Laboratory excercise written by Andrew J. Crawford <andrew@dna.ac>
with the support of CIES Fulbright Program and Fulbright Colombia.
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 6
Nested Clade Phylogeographic Analysis using *GeoDis* v. 2.4

04 marzo 2006

NB: Originally I had no intention of teaching NCPA to my students, but they demanded it. In the spirit of teaching both sides of an argument, I prepared this laboratry section. More recently, NCPA in general and this lab handout in particular has been rendered conceptually and methodologically outdated by the following paper:

> Panchal M, Beaumont MA (2007) The automation and evaluation of nested clade phylogeographic analysis. *Evolution*, 61, 1466–1480.

Proceed at your own risk….

Today we have one big goal: successfully execute a Nested Clade Phylogeographic Analysis (NCPA) using *GeoDis* version 2.4, and we will spend much of our 4-hour lab period preparing the complex infile for *GeoDis*. Before starting this lab you will need two pieces of data. First, you will need the statistical parsimony network you made using *TCS* last week. Second, you will need longitude & latitude data for each haplotype in your data set. *GeoDis* software is available once again from *el famoso* David Posada. The NCPA Inference Key is also updated regularly. Check for both the latest version of *GeoDis* and the latest edition of the NCPA Inference Key here:

Software

GeoDis version 2.6 (Java program, runs on Windows OS, Macintosh OSX, Linux and other types of Unix)
`http://darwin.uvigo.es/software/geodis.html`

This program runs in *Java*, so we need to make sure any PC's have a Java Virtual Machine installed.
1) if you have an internet connection, go to http://javatester.org/version.html
2) or at the Command Prompt (Símbolo del Sistema), type: `java -version`

If your machine lacks *Java*, install it from here: http://java.sun.com/webapps/getjava/BrowserRedirect
The latest edition is 1.5.0_02.

The appropriate citation for the *GeoDis* software package is:

> Posada D, Crandall KA, Templeton A (2000) GeoDis: a program for the cladistic nested analysis of the geographical distribution of genetic haplotypes. *Molecular Ecology* **9**, 487-488.

Some key reference on NCPA are:

Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in Drosophila. *Genetics* **117**, 343-351. (*Explains the basic "nesting rules" needed to create the nested clades which are central to NCPA analysis.*)

Templeton AR, Crandall A, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. *Genetics* **140**, 767-782. (*Explains "probability of parsimony" used to construct network that represents "95% set of plausible solutions*.")

Hudson RR, Boos DD, Kaplan NL (1992) A statistical test for detecting geographical subdivision. *Molecular Biology and Evolution* **9**, 138-151. (*Explains the permutation tests employed in NCPA.*)
Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. IV. Nested analyses with cladogram uncertainty and recombination. *Genetics* **140**, 767-782. (*Adds some new guidelines to the "nesting rules" needed to create the nested clades.*)
Templeton AR, Routman E, Phillips CA (1995) Separating population structure from population history: a cladistic analysis of the geographic distribution of mitochondrial DNA haplotypes in the tiger salamander, *Ambystoma tigrinum*. *Genetics* **140,** 767-782. (*First NCA in phylogeography*.)
Templeton AR (1998) Nested clade analyses of phylogeographic data: testing hypotheses about gene flow and population history. *Molecular Ecology* **7**, 381-397. (*Extensive review of NCPA approach.*)
Knowles LL, Maddison WP (2002) Statistical Phylogeography. *Molecular Ecology* **11**, 2623-2635. (*Oft-cited criticism of NCPA and appeal to use coalescent simulations to test hypotheses*.)
Templeton AR (2004) Statistical Phylogeography: Methods of evaluating and minimizing inference errors. *Molecular Ecology* **13**, 789-809. (*Extensive response to Knowles & Maddison 2002 criticism.*)
Castelloe, J., and A. R. Templeton. 1994. Root probabilities for intraspecific gene trees under neutral coalescent theory. *Molecular Phylogenetics and Evolution* **3**, 102-113. (*Concerning "outgroup weighting" option in* GeoDis*, which I have never tried.*)

The GeoDis file should also include the somewhat helpful documentation file, GeoDis2.4.pdf. Open this file now.

