Phylogeny

PhyKIT: a broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data

Jacob L. Steenwyk1,*, Thomas J. Buida III2, Abigail L. Labella1, Yuanning Li3, Xing-Xing Shen3, & Antonis Rokas1,*

1 Vanderbilt University, Department of Biological Sciences, VU Station B #35-1634, Nashville, TN 37235, United States of America
2 9 City Place #322, Nashville, TN 37209, United States of America
3 Ministry of Agriculture Key Lab of Molecular Biology of Crop Pathogens and Insects, Institute of Insect Sciences, Zhejiang University, Hangzhou 310058, China

*To whom correspondence should be addressed. jacob.steenwyk@vanderbilt.edu or antonis.rokas@vanderbilt.edu

Associate Editor: Russell Schwartz

Received on 06-Dec-2020; revised on 13-Jan-2021; accepted on 01-Feb-2021

Abstract

Motivation: Diverse disciplines in biology process and analyze multiple sequence alignments (MSAs) and phylogenetic trees to evaluate their information content, infer evolutionary events and processes, and predict gene function. However, automated processing of MSAs and trees remains a challenge due to the lack of a unified toolkit. To fill this gap, we introduce PhyKIT, a toolkit for the UNIX shell environment with 30 functions that process MSAs and trees, including but not limited to estimation of mutation rate, evaluation of sequence composition biases, calculation of the degree of violation of a molecular clock, and collapsing bipartitions (internal branches) with low support.

Results: To demonstrate the utility of PhyKIT, we detail three use cases: (1) summarizing information content in MSAs and phylogenetic trees for diagnosing potential biases in sequence or tree data; (2) evaluating gene-gene covariation of evolutionary rates to identify functional relationships, including novel ones, among genes; and (3) identify lack of resolution events or polytomies in phylogenetic trees, which are suggestive of rapid radiation events or lack of data. We anticipate PhyKIT will be useful for processing, examining, and deriving biological meaning from increasingly large phylogenomic datasets.


Contact: jacob.steenwyk@vanderbilt.edu or antonis.rokas@vanderbilt.edu

Supplementary information: Supplementary data are available on figshare (doi: 10.6084/m9.figshare.13118600) and are available at Bioinformatics online.

1 Introduction

Multiple sequence alignments (MSAs) and phylogenetic trees are widely used in numerous disciplines, including bioinformatics, evolutionary biology, molecular biology, and structural biology. As a result, the development of user-friendly software that enables biologists to process and analyze MSAs and phylogenetic trees is an active area of research (Kapli et al., 2020).

In recent years, numerous methods have proven useful for diagnosing potential biases and inferring biological events in genome-scale phylogenetic (or phylogenomic) datasets. For example, methods that evaluate sequence composition biases in MSAs (Phillips and Penny, 2003), signatures of clock-like evolution in phylogenetic trees (Liu et al., 2017), phylogenetic treeness (Lanyon, 1988; Phillips and Penny, 2003), taxa whose long branches may cause variation in their placement on phylogenetic trees (Struck, 2014), and others have assisted in summarizing the information content in phylogenomic datasets and improved phylogenetic inference (Salichos and Rokas, 2013; Walker et
functions help examine information content biases, gene-gene grouped into broad categories that assist in conducting analyses of MSAs. PhyKIT has 30 different functions that help process and analyze MSAs and facilitate the integration of additional functions in future versions. Detailed information about each one of PhyKIT’s functions and tutorials for using the software can be found in the online documentation (https://jlsteenwyk.com/PhyKIT). Here, we focus on three specific groups of functions implemented in PhyKIT that enable researchers to summarize information content in phylogenomic datasets, create gene-gene evolutionary rate covariation networks, and identifying polytomies in phylogenomic data.

