Building explicit hybridization networks using the maximum likelihood and Neighbor-Joining approaches

Matthieu Willems, Nadia Tahiri and Vladimir Makarenkov

Abstract Tree topologies are the simplest structures which can be used to represent the evolution of species. Over the two last decades more complex structures, called phylogenetic networks, have been introduced to take into account the mechanisms of reticulate evolution, such as species hybridization and horizontal gene transfer among bacteria and viruses. Several algorithms and software have been developed in this context, but most of them yield as output only an implicit network, which can be difficult to interpret. In this paper, we introduce a new algorithm for inferring explicit hybridization networks from binary data. In order to build our explicit hybridization networks, we use a maximum likelihood approach applied to Neighbor-Joining tree configurations.

Matthieu Willems
Département d'informatique, Université du Québec à Montréal, Case postale 8888, Succursale Centre-ville, Montréal (Québec) H3C 3P8, Canada
✉ matthieu.willems@polytechnique.org

Nadia Tahiri
Département d'informatique, Université du Québec à Montréal, Case postale 8888, Succursale Centre-ville, Montréal (Québec) H3C 3P8, Canada
✉ tahiri.nadia@uqam.ca

Vladimir Makarenkov
Département d'informatique, Université du Québec à Montréal, Case postale 8888, Succursale Centre-ville, Montréal (Québec) H3C 3P8, Canada
✉ makarenkov.vladimir@uqam.ca

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 4, No. 1, 2018

DOI 10.5445/KSP/1000085951/04
ISSN 2363-9881

Our algorithm takes as input a set of n binary sequences, e.g., presence/absence of restriction sites or presence/absence of certain genes in genomes, corresponding to a set of n species. We obtain as output a hybridization network in which terminal nodes represent the n input species and the hybrids are explicitly identified among them. The new algorithm was tested on various simulated and real datasets, and its efficiency was compared to a distance-based method developed in our previous study. Overall, the new algorithm provided better hybrid recovery results in terms of true positive and false positive rates than the distance-based method. The main novelty of our method is that it allows one to reconstruct explicit hybridization networks by combining both the distance (Neighbor-Joining) and maximum likelihood reconstruction approaches. It also provides the respective contributions of all parents to hybrids.

1 Introduction

Topological discordance among gene trees representing the evolution of a given set of species is commonly attributed to different reticulate evolutionary processes, including species hybridization, horizontal gene transfer, ancient gene duplication, gene loss and incomplete lineage sorting (Huson and Bryant, 2006). Interspecific gene exchange has been well documented and is very frequent across many groups of animals, plants and bacteria. Reticulate evolutionary processes, not following vertical inheritance of genetic material, cannot be adequately represented by traditional phylogenetic trees such as a species tree or the tree of life. Phylogenetic networks should be used instead to represent these evolutionary events.

Many efforts in the field of phylogenetics have been dedicated to the inference and statistical validation of phylogenetic trees, while effective methods and user-friendly software for reconstructing phylogenetic networks still remain limited or under development (Solís-Lemus and Ané, 2016). The main disadvantage of many hybridization networks building methods is that the networks they provide are rather implicit than explicit, thus rarely allowing an accurate identification of hybrid species and their parents (Willems et al, 2014). Several attempts to model reticulate evolutionary relationships using phylogenetic networks have been made by a number of research groups around the world. Bandelt and Dress (1992) were among the first authors to do it. They described the split

decomposition method that allows for data representation under the form of a split graph that reveals conflicting signals hidden in the data. Bryant and Moulton (2004) continued the work of Bandelt and Dress (1992) by proposing the NeighborNet network-building algorithm that reconstructs planar split networks. Despite the popularity of the split decomposition and NeighborNet methods, the networks they generate usually contain a very high number of splits from which explicit reticulation events cannot be deduced easily. Huson and Bryant (2006) reviewed the terminology and interpretations used when defining different types of phylogenetic networks. The authors showed how a split network can depict confidence sets of trees and introduced a statistical test for determining whether the conflicting signal in a network is tree-like. Huson and Bryant (2006) have also developed the popular SplitsTree program allowing for inferring different types of phylogenetic networks from sequences, distances and trees. Huson and Klöpper (2007) designed an algorithm to infer recombination events from binary sequences by using general reticulation networks and galled trees (explicit networks). Huson and Scornavacca (2012) developed the Dendroscope 3 program to study rooted phylogenetic trees and networks. This program includes a number of algorithms for drawing and comparing rooted phylogenetic networks, most of which are implicit, as well as for inferring them from a set of rooted trees. Albrecht et al (2012) developed a fast parallel algorithm to infer a minimum hybridization network from two input trees. Chen and Wang (2012) described an algorithm for constructing implicit phylogenetic networks from multiple conflicting gene trees.