Overview of today's NCPA lab:

1) Nest the clades found in your network obtained last week. Use Templeton's procedure.
2) Make input file for use in *GeoDis*. File contains clade structure and GPS coordinates.
3) Run *GeoDis* software to find significant associations between clades and geography.
4) Interpret significant clades using Inference Key.

Note, you need to have not only your network ready from last week (made with *TCS*) but you will also need to know which haplotypes in your network are represented by how many copies (samples). If these identical gene copies are not from the same locality, you will also have to note the name (identity) of each of these identical haplotypes. Of course, this information may not be obvious from your graphic image of your network. If not, hopefully you saved your network as a .gml file. If so, you can open this file again in *TCS* and check the name and number of haplotypes represented by each sampled node on your network.

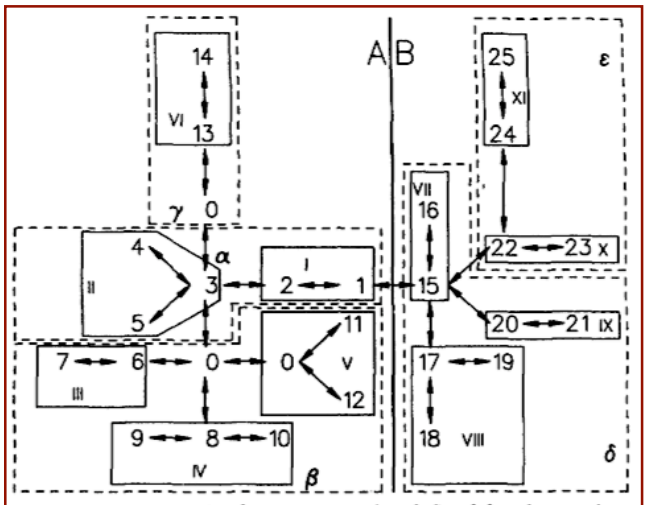

FIGURE 2.—The haplotype categories defined by the nesting algorithm discussed in the text. The haplotypes enclosed by *solid lines* indicate the 1-step clades and are designated by *Roman numerals*. The *dashed lines* enclose the 2-step clades that are designated by *Greek letters*. Finally, a *thick, solid line* indicates the partitioning of the cladogram into two 3-step clades, designated by *Roman letters*.

I. Nesting your clades

You will need to have your statistical parsimony network handy, either on paper or in a graphics program that will allow you to draw successive circles around haplotypes and nodes in your network. Using Templeton's rules for nesting clades are simple, but depending on your data, you may encounter unusual cases in which the rules for nesting may not be clear. Templeton and colleagues use the following terminology. A haplotype is referred to as a 0-step clade. 1-step clades include at least one haplotype, but may include more or it may include a node representing an unsampled haplotype. 2-step clades are groups of 1-step clades, etc. All *x*-step clades are included in *x*+1 step clades until *x*+1 becomes a clade that includes the entire network, a

large *n*-step clade. The nesting procedure distinguishes "terminal" and "internal" *x*-step clades. **Terminal clades** are those that have only one line connecting them to the rest of the network. **Internal clades** have two or more connections to other parts of the network. For example, in the figure below, haplotype 4 (a 0-step clade) is *terminal* but clade II (a 1-step clade) is *internal*. The process of nesting clades at any given level always starts with the terminal clades.

The following is my summary of Templeton and colleagues' rules. If you have more than one independent network produced by *TCS*, you will have to apply this procedure to each network. Do not attempt to link up these independent networks. However, using *PAUP** in a latter lab, you could calculate or estimate the number of parsimony steps between your independent clades, and use this information to place all your networks into one cohesive illustration.

