Algebraic Invariants for Inferring 4-leaf Semi-directed Phylogenetic networks

Samuel Martin[1,*], Vincent Moulton[2] and Richard M. Leggett[1]

[1] *Earlham Institute, Norwich Research Park, Norwich, NR4 7UZ, United Kingdom*

[2] *School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich, NR4 7TJ, United Kingdom*

* To whom correspondence should be addressed. Email: samuel.martin@earlham.ac.uk

Abstract A core goal of phylogenomics is to determine the evolutionary history of a set of species from biological sequence data. Phylogenetic networks are able to describe more complex evolutionary phenomena than phylogenetic trees but are more difficult to accurately reconstruct. Recently, there has been growing interest in developing methods to infer semi-directed phylogenetic networks. As computing such networks can be computationally intensive, one approach to building such networks is to puzzle together smaller networks. Thus, it is essential to have robust methods for inferring semi-directed phylogenetic networks on small numbers of taxa. In this paper, we investigate an algebraic method for performing phylogenetic network inference from nucleotide sequence data on 4-leaved semi-directed phylogenetic networks by analysing the distribution of leaf-pattern probabilities. On simulated data, we found that we can correctly identify with high accuracy semi-directed networks as sequences approach 10Mbp in length, and that we are able to use our approach to identify tree-like evolution and determine the underlying tree. We also applied our approach to published transcriptome data from swordtail fish to

compare its performance with a pseudolikelihood method for inferring semi-directed networks.

Keywords Phylogenetic network, semi-directed network, phylogenetic invariants,

1. Introduction

Phylogenetic networks describe the evolutionary history of taxa where reticulate evolution events, such as hybridization and horizontal gene transfer, have occurred (Bapteste et al. 2013). Biologists are becoming increasingly aware that such events are common in the evolutionary histories of many species, and so the development of methods for constructing phylogenetic networks from biological data is an active area of research.

Over the past decades many methods of phylogenetic network reconstruction have been suggested. One approach is to infer implicit or unrooted networks that do not aim to represent specific biological processes. For example, distance-based methods such as Neighbor-Net (Bryant, Moulton, 2004), construct split networks directly from a distance matrix without the need for sequence data. Other methods construct networks by analysing gene trees (Than et al. 2008), concordance factors (Allman et al. 2019), or quartets (Grünewald et al. 2013) and attempt to find the “best” network that displays these. Several maximum parsimony algorithms have also been developed for phylogenetic networks (Kanna, Wheeler, 2012). An alternative approach is to place an evolutionary model on an explicit network (where each vertex in the network represents a biological event or ancestral species), thereby creating a rooted, directed phylogenetic network (or Bayesian network). Methods such as maximum likelihood e.g. (Wen et al. 2018) or Markov chain Monte

Carlo e.g. (Zhang et al. 2018) can then be used to determine how well a set of data fits a certain model and thereby compute the network.

Recently, there has been increasing interest in inferring semi-directed phylogenetic networks for evolutionary analysis (e.g. (Solís-Lemus, Ané, 2016; Solís-Lemus, Bastide, 2017; Allman et al. 2019; Gross et al. 2021; Linz, Wicke, 2023)). These are networks in which only some of the edges are directed, and in which the directed edges usually indicate reticulate events (see e.g., Figure 1). Ideally, these could be computed by maximising a likelihood function, but for larger networks, performing a full search of the semi-directed model to determine the parameters that maximise a likelihood function is too computationally intensive to be practical. One solution to this problem is to use pseudolikelihood, which is based on the likelihood formulas of the 4-taxon subnetworks, as in (Solís-Lemus, Ané, 2016). Another approach is to build such networks from knowledge of the networks displayed by a small number of taxa. This approach has been used in the past for trees (e.g., (Schmidt et al. 2002)) as well as explicit networks (Oldman et al. 2016) and more recently for semi-directed networks (Huebler et al. 2023). One of the key challenges to this approach is to accurately determine each subnetwork displayed by small numbers of taxa.

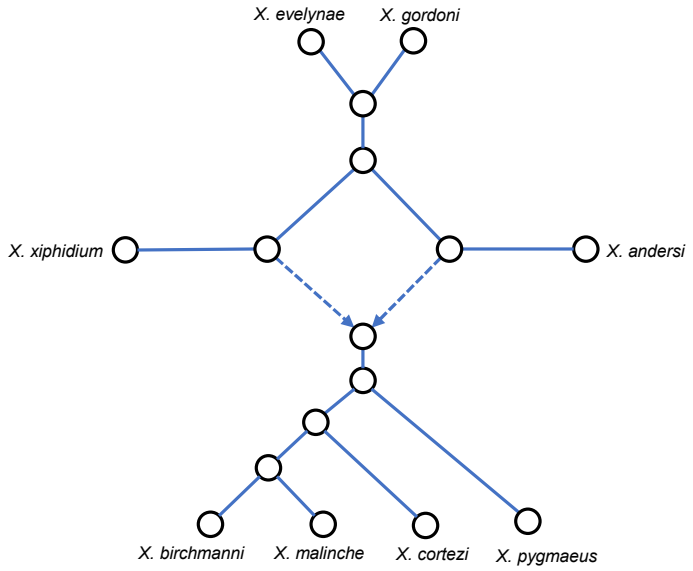

Figure 1. A semi-directed network with eight leaves, labelled by *Xiphophorus* species, constructed from Figure 10 of (Solís-Lemus, Ané, 2016).

For phylogenetic trees, algebraic techniques based on so-called phylogenetic invariants have been used successfully for both the understanding of evolutionary models and for methods of phylogenetic tree inference, e.g. (Casanellas, Fernández-Sánchez. 2007; Chifman, Kubatko 2014). Recently, algebraic methods have also been used to determine when hybridisation between species is likely to have occurred (Blischak et al. 2018) and combined with statistical learning techniques to infer small semi-directed networks (Barton et al. 2022). In this paper, we investigate a new method for determining 4-leaf semi-directed networks that uses algebraic invariants for “group-based” models of evolution, namely, the Jukes-Cantor (JC) model, and the Kimura 2-parameter (K2P) model. As well as performing simulations to understand the performance of the method, we show that it can be used to distinguish between tree-like and non-treelike evolution. We also present results obtained from running it on a data set that has been previously analysed using semi-

directed networks in (Solís-Lemus, Ané, 2016) and separately in (Blischak et al. 2018) to compare its performance with these methods.

