

ULTRAFast LEARNING OF 4-NODE HYBRIDIZATION CYCLES IN PHYLOGENETIC NETWORKS USING ALGEBRAIC INVARIANTS

Zhaoxing Wu

Wisconsin Institute for Discovery
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706

Claudia Solís-Lemus*

Wisconsin Institute for Discovery
Department of Plant Pathology
University of Wisconsin-Madison
Madison, WI 53706

Abstract

Motivation: The abundance of gene flow in the Tree of Life challenges the notion that evolution can be represented with a fully bifurcating process, as this process cannot capture important biological realities like hybridization, introgression, or horizontal gene transfer. Coalescent-based network methods are increasingly popular, yet not scalable for big data, because they need to perform a heuristic search in the space of networks as well as numerical optimization that can be NP-hard.

Results: Here, we introduce a novel method to reconstruct phylogenetic networks based on algebraic invariants. While there is a long tradition of using algebraic invariants in phylogenetics, our work is the first to define phylogenetic invariants on concordance factors (frequencies of 4-taxon splits in the input gene trees) to identify level-1 phylogenetic networks under the multispecies coalescent model. Our novel inference methodology is optimization-free as it only requires the evaluation of polynomial equations, and as such, it bypasses the traversal of network space, yielding a computational speed at least 10 times faster than the fastest-to-date network methods. We illustrate the accuracy and speed of our new method on a variety of simulated scenarios as well as in the estimation of a phylogenetic network for the genus *Canis*.

Availability and Implementation: We implement our novel theory on an open-source publicly available Julia package `PhyloDiamond.jl` available at <https://github.com/solislemuslab/PhyloDiamond.jl> with broad applicability within the evolutionary biology community.

Contact: solislemus@wisc.edu

Keywords Hybridization · Polynomials · Reticulate evolution · Coalescent model

1 Introduction

The Tree of Life is the graphical structure that represents the evolutionary process from single-cell organisms at the origin of life to present day biodiversity. Mathematically, a phylogenetic tree is a fully bifurcating graph in which its internal nodes represent speciation events that give rise to two children nodes. Recent evidence [13, 1, 33, 35, 44] has challenged the notion that evolution across the Tree of Life can be represented with a fully bifurcating process, as this process cannot capture important biological realities like hybridization, introgression, or horizontal gene transfer that require two fully separated branches to join again. These processes of reticulate evolution are more prevalent in certain groups like plants [11], fungi [43] and prokaryotes [15]. To accurately include these groups in the Tree of Life, recent years have seen an explosion of methods to reconstruct phylogenetic networks, which naturally account for reticulate evolution [14, 16, 7, 27]. Existing methods, however, are still not scalable enough to tackle the complexities of the present day's biological big data. The most scalable alternatives infer split (or implicit) networks [8, 26, 23] which are not biologically interpretable as internal nodes no longer represent biological events (speciation or hybridization), and the resulting network is completely unrooted, even with the inclusion of an outgroup.

Among methods to infer explicit networks (those whose internal nodes do represent speciation or hybridization events), likelihood-based approaches are among the most popular ones. `PhyloNet` [47] has been the pioneer of likelihood inference of phylogenetic networks, later expanding to Bayesian [46] and pseudolikelihood alternatives [48]. However, even in its most scalable option (pseudolikelihood), the method is still not suitable beyond dozens of taxa and hundreds

*Corresponding author: solislemus@wisc.edu

of genes. SNaQ [38] within the PhyloNetworks Julia package [40] has proven a scalable alternative to infer large phylogenetic networks from multilocus alignments. This method is based on a pseudolikelihood model under the coalescent, and its running time does not increase as a function of number of genes because the input data for SNaQ are split frequencies on subsets of 4 taxa (concordance factors). That is, after the estimation of the concordance factors, the running time is the same for data with 10 genes or 10,000 genes. Despite its popularity, SNaQ still lacks scalability as the number of taxa increases and remains a suitable alternative only for 50 or fewer taxa. Last, BEAST2 [49] provides the only co-estimation method that infers simultaneously the phylogenetic networks and the gene trees, and it allows users to infer a variety of relevant biological parameters (such as divergence times), yet the complexity of this method renders it unsuitable for anything beyond a handful of taxa.

Because of the complexity in inferring a phylogenetic network, the evolutionary biology community has embraced the use of hybrid detection methods such as ABBA-BABA test [32], MSCquartets [30, 36], and HyDe [28] to identify hybridization events on a fixed phylogeny. These methods take in subsets of taxa and test whether the current subset can be well-explained with a tree-like pattern or if a hybridization event exists in this subset. While fast, these methods can be inaccurate in the presence of multiple hybridization events affecting the same taxa [6] or by ghost lineages [31, 45, 6]. Furthermore, the process of first reconstructing a phylogenetic tree and then adding hybridization events can be flawed given the bias that gene flow causes on the inference of the backbone phylogenetic tree [39].

Here, we introduce a novel method to infer hybridization events using phylogenetic invariants. Our method only requires evaluation of polynomial equations, so it bypasses optimization in the space of networks and it is at least 10 times faster than current inference network methodologies. Using gene trees as input, our method exploits the signal in the frequencies of splits in the data, also well-known as *concordance factors*. Indeed, under the coalescent model, splits of taxa display certain probabilities of appearing in the sample of gene trees. Under the true network, these frequencies need to satisfy certain polynomial equations, denoted *phylogenetic invariants*. By plugging in the observed frequencies on the phylogenetic invariants of every possible placement of the hybridization cycle, we can identify the hybridization cycle that better agrees with the data as the one whose evaluated invariants are closest to zero.

Algebraic invariants have been widely used in phylogenetics to identify trees under a variety of models of evolution (Jukes-Cantor [17, 42], GTR [2, 18, 9, 10]), and more recently, to identify phylogenetic networks [22, 21, 3, 12, 4]. Nevertheless, existing work on phylogenetic invariants on networks has been restricted to models of evolution. Our work is the first to define phylogenetic invariants on concordance factors under the multispecies coalescent model on networks.

This work first introduces the phylogenetic invariants that concordance factors need to satisfy under level-1 phylogenetic networks with a 4-node hybridization cycle under the multispecies coalescent model. Next, we describe an inference methodology to use the phylogenetic invariants to infer the correct 4-node hybridization cycle among n taxa. Moreover, we demonstrate the performance of our method with simulated data, and we revisit the inference of the *Canis* phylogenetic network that had already been published in [19]. We show that our method is faster than all existing network inference methods, and it accurately identifies the correct hybridization cycle in the *Canis* genus. Additionally, in all simulating scenarios, our method recognizes the correct placement of hybridization events with high accuracy. Even in the few cases when our method identifies an incorrect network (usually cases when there is not sufficient sampling of taxa that descend from the hybrid node), the correct network is still ranked among top 5 optimal networks given the invariant score (evaluated phylogenetic invariants that need to vanish to zero). This means that, regardless of sampling, our method is always able to reduce the space of candidate networks that can later be tested with a likelihood-based approach.

We highlight that our method has two main limitations: 1) it is only suitable to identify level-1 hybridization cycles with 4 nodes, and 2) it sometimes fails to identify the correct network as the one with the top 1 ranked invariant score when there is only one taxon sampled that is descendant of the hybrid node. Despite these limitations, our method is a valuable innovation in the landscape of phylogenetic network methods, if anything, to reduce the space of candidate networks to be used in likelihood-based methodologies. Indeed, we do not consider our method as a replacement for other network inference approaches. On the contrary, we believe that our ultrafast learning methodology can serve to 1) identify the true phylogenetic network when it involves a simple hybridization event, and 2) provide several candidate networks to be later tested with a likelihood-based approach, effectively bypassing heuristic optimization in the space of networks. Furthermore, our method is open source, publicly available as the new Julia package `PhyloDiamond.jl` in <https://github.com/solislemuslab/PhyloDiamond.jl>.

The structure of the paper is as follows. In Materials and Methods, we introduce the phylogenetic invariants under the multispecies coalescent model on networks, and we describe an inference methodology that uses the phylogenetic invariants to identify the 4-node hybridization cycle that generates the data. In this section, we also depict the simulation study and the re-analysis of the *Canis* data from [19]. In Results, we present the results on the simulated data and the *Canis* phylogeny. Last, in Discussion, we explain the main limitations of our method and potential future work.

2 Materials and Methods

2.1 Phylogenetic invariants for 4-node hybridization cycles in level-1 phylogenetic networks

Under the coalescent model, the distribution of gene trees estimated from multilocus sequence alignments provides information on the true network that generated the data [47]. In [38], it was shown that the split frequencies on subsets of 4 taxa, namely the *concordance factors* (CF), also provide information on the true network. That is, a CF of a given quartet (or split) is the proportion of genes whose true tree displays that quartet (or split) [5]. For example, for a taxon set $s = \{a, b, c, d\}$, there are only three possible quartets, represented by the splits $q_1 = ab|cd$, $q_2 = ac|bd$ and $q_3 = ad|bc$. The CF for the split $ab|cd$ is the proportion of gene trees that display this split.

As in [38], our method uses the CFs as input data, and we focus on the case of 4-node hybridization cycles on level-1 phylogenetic networks as shown in Figure 1 (top). A level-1 phylogenetic network is a network with no vertex belonging to more than one hybridization cycle. The semi-directed network in Figure 1 (top) is unrooted, yet the direction of the hybrid edges (in blue) is known so that the root placement of the network is constrained. Namely, the root cannot be anywhere below the hybrid node (in blue). The hybridization cycle partitions the taxa into 4 subsets with n_0, n_1, n_2 and n_3 as the number of taxa in each subset. For example, the clade below the hybrid node in blue in Figure 1 (top), also known as the hybrid clade, has n_0 taxa, and the two sister clades to the hybrid clade have n_1 and n_2 taxa. The notation n_i not only represents the number of taxa in a specific clade, this terminology is also used to refer to that specific clade. In fact, we refer to the hybrid clade as the n_0 clade. This n_i notation will later help us represent different network structures and 4-taxon subsets. Moreover, each clade in the network could be a subtree or a subnetwork as long as the overall network is level-1 [25]. Last, we highlight that by 'taxa', we mean either different species or populations, or multiple individuals sampled from the same species.

In this work, we introduce a method to infer 4-node hybridization cycles in level-1 semi-directed networks. Since we focus on level-1 4-node hybridization cycles (as in Figure 1 (top)), we utilize a simplifying notation to represent each network. Let N represent an n -taxon semi-directed level-1 phylogenetic network, and we represent this network by the number of taxa in its four clades. For example, in Figure 1, the network denoted as N2222 has two taxa in each of the four clades, while the network N2112 contains two taxa in clades n_0 and n_3 , and one taxon in clades n_1 and n_2 .

Another relevant notation is the vector representation of every 4-taxon subset. For example, in Figure 1, the 4-taxon subset $(0, 2, 0, 2)$ drawn from network N2222 corresponds to the case when two taxa are taken from n_1 and n_3 and no taxon is taken from n_0 and n_2 , and this subset corresponds to the quartet $AB|CD$. Another example is the 4-taxon subset $(0, 1, 1, 2)$ drawn from network N2112. Note that since this network only has one taxon in clades n_1 and n_2 , any 4-taxon subset drawn from this network must have at most one taxon in the second and third value in the 4-taxon vector notation. In general, a 4-taxon subset is represented by a 4-dimensional vector so that each element in this vector corresponds to the number of taxa drawn from each of the four clades n_0, n_1, n_2, n_3 .

Next, we describe the formulas for the expected CFs under the coalescent model for every 4-taxon subset on a level-1 semi-directed network with branch lengths in coalescent units (t_i) and inheritance probability (γ). These formulas have already been derived in [38]. Let N be the network on Figure 1 (top), and let $(0, 0, 2, 2)$ be the 4-taxon subset (see Figure 1 (bottom right) for an example) for which we want to compute the CF formulas under the coalescent model. Let taxa $k_1, k_2 \in n_2$ and taxa $l_1, l_2 \in n_3$. The probability of taxa k_1 and k_2 coalescing in a branch with length $t_2 + t_{23} + t_3$ is given by $1 - \frac{2}{3}z_2z_{23}z_3$ where $z_i = \exp(-t_i)$. Then, the formulas for the three CF under the coalescent model are defined as:

$$\begin{aligned} P(k_1, k_2 | l_1, l_2) &= 1 - \frac{2}{3}z_2z_{23}z_3 = a_1 \\ P(k_1, l_1 | k_2, l_2) &= \frac{1}{3}z_2z_{23}z_3 = a_2 \\ P(k_1, l_2 | k_2, l_1) &= \frac{1}{3}z_2z_{23}z_3 = a_3 \end{aligned}$$

where a_1, a_2 and a_3 denote the values for the true CFs corresponding to the splits $k_1, k_2 | l_1, l_2$, $k_1, l_1 | k_2, l_2$ and $k_1, l_2 | k_2, l_1$ respectively. Thus, for the 4-taxon subset $(0, 0, 2, 2)$, there are three polynomial equations that represent the probabilities for each quartet under the coalescent model. The true CF values (a_1, a_2, a_3) derived from these three formulas will later be used by our algorithm as the observed CF estimated from the sample of gene trees.

In Supplementary Material, we list all of the CF formulas for all 4-taxon subsets on a level-1 semi-directed network with one 4-node hybridization cycle (Figure 1 (top)). Because of the theoretical work in [38, 41], we know that we only need two taxa per clade to represent all CF formulas involving the hybridization cycle, and thus, we can restrict to the case of $n = 8$ taxa for the definition of all CF formulas (yet our method is not restricted to 8 taxa, see Inference

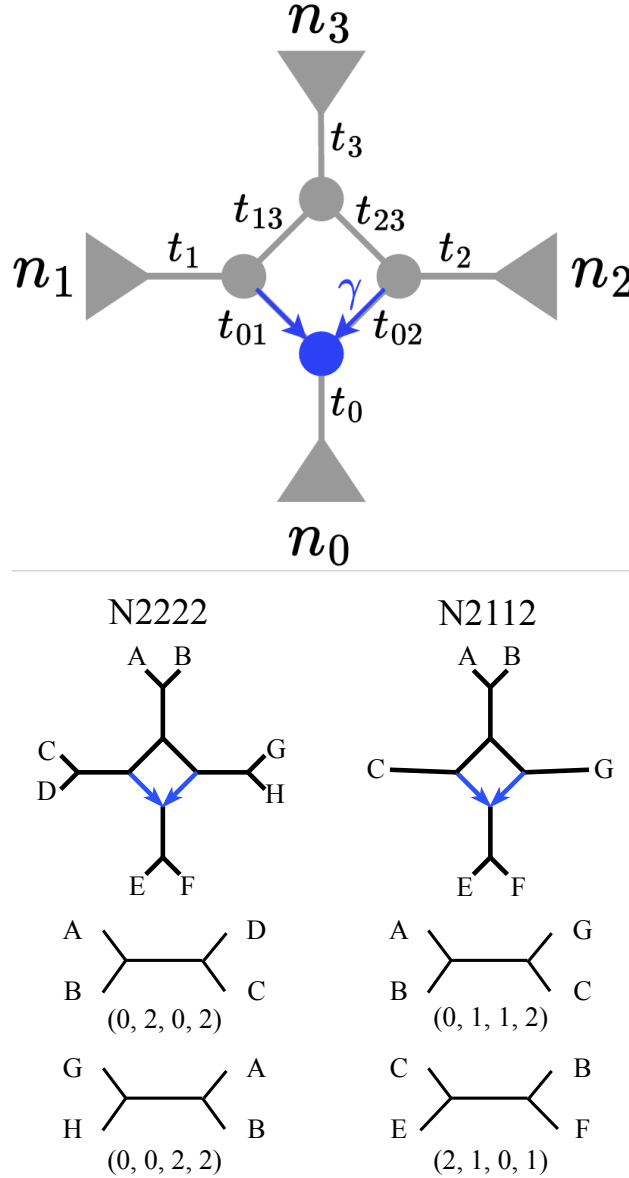


Figure 1: **Top:** 4-node hybridization cycle on a level-1 phylogenetic network where t_i represents the branch length in coalescent units. The hybrid node and hybrid edges are in blue. The minor hybrid edge is labelled with γ that represents the inheritance probability (or proportion of genes transferred through the minor edge). Denote this level-1 semi-directed network with 4-node hybridization cycles as N . **Middle:** Two examples of level-1 networks, each with one 4-node hybridization cycle, and their corresponding notation (N2222 for the 8-taxon network and N2112 for the 6-taxon network). **Bottom:** Examples of quartets drawn from each network in top right along with their vector notations. For example, the quartet $AB|CD$ drawn from the N2222 network corresponds to the 4-taxon subset $(0, 2, 0, 2)$ where each entry in this vector corresponds to the number of taxa drawn from each of the four clades: n_0, n_1, n_2, n_3 .

of n -taxon phylogenetic networks with one 4-node hybridization cycle using phylogenetic invariants). Furthermore, not all $\binom{8}{4} = 70$ 4-taxon subsets are required. Indeed, we ignore 4-taxon subsets of the form $(3, 1, 0, 0)$ since three or more taxa from one of the clades (in this case, 3 taxa from clade n_0) do not provide information about the hybridization cycle. Recall that only the length of internal branches appears in the CF formulas but not the length of external branches. Therefore, if there are three or more taxa on one clade (e.g. 3 taxa in clade n_0), then branches involved in the hybridization cycle will correspond to an external branch and thus, will not be included in any of the CF formulas. The same is true if all four taxa in the 4-taxon subset all come from the same clade. Thus, only 4-taxon subsets where at most

2 taxa come from a given clade contain information about the hybridization cycle. There are 19 such 4-taxon subsets in which $0 \leq i, j, k, l \leq 2$ for (i, j, k, l) , and thus, there are $19 \times 3 = 57$ concordance factor polynomial equations along with 57 true CF values (a_i).

