Molecular Phylogenetics and Evolution 54 (2010) 306–308

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Short Communication

`rely.py`, a python script to detect reliable clades

Blaise Li, Agnès Dettai, Guillaume Lecointre *

Équipe 'Phylogénie', UMR 7138 'Systématique, Adaptation, Évolution', Département Systématique et Évolution, Muséum National d'Histoire Naturelle. 57, rue Cuvier, CP26, 75231 Paris cedex 05, France

ARTICLE INFO

Article history:
Received 21 April 2009
Revised 24 September 2009
Accepted 25 September 2009
Available online 2 October 2009

Keywords:
Clades
Missing data
Phylogeny
Python scripts
Reliability
Taxonomic congruence

ABSTRACT

`rely.py` is a program implementing the method to detect independently repeated clades by comparing phylogenies as described in Li and Lecointre (2009) and adapted to incompletely overlapping datasets in Li et al. (2009). The comparison can be performed on trees obtained by any inference method (maximum parsimony, Bayesian inference, maximum likelihood). The program computes repetition indices, provides greedy summary trees for each validity domain and a nexus matrix representation of the clades weighted by their repetition indices. The additional script `concatnexus.py` assists the user in preparing the primary analyses, but it can also be used separately to concatenate nexus datasets.

1. Introduction

Trees inferred from different datasets often partially contradict one another (Edwards, 2009). Analysis of the combination of several datasets in the same matrix is expected to make historical signal emerge above individual biases (Miyamoto and Fitch, 1995); however in some cases, a dataset with a strong biased signal can erroneously impose part of the topology of the tree (Brinkmann et al., 2005).

To evaluate the reliability of individual clades, separate analyses can be performed to distinguish between biased and historical signals, or else to evaluate what proportion of individual gene trees exhibit the same history (Edwards, 2009). Marker-specific biases have various effects, and should not repeatedly produce the same artefactual clades (Miyamoto and Fitch, 1995). Therefore, species clades shared by trees obtained from separate, independent, datasets should represent the sought-after historical signal. A procedure based on this property was proposed in Li and Lecointre (2009) to compute a repetition index to estimate the reliability of clades. Nonetheless, that method could not use incompletely overlapping datasets. This was improved in the version used in Li et al. (2009). The python script `rely.py` makes this improved method readily available.

2. Description

2.1. Counting independent occurrences

The datasets are first separated into minimal independent units: the 'elementary datasets'. In this context, 'independent' means 'unlikely to be subject to the same biases'.[1] Let us suppose we have 3 elementary datasets, noted *A*, *B* and *C*. The repetition index is primarily based on the number of occurrences of a clade across the analyses of independent datasets such as the elementary datasets.

Biases are more likely to occur in separate analyses than in a combination, because of the smaller dataset size. Dettai and Lecointre (2004) proposed the 'partial combination approach' to take advantage of both combined and separate analyses. It consists in analyzing all possible partially combined datasets and counting clades over sets of independent combinations. Some combinations, by increasing the size of the datasets, will hopefully allow historical signal to overcome some of the marker-specific biases. An example of partial combination is the combined dataset $A \cup B$.

The script `concatnexus.py` generates concatenated files from the elementary datasets, which are used to perform the 'primary analyses'. Their results are called 'source-trees' in a supertree perspective.

To compute an improved repetition index based on this partial combination approach, all possible combinations of the elementary

* Corresponding author. Fax: +33 1 40 79 38 44.
E-mail addresses: blaise.li@normalesup.org (B. Li), adettai@mnhn.fr (A. Dettai), lecointr@mnhn.fr (G. Lecointre).

[1] In particular, datasets sharing some data cannot be considered independent.

1055-7903/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2009.09.034

datasets are analyzed and arranged into 'partitioning schemes'. A partitioning scheme is a set of datasets (elementary datasets or combinations thereof) that do not share elementary datasets. Examples of partitioning schemes are (A, B, C) and $(A \cup B, C)$: being elementary datasets, A, B and C are independent from one another, $A \cup B$ is independent from C because the two datasets do not share elementary datasets.

Within a partitioning scheme, occurrences of a clade may be legitimately counted among partitions as the partitions are independent. This should avoid counting several occurrences for a clade produced by a shared bias. The counting is done within every possible partitioning scheme, to explore the emergence of historical signal systematically, while still having independent datasets to compare. This might prove computationally intensive with many datasets. We then suggest that the user select and justify the partial combinations to include.

The repetition index of a clade is based on the largest number of occurrences found across the investigated partitioning schemes. The partitioning scheme allowing the largest number of independent occurrences presumably represents the optimal way of combining some of the elementary datasets with respect to the particular piece of phylogenetic information that is the clade under focus.

2.2. Comparing contradicting clades

Except in cases of reticulate evolution, two clades that are incompatible (i.e. that contradict one another) should not be both considered reliable: by definition, incompatible clades cannot cooccur in a tree.

The repetition index is refined by taking into account contradiction between clades. The list of contradictors is established for each clade. Among the contradictors of a clade, the one with the largest repetition index is called the 'best contradictor'. Because of its higher repetition index, it deserves priority over others.