(1) Starting with **terminal** *x*-step clades (*e.g.*, the terminal haplotypes), join these to the node one mutation step inward, *i.e.*, towards the interior of the network, into an *x*+1 clade. During the first round of nesting, you are nesting 0-step clades into 1-step clades. During the second round, you will be nesting 1-step clades into 2-step clades, and so on.
   (A) When multiple terminal clades connect to the same internal node or haplotype, unite all of these terminal haplotypes (or *x*-step clades) into a large *x*+1 clade.
   (B) In the rare case of a "symmetrically stranded clade," in which a node, haplotype, or *x*-step clade is all by itself and unable to be nested with another element like itself, then you need to make a decision about what to do with it.
      (i) If this stranded clade is just an unsampled haplotype, forget about it.
      (ii) Otherwise, this stranded *x*-step clade should be included with the neighboring *x*+1 step clade has fewer samples.
      (iii) If the neighboring *x*-step clades have the same number of samples, just flip a coin.
(2) For any internal haplotypes or nodes (*i.e.*, *x*-step clades) not yet incorporated into a terminal *x*+1 step clade, those haplotypes, nodes, or *x*-step clades located in the network next to (interior to) the *x*+1 clades are now considered terminal clades and you repeat the step 1 above until all the samples are group into *x*+1 clades.
(3) Next, repeat procedure (steps 1 & 2 above) to make *x*+2 groups, etc. The result might look like the example to the right, from Templeton et al. (1987).
(4) What about reticulations (loops) in your network? You nest the unambiguous portions of the network first. The entire loop itself is nested with the unambiguous portions of the network only when you reach that nesting level at which no matter how the ambiguous connects were resolved you know that the ambiguously connected haplotype or clade would have to be a member of that particular *x*-step clade. In the example at right (from Templeton & Sing [1992]), note that haplotype 12 is separated from haplotypes (3,8,15) by two mutation steps. Haplotype 12 is also separated from haplotype 5 by two mutational steps. Thus, haplotype 12 formed its own 1-step clade, but was combined with other haplotypes at the 2-step clade level. Likewise, haplotype 13 is separted from other haplotypes by

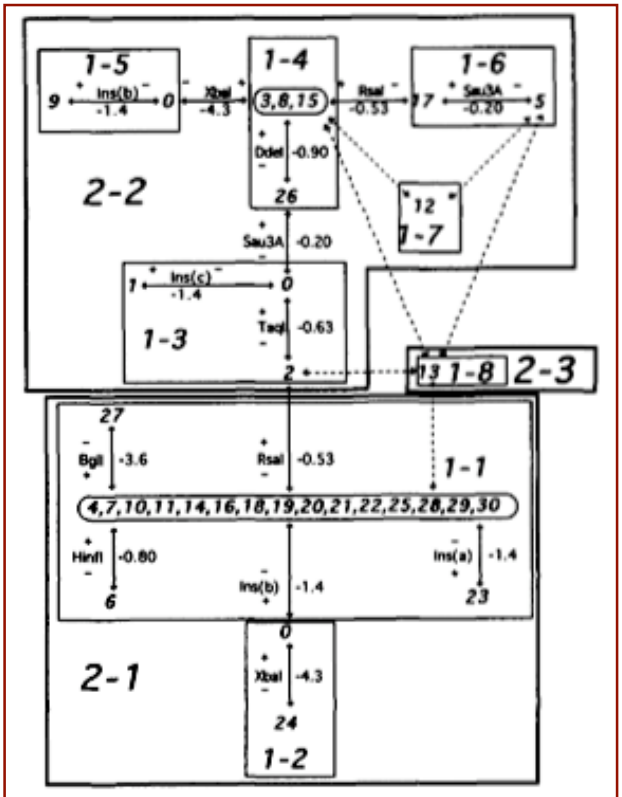

FIGURE 6.—The cladogram set estimated by the TEMPLETON, CRANDALL and SING (1992) algorithm as derived from the 42 lines described in GAME and OAKESHOTT (1990) for the 5′ subregion of the *Est-6* DNA region of *D. melanogaster*. The haplotypes are numbered as given in GAME and OAKESHOTT (1990). However, some of these haplotypes are identical for the variable sites in this 5′ subregion, and accordingly are pooled together as shown by the rounded boxes. Each arrow indicates one mutational event. The description of the event is indicated by the arrow, using the notation given in GAME and OAKESHOTT (1990). Solid arrows indicate transitions that are unambiguous whereas dashed arrows indicate alternative connections that are all likely, using the criteria given in TEMPLETON, CRANDALL and SING (1992). Single-width lines enclose haplotypes that are nested together to form 1-step clades, as designated by the notation *1-#*. Double-width lines enclose the 1- step clades that are nested together to from 2-step clades, as notated by *2-#*.

at least three mutation steps, thus it formed both its own 1-step clade and its own 2-step clade. Hopefully, the example at right will give you an idea of what is supposed to happen, according to Templeton.

