Surprising properties of Maximum Parsimony on phylogenetic networks

Mareike Fischer

Abstract Phylogenetic inference aims at reconstructing the evolutionary relationships of different species given some data (e.g. DNA, RNA or proteins). Traditionally, the relationships between species were assumed to be treelike, so the most frequently used phylogenetic inference methods like Maximum Parsimony were originally introduced to reconstruct phylogenetic trees. However, it has been well-known that some evolutionary events like hybridization or horizontal gene transfer cannot be represented by a tree but rather require a phylogenetic network. Therefore, current research seeks to adapt tree inference methods to networks. In the present paper, we analyze Maximum Parsimony on networks for various network definitions which have recently been introduced. For trees, there is a famous result by Tuffley and Steel which states that under a certain model of evolution, Maximum Parsimony always coincides with Maximum Likelihood. We now show that the various definitions Maximum Parsimony on networks can also be proven to be equivalent to certain functions which are similar to the likelihood concept, and we discuss their biological meaningfulness. We also present some complexity results of finding a most parsimonious network.

Keywords Maximum Parsimony · Maximum Likelihood · phylogenetic networks

PACS PACS 87.23.Kg · PACS 89.75.Hc

Mathematics Subject Classification (2000) MSC 92E10 · MSC 92D15 · MSC 92B05

1 Introduction

The evolutionary history of a set of species is usually described by a phylogenetic tree – in fact, the Tree of Life project (Maddison et al. 2007) even aims at reconstructing the tree of all living species on earth. However, it has been well known that reticulate events such as hybridization or horizontal gene transfer, which for example play an important role in the evolution of plants and bacteria (Arnold 1996; Bogart 2003; Koonin et al. 2001; McDaniel et al. 2010), make evolution non-treelike in the sense that such events cannot be adequately described by phylogenetic trees. Therefore, phylogenetic networks were introduced as a mathematical generalization of the tree concept accommodating reticulate evolution.

Mareike Fischer
Institute of Mathematics and Computer Science, Ernst-Moritz-Arndt-University Greifswald, Walther-Rathenau-Str. 47, 17487 Greifswald, Germany E-mail: email@mareikefischer.de
Mathematically, such networks are far more complex than trees, which can for example be seen by the fact that even problems which are polynomial-time solvable on trees often turn out to be NP-hard on networks. For example, one of the oldest tree reconstruction methods, namely Maximum Parsimony (or MP for short), has long been known to be easy at least for the small parsimony problem: Calculating the so-called parsimony score for a given tree is easy and can be done in polynomial time with the Fitch algorithm (Fitch, 1971), but the big parsimony problem, namely finding the best set of trees, i.e. the ones with the smallest parsimony score, is NP-hard (Foulds and Graham, 1982). But on networks, even the small parsimony problem has recently been found to be NP-hard (Fischer et al, 2015). This is true for various definitions of parsimony on networks, i.e. for different generalizations of the parsimony concept from trees to networks.

Moreover, when methods of phylogenetic inference, like Maximum Parsimony or Maximum Likelihood (ML for short), are generalized such that they can reconstruct not only trees but networks, too, it is mathematically intriguing to figure out which of their properties still hold under such generalizations. For example, in 1997, Tuffley and Steel showed that Maximum Parsimony and Maximum Likelihood actually are equivalent in the sense that they choose the same tree or set of trees, when a simple nucleotide substitution model is assumed and sites evolve according to the no common mechanism model and are therefore independent (Tuffley and Steel, 1997). In the present paper, we investigate the question whether this equivalence also holds for phylogenetic networks (again, assuming no common mechanism) – and we analyze this problem for various definitions of parsimony on networks that can be found in the literature (Fischer et al, 2015; Nakhleh, 2011). However, while it turns out that the equivalence of parsimony to certain functions of the underlying network indeed still holds in most cases (as long as the model under consideration is very simple), we also see that in some cases like the so-called hardwired parsimony score, the function to which parsimony is equivalent is not really a likelihood. This is why we instead speak of a pseudo-likelihood. This pseudo-likelihood has no underlying evolutionary model and is therefore of limited biological value. The fact that Maximum Parsimony is equivalent to such an implausible method shows that hardwired Maximum Parsimony on networks is a purely mathematical concept and probably should not be used with real data. Last but not least, we derive some interesting results concerning the relationship of networks and trees embedded in these networks both for the softwired and the hardwired parsimony definitions. We find that the hardwired parsimony is always optimized by a tree – so in fact, nothing is gained from going from tree to networks. On the other hand, the softwired parsimony seeks to do the opposite – it wants to include as many trees as possible in the optimal network, which also means that an optimal network for any data can be constructed by simply using a network containing all possible trees. While these findings allow for some complexity statements being derived from the complexity of finding the most parsimonious tree, they also show that MP on networks is not really biologically justified and should therefore be modified when applied to real data.

2 Preliminaries

Before we can start our analysis, we need to introduce some concepts and notation.

First, we need to define phylogenetic networks and trees. In this paper, when referring to a phylogenetic network $N$ or just network $N$, we mean a rooted binary phylogenetic hybridization network as defined in (Fischer et al, 2015): Let $X = \{1, \ldots, n\}$ be a finite set. A rooted binary phylogenetic hybridization network $N$ on a set $X$ of species (taxa) is a rooted acyclic directed graph, with no vertices of indegree 1 and outdegree 1, such that all inner nodes have a total degree of 3, except for precisely one node with indegree 0 and outdegree 2, which is called root $\rho$. The leaves have outdegree 0 and indegree 1 and are bijectively labelled...
by the elements of $X$. Vertices with indegree 2 and outdegree 1 are called reticulation vertices or reticulation nodes and the edges with reticulate vertices as head vertices are called reticulation edges. We refer to all other edges as tree edges. An example of a phylogenetic network with one reticulation vertex and four taxa ($X = \{1, 2, 3, 4\}$) is depicted by Figure 1. Note that a rooted binary phylogenetic $X$-tree, (or phylogenetic tree or tree for short), $T$, is a phylogenetic network with no reticulation vertex. In this paper, we use $V(N)$ and $E(N)$ to denote the node and edge set of a phylogenetic network $N$, respectively.

Let $T$ be a phylogenetic $X$-tree and $N$ be a phylogenetic network on $X$. We say that $N$ displays $T$ (or, equivalently, that $T$ is embedded in $N$), if $T$ can be obtained from $N$ by deleting one of the reticulation edges for each reticulation vertex and suppressing the resulting vertices of indegree 1 and outdegree 1. We denote $\tau(N)$ the set of all trees which are displayed by a network $N$. (Note that if there are $i$ reticulation vertices in the network, then there are at most $2^i$ phylogenetic $X$-trees displayed by the network, but the exact number cannot easily be calculated; see (Linz et al., 2013) for more details). Figure 1 shows an example of a phylogenetic network $N$ and all trees displayed by this network.

Next, we need to define the type of data we are relating to phylogenetic trees and networks. These data are given as characters: A function $\chi : X \rightarrow C$, where $C$ is a set of $r$ character states, is called a character. Note that if $|C| = r$, we say that $\chi$ is an $r$-state character. Assuming without loss of generality that $X = \{1, \ldots, n\}$, rather than explicitly writing $\chi(1) = c_1, \chi(2) = c_2, \ldots, \chi(n) = c_n$ for some states $c_i \in C$, we normally write $\chi = c_1c_2\ldots c_n$. Figure 1 depicts a character $\chi = \alpha\alpha\beta\beta$ on four taxa on a network $N$ (and on its embedded trees, respectively). Note that such characters are also often referred to as sites in biological literature, and often it is assumed that $r = 4$, referring to the four DNA nucleotides A, C, G and T. However, our results are not restricted in this way but hold for general $r$.

We say that a function $\hat{\chi} : V(N) \rightarrow C$ is an extension of a character $\chi$ on $N$ if it agrees with $\chi$ on the leaves of $N$, i.e. $\hat{\chi}(i) = \chi(i)$ for all $i$ in $X$. An extension is also depicted by Figure 1 – consider the states assigned to the inner nodes $i$ of $N$, $T_1$ and $T_2$, respectively. Note that if $\hat{\chi}$ is an extension of a character $\chi$ on some phylogenetic network $N$, and if $T$ is a phylogenetic tree displayed by $N$, then $\hat{\chi}_T := \hat{\chi}|_T$ is an extension of $\chi$ on $T$ which is induced by $\hat{\chi}$ if for every $v \in V(T)$, $\hat{\chi}_T(v) = \hat{\chi}(v)$. This means that we can derive an extension of a character on a tree displayed by a network when an extension of this character on the network is given, namely by considering only the nodes which are both in the tree and in the network. Figure 1 depicts an example of this setting. Moreover, for an extension $\hat{\chi}$ of $\chi$, we denote the number of edges $e = (u,v)$ in $N$ on which a substitution or change occurs by $ch_N(\hat{\chi})$, i.e. the number of edges $e = (u,v)$ for which $\hat{\chi}(u) \neq \hat{\chi}(v)$.

