A UNIVERSAL TREE-BASED NETWORK WITH THE MINIMUM NUMBER OF RETICULATIONS

MAGNUS BORDEWICH AND CHARLES SEMPLE

ABSTRACT. A tree-based network $\mathcal{N}$ on X is universal if every rooted binary phylogenetic X-tree is a base tree for $\mathcal{N}$. Hayamizu and, independently, Zhang constructively showed that, for all positive integers n, there exists an universal tree-based network on n leaves. For all n, Hayamizu's construction contains $\Theta(n!)$ reticulations, while Zhang's construction contains $\Theta(n^2)$ reticulations. A simple counting argument shows that an universal tree-based network has $\Omega(n \log n)$ reticulations. With this in mind, Hayamizu as well as Steel posed the problem of determining whether or not such networks exists with $O(n \log n)$ reticulations. In this paper, we show that, for all n, there exists an universal tree-based network on n leaves with $O(n \log n)$ reticulations.

1. INTRODUCTION

A *phylogenetic network* $\mathcal{N}$ *on* X is a rooted acyclic digraph with no edges in parallel satisfying the following properties:

(i) the root has out-degree two;
(ii) a vertex with out-degree zero has in-degree one, and the set of vertices with out-degree zero is X; and
(iii) all other vertices either have in-degree one and out-degree two, or in-degree two and out-degree one.

For technical reasons, if $|X| = 1$, then we additionally allow $\mathcal{N}$ to consist of the single vertex in X. As described here, a phylogenetic network is sometimes referred to as a binary phylogenetic network. Vertices of in-degree one and out-degree zero are called *leaves*, while vertices of in-degree one and out-degree two are *tree vertices* and vertices of in-degree two and out-degree one are *reticulations*. An edge directed into a reticulation is called a *reticulation edge*. All other edges are *tree edges*. A *rooted binary phylogenetic* X*-tree* is a phylogenetic network on X with no reticulations. Throughout the paper, we denote the size of X by n.

Date: March 14, 2018.
1991 *Mathematics Subject Classification.* 05C85, 92D15.
Key words and phrases. Phylogenetic network; Universal tree-based network, Base tree.

Let $\mathcal{T}$ be a phylogenetic X-tree and let $\mathcal{N}$ be a phylogenetic network on X. We say that $\mathcal{N}$ *displays* $\mathcal{T}$ if, up to suppressing degree-two vertices, $\mathcal{T}$ can be obtained from $\mathcal{N}$ by deleting edges and vertices, in which case, the resulting acyclic digraph together with a path to its root from the root of $\mathcal{N}$ is an *embedding* of $\mathcal{T}$ in $\mathcal{N}$. Note that if $\mathcal{S}$ is an embedding of $\mathcal{T}$ in $\mathcal{N}$, then the root of $\mathcal{S}$ is the root of $\mathcal{N}$ and so it may have out-degree one.

A phylogenetic network $\mathcal{N}$ on X is a *tree-based network* if there is an embedding $\mathcal{S}$ of a phylogenetic X-tree $\mathcal{T}$ in $\mathcal{N}$ such that $\mathcal{S}$ contains every vertex of $\mathcal{N}$. If this holds, then $\mathcal{S}$, as well as $\mathcal{T}$, is a *base tree* for $\mathcal{N}$. Tree-based networks were introduced by Francis and Steel [7] as a way of quantifying the concept of an 'underlying tree'. In particular, tree-based networks can be equivalently defined as those phylogenetic networks that can be obtained from a rooted binary phylogenetic tree $\mathcal{T}$ by simply adding edges whose end-vertices subdivided edges of $\mathcal{T}$. The concept of a tree-based network is relevant to the on-going debate concerning the extent to which the evolution of early life can be viewed as simply a phylogenetic tree with some additional edges or whether the concept of underlying tree is completely meaningless [1, 6]. Tree-based networks have generated considerable interest in the last year (for example, see [2, 8, 9, 12, 13]).

