Summary

*GeoPhyl* is an Excel workbook that allows students to explore evolution over space and time using published data on the invasive plant *Tamarix* (salt cedar). The user begins by choosing a specific portion of *Tamarix*’s phylogenetic tree. The program then plots the geographic location of samples representing that portion of the tree, as well as the location of samples representing the next higher level within the tree’s nested structure. The program also calculates statistics that describe the samples’ geographic distribution. Finally, *GeoPhyl* conducts a permutation test that allows the user to determine, both graphically and numerically, how well the observed data match a random geographic distribution.

Introduction

The development of methods for collecting and analyzing DNA and protein sequence data has revolutionized biology, allowing researchers to identify previously unknown evolutionary relationships, elucidate gene functions, and reconstruct the evolutionary history (or *phylogeny*) of genes and species. Still greater insight can be gained by analyzing this genetic data in a geographic context, a field known as *phylogeography*. With this added spatial information, we can begin to reconstruct past events such as migrations, range expansions and contractions, and fragmentation of populations.

![Figure 1A. A haplotype network. Each box represents a distinct genetic sequence, or haplotype. Line segments, or branches, connect related haplotypes; their length is proportional to the number of mutations separating those two haplotypes. The dotted line denotes a particularly long branch.](image1a)

![Figure 1B. A map of the sites from which individuals were sampled. Each circle represents a sample site. These figures have been modified from a published study of tiger salamanders (Templeton et al., 1995).](image1b)
A key part of phylogeographic analysis consists of determining what phylogeographic patterns would be observed under specific models of dispersal. One can then compare actual phylogeographic data to these predicted patterns, and rule out any processes inconsistent with the data. For example, under a panmixia model in which each individual disperses at random over the entire geographic range, haplotypes should be randomly distributed (Figure 2). By contrast, under a model of limited dispersal ability (isolation by distance), old haplotypes will be randomly distributed, but younger ones will be geographically restricted to a smaller region near their close relatives (Figure 3).

Several statistical procedures have been developed for quantifying phylogeographic patterns. One of these methods (Templeton et al. 1995) is called nested cladistic analysis.
NCA works by breaking the haplotype tree into small clusters, or *clades*\(^1\), of closely related haplotypes. These clades are grouped together into higher-level clades of less closely related haplotypes, which are grouped into still higher-level clades, and so on until the final clade encompasses the whole tree (Figure 4). Note that the clade structure depends only on the phylogeny and *not* on any geographic data.

Within each clade, NCA calculates two statistics based on the geographic locations of individuals bearing haplotypes within that clade (see Figure 5):

- \(D_C\), the *clade distance*. This is the average geographic distance of samples in the clade from that clade’s geographic center. \(D_C\) can be interpreted as how far the clade has dispersed outward from its center.

- \(D_N\), the *nested clade distance*. This is the average geographic distance of samples in the clade from the geographic center of the clade *one level up*. \(D_N\) can be interpreted as how far the clade has dispersed from its inferred evolutionary origin.

---

\(^1\) In phylogenetics, the term "clade" generally means any group including all the descendants of a single ancestor. NCA clades don’t necessarily fit this definition.
Figure 5. Depictions of $D_c$ and $D_N$. Blue triangles represent samples of our study clade. Suppose that this blue clade is nested with the red and green clades in a higher-level clade. Colored stars denotes individual clades’ centers, and a black star denotes the center of the higher-level clade. $D_c$ is the average distance of the blue samples from their (blue) center; $D_N$ is the average distance of those same samples from the nested clade’s (black) center.

We can then translate our earlier qualitative phylogeographic predictions into specific hypotheses about the values of $D_c$ and $D_N$. For example, we noted above that a model of isolation by distance predicts restricted distributions of younger haplotypes but broader distributions of older haplotypes. This corresponds to a prediction of small $D_c$ values for tip clades (those connected to only one other clade at that level of nesting) but not for interior clades (those connected to two or more other clades). Nonparametric statistical tests can be performed to assess the significance of observed $D_c$ and $D_N$ values and reject specific phylogeographic hypotheses: see the Model details section below for further information.

The specific data set included in GeoPhyl comes from a published study of Tamarix (Gaskin and Schaal 2002). This plant genus, native to Eurasia, comprises the second worst plant invasion in the United States, with its worst impact along desert riverbanks in the Southwest. Phylogeographic analysis can help reveal both the history of Tamarix introductions into the U.S. and patterns of gene flow between habitat patches. More details on Tamarix biology and additional data are available from the BEDROCK website (see Resources section below).