Note that biological data normally do not only consist of one character or site, but rather many of them. We denote by $S$ a sequence of $m$ characters, i.e. $S = \chi_1, \ldots, \chi_m$ (for some integer $m \geq 1$). Note that in biological contexts, such a sequence of characters or sites is often referred to as alignment.

Now that we have defined a structure, namely phylogenetic networks and trees, as well as a way to connect data like DNA with this structure via characters, we are finally in a position to define the two concepts of phylogenetic inference we are analyzing in this paper. We start with parsimony on trees. Recall that the parsimony score (PS for short) of a character $\chi$ on a phylogenetic tree $T$ is the minimal number of substitutions required by any extension $\hat{\chi}$ of $\chi$ on $T$, i.e. $PS(\chi, T) = \min_{\hat{\chi}} ch_T(\hat{\chi})$ (Fitch, 1971; Semple and Steel, 2003). A character $\chi : X \rightarrow C$ is called convex on $T$ if $PS(\chi, T) = |\chi(X)| - 1$. Note that this is the minimal possible parsimony score, because if $|\chi(X)|$ states are employed, one of them can be the root state, but at least one change is required to the leaves in the $|\chi(X)| - 1$ remaining states.
The parsimony score of a sequence $S$ of characters is defined as the sum of the parsimony scores of the individual characters (i.e. for $S = \chi_1, \ldots, \chi_m$, we have $PS(S, T) = \sum_{i=1}^{m} PS(\chi_i, T)$), and note that a Maximum Parsimony tree, or MP-tree for short, is a tree with minimal score for a given character or sequence of characters, respectively. In this sense, an MP-tree of a sequence of characters can be regarded as a consensus for the involved characters.

As explained in (Fischer et al., 2015), there are mainly two distinct ways to generalize the parsimony principle from trees to networks. The first one is the so-called softwired parsimony score of a phylogenetic network $N$ and an $r$-state character $\chi$ on $X$. It is defined by considering all trees $T$ displayed by $N$ and taking the minimum value of $ch_T(\hat{\chi})$ over all extensions $\hat{\chi}$ of $\chi$ and all such trees. So

$$PS_{sw}(\chi, N) = \min_{T \in T(N)} \min_{\hat{\chi}} ch_T(\hat{\chi}),$$

where the inner minimum is taken over all extensions $\hat{\chi}$ of $\chi$ to $V(T)$, respectively. The softwired parsimony score of a sequence of characters $S$ is again given by taking the sum: $PS_{sw}(S, N) = \sum_{i=1}^{m} PS_{sw}(\chi_i, N)$. A (not necessarily unique) network with minimal softwired parsimony score is called Softwired Maximum Parsimony network, or SMP-network for short (Fischer et al., 2015). Note that an SMP-network does not represent a consensus for the involved characters of a sequence $S$, because basically each character can independently choose its own MP tree, and an SMP-network by definition contains at least all these trees, but not necessarily an MP tree for the entire sequence $S$.

The softwired parsimony score reflects the biological idea of a generalization of trees to networks, because while in cases of hybridization it is true that parts of the genome come from one ancestral species and other parts from the other one, a single nucleotide can always be traced back to one parent. Therefore, the evolution of a single nucleotide is always treelike, and thus biologically it makes sense to consider all trees embedded in a network in order to calculate the parsimony score.

However, there is also a more mathematically motivated way to extend the parsimony concept from trees to networks: The hardwired parsimony score of a phylogenetic network $N$ and an $r$-state character $\chi$ on $X$ is defined as the minimum value of $ch_N(\hat{\chi})$ over all extensions $\hat{\chi}$ of $\chi$ to $V(N)$; i.e.

$$PS_{hw}(\chi, N) = \min_{\hat{\chi}} ch_N(\hat{\chi}),$$

where the minimum is taken over all extensions $\hat{\chi}$ of $\chi$ to $V(N)$. Again, the hardwired parsimony score of a sequence of characters $S$ is just the sum of the individual scores of its elements: $PS_{hw}(S, N) = \sum_{i=1}^{m} PS_{hw}(\chi_i, N)$. Finally, a (not necessarily unique) network with minimal hardwired parsimony score is called Hardwired Maximum Parsimony network, or HMP—network for short (Fischer et al., 2015). This definition of parsimony on a network does not consider the biological motivation of inheritance of nucleotides from ancestral species, but rather represents a purely graph theoretical extension of parsimony as it is defined on trees (i.e. the number of edges with substitutions is minimized).

Figure 1 depicts both parsimony concepts on networks: Here, the softwired parsimony score is 1, as $T_1$, which is embedded in $N$, only requires one change, and the hardwired parsimony score is 2, as is shown by the dashed edges in $N$. This example shows that the softwired and hardwired scores can differ. But note that it can be easily shown that for a rooted binary phylogenetic tree $T$ and a sequence $S$ of characters, we always have $PS(S, T) = PS_{sw}(S, T) = PS_{hw}(S, T)$; so for trees, the definitions are equivalent.
Next, we want to introduce the second phylogenetic inference method that we are going to analyze, namely Maximum Likelihood. Therefore, we first need to introduce an evolutionary nucleotide substitution model. The model we will consider here is called $N_r$-model \cite{Neyman1971}, which plays a role in various contexts \cite{Tuffley1997}. Note that in biology, the $N_r$-model, i.e. the special case where $r = 4$, is better known as the Jukes-Cantor model. The model is defined as follows: Let $N$ be a phylogenetic network and let $c_1, \ldots, c_r$ be $r$ distinct character states ($r \geq 2$). The $N_r$-model assumes a uniform distribution of states at the root, and moreover it assumes equal rates of substitutions between any two distinct character states \cite{Neyman1971}. Under the $N_r$-model, we denote by $p(e)$ the probability that a substitution of a character state $c_i$ by another character state $c_j$ occurs on some edge $e \in E(N)$ for $c_i \neq c_j$. Furthermore, let $q(e) = 1 - (r - 1)p(e)$ denote the probability that no substitution occurs on edge $e$. Then, in the $N_r$-model we have $0 \leq p(e) \leq \frac{1}{r}$ for all $e \in E(N)$ and $(r - 1)p(e) + q(e) = 1$. Note that the $N_r$-model is time-reversible, i.e. it does not matter where the root of a network is placed, and the rate of change from state $c_i$ to $c_j$ is the same as that from $c_j$ to $c_i$.

Note that in this paper, our model assumption is that whenever there is a sequence of characters rather than a single character, the different characters have evolved under the same as that from $c_j$ to $c_i$.

We will now turn our attention to likelihood concepts. Recall that on a tree $T$, the probability $P(\chi|T, P^T)$ of a character $\chi$ for a given probability vector $P^T$ for changes on the edges of $T$ is the probability that a root state evolves along $T$ to the joint assignment of leaf states induced by $\chi$. Moreover, we have $P(\chi|T, P^T) = \sum \hat{\chi} P(\hat{\chi} \mid T, P^T)$, i.e. the likelihood of a character on a tree can be calculated as the sum of the likelihoods of all possible extensions on this tree \cite{Felsenstein1981}. The Maximum Likelihood of a character on a tree, denoted by max $P(\chi|T)$, is the value of $P(\chi|T, P^T)$ maximized over all possible assignments of substitution probabilities $P^T$, i.e. max $P(\chi|T) = \max P(\chi|T, P^T)$. Moreover, the trees for which max $P(\chi|T)$ is maximal are called Maximum Likelihood trees or ML trees for short.

Next we define concepts similar to ML on trees for networks. However, as we will see in the following, these concepts are not really likelihoods in the mathematical sense, which is why we will talk about pseudo-likelihoods instead. So we are now going to define Maximum pseudo-Likelihood networks. Likelihoodlike functions on a phylogenetic network can be defined in various ways, but we are interested in those which will turn out to be related to parsimony. So we consider two definitions that we will focus on in this paper: We call these networks Softwired Maximum pseudo-Likelihood (SMP$\ell$) networks and Hardwired Maximum pseudo-Likelihood (HMP$\ell$) networks, respectively. We will define these concepts in the following.

