Journal Title Here, 2022, pp. 1–4

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

PAPER

SNaQ2: Improved Scalability for Phylogenetic Network Inference

Sungsik Kong,[1] Nathan Kolbow,[2] Tyler Chafin[3] and Claudia Solís-Lemus[3,*]

[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

A phylogenetic network is an acyclic directed graph that generalizes the bifurcating phylogenetic tree by allowing nodes to have an indegree of two, thereby creating a reticulation structure. Phylogenetic networks represent complex biological scenarios that phylogenetic trees cannot, such as hybrid speciation, introgression, allopolyploid speciation, and more. However, network inference is computationally demanding and often lacks scalability. It is not uncommon to conduct analysis with only a handful of taxa, which narrows the scope of biological investigation. In this study, we present SNaQ2, a new version of SNaQ—a popular summary-based network inference method that uses concordance factors. In SNaQ2, computational efficiency is enhanced through (1) weighted random selection of quartets, (2) parallelization of the quartet likelihood calculation during composite likelihood computation, and (3) probabilistic decision-making during network search heuristics. In this talk, I will first briefly introduce the essence of the original SNaQ, then describe the key improvements made in SNaQ2 in more detail. I will also present the results of our benchmarks that compare the performance of the two versions of SNaQ, along with the application of the new version to empirical datasets.

Key words: Composite likelihood, Parallelization, Phylogenetic Networks, Scalability

Introduction

Phylogenetic network is an acyclic directed graph that generalizes the bifurcating phylogenetic tree by allowing some nodes to have indegree of two and create 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, their wide use in practice is hindered by their lack of scalability (i.e., 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 is am NP-Hard problem (non-deterministic polynomial-time). Some attempts to ameliorate this issue has been made but the computational requirement is still exessively high even 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 (Species Network applying Quartets) [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 calculation 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 the 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 (1) weighted random selection of quartets, (2) parallelization of the quartet likelihood calculation during composite likelihood computation, and (3) probabilistic decision-making during network search 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

While details of SNaQ is available in Solís-Lemus and Ané [2016], we make a brief description here for readers to clearly see the improvemetns in SNaQ2 in the following subsection. To quantify the fit of the data on a network topolgy, SNaQ first extracts unrooted quarnets (i.e., networks with four tips) from the full network, and computes the expected concordance factor (CF) for each quarnet. Note CF represents the proportion of genes whose true relationship is the quartet under the coalescent model [Baum, 2007]. For a taxon set $X \in \{a, b, c, d\}$ has three possible ways to cluster four taxa into two groups of two (i.e., separated by a split [Chifman and Kubatko, 2014]). Thus, there are 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 $X = (X_{q_1}, X_{q_2}, X_{q_3})$. Assuming each loci is unlinked, X follows a multinomial distribution of the expected CF for each quarnet with probabilities $(CF_{q_1}, CF_{q_2}, CF_{q_3})$. For a level-1 network with $n \geq 4$ taxa, the composite likelihood of a network is computed using:

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

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

SNaQ heuristically searches the 'best' network topology using hill climbing. Five topological 'moves' to traverse the networks space and jump between different dimentions used are (i) nearest-neighbor interchange (NNI), (ii) addition of a reticulation, (iii) change direction of the reticulation edge, and move (iv) the target or (v) the origin of an existing hybridization edge. In brief, the search begins with the original topology N_0 that has the composite likelihood of L_0, and a new topology N_1 with the composite likelihood L_1 is proposed by applying randomly selected one of above five moves on N_0. If $L_0 < L_1$, N_0 is discarded and N_1 becomes N_0. Otherwise, N_1 is discarded and another move is applied to N_0. This process continues until an optima is reached. Typically, SNaQ executes multiple independent runs (i.e., searches).

Improvements in SNaQ2

Parallelization of the composite likelihood calculation

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

Sampling subset of quartets for composite likelihood

In SNaQ, all quartets extracted from a network were used to compute composite likelihood. While this computation is generally efficient, it may lead to the bottleneck as the number of taxa increases, since there are $\binom{n}{4}$ quartets in a network. In SNaQ2, a new argument `propQuartets`, which specifies the proportion of sampled quarnets from the full set of quarnets extracted for the composite likelihood calculation, is available in the main function `snaq!`. The value of `propQuartets` must be non-negative float ≤ 1. SNaQ2 currently subseamples quartets in a randomized manner, which is a common subsampling approach (e.g., SVDQuartets [Chifman and Kubatko, 2014, 2015]) in phylogenetic inference, although some studies show weighted subsampling can improve accuracy (e.g., ASTRAL [Zhang and Mirarab, 2022]).

Proposals using quartet weighting

All topological moves, except 'change direction of the reticulation edge', involve random selection of a tree edge in N_0 that acts as the point of moficiation. For example, to perform the topological move 'move the origin of an existing hybridization edge', a reticulation edge whose head is at a randomly selected reticulation node u is selected at the probability of γ (i.e., the inhertiance probability assigned to each reticulation edge), followed by a *random* selection of a tree edge that will have a new node u' in the middle. Then, the head of the selected reticulation edge becomes u' with γ and the original incoming branch becomes the other reticulation edge with $(1 - \gamma)$. All nodes with degree two are removed.

This stochasticity produced by a random selection of a tree branch to make a move can result in increased computational time to find a optima during the search 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. The rationale is that if an edge (or split) occurs in the frequency that deviates from the expectation, that edge is unlikely to present in the true network topology. A new argument `probQR` is added in the function `snaq!`, which must be non-negative float ≤ 1 that defines the probability of an edge chosen based on the weight (i.e, 0.0=full random and 1.0=full weighted).

