Journal Title Here, 2022, pp. 1–3

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Article Title

First Author,[1,*] Second Author,[2] Third Author,[3] Fourth Author[3] and Fifth Author[4]

[1]Department, Organization, Street, Postcode, State, Country, [2]Department, Organization, Street, Postcode, State, Country, [3]Department, Organization, Street, Postcode, State, Country and [4]Department, Organization, Street, Postcode, State, Country
[*]Corresponding author. email-id.com
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Abstracts must be able to stand alone and so cannot contain citations to the paper's references, equations, etc. An abstract must consist of a single paragraph and be concise. Because of online formatting, abstracts must appear as plain as possible.

Key words: keyword1, Keyword2, Keyword3, Keyword4

Introduction

Phylogenetic network is an acyclic directed graph that generalizes the bifurcating phylogenetic tree by allowing nodes to have indegree of two, thereby creates a reticulation structure [Huson et al., 2010, Kong et al., 2022a]. Phylogenetic networks depict complex biological scenarios that the trees cannot, such as hybrid speciation, introgression, allopolyploid speciation, and so on [Huson and Bryant, 2006][add the new review paper here]. A handful of computational methods that estimate networks from genomic data has been proposed, however, the wide use of networks in practice is hindered by their lack of scalability which refers to the ability of a system to process a growing amount of work in a decreasing or stable amount of time [Bondi, 2000]. More precisely, phylogenetic network estimation belongs to the class of NP-Hard problems (non-deterministic polynomial-time). Some attempts to ameliorate this issue has been made but the computational requirement is still much higher for the dataset size typically applied to the tree estimation (i.e., tens of taxa).

A common strategy to enhance the efficiency of the network inference is to summarize input sequence data into a set of gene trees in prior to the analysis as implemented in many functions in PhyloNet [Than et al., 2008, Wen et al., 2018] or SNaQ [Solís-Lemus and Ané, 2016] available in Julia package PhyloNetworks [Solís-Lemus et al., 2017]. Computational cost is further ameliorated by using composite likelihood (or pseudolikelihood) that involves decomposition of the network into a set of smaller problems (e.g., triplets or quartets), excecute likelihood computation on each of them, and combine them together to approximate the likelihood of the full network. This approach has been useful in both tree (e.g., MP-EST [Liu et al., 2010]) and network (e.g., SNaQ, PhyNEST [Kong et al., 2022b], PhyloNet [Yu and Nakhleh, 2015, Zhu and Nakhleh, 2018]) inference and shown to be much faster than the full likelihood or the Bayesian methods, without compromising accuracy [Hejase and Liu, 2016].

Nevertheless, network inference is still a computationally demanding procedure. It is not uncommon to conduct the analysis with a handful of taxa, which narrows the scope of biological investigation. In this study, we present SNaQ2, a new version of SNaQ with improved computational efficiency via parallelization of the composite likelihood computation and making probabilistic decisions during the network searching heuristics. In the following, we first briefly introduce the essence of the original SNaQ followed by the key improvemetns made in SNaQ2. Then, we present the result of our benchmarks that compares the performance of SNaQ and SNaQ2 and we apply SNaQ2 on empirical datasets. Our results clearly demonstrate improved efficiency in SNaQ2.

Methods

Original SNaQ

As mentioned, SNaQ estimates phylogenetic networks from multi-locus data using composite likelihood. While detailed description of the method is available in Solís-Lemus and Ané [2016], we make a brief description here for readers to understand the improvemetns in SNaQ2 in the following subsection. First, SNaQ extracts unrooted quarnet (i.e., networks with four tips) from the full network, and computes expected concordance factor (CF) for each quarnet that represents the proportion of genes whose true relationship is the quartet under the coalescent model. Note there are $\binom{4}{2}=3$ possible ways to cluster four taxa into two groups of two (i.e., separated by a split). In other words, for a taxon set $X \in \{a, b, c, d\}$ there can be three unrooted quarnets, q_1 that contains a and b on one side and c and d on the other, denoted by $q_1 = ab|cd$, $q_2 = ac|bd$, and $q_3 = ad|bc$.