Let $N$ be a phylogenetic network and $\chi$ be a character on its leaf set, and let $P^N := (p(e), e \in E(N))$ denote the vector of the probabilities $p(e)$ of a character state change on the edges $e$ of $N$ under the $N_r$-model. If $T$ is a tree which is displayed by $N$, then we define the substitution probabilities vector assigned to the edges of $T$ as $P^T := (p'(e_j), e_j \in E(T))$. Here, for every $e_j \in E(T)$ which is produced by suppressing vertices of indegree 1 and outdegree 1 by contracting edges $e_i$ and $e_k \in E(N)$, we set $p'(e_j) := p(e_i) + p(e_k) - r$. 


can be defined as:

\[ P_{\text{hw}}(\chi \mid N, P^N) = \max_{T \in \mathcal{E}(N)} P(\chi \mid T, P^T) = \max_{T \in \mathcal{E}(N)} \sum_{\hat{T}} P(\hat{T} \mid T, P^T), \]

where \( P^T \) is the restriction of \( P^N \) to \( T \) as explained above and where the summation is taken over all extensions \( \hat{T} \) of \( \chi \) on \( T \). We define the softwired maximum pseudo-likelihood value of \( \chi \) on \( N \) as the maximum value of \( P(\chi \mid T, P^T) \) of the most likely phylogenetic \( X \)-tree which is displayed by the phylogenetic network. This means, the softwired pseudo-likelihood is defined by:

\[ \max P_{\text{hw}}(\chi \mid N) = \max_{T \in \mathcal{E}(N)} \max_{P^T} P(\chi \mid T, P^T), \]

where the inner maximum is taken over all the vectors under the \( N_t \)-model of \( T \) (\( T \) is displayed by \( N \)) and \( P(\chi \mid T, P^T) = \sum_{\hat{T}} P(\hat{T} \mid T, P^T) \), where the summation is taken over all extensions \( \hat{T} \) of \( \chi \) on \( T \). Furthermore, a (not necessarily unique) Softwired Maximum pseudo-Likelihood network (or SMpL network for short) of \( \chi \) is a network for which this value is maximal (i.e. \( \arg\max_N \max P_{\text{hw}}(\chi \mid N) \)). Note that this definition of pseudo-likelihood on networks does not incorporate a probability distribution on the trees embedded in the network as it can be found in the literature (see, e.g. [Nakhleh 2011, Yu et al. 2012]). We will discuss this setting more in-depth in Section 4 in order to investigate how our results would change if this modified definition was assumed.

The second likelihood-like concept on networks which we are going to analyze in detail will be referred to as \textit{hardwired pseudo-likelihood}. Let \( \hat{T} \) be an extension of \( \chi \) on a phylogenetic network \( N \) and let \( p^N := (p(e) : e \in E(N)) \) be the substitution probabilities vector assigned to edges of \( N \) under the \( N_t \)-model. Then, the hardwired pseudo-likelihood of observing character \( \chi \) on \( N \) for the given vector \( p^N \) under the \( N_t \)-model can be defined as:

\[ P_{\text{hw}}(\chi \mid N, P^N) = \sum_{\hat{T}} P(\hat{T} \mid N, P^N), \]

where the summation is taken over all extensions \( \hat{T} \) of \( \chi \) on \( N \), and where we define

\[ P(\hat{T} \mid N, P^N) = \frac{1}{r} \prod_{e \in (u,v); \hat{T}(u) \neq \hat{T}(v)} p(e) \prod_{e \in (u,v); \hat{T}(u) = \hat{T}(v)} q(e). \]

Now, the hardwired maximum pseudo-likelihood value of \( \chi \) on \( N \), denoted by \( \max P_{\text{hw}}(\chi \mid N) \), is the maximum of \( P_{\text{hw}}(\chi \mid N, P^N) \) over all \( P^N \), i.e.

\[ \max_{P^N} P_{\text{hw}}(\chi \mid N) = \max P_{\text{hw}}(\chi \mid N, P^N). \]

Finally, a (not necessarily unique) Hardwired Maximum pseudo-Likelihood network (or HMpL network for short) of \( \chi \) is a network for which this value is maximal (i.e. \( \arg\max_N \max P_{\text{hw}}(\chi \mid N) \)).
Now, by definition of the \( N_r \)-model with no common mechanism, the softwired maximum pseudo-likelihood score and the hardwired maximum pseudo-likelihood score for a sequence of characters \( S = \chi_1, \ldots, \chi_m \) on \( N \) can be calculated as the product of the pseudo-likelihoods of the individual characters due to independence, i.e. \( \max P_{sw}(S \mid N) = \prod_{i=1}^{m} \max P_{sw}(\chi_i \mid N) \) and \( \max P_{hw}(S \mid N) = \prod_{i=1}^{m} \max P_{hw}(\chi_i \mid N) \), respectively.

Note that for a phylogenetic \( X \)-tree \( T \), the softwired and the hardwired definitions are equal, i.e. \( \max P_{sw}(S \mid T) = \max P_{hw}(S \mid T) \), and they also coincide with \( \max P(S \mid T) \).

As with parsimony, the softwired definition of pseudo-likelihood on networks is motivated mainly by biology, as a single nucleotide can be traced back to one ancestral species rather than two, i.e. each nucleotide evolves in a treelike fashion, and the network is considered due to different nucleotides choosing different trees. The hardwired definition, on the other hand, considers the entire network as a whole graph, thus providing a more mathematically motivated extension of the likelihood definition from trees to hybridization networks. In both cases, the described generalizations of the likelihood concept from trees to networks do not provide real likelihoods (which is why we talk about pseudo-likelihoods). In the hardwired case, this can be easily seen: The entire graph is considered, so for a reticulation node, both ancestors play a role. But the model does not specify how the descendant’s state is chosen; there is no probability distribution given for this node. If we included something like that in our model, i.e. if we said that there is a certain probability, say \( \alpha \) that the first reticulation edge is chosen and a certain probability, say \( 1 - \alpha \), that the second reticulation edge is chosen, we would basically be leaving the hardwired scenario, because then we would only consider one reticulation edge at a time rather than the whole graph. So why can this hardwired function not be considered a real likelihood? This is due to the fact that for some choices of \( P^X \), the sum of the hardwired pseudo-likelihood over all characters will be less than 1. The reason is basically that the likelihoods of all characters on a given tree sums up to 1 (as these are indeed likelihoods), and if new edges are added to the tree in order to form a network, then each likelihood will be multiplied with values less than 1 (namely with the substitution probabilities on each additional edge), which decreases the total sum. However, we find this model useful for our purposes, because we seek to analyze parsimony, and – as the present paper will show – hardwired parsimony is equivalent to our definition of hardwired pseudo-likelihood. Note that parsimony on trees is a simple way to explain evolutionary history (by figuring out on which edges in the tree changes must have occurred). But hardwired parsimony on networks does not describe any evolutionary process, because here, too, it is not clear from which parent a reticulation node inherits its state and why. Given that hardwired parsimony already has this property, it is not surprising that it turns out to be equivalent to hardwired pseudo-likelihood, which also does not model inheritance for reticulation vertices.

Note that the softwired case is biologically more reasonable, because a single nucleotide will always come from one parental species (as it cannot be divided). So evolution of a single nucleotide is always treelike. However, our softwired definition of pseudo-likelihood did again not take any distribution on the reticulation edges into account. This can lead to the sum of all characters’ softwired pseudo-likelihoods to be larger than 1, because the sum over the probabilities of all characters on one tree is exactly 1 (as they are real likelihoods), and now we consider the best tree for each character and do not stick to one given tree, which is why the sum can increase. So again, this is why we talk about a pseudo-likelihood. Note that this pseudo-likelihood has already appeared and proven useful in other, non-network, contexts, where it was referred to as “most parsimonious likelihood” (cf. Barry and Hartigan (1987); Steel and Penny (2000)). In Section 4, we will see what happens when we include a distribution on the reticulation edges in our softwired model. As explained above, this will make the model more biologically plausible, but it will (as we will show) no longer be equivalent to softwired parsimony, which we are analyzing in this paper.

As the main purpose here is to analyze some properties of parsimony, we stick to the softwired and hardwired definitions as given in this section. We will discuss some important properties of these two models.
in the following, also highlighting their differences – but, interestingly, the models also have a lot in common. In particular, we want to elaborate their respective close relationship with the introduced pseudo-likelihood concepts.

Fig. 1 A phylogenetic network \( N \) on taxon set \( X = \{1, 2, 3, 4\} \) with one reticulation vertex and the two trees \( T_1 \) and \( T_2 \) it displays. The character \( \alpha\alpha\beta\beta \) is also depicted together with an extension to the inner nodes. The dashed edges represent character state changes.

3 Results

3.1 Establishing the equivalence of MP and pseudo-ML in the softwired and hardwired cases

Before we can use the definitions of the previous section in order to present our results, we want to recall the following famous result by Tuffley and Steel, which basically states the equivalence of Maximum Parsimony and Maximum Likelihood under the \( N_r \)-model with no common mechanism for trees. Subsequently, we will establish similar results for the softwired and hardwired settings on networks, respectively.

**Theorem 1** (Theorem 5 in [Tuffley and Steel, 1997]) Let \( T \) be a phylogenetic tree and let \( S := \chi_1 \ldots \chi_m \) be a sequence of \( r \)-state characters on \( X \). Then, under the \( N_r \)-model with no common mechanism, we have:

\[
\max P(S \mid T) = r^{-PS(S,T) - m}.
\]

Thus, Maximum Likelihood and Maximum Parsimony both choose the same tree(s).
Note that Theorem 1 not only implies that both methods choose the same optimal sets of trees, but rather that both methods induce the same ranking of trees. This means that whenever a tree $T_1$ has a lower parsimony score than another tree $T_2$, the relationship stated in Equation (1) implies that the likelihood of $T_1$ is then higher than that of $T_2$, which means that $T_1$ will both be more parsimonious and more likely than $T_2$. We now use this fact in order to directly establish a similar equivalence result for the softwired setting, i.e. we assume now that a phylogenetic network is given, but we consider it in terms of the set of trees it displays. We use Theorem 1 to show that $\text{SMP}$ and $\text{SMpL}$ are equivalent in this case.

