PHYLOGENETIC TARGETING

Christian Arnold and Charles L. Nunn

Department of Anthropology

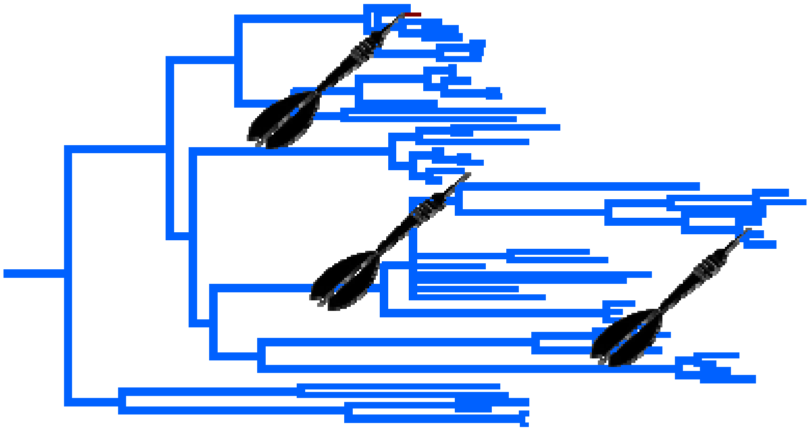
Harvard University

11 Divinity Avenue

Cambridge, MA 02138

http://www.fas.harvard.edu/~primecol

http://phylotargeting.fas.harvard.edu



Manual

Last updated: June 2009

In this document, we provide instructions, explanations and general comments for our method of phylogenetic targeting. The help is organized into different sections that correspond to the different steps in the web-implementation of *PhyloTargeting*. The help is not exhaustive; it is rather an addition to the help boxes that are provided throughout the application. More details can be found in Arnold and Nunn 2009 (in prep.), which will be available here in the near future.

If you have questions or comments, feel free to contact the administrator of the website and main author of the method, Christian Arnold (carnold@fas.harvard.edu). He will be happy to answer any questions related to phylogenetic targeting as well as questions related to the web-implementation of the program.

If you use the method of phylogenetic targeting, please cite the following reference:

Arnold, C. and C. L. Nunn. 2009. Phylogenetic Targeting of Research Effort in Evolutionary Biology. In prep.

0. General

The method of *phylogenetic targeting* requires a phylogeny, data for the trait(s) of interest and one or more explicit hypotheses that offer predictions for how variation in one trait (X_1) correlates with variation in another trait that is common to all the hypotheses and, because it is not known in all the species, is the "target" of the analysis (Y_t). We call this association between Y_t and X_1 the primary prediction. Additional predictions, if desired, are implemented through traits $X_2 \ldots X_t$, which relate to competing hypotheses or potentially confounding variables. The goal of the method is to identify species that should be studied with regard to Y_t by using information from already collected data for the *X* traits. Different targeting analyses are thus likely to focus on a primary main hypothesis and various combinations of alternative hypotheses. Scores are calculated so that higher values indicate more preferred species to study, based on user-defined criteria involving control of confounding variables, testing of alternative hypotheses, and availability of data on Y_t for one or more species in a clade.

1. Step 1

In step 1, the user may upload a dataset to the server to target species that should be studied using the method of phylogenetic targeting.

Dataset specifications

The dataset must be in the NEXUS format, and the following blocks are mandatory:

1. TAXA block with a list of species
2. at least one CHARACTER block with character data for the species
3. TREE block with at least one phylogenetic tree (with branch lengths)

NEXUS files that have been generated using the software Mesquite (http://mesquiteproject.org) (Maddison and Maddison 2006) are known to have full compatibility. Thus, we advise to either create the datasets in Mesquite or converting the datasets using Mesquite.

Supported characters

PhyloTargeting supports both discrete and continuous characters, although continuous characters offer more power for targeting species and should thus be preferred whenever possible. Discrete characters can be furthermore treated as ordered or unordered (see step 2) if they have three or more states.

Phylogenetic trees

The phylogenetic trees can have an arbitrary topology; however, nodes must not be labeled and branch lengths must be assigned for all branches (numbers must be non-negative). Missing branch lengths will be automatically assigned a branch length of 1. It is also important to know that hard and soft polytomies are treated differently in our approach. For details, see the maximal pairing section in step 4!

