User guide for
PhyloSuscept: Phylogenetic prediction of host ranges of plant pests and pathogens
cisr-stats.cisr.ucsc.edu/aphis/

by
Gregory S. Gilbert
Environmental Studies Department
University of California Santa Cruz
Santa Cruz, CA 95064 USA
ggilbert@ucsc.edu

27 Feb 2015

The host ranges of most plant pests and pathogens are somewhat predictable; they are more likely to be able to attack closely related plant species. This "phylogenetic signal" -- where close relatives are more likely to share a pest than are distant relatives (Gilbert *et al.* 2012) -- forms the basis of PhyloSuscept, a phylogenetically based analytical tool, designed to help predict which plant species are most likely to be susceptible to a particular pest or pathogen when available empirical data are limited.

For instance, imagine that an insect pest has just been detected in your state for the first time. From the literature, you learn that it has been reported on a number of plant hosts in its native range. Although most of these hosts species may not be present in your region, the pest may be pre-adapted to be able to attack local relatives. PhyloSuscept helps you quickly screen a list of important crops or locally important wild species of concern to help you identify which species are most likely to be susceptible and which may merit special attention for surveillance or testing.

PhyloSuscept comes in two flavors, built on the same underlying engine. The **two-file version** allows users to submit a file of known hosts for a particular pest and a list of plant species of concern. PhyloSuscept returns the probability that each of the species of concern would be susceptible to a pest of that known host range. The **phytosanitary risk version** comes with a default list of 1369 plant host taxa of agricultural, forestry, or environmental significance; the user provides a list of known hosts for a pest of interest, and the susceptibility of those phytosanitary host taxa is evaluated.

Remember, this is a statistical tool and still in development. It is designed to provide information and guidance, but probabilities are just that -- pests and pathogens have a tendency of beating the odds when you don't necessarily expect it. Use PhyloSuscept at your own risk.

Below are basic instructions on the use of the two-file and phytosanitary versions, as well as a description of how PhyloSuscept works. Please send comments, corrections, and suggestions to ggilbert@ucsc.edu.

How to use PhyloSuscept
(two-file version, cisr-stats.cisr.ucsc.edu/aphis/)

1. Create two comma-delimited (.csv) files containing genus and species names, one for known hosts and one for species of concern. You should include the headers Genus and species in the file, and if desired, infraspecific. If you do include infraspecific ranks it will be handled at the level of species in that genus, not as a subspecific taxon. A genus with sp. or

spp. is treated the same as one with a recognized species name. As an example, you could create the file *known.csv* for known hosts of your pest:

Genus	species
Acacia	confusa
Albizia	julibrissin
Albizia	sp.
Robinia	pseudoacacia
Cornus	foemina
Hydrangea	quercifolia
Solidago	sp.

Perhaps the species you are concerned about are the following, which you save as *myspp.csv.* Note that here the common names (or could be infraspecific taxa) are included in the third column for user convenience and labeling; they would be treated at the species level, not as true subspecific taxa).

Genus	species	infraspecific
Beta	vulgaris	Beet
Spinacia	oleracea	Spinach
Rheum	x_hybridum	Rhubarb
Solanum	tuberosum	Potato
Solanum	lycopersicum	Tomato
Daucus	carota	Carrot
Fagopyrum	esculentum	Buckwheat
Zea	mays	Corn
Lactuca	sativa	Lettuce

2. Upload each file by clicking the Browse buttons for "known species list" and "concerned species list", and then click Upload and Execute.

3. When complete (this may take several minutes for extensive lists of species), click where it says "Download a zip file with the output from the analyses HERE" to download a compressed bundle of output files. Double click on the resulting *deployr_directory_archive.zip* file to expand to a folder called "output". The output folder contains nine files, each with custom names built from the names of the files you submitted (here *known.txt* and *myspp.txt*):

known_to_myspp_outputprobs.csv (open in Excel) Rows are the species of concern. Columns are the major groups of plant pests. The values are the calculated probabilities that a pest with the given host range will be able to attack each of the species of concern. These are relative values that require user judgement to evaluate. Values close to one (1) are highly likely to be susceptible; values below 0.25 are at background levels with a low probability of susceptibility.

known_to_myspp_distancematrix.csv (open in Excel) Phylogenetic distances from (in millions of years of independent evolution) between your species of concern (in rows) and each known host (in columns). These are the values used to calculate the outputprobs.csv.

known_and_myspp_phylotaxa_out_dated_phydist.csv (open in Excel) Phylogenetic distances (in millions of years of independent evolution) between each pair of species in the combined

species list. These are calculated from the dated phylogenetic tree (.new) using the *cophenetic* function in the *Picante* package in R.

known_tree.png (open in an image viewer like Preview or Windows Photo Gallery) A phylogenetic tree of the known host species, including estimated phylogenetic distances (in millions of years).

myspp_tree.png (open in an image viewer like Preview) A phylogenetic tree of the species of concern, including estimated phylogenetic distances (in millions of years).