2. Materials and Methods

2.1 Background

For a rooted phylogenetic network, its *semi-directed network* is the mixed graph obtained by unrooting the network and undirecting all edges except for the reticulation edges. For the substitution models we define on rooted phylogenetic networks, only the semi-directed network is identifiable from the leaf pattern distribution (also called the marginal character distribution), that is, the distribution of nucleotides observed at the leaves (Gross et al. 2021). This is analogous for the case for phylogenetic trees, where only the unrooted tree topology is identifiable from the leaf pattern distribution.

We place a model of nucleotide substitution on a rooted phylogenetic network in the form of a directed graphical model, by assigning a 4x4 transition matrix to each edge in the network (where each entry in the matrix is the probability of a particular nucleotide substitution occurring along that edge, e.g. an A to a C), a distribution of nucleotides at the root of the network, and for each reticulation vertex, the probability that a particular position is inherited along either reticulation edge. This model allows us to obtain expressions for the probability of observing the leaf patterns (or marginal characters) of the network. Each leaf pattern is a sequence of nucleotides that can be observed at the leaves of the network at a single position in a sequence alignment.

For phylogenetic models, we require that the transition matrix for each edge in a phylogenetic network is given by a matrix P of the form $P = \exp(Qt)$, where Q is a rate matrix for the edge and t is the branch length. To make the model simpler, a single rate matrix Q is sometimes given to each edge in the network, and the branch length is allowed to vary. Such a model is called homogeneous. One of the key features of the models we study in this paper is that they are non-homogeneous, that is, different branches of the network may have different rate matrices Q.

By considering the numerical parameters of the model as free variables, we think of the distribution of leaf patterns as a multi-dimensional polynomial function. These functions are generally complex and difficult to calculate, but they can be simplified if we use certain evolutionary models called “group-based” models (e.g., Sullivant, 2018). For 4-state nucleotide models, there are three well known group-based models: the Jukes Cantor model (JC), the Kimura 2-parameter model (K2P), and the Kimura 3-parameter model (K3P). In each case, a Fourier transform can be applied to the model that simultaneously diagonalises the transition matrices. This makes the new, transformed distribution of marginal characters much simpler, although it is no longer probabilistic. In the case of phylogenetic trees, the new distribution functions are monomial, which makes them especially amenable for algebraic study, and indeed, these models have been well studied from an algebraic perspective (e.g., Sturmfels, Sullivant, 2005; Allman et al. 2011). Furthermore, the parameter space for group-based models is significantly smaller than for general Markov models, since in each group-based model, the form of the rate matrix is restricted. The forms of the rate matrices for the 4-state group-based models are given in Figure 2.

$$\begin{pmatrix} -3\alpha & \alpha & \alpha & \alpha \\ \alpha & -3\alpha & \alpha & \alpha \\ \alpha & \alpha & -3\alpha & \alpha \\ \alpha & \alpha & \alpha & -3\alpha \end{pmatrix}, \begin{pmatrix} -\alpha-2\beta & \beta & \alpha & \beta \\ \beta & -\alpha-2\beta & \beta & \alpha \\ \alpha & \beta & -\alpha-2\beta & \beta \\ \beta & \alpha & \beta & \alpha-2\beta \end{pmatrix},$$

$$\begin{pmatrix} -\alpha-\beta-\gamma & \beta & \alpha & \gamma \\ \beta & -\alpha-\beta-\gamma & \gamma & \alpha \\ \alpha & \gamma & -\alpha-\beta-\gamma & \beta \\ \gamma & \alpha & \beta & -\alpha-\beta-\gamma \end{pmatrix}$$

Figure 2. Form of rate matrices for the JC, K2P, and K3P evolutionary models with nucleotide ordering A, C, G, T.

For these models, any two distinct rooted phylogenetic networks that have the same underlying semi-directed network have the same leaf pattern distribution (Gross, Long, 2018; Gross et al. 2021). We can therefore only infer semi-directed phylogenetic networks from the leaf pattern distribution.

From a phylogenetic network and evolutionary model, one can obtain an algebraic object called a variety. This object can be thought of as a high-dimensional “surface”, and it consists of all possible distributions of leaf patterns (including non-probabilistic ones) from the model. Recent study of these objects has given identifiability results. For the JC model, it was shown that the semi-directed network topology of large cycle networks is generically identifiable from the distribution of leaf patterns (Gross, Long, 2018). Shortly afterwards, analogous results were proven for the Kimura 2-parameter and Kimura 3-parameter evolutionary models (Hollering, Sullivant, 2021), and then for all three evolutionary models on level-1, triangle-free phylogenetic

networks (Gross et al. 2021). Further algebraic properties have been determined for any triangle-free level-1 network under any group-based model (Gross et al. 2023).

Since the semi-directed networks of the phylogenetic networks we are interested in are generically identifiable, the probability that a randomly observed leaf pattern comes from two distinct semi-directed networks is 0, and so the true semi-directed network can be inferred from observation at the leaves. However, this is only true for points that lie exactly on the distribution defined by the model. In reality, data produced from real biological processes will not lie exactly on the distribution, even if the model describes the process exactly and the data is free from errors. In this paper, we adopt the frequentist approach and approximate the distribution of leaf patterns from a multiple sequence alignment. Each column in the alignment is a pattern of nucleotides observed at the leaves of the network, so by counting the number of distinct patterns that occur in the alignment we can approximate the distribution of leaf patterns. Since this is only an approximation, it is unlikely to lie exactly on the distribution defined by the model but it should be “close” in some sense. It is therefore essential to understand how “close” together the distributions defined by distinct models are, in order to know how reliably we can infer the network from data.

Algebraic invariants (also called phylogenetic invariants) are polynomials that evaluate to 0 only on all points of an algebraic variety given by a fixed phylogenetic network and model of evolution. (Note that the term “phylogenetic invariants” is sometimes used to mean only those algebraic invariants that lie in the zero ideal of exactly one semi-directed network. This is not the meaning we use and so we avoid

this term and use “algebraic invariants” instead.) They can be used to infer whether a set of data could have been produced by a given network without the need for parameter estimation, and they have been used to successfully infer 4-leaf trees from simulated data under the K3P model (Casanellas, Fernández-Sánchez. 2007), and 4-leaf networks under the JC model (Barton et al. 2022). As a method, they have several advantages. First, the invariants for a fixed phylogenetic tree or phylogenetic network and model of evolution need only be calculated once. For small trees, many invariants have already been calculated and are available online (Casanellas, Garcia, Sullivant. 2005). Second, using invariants is a statistically consistent method to infer an unrooted phylogenetic tree or semi-directed network topology. Third, once the invariants have been calculated, applying the method to a dataset is simply a case of evaluating a fixed number of polynomials, and so can be performed quicker than many other statistically consistent methods, such as maximum likelihood.