Given a set of estimated gene trees $G = \{G_1, G_2, \ldots, G_g\}$ for g loci, the number of gene trees that match with the each of the three quarnets is denoted $X = (X_{q_1}, X_{q_2}, X_{q_3})$. Considering X follows a multinomial distribution with probabilities $(CF_{q_1}, CF_{q_2}, CF_{q_3})$ (i.e., the expected CF for each quarnet) assuming unlinked loci. For a level-1 network with $n \geq 4$ taxa, the composite likelihood is computed using:

$$L = \prod_{s \in S} (CF_{q_1})^{X_{q_1}} (CF_{q_2})^{X_{q_2}} (CF_{q_3})^{X_{q_3}} \quad (1)$$

where S is the collection of all quarnets exctracted from the network.

To find the network topology that has the best fit of the observed data, the network space is searched using heuristics. Five 'moves' are implemented to traverse the networks and jump between different dimentions, which are (1) nearest-neighbor interchange (NNI), (2) addition (or deletion) of a reticulation, (3) change direction of the reticulation edge, and move (4) the target or (5) the origin of an existing hybridization edge. At each iteration, one of the randomly selected 'move' makes a slight modification to the original topology N_0 and propose the new topology N_1. SNaQ uses hill climbing algorithm to traverse the network space and to find an optimum. So if the composite likelihood of N_1 is greater than N_0, N_0 is discarded and N_1 becomes N_0; but the move does not result in N_1 with improved composite likelihood, N_1 is discarded and another move is applied to N_0. This process continues until an optima is reached.

Since hill climbing only guantees to find the local optimum rather than global optimum, SNaQ executes multiple independent runs, which is typical practice in phylogenetic heuristics. Furthermore, starting topology is modified at some probablility at the onset of the search for exhaustive search.

Improvements in SNaQ2

Parallelization of the pseudolikelihood calculation

Original SNaQ makes use of parallelization mechanism by allowing each run on different processors (or cores) using Julia package Distributed. SNaQ2 further improves the computational speed by multithreading the pseudolikelihood calculation. In particular, extraction of quartet topologies from a network, calculation of expected CFs of the extracted quartet, and computation of quarnet likelihood are parallelized. This setting allows to allocate all runs independently on seaprate high-performance computing nodess, with each node fully utilized to parallelize the pseudolikelihood calcuation for the run it is responsible for.

Sampling subset of quartets for pseudolikelihood

In the orginal SNaQ, all extracted quartets were used to compute pseudolikelihood of a network. While this computation is generally efficient, it may lead to the bottleneck as the number of taxa increases since there are *nchoose*4 quartets in a network. Several studies that also utilizes quartet has shown that subsampling some quartets lead to accurate estimation of a phylogeny (e.g., SVD Quartets). In SNaQ2, we added a new argument `propQuartets` in the network inference function `snaq!`. The argument `propQuartets` specifies the proportion of randomly sampled quartets for the pseudolikelihood calculation (i.e., nonnegative float equal to or smaller than 1).

Proposals using quartet weighting

All moves, except the move that changes direction of the selected reticulation edge, involves random selection of a tree edge in N_0 that will be modified. For example, to move the origin on an existing reticulation edge, a reticulation edge whose head is at a randomly selected reticulation node u is selected at prorbability of γ, followed by a *random* selection of a tree edge that will have a new reticulation node in the middle. This stochasticity can result in increased time requires to find the global optimum during the searching process.

We make an improvement in heuristics by selecting the edge via weighted random sampling where the weight is $\Delta CF = \sum_{i=1}^{3} |X_{q_i} - CF_{q_i}|$, calculated for every quartet extracted from a network. It can be done for a valid edges to choose from, although not implemented in the method. The rational here is... The new variable that is added to snaq! is probQR (varied from 0.0=full random to 1.0=full weighted).

Evaluation using simulated and empirical data

Simulation

We evaluate the performance of SNaQ2 using simulation. A set of species networks wher each network has $n = \{10, 20, 30\}$ tips with 1 or 3 reticulations when $n = 10$ and 1, 3, or 5 otherwise was generated. For each n, we first generated a species tree under a Yule process using R package phytools, then we sequentially added reticulation onto the topology at arbitrary position. We checked each network is level-1 considering that snaq! can only infer networks in this class, both manually and using R package SiPhyNetworks. For each species network, we set each branch length $= \{0.5, 1.0, 2.0\}$ colescent unit to represent high, medium, and low amount of incomplete lineage sorting.