**Evaluating information content and biases in phylogenomic datasets**

MSAs and phylogenetic trees are frequently examined to evaluate their information content and potential biases in characteristics such as sequence composition or branch lengths (Shen, Salichos, et al., 2016; Liu, et al., 2017; Phillips and Penny, 2003; Doyle et al., 2015; Smith et al., 2018; Philippe et al., 2011; Struck, 2014). PhyKIT implements numerous functions for doing so. We demonstrate the application of 14 functions:

1. **Alignment length.** The length of a multiple sequence alignment, which is associated with robust bipartition support and tree accuracy (Shen, Salichos, et al., 2016; Walker et al., 2019);
2. **Alignment length with no gaps.** The length of a multiple sequence alignment after excluding sites with gaps, which is associated with robust bipartition support and tree accuracy (Shen, Salichos, et al., 2016);
3. **Degree of violation of a molecular clock (DVMC).** A metric used to determine the clock-like evolution of a gene using the standard deviation of branch lengths for a single gene tree (Liu, et al., 2017). DVMC is calculated using the following formula:

\[ DVMC = \frac{1}{N-1} \sum_{j=1}^{N} (l_j - \bar{l})^2 \]

where \( N \) represents the number of tips in a phylogenetic tree, \( l_j \) being the distance between the root of the tree and species \( j \), and \( \bar{l} \) represents the average root to tip distance. DVMC can be used to identify genes with clock-like evolution for divergence time estimation (Liu, et al., 2017);
4. **Internal branch lengths.** Summary statistics of internal branch lengths in a phylogenetic tree are reported including mean, median, 25th percentile, 75th percentile, minimum, maximum, standard deviation, and variance values. Examination of internal branch lengths is useful in evaluating phylogenetic tree shape;
5. **Branch score.** A metric that examines the degree of taxon-specific long branch attraction (Struck, 2014; Weigert et al., 2014). Long branch scores of individual taxa are calculated using the following formula:

\[ LB_i = \left( \frac{PD_i}{PD_{all}} - 1 \right) \times 100 \]

where \( PD_i \) represents the average pairwise patristic distance of taxon \( i \) to all other taxa, \( PD_{all} \) represents the average patristic distance across all taxa, and \( LB_i \) represents the long branch score of taxon \( i \). Long branch scores can be used to evaluate heterogeneity in tip-to-root distances and identify taxa that may be susceptible to long branch attraction;
6. **Pairwise identity.** Pairwise identity is a crude approximation of the evolutionary rate of a gene and is calculated by determining the average number of sites in an MSA that are the same character state between all pairwise combinations of taxa. This can be used to group genes based on their evolutionary rates (e.g., faster-evolving genes vs. slower-evolving ones) (Chen et al., 2017);
Calculating gene-gene evolutionary rate covariation or coevolution

Patristic distances. Patristic distances refer to all distances between all pairwise combinations of tips in a phylogenetic tree (Fournet and Gibbs, 2006), which can be used to evaluate the rate of evolution in gene trees or taxon sampling density in species trees.

Parsimony-informative sites. Parsimony-informative sites are those sites in an MSA that have at least two character states (excluding gaps) that occur at least twice (Kumar et al., 2016); the number of parsimony-informative sites is associated with robust bipartition support and tree accuracy (Shen, Salichos, et al., 2016; Steenwyk et al., 2020).

Variable sites. Variable sites are those sites in an MSA that contain at least two different character states (excluding gaps) (Kumar et al., 2016); the number of variable sites is associated with robust bipartition support and tree accuracy (Shen, Salichos, et al., 2016).

Relative composition variability. Relative composition variability is the average variability in the sequence composition among taxa in an MSA. Relative composition variability is calculated using the following formula:

\[ \text{Relative composition variability} = \frac{1}{c} \sum_{i=1}^{c} \sum_{j=1}^{n} \frac{|c_{ij} - \bar{c}_i|}{s \times \pi} \]

where \( c \) is the number of different character states per sequence type, \( n \) is the number of taxa in an MSA, \( c_{ij} \) is the number of occurrences of the \( i \)th character state for the \( j \)th taxon, \( \bar{c}_i \) is the average number of the \( i \)th character state across \( n \) taxa, and \( s \) refers to the total number of sites (characters) in an MSA. Relative composition variability can be used to evaluate potential sequence composition biases in MSAs, which in turn violate assumptions of site composition homogeneity in standard models of sequence evolution (Phillips and Penny, 2003).