As just described, the semi-directed network N in Figure 1 (top) can be represented by a set of 57 polynomial equations, denoted $CF(N)$ in 66 variables: 9 variables corresponding to branch lengths and inheritance probability ($z_0, z_1, z_2, z_3, z_{01}, z_{02}, z_{13}, z_{23}, \gamma$ for $z_i = \exp(-t_i)$) and 57 variables corresponding to the true CF values (a_i for $i = 1, \dots, 57$). Here, we identify the relationships that the a_i values need to satisfy for the polynomial system $CF(N)$ to be consistent. These relationships are denoted *phylogenetic invariants*. For example, because the three concordance factors corresponding to a given 4-taxon subset (say $(0, 0, 2, 2)$ as described above) need to sum up to one, we have one phylogenetic invariant for these three numbers: $a_1 + a_2 + a_3 = 1$, or similarly $i_1(a_1, a_2, a_3) = a_1 + a_2 + a_3 - 1$ with $i_1(a_1, a_2, a_3) = 0$. In addition, since the two minor concordance factors are equal, we have a second phylogenetic invariant defined for these numbers: $a_2 = a_3$, or similarly $i_2(a_2, a_3) = a_2 - a_3$ with $i_2(a_2, a_3) = 0$. It turns out that the set of all concordance factor values for all 4-taxon subsets (the 57 polynomial equations denoted $CF(N)$) define a set of phylogenetic invariants (equations only in the a_i variables) that need to vanish to zero whenever the concordance factors come from the true network. In this way, we can evaluate whether the given CF values come from the true network or not without knowing (t_i) and (γ) .

Note that not all networks will have the same number of CF polynomial equations. For example, for the 8-taxon network $N = 2222$, all 19 4-taxon subsets can be extracted, and thus, all 57 CF equations are defined in $CF(N)$. For the 5-taxon network $N = 1112$, the 4-taxon subset $(2, 1, 1, 0)$ cannot be considered since this subset requires two taxa from the clade n_0 and the 5-taxon network only has one taxon in this clade. Therefore, the 5-taxon network $N = 1112$ defines a set of fewer polynomial equations (12 to be exact) in $CF(N)$ involving only the CF values: $a_7, a_8, a_9, a_{22}, a_{23}, a_{24}, a_{28}, a_{29}, a_{30}, a_{31}, a_{32}$ and a_{33} .

To obtain the phylogenetic invariants defined by N , denoted $\mathcal{I}(N)$, we get the Gröbner basis of $CF(N)$ on the a_i variables using any elimination method in Macaulay2 [20]. All Macaulay2 scripts (and output files) are publicly available in <https://github.com/solislemuslab/PhyloDiamond.jl>.

For example, for the network $N = 1112$, there are 10 phylogenetic invariants in the Gröbner basis of $CF(N)$ on the a_i variables:

1. $a_{32} - a_{33}$
2. $a_{31} + 2a_{33} - 1$
3. $a_{28} + a_{29} + a_{30} - 1$
4. $a_{23} - a_{24}$
5. $a_{22} + 2a_{24} - 1$
6. $a_8 - a_9$
7. $a_7 + 2a_9 - 1$
8. $3a_9 * a_{30} + a_9 - a_{24} - a_{33}$
9. $a_{24} * a_{29} + 2a_{24} * a_{30} + a_{29} * a_{33} - a_{30} * a_{33} - a_{33}$
10. $3a_9 * a_{29} - 2a_9 + 2a_{24} - a_{33}$

We consider the first 7 invariants as the “trivial” invariants related to the sum-to-one property and the equality of the minor CFs. The last three invariants, on the other hand, identify relationships that the true CFs need to satisfy if they originated from the $N = 1112$ network. We note that the phylogenetic invariants depend on CF values a_i but neither on branch lengths t_i nor on inheritance probabilities γ . This property is desirable because when inferring a network, CFs values a_i can be estimated from gene trees or genetic sequences, but both t and γ are unknown parameters. All the phylogenetic invariants corresponding to each n -taxon network ($5 \leq n \leq 8$) can be found in the Supplementary Material.

2.2 Inference of n-taxon phylogenetic networks with one 4-node hybridization cycle using phylogenetic invariants

The procedure to infer a phylogenetic network with one 4-node hybridization cycle using phylogenetic invariants starts with a table of estimated CFs. As shown on Figure 2 (left), the CF table contains rows representing different 4-taxon subsets and three columns labeled by grey-scale circles denoting three types of bipartitions. Each value in this table is mapped to the vector of $[a_1 \dots a_{57}]$ and this vector is plugged in the invariants for each candidate network $\mathcal{I}(N)$.

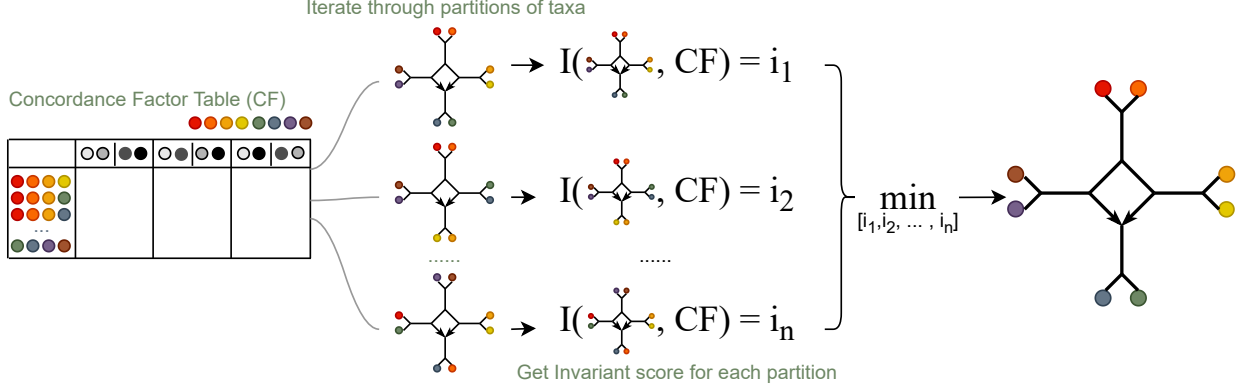


Figure 2: Graphical abstract of the methodology using phylogenetic invariants to identify the best placement of the 4-node hybridization cycle to fit the table of observed concordance factors. In this case, there are 8 taxa shown as colored circles. The invariant score is defined as the L_2 norm of the vector of evaluated invariants.

Here, all candidate networks are all possible partitions of given taxa. For example, as mentioned, the candidate network $N = 1112$ has 10 phylogenetic invariants ($|\mathcal{I}(N)| = 10$), and thus, after taking in the observed CFs, the invariants output a 10-dimensional vector. The *invariant score* is defined as the L_2 norm of the vector of evaluated invariants and the candidate network with the smallest invariant score is identified as the network that agrees best with the observed CFs. See Figure 2 for a graphical abstract of the procedure and Algorithm 1 for a more detailed description of the steps.

Algorithm 1: Inference of a 4-node hybridization cycle in an n -taxon phylogenetic network ($5 \leq n \leq 8$) with phylogenetic invariants

Input : Table of estimated concordance factors; optional: the number of optimal networks (m) to return (default $m = 5$)

Output : Top m optimal networks with smallest invariant score

scores \leftarrow an empty array;

nets \leftarrow an empty array;

for N_i , a possible partition of taxa **do**

append(nets, N_i);
map observed CF values to a_i values;
score $\leftarrow L2norm(\mathcal{I}(N_i))$;
append(scores, score);

rank(scores, nets);

return scores[1:m], nets[1:m];

Given that we only have phylogenetic invariants for the cases of $5 \leq n \leq 8$ taxa, if the true network has $5 \leq n \leq 8$ taxa, the procedure is just as described in Figure 2 and Algorithm 1. If the true network has more than 8 taxa, then the procedure is slightly different. In this case, our method needs to iterate through all possible subsets of 8 taxa, and fit the phylogenetic invariants on each subset to identify the top subnetwork with $n = 8$ taxa. Let N_8^* be the best 8-taxon network to fit the observed CFs. The missing taxa are added to N_8^* according to their placement in the second best network, or if the missing taxa are not in the second best network, the algorithm adds them according to their placement in the third best or the fourth best and so on. For example, assume that we have $n = 9$ taxa (A,B,C,D,E,F,G,H,I) and the best network with $n = 8$ taxa (N_8^*) contains A,B,C,D,E,F,G,H in the following partition: A,B in n_0 ; C,D in n_1 ; E,F in n_2 , and G,H in n_3 . We then find the best network containing taxon I. Assume that this network puts taxon I in the clade n_2 , then the resulting best network would have E,F, and I in n_2 . Note that our algorithm is unable to resolve the bifurcation structure within a given clade (e.g. the n_2 clade here with E,F, and I), but it is able to 1) correctly identify the partition of taxa among the four clades defined by the 4-node hybridization cycle and 2) correctly identify those are in the hybrid clade (n_0), those are in the sister clades to the hybrid clade (n_1 and n_2), and those are in the remaining fourth clade (n_3). See Algorithm in the Appendix for the steps in the procedure for $n > 8$ taxa.

2.3 Simulation study

2.3.1 Networks with eight or fewer taxa

In this section, we focus on the case of six taxa ($N = 2211, 2121, 2112, 1122, 1212, 1221$), seven taxa ($N = 2221, 2212, 2122, 1222$) and eight taxa ($N = 2222$). Next subsection describes the simulations with more than eight taxa.

For a given network N , we generate data under four settings: 1) true concordance factors (proof of concept); 2) true concordance factors perturbed by Gaussian error; 3) simulated gene trees, and 4) estimated gene trees. To generate the true concordance factors on a given network in the first scenario, we use the Julia package `PhyloNetworks.jl` [40], and this scenario is designed to demonstrate that the invariants method works under ideal conditions. For the second setting, we add Gaussian noise to the true concordance factors with zero mean and standard deviation of $\sigma = 0.0005, 0.00005, 0.000005$. We choose Gaussian noise as is standard practice. In the third setting, we use `ms-converter` that runs `ms` [24] to simulate $k = 100, 1000, 10000$ gene trees under the coalescent model on a given network N . All internal branches in the network are set to 1.0 coalescent unit so that the amount of incomplete lineage sorting is controlled, but not trivial, and we set an inheritance probability parameter of $\gamma = 0.3$. For the fourth setting, we simulate sequences of length $L = 500, 2000$ bp on each simulated gene tree with `seq-gen` [34] under the HKY model, scale the branch lengths by 0.036, and set nucleotide frequencies as 0.300414, 0.191363, 0.196748, and 0.311475. Then, we estimate gene trees with `IQ-Tree` [29] with `ModelFinder Plus` for model selection [37]. We repeat each simulation setting 30 times. All simulation scripts are in the GitHub repository <https://github.com/solislemuslab/PhyloDiamond.jl>.

2.3.2 Networks with more than 8 taxa

We also analyze cases of nine taxa ($N = 2223, 2232, 2322, 3222$) and ten taxa ($N = 3322, 3232, 3223, 2233, 2323, 2332$). Given that the algorithm to infer a network with more than 8 taxa relies entirely on the algorithm on 8 taxa, we focus on extensive simulations for the algorithm on 8 taxa in the previous section, and simply show proof-of-concept simulations for the case of more than 8 taxa in this section. In particular, we focus on testing the following scenarios: 1) true concordance factors, and 2) true concordance factors perturbed by Gaussian error as described above ($\sigma = 0.0005$).

2.4 Reticulate evolution in the Genus *Canis*

We further test our method on the *Canis* dataset from [19]. The original genomic dataset contained 12 gray wolves, 14 dogs, five coyotes, one Ethiopian wolf, three golden jackals, six African golden wolves, two dholes, four African hunting dogs, and one Andean fox. Given the widespread gene flow reported in the original study [19], we need to subsample the taxa so that there is one 4-node hybridization event among the selected taxa: one African hunting dog, one coyote, one dhole, one dog, one golden jackal, and one grey wolf. We use the estimated gene trees from [50] available in https://github.com/chaoszhang/Weighted-ASTRAL_data to infer the concordance factors with the Julia package `PhyloNetworks` [40]. We note that it is not necessary to subsample the taxa to use our method. The algorithm described for more than 8 taxa is capable of identifying the 4-node hybridization cycle for any number of taxa. However, as mentioned, its limitation is that we are unable to resolved the relationships within the four clades. In this case, we decided to subsample the species in order to obtain a fully resolved phylogeny.

To better evaluate our method's performance, we also run other methods on this dataset for comparison. In addition to our method, we also run `SNaQ` [38] on $h = 1$ hybridization event, 10 independent runs, and using one randomly selected gene tree as the starting tree. We use all 449,450 gene trees to infer the table of CFs and then map the allele names to species names. We also run `PhyloNet` on two options: 1) maximum likelihood (ML) and 2) maximum pseudolikelihood (MPL). We choose $h = 1$ hybridization event on both cases, with 10 independent runs for each. We choose one representative individual per species in the gene trees and removed gene trees that contained fewer than 5 taxa. In total, we use 448,758 gene trees in the `PhyloNet` analyses. We root the gene trees on the known outgroup (African hunting dog) or on dhole if African hunting dog is not in the gene tree.

3 Results

3.1 Simulation study on eight or fewer taxa

3.1.1 Identifying the correct network as the top 1 ranked network with the smallest invariant score

Figure 3 (first row) shows the proportion of times (out of the 30 replicates) that our invariants method identifies the true network as the top 1 ranked network with the smallest invariant score for the cases of true concordance factors and the Gaussian-perturbed CFs with increasing standard deviation (from left to right). We also quantify the proportion of times that our invariants method identifies the symmetric network (when clades n_1 and n_2 in Figure 1 (top) are switched) because in the 4-node hybridization cycle it is difficult for the method to distinguish clades n_1 and n_2 given that they both contribute to the hybridization node. Even for the case of true CFs, our method identifies the symmetric network as the top 1 network, instead of the true network, in two cases ($N = 1221$ and $N = 2121$), yet the difference in invariance score (L_2 -norm of evaluated polynomials) of the true and symmetric networks in these cases is of the order of 10^{-16} . For the noisiest setting ($\sigma = 0.0005$, far right), it is evident that some networks are difficult to be detected by our method, but it is worth highlighting that whenever there are two taxa sampled on each of the four clades (n_0, n_1, n_2, n_3) as in $N = 2222$, our method always identifies the true (or symmetric) network as the top 1. This condition can easily be satisfied whenever there are multiple individuals per species.

Figure 3 (second row) shows the proportion of times (out of the 30 replicates) that our invariants method identifies the true network as the top 1 ranked network with smallest invariant score for the case of true simulated gene trees with increasing number of gene trees (from left to right), where “g.t.” is abbreviation for “gene trees”. Unlike the case of true or Gaussian-perturbed CFs, our method only identifies the true (or symmetric) network when there are 2 taxa sampled per clade ($N = 2222$) or at least 2 taxa sampled on the hybrid clade and sister clades ($N = 2221$). With 1,000 gene trees or more, the networks $N = 2121$ and $N = 2211$ are also accurately identified by our method.

Figure 3 (third row) shows the proportion of times (out of the 30 replicates) that our invariants method identifies the true network as the top 1 ranked network with smallest invariant score for the case of estimated gene trees with increasing number of gene trees and sequence length (from left to right). Again, our method only identifies the true (or symmetric) network when there are 2 taxa sampled per clade ($N = 2222$) or at least 2 taxa sampled on the hybrid clade and sister clades ($N = 2221$). As before, with 1,000 gene trees, the networks $N = 2121$ and $N = 2211$ are also accurately identified by our method, and with 10,000 gene trees, the networks $N = 1221$ and $N = 1222$ are also accurately identified.

Summarizing Figure 3 related to our method’s ability to identify the true (or symmetric) network as top 1 ranked network by its invariant score, we can highlight that networks with only one sampled taxon in some of the clades are harder to be identified. As long as two taxa are sampled from each of the clades, our method is able to identify the network with high accuracy on all simulation settings. The number of genes is more important than the sequence length, and with at least 100 genes, our method is able to correctly identify networks with fewer than two taxa sampled from some clades (like 2221). We highlight that we are equally interested in the true and symmetric networks since a follow-up optimization of branch lengths and inheritance probability (γ) on the fixed network will estimate the correct γ and identify which of the sister clades is the major ($\gamma > 0.5$) and which is the minor ($\gamma < 0.5$).

3.1.2 Reducing the space of candidate networks using the invariant score

Figure 4 (first row) shows the proportion of times (out of the 30 replicates) that the true (or symmetric) network are within the top 5 ranked networks based on smallest invariant score for the cases of true concordance factors and the Gaussian-perturbed CFs with increasing standard deviation (from left to right). In all cases, our method includes the true (and symmetric) networks within the top 5 which means that our method accurately and in a fast manner reduces the space of candidate networks to just 5 candidates that can later be tested with another accurate methodology like likelihood.

Figure 4 (second row) shows the proportion of times (out of the 30 replicates) that the true (or symmetric) network are within the top 5 ranked networks based on smallest invariant score for the case of true simulated gene trees. Only the networks with one sampled taxon on the hybrid clade (n_0) cannot be recovered with 100 or 1000 gene trees. Whenever there are 10,000 gene trees in the sample, the networks $N = 1221$ and $N = 1222$ can now be recovered as part of the top 5, yet the networks $N = 1122$ or $N = 1212$ (when there is one sampled taxon in the hybrid clade and one of the sister clades) failed to be detected. The same conclusions are true for the case of estimated gene trees (Figure 4 (third row)) with number of gene trees having more weight on the accuracy than sequence length.