A tree-based network on X is *universal* if every rooted binary phylogenetic X-tree is a base tree for $\mathcal{N}$. Not all phylogenetic networks are tree-based and so, *a priori*, it is not clear whether a universal tree-based network on X exists for all positive integers n. Francis and Steel [7] showed that if $|X| \leq 3$, then there exists a universal tree-based network on X, and posed the problem of determining whether or not such a network exists for all n. Hayamizu [8] and, independently, Zhang [13] constructively showed that, for all n, there is indeed a universal tree-based network on n leaves. For all n, the construction in [8] contains $\Theta(n!)$ reticulations, while the construction in [13] contains $\Theta(n^2)$ reticulations. As a consequence of this, Hayamizu [8] and Mike Steel (private communication) asked the very natural question: What is the minimum number of reticulations in a universal tree-based network?

Let b_n denote the number of rooted binary phylogenetic X-trees. A classical result in phylogenetics dates back to Schröder (1870) who showed that

$$b_n = 1 \times 3 \times 5 \times \cdots \times (2n-3) = \frac{(2n-2)!}{(n-1)!2^{n-1}}.$$

Therefore, by Stirling's approximation,

$$b_n \sim \frac{1}{\sqrt{2}} \left(\frac{2}{e}\right)^n n^{n-1}. \tag{1}$$

Now, if $\mathcal{N}$ is a phylogenetic network on X with r reticulations, then $\mathcal{N}$ displays at most 2^r distinct rooted binary phylogenetic trees, as each embedding of such a tree is realised by choosing exactly one of the reticulation edges directed into each reticulation. Equating this with (1) and solving for r, it follows that a phylogenetic network on X that displays every

rooted binary phylogenetic X-tree has $\Omega(n \log n)$ reticulations. Hayamizu's and Steel's question is therefore more precisely stated as the following: for all positive integers n, does there exist an universal tree-based network on n leaves with $O(n \log n)$ reticulations? In this paper, we affirmatively answer this question. In particular, we establish the following theorem, the proof of which is given in the next section. Note that all logarithms in this paper are base 2.

Theorem 1.1. *For all positive integers n, there exists a universal tree-based network on n leaves with $O(n \log n)$ reticulations.*

We end the introduction with two remarks. The constructions in [8] and [13] are very similar. Intuitively, for each n, they construct a tree-based network consisting of two halves. The first half, which contains the root, embeds all tree shapes, that is, all (unlabelled) rooted binary trees. The second half, which contains the leaves, then reorders the leaves of these trees to produce the desired rooted binary phylogenetic tree. The first half constructions of both papers are identical and consists of $\Theta(n^2)$ reticulations. The difference between the two constructions lies in the second halves. Although the approaches are the same, Zhang [13] uses $\Theta(n^2)$ reticulations, while Hayamizu [8] use $\Theta(n!)$ reticulations. We adopt a similar overall approach in this paper, except that both our first and second half constructions use $O(n \log n)$ reticulations.

The second remark is that, like the universal tree-based networks in [8] and [13], the universal tree-based network constructed in this paper is both temporal and stack-free. A phylogenetic network $\mathcal{N}$ is *temporal* if there is a mapping from the set of vertices of $\mathcal{N}$ to the non-negative integers such that if (u, v) is a tree edge, then $t(u) < t(v)$, while if (u, v) is a reticulation edge, then $t(u) = t(v)$. Biologically, if $\mathcal{N}$ is temporal, then $\mathcal{N}$ satisfies two natural timing constraints. In particular, successively occurring speciation events, and contemporaneously occurring reticulation events. Lastly, a phylogenetic network is *stack-free* if it has no two reticulations one of which is a parent of the other.

