Overlaid species forests

K.T. Huber[1], V. Moulton[1], G. E. Scholz[2]

Abstract

Introgression is an evolutionary process in which genes or other types of genetic material are introduced into a genome. It is an important evolutionary process that can, for example, play a fundamental role in speciation. Recently the concept of an overlaid species forest (OSF) was introduced as a discrete way to model introgression. Basically, an OSF consists of a gene history in the form of a phylogenetic tree, a collection of lineage trees or forest for some species of interest, and a map that overlays the gene tree onto the forest. In this paper we shall study mathematical properties of OSFs and their relationship with other structures in phylogenetics, such as lateral gene transfer models, subtree prune and regraft operations, and phylogenetic networks. In particular, we show that a certain algorithm called OSF-BUILDER for constructing an OSF is guaranteed to produce a special type of OSF with a minimum number of introgressions, as well as providing some characterizations for networks that can arise from OSFs. We also give bounds on how much an OSF can change when the underlying gene tree or forest is perturbed. We expect that these results will be useful in developing new algorithms for deriving introgression histories, a rapidly growing area of interest in phylogenomics.

Keywords: phylogenetic network, introgression model, overlaid species forest (OSF), unfolding
2008 MSC: 05C90, 92D15

1. Introduction

Introgression is an evolutionary process in which foreign genetic material, such as genes, are introduced into a genome [16, Glossary, p. 230]. It is an important process since it can, for example, help species adapt to or expand into new environments [9]. Introgression is a wide-spread phenomenon in plants and animals, and can occur as the result of sexual contact or hybridization [16]. Some striking examples of introgression include genes introduced by Neanderthals into modern humans [18] and genetically modified crop genes moving into their wild relatives [23]. Other examples are given in e.g. [16, 28]. Although several approaches have

*Corresponding author: K.T. Huber

Email addresses: k.huber@uea.ac.uk (K.T. Huber), v.moulton@uea.ac.uk (V. Moulton), gllm.scholz@gmail.com (G. E. Scholz)

[1] School of Computing Sciences, University of East Anglia, UK

[2] Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, Leipzig University, Leipzig, Germany.

been introduced to detect introgression (e.g. the ABBA/BABA test [22]), to date relatively few approaches have been proposed for reconstructing explicit evolutionary scenarios which involve introgression (see [21, 26] for some examples).

One approach to derive introgression histories, illustrated for example by a study of how gene introgression leads to variation in butterfly wing patterns [25, Figure 5], is to trace how genes move between different lineages of a species over time. This approach was recently formalised in [19] in which a model is presented that involves overlaying a species forest F with a gene tree G. In this method, F represents a collection of lineage trees for certain species (e.g. butterflies), and G the evolutionary history of some gene which jumps or introgresses between the lineages (e.g. genes which affect butterfly wing colouring). The lineage trees can be derived by, for example, considering a species tree and taking subtrees labelled by species that share a common trait (see e.g. [1, 17]). The gene tree and species forest are linked via a natural map ϕ from the leaf-set of G to the leaf-set of F given by taking a gene to the species in which it resides. We call (G,F,ϕ) a *forest triple* (see e.g. Figure 1). Based on the triple (G,F,ϕ), we look for ways to overlay G onto the forest F, i.e. map the vertex set of G to the vertex set of F whilst respecting ϕ and other conditions of ancestry, so as to represent the introgression history of the gene in question. For example, in Figure 1(c), we present one possible mapping ψ of the vertex set of the depicted gene tree G to the vertex set of the species forest F also depicted in the figure. We call such a mapping an *overlaid species forests* or OSF for short.

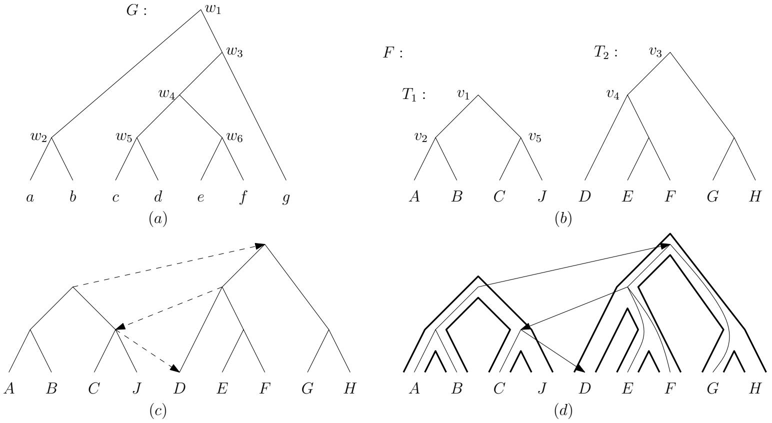

Figure 1: A forest triple where G is the tree in (a), F is the forest $\{T_1,T_2\}$ in (b) and ϕ is the map that takes every leaf of G in lower case labels to the corresponding leaf of F in capitals (e.g. $\phi(b)=B$). (c) A representation of an OSF ψ for (G,F,ϕ) where $\psi(w_i)=v_i$ for $i=1,2,3,5$ and $\psi(w_6)=\psi(w_4)=v_4$. The dashed arcs are contact arcs. This representation is visualised in more detail in (d).

OSFs provide an interesting new discrete model of introgression, and in this paper we shall study some of their basic properties and relationships with other structures in phylogenetics. We now summarize the rest of the paper. We begin by introducing some notation in Section 2, and reviewing the OSF-Builder algorithm. This algorithm is introduced in [19] and, for a given forest triple, aims to produce an OSF that minimizes the number of times that a gene introgresses between lineages. The rationale behind this is that introgression is thought to be relatively infrequent in nature [2, page 82:1]. In Section 3, we show that the OSF-BUILDER algorithm is guaranteed to produce an optimal OSF, and that this OSF is of a special form, called a *strict OSF* (Theorem 3.3). This type of OSF is related to a well-studied tree reconciliation model for representing lateral gene transfer for trees introduced in [24]. The main difference between this model and ours is that, to account for the fact that introgression occurs between lineages and that a lineage may be thought of as a rooted subtree (or clade) of a species tree, we have replaced a tree by a forest. Our approach to proving that we obtain optimal OSFs uses the close relationship between the OSF-BUILDER algorithm and the Fitch-Hartigan algorithm for computing most parsimonious trees [7, 10].

In Section 4, we then look more deeply into structural properties of OSFs. Since each OSF can be visualized in terms of a network, as depicted in e.g. Figure 1(c), it is thus of interest to find characterizations for those networks which arise from OSFs. We call such networks *valid* networks. Note that since an OSF can be viewed as a forest with some additional contact arcs (see Figure 1(c)), characterizing valid networks is related to the problem of characterizing *tree-based networks*, phylogenetic networks that arise by adding arcs to a phylogenetic tree [8] (see Section 7 for more details). We show that a network is a valid network if and only if it can be obtained by applying the OSF-BUILDER algorithm to some forest triple (Theorem 4.2). Then, in Section 5, we present a structural characterization of valid networks (Theorem 5.2). To prove this result, we use a variation of the process of "unfolding" a phylogenetic network [13].

In Section 6, we consider the effect a change in the gene tree and species forest has on the number of arcs contained in an optimal network corresponding to a forest triple. This is of interest because it helps to shed light on how noise in a forest triple can affect the OSFs generated by the OSF-BUILDER algorithm. In particular, we provide upper bounds on the amount that the optimal score for an OSF can change by in terms of the number of *subtree prune and regraft (SPR) operations* that alter a gene tree or forest (Theorems 6.1 and 6.2). We do this by exploiting some results in [6] which relate the so-called maximum parsimony and SPR-distances between two phylogenetic trees. In the last section, we conclude by briefly discussing some possible future directions.

2. Preliminaries

In what follows, we assume that X is a finite set with $|X| \geq 2$ and that all graphs are directed unless otherwise stated, and without loops.

2.1. *Networks*

Let G be a directed graph. We denote the vertex set of G by $V(G)$ and its set of arcs by $A(G)$. We denote an arc a from a vertex $u \in V(G)$ to a vertex $v \in V(G)$ by $a = (u, v)$ and refer to u as $tail(a)$ and to v as $head(a)$, respectively. A *walk* in G is a sequence of vertices $v_1, \dots, v_m$ in $V(G)$, $m \geq 1$, such that $(v_i, v_{i+1}) \in A(G)$, $1 \leq i \leq m-1$. A *trail* in G is a walk with no repeated arcs, and a *path* is a trail with no repeated vertices. A *cycle* in G is a trail of the form $v_1, \dots, v_m, v_1$, $m \geq 2$ where $v_1, \dots, v_m$ is a path. We say that G is *connected* if its underlying (undirected) graph is connected (note that the underlying graph might contain multi-edges).