Example dataset

In what follows, we list an example data file that has full compatibility with *PhyloTargeting*:

```
#NEXUS
[comments]

BEGIN TAXA;
      DIMENSIONS NTAX=3;
      TAXLABELS
            Allocebus_trichotis Arctocebus_calabarensis Avahi_laniger
      ;
END;

BEGIN CHARACTERS;
      DIMENSIONS  NCHAR=3;
      FORMAT DATATYPE = CONTINUOUS;
      CHARSTATELABELS
            1 Group_Size,
            2 Home_Range,
            3 Longevity ;
      MATRIX
      Allocebus_trichotis       4 ? ?
      Arctocebus_calabarensis   1 ? 13
      Avahi_laniger             2 2 ?
;
END;
BEGIN CHARACTERS;
      DIMENSIONS  NCHAR=2;
      FORMAT DATATYPE = STANDARD GAP = - MISSING = ? SYMBOLS = " 0 1 ";
      CHARSTATELABELS
            1 CognitiveStudy, 2 ActivityPeriod ;
      MATRIX
      Allocebus_trichotis      01
      Arctocebus_calabarensis  01
      Avahi_laniger            10
;

END;
BEGIN TREES;
      TRANSLATE
            1 Allocebus_trichotis,
            2 Arctocebus_calabarensis,
            3 Avahi_laniger;
      TREE 'mammals primates upperDates++' =
((1:55.10000000000001,3:55.100001000000006):24.5,2:79.6):10.8;

END;
```

We also provide a more complex example dataset in step 1. If you are not familiar with the program and its settings, loading the example dataset and exploring the options in the program is a good starting point.

2. Step 2

In step 2, the user has to specify some settings, which include the following:

1. Selection of species
2. Selection of the phylogenetic tree
3. Selection of traits
4. Selection of an availability variable (optional)
5. Choice if contrasts should be standardized

Here, we will describe options 3 to 5 in more detail. The first two options are sufficiently described in the help boxes in step 2.

Selection of traits

For predictions that only involve a primary hypothesis (i.e., only one independent variable), phylogenetic targeting uses a scoring system that maximizes the variability in X_1. In other words, species pairs are targeted that differ the most in X_1, as this increases the available range of variation and also enhances statistical power to test the hypotheses (Westoby et al. 1998; Westoby 1999; Garland 2001; Garland et al. 2005). If we were interested in hypotheses that involve body mass as an independent variable, for example, phylogenetic targeting gives pairs with the largest differences in body mass higher scores. Thus, pairwise comparisons with big differences in X_1 are scored more positively, whereas smaller differences are scored less positively. These contrasts are then standardized to the scale 0 to 1, with a difference of 0 assigned a score of 0 and the largest difference in all considered pairs assigned a score of 1. All other differences are assigned a score between 0 and 1 by applying a linear scaling transformation. We call this the score of X_1. Models that incorporate additional traits enable the testing of different kinds of hypotheses (e.g., mutually exclusive and non-mutually exclusive), and they are often used to control for confounding variables. For each $X_2 \ldots X_n$, a separate scoring mechanism is defined in which larger contrasts can have either a negative or a positive influence on the overall score. The decision for whether larger differences in each of the $X_2 \ldots X_n$ variable is scored higher or lower depends on whether the variables reflect confounding variables or a desire to distinguish among competing hypotheses. To simplify discussion in what follows, we consider a case in which only one additional variable is included; thus $Y_t = f(X_1, X_2)$.

To control for confounding variables, the goal is to minimize variation in the predictor variable that corresponds to the confounding variable of interest, i.e. X_2. Thus, pairwise comparisons in X_2 that make the absolute value of change in a particular confounding variable as small as possible are scored higher, whereas pairwise comparisons with bigger differences are scored lower. The smallest pairwise contrast is assigned a score of 1, whereas the maximum pairwise contrast is assigned a score of 0. All other differences are assigned a score between 0 and 1. The smallest pairwise contrast is always assigned 0 even if no pairwise comparison has a difference of 0 in this trait, as this ensures that non-zero differences are assigned to a score different from 0. To address mutually exclusive hypotheses, the goal is to maximize scores for X_2 that differ maximally from contrasts in X_1. Two different scoring options can be applied that both target big differences, but differ in how they score these differences. The first option scores differences in X_2 in the opposite direction as the difference in X_1 positively and differences in the same direction as X_1 negatively. The biggest difference in the opposite direction is assigned a score of 1, whereas the biggest difference in the same direction is assigned a score of -1. A difference of 0 is assigned a score of 0. All other differences are assigned a score between -1 and 1 by applying a linear scaling transform, which is calculated separately for positive and negative contrasts. The second option is exactly the opposite of the first option; that is, differences in the opposite direction from the difference in X_1 are scored negatively and differences in the same direction are scored positively. For example, this option might be useful if an increase in X_1 is predicted to reduce Y_t while an increase in X_2 is predicted to increase Y_t. Thus, it is necessary to give higher scores to contrasts in the same direction for X_1 and X_2 to distinguish among the hypotheses.