Here, we investigate the *practical identifiability* of semi-directed networks and the effectiveness of using algebraic invariants for network inference under the 4-state group-based models. We have focussed on the 4-leaf 4-cycle network, a directed example of which is depicted in Figure 3. This network has particular biological relevance in that it can represent the evolutionary relationship between two species, their hybridisation, and an outgroup; and it is generically identifiable.

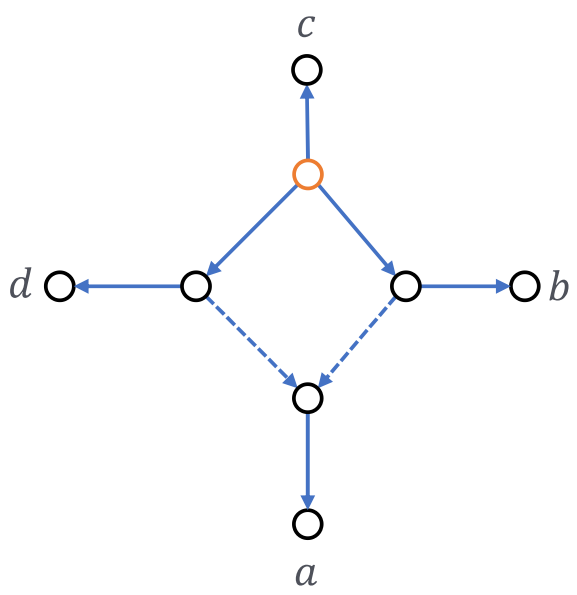

Figure 3. A directed cycle network with cycle of length 4. Dashed edges represent reticulation edges. The root vertex is highlighted in orange.

2.2 Algorithm

We developed an algorithm utilising phylogenetic invariants to infer the correct 4-leaf, 4-cycle network from a multiple sequence alignment (MSA) of 4 taxa. We use the notation $(abcd)$ to denote the 4-leaf 4-cycle network with taxon "a" at the leaf below the reticulation vertex and taxa "b", "c", and "d" at the leaves going anti-clockwise from "a" (as in Figure 2). There are $12 = {}^{4!}/_{2}$ possible 4-leaf 4-cycle networks, since $(abcd)$ and $(adcb)$ are the same. Note that for each network the underlying directed graph is the same, but the taxa labels at the leaves are permuted.

The aim of our algorithm is to determine which of the 4-leaf, 4-cycle networks is most likely to have generated a given MSA. Each of the 4-leaf, 4-cycle networks is represented by a "surface" giving the leaf pattern distributions that can be obtained from that network and substitution model. From an MSA we obtain an approximation

of a leaf pattern distribution, and we use algebraic invariants to determine how close this is to each of the surfaces. A depiction of this process is given in Figure 4.

First, using the software Macaulay2 (Grayson, Stillman), we calculated invariants for the 4-leaf, 4-cycle phylogenetic network depicted in Figure 2, for each of the JC, K2P, and K3P models. Each invariant is a homogeneous polynomial that lies in the vanishing ideal associated to the phylogenetic network (see e.g. (Sullivant, 2018)). After 100 days on a single core of an HP MC990X system, computations for the K2P and K3P models had not finished and were terminated. The number of invariants we obtained, and their total degrees are shown in Table 1.

Table 1. Number of polynomials of each degree in calculated set of invariants.

Degree	JC	K2P	K3P
1	51	29	0
2	2	6	11
3	15	67	-
4	17	-	-
5	14	-	-
6	5	-	-
≥7	1	-	-
Total	105	-	-

For a given MSA, we score each of the twelve networks by permuting the sequences in the MSA and applying invariants to the resulting leaf pattern distribution. The first step is to read the alignment and count the number of columns that occur for each leaf pattern. This allows us to approximate the leaf pattern distribution which we store as a single vector p. Our algorithm next reads in a file of invariants that have undergone the Fourier transformation described above. The next step is to transform the vector p using the same Fourier transformation and evaluate each invariant at the transformed vector q. This gives us a list of numbers from which a score for the

corresponding network is given. As in (Casanellas, Fernández-Sánchez, 2015) we found that scoring using the 1-Norm gave us the best results. In this case, the score for network $\mathcal{N}$ is given by the formula

$$S_{\mathcal{N}} = \sum_{f \in G} |f(q)|,$$

where G is a set of invariants, and q is the transformed data point obtained from the permuted MSA. The entries of q are small real numbers close to 0, so we expect $|f(q)|$ to be smaller when the degree of f is higher, and it follows that the invariants with higher degree will contribute less to the score than those with lower degree. We therefore chose sets of invariants so that each invariant in the set had the same total degree. Whilst this raises a mathematical question about identifiability (that is, we do not know whether these invariants are enough to uniquely determine the semi-directed network topology from a generic marginal character distribution), initial experimentation for the JC model showed that this gave us better results than using the complete set of invariants. Once each network has been scored, the networks are ordered by score in ascending order, and the network with the smallest score is chosen as the “correct” network.