We denote by $outdeg_G(v) = outdeg(v)$ the number of outgoing arcs of a vertex $v \in V(G)$ and by $indeg_G(v) = indeg(v)$ its number of incoming arcs. We call a vertex $v \in V(G)$ with $indeg(v) = 1$ and $outdeg(v) = 0$ a *leaf* of G and denote by $L(G)$ the set of leaves of G. We call a vertex $v \in V(G)$ with $indeg(v) = 0$ and $outdeg(v) \geq 2$ a *root* of G.

A *network (on X)* is a directed graph G which satisfies:

(N1) $X \subseteq V(G)$,

(N2) if $v \in V(G)$ such that $indeg(v) = 1 = outdeg(v)$ then $v \in X$,

(N3) if $v \in V(G)$ such that $indeg(v) = 0$ then $v \notin X$, and

(N4) if $v \in V(G)$ such that $outdeg(v) = 0$ then $v \in X$.

Note that our definition of a network differs from the standard definition of a (phylogenetic) network in that we do not require it to be connected, and we allow it to have multiple roots as well as interior vertices that are contained in X. Two networks N and N' on X are *isomorphic* if there exists a bijective map $\psi : V(N) \to V(N')$ that induces a graph isomorphism between N and N' that is the identity on X.

Suppose G is a network and $u, v \in V(G)$. Then we put $u \preceq_G v$ (or just $u \preceq v$) if there is a directed path in G starting at u and ending at v. If $u \preceq_G v$ then we call u an *ancestor* of v (in G), and say that v lies *below* u (in G). In that case, we also call v the *descendant* of u. Note that a vertex can be its own ancestor. We call u a *strict ancestor* of v if u is an ancestor of v and $v \neq u$. In that case, we also call v a *strict descendant* of u. The vertex u is a *child* of v if $(v, u) \in A(G)$.

A *(rooted) phylogenetic tree T (on X)* is a network T whose underlying graph is a tree, that has a single vertex with indegree zero denoted ρ_T, and leaf set X (and so $L(T)$ has size at least 2, by our assumption on X). We denote the set of leaves of T below u by $\mathscr{C}_T(u)$. For any non-empty subset $Y \subseteq X$, we denote by $lca_T(Y)$ the unique vertex $v \in V(T)$ that is an ancestor of every element in Y such that no vertex below v and distinct from v is an ancestor of every element of Y. If $Y = \{y_1, \ldots, y_k\}$, $k \geq 1$, then we sometimes write $lca_T(y_1, \ldots, y_k)$ rather than $lca_T(\{y_1, \ldots, y_k\})$. Note that $lca_T(x) = x$, for all $x \in X$.

2.2. *Overlaid species forests*

A *forest* F is a directed graph whose set of components is non-empty and every component is a phylogenetic tree. In case F has a single component, we view F as a phylogenetic tree. Note that the leaf set of a forest F is $\bigcup_{T \in F} L(T)$. A vertex in a forest F that is not a leaf is called an *interior vertex* of F. We let $V^0(F)$ denote the set of interior vertices of F. We say that a forest F is *binary* if $outdeg(v) = 2$ holds for all interior vertices v of F.

A triple $\mathscr{F} = (G, F, \phi)$ consisting of a phylogenetic tree G, a forest F, and a (leaf) map $\phi = \phi_{G,F} : L(G) \to L(F)$ is called a *forest triple*. We say that $\mathscr{F}$ is *binary* if both G and F are binary. For ease of readability of our examples, we usually denote the leaves of G by lower-case letters (also with indices), and the leaves in F which they map to under ϕ by the corresponding capital letter.

Given a forest triple $\mathscr{F} = (G, F, \phi)$, an *overlaid species forest or OSF (for $\mathscr{F}$)*, is a map $\psi : V(G) \to V(F)$ which satisfies:

(P1) $\psi|_{L(G)} = \phi$.

(P2) If $u, v \in V(G)$ satisfy $u \preceq_G v$ and $\psi(u), \psi(v) \in V(T)$ holds for some $T \in F$ then $\psi(u) \preceq_T \psi(v)$.

(P3) For all $u \in V^0(G)$, there exists some leaf $v \in \mathscr{C}_G(u)$ such that $\phi(v)$ and $\psi(u)$ belong to the same tree in F.

To illustrate the definition of an OSF, note that, for example, the map $\psi : V(G) \to V(F)$ given in the caption of Figure 1 is an OSF for the binary forest triple (G, F, ϕ) given in that figure.

As mentioned in the introduction, the definition of an OSF is related to tree reconciliation models for lateral gene transfer and, more generally, to cophylogeny models (see e.g. [3]). One of the first models of this type for lateral gene transfer was introduced in [24]. This model is based on the concept of a *DTL-scenario* [24, p.519], which is essentially a map ψ from the vertex set $V(G)$ of a given gene tree G to the vertex set $V(T)$ of a given species tree T which satisfies the following properties (see also [12]):

(1) $\psi|_{L(G)} = \phi$ where $\phi : L(G) \to L(T)$ denotes the map which takes a gene to the species that it is contained in.

(2) If $u \in V^0(G)$ then

  (a) there is no child v of u in G such that $\psi(v)$ is a strict ancestor of $\psi(u)$, and

  (b) there is a child v of u such that $\psi(v)$ is a descendant of $\psi(u)$.

As can be seen, these properties are similar to (P1)–(P3) for an OSF; in fact note that Property (1) is the same as (P1) (expect we are replacing a tree T with a forest F). As well as modelling transfer events, DTL-scenarios model so-called duplication and speciation events. Note that costs for each of these three possible events are often also incorporated into these models, and algorithms have been developed to find DTL-scenarios with minimal cost (see e.g. [15]). To fully explain these concepts is beyond the scope of this paper, but as an illustration, in Figure 2 we present a simple example to elucidate the difference between OSFs and DTL-scenarios. In particular, in this example we see that the DTL-scenario for G and the species tree T obtained from the trees T_1 and T_2 as indicated does not give rise to an OSF. This illustrates why we cannot directly apply DTL-scenarios to model introgression in case we are interested in modelling introgression between (and not within) lineages.

2.3. *The* OSF-BUILDER *algorithm*

Suppose that ψ is an OSF for a forest triple $\mathscr{F} = (G, F, \phi)$. A *contact arc of* ψ is a pair $(\psi(u), \psi(v))$ where $(u, v) \in A(G)$ and $\psi(u) \in V(T)$ and $\psi(v) \in V(T')$ for $T, T' \in F$ distinct. We let $C(\psi)$ denote the multi-set of contact arcs of ψ, with multi-set cardinality $|C(\psi)|$, and we set

$$t(\mathscr{F}) = min\{|C(\psi)| \,:\, \psi \text{ an OSF for } \mathscr{F}\}.$$

In addition, we let $C^*(\psi)$ denote the underlying set of $C(\psi)$. For example, for the OSF ψ in Figure 3, we have

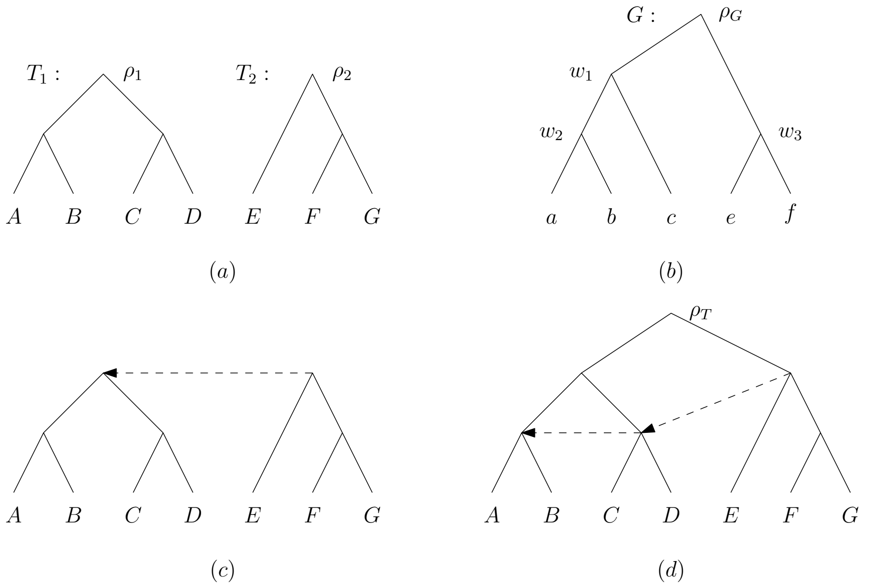

Figure 2: (a) A species forest $F = \{T_1, T_2\}$. (b) A gene tree G. (c) An OSF $\psi : V(G) \to V(F)$ for (F, G, ϕ) defined by $\psi(\rho_G) = \psi(w_3) = \rho_2$, $\psi(w_1) = \rho_1$, and $\psi(w_2) = lca_{T_1}(A, B)$. (d) For G and the tree T obtained by combining T_1 and T_2 as indicated, a representation of a DTL-scenario λ given by $\lambda(\rho_G) = \lambda(w_3) = \rho_2$, $\lambda(w_1) = lca_T(C, D)$, and $\lambda(w_2) = lca_T(A, B)$. Note that one of the arcs indicating transfer (dashed) stays within the subtree T_1 and so λ does not correspond to an OSF for G and F.

$$C(\psi) = \{(\psi(\rho), \psi(w_2)), (\psi(w_2), \psi(w_4)), (\psi(w_4), \psi(w_6))\} = \{(\rho_1, \rho_2), (\rho_2, \rho_1), (\rho_1, \rho_2)\}.$$

So $|C(\psi)| = 3$, but $|C^*(\psi)| = 2$. In [19] an algorithm called OSF-BUILDER is introduced for computing an OSF ψ for $\mathscr{F}$ with $|C(\psi)| = t(\mathscr{F})$ which we now briefly review. It is based on the Fitch-Hartigan algorithm for computing the parsimony score of a character on a phylogenetic tree [10] (see also [7]).