**Theorem 2 (Equivalence of MP and pseudo-ML for softwired networks)** Let $N$ be a phylogenetic network and $S = \chi_1 \ldots \chi_m$ be a sequence of $r$-state characters on $X$. Then, under the $N_r$-model with no common mechanism, we have:

$$\max P_{\text{sw}}(S \mid N) = r^{-PS_{\text{sw}}(S,N) - m}.$$  

Thus, Softwired Maximum Parsimony and Softwired Maximum pseudo-Likelihood both choose the same network(s).

**Proof** We first consider the case $m = 1$, i.e. a single character $\chi$ and only use our definitions from Section 2 as well as Theorem 1:

$$\max P_{\text{sw}}(\chi \mid N) \overset{\text{def}}{=} \max_{T \in \tau(N)} P(\chi \mid T, T^*) \overset{\text{def}}{=} \max_{T \in \tau(N)} P(\chi \mid T) \overset{\text{Th. 1}}{=} \max_{T \in \tau(N)} r^{-PS(\chi, T) - 1} \overset{\text{Th. 1}}{=} r^{-PS_{\text{sw}}(\chi, N) - 1}.$$  

Now for a sequence $S = \chi_1 \ldots \chi_m$ of characters we use the fact that in the no common mechanism (ncm) model, the characters are independent:

$$\max P_{\text{sw}}(S = \chi_1 \ldots \chi_m \mid N) = \prod_{i=1}^{m} r^{-PS_{\text{sw}}(\chi_i, N) - 1} \overset{\text{ncm}}{=} r^{- \sum_{i=1}^{m} (PS_{\text{sw}}(\chi_i, N) + 1)} \overset{\text{Th. 1}}{=} r^{-PS_{\text{sw}}(S, N) - m}.$$  

This completes the proof. □

So in the softwired case the equivalence of the two methods on networks is a direct consequence of the equivalence on trees, because Theorem 1 does not only state that the optimal trees are equal, but the entire ranking induced by parsimony and likelihood is identical. So even if the given network does not contain a globally optimal tree either for parsimony or for the likelihood, it will consider a (not necessarily unique) tree which is best in the set of displayed trees, and this tree will again be the same for parsimony and likelihood.

However, in the hardwired case, things are not so easy, but we will show that nevertheless equivalence still holds. We now state the result before we derive some properties needed to prove it.

**Theorem 3 (Equivalence of MP and pseudo-ML for hardwired networks)** Let $N$ be a phylogenetic network and $S := \chi_1 \ldots \chi_m$ be a sequence of $r$-state characters on $X$. Then, under the $N_r$-model with no common mechanism, we have:

$$\max P_{\text{hw}}(S \mid N) = r^{-PS_{\text{hw}}(S,N) - m}.$$  

Thus, Hardwired Maximum Parsimony and Hardwired Maximum pseudo-Likelihood both choose the same network(s).

In the following, we first stick to the case $m = 1$, i.e. a single character $\chi$. Using the no common mechanism property, it will be easy to derive the required statements for a sequence of characters later on just as we did in the proof of Theorem 2. However, before we can prove Theorem 3 we need some technical preliminaries concerning the hardwired pseudo-likelihood function. The following lemma can be found in Fischer (2009), where also a proof is given. It basically states that multilinear functions on bounded variables have a trivial maximum. The relevance of this lemma for (pseudo-)ML will become apparent subsequently.
Corollary 1 Let $\chi$ be a character on a phylogenetic network $N$. Then under the $N_r$-model, the hardwired pseudo-likelihood $P_{hw}(\chi \mid N, P^N)$ can be maximized at a point where all substitution probabilities are either 0 or $\frac{1}{2}$.

Proof Note that in the $N_r$-model, $P_{hw}(\chi \mid N, P^N) = \frac{1}{2} P_{hw}(\chi \mid N, P^N, \chi(1) = c_1)$, where $c_1$ denotes the character state assigned to taxon 1. This is due to the fact that the $N_r$-model is time-reversible (as explained in Section 2) and therefore an arbitrary node, like e.g. leaf 1, can be chosen to be the root, and the root state is chosen with probability $\frac{1}{2}$. Moreover, by definition of the $N_r$-model, we have $p(e) \leq \frac{1}{2}$. Now,

\[ P_{hw}(\chi \mid N, P^N, \chi(1) = c_1) = \sum_{\hat{\chi}} P(\hat{\chi} \mid N, P^N, \chi(1) = c_1), \]

where the summation is taken over all possible extensions $\hat{\chi}$ of $\chi$, and the state of taxon 1 remains fixed. Now let $v$ be the node adjacent to leaf 1 in $N$, and let $e$ be the edge $(1, v)$. Then, the hardwired pseudo-likelihood may be computed by the recursion

\[ P_{hw}(\chi \mid N, P^N, \chi(1) = c_1) = \sum_{\hat{\chi}(v) = c_1} P(\hat{\chi} \mid N, P^N, \chi(1) = c_1) q(e) + \sum_{\hat{\chi}(v) \neq c_1} P(\hat{\chi} \mid N, P^N, \chi(1) = c_1) p(e), \tag{2} \]

where $p(e)$ is the probability of a substitution on edge $e$, and $q(e) = 1 - (r - 1)p(e)$ is the probability of having no substitution on edge $e$. Clearly, Equation (2) is linear in each $p(e)$. Therefore, the hardwired pseudo-likelihood function $P_{hw}(\chi \mid N, P^N) = \frac{1}{2} P_{hw}(\chi \mid N, P^N, \chi(1) = c_1)$ is multilinear, and the claim follows from Lemma 1 and the fact that $0 \leq p(e) \leq \frac{1}{2}$ in the $N_r$-model. $\square$

So by Corollary 1, $P_{hw}(\chi \mid N, P^N)$ on a phylogenetic network $N$ is maximized under the $N_r$-model by assigning some edges the substitution probability 0 and all others the substitution probability $\frac{1}{2}$. However, we still need to relate this property to parsimony. In some sense, this means that we have to find out how many edges we have to assign $\frac{1}{2}$. We will elaborate this in the following proof of Theorem 3.

Proof (Theorem 3) We first show that $\max P_{hw}(S \mid N) \leq r^{-P_{hw}(S, N)} - m$. We begin with a single character $\chi$ rather than a sequence of characters $S$. So let $N$ be an phylogenetic network, and let the character $\chi$ be a character on the same taxon set. By Corollary 1 we can assume without loss of generality that $P^N$ has the property that all edges of $N$ are assigned substitution probabilities either 0 or $\frac{1}{2}$, because we are considering $\max P_{hw}(\chi \mid N)$, i.e. we are interested in the optimal value of $P_{hw}(\chi \mid N, P^N)$. We now partition the edge set $E(N)$ of $N$ into two sets $E_1$ and $E_0$, such that edges in $E_1$ have substitution probability $\frac{1}{2}$ and all edges in $E_0$ have substitution probability 0. Let $k_N = |E_1|$. Note that if an extension $\hat{\chi}$ of $\chi$ has a substitution on an edge $e$ in $E_0$, then $P_{hw}(\hat{\chi} \mid N, P^N) = 0$, i.e. $\hat{\chi}$ does not contribute to the pseudo-likelihood calculation. This is due to the fact that on all edges in $E_0$, the substitution probability is 0. Note that if all edges of $N$ were in $E_1$, the pseudo-likelihood of all extensions of $\chi$ would be zero. So it is possible to choose $E_1$ and $E_0$ such that not all extensions have a zero pseudo-likelihood, and as the assignment of 0 and $\frac{1}{2}$ was done to maximize $P_{hw}(\chi \mid N, P^N)$, we know that there must be an extension with a positive pseudo-likelihood. Moreover, $k_N \geq P_{hw}(\chi, N)$, since any extension $\hat{\chi}$ of $\chi$
has, by definition of the parsimony score $PS_{hw}(\chi, N)$, at least $PS_{hw}(\chi, N)$ substitutions, all of which must occur on edges in $E_1$ for those extensions which have a non-zero pseudo-likelihood (and as we have shown, there is at least one such extension).

If for an extension $\hat{\chi}$ we have $P(\hat{\chi} | N, P^N) \neq 0$, then $P(\hat{\chi} | N, P^N) = \left( \frac{1}{r} \right)^{k_{N}+1}$ by definition of $E_1$, as on these edges the substitution probabilities $p_e$ as well as the probability $q_e$ for no substitution are all $\frac{1}{r}$, and an additional factor of $\frac{1}{r}$ is needed for the choice of the root state under the $N_r$-model. Therefore, $P(\chi | N, P^N) = \frac{a}{r_{N}^{k_{N}+1}}$, where $a$ is the number of extensions $\hat{\chi}$ that have a non-zero pseudo-likelihood. We now show that $a \leq r_{N}^{k_{N}+PS_{hw}(\chi, N)}$.