Saturation. Saturation refers to when an MSA contains many sites that have experienced multiple substitutions in individual taxa. Saturation is estimated from the slope of the regression line between patristic distances and pairwise identities. Saturated MSAs have reduced phylogenetic information and can result in issues of long branch attraction (Philippe et al., 2011; Lake, 1991).

Total tree length. Total tree length refers to the sum of internal and terminal branch lengths and is calculated using the following formula:

\[ \text{total tree length} = \sum_{i=1}^{a} l_i + \sum_{j=1}^{b} l_j \]

where \( l_i \) is the branch length of the \( i \)th branch of a internal branches and \( l_j \) is the branch length of the \( j \)th branch of \( b \) terminal branches. Total tree length measures the inferred total amount or rate of evolutionary change in a phylogenetic tree.

Treeness. Treeness (also referred to as stemminess) is a measure of the inferred relative amount or rate of evolutionary change that has taken place on internal branches of a phylogenetic tree (Lanyon, 1988; Phillips and Penny, 2003) and is calculated using the following formula:

\[ \text{treeness} = \sum_{u=1}^{v} l_u \]

where \( l_u \) is the branch length of the \( u \)th branch of \( b \) terminal branches, and \( \bar{l} \) refers to the total branch length of the phylogenetic tree. Treeness can be used to evaluate how much of the total tree length is observed among internal branches.

Treeness divided by relative composition variability. This function combines two metrics to measure both composition bias and other biases that may negatively influence phylogenetic inference. High treeness divided by relative composition variability values have been shown to be less susceptible to sequence composition biases and are associated with robust bipartition support and tree accuracy (Phillips and Penny, 2003; Shen, Salichos, et al., 2016).

Calculating gene-gene evolutionary rate covariation or coevolution

To measure gene-gene evolutionary rate covariation, PhyKIT implements the mirror tree method (Sato et al., 2005; Pazos and Valencia, 2001), which examines whether two trees have correlated branch lengths. Specifically, PhyKIT calculates the Pearson correlation coefficient between branch lengths in two phylogenetic trees that share the same tips and topology. To account for differences in taxon representation between the two trees, PhyKIT first automatically determines which taxa are shared and prunes one or both such that the same set of taxa is present in both trees. PhyKIT requires that the two input trees have the same topology, which is typically the species tree topology inferred from whole genome or proteome data. Thus, the user will typically first estimate a gene’s branch lengths by constraining the topology to match that of the species tree. When running this function, users should be aware that many biological factors, such as horizontal transfer (Doolittle and Baptiste, 2007), incomplete lineage sorting (Degnan and Salter, 2005), and introgression/hybridization (Sang and Zhong, 2000), can lead to gene histories that deviate from the species tree. In these cases, constraining a gene’s history to match that of a species may lead to errors in the covariation analysis.

Due to factors including time since speciation and mutation rate, correlations between uncorrected branch lengths result in a high frequency of false positive correlations (Sato et al., 2005; Clark et al., 2012; Chikina et al., 2016). To ameliorate the influence of these factors, PhyKIT first transforms branch lengths into relative rates. To do so, branch lengths are corrected by dividing the branch length in the gene tree by the corresponding branch length in the species tree. Previous work revealed that one or a few outlier branch length values can be responsible for false positive correlations and should be removed prior to analysis (Clark et al., 2012). Thus, PhyKIT removes outlier data points defined as having corrected branch lengths greater than five (i.e., removing gene tree branch lengths that are five or more times greater than their corresponding species tree branch lengths). Lastly, values are converted into relative rates using a Z-transformation. The resulting relative rates are used when calculating Pearson correlation coefficients.