Most of the existing network-building algorithms rely on the distance-based approach, but recently some new methods based on the maximum likelihood approach which usually provides more accurate results have started to appear. For instance, Solís-Lemus and Ané (2016) presented a new method for inferring explicit pseudolikelihood phylogenetic networks from multi-locus genetic data. The main advantage of their method is that it accounts for incomplete lineage sorting in the framework of a coalescent model as well as for horizontal inheritance of genes through reticulation nodes in the network. This method proceeds by calculating the concordance factor of any given quartet (or split) of species that is the proportion of genes whose true tree displays that quartet. Olave et al (2018) have also proposed a method to detect hybridization explicitly in the presence of incomplete lineage sorting by evaluating the likelihood of various models with different levels of gene flow and assessing the expected

gene tree discrepancy. It is worth noting that the methods of Solís-Lemus and Ané (2016) and Olave et al (2018) build hybridization, or gene flaw, networks from sets of different gene trees. In this article, we extend our previous distance-based algorithm for inferring explicit hybridization networks (Willems et al, 2014) by considering a maximum likelihood approach which will be applied to Neighbor-Joining (Saitou and Nei, 1987) tree configurations.

2 Methods

In this section, we recall the terminology which is necessary to define our new algorithm.

Trees and networks
A phylogenetic network is a graph used to represent evolutionary relationships between a set of species which are associated with some of the graph nodes. A phylogenetic network is said to be *explicit* if it explicitly represents some reticulate evolutionary events, such as hybridization or horizontal gene transfer, in a way that hybrid species and their parents (for a hybridization event) or donors and recipients of genetic material (for a horizontal gene transfer) can be identified explicitly. A phylogenetic network is said to be *implicit* if it only allows visualization of certain evolutionary incompatibilities, usually compared to a phylogenetic tree, without explicitly identifying the species involved in reticulate evolutionary events and their roles in these events.

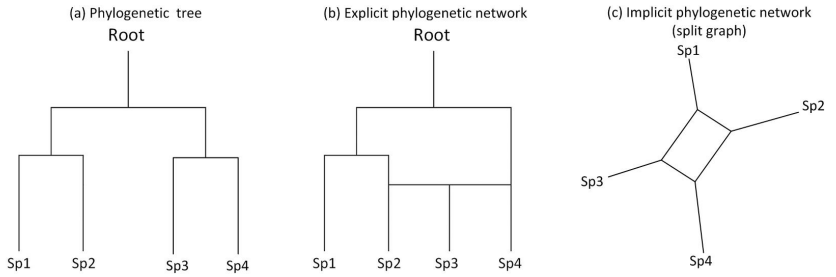

Figure 1: (a) A rooted phylogenetic tree; (b) A rooted phylogenetic (hybridization) network - here, Species 3 is a hybrid of Species 2 and 4; (c) An implicit phylogenetic network (split graph).

Figure 1 presents examples of a traditional phylogenetic tree defined on a set of 4 species (case a), of an explicit phylogenetic (hybridization) network (here Species 3 is a hybrid of Species 2 and 4, case b), and of a split graph, which is an implicit phylogenetic network (case c).

Neighbor Joining (NJ)
Neighbor Joining (Saitou and Nei, 1987) is the most popular distance-based algorithm for inferring phylogenetic trees. Starting from a star tree, this clustering algorithm selects at each step the best neighbors i and j (according to the minimum evolution criterion) and replaces them by their common ancestor X. The minimum evolutionary criterion states that the optimal tree is the tree with the shortest total sum of edge lengths. The NJ algorithm requires as input a matrix of evolutionary distances between species at hand and returns as output a tree metric matrix (i.e., a metric that can be uniquely represented by a tree). NJ infers the correct phylogenetic tree if the input distances between species are sufficiently close to the true evolutionary distances.

Maximum likelihood
The maximum likelihood in a phylogenetic context can be defined as follows.

Input: n binary sequences of length L corresponding to n species.

