

PHYLOGENETIC CLOCKS

C Semple and M Steel

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

Report Number: UCDMS2004/9

MAY 2004

PHYLOGENETIC CLOCKS

CHARLES SEMPLE AND MIKE STEEL

ABSTRACT. Graphs obtained from a binary leaf labelled ('phylogenetic') tree by adding an edge so as to introduce a cycle provide a useful representation of hybrid evolution in biology. This class of graphs (which we call 'phylogenetic clocks') also has some attractive combinatorial properties, which we present. We characterize when a set of binary phylogenetic trees is displayed by a phylogenetic clock in terms of tree rearrangement operations. This leads to a triple-wise compatibility theorem, and a simple, fast algorithm to determine clock compatibility. We also use generating function techniques to provide closed-form expressions that enumerate phylogenetic clocks with specified or unspecified cycle length.

1. INTRODUCTION

In areas of classification such as evolutionary biology and linguistics, trees and, more recently, graphs provide a useful way of representing the relationships between a set X of objects [2, 8]. In evolutionary biology, X is generally a set of extant species whose elements correspond to a distinguished subset of the vertices of the tree or graph, while the remaining vertices and edges describe speciation events and ancestral relationships between the elements of X . Traditionally, trees (both rooted and unrooted) are used for this purpose. However, there is an increasing interest in graphs that contain cycles to represent 'reticulate' evolution, as arising from biological processes that include horizontal gene transfer and the formation of hybrid species [6]. The simplest types of such a graph are those that contain a single cycle, and it is this class that we study here. To describe this further, we introduce some definitions.

A *binary phylogenetic tree (on X)* is a tree T in which every interior vertex has degree three and whose leaf set is X . The set X is often referred to as the *label set* of T and its elements as *labels*. For example, a binary phylogenetic tree is shown in Fig. 1. Here $X = \{a, b, \dots, l\}$. A *phylogenetic clock (on X)* is a graph \mathcal{G} that has exactly one cycle, every interior vertex has degree three, and the set of degree-one vertices is X . Thus, by deleting a single edge of the cycle in \mathcal{G} and suppressing the resulting degree-two vertices, we obtain a binary phylogenetic X -tree. Indeed, we say \mathcal{G} *displays* a binary phylogenetic X -tree T if T can be obtained from \mathcal{G} in this way. In general, let \mathcal{P} be a collection of phylogenetic X -trees. Then \mathcal{G} *displays*

Date: 4 May 2004.

Key words and phrases. Phylogenetic tree, compatibility, circular orderings, generating function.

We thank the New Zealand Marsden Fund (UOC310) for supporting this research.

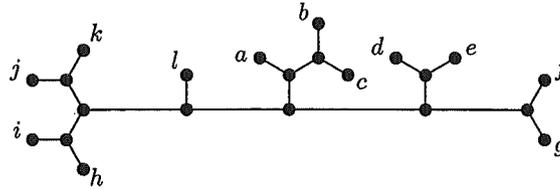


FIGURE 1. A binary phylogenetic tree.

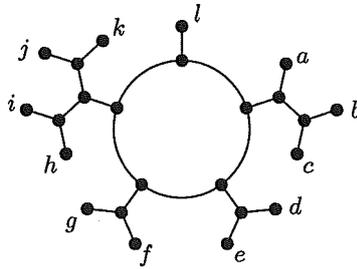


FIGURE 2. A phylogenetic clock.

\mathcal{P} if \mathcal{G} displays each tree in \mathcal{P} , in which case we say that \mathcal{P} is *clock compatible*. To illustrate these definitions, the phylogenetic clock shown in Fig. 2 displays the binary phylogenetic tree shown in Fig. 1. Two phylogenetic clocks \mathcal{G} and \mathcal{G}' on X are *isomorphic* if there is a graph isomorphism from \mathcal{G} to \mathcal{G}' which when restricted to X is the identity map.

One of the main questions that motivates this study is the following: Given a collection \mathcal{P} of binary phylogenetic trees on X when is \mathcal{P} clock compatible? For $|\mathcal{P}| = 2$, this question is closely related to tree rearrangement operations, and the number of possible phylogenetic clocks that display \mathcal{P} is either 0, 1, or 3. When \mathcal{P} has arbitrary size, the clock compatibility question can be reduced to consideration of triples of trees from \mathcal{P} , allowing for a simple polynomial-time algorithm. These results may provide a basis for biologists to move from tree-based representations of evolutionary relationships to situations where there has been a single hybridisation event. This is currently a topical problem in systematic biology, although most studies to date have dealt only with rooted trees as their input ([4, 5, 6]). In contrast, the approach described here deals with unrooted trees which are important for applications as these are typically what tree reconstruction techniques (such as neighbour joining and maximum likelihood) output. We remark here that the emphasis in this paper is on providing an attractive mathematical foundation for a simple model of reticulate evolution, rather than an algorithmic analysis of a more complex scenario.

