NetCombin: An algorithm for optimal level-k network construction from triplets

Hadi Poormohammadi¹,²* Mohsen Sardari Zarchi¹

¹ Department of Computer Engineering, Meybod University, Meybod, Iran
² School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

* pormohamadi@meybod.ac.ir

Abstract

Phylogenetic networks construction is one the most important challenge in phylogenetics. These networks can present complex non-treelike events such as gene flow, horizontal gene transfers, recombination or hybridizations. Among phylogenetic networks, rooted structures are commonly used to represent the evolutionary history of a species set, explicitly. Triplets are well known input for constructing the rooted networks. Obtaining an optimal rooted network that contains all given triplets is main problem in network construction. The optimality criteria include minimizing the level and the number of reticulation nodes. The complexity of this problem is known to be NP-hard. In this research, a new algorithm called Netcombin is introduced to construct an optimal network which is consistent with input triplets. The innovation of this algorithm is based on binarization and expanding processes. The binarization process innovatively uses a measure to construct a binary rooted tree $T$ consistent with the maximum number of input triplets. Then $T$ is expanded in an intellectual process by adding minimum number of edges to obtain final network with the minimum number of reticulation nodes. In order to evaluate the proposed algorithm, NetCombin is compared with four state of the art algorithms, RPNCH, NCHB, TripNet, and SIMPLISTIC. The experimental results on real data indicate that by considering the trade-off between speed and precision, the NetCombin outperforms the others.

Author summary

Hadi Poormohammadi got his PhD in Mathematics, Applied combinatorics from Shahid Beheshti University, Tehran, Iran in 2013. He is now working as an assistant professor at the Faculty of Computer Engineering, Meybod University. His research interests include Combinatorics, Graph theory and Bioinformatics.

Mohsen Sardari Zarchi got his PhD in computer engineering, Artificial Intelligence from University of Isfahan in 2015. He is now working as an assistant professor at the Faculty of Computer Engineering, Meybod University. His research interests include Deep learning, Image processing, Artificial intelligence and Bioinformatics.

Introduction

Phylogenetics is a branch of bioinformatics that studies and models the evolutionary relations between a set of species or organisms (formally called taxa) [1,2]. The tree
structure is the basic model which can show the history of tree-like events such as mutation, insertion and deletion appropriately [1–5]. The main disadvantage of the tree model is its disability to show non-tree-like events (more abstractly, reticulate events) like recombination, hybridization and horizontal gene transfer [2,12]. To overcome this weakness, phylogenetic networks are introduced to generalize phylogenetic trees and represent reticulate events [1,2,6–11].

The structures of trees and networks can be divided into two groups, rooted and un-rooted. Rooted structures can show reticulate events, explicitly. Hence this structure has received more attention recently for constructing networks. The rooted structures are directed graphs which contains a unique vertex called root with in-degree 0 and out-degree at least 2 [1,2,6]. Figure 1a shows an example of a rooted tree.

**Fig 1.** (a) A rooted tree in which the out-degree of the root is 3. (b) A rooted binary tree.

Usually rooted structures are represented in the binary form, i.e. the out-degree of each vertex is at most 2. In a rooted binary tree, the out-degree of all vertices except the leaves are 2, the out-degree of leaves are 0, the in-degree of all vertices except the root, is 1 (Figure 1b). Formally in rooted structures a common ancestor for a given set of taxa is considered as root [1,2,6].

One famous approach to build phylogenetic networks is constructing them up from small trees or networks [1,2]. Triplet is the smallest tree structure which shows the evolutionary relation between three taxa. The symbol $ij|k$ is used to show a triplet $t$ with $i,j$ in one side of $t$ and $k$ in another side of $t$ (Figure 2) [2,6].

**Fig 2.** A triplet $ij|k$.

Triplets are commonly considered as a standard input for building rooted structures [2,6]. These small tree structures are usually obtained from a set of biological sequence data using standard methods such as Maximum Likelihood (ML) and Maximum Parsimony (MP) [5,12]. Also in some cases, the output of some biological experiments is directly in the form of triplets [13]. Moreover, in some experiments, triplets are generated randomly to evaluate a model [5].

For a given set of triplets as input, the main challenge is to construct a rooted structure (tree or network) which contains all triplets or equivalently, all triplets are consistent with the obtained structure [2,6,11,12]. Formally, a triplet is consistent with a rooted structure when the triplet is a subgraph of that structure [2,5,6,11,12]. The majority preference is to obtain a rooted tree structure. However, as mentioned before, the tree structure can’t represent reticulate events and eventually the reticulate events can’t be covered. So, in this case the network construction should be considered.

In order to build a rooted network from a set of triplets, several algorithms were introduced recently [2,6,7,11,12,14]. The well-known algorithms are TripNet [6], SIMPLISTIC [12], NCHB [14] and RPNCH [7]. These algorithms try to find an optimal rooted phylogenetic network that is consistent with a given set of triplets. Formally, a rooted phylogenetic network (network for short) is a directed acyclic graph (DAG) (Figure 3) that is connected and each vertex satisfies one of the following four categories:

(i) A unique root node with in-degree 0 and out-degree 2. (ii) Tree nodes with in-degree 1 and out-degree 2. (iii) Reticulation nodes with in-degree 2 and out-degree 1. (iv) Leaves with in-degree 1 and out-degree 0 (Figure 4a). A network is called a network on $X$ if the set of its leaves is $X$. For example the network of figure 4a is a network on $X = \{l_1, l_2, \ldots, l_7\}$.

The optimal network is defined based on the two optimality criterions i.e. minimizing the number of reticulation nodes and the level of the final network [2,6,12].
The desirable goal is to construct a rooted network with no reticulate events i.e. a rooted phylogenetic network that contains all given triplets. However, usually there is at least one information in the set of input triplets \([6,16]\). More precisely, a set of triplets \(\tau\) is called dense if for each subset of three taxa there is at least one information in the set of triplets \([6,16]\). For example for a given set of taxa \(X = \{a, b, c, d, e\}\), the set of triplets \(\tau = \{ab|c, ad|b, be|a, ac|d, de|a, bd|c, be|d, de|c\}\) is dense.