There exist several evolutionary models based on Markov processes. Let $\Pr(t, N_1, N_2)$ be the probability that character N_1 (DNA, amino acid or binary) changes into character N_2 during evolutionary time t. The likelihood of a tree T can be given by the following formula:

$$\mathcal{L}(T) = \prod_{l=1}^{L} \mathcal{L}_l(T), \tag{1}$$

where $\mathcal{L}_l(T)$ is the sum of the probabilities of all possible evolutionary scenarios at position l.

F81 Model (Felsenstein (1981))
We will use the F81 evolutionary model originally defined by Felsenstein because it perfectly suits for describing the evolution of binary sequences (e.g., presence/absence of restriction sites or presence/absence of certain genes in genomes). Let π_0 (respectively π_1) be the proportion of 0's (respectively, 1's) in the input data. If $\beta = \frac{1}{1-{\pi_0}^2-{\pi_1}^2}$, then the transition probabilities of the F81

Markov process are given by the following formulas, where t is the evolutionary time between two binary sequences:

$$\begin{cases} \Pr(t,0,0) = e^{-\beta t} + \pi_0 \left(1 - e^{-\beta t}\right), \\ \Pr(t,1,1) = e^{-\beta t} + \pi_1 \left(1 - e^{-\beta t}\right), \\ \Pr(t,0,1) = \pi_1 \left(1 - e^{-\beta t}\right), \\ \Pr(t,1,0) = \pi_0 \left(1 - e^{-\beta t}\right). \end{cases} \tag{2}$$

Probability vectors
For each species i, we consider its binary sequence as a probability vector of dimension L (probability of having 1 at position l): $P(i,l) = 0$ (respectively, 1), if the lth character of sequence i is equal to 0 (respectively, 1).

Likelihood of an NJ tree
An NJ tree configuration is shown in Figure 2. Species 1 and 2 are neighbors. The intermediate Species X and Y correspond to the two internal nodes. The maximum likelihood function to be used to estimate the likelihood of this configuration T (i.e., configuration considered at a certain step of the NJ algorithm) can be defined as follows:

$$\mathcal{L}_{1,2}^{T} = \prod_{l=1}^{L} \left(\sum_{(\epsilon_X, \epsilon_Y) \in \{0,1\}^2} \left(P_{1X} P_{2X} P_{XY} \prod_{k=3}^{n} P_{Yk} \right) \right), \tag{3}$$

where:

- $P_{1X} = (1 - P(1,l))\Pr(t_1, 0, \epsilon_X) + P(1,l)\Pr(t_1, 1, \epsilon_X)$,
- $P_{2X} = (1 - P(2,l))\Pr(t_2, 0, \epsilon_X) + P(2,l)\Pr(t_2, 1, \epsilon_X)$,
- $P_{XY} = \Pr(t_{XY}, \epsilon_X, \epsilon_Y)$,
- $P_{Yk} = (1 - P(k,l))\Pr(t_k, 0, \epsilon_Y) + P(k,l)\Pr(t_k, 1, \epsilon_Y)$,

and all the notations correspond to Figure 2. Indeed, for each position l, there are four possible scenarios ($\epsilon_X = 0$ or 1, and $\epsilon_Y = 0$ or 1) for each Species X and Y. The probability of each scenario is equal to the product of the probabilities along all the $n + 1$ branches of the NJ configuration shown in Figure 2.

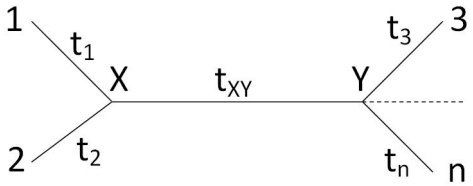

Figure 2: An intermediate NJ tree configuration used to compute the likelihood that Species 1 and 2 are neighbors. X and Y are the internal nodes of the presented intermediate NJ tree with n leaves and $n+1$ edges.

Hybrid Likelihood

Now, we can define the likelihood function to estimate the likelihood that Species 1, 2 and 3 are linked by a parents-hybrid relationship. In this case we define the probability that Species 3 is a hybrid of Species 1 and 2 by:

$$\mathcal{L}^{H}_{3,1,2} = \prod_{l=1}^{L} \left(\sum_{i \in \{1,2\}} \sum_{(\epsilon_i, \epsilon_3) \in \{0,1\}^2} P_i P_3 \Pr(t_i, \epsilon_i, \epsilon_3) \right), \tag{4}$$

where $P_i = (1-\epsilon_i)(1-P(i,l)) + \epsilon_i P(i,l)$, for $1 \leq i \leq 3$, and all the notations correspond to Figure 3. In this case, we have two possible scenarios for each position l, since each character of Species 3 may come either from Species 1 or from Species 2.

