BMC Evolutionary Biology

SOFTWARE Open Access

PhySpeTree: an automated pipeline for reconstructing phylogenetic species trees

Yang Fang[1†], Chengcheng Liu[2], Jiangyi Lin[3], Xufeng Li[1], Kambiz N. Alavian[4,5], Yi Yang[1*] and Yulong Niu[1*†]

Abstract

Background: Phylogenetic species trees are widely used in inferring evolutionary relationships. Existing software and algorithms mainly focus on phylogenetic inference. However, less attention has been paid to intermediate steps, such as processing extremely large sequences and preparing configure files to connect multiple software. When the species number is large, the intermediate steps become a bottleneck that may seriously affect the efficiency of tree building.

Results: Here, we present an easy-to-use pipeline named PhySpeTree to facilitate the reconstruction of species trees across bacterial, archaeal, and eukaryotic organisms. Users need only to input the abbreviations of species names; PhySpeTree prepares complex configure files for different software, then automatically downloads genomic data, cleans sequences, and builds trees. PhySpeTree allows users to perform critical steps such as sequence alignment and tree construction by adjusting advanced options. PhySpeTree provides two parallel pipelines based on concatenated highly conserved proteins and small subunit ribosomal RNA sequences, respectively. Accessory modules, such as those for inserting new species, generating visualization configurations, and combining trees, are distributed along with PhySpeTree.

Conclusions: Together with accessory modules, PhySpeTree significantly simplifies tree reconstruction. PhySpeTree is implemented in Python running on modern operating systems (Linux, macOS, and Windows). The source code is freely available with detailed documentation (https://github.com/yangfangs/physpetools).

Keywords: Species tree, Automatic construction, Pipeline

Background

The reconstruction of phylogenetic species trees is of central importance in many biological disciplines. For example, the tree of life provides a remarkable view of organizing principles in biology [1, 2]. In addition, many new genomes are being sequenced, and their taxonomic identities can be determined by inserting them into prebuilt species trees [3]. Moreover, combined with species trees, phylogenetic profiling using gain and loss patterns of homologs achieves high performance in predicting protein linkages [4–8].

Toolkits and pipelines have been developed for phylogenetic reconstruction (Table 1). Toolkits such as BuddySuite [9], ETE3 [10], and MEGA [11] are widely used for phylogenetic inference and tree manipulation. BuddySuite and ETE3 provide rich interfaces that allow researchers to carry out secondary development. BuddySuite includes a pipeline with which to reconstruct gene or species trees, but third-party software needs to be specified and manually installed in the local running environment, which may be inconvenient for users on different platforms. MEGA is a standalone and cross-platform program, and it also provides a user-friendly graphical interface. BIR [12], Agalma [13], PhyloPlAn [14], and AMPHORA [12] are designed for phylogenomic analysis. BIR is particularly useful for preparing gene sequences for phylogenomic inference. Agalma has a command-line interface for phylogenomic analyses based on genomic and transcriptome data. PhyloPlAn and AMPHORA (AMPHORA2 [14]) are effective pipelines for large-scale phylogenetic inference based on thoroughly tested marker genes, and other operations such as

* Correspondence: yangyi528@scu.edu.cn; yulong.niu@hotmail.com
†Yang Fang and Yulong Niu contributed equally to this work.
[1]Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, People's Republic of China
Full list of author information is available at the end of the article

Table 1 Comparison of phylogenetic tree construction software

Software	Auto-pipeline[a]	Alignment			Trim		Tree building			Tree merge	
		MUSCLE	MAFFT	ClustalW	Gblocks	trimAI	RAxML	FastTree	IQ-TREE	ASTRAL	SPRSupertrees
BuddySuite [9]	×	√	√	√	×	√	√	√	√	×	×
ETE3 [10]	×	√	√	×	×	√	√	√	×	×	×
MEGA [11]	×	√	×	√	×	×	×	×	×	×	×
BIR [12]	×	√	√	√	√	√	√	√	×	×	×
Agalma [13]	×	×	×	×	√	×	√	×	×	×	×
PhyloPhlAn [14]	×	√	×	×	×	×	√	√	×	×	×
AMPHORA [15]	×	×	×	√	√	×	√	×	×	×	×
PhySpeTree	√	√	√	√	√	√	√	√	√	√	√

[a] The automation refers to sequences download, preprocess, and preparation of configure files. Critical steps such as sequence alignment and tree construction can be manually adjusted with advanced options in PhySpeTree

taxonomic curation, estimation, and insertion are also available. The marker genes, however, are conserved only between microbial genomes, so PhyloPlAn and AMPHORA are limited to reconstructing bacterial and archaeal species trees.

Although the software mentioned above are powerful in inferring phylogenies, most require users to manually download genomic data, clean and align sequences, or prepare complex configure files. These laborious and time-consuming steps may impede tree reconstruction, especially when the number of species becomes large. Hence, there is a clear need for a flexible and efficient pipeline that can reduce the time required for species tree building processes.