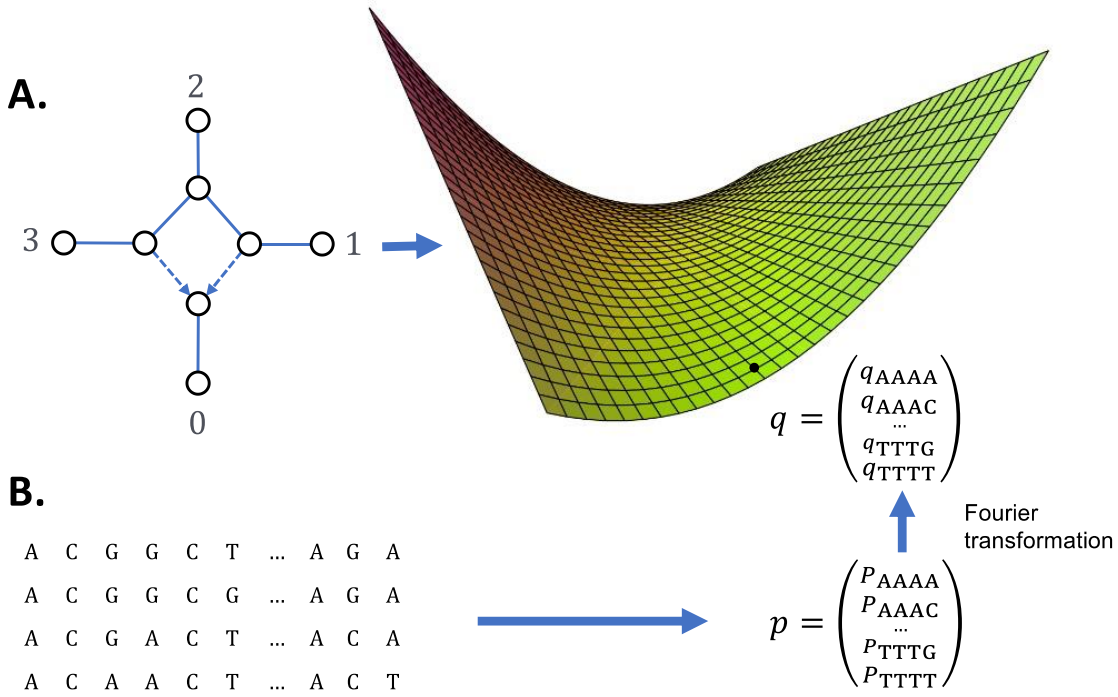

Figure 4. **A.** Each leaf-labelled, semi-directed network is converted into a high-dimensional "surface" in a high-dimensional space (depicted here as a 2D surface in a 3D space), representing all leaf pattern distributions possible from the model. **B.** A multiple sequence alignment is converted into a point q in this space, and this lies exactly on the surface (i.e., the alignment could have been generated by the model) if and only if $f(q) = 0$ for all invariants f. Since each f is a continuous polynomial, for points q close to the surface, $f(q)$ will be close to 0. Depiction of the surface was created using CalcPlot3D available at https://c3d.libretexts.org/CalcPlot3D/index.html.

We implemented this algorithm and assessed its performance on MSA data simulated under 4-leaf 4-cycle networks, and the JC and K2P evolutionary models. We also applied our method to real transcriptome data from 24 swordtail fish and platyfish species (genus Xiphophorus) and two outgroups (*Pseudoxiphophorus*

jonesii and *Priapella compressa*). The transcriptome data was generated in (Cui et al. 2013) and alignments were kindly provided to us by the authors of (Blischak et al. 2018).

3. Results

In this section all results are obtained using the 67 degree 3 invariants from the K2P model to perform network inference as described in Section 2. Since the JC model is a submodel of K2P, these invariants are also invariants for the JC model. This again raises a question of identifiability, but our results suggest that the semi-directed network is still identifiable under the JC model using these invariants.

3.1 Non-homogeneous simulated data

We generated multiple sequence alignment data from each of the twelve distinct leaf permutations of the directed network depicted in Figure 2, each of which has a distinct semi-directed network. For each network we generated 100 MSAs of lengths 1kbp, 10kbp, 100kbp, 1mbp, and 10mbp under both the JC and K2P models. Substitution rates were randomly generated for each edge between 0 and 0.1 for JC, and 0 and 0.15 for K2P. The tree ratio (γ) was fixed at 0.5. Figure 5.A shows the confusion matrices for each of these datasets for the JC model, and Figure 5.B shows the confusion matrices for the K2P model.

We found a clear distinction between scores for the true network, and scores for the other networks. Figure 5.C shows the distribution of scores obtained for each network when the true network was (0123), for each alignment length.

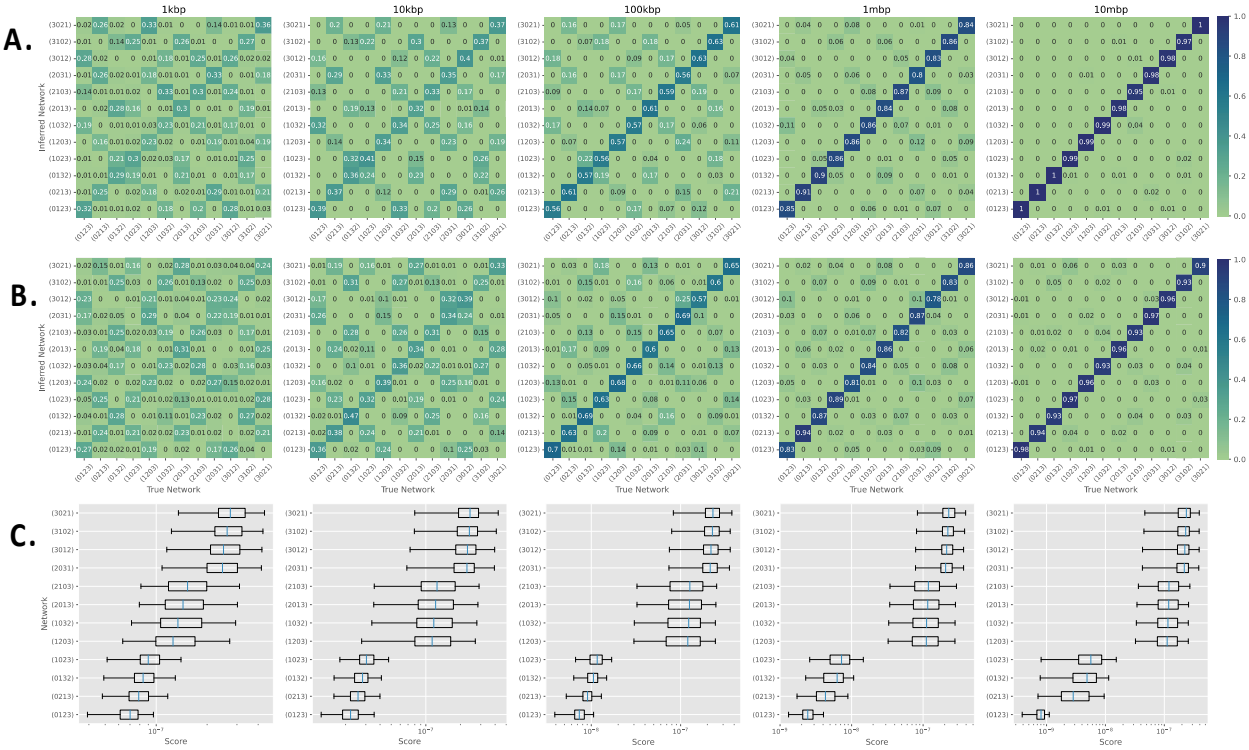

Figure 5. **A.** Confusion matrices for inference of 4-leaf, 4-cycle networks from data simulated under the Jukes-Cantor (JC) model. **B.** Confusion matrices for inference of 4-leaf, 4-cycle networks from data simulated under the Kimura 2-parameter (K2P) model. **C.** Distribution of scores from data generated by the network (0123) under the JC model.

3.2 Varying the tree ratio

We generated further datasets where the 4-leaf, 4-cycle network was fixed and the tree ratio γ was varied from 0.0 to 1.0 in intervals of 0.05. As before, we generated 100 MSAs of length 1kbp, 10kbp, 100kbp, 1mbp, and 10mbp for both JC and K2P models. Figure 6 shows the number of correctly inferred networks in each case for the JC model (6.A) and K2P model (6.B). Observe that when $\gamma = 0$ or 1 we obtain a low percent of correctly identified networks regardless of MSA length.