Selection of an availability variable (optional)

In addition to manually excluding species from an analysis, it is possible to define an “availability variable” to automatically exclude species or pairs in relation to the availability of data for Y_t. One can thus use the availability variable to identify other species that should be studied in the context of existing data on Y_t. An availability variable also provides a way to quickly “pinpoint” where the missing data points are in a phylogenetic context, which can help to identify biases in the distribution of the studied species. The availability variable must be a discrete binary variable because it identifies whether or not data are available for Y_t for a particular species. Possible options would be to only consider pairs where data are available for both species that form the pair, for one, for at least one species and for none of the species. These options are intuitive; for example, if the availability variable is a variable where data are available for all species, no pair will be excluded if the option is chosen to consider only pairs where one species has already been studied and data are needed for the other species. If only a small number of species have been studied in relation to Y_t, however, most of the pairs will be excluded and only those containing one studied species and one that has yet to be studied remain. It can thus be seen as an additional selection factor that effectively constrains the species that will be targeted.

Choice if contrasts should be standardized

Regardless of the scoring model, the summed score of a pairwise comparison can sometimes be uninformative when compared among different pairs because the more divergent two species are, the more likely it is that they evolved bigger differences. In other words, different pairs will have different expected amounts of change (i.e., variance). In our approach, we provide an option to overcome this problem by normalizing the summed score by its expected variance (square root of the sum of the branch lengths that connect the two species) (Felsenstein 1985; Garland et al. 1992). We call this the standardized summed score. By doing so, all pairwise comparisons have a common variance as required by most statistical tests.

Changing this option alters the summed scores of pairwise comparisons substantially, which ultimately also alters the maximal pairing. Generally, standardizing contrasts greatly reduces the summed score of divergent species pairs. Thus, they are less likely selected in the maximal pairing.

3. Step 3

In step 3, a summary of all calculated pairwise comparisons is shown. In addition to the help boxes in step 3, we describe some features in more detail.

Pairing statistics

If the user selected at least one pairwise comparison, some basic statistics are displayed above the actual summary table with details about the current pairing (set of selected pairwise comparisons). For more information, see the help boxes in step 3!

Alternative representations for the summary table

In addition to the summary table, we provide alternative representations of the calculated pairwise comparisons, their scores, and their phylogenetic dependence. The user can display a graphical representation of the phylogenetic tree and the current pairing, as well as a PDF version and a text version of the summary table. The latter may be better suited for further analysis (e.g. for importing them into spreadsheet programs) and printing than the HTML version of the table. For more details, see the help boxes in step 3!

Summary table

The summary table shows all calculated pairwise comparisons, trait differences and scores, phylogeny-related information, summed scores, and the possibility to manually select individual pairwise comparisons for selecting data points that are phylogenetically (and thus also statistically) independent. In what follows, we provide additional information to the help boxes in step 3.

For each of the traits that have been selected in step 2, the website displays a column that prints the actual trait differences for all pairwise comparisons as well as the score of those differences, based on the scoring system defined in step 2. *PhyloTargeting* forces the difference in X_1 to be positive and achieve consistency with other widely-used programs, such as CAIC (Purvis and Rambaut 1995) and PDAP-Mesquite (Midford et al. 2005). This “positivization assumption” also helps to make sense of the other trait differences and their directions, as it becomes possible to determine whether other pairwise comparisons are consistently positively or negatively associated with X_1 (e.g., if X_2 is positive, it must be in the same direction as X_1). This helps to guide the manual selection of contrasts. As highlighted in step 2, the user can also include additional traits $X_{2\ldots}$ X_n. The direction of change for $X_{2\ \ldots}$ X_n always refers to the direction of change in X_1, e.g. a positive value means that the direction of change is the same as in X_1. For more details on the scoring system, see step 2.

Lastly, the scoring columns show the “summed score” and the “standardized summed score” for each pairwise comparison. For each pairwise comparison, the scores for all traits $X_{1\ \ldots}$ X_n are summed up to define the summed score. The summed score combines the information from all traits and thus represents the strength of a pair for testing the hypotheses. For models with only Y_t and X_1, the summed score equals the score of X_1. If contrasts are not standardized, “summed score” and “standardized summed score” are identical. The maximal pairing (see step 4), however, is always based on the latter.

4. Step 4

In step 4, the user finds some additional features that facilitate targeting species. In general, those options further constraint the set of pairwise comparisons as shown in step 3. In particular, the user can calculate the maximal pairing. The following three options are available in step 4:

Display only pairwise comparisons from the following species

This option shows only pairwise comparisons from a particular species. After the user selected the species of interest, the summary table from step 3 will be displayed. The only difference to step 3 is that only pairwise comparisons with the species of interest are shown. This option is thus helpful for larger datasets with a high number of pairwise comparisons.