Figure 2 illustrates $E_1$ and $E_0$ by dotted and solid edges, respectively. The groups of vertices that are connected by edges of $E_0$ must be assigned the same state by any extension $\hat{\chi}$ of $\chi$ that contributes to the pseudo-likelihood, because there the substitution probabilities are 0. Note that for such extensions, substitutions can only occur on edges of $E_1$, but it is not required that on all such edges there is a substitution. So even though in a phylogenetic network $N$ there may be various paths from one leaf to another one, if both leaves are in a different state, each of these paths must contain at least one edge of $E_1$, because otherwise a change on an edge in $E_0$ would be needed, but this has probability 0.

Fig. 2 Illustration of a phylogenetic network with two reticulation nodes and an assignment of substitution probabilities either 0 (represented by solid edges, set $E_0$) or $\frac{1}{r}$ (represented by dotted edges, set $E_1$) to all edges. When disregarding the edges in $E_1$, blocks containing only edges with substitution probability 0 remain. These blocks are highlighted in the figure. Whenever such a block contains a leaf, it is called labelled. In the above illustration, there are six blocks, four of which are labeled and two of which are unlabeled.

If you disregard all edges in $E_1$, network $N$ gets decomposed into different edge-disjoint components which only contain edges of $E_0$. The vertex set of such components will be referred to as blocks in the following, and these blocks are highlighted in Figure 2. We call a block labeled whenever it contains a leaf. As explained before, any extension $\hat{\chi}$ of $\chi$ that contributes to the pseudo-likelihood $P(\hat{\chi} | N, P^N)$ only allows for changes on edges of $E_1$. Therefore, whenever a block contains a leaf vertex $i$, all vertices in this labeled block must be assigned the same state $\chi(i)$ by such an extension $\hat{\chi}$. So the states of the labeled blocks are fixed by their leaf states, and this is true for all extensions which have a non-zero pseudo-likelihood and thus contribute to the pseudo-likelihood of $\chi$. Thus, the number $a$ of such extensions only depends on the number of unlabeled blocks, which we will call $u_N$, and not on the number $l_N$ of labeled blocks. In particular, as all unlabeled blocks can choose any of the $r$ character states, we have $a = r^{u_N}$.
We now show that $u_N \leq k_N - PS_{hw}(\chi | N)$. This is the crucial part of our proof. We prove this by induction on the number $\hat{r}$ of reticulation nodes in $N$. If $\hat{r} = 0$, $N$ is a phylogenetic tree. In a tree, we know that there are exactly $k_N + 1$ blocks (because there are $k_N$ edges in $E_1$, and if you disregard these, exactly $k_N + 1$ components remain). So we know

$$u_N + l_N = k_N + 1.$$  \hspace{1cm} (3)

Now in a tree, we know that the function $PS_{hw} = PS$ can at most equal $l_N - 1$, because even if all labeled blocks are in different states, Maximum Parsimony will choose one of these states as the root state and require only one change to all other blocks (as there is only one path from the root to each block in a tree). So we have $PS_{hw}(\chi, N) = PS(\chi, N) \leq l_N - 1$ and thus $l_N \geq PS_{hw}(\chi, N) + 1$. Using (3), this leads to $u_N \leq (k_N + 1) - (PS_{hw}(\chi, N) + 1) = k_N - PS_{hw}(\chi, N)$, which is what we wanted to show. It remains to show that we can derive the same inequality for $\hat{r} + 1$ reticulation nodes if we have it for $\hat{r}$.

In the following, let $\hat{\chi}$ denote an extension of $\chi$ on $N$ with a non-zero pseudo-likelihood, and let $\hat{\chi}|_{\hat{N}}$ denote the restriction of $\hat{\chi}$ to a network $\hat{N}$ which results from $N$ by deleting one reticulation edge and suppressing the resulting two nodes of degree 2. Note that by doing so, we get $|E(\hat{N})| = E(N) - 3$. This is due to the fact that if you have, for example, the scenario depicted in Figure 3 five edges from $N$ get replaced by only three edges in $\hat{N}$. Without loss of generality, we assume in the following that the reticulation edge which is deleted when deriving $\hat{N}$ from $N$ is called $e_c$, and that edges $e_a$ and $e_b$ are replaced by $e_{ab}$ and $e_d$ and $e_e$ are replaced by $e_{de}$, as depicted by Figure 3. Moreover, recall that deleting a reticulation edge and suppressing nodes of degree 2 implicitly leads to the deletion of the adjacent reticulation edge and the corresponding reticulation node.

![Fig. 3](image.png)

Now, we distinguish several cases.

**Case 1:** $e_c \in E_0$. As $\hat{N}$ has one less reticulation node than $N$, by the inductive assumption we have $u_{\hat{N}} \leq k_{\hat{N}} - PS_{hw}(\chi, \hat{N})$. Moreover, we have $u_N \leq u_{\hat{N}}$, because by removing at least $e_c$ from $E_0$, it is possible that one block which is labeled in $N$ gets disconnected into two blocks, and one of them might be unlabeled, in which case $u_{\hat{N}}$ would be larger than $u_N$, but it is not possible to decrease $u_N$ by deleting edges. Now we consider all possible subcases depending on $e_a$, $e_b$, $e_d$ and $e_e$:

- If $e_a$, $e_b$, $e_d$ and $e_e$ were in $E_0$, we have $e_{ab}$ and $e_{de}$ both in $E_0$ now, too, and thus $k_N = k_{\hat{N}}$. Note that also $PS_{hw}(\hat{\chi}|_{\hat{N}}, \hat{N}) = PS_{hw}(\hat{\chi}, N)$ for all extensions $\hat{\chi}$ of $\chi$ which have a non-zero pseudo-likelihood, because such extensions cannot have a change on edges in $E_0$ and thus by deleting the reticulation edge we did not modify the number of changes. So, altogether, this leads to:

$$u_N \leq u_{\hat{N}} \leq k_{\hat{N}} - PS_{hw}(\chi, \hat{N}) = k_N - PS_{hw}(\chi, N).$$  \hspace{1cm} (4)
If one of $e_a$ or $e_b$ (or both) were in $E_1$, so is now $e_{ab}$. The same applies to $e_d$, $e_e$ and $e_{de}$, respectively. So $E_1$ has decreased by at least one if we are not in the above case (because then at least one of the edges $e_a$, $e_b$, $e_d$ and $e_e$ was in $E_1$), but by at most 2, because this is the maximum which is achieved when all edges $e_a$, $e_b$, $e_d$ and $e_e$ were in $E_1$, so these four edges from $N$ get replaced by the two $E_1$ edges $e_{ab}$ and $e_{de}$ in $\hat{N}$. So let us now assume that $E_1$ in $N$ as $i$ more edges than $E_1$ in $\hat{N}$ for $i \in \{1, 2\}$. This means we have $k_{\hat{N}} = k_N - i$. Note that then $PS_{hw}(\hat{\chi}, \hat{N}) \geq PS_{hw}(\hat{\chi} | N, \hat{N}) \geq PS_{hw}(\hat{\chi}, N) - i$ for all extensions $\hat{\chi}$ of $\chi$ which have a non-zero pseudo-likelihood, because such extensions can only have changes on edges in $E_1$ and there are $i$ fewer such edges in $\hat{N}$ than in $N$. However, by deleting $i$ edges, the parsimony score can only decrease by at most $i$. So, altogether, this leads to:

$$u_N \leq u_{\hat{N}} \leq k_{\hat{N}} - PS_{hw}(\chi, \hat{N}) \leq (k_N - i) - (PS_{hw}(\hat{\chi}, N) - i) = k_N - PS_{hw}(\hat{\chi}, N).$$

(5)

So, altogether, by Equations (4) and (5), we have $u_N \leq k_N - PS_{hw}(\hat{\chi}, N)$, which is what we wanted to show.

**Case 2:** $e_e$ is in $E_1$. Again, as $\hat{N}$ has one less reticulation node than $N$, by the inductive assumption we have $u_{\hat{N}} \leq k_{\hat{N}} - PS_{hw}(\chi, \hat{N})$. We now consider all possible cases for $e_a$, $e_b$, $e_d$ and $e_e$.

- If $e_a$, $e_b$, $e_d$ and $e_e$ are all in $E_1$, then both the reticulation node as well as the node between $e_a$ and $e_b$ form unlabelled blocks in $N$ which are not contained in $\hat{N}$, so $u_{\hat{N}} = u_N - 2$. Moreover, we have $k_{\hat{N}} = k_N - 3$, because $e_a$ and $e_b$ get replaced by $e_{ab}$, $e_d$ and $e_e$ get replaced by $e_{de}$ and $e_e$ is deleted. In this case, we also have $PS_{hw}(\hat{\chi}, N) \geq PS_{hw}(\hat{\chi} | N, \hat{N}) \geq PS_{hw}(\hat{\chi}, N) - 3$ for all extensions $\hat{\chi}$ of $\chi$ which have a non-zero pseudo-likelihood, because such extensions can only have changes on edges in $E_1$ and there are now 3 fewer such edges in $\hat{N}$ than in $N$. However, by deleting 3 edges, the parsimony score can only decrease by at most 3.