To reflect the uncertainty resulting from conflicting hypotheses, the repetition index of a clade is updated by subtracting the largest number of occurrences of its best contradictor from the largest number of occurrences of the clade itself. The update of the repetition index is made for all clades which may change the rankings of the contradictors. If the best contradictor of a clade has changed according to the updated repetition index, all indices are updated once again, taking into account the new best contradictors. This updating procedure is repeated until stabilization, leading to final repetition indices. In some cases the repetition indices vary periodically instead of stabilizing; the mean value over a period is then taken as the final repetition index.

2.3. Defining validity domains

All this supposes that all datasets have the same taxa (Li and Lecointre, 2009). If trees do not contain exactly the same taxa, the clades cannot be directly compared. We need to take into account the set of taxa on which a clade and its repetition index are defined ('validity domain').

A clade is an oriented bipartition defined by two items: the set of taxa that are in the 'internal' part of the bipartition and the set of taxa that are in the 'external' part of the bipartition (containing the root of the tree). The union of these two parts is the validity domain on which the clade is defined. A clade could as well be defined by its internal part and its validity domain, the external part being deduced by making the difference between the validity domain and the internal part. Two clades are the same if their two defining items are the same. If all datasets have the same taxa, comparing two clades amounts to comparing their internal parts, because their validity domains are the same. If the validity domains are not the same, one cannot compare the clades directly. In Li et al. (2009), we proposed to restrict the comparison between two clades to the set of taxa for which both clades are informative: the intersection of their validity domains. Thus, before comparing two clades, each clade is pruned by eliminating from it taxa that are not in the other's validity domain. This operation leads to three levels of validity domains (Fig. 1).

To each primary analysis (dataset, or source-tree obtained by analyzing this dataset) is associated a 'first-level validity domain': the set of taxa on which the source-tree was built.

If the independent datasets constituting a given partitioning scheme do not have exactly the same validity domain, the clades are counted on a reduced set of taxa that is the intersection of the first-level validity domains of the datasets. Each partitioning scheme is thus associated to a 'second-level validity domain': the taxonomic sampling common to the trees on which clades are counted. Before occurrences are counted, clades are pruned by

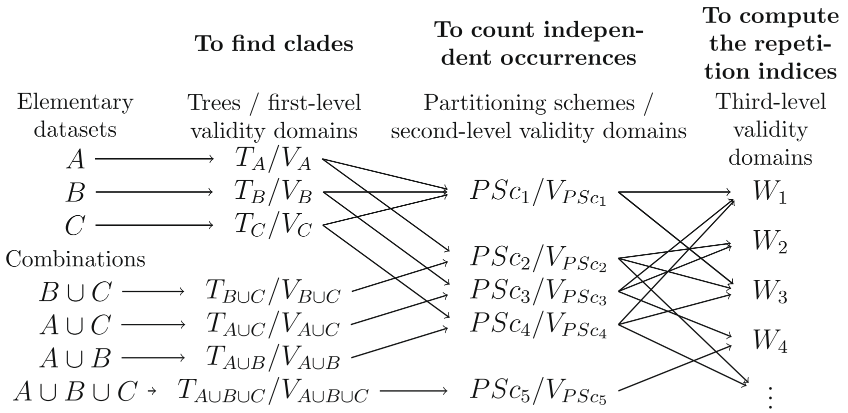
To find clades
To count independent occurrences
To compute the repetition indices
Elementary datasets
Trees / first-level validity domains
Partitioning schemes / second-level validity domains
Third-level validity domains
A
B
C
T_A/V_A
T_B/V_B
T_C/V_C
Combinations
$B \cup C$
$A \cup C$
$A \cup B$
$A \cup B \cup C$
$T_{B\cup C}/V_{B\cup C}$
$T_{A\cup C}/V_{A\cup C}$
$T_{A\cup B}/V_{A\cup B}$
$T_{A\cup B\cup C}/V_{A\cup B\cup C}$
PSc_1/V_{PSc_1}
PSc_2/V_{PSc_2}
PSc_3/V_{PSc_3}
PSc_4/V_{PSc_4}
PSc_5/V_{PSc_5}
W_1
W_2
W_3
W_4
⋮

Fig. 1. The three levels of validity domains. The first-level validity domains (V_X) are the sets of terminal taxa of the trees (T_X) obtained by the analyses of the datasets (X). In this example, 3 elementary datasets are used, which leads to 7 datasets, and thus to 7 trees and 7 first-level validity domains. The second-level validity domains (V_{PSc_i}) are the intersections of the validity domains of the independent datasets involved in the partitioning schemes (PSc_i). Only the full partitioning schemes are shown here. The occurrences of the clades are counted within a partitioning scheme across its constituting datasets, after pruning the corresponding trees by eliminating taxa outside the relevant second-level validity domain. The third-level validity domains (W_i) are the intersections of all possible combinations of second-level validity domains. The repetition indices are attached to such third-level validity domains. They are based on the maximum number of occurrences (for the clades once pruned by eliminating taxa outside the third-level validity domain) found among the partitioning schemes whose validity domains span at least the entire third-level validity domain. Only some of the possible third-level validity domains are shown here.

eliminating from them taxa not present in the second-level validity domain.

