Phylogenetic Trees and Networks Reduce to Phylogenies on Binary States: Does It Furnish an Explanation to the Robustness of Phylogenetic Trees against Lateral Transfers?

Marc Thuillard, Didier Fraix-Burnet

Phylogenetic Trees and Networks Reduce to Phylogenies on Binary States: Does It Furnish an Explanation to the Robustness of Phylogenetic Trees against Lateral Transfers?

Marc Thuillard[1] and Didier Fraix-Burnet[2,3]

[1]La Colline, Saint-Blaise, Switzerland. [2]Université Grenoble Alpes, IPAG, Grenoble, France. [3]CNRS, IPAG, Grenoble, France.

ABSTRACT: This article presents an innovative approach to phylogenies based on the reduction of multistate characters to binary-state characters. We show that the reduction to binary characters' approach can be applied to both character- and distance-based phylogenies and provides a unifying framework to explain simply and intuitively the similarities and differences between distance- and character-based phylogenies. Building on these results, this article gives a possible explanation on why phylogenetic trees obtained from a distance matrix or a set of characters are often quite reasonable despite lateral transfers of genetic material between taxa. In the presence of lateral transfers, outer planar networks furnish a better description of evolution than phylogenetic trees. We present a polynomial-time reconstruction algorithm for perfect outer planar networks with a fixed number of states, characters, and lateral transfers.

KEYWORDS: phylogenetic trees, phylogenetic networks, outer planar network, lateral transfer, multistate characters

CITATION: Thuillard and Fraix-Burnet. Phylogenetic Trees and Networks Reduce to Phylogenies on Binary States: Does It Furnish an Explanation to the Robustness of Phylogenetic Trees against Lateral Transfers?. *Evolutionary Bioinformatics* 2015:11 213–221 doi: 10.4137/EBO.S28158.

TYPE: Original Research

RECEIVED: May 15, 2015. **RESUBMITTED:** August 23, 2015. **ACCEPTED FOR PUBLICATION:** August 31, 2015.

ACADEMIC EDITOR: Jike Cui, Associate Editor

PEER REVIEW: Four peer reviewers contributed to the peer review report. Reviewers' reports totaled 1,066 words, excluding any confidential comments to the academic editor.

FUNDING: Authors disclose no funding sources.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: Thuillweb@hotmail.com

Introduction

Among phylogenetic methods, distance-based approaches have gained the somewhat ambiguous reputation to be computationally efficient and to furnish surprisingly good results when compared to the conceptually better character-based approaches. Distance-based approaches are therefore often regarded as a useful evil for example in very large phylogenies when the computational complexity of character-based approaches becomes rapidly so large that even the best heuristics may lead to suboptimal results. In this article, we would like to take a different stand by not merely opposing character- versus distance-based approaches, but by exploring the complementarity of the two approaches. Stevens and Gusfield[1] have shown that any phylogeny on k-state characters reduces to a phylogeny on binary-state characters with possibly some missing character states. This leads to the first question that one would like to address. Can the reduction procedure be applied to distance-based phylogenies? We will find out that distance- and character-based approaches are complementary methods that fit in a common theoretical framework. This raises another question, namely, to what extent do properties of distance-based approaches generalize to multistate characters. The question seems quite legitimate as on a set of binary characters both character- and distance-based phylogenies can be applied to reconstruct a phylogeny. In this article, we concentrate on the question of the robustness of a phylogenetic tree against lateral transfers. We show that lateral transfers between consecutive taxa on a circular order preserve the circular order of the taxa and may lead to data that cannot be well described by a phylogenetic tree but are still well described by a phylogenetic network.

With the availability of many complete genomes, the importance of lateral transfers in evolution became clearly recognized and increasingly doubts have arisen about the feasibility of drawing a tree of life based on genome evolution.[2] One is today in a somewhat paradoxical situation. On the one hand, almost nobody disputes the existence and importance of lateral transfers in evolution. On the other hand, there are numerous studies that show a good level of consistency between character-based phylogenies and accepted phylogenies obtained with different methods when such phylogenies do exist. An answer to this apparent paradox was suggested for distance-based approaches,[3] namely, that the circular order of the taxa on a tree is quite robust against lateral transfers. The circular order of the taxa on a tree is the order at which the end nodes are encountered in a clockwise scanning of a tree. If lateral transfers are only between consecutive end nodes, the tree reconstructed with the Neighbor-Joining algorithm furnishes a circular order of the nodes.[4,5] The order of the nodes corresponds to one of the possible orders of the tree prior to lateral

transfers. By using the degrees of freedom on the order of the taxa in a tree, a large number of lateral transfers can be accommodated by a tree while preserving the circular order. To what extent is this result still valid in the case of k-state characters? We will learn that multistate characters are quite robust against lateral transfer. In comparison to a distance-approach, the situation is complicated by the fact that the reconstruction of a phylogeny on multistate characters is not unique.

In the case of lateral transfers, an outer planar network is a better representation of phylogenetic data than a phylogenetic tree. An outer planar network is a special type of phylogenetic networks that can be reconstructed from a distance matrix using NeighborNet.[6] In distance-based approaches, lateral transfers between consecutive end nodes preserve the circular order of a phylogenetic tree and the distance matrix fulfills the so-called Kalmanson inequalities.[5] If a distance matrix fulfills the Kalmanson inequalities, then the matrix can be exactly described by an outer planar network. The question arises whether there is an efficient method to reconstruct a perfect outer planar network from characters. For a perfect phylogeny, the problem of reconstructing a perfect tree is nondeterministic polynomial time.[7,8] If the number of states is a fixed constant k, then the problem is referred to as the k-perfect phylogeny problem. In this case, the problem has polynomial time-complexity.[9,10] The best algorithm has a $O(2^{2k}m^2n)$ time-complexity for a problem with m characters defined on n taxa.[11] Can the problem of reconstructing a perfect outer planar network in multistate characters be solved in polynomial time at fixed k, as is the case for the perfect phylogeny problem? We will find out that there is a time-polynomial reconstruction algorithm for a p-level outer planar network on a fixed number of k-state characters.