Suppose $\mathscr{F} = (G, F, \phi)$ is a forest triple. Consider the map $f = f_{\mathscr{F}} : L(G) \to F$, which takes each $v \in L(G)$ to the tree $T \in F$ with $\phi(v) \in V(T)$, as being a character[3] on $L(G)$. An *extension $\bar{f}$ of f to $V(G)$* is a map $\bar{f} : V(G) \to F$ such that $\bar{f}(v) = f(v)$ for all $v \in L(G)$. The *parsimony score $l_f(G)$ of f on G* is then given by taking the minimum, over all extensions $\bar{f}$ of f, of the number of arcs $(u, v) \in A(G)$ with $\bar{f}(u) \neq \bar{f}(v)$ (cf. [20, p.84]).

The OSF-BUILDER algorithm works by first associating to every vertex $v \in V^0(G)$ the set $\sigma(v)$ of trees in F which are assigned most frequently to its children in a bottom-up fashion, starting at the leaves. Then it computes an OSF ψ in a top-down phase as follows. It begins by computing an extension $\bar{f}$ of f. To do this it initializes first $\bar{f}(\rho_G)$ to be any tree in $\sigma(\rho_G)$. Then, moving down G, for any vertex v in $V^0(G)$ with $\bar{f}$ defined on v but not yet on its children, for u a child of v it sets $\bar{f}(u) = \bar{f}(v)$ if the tree $\bar{f}(v)$ is contained in $\sigma(u)$, and otherwise it sets $\bar{f}(u)$ to

[3] For an arbitrary set Y, a *character on Y* is a map from Y to some finite set.

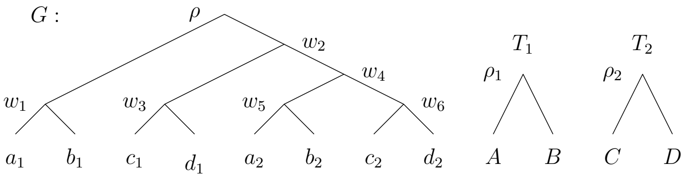

Figure 3: A gene tree G and a forest $F = \{T_1, T_2\}$, for which the map $\psi : V(G) \to V(F)$ defined as $\psi(\rho) = \psi(w_i) = \rho_1$ for all $i = 1,4,5$, and $\psi(w_i) = \rho_2$, for all $i = 2,3,6$ is an OSF.

be any tree in $\sigma(u)$. The OSF ψ is then defined by setting

$$\psi(v) = lca_{\bar{f}(v)}(\{w \in L(\bar{f}(v)) : \text{ there exists some leaf } x \in L(G) \text{ below } v \text{ such that } w = \phi(x)\}),$$

for all $v \in V(G)$. In other words, $\psi(v)$ is the last common ancestor of all leaves in the tree $\bar{f}(v)$ that are the image of some leaf in $\mathscr{C}_G(v)$ under ϕ. It follows that OSF-BUILDER computes an OSF ψ on $\mathscr{F}$ such that $|C(\psi)| = l_{f_{\mathscr{F}}}(G)$.

3. Strict OSFs

In this section, we show that the OSF-BUILDER algorithm described in the last section produces a special type of OSF ψ for $\mathscr{F}$ with $|C(\psi)| = t(\mathscr{F})$ which is defined as follows (see also [19, Supp. Mat., Theorem 2] for a related, weaker result). Suppose that $\mathscr{F} = (G, F, \phi)$ is a forest triple. A *strict overlaid species forest or sOSF (for $\mathscr{F}$)* is a map $\psi : V(G) \to V(F)$ which satisfies (P1), (P2) and

(P3') For all $u \in V(G)$, there exists a child v of u in G such that $\psi(v)$ and $\psi(u)$ belong to the same tree in F.

Note that every sOSF for $\mathscr{F}$ is also an OSF for $\mathscr{F}$ (see Lemma 3.1) but not conversely (see e.g. Figure 4). Also note the similarity between Property (P3') and Property (2b) of a DTL-scenario as introduced in Section 2.3.

Biologically speaking, the OSF requirement that a descendant of a gene in lineage T must be found in at least one of the extant species in T (Property (P3)) is is replaced in a sOSF with the requirement that this must hold for at least one of its children. As a consequence, in an sOSF for every non-leaf vertex u in G there does not only need to exist a leaf l_u in G such that u and l_u are mapped to the same lineage in F, but also every vertex on the path from u to l_u must be mapped to that lineage. This is not only convenient for making computations, but it is clearly also a much stronger assumption on how genes introgress which may not always apply.

We now present some basic properties of strict OSFs.

Lemma 3.1. *Suppose $\mathscr{F} = (G, F, \phi)$ is forest triple.*

(i) If ψ is an sOSF for $\mathscr{F}$, then it is an OSF for $\mathscr{F}$.

(ii) If ψ is an OSF for $\mathscr{F}$ output by OSF-BUILDER, *then it is an sOSF.*

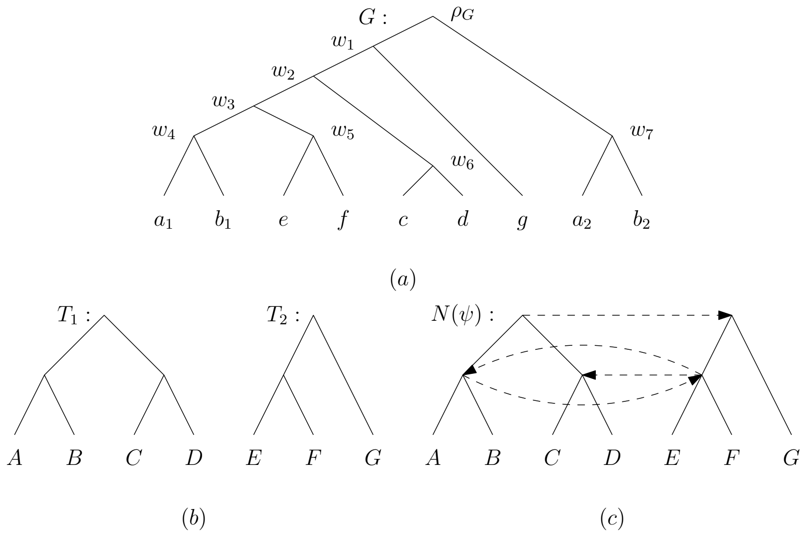
$G:$
ρ_G
w_1
w_2
w_3
w_4
w_5
w_6
w_7
a_1 b_1 e f c d g a_2 b_2
(a)
$T_1:$
$T_2:$
$N(\psi):$
A B C D E F G
A B C D E F G
(b)
(c)

Figure 4: For G the tree in (a), $F = \{T_1, T_2\}$ the forest in (b), and ϕ the corresponding leaf-map, we present in (c) the representation $N(\psi)$ of the OSF $\psi : V(G) \to V(F)$ for the forest triple (G, F, ϕ) where ψ maps the root of G to the root of T_1, w_1 to the root of T_2, w_3 and w_4 and w_7 to the parent of A and B, w_6 to the parent of C and D, and w_2 and w_5 to the parent of E and F. The dashed arcs are the contact arcs of $N(\psi)$. Since ψ does not satisfy Property (P3') for w_2, it is not an sOSF for $\mathscr{F}$.

Proof: (i) We need to show that (P3) holds for ψ, that is, if $u \in V(G)$ is such that $\psi(u) \in V(T)$ for some $T \in F$, then there exists some leaf $v \in \mathscr{C}_G(u)$ such that $\psi(v)$ also belongs to T. But this clearly holds since if $u \in V(G)$, by (P3'), we can take a child v of u in G that is mapped by ψ to a vertex below $\psi(u)$, and then repeat this process of applying (P3') for v and its children, until we reach a leaf of G.