2. Proof of Theorem 1.1

For all positive integers n, we first construct a phylogenetic network $\mathcal{U}_n$, which we eventually establish is an universal tree-based network on n leaves with $O(n \log n)$ reticulations. We begin by describing the structure of the top half of $\mathcal{U}_n$, which we denote by A_n. The *rooted binary caterpillar* $(1, 2, \ldots, n)$ is the rooted binary phylogenetic tree on $\{1, 2, \ldots, n\}$ such that the parents of leaves 1 and 2 are the same, the parent of leaf n is the root, and $q_n, q_{n-1}, \ldots, q_2, 1$ is a directed path from the root to leaf 1 where, for each $i \in \{2, 3, \ldots, n\}$, q_i is the parent of leaf i (see Fig. 1, solid edges). The path $q_n, q_{n-1}, \ldots, q_2, 1$ is the *spine* of the caterpillar. The digraph A_n can be viewed as the rooted binary caterpillar $(1, 2, \ldots, n)$ with additional edges whose end-vertices subdivide 'neighbouring' pendant edges of the caterpillar. More precisely, take the rooted binary caterpillar $(1, 2, \ldots, n)$ and, for each

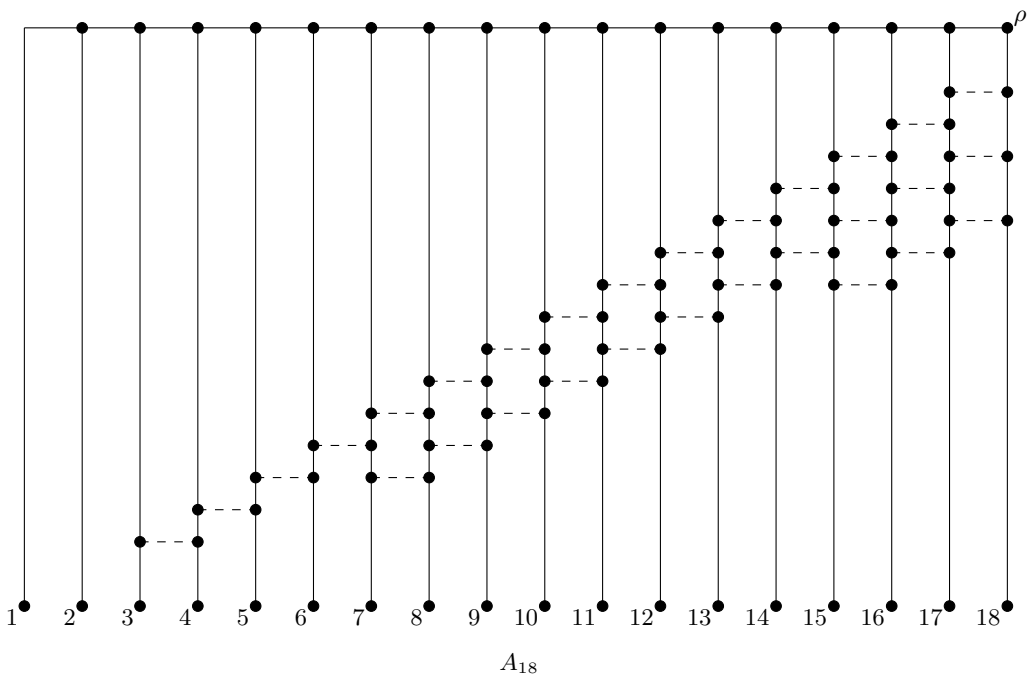

FIGURE 1. The construction A_{18}, where the root is denoted by ρ. The rooted binary caterpillar $(1, 2, \ldots, 18)$ is shown as solid edges, while dashed edges are the additional edges. Vertical edges are directed down the page, and horizontal edges are directed right to left.

$i \geq 2$, subdivide the pendant edge incident with leaf i with

$$\left\lfloor \log\left(\frac{i}{2}\right)\right\rfloor + \left\lfloor \log\left(\frac{i+1}{2}\right)\right\rfloor$$

new vertices. New edges are now added between the new vertices as follows. For each i, denote the path from the spine to leaf i by P_i. Numbering the new vertices on each P_i starting at the spine, add an edge directed from the $2j$-th new vertex on P_i to the $(2j-1)$-th new vertex on P_{i-1}. Lastly, suppress the $\left\lfloor \log\left(\frac{n+1}{2}\right)\right\rfloor$ degree-two vertices on P_n. For an example, see Fig. 1 which shows the resulting construction for when $n = 18$ where, as also in Fig. 5, vertical edges are directed down the page, and horizontal edges are directed from right to left.