If no outer planar network describes exactly the input data, then one needs methods to judge the quality of a given reconstruction. The reduction to binary characters with missing states allows the introduction of measures characterizing how well a particular phylogenetic tree or outer planar network describes a set of k-state characters defined on a set of taxa. The measures are obtained by combining the deviation to the four gamete rules and the deviation to the circular consecutive-ones conditions or alternately the contradiction to the Kalmanson inequalities. Both the deviation to the circular consecutive-ones and the contradiction are zero in case of a perfect tree or outer planar network, but for nonperfect trees, these two measures may lead to different solutions.

Phylogenetic Trees and Networks

Phylogenetic methods are divided into two main approaches, character- and distance-based approaches. Character-based approaches are conceptually well suited to evolutionary studies, whereas the low complexity of distance-based approaches permits to deal with a large amount of data. Phylogenetic data are commonly represented under the form of a tree or in the form of a combination of trees as in phylogenetic networks. In this section, the main results that are used in the following sections are presented succinctly, since they have been published elsewhere.[1–17] New results are presented in the subsequent section "A common framework for character- and distance-based phylogenies" as well as in Annexes 2 and 3. Annex 1 is a reformulation of known results that is necessary to understand Annex 2.

Distance-based approach to phylogeny. A graph G consists of a set of nodes $V(G)$ and a set of edges $E(G)$ with $e(x,y)$ denoting the edge containing the two nodes x and y. A weighted phylogenetic tree T is a graph with X as its set of end nodes (or leaves) with internal nodes of degree at least 3 and a unique path between any two distinct end nodes. A positive weight is associated with each edge, so that the tree can be reconstructed univocally from the distance matrix D. If a distance matrix satisfies the four-point condition, then D can be exactly represented by a weighted phylogenetic tree.[12,13] The matrix $Y^n_{i,j} = 1/2 \cdot (D_{i,n} + D_{j,n} - D_{i,j})$ is often called the distance in the Farris space. It corresponds to the shortest distance between a reference end node n and the first ancestor node common to both the end nodes i and j. A circular order on a phylogenetic tree corresponds to an indexing of the n end nodes on a circular (clockwise or anti-clockwise) scanning of the end nodes in T. For taxa indexed according to a circular order of the distance matrix in the Farris space, $Y^n_{i,j}$ fulfills the Kalmanson inequalities[14]:

$$Y^n_{i,j} \geq Y^n_{i,k}; Y^n_{i,k} \leq Y^n_{j,k} \; (i \leq j \leq k) \tag{1}$$

with $Y^n_{i,j} = 1/2 \cdot (D_{i,n} + D_{j,n} - D_{i,j})$. Given a distance D on X, we shall call a perfect order, an order $O = (x_1, x_2, \ldots, x_n)$ of X for which all the Kalmanson inequalities are satisfied. In this case, the distance matrix D can be exactly represented by a phylogenetic tree if the four-point conditions are also fulfilled or otherwise by an outer planar network.[15] The contradiction level C on the circular order of the taxa in a tree or a network is a measure of the deviation to a perfect order.[16] For an unrooted tree or network, the contradiction is

$$C = \frac{\sum_n \left(\sum_{k>j\geq i} \max(Y^n_{i,k} - Y^n_{i,j}, 0)^2 + \sum_{k\geq j>i} \max(Y^n_{i,k} - Y^n_{j,k}, 0)^2 \right)}{\sum_n \left(\sum_{k>j\geq i} (Y^n_{i,k} - Y^n_{i,j})^2 + \sum_{k\geq j>i} (Y^n_{i,k} - Y^n_{j,k})^2 \right)}. \tag{2}$$

For binary characters, the Kalmanson inequalities are fulfilled if there is a circular order of the taxa so that each binary character (labeled with either state 0 or 1) fulfills the circular consecutive-ones condition.[17] The circular consecutive-ones condition is fulfilled, if for any binary state, the taxa with the 1 state are consecutive on the circular order. Generally, the contradiction takes a value between 0 and 1.

Character-based approach to phylogenetic trees and networks. A phylogeny defined by a set of characters is

referred to as a character-based phylogeny. A perfect phylogeny problem has typically as input a character matrix M with $M_{i,j}$ the state of the jth character on the ith taxon. The convexity of all characters on the phylogenetic tree is a necessary condition to having a perfect phylogeny. A character is convex on a phylogenetic tree T if there exists a labeling of the interior nodes, so that the subgraph of T induced by any character state α of a character C is connected.[13] Any k-state phylogeny reduces to a binary-state phylogeny, and therefore, a tentative reconstruction of a perfect phylogeny can be evaluated using the four gamete rules. In the four-gamete rules, a gamete is defined as a pair of binary characters defined on each taxon. For phylogenies defined on binary characters, the four-gamete rule states that a perfect phylogeny exists for binary input sequences if and only if no pair of characters contains all four possible binary pairs: (0,0), (0,1), (1,0), and (1,1).[18]

An unrooted phylogenetic network N on a set of taxa X is any undirected graph whose end nodes are bijectively labeled by the taxa in X. An edge whose removal disconnects the graph is a bridge, and a blob is defined as a maximal bridgeless component of a network. A level-p network on a set X of taxa is such that an unrooted tree connecting all nodes can be obtained by removing at most p edges per blob.[19] Figure 1 shows an example of a level-1 network with a single blob. Studies on unrooted networks have mostly focused on level-1 networks. The reconstruction of phylogenetic networks from a set of quartets[20,21] or triplets[22,23] has been discussed in several papers. The reconstruction of unrooted level-1 phylogenetic networks completely defined by a set of quartets was shown to be $O(n^4)$.[19] While the reconstruction of unrooted phylogenetic network from binary characters is well studied,[24–27] we do not know of any polynomial reconstruction method for k-state characters.

