`PhyloNetworks` Documentation

Version 0.0.1

Claudia Solís-Lemus, Cécile Ané and John Spaw

September 25, 2015

1 Introduction

`PhyloNetworks` is a `Julia` package with several functions for phylogenetic networks, among which we can highlight reading from and writing in parenthetical format, re-rooting and plotting. However, the main function in `PhyloNetworks` is the implementation of the method `SNaQ` (Solís-Lemus and Ané, 2015): a statistical method to infer a phylogenetic network from input gene trees.

It is important to notice that a phylogenetic network has several conditions on the nodes and edges.

For the present work, we will use the following definition (but refer to Huson et al. (2010) for other types of evolutionary networks). A *rooted phylogenetic network* for a set of taxa X is a connected directed acyclic graph with vertices $V = V_L \cup V_H \cup V_T$, edges $E = E_H \cup E_T$ and a bijective leaf-labeling function $f : V_L \to X$ with the following characteristics:

- The root r has indegree$(r) = 0$ and outdegree$(r) = 2$.
- For any $v \in V_L$ (leaf), $indegree(v) = 1$ and $outdegree(v) = 0$.
- For any $v \in V_T$ (tree node), $indegree(v) = 1$ and $outdegree(v) = 2$.
- For any $v \in V_H$ (hybrid node), $indegree(v) = 2$ and $outdegree(v) = 1$.
- A tree edge $e \in E_T$ is an edge whose child is a tree node.
- A hybrid edge $e \in E_H$ is an edge whose child is a hybrid node.

Thus, we are not allowing internal nodes with only two edges, nor polytomies. We also do not allow a leaf to be hybrid node, and only 2 hybrid edges per hybrid node. These restrictions are enforced in the optimization, but not in the read/write/plot/root functions. However, the only rule strictly enforced for all functions is that a hybrid node cannot have more than two hybrid edges pointing at it. This is a restriction that we hope to eliminate in the future.

If we also assume that the hybridization cycles do not intersect, then these networks are called *level-1 networks* (Huson et al., 2010), and have been shown to be identifiable (Pardi and Scornavacca, 2015; Solís-Lemus and Ané, 2015).

2 Setup

2.1 Installation of `Julia`

To install `Julia` version 0.3.* (if Julia is already installed, skip this step): go to `http://julialang.org/downloads/`. `PhyloNetworks` was developed under `Julia` version 0.3.5, and has been tested on different versions of 0.3.X. We have not tested its robustness on `Julia` version 0.4 or above.

2.2 Installation of `PhyloNetworks`

To install the package, open `Julia` and type:

```
Pkg.clone("https://github.com/crsl4/PhyloNetworks.git")
Pkg.build("PhyloNetworks")
```

The `PhyloNetworks` package has the following dependencies, but everything is installed when `PhyloNetworks` is added.

- `GraphViz` (version 0.0.3)
- `NLopt` (version 0.2.0)

The version in parenthesis correspond to the ones used when implementing `PhyloNetworks`. To test that you installed correctly the `PhyloNetworks` package, try the following example:

2.2.1 Example to test correct `PhyloNetworks` installation

Open `Julia` and type

```
Pkg.test("PhyloNetworks")
```

This will take a couple of minutes as the package needs to compile all the functions. If the installation was successful, you will see a message at the end: **Tests passed**. Otherwise, an error will be thrown.

3 Usage of PhyloNetworks

Before each session, need to type in `Julia`:

```
using PhyloNetworks
```

This will take a couple of minutes as it needs to precompile the functions. The first run of a function in `Julia` will also compile, so it will be slower than any subsequent runs.

4 Update of PhyloNetworks

It is important to regularly update the version of the Julia packages:

```
Pkg.update()
```

This will take a couple of minutes as it needs to precompile the functions in every package. This is particulary important for the package `PhyloNetworks` since it is a new package under constant development.