We next make three observations concerning A_n:

(i) For all i, the number of new vertices subdividing P_i is a function of i only, and so the number of edges joining P_i and P_{i-1} and the placement of these edges is determined by i and not n. Thus, up to degree-two vertices on P_i, for all $i \in \{1, 2, \ldots, n\}$, the

restriction of A_n to the paths $P_1, P_2, \ldots, P_i$ and all edges joining two vertices on these paths is isomorphic to A_i.

(ii) For each i, the odd numbered new vertices on P_i are reticulations, and the even numbered new vertices on P_i are tree vertices.

(iii) The total number of reticulations in A_n is

$$\sum_{i=2}^{n} \left\lfloor \log\left(\frac{i+1}{2}\right) \right\rfloor \leq \log\left(\prod_{i=2}^{n}\left(\frac{i+1}{2}\right)\right) \leq \log\left(n^n\right) = n\log n.$$

Ignoring the leaf labels, we will later show that if T is a rooted binary tree, then there is an embedding of T in A_n using every vertex.

The bottom half of $\mathcal{U}_n$, denoted B_n, uses a *Beneš network* construction to enable any permutation of the leaves. A Beneš network [3, 4] is an example of a "rearrangeable non-blocking" network. In particular, it is an arrangement of wires (transmission links) and switches which realises all possible permutations between the input and output terminals. Each switch corresponds to a permutation of size two. The original motivation for such networks is in providing a telephone service in which it is possible to rearrange existing calls to allow any new call into the system. For further background on Beneš networks, we refer the interested reader to [10]. For the construction of B_n, we extend, for each i, the path P_i of A_n, each extension corresponds to a wire and, whenever there is a switch between wires i and j in the Beneš network, we insert a tree vertex followed by a reticulation in each corresponding path, making the new tree vertices on paths P_i and P_j the parents of the new reticulations on paths P_j and P_i, respectively. To illustrate, Fig. 2 shows (i) the construction of B_4 from the Beneš network on four wires, and (ii) the embedding of a permutation that reorders the leaves $(4, 1, 3, 2)$ coming out of A_4 to the ordering $(1, 2, 3, 4)$. Note that, for each ordering of the elements in $\{1, 2, 3, 4\}$, there is an embedding of the permutation that reorders it to the ordering $(1, 2, 3, 4)$. Furthermore, for $n = 4$, Fig. 3 shows how the two halves are combined.

Beneš networks were originally constructed for when n is a power of two. However, more recently, Beneš networks have been constructed for arbitrary n [5]. Both constructions are recursive in that the Beneš network of size n (that is, with n input and n output terminals) is built from the Beneš network of size $\left\lfloor \frac{n}{2} \right\rfloor$ and the Beneš network of size $\left\lceil \frac{n}{2} \right\rceil$. For example, the Beneš network of size 8 is constructed from two Beneš networks of size 4 as shown in Fig. 4, where each switch has been replaced with a pair of crossover edges. Moreover, both when n is a power of two and in the general case, the constructions have $O(n \log n)$ switches. Thus, our network B_n has $O(n \log n)$ reticulations and so, by construction, for all n, the phylogenetic network $\mathcal{U}_n$ has $O(n \log n)$ reticulations. Furthermore, for all n, we have that $\mathcal{U}_n$ is tree-based since we can take the vertical edges (that is, the caterpillar in A_n and the identity permutation in B_n) as a base tree. Since the embedding of any permutation contains every vertex of B_n, to complete the proof of Theorem 1.1 it suffices