As mentioned above, density is a critical constraint concerning with constructing a rooted phylogenetic network that contains all given triplets. However, usually there is no constraint on the input triplets and in most cases the input triplets might not be dense. So, introducing efficient heuristic methods to solve this problem is demanding. The desirable goal is to construct a rooted network with no reticulate events i.e. a rooted tree structure. BUILD is the algorithm that was introduced for obtaining a tree structure from a given set of triplets if such a tree exists \([17]\). In fact, BUILD algorithm decides in polynomial time order if there is a rooted phylogenetic tree that contains all given triplets and then produces an output if such a tree exists. Figure 5, indicates an example of BUILD algorithm steps for the given set of triplets and then produces an output if such a tree exists. Figure 5, indicates an example of BUILD algorithm steps for the given set of triplets. In tree construction process for a given set of triplets, BUILD stops, it means that there is no tree structure for the given set of triplets. Figure 6 shows an example for the set of triplets \(\tau = \{bc|a, bd|a, cd|a, bc|d, cd|b\}\) in which BUILD algorithm stops. In this case, the main goal is to construct a network structure similar to a tree as much as possible. In other words, constructing a rooted phylogenetic network with the minimum reticulate events is the main challenge.

The simplest possible non-treelike structure (network structure) is level-1 rooted phylogenetic network which also known as galled tree \([15,16]\). Figure 7 shows an example of building a level-1 rooted phylogenetic network. The level of a rooted phylogenetic network is defined as the maximum number of reticulation nodes in each biconnected component (Figure 4b).

Generally the problem of constructing an optimal rooted phylogenetic network consistent with a given set of triplets is known to be NP-hard \([15,16]\). When the structure of the input triplets is dense, this problem can be solved in polynomial time order \([16]\). A set of triplets \(\tau\) is dense if for a given set of taxa \(X\) and each subset of three taxa \(\{i, j, k\}\) one of the triplets \(ij|k\) or \(ik|j\) or \(jk|i\) belongs to \(\tau\) \([6,16]\).

The Aho graph \(AG(\tau)\) is defined based on \(\tau\). The set of nodes are \(X = L(\tau) = \{a, b, c, d, e, f\}\) and two nodes \(i, j \in X\) are connected iff there is a node \(x \in X\) such that \(ij|x \in \tau\). Also \(AG(\tau|A)\) in which \(A \subseteq X\) is defined in a similar way. Note that in this way the induced graph of \(AG(\tau)\) on the set of nodes \(A \subseteq X\) is considered i.e. the set of nodes are \(A\) and \(i, j \in A\) are connected iff there is a node \(x \in A\) such that \(ij|x \in \tau\). (a) \(AG(\tau)\). (b) Based on \(AG(\tau)\) the resulting tree is obtained. (c) \(AG(\tau|e, f)\). (d) \(AG(\tau|a, b, c, d)\). (e) Based on \(AG(\tau|e, f)\) the resulting tree is obtained. (f) Based on \(AG(\tau|a, b, c, d)\) the resulting tree is obtained. (g) \(AG(\tau|b, c, d)\). (h) Based on \(AG(\tau|b, c, d)\) the resulting tree is obtained. (i) \(AG(\tau|e, d)\). (j) Based on \(AG(\tau|c, d)\) the resulting tree is obtained. (k) The final tree consistent with given \(\tau\) that is obtained from BUILD algorithm by reversing its steps.
Fig 6. (a) The simplest possible network consistent with the given \( \tau = \{bc|a, bd|a, cd|a, bc|d, cd|b\} \). (b) \( AG(\tau) \). (c) Based on \( AG(\tau) \) the resulting tree is obtained. (d) \( AG(\tau|b, c, d) \). The graph is still connected. (e) BUILD algorithm stops based on step d.

event of a galled tree. If level-1 networks can not represent all input triplets, more complex (higher level) networks are considered to achieve consistency. LEV1ATHAN is a well-known algorithm to construct level-1 networks [18]. In [12] an algorithm is introduced that produces at most a level-2 network (Figure 4).

Fig 7. A level-1 network (galled tree) with two reticulation nodes.

When more complex networks are needed (e.g. figure 8), not restricted algorithms such as NCHB, TripNet, RPNCH and SIMPLISTIC are applicable which try to construct a consistent network with the optimality criterions (the level and the number of reticulation nodes) [6,7,12,14]. Among the above four mentioned algorithms, SIMPLISTIC is not exact and just works for dense sets of triplets while for the other three methods there is no constraint on the input triplets. This is one of the SIMPLISTIC disadvantages. Moreover for complex networks SIMPLISTIC is very time consuming and has not the ability to return an output in an appropriate time [6].

Fig 8. Two different level-3 networks with three reticulation nodes \( r_1, r_2, r_3 \).

TripNet has three speed options: slow, normal, and fast. The slow option returns a network near to an optimal network. Normal option works faster compared to slow option, but its network is more complex compared to slow option. Note that slow and normal options return an output in an appropriate time for input triplets consistent with simple and low level networks. However these two options are not appropriate for large data, because by increasing the number of taxa, the set of triplets corresponds to them are consistent with high level networks. Fast option usually output a network in an appropriate time but its network is more complex compared to the two other options. This option is used when the slow and normal options have not the ability to return a network in an appropriate time. It means that fast option just try to output a network and does not consider the optimality criterions. In summary, TripNet has not the ability to return an optimal network in an appropriate time, when input data is large [6]. NCHB is an improvement of TripNet which tries to improve the complexity of the TripNet networks but like TripNet it has not the ability to return an optimal network in an appropriate time for large data [14].

RPNCH is a fast method for constructing a network consistent with a given set of triplets, but its output is usually more complex considering the two optimality criterions compared to SIMPLISTIC and TripNet networks. It means that although RPNCH is fast but on average, the RPNCH networks are far away from the optimality criterions [7].