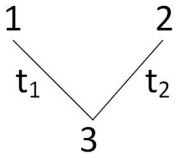

Figure 3: Configuration used to compute the likelihood that Species 3 is a hybrid of Species 1 and 2.

3 Main algorithm

In this section we describe the main algorithm 1 allowing us to infer either an explicit hybridization network or a phylogenetic tree (if no hybrids are present in the data) from a set of n binary sequences (e.g., encoding presence-absence of certain genes in species genomes).

Algorithm 1 Main Algorithm

Input: n binary sequences of size L corresponding to n species.

Output: An explicit hybridization network with terminal nodes corresponding to the input species. Some of these terminal nodes will be identified as hybrids.

$n_A = n$
while $n_A > 3$ **do**
    Determine the best pair of neighbors (i^*, j^*) according to NJ over all possible pairs of remaining species i and j.
    Compute the likelihood $\mathcal{L}^T_{i^*,j^*}$ of the NJ tree topology in which i^* and j^* are neighbors (see Figure 2 and Formula 3).
    Determine the best hybrid triplet h', i', j' and compute its maximum likelihood $\mathcal{L}^H_{h',i',j'}$ (where h' is a hybrid of i' and j'; see Figure 3 and Formula 4) over all possible triplets of current species h, i and j.
    if $\mathcal{L}^H_{h',i',j'} > \mathcal{L}^T_{i^*,j^*}$ **then**
        Species h' is identified as a hybrid of i' and j', and is removed from the dataset.
    else
        Species i^* and j^* are considered as neighbors, and we replace them by their direct common ancestor X determined by NJ.
        if $\mathcal{L}^H_{h',i',j'} > \mathcal{L}^T_{i^*,j^*}$ **then**
            Species h' is identified as a hybrid of i' and j', and is removed from the dataset.
        else
            Species i^* and j^* are considered as neighbors, and we replace them by their direct common ancestor X determined by NJ.
        end if
    end if
    $n_A = n_A - 1$
end while
We merge the three remaining species.

Remarks
When we replace i and j by their common ancestor X, the l-th component of the probability vector of X is computed by dividing $\mathcal{L}(X, l, 1)$ (the likelihood of having 1 at position l) by $\mathcal{L}(X, l, 1) + \mathcal{L}(X, l, 0)$. The likelihoods $\mathcal{L}(X, l, \epsilon_X)$, for $\epsilon_X = 0$ or 1, are computed as follows:

$$\mathcal{L}(X, l, \epsilon_X) = \sum_{\epsilon_Y \in \{0,1\}} \left(P_{1X} P_{2X} P_{XY} \prod_{k=3}^{n} P_{Yk} \right), \tag{5}$$

where $P_{1X}, P_{2X}, P_{XY}, P_{Yk}$ are computed according to Equation 3. For each position, we compute the probability that the hybrid character comes from each of the parents (degree of hybridization). Edge lengths are optimized using the Newton-Raphson method in the maximum likelihood computations.

The second version of our algorithms proceeds by correcting the likelihood values using a Bayesian information criterion (Schwarz and Gideon, 1978) in the following way:

- $2\mathcal{L}^T_{i,j} + (n+1)\ln(2L)$,
- $2\mathcal{L}^H_{h,i,j} + 2\ln(3L)$,

since there are $n+1$ degrees of freedom and $2L$ data in the NJ tree with n leaves (see Figure 2), and 2 degrees of freedom and $3L$ data in the hybrid configuration (see Figure 3). This refined version of our algorithm was also tested in our simulations (see section 4).

