SETS` block for a data set with 40 sequences, two populations, and two species, plus an outgroup:

```
begin SETS;
      TaxSet Santafe_de_Bogota = 1-23 ;
      TaxSet Los_Llanos = 24-38 ;
      TaxSet G_mutisi = 1-10 24-30 ;
      TaxSet G_bogotana = 11-23 31-38;
      TaxSet ingroup = 1-38 ;
      TaxSet outgroup_muestras = 39-40 ;
      TaxSet bichos_rojos = 2 7 10-15 22 24-28 ;
END;
```

The numbers refer to the taxa in the order they are listed in the NEXUS file. Notice that the numbers can be listed as a hyphenated range, or as a simple list of numbers separated by spaces (no commas). The program *PAUP** will accept either a number or the actual names of each taxon (DNA sequence). Unfortunately, the software *dnaSP* will ONLY accept numbers, not names, in `taxset` commands. Including `[ ]` symbols inside a taxset command may be mis-interpreted as a "`;`" by *dnaSP*, so be careful when using square brackets (or avoid including them when using *dnaSP*).

If you used *MacClade* or *Mesquite* to identify codons or translate your DNA sequences into amino acid sequences (i.e., looking for pre-mature stop codons), then your data may already contain a `CODONS` block that looks something like this:

```
BEGIN CODONS;
CODONPOSSET * CodonPositions =
      1: 1-355\3,
      2: 2-356\3 356,
      3: 3-354\3;
CODESET  * UNTITLED = mtDNA.mam.ext: all ;
END;
```

DnaSP will read the information in both the `SETS` block and the `CODONS` block. If you do not have a `CODONS` block, you should calculate your codon positions using *MEGA*, *Mesquite*, or *MacClade*, though for today's lab you will not need a `CODONS` block, but later you will want one.

For *dnaSP* you can add one more block, a `DNASP` block:

```
begin DNASP ;
      CHROMOSOMALLOCATION= mitochondrial ;
      GENOME= Haploid ;
END;
```

The new version of *dnaSP* is very flexible and accepts data files in the following formats: FASTA, MEGA, NBRF, NEXUS, PHYLIP. However, *dnaSP* might have problems with certain non-standard blocks that may be in your data file, such as a `MacClade` block, a `Mequite` block, or a `MrBayes` block. You may need to delete these blocks, but do so only AFTER saving your file under a new name, perhaps a name ending with "_dnaSP.nex" or something like that. Also, *dnaSP* does not accept any degenerate bases in DNA sequences, only `A`, `G`, `C`, `T`. Therefore, if your data contain any degenerate bases such as `N`, `Y`, `R`, `S`, `W`, `M`, `K`, you will have to save your data file under a new name (perhaps add the ending "_dnaSP.nex" or something like that), and then you will have to change all degenerate bases to "**?**" i.e., missing data. If your data file came out of *BioEdit*, *MacClade*, etc., degenerate bases may be written as `{AG}`. If so, replace all four characters,`{`, `A`, `G`, and `}` with a single `?` character.

The program *dnaSP* is going to ignore sites with gaps (and ignore sites with missing data, maybe?). If you are analyzing a data set with multiple genes, and the individuals missing one particular gene are still included in that data set but at all positions they have only gap symbols (-) or missing-data symbols (?), then when you open these data in *dnaSP*, the program will read all the data but will count and analyze nothing. Therefore, if you have any individuals in the alignment that totally lack data for a given gene, you should remove these individuals from the data set. Again, this problem could only arise if you had multi-genic data that were incomplete and that you previously placed into one master NEXUS file.

II. Loading your data into *dnaSP*

Launch *dnaSP*

Select File > Open Data File... (or just click on the open file icon or hit Ctrl-O) and navigate to your data file and select it.

The "Data Information" window will appear. Check to make sure that the data are as you expected.

If something goes wrong, open your data file in a text editor and make any corrections. Repeat this process as often as necessary.

In addition to describing your data using blocks within your NEXUS file, you can also use various options and controls under the Data menu.

If you would like to see your data from within *dnaSP*, select Display > View Data .

III. Describing your data

Your assignment for this laboratory is to create a data summary table that will go into your growing manuscript. You will use *dnaSP* to report population genetics summary statistics of interest. Your table will follow the model below, following, e.g., Kliman et al. (2000) *Genetics* **156**, 1913, and Carstens et al. (2005) *Mol. Ecol.* **14**, 255. You may want to create a speadsheet now and fill in the data as you run each analysis.

Population	n	Prob.	h	S	θ ± 1 S.D.	π ± 1 S.D.	D_T	Prob. (|D_T|)>0	D_T 95% C.I.	Prob. by coal. simul	R_2	max. k
Pop1 name												
Pop2 name												
Pop3 name												
Etc. etc.												

n = number of samples
Prob. = probability of having captured the deepest coalescent event: (n-1)/(n+1)
h = number of haplotypes
S = number of segregating sites
θ ± 1 S.D. = population mutation rate, plus or minus one standard deviation.
π ± 1 S.D. = average pairwise distance, plus or minus one standard deviation.
D_T = Tajima's D (π vs. θ)
Prob.(|D_T|)>0 = Probability $D_T \neq 0$ as determined by coalescent simulation.
D_T 95% C.I. = 95% confidence interval around D_T as estimated by coalescent simulation.
R_2 = Ramos-Onsins & Rozas' R2 ($\pi L/2$ vs. η_1 within a DNA sequence).
k max. = maximum number of nucleotide differences between any two sequences within a population ($k = \pi L$).