Summarizing Figure 4 related to our method’s ability to identify the true (or symmetric) network within the top 5 ranked networks by its invariant score, we can highlight that as long as there are at least 2 taxa sampled from the hybrid clade,

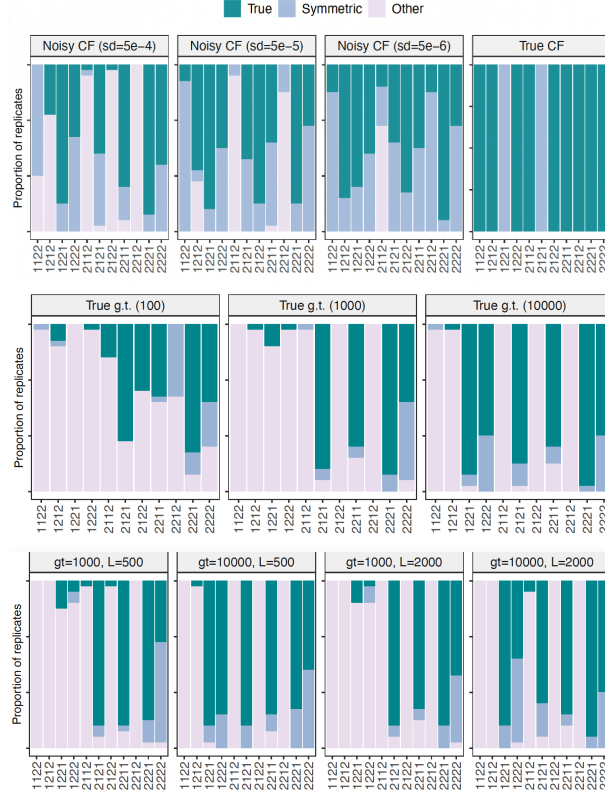


Figure 3: Proportion of times that the true hybridization cycle or the symmetric cycle (with clades n_1 and n_2 inverted) are in the top 1 ranked hybridization cycles identified by the phylogenetic invariants on 3 different cases. Y-axis is between 0 and 100. **First row:** true and Gaussian-perturbed CFs. Each panel corresponds to a type of simulation: using true concordance factors (left) and using concordance factors with added Gaussian noise (with increasing standard deviation for noise from left to right). Whenever there are at least two taxa in each of the four clades (network 2222), our method is very accurate in detecting the true (or its symmetric) network. Our method is also accurate to identify the true network in the top 5 ranked networks (see Figure 4 (first row)) which will greatly reduce the space of candidate networks to be compared with a likelihood approach. **Second row:** true simulated gene trees (“g.t.”). Whenever there are at least two taxa in each of the four clades (network 2222), our method is very accurate in detecting the true (or its symmetric) network. Our method is also accurate to identify the true network in the top 5 ranked networks (see Figure 4 (second row)) which will greatly reduce the space of candidate networks to be compared with a likelihood approach. **Third row:** estimated gene trees. Each panel corresponds to a number of gene trees (g.t. from 1000 to 10,000) and sequence length (L from 500bp to 2000bp). Whenever there are at least two taxa in each of the four clades (network 2222) or only clade n_3 having one taxon ($N = 2221$), our method is very accurate in detecting the true (or its symmetric) network. Our method is also accurate to identify the true network in the top 5 ranked networks (see Figure 4 (third row)) which will greatly reduce the space of candidate networks to be compared with a likelihood approach.

then our method is able to place the true (and symmetric) network within the top 5 networks. This effectively reduces the space of candidate networks to compare with a likelihood approach. As mentioned, we do not view our method as a substitute to existing network inference methods. On the contrary, we believe that our invariants method will serve as a complement to identify a small subset of network possibilities that will help bypass the optimization on network space.

Figure 5 (first row) shows the rank in the invariant scores of the true and symmetric networks on the cases of true and Gaussian-perturbed CFs. Ideally, the true network (or at least its symmetric version) should be ranked as 1 by our invariants method. While the true and symmetric networks are not always ranked in number 1, they are ranked within the top 5 for most of the cases. This further confirms the advantage of our current method. It is a fast way to reduce the space of candidate networks (to 5) that can later be tested using a likelihood approach.

Figure 5 (second row) shows the rank in the invariant scores of the true and symmetric networks for the case of true simulated gene trees (increased number of gene trees from left to right). The true (and symmetric) networks are not

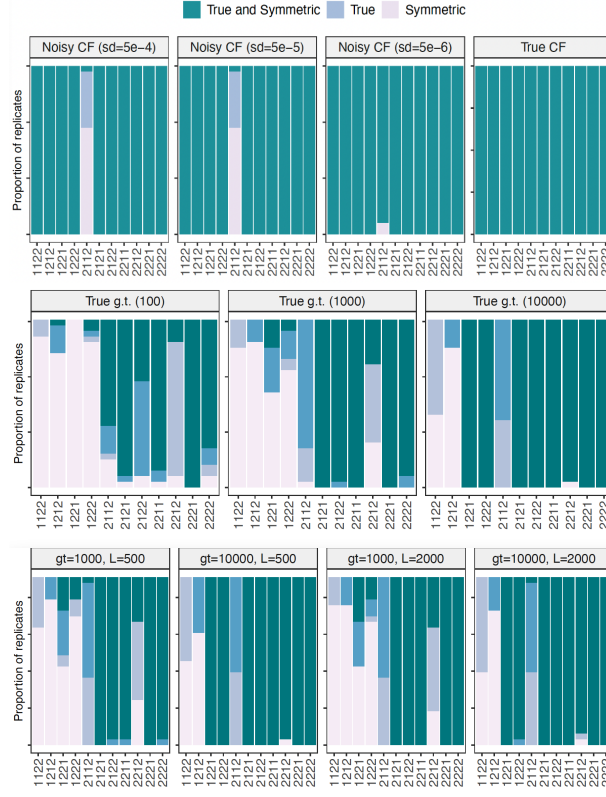


Figure 4: Proportion of times that the true network or the symmetric network (with clades n_1 and n_2 inverted) are in the top 5 ranked networks identified by the phylogenetic invariants on 3 different cases. Y-axis is between 0 and 100. **First row**: true and Gaussian-perturbed CFs. Each panel corresponds to a type of simulation: using true concordance factors (left) and using concordance factors with added Gaussian noise (with increasing standard deviation for noise from left to right). All networks are accurately identified in the top 5 ranked based on invariant score which provides evidence that our method fast and accurately reduce the space of candidate networks to just 5 alternatives. **Second row**: true simulated gene trees (“g.t.”). Only the networks that have one taxon below the hybrid node (1122, 1212, 1221, 1222) do not allow accurate reconstruction with fewer than 10,000 gene trees which brings into attention the importance of taxon sampling for this method. **Third row**: estimated gene trees. Each panel corresponds to a number of gene trees (g.t. from 1000 to 10,000) and sequence length (L from 500bp to 2000bp). Only the networks that have one taxon below the hybrid node (1122, 1212, 1221, 1222) do not allow accurate reconstruction with fewer than 10,000 gene trees which brings into attention the importance of taxon sampling for this method.

within the top 5 networks only for the cases of one sampled taxon from the hybrid clade ($N = 1122, 1212, 1221, 1222$). This same behavior is evident in Figure 5 (third row) for estimated gene trees. Again, these figures show our method’s ability to reduce the possible networks that fit the data.

In Supplementary Material, we also show plots with the values of the invariant scores for the true and symmetric networks.

Last, we present a comparison of the running times of four network methods in the Appendix. While all methods are able to identify the correct network, our phylogenetic invariants method (top row) only takes 7.17 seconds to correctly infer the network $N = 2222$ which is over twice as fast as the second fastest (PhyloNet MPL [48]). The difference in running times (and even accuracy) is more evident in the results on the *Canis* dataset (Section 3.3).

3.2 Simulation study on more than eight taxa

Table in the Supplementary Material shows the proof-of-concept results on simulated data with more than 8 taxa. Since the algorithm for more than 8 taxa (Algorithm in the Supplementary Material) relies entirely on the algorithm for 8 or fewer taxa (Algorithm 1), we simply show here that our steps to append taxon to the optimal subset of 8 taxa identified by the method indeed yields the desired network with more than 8 taxa.

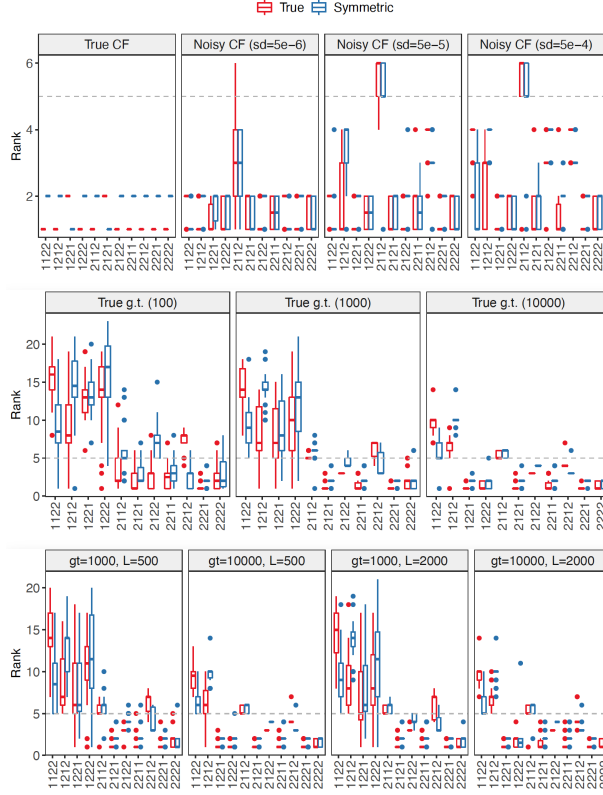


Figure 5: Rank (y-axis) in the invariant score for each network (x-axis) for the true network (red) and its symmetric network (inverted clades n_1 and n_2 , in blue) on 3 different cases. Dashed line corresponds to rank 5. **First row:** true and Gaussian-perturbed CFs. Each panel corresponds to a type of simulation: using true concordance factors (left) and using concordance factors with added Gaussian noise (with increasing standard deviation for noise from left to right). Both the true and symmetric networks are within the top 5 ranked networks by the method in all cases, and are thus, easy to distinguish from wrong networks. **Second row:** true simulated gene trees (“g.t”). Each panel corresponds to a number of simulated gene trees from 100 (left) to 10,000 (right). Both the true and symmetric networks are within the top 5 ranked networks by the method as the number of genes increases, and are thus, easy to distinguish from wrong networks. Only the networks that have one taxon below the hybrid node (1122, 1212, 1221, 1222) do not allow accurate reconstruction which brings into attention the importance of taxon sampling for this method. **Third row:** estimated gene trees. Each panel corresponds to a number of gene trees (g.t. from 1000 to 10,000) and sequence length (L from 500bp to 2000bp). Only the networks that have one taxon below the hybrid node (1122, 1212, 1221, 1222) do not allow accurate reconstruction which brings into attention the importance of taxon sampling for this method.

3.3 Phylogenetic network for the *Canis* genus

The original analysis of the *Canis* genus identified nine gene flow events using the D statistics [32] (Figure 3a in [19]). Here, we are able to replicate the hybridization event involving gene flow from the ancestor of dog and grey wolf into the ancestor of golden jackal (Figure 6 (top)). The same network is correctly inferred by SNaQ and by PhyloNet ML, but both methods take much longer to run (Table 1). PhyloNet MPL identifies a different hybridization event that has not been reported in previous studies, so it is impossible to validate it at this point. This hybridization involves an unsampled or extinct taxon (Figure 6 (bottom)).

In terms of running time, our invariants method is able to identify the correct hybridization in under 7 seconds while SNaQ (which also identifies the correct hybridization) takes over 140 seconds. PhyloNet MPL takes 40 times longer than our invariants method (with running time of over 281 seconds) and identifies a different hybridization event to those originally published (so, not possible to validate at this point). PhyloNet ML is the slowest method with over 45 minutes of running time on this small network of only 6 taxa. The inferred network by PhyloNet ML also agrees with the one originally published in [19].

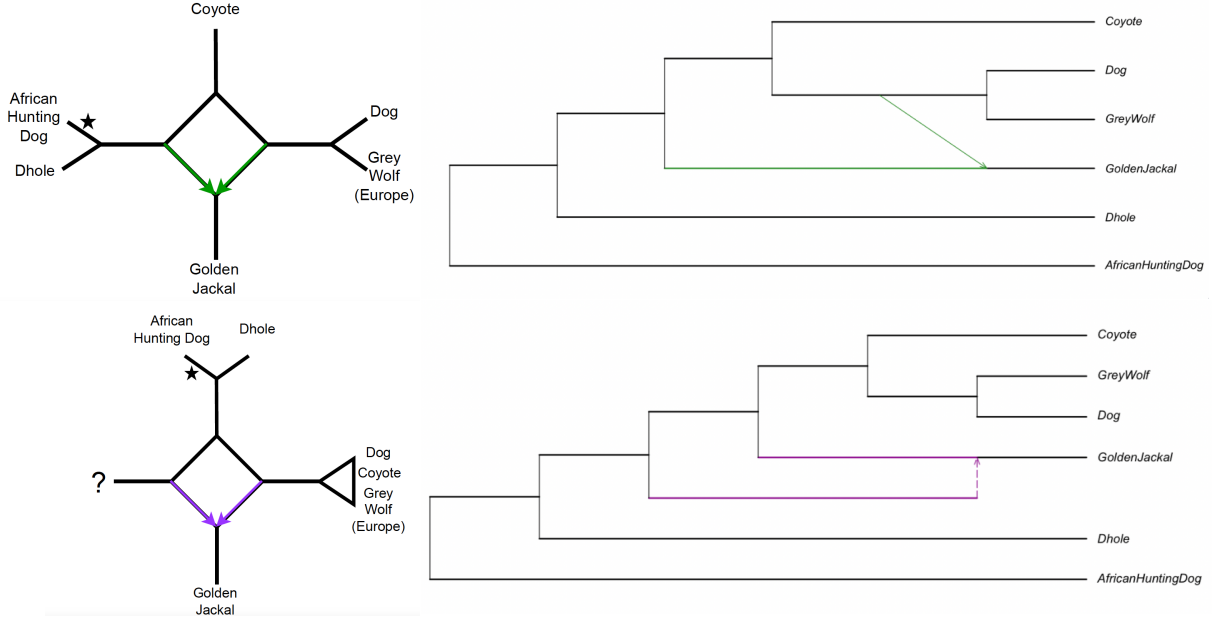


Figure 6: **Left top:** Semi-directed phylogenetic network estimated by SNaQ [38], PhyloNet ML [47], and our method based on phylogenetic invariants. **Left bottom:** Semi-directed phylogenetic network estimated by PhyloNet MPL [48]. The star marks the place where the root should be. **Right:** Rooted version of the semi-directed networks. The network estimated with PhyloNet MPL (bottom) is different from the network inferred by the other methods, and it involves an unsampled (or extinct) taxon in the hybridization. SNaQ, PhyloNet ML and our phylogenetic invariants method (top) identify the same hybridization event as the original publication [19] between the ancestral species to dog and grey wolf and the ancestral species of golden jackal, yet PhyloNet ML takes 389 times longer and SNaQ takes 20 times longer than our invariants method (see running times in Table 1).

Method	Time (seconds)
Phylogenetic invariants (our method)	6.78
SNaQ	140.58
PhyloNet ML	2723.99
PhyloNet MPL	281.25

Table 1: Running times (in seconds) on the *Canis* dataset. Our method based on evaluation of phylogenetic invariants is 20 times faster than the second fastest method (SNaQ).

4 Discussion

Here, we introduce a novel method to infer 4-node hybridization cycles on level-1 phylogenetic networks of over five taxa based on ultrafast evaluations of phylogenetic invariants. Our method bypasses optimization on network space and is able to accurately detect the hybridization cycles based on simulated and real data, especially when there are at least 2 taxa sampled from each of the four clades defined by the hybridization cycle (Figure 1 (top)). While our method is at least ~ 10 times faster than other existing network methods, it is not meant to replace existing methods. On the contrary, we believe that our ultrafast algorithm can work in conjunction with existing methodologies by reducing the space of candidate networks that can later be evaluated based on likelihood or Bayesian approaches.

Nevertheless, there are limitations in our current method. Specifically, we are only able to detect hybridization cycles with 4 nodes on level-1 networks. For the case of three nodes in the hybridization cycle, it has been investigated already in [38] that the set of phylogenetic invariants is empty. However, there is still room to extend the current methodology to hybridization cycles of five nodes or more. In addition, even when our method can infer networks with any number of taxa, it is unable to resolve the topology inside any of the four clades defined by the hybridization cycle (Figure 1 (top)). That is, our method is only able to identify the placement of the 4-node hybridization cycle. For this reason, we view our method as in line with other hybrid detection methods such as ABBA-BABA test [32], MSCquartets [30, 36], and HyDe [28] that are also only able to detect specific hybridization patterns in subsets of taxa.

Future work will involve the generation of the phylogenetic invariants related to hybridization cycles of five nodes or more, and the development of a merging algorithm that can produce a fully resolved n -taxon phylogenetic network from the 8-taxon estimated networks inferred by our invariants method. In addition, our proposed method is unable to infer biological parameters like inheritance probability or branch length, but future work could exploit the CF formulas (which depend on branch lengths and inheritance probabilities) to provide estimated values to these parameters from the observed CFs.

5 Data and code availability

The *Canis* dataset was made publicly available by the original publication [19] and can be accessed through the GitHub repository of [50] here https://github.com/chaoszhang/Weighted-ASTRAL_data. All the scripts for our work are publicly available in the GitHub repository <https://github.com/solislemuslab/PhyloDiamond.jl>. The new Julia package *PhyloDiamond.jl* is open-source, publicly available in <https://github.com/solislemuslab/PhyloDiamond.jl>.

6 Acknowledgements

This work was supported by the National Science Foundation [DEB-2144367 to CSL]. This was also funded by the UW-Madison Fall Competition [to CSL]. The authors thank Marianne Bjørner and Nathan Kolbow for help with phylogenetics code, and Sam Ozminkowski for help with building the Julia package.