(ii) In the top-down phase of OSF-Builder, for a vertex u, at least one child v of u will be mapped to the same tree as u. Hence (P3') holds for any OSF returned by OSF-Builder ∎

Before proving the main result of this section (Theorem 3.3), we give a useful characterization of the subsets of $A(G)$ of G that correspond to the multiset of contact arcs of some sOSF. For a set $I \subseteq A(G)$, we let $G - I$ denote the graph obtained by removing all arcs in I from $A(G)$. Clearly, $G - I$ is a set of (not necessarily phylogenetic) trees. Suppose we are given a forest triple $\mathscr{F} = (G, F, \phi)$. We call a set $I \subseteq A(G)$ an *introgression set for* $\mathscr{F}$ if the following hold: (i) if u is a tail of some arc in I, then u is also a tail of some arc in $A(G)$ that is not in I, (ii) if M is a tree in $G - I$, then ϕ maps every leaf of $L(G)$ contained in $V(M)$ to the same tree T_M in F, and (iii) if $M \neq M' \in G - I$ and $u \in V(M)$, $v \in V(M')$ with (u, v) or (v, u) in I, then $T_M \neq T_{M'}$. For example, the set $\{(w_1, g), (w_3, w_5)\}$ is an introgression set for the forest triple depicted in Figure 4. Note that analogous sets for gene/species tree reconciliations (called transfer sets) are defined in [24].

It is straight-forward to check that for any sOSF ψ for $\mathscr{F}$, the set of arcs in $A(G)$ which map under ψ to $C(\psi)$ is an introgression set. The following proposition which is analogous to [24, Lemmas 4 and 5] for the DTL-scenario shows that the converse holds. It implies that introgression sets induce a partition of the set of strict OSFs for $\mathscr{F}$, where two strict OSFs are in the same part if and only if they give rise to the same introgression set for $\mathscr{F}$.

Proposition 3.2. *Suppose $\mathscr{F} = (G, F, \phi)$ is a forest triple and $I \subseteq A(G)$. Then I is an introgression set for $\mathscr{F}$ if and only if there exists an sOSF ψ for $\mathscr{F}$ such that $C(\psi)$ coincides with the multiset $\{(\psi(u), \psi(v)) : (u, v) \in I\}$.*

Proof: The if statement follows from the remark preceding the proposition. To see the converse, suppose that I is an introgression set for $\mathscr{F}$, and let $\psi_I : V(G) \to V(F)$ be the map given by setting, for all $u \in V(G)$,

$$\psi_I(u) = lca_{T_M}(\{\phi(g) \in L(T_M) : g \in \mathscr{C}_G(u) \cap V(M)\}),$$

where M is the tree in $G - I$ containing u. Then ψ_I is an sOSF. ∎

To illustrate Proposition 3.2, consider the introgression set $I = \{(w_1, a), (w_2, c)\}$ mentioned above for the forest triple (G, F, ϕ) considered in Fig. 4. Then there must exist an sOSF ψ such that the multiset $C(\psi)$ is induced by I. As is easy to check, ψ is the map given by assigning $\psi(u)$ to the root of T_2 in case $u \in \{w_1, w_2, \rho_G\}$ and $\psi(u) = \phi(u)$, for all leaves u of G.

As mentioned above, for a forest triple $\mathscr{F}$ there exist OSFs for $\mathscr{F}$ that are not sOSFs. Moreover, there exist sOSFs ψ for $\mathscr{F}$ with $|C(\psi)| = t(\mathscr{F})$ which OSF-BUILDER is not able to construct ([19, Fig. 1, Supp. Mat.]). Even so, we now show that, for any forest triple, OSF-BUILDER is guaranteed to produce an optimal OSF that is also an sOSF.

Theorem 3.3. *Suppose $\mathscr{F} = (G, F, \phi)$ is a forest triple. Then*

$$t(\mathscr{F}) = min\{|C(\psi)| : \psi \textit{ is a } \text{strict } \textit{OSF for } \mathscr{F}\}. \tag{1}$$

Moreover, OSF-BUILDER *constructs an sOSF ψ for $\mathscr{F}$ with $|C(\psi)| = t(\mathscr{F})$.*

Proof: We begin by showing that Equation (1) holds. Given an OSF ψ for $\mathscr{F}$, we put

$$U(\psi) = \{u \in V^0(G) : (\psi(u), \psi(v)) \in C^*(\psi) \text{ for every child } v \text{ of } u\}.$$

Note that ψ is an sOSF if and only if $U(\psi) = \emptyset$. We claim that if ψ is an OSF for $\mathscr{F}$ that is not an sOSF, then there exists an OSF ψ' for $\mathscr{F}$ such that $|C(\psi)| \geq |C(\psi')|$ and $|U(\psi)| > |U(\psi')|$. Equation (1) then follows since for any OSF ψ for $\mathscr{F}$ we can keep applying this claim until we obtain an OSF ψ'' for $\mathscr{F}$ with $|C(\psi)| \geq |C(\psi'')|$ and $|U(\psi'')| = 0$.

To prove that claim, suppose that ψ is an OSF for $\mathscr{F}$ that is not an sOSF. Choose some $u \in U(\psi)$ such that no vertex below u but distinct from u is contained in $U(\psi)$. Then there exists a subset $\{u_1 \ldots, u_k\}$, $k \geq 1$, of children of u whose elements are all mapped by ψ to some tree $T \in F$ which is different from the tree in F containing $\psi(u)$. Define the map $\psi_u : V(G) \to V(F)$ by setting $\psi_u(w) = \psi(w)$ for all $w \in V(G) - \{u\}$ and $\psi_u(u) = lca_T(\psi(u_1), \ldots, \psi(u_k))$. Then ψ_u is an OSF for $\mathscr{F}$. Note that this might have rendered the parent of u an element in $U(\psi_u)$ so that $|U(\psi_u)| = |U(\psi)|$ holds. We put $\psi = \psi_u$ and apply this construction of ψ_u to a vertex $u \in U(\psi)$ such that no vertex below but distinct from u is contained in $U(\psi)$ and so on. Since G is finite this implies that there must exist some OSF ψ' for $\mathscr{F}$ and some $u \in U(\psi')$ such that $\psi' = \psi_u$

is an OSF for $\mathscr{F}$ and $|C(\psi)| \geq |C(\psi')|$ and $|U(\psi)| > |U(\psi')|$ holds since, eventually, we will reach the root of G which does not have a parent. This concludes the proof of the claim.

Now, to see that the second statement in the theorem holds, note that by the proof of the previous claim it follows that if ψ is any sOSF for $\mathscr{F}$ with $|C(\psi)|$ minimum, then $|C(\psi)| = t(\mathscr{F})$. Hence, to complete the proof, it suffices to show that if ψ' is any sOSF for $\mathscr{F}$ output by OSF-BUILDER, then $|C(\psi)| = |C(\psi')|$.

To see this, assume that ψ is an sOSF for $\mathscr{F}$ such that $|C(\psi)|$ is minimum and that ψ' is a sOSF for $\mathscr{F}$ returned by OSF-BUILDER. Then clearly $|C(\psi')| \geq |C(\psi)|$. To show that the reverse inequality holds, first note that any introgression set I in $A(G)$ gives rise in a natural way to an extension $\bar{f} : V(G) \to F$ of the character $f_{\mathscr{F}} : L(G) \to F$. Indeed, we just extend $f_{\mathscr{F}}$ to $V(G)$ by taking the value of $v \in V(G)$ to be equal to that of $f_{\mathscr{F}}$ on the leaves of the tree in $G - I$ which contains v in its vertex set. Hence, $|C(\psi)| \geq l_{f_{\mathscr{F}}}(G)$. But $l_{f_{\mathscr{F}}}(G) = |C(\psi')|$ since ψ' is constructed by OSF-BUILDER. Thus, $|C(\psi)| \geq |C(\psi')|$ holds too which implies $|C(\psi)| = |C(\psi')|$. ∎

4. Valid networks

In applications, it is important to visualize OSFs using networks as in Figure 1(c) in the introduction (see [19]). In this section, we present some properties of networks that arise from OSFs. These are defined as follows. Given a forest triple $\mathscr{F} = (G, F, \phi)$ and an OSF ψ for $\mathscr{F}$, we define the network $N(\psi)$ to be the graph whose vertex set is $V(F)$ and whose arc set is $A(F) \cup C^*(\psi)$. An arc in $C^*(\psi)$ is called a *contact arc of* $N(\psi)$. We illustrate these concepts in Figure 5.

Note that the network of an OSF may contain no roots and/or no leaves and it may also contain directed cycles. In principle these could cause issues with modelling introgression since, for example, we want to exclude the possibility of a species containing a gene as well as a descendant of the gene which might arise in a network of an OSF containing a cycle. For DTL-scenarios where this issue also arises, a special class of scenario is often considered to resolve it (an *acyclic* scenario [24, p.520]), although this can be problematic as it leads to NP-completeness when optimising costs [24, Theorem 1]. For OSFs there is a somewhat simpler solution; as we shall show in the Appendix there is always a so-called *binary resolution* of a network of an OSF which may still contain directed cycles but that has the property that none of these cycles corresponds to the image of some path in the gene tree. In other words, these cycles do not correspond to a contradictory evolutionary scenario where a gene and its descendant are contained in the same species. An example of a binary resolution is presented in Figure 5.