(5) Following the above rules on reticulations, you might well end up with two clades with the same content, one nested inside the other. However, it appears that *GeoDis* does not actually want to analyze the "same" clade twice, so it does not matter much how many times you encircle the same clades, waiting for the level at which the reticulation (ambiguity) is resolved. So, do not worry too much about this. Also, with long or complex reticulations, Templeton does not give us much guidance. If your network is highly reticulated, either just try your best, or decide to analyze a reasonable subset of the data.

(6) Do not forget to label your nested clades. Each clade is designated by a code number, "*X*-*N*". *X* is the clade level (*e.g.*, 1-step, 2-step, etc.). *N* represents an arbitrary but sequential numbering scheme from 1 to *N* number of clades at a given level. (Clades with no observed haplotypes are not numbered or included in the analysis.) Again, see the example at right (above).

II. Creating the infile for *GeoDis*

If you have not already opened a copy of the documentation file, GeoDis2.4.pdf, do so now. You will need to create your data input file using a basic text editor such as *WordPad* or *TextWrangler*. Create this file now, save and name it appropriately. Each independent network will need its own data input file. Note, do not leave blank spaces at the end of lines. As you read the instructions below, also refer to the documentation, GeoDis2.4.pdf.

(1) The first line of text is the name of your data. Note, if you have more than one network, the data file will include only that corresponding subset of your data. **Separate networks must be analyzed separately**.

(2) The second line of your data file will be the total number of localities you have in the network to be analyzed.

(3) Next you will specify the geographic information about your sampling localities. Each locality will have two consecutive lines of text. The first line will contain the number of the locality (starting with 1, and numbering arbitrarily but consecutively) and then the name of the locality. The second line for each locality will include the number of samples from that locality followed by the latitude and then longitude for the locality. Latitude and longitude may be recorded as degrees minutes seconds or as decimal degrees. For more information on how to record the GPS data, see the bottom of page 3 of the *GeoDis* documentation. See page 5 of the documentation for an example data file. Note, *GeoDis* gives you the option of defining your own distances (instead of using lat.-long. data), in which case you could incorporate habitat heterogeneity, define distance by rivers, account for direction of ocean currents, etc. For this lab we are just using GPS points.

(4) Next you will define the clade structure. **Note, that a clade without BOTH geographic and genetic variation is NOT included in the input file** for *GeoDis*, though these samples may be a part of a higher (*x*+1) clade. The next line of your input file will contain the number of nesting clades that have geographic and genetic variation, **not** counting the total network. The total network will be an additional clade at the end of your input file.

(5) Next you will write a series of lines of text for each of your nesting clades describing the contents of that clade. We can refer to this nesting clade as an *x*+1 step clade. Each of the following items goes on one more lines.

(A) First line to describe a nesting (*x*+1 step) clade is of course the name, e.g., "Clade 1-3"

(B) Second line = number of *x*-step clades contained within this nesting clade. Be sure to count even the *x*-step clades that have no geographic or genetic variation. Do not count any *x*-1 step clades, of course.

(C) Third line contains the names of the clades contained with this nesting clade. Note, haplotypes are 0-step clades, and these can be given proper names or roman numerals, as you wish. 1-step clades will contain haplotypes while all higher clades will contain nested clades.

(D) The forth line will indicate for each of the nested (contained) clades whether it is terminal or interior. Terminal clades are designated by the number 1, interior clades are designated by 0. Simply list the 0's and 1's in the same order as the named clades in the above line, e.g., "1 0 1"

(E) The fifth line indicates the number of locations contained (represented) in the nesting (*x*+1 step) clade.

(F) Now specify these locations by their numbers (see item 3 above), again in the same order.

(G) Now for the confusing part! Basically, for each nesting (*x*+1 step) clade you need to make a tiny table with (*x*-step) clades as rows and localities as columns. In this tiny table you will report the number of samples that each (*x*-step) clades has from each of the locations you listed (by number) in the line immediately above. The number of rows will be equal to the number of clades (or haplotypes) list in (B) (second line) above. See documentation for more information on how to do this.

(6) Repeat (A) through (G) for each clade that has genetic or geographic variation, and add a final line with the word "END". Remember, do not leave any trailing blank spaces on any lines of text. Your input file is ready to run!

