Phylogenetic Tree HOWTO

Jason Stajich, Dept Molecular Genetics and Microbiology, Duke University
<jason AT bioperl.org>

2003-12-01

Revision History

Revision 0.1	2003-12-01	JES
	First version	
Revision 0.2	2004-11-05	BIO
	Add SVG section and links	
Revision 0.3	2005-07-11	JES
	Explore Node objects more	

This HOWTO intends to show how to use the Bioperl Tree objects to manipulate phylogenetic trees. It shows how to read and write trees, query them for information about specific nodes or overall statistics, and create pictures of trees. Advanced topics include discussion of generating random trees and extensions of the basic structure for integration with other modules in Bioperl.

Table of Contents

1. Introduction

Generating and manipulating phylogenetic trees is an important part of modern systematics and molecular evolution research. The construction of trees is the subject of a rich literature and active research. This HOWTO and the modules described within are focused on querying and manipulating trees once they have been created.

The data we intend to capture with these objects concerns the notion of Trees and their Nodes. A Tree is made up of Nodes and the relationships which connect these nodes. The basic representation of parent and child nodes is intended to represent the directionality of evolution. This is to capture the idea that some ancestral species gave rise, through speciation events, to a number of child species. The data in the trees need not be a strictly bifurcating tree (or binary trees to the CS types), and a parent node can give rise to 1 or many child nodes.

In practice there are just a few main objects, or modules, you need to know about. There is the main Tree object `Bio::Tree::Tree` which is the main entry point to the data represented by a tree. A Node is represented generically by `Bio::Tree::Node`, however there are subclasses of this object to handle particular cases where we need a richer object (see `Bio::PopGen::Simulations::Coalescent` for more information). The connections between Nodes are described using a few simple concepts. There is the concept of pointers or references where a particular Node keeps track of who its parent is and who its children are. A Node can only have 1 parent and it can have 1 or many children. In fact all of the information in a tree pertaining to the relationships between

Nodes and specific data, like bootstrap values and labels, are all stored in the Node objects while the `Bio::Tree::Tree` object is just a container for some summary information about the tree and a description of the tree's root node.

2. Reading and Writing Trees

Trees are used to represent the ancestry of a collection of taxa, sequences, or populations.

Using `Bio::TreeIO` one can read trees from files or datastreams and create `Bio::Tree::Tree` objects. This is analagous to how we read sequences from sequence files with `Bio::SeqIO` to create Bioperl sequence objects which can be queried and manipulated. Similarly we can write `Bio::Tree::Tree` objects out to string representations like the Newick or New Hampshire format which can be printed to a file, a datastream, stored in database, etc.

The main module for reading and writing trees is the `Bio::TreeIO` factory module which has several driver modules which plug into it. These drivers include `Bio::TreeIO::newick` for New Hampshire or Newick format, `Bio::TreeIO::nhx` for the New Hampshire eXtended format from Sean Eddy and Christian Zmeck as part of their RIO and ATV system [reference here]. The driver `Bio::TreeIO::nexus` supports parsing tree data from PAUP's Nexus format. However this driver currently only supports parsing, not writing, of Nexus format tree files. There are also modules for lintree and Page1 output formats.

3. Example Code

Here is some code which will read in a Tree from a file called "tree.tre" and produce a Bio::Tree::Tree object which is stored in the variable

```
$tree
```

.

Like most modules which do input/output you can also specify the argument -fh in place of -file to provide a glob or filehandle in place of the filename.

```
use Bio::TreeIO;
# parse in newick/new hampshire format
my $input = new Bio::TreeIO(-file   => "tree.tre",
                            -format => "newick");
my $tree = $input->next_tree;
```

Once you have a Tree object you can do a number of things with it. These are all methods required in `Bio::Tree::TreeI`.

4. Bio::Tree::TreeI methods

Request the taxa (leaves of the tree).

```
my @taxa = $tree->get_leaf_nodes;
```

Get the root node.

```
my $root = $tree->get_root_node;
```

Get the total length of the tree (sum of all the branch lengths), which is only useful if the nodes actually have the branch length stored, of course.

```
my $total_length = $tree->total_branch_length;
```

5. Bio::Tree::TreeFunctionsI

An additional interface was written which implements utility functions which are useful for manipulating a Tree.