Here, we present an easy-to-use Python package named PhySpeTree, which provides an automated solution for the entire process of species tree reconstruction, from data collection to tree building. PhySpeTree has two parallel pipelines based on either the most commonly adopted small subunit ribosomal RNA (SSU rRNA) [15] or concatenated highly conserved proteins (HCPs) [16]. The distinguishing feature of PhySpeTree is its automated design. Users need only to input the abbreviations of species names, and then PhySpeTree can automatically download and analyze sequences. Some critical steps, such as multiple sequence alignment and tree construction, can be manually adjusted. Moreover, PhySpeTree contains modules to facilitate downstream analysis. For example, users can apply the "autobuild" module to extend prebuilt trees by inserting new organisms. The "iview" and "combine" modules are designed for tree visualization in iTOL [17] and consensus tree construction [18], respectively. Together with accessory modules, PhySpeTree significantly simplifies tree reconstruction.

Implementation

PhySpeTree is implemented in Python and distributed as an independent package. PhySpeTree integrates multiple tools and provides an automated solution for reconstructing species trees (Table 1). The workflow of PhySpeTree is shown in Fig. 1. First, users input the abbreviations of species names (Additional file 1: Figure S1 and Additional file 2: Table S2) and choose the sequence type (SSU rRNA or HCP) to build species trees. If the HCP option is selected, PhySpeTree retrieves and concatenates HCP sequences from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [19]. Otherwise, PhySpeTree uses SSU rRNA sequences from the SILVA database [20]. For unannotated organisms, users can prepare FASTA format files containing either HCP or SSU rRNA sequences and then insert them into prebuilt databases. Second, multiple sequence alignment is conducted by MUSCLE [21], MAFFT [22], or ClustalW [23], and conserved blocks are selected by Gblocks [24] or trimAI [25]. Finally, PhySpeTree reconstructs species trees by RAxML [18], IQ-TREE [26], or FastTree [27]. In addition, PhySpeTree provides flexible modules to facilitate downstream analysis, such as generating visualization files for iTOL [17] and tree combination (Fig. 1).

SSU rRNA option

For bacterial and archaeal organisms, SSU rRNA sequences are widely used to build species trees [2]. We prebuilt a dataset according to the latest version of the SILVA database (Release 132, Dec. 13, 2017) [20]. The dataset contains truncated SSU rRNA sequences from 140,662 species, and nucleotides that are not aligned are removed (Additional file 1: Figure S1A and Additional file 2: Table S1). When the SSU rRNA option is selected, PhySpeTree automatically fetches related sequences.

HCP option

It has been reported that HCP-based species trees have a higher resolution than the ones built based on a single gene [15]. Hence, PhySpeTree also provides the HCP option. First, we chose 31 single-copy HCPs without horizontal transfer from Ciccarelli et al [16]. Then, we