Identifying polytomies in phylogenomic data

Rapid radiations or diversification events have occurred throughout the tree of life including among mammals, birds, plants, and fungi (Liu et al., 2017; Jarvis et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020). Polytomies correspond to internal branches whose length is 0 (or statistically indistinguishable from 0) and can be driven either by biological (e.g., rapid radiations) or analytical (e.g., low amount of data) factors. Thus, polytomies are useful for inferring rapid radiation or diversification events and exploring incongruence in phylogenies (One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020; Sayyari and Mirarab, 2018).

To identify polytomies, a modified approach to a previous strategy was implemented (Sayyari and Mirarab, 2018). More specifically, the support for three alternative topologies is calculated among all gene trees from a phylogenomic dataset. For example, in species tree ((A,B),C), D), if examining the presence of a polytomy at the ancestral bipartition of tips A, B, and C, PhyKIT will determine the number of gene trees that
support \(((A,B),C)\), \(((A,C),B)\), and \(((B,C),A)\); using the rooted gene trees provided by the user. Equal support for the three topologies (i.e., the presence of a polytomy) among a set of gene trees is assessed using a Chi-squared test. Failing to reject the null hypothesis is indicative of a polytomy (Sayyari and Mirarab, 2018). Note that this approach is distinct from the approach of Sayyari and Mirarab to identify polytomies because PhyKIT uses a gene-based signal rather than a quartet-based signal. The difference between the two methods is that each gene contributes equally to the inference of a polytomy when a gene-based signal is used, whereas genes with greater taxon representation (which contain a greater number of quartets) will contribute a greater signal during polytomy identification when a quartet-based signal is used. From a technical perspective, both approaches are simple to implement and require only a single line of code in the command-line.

3 Results

We outline three example uses of PhyKIT: 1) summarizing information content and identifying potential biases in animal, plant, yeast, and filamentous fungal phylogenomic datasets (One Thousand Plant Transcriptomes Initiative, 2019; Shen, Zhou, et al., 2016; Steenwyk et al., 2019; Laumer et al., 2019), 2) constructing a network of significant gene-gene covariation, which reveals genes of shared functions from empirical data spanning ~550 million years of evolution among fungi (Shen et al., 2020), and 3) illustrating how to identify polytomies using simulated and empirical data (Steenwyk et al., 2019).

3.1 Summarizing information content and biases in phylogenomic data

Examining information content in phylogenomic datasets can help diagnose potential biases that stem from low signal-to-noise ratios, multiple substitutions, non-clocklike evolution, and other biological or analytical factors. To demonstrate the utility of PhyKIT to summarize the information content in phylogenomic datasets, we calculated 14 different metrics known to help diagnose potential biases in phylogenomic datasets or be associated with accurate and well supported phylogenetic inferences (Smith et al., 2018; Doyle et al., 2015; Liu et al., 2017; Felsenstein, 1978; Philippe et al., 2011; Struck, 2014; Shen, Salichos, et al., 2016; Phillips and Penny, 2003) using four empirical phylogenomic datasets from animals (201 tips; 2,891 genes) (Laumer et al., 2019), budding yeast (332 taxa; 2,408 genes) (Shen et al., 2018), filamentous fungi (93 taxa; 1,668 genes) (Steenwyk et al., 2019), and plants (1,124 taxa; 403 genes) (One Thousand Plant Transcriptomes Initiative, 2019) (Figure 1, Supplementary table 1).

Examining the distributions of the values of the 14 different metrics revealed inter- and intra-dataset heterogeneity (Figure 1). For example, inter-dataset heterogeneity was observed among animal and plant datasets, which had the lowest and highest average pairwise identity across alignments, respectively; intra-dataset heterogeneity was observed in the uniform distribution of pairwise identities in the budding yeast datasets. Similarly, inter-dataset heterogeneity was observed in estimates of saturation where the budding yeast and filamentous fungal MSAs were less saturated by multiple substitutions than the plant and animal datasets; intra-data heterogeneity was also observed in all four datasets. Varying degrees of inter- and intra-dataset heterogeneity was observed for other information content statistics, which may be due biological (e.g., mutation rate) or analytical factors (e.g., taxon sampling, distinct alignment, trimming, and tree inference strategies).