4 Simulation study

In our simulations, we first generated random phylogenetic trees using the tree generation algorithm available on the T-REX website (Boc et al, 2012). This algorithm requires as input the number of species n as well as the average tree edge length, and returns as output a random binary phylogenetic tree with n leaves that is built according to the method of Kuhner and Felsenstein (1994). In total, we generated 1000 unrooted phylogenetic trees for each of the following tree sizes: $n = 8$, $n = 16$ and $n = 32$, with the average edge length of 0.1. Then, we

simulated the evolution of binary sequences of length $L = 1000$ along these trees using the F81 evolutionary model (Felsenstein, 1981). For each phylogenetic tree generated this way, we obtained n binary sequences (corresponding to the tree leaves) of size L. The hybrids were added to the data afterwards. To add the hybrids, we first randomly selected two integers $1 \leq i < j \leq n$. Let α be the selected degree of hybridization, i.e., the contribution of a parent to a hybrid expressed as a proportion and varying from 0 to 1 (see Willems et al (2014) for more details on this hybridization parameter). We generated a new hybrid sequence with the first $\alpha \times L$ binary characters of sequence i to which we added the remaining $(1 - \alpha) \times L$ binary characters of sequence j. This new hybrid sequence was added to the n original sequences. In our simulations, we considered the following values of the hybridization parameter: $\alpha = 0.3$, $\alpha = 0.4$ and $\alpha = 0.5$ (for each tree considered in our simulations, the value of α was selected randomly, following the uniform distribution). Obviously, α and $1 - \alpha$ play a symmetric role in our model, and the hybridization degrees of 0.3 and 0.4 correspond to the hybridization degrees of 0.7 and 0.6, respectively. In general, the hybrid detection rate decreases as the value of α becomes closer to 0 or to 1 because one of the two parents becomes closer to the hybrid. The number of hybrids added to trees ranged from 0 to 5. Thus, for each considered tree size, n, we obtained 1000 binary sequence alignments corresponding to the original trees and 6000 matrices corresponding to phylogenetic networks having 0 to 5 hybrids.

Our first simulation was carried out using: (1) our previous distance-based method (Willems et al, 2014), (2) our new algorithm described in this paper, and (3) its refined version in which we corrected the likelihood function by means of a Bayesian information criterion (see section 3). The results of this first simulation are shown in Figure 4.

Moreover, for the trees with 8 leaves (i.e., species), we also simulated the data with different lengths of the binary sequences. Specifically, sequences of the following sizes: 20, 50, 100, 200, 500 and 1000, were considered. This second simulation was also conducted with phylogenetic networks including 0 to 5 hybrids. The results of this simulation are presented in Figure 5.

The results of the first simulation demonstrate that the new hybrid detection method presented in the previous section clearly outperformed our previous distance-based method (Willems et al (2014)) in terms of the false positive rate. However, the two methods showed very close results in terms of the true

positive rate. Furthermore, the refined version of our original method, in which we corrected the likelihood with a Bayesian information criterion, provided the best overall results in terms of the true positive rate and the second best overall results, after our new method, in terms of the false positive rate. Our second simulation confirmed the trends observed in the first one: The refined version of our new (original) method was much better than the original method in terms of the true positive rate, but was not good enough in terms of the false positive rate. These trends are especially noticeable when short binary sequences (i.e., when L ranged between 20 and 200) are considered.

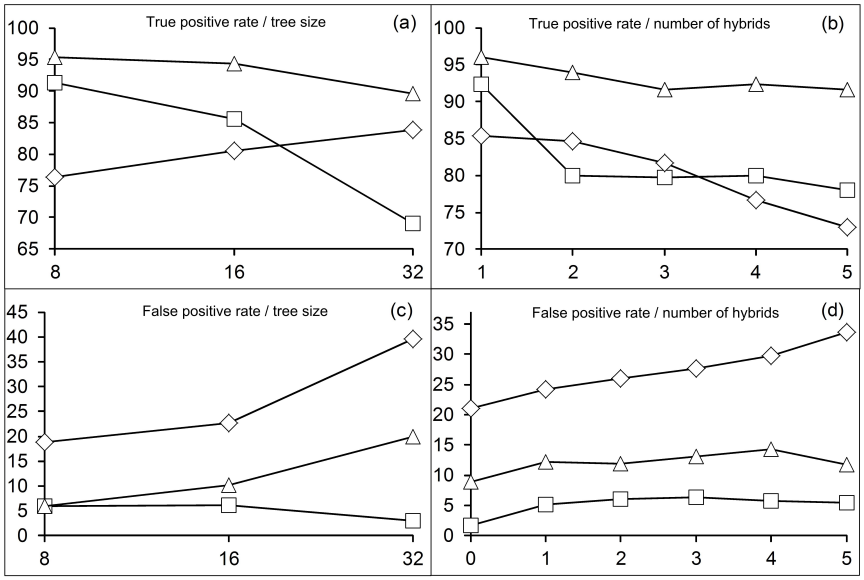

Figure 4: True positive and false positive rates (in %) with respect to the tree size (cases a and c) and the number of hybrids (cases b and d) obtained in simulations with 0 to 5 hybrids for trees with 8, 16 and 32 leaves and binary sequences of size 1000 using: (1) our previous distance-based method (Willems et al (2014); (◇), (2) our original ML method (□), and (3) the refined version of the ML method in which we corrected the likelihood with a Bayesian information criterion (△). The averages over all parameter combinations except the fixed one (tree size or number of hybrids) are shown.