Output: What do I see?
There are two files associated with GeoPhyl. The first, Tamarix.jpg, is a JPEG file showing the evolutionary network of Tamarix haplotypes from a published study (Gaskin and Schaal 2002). The second file is GeoPhyl itself, an Excel workbook containing both genetic and geographic data for viewing and analysis. We will begin with a description of Tamarix.jpg and then explore GeoPhyl.

Tamarix.jpg
This file shows the haplotype network of *Tamarix* samples in the Gaskin and Schaal study. Each numbered box corresponds to one observed haplotype. The size of the box indicates how many copies of that haplotype that were observed. The box pattern indicates the geographic locations in which the haplotype was observed: in both Asia and the U.S. (shaded), in the U.S. only (dashed), in Asia only (solid), or in neither Asia nor the U.S. (crossed). The very small boxes indicate missing links: haplotypes that were not observed but which lie between observed haplotypes and are therefore inferred to have occurred during the evolution of *Tamarix*. Each branch in this diagram represents a single mutation event. For example, in the upper right-hand corner, there are two mutations between haplotypes 11 and 26, but only one between haplotypes 7 and 11.

The colored blobs, called *clades*, are groups of related haplotypes. The basic units are individual haplotypes, which we can also think of as level-zero clades (for example, we can denote haplotype 1 as clade 0-1, haplotype 2 as clade 0-2, and so on). Moving up one level in the nesting scheme, the level-one clades shown in red contain two or more closely related level-zero clades. We assign identification labels to each level-one clade that contains at least one observed haplotype, following the same scheme as for the level-zero clades (e.g., 1-1, 1-2, etc.). We then move up another level and draw the level-two clades, shown in orange, which contain closely related level-one clades. This process is repeated until the final clade (sometimes omitted) would encompass the entire network. (More details on the nesting procedure are given in Templeton *et al.* 1995 and BEDROCK 2004.)
The main interface of the *GeoPhyl* workbook is on the "Map" worksheet. The clade currently selected for analysis (the *reference clade*) is displayed in cell K2. This cell is outlined in red for easy reference. Immediately to the right in cell L2 is the *nested clade*, the clade one level up that contains the reference clade. Immediately below each of these clades is the sample size $N$, the number of haplotype samples lying within that clade. For example, 13 copies of haplotype 53 were observed, so $N = 13$ for the clade 0-53. If a particular clade doesn't contain any haplotypes found in the U.S., then $N = 0$ for that clade. The final row in this top section displays the tip/interior status of both the reference and the nested clades. Recall that a tip clade is one that connects to only one other clade of that level of nesting, whereas an interior clade connects to two or more such clades.

The large map in the center of the worksheet displays the geographic locations of samples within these two clades. Pink diamonds represent samples within the reference clade; blue diamonds represent samples within the nested clade. Note that some samples may come from the same population or even the same individual *Tamarix* plant, so their location symbols may overlap. The map also displays the geographic centers of the reference and nested clades, respectively shown as the red and blue solid circles. We will explain the other symbols shortly.

The "Actual Data" table on the center right-hand side shows the latitude and longitude of both the geographic centers of both the reference clade (red) and the nested clade (blue). The table also shows the clade distances (in miles) $D_C$ for both clades and the nested clade distance $D_N$ for the reference clade. Recall that $D_C$ is the average geographic distance of samples in a clade from that clade's geographic center, while $D_N$ is the average distance of samples in the reference clade from the *nested* clade's center.
Now that we know the actual values of $D_C$ and $D_N$, how can we interpret them? For example, is a $D_C$ value of 407.04 miles for a given clade unusually large or small, or is it pretty much what we would expect? To answer this question, we must determine the underlying distribution from which the observed value is sampled. GeoPhyl does this using a permutation procedure in which individual haplotype samples are randomly shuffled around among the sample locations.

On the map, the permuted locations of samples in the reference and nested clades are respectively indicated by red and blue hollow squares, while the geographic centers of these permuted locations are indicated by the red and blue hollow circles. The "Permuted Data" table in the lower right-hand corner displays the geographic co-ordinates of the permuted clades' centers, as well as $D_C$ and $D_N$ statistics for the permuted data. The "Calculate Now" command (Command + '=' on Macs) makes it easy to run new permutations quickly. We can then compare the actual data to the results of multiple permutations, either on the map itself or by analyzing the tabulated statistics. For example, looking at clade 0-53, we see that the center of the permuted reference clade (hollow red circle) is consistently to the northwest of the clade's actual center (solid red circle), indicating a marked southeastern distribution for this clade. Moreover, permuted $D_C$ values for this clade are generally above the actual $D_C$ value, suggesting that the clade's range may be geographically restricted.