Generally none of the above four methods have the ability to return a network near to an optimal network consistent with a given set of input triplets in an appropriate time. So the focus of this paper is to introduce a new method called NetCombin for constructing a network near to optimal in an appropriate time without any constraint on the input triplets. In this research our innovation is based on the binarization and expanding processes. In the binarization process nine measures are used innovatively to construct binary rooted trees consistent with the maximum number of input triplets. These measures are computed based on the structure of the tree and the relation between input triplets. In the expansion process which converts obtained binary tree
Definitions and notations

In this section the basic definitions that are used in the proposed algorithm, are presented formally. From here, a set of triplets and a network are indicated by \( \tau \) and \( N \) respectively.

A rooted phylogenetic tree (tree for short) on a given set of taxa \( X \) is a rooted directed tree that contains a unique node \( r \) (root) with in-degree zero and out-degree at least 2. In a tree, leaves are with in-degree 1 and out-degree 0 and are distinctly labeled by \( X \). Also inner nodes i.e. nodes except root and leaves, has in-degree 1 and out-degree at least 2 \([2,6]\). Figure 1, indicates an example of a tree on \( X = \{a, b, c, d, e, f\} \).

The symbol \( L_N \) denotes the set of all leaf labels of \( N \). \( N \) is a network on \( X \) if \( \text{L}_N = X \). A triplet \( ij|k \) is consistent with \( N \) or equivalently \( N \) is consistent with \( ij|k \) if \( \{i,j,k\} \subseteq L_N \) and \( N \) contains two distinct nodes \( u \) and \( v \) and pairwise internally node-disjoint paths \( u \rightarrow i, u \rightarrow j, v \rightarrow u \), and \( v \rightarrow k \). For example, figure 9 shows that triplets \( ij|k \) and \( jk|i \) are consistent with the given network, but \( ik|j \) is not consistent. A set of triplets \( \tau \) is consistent with a network \( N \) (or equivalently \( N \) is consistent with \( \tau \)) if all the triplets in \( \tau \) are consistent with \( N \). \( \tau(N) \) denotes the set of all triplets that are consistent with \( N \). Let \( L(\tau) = \bigcup_{t \epsilon \tau} L_t \). \( \tau \) is a set of triplets on \( X \) if \( L(\tau) = X \) \([6]\).

Fig 9. (a) Triplet \( ij|k \) is consistent with the network. (b) Triplet \( jk|i \) is consistent with the network.

Binarization is a basic concept, defined as follows. Let \( T \) be a rooted tree and \( x \) be a node with \( x_1, x_2, \ldots, x_k, k \geq 3 \) children. These \( k \) children are partitioned into two disjoint subsets \( X_l \) and \( X_r \). Let \( X_l = \{x'_1, x'_2, \ldots, x'_l\} \) and \( X_r = \{x'_{l+1}, \ldots, x'_k\} \) in which \( x'_1, x'_2, \ldots, x'_k \) is an arbitrary relabeling of \( x_1, x_2, \ldots, x_k \). If \( |X'_j| > 1 \) then create a new node \( x_l \), remove the edges \((x, x'_1), (x, x'_2), \ldots, (x, x'_l)\) and create the edges \((x, x_{l+1}), (x, x_{l+2}), \ldots, (x, x'_k)\). Do the same process if \( |X'_r| > 1 \). Continue the process until the out-degree of all nodes except the leaves, be 2. The new tree is called a binarization of \( T \). Figure 10 shows an example of a non-binary tree and its two different binarizations. If \( T_2 \) is a binarization of \( T_1 \) then \( \tau(T_1) \subseteq \tau(T_2) \) \([6]\).

Fig 10. (a) A non-binary tree. (b,c) Two binarizations of figure 10a.

\( G_\tau \) the directed graph related to \( \tau \), is defined by \( V(G_\tau) = \{(i,j) : i, j \in L(\tau), i \neq j\} \) \((i,j) \) is denoted by \( ij \) for short and
\( E(G_\tau) = \{(ij, ik) : ij|k \in \tau\} \cup \{(ij, jk) : ij|k \in \tau\} \) \([6]\). (E.g. figures 13b and 14b). The height function of a tree and network is defined as follows. Let \( \binom{X}{2} \) denotes the set of all subsets of \( X \) of size 2. A function \( h : \binom{X}{2} \rightarrow N \) is called a height function on \( X \) \([6]\). let \( T \) be a tree on \( X \), with the root \( r \), \( c_{ij} \) be the lowest common ancestor of \( i, j \in X \), and \( l_T \) denotes the length of the longest directed path in \( T \). Let \( x, y \) be two
Let \( G_\tau \) be a DAG and \( l_{G_\tau} \) be the length of the longest directed path in \( G_\tau \). Assign \( l_{G_\tau} + 1 \) to the nodes with out-degree=0 and remove them. Assign \( l_{G_\tau} \) to the nodes with out-degree=0 and remove them. Define \( h_{G_\tau}(a,b) \), \( a, b \in L(\tau) \) and \( a \neq b \) as the value that is assigned to the node \( ab \in V(G_\tau) \) and call it the height function related to \( G_\tau \) [6]. For example figures 13a to 13d. If \( \tau \) is consistent with a tree then \( G_\tau \) is a DAG and \( h_{G_\tau} \) is well defined [6]. Let \( r \) be the root of a given network \( N \) and \( l_N \) be the length of the longest directed path in \( N \). For each node \( a \) let \( d(r,a) \) be the length of the longest directed path from \( r \) to \( a \). For any two nodes \( a \) and \( b \), \( a \) is an ancestor of \( b \) if there is a directed path from \( a \) to \( b \). In this case \( b \) is lower than \( a \). For any two nodes \( a, b \in L(N) \) a node \( c \) is called a lowest common ancestor of \( a \) and \( b \) if \( c \) is a common ancestor of \( a \) and \( b \) and there is no common ancestor of \( a \) and \( b \) lower than \( c \). For any two \( a, b \in L(N), a \neq b \), let \( C_{ab} \) denotes the set of all lowest common ancestor of \( a \) and \( b \). For each \( a, b \in L(N) \), define \( h_N(a,b) = \min\{l_N - d(r,c) : c \in C_{ab}\} \) and call it the height function of \( N \) [6]. For example for the network of figure 6a, \( l_N = 4 \) and \( h_N(a,b) = h_N(a,c) = h_N(a,d) = 4 \), \( h_N(b,d) = 3 \), and \( h_N(b,c) = h_N(c,d) = 2 \).