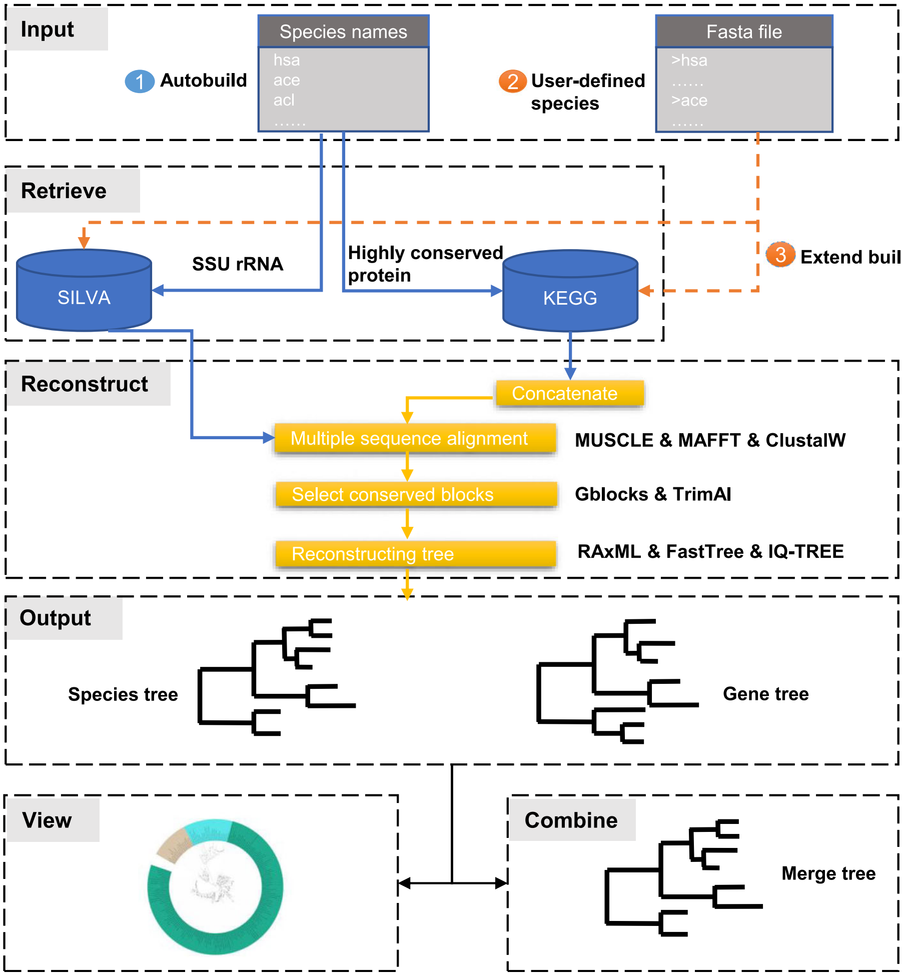


Fig. 1 The workflow of PhySpeTree. PhySpeTree retrieves and downloads corresponding sequences following with multiple sequence alignments, conserved blocks selection, and tree reconstruction. PhySpeTree also allows users to insert their own sequences of HCP or SSU rRNA. Output trees are in the "newick" format files. ①Automatic tree reconstruction ②Processing user-defined fasta files for unannotated organisms ③Reconstructing species trees with unannotated organisms

manually mapped them to KEGG orthologues (Release 90.1, May 1, 2019) [19] (Additional file 2: Table S3). When users choose the HCP option, PhySpeTree directly retrieves HCP sequences from the KEGG database. The HCP option currently supports 5943 organisms (Additional file 1: Figure S1B and Additional file 2: Table S2).

Sequence alignment and tree reconstruction

PhySpeTree integrates various tools for multiple sequence alignment and tree reconstruction. For sequence alignment, MUSCLE [21], MAFFT [22], and Clustal [22] are provided. To infer accurate phylogenies, the maximum likelihood-based method RAxML is set as the default option [18]. In addition, IQ-TREE [26] and FastTree [27] are alternatives to accelerate tree reconstruction. Advanced parameters of integrated tools can be specifically set and passed in PhySpeTree, allowing users to manipulate critical steps in sequence alignment and tree reconstruction.

Result

Modules of PhySpeTree

PhySpeTree contains five modules. The main module "**autobuild**" is developed to automatically build species trees. With this module, users do not need to prepare sequences in advance. Instead, the abbreviations of species names are the only required inputs. The intermediate steps, e.g., sequence download, cleaning, alignment, and tree reconstruction, are automatically handled by PhySpeTree. The following command line shows an example:

```
$ PhySpeTree autobuild -i species_names.txt --hcp
```

where "species_names.txt" is the file of abbreviated organism names; for example, "hsa" represents *Homo sapiens* (Additional file 2: Table S2). "--hcp" indicates that the HCP option is selected.

Moreover, the "**autobuild**" module can be used to extend prebuilt trees by inserting new organisms whose genome annotations may be incomplete. For the new organisms, the SSU rRNA may come from experiments, while orthologous databases such as eggNOG [28] and OMA [29] are good resources for searching for corresponding HCP sequences. FetchMG [30] is also available to identify HCP sequences in reference genomes and metagenomes. The following commands illustrate how to insert a new organism into trees:

```
$ PhySpeTree autobuild -i species_names.txt -e
new_hcp.fasta --ehcp
```

where "new_hcp.fasta" is the HCP sequence of the new organism. The file should be prepared by users. "--ehcp" indicates that the tree is extended according to HCP sequences.

Instead of using default settings, in the "**autobuild**" model, users can adjust advanced options to control critical steps, such as sequence alignment, conserved block selection, and tree building. The following command shows how to set advanced options of RAxML:

```
$ PhySpeTree autobuild -i species_names.txt --srna --raxml
--raxml_p '-f a -m GTRGAMMA -p 12345 -× 12,345 -# 100 -n T1'
```

where "--raxml_p" indicates advanced options passed to RAxML.

The module "**build**" is developed for advanced users to directly reconstruct trees from protein or gene sequences. It is practically useful to reconstruct trees by user-defined sequences other than SSU rRNA or HCP sequences. This function may overlap with ETE3 and MEGA and is executed as:

```
$ PhySpeTree build -i defined_seq.fasta --single
```

where "defined_seq.fasta" is a FASTA file containing user-defined sequences. "--single" indicates a single sequence for each organism.

The "**iview**" module is designed to facilitate tree visualization. It provides a convenient interface used to generate configure files for iTOL, which is a powerful online tool for tree display, annotation, and manipulation [17]. The taxonomy of species is directly retrieved from the KEGG database. The following command annotates input species at the phylum level:

```
$ PhySpeTree iview -i species_names.txt --range -a phylum
```

where "species_names.txt" is the same file as in the "autobuild" module.

To reconstruct a consensus tree, PhySpeTree uses the "**combine**" module to merge multiple trees. This module is useful for comparing and selecting conserved branches from trees generated by different sequences or tree building methods. It is implemented as follows:

```
$ PhySpeTree combine -i combine.tree
```

where "combine.tree" is a file containing multiple trees.