The other worksheets in GeoPhyl can be accessed through the tabs at the bottom left. These sheets also contain useful information but are generally of secondary importance. "DC, DN Table," the first of these sheets, shows sample sizes, geographic centers, and other statistics for every clade in the U.S. sample. Note that when two or more clades contain exactly the same individual samples, values are listed only for the highest-level clade. For example, clade 0-52 contains five samples. It is nested within clade 1-17, which contains only those five same samples, so a comparison of clade 0-52 vs. clade 1-17 is not informative. However, 1-17 is nested within 2-8, which contains 119 additional samples from clade 1-18. Thus, comparisons of 1-17 vs. 2-8 are informative and are listed in the table.

The last two sheets in GeoPhyl contain the actual and permuted geographic data, as well as the underlying calculations. Two columns on the "Actual Data" sheet are of particular interest: column D lists the genotype of each individual sample (i.e., that sample's two haplotypes), and column E lists the raw geographic data from which latitude and longitude co-ordinates were determined. Most of the remaining columns are self-explanatory and/or redundant with the results tabulated on the "Map" sheet.

Controls: What can I do?
While GeoPhyl's complex output can sometimes overwhelm a new user, the controls are refreshingly simple. The primary control is the choice of a reference clade, which can be entered in cell I2 on the "Map" worksheet. (Remember that haplotypes should be entered as zero-level clades; for example, 0-1 or 0-52.) The program will then automatically determine the appropriate nested clade, calculate all sample sizes and other statistics, and
update the map. If you want to run multiple permutations on the same clade, simply use the "Calculate Now" command (Command + "=" on Macs).

The workbook is protected so that users will not unintentionally overwrite crucial formulas or data. If you want to make changes beyond those outlined above, such as hiding or un-hiding specific columns or changing the model, you must first use the "Unprotect Sheet" command. In Excel 2001, this command is on the Tools menu.

A customizable version of GeoPhyl, in which the user can enter his or her own data set and haplotype tree for phylogeographic analysis, is currently in the planning stages.

**How it works: Model details**

Columns Z – AC of the "Actual Data" worksheet contain information on the structure of the haplotype network. When the user chooses a reference clade, GeoPhyl uses this table to determine the appropriate nested clade and to calculate the sample size and tip/interior status of each clade. It also determines which haplotypes are contained within the two clades.

Next, GeoPhyl separates the genotype of each individual sample into its two constituent haplotypes and determines how many of those are contained within the reference clade. This number is entered in column L. The latitude and longitude of the clade's geographic center is then computed as the arithmetic mean of each sample's latitude and longitude, weighted by the number of haplotypes from that sample are contained within the reference clade. Next, each sample's great-circle distance from this geographic center is calculated as $D_i = 2R \arcsin (k/2R)$, where $R$ is the Earth's radius and $k$ is the sample's straight-line distance from the geographic center, given by

$$k = R \sqrt{(\sin \theta_i \cos \phi_i - \sin \bar{\theta} \cos \bar{\phi})^2 + (\cos \theta_i \cos \phi_i - \cos \bar{\theta} \cos \bar{\phi})^2 + (\sin \phi_i - \sin \bar{\phi})^2}.$$  

Here $\theta_i$ and $\phi_i$ are the sample's longitude and latitude, while $\bar{\theta}$ and $\bar{\phi}$ are the longitude and latitude of the reference clade's geographic center (see Figure 8). The clade distance $D_c$ for the reference clade is then calculated as the average of these $D_i$ sample distances, again weighted by the number of haplotypes from that sample contained within the reference clade.

---

2 This method is equivalent to assuming that the sample points lie within a plane; i.e., that the world is flat. This assumption obviously breaks down for species with very broad and/or circumpolar distributions. Nonetheless, the planar approximation has two major advantages. First, determining the center of a group of points on a sphere's surface is a surprisingly difficult problem with no known algebraic solution. The planar method is vastly easier to implement. Second, many circumpolar species are confined to narrow bands of latitude: the planar approximation reflects this biological fact better than the "more accurate" spherical method, which tends to place their center outside the limits of their actual distribution.
\[ x = R \sin \theta \cos \phi \]
\[ y = R \cos \theta \cos \phi \]
\[ z = R \sin \phi \]

\[ k = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2} \]