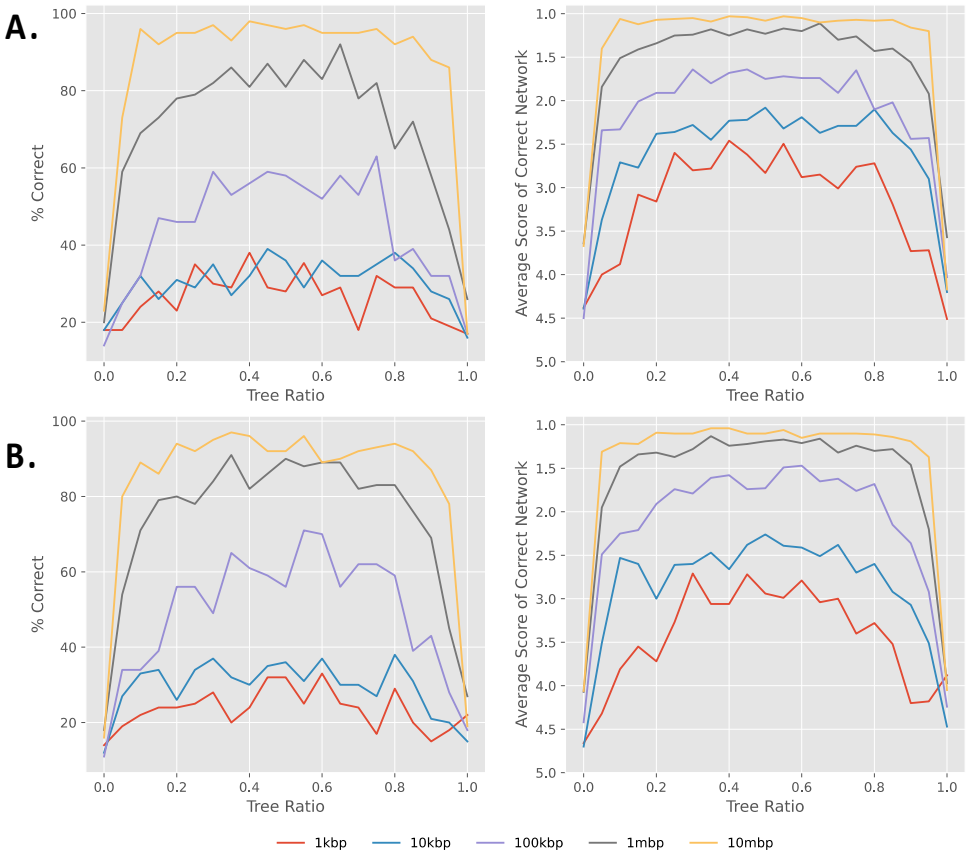

Figure 6. **A.** Percent of tree-ratio datasets where the network was correctly identified (left) and average score over all datasets that the correct network achieved (right) for JC. **B.** Analogous plots for K2P datasets.

3.3 Identification of Treelike versus Non-treelike Evolution

Our efforts so far have focussed on identifying the underlying 4-leaf, 4-cycle networks from a multiple sequence alignment. However, in many cases, it may not be known whether the evolution of a set of taxa has been tree-like or not. In this section, we focus on using our invariants-based method to help understand whether data from four taxa has been generated from a tree (treelike) or a 4-leaf, 4-cycle network (non-treelike).

To do this, we make the following observation: For a fixed model of evolution, the model for a 4-leaf unrooted tree is contained in precisely eight 4-leaf, 4-cycle, network models, and is not contained in the remaining 4. Therefore, if a set of data is generated by a tree, we expect the score that this data obtains via our invariants method to be low for eight of the networks, and high for the remaining 4 networks. Furthermore, from the 4 networks for which we expect a high score we can determine the tree that generated the data.

We found that this signature is identifiable even for short alignment lengths (see Figure 7.C), so we updated our algorithm to look for this signature and re-evaluated the simulated data from section 3.2. In this case, when $\gamma = 0$, evolution has been treelike, along a tree which we refer to as tree 1. When $\gamma = 1$, evolution has also been treelike, along a different tree which we refer to as tree 2. In all other cases evolution has been non-treelike. Figure 7.A shows that for the data simulated under the JC model, we achieve a 100% TP rate and 0% FP rate for alignments of length 10mbp. Figure 7.B shows a similar picture for the data simulated under the K2P model.