We now present some basic properties of the network $N(\psi)$.

Lemma 4.1. *If ψ is an OSF for a forest triple (G, F, ϕ), then $N(\psi)$ is a network on $L(F)$. Moreover, $N(\psi)$ is connected if and only if $\phi(L(G)) \cap L(T) \neq \emptyset$ for all $T \in F$.*

Proof: To show that $N(\psi)$ is a network, first note that Property (N1) clearly holds, and that Property (N2) holds because the trees in F are phylogenetic trees whose leaf-set union is $L(F)$. Property (N3) holds since every vertex in $L(F)$ is the head of some arc in F (as a phylogenetic

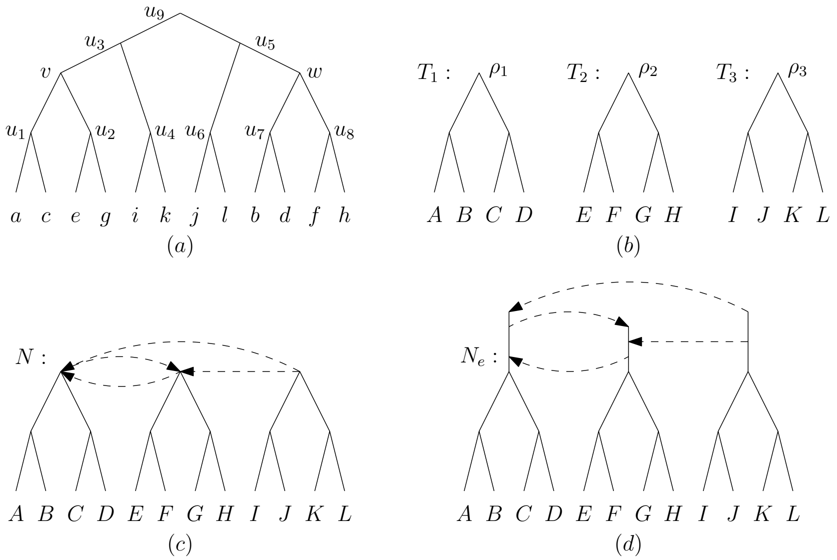
u_9
u_3
u_5
v
w
u_1
u_2
u_4
u_6
u_7
u_8
a c e g i k j l b d f h
(a)
T_1 : ρ_1
T_2 : ρ_2
T_3 : ρ_3
A B C D E F G H I J K L
(b)
N :
A B C D E F G H I J K L
(c)
N_e :
A B C D E F G H I J K L
(d)

Figure 5: A forest triple $\mathscr{F}$ consisting of the tree G depicted in (a), the forest $F = \{T_1, T_2, T_3\}$ in (b), and leaf-map defined as usual. The network N in (c) is $N(\psi)$ and $N(\psi')$ for the OSFs $\psi, \psi' : V(G) \to V(F)$ on $\mathscr{F}$ given by $\psi(u_i) = \psi'(u_i) = \rho_1$ if $i \in \{1,7\}$, $\psi(u_i) = \psi'(u_i) = \rho_2$ if $i \in \{2,8\}$, $\psi(u_i) = \psi'(u_i) = \rho_3$ for all other i, and $\psi(v) = \psi'(w) = \rho_2$ and $\psi(w) = \psi'(v) = \rho_1$. The set of dahed arcs is $C^*(\psi)$ and also $C^*(\psi')$. The network N_e in (d) is a binary resolution of N (see Appendix).

tree has at least 2 leaves). Property (N4) holds since every vertex v in $N(\psi)$ with $outdeg(v) = 0$ must clearly be contained in $L(F)$.

We now show that the second statement holds. Suppose first that $\phi(L(G)) \cap L(T) \neq \emptyset$ for all $T \in F$. By Property (P2), every arc (u, v) in G is either mapped under ψ to an element in $C^*(\psi)$ or $\psi(u)$ and $\psi(v)$ are both contained in the same tree T of F with $\psi(v)$ below $\psi(u)$ in T. Hence, if for any tree in F we pick some leaf $v \in \phi(L(G))$ (which is possible as $\phi(L(G)) \cap L(T) \neq \emptyset$ for all trees $T \in F$), it follows that under ψ the path in G from ρ_G to any $u \in L(G)$ with $\phi(u) = v$ yields a path in $N(\psi)$ from $\psi(\rho_G)$ to v. In particular, for each tree T in F, there is an undirected path in the underlying graph of $N(\psi)$ from $\psi(\rho_G)$ to some vertex in T in F. Hence, $N(\psi)$ is connected.

Conversely, suppose that $N(\psi)$ is connected. Let $T \in F$. If $\psi(\rho_G) \in V(T)$ then, by Property (P3), there is some $v \in \mathscr{C}_G(\rho_G) = L(G)$ such that $\psi(v) = \phi(v)$ also belongs to T. So $\phi(L(G)) \cap L(T) \neq \emptyset$. So assume that $\psi(\rho_G) \notin V(T)$. Let T' denote the tree in F that contains $\psi(\rho_G)$ in its vertex set. As $N(\psi)$ is connected and $|F| \geq 2$, the construction of $N(\psi)$ implies that there must exist a path from $\psi(\rho_G)$ to a vertex v in T. Without loss of generality we may assume that v is the head of some arc in $C^*(\psi)$. Then, by the definition of a contact arc, there must exist some arc (w, u) in G such that $\psi(u) = v$. Since $\psi(u) \in V(T)$, Property (P3) implies that there must exist some leaf $x \in \mathscr{C}_G(u)$ such that $\psi(x) = \phi(x)$ also belongs to T. Hence, $\phi(L(G)) \cap L(T) \neq \emptyset$ must hold in this case too. ∎

Now, we say that a network N is *a representation* of an OSF ψ on (G, F, ϕ) if it is isomorphic to $N(\psi)$ or – equivalently – there is a set $A \subseteq A(N)$ such that $N - A = (V(N), A(N) - A)$ is isomorphic to F and $A = C^*(\psi)$ (under the isomorphism between $N - A$ and F). In addition, we call a network N *valid* if there exists a forest triple $\mathscr{F}$ such that N is a representation of some OSF on $\mathscr{F}$.

Note that a network can be a representation for more than one OSF (see e.g. Figure 5). Also note that a valid network is always a representation of some sOSF. Indeed, suppose N is of the form $N(\psi)$ for some OSF ψ for some forest triple (G, F, ϕ). Then we can create a new forest triple (G', F, ϕ') by inserting a new leaf vertex g_u in G pendant to each $u \in V^0(G)$ to create G' and extending ϕ to a leaf-map on $L(G')$ by mapping each new vertex g_u to a leaf in the tree to which u is mapped under ψ that is below $\psi(u)$. Then it is straight-forward to check that the map ψ' obtained by extending ψ in the natural way to a map $V(G') \to V(F)$ is an sOSF for (G', F, ϕ') such that $N(\psi)$ is isomorphic to $N(\psi')$.

Interestingly, as we shall now show, an even stronger result holds:

Theorem 4.2. *Suppose that ψ is an OSF for some forest triple $\mathscr{F} = (G, F, \phi)$. Then there is an sOSF ψ' for some forest triple $\mathscr{F}' = (G', F, \phi')$ that is output by* OSF-BUILDER *such that $N(\psi)$ is isomorphic to $N(\psi')$. In particular, a network N is valid if and only if there is some OSF ψ output by* OSF-BUILDER *such that N is a representation of ψ.*

Proof: Consider the tree G. Construct a new phylogenetic tree G' by inserting, for each $u \in V^0(G)$, $outdeg(u) + 1$ new arcs into G of the form (u, v) (so that in particular v is a leaf in G'). Extend the leaf-map ϕ on $L(G)$ to a leaf-map ϕ' on $L(G')$ by, for each u in $V^0(G)$, mapping the new children of u arbitrarily onto a set S of leaves in $L(T)$, where $T \in F$ is the tree with $\psi(u) \in V(T)$, so that $lca_T(S) = \psi(u)$. Let ψ' be the extension of ψ to $V(G')$ given by putting, for all $v \in V(G')$, $\psi'(v) = \psi(v)$ if $v \in V(G)$ and $\psi'(v) = \phi'(v)$ otherwise.

It is straight-forward to check that ψ' is an sOSF for (G', F, ϕ'), and that $N(\psi')$ is isomorphic to $N(\psi)$. Moreover, for all $u \in V^0(G')$, the definition of G' implies that the set $\sigma(u)$ computed by OSF-BUILDER must have size 1 and that it consists of the tree in F in which $\psi(u)$ is contained. It follows that the necessarily unique map obtained by OSF-BUILDER by applying its top-down phase is equal to ψ'. ∎

5. A characterization of valid networks

In this section, we give a characterization for valid networks. To do this we first show that a valid network can be represented by a special type of OSF.