The possible contradictors of a clade may be collected from different partitioning schemes, and thus be defined on different second-level validity domains. Determining whether clades defined on different sets of taxa are compatible or contradictory may be tricky (see Bininda-Edmonds, 2003, p. 840). This comparison of clades is thus made with reduced counterparts of the clades, within 'third-level validity domains': the intersections of the second-level validity domains on which the clades are defined. Before establishing lists of contradictors, clades are pruned by eliminating from them taxa not present in the third-level validity domain.

These pruning steps lead to a loss of information about taxa that are present in only a few datasets. The more disparate the datasets in a partitioning scheme, the smaller its validity domain. Therefore, 'partial partitioning schemes' are also taken into account: partitioning schemes in which not all elementary datasets are represented.[2] For example, (A, B) is a partial partitioning scheme: C is not represented.

Similarly, the more disparate the partitioning schemes from which possible contradictors are collected, the smaller the resulting third-level validity domain. If one restricts the contradiction-and-update step to the taxa common to all second-level validity domains, many taxa are lost by reducing the study to the common sampling. To compute a repetition index for clades defined on wider sets of taxa, the contradiction-and-update step is made within every possible third-level validity domain: some involving clades coming from only one (or just a few) partitioning scheme(s) but with many taxa, and some involving clades collected from diverse partitioning schemes, but with only few taxa.

These procedures lead to a variety of partitioning schemes in which a certain number of clades are defined and associated with repetition indices.

2.4. Building a summary tree

To build a supertree combining all taxa, clades coming from all third-level validity domains are gathered in a same 'taxon × clade' matrix. Each clade is represented as a character weighted by its repetition index, and with three states: '0' if the taxon is in the external part of the corresponding bipartition, '1' if it is in the internal part, and '?' if it is not in the third-level validity domain on which the clade is defined. A parsimony analysis of this matrix, using the repetition indices as character weights, should produce a tree including mostly reliable relationships. An example of the use of this method can be seen in Li et al. (2009).

Within a third-level partitioning scheme of interest (such as the set of taxa common to all analyses), a greedy procedure can also be used to produce a tree including clades with a high repetition index first (see Li and Lecointre, 2009). On such a summary tree, repetition indices can be displayed on the branches. The output trees of `rely.py` are in nexus or treegraph (Müller and Müller, 2004) formats.

3. Conclusion

As the use of multiple markers to resolve complex phylogenetic problems becomes commonplace, the incongruence among datasets is more and more conspicuous and methods are being proposed to detect reliable clades. Among these, the repetition index of Li and Lecointre (2009) lacked a practical and documented implementation. The script `concatnexus.py` can be useful for those wanting to perform analyses of various data combinations. The script `rely.py` is more specialized and implements the reliability analysis described in the present paper. Its use is facilitated if the primary analyses have been prepared with `concatnexus.py`. Both scripts are available at http://www.normalesup.org/~bli/Programs/programs.html.

Acknowledgments

We thank L. Corbari for comments on a first version of this paper and F.-J. Lapointe and T. Silberfeld for testing the Python scripts and their documentation.

Fig. 1 was prepared using TikZ (http://sourceforge.net/projects/pgf/).

References

Bininda-Edmonds, O., 2003. Novel versus unsupported clades: assessing the qualitative support for clades in MRP supertrees. Systematic Biology 52 (6), 839–848.

Brinkmann, H., Van der Giezen, M., Zhou, Y., Poncelin de Raucourt, G., Philippe, H., 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Systematic Biology 54 (5), 743–757.

Dettai, A., Lecointre, G., 2004. In search of notothenioid (Teleostei) relatives. Antarctic Science 16 (1), 71–85. Available from: <http://dx.doi.org/10.1017/S095410200400183X>.

Edwards, S.V., 2009. Is a new and general theory of molecular systematics emerging? Evolution 63 (1), 1–19. Available from: <http://dx.doi.org/10.1111/j.1558-5646.2008.00549.x>.

Li, B., Dettai, A., Cruaud, C., Couloux, A., Desoutter-Meniger, M., Lecointre, G., 2009. RNF213, a new nuclear marker for acanthomorph phylogeny. Molecular Phylogenetics and Evolution 50, 345–363. Available from: <http://dx.doi.org/10.1016/j.ympev.2008.11.013>.

Li, B., Lecointre, G., 2009. Formalizing reliability in the taxonomic congruence approach. Zoologica Scripta 38 (1), 101–112. Available from: <http://dx.doi.org/10.1111/j.1463-6409.2008.00361.x>.

Miyamoto, M., Fitch, W., 1995. Testing species phylogenies and phylogenetic methods with congruence. Systematic Biology 44 (1), 64–75.

Müller, J., Müller, K., 2004. TreeGraph: automated drawing of complex tree figures using an extensible tree description format. Molecular Ecology Notes 4, 786–788.

[2] These are opposed to 'full partitioning schemes', where all elementary datasets are represented. In mathematical language, full partitioning schemes would be called 'partitions' of the set of the elementary datasets.