5 Description of main functions

Functions in `PhyloNetworks`:

- **readTopology:** Function to read a tree or network from parenthetical format. Input can be a string or a name of a text file (text file should contain only one line with the tree and end in ;, no quotes. To read more than one tree, see `readInputTrees` funcion). This function returns a `HybridNetwork` object. This function allows for many topologies that can be forbidden with the restriction described above. That is, this function allows for polytomies in tree nodes (tree node with more than two children) and internal nodes with only two edges.
  Usage:

  ```
  net=readTopology(filename);
  net=readTopology(string);
  ```

  WARNING: it is preferable to end the command with `;` to avoid printing to screen some inner details on the network object.

- **readTopologyLevel1:** Same as `readTopology`, but this function enforces the restrictions on the network. `readTopology` should be used when you want more variety of topologies, but these topologies cannot be used directly as starting point in the `SNaQ` optimization method (which is taken care automatically inside the optimization function). This function also returns a `HybridNetwork` object.
  Usage:

  ```
  net=readTopologyLevel1(filename);
  net=readTopologyLevel1(string);
  ```

WARNING: it is preferable to end the command with ; to avoid printing to screen some inner details on the network object.

- **tipLabels:** Prints list of taxon names in the `HybridNetwork` object.
  Usage:

```
tipLabels(net)
```

- **writeTopology:** Function to write the parenthetical format of a `HybridNetwork` object. It has four options:
  - *di=true*: to print in a format that Dendroscope can read, that is, without the γ values (default false)
  - *outgroup=taxon name*: to root by the outgroup before printing (default none). The outgroup has to be a single taxon. We will update to allow a clade in the future.
  - *names=true*: to print the taxon names in the leaves as opposed to the node numbers (default true).

  Usage:

```
writeTopology(net)
writeTopology(net,di=true)
```

- **root:** Function to root a network at a node or outgroup (single taxon). When a node number is used as parameter, you have the *resolve* option. If true, a branch with zero length is added. The node numbers can be known with the plot function described in section 7.
  Usage:

```
root!(net,nodeNumber,resolve)
root!(net,outgroup)
```

- **deleteLeaf:** Function to delete a leaf from a `HybridNetwork` object.
  Usage:

```
deleteLeaf!(net,taxonName)
```

- **printEdges:** Function to print out the information on all the edges in a `HybridNetwork` object. Not all information is useful for the user, but information like *edge length* and *gamma* are printed.
  Usage:

```
printEdges(net)
```

- **printNodes:** Function to print out the information on all the nodes in a `HybridNetwork` object. Not all information is useful for the user, but information like *node number* is printed.
  Usage:

```
printNodes(net)
```

6 SNaQ: estimation of phylogenetic network

6.1 Read your data

`SNaQ` estimates a phylogenetic network with the statistical methodology described in (Solís-Lemus and Ané, 2015). There are two possible input data:

- List of estimated gene trees: Estimated gene trees can be obtained from sequence data using RAxML (Stamatakis, 2014) or MrBayes (Huelsenbeck and Ronquist, 2001). Trees need to be in parenthetical format. Other formats will be available in future versions of the package.
- Table of estimated concordance factors (CF): Estimated CF can be obtained from estimated gene trees with BUCKy Ané et al. (2007).

The methodology also requires a starting topology for the search. This can be read from parenthetical format.
Functions to read data into `Julia`:

- **readTrees2CF:** function that read a text file with a list of trees in parenthetical format (one tree per line, but it can have extra lines like headline that the function will ignore. It will assume that any line starting with "(" is a tree). The function will calculate the observed CF for the corresponding 4-taxon subsets. There are two possibilities for the 4-taxon subsets: one, the list of desired 4-taxon subsets to analyze can be given in a text file (*quartetfile name*), and two, if no such file is specified, then the user can set a number of 4-taxon subsets (*numQ*) to be chosen randomly (*whichQ=:rand*) from the set of all possible 4-taxon subsets: $\binom{n}{x}$ or else all the quartets will be considered. *numQ* needs to be smaller than the total number of quartets. This function will also write by default a text file with a table of CF called "tableCF.txt", but the name can be specified by *filename* and if you do not want to write such table, set *writetab=false*. If the quartet file is not specified, then by default it will find quartets from the union of all taxa in the tree file, but you can choose the set of taxa for the quartets with the argument *taxa*. The function does not need all the arguments, see below examples for details.