References

- [1] Roya Adavoudi and Małgorzata Pilot. Consequences of hybridization in mammals: A systematic review. *Genes*, 13(1):50, 2021.
- [2] Elizabeth S Allman and John A Rhodes. Phylogenetic invariants for the general Markov model of sequence mutation. *Mathematical Biosciences*, 186(2):113–144, 2003.
- [3] Muhammad Ardiyansyah. Distinguishing Level-2 Phylogenetic Networks Using Phylogenetic Invariants. *arXiv preprint arXiv:2104.12479*, 2021.
- [4] Travis Barton et al. Statistical learning with phylogenetic network invariants. *arXiv preprint arXiv:2211.11919*, 2022.
- [5] David A Baum. Concordance trees, concordance factors, and the exploration of reticulate genealogy. *Taxon*, 56(May):417–426, 2007.
- [6] Marianne Bjørner et al. Detectability of Varied Hybridization Scenarios using Genome-Scale Hybrid Detection Methods. *arXiv preprint arXiv:2211.00712*, November 2022.
- [7] Christopher Blair and Cécile Ané. Phylogenetic trees and networks can serve as powerful and complementary approaches for analysis of genomic data. *Syst. Biol.*, 69(3):593–601, 2020.
- [8] David Bryant and Vincent Moulton. Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol Biol Evol*, 21(2):255–265, 2004.
- [9] Marta Casanellas et al. "SAQ: Semi-Algebraic quartet reconstruction". *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 18(6):2855–2861, December 2021.
- [10] Marta Casanellas et al. "designing weights for Quartet-Based methods when data are heterogeneous across lineages". *Bull. Math. Biol.*, 85(7):68, June 2023.
- [11] Kimberly M Charles and Ivana Stehlik. Assisted species migration and hybridization to conserve cold-adapted plants under climate change. *Conservation Biology*, 35(2):559–566, 2021.
- [12] Joseph Cummings et al. Invariants for level-1 phylogenetic networks under the Cavendar-Farris-Neyman Model. *arXiv preprint arXiv:2102.03431*, 2021.
- [13] Vanessa De Santis et al. Biological and trophic consequences of genetic introgression between endemic and invasive *Barbus* fishes. *Biological Invasions*, 23(11):3351–3368, 2021.
- [14] James H Degnan. Modeling Hybridization Under the Network Multispecies Coalescent. *Syst. Biol.*, 67(5):786–799, 05 2018.
- [15] Awa Diop et al. Gene flow and introgression are pervasive forces shaping the evolution of bacterial species. *Genome Biology*, 23(1):1–19, 2022.

- [16] RA Elworth et al. Advances in computational methods for phylogenetic networks in the presence of hybridization. In *Bioinformatics and phylogenetics*, pages 317–360. Springer, 2019.
- [17] Joseph Felsenstein. Counting phylogenetic invariants in some simple cases. *Journal of Theoretical Biology*, 152(3):357–376, 1991.
- [18] Jesús Fernández-Sánchez and Marta Casanellas. "invariant versus classical quartet inference when evolution is heterogeneous across sites and lineages". *Syst. Biol.*, 65(2):280–291, March 2016.
- [19] Shyam Gopalakrishnan et al. "interspecific gene flow shaped the evolution of the genus *canis*". *Current Biology*, 28(3), 2018.
- [20] DR Grayson and ME Stillman. Macaulay2, a software system for research in algebraic geometry. Available at <http://www.math.uiuc.edu/Macaulay2/>.
- [21] Elizabeth Gross et al. Distinguishing level-1 phylogenetic networks on the basis of data generated by Markov processes. *Journal of Mathematical Biology*, 83(3):1–24, 2021.
- [22] Elizabeth Gross and Colby Long. Distinguishing phylogenetic networks. *SIAM Journal on Applied Algebra and Geometry*, 2(1):72–93, 2018.
- [23] Stefan Grünwald et al. QNet: an agglomerative method for the construction of phylogenetic networks from weighted quartets. *Mol Biol Evol*, 24(2):532–538, 2007.
- [24] Richard R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, February 2002.
- [25] Daniel Huson et al. *Phylogenetic Networks*. Cambridge University Press, New York, NY, first edition, 2010.
- [26] Daniel H Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol*, 23(2):254–267, 2006.
- [27] Sungsik Kong et al. Classes of explicit phylogenetic networks and their biological and mathematical significance. *Journal of Mathematical Biology*, 84(6):1–44, 2022.
- [28] Laura S. Kubatko and Julia Chifman. An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evolutionary Biology*, 19(1):112, 2019.
- [29] Bui Quang Minh et al. "iq-tree 2: New models and efficient methods for phylogenetic inference in the genomic era". *Mol Biol Evol*, 37(5):1530–1534, May 2020.
- [30] Jonathan D Mitchell, Elizabeth S Allman, and John A Rhodes. Hypothesis testing near singularities and boundaries. *Electron J Stat*, 13(1):2150–2193, 2019.
- [31] Xiao-Xu Pang and Da-Yong Zhang. Impact of Ghost Introgression on Coalescent-based Species Tree Inference and Estimation of Divergence Time. January 2022.
- [32] Nick Patterson et al. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 11 2012.
- [33] Oscar A Pérez-Escobar et al. Molecular clocks and archeogenomics of a late period Egyptian date palm leaf reveal introgression from wild relatives and add timestamps on the domestication. *Mol Biol Evol*, 38(10):4475–4492, 2021.
- [34] Andrew Rambaut and Nicholas C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13(3):235–238, June 1997.
- [35] Olivier Rey et al. Diverging patterns of introgression from *Schistosoma bovis* across *S. haematobium* African lineages. *PLoS Pathogens*, 17(2):e1009313, 2021.
- [36] John A Rhodes et al. MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R. *Bioinformatics*, 37(12):1766–1768, Jul 2021.
- [37] Kalyaanamoorthy S et al. "modelfinder: fast model selection for accurate phylogenetic estimates". *Nat Methods*, 14(6):587–589, June 2017.
- [38] Claudia Solís-Lemus and Cécile Ané. "inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting". *PLoS Genet.*, 12(3):e1005896, March 2016.
- [39] Claudia Solís-Lemus et al. "inconsistency of species tree methods under gene flow". *Syst. Biol.*, 65(5):843–851, September 2016.
- [40] Claudia Solís-Lemus et al. "PhyloNetworks: A package for phylogenetic networks". 34(12):3292–3298, December 2017.
- [41] Claudia R. Solís-Lemus, Arrigo Coen, and Cécile Ané. On the Identifiability of Phylogenetic Networks under a Pseudolikelihood model. *arXiv preprint arXiv:2010.01758*, 2020.

- [42] Mike A Steel and YX Fu. Classifying and counting linear phylogenetic invariants for the Jukes–Cantor model. *Journal of Computational Biology*, 2(1):39–47, 1995.
- [43] Jan Steensels et al. Interspecific hybridization as a driver of fungal evolution and adaptation. *Nature Reviews Microbiology*, 19(8):485–500, 2021.
- [44] Anton Suvorov et al. Deep ancestral introgression shapes evolutionary history of dragonflies and damselflies. *Syst. Biol.*, 71(3):526–546, 2022.
- [45] Théo Tricou et al. "ghost lineages highly influence the interpretation of introgression tests". *Syst. Biol.*, February 2022.
- [46] Dingqiao Wen et al. Bayesian Inference of Reticulate Phylogenies under the Multispecies Network Coalescent. *PLoS Genet.*, 12(5):e1006006–, 05 2016.
- [47] Yun Yu et al. Maximum Likelihood Inference of Reticulate Evolutionary Histories. *PNAS*, 111(46):16448–16453, 2014.
- [48] Yun Yu and Luay Nakhleh. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, 16(10):1–10, 2015.
- [49] Chao Zhang et al. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, 19(6):153, 2018.
- [50] Chao Zhang and Siavash Mirarab. Weighting by Gene Tree Uncertainty Improves Accuracy of Quartet-based Species Trees. *Mol Biol Evol*, 39(12):msac215, December 2022.

A Representation of a level-1 semi-directed phylogenetic network with a set of polynomial equations corresponding to the expected concordance factors under the coalescent model

Notation:

- n corresponds to the number of individuals from each of the 4 clades: (n_0, n_1, n_2, n_3) . For example, $n = (0, 0, 2, 2)$ means that you have the quartet with 2 individuals in n_2 and 2 individuals in n_3
- "Individuals" list the individuals taken from each of the four clades. For example, $i_1, i_2 \in n_0$ means that we took two species from clade n_0 denoted i_1 and i_2 . We use these species names to define the splits for the formulas: e.g. $P(i_1, i_2 | j_1, j_2)$ represents the probability that i_1 and i_2 are together in one side of the split (and j_1, j_2 together)
- "Type" corresponds to the type of quartet which matches the types in [38]
- "CF Formula" corresponds to the formula of the expected CF for that given quartet under the multispecies coalescent model
- "CF value" is the variable we give to the observed CF we will read from the data table

By [41], we know that we only need to select at most two individuals per subgraph (n_0, n_1, n_2, n_3) to define all the CF formulas that involve the hybridization cycle. We list all the CF formulas in the table below assuming that we do have 2 individuals per clade (n_0, n_1, n_2, n_3) (that is, we have at least eight species).

n	Type	CF Formula	CF value
(0, 0, 2, 2)	5	$1 - \frac{2}{3}z_2z_{2,3}z_3$	$a1$
		$\frac{1}{3}z_2z_{2,3}z_3$	$a2$
		$\frac{1}{3}z_2z_{2,3}z_3$	$a3$
(0, 1, 2, 1)	5	$1 - \frac{2}{3}z_2z_{3,2}z_2$	$a4$
		$\frac{1}{3}z_2z_{3,2}z_2$	$a5$
		$\frac{1}{3}z_2z_{3,2}z_2$	$a6$
(0, 1, 1, 2)	5	$1 - \frac{2}{3}z_3$	$a7$
		$\frac{1}{3}z_3$	$a8$
		$\frac{1}{3}z_3$	$a9$
(0, 2, 2, 0)	5	$1 - \frac{2}{3}z_2z_{2,3}z_{1,3}z_1$	$a10$
		$\frac{1}{3}z_2z_{2,3}z_{1,3}z_1$	$a11$
		$\frac{1}{3}z_2z_{2,3}z_{1,3}z_1$	$a12$
(0, 2, 1, 1)	5	$1 - \frac{2}{3}z_{1,3}z_1$	$a13$
		$\frac{1}{3}z_{1,3}z_1$	$a14$
		$\frac{1}{3}z_{1,3}z_1$	$a15$
(0, 2, 0, 2)	5	$1 - \frac{2}{3}z_3z_{1,3}z_1$	$a16$
		$\frac{1}{3}z_3z_{1,3}z_1$	$a17$
		$\frac{1}{3}z_3z_{1,3}z_1$	$a18$
(1, 0, 2, 1)	2	$(1 - \gamma) \left(1 - \frac{2}{3}z_2z_{3,2} \right) + \gamma \left(1 - \frac{2}{3}z_2 \right)$	$a19$

		$(1-\gamma)\frac{1}{3}z_{2,3}z_2 + \gamma\frac{1}{3}z_2$	a_{20}
		$(1-\gamma)\frac{1}{3}z_{2,3}z_2 + \gamma\frac{1}{3}z_2$	a_{21}
(1, 0, 1, 2)	2	$(1-\gamma)\left(1 - \frac{2}{3}z_3\right) + \gamma\left(1 - \frac{2}{3}z_{2,3}z_3\right)$	a_{22}
		$(1-\gamma)\frac{1}{3}z_3 + \gamma\frac{1}{3}z_{2,3}z_3$	a_{23}
		$(1-\gamma)\frac{1}{3}z_3 + \gamma\frac{1}{3}z_{2,3}z_3$	a_{24}
(1, 1, 2, 0)	2	$(1-\gamma)\left(1 - \frac{2}{3}z_{1,3}z_{2,3}z_2\right) + \gamma\left(1 - \frac{2}{3}z_2\right)$	a_{25}
		$(1-\gamma)\frac{1}{3}z_{1,3}z_{2,3}z_2 + \gamma\frac{1}{3}z_2$	a_{26}
		$(1-\gamma)\frac{1}{3}z_{1,3}z_{2,3}z_2 + \gamma\frac{1}{3}z_2$	a_{27}
(1, 1, 1, 1)	3	$(1-\gamma)\left(1 - \frac{2}{3}z_{1,3}\right) + \gamma\frac{1}{3}z_{2,3}$	a_{28}
		$(1-\gamma)\frac{1}{3}z_{1,3} + \gamma\left(1 - \frac{2}{3}z_{2,3}\right)$	a_{29}
		$(1-\gamma)\frac{1}{3}z_{1,3} + \gamma\frac{1}{3}z_{2,3}$	a_{30}
(1, 1, 0, 2)	2	$(1-\gamma)\left(1 - \frac{2}{3}z_{1,3}z_3\right) + \gamma\left(1 - \frac{2}{3}z_3\right)$	a_{31}
		$(1-\gamma)\frac{1}{3}z_{1,3}z_3 + \gamma\frac{1}{3}z_3$	a_{32}
		$(1-\gamma)\frac{1}{3}z_{1,3}z_3 + \gamma\frac{1}{3}z_3$	a_{33}
(1, 2, 1, 0)	2	$(1-\gamma)\left(1 - \frac{2}{3}z_1\right) + \gamma\left(1 - \frac{2}{3}z_{2,3}z_{1,3}z_1\right)$	a_{34}
		$(1-\gamma)\frac{1}{3}z_1 + \gamma\frac{1}{3}z_{2,3}z_{1,3}z_1$	a_{35}
		$(1-\gamma)\frac{1}{3}z_1 + \gamma\frac{1}{3}z_{2,3}z_{1,3}z_1$	a_{36}
(1, 2, 0, 1)	2	$(1-\gamma)\left(1 - \frac{2}{3}z_1\right) + \gamma\left(1 - \frac{2}{3}z_{1,3}z_1\right)$	a_{37}
		$(1-\gamma)\frac{1}{3}z_1 + \gamma\frac{1}{3}z_{1,3}z_1$	a_{38}
		$(1-\gamma)\frac{1}{3}z_1 + \gamma\frac{1}{3}z_{1,3}z_1$	a_{39}
(2, 0, 2, 0)	4	$(1-\gamma)^2\left(1 - \frac{2}{3}z_2z_0z_{0,1}z_{1,3}z_{2,3}\right) + 2\gamma(1-\gamma)\left(1 - \frac{2}{3}z_2z_0\right) + \gamma^2\left(1 - \frac{2}{3}z_2z_0z_{0,2}\right)$	a_{40}
		$(1-\gamma)^2\frac{1}{3}z_2z_0z_{0,1}z_{1,3}z_{2,3} + 2\gamma(1-\gamma)\frac{1}{3}z_2z_0 + \gamma^2\frac{1}{3}z_2z_0z_{0,2}$	a_{41}
		$(1-\gamma)^2\frac{1}{3}z_2z_0z_{0,1}z_{1,3}z_{2,3} + 2\gamma(1-\gamma)\frac{1}{3}z_2z_0 + \gamma^2\frac{1}{3}z_2z_0z_{0,2}$	a_{42}
(2, 0, 1, 1)	1	$(1-\gamma)^2\left(1 - \frac{2}{3}z_0z_{1,3}z_{0,1}\right) + 2\gamma(1-\gamma)\left(1 - z_0 + \frac{1}{3}z_0z_{2,3}\right) + \gamma^2\left(1 - \frac{2}{3}z_0z_{0,2}\right)$	a_{43}
		$(1-\gamma)^2\frac{1}{3}z_0z_{1,3}z_{0,1} + \gamma(1-\gamma)z_0\left(1 - \frac{1}{3}z_{2,3}\right) + \gamma^2\frac{1}{3}z_0z_{0,2}$	a_{44}
		$(1-\gamma)^2\frac{1}{3}z_0z_{1,3}z_{0,1} + \gamma(1-\gamma)z_0\left(1 - \frac{1}{3}z_{2,3}\right) + \gamma^2\frac{1}{3}z_0z_{0,2}$	a_{45}
(2, 0, 0, 2)	4	$(1-\gamma)^2\left(1 - \frac{2}{3}z_3z_0z_{1,3}z_{0,1}\right) + 2\gamma(1-\gamma)\left(1 - \frac{2}{3}z_3z_0\right) + \gamma^2\left(1 - \frac{2}{3}z_3z_0z_{2,3}z_{0,2}\right)$	a_{46}
		$(1-\gamma)^2\frac{1}{3}z_3z_0z_{1,3}z_{0,1} + 2\gamma(1-\gamma)\frac{1}{3}z_3z_0 + \gamma^2\frac{1}{3}z_3z_0z_{2,3}z_{0,2}$	a_{47}
		$(1-\gamma)^2\frac{1}{3}z_3z_0z_{1,3}z_{0,1} + 2\gamma(1-\gamma)\frac{1}{3}z_3z_0 + \gamma^2\frac{1}{3}z_3z_0z_{2,3}z_{0,2}$	a_{48}

(2, 1, 1, 0)	1	$(1 - \gamma)^2 \left(1 - \frac{2}{3} z_0 z_{0,1}\right) + 2\gamma(1 - \gamma) \left(1 - z_0 + \frac{1}{3} z_0 z_{2,3} z_{1,3}\right) + \gamma^2 \left(1 - \frac{2}{3} z_0 z_{0,2}\right)$	a49
		$(1 - \gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1 - \gamma) z_0 \left(1 - \frac{1}{3} z_{2,3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2}$	a50
		$(1 - \gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1 - \gamma) z_0 \left(1 - \frac{1}{3} z_{2,3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2}$	a51
(2, 1, 0, 1)	1	$(1 - \gamma)^2 \left(1 - \frac{2}{3} z_0 z_{0,1}\right) + 2\gamma(1 - \gamma) \left(1 - z_0 + \frac{1}{3} z_0 z_{1,3}\right) + \gamma^2 \left(1 - \frac{2}{3} z_0 z_{0,2} z_{2,3}\right)$	a52
		$(1 - \gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1 - \gamma) z_0 \left(1 - \frac{1}{3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} z_{2,3}$	a53
		$(1 - \gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1 - \gamma) z_0 \left(1 - \frac{1}{3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} z_{2,3}$	a54
(2, 2, 0, 0)	4	$(1 - \gamma)^2 \left(1 - \frac{2}{3} z_1 z_0 z_{0,1}\right) + 2\gamma(1 - \gamma) \left(1 - \frac{2}{3} z_1 z_0\right) + \gamma^2 \left(1 - \frac{2}{3} z_1 z_0 z_{0,2} z_{2,3} z_{1,3}\right)$	a55
		$(1 - \gamma)^2 \frac{1}{3} z_1 z_0 z_{0,1} + 2\gamma(1 - \gamma) \frac{1}{3} z_1 z_0 + \gamma^2 \frac{1}{3} z_1 z_0 z_{0,2} z_{2,3} z_{1,3}$	a56
		$(1 - \gamma)^2 \frac{1}{3} z_1 z_0 z_{0,1} + 2\gamma(1 - \gamma) \frac{1}{3} z_1 z_0 + \gamma^2 \frac{1}{3} z_1 z_0 z_{0,2} z_{2,3} z_{1,3}$	a57

Table 2: Concordance factor equations for all 4-taxon subsets that involve the hybridization cycle.

