Bulletin of Mathematical Biology (2010) 72: 1783–1798
DOI 10.1007/s11538-010-9506-7

ORIGINAL ARTICLE

Quantifying the Extent of Lateral Gene Transfer Required to Avert a 'Genome of Eden'

Leo van Iersel*, Charles Semple, Mike Steel

Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

Received: 5 November 2009 / Accepted: 8 January 2010 / Published online: 20 January 2010

Abstract The complex pattern of presence and absence of many genes across different species provides tantalising clues as to how genes evolved through the processes of gene genesis, gene loss, and lateral gene transfer (LGT). The extent of LGT, particularly in prokaryotes, and its implications for creating a 'network of life' rather than a 'tree of life' is controversial. In this paper, we formally model the problem of quantifying LGT, and provide exact mathematical bounds, and new computational results. In particular, we investigate the computational complexity of quantifying the extent of LGT under the simple models of gene genesis, loss, and transfer on which a recent heuristic analysis of biological data relied. Our approach takes advantage of a relationship between LGT optimization and graph-theoretical concepts such as tree width and network flow.

Keywords Tree · Phylogenetic network · Lateral gene transfer · Tree-width

1. Introduction

Modern sequencing technology is providing an increasingly detailed picture of the distribution of genes across a wide array of taxa. Some molecular biologists have used these data to argue that unless ancestral genomes were considerably larger than present-day ones, extensive lateral gene transfer (LGT) must be invoked to explain the current distribution of genes (Dagan and Martin, 2007; Dagan et al., 2008; Mirkin et al., 2003). LGT is a process by which a gene (or genes) from one species is transferred into the genotype of another species by various genetic mechanisms. The extent of LGT is controversial, but it has been argued to be widespread in prokaryotes (e.g. bacteria) and, to some extent, in other domains of life (Andersson, 2005), suggesting in turn that a network, rather than a tree, best describes the evolution of life (Doolittle and Bapteste, 2007).

*Corresponding author.
E-mail addresses: l.j.j.v.iersel@gmail.com (Leo van Iersel), c.semple@math.canterbury.ac.nz (Charles Semple), m.steel@math.canterbury.ac.nz (Mike Steel).
We thank the Allan Wilson Centre for Molecular Ecology and Evolution, and the New Zealand Marsden Fund for helping fund this work.

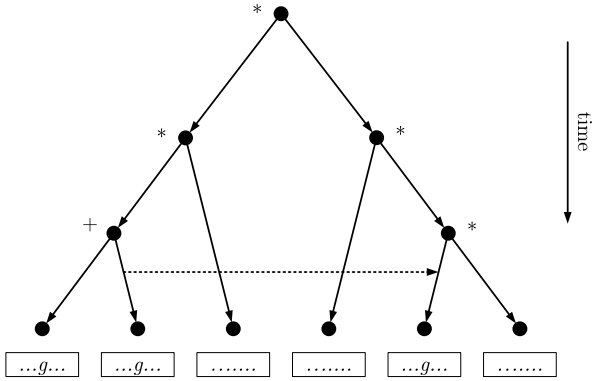

Fig. 1 The dilemma of ancestral genome inflation: If gene g, distributed as shown, is not transferred laterally then under the model, g must be in five ancestral genomes $(*, +)$ not just at $+$.

Although the pattern of presence and absence of different genes across a set of species can suggest that LGT events occurred in the evolution of these species, another explanation is that certain genes are simply lost in different lineages. As a result, various attempts to quantify the extent of LGT based on gene content have been developed, typically based either on most-parsimonious scenarios or on stochastic models of gene genesis, loss, and transfer (see, for example, Dagan and Martin, 2007; Jin et al., 2007; Spencer et al., 2006). Attempts to reconstruct evolutionary histories under the assumption that no LGT events have occurred (and that genes arise just once) imply that some common ancestors of the considered species must have had far more genes than their current-day descendants. Doolittle et al. (2003) refer to such an unlikely all-encompassing ancestral genome as the 'genome of Eden' hypothesis. Allowing LGT events reduces the need for genes to be present at earlier species, as illustrated for a single gene in Fig. 1.

In this paper, we exploit the combinatorial structure that underlies a key biological insight on which a recent heuristic analysis of data was based by Dagan and Martin (2007) (see also Dagan et al., 2008; Mirkin et al., 2003). This insight is that simple models of gene evolution, in which a gene typically arises just once (gene genesis) but can be lost multiple times, imply lower bounds on the extent of LGT simply to prevent hypothetical ancestral genomes from becoming unfeasibly large (e.g. if no LGT events are allowed, some ancestral genomes in Dagan and Martin (2007) were at least five times larger than modern prokaryote genomes). For such a model, we aim to bound the number of gene transfer events that have occurred in the evolution of a set of taxa, based on the presence/absence patterns of genes in each of these taxa, assuming that ancestral genomes are bounded by a given size.

Notice that we wish to count transfer events (rather than the total number of genes that are transferred), since in each transfer event, several genes may be transferred from one species into another. Thus, our count of LGTs is conservative, and recognizes that genes are not independently transferred and that a transfer event may insert a section of the genome (with several genes) into an individual organism of a different species.

The structure of this paper is as follows. In the next section, we define the model of gene genesis, loss, and transfer precisely, and summarize our main results. We then

provide proofs of these results in subsequent sections, and end with some concluding comments and a conjecture.

2. Mathematical model and summary of main results

2.1. Definitions and model specification

We begin by recalling some notation concerning digraphs, and phylogenetic trees and networks.

Let v be a vertex of a digraph D. The *indegree* of v is the number of arcs directed into v, while the *outdegree* of v is the number arcs directed out of v. The indegree of v is denoted by $d^-(v)$ and the outdegree of v is denoted by $d^+(v)$. The *degree* of v is $d^-(v) + d^+(v)$. Furthermore, u is an *in-neighbour* of v if (u, v) is an arc in D, while w is an *out-neighbour* of v if (v, w) is an arc in D. A digraph D is *rooted* if there exists a vertex, ρ say, of indegree zero such that, for each vertex v in D, there exists a directed path from ρ to v.

Throughout the paper, $\mathcal{X}$ will denote a finite set of taxa and $\mathcal{G}$ will denote a finite set of genes. A *phylogenetic tree* (*on* $\mathcal{X}$) is a rooted tree whose root has degree at least two and all other internal vertices have degree at least three, and whose leaf set is $\mathcal{X}$. More generally, a *phylogenetic network* N (*on* $\mathcal{X}$) is a rooted acyclic digraph with the following properties:

(i) the root has outdegree at least two and, for all vertices v with $d^+(v) = 1$, we have $d^-(v) \geq 2$; and
(ii) the set of vertices of outdegree zero is $\mathcal{X}$.