III. Launch *GeoDis*

Double click on the *GeoDis* application. Do not select outgroup weights. If you recorded locality GPS coordinates as degrees minutes seconds, then do NOT select decimal degrees. Click on Output file and name your so-to-be-created output file. Increase the number of permutations to 5,000 or 10,000. Then click RUN.

If your file did NOT run, here are some possible problems. (A) Review your input file and make sure you have the right number of clades and localities correctly indicated at all levels. (B) Do not leave blank spaces at the ends of any lines. (C) Your file might be in Unicode format, in which case change it to (ASCII) text. (D) Your localities must be numbered sequentially, 1-to-N localities. Do not try to use one master list of localities if you have multiple independent clades for your data. (E) Do not try to analyze just one clade. (F) Try quitting *GeoDis* and restarting it.

IV. Reading the output of *GeoDis*

For each clade in the input file (those containing geographic and genetic variation), *GeoDis* calculates clade distances and applies a set of three statistical tests of association of geographic location versus clades within a nesting level. The first test treats locations as categorical variables and applies a permutational contingency test. Significance is evaluated using a χ^2 statistic to the permutated data in which rows are genetic clades and columns are geographical locations (Templeton & Sing 1993).

A second set of statistical tests of association is applied to each clade, this time using information on geographic distances. For each test, the null hypothesis states that there is no significant association between geographic clade distances and clade membership. The null distribution of clade distances is obtained by a permutation test (randomization or Monte Carlo procedure) of clades against sampling location (following Hudson, Boos & Kaplan 1992). *GeoDis* tests for significantly small or large clade distances. Recall from lecture and from Templeton et al. (1995) that the WITHIN CLADE distance for clade X, $D_C(X)$, is the average geographic distance of haplotype sampling localities to the geographic center of all the haplotypes within clade X. Therefore, $D_C(X)$ measures the geographic spread of clade X. NESTED CLADE distance of clade X, $D_N(X)$, is calculated by first located the geographic center of all haplotypes in that clade containing clade X, e.g., clade Y. In other words, clade X is nested within the more inclusive clade Y. $D_N(X)$, is then calculated as the average geographic distance of all haplotypes in clade X from the geographic center of clade Y. Thus, $D_N(X)$ measures how far away clade X is located from other clades at the same hierarchical level in the network.

The third set of tests conducted by *GeoDis* involves a comparison of the geographic distance (or geographic extent) of terminal (or tip) clades versus interior clades. The test is applied to each nesting clade containing both terminal and interior clades. Significant differences are again evaluated by a permutation test. Recall, for example, that the model of "restricted but recurrent gene flow" predicts that older clades (*i.e.*, interior clades) should be more geographically widespread than terminal clades (Templeton et al. 1995).

IV. Applying the Inference Key

Look for any significant *p*-values. For those clades with significant *p*-values, apply the Inference Key from November 2005, available on Posada's webpage: `http://darwin.uvigo.es/software/geodis.html`

Lab Report

Please answer the following questions.

(1) Did you have any *a priori* expectations regarding whether your species, or certain populations within your species, might conform to one of the three scenarios proposed by Templeton and colleagues:
(A) Restricted but recurrent gene flow?
(B) Range expansion?
(C) Allopatric fragmentation (range contraction)?

(2) Did you uncover any significant associations between clades, clade age and geographic distances? If so, please create a table based on the following example:

Clade	Test (distance or age)	*P*-value	Inference
Haplotypes nested in 1-2			
Clades nested in 2-3, etc			

and provide any biological, geographical or historical interpretations for these results.

If you did not find any significant associations using *GeoDis*, please explain why you think you obtained these non-significant results.

(3) Please illustrate at least one statistical parsimony network for your data set, and include this network in your final report. If you have multiple networks, choose one with significant results or else choose the one with the most haplotypes. In your figure indicate your nesting scheme and label clades according to your *GeoDis* analysis. Also label haplotypes or clades by geographic location, as necessary. Be sure the reader understands the association between the genealogical information provided by your network and the geographical location of the sampling localities from which the haplotypes originated. You probably should NOT use your output from *TCS*. Rather, you should create simpler, more readable network diagrams using a graphics program. Check the literature for nice examples, or ask the Profe.

Your answers, the table (if necessary) and at least one professional-looking network illustration are due next Saturday, March 11, 2006.

Professor Alan R. Templeton thanks you!