Find a particular node, either by name or by some other field that is stored in a Node. The field type should be the function name we can call on all of the Nodes in the Tree.

```
# find all the nodes named 'node1' (there should be only one)
my @nodes = $tree->find_node(-id => 'node1');
# find all the nodes which have description 'BMP'
my @nodes = $tree->find_node(-description => 'BMP');
# find all the nodes with bootstrap value of 70
my @nodes = $tree->find_node(-bootstrap => 70);
```

If you would like to do more sophisticated searches, like "find all the nodes with bootstrap values better than 70", you can easily implement this yourself.

```
  my @nodes = grep { $_->bootstrap > 70 } $tree->get_nodes;
```

Remove a Node from the Tree and update the children/ancestor links where the Node is an intervening one.

```
# provide the node object to remove from the Tree
$tree->remove_Node($node);
# or specify the node Name to remove
$tree->remove_Node('Node12');
```

Get the lowest common ancestor for a set of Nodes. This method is used to find an internal Node of the Tree which can be traced, through its children, to the requested set of Nodes. It is used in the calculations of monophyly and paraphyly and in determining the distance between two nodes.

```
# Provide a list of Nodes that are in this tree
# This only works when @nodes is 2 sequences
my $lca = $tree->get_lca(-nodes => \@nodes);
```

In order to get the LCA for multiple nodes we have to iterate until they converge on a single node.

```
use strict;
use Bio::TreeIO;
my $tree = Bio::TreeIO->new(-format => 'newick', -fh => \*DATA)->next_tree;
my @nodes = grep { $_->id =~ /c|d|f/ } $tree->get_nodes;
my @orig = @nodes;
while( @nodes > 1 ) {
  my $lca = $tree->get_lca(-nodes => [shift @nodes, shift @nodes]);
  push @nodes, $lca;
}
my $lca = shift @nodes;
```

```
print "lca is ",$lca->id, " for ", join(",",map { $_->id } @orig), "\n";

@nodes = grep { $_->id =~ /a|z/ } $tree->get_nodes;
@orig = @nodes;
while( @nodes > 1 ) {
  my $lca = $tree->get_lca(-nodes => [shift @nodes, shift @nodes]);
  push @nodes, $lca;
 }
 $lca = shift @nodes;
 print "lca is ",$lca->id, " for ", join(",",map { $_->id } @orig), "\n";

 __DATA__
 (a,((c,d)z,(e,f)y)x)root;
```

Get the distance between two nodes by adding up the branch lengths of all the connecting edges between two nodes.

```
my $distances = $tree->distance(-nodes => [$node1,$node2]);
```

Perform a test of monophyly for a set of nodes and a given outgroup node. This means the common ancestor for the members of the internal_nodes group is more recent than the common ancestor that any of them share with the outgroup node.

```
if( $tree->is_monophyletic(-nodes    => \@internal_nodes,
                           -outgroup => $outgroup) ) {
  print "these nodes are monophyletic: ",
       join(",",map { $_->id } @internal_nodes ), "\n";
}
```

Perform a test of paraphyly for a set of nodes and a given outgroup node. This means that a common ancestor 'A' for the members of the ingroup is more recent than a common ancestor 'B' that they share with the outgroup node *and* that there are no other nodes in the tree which have 'A' as a common ancestor before 'B'.

```
if( $tree->is_paraphyletic(-nodes    => \@internal_nodes,
                           -outgroup => $outgroup) > 0 ) {
  print "these nodes are monophyletic: ",
       join(",",map { $_->id } @internal_nodes ), "\n";
}
```

Re-root a tree, specifying a different node as the root (and a different node as the outgroup).

```
# node can either be a Leaf node in which case it becomes the
# outgroup and its ancestor is the new root of the tree
# or it can be an internal node which will become the new
# root of the Tree
$tree->reroot($node);
```

6. Making Images of Trees

You can also make images of trees. If you have the module SVG::Graph installed you can create an SVG image of your tree. The example below uses TreeIO to get a Tree object and then its tree is written to an image file.

```
use Bio::TreeIO;
my $in = new Bio::TreeIO(-file => 'input',
                         -format => 'newick');
my $out = new Bio::TreeIO(-file => '>mytree.svg',
                          -format => 'svggraph');

while( my $tree = $in->next_tree ) {
    $out->write_tree($tree);
}
```

Alternatively you could use an output format of "tabtree", this option will create an ASCII drawing of the tree.

7. Constructing Trees

Pairwise distances for all sequences in an alignment can be computed with `Bio::Align::DNAStatistics` and and `Bio::Align::ProteinStatistics`. There are several different methods implemented. For DNA alignments, Jukes-Cantor (1969), Juke-Cantor uncorrected, Kimura 2-parameter (1980), Felsenstein (1981), Tajima-Nei (1984), and Tamura (1992) are currently implemented. In addition, for coding sequences, synonymous and non-synonymous counts can be computed with the `calc_KaKs_pair`. For Protein sequences alignments only Kimura (1983) is currently supported although other methods will be added.

To use these methods simply initialize a statistics module, and pass in an alignment object (`Bio::SimpleAlign`) and the type of distance method to use and the module will return a `Bio::Matrix::PhylipDist` matrix object of pairwise distances. The code example below shows how this could be done.

Given the matrix of pairwise distances one can build a phylogenetic tree using 2 simple methods provided in the `Bio::Tree::DistanceFactory`. Simple request either Neighbor-Joining (NJ) trees or Unweighted Pair Group Method with Arithmatic Mean (UPGMA) clusters. There are caveats with these methods and whether or not the distances are additive. The method `check_additivity` in `Bio::Tree::DistanceFactory` is provided to calculate whether or not additivity holds for the data.