After the additional tests that we conducted with larger trees (with 50 to 100 leaves), we can conclude that in terms of the true positive rate our algorithm based on the Bayesian information criterion becomes equivalent to the distance-

based method described in Willems et al (2014), but in terms of the false positive rate it clearly outperforms the distance-based method.

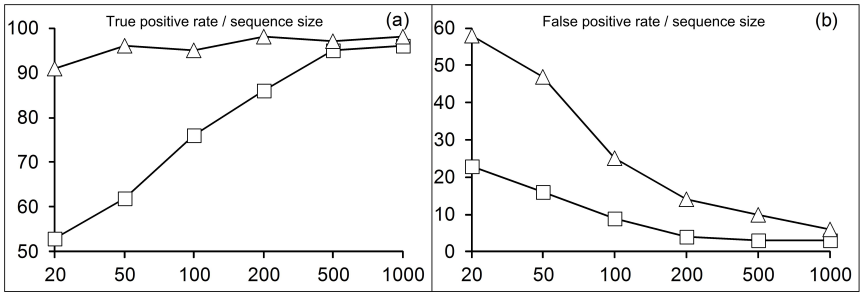

Figure 5: True positive (a) and false positive (b) rates (in %) shown with respect to the sequence size. These results were obtained from simulations with 0 to 5 hybrids and trees with 8 leaves. The sequences of sizes: 20, 50, 100, 200, 500 and 1000 were analyzed. The simulations were conducted using: (1) our original ML method (□) and (2) the refined version of the ML method in which we corrected the likelihood with a Bayesian information criterion (△). The averages over all parameter combinations except the fixed one (sequence size) are shown.

5 Analysis of the mosquitoes data

To test our new method on real data, we considered the dataset of restriction maps of the rDNA cistron for 16 species of mosquitoes constructed using eight recognition restriction enzymes (Kumar et al, 1998).A total of 26 sites were scored. The original binary sequence data are reported in Table 1 in Kumar et al (1998).

Huson and Klöpper (2007) have constructed a galled network (i.e., a specific type of recombination phylogenetic network) for this dataset and found that these data include 4 hybrids. Later on, Willems et al (2014) used the Hamming distances to transform the original mosquitoes data into a distance matrix before applying their distance-based network reconstruction method (see Figure 15 in Willems et al (2014)). The method by Willems et al (2014) also returned a network with 4 hybrids (i.e., 4 reticulations in the network).

After the application of the refined version of our method based on the BIC correction, we obtained an explicit phylogenetic network with 5 hybrids (see Figure 6). It is worth noting that the general structure of our network is very

similar to that of the galled network by Huson and Klöpper (2007) as well as to that of the reticulation network by Willems et al (2014). However, we can also observe some permutations between hybrids and parents in these networks. For example, the species *Aedes polynesiensis* and *Aedes triseriatus* have been identified as parent species in the network by Willems et al (2014), but have been identified as hybrids in the network provided by our new method (see Figure 6). Also, the species *Tripteroides bambusa* has been identified as a hybrid in our phylogenetic network, but not in that by Willems et al (2014).