The elements of $\mathcal{X}$ are the *leaves* of N. For a subset U of the vertex set of N, the sub-digraph of $N = (V, A)$ *induced by* U is the digraph whose vertex set is U, and whose arc set is the subset $\{(u, v) : u, v \in U \text{ and } (u, v) \in A\}$ of A.

We now describe the model of gene genesis, loss, and transfer. For each taxon $x \in \mathcal{X}$, assume that the subset $G(x)$ of $\mathcal{G}$ consisting of the genes in $\mathcal{G}$ that have been observed in taxon x is known. We refer to the associated map $G : \mathcal{X} \to 2^{\mathcal{G}}$ as a *genome assignment*. Let $N = (V, A)$ be a phylogenetic network on $\mathcal{X}$. For a fixed positive integer k, and a genome assignment $G : \mathcal{X} \to 2^{\mathcal{G}}$, a (G, k)-*gene labelling of* N is a mapping $F : V \to 2^{\mathcal{G}}$ such that the following hold:

(I) $F(x) = G(x)$ for each $x \in \mathcal{X}$;
(II) $|F(v)| \leq k$ for all $v \in V$;
(III) For each gene $g \in \mathcal{G}$, the sub-digraph of N induced by $\{v \in V : g \in F(v)\}$ is rooted (and, therefore, connected).

Note that if $x \in \mathcal{X}$ and $|G(x)| > k$, then N has no (G, k)-labelling. If N has a (G, k)-labelling, we say that N *exhibits* such a labelling. A gene labelling describes a possible evolution of the genes observed in the taxa under consideration. Property (I) says that each leaf of the network is labelled by the set of genes observed in the corresponding taxon. Property (II) demands that each vertex is labelled by a set of at most k genes; the parameter k thus bounds the sizes of the ancestral genomes. Lastly, (III), means that each

gene in $\mathcal{G}$ is created once at most. There is no restriction on the number of times a gene is lost.

Any function F which satisfies properties (I) and (III) we will call a *G-gene labelling*. With these definitions in hand, we can now state the main results of this paper.

2.2. *Bounding the number of gene transfers required*

Our first result establishes lower and upper bounds on the number of LGT events required to explain a given data set. Suppose our input is given by a rooted phylogenetic tree T on $\mathcal{X}$ ("species tree"), a genome assignment $G : \mathcal{X} \to 2^{\mathcal{G}}$, and a positive integer k. Given a phylogenetic network N, we say that N can be obtained from T by *adding h arcs*, if there is a subgraph T' of N that is a subdivision of T (i.e. T' can be obtained from T by replacing arcs by directed paths) and at most h arcs of N are not arcs of T'. Here, one views these added arcs as LGT events.

We are interested in the minimum number of LGT events that must be added to T in order for the resulting network to exhibit a (G, k)-gene labelling. We denote this minimum number by $\ell(T, G, k)$. Given the above input, Theorem 1 provides lower and upper bounds for $\ell(T, G, k)$. For a vertex v of T, let $n(v)$ denote the number of genes $g \in \mathcal{G}$ for which there exist two leaves $x_1, x_2 \in \mathcal{X}$ such that $g \in G(x_1)$, $g \in G(x_2)$ and the most recent common ancestor of x_1 and x_2 in T is v.

Theorem 1. *Let $T = (V, E)$ be a rooted phylogenetic tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer such that $|G(x)| \leq k$ for all $x \in \mathcal{X}$. Then*

(i) $\ell(T, G, k) \geq \sqrt{\frac{2}{3}|\{v \in V : n(v) > k\}|}$.

(ii) $\ell(T, G, k) \leq \lceil \frac{|\mathcal{G}|-k}{k} \rceil \cdot (|\mathcal{X}| + 1)$.

In particular, it follows that, to any phylogenetic tree, LGT events can be added in order for the resulting network to exhibit a (G, k)-gene labelling (for any G, k with $|G(x)| \leq k$ for all $x \in \mathcal{X}$). The proof of Theorem 1 is given in Section 3.

2.3. *Hardness results*

The next two results show that two fundamental decision questions concerning the existence of (G, k)-labellings are NP-complete. First, consider the following problem:

GENE LABELLING

Given: A phylogenetic network N on $\mathcal{X}$, a finite set $\mathcal{G}$ of genes, a genome assignment $G : \mathcal{X} \to 2^{\mathcal{G}}$, and a positive integer k.
Question: Does N exhibit a (G, k)-labelling?

Theorem 2. *The decision problem* GENE LABELLING *is NP-complete even if $k = 1$.*

A related problem, but concerning rooted phylogenetic trees, is the following:

(G, k)-TREE

Given: A finite set $\mathcal{X}$ of taxa, a finite set $\mathcal{G}$ of genes, a genome assignment $G : \mathcal{X} \to 2^{\mathcal{G}}$, and a positive integer k.

Question: Does there exist a rooted phylogenetic tree N on $\mathcal{X}$ that exhibits a (G, k)-labelling?

Theorem 3. *The decision problem* (G, k)-TREE *is NP-complete.*

The proofs of these two theorems are established in Section 4.

2.4. Algorithms

Despite the apparent intractability of the two problems described above, there are instances for which there exist polynomial-time algorithms. Several such instances are described in Section 5. One in particular is given next.

Let N be a phylogenetic network on $\mathcal{X}$. A sequence of vertices and arcs is an *underlying cycle* of N if it is a cycle of the underlying graph (i.e the undirected graph obtained by ignoring the directions of the arcs). A phylogenetic network N on $\mathcal{X}$ is a *galled tree* (Gusfield et al., 2004) if, for each pair C and D of underlying cycles, the vertex sets of C and D are disjoint. Each such cycle is called a *gall*. Theorem 4 shows that restricting the phylogenetic networks in GENE LABELLING to galled trees, the decision problem becomes polynomial-time solvable.

Theorem 4. *Let N be a galled tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer. Then there is a polynomial-time algorithm for deciding whether or not N exhibits a (G, k)-gene labelling.*

Theorem 4, together with the following corollary, is established in Section 5.

Corollary 1. *Let T be a rooted phylogenetic tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer. If h is a fixed positive integer, then there is a polynomial-time algorithm for deciding whether or not there is a galled tree N on $\mathcal{X}$ that can be obtained from T by adding at most h arcs and which exhibits a (G, k)-gene labelling.*

3. How many gene transfers are needed?

In this section, we prove Theorem 1.