Reducing a phylogeny on k-state characters to a phylogeny on binary characters. The transformation from multistate to binary characters is done by defining a character $C_{p,q}$ for each pair $(p > q)$ of character states (α_p, α_q). Given a character state α of a multistate character, the state of $C_{p,q}$ is given by[1]

$$C_{p,q} = \begin{cases} 1, \text{ if } \alpha = \alpha_p \\ 0, \text{ if } \alpha = \alpha_q \\ ?, \text{ otherwise} \end{cases} \quad (3)$$

Figure 2 illustrates the transformation defined by Eq. (3) with an example for ($\alpha_p = A$, $\alpha_q = E$).

The transformation defined by Eq. (3) applies to any number of character states. For a small number of character states, there is a limited amount of possibilities for completing the missing character states in Eq. (3) to binary characters. For a phylogenetic tree defined by three-state characters, there are three different ways to complete each character;[28] two of them are compatible with the phylogenetic tree. For a set of binary characters, the fulfillment of all four gamete rules is a necessary and sufficient condition for a perfect phylogeny. Is the fulfillment of four gamete rules also a sufficient condition to obtain a perfect phylogeny in binary characters obtained by reduction of k-state characters to binary characters? Fulfilling the four gamete rules is not a sufficient condition to obtain a perfect phylogeny.[1] A second necessary condition is that there is a binary character associated with each different pair of states α,β using Eq. (3).

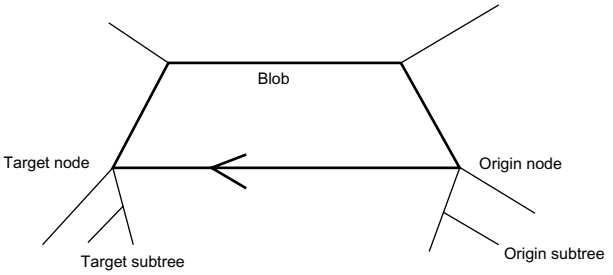

Figure 1. A phylogenetic network can be used to represent the effect of a lateral transfer between an origin node and a target node. The lateral transfer may result into a local deviation of a perfect phylogenetic tree represented in the form of a blob.

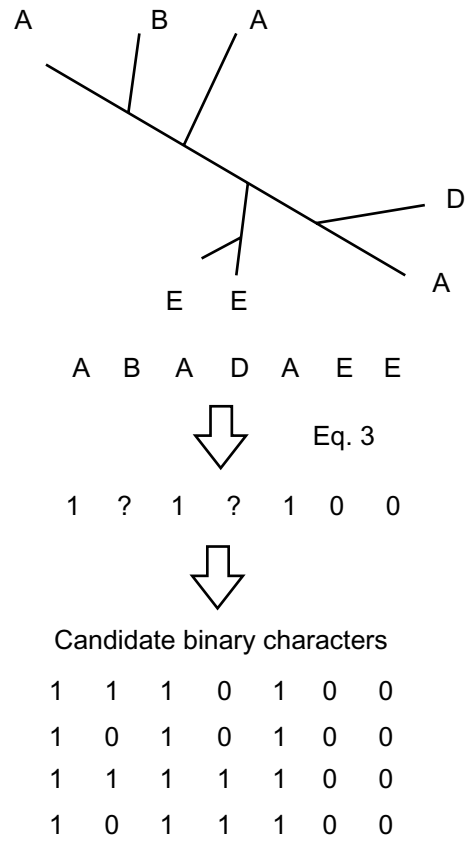

Figure 2. A k-state character compatible with a phylogenetic tree reduces to a binary character with missing states (using Eq. (3)). In a perfect tree, the missing characters can be defined so as to obtain binary characters compatible with a phylogenetic tree (Eq. 3). In this example, the character A is transformed into 1 and E into 0.

We suggest defining the four gamete rules together with this second condition as the extended four gamete rules. The fulfillment of the extended four gamete rules is a necessary and sufficient condition for a perfect phylogeny defined on k-state characters to reduce to a perfect phylogeny on binary states using Eq. (3).

A Common Framework for Character- and Distance-Based Phylogenies

Let us show that a phylogenetic tree defined by a distance matrix can be considered a special case of a phylogenetic tree defined by characters. To do so, the reduction to binary characters approach is extended to evolution models defined by a transition model between bases. Let us start with a simple example: the Jukes–Cantor model. Replacing the missing states '?' in Eq. (3) by 0 leads to the Jukes–Cantor model of DNA evolution. (The Jukes and Cantor[29] model assumes equal base frequencies and mutation rates.) The distance between two taxa is defined as

$$D_{i,j} = \frac{3}{4}\ln\left(1-\frac{4}{3}p_{ij}\right) \tag{4}$$

with $p_{ij}=\sum_{p=1,\ldots,m} d_{i,j}(p)\Big/\left(\sum_{\substack{p=1,\ldots,m\\ \text{all taxa},i\geq j}} d_{i,j}(p)\right)$, and $d_{i,j}(p)=1$ if the states of the pth character on the taxa i and j are different and $d_{ij}(p) = 0$ otherwise.