known_and_myspp_cleannames.txt (open in Excel) Corrected names and plant families (after checking in www.theplantlist.org) for the submitted plant species from both lists. You should review this list to make sure that there were no incorrect substitutes for your names. In such a case, consult The Plant List for the currently accepted name for your species, and then resubmit your corrected species lists.

known_and_myspp_phylotaxa_out_dated.png (open in an image viewer like Preview) A phylogenetic tree of the known hosts and species of concern combined, including estimated phylogenetic distances (in millions of years). Symbols indicate which list each species comes from (or if it is on both lists).

known_and_myspp_phylotaxa_out_dated.nolabel.png (open in an image viewer like preview) A phylogenetic tree of the known hosts and species of concern combined. including estimated phylogenetic distanced (in millions of years). Symbols indicated which list each species come from (or if it is on both lists). Same as the previous figure but family and node labels are removed for clarity.

known_and_myspp_phylotaxa_out_dated.new (open in text editor or a Newick viewer) Dated Newick file of the phylogenetic tree for known hosts and species of concern combined. Can view with any of a number of online Newick viewers such as Newick Viewer (http://www.trex.uqam.ca)or using R (must first install and load the *ape* library) with this code: plot(read.tree(file.choose())); axisPhylo()

How to use PhyloSuscept (phytosanitary risk version, cisr-stats.cisr.ucsc.edu/aphis/)

The basic use of the phytosanitary risk version is the same as that for the two-file version, above, except the user provides only a list of known hosts. The list of phytosanitary species of concern is a fixed list that includes 116 important crop species in the National Agriculture Survey Statistics (NASS), the 447 tree species included in the Forest Inventory Analysis Species (FIAS), and 806 plant species listed under the US Federally Endangered Species Act (FESA).

1. Create a comma-delimited (.csv) file containing genus and species names for known hosts of the pest of interest. Do not include headers in the files, just genus and species. You can include infraspecific ranks, but they will be treated as a species within that genus, not as a true infraspecific taxon. A genus with sp. or spp. is treated the same as one with a recognized species name. As an example, you could create the file *known.txt* for known hosts of your pest:

Genus	species	infraspecific

Amaranthus	*retroflexus*	
Atriplex	*laciniata*	
Atriplex	*tatarica*	
Beta	*vulgaris*	*vulgaris*
Chenopodium	*album*	
Chenopodium	*capitatum*	
Salsola	*kali*	
Suaeda	sp.	
Polygonum	*aviculare*	

2. Select your known-hosts file by clicking the Browse buttons for "known species list", click on the “APHIS” button for Concerned species list, and then click Upload and Execute.

3. When analysis is complete (this may take several minutes for extensive lists of species), click where it says "Download a zip file with the output from the analyses HERE" to download a compressed bundle of output files. Double click on the resulting *deployr_directory_archive.zip* file to expand to a folder called "output". The output folder contains seven files, each with custom names built from the name of the known-host file you submitted (here *known.txt*):

knownhosts_to_APHISconcern_README.txt (open in a text editor) This is a documentation of the analysis and overview of the results. It includes descriptions of the number of taxa included, name changes, a brief description of included files, and a table of taxa from the APHIS phytosanitary species of concern with "elevated" probability of being susceptible to a pest with the known host range. Here "elevated" means probability index > 0.5. However, this should be treated as a general overview; consult the outputprobs.csv file for pest-group specific estimates for all species of concern.

known_to_APHISconcern_outputprobs.csv (open in Excel) Rows are the phytosanitary species of concern. Columns are the major groups of plant pests. The values are the calculated probabilities that a pest with the given host range will be able to attack each of the species of concern. These are relative values that require user judgement to evaluate. Values close to one (1) are highly likely to be susceptible; as a general rule, taxa with values below 0.25 are at background levels with a low probability of susceptibility.

known_to_APHISconcern_distancematrix.csv (open in Excel) Phylogenetic distances from (in millions of years of independent evolution) between your species of concern (in rows) and each known host (in columns). These are the values used to calculate the outputprobs.csv.

known_tree.png (open in an image viewer like Preview or Windows Photo Gallery) A phylogenetic tree of the known host species provided in your list, including estimated phylogenetic distances (in millions of years).

known_and_APHISconcern_cleannames.txt (open in Excel) Corrected names and plant families (after checking in www.theplantlist.org) for the submitted plant species from both lists. You should review this list to make sure that there were no incorrect substitutes for your names. In such a case, consult The Plant List for the currently accepted name for your species, and then resubmit your corrected species lists.

known_and_APHISconcern_phylotaxa_out_dated_fan.png (open in an image viewer like Preview) A fan-shaped phylogenetic tree of the known hosts and APHIS phytosanitary species

of concern combined, including estimated phylogenetic distances (in millions of years). Symbols indicate which list each species comes from (or if it is on both lists).

known_and_APHISconcern_phylotaxa_out_dated.new.txt (open in text editor or a Newick viewer) Dated Newick file of the phylogenetic tree for known hosts and species of concern combined. Can view with any of a number of online viewers such as Newick Viewer (http://www.trex.uqam.ca) or using R (must first install and load the *ape* library) with this code: plot(read.tree(file.choose())); axisPhylo().