For the mapping of observed CFs to a_i values, we need to know to which of the three splits each of the a_i 's corresponds to. So, next, we elaborate on which specific split each of the a_i corresponds to:

$n = (0, 0, 2, 2)$

- Individuals $k_1, k_2 \in n_2$ and $l_1, l_2 \in n_3$
- $P(k_1, k_2 | l_1, l_2) = 1 - \frac{2}{3} z_2 z_{2,3} z_3 = a_1$
- $P(k_1, l_1 | k_2, l_2) = \frac{1}{3} z_2 z_{2,3} z_3 = a_2$
- $P(k_1, l_2 | k_2, l_1) = \frac{1}{3} z_2 z_{2,3} z_3 = a_3$

$n = (0, 1, 2, 1)$

- Individuals $j_1 \in n_1; k_1, k_2 \in n_2$ and $l_1 \in n_3$
- $P(k_1, k_2 | j_1, l_1) = 1 - \frac{2}{3} z_2 z_{2,3} = a_4$
- $P(k_1, j_1 | k_2, l_1) = \frac{1}{3} z_2 z_{2,3} = a_5$
- $P(k_1, l_1 | k_2, j_1) = \frac{1}{3} z_2 z_{2,3} = a_6$

$n = (0, 1, 1, 2)$

- Individuals $j_1 \in n_1; k_1 \in n_2$ and $l_1, l_2 \in n_3$
- $P(l_1, l_2 | j_1, k_1) = 1 - \frac{2}{3} z_3 = a_7$
- $P(l_1, j_1 | l_2, k_1) = \frac{1}{3} z_3 = a_8$
- $P(l_1, k_1 | l_2, j_1) = \frac{1}{3} z_3 = a_9$

$$n = (0, 2, 2, 0)$$

- Individuals $j_1, j_2 \in n_1; k_1, k_2 \in n_2$
- $P(j_1, j_2 | k_1, k_2) = 1 - \frac{2}{3} z_2 z_{2,3} z_{1,3} z_1 = a_{10}$
- $P(j_1, k_1 | j_2, k_2) = \frac{1}{3} z_2 z_{2,3} z_{1,3} z_1 = a_{11}$
- $P(j_1, k_2 | j_2, k_1) = \frac{1}{3} z_2 z_{2,3} z_{1,3} z_1 = a_{12}$

$$n = (0, 2, 1, 1)$$

- Individuals $j_1, j_2 \in n_1; k_1 \in n_2; l_1 \in n_3$
- $P(j_1, j_2 | k_1, l_1) = 1 - \frac{2}{3} z_{1,3} z_1 = a_{13}$
- $P(j_1, k_1 | j_2, l_1) = \frac{1}{3} z_{1,3} z_1 = a_{14}$
- $P(j_1, l_1 | j_2, k_1) = \frac{1}{3} z_{1,3} z_1 = a_{15}$

$$n = (0, 2, 0, 2)$$

- Individuals $j_1, j_2 \in n_1; l_1, l_2 \in n_3$
- $P(j_1, j_2 | l_1, l_2) = 1 - \frac{2}{3} z_3 z_{1,3} z_1 = a_{16}$
- $P(j_1, l_1 | j_2, l_2) = \frac{1}{3} z_3 z_{1,3} z_1 = a_{17}$
- $P(j_1, l_2 | j_2, l_1) = \frac{1}{3} z_3 z_{1,3} z_1 = a_{18}$

$$n = (1, 0, 2, 1)$$

- Individuals $i_1 \in n_0; k_1, k_2 \in n_2; l_1 \in n_3$
- $P(k_1, k_2 | i_1, l_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_{2,3} z_2 \right) + \gamma \left(1 - \frac{2}{3} z_2 \right) = a_{19}$
- $P(k_1, i_1 | k_2, l_1) = (1 - \gamma) \frac{1}{3} z_{2,3} z_2 + \gamma \frac{1}{3} z_2 = a_{20}$
- $P(k_1, l_1 | k_2, i_1) = (1 - \gamma) \frac{1}{3} z_{2,3} z_2 + \gamma \frac{1}{3} z_2 = a_{21}$

$$n = (1, 0, 1, 2)$$

- Individuals $i_1 \in n_0; k_1 \in n_2; l_1, l_2 \in n_3$
- $P(l_1, l_2 | i_1, k_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_3 \right) + \gamma \left(1 - \frac{2}{3} z_{2,3} z_3 \right) = a_{22}$
- $P(l_1, i_1 | l_2, k_1) = (1 - \gamma) \frac{1}{3} z_3 + \gamma \frac{1}{3} z_{2,3} z_3 = a_{23}$
- $P(l_1, k_1 | l_2, i_1) = (1 - \gamma) \frac{1}{3} z_3 + \gamma \frac{1}{3} z_{2,3} z_3 = a_{24}$

$$n = (1, 1, 2, 0)$$

- Individuals $i_1 \in n_0; j_1 \in n_1; k_1, k_2 \in n_2$
- $P(k_1, k_2 | i_1, j_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_{1,3} z_{2,3} z_2 \right) + \gamma \left(1 - \frac{2}{3} z_2 \right) = a_{25}$
- $P(k_1, i_1 | k_2, j_1) = (1 - \gamma) \frac{1}{3} z_{1,3} z_{2,3} z_2 + \gamma \frac{1}{3} z_2 = a_{26}$

- $P(k_1, j_1 | k_2, i_1) = (1 - \gamma) \frac{1}{3} z_{1,3} z_{2,3} z_2 + \gamma \frac{1}{3} z_2 = a_{27}$

$$n = (1, 1, 1, 1)$$

- **Individuals** $i_1 \in n_0; j_1 \in n_1; k_1 \in n_2; l_1 \in n_3$
- $P(i_1, j_1 | k_1, l_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_{1,3}\right) + \gamma \frac{2}{3} z_{2,3} = a_{28}$
- $P(i_1, k_1 | j_1, l_1) = (1 - \gamma) \frac{1}{3} z_{1,3} + \gamma \left(1 - \frac{2}{3} z_{2,3}\right) = a_{29}$
- $P(i_1, l_1 | j_1, k_1) = (1 - \gamma) \frac{1}{3} z_{1,3} + \gamma \frac{1}{3} z_{2,3} = a_{30}$

$$n = (1, 1, 0, 2)$$

- **Individuals** $i_1 \in n_0; j_1 \in n_1; l_1, l_2 \in n_3$
- $P(l_1, l_2 | i_1, j_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_{1,3} z_3\right) + \gamma \left(1 - \frac{2}{3} z_3\right) = a_{31}$
- $P(l_1, i_1 | l_2, j_1) = (1 - \gamma) \frac{1}{3} z_{1,3} z_3 + \gamma \frac{1}{3} z_3 = a_{32}$
- $P(l_1, j_1 | l_2, i_1) = (1 - \gamma) \frac{1}{3} z_{1,3} z_3 + \gamma \frac{1}{3} z_3 = a_{33}$

$$n = (1, 2, 1, 0)$$

- **Individuals** $i_1 \in n_0; j_1, j_2 \in n_1; k_1 \in n_2$
- $P(j_1, j_2 | i_1, k_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_1\right) + \gamma \left(1 - \frac{2}{3} z_{2,3} z_{1,3} z_1\right) = a_{34}$
- $P(j_1, i_1 | j_2, k_1) = (1 - \gamma) \frac{1}{3} z_1 + \gamma \frac{1}{3} z_{2,3} z_{1,3} z_1 = a_{35}$
- $P(j_1, k_1 | j_2, i_1) = (1 - \gamma) \frac{1}{3} z_1 + \gamma \frac{1}{3} z_{2,3} z_{1,3} z_1 = a_{36}$

$$n = (1, 2, 0, 1)$$

- **Individuals** $i_1 \in n_0; j_1, j_2 \in n_1; l_1 \in n_3$
- $P(j_1, j_2 | i_1, l_1) = (1 - \gamma) \left(1 - \frac{2}{3} z_1\right) + \gamma \left(1 - \frac{2}{3} z_{1,3} z_1\right) = a_{37}$
- $P(j_1, l_1 | j_2, i_1) = (1 - \gamma) \frac{1}{3} z_1 + \gamma \frac{1}{3} z_{1,3} z_1 = a_{38}$
- $P(j_1, i_1 | j_2, l_1) = (1 - \gamma) \frac{1}{3} z_1 + \gamma \frac{1}{3} z_{1,3} z_1 = a_{39}$

$$n = (2, 0, 2, 0)$$

- **Individuals** $i_1, i_2 \in n_0; k_1, k_2 \in n_2$
- $P(i_1, i_2 | k_1, k_2) = (1 - \gamma)^2 \left(1 - \frac{2}{3} z_2 z_0 z_{0,1} z_{1,3} z_{2,3}\right) + 2\gamma(1 - \gamma) \left(1 - \frac{2}{3} z_2 z_0\right) + \gamma^2 \left(1 - \frac{2}{3} z_2 z_0 z_{0,2}\right) = a_{40}$
- $P(i_1, k_1 | i_2, k_2) = (1 - \gamma)^2 \frac{1}{3} z_2 z_0 z_{0,1} z_{1,3} z_{2,3} + 2\gamma(1 - \gamma) \frac{1}{3} z_2 z_0 + \gamma^2 \frac{1}{3} z_2 z_0 z_{0,2} = a_{41}$
- $P(i_1, k_2 | k_1, i_2) = (1 - \gamma)^2 \frac{1}{3} z_2 z_0 z_{0,1} z_{1,3} z_{2,3} + 2\gamma(1 - \gamma) \frac{1}{3} z_2 z_0 + \gamma^2 \frac{1}{3} z_2 z_0 z_{0,2} = a_{42}$

$$n = (2, 0, 1, 1)$$

- Individuals $i_1, i_2 \in n_0; k_1 \in n_2; l_1 \in n_3$
- $P(i_1, i_2 | k_1, l_1) = (1-\gamma)^2 \left(1 - \frac{2}{3} z_0 z_{1,3} z_{0,1}\right) + 2\gamma(1-\gamma) \left(1 - z_0 + \frac{1}{3} z_0 z_{2,3}\right) + \gamma^2 \left(1 - \frac{2}{3} z_0 z_{0,2}\right) = a_{43}$
- $P(i_1, k_1 | l_1, i_2) = (1-\gamma)^2 \frac{1}{3} z_0 z_{1,3} z_{0,1} + \gamma(1-\gamma) z_0 \left(1 - \frac{1}{3} z_{2,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} = a_{44}$
- $P(i_1, l_1 | i_2, k_1) = (1-\gamma)^2 \frac{1}{3} z_0 z_{1,3} z_{0,1} + \gamma(1-\gamma) z_0 \left(1 - \frac{1}{3} z_{2,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} = a_{45}$

$$n = (2, 0, 0, 2)$$

- Individuals $i_1, i_2 \in n_0; l_1, l_2 \in n_3$
- $P(i_1, i_2 | l_1, l_2) = (1-\gamma)^2 \left(1 - \frac{2}{3} z_3 z_0 z_{1,3} z_{0,1}\right) + 2\gamma(1-\gamma) \left(1 - \frac{2}{3} z_3 z_0\right) + \gamma^2 \left(1 - \frac{2}{3} z_3 z_0 z_{2,3} z_{0,2}\right) = a_{46}$
- $P(i_1, l_1 | i_2, l_2) = (1-\gamma)^2 \frac{1}{3} z_3 z_0 z_{1,3} z_{0,1} + 2\gamma(1-\gamma) \frac{1}{3} z_3 z_0 + \gamma^2 \frac{1}{3} z_3 z_0 z_{2,3} z_{0,2} = a_{47}$
- $P(i_1, l_2 | i_2, l_1) = (1-\gamma)^2 \frac{1}{3} z_3 z_0 z_{1,3} z_{0,1} + 2\gamma(1-\gamma) \frac{1}{3} z_3 z_0 + \gamma^2 \frac{1}{3} z_3 z_0 z_{2,3} z_{0,2} = a_{48}$

$$n = (2, 1, 1, 0)$$

- Individuals $i_1, i_2 \in n_0; j_1 \in n_1; k_1 \in n_2$
- $P(i_1, i_2 | j_1, k_1) = (1-\gamma)^2 \left(1 - \frac{2}{3} z_0 z_{0,1}\right) + 2\gamma(1-\gamma) \left(1 - z_0 + \frac{1}{3} z_0 z_{2,3} z_{1,3}\right) + \gamma^2 \left(1 - \frac{2}{3} z_0 z_{0,2}\right) = a_{49}$
- $P(i_1, j_1 | i_2, k_1) = (1-\gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1-\gamma) z_0 \left(1 - \frac{1}{3} z_{2,3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} = a_{50}$
- $P(i_1, k_1 | i_2, j_1) = (1-\gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1-\gamma) z_0 \left(1 - \frac{1}{3} z_{2,3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} = a_{51}$

$$n = (2, 1, 0, 1)$$

- Individuals $i_1, i_2 \in n_0; j_1 \in n_1; l_1 \in n_3$
- $P(i_1, i_2 | j_1, l_1) = (1-\gamma)^2 \left(1 - \frac{2}{3} z_0 z_{0,1}\right) + 2\gamma(1-\gamma) \left(1 - z_0 + \frac{1}{3} z_0 z_{1,3}\right) + \gamma^2 \left(1 - \frac{2}{3} z_0 z_{0,2} z_{2,3}\right) = a_{52}$
- $P(i_1, j_1 | i_2, l_1) = (1-\gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1-\gamma) z_0 \left(1 - \frac{1}{3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} z_{2,3} = a_{53}$
- $P(i_1, l_1 | i_2, j_1) = (1-\gamma)^2 \frac{1}{3} z_0 z_{0,1} + \gamma(1-\gamma) z_0 \left(1 - \frac{1}{3} z_{1,3}\right) + \gamma^2 \frac{1}{3} z_0 z_{0,2} z_{2,3} = a_{54}$

$$n = (2, 2, 0, 0)$$

- Individuals $i_1, i_2 \in n_0; j_1, j_2 \in n_1;$
- $P(i_1, i_2 | j_1, j_2) = (1-\gamma)^2 \left(1 - \frac{2}{3} z_1 z_0 z_{0,1}\right) + 2\gamma(1-\gamma) \left(1 - \frac{2}{3} z_1 z_0\right) + \gamma^2 \left(1 - \frac{2}{3} z_1 z_0 z_{0,2} z_{2,3} z_{1,3}\right) = a_{55}$
- $P(i_1, j_1 | i_2, j_2) = (1-\gamma)^2 \frac{1}{3} z_1 z_0 z_{0,1} + 2\gamma(1-\gamma) \frac{1}{3} z_1 z_0 + \gamma^2 \frac{1}{3} z_1 z_0 z_{0,2} z_{2,3} z_{1,3} = a_{56}$
- $P(i_1, j_2 | i_2, j_1) = (1-\gamma)^2 \frac{1}{3} z_1 z_0 z_{0,1} + 2\gamma(1-\gamma) \frac{1}{3} z_1 z_0 + \gamma^2 \frac{1}{3} z_1 z_0 z_{0,2} z_{2,3} z_{1,3} = a_{57}$

B Phylogenetic invariants for n -taxon phylogenetic networks with one 4-node hybridization cycle

Below, we present the invariants for different networks N all with one 4-cycle, but with different number of species on the clades n_0, n_1, n_2, n_3 . For example, the network $N = 1112$ corresponds to a network with 5 species: one in n_0 , one in n_1 , one in n_2 and two in n_3 . The number of species defines the number of CF formulas. For example, for 6 species, there are $\binom{6}{4} = 15$ 4-taxon subsets, each with 3 CF formulas. So, for 6 species, we have 45 CF formulas and thus, 45 CF values. However, we only want to focus on the 4-taxon subsets that involve the hybridization cycle. For the case of $N = 1112$, they are only 4: $(0, 1, 1, 2), (1, 0, 1, 2), (1, 1, 1, 1), (1, 1, 0, 2)$. Note that our table of CF formulas has 57 different CF formulas (and therefore, values). This discrepancy is due to the fact that the table is listing all possible CF formulas and we will have fewer formulas if we have less than 8 species.

For some examples, for computational restrictions, we had to include just a subset of the original CF equations to obtain the Gröbner basis in a_i . We denote these cases with "subset". All Macaulay2 scripts (and output) can be found in the GitHub repository: <https://github.com/solislemuslab/phylo-diamond.jl>.