  WARNING: This function has not yet been tested with missing data. That is, it has been tested in examples where all the trees have the same taxa.
  Usage:

```
readTrees2CF(treefile, quartetfile, whichQ, numQ,
             writetab, tablename)
readTrees2CF(treefile, whichQ, numQ, taxa, writetab,
             tablename)
```

- **readTableCF:** Function to read a table of observed CF. The table should contain 7 columns as in Table 1.

Taxon1	Taxon2	Taxon3	Taxon4	$CF12\|34$	$CF13\|24$	$CF14\|23$

Table 1: Example of column format for the table of CF to be used as input data

Usage:

```
readTableCF(filename);
```

Both functions `readTrees2CF` and `readTableCF` return a data structure called `DataCF` that contains the following attributes:

  - quartet: Array with the 4-taxon subsets either read from a table of CF or chosen to be analyzed from a list of trees.
  - numQuartets: Number of 4-taxon subsets
  - tree: Array of `HybridNetwork` objects that represent the list of estimated trees read. If the input data was not a list of trees, this attribute will be empty.
  - numTrees: Number of trees read.

- **readInputTrees:** read a text file with a list of trees in parenthetical format (one per line, ignores any line that does not start on "(") and returns a vector of trees.
Usage:

```
readInputTrees(filename)
```

- **summarizeCFdata:** takes as input a data structure `DataCF` (from previous functions) and provides a few descriptive information. By default, it prints the information on the screen, but it can be saved to a file. The option *pc* is used only when CF were computed from a collection of gene trees: CF for 4-taxon subsets that were computed with less than *pc* percentage of the gene trees will be listed. The option *pc* should be a number between 0 and 1.
Usage:

```
summarizeCFdata(d)
summarizeCFdata(d,filename)
summarizeCFdata(d,filename,pc)
```

- **readStartTop:** read a tree in parenthetical format from a text file and updates its branch lenghts with the CF data on the data structure `DataCF`. It is equivalent to run readTopologyUpdate first and then updateBL!(net,d)
  WARNING: updateBL only works for a tree topology, not proven for a network yet.
  Usage:

  ```
  readStartTop(treefile, d)
  ```

6.1.1 Small example on reading data

All the files for this section can be downloaded from the *examples* folder in github repository. You need to have them in your working directory. Suppose you have a file with a list of gene trees called *treefile.txt* and you want to use all the possible 4-taxon subsets for the taxa in those trees to calculate the CF:

```
d1=readTrees2CF("treefile.txt")
```

If you want to use a random sample of 100 4-taxon subsets:

```
d2=readTrees2CF("treefile.txt",whichQ=:rand,numQ=100)
```

On the contrary, if you have already the CF in a file *tableCF.txt* in the format on Table 1, you would read it like:

```
d=readTableCF("tableCF.txt")
```

If you have a tree *startTree.tre* in parenthetical format to use as starting point for the optimization and want to update the branch lengths according to the CF already read in the data structure *d*:

```
T=readStartTop("startTree.tre",d);
```

6.2 Estimation method

The function `snaq` runs the estimation method described in (Solís-Lemus and Ané, 2015) and it needs two parameters: starting topology and a data object `DataCF`:

```
snaq(startingTopology,data)
```

The function has the following options (with default values in parenthesis):

- **Nfail** *(100)*: number of proposal failed allowed before stopping the optimization
- **hmax** *(1)*: maximum number of hybridizations allowed