Proof of Theorem 1: For the proof of (i), suppose that a network N admitting a (G, k)-gene labelling can be obtained by adding $\ell(T, G, k)$ arcs to T. It follows that there exists a tree T' that is a subdivision of T and a subgraph of N. In other words, T' is a topological embedding of T in N. An arc of N is said to be an *lgt-arc* if it is not an arc of T'. Consider two leaves x_1, x_2 and their lowest common ancestor v in T'. Suppose that for a gene $g \in \mathcal{G}$ we have $g \in G(x_1)$ and $g \in G(x_2)$. Since network N admits a (G, k)-gene

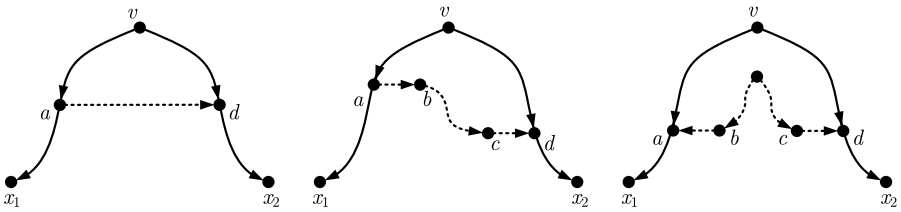

Fig. 2 Illustration for the proof of Theorem 1. The three cases apply, without loss of generality, whenever $g \in G(x_1)$, $g \in G(x_2)$, but $g \notin F(v)$, where v is the lowest common ancestor of x_1 and x_2 in T'. *Straight lines* denote arcs, while *curves* denote paths. *Solid curves* are in T', while *dotted lines/curves* can be either in T' or only in N.

labeling F, there has to be an undirected path between x_1 and x_2 in N containing only vertices u with $g \in F(u)$. Furthermore, at least one such undirected path has to consist of two directed paths, one ending in x_1 and one ending in x_2, since the subgraph of N induced by $\{v \in V \mid g \in F(v)\}$ is rooted, and hence contains a rooted tree. There are four possibilities. Firstly, it is possible that this undirected $x_1 - x_2$-path passes through v, implying that $g \in F(v)$. The remaining three cases are illustrated in Fig. 2. The first case is that the undirected $x_1 - x_2$-path uses an lgt-arc (a, d) between two vertices a, d that have v as their lowest common ancestor in T'. A second possibility is that the path uses two lgt-arcs (a, b) and (c, d) such that v is the lowest common ancestor of a and d in T' (a does not have to be on the path $v \to x_1$ as in the figure). Finally, it is also possible that the path uses two lgt-arcs (b, a) and (c, d) such that v is the lowest common ancestor of a and d in T' (other logical possibilities can be obtained from the given possibilities by relabeling vertices). Thus, for any vertex v with $n(v) > k$, there has to be either an lgt-arc (a, d) or two lgt-arcs $(a, b), (c, d)$ or two lgt-arcs $(b, a), (c, d)$, with a and d two vertices that have v as their lowest common ancestor in T'.

Given a vertex v of T, we say that an lgt-arc (s, t) *satisfies* v if v is the lowest common ancestor of s and t in T'. Since in a tree there is a unique lowest common ancestor, each single lgt-arc satisfies at most one vertex. Furthermore, we say that a pair of lgt-arcs $\{(s, t), (s', t')\}$ *satisfies* v if v is the lowest common ancestor of either s and t', or of s' and t or of t and t' in T'. It follows directly that each pair of lgt-arcs satisfies at most three vertices. Since there are $\ell(T, G, k)$ lgt-arcs, in total at most $3\binom{\ell(T,G,k)}{2} + \ell(T, G, k)$ vertices v with $n(v) > k$ can be satisfied. From the previous paragraph we know that each vertex v with $n(v) > k$ needs to be satisfied, either by a single lgt-arc or by a pair of lgt-arcs. It follows that there can be at most $3\binom{\ell(T,G,k)}{2} + \ell(T, G, k)$ vertices v with $n(v) > k$. Part (i) follows by generously bounding $3\binom{\ell(T,G,k)}{2} + \ell(T, G, k)$ by $\frac{3}{2}\ell(T, G, k)^2$.

For (ii), we can construct a network N admitting a (G, k)-gene labelling as follows. We select a set G^0 of k arbitrary genes in $\mathcal{G}$ and set $F(v) = G^0$ for each internal vertex v of T. The third property of a (G, k)-gene labelling is now satisfied for the genes in G^0. For the remaining $|\mathcal{G}| - k$ genes, we do the following. We introduce $f = \lceil \frac{|\mathcal{G}|-k}{k} \rceil$ additional isolated vertices $v_1, \ldots, v_f$ and label these vertices by disjoint sets $F(v_1), \ldots, F(v_f)$ that partition $\mathcal{G} \setminus G^0$ and contain at most k genes each. Finally, we add arcs from the root to each v_i and from each v_i to each leaf x with $G(x) \cap F(v_i) \neq \emptyset$. This leads to the claimed upper bound. □

To improve upon this, simple upper bound turns out to be challenging. This can perhaps be explained by the results in the next section, in which we show that, even if the network N is given and $k = 1$, it is NP-complete to decide if a (G, k)-gene labelling of N exists.

4. Unravelling lateral gene transfer is hard

We begin this section by first showing that GENE LABELLING is NP-complete. First, consider the following decision problem:

DIRECTED ACYCLIC SUBGRAPH HOMEOMORPHISM (DASH)

Given: Directed acyclic graphs $D = (V_D, E_D)$ and $P = (V_P, E_P)$ with $V_P \subseteq V_D$.
Question: Is P homeomorphic to a subgraph of D?

A graph P is *homeomorphic* to a graph H if H can be obtained from P by replacing arcs (u, v) by internally vertex-disjoint directed $u - v$ paths. Hence, DASH can be seen as a disjoint-paths problem. The graph P is called the "pattern graph". It was observed by Fortune et al. (1980) that NP-hardness of DASH follows from a result of Even et al. (1972) on multi-commodity flows.

Theorem 2. *The decision problem* GENE LABELLING *is NP-complete even if* $k = 1$.

