Rooted versus Unrooted

Phylogenetic trees are either rooted or unrooted, depending on the research questions being addressed. The root of the phylogenetic tree is inferred to be the oldest point in the tree and corresponds to the theoretical last common ancestor of all taxonomic units included in the tree. The root gives directionality to evolution within the tree (Baldauf, 2003). Accurate rooting of a phylogenetic tree is important for directionality of evolution and increases the power of interpreting genetic changes between sequences (Pearson et al., 2013).

Many techniques such as molecular clock, Bayesian molecular clock, outgroup rooting, or midpoint rooting methods tend to estimate the root of a tree using data and assumptions (Boykin et al., 2010). However, Steel (2012) discusses root location in random trees and points out that information in the prior distribution of the topology alone can convey the location of the root of the tree. These results show that the tree models that treat all taxa equally and are sampling consistently convey information about the location of the ancestral root in unrooted trees (Steel, 2012).

Why Do We Need a Rooted Tree?

We are interested in rooting a phylogenetic tree in order to show the path of evolution of biological species. Therefore most users of phylogenetic trees want rooted trees because they give an indication of the directionality of evolutionary change. The root of phylogenetic tree is crucial in evolutionary interpretation of the tree (Williams, 2014), because an unrooted tree species shows only the relationships among the taxa and does not define the evolutionary path (Figure 1(a)).

When Do We Need an Unrooted Tree?

An unrooted tree is desired when we do not have a distantly related group (sequence) for comparison or when primary interest is focused only on relationships among the taxa rather than on the directionality of evolutionary change. Unrooted trees are beneficial in depicting clusters of related sequences. Unrooted gene trees have also become more prevalent within the multispecies coalescent phylogenetic framework, leading...
to systematic approaches for inferring unrooted species trees from unrooted gene tree topologies (Liu and Yu, 2011). Gene trees and species trees can have similar topologies but often there is considerable discordance between gene tree and species trees (Degnan, 2013). For this reason, understanding how to root gene trees will have implications for accurate species tree inference (Figures 1(b) and 1(c)).

### How Do You Root a Phylogenetic Tree?

#### Outgroup Rooting

There are several rooting methods (Table 1), however the most popular and widely used is the outgroup method (Wheeler, 1990; Tarrío et al., 2000; Hess and De Moraes Russo, 2007; Boykin et al., 2010). Outgroup method assumes that one or more of the taxa are divergent from the rest of the taxa (ingroup). The branch linking the ingroup and outgroup becomes the starting point, and defines all subsequent evolutionary events within the tree (Brady et al., 2011; Williams, 2014). In addition to providing evolutionary information of the ingroup, the outgroup has other additional functions. It allows the identification of distinct features within ingroup sequences (Wheeler, 1990). An important aspect of outgroup method is the need for a priori knowledge on the appropriate outgroup to use for the set of sequences (Wheeler, 1990; Hess and De Moraes Russo, 2007). However, this is also the main bottleneck for this method, especially within higher taxonomic groups such as angiosperms, birds, and mammals where a consensus outgroup is lacking (Qiu et al., 2001). As a consequence, many authors are forced to choose between different sorts of outgroups that are either phylogenetically close or phylogenetically distant (Rota-Stabelli and Telford, 2008).

Lack of an appropriate outgroup results in drawbacks such as the long branch attraction (LBA). LBA occurs mainly when the outgroup taxa are distantly related to the ingroup due to either large divergence time and/or increased rate of evolution (Tarrio et al., 2000). This results in homoplastic changes occurring at rapidly evolving sites thus resulting in artificial rooting (random rooting) (Wheeler, 1990; Hendy and Penny, 2011; Maddison et al., 1984). Several criteria have been proposed to prevent LBA within phylogenetic trees, through a multistep process as proposed by Rota-Stabelli and Telford (2008) to assist in outgroup selection especially in the case of arthropod classes. They include: (1) low substitution rate; (2) ingroup like G + C composition; (3) new strand bias estimators ‘skew index’; (4) the tendency of the outgroup to avoid ‘random branding effect’; and (5) phylogenetic proximity to the arthropod.

