Crimson Publishers
Wings to the Research

Mini Review

SHERLOCK – an Automatized Analysis of Molecular Sequence Variation in Species Communities Using Statistical Tests on Patristic Tree Distances

Seidel NI, Geiger MF and Kück P*

Leibniz Institute for the Analysis of Biodiversity Change, Germany

ISSN: 2770-6745

***Corresponding author:** Patrick Kück, Leibniz Institute for the Analysis of Biodiversity Change, Adenauerallee 160, 53113 Bonn, Germany

Submission: May 10, 2022

Published: May 26, 2022

Volume 2 - Issue 4

How to cite this article: Seidel NI, Geiger MF, Kück P*. SHERLOCK – an Automatized Analysis of Molecular Sequence Variation in Species Communities Using Statistical Tests on Patristic Tree Distances. Biodiversity Online J. 2(4). BOJ. 000541. 2022. DOI: 10.31031/BOJ.2022.02.000541

Abstract

Phylogenetic trees are commonly used to gain information on organisms evolutionary relationships based on molecular sequences (e.g. genes, proteins, genomes). In a reconstructed tree, it can be assumed that the additive branch lengths from one sequence to another reflect the amount of evolutionary change between these two sequences. The sum of branch lengths that link two nodes in a tree can be used to calculate the overall so called phylogenetic diversity of a tree, i.e. the total evolutionary change inferred for a set of taxa. With Sherlock, we provide a simple and efficient tool to statistically analyse phylogenetic diversity in sequence data in comparison with a null-model distribution based on randomly drawn sequences of the original data set. SHERLOCK incorporates external alignment and tree reconstruction software, which allows for the first time a fully automatized analysis and visualization of patristic distances on the basis of raw sequence data.

Keywords: Phylogenetics; Phylogenetic diversity; Molecular sequence analysis; Automatized pipeline

Main Method

A phylogenetic tree represents a hypothesis on the evolutionary history and diversity of a set of taxa where branch lengths are estimates of the number of character changes that occurred for a certain branch. A patristic or Phylogenetic Distance (PD) is defined as the sum of the lengths of the branches that link two nodes in a tree. The overall PD of a tree summarizes the total evolutionary change inferred for the set of taxa. Comparing the observed overall PD to a null-model distribution or between trees obtained from different sets of taxa can provide the basis for serving a wide range of research fields: prioritization of conservation areas [1] or target taxa ('EGDE approach' [2]), species communities (β diversity, [3]) or trait variation [4]. Local communities often (according to theory) should consist of rather distantly related species in order to reduce competition between closely related species, whereas if relatives share similar environmental tolerances local communities should contain more closely related species [5]. The overall PD of an (ideally multigene) tree represents a proxy for the scale of phenotypic differences expected between any two species of a tree across a large number of traits [6]. Data sets of phylogenetically distantly related species have a high overall PD (normalized for the number of taxa) in comparison with closely related species.

We applied the PD metric in SHERLOCK to characterize species communities (e.g. regional subsets) from within a large data set of species from a large geographic region. Mapping the observed and normalized overall PD of a particular community to a PD null-model distribution based on random subsets allows to test whether the taxonomic structure of an individual data set is significantly different from a null-model expectation. The extent of clustering or equipartition of a community is thereby reflected by the total, the mean, and the median branch length of the inferred community tree in proportion to a corresponding null-model distribution of random PD's. Whereas there are tools available for calculating patristic distances from trees in general [7,8], SHERLOCK allows for the first time a fully automatized analysis and visualization of the distribution of branch lengths, incorporating external software for alignment processing [9] and two Maximum likelihood (ML) tree reconstruction methods [10,11]. Statistical tests and result plots are generated with R-ggplot2 [12] and gridExtra [13].