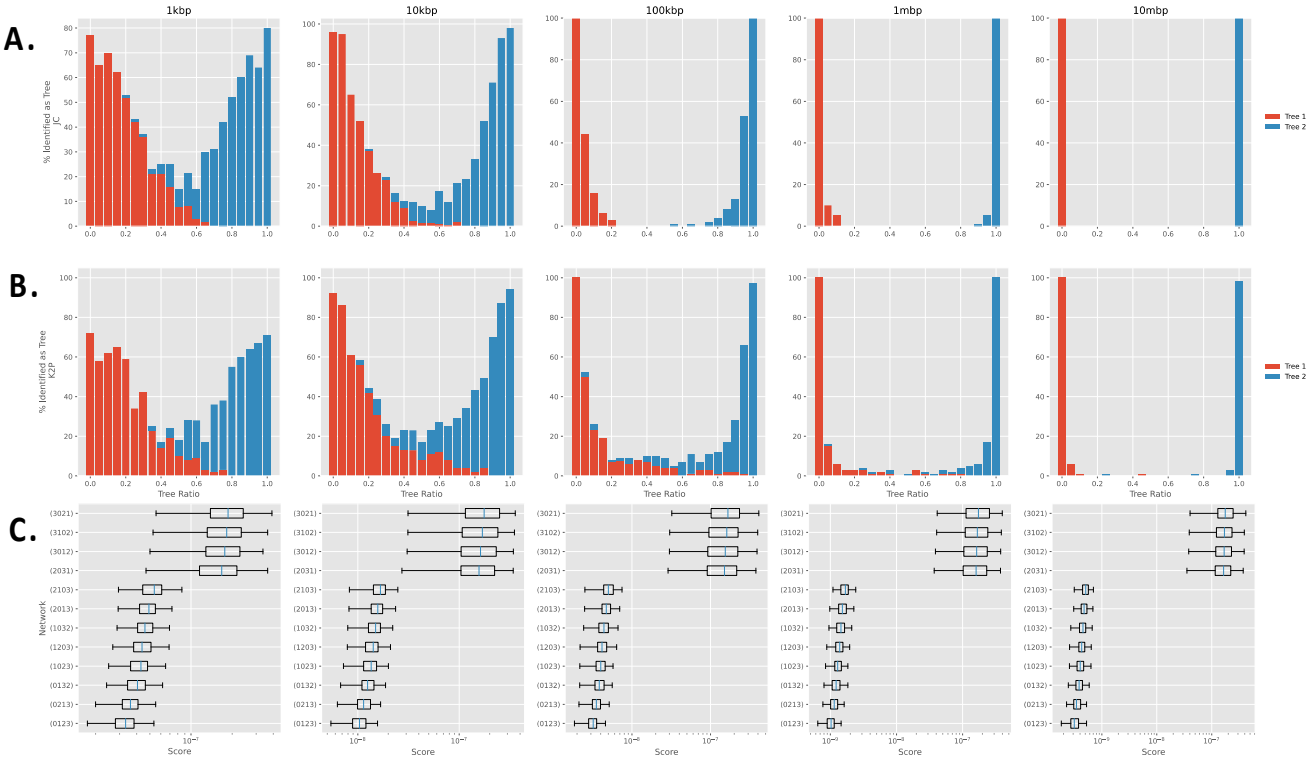

Figure 7. A. Percent of JC datasets identified as evolving along either tree 1 or tree 2. **B.** Percent of K2P datasets identified as evolving along either tree 1 or tree 2. **C.** Scores assigned to networks generated along tree 1 (i.e. $\gamma = 0$) under the JC model. The low scores of eight of the networks are clearly identifiable.

3.4 *Inference of networks from real datasets with many taxa*

We used our method to evaluate each subset of four taxa from a multiple sequence alignment of transcriptome data from 25 swordtail fish and platyfish species (genus Xiphophorus) and two outgroups (*Pseudoxiphophorus jonesii* and *Priapella compressa*) giving a total of 17,553 subsets.

Each of the Xiphophorus species belongs to one of three distinct clades; southern swordtails, northern swordtails, and platyfishes (split further into southern platyfishes and northern platyfishes). We concatenated all alignments of all 27 taxa and

extracted the concatenated alignment for each subset of four taxa. Each of these subsequent alignments was analysed by our invariants script, which ignores columns in the alignment containing the gap character "-". Without gaps, alignments between subsets of four-taxa ranged between 180kbp and 3.37mbp.

Of the 17,553 subsets, 131 had no discernible signal, that is, the difference between the highest and lowest scores was too small to be considered significant (all scores were within 10% of the lowest score) and were rejected from further analysis. Of the remaining subsets 2,557 were identified as having tree-like evolution, suggesting that there has been widespread hybridisation between Xiphophorus species, as demonstrated in previous analyses of this dataset (Cui et al. 2013; Solís-Lemus and Ané 2016; Blischak et al. 2018).

Following (Solís-Lemus and Ané, 2016) and (Blischak et al. 2018), we looked for support for an ancestral hybridisation event between *X. xiphidium* and the northern swordtail clade. We examined all 4-subsets consisting of *X. xiphidium,* a single taxon from the northern swordtail clade, and two further Xiphophorus taxa that weren't from the northern swordtail clade, giving us 708 subsets. Of these, 55 were determined to have undergone tree-like evolution, and 545 had a network consistent with the ancestral hybridisation event in the top four scoring networks for those taxa. Of these 545 subsets, 127 had either a best scoring network consistent with this hybridisation event, or a second-best scoring network consistent with it, and the second-best score within 1% of the best score. Interestingly, the nine networks containing a single northern swordfish below the reticulation, and *X. xiphidium, X. evelynae,* and *X. andersi* adjacent to the other cycle vertices had the lowest scores. Some of these

are displayed by the semi-directed network in Figure 1, and the three lowest scoring networks are given explicitly in Figure 8.

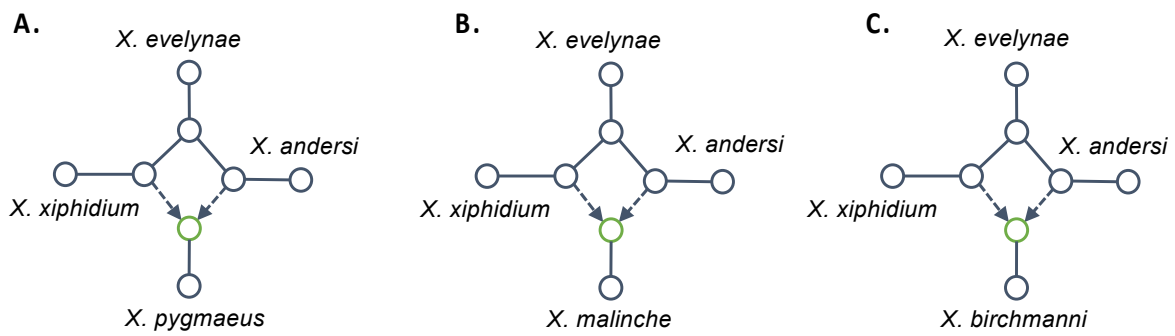

Figure 8. The lowest scoring networks from 4-subsets consisting of *X. xiphidium,* a single taxon from the northern swordtail clade, and two further Xiphophorus taxa that weren't from the northern swordtail clade.

3.5 *Timings*

Figure 9 shows the time taken for the analysis of the simulated data from Section 3.1 and the real data from Section 3.4. Each analysis was run on a single CPU with 32GB of RAM. Since each dataset is evaluated on a fixed set of invariants (that have been pre-computed and are stored in a text file), most of the time is taken on reading the alignments to obtain the frequencies of occurrence for each leaf pattern, and then performing a Fourier transform of this data. The time therefore scales with the length of the alignment. Shorter alignments perform quickly (<1 minute), but longer alignments can take several hours.