Proof: The reduction is from DASH. Let (D, P) be an instance of DASH. We begin by showing that we may assume, for each vertex u in P, we have $d_P^-(u) + d_P^+(u) = 1$. To see this, let D' and P' be the digraphs obtained from D and P, respectively, by iteratively doing the following for each vertex v in P:

(i) Let $\{s_1, s_2, \ldots, s_i\}$ be the set of in-neighbours of v in P and let $\{t_1, t_2, \ldots, t_j\}$ be the set of out-neighbours of v in P.
(ii) In P, replace v and the arcs $(s_1, v), \ldots, (s_i, v)$ and $(v, t_1), \ldots, (v, t_j)$ with the new vertices $v_1, v_2, \ldots, v_{i+j}$ and the new arcs $(s_1, v_1), \ldots, (s_i, v_i)$ and $(v_{i+1}, t_1), \ldots, (v_{i+j}, t_j)$.
(iii) Let $\{x_1, x_2, \ldots, x_r\}$ be the set of in-neighbours of v in D and let $\{y_1, y_2, \ldots, y_s\}$ be the set of out-neighbours of v in D.
(iv) In D, replace v and the arcs $(x_1, v), \ldots, (x_r, v)$ and $(v, y_1), \ldots, (v, y_s)$ with the new vertices $v_1, v_2, \ldots, v_{i+j}$ and the new arcs

$$(x_1, v_1), (x_2, v_1), \ldots, (x_r, v_1), (x_1, v_2), (x_2, v_2), \ldots, (x_r, v_2),$$
$$\ldots, (x_1, v_i), (x_2, v_i), \ldots, (x_r, v_i)$$

and

$$(v_{i+1}, y_1), (v_{i+1}, y_2), \ldots, (v_{i+1}, y_s), (v_{i+2}, y_1), (v_{i+2}, y_2), \ldots, (v_{i+2}, y_s),$$
$$\ldots, (v_{i+j}, y_1), (v_{i+j}, y_2), \ldots, (v_{i+j}, y_s).$$

At the end of this iterative construction, for each vertex u in P', we have $d^-_{P'}(u) + d^+_{P'}(u) = 1$. Moreover, it is straightforward to check that P' is homeomorphic to a subgraph of D' if and only if P is homeomorphic to a subgraph of D. It now follows that we may assume that our given instance (D, P) of DASH is of the form at the completion of this construction.

We next describe a polynomial-time transformation of our instance (D, P) of DASH into an instance of GENE LABELLING with $k = 1$. Set $k = 1$. We define N, $\mathcal{X}$, $\mathcal{G}$, and the function $G : \mathcal{X} \to 2^{\mathcal{G}}$ iteratively as follows. Initially, set $\mathcal{X}$ and $\mathcal{G}$ to be both empty. Let N be the phylogenetic network obtained from $D = (V, A)$ by applying the following sequence of operations:

(O-I) For each arc $a = (u, v)$ of P, add a new gene g_a to $\mathcal{G}$, add new leaf vertices ℓ_u, ℓ_v to V and to $\mathcal{X}$, add new arcs (u, ℓ_u) and (v, ℓ_v) to A, and set $G(\ell_u) = G(\ell_v) = \{g_a\}$. Furthermore, delete all incoming arcs of u from A. At the end of (I), the constructions of the sets $\mathcal{X}$ and $\mathcal{G}$, and the function $G : \mathcal{X} \to 2^{\mathcal{G}}$ are completed.

(O-II) Repeatedly remove all leaves of the resulting network not in $\mathcal{X}$ and repeatedly remove all vertices of indegree zero that do not have an element of $\mathcal{X}$ as a child.

(O-III) Finally, root the resulting network by choosing a vertex of indegree zero as a root and then adding an arc from this root to each other vertex of indegree zero. Setting N to be the resulting phylogenetic network on $\mathcal{X}$, we have now constructed the desired instance of GENE LABELLING.

An example of this construction is shown in Fig. 3. Note that, while D may not be connected, N is connected because of (O-III). We complete the proof by showing that N admits a $(G, 1)$-gene labelling if and only if P is homeomorphic to a subgraph of D.

Suppose that P is homeomorphic to a subgraph of D. Then, for each arc $a = (u, v)$ of P, there exists a directed $u - v$ path in D such that all these directed paths are pairwise vertex disjoint. We first claim that for each such $u - v$ path in D, there exists a corresponding $u - v$ path in N. To see this, observe that, in the construction of N from D, the

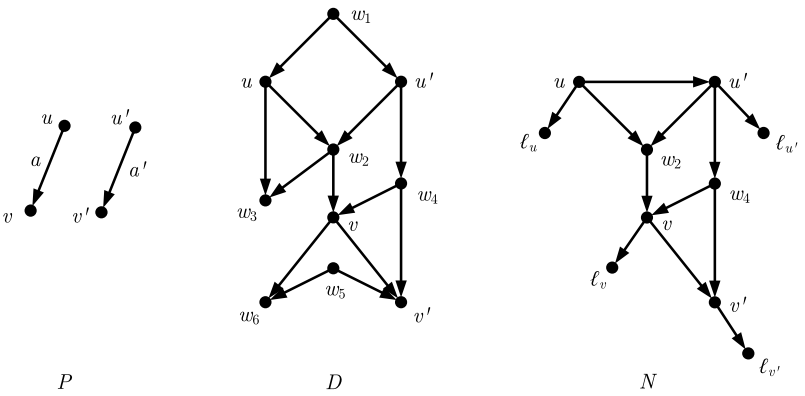

Fig. 3 An example of the reduction in the proof of Theorem 2. From an instance (P, D) of DASH, a phylogenetic network N is constructed with leaf-labelling $G(\ell_u) = G(\ell_v) = \{g_a\}$ and $G(\ell_{u'}) = G(\ell_{v'}) = \{g_{a'}\}$. Disjoint paths $u \to w_2 \to v$ and $u' \to w_4 \to v'$ in D correspond to a labelling $F(\ell_u) = F(u) = F(w_2) = F(v) = F(\ell_v) = \{g_a\}$, $F(\ell_{u'}) = F(u') = F(w_4) = F(v') = F(\ell_{v'}) = \{g_{a'}\}$.

only arcs deleted are those arcs directed into a vertex, u say, for which u is a vertex in P, and arcs incident with a vertex, w say, for which either there is no directed path from w to a vertex in $\mathcal{X}$ or there is no directed path from a parent of a vertex in $\mathcal{X}$ to w.

None of these deletions deletes an arc on any $u - v$ path in D and so the claim holds. Now, for each arc $a = (u, v)$ of P and for each vertex w on the associated $u - v$ path in N, set $F(w) = \{g_a\}$. Since the children ℓ_u of u and ℓ_v of v are the only other vertices with a label containing g_a, the subgraph of $N = (V, A)$ induced by $\{w \in V \mid g_a \in F(w)\}$ is rooted and connected. Labelling all remaining vertices w by $F(w) = \emptyset$ thus leads to a $(G, 1)$-gene labelling of N.