Evaluation using simulated and empirical data

Simulation

We evaluate the performance of SNaQ2 using simulation. A set of species networks that has $n = \{10, 20, 30\}$ tips with h reticulations where $h = \{1, 3\}$ when $n = 10$, and $h = \{1, 3, 5\}$ otherwise are manually generated. More specifically, we generated a species tree under a Yule process using R package phytools [Revell, 2012] for each n, then we sequentially added reticulation onto the topology at arbitrary position. We checked each network is level-1, both manually and using R package SiPhyNetworks [Justison et al., 2023], considering that both versions of SNaQ assume the true networks belongs to the class of level-1 networks. For each species network, we set every branch length 0.5, 1.0, or 2.0 colescent unit to represent high, medium, and low amount of incomplete lineage sorting.

Using Julia package PhyloCoalSimulations [Fogg et al., 2023], we generate a set of $g \in \{300, 1000, 3000\}$ gene trees for each species network. For each gene tree, we generate multiple sequence alignment that is 10^3 bp long using ms [Hudson, 2002], setting the branch length scale parameter to 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 sequence alignment is then used as an input file to estimate gene tree using IQ-TREE 1.6.12 [Nguyen et al., 2015] with the best substituion model being identified withtin the softtware with default parameters.

Gene tree estimation error is measured using Python package FastMulRFS [Molloy and Warnow, 2020].

A table of CFs computed from the set of estimated gene trees are used for network estimation using SNaQ and SNaQ2. The starting topology is randomly selected among the estimated gene trees and the true h is specified. All parameters were set default and identically in both versions of SNaQ, but we additionally specified `probQuartets` $\in \{1.0, 0.9, 0.7\}$ and `proqQR` $\in \{0, 0.5, 1.0\}$ for SNaQ2. We recorded runtime for each network analysis. We computed the hardwired cluster dissimilarity metric between the estimated network with the true network as well as the major trees of the estimated network and the true network using Julia package PhyloNetworks. All computations are conducted using Condor at University of Wisconsin-Madison. The analyses are executed using various number of processors $\in \{4, 8, 16\}$ to compare the efficiency in different computing power. One hundred replicates are made.

Empirical data

Using SNaQ2, we reanalyze the table of CFs used in Solís-Lemus and Ané [2016] that is obtained from the transcriptome data in Cui et al. [2013] to reconstruct the evolutionary history of 24 swordtails and playfishes (*Xiphophorus*: Poeciliidae). We set `probQuartets`=0.7, `proqQR`=1.0, and $h = 2$ (as specified in the Solís-Lemus and Ané [2016]). We used 16 processors.

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

D. A. Baum. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *TAXON*, 56(2):417–426, May 2007. ISSN 00400262. doi: 10.1002/tax.562013.

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.

J. Chifman and L. Kubatko. Quartet inference from SNP data under the coalescent model. *Bioinformatics*, 30(23):3317–3324, Dec. 2014. ISSN 1460-2059, 1367-4803. doi: 10.1093/bioinformatics/btu530.

J. Chifman and L. Kubatko. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. *Journal of Theoretical Biology*, 374:35–47, June 2015. ISSN 00225193. doi: 10.1016/j.jtbi.2015.03.006.

R. Cui, M. Schumer, K. Kruesi, R. Walter, P. Andolfatto, and G. G. Rosenthal. PHYLOGENOMICS REVEALS EXTENSIVE RETICULATE EVOLUTION IN *XIPHOPHORUS* FISHES: PHYLOGENOMICS OF *XIPHOPHORUS* FISHES. *Evolution*, 67(8):2166–2179, Aug. 2013. ISSN 00143820. doi: 10.1111/evo.12099.

J. Fogg, E. S. Allman, and C. Ané. PhyloCoalSimulations: A Simulator for Network Multispecies Coalescent Models, Including a New Extension for the Inheritance of Gene Flow. *Systematic Biology*, 72(5):1171–1179, Nov. 2023. ISSN 1063-5157, 1076-836X. doi: 10.1093/sysbio/syad030.

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.

R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, Feb. 2002. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/18.2.337.

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.

J. A. Justison, C. Solis-Lemus, and T. A. Heath. SiPhyNetwork : An R package for simulating phylogenetic networks. *Methods in Ecology and Evolution*, 14(7):1687–1698, July 2023. ISSN 2041-210X, 2041-210X. doi: 10.1111/2041-210X.14116.

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.

E. K. Molloy and T. Warnow. FastMulRFS: Fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics*, 36(Supplement_1):i57–i65, July 2020. ISSN 1367-4803, 1367-4811. doi: 10.1093/bioinformatics/btaa444.

L.-T. Nguyen, H. A. Schmidt, A. von Haeseler, and B. Q. Minh. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, 32(1):268–274, Jan. 2015. ISSN 1537-1719, 0737-4038. doi: 10.1093/molbev/msu300.

L. J. Revell. Phytools: An R package for phylogenetic comparative biology (and other things). *Methods in Ecology* *and Evolution*, 3(2):217–223, Apr. 2012. ISSN 2041-210X, 2041-210X. doi: 10.1111/j.2041-210X.2011.00169.x.

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.

C. Zhang and S. Mirarab. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Molecular Biology and Evolution*, 39(12):msac215, Dec. 2022. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msac215.

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.