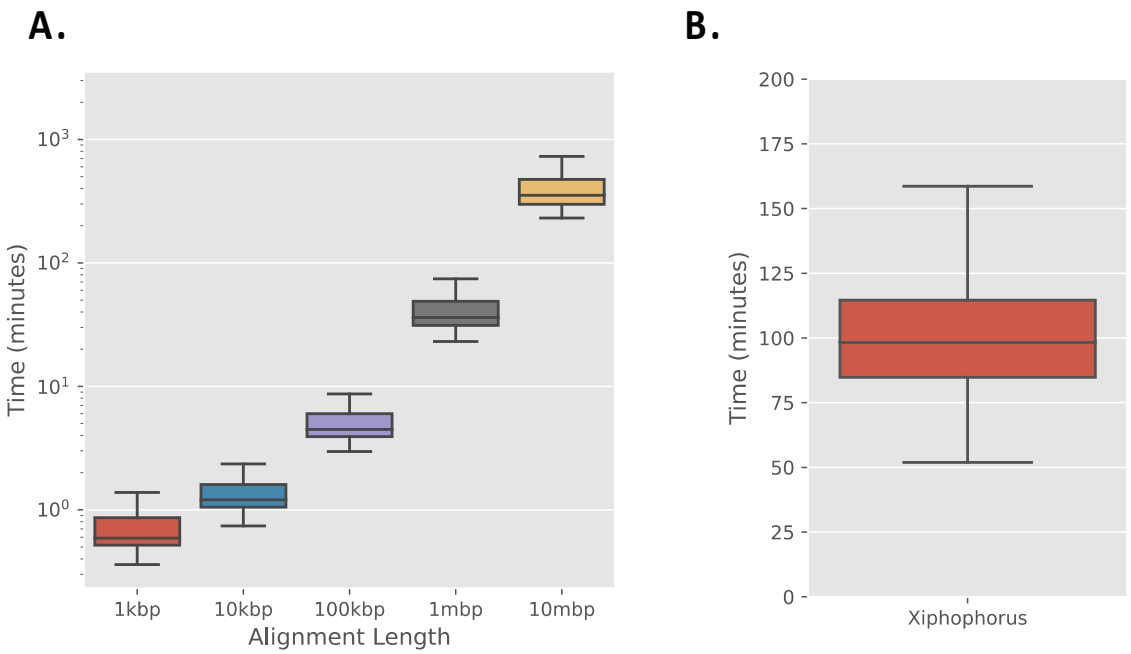

Figure 9 A. Timings for simulated JC data split by alignment length. **B.** Timings for all Xiphophorus datasets, where alignments (without gaps) ranged between 180kbp and 3.37mbp.

4 Discussion

We have developed a novel method for inferring a semi-directed network topology from aligned sequence data between four taxa. We demonstrate its use in identifying 4-leaf, 4-cycle networks from simulated data, and in identifying whether hybridisation is likely to have occurred. We have shown that the method performs well on simulated data and converges to 100% accuracy with alignments of length 10mbp. Furthermore, we show that our method can detect when evolution between taxa has been treelike, converging quickly to a high true positive rate and low false negative rate as alignment length increases. We also applied our method to aligned transcriptome data from swordfish species and found previously identified hybridisation events.

On simulated data we found that the method performed significantly better on longer sequences. This is no surprise since the method we have used is known to be statistically consistent. However, we observe a rate of convergence that is much less than the analogous rate for trees. In (Casanellas, Fernández-Sánchez. 2007), the authors observe almost 100% accuracy for alignments of length 10kbp on 4-leaf trees under the K3P model. For 4-leaf, 4-cycle networks under the K2P model, we do not achieve close to 100% accuracy until alignment lengths are in the order of 10mbp, although Figure 5C suggests that for a subset of 4 taxa displaying 4-cycle evolution, the true network representing these is likely to be in the top 4 for alignments of length 10kbp. In both studies, the space from which we are sampling is the same, and whilst there are only three 4-leaf trees, there are twelve 4-leaf, 4-cycle networks. We conjecture that the leaf-pattern distribution varieties of the 4-leaf 4-cycle networks are “close together” in some sense, which will make inference difficult for any method. Indeed, the varieties corresponding to any two of our 4-leaf-4-cycle networks contain exactly one variety corresponding to a phylogenetic tree in their intersection, and so for data where the tree ratio is close to 0 or 1, the true semi-directed network topology will be more difficult to infer. This can be observed in Figure 7.

On real data we were able to identify reticulation events consistent with previous studies, although in this case the placement of some taxa was unclear, and in some cases the scores from all possible networks were very close together. This is likely due to inaccuracy of the K2P model for this dataset, as well as noise that is present in real data.

Our tool performs moderately quickly on all datasets, with time growing with alignment length. The computations that take the most time are calculating the frequencies of each leaf pattern and performing a linear transformation of these frequencies. Both tasks are parallelisable, and implementing this could increase the speed by up to 12x, although we have not explored this yet. The remaining time is spent evaluating the polynomial invariants on the transformed frequency data, and this is also parallelisable. Thus, there is potential for our tool to be significantly faster. The reason we can perform network inference relatively quickly is that the most difficult computations (computing the invariants of the networks) need only be done once. We have already done them and distributed the results with the tool.

One of the biggest challenges of this work was calculating invariants. We used methods in elimination theory to find a Gröbner basis with the software Macaulay2, but this does not scale well. Indeed, we were unable to calculate higher degree invariants for the K3P model. This model is more versatile than K2P and JC, so we hope our results provide motivation for developing better methods of calculating invariants in this case. In (Cummings et al. 2021), the authors reduce the calculations for finding quadratic invariants for sunlet networks under the Cavender-Farris-Neyman (CFN) 2-state model to finding the kernel of a linear map. The result is a much faster method of calculating invariants than using Gröbner basis methods, and may be extendable to higher degree invariants and other group-based models. However, even the K3P model is somewhat simplistic, so we would like to be able to calculate invariants for more complex substitution models such as the generalised time reversible (GTR) model, or models that incorporate a molecular clock. Work in

this direction has been recently performed for the CFN model on phylogenetic trees in (Coons, Sullivant, 2021).

Our method looked only at small networks with 4-taxa. In principle, one can apply the same method to larger networks with more taxa, but the problem of calculating invariants for larger networks is currently intractable. Alternatively, one can construct a larger network by considering the networks on smaller numbers of taxa, e.g. all 4-leafed networks in (Huber et al. 2018). An algorithm to do this for phylogenetic trees based on quartet puzzling and using phylogenetic invariants was given in (Rusinko, Hipp, 2012). For group-based models, 3-cycle networks with 4 taxa are not fully identifiable, however it is possible to identify when a 3-cycle has occurred, as in (Barton et al. 2022), thus there is some hope of developing a method that can build a phylogenetic network from MSA data on many taxa, by examining all the 4-taxa subsets. A similar approach was taken in (Oldman et al. 2016) by identifying the subnetworks on 3-taxa.

References

[Bapteste et al. 2013] E. Bapteste, L. van Iersel, A. Janke, S. Kelchner, S. Kelk, J.O. McInerney, D.A. Morrison, L. Nakhleh, M. Steel, L. Stougie, J. Whitfield. Networks: expanding evolutionary thinking. *Trends Genet.* 29(8):439-441, 2013.

[Bryant, Moulton, 2004] D. Bryant, V. Moulton. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* 21(2):255-265. 2004.

[Than et al. 2008] C. Than, D. Ruths. L. Nakhleh. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics*. 9:32, 2008.