Now suppose that F is a $(G, 1)$-gene labelling of N. It remains to show that P is homeomorphic to a subgraph of D. Consider a gene $g_a \in \mathcal{G}$, and let $a = (u, v)$ be the associated arc of P. Since F is a $(G, 1)$-gene labelling, the subgraph of N induced by $\{w \in V(N) : g_a \in F(w)\}$ is connected. Furthermore, each of the arcs added in (O-III) in the construction of N joins two vertices that are assigned distinct genes in $\mathcal{G}$ by F as F is a $(G, 1)$-labelling of N. Thus, none of these arcs are contained in the subgraph of N induced by $\{w \in V(N) : g_a \in F(w)\}$. Since u has no other incoming arcs, u has indegree zero in this subgraph. Since the child ℓ_v of v is also labelled $F(\ell_v) = \{g_a\}$, it follows that N contains a directed path from u to v whose vertices are assigned $\{g_a\}$ under F. This path is also a directed path in D. Moreover, for two distinct genes $g_a, g_b \in \mathcal{G}$, these paths are pairwise disjoints and so they are pairwise disjoint in D. The union of these paths in D forms a subgraph H of D such that P is homeomorphic to H. This completes the proof of the theorem. □

We turn now to the proof of Theorem 3, which is based on the concepts of tree-width and tree-decomposition from graph theory – we define these notions now; for further background the interested reader may wish to consult (Diestel, 2006).

A *tree decomposition* of a graph $H = (V_H, E_H)$ is a pair $(T, \{X_i : i \in I\})$ where $T = (I, E_T)$ is a tree and, for all $i \in I$, the set X_i is a subset of V_H such that:

(i) $\bigcup_{i \in I} X_i = V_H$;
(ii) for each $(u, v) \in E_H$, there exists an $i \in I$ with $u, v \in X_i$;
(iii) for each $v \in V_H$, the subgraph of T induced by $\{i \in I : v \in X_i\}$ is connected.

The *width* of the tree decomposition is defined as $\max_{i \in I} |X_i| - 1$.

We use the following NP-complete problem for the reduction in the proof of the theorem.

TREEWIDTH

Given: An undirected graph $H = (V_H, E_H)$ and a natural number k'.
Question: Does there exist a tree decomposition of H with width at most k'?

Theorem 3. *The decision problem* (G, k)-TREE *is NP-complete.*

Proof: The reduction is from TREEWIDTH. Let (H, k') be an instance of TREEWIDTH, and set $\mathcal{X} = E_H$, $\mathcal{G} = V_H$, $G(x) = \{u, v\}$ for each edge $x = \{u, v\} \in E_H$, and $k = k' + 1$. We complete the proof by showing that there exists a tree decomposition of H with width at most k' if and only if there exists a phylogenetic tree N on $\mathcal{X}$ that admits a (G, k)-gene labelling.

Firstly, let $(T, \{X_i : i \in I\})$ be a tree decomposition of H with width k'. For each $\{u, v\} \in E_H$, there exists an $i \in I$ with $u, v \in X_i$. Hence, for each taxon $x \in \mathcal{X}$, there exists a vertex i of T with $G(x) \subseteq X_i$. We construct N from T by choosing an arbitrary vertex as a root, directing all edges away from the root and, for each $x \in \mathcal{X}$, adding a leaf x and an arc (i, x) where i is an arbitrary vertex of T with $G(x) \subseteq X_i$. Repeatedly deleting leaves not in $\mathcal{X}$, set N to be the resulting rooted phylogenetic tree on $\mathcal{X}$. We can now obtain a (G, k)-gene labelling F of N by setting $F(x) = G(x)$ for each leaf $x \in \mathcal{X}$ and $F(i) = X_i$ for each other vertex. For each gene $g \in \mathcal{G}$, the subgraph of $N = (V, A)$ induced by $\{v \in V : g \in F(v)\}$ is connected by property (iii) of a tree decomposition, and is rooted as N is a rooted phylogenetic tree.

Now suppose that there exists a phylogenetic tree N on $\mathcal{X}$ and a (G, k)-gene labelling F of $N = (V, A)$. Then we can obtain a tree decomposition $(T, \{X_i : i \in I\})$ of H by setting $I = V$ and $X_i = F(i)$ for all $i \in I$, and defining T to be the tree obtained from N by ignoring the rooting and thus orientation of each of the arcs. All properties of a tree decomposition are clearly satisfied, and the width is at most $k' = k - 1$ because $|F(i)| \leq k$ by the definition of a (G, k)-gene labelling. □

5. . . . But sometimes it is easy

Let N be a galled tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer. The main result of this section shows that there is a polynomial-time algorithm for deciding whether N exhibits a (G, k)-labelling. If N is a phylogenetic tree, then this problem is equivalent to deciding if $\ell(N, G, k) = 0$.

Proposition 1. *Let T be a phylogenetic tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer. Then there is a polynomial-time algorithm for deciding whether $\ell(T, G, k) = 0$.*

Proof: Deciding whether $\ell(T, G, k) = 0$ is equivalent to deciding if T has a (G, k)-gene labelling. With this in mind, it is easily seen that the following G-gene labelling function F of T minimizes k. For all $v \in V$, the gene $g \in \mathcal{G}$ is in $F(v)$ precisely if v is a vertex of the minimal subtree of T that connects those leaves x for which $g \in G(x)$. If $|F(v)| \leq k$ for v, then F is a (G, k)-gene labelling; otherwise there is no such gene labelling of T. □

Proposition 2 (below) establishes the main result when N has exactly one gall. We will use this proposition as the base case for an inductive proof of the main result. The proof of this proposition relies on the following construction. Let N be a galled tree on $\mathcal{X}$ with exactly one gall. Thus the undirected graph underlying N has exactly one cycle. Label (in order) the vertices of this cycle $w_1, w_2, \ldots, w_p$, where w_p is the unique vertex in N with two arcs directed into it.

Let F^* be the following map from the vertex set V of N to $2^{\mathcal{G}}$. For each $v \in V$, the gene $g \in \mathcal{G}$ is in $F^*(v)$ precisely if, ignoring the direction of the arcs, either:

(i) there is a pair of leaves x_1 and x_2 with $g \in G(x_1)$ and $g \in G(x_2)$, and v is on a path between x_1 and x_2 that avoids w_p, or
(ii) there is a pair of leaves x_1 and x_2 with $g \in G(x_1)$ and $g \in G(x_2)$, and v is on *all* paths between x_1 and x_2.