- $N = 1112$
 1. $a_{32} - a_{33}$
 2. $a_{31} + 2a_{33} - 1$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{23} - a_{24}$
 5. $a_{22} + 2a_{24} - 1$
 6. $a_8 - a_9$
 7. $a_7 + 2a_9 - 1$
 8. $3a_9 * a_{30} + a_9 - a_{24} - a_{33}$
 9. $a_{24} * a_{29} + 2a_{24} * a_{30} + a_{29} * a_{33} - a_{30} * a_{33} - a_{33}$
 10. $3a_9 * a_{29} - 2a_9 + 2a_{24} - a_{33}$
- $N = 1121$
 1. $a_{28} + a_{29} + a_{30} - 1$
 2. $a_{26} - a_{27}$
 3. $a_{25} + 2a_{27} - 1$
 4. $a_{20} - a_{21}$
 5. $a_{19} + 2a_{21} - 1$
 6. $a_5 - a_6$
 7. $a_4 + 2a_6 - 1$
 8. $a_6 * a_{29} + 2a_6 * a_{30} - a_6 + a_{21} - a_{27}$
- $N = 1122$
 1. $a_{32} - a_{33}$
 2. $a_{31} + 2a_{33} - 1$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{26} - a_{27}$
 5. $a_{25} + 2a_{27} - 1$
 6. $a_{23} - a_{24}$
 7. $a_{22} + 2a_{24} - 1$
 8. $a_{20} - a_{21}$
 9. $a_{19} + 2a_{21} - 1$
 10. $a_8 - a_9$
 11. $a_7 + 2a_9 - 1$
 12. $a_5 - a_6$
 13. $a_4 + 2a_6 - 1$
 14. $a_2 - a_3$
 15. $a_1 + 2a_3 - 1$

16. $3a_9 * a_{30} + a_9 - a_{24} - a_{33}$
 17. $a_{24} * a_{29} + 2a_{24} * a_{30} + a_{29} * a_{33} - a_{30} * a_{33} - a_{33}$
 18. $3a_9 * a_{29} - 2a_9 + 2a_{24} - a_{33}$
 19. $a_6 * a_{29} + 2a_6 * a_{30} - a_6 + a_{21} - a_{27}$
 20. $a_3 * a_{29} + 2a_3 * a_{30} - 3a_6 * a_{33}$
 21. $3a_6 * a_{24} - 3a_3 * a_{30} + 3a_6 * a_{33} - a_3$
 22. $3a_9 * a_{21} - 3a_9 * a_{27} + 3a_6 * a_{33} - a_3$
 23. $3a_6 * a_9 - a_3$
- $N = 1211$
 1. $a_{38} - a_{39}$
 2. $a_{37} + 2a_{39} - 1$
 3. $a_{35} - a_{36}$
 4. $a_{34} + 2a_{36} - 1$
 5. $a_{28} + a_{29} + a_{30} - 1$
 6. $a_{14} - a_{15}$
 7. $a_{13} + 2a_{15} - 1$
 8. $a_{15} * a_{29} - a_{15} * a_{30} + a_{36} - a_{39}$
 - $N = 1212$
 1. $a_{38} - a_{39}$
 2. $a_{37} + 2a_{39} - 1$
 3. $a_{35} - a_{36}$
 4. $a_{34} + 2a_{36} - 1$
 5. $a_{32} - a_{33}$
 6. $a_{31} + 2a_{33} - 1$
 7. $a_{28} + a_{29} + a_{30} - 1$
 8. $a_{23} - a_{24}$
 9. $a_{22} + 2a_{24} - 1$
 10. $a_{17} - a_{18}$
 11. $a_{16} + 2a_{18} - 1$
 12. $a_{14} - a_{15}$
 13. $a_{13} + 2a_{15} - 1$
 14. $a_8 - a_9$
 15. $a_7 + 2a_9 - 1$
 16. $a_{18} * a_{30} - a_{15} * a_{33} - a_9 * a_{36} + a_9 * a_{39}$
 17. $3a_9 * a_{30} + a_9 - a_{24} - a_{33}$
 18. $a_{24} * a_{29} + 2a_{24} * a_{30} + a_{29} * a_{33} - a_{30} * a_{33} - a_{33}$
 19. $a_{18} * a_{29} - a_{15} * a_{33} + 2a_9 * a_{36} - 2a_9 * a_{39}$
 20. $a_{15} * a_{29} - a_{15} * a_{30} + a_{36} - a_{39}$
 21. $3a_9 * a_{29} - 2a_9 + 2a_{24} - a_{33}$
 22. $3a_{15} * a_{24} - 3a_9 * a_{36} + 3a_9 * a_{39} - a_{18}$
 23. $3a_9 * a_{15} - a_{18}$
 - $N = 1221$
 1. $a_{38} - a_{39}$
 2. $a_{37} + 2a_{39} - 1$
 3. $a_{35} - a_{36}$
 4. $a_{34} + 2a_{36} - 1$
 5. $a_{28} + a_{29} + a_{30} - 1$
 6. $a_{26} - a_{27}$
 7. $a_{25} + 2a_{27} - 1$
 8. $a_{20} - a_{21}$

9. $a_{19} + 2a_{21} - 1$
 10. $a_{14} - a_{15}$
 11. $a_{13} + 2a_{15} - 1$
 12. $a_{11} - a_{12}$
 13. $a_{10} + 2a_{12} - 1$
 14. $a_5 - a_6$
 15. $a_4 + 2a_6 - 1$
 16. $a_{15} * a_{29} - a_{15} * a_{30} + a_{36} - a_{39}$
 17. $a_{12} * a_{29} - a_{12} * a_{30} + 3a_6 * a_{36} - 3a_6 * a_{39}$
 18. $a_6 * a_{29} + 2a_6 * a_{30} - a_6 + a_{21} - a_{27}$
 19. $3a_{15} * a_{21} - 3a_{15} * a_{27} + 3a_{12} * a_{30} - 3a_6 * a_{36} + 3a_6 * a_{39} - a_{12}$
 20. $3a_6 * a_{15} - a_{12}$
 21. $a_{21} * a_{29} * a_{36} + 2a_{21} * a_{30} * a_{36} - a_{27} * a_{29} * a_{39} + a_{27} * a_{30} * a_{39} - a_{15} * a_{27} + a_{12} * a_{30} - a_6 * a_{36} - a_{21} * a_{36} - a_{27} * a_{36} + 2a_{21} * a_{39}$
- $N = 1222$
 1. $a_{38} - a_{39}$
 2. $a_{37} + 2a_{39} - 1$
 3. $a_{35} - a_{36}$
 4. $a_{34} + 2a_{36} - 1$
 5. $a_{32} - a_{33}$
 6. $a_{31} + 2a_{33} - 1$
 7. $a_{28} + a_{29} + a_{30} - 1$
 8. $a_{26} - a_{27}$
 9. $a_{25} + 2a_{27} - 1$
 10. $a_{23} - a_{24}$
 11. $a_{22} + 2a_{24} - 1$
 12. $a_{20} - a_{21}$
 13. $a_{19} + 2a_{21} - 1$
 14. $a_{17} - a_{18}$
 15. $a_{16} + 2a_{18} - 1$
 16. $a_{14} - a_{15}$
 17. $a_{13} + 2a_{15} - 1$
 18. $a_{11} - a_{12}$
 19. $a_{10} + 2a_{12} - 1$
 20. $a_8 - a_9$
 21. $a_7 + 2a_9 - 1$
 22. $a_5 - a_6$
 23. $a_4 + 2a_6 - 1$
 24. $a_2 - a_3$
 25. $a_1 + 2a_3 - 1$
 26. $a_{18} * a_{30} - a_{15} * a_{33} - a_9 * a_{36} + a_9 * a_{39}$
 27. $3a_9 * a_{30} + a_9 - a_{24} - a_{33}$
 28. $a_{24} * a_{29} + 2a_{24} * a_{30} + a_{29} * a_{33} - a_{30} * a_{33} - a_{33}$
 29. $a_{18} * a_{29} - a_{15} * a_{33} + 2a_9 * a_{36} - 2a_9 * a_{39}$
 30. $a_{15} * a_{29} - a_{15} * a_{30} + a_{36} - a_{39}$
 31. $a_{12} * a_{29} - a_{12} * a_{30} + 3a_6 * a_{36} - 3a_6 * a_{39}$
 32. $3a_9 * a_{29} - 2a_9 + 2a_{24} - a_{33}$
 33. $a_6 * a_{29} + 2a_6 * a_{30} - a_6 + a_{21} - a_{27}$
 34. $a_3 * a_{29} + 2a_3 * a_{30} - 3a_6 * a_{33}$
 35. $3a_{15} * a_{24} - 3a_9 * a_{36} + 3a_9 * a_{39} - a_{18}$

36. $3a_6 * a_{24} - 3a_3 * a_{30} + 3a_6 * a_{33} - a_3$
 37. $a_{18} * a_{21} - a_{12} * a_{24} - a_{18} * a_{27} + a_{12} * a_{33} + a_3 * a_{36} - a_3 * a_{39}$
 38. $3a_{15} * a_{21} - 3a_{15} * a_{27} + 3a_{12} * a_{30} - 3a_6 * a_{36} + 3a_6 * a_{39} - a_{12}$
 39. $3a_9 * a_{21} - 3a_9 * a_{27} + 3a_6 * a_{33} - a_3$
 40. $a_6 * a_{18} - a_{12} * a_{24} + a_3 * a_{36} - a_3 * a_{39}$
 41. $3a_9 * a_{15} - a_{18}$
 42. $3a_6 * a_{15} - a_{12}$
 43. $a_3 * a_{15} - a_{12} * a_{24} + a_3 * a_{36} - a_3 * a_{39}$
 44. $a_9 * a_{12} - a_{12} * a_{24} + a_3 * a_{36} - a_3 * a_{39}$
 45. $3a_6 * a_9 - a_3$
 46. $a_{21} * a_{29} * a_{36} + 2a_{21} * a_{30} * a_{36} - a_{27} * a_{29} * a_{39} + a_{27} * a_{30} * a_{39} - a_{15} * a_{27} + a_{12} * a_{30} - a_6 * a_{36} - a_{21} * a_{36} - a_{27} * a_{36} + 2a_{21} * a_{39}$
 47. $6a_9 * a_{27} * a_{36} - 3a_6 * a_{33} * a_{36} - 3a_{21} * a_{33} * a_{36} - 3a_9 * a_{27} * a_{39} - 3a_{24} * a_{27} * a_{39} + 6a_6 * a_{33} * a_{39} + a_{18} * a_{27} - a_{12} * a_{33} + a_3 * a_{36} - a_3 * a_{39}$
- $N = 2111$
 1. $a_{53} - a_{54}$
 2. $a_{52} + 2a_{54} - 1$
 3. $a_{50} - a_{51}$
 4. $a_{49} + 2a_{51} - 1$
 5. $a_{44} - a_{45}$
 6. $a_{43} + 2a_{45} - 1$
 7. $a_{28} + a_{29} + a_{30} - 1$
 - $N = 2112$
 1. $a_{53} - a_{54}$
 2. $a_{52} + 2a_{54} - 1$
 3. $a_{50} - a_{51}$
 4. $a_{49} + 2a_{51} - 1$
 5. $a_{47} - a_{48}$
 6. $a_{46} + 2a_{48} - 1$
 7. $a_{44} - a_{45}$
 8. $a_{43} + 2a_{45} - 1$
 9. $a_{32} - a_{33}$
 10. $a_{31} + 2a_{33} - 1$
 11. $a_{28} + a_{29} + a_{30} - 1$
 12. $a_{23} - a_{24}$
 13. $a_{22} + 2a_{24} - 1$
 14. $a_8 - a_9$
 15. $a_7 + 2a_9 - 1$
 16. $3a_9 * a_{30} + a_9 - a_{24} - a_{33}$
 17. $a_{24} * a_{29} + 2a_{24} * a_{30} + a_{29} * a_{33} - a_{30} * a_{33} - a_{33}$
 18. $3a_9 * a_{29} - 2a_9 + 2a_{24} - a_{33}$
 - $N = 2121$
 1. $a_4 + 2a_6 - 1$
 2. $a_5 - a_6$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{19} + 2a_{21} - 1$
 5. $a_{20} - a_{21}$
 6. $a_{25} + 2a_{27} - 1$
 7. $a_{26} - a_{27}$
 8. $a_{40} + 2a_{42} - 1$

9. $a_{41} - a_{42}$
 10. $a_{43} + 2a_{45} - 1$
 11. $a_{44} - a_{45}$
 12. $a_{49} + 2a_{51} - 1$
 13. $a_{50} - a_{51}$
 14. $a_{52} + 2a_{54} - 1$
 15. $a_{53} - a_{54}$
 16. $a_5 * a_{29} + 2a_5 * a_{30} - a_5 + a_{20} - a_{26}$
 17. $2a_{20} * a_{29}^3 a_{41} + a_{26} * a_{29}^3 a_{41} + 3a_{20} * a_{29}^2 a_{30} * a_{41} - 3a_{20} * a_{29} * a_{30}^2 a_{41} - 3a_{26} * a_{29} * a_{30}^2 a_{41} - 2a_{20} * a_{30}^3 a_{41} + 2a_{26} * a_{30}^3 a_{41} - 3a_{20} * a_{26} * a_{29}^2 a_{44} - 3a_{26}^2 a_{29}^2 a_{44} - 3a_{20} * a_{26} * a_{29} * a_{30} * a_{44} + 6a_{26}^2 a_{29} * a_{30} * a_{44} + 6a_{20} * a_{26} * a_{30}^2 a_{44} - 3a_{26}^2 a_{30}^2 a_{44} - 6a_{20}^2 a_{29}^2 a_{50} + 3a_{20} * a_{26} * a_{29}^2 a_{50} - 15a_{20}^2 a_{29} * a_{30} * a_{50} - 6a_{20} * a_{26} * a_{29} * a_{30} * a_{50} - 6a_{20}^2 a_{30}^2 a_{50} + 3a_{20} * a_{26} * a_{30}^2 a_{50} + 3a_{20} * a_{26} * a_{29}^2 a_{53} + 3a_{20} * a_{26} * a_{29} * a_{30} * a_{53} - 6a_{20} * a_{26} * a_{30}^2 a_{53} - 4a_{20} * a_{29}^2 a_{41} - 2a_{26} * a_{29}^2 a_{41} - a_{20} * a_{29} * a_{30} * a_{41} + a_{26} * a_{29} * a_{30} * a_{41} + 5a_{20} * a_{30}^2 a_{41} + a_{26} * a_{30}^2 a_{41} + 3a_{20}^2 a_{29} * a_{44} + 6a_{20} * a_{26} * a_{29} * a_{44} + 3a_{26}^2 a_{29} * a_{44} + 6a_{20}^2 a_{30} * a_{44} + 18a_5 * a_{26} * a_{30} * a_{44} - 6a_{20} * a_{26} * a_{30} * a_{44} - 3a_{26}^2 a_{30} * a_{44} + 6a_{20} * a_{26} * a_{29} * a_{50} - 9a_5 * a_{20} * a_{30} * a_{50} + 9a_{20}^2 a_{30} * a_{50} - 9a_5 * a_{26} * a_{30} * a_{50} + 12a_{20} * a_{26} * a_{30} * a_{50} - 3a_{20}^2 a_{29} * a_{53} - 3a_{26}^2 a_{29} * a_{53} - 6a_{20}^2 a_{30} * a_{53} + 3a_{26}^2 a_{30} * a_{53} + 3a_{26} * a_{29} * a_{41} - 3a_{26} * a_{30} * a_{41} - 3a_5^2 a_{44} + 3a_5 * a_{20} * a_{44} - 3a_5 * a_{26} * a_{44} - 6a_{26}^2 a_{44} + 3a_5^2 a_{50} - 6a_5 * a_{20} * a_{50} + 3a_{20}^2 a_{50} + 6a_5 * a_{26} * a_{50} - 6a_{20} * a_{26} * a_{50} + 3a_5 * a_{20} * a_{53} - 3a_{20}^2 a_{53} - 3a_5 * a_{26} * a_{53} + 6a_{20} * a_{26} * a_{53}$
- $N = 2211$
 1. $a_{13} + 2a_{15} - 1$
 2. $a_{14} - a_{15}$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{34} + 2a_{36} - 1$
 5. $a_{35} - a_{36}$
 6. $a_{37} + 2a_{39} - 1$
 7. $a_{38} - a_{39}$
 8. $a_{43} + 2a_{45} - 1$
 9. $a_{44} - a_{45}$
 10. $a_{49} + 2a_{51} - 1$
 11. $a_{50} - a_{51}$
 12. $a_{52} + 2a_{54} - 1$
 13. $a_{53} - a_{54}$
 14. $a_{55} + 2a_{57} - 1$
 15. $a_{56} - a_{57}$
 16. $a_{14} * a_{29} - a_{14} * a_{30} + a_{35} - a_{38}$
 17. $3a_{29}^2 a_{35} * a_{38} * a_{44} + 3a_{29} * a_{30} * a_{35} * a_{38} * a_{44} - 6a_{30}^2 a_{35} * a_{38} * a_{44} + 3a_{29}^2 a_{35} * a_{38} * a_{50} + 12a_{29} * a_{30} * a_{35} * a_{38} * a_{50} + 12a_{30}^2 a_{35} * a_{38} * a_{50} - 6a_{29}^2 a_{38}^2 a_{50} + 3a_{29} * a_{30} * a_{38}^2 a_{50} + 3a_{30}^2 a_{38}^2 a_{50} - 3a_{29}^2 a_{35}^2 a_{53} - 12a_{29} * a_{30} * a_{35}^2 a_{53} - 12a_{30}^2 a_{35}^2 a_{53} - 3a_{29}^2 a_{35} * a_{38} * a_{53} - 3a_{29} * a_{30} * a_{35} * a_{38} * a_{53} + 6a_{30}^2 a_{35} * a_{38} * a_{53} - a_{29}^3 a_{35} * a_{56} - 3a_{29}^2 a_{30} * a_{35} * a_{56} + 4a_{30}^3 a_{35} * a_{56} - 2a_{29}^3 a_{38} * a_{56} - 3a_{29}^2 a_{30} * a_{38} * a_{56} + 3a_{29} * a_{30}^2 a_{38} * a_{56} + 2a_{30}^3 a_{38} * a_{56} + 3a_{29} * a_{35}^2 a_{44} + 6a_{30} * a_{35}^2 a_{44} - 6a_{29} * a_{35} * a_{38} * a_{44} - 3a_{30} * a_{35} * a_{38} * a_{44} + 3a_{29} * a_{38}^2 a_{44} - 3a_{30} * a_{38}^2 a_{44} - 9a_{14} * a_{30} * a_{35} * a_{50} - 9a_{14} * a_{30} * a_{38} * a_{50} - 12a_{29} * a_{35} * a_{38} * a_{50} - 6a_{30} * a_{35} * a_{38} * a_{50} + 12a_{29} * a_{38}^2 a_{50} + 6a_{30} * a_{38}^2 a_{50} + 18a_{14} * a_{30} * a_{35} * a_{53} + 3a_{29} * a_{35}^2 a_{53} + 6a_{30} * a_{35}^2 a_{53} - 9a_{30} * a_{35} * a_{38} * a_{53} - 3a_{29} * a_{38}^2 a_{53} + 3a_{30} * a_{38}^2 a_{53} + a_{29}^2 a_{35} * a_{56} + a_{29} * a_{30} * a_{35} * a_{56} - 2a_{30}^2 a_{35} * a_{56} + 2a_{29}^2 a_{38} * a_{56} - a_{29} * a_{30} * a_{38} * a_{56} - a_{30}^2 a_{38} * a_{56} - 3a_{14} * a_{35} * a_{44} - 3a_{35}^2 a_{44} + 3a_{14} * a_{38} * a_{44} + 9a_{35} * a_{38} * a_{44} - 6a_{38}^2 a_{44} + 3a_{14}^2 a_{50} + 6a_{14} * a_{35} * a_{50} - 6a_{14} * a_{38} * a_{50} + 3a_{35} * a_{38} * a_{50} - 3a_{38}^2 a_{50} - 3a_{14}^2 a_{53} - 3a_{14} * a_{35} * a_{53} - 6a_{35}^2 a_{53} + 3a_{14} * a_{38} * a_{53} + 3a_{35} * a_{38} * a_{53} + 3a_{35}^2 a_{53} - 2a_{29} * a_{35} * a_{56} - 4a_{30} * a_{35} * a_{56} + 2a_{29} * a_{38} * a_{56} + 4a_{30} * a_{38} * a_{56} + 2a_{35} * a_{56} - 2a_{38} * a_{56}$
 - $N = 2212$
 1. $a_{16} + 2a_{18} - 1$
 2. $a_{17} - a_{18}$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{13} + 2a_{15} - 1$
 5. $a_{14} - a_{15}$