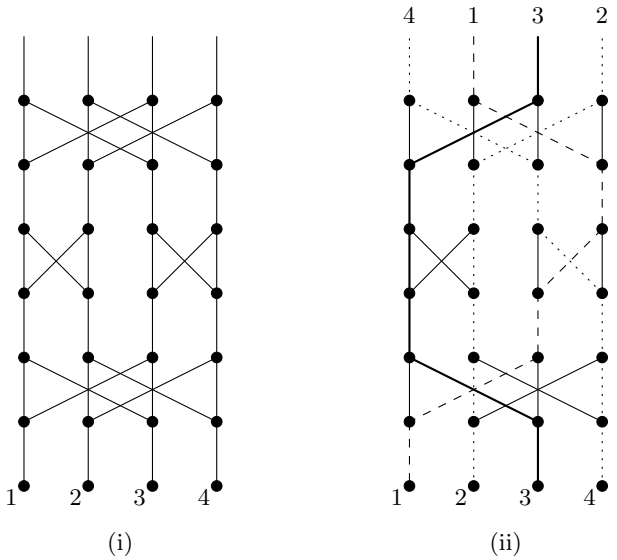

FIGURE 2. (i) The construction B_4. The vertical edges and the six pairs of "crossover" edges correspond to the four wires and the six switches of the Beneš network of size four, respectively. (ii) An embedded reordering of the permutation $(4, 1, 3, 2)$. In (i) and (ii), edges are directed down the page.

to show that, up to leaf labels, every rooted binary phylogenetic tree on $\{1, 2, \ldots, n\}$ is a base tree of A_n.

We use induction on n, to show that A_n embeds any tree shape on n leaves and that every vertex of A_n is contained in the embedding. Since there is exactly one tree shape on $n \leq 3$ leaves, the base case holds for all $n \leq 3$. Now suppose that $n \geq 4$ and that, for all $m \leq n-1$, the construction A_m embeds any tree shape with m leaves and every vertex of A_m is contained in the embedding.

Let T be a tree shape on n leaves, and denote the two maximal rooted binary subtrees whose roots are children of the root of T by T_1 and T_2. Let $t_1 = |T_1|$ and $t_2 = |T_2|$. Without loss of generality we may assume that $t_1 \geq t_2$. Note that $t_2 \leq n/2$. Consider the network, denoted A_{n,t_1}, obtained from A_n by removing all of the additional edges added between P_{t_1} and P_{t_1+1} in the construction of A_n, and also removing the first edge on each of the paths $P_{t_1+1}, P_{t_1+2}, \ldots, P_{n-1}$. The network $A_{18,10}$ is shown in Fig. 5.

Up to degree-two vertices, the root of A_{n,t_1} has two children. The first is the root of a subnetwork D_{t_1} which is isomorphic to A_{t_1} since no edge is removed between any two of the paths $P_1, P_2, \ldots, P_{t_1}$ and, as observed earlier, the restriction of A_n to the paths $P_1, P_2, \ldots, P_{t_1}$ and all edges joining two vertices on these paths is isomorphic to A_{t_1}. The second is the root of a subnetwork D_{t_2}. Now, up to isomorphism, A_{t_2} can be obtained from

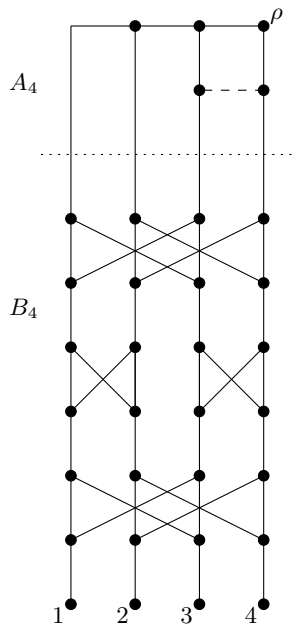

FIGURE 3. Illustrating how A_4 and B_4 combine to give $\mathcal{U}_4$. The dotted line separates the two halves. Horizontal edges are directed right to left, while all other edges are directed down the page.