A quartet is an un-rooted binary tree with four leaves. The symbol \( ij|kl \) is used to show a quartet in which \( i, j \) and \( k, l \) are its two pairs. Each quartet contains a unique edge for which two its endpoints are not leaves. This edge is called the inner edge of the quartet (See figure 12) [6].

**Fig 12.** A quartet \( ij|kl \) with the leaves \( \{i, j, k, l\} \).

**Method**

In order to build a network \( N \) consistent with a given set of triplets \( \tau \), the height function \( h_N \) related to \( \tau \) is defined [6]. The height function is a measure that is used to obtain a basic structure of the final network (\( N \)) [6]. This basic structure is in the form of a rooted tree. The height function enforce that the obtained rooted tree be consistent with maximum number of triplets of \( \tau \). In this research, the height function is obtained as follow [6]:

**Constructing tree using height function**

Let \( T \) be a tree with its unique height function \( h_T \) and \( i, j \in L_T \). The triplet \( ij|k \) is consistent with \( T \) iff \( h_T(i, j) < h_T(i, k) \) or \( h_T(i, j) < h_T(j, k) \) [6]. Moreover for a given network \( N \) and \( i, j, k \in L_N \) with the height function \( h_N \) if \( h_N(i, j) < h_N(i, k) \) or \( h_N(i, j) < h_N(j, k) \) then the triplet \( ij|k \) is consistent with \( N \) [6]. The above two items imply that the following Integer Programming (IP) \( IP(\tau, s) \) is established for a given triplets \( \tau \) with \( |L(\tau)| = n \) [6].

Maximize 
\[
\sum_{1 \leq i, j \leq n} h(i, j)
\]

Subject to: 
\[
\begin{align*}
  h(i, k) - h(i, j) & > 0 & & ij|k \in \tau, \\
  h(j, k) - h(i, j) & > 0 & & ij|k \in \tau, \\
  0 & < h(i, j) & & 1 \leq i, j \leq n.
\end{align*}
\]
The solution of the above IP provides a criterion to obtain the basic tree structure. Ideally it is expected that the above IP has a feasible solution i.e. a solution that satisfies all its constraints. If there is a tree consistent with a given \( \tau \) then the above IP has a feasible solution and the solution that maximizes the above IP is the height function of a tree that is consistent with \( \tau \). More precisely in this case \( h_{T_\tau} \) is the unique optimal solution to the IP(\( \tau, l_{G_\tau} + 1 \)) in which \( T_\tau \) is the unique tree that is constructed by BUILD (or equivalently HBUILD) [6]. So in this case by using HBUILD the desired tree consistent with \( \tau \) can be constructed in polynomial time based on the optimal solution [6] Figure 13 indicates an example of the HBUILD process for the given \( \tau = \{cd|b, cd|a, bd|a, bc|a\} \).

Fig 13. The HBUILD process for \( \tau = \{cd|b, cd|a, bd|a, bc|a\} \). (a) Firstly a tree \( T \) is considered. Then the set \( \tau \) of all triplets consistent with \( T \) is obtained. (b) \( G_\tau \) is obtained from \( \tau \). \( l_{G_\tau} = 2 \) and \( h(a, b) = h(a, c) = h(a, d) = 3 \) since the out-degree of the nodes \( ab, ac, ad \) are zero. (c) The nodes with out-degree zero and their related edges are removed (removed edges are dashed). \( h(b, c) = h(b, d) = 2 \). (d) The nodes with out-degree zero and their related edges are removed. \( h(c, d) = 1 \). (e) The weighted complete graph \( (G, h) \) related to the obtained height function in which the edge \( \{i, j\} \) has weight \( h(i, j) \). (f) Removing the edges with maximum weight from \( (G, h) \). (g) The tree that is obtained based on the graph of figure 13f. (h) Removing the edges with maximum weight from the resulting graph on the set of nodes \( b, c, d \). (i) The tree that is obtained based on the graph of figure 13h. (j) Removing the edges with maximum weight from the resulting graph on the set of nodes \( c, d \). (k) The tree that is obtained based on the graph of figure 13j. (l) The final tree consistent with given \( \tau \) that is obtained from HBUILD algorithm by reversing its steps.

Generally the above IP has a feasible solution if the graph \( G_\tau \) is a DAG and in this case the minimum \( s \) that gives a feasible solution for IP(\( \tau, s \)) is \( l_{G_\tau} + 1 \) [6]. So for a given \( \tau \) the IP might have a feasible solution although there is no tree consistent with \( \tau \). In the worst case, there is no tree consistent with a given \( \tau \) and no feasible solution for the above IP i.e. equivalently the graph \( G_\tau \) is not a DAG. To overcome this flaw, the goal is to remove minimum number of edges from \( G_\tau \) (minimum number of criterions from the IP) to lose minimum information. The problem of removing minimum number of edges from a directed graph to obtain a DAG is known as the Minimum Feedback Arc Set problem, MFAS problem for short. MFAS is NP-hard [19]. The heuristic method that is introduced in [14] is used to obtain a DAG from \( G_\tau \). For simplicity the new graph that is a DAG is called \( G_\tau \) again. Now the height function \( h_{G_\tau} \) related to \( G_\tau \) is the desired solution.

In the following the goal is to obtain a tree structure from the obtained \( h_{G_\tau} \). In the initial step HBUILD is applied on \( h_{G_\tau} \). The ideal situation is when HBUILD continues until a tree structure is obtained. However, HBUILD may stop in one of its subsequence steps. More precisely, Let \( (G, h) \) be the weighted complete graph related to \( h_{G_\tau} \). HBUILD algorithm removes the edges with maximum weight from \( (G, h) \). If by removing the edges with maximum weight from each connected component, the resulting graph becomes disconnected then this process continues iteratively until each connected component contains only one node. The basic tree structure (??TTT) is obtained by reversing the above disconnecting process in HBUILD (See figure 13).