[Allman et al. 2019] E.S. Allman, H. Baños, J.A. Rhodes. NANUQ: A method for inferring species networks from gene trees under the coalescent model. *Algorithms Mol. Biol.* **14**,24, 2019.

[Grünewald et al. 2013] S. Grünewald, A. Spillner, S. Bastkowski, A. Bogershausen, V. Moulton. SuperQ: Computing supernetworks from quartets. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 10(1):151-160, 2013.

[Kanna, Wheeler, 2012] L. Kannan, W.C. Wheeler. Maximum parsimony on phylogenetic networks. *Algorithms Mol. Biol.* 7:9, 2012.

[Wen et al. 2018] D. Wen, Y. Yu, J, Zhu, L. Nakhleh. Inferring phylogenetic networks using PhyloNet. *Syst. Biol.* 67(4):735-740, 2018.

[Zhang et al. 2018] C. Zhang, H.A. Ogilvie, A.J. Drummon, T. Stadler. Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.* 35(2):504-517, 2018.

[Solís-Lemus, Ané, 2016] C. Solís-Lemus, C. Ané. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12(3): e1005896, 2016.

[Solís-Lemus, Bastide, 2017] C. Solís-Lemus, P. Bastide, C. Ané. PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.* 34(12):3292-3298, 2017.

[Linz, Wicke, 2023] S. Linz, K. Wicke. Exploring spaces of semi-directed phylogenetic networks. *arXiv* 2301.10818, 2023.

[Gross et al. 2021] E. Gross, L. van Iersel, R. Janssen, M. Jones, C. Long, Y. Murakami. Distinguishing Level-1 Phylogenetic Networks on the Basis of Data Generated by Markov Processes. *J. Math. Biol.* 83:32, 2021.

[Chifman, Kubatko. 2014] J. Chifman, L. Kubatko. Quartet inference from SNP data under the coalescent model. *Bioinformatics*. 30(23):3317-3324, 2014.

[Oldman et al. 2016] J. Oldman, T. Wu, L. van Iersel, V. Moulton. TriLoNet: Piecing together small networks to reconstruct reticulate evolutionary histories. *Mol. Biol. Evol*. 33(8):2151-2162, 2016.

[Schmidt et al. 2002] H.A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics.* 18(3):502-504, 2002.

[Huebler et al. 2023] S. Huebler, R. Morris, J. Rusinko, Y. Tao. Constructing semi-directed level-1 phylogenetic networks from quarnets. *arXiv* 1910.00048, 2023.

[Gross, Long, 2018] E. Gross, C. Long. Distinguishing Phylogenetic Networks. *SIAM J. Appl. Algebra Geometry.* 2(1):72-93, 2018.

[Sturmfels, Sullivant, 2005] B. Sturmfels, S. Sullivant. Toric ideals of phylogenetic invariants. *J. Comput. Biol*. 12(4):457-481, 2005.

[Allman et al. 2011] E.S. Allman, S. Petrović, J.A. Rhodes, S. Sullivant. Identifiability of Two-Tree Mixtures for Group-Based Models. *TCBB.* 8(3)710-722, 2011.

[Hollering, Sullivant. 2021] B. Hollering, S. Sullivant. Identifiability in Phylogenetics using Algebraic Matroids. *J. Symb. Comput.* 104:142-158, 2021.

[Gross et al. 2023] E. Gross, R. Krone, S. Martin. Dimensions of Level-1 Phylogenetic Networks. *Manuscript in Preparation*, 2023.

[Casanellas, Fernández-Sánchez. 2007] M. Casanellas, J. Fernández-Sánchez. Performance of a New Invariants Method on Homogeneous and Nonhomogeneous Quartet Trees. *Mol. Biol. Evol.* 24(1): 288-293, 2007.

[Barton et al. 2022] T. Barton, E. Gross, C. Long, J. Rusinko. Statistical learning with phylogenetic network invariants. *arXiv* 2211.11919, 2022.

[Casanellas, Garcia, Sullivant. 2005] M. Casanellas, L.D. Garcia, S. Sullivant. Catalog of Small Trees. In: L. Pachter, B. Sturmfels, editors. Algebraic statistics for computational biology. Chapter 15. Cambridge University Press, 2005.

[Grayson, Stillman] D.R. Grayson, M.E. Stillman. Macaulay2, a software system for research in algebraic geometry. Available at http://www.math.uiuc.edu/Macaulay2/.

[Sullivant, 2018] S. Sullivant. Algebraic Statistics. *Graduate Studies in Mathematics*, Volume 194. American Mathematical Society, 2018.

[Cui et al. 2013] R. Cui, M. Schumer, K. Kruesi, R. Walter, P. Andolfatto, G.G. Rosenthal. Phylogenomics reveals extensive reticulate evolution in Xiphophorus fishes. *Evol.* 67(8): 2166-2179, 2013.

[Blischak et al. 2018] P.D. Blischak, J. Chifman, A.D. Wolfe, L.S. Kubatko. HyDe: A python package for genome-scale hybridization detection. *Systematic Biology.* 67(5): 821-829, 2018.

[Coons, Sullivant, 2021] J.I. Coons, S. Sullivant. Toric Geometry of the Cavender-Farris-Neyman model with a molecular clock. *Adv. Appl. Math.* 123:102119, 2021.

[Cummings et al. 2021] J. Cummings, B. Hollering, C. Manon. Invariants for level-1 phylogenetic networks under the Cavendar-Farris-Neyman Model. *arXiv*. arXiv:2102.03431v1, 2021.

[Huber et al. 2018] K.T. Huber, V. Moulton, C. Semple, T. Wu. Quarnet Inference Rules for Level-1 Networks. *Bull. Math. Biol*. 80: 2137-2153, 2018.

[Rusinko, Hipp, 2012] J.P. Rusinko, B. Hipp. Invariant based quartet puzzling. *Algorithms Mol. Biol.* 7(35), 2012.

Funding

This work was supported by EPSRC Mathematical Sciences Small Grant award EP/W007134/1. SM and RL are supported by the Biotechnology and Biological Sciences Research Council (BBSRC), part of UK Research and Innovation, through the Core Capability Grant BB/CCG1720/1 at the Earlham Institute and are grateful for further funding from BBSRC (grant number BB/X005186/1).

Acknowledgments

The authors are grateful to Paul Blischak for providing alignment data of the Xiphophorus species, and for HPC support from NBI's Research Computing group.

Data availability

All scripts used in this project, including scripts for simulating data and Macaulay2 scripts for calculating invariants, are available at the GitHub repository https://github.com/SR-Martin/4cycle_invariants.