The Kimura two-parameter model distinguishes between transitions and transversions (A or G ↔ C or T). The Kimura two-parameter distance is given by[30]

$$D_{i,j} = -\frac{1}{2}\ln\left(1-2p_{i,j}-q_{i,j}\right)-\frac{1}{4}\ln\left(1-2q_{i,j}\right), \tag{5}$$

where $p_{i,j}$ is the proportion of sites with transitions and $q_{i,j}$ is the proportion of transversions. The proportion of transitions and transversions can be obtained using Eq. (6), which is described below. The computation of a distance matrix from aligned sequences (without gaps) requires computing the number of state changes between two states A_1 and A_2. The number of state changes, N $((A_1, A_2) \leftrightarrow (A_3,\ldots,A_k))$ and N $(A_1 \leftrightarrow (A_2,A_3,\ldots,A_k))$, can be computed using Eq. (3). For instance, by setting $A_1 = A_2 = 1$ and all other states to 0, then one obtains N $((A_1, A_2) \leftrightarrow (A_3,\ldots,A_k))$ by computing the total number of transitions between the binary state 1 and the binary state 0. Given the number of state changes between any of the two states A_1,A_2 and any other state and the transition from one state A_1 to any other state, one has

$$\begin{aligned}&N\left(A_1 \leftrightarrow A_2\right) = 1/2\left(N\left(A_1 \leftrightarrow \left(A_2,A_3,\ldots,A_k\right)\right)\right.\\&\left.+\left(A_2 \leftrightarrow \left(A_1,A_3,\ldots,A_k\right)\right)-N\left(\left(A_1,A_2\right)\leftrightarrow\left(A_3,\ldots,A_k\right)\right)\right)\end{aligned} \tag{6}$$

from which the number of state changes between the two states A_1 and A_2 and the distance matrix can be computed.

In summary, the reduction to binary characters' approach can be applied to both character- and distance-based phylogenies. It provides a unifying framework to explain the differences and similarities between distance- and character-based phylogenies. The main difference between the character-approach and the distance-approach is that in the distance-approach, the transformation given by Eq. (3) does not depend on the set of data. In a distance-based approach, all missing states, ?, are defined for a given model of character evolution. In a character-based approach, a missing state takes a binary value that depends on the input data. It is therefore not surprising that distance-based approaches are computationally less demanding as the search space is much smaller than in character-based approaches. In the reduction to binary characters' framework, distance-based approaches can be regarded as special cases of character-based approaches. The two approaches are indeed complementary. Character- and distance-based approaches can be combined. Once multistate characters are transformed into binary-state characters, a distance matrix can be computed and a phylogenetic tree or network can be reconstructed using a distance-based approach, such as Neighbor-Joining[31] or NeighborNet.[6] Figure 3 summarizes the last aspect.

How Lateral Transfers Transform a Perfect Phylogeny into a Phylogenetic Network?

Given a tree on multistate characters, a lateral transfer is modeled as an edge (with some direction) from an origin node to a target node on a phylogenetic tree together with possibly a labeling that indicates which characters are transferred. A lateral transfer corresponds to the replacement of some states on the target node by the corresponding character states of the origin node of the transfer. Let us define the target (respectively origin) subtree as the subtree attached to the blob at the target (respectively origin) node. The effect of a unique lateral transfer can be modeled using the assumptions that (i) the target subtree is a perfect phylogeny on the subset of characters defined on the set S_t of end nodes completed by the target node and (ii) the state of the target node is the only state with possibly one character state common to both the target subtree and the remaining nodes. Figure 4 shows an illustrative example of a lateral transfer. The combined effect of several transfers is described below.

In the Jukes–Cantor distance-based model of evolution, the circular order of an outer planar network is quite robust against lateral transfers. One can show that the circular order of the taxa on an outer planar network describing a tree after lateral transfer between consecutive end nodes corresponds to a circular order of the tree.[3] Furthermore, if the tree with lateral transfers is reconstructed with Neighbor-Joining, there is a circular order of the tree that is the same as one of the perfect outer planar network.[4,5] To what extent do these results apply to the case of k-state characters? For k-state characters,

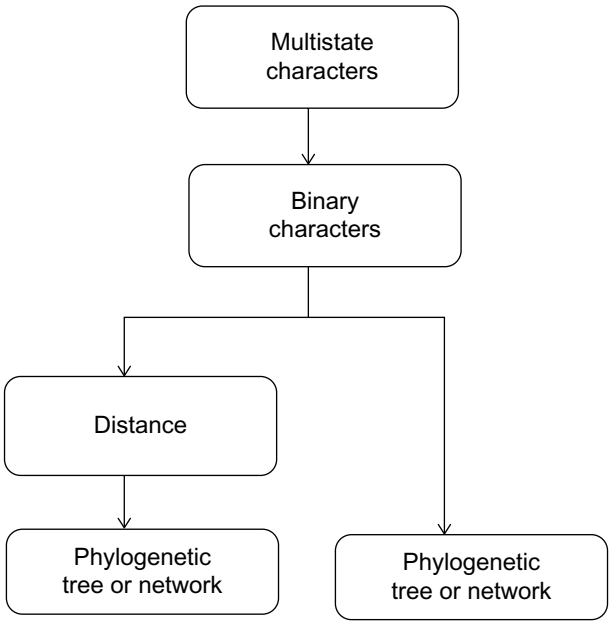

Figure 3. The reduction of *k*-state characters to binary characters offers a common framework to distance-based and character-based approaches. Using Eq. (3), the *k*-state characters are reduced to binary characters. In this framework, a distance-approach is a special case of Eq. (3) in which the missing state? is given by the model of base evolution. After transformation to binary characters, the data can be processed either with a distance-approach or a character-based approach.