If by removing the edges with maximum weight from a connected component \( C \), the resulting graph \( C' \) remains connected, then HBUILD halts. Hence, the goal is to disconnect the obtained connected component (\( C' \)). In order to disconnect \( C' \), similar to RPNCH [7] three different processes can be performed as follow (See figure 14) :

1. The process of removing the edges with maximum weights from \( C' \) is continued until \( C' \) becomes disconnected.
ii. The Min-Cut method is applied on $C'$. Min-Cut is a method that removes minimum weights sum of removed edges in a way that the resulting graph is converted into two connected components.

iii. Let $w$ be maximum weight of all edges in $C'$. The new weights are computed based on the current weights and $w$. For each edge with weight $m$, its new weight is assigned as:

$$m_{new} = w - m + 1.$$ 

Then Min-Cut method is applied on the updated graph.

In this research, for each connected component, the above three processes is applied and then by using HBUILD, three possible tree structures are obtained. From here, without loss of generality, the symbol $T_{int}$ is used to show the tree structure obtained from HBUILD or the three tree structures gained from the above processes.

Fig 14. A basic tree structure construction process for a given

$\tau = \{bc|a, bd|a, cd|a, bc|d, cd|h\}$. $\tau$ is not consistent with a tree and HBUILD stops in some steps. (a) $\tau$ is obtained from the given network and the network is the optimal network consistent with $\tau$. (b) $G_\tau$ related to $\tau$. (c) $G_\tau$ is not a DAG and contains a cycle. The edge $(cd, bc)$ is removed to obtain a DAG. For simplicity the new DAG is called $G_\tau$ again. $|G_\tau| = 3$. (d) The nodes $ab, ac, ad$ with out-degree zero and their related edges are removed. $h(a, b) = h(a, c) = h(a, d) = 4$. (e) The node $bd$ with out-degree zero and its related edges are removed. $h(b, d) = 3$. (f) The node $cd$ with out-degree zero and its related edges are removed. $h(c, d) = 2$. Finally the node $bc$ with out-degree zero is removed. $h(b, c) = 1$. (g) The graph $(G, h)$ based on the obtained $h_{G_\tau}$ related to $G_\tau$. (h) The edges with maximum weight are removed from $(G, h)$. The resulting graph is disconnected. So HBUILD continues. (i) The tree structure related to the figure 14h. (j) The edges with maximum weight are removed from the graph of figure 14h. The resulting graph still connected. So HBUILD stops. From here three criterions are applied to disconnect the connected subgraphs. (k) To disconnect the connected component of figure 14j the process of removing the edges with maximum weight (the edges with weight 2) continues. The resulting graph becomes disconnected. In the remaining steps, HBUILD is applied and finally a tree structure is obtained by reversing the steps of the algorithm. (l) To disconnect the connected component of figure 14j, Min-Cut is applied. So the graph nodes are partitioned into two parts $\{c, d\}$ and $\{b\}$ and the edge $cb$ is removed. The resulting graph becomes disconnected. In the remaining steps, HBUILD is applied and finally a tree structure is obtained by reversing the steps of the algorithm. (m) To disconnect the connected component of figure 14j, Max-Cut is applied. So the graph nodes are partitioned into two parts $\{c, b\}$ and $\{d\}$ and the edge $cd$ is removed. The resulting graph becomes disconnected. In the remaining steps, HBUILD is applied and finally a tree structure is obtained by reversing the steps of the algorithm. (n) The two tree structures that finally are obtained (Three tree structures are obtained. the two structures are the same). (o) The two obtained tree structures of figure 14n are two spanning tree structures of figure 14a and the algorithm finally obtained these two tree structures.

**Binarization**

Let $T$ be a rooted tree and $\tau(T)$ be the set of triplets consistent with $T$. Also let $T_{binary}$ be a binarization of $T$ and $\tau(T_{binary})$ be the set of triplets consistent with $T_{binary}$. Then $\tau(T) \subseteq \tau(T_{binary})$. It means that binarization is an effective tool to make the tree structure more consistent with the given triplets. To perform binarization on $T_{int}$, the following heuristic algorithm is proposed.
For a given set of triplets $\tau$ and $T_{int}$ a binary tree structure $T_{intBin}$ is demanded. Binarization can be performed simply with a random approach [6,7]. In order to make binarization more efficient, a new heuristic algorithm is introduced innovatively in this research. This algorithm is originally based on the three parameters, $w$, $t$, and $p$ [14,20].

Let $\tau$ be a set of triplets and $V_i, V_j \subseteq L(\tau)$ and $V_i \cap V_j = \emptyset$. Let $W(V_i, V_j) = \{v, v_j | v \in \tau \wedge v_i \in V_i, v_j \in V_j \text{ and } v \notin V_i \cup V_j\}$, $P(V_i, V_j) = \{v_j | v \in \tau \wedge v_i \in V_i, v_j \in V_j \text{ and } v \notin V_i \cup V_j\}$, and $T(V_i, V_j) = \{v_j \in \tau | v_i \in V_i, v_j \in V_j\}$. Also let $w(V_i, V_j)$, $p(V_i, V_j)$, and $t(V_i, V_j)$ be the cardinality of $W(V_i, V_j)$, $P(V_i, V_j)$, and $T(V_i, V_j)$, respectively [14,20]. Based on the three parameters ($w$, $t$, and $p$), nine different measure are defined [5]. The measures $M = \{m_1, m_2, \ldots, m_9\}$ are defined as: $m_1 = t(V_i, V_j)$, $m_2 = w(V_i, V_j)$, $m_3 = (w - p)(V_i, V_j)$, $m_4 = \frac{w}{w + p}(V_i, V_j)$, $m_5 = \frac{w}{t}(V_i, V_j)$, $m_6 = \frac{w - p}{w + p}(V_i, V_j)$, $m_7 = \frac{w - p}{w}(V_i, V_j)$, $m_8 = (w - p + \frac{w}{p})(V_i, V_j)$, and $m_9 = (\frac{w - p}{t} + \frac{w}{t + p})(V_i, V_j)$. By using these measures, nine binary tree structures ($T_{intBin}$) are built from $T_{int}$.