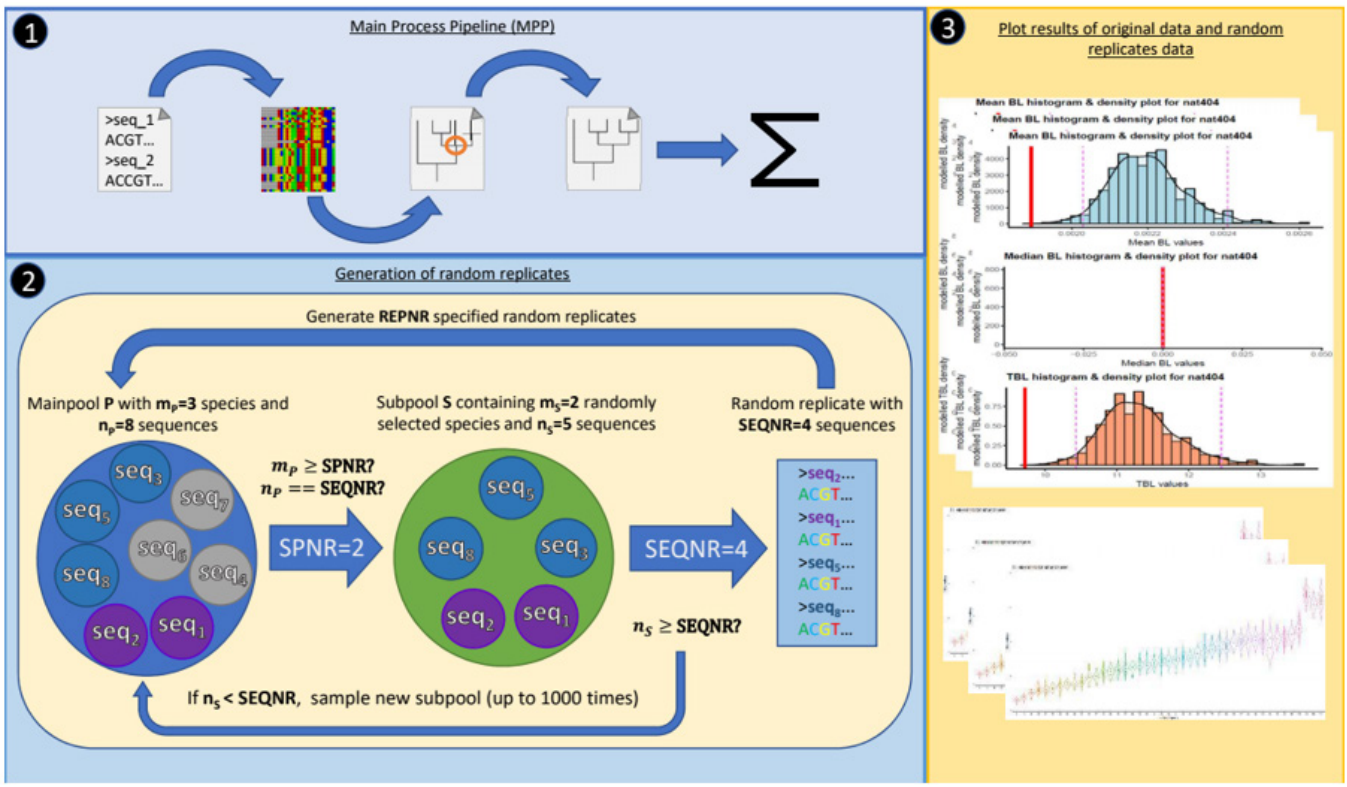

Figure 1: Process steps in SHERLOCK

1) Main process step (MPP), focusing (left to right) on raw data preparation (exclusion of potential gaps), alignment generation, ML tree reconstruction and resolution of polytomies, and patristic tree distance calculation. Different alignment and ML methods are available. Both, original and randomized data, are looped through the MPP.