6. $a_{31} + 2a_{33} - 1$
7. $a_{32} - a_{33}$
8. $a_7 + 2a_9 - 1$
9. $a_8 - a_9$
10. $a_{34} + 2a_{36} - 1$
11. $a_{35} - a_{36}$
12. $a_{37} + 2a_{39} - 1$
13. $a_{38} - a_{39}$
14. $a_{22} + 2a_{24} - 1$
15. $a_{23} - a_{24}$
16. $a_{43} + 2a_{45} - 1$
17. $a_{44} - a_{45}$
18. $a_{46} + 2a_{48} - 1$
19. $a_{47} - a_{48}$
20. $a_{49} + 2a_{51} - 1$
21. $a_{50} - a_{51}$
22. $a_{52} + 2a_{54} - 1$
23. $a_{53} - a_{54}$
24. $a_{55} + 2a_{57} - 1$
25. $a_{56} - a_{57}$
26. $a_{17} * a_{30} - a_{14} * a_{32} - a_8 * a_{35} + a_8 * a_{38}$
27. $3a_8 * a_{30} + a_8 - a_{23} - a_{32}$
28. $a_{23} * a_{29} + 2a_{23} * a_{30} + a_{29} * a_{32} - a_{30} * a_{32} - a_{32}$
29. $a_{17} * a_{29} - a_{14} * a_{32} + 2a_8 * a_{35} - 2a_8 * a_{38}$
30. $a_{14} * a_{29} - a_{14} * a_{30} + a_{35} - a_{38}$
31. $3a_8 * a_{29} - 2a_8 + 2a_{23} - a_{32}$
32. $3a_{14} * a_{23} - 3a_8 * a_{35} + 3a_8 * a_{38} - a_{17}$
33. $3a_8 * a_{14} - a_{17}$
34. $3a_{32} * a_{38} * a_{44} - 2a_{29} * a_{38} * a_{47} - a_{30} * a_{38} * a_{47} - 3a_{32} * a_{38} * a_{50} + 3a_{32} * a_{35} * a_{53} + 3a_8 * a_{38} * a_{53} - 3a_{23} * a_{38} * a_{53} + a_{29} * a_{32} * a_{56} - a_{30} * a_{32} * a_{56} - a_{17} * a_{44} + a_{14} * a_{47} - a_{35} * a_{47} + a_{38} * a_{47} + a_{17} * a_{50} - a_{17} * a_{53} - a_8 * a_{56} + a_{23} * a_{56}$
35. $3a_8 * a_{35} * a_{44} - 3a_{32} * a_{35} * a_{44} - 3a_8 * a_{38} * a_{44} + a_{29} * a_{35} * a_{47} + 2a_{30} * a_{35} * a_{47} + 3a_8 * a_{35} * a_{50} - 3a_{23} * a_{38} * a_{50} + a_{29} * a_{32} * a_{56} - a_{30} * a_{32} * a_{56} + a_{17} * a_{44} - a_{14} * a_{47} - a_{35} * a_{47} + a_{38} * a_{47} - a_8 * a_{56} + a_{23} * a_{56}$
36. $3a_{17} * a_{32} * a_{44}^2 - 6a_{14} * a_{32} * a_{44} * a_{47} + 3a_{32} * a_{35} * a_{44} * a_{47} + 3a_{14} * a_{30} * a_{47}^2 - a_{29} * a_{35} * a_{47}^2 - 2a_{30} * a_{35} * a_{47}^2 - 3a_8 * a_{17} * a_{44} * a_{50} - 3a_{17} * a_{32} * a_{44} * a_{50} + 3a_{14} * a_{32} * a_{47} * a_{50} - 3a_8 * a_{35} * a_{47} * a_{50} + 3a_{23} * a_{38} * a_{47} * a_{50} + 3a_8 * a_{17} * a_{50}^2 - 3a_{17} * a_{23} * a_{44} * a_{53} + 3a_8 * a_{35} * a_{47} * a_{53} - 3a_8 * a_{38} * a_{47} * a_{53} - 3a_{17} * a_{23} * a_{50} * a_{53} + 6a_8 * a_{32} * a_{44} * a_{56} - a_{29} * a_{32} * a_{47} * a_{56} + a_{30} * a_{32} * a_{47} * a_{56} - 3a_8^2 * a_{50} * a_{56} + 3a_8 * a_{23} * a_{50} * a_{56} - 3a_8 * a_{32} * a_{50} * a_{56} + 3a_8^2 * a_{53} * a_{56} - 3a_8 * a_{23} * a_{53} * a_{56} + 3a_8 * a_{32} * a_{53} * a_{56} - a_{35} * a_{47}^2 + a_{38} * a_{47}^2 + a_{17} * a_{47} * a_{50} + a_{17} * a_{47} * a_{53} - 2a_{32} * a_{47} * a_{56}$
37. $9a_8 * a_{23} * a_{38} * a_{44} * a_{50} - 9a_8 * a_{23} * a_{38} * a_{50}^2 + 9a_{23} * a_{32} * a_{35} * a_{44} * a_{53} + 9a_8 * a_{23} * a_{38} * a_{44} * a_{53} + 3a_{29} * a_{32} * a_{35} * a_{47} * a_{53} - 3a_{30} * a_{32} * a_{35} * a_{47} * a_{53} + 9a_{23}^2 * a_{38} * a_{50} * a_{53} + 9a_{23} * a_{30} * a_{32} * a_{53} * a_{56} + 3a_{29} * a_{32}^2 * a_{53} * a_{56} - 3a_{30} * a_{32}^2 * a_{53} * a_{56} - 3a_{17} * a_{23} * a_{44}^2 - 3a_{23} * a_{35} * a_{44} * a_{47} + 6a_{32} * a_{35} * a_{44} * a_{47} - 3a_{29} * a_{35} * a_{47}^2 - 3a_{30} * a_{35} * a_{47}^2 + 3a_{17} * a_{23} * a_{44} * a_{50} - 9a_8 * a_{35} * a_{47} * a_{50} + 3a_8 * a_{38} * a_{47} * a_{50} + 3a_{23} * a_{38} * a_{47} * a_{50} - 3a_{17} * a_{23} * a_{44} * a_{53} + 3a_8 * a_{35} * a_{47} * a_{53} - 3a_{32} * a_{35} * a_{47} * a_{53} - 3a_8 * a_{38} * a_{47} * a_{53} - 3a_{23} * a_{38} * a_{47} * a_{53} - 6a_8 * a_{23} * a_{44} * a_{56} - 3a_{23} * a_{30} * a_{47} * a_{56} - 3a_{29} * a_{32} * a_{47} * a_{56} + 3a_{30} * a_{32} * a_{47} * a_{56} + 6a_8 * a_{23} * a_{50} * a_{56} - 3a_{23}^2 * a_{50} * a_{56} - 3a_8 * a_{23} * a_{53} * a_{56} - 3a_{32}^2 * a_{53} * a_{56} + a_{14} * a_{47}^2 + 2a_{35} * a_{47}^2 - a_{38} * a_{47}^2 - a_{17} * a_{47} * a_{50} + a_{17} * a_{47} * a_{53} + 2a_8 * a_{47} * a_{56} + a_{32} * a_{47} * a_{56}$
38. $9a_{23} * a_{32} * a_{35} * a_{44}^2 + 9a_{23} * a_{30} * a_{35} * a_{44} * a_{47} + a_{29}^2 * a_{35} * a_{47}^2 + 4a_{29} * a_{30} * a_{35} * a_{47}^2 - 5a_{30}^2 * a_{35} * a_{47}^2 - 18a_{23} * a_{32} * a_{35} * a_{44} * a_{50} - 6a_{29} * a_{32} * a_{35} * a_{47} * a_{50} + 6a_{30} * a_{32} * a_{35} * a_{47} * a_{50} + 27a_{23} * a_{30} * a_{38} * a_{47} * a_{50} + 9a_{29} * a_{32} * a_{38} * a_{47} * a_{50} - 9a_{30} * a_{32} * a_{38} * a_{47} * a_{50} + 18a_8 * a_{23} * a_{35} * a_{50}^2 - 9a_8 * a_{23} * a_{38} * a_{50}^2 - 9a_{23}^2 * a_{38} * a_{50}^2 - 9a_{23}^2 * a_{35} * a_{44} * a_{53} + 9a_{23} * a_{32} * a_{35} * a_{44} * a_{53} + 9a_8 * a_{23} * a_{38} * a_{44} * a_{53} + 9a_{23} * a_{30} * a_{35} * a_{47} * a_{53} + 6a_{29} * a_{32} * a_{35} * a_{47} * a_{53} - 6a_{30} * a_{32} * a_{35} * a_{47} * a_{53} - 9a_{23}^2 * a_{35} * a_{50} * a_{53} + 9a_{23}^2 * a_{38} * a_{50} * a_{53} + 9a_{23} * a_{30} * a_{32} * a_{44} * a_{56} + 3a_{29} * a_{32}^2 * a_{44} * a_{56} - 3a_{30} * a_{32}^2 * a_{44} * a_{56} + 9a_{23} * a_{30}^2 * a_{47} * a_{56} + a_{29}^2 * a_{32} * a_{47} * a_{56} + a_{29} * a_{30} * a_{32} * a_{47} * a_{56} - 2a_{30}^2 * a_{32} * a_{47} * a_{56} + 9a_{23}^2 * a_{30} * a_{50} * a_{56} - 27a_{23} * a_{30} *$

- $$\begin{aligned}
& a_{32} * a_{50} * a_{56} - 9a_{29} * a_{32}^2 * a_{50} * a_{56} + 9a_{30} * a_{32}^2 * a_{50} * a_{56} - 9a_{23}^2 * a_{30} * a_{53} * a_{56} + 18a_{23} * a_{30} * a_{32} * \\
& a_{53} * a_{56} + 6a_{29} * a_{32}^2 * a_{53} * a_{56} - 6a_{30} * a_{32}^2 * a_{53} * a_{56} - 3a_{17} * a_{23} * a_{44}^2 - 9a_{23} * a_{35} * a_{44} * a_{47} - 3a_{32} * \\
& a_{35} * a_{44} * a_{47} + 9a_{23} * a_{38} * a_{44} * a_{47} - 3a_{29} * a_{35} * a_{47}^2 + 3a_{29} * a_{38} * a_{47}^2 - 3a_{30} * a_{38} * a_{47}^2 + 6a_{17} * a_{23} * \\
& a_{44} * a_{50} - 3a_8 * a_{35} * a_{47} * a_{50} - 9a_{23} * a_{35} * a_{47} * a_{50} + 6a_{32} * a_{35} * a_{47} * a_{50} + 6a_8 * a_{38} * a_{47} * a_{50} + 6a_{23} * \\
& a_{38} * a_{47} * a_{50} - 9a_{32} * a_{38} * a_{47} * a_{50} - 3a_{17} * a_{23} * a_{44} * a_{53} + 3a_8 * a_{35} * a_{47} * a_{53} + 3a_{23} * a_{35} * a_{47} * a_{53} - \\
& 6a_{32} * a_{35} * a_{47} * a_{53} - 3a_8 * a_{38} * a_{47} * a_{53} - 3a_{23} * a_{38} * a_{47} * a_{53} - 3a_8 * a_{23} * a_{44} * a_{56} + 3a_{23}^2 * a_{44} * a_{56} - \\
& 3a_{32}^2 * a_{44} * a_{56} - 12a_{23} * a_{30} * a_{47} * a_{56} - 4a_{29} * a_{32} * a_{47} * a_{56} + a_{30} * a_{32} * a_{47} * a_{56} - 3a_{23} * a_{32} * a_{50} * a_{56} + \\
& 9a_{32}^2 * a_{50} * a_{56} - 3a_8 * a_{23} * a_{53} * a_{56} + 3a_{23}^2 * a_{53} * a_{56} + 3a_{23} * a_{32} * a_{53} * a_{56} - 6a_{32}^2 * a_{53} * a_{56} + a_{14} * a_{47}^2 + \\
& 4a_{35} * a_{47}^2 - 4a_{38} * a_{47}^2 - 2a_{17} * a_{47} * a_{50} + a_{17} * a_{47} * a_{53} - a_8 * a_{47} * a_{56} + a_{23} * a_{47} * a_{56} + 4a_{32} * a_{47} * a_{56} \\
39. & 3a_{14} * a_{17}^2 * a_{44}^2 - 6a_{14}^2 * a_{17} * a_{44} * a_{47} + 3a_{14} * a_{17} * a_{35} * a_{44} * a_{47} + 3a_{14}^3 * a_{47}^2 - 3a_{14}^2 * a_{35} * a_{47}^2 - \\
& 3a_{14} * a_{17}^2 * a_{44} * a_{50} - 3a_{17}^2 * a_{38} * a_{44} * a_{50} + 3a_{14}^2 * a_{17} * a_{47} * a_{50} + 3a_{14} * a_{17} * a_{38} * a_{47} * a_{50} + 3a_{17}^2 * \\
& a_{38} * a_{50}^2 - 3a_{17}^2 * a_{35} * a_{44} * a_{53} + 3a_{14} * a_{17} * a_{35} * a_{47} * a_{53} - 3a_{17}^2 * a_{35} * a_{50} * a_{53} + 3a_8 * a_{17} * a_{35} * \\
& a_{50} * a_{56} - 3a_8 * a_{17} * a_{38} * a_{50} * a_{56} - 3a_8 * a_{17} * a_{35} * a_{53} * a_{56} + 3a_8 * a_{17} * a_{38} * a_{53} * a_{56} + 2a_{17}^2 * a_{44} * \\
& a_{56} - 2a_{14} * a_{17} * a_{47} * a_{56} + a_{17} * a_{35} * a_{47} * a_{56} - a_{17} * a_{38} * a_{47} * a_{56} - a_{17}^2 * a_{50} * a_{56} + a_{17}^2 * a_{53} * a_{56} \\
40. & 3a_{29}^2 * a_{35} * a_{38} * a_{44} + 3a_{29} * a_{30} * a_{35} * a_{38} * a_{44} - 6a_{30}^2 * a_{35} * a_{38} * a_{44} + 3a_{29}^2 * a_{35} * a_{38} * a_{50} + 12a_{29} * a_{30} * \\
& a_{35} * a_{38} * a_{50} + 12a_{30}^2 * a_{35} * a_{38} * a_{50} - 6a_{29}^2 * a_{38}^2 * a_{50} + 3a_{29} * a_{30} * a_{38}^2 * a_{50} + 3a_{30}^2 * a_{38}^2 * a_{50} - 3a_{29}^2 * a_{35}^2 * \\
& a_{53} - 12a_{29} * a_{30} * a_{35}^2 * a_{53} - 12a_{30}^2 * a_{35}^2 * a_{53} - 3a_{29}^2 * a_{35} * a_{38} * a_{53} - 3a_{29} * a_{30} * a_{35} * a_{38} * a_{53} + 6a_{30}^2 * \\
& a_{35} * a_{38} * a_{53} - a_{29}^3 * a_{35} * a_{56} - 3a_{29}^2 * a_{30} * a_{35} * a_{56} + 4a_{30}^3 * a_{35} * a_{56} - 2a_{29}^3 * a_{38} * a_{56} - 3a_{29}^2 * a_{30} * a_{38} * \\
& a_{56} + 3a_{29} * a_{30}^2 * a_{38} * a_{56} + 2a_{30}^3 * a_{38} * a_{56} + 3a_{29} * a_{35}^2 * a_{44} + 6a_{30} * a_{35}^2 * a_{44} - 6a_{29} * a_{35} * a_{38} * a_{44} - 3a_{30} * \\
& a_{35} * a_{38} * a_{44} + 3a_{29} * a_{38}^2 * a_{44} - 3a_{30} * a_{38}^2 * a_{44} - 9a_{14} * a_{30} * a_{35} * a_{50} - 9a_{14} * a_{30} * a_{38} * a_{50} - 12a_{29} * \\
& a_{35} * a_{38} * a_{50} - 6a_{30} * a_{35} * a_{38} * a_{50} + 12a_{29} * a_{38}^2 * a_{50} + 6a_{30} * a_{38}^2 * a_{50} + 18a_{14} * a_{30} * a_{35} * a_{53} + 3a_{29} * \\
& a_{35}^2 * a_{53} + 6a_{30} * a_{35}^2 * a_{53} - 9a_{30} * a_{35} * a_{38} * a_{53} - 3a_{29} * a_{38}^2 * a_{53} + 3a_{30} * a_{38}^2 * a_{53} + a_{29}^2 * a_{35} * a_{56} + a_{29} * \\
& a_{30} * a_{35} * a_{56} - 2a_{30}^2 * a_{35} * a_{56} + 2a_{29}^2 * a_{38} * a_{56} - a_{29} * a_{30} * a_{38} * a_{56} - a_{30}^2 * a_{38} * a_{56} - 3a_{14} * a_{35} * a_{44} - \\
& 3a_{35}^2 * a_{44} + 3a_{14} * a_{38} * a_{44} + 9a_{35} * a_{38} * a_{44} - 6a_{38}^2 * a_{44} + 3a_{14}^2 * a_{50} + 6a_{14} * a_{35} * a_{50} - 6a_{14} * a_{38} * a_{50} + \\
& 3a_{35} * a_{38} * a_{50} - 3a_{38}^2 * a_{50} - 3a_{14}^2 * a_{53} - 3a_{14} * a_{35} * a_{53} - 6a_{35}^2 * a_{53} + 3a_{14} * a_{38} * a_{53} + 3a_{35} * a_{38} * a_{53} + \\
& 3a_{38}^2 * a_{53} - 2a_{29} * a_{35} * a_{56} - 4a_{30} * a_{35} * a_{56} + 2a_{29} * a_{38} * a_{56} + 4a_{30} * a_{38} * a_{56} + 2a_{35} * a_{56} - 2a_{38} * a_{56}
\end{aligned}$$
- $N = 2122$ (subset)
 1. $a_7 + 2a_9 - 1$
 2. $a_8 - a_9$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{22} + 2a_{24} - 1$
 5. $a_{23} - a_{24}$
 6. $a_{31} + 2a_{33} - 1$
 7. $a_{32} - a_{33}$
 8. $a_4 + 2a_6 - 1$
 9. $a_5 - a_6$
 10. $a_{19} + 2a_{21} - 1$
 11. $a_{20} - a_{21}$
 12. $a_{25} + 2a_{27} - 1$
 13. $a_{26} - a_{27}$
 14. $a_1 + 2a_3 - 1$
 15. $a_2 - a_3$
 16. $3a_8 * a_{30} + a_8 - a_{23} - a_{32}$
 17. $a_{23} * a_{29} + 2a_{23} * a_{30} + a_{29} * a_{32} - a_{30} * a_{32} - a_{32}$
 18. $3a_8 * a_{29} - 2a_8 + 2a_{23} - a_{32}$
 19. $a_5 * a_{29} + 2a_5 * a_{30} - a_5 + a_{20} - a_{26}$
 20. $a_2 * a_{29} + 2a_2 * a_{30} - 3a_5 * a_{32}$
 21. $3a_5 * a_{23} - 3a_2 * a_{30} + 3a_5 * a_{32} - a_2$
 22. $3a_8 * a_{20} - 3a_8 * a_{26} + 3a_5 * a_{32} - a_2$
 23. $3a_5 * a_8 - a_2$
 - $N = 2221$ (subset)
 1. $a_{13} + 2a_{15} - 1$
 2. $a_{14} - a_{15}$