The binarization process is performed as follow:

```
Binarization Pseudocode
1: Input: $T_{int}
2: T_{intBin} = T_{int}
3: If $T_{intBin}$ is binary
4: Do nothing
5: else
6: for each vertex $v$ from $T_{intBin}$ with $c_1, c_2, \ldots, c_n$ children, $n > 2$. Initialize a set $C$ with $\{c_1, c_2, \ldots, c_n\}.
7: while $|C| > 1$ do
8: Find and remove two vertices $c_i, c_j \in C$ with maximum measure values.
9: Merge $c_i$ and $c_j$ and obtain $c_{new0}$.
10: Generate 6 new structures using SPR with roots $c_{new1}, c_{new2}, \ldots, c_{new6}$.
11: Among 6+1 structures, select the more consistent structure and add its root to $C$.
12: Update $T_{intBin}$ respect to selected structure.
13: end while
14: end for
16: Output: $T_{intBin}$
```

The binarization process is performed based on using nine different defined measures and Subtree-Pruning-Regrafting (SPR) [21]. SPR is a method in tree topology search [21]. In the binarization process SPR helps to obtain a tree from $T_{int}$ more consistent with input triplets. If $T_{int}$ is binary there is nothing to do; Else there is at least a vertex $v$ in $T_{int}$ with $c_1, c_2, \ldots, c_n$ children, $n > 2$. In this case the goal is replacing this part of the tree with a binary structure (a binary subtree). For this purpose in the first step there are $n$ sets each contains one $c_i, 1 \leq i \leq n$. Then iteratively in each step, two sets with the maximum measure values (according to the one of the nine defined measures) are selected. Let $c_i$ and $c_j$, $1 \leq i, j \leq n$ be two nodes with maximum measure value. By merging $c_i$ and $c_j$, a new vertex $c_{new}$ is created (See figure 15). Here, SPR is used innovatively to improve the merging consistency.

Suppose that $c_{lk}$ and $c_{rk}$ are the roots of the two left and right subtrees of $c_k$, $k \in \{i, j\}$. The idea behind SPR is replacing subtrees to achieve a new binary tree structure with higher consistency. In this work the potential replacement are introduced in six different ways as follow (See figure 16). i) $c_i \leftrightarrow c_{lj}$, ii) $c_l \leftrightarrow c_{rj}$, iii) $c_j \leftrightarrow c_{li}$, iv) $c_j \leftrightarrow c_{ri}$, v) $c_l \leftrightarrow c_{lj}$, vi) $c_l \leftrightarrow c_{rj}$. By using these SPRs, six new structures are
obtained. Among these tree structures and the structure without replacement, the best tree structure consistent with more input triplets is selected.

Fig 15. (a) A non-binary tree. The node $v$ contains four children $c_1, c_2, c_3, c_4$. (b) The nodes $c_i, c_j, c_k, c_l$ are an arbitrary relabeling of the nodes $c_1, c_2, c_3, c_4$. Firstly two nodes are merged. (c) Secondly two structures can be obtained from the structure of figure 15b. (d) Finally for each structure of figure 15c, a binary structure is obtained. These two binary structures are replaced with the non-binary part of the tree of figure 15a.

Fig 16. SPR is used to obtain six different structures from a given tree structure. (a) The structure that is obtained by merging $c_i, c_j$ and connecting them to a new node $c_{new}$. (b to g) Six different tree structures that are obtained from figure 16a and by using SPR with replacing $c_i \leftrightarrow c_{lj}, c_i \leftrightarrow c_{rj}, c_j \leftrightarrow c_{li}, c_j \leftrightarrow c_{ri}, c_{li} \leftrightarrow c_{lj}, c_{li} \leftrightarrow c_{rj}$, respectively.

Network construction

Let $\tau' \subseteq \tau$ be the set of triplets that are not consistent with $T_{intBin}$. Here, the goal is to add some edges to $T_{intBin}$ in order to construct a network consistent with input $\tau'$.

In the network construction process, edges are added incrementally to obtain the final network consistent with $\tau$. In order to add edges, we use innovatively a heuristic criterion to select edges rather than random selection. The heuristic criterion is depended on the current non-consistent triplets $\in \tau'$ and the current network structure.

To this purpose, for each pair of edges of the current network structure, a value is assigned. To compute the value of each pair $\{e, f\}$ of edges, a new edge is added by connecting $e$ and $f$ via two new nodes $n_e, n_f$ (See figure 17). The value is the number of triplets in $\tau'$ that are consistent with the new network structure. In each step of adding an edge, the set of triples $\tau'$, are updated by removing consistent triplets.

The tree construction order $O_{T_{intBin}} = O(mn^2 + n^3\log(n))$ where $|L(\tau)| = n$ and $|\tau| = m$.

Fig 17. (a) $T_{intBin}$ for $\tau' = \{ij|k, ij|l, ij|m, lm|k, lm|j, lm|i, mk|i, mk|j, lk|i, lk|j\}$ and $\tau = \tau' \cup \{lk|m\}$. (b) Two edges $e$ and $f$ are selected to obtain a network consistent with $\tau$. (c) Final network consistent with $\tau$.

Experiments

The RPNCH, NCHB, SIMPLISTIC and TripNet are famous algorithms in constructing phylogenetic networks from given triplets. The SIMPLISTIC algorithm just works for dense triplets [12], while there is no constraints on the NCHB, TripNet, and RPNCH inputs [6, 7, 14]. In order to evaluate the performance of NetCombin, the following scenario is designed.

Data generation

There is two standard approaches to generate triplets data. Firstly, triplets can be generated randomly which is the simplest way. Secondly, triplets can be obtained from real data. Real data usually are in the form of (biological) sequences which is used in this research. There are standard methods for converting sequences into triplets.
Maximum Likelihood (ML) is the well-known method which constructs tree from sequence data \([5, 12]\). For this reason, TREEVOLVE is used which is a software for generating biological sequences \([22]\). TREEVOLVE has different parameters that can be adjusted manually. In this research we set the parameters, the number of samples, the number of sequences, and the length of sequences. For the other parameters, the default values are used. The number of sequences is set to 10, 20, 30, and 40 and the length of sequences is set to 100, 200, 300, and 400. For each case, the number of samples is 10. So totally 160 different sets of sequences are generated. Then PhyML software is used which works based on Maximum Likelihood (ML) criterion. For each set of sequences, all subsets of three sequences are considered and for each of them, an outgroup is assigned. Each subset of three sequences plus the assigned outgroup, are considered as input for PhyML and for these data the output of PhyML is a quartet. Finally by removing the outgroup from each quartet, the set of triplets is obtained. In this research, each triplet information related to a quartet in which the weight of its unique inner edges is zero, is removed. This is because of these types of triplets contains no information and are stars. The way of generating triplets may give non-dense sets of triplets. SIMPLISTIC is used as a method for comparison and its output should be dense. So by adding a random triplet correspond to each star, each non-dense set is converted to a dense set and is used as the input.