In summary, PhyKIT is useful for examining the information content of phylogenomic datasets. For example, the generation of different phylogenomic data submatrices by selecting subsets of genes or taxa with certain properties (e.g., retention of genes with the highest numbers of parsimony-informative sites or following removal of taxa with high long branch scores) can facilitate the exploration of the robustness of species tree inference or estimating time since divergence (Walker et al., 2019; Shen et al., 2018; Steenwyk et al., 2019; Li et al., 2020; Salichos and Rokas, 2013; Liu et al., 2017; Shen et al., 2020).
3.2 A network of gene-gene covariation reveals neighborhoods of genes with shared function

Genes with similar evolutionary histories often have shared functions, are co-expressed, or are parts of the same multimeric complexes (Clark et al., 2012; Sato et al., 2005). Using PhyKIT, we examined gene-gene covariation using 815 genes spanning 1,107 genomes and ~563 million years of evolution among fungi (Shen et al., 2020). By examining 331,705 pairwise combinations of genes, we found 298 strong signatures of gene-gene covariation (defined as $r > 0.825$). The two genes with the strongest signatures of covariation were $SEC7$ and $TAO3$ ($r = 0.87$), suggesting that their protein products have similar or shared functions. Supporting this hypothesis, Sec7p contributes to cell-surface growth in the model yeast Saccharomyces cerevisiae (Novick and Schekman, 1979) and genes with the Sec7 domain are transcriptionally coregulated with yeast-hyphal switches in the human pathogen C. albicans (Song et al., 2008). Similarly, Tao3p in both S. cerevisiae and C. albicans is part of a RAM signaling network, which controls hyphal morphogenesis, polarized growth, and cell-cycle related processes including cell separation, cell proliferation, and phase transitions (Bogomolnaya et al., 2006; Song et al., 2008).

Complex relationships of gene-gene covariation can be visualized as a network (Figure 2). Examination of network neighborhoods identified groups of genes that have shared functions and are parts of the same multimeric complexes. For example, the proteins encoded by $NDC80$ and $NUF2$ are part of the same kinetochore-associated complex termed the NDC80 complex—which is required for efficient mitosis (Sundin et al., 2011)—and significantly covary with one another ($r = 0.84$). Similarly, multiple genes that encode proteins involved in DNA replication and repair (i.e., $POL2$, $MSH6$, $RAD26$, $CDC9$, and $EXO1$) were part of the same network neighborhood, consistent with previous work suggesting an intimate interplay between DNA replication and multiple DNA repair pathways (Boiteux and Jinks-Robertson, 2013; Lujan et al., 2012; Tsubouchi and Ogawa, 2000). Other network neighborhoods of genes with shared function such as ribosome biogenesis, Golgi apparatus-related transport, and control of DNA replication were identified (Figure 2).

![Fig. 2. Gene-gene covariation network inferred from ~550 million years of evolution across 1,107 fungi. A network of significant gene-gene coevolution identifies network neighborhoods representative of associated functional categories. For example, the NDC80 and NUF2 genes (toward the top right of the network) were identified to be significantly coevolving with one another ($r = 0.84, p < 0.01$, Pearson’s correlation test); they both encode proteins that are part of the same multimeric kinetochore-associated complex (green). Similarly, genes that are DNA replication factors (orange), contribute to DNA replication and repair processes (yellow), participate in Golgi apparatus-related transport (brown), or ribosome biogenesis (pink) were found to be neighbors in the network. Network visualization was done with the igraph package, v1.2.4.2 (Hunter and Cohen, 2007), in R, v3.6.2 ([https://www.r-project.org/](https://www.r-project.org/)).](https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btab096/6131675)

Taken together, these results indicate PhyKIT is a useful tool for evaluating gene-gene covariation and predicting genes’ functions (Brunette et al., 2019; Clark et al., 2012; Sato et al., 2005). Thus, we anticipate PhyKIT will be helpful for evaluating gene-gene covariation and conducting evolution-based screens for gene functions across the tree of life.