3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{34} + 2a_{36} - 1$
 5. $a_{35} - a_{36}$
 6. $a_{37} + 2a_{39} - 1$
 7. $a_{38} - a_{39}$
 8. $a_{10} + 2a_{12} - 1$
 9. $a_{11} - a_{12}$
 10. $a_4 + 2a_6 - 1$
 11. $a_5 - a_6$
 12. $a_{19} + 2a_{21} - 1$
 13. $a_{20} - a_{21}$
 14. $a_{25} + 2a_{27} - 1$
 15. $a_{26} - a_{27}$
 16. $a_{14} * a_{29} - a_{14} * a_{30} + a_{35} - a_{38}$
 17. $a_{11} * a_{29} - a_{11} * a_{30} + 3a_5 * a_{35} - 3a_5 * a_{38}$
 18. $a_5 * a_{29} + 2a_5 * a_{30} - a_5 + a_{20} - a_{26}$
 19. $3a_{14} * a_{20} - 3a_{14} * a_{26} + 3a_{11} * a_{30} - 3a_5 * a_{35} + 3a_5 * a_{38} - a_{11}$
 20. $3a_5 * a_{14} - a_{11}$
 21. $a_{20} * a_{29} * a_{35} + 2a_{20} * a_{30} * a_{35} - a_{26} * a_{29} * a_{38} + a_{26} * a_{30} * a_{38} - a_{14} * a_{26} + a_{11} * a_{30} - a_5 * a_{35} - a_{20} * a_{35} - a_{26} * a_{35} + 2a_{20} * a_{38}$
- $N = 2222$ (subset)
 1. $a_{16} + 2a_{18} - 1$
 2. $a_{17} - a_{18}$
 3. $a_{28} + a_{29} + a_{30} - 1$
 4. $a_{13} + 2a_{15} - 1$
 5. $a_{14} - a_{15}$
 6. $a_{31} + 2a_{33} - 1$
 7. $a_{32} - a_{33}$
 8. $a_7 + 2a_9 - 1$
 9. $a_8 - a_9$
 10. $a_{34} + 2a_{36} - 1$
 11. $a_{35} - a_{36}$
 12. $a_{37} + 2a_{39} - 1$
 13. $a_{38} - a_{39}$
 14. $a_{22} + 2a_{24} - 1$
 15. $a_{23} - a_{24}$
 16. $a_{10} + 2a_{12} - 1$
 17. $a_{11} - a_{12}$
 18. $a_4 + 2a_6 - 1$
 19. $a_5 - a_6$
 20. $a_{19} + 2a_{21} - 1$
 21. $a_{20} - a_{21}$
 22. $a_{25} + 2a_{27} - 1$
 23. $a_{26} - a_{27}$
 24. $a_1 + 2a_3 - 1$
 25. $a_2 - a_3$
 26. $a_{17} * a_{30} - a_{14} * a_{32} - a_8 * a_{35} + a_8 * a_{38}$
 27. $3a_8 * a_{30} + a_8 - a_{23} - a_{32}$
 28. $a_{23} * a_{29} + 2a_{23} * a_{30} + a_{29} * a_{32} - a_{30} * a_{32} - a_{32}$
 29. $a_{17} * a_{29} - a_{14} * a_{32} + 2a_8 * a_{35} - 2a_8 * a_{38}$

$$30. a_{14} * a_{29} - a_{14} * a_{30} + a_{35} - a_{38}$$

$$31. a_{11} * a_{29} - a_{11} * a_{30} + 3a_5 * a_{35} - 3a_5 * a_{38}$$

$$32. 3a_8 * a_{29} - 2a_8 + 2a_{23} - a_{32}$$

$$33. a_5 * a_{29} + 2a_5 * a_{30} - a_5 + a_{20} - a_{26}$$

$$34. a_2 * a_{29} + 2a_2 * a_{30} - 3a_5 * a_{32}$$

$$35. 3a_{14} * a_{23} - 3a_8 * a_{35} + 3a_8 * a_{38} - a_{17}$$

$$36. 3a_5 * a_{23} - 3a_2 * a_{30} + 3a_5 * a_{32} - a_2$$

$$37. a_{17} * a_{20} - a_{11} * a_{23} - a_{17} * a_{26} + a_{11} * a_{32} + a_2 * a_{35} - a_2 * a_{38}$$

$$38. 3a_{14} * a_{20} - 3a_{14} * a_{26} + 3a_{11} * a_{30} - 3a_5 * a_{35} + 3a_5 * a_{38} - a_{11}$$

$$39. 3a_8 * a_{20} - 3a_8 * a_{26} + 3a_5 * a_{32} - a_2$$

$$40. a_5 * a_{17} - a_{11} * a_{23} + a_2 * a_{35} - a_2 * a_{38}$$

$$41. 3a_8 * a_{14} - a_{17}$$

$$42. 3a_5 * a_{14} - a_{11}$$

$$43. a_2 * a_{14} - a_{11} * a_{23} + a_2 * a_{35} - a_2 * a_{38}$$

$$44. a_8 * a_{11} - a_{11} * a_{23} + a_2 * a_{35} - a_2 * a_{38}$$

$$45. 3a_5 * a_8 - a_2$$

$$46. a_{20} * a_{29} * a_{35} + 2a_{20} * a_{30} * a_{35} - a_{26} * a_{29} * a_{38} + a_{26} * a_{30} * a_{38} - a_{14} * a_{26} + a_{11} * a_{30} - a_5 * a_{35} - a_{20} * a_{35} - a_{26} * a_{35} + 2a_{20} * a_{38}$$

$$47. 6a_8 * a_{26} * a_{35} - 3a_5 * a_{32} * a_{35} - 3a_{20} * a_{32} * a_{35} - 3a_8 * a_{26} * a_{38} - 3a_{23} * a_{26} * a_{38} + 6a_5 * a_{32} * a_{38} + a_{17} * a_{26} - a_{11} * a_{32} + a_2 * a_{35} - a_2 * a_{38}$$

C Algorithms

Algorithm 2: Inference of n-taxon phylogenetic network ($n > 8$) with phylogenetic invariants

Input : Table of estimated concordance factors; optional: the number of optimal networks (m) to return (default $m = 5$)

Output : Top m optimal networks with smallest invariant score

scores \leftarrow an empty array;

subnets \leftarrow an empty array;

for P_i , a subset of 8 taxa **do**

```
  scoresi, subnetsi  $\leftarrow$  Algorithm 1(CF,  $P_i$ ,  $m=2520$ ); /* Note total number of subnets (partitions
    for 8 taxa) = 2520 */
  append(scores, scoresi);
  append(subnets, subnetsi);
```

subnets_sorted \leftarrow sort subnets in descending order according to its score;

result \leftarrow an empty array;

for i in $1:\text{length}(\text{subnets_sorted})$ **do**

```
  miss_taxa  $\leftarrow$  array of missing taxa in subnets_sorted[i];
```

```
  for  $t$  in miss_taxa do
```

```
    for  $j$  in  $(i+1):\text{length}(\text{subnets\_sorted})$  do
```

```
      if  $t$  in  $n_0$  of subnets_sorted[j] then
```

```
        add  $t$  to  $n_0$  of subnets_sorted[i];
        break;
```

```
      else if  $t$  in  $n_3$  of subnets_sorted[j] then
```

```
        add  $t$  to  $n_3$  of subnets_sorted[i];
        break;
```

```
      else if  $t$  in  $n_1$  of subnets_sorted[j] then
```

```
        if ( $n_2$  of subnets_sorted[j] ==  $n_1$  of subnets_sorted[i]) & ( $n_1$  of subnets_sorted[j] without  $t$  is in
           $n_2$  of subnets_sorted[i]) then
          add  $t$  to  $n_2$  of subnets_sorted[i];
          break;
```

```
        else
```

```
          add  $t$  to  $n_1$  of subnets_sorted[i];
          break;
```

```
      else if  $t$  in  $n_2$  of subnets_sorted[j] then
```

```
        if ( $n_1$  of subnets_sorted[j] ==  $n_2$  of subnets_sorted[i]) & ( $n_2$  of subnets_sorted[j] without  $t$  is in  $n_1$ 
          of subnets_sorted[i]) then
          add  $t$  to  $n_1$  of subnets_sorted[i];
          break;
```

```
        else
```

```
          add  $t$  to  $n_2$  of subnets_sorted[i];
          break;
```

```
  if subnets_sorted[i] not in result then
```

```
    append(result, subnets_sorted[i])
```

```
  if  $\text{length}(\text{result})=m$  then
```

```
    break;
```

```
return result;
```

D Simulation study

Method	Time (seconds)
Phylogenetic invariants (our method)	7.17
SNaQ	80.23
PhyloNet ML	84.34
PhyloNet MPL	17.99

Table 3: Running times (in seconds) on the inference of network $N = 2222$ from 100 true simulated gene trees by four network methods: 1) our phylogenetic invariants (top row); 2) SNaQ [38]; 3) PhyloNet ML [47], and 4) PhyloNet MPL [48].

Network	True CF	Noisy CF
2223	1	1
2232	1	1 (symmetric)
2322	1	2
3222	1	1
2233	1	1 (symmetric)
2323	1	2 (symmetric)
3223	1	1
2332	1	2 (symmetric)
3232	1	2 (symmetric)
3322	1	1 (symmetric)

Table 4: Rank of the true (or symmetric) network under the two types of simulations: true CFs (left column) and Gaussian-perturbed CFs (right column) for a level of noise of $\sigma = 0.0005$. In all cases, either the true network or the symmetric networks are within the top 2 of networks identified by the method which proves that our algorithm for more than 8 taxa that builds on Algorithm 1 works appropriately.

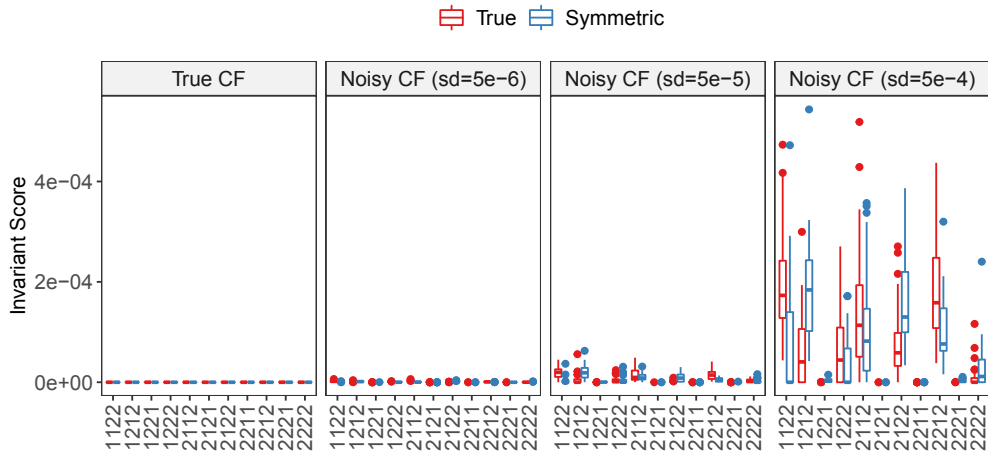


Figure 7: Invariant score (y-axis) measured as L_2 -norm of the phylogenetic invariants for each network (x-axis) for the true network (red) and its symmetric network (inverted clades n_1 and n_2 , in blue) on the cases of true and Gaussian-perturbed CFs. Each panel corresponds to a type of simulation: using true concordance factors (left) and using concordance factors with added Gaussian noise (with increasing standard deviation for noise from left to right). Both the true and symmetric networks have invariant score close to zero, and are thus, easy to distinguish from wrong networks (whose invariant score is far from zero). As the noise increases, the invariant scores moves away from zero, but still within 10^{-4} .

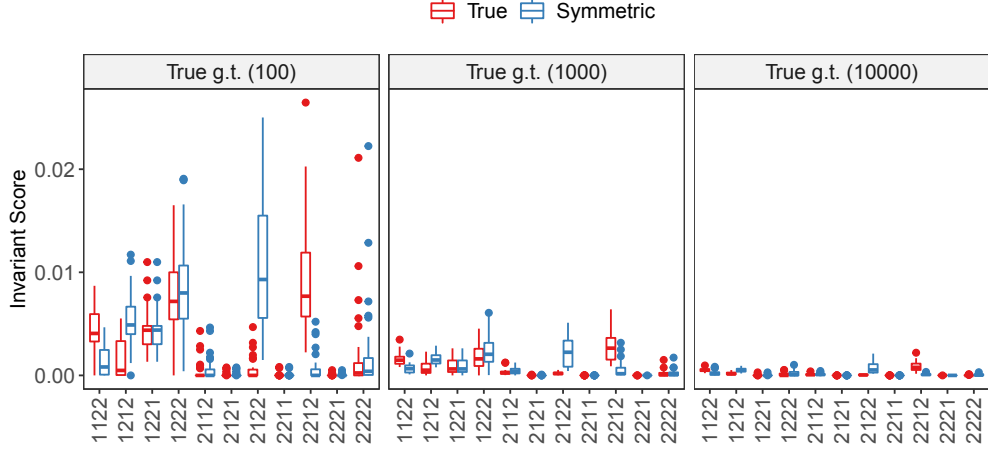


Figure 8: Invariant score (y-axis) measured as L_2 -norm of the phylogenetic invariants for each network (x-axis) for the true network (red) and its symmetric network (inverted clades n_1 and n_2 , in blue) for the case of true simulated gene trees (“g.t.”). Each panel corresponds to a number of simulated gene trees from 100 (left) to 10,000 (right). With true and perturbed concordance factors (top left and top right, respectively). As the number of gene trees increases, the invariant scores of the true and symmetric networks converge to zero, and they are thus, easy to distinguish from wrong networks (whose invariant score is far from zero).

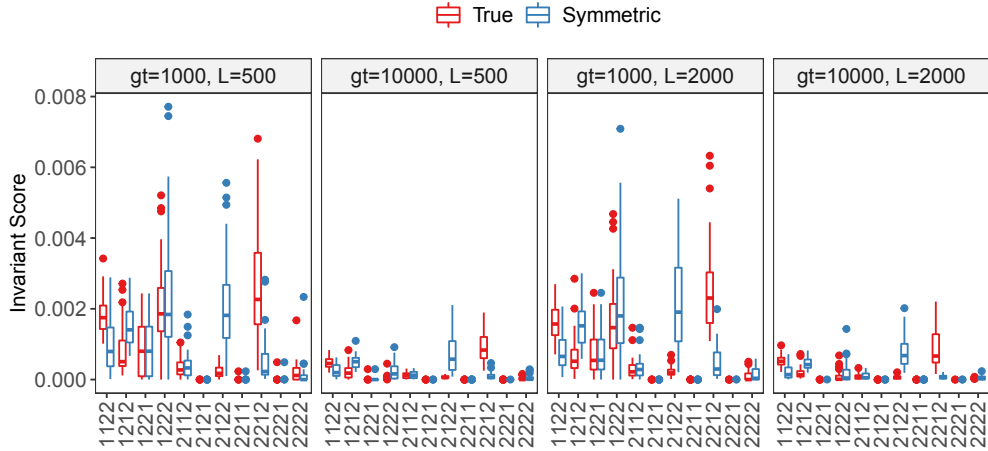


Figure 9: Invariant score (y-axis) measured as L_2 -norm of the phylogenetic invariants for each network (x-axis) for the true network (red) and its symmetric network (inverted clades n_1 and n_2 , in blue) for the case of estimated gene trees. Each panel corresponds to a number of gene trees (g.t. from 1000 to 10,000) and sequence length (L from 500 to 2000). As the number of gene trees increases, the invariant scores of the true and symmetric networks converge to zero, and they are thus, easy to distinguish from wrong networks (whose invariant score is far from zero).