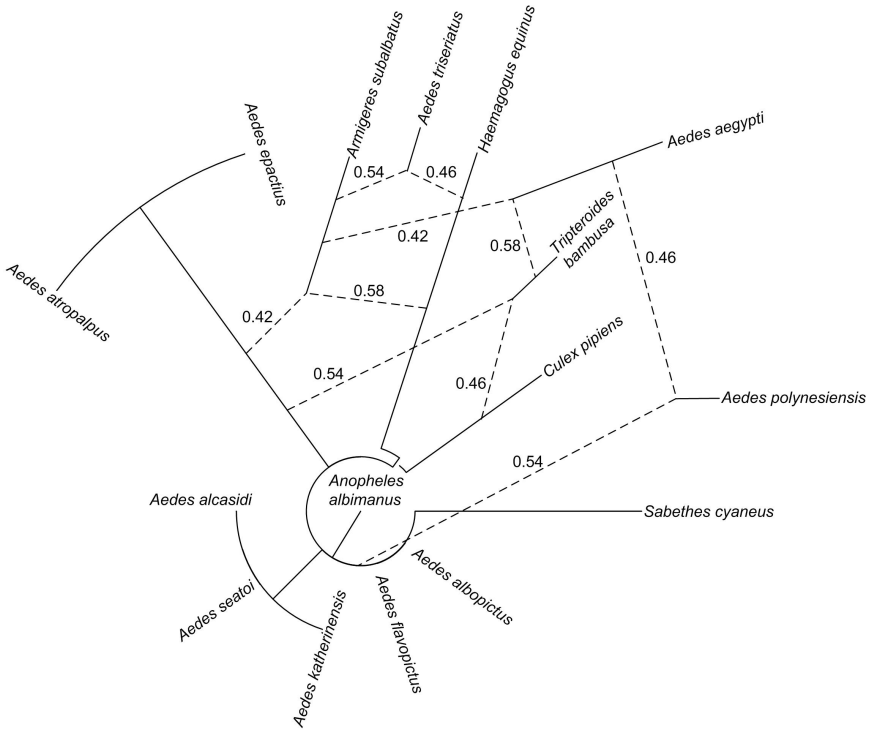

Figure 6: Network obtained for the restriction map of the 16-species mosquitoes dataset (Kumar et al, 1998) using our new method based on the BIC correction. Five hybrids, linked to the rest of the network by dashed lines, were identified. The numbers on dashed network edges represent the respective contributions of parents to hybrids.

On the other hand, Huson and Klöpper (2007) found four hybrids for the mosquitoes data. For example, as in our hybridization network the species *Aedes triseriatus* was found to be a hybrid of the species *Haemagogus equinus* and a group of several species. On the contrary, the species *Aedes polynesiensis* was found to be a hybrid in our network and not a hybrid in the network inferred by Huson and Klöpper (2007), whereas the cluster including the species *Aedes katherinensis*, *Aedes alasidi* and *Aedes seatoi* was found to be a cluster of hybrids in the network of Huson and Klöpper (2007), but not in our network.

The key advantage of our network representation, as well as of that by Willems et al (2014), over galled networks and split graphs is that our methods identify hybrids and their parents explicitly. Moreover, the respective contributions of parents to hybrids (i.e., hybridization degree) were also determined by our method (see the numbers on dashed edges in Figure 6).

6 Conclusion

In this paper, we have introduced a new accurate method for inferring explicit hybridization networks from presence-absence data (i.e., binary sequences). These networks can be used to represent adequately reticulate phylogenetic relationships between species, including for example hybridization events and horizontal gene transfers between species. Most of the existing algorithms developed in this context return as output only an implicit phylogenetic network, which is often very difficult to interpret. Our new method infers explicit hybridization networks using both the maximum likelihood and Neighbor-Joining approaches. The new method can be applied for building and interpreting phylogenetic networks for different types of binary sequences associated with species at hand, e.g., presence/absence of certain genes in genomes or presence/absence of restriction sites. Our simulations showed that the new method, and especially its modified version that uses a maximum likelihood correction by a Bayesian information criterion, outperforms the distance-based technique of Willems et al (2014) in terms of both the true positive and false positive rates, regardless of the number of species and hybrids in the dataset. Another advantage of the method presented here, compared to the distance-based method of Willems et al (2014), is that our new method does not require an additional hybrid selection threshold parameter that should be specified by the user in the distance-based

method. The main drawback of the new method is that it is slower than the distance-based method of Willems et al (2014). For example, for a dataset with 8 species (respectively, 64 species), our hybrid detection program written in C++, executed on an IBM PC computer equipped with an Intel i7 processor and 8GB of RAM, takes on average less than a second (respectively, 1.2 seconds) to carry out the distance-based method described in Willems et al (2014), while it takes on average 1.5 seconds (respectively, 13 minutes and 56 seconds) to carry out the maximum likelihood-based method described in this paper for the sequences of length $L = 100$. The time complexity of our distance-based method is $O(n^3)$. It is asymptotically equivalent to the time complexity of the NJ algorithm (Saitou and Nei, 1987). In general, our maximum likelihood-based method is more accurate (see Figures 4-5) but also slower (its time complexity is linear in terms of the sequence length, but is exponential in terms of the number of species in the worst case) than our distance-based method, as it is usually the case in phylogenetic analysis (Felsenstein, 2003). Thus, the maximum likelihood-based method presented in this paper can be recommended for use with phylogenetic networks having up to 100 species, while the distance-based method described in Willems et al (2014) can be applied to larger genomic datasets.