- If $e_a$ and $e_b$ are both in $E_1$ but either $e_d$ or $e_e$ (or both) are not, then the node between $e_a$ and $e_b$ forms an unlabelled block in $N$ which is not contained in $\hat{N}$, so $u_{\hat{N}} = u_N - 1$. Moreover, we have $k_{\hat{N}} = k_N - 2$. By the same argument as above, we have $PS_{hw}(\hat{\chi}, N) \geq PS_{hw}(\hat{\chi} | N, \hat{N}) \geq PS_{hw}(\hat{\chi}, N) - 2$ for all extensions $\hat{\chi}$ of $\chi$ which have a non-zero pseudo-likelihood.

- If $e_d$ and $e_e$ are both in $E_1$ but either $e_a$ or $e_b$ (or both) are not, then the reticulation node forms an unlabelled block in $N$ which is not contained in $\hat{N}$, so $u_{\hat{N}} = u_N - 1$, and we have $k_{\hat{N}} = k_N - 2$, and as in the previous case we have $PS_{hw}(\hat{\chi}, N) \geq PS_{hw}(\hat{\chi} | N, \hat{N}) \geq PS_{hw}(\hat{\chi}, N) - 2$ for all extensions $\hat{\chi}$ of $\chi$ which have a non-zero pseudo-likelihood.

- If all of the above cases do not apply, we have $u_{\hat{N}} = u_N$, because no unlabelled block was deleted when going from $N$ to $\hat{N}$. In this case, we have $k_{\hat{N}} = k_N - 1$ (as $e_e$ was deleted), and by the same argument as above we have $PS_{hw}(\hat{\chi}, N) \geq PS_{hw}(\hat{\chi} | \hat{N}) \geq PS_{hw}(\hat{\chi}, N) - 1$ for all extensions $\hat{\chi}$ of $\chi$ which have a non-zero pseudo-likelihood.

Now we summarize our findings: Note that in all cases, we have $u_{\hat{N}} = u_N - i$, $k_{\hat{N}} = k_N - (i + 1)$ and $PS_{hw}(\hat{\chi} | N, \hat{N}) \geq PS_{hw}(\hat{\chi}, N) - i$ for some $i \in \{0, 1, 2\}$. Using the inductive assumption, this leads to

$$u_N = u_{\hat{N}} - i \leq k_{\hat{N}} - PS_{hw}(\chi, \hat{N}) \leq k_N - (i + 1) - (PS_{hw}(\chi, N) - i) < k_N - PS_{hw}(\hat{\chi}, N).$$

(6)

This completes the induction.

So we have shown inductively that in all cases, $u_N \leq k_N - PS_{hw}(\hat{\chi}, N)$, and therefore $u_N - k_N \leq -PS_{hw}(\hat{\chi}, N)$. Also, we already know that $P(\chi \mid N, P^N) = \frac{a}{P_{N^T}} = \frac{P_{N^T}}{P_{N^T}}$. Combining these results leads to $P(\chi \mid N, P^N) = \frac{P_{N^T}}{r^{-PS_{hw}(\hat{\chi}, N)-1}} \leq \frac{P_{N^T}}{P_{N^T}} = \frac{P_{N^T}}{r^{-PS_{hw}(\hat{\chi}, N)-1}}$. So for a single character $\chi$ we have $P(\chi \mid N, P^N) \leq \frac{P_{N^T}}{P_{N^T}}$. 

\[r^{-PS_{hw}(\hat{\chi}, N)-1} \leq \frac{P_{N^T}}{P_{N^T}} = \frac{P_{N^T}}{r^{-PS_{hw}(\hat{\chi}, N)-1}}.\]
But note that when we maximize the pseudo-likelihood, i.e. when we consider max \( P(\chi \mid N) \), we at least get \( r^{-PS_{hw}(\hat{\chi}, N)} - 1 \). This can be achieved by taking a most parsimonious extension \( \hat{\chi} \) of \( \chi \) on \( N \) and setting the substitution probabilities of those edges to \( \frac{1}{r} \) where \( \hat{\chi} \) has a substitution and those of all other edges 0. Then, there are \( PS_{hw}(\hat{\chi}, N) \) edges with substitution probability \( \frac{1}{r} \), and there is at least one extension (namely \( \hat{\chi} \)) which has a non-zero pseudo-likelihood, so \( \max P(\chi \mid N) \geq \frac{1}{r} \cdot \left( r^{-PS_{hw}(\hat{\chi}, N)} - 1 \right) = r^{-PS_{hw}(\hat{\chi}, N)} - 1 \), where the first factor \( \frac{1}{r} \) is the probability of the root state.

Combining the facts that \( P(\chi \mid N, P^N) \leq r^{-PS_{hw}(\hat{\chi}, N)} - 1 \) and \( \max P(\chi \mid N) \geq r^{-PS_{hw}(\hat{\chi}, N)} - 1 \), we conclude that \( \max P(\chi \mid N) = r^{-PS_{hw}(\hat{\chi}, N)} - 1 \), which shows that Theorem 3 is true for a single character \( \chi \). Now for a sequence \( S = \chi_1 \ldots \chi_m \) of characters we use the fact that in the no common mechanism (ncm) model, the characters are independent:

\[
\max P_{hw}(S = \chi_1 \ldots \chi_m \mid N) = \prod_{i=1}^{m} r^{-PS_{hw}(\chi_i, N) - 1} \overset{\text{ncm}}{=} r^{-\sum_{i=1}^{m} (PS_{hw}(\chi_i, N) + 1)} \overset{\text{def}}{=} r^{-PS_{sw}(S, N) - m}.
\]

This completes the proof of Theorem 3. □

With Theorem 2 and Theorem 3, we have established the fact that the equivalence of MP and pseudo-ML holds both in the softwired as well as in the hardwired case just as it does for the tree case, which was proven by Tuffley and Steel (Tuffley and Steel, 1997). However, these results do not state any characteristics of the optimal networks for MP and pseudo-ML. In the following section, we will show how optimal networks both for MP and pseudo-ML can be traced back to trees.

### 3.2 Optimal MP and pseudo-ML networks

Theorem 3 establishes an equivalence of MP and pseudo-ML for hardwired networks in the same sense as Theorem 1 does for trees, because for every phylogenetic network, by the equation given in the theorem, the maximum pseudo-likelihood can be calculated once the parsimony score is known or vice versa, and indeed the pseudo-likelihood is maximal whenever the parsimony score is minimal. In this sense, the theorem gives a ranking of the entire space of all phylogenetic networks, i.e. a network with a higher pseudo-likelihood will also have a lower parsimony score and vice versa. However, this does not yet characterize optimal networks. We will show in the following that the hardwired parsimony score of a phylogenetic network is actually minimized on a tree and that the softwired parsimony score of a phylogenetic network is minimized on a network which contains all possible trees on the underlying taxon set, and we subsequently derive the implications of this theorem for maximum pseudo-likelihood.

The following theorem was independently discovered also by C. Bryant, S. Linz and C. Semple (unpublished work, personal correspondence), and the second part of this theorem is closely related to findings in (Nakhleh, 2011).

**Theorem 4** Let \( N \) be a phylogenetic network on taxon set \( X \) and let \( T \) be displayed by \( N \). Let \( S = \chi_1, \ldots, \chi_m \) be a sequence of characters on \( X \). Then,

1. \( PS(S, T) \leq PS_{sw}(S, N) \), and
2. \( PS(S, T) \geq PS_{sw}(S, N) \).

**Proof**
1. Let $T \in \tau(N)$. As explained in Section 2 this implies for all extensions $\hat{\chi}$ of a character $\chi$ on $N$ that $\hat{\chi}_T(v) = \hat{\chi}(v)$ for all nodes $v$ which are contained both in $T$ and $N$. Therefore, it follows that $ch_N(\hat{\chi}) \geq ch_T(\hat{\chi})$, because the vertex set of $N$ contains that of $T$, but possibly also more vertices. Now we can conclude:

$$PS_{hw}(S, N) = \sum_{i=1}^{m} PS_{hw}(\chi_i, N) = \sum_{i=1}^{m} \min ch_N(\chi_i) \geq \sum_{i=1}^{m} \min ch_T(\hat{\chi}_iT) = \sum_{i=1}^{m} PS(\chi_i, T) = PS(S, T).$$

Here, the minimum is taken over all extensions of $\chi_i$ on $N$, and $\hat{\chi}_iT$ denotes the restriction of $\hat{\chi}_i$ to $T$. This completes the proof.