the situation is complicated by the fact that there are possibly several perfect phylogenies on a given set of character states. Owing to the possible nonuniqueness of the reconstruction, several different hypotheses on the lateral transfer may be found. Another difference is that a phylogenetic tree in a distance-based approach has a tree representation that fulfills the master tour property.[32] In other words, if some taxa are removed, then any circular order can be generated by removing these taxa from a circular order of the complete phylogenetic tree.[16] In a multistate phylogeny, the master tour property

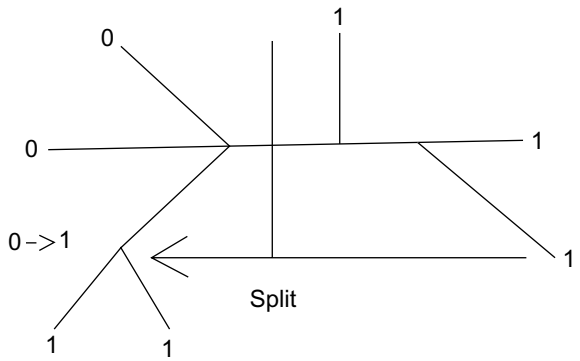

Figure 4. An example showing a split of a perfect phylogeny obtained after removing the target subtree. The lateral transfer preserves the circular consecutive-ones property in the resulting outer planar network obtained after a lateral transfer between adjacent nodes: (0, 0, 1, 1, 1, 1, 1).

is not always fulfilled. After these important remarks, let us discuss the robustness of the circular order of the taxa on a tree defined on multistate characters against lateral transfers between adjacent nodes (Two nodes are adjacent if they are on the shortest paths between two consecutive end nodes on a circular order.). Let us discuss first the case of a unique lateral transfer.

Proposition 1: If a lateral transfer is between two adjacent nodes on a perfect tree, then there is a circular order of the taxa for which the character states after reduction of the multistate characters to binary-state characters fulfill both the circular consecutive-ones properties and the Kalmanson inequalities and can be represented exactly by an outer planar network.

Proof: The target subtree is a perfect phylogeny on the subset of characters defined on the set S_t consisting of the target node and the end nodes of the target subtree. Any split in the target subtree is described using Eq. (3) setting all states in $S - S_t$ to the same binary state as the target state. It is always possible as the state of the target node is the only state with possibly a state common to both the target subtree and the single blob with the removed target subtree. For the remaining pairs of character states, let us set all nodes in the target subtree to the same state. For the rest of the proof, it is therefore sufficient to consider the subtree obtained after removing the target subtree but keeping the target node. Any split in the subtree on the taxa $S - S_t$ with a splitting vector that is not on the direct path between the origin and the target nodes of the lateral transfer is not affected by the lateral transfer. The only remaining splits that are possibly affected by the lateral transfer are splits with a splitting vector on the path between origin and target nodes. As in this case, all nodes in the target and the origin subtrees have the same state, it follows that a lateral transfer between consecutive nodes preserves the circular consecutive-ones property and consequently the Kalmanson inequalities.

Since for any tree and any two nodes there is always a planar tree representation in which the two nodes are adjacent,[3] the domain of validity of Proposition 1 is quite broad.

Proposition 1 can be generalized to several lateral transfers. In particular, for a level-1 network, the generalization follows from the fact that a level-1 network can be decomposed into single level-1 outer planar networks on which the proof of Proposition 1 applies independently of the other blobs. Let us now consider lateral transfers between adjacent nodes. In order to have a causal model of evolution, one assumes that no edge describing a lateral transfer in a planar representation of the tree crosses another edge and no node is at the origin or target of two lateral transfers.

Proposition 2: If lateral transfers on a perfect phylogeny are such that (i) all lateral transfers are between adjacent nodes, (ii) no edge describing a lateral transfer in a planar representation of the tree crosses another edge, and (iii) no

node is at the origin or target of two lateral transfers, then there is a circular order of the tree for which the data fulfill both the circular consecutive-ones properties and can be represented exactly by an outer planar network.

Proof: Given four end nodes in the tree and the shortest paths between the four nodes on the tree, assume that the four nodes have states α, β. A circular order of the four end nodes on the tree with the ordered states α, β, α, β is not allowed on a perfect phylogeny. It follows that such an order may possibly only result from some lateral transfers. This case can also be excluded as lateral transfers are between adjacent nodes, no node is at the origin or target of two lateral transfers preventing crossover, and the target node is the only node with possibly one character state common to both the target subtree and the remaining nodes. Let us show that the reduction to binary states for the pair (α, β) fulfills the circular consecutive-ones properties under a proper choice of the missing states. Consider on the circular order of the taxa the interval between the first node and the last end node with β state, which does not contain the state α, and set all binary states on the interval to zero. The remaining nodes including the nodes with α state are set to one. The reduced states on the circular order of the taxa fulfill by construction of the circular consecutive-ones property.

Proposition 2 shows that given a perfect phylogeny and some lateral transfers between adjacent nodes, the data can be described exactly by an outer planar network on binary-state characters provided some constraints (i–iii) are satisfied. As already known from binary-state characters, a phylogenetic tree constructed with Neighbor-Joining will have a planar tree representation with the same circular order as a perfect outer planar network.[4] In this sense, character-based phylogenies are quite robust against lateral transfers. Contrarily to a perfect phylogeny obtained from a distance matrix, the circular order may not correspond to a circular order of the perfect phylogeny. Quite surprisingly, a phylogenetic tree described exactly by a distance matrix is therefore even more robust against lateral transfer than a character-based phylogeny and may furnish more information on lateral transfers events.