3.3 Identifying polytomies in phylogenomic datasets

Rapid radiations or diversification events have occurred throughout the tree of life (Liu et al., 2017; Jarvis et al., 2014; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020). One approach to identifying rapid radiations is by testing for the existence of polytomies in species trees (Sayyari and Mirarab, 2018; One Thousand Plant Transcriptomes Initiative, 2019; Li et al., 2020). Polytomies can also arise when the amount of data at hand is insufficient for resolution (Walsh et al., 1999). To demonstrate the utility of PhyKIT to identify polytomies, we examined the ability of our approach to identify a simulated polytomy (Figure 3A). PhyKIT was able to conservatively identify the simulated polytomy demonstrating the efficacy of our approach.
We next examined if there is evidence of polytomies in the evolutionary history of filamentous fungi from the genera Aspergillus and Penicillium. We examined three branches. The first two branches—one dating back ~110 million years ago (Figure 3B), and another dating back ~25 million years ago (Figure 3C)—were not polytomies. In contrast, examination of a ~60-million-year-old branch involving Lanata-divaricata, Citrina, and Exilicaulis (Figure 3D), which are major lineages (or sections) in the genus Penicillium, was consistent with a polytomy. Given the large number of gene trees used in our analysis (n=1,668), these results are consistent with a rapid radiation or diversification event in the history of Penicillium species.

In summary, these results suggest that PhyKIT is useful in identifying polytomies in simulated and empirical datasets. More broadly, these results support the notion that polytomies can be used to identify rapid radiation events. Beyond polytomy identification, PhyKIT can be used for exploring incongruence in phylogenies by calculating gene support frequencies. Calculations of gene-based support among different topologies can be used in diverse applications, including identifying putative introgression / hybridization events and conducting phylogenetically-based genome-wide association (PhyloGWAS) studies (Steenwyk et al., 2019; Pease et al., 2016).

4 Conclusion

We developed PhyKIT, a comprehensive toolkit for processing and analyzing MSAs and trees in phylogenomic datasets. Executing functions implemented in PhyKIT would otherwise require extensive programming, multiple software, and/or web-based applications (Supplemental table 1); thus, PhyKIT offers users a way to streamline approaches and pipelines by relying on only one software. PhyKIT is freely available on GitHub ([https://github.com/JLSteenwyk/PhyKIT](https://github.com/JLSteenwyk/PhyKIT)), PyPi ([https://pypi.org/project/phykit/](https://pypi.org/project/phykit/)), and the Anaconda Cloud ([https://anaconda.org/JLSteenwyk/phykit](https://anaconda.org/JLSteenwyk/phykit)) under the MIT license with extensive documentation and user tutorials ([https://jsteenwyk.com/PhyKIT](https://jsteenwyk.com/PhyKIT)). PhyKIT is a fast and flexible toolkit for the UNIX shell environment, which allows it to be easily integrated into bioinformatic pipelines. We anticipate PhyKIT will be of interest to biologists from diverse disciplines and with varying degrees of experience in analyzing MSAs and phylogenies. In particular, PhyKIT will likely be helpful in addressing one of the greatest challenges in biology, building, understanding, and deriving meaning from the tree of life.

Acknowledgements

We thank the Rokas lab for helpful discussion and feedback.

Funding

J.L.S. and A.R. were funded by the Howard Hughes Medical Institute through the James H. Gilliam Fellowships for Advanced Study program. X.X.S. was supported by National Natural Science Foundation of China (No. 32071665) and the Fundamental Research Funds for the Central Universities (No. 2020QNA6019).

Conflict of Interest: none declared.

References


Chikina, M. et al. (2016) Hundreds of Genes Experienced Convergent
PhyKIT: a toolkit for examining phylogenomic data

Suyama, M. et al. (2006) PAL2NAL: Robust conversion of protein
sequence alignments into the corresponding codon alignments.

*Nucleic Acids Res.*, **34**.