- **ftolRel** *(1e-5)*: relative tolerance on the function for the numerical optimization
- **ftolAbs** *(1e-6)*: absolute tolerance on the function for the numerical optimization
- **xtolRel** *(1e-3)*: relative tolerance on the parameters for the numerical optimization
- **xtolAbs** *(1e-4)*: absolute tolerance on the parameters for the numerical optimization
- **verbose** *(false)*: if true, it prints the details of the numerical optimization
- **runs** *(10)*: number of independent starting points. The first two runs start with the starting topology, and subsequent runs modify it by an NNI move. The addition of the initial hybridizations (if the starting topology has fewer hybridizations than *hmax* are done at random and the seeds are saved)
- **outgroup** *(none)*: name of outgroup taxon to root the estimated network at the end
- **rootname** *(snaq)*: root name for the output files: .log, .out, .err
- **returnNet** *(true)*: if true, the `snaq` function returns the resulting `HybridNetwork` object. If false, the resulting network is only written in the .out file.
- **seed** *(0)*: seed to replicate the analyses. This is the main seed from which one seed per run will be drawn randomly. To replicate the results for all the runs, simply set the same seed. If you want to replicate the results of a given run, set *runs=1* and as seed the seed reported in the log file for the given run. With default, the clock time is used to define the main seed.
- **probST** *(0.3)*: probability to use the starting topology as the starting point of each run. To improve the optimization, it is important that each run starts in a different place. At the beginning of every run, a biased coin is thrown so that with probability *probST*, the starting topology is not changed and with probability *1-probST*, an NNI move is performed on the starting topology. If on top the starting topology is a network, a second biased coin is flipped with the same probability so that with probabiliy *1-probST* we propose half the times move origin or and half the times move target.

WARNING: the method does not currently have a way to control for the complexity of the network, so it is important to try to avoid overparametrizing the model by selecting a very big `hmax` from start. It is best to start with `hmax=1`, and increase arithmetically trying to keep the estimated network interpretable. Keep in mind also that the fact that we estimate *level 1 networks* means that the hybridization cycles cannot overlap. If the number of taxa is small, and `hmax` is set very big, then the method will not be able to place that many hybridizations.

6.2.1 Small example on estimating phylogenetic network

You have d, T from previous example (6.1.1), you can estimate the phylogenetic network by:

```
net=snaq(T,d);
net=snaq(T,d,hmax=2);
```

These runs might take a few minutes. The estimated network will be stored in `net`, so that the user that write in parenthetical format, re root or plot:

```
PhyloNetworks.plotPhylonet(net)
```

More on the plot function below. There will also be output files created with information on the estimation procedure.

6.3 Read .out file

- **readSnaqNetwork:** reads the .out file generated by the `snaq` function and returns a `HybridNetwork` object.
  Usage:

  ```
  readSnaqNetwork(outfile)
  ```

The .log file contains information on the heuristic optimization such as the seed for each run, the reason why the optimization stopped (there are different criteria, mostly either there have been too many failed proposals or the likelihood is not changing anymore), the number of iterations needed for the algorithm to converge, the value of the likelihood for the estimated network, and the number of moves proposed and accepted.

6.4 Debugging: the .err file

`SNaQ` is a complex computational algorithm, so despite our best efforts, there are probably many undetected bugs and errors. The user can be extremely helpful in fixing this. After you do any analysis, please check the .err file to check how many runs failed because of a bug:

```
Total errors: 1 in seeds [4545]
```

The seed that caused the error and the description of the error (which will not necessarily be informative for the user) will be listed in the .log file. To help us out to debug `SNaQ`, please use the same settings under which you found the error to run the function and the seed in the .err file associated to the bug:

```
const DEBUG = true
const REDIRECT = true
snaqDebug(T,d,hmax=2,seed=4545);
```

This will create two text files: *snaqDebug.log* and *debug.log*.