**Experimental Results**

In order to show the performance of NetCombin we compare it with TripNet, SIMPLISTIC, NCHB, and RPNCH on the data that are generated in the previous subsection. Since for large size data, SIMPLISTIC has not the ability to return a network in an appropriate time, the time restriction 6 hours is considered. Let \(N_{finite}\) be the set of networks for which the running time of the method is at most 6 hours. Let \(S_{sequence}\) shows the number of sequences where \(S_{sequence} \in \{10, 20, 30, 40\}\). The output of TripNet, SIMPLISTIC, NCHB, and RPNCH is a unique network, but NetCombin outputs 27 networks and the best network is reported. Since the process of constructing these 27 NetCombin networks is independent, we apply NetCombin in a parallel way to obtain 27 networks simultaneously.

The results of comparing these methods on the two optimality criterions and running time are available in tables 1 to 4.

Table 1 and 2 show the results of the number of networks that belong to \(N_{finite}\), and the average of running time of the networks that belong to \(N_{finite}\). These results show that when the number of taxa is 10, in all cases, all methods on average give an output in at most 2 seconds. When the number of taxa is 20, in 5% of the cases, SIMPLISTIC has not the ability to return a network in less than 6 hours. For the remaining 95% of the cases SIMPLISTIC on average gives an output in 310 seconds. For these data the other four methods on average construct a network in less than 4 seconds. When the number of taxa is 30, in 32.5% of the cases, SIMPLISTIC outputs a network in 2600 seconds. For the remaining 67.5% of the cases, SIMPLISTIC has not the ability to return a network in less than 6 hours. For these data, on average NetCombin and RPNCH output a network in at most 15 seconds, while NCHB and TripNet on average output a network in 203 and 210 seconds, respectively. When the number of input taxa is 40, in all cases SIMPLISTIC does not return an output in time restriction 6 hours. In this case NetCombin and RPNCH on average output a network in at most 44 seconds while NCHB and TripNet return a network in time at least 740 seconds.

Tables 3 and 4 indicate the results for the two optimality criterions i.e. the number of reticulation nodes and level for the networks that belong to \(N_{finite}\). The results show that when the number of taxa is 10, on average the number of reticulation nodes for the TripNet and NCHB networks is at most 0.9, while for these data on average the
Table 1. The number of NetCombin, TripNet, NCHB, SIMPLISTIC, and RPNCH networks that belong to $N_{finite}$

| Number of sequences ($s_{sequence}$) | 10  | 20  | 30  | 40  |
|--------------------------------------|-----|-----|-----|-----|
| Number of the NetCombin networks $\in N_{finite}$ | 40  | 40  | 40  | 40  |
| Number of the TripNet networks $\in N_{finite}$ | 40  | 40  | 40  | 40  |
| Number of the NCHB networks $\in N_{finite}$ | 40  | 40  | 40  | 40  |
| Number of the SIMPLISTIC networks $\in N_{finite}$ | 40  | 38  | 13  | 0  |
| Number of the RPNCH networks $\in N_{finite}$ | 40  | 40  | 40  | 40  |

Table 2. The average running time results for the networks that belong to $N_{finite}$ for NetCombin, TripNet, NCHB, SIMPLISTIC, and RPNCH.

| Number of sequences ($s_{sequence}$) | 10  | 20  | 30  | 40  |
|--------------------------------------|-----|-----|-----|-----|
| NetCombin avg running time for networks $\in N_{finite}$ (Sec) | 2   | 4   | 15  | 44  |
| TripNet avg running time for networks $\in N_{finite}$ (Sec) | 1   | 1.7 | 210 | 740 |
| NCHB avg running time for networks $\in N_{finite}$ (Sec) | 1   | 1.8 | 203 | 745 |
| SIMPLISTIC avg running time for networks $\in N_{finite}$ (Sec) | 1   | 310 | 2600 | -  |
| RPNCH avg running time for networks $\in N_{finite}$ (Sec) | 1   | 2   | 10  | 30  |

NetCombin, RPNCH, and SIMPLISTIC number of reticulation nodes is at least 2 and at most 3. Also for these data, on average the level of the NCHB and TripNet networks, is not more than 0.9, while the level of NetCombin, SIMPLISTIC, and RPNCH networks, on average is at least 2 and at most 2.8. When the number of input taxa is 20 on average the TripNet and NCHB number of reticulation nodes is 2.6 and 1.8, respectively. For these data the NetCombin number of reticulation nodes on average is 4, while for SIMPLISTIC and RPNCH, on average the number of reticulation nodes is 6.95 and 9, respectively. Also for these data the level of the NCHB and TripNet networks on average are 1.5 and 2.3, respectively. For these data on average the level of the NetCombin networks is 3, while for the SIMPLISTIC and RPNCH networks the level is 4.2 and 6.4, respectively. When the number of taxa is 30, on average the number of NCHB, TripNet, and NetCombin reticulation nodes, are at least 7.2 and at most 9, while for the SIMPLISTIC and RPNCH networks on average this number is 11.275 and 13, respectively. For these data on average the level of the NCHB, TripNet, NetCombin, and SIMPLISTIC networks is at least 6 and at most 7.4 while on average the RPNCH networks level is 10.5. When the number of taxa is 40, on average the, NCHB, NetCombin, and TripNet number of reticulation nodes are 15.2, 15.5, and 16.3, respectively, while RPNCH networks on average contain 20 reticulation nodes. For these data on average the level of NetCombin, NCHB, and TripNet networks are 15, 15, and 16, respectively, while the level of the RPNCH networks on average is 19.