We start with a definition. Suppose that (G, F, ϕ) is a forest triple and that ψ is an OSF for (G, F, ϕ). From a path $\gamma = w_1, w_2, \ldots, w_k$, $k \geq 2$, in G, we derive a walk γ_ψ in $N(\psi)$ (possibly of length 0) that starts at $\psi(w_1)$ and ends at $\psi(w_k)$. Central to the definition of γ_ψ is the observation that if (w_i, w_{i+1}), $1 \leq i \leq k-1$, is a contact arc of ψ, then $(\psi(w_i), \psi(w_{i+1}))$ is an arc in $N(\psi)$ but not in F. Otherwise $\psi(w_i)$ and $\psi(w_{i+1})$ are both vertices in some tree T of F, and, therefore, there exists a path in T (possibly of length 0) starting at $\psi(w_i)$ and ending at $\psi(w_{i+1})$. To obtain γ_ψ we then take the walk obtained by inserting these paths into the sequence $\psi(w_1), \psi(w_2), \ldots, \psi(w_k)$, where any consecutive repeats are suppressed.

Proposition 5.1. *Suppose that ψ is an OSF for some forest triple (G,F,ϕ). Then there exists a phylogenetic tree G_0 on $L(G)$ and an OSF ψ_0 for (G_0,F,ϕ) such that, for every path γ in G_0, the walk γ_{ψ_0} is in fact a trail in $N(\psi_0)$ and $N(\psi)$ is isomorphic to $N(\psi_0)$.*

Proof. We first remark that if there exists a path γ in G such that the walk γ_ψ is not a trail in $N(\psi)$, then γ_ψ must contain an arc of $N(\psi)$ that is crossed by γ_ψ twice. Hence, there must exist two arcs (u_1,v_1) and (u_2,v_2) of G distinct such that $\psi(u_1)=\psi(u_2)\neq\psi(v_1)=\psi(v_2)$, and v_1 is an ancestor of u_2 in G. Denoting by $\Gamma(G,\psi)$ the set of such pairs of arcs, it then suffices to construct a phylogenetic tree G_0 on $L(G)$ from G and an OSF ψ_0 on (G_0,F,ϕ) such that $N(\psi_0)$ and $N(\psi)$ are isomorphic and the analogously defined set $\Gamma(G_0,\psi_0)$ is empty.

Now, let G' be the phylogenetic tree obtained from G by removing the arc (u_2,v_2), subdividing the arc (u_1,v_1) by inserting a new vertex w, adding the arc (w,v_2), and suppressing u_2 as this has rendered it a vertex with indegree and outdegree one (See Figure 6 for an example with $(u_1,v_1)=(\rho,w_2)$ and $(u_2,v_2)=(w_4,w_6)$). We define the OSF ψ' on (G',F,ϕ) by putting $\psi'(v)=\psi(v)$ for all vertices $v\in V(G')-\{w\}$ and $\psi'(w)=\psi(v_2)$. Clearly, $N(\psi')$ and $N(\psi)$are isomorphic by construction. Moreover, it can be easily checked that $|\Gamma(G',\psi')|<|\Gamma(G,\psi)|$. Since $\Gamma(G,\psi)$ is finite repeating this process must yield a phylogenetic tree G_0 and an OSF ψ_0 on (G_0,F,ϕ) such that $|\Gamma(G_0,\psi_0)|=0$. □

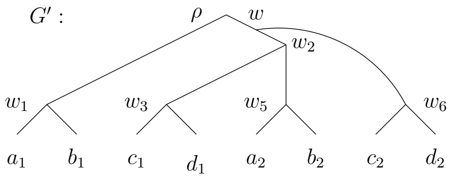

Figure 6: The phylogenetic tree G' obtained from G in the construction presented in Proposition 5.1 applied to the OSF $(G,F\phi)$ presented in Figure 3.

We now present the aforementioned characterization of valid networks. Our proof uses the idea of unfolding a network, which was introduced in [11].

Theorem 5.2. *Suppose that N is a network on X. Then N is valid if and only if there is some vertex $\rho\in V(N)$ and some $A\subseteq A(N)$ such that*

(V1) $N-A=(V(N),A(N)-A)$ is a forest F with $|F|\geq 2$ and leaf-set X in which every arc in A has its ends in different trees of F, and

(V2) every arc in A is contained in some trail $\rho=v_1,v_2,\dots,v_m$, $m\geq 2$, of N such that if v_i,v_j, $i\leq j$ are vertices in the same tree T in F then $v_i\preceq_T v_j$.

Proof: Suppose first that N is valid, that is, $N=N(\psi)$ for some OSF ψ of a forest triple $\mathscr{F}=(G,F,\phi)$. Put $A=C^*(\psi)$. As ψ is an OSF for $\mathscr{F}$, it follows in view of the definition of a contact arc that $N-A$ is a forest F such that every arc in A has its ends in different trees of F. In view of Properties (N2) and (N4), we also have $L(F)=X$. Hence, Property (V1) holds.

We now show that Property (V2) holds relative to the vertex $\rho = \psi(\rho_G)$ and the set $A = C^*(\psi)$. Note first that by Proposition 5.1, we may assume for every path γ in G that the associated walk γ_ψ in $N(\psi)$ is a trail. Now, suppose we are given some arc (u', v') in A. Then by definition of $C^*(\psi)$, there must exist some arc (u, v) in G such that $\psi(u) = u'$ and $\psi(v) = v'$. Let T denote the tree in F that contains v' in its vertex set. Then Property (P3) implies that there must exist some leaf $l \in \mathscr{C}_G(v)$ such that $\psi(l) = \phi(l)$ also belongs to T. Hence, there must exist a path γ in G from ρ_G to l that contains (u, v). By Property (P2), it follows that the trail γ_ψ in $N(\psi)$ has the required properties. So Property (V2) holds, as required.

Conversely, suppose N is a network on X, that contains an arc set A and a vertex ρ such that (V1) and (V2) are satisfied. Let F denote the forest $N - A$. Then $|F| \geq 2$ and $L(F) = X$.

We start with constructing a forest triple (G, F, ϕ). Let $G = G(N)$ be the graph obtained by "unfolding N at ρ", which we define as follows (cf. [11, p.617]):

- the vertices of G are trails $\rho = v_1, v_2, \ldots, v_m$, $m \geq 1$, in N such that if v_i, v_j, $i \leq j$, are both vertices in some tree T of F then $v_i \preceq_T v_j$;
- for all vertices γ and γ' in G there is an arc from γ to γ' in G if and only if $\gamma' = \gamma a$ holds for some arc a in $A(N)$;
- the vertices in G that start at ρ and end at a vertex $v \in X$ are labelled by distinct elements in the set $X_v = \{v_i : 1 \leq i \leq n_v\}$, where n_v is the number of vertices in G which end at v.

By construction it immediately follows that G is a tree with root ρ (considered as a path with length 0).

To see that G is a phylogenetic tree suppose that γ is an element in $V(G)$ whose end vertex v is not a leaf in N. Then, by Property (V1), v must be an interior vertex of some tree T in F. Since $outdeg(v) \geq 2$ there must be at least 2 possibilities to extend γ by adding an arc in T whose tail is v. Hence, γ is an interior vertex of G with outdegree at least 2. Moreover, the leaf-set of G is equal to the set obtained by taking the union Y of the sets X_v, where the union is taken over all $v \in X$. Since $|Y| \geq 2$, it follows that G is a phylogenetic tree on Y.

To obtain the leaf-map $\phi : L(G) \to L(F)$ note that there is a natural map from $V(G)$ into $V(N)$ which takes each element in $V(G)$ to its end vertex in N. This gives rise to a map ψ from $V(G)$ to $V(F)$, which maps the vertex set of G to the vertex set of N. We let $\phi = \psi|_{L(G)}$ be the map from $L(G)$ to $L(F)$ induced by ψ. This completes the construction of the forest triple (G, F, ϕ).

We now claim that ψ is an OSF for (G, F, ϕ) such that $A = C^*(\psi)$, which will complete the proof the theorem. Property (P1) holds by definition of ϕ. That Property (P2) holds follows immediately from Property (V2) and the definition of G and ψ. Property (P3) holds since if γ is an element in $V(G)$ with end vertex v in N, then v is contained in some tree T of F. In case v is not a leaf of T there must exist a leaf $l \in V(T)$ below v. In that case, let $\gamma_1 \in V(G)$ denote the path obtained by extending γ by some path in T from v to l. Otherwise, put $\gamma_1 = \gamma$. In particular, it follows that $\gamma_1 \in \mathscr{C}_G(\gamma)$, and that $\psi(\gamma_1)$ belongs to T.