The module "**check**" is designed to check whether input species are supported in PhySpeTree. It can also be used to check sequence information that is needed to extend the current tree. The following command will return species that are not supported by the HCP option:

```
$ PhySpeTree check -i species_names.txt –hcp
```

Installation The PhySpeTree pipeline is implemented in Python and has been tested on Linux systems such as Fedora, Ubuntu, and CentOS. We also released a Docker image to support Windows and macOS. The latest version can be installed as follows:

```
$ pip install PhySpeTree
```

Alternatively, PhySpeTree can be directly installed from the GitHub repository. Code is available at https://github.com/yangfangs/physpetools/releases, and PhySpeTree is installed by a local command as follows (executed in the PhySpeTree directory):

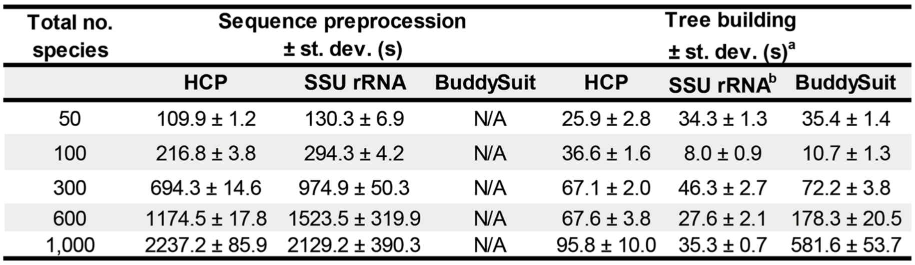
```
$ python setup.py install
```

Usage and tutorial To facilitate the use of PhySpeTree, we distribute a detailed tutorial (https://yangfangs.github.io/physpetools/) (Additional file 3). The tutorial provides step-by-step examples to show how to use the modules mentioned above.

Benchmark test of the efficiency and consistency of PhySpeTree

To test the efficiency of PhySpeTree, we simulated five data sets with different numbers of species (50, 100, 300, 600, and 1000) by randomly selecting taxa from our prebuilt HCP and SSU rRNA databases. Each data set was independently generated three times, and the mean and standard deviation of the run time were recorded. One of the great advantages of using PhySpeTree is that it provides automated sequence preprocessing (e.g., querying databases, downloading sequences, and formatting). The time required for this process showed linear growth with an increase in the number of species (Table 2). It took approximately 3 s to preprocess one species, and most of the time was spent on querying remote prebuilt databases. The prebuilt databases can be downloaded and easily deployed, so we provided a special option,

Table 2 Run time test of PhySpe Tree



Total no. species	Sequence preprocession ± st. dev. (s)			Tree building ± st. dev. (s)[a]		
	HCP	SSU rRNA	BuddySuit	HCP	SSU rRNA[b]	BuddySuit
50	109.9 ± 1.2	130.3 ± 6.9	N/A	25.9 ± 2.8	34.3 ± 1.3	35.4 ± 1.4
100	216.8 ± 3.8	294.3 ± 4.2	N/A	36.6 ± 1.6	8.0 ± 0.9	10.7 ± 1.3
300	694.3 ± 14.6	974.9 ± 50.3	N/A	67.1 ± 2.0	46.3 ± 2.7	72.2 ± 3.8
600	1174.5 ± 17.8	1523.5 ± 319.9	N/A	67.6 ± 3.8	27.6 ± 2.1	178.3 ± 20.5
1,000	2237.2 ± 85.9	2129.2 ± 390.3	N/A	95.8 ± 10.0	35.3 ± 0.7	581.6 ± 53.7

[a] The tree reconstruction pipeline was conducted by MAFFT (alignment), Gblocks (trim), and FastTree (tree building). Benchmark test was conducted with i7-4790 3.6GHz CPU (parallel on 6 threads) with 16GB memory on Fedora operating system. [b] "--auto" option was turn on in MAFFT. The alignment strategy was automatically chose according to the number and length of sequences

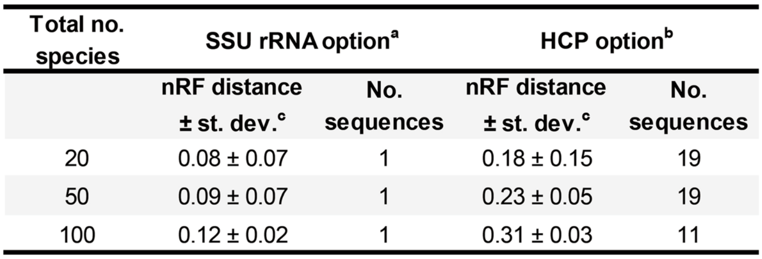
"-db", to further improve efficiency by manipulating sequences on local computers. Another advanced feature of PhySpeTree is the fully optimized configuration of software streams. We then compared the run time of tree building in PhySpeTree with that of a pipeline in BuddySuit [9]. The same third-party software and parameters were used. In PhySpeTree, the SSU rRNA option was slightly better than the HCP option, mainly because more sequences were processed with the HCP option. The run time of BuddySuit was comparable to that of PhySpeTree when building small tress (fewer than 300 species), whereas PhySpeTree outperformed BuddySuit when the number of species increased. For example, in building a tree with 1000 species, the optimized configuration of PhySpeTree resulted in more than a 5X speed gain (Table 2). Overall, our benchmark tests showed that PhySpeTree is a highly efficient pipeline in the reconstruction of large-scale trees.