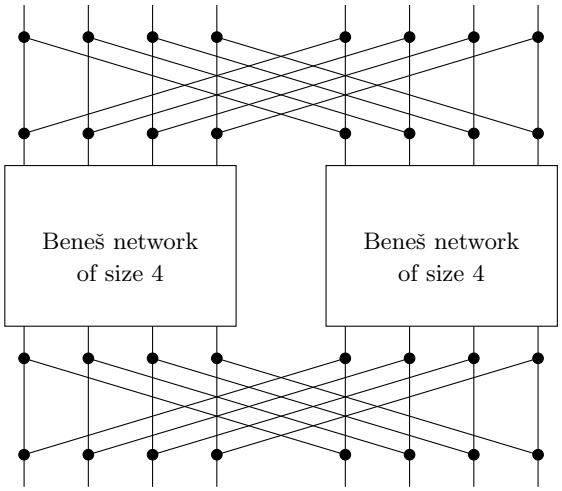

FIGURE 4. The Beneš network of size 8 constructed from two Beneš networks of size 4. Edges are directed down the page.

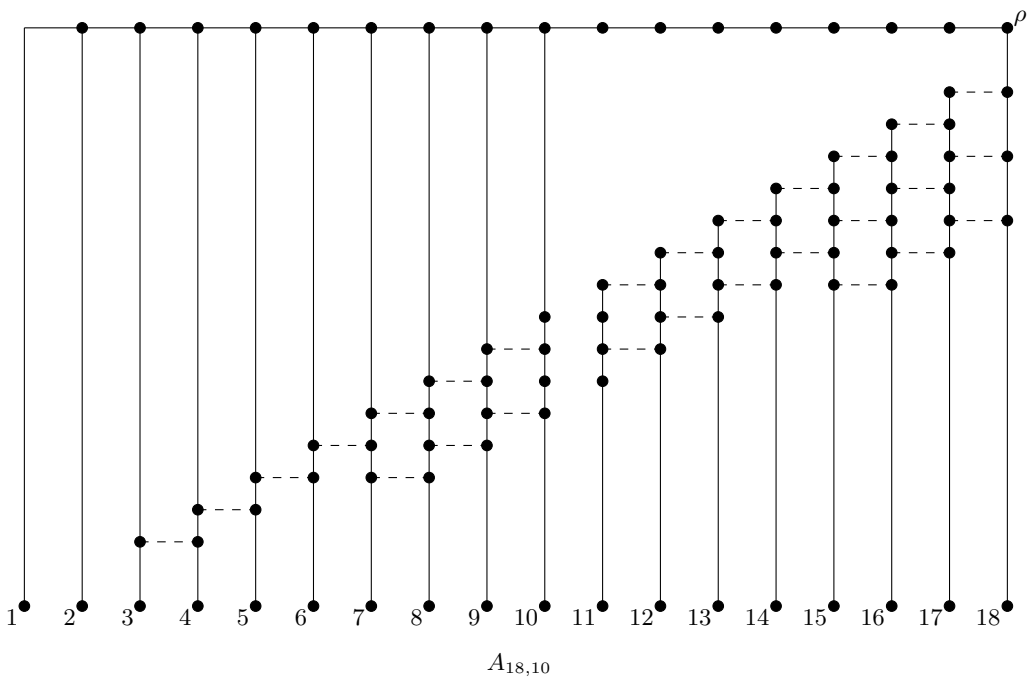

FIGURE 5. The network $A_{18,10}$, where the root is denoted by ρ. Vertical edges are directed down the page, and horizontal edges are directed from right to left.