**Discussion**

In this paper we introduced NetCombin which is a method for producing an optimal network consistent with a given set of triplets. In order to show the performance of NetCombin we compared it with NCHB, TripNet, SIMPLISTIC, and RPNCH on the 160 different sets of triplets that are generated in the process that is introduced in subsection 4-1.

The results show that although, on average RPNCH is the fastest method, but the level and the number of reticulation nodes of its results are highest. More over on...
average the differences between NetCombin, NCHB, and TripNet results for the two optimality criterions with RPNCH results are significant.

The results show that on average for small size data SIMPLISTIC is appropriate. But by increasing the number of taxa and for large size data it has not the ability to return a network in an appropriate time and its running time is highest. Also in all cases on average the SIMPLISTIC number of reticulation nodes and levels are just better than RPNCH. Note that SIMPLISTIC just works for dense sets of input triplets. The results show that by increasing the number of taxa, the running time of SIMPLISTIC increases exponentially. In more details when the number of taxa is 40, in time less than 6 hours it does not return any network, while the other 4 methods in at most 745 seconds output a network.

Also the results show that on average NCHB and TripNet running time results are nearly the same, but on average the two optimality criterions for NCHB results are better compared to TripNet. Note that the differences between TripNet and NCHB results for the optimality criterions are not significant.

The results show that for small size data TripNet and NCHB are appropriate and their results for the optimality criterions and running time are on average the best. But by increasing the number of taxa, the running time of these methods exceeds significantly compared to NetCombin, while the two optimality criterions for their networks are nearly the same with NetCombin networks results.

The results show that generally and by considering the running time, the level, and the number of reticulation nodes of the final networks, on average NetCombin is a valuable method that returns an appropriate network in an appropriate time.

Acknowledgement

The second author would like to thank the Institute for Research in Fundamental Sciences (IPM), Tehran, Iran. "The authors declare no conflict of interest"
References

1. Felsenstein Joseph. Inferring phylogenies. Sinauer associates Sunderland, MA.; 2004.

2. Huson Daniel H, Rupp Regula, Scornavacca Celine. Phylogenetic networks: concepts, algorithms and applications. Cambridge University Press.; 2010.

3. Poormohammadi Hadi. A New Heuristic Algorithm for MRTC Problem. Journal of Emerging Trends in Computing and Information Sciences. 2012 ;3(7).

4. Jahangiri Soheil, Hashemi Seyed Naser, Poormohammadi Hadi. New heuristics for rooted triplet consistency. Algorithms. 2013 ;6(3).

5. Poormohammadi Hadi, Sardari Zarchi Mohsen. CBTH: a new algorithm for MRTC problem. Iranian Journal of Biotechnology (IJB). 2019, Accepted.

6. Poormohammadi Hadi, Sardari Zarchi Mohsen. TripNet: a method for constructing rooted phylogenetic networks from rooted triplets. PloS one. 2014 ;6(6):e106531.

7. Reyhani Mohammad Hossein, Poormohammadi Hadi. RPNCH: A method for constructing rooted phylogenetic networks from rooted triplets based on height function. Journal of Paramedical Sciences. 2017 ;8(4):14-20.

8. Linder C Randal, Moret Bernard ME, Nakhleh Luay, Warnow Tandy. Network (reticulate) evolution: biology, models, and algorithms. The Ninth Pacific Symposium on Biocomputing (PSB). 2004.

9. Bordewich Magnus, Semple Charles. Computing the minimum number of hybridization events for a consistent evolutionary history. Discrete Applied Mathematics. 2007;155(8):914-928.

10. Gambette Philippe, Huber Katharina T, Kelk Steven. On the challenge of reconstructing level-1 phylogenetic networks from triplets and clusters. Journal of mathematical biology. 2017;74(7):1729-1751.

11. Tusserkani Ruzbeh, Poormohammadi Hadi, Azadi Azin, Eslahchi Changiz. Inferring phylogenies from minimal information. Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB). 2017; 202-206.

12. Van Iersel Leo, Kelk Steven. Constructing the simplest possible phylogenetic network from triplets. Algorithmica. 2011:60(2):207-235.

13. Kannan Sampath K, Lawler Eugene L, Warnow Tandy J. Determining the evolutionary tree using experiments. Journal of Algorithms. 1996:21(1):26-50.

14. Poormohammadi Hadi, Sardari Zarchi Mohsen, Ghanei Hossein. NCHB: A Method for Constructing Rooted Phylogenetic Networks from Rooted Triplets based on Height Function and Binarization. Journal of Theoretical Biology. Submitted.

15. Jansson Jesper, Sung Wing-Kin. Algorithms for combining rooted triplets into a galled phylogenetic network. Encyclopedia of Algorithms. 2016; 48-52.

16. Jansson Jesper, Sung Wing-Kin. Inferring a level-1 phylogenetic network from a dense set of rooted triplets. Theoretical Computer Science. 2006;363(1):60-68.
17. Aho Alfred V, Sagiv Yehoshua, Szymanski Thomas G, Ullman Jeffrey D. Inferring a tree from lowest common ancestors with an application to the optimization of relational expressions. SIAM Journal on Computing. 1981;10(3):405-421.

18. Huber Katharina T, Van Iersel Leo, Kelk Steven, Suchecki Radoslaw. A practical algorithm for reconstructing level-1 phylogenetic networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2011;8(3):635-649.

19. Karp Richard M. Reducibility among combinatorial problems. Complexity of computer computations. 1972;85-103.

20. Wu Bang Ye. Constructing the maximum consensus tree from rooted triples. Journal of Combinatorial Optimization. 2004;8(1):29-39.

21. Wu Yufeng. A practical method for exact computation of subtree prune and regraft distance. Bioinformatics. 2008;25(2):190-196.

22. Grassly N, Rambaut A. Treevole: a program to simulate the evolution of DNA sequences under different population dynamic scenarios. 1.3. Wellcome Centre for Infectious Disease, Department of Zoology. 1997.
Figure 1
Figure 4
Figure 5
Figure 6
Figure 9
$C_{ad} = C_{ae} = C_{bd} = C_{be} = C_{cd} = C_{ce}$

Figure 11
Figure 12

inner edge
Figure 16
Figure 17