Display only selected pairwise comparisons from step 3

This option shows only pairwise comparisons that have been already selected in step 3 (by using the SELECT link). It is thus possible to display only the pairwise comparisons that belong to the current pairing.

Calculate the maximal pairing

The actual selection of species is performed by a dynamic programming algorithm that we call maximal pairing. The maximal pairing, which selects pairs of species that are phylogenetically independent is thus the central component of phylogenetic targeting. The algorithm selects data points that are independent in a statistical sense, making it possible to analyze the data using standard statistical methods (Felsenstein 1985; Harvey and Pagel 1991; Garland et al. 1992). Unlike phylogenetically independent contrasts, however, only pairs between the tips of the tree are selected. The selection of pair is based on the summed score for each pair, and the algorithm determines the set of phylogenetically independent pairs (which we call a pairing) that maximizes the sum of the individual summed scores (maximal pairing). This criterion is thus assumed to maximize the power to test the hypotheses, given the phylogenetic constraints. For models that involve only X_1, for example, the maximal pairing generally selects pairs of closely related species that maximize differences in X_1, and those pairs are only distantly related to each other. In a comparative test, such a design is considered to be especially powerful (Garland et al. 2005). If, however, an additional trait X_2 is used to control for confounding variables (thus scoring small differences in X_2 higher), the algorithm both maximizes differences in X_1 and minimizes differences in X_2. Conversely, if one aims to maximize differences in X_2 (thus scoring larger differences in X_2 opposite to X_1 higher), the algorithm maximizes differences in X_1 and maximizes differences in X_2 opposite in sign to X_1. For more details, we refer to Arnold (2008).

Hard and soft polytomies are treated differently, as follows. Polytomies that are defined as a series of zero-branches (soft polytomies) are treated as a series of true dichotomies. Here, in most cases, fewer pairs can be selected, due to the fact that no branch can be shared twice. If the polytomy is defined as hard (i.e. split into more than two lineages), multiple pairs can go through the polytomous node without violating phylogenetic independence in the maximal pairing (see step 4). Zero-length branches should be nevertheless treated with caution, since the arbitrary order of zero-branches might change the maximal pairing considerably. Thus, if possible, polytomies should be treated as hard.

5. Step 5

In step 5, the user can save the current analysis (settings, calculated pairwise comparisons, etc.) to a file. The analysis may be continued at a later date by uploading the saved file to the server in step 1. For security reasons, the file is stored encrypted, and any modification results in an error message when trying to restore a previous analysis, making it impossible to proceed.

6. References

Arnold, C. 2008. Phylogenetic Targeting: A Systematic Approach and Computer Program for Targeting Research Effort in Comparative Evolutionary Biology, University of Leipzig, Leipzig.

Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

Garland, T. 2001. Phylogenetic comparison and artificial selection - Two approaches in evolutionary physiology, Pages 107-132 *in* R. C. Roach, P. D. Wagner, and P. H. Hackett, eds. From Genes to the Bedside (Advances in Experimental Biology and Medicine). New York, Kluwer Academic/Plenum Publishers.

Garland, T., A. F. Bennett, and E. L. Rezende. 2005. Phylogenetic approaches in comparative physiology. Journal of Experimental Biology 208:3015-3035.

Garland, T., P. H. Harvey, and A. R. Ives. 1992. Procedures for the analysis of comparative data using phylogenetically independent contrasts. Systematic Biology 4:18-32.

Harvey, P. H., and M. D. Pagel. 1991, The Comparative Method in Evolutionary Biology: Oxford Series in Ecology and Evolution. Oxford, Oxford University Press.

Maddison, W. P., and D. R. Maddison. 2006.Mesquite: a modular system for evolutionary analysis, version 2.5.http://mesquiteproject.org.

Midford, P. E., T. Garland, and W. P. Maddison. 2005.PDAP Package of Mesquite, version 1.08(2).

Purvis, A., and A. Rambaut. 1995. Comparative analysis by independent contrasts (CAIC): an Apple Macintosh application for analysing comparative data. Computer Applications in the Biosciences 11:247-251.

Westoby, M. 1999. Generalization in functional plant ecology: the species sampling problem, plant ecology strategy schemes, and phylogeny, Pages 847–872 *in* F. I. Pugnaire, and F. Valladares, eds. Handbook of functional plant ecology. New York, M. Dekker.

Westoby, M., S. A. Cunningham, C. Fonseca, J. Overton, and I. J. Wright. 1998. Phylogeny and variation in light capture area deployed per unit investment in leaves: designs for selecting study species with a view to generalizing, Pages 539–566 *in* H. Lambers, H. Poorter, and M. M. I. V. Vuren, eds. Variation in growth rate and productivity of higher plants. Leiden, The Netherlands, Backhuys Publishers.