Reconstruction Algorithms for Phylogenetic Trees and Outer Planar Networks

In this section, a reconstruction algorithm for perfect outer planar networks is presented. The main ideas are presented below, although the reader is referred to the Supplementary File for a more detailed description of the algorithm. Let us recall that in a level-1 network, a perfect phylogeny is obtainable by removing at most one edge per blob. A perfect level-1 outer planar network is reconstructed by assembling level-1 phylogenetic networks containing a single blob. The algorithm is an extension of the reconstruction algorithm for perfect phylogenies.

Figure 5B–E illustrates graphically the algorithm. The network is first decomposed into the union of single blob level-1 outer planar networks (Fig. 5B). Each single blob is then decomposed into a blob with single edges attached to it and some phylogenetic subtrees (Fig. 5C). By removing the target node, each blob with single edges transforms into a perfect phylogeny (Fig. 5D) that is reduced into a perfect phylogeny on binary-state characters. The perfect subphylogeny is transformed into a blob by reinserting the taxon at different possible positions within the circular order of the tree and searching for a circular order on which the circular consecutive-ones conditions and consequently all Kalmanson inequalities are fulfilled and the deviation to the four gamete rules is the lowest (It may not be unique.).

Dealing with level-1 networks limits the complexity of the reconstruction algorithm as each single blob N_b can be reconstructed independently of the other blobs. The decomposition of a level-1 network into single blobs uses a procedure very similar to the tree decomposition and has a $O(2^{3k}m^3n)$ time-complexity (The higher time-complexity is due to the use of Agarwala and Fernández-Baca's procedure). The phylogeny obtained after removing the target node from a single blob level-1 network with single edges is a caterpillar tree. The target node can be connected to the caterpillar in four different manners. Verifying the circular consecutive-ones condition requires testing four possibilities per blob at the end of the caterpillar tree. The maximum number of blobs is smaller than half the number of taxa. Testing the circular consecutive-ones property has therefore a O(n) complexity. The time-complexity of the algorithm is essentially given by the complexity of reconstructing the single blobs. The time-complexity is quadratic with the number of end nodes. The time-complexity corresponds to the $O(2^{2k}m^2n)$ time-complexity to reconstruct perfect phylogenies multiplied by the number of possibilities to remove a node resulting in a $O(2^{2k}m^2n^2)$ time-complexity. The low complexity results from the fact that for each candidate blob with single edges different subsets of taxa are tested, so that considering all blobs, each proper cluster is tested at most 2n+1 times. The second reason for the low complexity is that the fast decomposition procedure[11] can be used here. The perfect reconstruction of a perfect outer planar network has therefore a $O(2^{2k}m^2n(n + 2^km))$ time-complexity.

There is a fundamental difference between level-1 and level-p networks with $p > 1$. In a level-1 network, there is always a circular order of the taxa so that each lateral transfer is between adjacent nodes. For level-p networks, this is not always the case. If the three conditions on lateral transfers are fulfilled ((i) all lateral transfers are between adjacent nodes, (ii) no edge describing a lateral transfer in a planar representation of the tree crosses another edge, and (iii) no node is at the origin or target of two lateral transfers), then an outer planar network describes exactly the data as Proposition 2 is not limited to level-1 networks. The reconstruction algorithm for perfect outer planar network can be adapted by reconstructing for

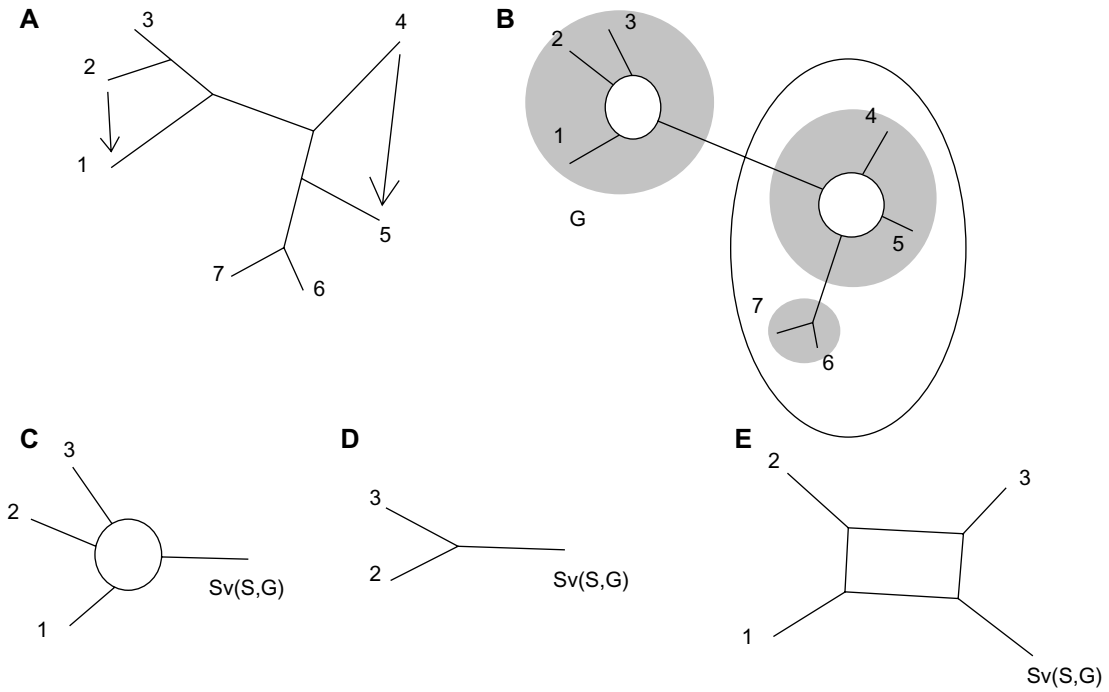

Figure 5. (**A**) A perfect phylogeny is transformed by lateral transfers between adjacent end nodes into a level-1 outer planar network. (**B**) The network contains two single blobs. (**C** and **D**) Each single blob network is a perfect subphylogeny after removing the target node (there may be several possibilities to remove a taxon and obtain a perfect phylogeny, and in those cases, the algorithm cannot be used to determine the origin and target of the lateral transfer). (**E**) A single blob description is obtained after reinserting taxon 1.

each single level-p blob first a perfect phylogeny after having removed p taxa on a single level-p blob. The time-complexity of the algorithm is $O(2^{2k}m^2n^p(n + 2^km))$. Figure 6 illustrates the procedure.