***n* = number of samples** This number you probably already know from working with your data. If you made a SETS block with TaxSet commands, then n is also available in *dnaSP* but checking Analysis > DNA Polymorphism... and then clicking on the menu Data Set, where *dnaSP* lets you know **(*n* =__)** for each population you have defined.

Prob. = probability of having captured the deepest coalescent event: (*n*-1)/(*n*+1) This value you will have to calculate by hand.

By clicking on Analysis > DNA Polymorphism... and selecting the first two Options: [ ] Compute Variance of Pi, [ ] Compute Pi (Jukes and Cantor), and finally selecting the appropriate population and clicking OK, you will get the following information: ***h***, ***S***, **Pi**, **Standard deviation of Pi, theta** and **standard deviation of theta**. Leave Region to Analyze in the default state (complete sequence) for these and all analyses, and we do not want to do any sliding windows today.

D_T* = *Tajima's D Click on Analysis > Tajima's Test... and under Data Set: select the appropriate population. Under Nucleotide Substitutions Considered click on Segregating sites. The output will give you Tajima's D, and Statistical significance as estimated by Tajima's (1989) original analytical method, which is not very powerful. Instead we will calculate significance level using coalescent simulation. The coalescent simulation should be done immediately after calculating D_T for each population. It will be much easier this way.

Prob.(|*D_T*|)>0 = Probability *D_T* ≠ 0 as determined by coalescent simulation. Immediately after obtaining D_T for a given population close the output window and next click on Tools > Coalescent Simulations... and note that in the new window your sample size for that population is already entered in this window automatically. Under Simulations Given... select Segregating Sites.

[For more information on coalescent simulations see Wall JD, Hudson RR (2001) *Mol. Biol. Evol.* 18:1134.] (Note, your observed number of segregating sites should now be shown in the box at the top right of the window.) Under Recombination, if you are analyzing mitochondrial DNA sequence data, select No Recombination. For nuclear gene sequences, select either Free Recombination or estimate the recombination rate using *dnaSP* and enter this value in Intermediate level, R (per gene). Under Compute... select Tajima's D. Note, your observed D_T should now be shown in the box labeled D(obs), value. Increase the No. Replicates to 5000, and hit Run. In the output window, note the Lower limit and Upper limit values of D_T and note the probability of your observed value of D_T in the third box down. If this number is less than 0.05 your D_T is significantly non-zero and you can reject the neutral Wright-Fisher population model. In your table, indicate this *p* value.

***D*$_T$ 95% C.I. = 95% confidence interval around D_T as estimated by coalescent simulation.** Report the minimum and maximum values shown in the output window.

R_2 = Ramos-Onsins & Rozas' R_2 (πL vs. η_1 within a DNA sequence). This value is calculated under Analysis > Population Size Changes... Again select your population of interest under Data Set. Leave the default Analysis as Pairwise No. of Differences (Mismatch Distribution). Also leave Model for Expected Values as its default, Constant Population size. The Output window will show you the R2 statistic. If you want to check significance of your R_2 value, you must run a coalescent simulation, as above. You also get another output window labeled Output.grid. Be sure you hit Graph > Select Graph > Pairwise Differences so you can see a graph of a mismatch distribution for your population along with a curve of the "expected" distribution under the assumption of Constant Population size. Read the next paragraph before you close this window!

***k* max. = maximum number of nucleotide differences between any two sequences within a population.** We can read this value (*k*) off of the Mismatch distribution from the table or from the graph. See above. It's the maximum of the Pairwise Differences that occurs with non-zero frequency in the data. The graph is more interesting, but with many DNA sequences in your sample, it will be easier to read this number from the table.

IV. Optional task (required for nuclear gene data sets)

Especially if you have nuclear gene sequences you should test for evidence of recombination in your data using the 4-gamete test (Hudson & Kaplan 1985). This test should be applied to within-population samples. If you apply this test to very divergent sequences, you may be confounding recombination with multiple substitutions. To test your data for recombination, go to Analysis > Recombination... and chose your Data Set. Be sure to analyze all sites. The output window will also give you an estimate of the recombination rate parameter, *C* (Hudson 1987), which you may need for additional coalescent simulations, but be aware that this parameter is notoriously difficult to estimate reliably. Citations are given in the *dnaSP* Help files.

REPORT:

Remember, the purpose of each lab report is so that by the end of the course each team of two students 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 what parameters you estimated and how you estimated it if there were any options or decisions you had to make. Report only as many details as you need to such that another person could repeat your analysis. Also report what each parameter means and give citation. Citations can be found in the *dnaSP* Help > Contents menu – or on the *dnaSP* webpage.

If your data were collected from across the landscape (species' range) instead of from within discrete populations, please explain which samples you had to combine into groups to run today's analysis and why are you analyzing these groups.

Your Results section should briefly summarize important parameters. Also, did your data conform to the equilibrium neutral model as far as you could tell? Continue expanding your Literature Cited section, including ones for *dnaSP* and for each summary statistic or parameter you estimated today.

This brief report will be due on SATURDAY, the 18th of February. (Saturday seems to be better, no?)