An alternative approach to assess the importance of an outgroup in rooting the tree is explored by Graham et al. (2002); this is by establishing whether the outgroup provides sufficient signal in response to root location, indicative of historic linkage or due to LBA. Using Pontederiaceae, an aquatic monocot, as the case study they assessed how the nearest outgroup provides for rooting Pontederiaceae compared to those less closely related relatives and further investigate the role of LBA when determining the optimal rooting of Pontederiaceae. However, they concluded that LBA may influence rooting, and may be supporting the wrong outgroup. To further reduce LBA and to ensure robustness of the outgroup rooting method they recommend multiple sampling of outgroups within the sister group rather than sampling within less closely related taxa.
Midpoint Rooting

Midpoint rooting calculates tip to tip distances and then places the root halfway between the two longest tips (Swofford et al., 1996). The ancestral point will be identified if the tree has constant rates of evolution. The method is exclusively dependent on the branch length of the phylogenetic tree and the assumption of the molecular clock (Holland et al., 2003). Homogeneity is assumed across the branch and that the two most divergent taxa evolved at equal rates (Holland et al., 2003; Swofford et al., 1996). If the tree is balanced, midpoint rooting works well. However, a major limitation of midpoint rooting is the dependence on having clocklike data and a balanced topology.

The midpoint rooting method is often applied to viral genetic datasets because in many cases outgroups are unknown. For example, Stavrinides and Guttman (2004) utilize midpoint rooting to establish the evolutionary relationship of severe acute respiratory syndrome (SARS) coronaviruses. They carried out phylogenetic analysis of the viral genes encoding for viral structural proteins specifically, envelope matrix (M) and nucleocapsid (N) proteins. The midpoint rooting of the
trees generated using the M protein data shows two groups, one that consisted of porcine, feline, and canine and one that contained bovine and murine coronaviruses (Coronaviridae). On the other hand, the N protein tree is midpoint rooted on the branch leading to the group 1 coronaviruses. Moreover, the appropriateness of midpoint rooting is supported by the results of Tajima’s relative rate test (Tajima, 1993), indicating no rate heterogeneity among the coronavirus groups (Stavrinides and Gutman, 2004).

This method should be used as an alternative to the outgroup rooting method and could be adopted as the default method when the outgroup method is difficult to apply either due to problems with available outgroups, such as LBA or lack of a priori knowledge of the outgroup (Hess and De Moraes Russo, 2007).

### Molecular Clock Rooting

The molecular clock rooting method has one assumption: the rate of evolution is constant for the sequences of interest (Yang and Rannala, 2012). The rate is typically expressed in substitutions per site per year or substitutions per site per million years (Brown and Yang, 2011). The strict clock is often used in analyses of sequences sampled at the intraspecific level, for which usually there is an exceptionally low rate of variation (Brown and Yang, 2011; Ho and Duchène, 2014). The molecular clock assumption becomes problematic for distantly related species because there is a linear relationship between the genetic distances and approximate divergence. The slope of the line directly corresponds to the evolutionary rate variation among species especially among divergent taxa (Welch and Bromham, 2005).

Before utilizing the molecular clock method for rooting a phylogenetic tree users should test if a molecular clock is appropriate to describe the data. Testing for the molecular clock entails generating two maximum likelihood trees, one computed with the molecular clock enforced and one without the molecular clock enforced and then utilizing the likelihood ratio test (Felsenstein, 1983; Holder and Lewis, 2003).

### Bayesian Molecular Clock Rooting

Huelsenbeck et al. (2002) proposed the use of Bayesian inference under the molecular clock assumption to infer the root of a phylogenetic tree. After obtaining the posterior distribution of trees under Bayesian inference, the root of the tree is inferred to be the root position with the highest posterior probability. This method also provides the posterior probability that the root lies on any branch of the ingroup topology. Another advantage of the Bayesian method is that it allows the user to evaluate alternative rootings. Other rooting methods only return one rooting for a particular dataset, without any numerical assessment of confidence in that rooting. A Bayesian molecular clock analysis successfully identified the root of Orcuttieae (Poaceae) (Boykin et al., 2010) when all other methods failed. Post_root was developed to analyze the output from MrBayes (Ronquist et al., 2012) or ExaBayes (Aberer et al., 2014) runs. The output from Post_Root will give the number of unique roots and also the most probable root position.

Most recently, Calvignac-Spencer et al. (2014) have further developed Post_Root to a web-based interface in their quest to identify the branch root posterior probability (RPP) of the most recent Ebola outbreak in West Africa. They were forced to rely on Bayesian molecular clock rooting because there is no known outgroup for Ebola. It is often the case when analyzing viral sequences that no outgroup is known; therefore the Bayesian molecular clock rooting is a very useful alternative, especially when rooting is crucial for viral outbreaks.

### Acknowledgment

Research in the Boykin lab is provided by the Bill & Melinda Gates Foundation via a subcontract (B04265) from the National Research Institute, University of Greenwich, UK. An Australian PhD Award to J Wainaina. The Plant Biosecurity Cooperative Research Centre, Australia (61056) supported S Maina.
See also: Bayesian Phylogenetic Methods. Phylogenetic Tree

References