If the lateral transfer leads to a deviation of the circular consecutive-ones condition, then polynomial-time reconstruction algorithms are not known to us. For level-2 networks, possible lateral transfers explaining the data can be discovered using an algorithm[33] combining the minimum contradiction approach[5] with NeighborNet once the multistate characters are reduced to binary-state characters. As a side comment, let us point out that detecting lateral transfers is important, since a lateral transfer may have a large influence on the apparent rate of evolution if the evolution rate is not constant on all sites.

The perfect reconstruction algorithm for level-1 outer planar networks relies heavily on the reconstruction algorithm for perfect phylogenies, and therefore, the reconstruction algorithm for perfect phylogenies is given in Annex 1 (Supplementary File). In Annex 2 (Supplementary File),

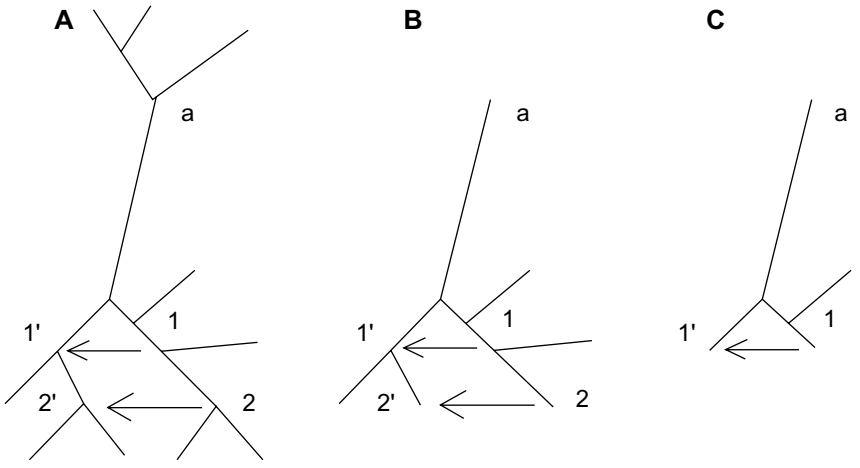

Figure 6. An example showing how a level-2 outer planar network is decomposed into a level-1 outer planar network: (**A**) phylogenetic tree with two lateral transfers, (**B**) single level-2 blob with single edges attached to it, and (**C**) by removing node 2′, the blob with single edges reduces to a level-1 network.

the reconstruction algorithm for perfect level-1 outer planar network is presented.

Outlook

Phylogenetic networks are typically used in case of data conflicting with evolution described by a tree. While this article focuses on lateral transfer, let us mention other effects. A crossover between consecutive taxa does not always preserve the circular order. Removing one of the taxa involved in crossover results in a phylogeny fulfilling the Kalmanson inequalities. Discussing gene fusion, complex hybridization schemes, or homologous recombination goes beyond the scope of this article as the effect on a phylogeny will very much depend on how the characters are defined. It could be an interesting research topic.

The extended four gamete rules and the Kalmanson inequalities are two fundamental properties of phylogenies. A perfect phylogeny defined on k-state characters reduces to binary characters using Eq. (3) that fulfill the extended four gamete rules, while a perfect outer planar network reduces to binary characters that fulfill the circular consecutive-ones conditions. In practical applications, deviations to perfect phylogenies or outer planar networks are quite common. It is quite rare in real-world applications that phylogenetic data correspond to a perfect phylogeny or outer planar network. For a limited number of characters and states, an exhaustive search is possible trying out all 2^{km} possibilities of obtaining binary characters. For each possibility, a circular order is obtained using NeighborNet and the deviation to the four gamete rules and the circular consecutive-ones properties (alternatively, the Kalmanson inequalities) may be used to estimate the quality of the reconstruction. Annex 3 shows an example using this approach on a dataset for rubber trees (*Hevea brasiliensis*) collected in South America from a polymorphic mitochondrial DNA region[34] using an exhaustive search. The next big challenge is the development of heuristics that guide the search toward better outer planar networks when an exhaustive search is not possible.

Acknowledgments

We would like to thank the anonymous referees for their pertinent comments and suggestions.

Author Contributions

Original idea and first draft: MT. Contributed to the writing of the manuscript: MT, DFB. Agree with manuscript results and conclusions: MT, DFB. Jointly developed the structure and arguments for the paper: MT, DFB. Made critical revisions and approved final version: MT, DFB. Both authors reviewed and approved of the final manuscript.

Supplementary Material

Supplementary Figure 1. The best outer planar network on a dataset for rubber trees from a polymorphic mitochondrial DNA region from rubber tree (*H. brasiliensis*): (**A**) obtained with the reduction to binary characters approach and (**B**) using the Hamming distance.

Annex 1. Perfect tree reconstruction algorithms.

Annex 2. Perfect outer planar network reconstruction algorithm.