2. Let $T \in \tau(N)$. Then, if there is a character $\chi$ in $S$ and a tree $\tilde{T} \in \tau(N)$ such that $PS(\chi, \tilde{T}) < PS(\chi, T)$, then we have $PS_{hw}(S, N) < PS(S, T)$, as character $\chi$ will strictly prefer $\tilde{T}$ to $T$ and thus contribute less to the softwired parsimony score of $N$ than to the score of $T$. On the other hand, if there is no character in $S$ which is optimized by a tree other than $T$, then clearly all characters in $S$ can choose $T$ for the softwired parsimony score and we get $PS_{hw}(S, N) = PS(S, T)$. Altogether, we can conclude $PS_{hw}(S, N) \leq PS(S, T)$. This completes the proof.  

So in fact, it follows from Theorem 4(1) that the set $M^*$ of all HMP networks contains, for each network in $M^*$, also all trees embedded in this network. Intuitively, this is rather obvious, as by deleting edges from a network, one can only decrease the parsimony score but not increase it. Note that this is entirely contrary to the softwired case, as stated in Theorem 4(2): Here, the parsimony score ranges over all trees embedded in the network, and thus having more edges in the network (and therefore more embedded trees) can decrease the parsimony score but never increase it. So Theorem 4(2) implies that the set $M^{**}$ of all SMP networks also contains, for each network $N$ in $M^{**}$, also all networks which contain $N$ – and thus in particular the network containing all binary phylogenetic trees on taxon set $X$ is guaranteed to be in $M^{**}$. Note that this means that we can find an optimal SMP network for $S$ without considering $S$ at all.

So the viewpoints of softwired and hardwired parsimony are quite opposite, and by Theorems 2 and 3 these viewpoints can be directly transferred to pseudo-likelihood, too.

**Corollary 2.** Let $N$ be a phylogenetic network on taxon set $X$ and let $T$ be displayed by $N$. Let $S = \chi_1, \ldots, \chi_m$ be a sequence of characters on $X$. Then,

1. $\max P(S|T) \geq \max P_{hw}(S|N)$, and
2. $\max P(S|T) \leq \max P_{sw}(S|N)$.

**Proof**

1. Using Theorems 1, 3 and 4(1) establishes the following inequality:

$$\max P(S|T) = \underbrace{\frac{1}{m}}_{\text{Th. 1}} \cdot \frac{1}{r^{PS(S,T)-m}} \leq \underbrace{\frac{1}{r^{PS_{hw}(S,N)-m}}}_{\text{Th. 2}} \max P_{hw}(S|N).$$

2. Using Theorems 1, 2 and 4(2) establishes the following inequality:

$$\max P(S|T) = \underbrace{\frac{1}{m}}_{\text{Th. 1}} \cdot \frac{1}{r^{PS(S,T)-m}} \leq \underbrace{\frac{1}{r^{PS_{sw}(S,N)-m}}}_{\text{Th. 3}} \max P_{sw}(S|N).$$

This completes the proof.  

So, interestingly, only the hardwired definition of parsimony and pseudo-likelihood allows for a reduction of the problem of finding an optimal phylogenetic network to finding an optimal phylogenetic tree both for MP and ML. In the softwired case, even though this definition is based on the trees displayed by the network, when seeking an optimal MP network each character in the given data can basically choose its own tree, i.e. a tree on which it has minimal score. All these trees are then combined into a phylogenetic network $N$ with minimal softwired parsimony score. As already mentioned in Section 2 this does not necessarily mean, however, that $N$ contains an MP tree, as an MP tree represents a kind of consensus – if one tree needs to be chosen for all characters at once, then the MP tree is the outcome. But it is not necessary that this compromise is contained in an SMP network. On the other hand, in the hardwired case, any optimal tree $T$ will also be an optimal network, and any optimal network displays only optimal trees. The latter must be true, because if there existed another tree not displayed by $N$ with a strictly better score than the trees displayed by $N$, $N$ would not be optimal; and for all trees in $\tau(N)$ we already know by Theorem 4 that they are parsimonywise at least as good as $N$, so they must be optimal if $N$ is optimal. So we conclude that the equivalence of MP and ML in the hardwired sense leads the more complicated network case back to the tree case.

We finish this section with some complexity results which follow directly from the above observations.

**Theorem 5** For a sequence $S$ of characters, finding an HMP network is NP-hard.

*Proof* Suppose there is a polynomial time algorithm $A$ to find an HMP network $N$ for $S$. We now describe a polynomial time algorithm $B$ to derive an MP tree $T$. First, we consider all edges in $N$ to be directed away from the root. We find a directed spanning tree in $N$, which can be done in linear time (Hastings, 2006). If this spanning tree contains leaves which are not in the leaf set of $N$, we delete them. Moreover, we suppress all resulting nodes of degree 2. All this can be done in time polynomial in the number of leaves $n$ of $N$ as long as the total number of edges and nodes in $N$ is polynomial in $n$. The result of this reduction is a binary phylogenetic tree $T$ displayed by $N$. By Theorem 4 all trees displayed by $N$ are MP trees. So in particular, $T$ is an MP tree, and it was found by combining the polynomial time algorithms $A$ and $B$. However, finding an MP tree is NP-hard (Foulds and Graham, 1982). So the assumption is wrong (unless $P = NP$) and thus, finding a hardwired MP network $N$ is NP-hard. □

**Corollary 3** For a sequence $S$ of characters, finding an HMpL network under the $N_r$-model with no common mechanism is NP-hard.

*Proof* By Theorem 5 finding an HMpL network is equivalent to finding an HMP network. By Theorem 5 the required conclusion follows. □

Note that, as opposed to an HMP network, an SMP network can be found in polynomial time, as we state in the following proposition.

**Proposition 1** For a sequence $S$ of characters, finding an SMP network is polynomial-time solvable (with respect to $|X|$ and $|S|$).

*Proof* Note that for an individual $r$-state character $\chi$ on taxon set $X$ with $|X| = n$, an MP tree can be found in polynomial time by the following procedure. Assume the character employs $k$ states, i.e. $|\chi(X)| = k$. The character partitions the leaf set by its assigned states, i.e. $X = X_1 \cup X_2 \cup \ldots \cup X_k$, where $X_i \neq \emptyset$, $X_i \cap X_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^k X_i = X$, and we have $\chi(x) = \chi(y)$ if and only if there is an $i \in \{1, \ldots, k\}$ such that $x, y \in X_i$. 

Mareike Fischer
Now for each \( i \in \{1, \ldots, k\} \) we choose an arbitrary rooted binary phylogenetic tree \( T_i \) on leaf set \( X_i \). Then we put all these trees into a common tree set \( \mathcal{T} \). Now we combine these trees into one tree on \( X \) as follows: We take two trees \( T \) and \( \tilde{T} \) from \( \mathcal{T} \) and add a new node as the root of these two trees and connect this new root with the roots of \( T \) and \( \tilde{T} \). We add the resulting tree to \( \mathcal{T} \) and delete \( T \) and \( \tilde{T} \) from \( \mathcal{T} \). We repeat this procedure until \( |\mathcal{T}| = 1 \). Then the remaining tree \( T \) in \( \mathcal{T} \) is a rooted binary phylogenetic \( X \)-tree, and by construction (as the partitioning induced by \( \chi \) has been kept), \( \chi \) is convex on \( T \). So for a single character, an MP tree can be constructed like this in polynomial time. We repeat this procedure for all \( m \) characters in \( S \) and then have a collection of at most \( m \) different rooted binary phylogenetic trees, where \( S = \chi_1, \ldots, \chi_m \), i.e. \( m \) is the number of characters in \( S \). We now want to combine these trees into a softwired network in order to derive an SMP network. This can be done in polynomial time: A naive way would be to start with two trees, to identify the leaves and to add a new root and two edges, namely one edge leading from the new root to each of the trees we started with. In the next step, another tree could be added to this network in the same manner. This approach is not elegant, but clearly polynomial. Better ways of combining trees into a common network are e.g. described in (Wu, 2010). In any case, the result is a network which contains for each character \( \chi_i \) in \( S \) a tree on which \( \chi_i \) is convex. As explained in Section 2 this is the best possible case for parsimony, and thus this network is an SMP network, and it was constructed in polynomial time. This completes the proof. □

We end this section with the following corollary, which is a direct consequence of Theorem 2 and Proposition 1.

**Corollary 4** For a sequence \( S \) of characters, finding an SMpL network under the \( N_r \)-model with no common mechanism is polynomial-time solvable.

**Remark 1** Note that any network \( N \) on a taxon set \( X \) which contains all possible trees will be an SMP network for any character sequence \( S \), because such a network in particular contains for each character at least one tree on which it is convex. So SMP networks can be constructed even without considering the data, and the trick is to include many reticulation nodes (and therefore many embedded trees). As we have shown, this is the exact opposite of HMP, which by Theorem 4 is optimized by a tree, which in turn is hard to find for a given sequence \( S \) as long as this sequence contains more than one character.