The following is a basic code snippet which describes how to use the pairwise distance and tree building modules in Bioperl.

```
use Bio::AlignIO;
use Bio::Align::DNAStatistics;
use Bio::Tree::DistanceFactory;
# for a dna alignment
# can also use ProteinStatistics
my $aln = Bio::AlignIO->new(-file => 'filename', -format=>'clustalw');
my $stats = Bio::Align::DNAStatistics->new;
my $mat = $stats->distance(-method => 'Kimura',
      -align  => $aln);
my $dfactory = Bio::Tree::DistanceFactory->new(-method => 'NJ');
my $tree = $dfactory->make_tree($mat);
```

TODO: Using external programs: phylip,MrBayes,paup,puzzle,protml

Non-parametric bootstrapping is one method to test the consistency of the data with the optimal tree. A set of subreplicates are generated from the alignment using the method from `Bio::Align::Utilities` called `bootstrap_replicates`. One passes in an alignment object and the count of the number of replicates to generate.

```
use Bio::Align::Utilities qw(:all);
my $replicates = bootstrap_replicates($aln,$count);
```

8. Advanced Topics

It is possible to generate random tree topologies with a Bioperl object called `Bio::Tree::RandomFactory`. The factory only requires the specification of the total number of taxa in order to simulate a history. One can request different methods for generating the random phylogeny. At present, however, only the simple Yule backward is implemented and is the default.

The trees can be generated with the following code. You can either specify the names of taxa or just a count of total number of taxa in the simulation.

```
use Bio::TreeIO;
use Bio::Tree::RandomFactory;
# initialize a TreeIO writer to output the trees as we create them
my $out = Bio::TreeIO->new(-format => 'newick',
                           -file   => ">randomtrees.tre");
my @listoftaxa = qw(A B C D E F G H);
my $factory = new Bio::Tree::RandomFactory(-taxa => \@listoftaxa);
# generate 10 random trees
for( my $i = 0; $i < 10; $i++ ) {
   $out->write_tree($factory->next_tree);
}
# One can also just request a total number of taxa (8 here) and
# not provide labels for them
# In addition one can specify the total number of trees
# the object should return so we can call this in a while
# loop
$factory = new Bio::Tree::RandomFactory(-num_taxa => 8
                                        -max_count=> 10);
while( my $tree = $factory->next_tree) {
  $out->write_tree($tree);
}
```

There are more sophisticated operations that you may wish to pursue with these objects. We have tried to create a framework for this type of data, but by no means should this be looked at as the final product. If you have a particular statistic or function that applies to trees that you would like to see included in the toolkit we encourage you to send details to the Bioperl list, bioperl-l@bioperl.org.

9. References and More Reading

For more reading and some references for the techniques above see these titles.

J. Felsenstein, "Infering Phylogenies" 2003. Sinuar and Associates.
D. Swoffrod, Olsen, Waddell and D. Hillis, "Phylogenetic Inference" 1996. in Mol. Systematics, 2nd ed, 1996, Ch 11.

Eddy SR, Durbin R, Krogh A, Mitchison G, "Biological Sequence Analysis" 1998. Cambridge Univ Press, Cambridge, UK.

10. Additional Information

Here's a list of the relevant modules. If you have questions or comments that aren't addressed herein then write the Bioperl community at bioperl-l@bioperl.org.

Related Modules

Bio/TreeIO.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO.html]
Bio/Tree/Tree.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/Tree.html]
Bio/Align/DNAStatistics.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Align/DNAStatistics.html]
Bio/Align/ProteinStatistics.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Align/ProteinStatistics.html]
Bio/Align/Utilities.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Align/Utilities.html]
Bio/Matrix/PhylipDist.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Matrix/PhylipDist.html]
Bio/PopGen/Simulation/Coalescent.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/PopGen/Simulation/Coalescent.html]
Bio/SimpleAlign.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/SimpleAlign.html]
Bio/Tree/DistanceFactory.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/DistanceFactory.html]
Bio/Tree/Node.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/Node.html]
Bio/Tree/RandomFactory.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/RandomFactory.html]
Bio/Tree/TreeI.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/TreeI.html]
Bio/Tree/AlleleNode.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/AlleleNode.html]
Bio/Tree/NodeI.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/NodeI.html]
Bio/Tree/NodeNHX.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/NodeNHX.html]
Bio/Tree/TreeFunctionsI.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/TreeFunctionsI.html]
Bio/Tree/Statistics.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/Tree/Statistics.html]
Bio/TreeIO/newick.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/newick.html]
Bio/TreeIO/nexus.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/nexus.html]
Bio/TreeIO/nhx.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/nhx.html]
Bio/TreeIO/pag.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/pag.html]
Bio/TreeIO/svggraph.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/svggraph.html]
Bio/TreeIO/lintree.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/lintree.html]
Bio/TreeIO/tabtree.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/tabtree.html]
Bio/TreeIO/TreeEventBuilder.pm [http://doc.bioperl.org/releases/bioperl-1.4/Bio/TreeIO/TreeEventBuilder.html]