You can send them to *claudia@stat.wisc.edu* with subject *Snaq bug found* or something similar. I will not have access to any of your data, the files simply print the steps I need to retrace the bug and hopefully, fix it.

Errors with the other functions are not straight-forward to detect, but if you encounter a problem with them, simply type:

```
const DEBUG = true
```

and re run the function saving the output to a file and send me that file.

7 Phylogenetic Network Visualization

This package contains functionality for creating visual plots of phylogenies generated using the `SNaQ` estimation method (Solís-Lemus and Ané, 2015). The included visualization tools allow the user to plot entire networks with many hybridization events or the underlying tree structure. Plots created by this package include a dynamic representation of probability of inheritance for hybridization events, which are represented by variation in hybrid edge thickness. Numerous customization options are available and are described in detail below.

7.1 Basic Network Plotting

Plotting with the `PhyloNetworks` package is simple, being entirely contained in the function `plotPhylonet`. Although there are many optional arguments available for this function, the only *required* input is the network itself. Networks may be input into `plotPhylonet` as one of two possible formats, depending on the user's particular preference:

1. Newick parenthetical format
2. HybridNetwork data type

Calling this function on a network will generate a .svg image file in the user's working directory as well as the corresponding .dot file used for rendering.

The .svg file type can be opened and viewed using a number of different programs including most web browsers (Inkscape, Chrome, Safari, etc.). If the user wishes to use a different image type, there are a number of options available (two of which we will describe here). First, the user can access one the many conversion tools that are available on the web. Many of these websites can convert a .svg image into a wide variety of standard image file types. Another option is to open the corresponding .dot file using GraphViz (which should already be installed) and exporting into the desired format.

7.2 Customization options

The `plotPhylonet` function contains a variety of optional arguments that may be used to tailor the output image to a particular use. A complete list of optional arguments is given below along with default values and argument descriptions.

- **mainTree** *(false)*: When true, only the underlying tree structure is plotted as determined by gammaThreshold. Otherwise, the entire network is shown.
- **imageName** *(netImage)*: Name for the output plot.
- **gammaThreshold** *(0.5)*: Set's the lowest gamma value to be included when plotting the underlying tree structure.
- **width** *(6.0)*: Sets the width of the image in inches.
- **height** *(8.0)*: Sets the height of the image in inches.
- **vert** *(true)*: When true, the hierarchy of the plot is directed vertically with the root node being place on top and the leaf nodes on bottom. Otherwise, the hierarchy is directed horizontally with the root on the left and leaf nodes on the right.
- **internalLabels** *(false)*: When true, node labels are included on all internal nodes. Otherwise, they are only included for leaf nodes.
- **fontSize** *(16.0)*: Sets the font size for node and edge labels in points.
- **layoutStyle** *("dot")*: Chooses the layout engine used by GraphViz for determining node and edge placement (more details can be found at `http://www.graphviz.org/Home.php`. Alternative options include "neato", "fdp", "sfdp", "circo", and "twopi".
- **hybridColor** *("green4")*: Sets the color for hybrid edges. Complete list of color options can be found at `http://www.graphviz.org/doc/info/colors.html`
- **forcedLeaf** *(true)*: When true, leaf nodes are placed on the same level, ranked at the bottom of the network.
- **unrooted** *(false)*: Plots an unrooted network or tree using the *neato* engine.
- **nodeSeparation** *(0.8)*: Sets the minimum distance between any two nodes in inches.
- **edgeStyle** *("line")*: Chooses the edge style used in the plot. Additional options include "ortho", "curved", "composite", "spline", and "false".
- **labelAngle** *(180.0)*: Sets the angle of leaf label placement relative to its parent edge.
- **labelDistance** *(3.0)*: Sets the distance of leaf label placement relative to its corresponding node.

- **includeGamma** *(false)*: When true, gamma labels are included on hybrid edges.
- **IncludeLength** *(false)*: When true, length labels are included on all edges.