### 4 Discussion

In the present paper, we have shown that the famous equivalence between Maximum Parsimony and Maximum Likelihood on phylogenetic trees (assuming the \( N_r \)-model with no common mechanism), which was discovered by Tuffley and Steel (Tuffley and Steel, 1997), also holds for phylogenetic networks in the softwired and hardwired sense, respectively – but only if we leave the world of likelihoods and consider pseudo-likelihoods instead. We also derived some interesting implications of this equivalence on the complexity of finding the best phylogenetic networks regarding MP and pseudo-ML. In this respect, it is maybe not surprising that the so-called big parsimony problem, namely finding the most parsimonious network, is NP-hard in the hardwired case (as this problem is also hard for trees). But amazingly, this problem is polynomial-time solvable in the softwired case. We showed that all trees displayed by an HMP network must be MP trees, but an SMP network does not need to contain any MP tree, which simplifies the problem. This highlights the fact that the softwired parsimony concept does not seek any consensus, it basically just represents all conflicts by hybridization nodes and edges, whereas parsimony on trees and also on hardwired networks tries to find a compromise for all input data. Due to our equivalence results, these statements could be shown to also be true for Maximum pseudo-Likelihood.
However, it is important to note that while there are various definitions of likelihood on networks to be found in the literature (see Nakhleh [2011] for an overview), our definition of softwired pseudo-likelihood is not among them (but note that our definition occurs in other contexts, e.g., in Steel and Penny [1987] or [Barry and Hartigan, 1987]), because biologically it is more sensible to multiply each tree in the network with the probability of choosing this tree, whereas our definition disregards this probability (and thus is not really a likelihood). One concept, which is related to our softwired pseudo-likelihood definition, can be found e.g. in (Nakhleh [2011]) and is defined as follows:

\[
P_{\text{sw weighted}}(\chi | N, P^N) = \max_{T \in \tau(N)} [P(T | N, \chi) \cdot P(\chi | T, P^T)].
\]  

We will call this the weighted softwired likelihood, and the maximum of the weighted softwired likelihood over all probability assignments \( P^N \) will be denoted \( \max P_{\text{sw weighted}}(\chi | N) \). The difference between the softwired pseudo-likelihood and the weighted softwired likelihood is that in the latter, each tree likelihood is multiplied by the probability that this tree is chosen amongst the trees in the network. Biologically, it makes sense to distinguish between trees which are likely to be chosen and those which are not. In this case, i.e., when the weighted definition of softwired likelihood is applied, parsimony and likelihood are no longer equivalent. To see this, consider the network and its embedded trees shown in Figure 1. Here, the parsimony scores of \( T_1 \) and \( T_2 \) are 1 and 2, respectively, so assuming the \( N_2 \)-model, by Theorem 2 we have \( \max P_{\text{sw}}(\chi | N) = 2^{-1-1} = \frac{1}{4} \), and the most likely tree in \( N \) is \( T_1 \). For the maximum of the weighted softwired likelihood we get:

\[
\max_{T \in \{T_1, T_2\}} [P(T | N, \chi) \cdot P(\chi | T)] = \max \left\{ P(T_1 | N, \chi) \cdot \frac{1}{4}, P(T_2 | N, \chi) \cdot \frac{1}{8} \right\},
\]

Now if we assume for example that \( T_2 \) is chosen three times as often as \( T_1 \), i.e., \( P(T_1 | N, \chi) = \frac{1}{4} \) and \( P(T_2 | N, \chi) = \frac{3}{4} \), then we have for the maximum of the weighted softwired likelihood:

\[
\max P_{\text{sw weighted}}(\chi | N, P^N) = \max \left\{ \frac{1}{4}, \frac{1}{4}, \frac{3}{8} \right\} = \frac{3}{8}.
\]

Not only is this maximum likelihood value unequal to \( 2^{-P_{\text{sw}}(\chi, N)} = \frac{1}{4} \) as suggested by Theorem 2, it is also achieved by tree \( T_2 \), which maximizes the weighted softwired likelihood in \( N \), whereas \( T_1 \) is strictly better than \( T_2 \) in the parsimony sense. So under the weighted definition of softwired likelihood, the equivalence fails. However, it can be easily seen that as long as all trees have the same probabilities \( P(T | N, \chi) \), the rankings suggested by parsimony and likelihood will still be identical and thus MP and ML will still choose the same networks. On the other hand, if you want to employ a non-uniform distribution on the trees displayed by the network, we conjecture that the equivalence to parsimony can be restored by changing the definition of softwired parsimony accordingly by weighing each tree by a suitable scaling factor.

For future research, it will be interesting to prove this conjecture and to analyze other definitions of likelihood on networks which can be found in the literature concerning their relationship to parsimony. Also, we plan to investigate the impact of scenarios like an imposed molecular clock or bounded substitution probabilities as in Fischer and Thatte [2010] on the equivalence of MP and ML when considering phylogenetic networks. All this will lead to a deeper understanding of the relationship between parsimony and likelihood under the \( N \)-model with no common mechanism on phylogenetic networks.

Our conclusion for this present paper is that our findings basically highlight the fact that MP on networks as it has been defined in the literature so far is not really biologically plausible. For trees, the fact that MP...
is equivalent to ML at least for the $N_r$-model with no common mechanism can maybe be viewed as a justification for MP in the sense that parsimony, which does not require any predefined evolutionary model, can by this equivalence be traced back to a model. On networks, this is different. Both the hardwired and the softwired parsimony are equivalent only to pseudo-likelihoods rather than likelihoods. And even more importantly, we showed that hardwired parsimony always leads to optimal trees, which basically means that nothing can be gained from considering networks; and the softwired parsimony can always be optimized by including as many trees as possible. In fact, any network containing all possible trees will be optimal for any input data, so that we can find an optimal network without even considering the data. So by all means, this shows that the standard definitions of parsimony on networks correspond to biologically implausible concepts. Therefore, we suggest that for future research concerning biological contexts, a weighted softwired parsimony concept similar to the weighted likelihood concept should be used on networks instead of the softwired or hardwired parsimony concepts.

References

Arnold M (1996) Natural Hybridization and Evolution. Oxford University Press, New York
Barry D, Hartigan J (1987) Statistical analysis of hominoid molecular evolution. Statistical Science 2(2):191–207
Bogart J (2003) Genetics and systematics of hybrid species, Science Publishers, Inc., Enfield, New Hampshire, pp 109–134
Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376
Fischer M (2009) Novel mathematical aspects of phylogenetic estimation. PhD thesis, University of Canterbury (University of Canterbury Research Repository).
Fischer M, Thatte B (2010) Revisiting an equivalence between maximum parsimony and maximum likelihood methods in phylogenetics. Journal of Mathematical Biology 72(1):208–220
Fischer M, van Iersel L, Kelk S, Scornavacca C (2015) On computing the maximum parsimony score of a phylogenetic network. SIAM J Discrete Math (SIDMA) 29(1):559–585
Fitch WM (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. Syst Zool 20(4):406–416
Foulds L, Graham R (1982) The Steiner problem in phylogeny is NP-complete. Advances in applied mathematics 3:43–49
Hastings K (2006) Introduction to the Mathematics of Operations Research with Mathematica, 2nd edition. Taylor & Francis Ltd
Koonin E, Makarova K, Aravind L (2001) Horizontal gene transfer in prokaryotes: quantification and classification. Annu Rev Microbiol 55:709–742
Linz S, St John K, Semple C (2013) Counting trees in a phylogenetic network is \#p-complete. SIAM journal on computing 42:1768–1776
Maddison D, Schulz KS, Maddison W (2007) The tree of life web project. In: Zhang ZQ, Shear W (eds) Linnaeus Tercentenary: Progress in Invertebrate Taxonomy., vol 1668 (1–766), Zootaxa, pp 19–40
McDaniel L, Young E, Delaney J, Ruhnau F, Ritchie K, Paul J (2010) High frequency of horizontal gene transfer in the oceans. Science 330:6000:50
Nakhleh L (2011) Evolutionary phylogenetic networks: models and issues. In: Problem solving handbook in computational biology and bioinformatics, Springer, pp 125–158
Neyman J (1971) Molecular studies of evolution: a source of novel statistical problems. Statistical decision theory and related topics pp 1–27
Semple C, Steel M (2003) Phylogenetics. Oxford University Press

Steel M, Penny D (2000) Parsimony, likelihood, and the role of models in molecular phylogenetics. Molecular Biology and Evolution 17(6):839–850

Tuffley C, Steel M (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bulletin of mathematical biology 59:581–607

Wu Y (2010) Close lower and upper bounds for the minimum reticulate network of multiple phylogenetic trees. Bioinformatics 26(12):i140–i148

Yu Y, Degnan J, Nakhleh L (2012) The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. PLoS Genet 8(4) 8(4):e1002,660

Acknowledgements. I wish to thank Simone Linz for helpful discussions concerning the big parsimony problem on softwired and hardwired networks, respectively, and Steven Kelk for helpful comments on some complexity issues. Moreover, I want to thank two anonymous reviewers for their very helpful comments and suggestions.