The following two observations are important for what follows. First, if F is a G-gene labelling of N, then it is easily seen that $F^*(v) \subseteq F(v)$ for all $v \in V$. Second, F^* is not necessarily a G-gene labelling of N. The exact reason for this is that there can be a gene $g \in \mathcal{G}$ such that the sub-digraph of N induced by $\{v \in V : g \in F^*(v)\}$ consists of two rooted connected components; one lying below w_p (more precisely, in the subgraph of N induced by the vertices that are reachable from w_p by a directed path) and one lying above w_p (more precisely, in the subgraph of N induced by the vertices that are not reachable from w_p by a directed path).

Now let $\mathcal{G}'$ be the subset of genes $g \in \mathcal{G}$ for which the sub-digraph of N induced by $\{v \in V : g \in F^*(v)\}$ is disconnected. We extend F^* to a G-gene labelling F of N by reformulating the problem as an undirected network flow problem and then using its solution to identify the extension. Here, one can view each edge $\{a, b\}$ as the two arcs (a, b) and (b, a). We construct an undirected graph U from N by starting with the sub-digraph of N induced by $\{w_1, w_2, \ldots, w_p\}$ and ignoring the direction of the arcs, adding a source vertex s, and, for each gene $g \in \mathcal{G}'$, adding a new vertex s_g and the three edges $\{s, s_g\}$, $\{s_g, w_{i_1-1}\}$, and $\{s_g, w_{i_2+1}\}$, where i_1 and i_2 are, respectively, the smallest and largest index $i \neq p$ for which $g \in F^*(w_i)$. Now assign each s_g capacity 1 and, for each $i \in \{1, 2, \ldots, p-1\}$, assign w_i capacity $k - |F^*(w_i)|$.

To illustrate the above construction, consider the galled tree N shown in Fig. 4(a). Each leaf x of N is labelled by the set $G(x)$ of input genes observed in the corresponding taxon. The map F^* is shown in Fig. 4(b). The undirected graph U with $k = 3$ is shown in Fig. 4(c).

Lemma 1. *There exists an integer flow in U from s to w_p with value $|\mathcal{G}'|$ if and only if there exists a (G, k)-gene labelling of N. Moreover, if there is such an integer flow, then it leads to a (G, k)-gene labelling of N.*

Proof: First suppose that there exists such a flow f with value $|\mathcal{G}'|$. Based on f, we show that there exists a (G, k)-labelling F of N. For this existence proof, we assume that we know the path that each unit of flow takes. We will conclude the proof by showing how an actual (G, k)-labelling can be constructed.

Initially set $F = F^*$. Since f has value $|\mathcal{G}'|$ and each s_g has capacity 1, there is exactly one unit of flow passing through s_g from s to w_p. Furthermore, as f is integer, it uses exactly one of the two edges $\{s_g, w_{i_1-1}\}$ and $\{s_g, w_{i_2+1}\}$. If f uses $\{s_g, w_{i_1-1}\}$, then the corresponding unit of flow either uses the vertices on the path from w_{i_1-1} to w_p through $\{w_1, w_p\}$ or the vertices on the path from w_{i_1-1} to w_p through $\{w_{p-1}, w_p\}$. Depending on which of these paths this unit of flow takes, add g to $F(w_i)$ for each of the vertices on this path. Similarly, if f uses $\{s_g, w_{i_2+1}\}$, then the corresponding unit of flow either uses the vertices on the path from w_{i_2+1} to w_p through $\{w_1, w_p\}$ or the vertices on the path from w_{i_2+1} to w_p through $\{w_{p-1}, w_p\}$. Depending on which of these paths this unit of flow takes, add g to $F(w_i)$ for each of the vertices on this path. Doing this for each $g \in \mathcal{G}'$, we claim that the resulting map $F : V \to 2^{\mathcal{G}}$ is a (G, k)-labelling of N. Clearly, F satisfies (III). Furthermore, as each vertex w_i has capacity $k - |F^*(w_i)|$, the cardinality of $F(w_i)$ is at most k. Thus F satisfies (II). It now follows that F is a (G, k)-labelling of N.

Now suppose that there exists a (G, k)-gene labelling F of N. By one of the two observations earlier, $F^*(v) \subseteq F(v)$ for all $v \in V$. Consider a gene $g \in \mathcal{G}'$. The sub-digraph of N induced by $\{v \in V : g \in F^*(v)\}$ consists of two rooted connected components. However,

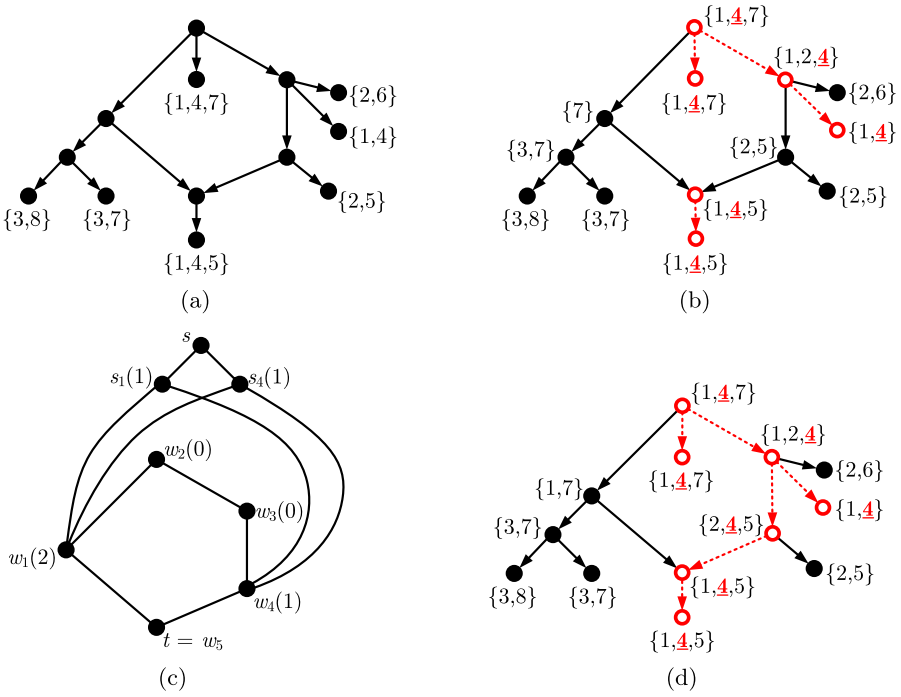

Fig. 4 (**a**) A galled tree N with one gall. Each leaf x of N is labelled by $G(x)$. (**b**) The initial labelling F^* in which, for example, the sub-digraph of N induced by $\{v \in V : 4 \in F^*(v)\}$ (displayed by the *dashed arcs* and *unfilled vertices*) consists of two connected components. (**c**) Auxiliary graph U with capacities in parentheses. (**d**) A $(G, 3)$-gene labelling of N. This gene labelling corresponds to a maximum flow in U which sends one unit of flow through s_1 and w_1 and one unit of flow through s_4 and w_4.

by (III), the sub-digraph of N induced by $\{v \in V : g \in F(v)\}$ is rooted and connected. Therefore, there is a path on the cycle consisting of vertices w_i with $g \in F(w_i) - F^*(w_i)$ that connects the two components. Sending one unit of flow from s to the first vertex on this path via s_g, and then along this path to w_p for each $g \in \mathcal{G}'$ gives the desired integer flow.