Using Julia package PhyloCoalSimulations, we generate a set of $g \in \{300, 1000, 3000\}$ gene trees for each species network. For each gene tree, we generated multiple sequence alignment that is 10^3 bp long, setting the scale branch parameter=0.03 and base frequency of nucleotides as A=0.3, C=0.2, G=0.2, and T=0.3 under the HKY model. The generated sequence alignment was used to estimate a gene tree using IQ-TREE 1.6.12 with the best substituion model being identified withtin the sotftware with default parameters. Gene tree estimation error was measured using python package FastMulRFS.

The set of estimated gene trees were subsequently used as an input file for network estimation using SNaQ1 and SNaQ2. For SNaQ1, a randomly selected estimated gene tree was used as the starting topology, a table of CFs computed from the set of estimated gene trees were used as input, and the true number of reticulations were provided. For SNaQ2, all parameter setting was identical to SNaQ1 with additional parameters probQuartets$\in \{1.0, 0.9, 0.7\}$ and proqQR$\in \{0, 0.5, 1.0\}$. We recorded runtime for each network analysis. We evaluated the accuracy of the estimated network using hardwired cluster dissimilarity metric in Julia package PhyloNetworks. More specifically, we compared between the estimated network with the true network as well as the major trees of the estimated network and the true network.

All computations were done using condor at University of Wisconsin-Madison. We ran the analyses in different computing power setting the number of processors $\in \{4, 8, 16\}$ to compare the efficiency in various conditions. One hundred replicates were made.

Empirical data

Empirical data: 1. the fish data in snaq1; 2. find something else

Results and discussion

Simulation

GTEE

Empirical data

Conclusion

Competing interests

No competing interest is declared.

Author contributions statement

NK and SK designed and consucted analysis, prepared the manuscript. TC implemented improvements into SNaQ.

Acknowledgments

WID Server?

References

A. B. Bondi. Characteristics of scalability and their impact on performance. In *Proceedings of the 2nd International Workshop on Software and Performance*, pages 195–203, Ottawa Ontario Canada, Sept. 2000. ACM. ISBN 978-1-58113-195-6. doi: 10.1145/350391.350432.

H. A. Hejase and K. J. Liu. A scalability study of phylogenetic network inference methods using empirical datasets and simulations involving a single reticulation. *BMC Bioinformatics*, 17(1):422, Dec. 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-1277-1.

D. H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–267, Feb. 2006. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/msj030.

D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 1 edition, Dec. 2010. ISBN 978-0-521-75596-2 978-0-511-97407-6. doi: 10.1017/CBO9780511974076.

S. Kong, J. C. Pons, L. Kubatko, and K. Wicke. Classes of explicit phylogenetic networks and their biological and mathematical significance. *Journal of Mathematical Biology*, 84(6):47, May 2022a. ISSN 0303-6812, 1432-1416. doi: 10.1007/s00285-022-01746-y.

S. Kong, D. L. Swofford, and L. S. Kubatko. Inference of Phylogenetic Networks from Sequence Data using Composite Likelihood. Preprint, Evolutionary Biology, Nov. 2022b.

L. Liu, L. Yu, and S. V. Edwards. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evolutionary Biology*, 10(1):302, 2010. ISSN 1471-2148. doi: 10.1186/1471-2148-10-302.

C. Solís-Lemus and C. Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLOS Genetics*, 12(3):e1005896, Mar. 2016. ISSN 1553-7404. doi: 10.1371/journal.pgen.1005896.

C. Solís-Lemus, P. Bastide, and C. Ané. PhyloNetworks: A package for phylogenetic networks. *Molecular Biology and Evolution*, 34(12):3292–3298, Dec. 2017. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msx235.

C. Than, D. Ruths, and L. Nakhleh. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*, 9(1):322, Dec. 2008. ISSN 1471-2105. doi: 10.1186/1471-2105-9-322.

D. Wen, Y. Yu, J. Zhu, and L. Nakhleh. Inferring phylogenetic networks using phylonet. *Systematic Biology*, 67(4):735–740, July 2018. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syy015.

Y. Yu and L. Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16 (S10):S10, Dec. 2015. ISSN 1471-2164. doi: 10.1186/1471-2164-16-S10-S10.

J. Zhu and L. Nakhleh. Inference of species phylogenies from bi-allelic markers using pseudo-likelihood. *Bioinformatics*, 34 (13):i376–i385, July 2018. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/bty295.