7.3 Visualization Examples

This section will provide explicit examples for using the visualization tools provided in the PhyloNetworks package. Plotting examples will use $(((((1,2))\#H1,(\#H1,(3,4))),5),6)$; as an example network.

Example 1: As previously mentioned, the most simple way to use the included visualization tools is to call the `plotPhylonet()` function with the network as an argument. Using our example network, this would be done by typing in `Julia`

```
PhyloNetworks.plotPhylonet("(((((1,2))#H1,(#H1,(3,4))),5),6);")
```

This will result in the image below being saved in your working directory as `netImage.svg`. The file name can be pre-specified when calling `plotPhylonet` by including the argument `imageName="newfilename"`, which saves the plot as `newfilename.svg`.

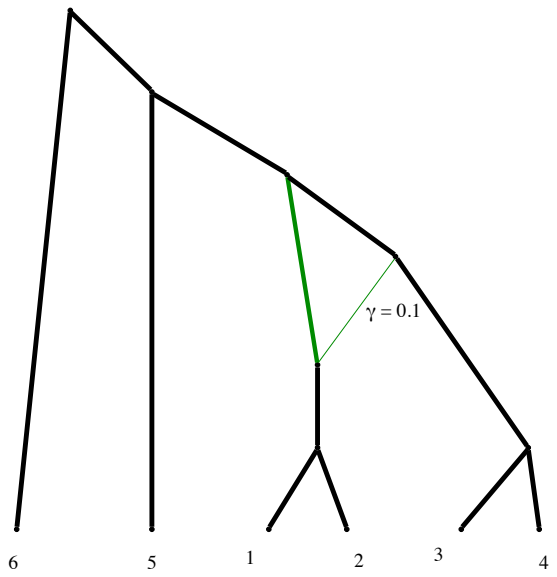


Figure 1: Basic plot using the `plotPhylonet` function. Note that if the gamma value for a hybrid edge is not explicitly defined, it will assume a default value of 0.1.

Calling additional customization options is as simple as including additional arguments separated by a comma. The user can combine any number of the available arguments to tailor

a plot to their particular example. Given below are a few examples that exhibit the different visualization options available.

Example 2: For the sake of tidiness, we define the network as its own variable, `net`. In addition, we include the arguments `vert = false`, which plots the hierarchy horizontally, `mainTree=true`, which only plots the underlying tree structure, and `fontSize = 20.0`, which increases the label font size from 16.0 to 18.0.

```
net = "(((((1,2))#H1,(#H1,(3,4))),5),6);"
PhyloNetworks.plotPhylonet(net, vert = false,
                           mainTree = true, fontSize = 20.0)
```

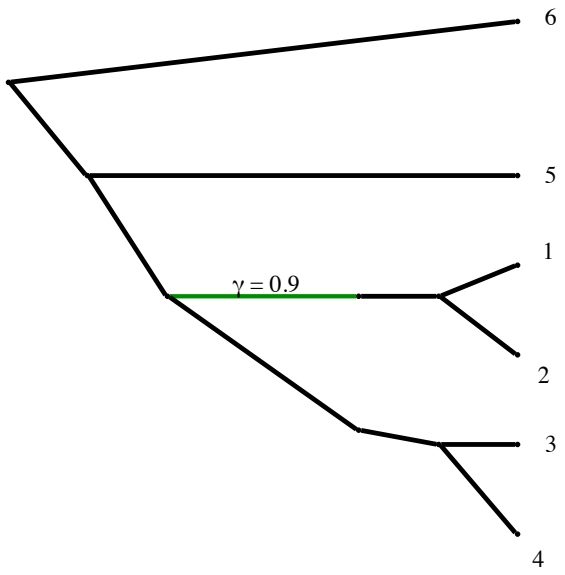


Figure 2: Plot of the underlying tree structure, oriented horizontally, with a changed font size.