We are currently working on a mixed version of the two methods. For a large number of species, we can first infer a preliminary hybridization network using the distance-based method. Then, we can refine it locally by using the maximum-likelihood-based method. At the same time, we also plan to incorporate into the new method *a priori* knowledge, consisting of known probabilities for species h to be a hybrid of species i and j ($1 \leq h, i, j \leq n$), based for example on the species dispersal areas.

The program implementing our new algorithm was implemented in the C++ language. It is freely available to the research community at the following URL address: `http://www.info2.uqam.ca/~makarenkov_v/makarenv/hybrids_detection.zip`.

Acknowledgements This work was supported by Natural Sciences and Engineering Research Council of Canada.

References

Albrecht B, Scornavacca C, Cenci A, Huson D (2012) Fast computation of minimum hybridization networks. Bioinformatics 28(2):191–197, DOI 10.1093/bioinformatics/btr618

Bandelt H, Dress A (1992) Split decomposition: A new and useful approach to phylogenetic analysis of distance data. Molecular Phylogenetics and Evolution 1(3):242–252, DOI 10.1016/1055-7903(92)90021-8

Boc A, Diallo A, Makarenkov V (2012) T-REX: A web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Research 40(W1):W573–W579, DOI 10.1093/nar/gks485

Bryant D, Moulton V (2004) Neighbor-net: An agglomerative method for the construction of phylogenetic networks. In: Algorithms in Bioinformatics, Guigó R, Gusfield D (eds), Springer, Berlin, vol. 21, p. 255–365, ISBN: 978-3-540457-84-8, DOI 10.1007/3-540-45784-4_28

Chen Z, Wang L (2012) Algorithms for Reticulate Networks of Multiple Phylogenetic Trees. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(2):372–384, DOI 10.1109/TCBB.2011.137

Felsenstein J (1981) Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17(6):368–376, DOI 10.1007/BF01734359

Felsenstein J (2003) Inferring phylogenies. Sinauer Associates, Sunderland. ISBN: 978-0-878931-77-4

Huson D, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. Molecular Biology and Evolution 23(2):254–267, DOI 10.1093/molbev/msj030

Huson D, Klöpper T (2007) Beyond Galled Trees: Decomposition and Computation of Galled Networks. In: Research in Computational Molecular Biology, Springer, Berlin, p. 211–225, ISBN: 978-3-540716-81-5, DOI 10.1007/978-3-540-71681-5_15

Huson D, Scornavacca C (2012) Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. Systematic Biology 61(6):1061–1067, DOI 10.1093/sysbio/sys062

Kuhner M, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11(3):459–468, DOI 10.1093/oxfordjournals.molbev.a040126

Kumar A, Black W, Rai K (1998) An estimate of phylogenetic relationships among culicine mosquitoes using a restriction map of the rDNA cistron. Insect Molecular Biology 7(4):367–373, DOI 10.1046/j.1365-2583.1998.740367.x

Olave M, Avila LJ, Sites Jr JW, Morando M (2018) Detecting hybridization by likelihood calculation of gene tree extra lineages given explicit models. Methods in Ecology and Evolution 9(1):121–133, Wiley Online Library, DOI 10.1111/2041-210X.12846

Saitou N, Nei M (1987) The neighbor-joining method: A new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4(4):406–425, DOI 10.1093/oxfordjournals.molbev.a040454

Schwarz G, Gideon E (1978) Estimating the dimension of a model. Annals of statistics 6(2):461–464, DOI 10.1214/aos/1176344136

Solís-Lemus C, Ané C (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. PLoS genetics 12(3):e1005896, Public Library of Science, DOI 10.1371/journal.pgen.1005896

Willems M, Tahiri N, Makarenkov V (2014) A new efficient algorithm for inferring explicit hybridization networks following the Neighbor-Joining principle. Journal of Bioinformatics and Computational Biology 12(05):1450024, DOI 10.1142/S0219720014500243