It remains to show that $A = C^*(\psi)$. Clearly $C^*(\psi) \subseteq A$ by Property (V1). Conversely, suppose $a \in A$. By Property (V2), a is contained in some vertex γ in $V(G)$, which we may assume without loss of generality to have a as the last arc. But then for the path γ_1 in $V(G)$ with $\gamma = \gamma_1 a$

we have $(\gamma_1, \gamma) \in A(G)$ and $a = (\psi(\gamma_1), \psi(\gamma))$. Thus $a \in C^*(\psi)$. ∎

6. Stability of optimal scores under tree and forest alterations

In this section, we are motivated by the following question. If we alter G or F in a forest triple $\mathscr{F} = (G, F, \phi)$, then how much does this change the minimum number of contact arcs of an OSF for $\mathscr{F}$ (i.e. by how much can $t(\mathscr{F})$ change)? This is important since it can help indicate, for example, how changes to the input of OSF-BUILDER can effect its output. In [19], the effect of altering G and F by SPR operations on $t(\mathscr{F})$ was studied empirically using simulations. Here we investigate this question from a theoretical point of view.

From now on all forest triples are assumed to be binary. We begin by recalling some facts concerning subtree prune and regraft (SPR) operations for rooted phylogenetic trees as defined in [4]. An *SPR operation* on a tree T is defined as cutting any arc (u, v) in $A(T)$ (so pruning off the subtree of T rooted at v), and then regrafting this pruned subtree into a subdivided arc of the pruned tree. Note that some care has to be taken when pruning off a subtree adjacent to the root; the reader can find full details in [4]. The *rooted SPR distance* between two phylogenetic trees T and T' on the same leaf set, denoted $d_{rSPR}(T, T')$, is the minimum number of SPR operations requited to transform one tree into the other.

We now ask the above question more precisely: If G and G' are two phylogenetic trees that differ by k SPR operations and $\mathscr{F} = (G, F, \phi)$ and $\mathscr{F}' = (G', F, \phi)$ are forest triples, then how different can $t(\mathscr{F})$ and $t(\mathscr{F}')$ be in terms of k? We next give an upper bound for this difference:

Theorem 6.1. *If G and G' are two binary phylogenetic trees on the same leaf set, and $\mathscr{F} = (G, F, \phi)$ and $\mathscr{F}' = (G', F, \phi)$ are forest triples, then $|t(\mathscr{F}) - t(\mathscr{F}')| \leq d_{rSPR}(G, G')$.*

This theorem follows immediately from [6, Corollary 3.12]. This corollary states that if G and G' are two binary phylogenetic trees on X, and f is a character on X, then $|l_f(G) - l_f(G')| \leq d_{rSPR}(G, G')$. The theorem thus follows since if $\mathscr{F}$ is a forest triple, then $t(\mathscr{F}) = l_{f_{\mathscr{F}}}(G)$.

The results in [6] also provide the following additional bounds on $|t(\mathscr{F}) - t(\mathscr{F}')|$. In case there are r trees in F whose leaf-sets intersect the image $\phi(L(G))$, and $r \leq n = |L(G)|$, then by [6, Lemma 3.14]

$$|t(\mathscr{F}) - t(\mathscr{F}')| \leq \lfloor (r-1)(\frac{n}{r} - 1) \rfloor,$$

and by [6, Theorem 3.15]

$$|t(\mathscr{F}) - t(\mathscr{F}')| \leq n - 2\sqrt{n} + 1,$$

a bound which is tight for $n = 9$ ([6, Figure 4]).

We now present a result which describes how applying an SPR operation to F in a forest triple $\mathscr{F} = (G, F, \phi)$ can affect $t(\mathscr{F})$. Note that if we alter any tree in F by a SPR operation then this has no effect on $t(\mathscr{F})$, as the associated character $f_{\mathscr{F}}$ remains unchanged.

Theorem 6.2. *Let $\mathscr{F} = (G, F, \phi)$ be a forest triple. Suppose a forest F' is obtained from F by pruning off some subtree T_0 in a tree $T \in F$ and grafting T_0 into a tree in $F - \{T\}$. Let $\mathscr{F}' = (G, F', \phi)$ denote the resulting forest triple. Then*

$$|t(\mathscr{F}) - t(\mathscr{F}')| \leq |\{v \in L(G) : \phi(g) \in L(T_0)\}|.$$

The theorem follows immediately from the following lemma, which generalizes [6, Observation 4.2].

Lemma 6.3. *If f is a character on a set X, T is a phylogenetic tree on X, and f' is a character on X obtained by changing the value of f on exactly $k \geq 1$ elements in X, then $|l_{f'}(T) - l_f(T)| \leq k$.*

Proof: We use induction on k. The base case, $k = 1$, is [6, Observation 4.2].

Now, suppose the inequality in the lemma holds for all $k \leq L-1$, some $L \geq 2$. Let f be a character on X, let T be a phylogenetic tree on X, and let f' be a character on X obtained by changing the value of f on exactly L elements in X. Let f'' be a character on X obtained by changing the value of f' on exactly one element $x \in X$ to the value $f(x)$. Then $|l_{f''}(T) - l_{f'}(T)| \leq 1$, and by induction $|l_{f'}(T) - l_f(T)| \leq L-1$. The lemma follows. ∎

7. Conclusion

There are several new directions that could be of potential interest. One possibility following on from the results presented in the last section could be to try and understand the effect that changing the root location of the gene tree can have on the optimal score for an OSF for a forest triple. This is important as it can be difficult to accurately root the tree in practice, and its location is known to have an impact on duplication-transfer-loss models for lateral gene transfer (see e.g.[14]).

In another direction, it could be worth defining and studying spaces of OSFs for a given forest triple – such spaces have been intensively studied for gene/species tree reconciliation models and can, for example, give important insights on how optimal OSFs are distributed over the collection of all possible OSFs (see e.g. [5]). Defining such spaces could also lead to new metrics that could be used to compare OSFs such as the ones defined in e.g. [12].

Finally, note that a network for an OSF can be thought of a special example of a forest with extra edges added. There has been much work recently on understanding the structure of networks that arise from adding some edges into a tree (so-called tree-based networks) [8, 27]. It would thus be of interest to see what properties "forest-based" networks have in common (or not) with tree-based networks. Along similar lines it might also be of interest to extend our approach by allowing contact arcs to also be added between vertices within the same tree of a forest, i.e. to consider a more general model of gene introgression in which genes can introgress both within and between lineages.

Appendix

As mentioned in Section 4, the network $N(\psi)$ for an OSF ψ may have no roots and/or no leaves and it may also contain directed cycles which can make it difficult to interpret in practice. In this Appendix we present a way to produce a binary resolution N_ψ of $N(\psi)$ which helps to circumvent these issues (cf. [19]). In particular, we show that this resolution has the attractive property that it only contains *incidental cycles* in $N(\psi)$ that is, cycles that are *not* of the form γ_ψ

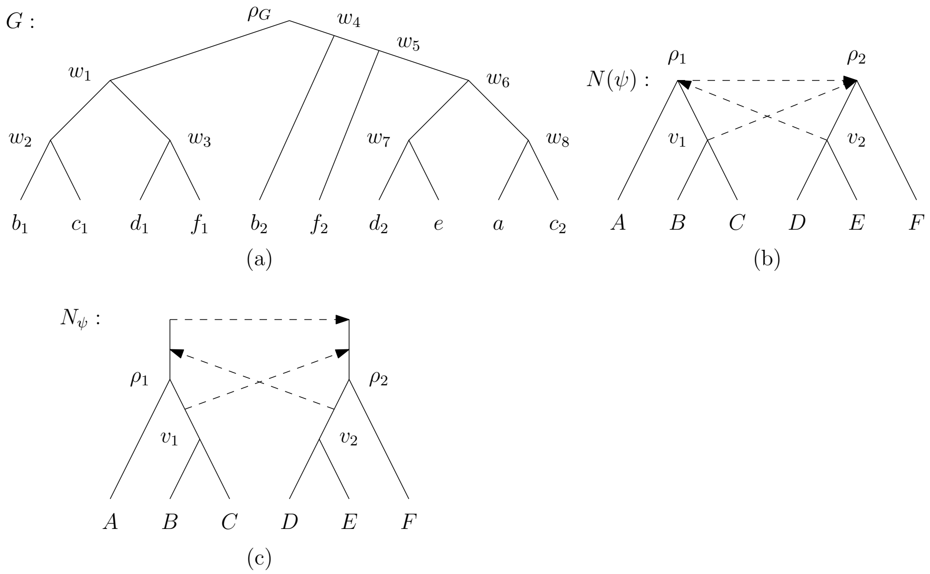

Figure 7: (a) A gene tree G. (b) The representation $N(\psi)$ of the OSF ψ mapping ρ_G, w_4 and w_8 to ρ_1, w_5 and w_3 to ρ_2, w_1 and w_2 to v_1, and w_6 and w_7 to v_2. The network in (b) contains an incidental cycle $\rho_1, v_1, \rho_2, v_2, \rho_1$. The second cycle in $N(\psi)$, $\rho_1, \rho_2, v_2, \rho_1$, is *not* incidental as it is the image under ψ of the path w_4, w_5, w_6, w_8 in G. (c) A resolution N_ψ of $N(\psi)$. The only cycle in N_ψ comes from the incidental cycle of $N(\psi)$.

for some path γ in G (see Section 5 for a definition of γ_ψ). We refer to Figure 7 for an illustration of the definitions and results in this section.