2) Generation of randomized data replicates of the main pool of original data (P). Sampling conditions of random data follow user specifications about the total number of sampled species (SPNR), species related sequences (SEQNR), and the total number of replicates (REPNR). In the example above, P consists of eight sequences (np = 8; seq1 to seq8) falling under three different species (mp = 3; circled blue (seq3, seq5, seq8), grey (seq4, seq6, seq7), and violet (seq1, seq2)). First, the software checks in advance if P generally satisfies the specifications of SPNR and SEQNR, and aborts the analysis if SPNR > mp or if SEQNR > np. A subpool of sequences (S) is randomly generated from P, whereby the number of drawn species in S (ms) follows the specified number of allowed species (ms = SPNR). The software checks then if the set of randomly drawn species can satisfy the number of sequences in S (ns ≥ SEQNR). Otherwise (ns < SEQNR), S will be rejected and randomly re-sampled until a random set of species satisfies the SEQNR condition. SHERLOCK determines a fix set of ms with ns ≥ SEQNR if ns is < SEQNR in 1000 random re-sampling attempts of ms. Afterwards, the final replicate is randomly generated from S following the SEQNR condition (in our example until SEQNR = 4). This procedure is repeated for each random data until the number of random data is equal the number of defined replicates (REPNR). All random replicates are subsequently forwarded to the MPP chain processes, and the resulting PDs referenced to the PD of the original data.

3) Graphical outputs are:

i. histograms for all original data partitions, plotting the original PD against its corresponding random null distribution.

ii. Separate boxplots of the random data PD's according to SEQNR and SPNR.

SHERLOCK reads sequence data of different species communities in fasta format. Process settings for each analysis (number of random replicates, sequence composition of the null-model distribution (i.e., number of entities/taxa and specimens/sequences) and requested alignment and tree reconstruction methods) have to be defined by a text file. The null-model distribution sampling is specified by the number of entities/taxa and number of DNA sequences to be drawn from a main sequence pool, containing all sequences from coherent species communities underlying sub pools (identical sequences of the same taxon are sampled only once). As main output, SHERLOCK prints a histogram, a density, and a violin plot of PD measures of each community analysis. A more detailed list of actually identified and randomly expected PD values of analyzed communities, including 0.975 and 0.025-quantile limits, are printed as separate text files. An additional off-range file lists if the identified PD of a community is significant different

from the PD distribution of randomly drawn sequences. SHERLOCK identifies, excludes, and lists all redundant sequence names of given input data in advance of the analysis. A workflow of SHERLOCK's main process steps is shown in Figure 1. SHERLOCK is written in Perl, open source, and usable as a command line application on Linux systems. We provide a comprehensive manual describing all process steps, software implementations, script commands and input/result files of an exemplary PD analysis. SHERLOCK, the manual, and all additional files are free downloadable at GitHub: https://github.com/NathanSeidel/Sherlock.

References

1. Winter M, Devictor V, Schweiger O (2013) Phylogenetic diversity and nature conservation: Where are we?. Trends Ecol Evol 28(4): 199-204.
2. Isaac NJ, Turvey ST, Collen B, Waterman C, Baillie JE (2007) Mammals on the EDGE: Conservation priorities based on threat and phylogeny. PLoS ONE 2(3): e296.
3. Stegen JC, Hurlbert AH (2011) Inferring ecological processes from taxonomic, phylogenetic, and functional trait β-diversity. PloS ONE 6(6): e20906.
4. Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biol Conserv 61(1): 1-10.
5. Vellend M, Cornwell WK, Magnuson FK, Mooers AO (2011) Measuring phylogenetic biodiversity. Biological Diversity: Frontiers in Measurement and Assessment. Oxford University Press, United Kingdom.
6. Cavender BJ, Kozak KH, Fine PVA, Kembel SW (2009) The merging of community ecology and phylogenetic biology. Ecology Letters 12(7): 693-715.
7. Fourment M, Gibbs MJ (2006) PATRISTIC: A program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol Biol 6: 1-5.
8. Xia X (2013) DAMBE5: A comprehensive software package for data analysis in molecular biology and evolution. Mol Biol Evol 30(7): 1720-1728.
9. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30(4): 772-780.
10. Nguyen LT, Schmidt HA, Haesseler A, Minh BQ (2015) A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies. Mol Biol Evol 32(1): 268-274.
11. Price MN, Dehal PS, Arkin AP (2010) FastTree2- Approximately maximum-likelihood trees for large alignments. PLoS ONE 5(3): e9490.
12. Wickham H (2016) Ggplot2: Elegant graphics for data analysis. Springer-Verlag.
13. Auguie B (2017) GridExtra: Miscellaneous functions for ''grid'' graphics.