There is a lack of ground truth for evaluating the topological accuracy of phylogenies across a wide range of species. As a surrogate, we quantitatively assessed the consistency of species trees from PhySpeTree with respect to the updated tree of life [31]. Because most organisms in the tree of life were uncultured or newly identified, we manually checked and filtered species names from our prebuilt databases. Finally, the SSU rRNA and HCP options matched 154 and 122 species, respectively (Additional file 2: Table S4 and Table S5). We then randomly selected 20, 50, and 100 species and used PhySpeTree to reconstruct species trees with both the SSU rRNA and HCP options. Normalized Robinson-Foulds (nRF) distances, ranging from 0 (identical) to 1 (most unlikely), were calculated to measure topological similarity (Table 3) [10, 32]. Unsurprisingly, SSU rRNA trees archived near perfect consistency (mean nRF distance < 0.13) with the tree of life, as almost identical SSU rRNA sequences were used. For up to 100 species, we found that the HCP option of PhySpeTree was also feasible (mean nRF distance: 0.18 ~ 0.32). Notably, increasing the number of species did not significantly reduce the accuracy of tree reconstruction. The topological dissimilarity between HCP trees and the tree of life was mainly due to the number and type of conserved proteins used to reconstruct the trees.

Case study: the evolutionary position of the archaeal phylum *Lokiarchaeota*

A recent study reported a novel archaeal phylum, *Lokiarchaeota*. Genomes in this phylum encode various eukaryotic signature proteins. Further phylogenetic analysis revealed a close relationship between *Lokiarchaeota* and Eukarya [3]. However, debates about the *Lokiarchaeota*-Eukarya affiliation arose mainly due to the number of species and HCPs used in tree reconstruction [3, 33]. PhySpeTree can be conveniently applied to investigate the evolutionary position of newly identified organisms. Thus, here, we provide

Table 3 Consistency test of PhySpe Tree comparied with the updated tree of life [33]



Total no. species	SSU rRNA option[a]		HCP option[b]	
	nRF distance ± st. dev.[c]	No. sequences	nRF distance ± st. dev.[c]	No. sequences
20	0.08 ± 0.07	1	0.18 ± 0.15	19
50	0.09 ± 0.07	1	0.23 ± 0.05	19
100	0.12 ± 0.02	1	0.31 ± 0.03	11

[a] SSU rRNA sequences were retrieved by PhySpeTree, aligned by SINA, and tree reconsturction by RAxML (GTRCAT model). [b] HCP sequences were retrieved by PhySpTree, aligned by MUSCLE, and tree reconstruction by RAxML (PROTGAMMAJTTX model). [c] Normalized Robinson-Foulds (nRF) distance was calculated by ETE3 [10]

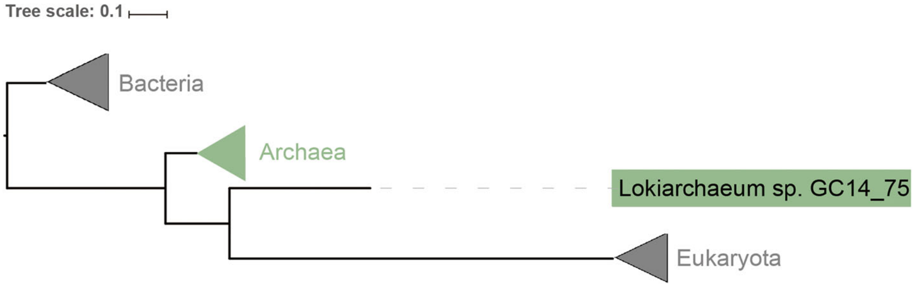


Fig. 2 Extended tree of life with *Lokiarchaeota*. The *Lokiarchaeum sp.* GC14_75 was inserted into the tree of life. The bacterial, eukaryotic, and archaeal branches were clasped

an example to show how to insert *Lokiarchaeum sp.* GC14_75 (*loki*) into prebuilt species trees.

At first, we randomly chose 1246 species, including 440 eukaryotic, 544 bacterial, and 280 archaea species, from the prebuilt HCP database, then used PhySpeTree to reconstruct a species tree with the HCP option (Additional file 2: Table S6). Next we prepared 25 HCPs of *loki* (ribosomal protein L1/L3/L5/L11/L13/L14/L22/ S2/S3/S4/S5/S7/S8/S9/S11/S13/S15/S17, phenylalanine−/seryl−/leucyl−/arginyl-tRNA synthetase, metal-dependent proteases with chaperone activity, predicted GTPase probable translation factor, and preprotein translocase subunit SecY) and used the "autobuild" module (with the "-e" option) to expand the tree of life with *loki* (Fig. 2 and Additional file 4). In accordance with the previous species trees [34, 35] reconstructed based on 55 concatenated ribosomal proteins, our results indicated phylogenetic affiliation between *loki* and eukaryotes. Although our results did not support that *loki* and other archaeal lineages were monophyletic [36], various tree topologies can be easily explored by PhySpeTree with different HCPs.