We have now shown that if there is an integer flow f from s to w_p with value $|\mathcal{G}'|$, then there is a (G, k)-labelling of N. This does not directly give such a labelling as we can make no distinction on the flow units. In particular, it is not directly clear which of the two paths a flow unit takes once it reaches a vertex w_i in the cycle. This can be rectified as follows. Let f be such a flow and let $g \in \mathcal{G}'$. Ignoring the vertices $w_{i_1}, \ldots, w_{i_2}$, either the flow unit through s_g takes the path from w_{i_1-1} to w_p via w_1 or the path from w_{i_2+1} to w_p via w_{p-1}. To make this decision, consider the following modification of the integer flow problem. Extend F^* to F^*_g by adding g to each of $F^*(w_{i_1-1}), \ldots, F^*(w_1)$ and, for each of these vertices, subtract one from their capacities. If there is an integer flow from s to w_p in $U \backslash s_g$ of $|\mathcal{G}'| - 1$ units, then we may assume that the unit of flow through s_g in U follows the path from w_{i_1-1} to w_p via w_{p-1}. In this case, replace F^* with F^*_g and U with $U \backslash s_g$, and repeat for another element in $\mathcal{G}' - g$. If there is no such integer flow in $U \backslash s_g$, then the unit of flow through s_g in U follows the path from w_{i_2+1} to w_p via w_{p-1}. In this second

case, replace F^* with that obtained by adding g to each of $F^*(w_{i_1-1}), \ldots, F^*(w_1)$ and, for each of these vertices, subtract one from their capacities, and replace U with $U \backslash s_g$. Continuing in this way, we eventually obtain a (G, k)-labelling of N. □

To illustrate Lemma 1 and its proof, consider the example prior to the lemma, illustrated in Fig. 4. In U, a maximum flow could send either two units of flow through w_1 or one unit of flow through vertex w_1 and one unit of flow through vertex w_4. From the latter option, one can for example obtain the $(G, 3)$-gene labelling shown in Fig. 4(d).

Proposition 2. *Let N be a phylogenetic network on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer.*

(i) *If N is a galled tree with exactly one gall, then there is a polynomial-time algorithm for deciding whether N exhibits a (G, k)-labelling, in which case, such a labelling can also be found in polynomial time.*
(ii) *If T is a phylogenetic tree, then there is a polynomial-time algorithm for deciding whether $\ell(T, G, k) = 1$.*

Proof: First note that a maximum-valued integer flow can be found in $O(n^{1.5} \log(n \cdot k))$ time (Goldberg and Rao, 1998). Thus, (i) follows from Lemma 1. For (ii), if $|\mathcal{X}| = n$, then there are $O(n^2)$ possible ways of adding a single arc to T. Applying Lemma 1 to each such way gives the desired algorithm. This completes the proof of the proposition. □

We now extend Proposition 2(i) to all galled trees using induction on the number of galls. Let N be a galled tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment and let k be a positive integer. If N has either no galls or exactly one gall, then we have such an algorithm by Propositions 1 and 2, so we may assume that N has at least two galls. In this case, there exists a vertex u_1 of N with the property that, for some gall, each of the vertices in the vertex set of this gall are descendants of u_1 and no vertex that is a proper descendant of u_1 has this property. Let N_1 be the phylogenetic network obtained from N by replacing u_1 and all of its descendants with a single vertex q_1. Let Q_1 be the phylogenetic network obtained from N by deleting all of the vertices of N that are not descendants of u_1 and adjoining a parent vertex r_1 to u_1 with one further child other than u_1. Call the additional child vertex v_1. Let L_{Q_1} denote the leaf set of Q_1. Effectively, we have partitioned N into two phylogenetic networks N_1 and Q_1. See Fig. 5 for an example. Let

$$G(q_1) = G(v_1) = \left(\bigcup_{x \in L_{Q_1} - \{v_1\}} G(x) \right) \cap \left(\bigcup_{x \in \mathcal{X} - L_{Q_1}} G(x) \right).$$

The proof of the following lemma is straightforward, and so the details are omitted.

Lemma 2. *The galled tree N has a (G, k)-labelling if and only if each of N_1 and Q_1 has a (G, k)-labelling.*

By Proposition 2(i), there is a polynomial-time algorithm for deciding whether or not Q_1 has a (G, k)-labelling. If there is no such labelling, then, by Lemma 2, N has no

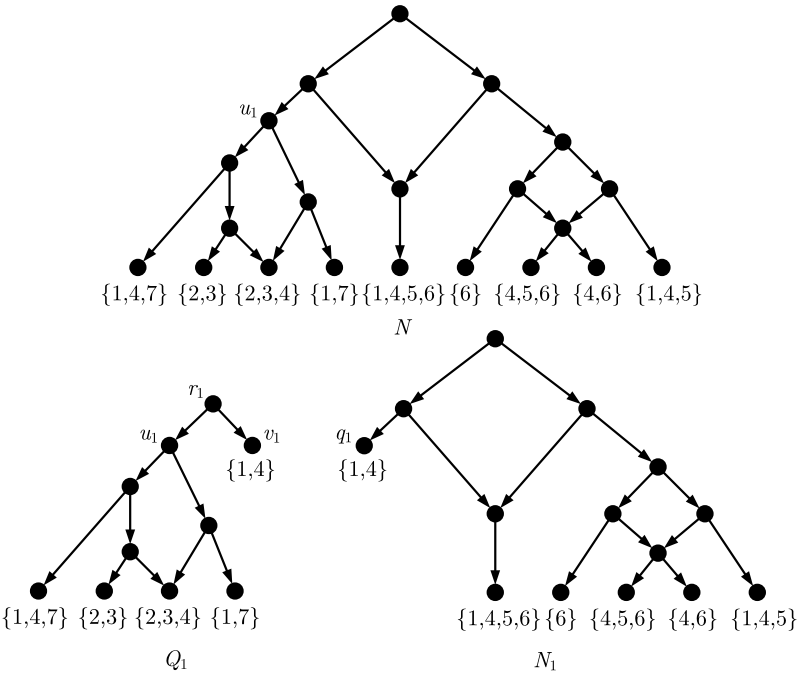

Fig. 5 A galled tree N and the decomposition of N into Q_1 and N_1 described in the text.