To prove the main theorem of this section we start by making a simple observation that is implied by Property (P2).

Observation *Suppose that ψ is an OSF for $\mathscr{F} = (G, F, \phi)$ and that γ is a path in G such that γ_ψ is a cycle in $N(\psi)$, then the first and last arcs in γ_ψ are both contact arcs for ψ.*

We now explain how to generate a binary resolution N_ψ of $N(\psi)$ – see Figure 7 for an example. Suppose ψ is an OSF for some forest triple $\mathscr{F} = (G, F, \phi)$. Without loss of generality we may assume that ψ is an sOSF. Let $I \subseteq A(G)$ denote the introgression set induced by $C(\psi)$ on G. We start with associating a directed graph N' to $N(\psi)$. To do this, we first attach an incoming arc to every root of $N(\psi)$. We consider these arcs also as arcs in F. For every vertex $v \in V(N(\psi))$ such that there exists some $w \in V(G)$ with $v = \psi(w)$ we then add $|\{a \in I : w \in \{head(a), tail(a)\}\}|$ subdivision vertices to the incoming arc of v in F all of which we label w. If v is in the image of more than one vertex of G under ψ, say w and w', then we also ensure that the ancestral relationships between the subdivision vertices labelled w and w' are preserved. Next, we attach every contact arc in $C^*(\psi)$ to a pair of correspondingly labelled subdivision vertices so that every arc in G is an arc in the resulting graph. Finally, we remove all vertices of indegree zero and outdegree one (and their outgoing arcs and labels).

Note that the resulting graph N' has at least one root, leaf set $L(F)$, and potentially still vertices that are involved in two or more outgoing arcs in F. Also note that N' might not be a representation of ψ since the trees in $N' - C^*(\psi)$ might contain vertices with indegree and outdegree one. However N' can be easily transformed into a representation of an OSF ψ' for some forest triple (G, F', ϕ') with $C^*(\psi') = C^*(\psi)$ by attaching to each subdivision vertex of N' a new leaf to obtain a new network N'' and defining F', ϕ and ψ' in the canonical way.

Note that by our observation, any cycle in $N(\psi)$ that is *not* incidental does not give rise to a cycle in N'' (i.e. non-incidental cycles in $N(\psi)$ are broken in N''). Resolving potential vertices in $N(\psi')$ that have three or more outgoing arcs in F results in a binary resolution for $N(\psi')$. The directed graph obtained by removing all leaves from that resolution that are not also contained in $N(\psi)$ (plus their incident arcs) is N_ψ.

In this construction some choices might be made and so there could be several binary resolutions. However, every binary resolution has the following property, which follows from the above construction (as an illustration see the binary resolution in Figure 7(c)):

Theorem *Suppose that ψ is an OSF. Then any cycle in N_ψ must come from an incidental cycle in $N(\psi)$.*

This theorem highlights the way in which the OSF model differs from the model of lateral gene transfer presented in [24]. In the latter model biologically infeasible DTL-scenarios can potentially arise [24, cf. p.520] (essentially a gene that transfers into an ancestor of the species from which it arose). However, by the last result we can exclude this situation for the OSF model, since we only obtain incidental cycles in N_ψ.

Acknowledgement The authors would like to thank Martin Taylor and Andrei-Alin Popescu for stimulating discussion on the topic of the paper. GS would like to thank the School of Computing Sciences, University of East Anglia where some of the ideas underlying this paper were conceived. All authors would like to thank the two anonymous referees for their helpful comments which helped them to improve the paper.

[1] M. A. Alexandrou, C. Oliveira, M. Maillard, R. A. R. McGill, J. Newton, S. Creer, and M.I. Taylor. Competition and phylogeny determine community structure in müllerian co-mimics. *Nature*, 469:84–88, 2011.

[2] M. L. Arnold and N. H. Martin. Minireview: Adaption by introgression. *Journal of Biology*, 8:82, 2009.

[3] C. Baudet, B. Donati, B. Sinaimeri, P. Crescenzi, C Gautier, C. Matias, and M-F Sagot. Cophylogeny reconstruction via an approximate bayesian computation. *Systematic Biology*, 64(3):416–431, 2015.

[4] M. Bordewich and C. Semple. On the computational complexity fo the rooted subtree prune and regraft distance. *Annals of Combinatorics*, 8:409–423, 2004.

[5] J.-P. Doyon, C. Chauve, and S. Hamel. Space of gene/species trees reconciliations and parsimonious models. *Journal of Computational Biology*, 16(10):1399–1418, 2009.

[6] M. Fischer and S. Kelk. On the maximum parsimony distance between phylogenetic trees. *Annals of Combinatorics*, 20(1):87–113, 2016.

[7] W. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Systematic Zoology*, 20:406–416, 1971.

[8] A. R. Francis and M. Steel. Which phylogenetic networks are merely trees with additional arcs? *Systematic Biology*, 64(5):768–777, 2015.

[9] P. R. Grant and B. R. Grant. Introgressive hybridization and natural selection in darwin's finches. *Biological Journal of the Linnean Society*, 117(4):812–822, 2016.

[10] J. A. Hartigan. Minimum mutation fits to a given tree. *Biometrics*, 29:53–65, 1973.

[11] K. T. Huber and V. Moulton. Phylogenetic networks from multi-labelled trees. *Journal of Mathematical Biology*, 52:613–632, 2006.

[12] K. T. Huber, V. Moulton, M.-F. Sagot, and B. Sinaimeri. Exploring and visualizing spaces of tree reconciliations. *Systematic Biology*, 68(4):607–618, 2019.

[13] K. T. Huber, V. Moulton, M. Steel, and T. Wu. Folding and unfolding phylogenetic trees and networks. *Journal of Mathematical Biology*, 73(6-7):1761–1780, 2016.

[14] S. Kundu and M. S. Bansal. On the impact of uncertain gene tree rooting on duplication-transfer-loss reconciliation. *BMC Bioinformatics*, 19(9):21–31, 2018.

[15] R. Libeskind-Hadas, Y.-C. Wu, M. S. Bansal, and M. Kellis. Pareto-optimal phylogenetic tree reconciliation. *Bioinformatics*, 30(12):i87–i95, 2014.

[16] J. Mallet. Hybridization as an invasion of the genome. *Trends in Ecology & Evolution*, 20(5):229–237, 2005.

[17] J. Parker, A. Rambaut, and O. Pybus. Correlating viral phenotypes with phylogeny: Accounting for phylogenetic uncertainty. *Infection, Genetics and Evolution*, 8(3):239–46, 2008.

[18] S. Sankararaman, S. Mallick, M. Dannemann, K. Prüfer, J. Kelso, S. Pääbo, N. Patterson, and D. Reich. The genomic landscape of neanderthal ancestry in present-day humans. *Nature*, 507(7492):354–357, 2014.

[19] G. E. Scholz, A.-A. Popescu, M. I. Taylor, V. Moulton, and K. T. Huber. Osf-builder: A new tool for constructing and representing evolutionary histories involving introgression. *Systematic Biology*, 68(5):717–729, 2019.

[20] C. Semple and M. Steel. *Phylogenetics*, volume 24. Oxford University Press, 2003.

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

[22] V. Sousa and J. Hey. Understanding the origin of species with genome-scale data: modelling gene flow. *Nature Reviews Genetics*, 14(6):404–414, 2013.

[23] C. N. Stewart, M. D. Halfhill, and S. I. Warwick. Transgene introgression from genetically modified crops to their wild relatives. *Nature Reviews Genetics*, 4(10):806–817, 2003.

[24] A. Tofigh, M. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):517–535, 2011.

[25] R. W. R. Wallbank, S. W. Baxter, C. Pardo-Diaz, J. J. Hanly, Martin S. H., J. Mallet, K. K. Dasmahapatra, C. Salazar, M. Joron, N. Nadeau, W. O. McMillan, and C. D. Jiggins. Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol.*, 14(1):e1002353, 2016.

[26] D. Wen, Y. Yu, M. W. Hahn, and L. Nakhleh. Reticulate evolutionary history and extensive introgression in mosquito species revealed by phylogenetic network analysis. *Molecular Ecology*, 25(11):2361–2372, 2016.

[27] L. Zhang. On tree-based phylogenetic networks. *Journal of Computational Biology*, 23(7):553–565, 2016.

[28] W. Zhang, K. K. Dasmahapatra, J. Mallet, G. R.P. Moreira, and M. R. Kronforst. Genome-wide introgression among distantly related heliconius butterfly species. *Genome Biology*, 17(1):25, 2016.