Conclusions

We developed an automated pipeline named PhySpeTree to reconstruct species trees across bacteria, archaea, and eukaryotes. The PhySpeTree pipeline contains as many options as other tree-building tools (detailed comparison in Table 1). However, another feature sets PhySpeTree apart: it automates intermediate processes, including retrieving sequences from public databases, preparing complex configure files to run different software, aligning sequences, and building trees. The inputs of PhySpeTree are simple that users need only to prepare the abbreviations of species names. For unannotated organisms, users can apply the "check" and "autobuild" modules in PhySpeTree to prepare sequence files. Because PhySpeTree is frequently synchronized with the most recent public databases, the number of unannotated organisms is expected to be small.

PhySpeTree provides both the traditional SSU rRNA option and the HCP option to reconstruct species trees. Benefiting from comprehensive rRNA databases (e.g., SILVA and RDP) [20, 37] and high-throughput rRNA amplicon sequencing [38], SSU rRNA has been widely used as a phylogenetic marker for taxonomic identification [31, 39]. However, inferring taxonomies based on a single marker gene is challenging, given that chimeric sequences arising from PCR and sequencing errors can corrupt tree topologies [40] as well as the limited resolution of SSU rRNA in closely related species [41]. Compared with trees obtained from a single marker gene, those reconstructed by the concatenation of highly conserved single-copy proteins show a higher resolution [16, 31, 42]. For example, to explore the phylogenetic history of organisms, a species tree across all three domains of life was generated based on HCPs [16]. The same set of HCPs was applied for the species assignment of prokaryotic genomes [43] and to establish metagenomic operational taxonomic units [44] and is applied in the HCP option in PhySpeTree. Recently, several revised species trees have been inferred by the concatenation of 16 ribosomal proteins [31] or 120 bacterial proteins [45, 46] to explore the tree of life. Although HCPs are extensively used, when applying the HCP option of PhySpeTree, users should be aware of the limitations of HCPs, such as recombination [42] and potential lateral gene transfer [47].

PhySpeTree was developed in Python and is executed as command lines, so it is easy for advanced developers to expand its modules or integrate PhySpeTree with other phylogenetic tools. For example, PUmPER [48] updates existing trees with new gene sequences. PhySpeTree may work as a complementary tool in terms of building the initial tree and automatically preparing updated sequences from public databases. On the other hand, users of PhySpeTree are reminded that phylogenetic discordance mainly caused by different evolutionary processes affects species tree accuracy [49]. Coalescent-based methods are broadly used to address incongruence [50]. ASTRAL [51] and NJst [52] are efficient tools for handling incomplete lineage sorting. They are also robust to branch length errors, which may result from rate heterogeneity. PhyloNet [53, 54], iGTP [55], Guenomu

[56], and SPRSupertrees [57] consider gene flow, gene duplication and loss, or horizontal gene transfer when inferring species trees. The coalescent-based methods mentioned above take gene trees as inputs, which can be conveniently estimated by PhySpeTree ("build" module) or any other tool listed in Table 1. Moreover, species trees inferred from PhySpeTree can benefit from other types of evidence; for example, fossils and ancient DNA can be incorporated into node-based and tip-based calibration [58, 59].

Availability and requirements

Project name: PhySpeTree

Project home page: https://yangfangs.github.io/physpetools/

Operating systems: Linux (Docker image for Windows and macOS)

Programming language: Python 2.7+ and python 3+

License: GNU General Public License v3.0

Other requirements: None

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12862-019-1541-x.

Additional file 1: Figure S1. The taxonomic distribution of species supported by the HCP (A) and SSU rRNA options (B).

Additional file 2: Table S1. The list of 140,662 species supported in the SSU rRNA option. **Table S2.** The list of 5943 species supported in the HCP option. **Table S3.** The list of 31 highly conserved proteins and corresponding KEGG IDs. **Table S4** and **Table S5.** The lists of SSU rRNA and HCP matched species between prebuilt databases of PhySpeTree and the updated tree of life, respectively. **Table S6.** Species used to reconstruct the tree of life in Fig. 2.

Additional file 3. The step by step usage and tutorial for PhySpeTree.

Additional file 4. Data used to extend tree-of-life with *Lokiarchaeum sp.* GC14_75. "FastTree.tree" is the output tree file. "tree_of_life_species_names_abb.txt" contains the species abbreviated names to use reconstruct tree-of-life. "highly_conserved_protein_loki" contains *Lokiarchaeum sp.* GC14_75 HCP sequences. "parameter.txt" contains parameter commands.