(G, k)-labelling. On the other hand, if Q_1 has a (G, k)-labelling, then one needs to check if N_1 has a (G, k)-labelling. Now repeat the above construction with N replaced by N_1. Continuing in this way, we either find a galled tree with a single gall that does not exhibit a (G, k)-labelling, and thereby show that N has no such labelling, or we find no such galled tree and conclude that N has a (G, k)-labelling. Note that the number of galls in N is polynomial in the size of the vertex set of N. In particular, we have established the following results.

Theorem 4. *Let N be a galled tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer. Then there is a polynomial-time algorithm for deciding whether N exhibits a (G, k)-gene labelling.*

Corollary 2. *Let T be a rooted phylogenetic tree on $\mathcal{X}$, let $\mathcal{G}$ be a set of genes, let $G : \mathcal{X} \to 2^{\mathcal{G}}$ be a genome assignment, and let k be a positive integer. If h is a fixed non-negative integer, then there is a polynomial-time algorithm for deciding whether or not there is a galled tree N on $\mathcal{X}$ that can be obtained from T by adding at most h arcs and which exhibits a (G, k)-gene labelling.*

Proof: Suppose that N is a galled tree on $\mathcal{X}$ that can be obtained from T by adding at most h arcs. Then there is an embedding T' of T in N. Notice that since N is a galled tree, it follows that all vertices of N are contained in T and thus that N can be obtained from T by subdividing at most $2h$ arcs and adding at most h arcs.

Hence, given T, we can try each possible way of subdividing at most $2h$ arcs and adding at most h arcs. For each such possibility, we check if the resulting network is a galled tree. In each such case, we can check if a (G, k)-gene labelling of this network exists, by Theorem 4. The time needed is polynomial in the size of the input, for each fixed h. □

6. Concluding comments

The analysis of this paper rests on a number of assumptions concerning gene evolution. Perhaps the most restrictive is the requirement that gene genesis is a unique event. This requirement reflects the fact that a gene is typically a long and fairly precise sequence of nucleotides, and the probability that a similar sequence could evolve independently in a different part of the tree is small. This seems reasonable if DNA sequence evolution is described by a neutral model (Kimura, 1983), but in some cases, natural selection will, no doubt, direct the evolution of DNA sequences towards certain genes that confer higher fitness. Thus, simple arguments based on neutrality need to be treated with caution. It would be interesting to extend the analysis of this paper to allow for a small frequency of independent gene genesis events.

A related question is what degree of sequence similarity is required in order to classify two sequences as coding for the same gene. Insisting on exact sequence identity is too severe, since it is well known that different species typically encode a gene with slightly different sequences that result from random site substitutions (indeed these differences have been the main signal used for phylogenetic tree reconstruction, Felsenstein, 2004). This question of gene identity is also relevant to the probability of independent gene genesis: a region of DNA that codes for a gene could, in principle, accumulate sufficient site mutations to put it just outside the range of being identified with that gene, but could then mutate back within range, giving the appearance of a second gene genesis event.

Other aspects of the model that may be criticized are the assumptions that the species tree is known with certainty (or, indeed, that it is meaningful to talk of a 'species tree', Doolittle and Bapteste, 2007), and that the model does not penalize gene losses at all.

Our computational complexity results highlight that many problems are surprisingly difficult, even for a tree, and some questions still remain to be explored further. One that seems particularly interesting is described as follows, along with our conjecture as to its possible resolution.

Given a rooted phylogenetic tree T, a set of genes $G(x)$ for each leaf x of T, and natural numbers k and h, consider the problem of deciding whether it is possible to add at most h arcs to T to obtain a phylogenetic network N that admits a (G, k)-gene labelling.

Conjecture 1. *This problem is NP-complete in general, but for each fixed h, it admits a polynomial-time algorithm.*

References

Andersson, J.O., 2005. Lateral gene transfer in eukaryotes. Cell. Mol. Life Sci. 62(11), 1182–1197.

Dagan, T., Martin, W., 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. Proc. Natl. Acad. Sci. USA 104, 870–875.

Dagan, T., Artzy-Randrup, Y., Martin, W., 2008. Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. Proc. Natl. Acad. Sci. USA 105, 10039–10044.
Diestel, R., 2006. Graph Theory, 3rd edn. Springer, Berlin.
Doolittle, W.F., Bapteste, E., 2007. Pattern pluralism and the Tree of Life hypothesis. Proc. Natl. Acad. Sci. USA 104, 2043–2049.
Doolittle, W.F., Boucher, Y., Nesbø, C., Douady, C.J., Andersson, J.O., Roger, A.J., 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? Philos. Trans. R. Soc. B 358, 39–58.
Goldberg, A.V., Rao, S., 1998. Beyond the flow decomposition barrier. J. ACM 45(5), 783–797.
Even, S., Itai, A., Shamir, A., 1972. On the complexity of timetable and multi-commodity flow problems. SIAM J. Comput. 1(2), 188–202.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Press, Sunderland.
Fortune, S., Hopcroft, J., Wyllie, J., 1980. The directed subgraph homeomorphism problem. Theor. Comput. Sci. 10, 111–121.
Gusfield, D., Eddhu, S., Langley, C., 2004. Optimal, efficient reconstruction of phylogenetic networks with constrained recombination. J. Bioinform. Comput. Biol. 2, 173–213.
Jin, G., Nakhleh, L., Snir, S., Tamir, T., 2007. Inferring phylogenetic networks by the maximum parsimony criterion: A case study. Mol. Biol. Evol. 24, 324–337.
Kimura, M., 1983. The Neutral Theory of Evolution. Cambridge University Press, Cambridge.
Mirkin, B.G., Fenner, T.I., Galperin, M.Y., Koonin, E.V., 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of lateral gene transfer in the evolution of prokaryotes. BMC Evol. Biol. 3, 2.
Spencer, M., Susko, E., Roger, A.J., 2006. Modelling prokaryote gene content. Evol. Bioinf. Online 2, 157–178.