How PhyloSuscept works

PhyloSuscept has four components.

1. *Find current names and family placement for submitted taxa.* The two species lists are submitted to www.theplantlist.org using the *taxize* function *plantminer* in R. *Plantminer* submits each *Genus species* to the plant list lookup, and it if is an accepted name returns the current family placement. If the submitted name is a synonym of a currently accepted taxon it substitutes the current name into the list. If there is no match for the name, but is a close spelling variant of an accepted *Genus* or *species*, it will substitute in the apparently correct name. It is crucial that the user inspect the output *cleannames.txt* file for any name changes in the analysis and correct the files as needed for resubmission. The Plant List is an excellent, authoritative repository of currently accepted names - it provide the names that will be used in the analysis. Note that if a plant is submitted as *Genus* sp. or *Genus* spp. or if the genus name is accepted but not the specific epithet, it is converted to *Genus* sp and handled as any other species within that genus for analysis. Note that The Plant List (and thus the underlying tree used here) retain the conserved names of Compositae for Asteraceae and Leguminosae for Fabaceae.

2. *Create a phylogenetic tree for the combined species list.* The two species lists are joined (first removing any duplicates) into a single combined species list. PhyloSuscept then calls to the *phylomatic* function in *phylocom* v 4.2 (available at http://phylodiversity.net) to create an ultrametric phylogenetic tree (in newick format) based on the topology of a hand-constructed supertree R2G2_20140601.new (Parker et al., in press). The R2G2 tree includes all families of angiosperms, gymnosperms, ferns, and lycophytes listed as accepted in The Plant List in April 2014. The Angiosperm tree is based on the latest APGIII tree, and the other groups on consensus phylogenetic topologies as given by the Missouri Botanic Garden and Tree of Life projects. In addition, major, diverse families were grafted onto the tree with resolution to the genus level for all recognized genera in the families. These include Pinaceae, Fabaceae (Leguminosae), Asteraceae (Compositae), Poaceae, Araceae, Arecaceae, Ericaceae. Every node in the R2G2 tree was dated using the *bladj* function in *phylocom,* interpolating from a list of ages of major nodes, based on the Wikstrom ages provided in phylocom and expanded using fossil dates given on the Missouri Botanic Garden web site. *Bladj* interpolates ages for all other nodes in the supertree. All nodes in the R2G2 tree were named, and the ages associated with all nodes in the tree were saved. This all-node ages file was then used to date the extracted phylogenetic tree for the submitted species lists. The phylogenetic tree for the submitted species, with dated nodes, is provided in the output as a Newick file that can be viewed in any of a number of online or standalone newick viewers, such as the trex Newick Viewer (http://www.trex.uqam.ca). Images of phylogenetic trees for the known hosts, species of concern, and the combined lists are given as .png files for convenience.

3. *Calculate the phylogenetic distance between every combination of known host and species of concern.* The *cophenetic* function from *Picante* in *R* was used to calculate the phylogenetic distance between each pair of species in the combined list. The phylogenetic distances are in millions of year (My) and represent time of independent evolution (twice the time to most recent common ancestor). This full distance matrix (allspp x allspp) is provided in the output, as well as a subset matrix with species of concern in rows and known host species in columns.

4. *Calculate a probability index that each species of concern would be susceptible to a pest of that known host range.* Gilbert et al. 2012 established logistic regression coefficients for the probability that two host species of a given phylogenetic distance are likely to share a pest. These were updated with the new R2G2 tree in Parker et al. *in press.* Coefficients were determined separately for fungi, bacteria, oomycetes, insects, mites, nematodes, viruses, and plant parasites. For a given species of concern, PhyloSuscept calculates the probability of sharing a pest with each of the known hosts independently, based on the paired phylogenetic distances. The probability of *not* sharing a pest would then be one minus the probability of sharing [p(notS) = 1-p(S)]. The product of all those pairwise p(notS) would then be the probability on not sharing a pest with any of the hosts, and so one minus that product is an index of the likelihood that the species of concern would share a pest with that known host list. This index is calculated separately using the coefficients for each of the eight groups of pests.

References

Gilbert, G.S., R. Magarey, K. Suiter, and C.O. Webb. 2012. Evolutionary tools for phytosanitary risk analysis: phylogenetic signal as a predictor of host range of plant pests and pathogens. Evolutionary Applications doi:10.1111/j.1752-4571.2012.00265.x.

Parker, Ingrid M., Megan Saunders, Megan Bontrager, Andrew P. Weitz, Rebecca Hendricks, Roger Magarey, Karl Suiter, Gregory S. Gilbert. *In press.* Phylogenetic structure and host abundance drive disease pressure in communities. *Nature.*