D_{t_2} by deleting a subset of the additional edges which were added to the rooted binary caterpillar $(1, 2, \ldots, n)$ to create A_n, and suppressing degree-two vertices. To see this and, in particular, that we have enough 'additional' edges, observe that, as $t_2 \leq n/2$, the l-th leaf of D_{t_2} corresponds to at least the $2l$-th leaf of A_n. Hence, there are at least

$$\left\lfloor \log\left(\frac{2(l+1)}{2}\right)\right\rfloor - 1 = \left\lfloor \log\left(\frac{l+1}{2}\right)\right\rfloor$$

edges between the paths P'_l and P'_{l+1} of D_{t_2}, where, for all $l \in \{t_1+1, t_2+2, \ldots, n-1\}$, the path P'_i is the path obtained from P_i (in A_n) by deleting the first edge. Thus we can delete edges of D_{t_2} starting with those nearest the leaves until we obtain the correct number of edges between P'_l and P'_{l+1} for A_{t_2}. Finally, we can suppress the resulting degree-two vertices which were created in one of two ways. Firstly, by the initial deletion of edges adjacent to the spine, and are therefore now on the unique path from the root to leaf t_1+1 and, secondly, created in the later deletions, and are therefore on paths of degree-two vertices between a leaf and its first ancestor which has degree greater than two (thus on every path from the root to that leaf). Hence, if we take any embedding of T_2 in A_{t_2} which contains every vertex of A_{t_2}, it corresponds to an embedding of T_2 in D_{t_2} which contains every vertex of D_{t_2}.

By induction, D_{t_1} and D_{t_2} contain an embedding of T_1 and T_2 that uses each of its vertices, respectively. Furthermore, extending these embeddings in A_n by including the vertices in the unique paths from the root of A_n to the roots of D_{t_1} and D_{t_2} gives an embedding of T in A_n that contains every vertex of A_n. This completes the proof of Theorem 1.1.

Acknowledgements

We thank one of the anonymous referees for suggesting the simpler Beneš networks as an alternative to sorting networks for constructing the bottom half of $\mathcal{U}_n$.

References

[1] S. S. Abby, E. Tannier, M. Guy, D. Vincent, Lateral gene transfer as a support for the Tree of Life, Proc. Natl. Acad. Sci. USA 109 (2012) 4962–4967.
[2] M. Anaya, O. Anipchenko-Ulaj, A. Ashfaq, J. Chiu, M. Kaiser, M. S. Ohsawa, M. Owen, E. Pavlechko, K. St. John, S. Suleria, K. Thompson, C. Yap, On determining if tree-based networks contain fixed trees, Bull. Math. Biol. 78 (2016) 961–969.
[3] V. E. Beneš, Permutation groups, complexes, and rearrangeable multistage connecting networks, Bell System Technical Journal 43 (1964) 1619–1640.
[4] V. E. Beneš, Optimal rearrangeable multistage connecting networks, Bell System Technical Journal 43 (1964) 1641–1656.
[5] C. Chang, R. Melhem, Arbitrary size Benes networks, Parallel Processing Letters 7 (1997) 279.
[6] T. Dagan, W. F. Martin, The tree of one percent, Genome Biol. 7 (2006) 118.
[7] A. R. Francis, M. Steel, Which phylogenetic networks are merely trees with additional arcs?, Syst. Biol. 64 (2015) 768–777.
[8] M. Hayamizu, On the existence of infinitely many universal tree-based networks, J. Theor. Biol. 396 (2016) 204–206.
[9] L. Jetten, L. van Iersel, Non-binary tree-based phylogenetic networks, IEEE/ACM Trans. Comput. Biol. Bioinform., in press.
[10] F. T. Leighton, Introduction to Parallel Algorithms and Archtectures: Arrays, Trees, Hypercubes, Morgan Kaufmann Publishers, MIT, 1992.
[11] E. Schröder, Vier combinatorische problem, Zeitschrift für Mathematik und Physik, 15 (1870) 361–376.
[12] C. Semple, Phylogenetic networks with every embedded phylogenetic tree a base tree, Bull. Math. Biol. 78 (2016) 132–137.
[13] L. Zhang, On tree-based phylogenetic networks, J. Comput. Biol. 23 (2016) 553–565.

School of Engineering Computer Sciences, Durham University, Durham DH1 3LE, United Kingdom

E-mail address: m.j.r.bordewich@durham.ac.uk

School of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand

E-mail address: charles.semple@canterbury.ac.nz