Annex 3. An example of an outer planar network reconstruction algorithm.

REFERENCES

1. Stevens K, Gusfield D. Reducing multistate to binary perfect phylogeny with applications to missing, removable, inserted, and deleted data. In: Moulton V, Singh M, eds. *Algorithms in Bioinformatics*. Berlin: Springer; 2010:274–87.
2. Doolittle WF. Uprooting the tree of life. *Sci Am*. 2000;2:90–5.
3. Thuillard M. Minimizing contradictions on circular order of phylogenic trees. *Evol Bioinform Online*. 2007;3:267–77.
4. Thuillard M. Why phylogenetic trees are often quite robust against lateral transfers. In: Pontarotti P, ed. *Evolutionary Biology. Concept, Modelization and Application*. Berlin: Springer; 2009:269–85.
5. Thuillard M. Minimum contradiction matrices in whole genome phylogenies. *Evol Bioinform Online*. 2008;4:237–47.
6. Bryant D, Moulton V. NeighborNet: an agglomerative method for the construction of planar phylogenetic networks. In: Guigó R, Gusfield D, eds. *Algorithms in Bioinformatics*. Berlin: Springer; 2002:375–91.
7. Bodlaender HL, Fellows MR, Warnow TJ. *Two Strikes Against Perfect Phylogeny*. Berlin: Springer; 1992:273–83.
8. Steel M. The complexity of reconstructing trees from qualitative characters and subtrees. *J Classif*. 1992;9:91–116.
9. Agarwala T, Fernández-Baca D. A polynomial-time algorithm for the perfect phylogeny problem when the number of character states is fixed. *SIAM J Comput*. 1994;23:1216–24.
10. Fernández-Baca D. The perfect phylogeny problem. In: Cheng X, Du DZ, eds. *Steiner Trees in Industry*. Berlin: Springer; 2001:203–34.
11. Kannan S, Warnow TJ. A fast algorithm for the computation and enumeration of perfect phylogenies. *SIAM J Comput*. 1997;26:1749–63.
12. Zaretskii KA. Constructing trees from the set of distances between pendant vertices. *Uspehi Matematiceskih Nauk*. 1965;20:90–2.
13. Semple C, Steel M. *Phylogenetics*. Oxford: Oxford University Press; 2003. [Oxford Lecture Series In Mathematics And Its Applications].
14. Kalmanson K. Edgeconvex circuits and the traveling salesman problem. *Can J Math*. 1975;27:1000–10.
15. Bandelt HJ, Dress AW. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Mol Phylogenet Evol*. 1992;1:242–52.
16. Thuillard M, Fraix-Burnet D. Phylogenetic applications of the minimum contradiction approach on continuous characters. *Evol Bioinform Online*. 2009;5:33–46.
17. Huson DH, Rupp R, Scornavacca C. *Phylogenetic Networks*. Cambridge: Cambridge University Press; 2010.
18. Estabrook GF, Johnson CS, McMorris FR. A mathematical foundation for the analysis of cladistic character compatibility. *Math Biosci*. 1976;29:181–7.
19. Gambette P, Berry V, Paul C. Quartets and unrooted phylogenetic networks. *J Bioinform Comput Biol*. 2012;10(4):1250004.
20. Grünewald S, Moulton V, Spillner A. Consistency of the QNet algorithm for generating planar split networks from weighted quartets. *Discrete Appl Math*. 2009;157:2325–34.
21. Yang J, Grunewald S, Xu Y, Wan XF. Quartet-based methods to reconstruct phylogenetic networks. *BMC Syst Biol*. 2014;8(1):21.
22. Jansson J, Sung WK. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. *Theor Comp Sci*. 2006;3106:60–8.
23. Van Iersel L, Keijsper J, Kelk S, Stougie L, Hagen F, Boekhout T. Constructing level-2 phylogenetic networks from triplets. *IEEE/ACM Trans Comput Biol Bioinform*. 2009;6(4):667–81.
24. Wang L, Zhang K, Zhang L. Perfect phylogenetic networks with recombination. *J Comput Biol*. 2001;8:69–78.
25. Gusfield D, Bansal V. A fundamental decomposition theory for phylogenetic networks and incompatible characters. In: Miyano S, Mesirov J, Kasif S, Istrail S, Pevzner PA, Waterman M, eds. *Research in Computational Molecular Biology*. Berlin: Springer; 2005.
26. Gusfield D, Bansal V, Bafna V, Song YS. A decomposition theory for phylogenetic networks and incompatible characters. *J Comput Biol*. 2007;14(10):1247–72.
27. Gusfield D. ReCombinatorics: The Algorithmics of Ancestral Recombination Graphs and Explicit Phylogenetic Networks. Boston: MIT Press; 2014.
28. Dress A, Steel M. Convex tree realizations of partitions. *Appl Math Lett*. 1992;5:3–6.
29. Jukes TH, Cantor CR. Evolution of protein molecules. In: Munro H, ed. *Mammalian Protein Metabolism*. New York, NY: Academic Press; 1969:21–132.

30. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 1980; 16:111–20.
31. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol*. 1987;4:406–25.
32. Deineko V, Rudolf R, Woeginger G. Sometimes traveling is easy: the master tour problem. *SIAM J Discrete Math*. 1995;11:81–93.
33. Thuillard M, Moulton V. Identifying and reconstructing lateral transfers from distance matrices by combining the minimum contradiction method and neighbor-net. *J Bioinform Comput Biol*. 2011;9:453–70.
34. Luo H, Boutry M. Phylogenetic relationships within *Hevea brasilensis* as deduced from a polymorphic mitochondrial DNA region. *Theor Appl Genet*. 1995;91:876–84.