7.4 Style Notes

Although there are default argument values given by `plotPhylonet`, they do not always result in the ideal plot for a particular example. Many of the included arguments were included for the purpose of the user being able to adjust certain layout parameters to best fit their own network. In particular, there is sometimes difficulty in neatly placing and orienting leaf labels and gamma labels. This is especially noticeable as the number of leaf

nodes becomes large or if the the names associated with leaf nodes are long. We have included some tips for fixing common issues below.

- To avoid edges overlapping gamma labels, include the argument `edgeStyle = true`. This will allow the layout engine to include curved splines, which will avoid overlaps.
- If leaf names are long, include the `vert = false` argument to set horizontal hierarchy.
- Label overlap can also be finely altered by changing the `labelDistance` and `labelAngle` arguments.
- Issues between in text readability can be fixed by adjusting the `fontSize`, `height`, or `width`.
- Although the arguments `layoutStyle` and `edgeStyle` have been included, some of options available are not guaranteed to be ideal for certain network plots.

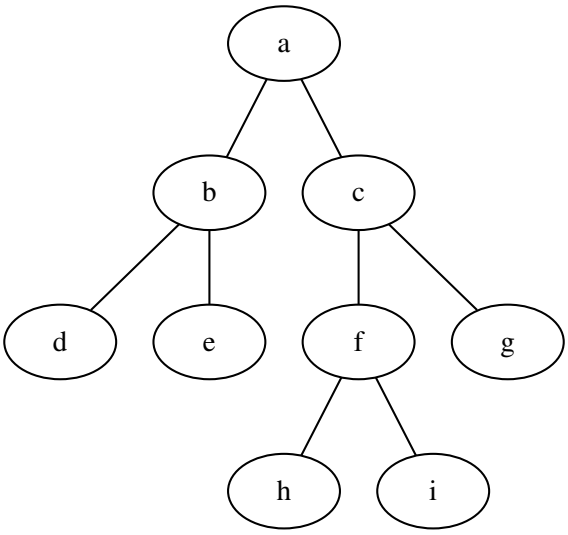
When plotting, a known issue can occur if Julia and GraphViz are not linked. To test the GraphViz installation, see the following section:

7.4.1 Example to test correct `GraphViz` installation

The `GraphViz` package installs the program `GraphViz`. To verify that the link between `Julia` and `GraphViz` is working properly. Everything should be done automatically, but it is worth testing. Open `Julia` and type

```
s=open("graph.dot","w")
write(s,"graph {
                  a -- b;
                  a -- c;
                  b -- d;
                  b -- e;
                  c -- f;
                  c -- g;
         f -- h;
         f -- i;
         }")
close(s)
PhyloNetworks.generalExport("graph.dot")
```

This will turn `graph.dot` into a file called `genImage.svg` in the working directory.
If this worked correctly, the console should display a series of prompts indicating its completion. A file called `scratchimage.svg` should be located in your working directory.



Plot of the underlying tree structure, oriented horizontally, with a changed font size.

WARNING: There is a known bug in the `plotPhylonet` function, see the issue in the `PhyloNetworks` Github repository for details. The error can be sometimes fixed by changing the position of the root with the `root` function.

References

Ané C, Larget B, Baum DA, Smith SD, Rokas A. 2007. Bayesian estimation of concordance among gene trees. *Molecular biology and evolution*. 24:412–26.

Huelsenbeck JP, Ronquist F. 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics*. 17:754–755.

Huson D, Rupp R, Scornavacca C. 2010. Phylogenetic Networks. New York, NY: Cambridge University Press, first edition.

Pardi F, Scornavacca C. 2015. Reconstructible Phylogenetic Networks: Do Not Distinguish the Indistinguishable. *PLOS Computational Biology*. 11:e1004135.

Solís-Lemus C, Ané C. 2015. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *arXiv*. pp. 1–32.

Stamatakis A. 2014. {RAxML} version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*. 30:1312–1313.