Abbreviations

HCP: Highly Conserved Protein; RNA; KEGG: Kyoto Encyclopedia of Genes and Genomes; SSU rRNA: Small subunit ribosomal

Acknowledgements

We thank Yaodong Liu for testing PhySpeTree.

Authors' contributions

Y.L.N and Y.Y. conceived and designed the experiments. Y.F. implemented the software. Y.L.N, J.Y.L. and K.N.A. tested the software. Y.F. and Y.L.N. wrote the paper. C.C.L and X.F.L revised the manuscript. All authors read and approved the final manuscript.

Funding

This study was funded by the Fundamental Research Funds for the Central Universities 2019SCU12062 (Y.L.N), National 973 Projects 2015CB755702 (Y.Y.), and NSFC31671455 (Y.Y.).

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Key Laboratory of Bio-Resources and Eco-Environment of Ministry of Education, College of Life Sciences, Sichuan University, Chengdu, People's Republic of China. [2]State Key Laboratory of Oral Diseases & National Clinical Research Center for Oral Diseases &Department of Periodontics, West China Hospital of Stomatology, Sichuan University, Chengdu, China. [3]Wu YuZhang Honors College of Sichuan University, Chengdu, People's Republic of China. [4]Department of Medicine, Division of Brain Sciences, Imperial College London, London, UK. [5]Department of Internal Medicine, Endocrinology, Yale University, New Haven, USA.

Received: 19 June 2019 Accepted: 13 November 2019
Published online: 02 December 2019

References

1. Pace NR. Mapping the tree of life: progress and prospects. Microbiol Mol Biol Rev. 2009;73(4):565–76.
2. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990;87(12):4576–9.
3. Spang A, Saw JH, Jorgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJ. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature. 2015;521(7551):173–9.
4. Barker D, Pagel M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol. 2005;1(1):e3.
5. Craig RA, Liao L. Phylogenetic tree information aids supervised learning for predicting protein-protein interaction based on distance matrices. BMC Bioinformatics. 2007;8:6.
6. Li Y, Calvo SE, Gutman R, Liu JS, Mootha VK. Expansion of biological pathways based on evolutionary inference. Cell. 2014;158(1):213–25.
7. Niu Y, Liu C, Moghimyfiroozabad S, Yang Y, Alavian KN. PrePhyloPro: phylogenetic profile-based prediction of whole proteome linkages. PeerJ. 2017;5:e3712.
8. Niu Y, Moghimyfiroozabad S, Safaie S, Yang Y, Jonas EA, Alavian KN. Phylogenetic profiling of mitochondrial proteins and integration analysis of bacterial transcription units suggest evolution of F1Fo ATP synthase from multiple modules. J Mol Evol. 2017;85(5–6):219–33.
9. Bond SR, Keat KE, Barreira SN, Baxevanis AD. BuddySuite: command-line toolkits for manipulating sequences, alignments, and phylogenetic trees. Mol Biol Evol. 2017;34(6):1543–6.
10. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of Phylogenomic data. Mol Biol Evol. 2016;33(6):1635–8.
11. Kumar S, Nei M, Dudley J, Tamura K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief Bioinform. 2008;9(4):299–306.
12. Wu M, Eisen JA. A simple, fast, and accurate method of phylogenomic inference. Genome Biol. 2008;9(10):R151.
13. Dunn CW, Howison M, Zapata F. Agalma: an automated phylogenomics workflow. BMC Bioinformatics. 2013;14:330.
14. Segata N, Bornigen D, Morgan XC, Huttenhower C. PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat Commun. 2013;4:2304.
15. Hug LA, Castelle CJ, Wrighton KC, Thomas BC, Sharon I, Frischkorn KR, Williams KH, Tringe SG, Banfield JF. Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling. Microbiome. 2013;1(1):22.

16. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. Toward automatic reconstruction of a highly resolved tree of life. Science. 2006; 311(5765):1283–7.
17. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44(W1):W242–5.
18. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
19. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28(1):27–30.
20. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glockner FO. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41(Database issue):D590–6.
21. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32(5):1792–7.
22. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30(14):3059–66.
23. Thompson JD, Gibson TJ, Higgins DG. Multiple sequence alignment using ClustalW and ClustalX. Curr Protocols Bioinformatics. 2002;Chapter 2:Unit 2–3.
24. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 2007;56(4):564–77.
25. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
26. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.
27. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5(3):e9490.
28. Jensen LJ, Julien P, Kuhn M, von Mering C, Muller J, Doerks T, Bork P. eggNOG: automated construction and annotation of orthologous groups of genes. Nucleic Acids Res. 2008;36(Database issue):D250–4.
29. Dessimoz C, Cannarozzi G, Gil M, Margadant D, Roth A, Schneider A, Gonnet GH. OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: introduction and first achievements. In: RECOMB Workshop on Comparative Genomics. Berlin, Heidelberg: Springer; 2005. p. 61–72.
30. Kultima JR, Sunagawa S, Li J, Chen W, Chen H, Mende DR, Arumugam M, Pan Q, Liu B, Qin J, et al. MOCAT: a metagenomics assembly and gene prediction toolkit. PLoS One. 2012;7(10):e47656.
31. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048.
32. Robinson DF, Foulds LR. Comparison of phylogenetic trees. Math Biosci. 1981;53(1–2):131–47.
33. Nasir A, Kim KM, Da Cunha V, Caetano-Anolles G. Arguments reinforcing the three-domain view of diversified cellular life. Archaea. 2016;2016:1851865.
34. Spang A, Stairs CW, Dombrowski N, Eme L, Lombard J, Caceres EF, Greening C, Baker BJ, Ettema TJG. Proposal of the reverse flow model for the origin of the eukaryotic cell based on comparative analyses of Asgard archaeal metabolism. Nat Microbiol. 2019;4(7):1138–48.
35. Zaremba-Niedzwiedzka K, Caceres EF, Saw JH, Backstrom D, Juzokaite L, Vancaester E, Seitz KW, Anantharaman K, Starnawski P, Kjeldsen KU, et al. Asgard archaea illuminate the origin of eukaryotic cellular complexity. Nature. 2017;541(7637):353–8.
36. Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P. Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS Genet. 2017;13(6):e1006810.
37. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM. Ribosomal database project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42(Database issue):D633–42.
38. Medlar A, Aivelo T, Loytynoja A. Seance: reference-based phylogenetic analysis for 18S rRNA studies. BMC Evol Biol. 2014;14:235.
39. Fontaneto D, Wu S, Xiong J, Yu Y. Taxonomic resolutions based on 18S rRNA genes: a case study of subclass Copepoda. PLoS One. 2015;10(6): e0131498.
40. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl Environ Microbiol. 2006;72(7):5069–72.
41. Poretsky R, Rodriguez RL, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One. 2014;9(4):e93827.
42. Thiergart T, Landan G, Martin WF. Concatenated alignments and the case of the disappearing tree. BMC Evol Biol. 2014;14:266.
43. Sunagawa S, Mende DR, Zeller G, Izquierdo-Carrasco F, Berger SA, Kultima JR, Coelho LP, Arumugam M, Tap J, Nielsen HB, et al. Metagenomic species profiling using universal phylogenetic marker genes. Nat Methods. 2013; 10(12):1196–9.
44. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh HJ, Cuenca M, Hingamp P, Alves R, Costea PI, Coelho LP, et al. Microbial abundance, activity and population genomic profiling with mOTUs2. Nat Commun. 2019;10(1):1014.
45. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil PA, Hugenholtz P. A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nat Biotechnol. 2018;36(10): 996–1004.
46. Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, Hugenholtz P, Tyson GW. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2(11): 1533–42.
47. Ku C, Martin WF. A natural barrier to lateral gene transfer from prokaryotes to eukaryotes revealed from genomes: the 70% rule. BMC Biol. 2016;14(1):89.
48. Izquierdo-Carrasco F, Cazes J, Smith SA, Stamatakis A. PUmPER: phylogenies updated perpetually. Bioinformatics. 2014;30(10):1476–7.
49. Maddison WP. Gene trees in species trees. Syst Biol. 1997;46(3):523–36.
50. Mallo D, Posada D. Multilocus inference of species trees and DNA barcoding. Philos Trans R Soc Lond B Biol Sci. 2016;371(1702). https://doi.org/10.1098/rstb.2015.0335.
51. Mirarab S, Reaz R, Bayzid MS, Zimmermann T, Swenson MS, Warnow T. ASTRAL: genome-scale coalescent-based species tree estimation. Bioinformatics. 2014;30(17):i541–8.
52. Liu L, Yu L. Estimating species trees from unrooted gene trees. Syst Biol. 2011;60(5):661–7.
53. Solis-Lemus C, Yang M, Ane C. Inconsistency of species tree methods under gene flow. Syst Biol. 2016;65(5):843–51.
54. Yu Y, Dong J, Liu KJ, Nakhleh L. Maximum likelihood inference of reticulate evolutionary histories. Proc Natl Acad Sci U S A. 2014;111(46):16448–53.
55. Chaudhary R, Bansal MS, Wehe A, Fernandez-Baca D, Eulenstein O. iGTP: a software package for large-scale gene tree parsimony analysis. BMC Bioinformatics. 2010;11:574.
56. De Oliveira Martins L, Mallo D, Posada D. A Bayesian Supertree model for genome-wide species tree reconstruction. Syst Biol. 2016;65(3):397–416.
57. Whidden C, Zeh N, Beiko RG. Supertrees based on the subtree prune-and-Regraft distance. Syst Biol. 2014;63(4):566–81.
58. Donoghue PC, Yang Z. The evolution of methods for establishing evolutionary timescales. Philos Trans R Soc Lond B Biol Sci. 2016;371(1699). https://doi.org/10.1098/rstb.2016.0020.
59. Rieux A, Balloux F. Inferences from tip-calibrated phylogenies: a review and a practical guide. Mol Ecol. 2016;25(9):1911–24.

Publisher's Note