In this paper we also consider the enumeration of phylogenetic clocks, where the cycle length is either specified or left unspecified, and we use this to derive further enumerative results. Throughout the paper, the notation and terminology follows

[8]. We end this section with some preliminaries that will be used throughout the paper.

Preliminaries. An X -split is a partition of X into two non-empty sets. We denote the X -split whose blocks are A and B by $A|B$. Associated with every phylogenetic X -tree \mathcal{T} is a particular collection of X -splits. This collection consists of those X -splits $A|B$ that are induced by the components of the graph resulting from the deletion of a single edge e of \mathcal{T} . We say that the X -split $A|B$ corresponds to e and let $\Sigma(\mathcal{T})$ denote the set of X -splits that correspond to the edges of \mathcal{T} .

Let $\pi = (x_1, x_2, \dots, x_n)$ be a cyclic permutation of X . For all $1 \leq i \leq j \leq n$, let $A_{ij} = \{x_k : i \leq k \leq j\}$ and let $\Sigma^\circ(\pi)$ denote the set

$$\Sigma^\circ(\pi) = \{A_{ij}|(X - A_{ij}) : 1 \leq i \leq j \leq n - 1\}$$

of X -splits. Arranging the elements x_1, x_2, \dots, x_n clockwise in a circle in the plane, we may view $\Sigma^\circ(\pi)$ as the set of X -splits that can be obtained by separating these elements according to which side of a line segment in the plane they lie on. Consequently, $|\Sigma^\circ(\pi)| = \binom{n}{2}$. A collection Σ of X -splits is said to be *circular* if $\Sigma \subseteq \Sigma^\circ(\pi)$ for some cyclic permutation π of X . In case $\Sigma(\mathcal{T}) \subseteq \Sigma^\circ(\pi)$ for some phylogenetic X -tree \mathcal{T} , we say that π provides a *circular ordering* for \mathcal{T} . This last definition has an equivalent formulation as follows. Suppose we embed \mathcal{T} in the plane, and trace around the outside of \mathcal{T} beginning at some leaf $x \in X$ and eventually returning to x (in this way each edge of \mathcal{T} is traversed exactly twice—once in each direction). The order in which the elements of X are met in this tracing induces a circular ordering for \mathcal{T} . The set of circular orderings for \mathcal{T} is precisely the set of orderings on X that are induced by tracing across all planar embeddings of \mathcal{T} . Similarly, we have an analogous notion of a *circular ordering* for a phylogenetic clock.

2. CLOCK COMPATIBILITY

In this section, we investigate the problem of determining precisely when a collection \mathcal{P} of binary phylogenetic X -trees is clock compatible. In the case $|\mathcal{P}| = 2$, this problem has an attractive solution in terms of tree rearrangements which we describe next. This solution will enable us to handle the case $|\mathcal{P}| \geq 3$ later in the section.

Let \mathcal{T} be a binary phylogenetic X -tree and let $e = \{u, v\}$ be an edge of \mathcal{T} . Let \mathcal{T}' be the binary phylogenetic X -tree that is obtained from \mathcal{T} by deleting e , and then attaching the component C_v that contains v to the component C_u that contains u by adjoining a new edge f from C_v to C_u so that, once degree-two vertices are suppressed, the resulting tree is a binary phylogenetic X -tree. The two tree rearrangement operations that we now describe are restricted by how this new edge is adjoined. We begin with the least restrictive operation.

- (i) We say that \mathcal{T}' has been obtained from \mathcal{T} by a *tree bisection and reconnection* (TBR) if there is no restriction on f .

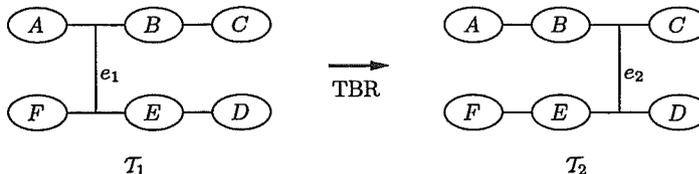


FIGURE 3. A schematic diagram of a TBR operation.

- (ii) We say that T' has been obtained from T by an (unrooted) *subtree prune and regraft* (SPR) if one end-vertex of f is v .

Observe that SPR is a special case of TBR. For further details of tree rearrangement operations, see [8].

The diagram shown in Fig. 3 is a schematic representation of a single TBR operation, where T_1 and T_2 are two binary phylogenetic X -trees. If B and E are both empty, then T_1 is isomorphic to T_2 , and so the TBR operation is redundant. Furthermore, it is easily checked that the TBR operation is an SPR operation precisely if either $|A \cup B \cup C| = 1$ or $|D \cup E \cup F| = 1$, or one of B or E is empty. We will make use of this diagram in the next section. To this end, we will make the valid assumption that, provided $|A \cup B \cup C|, |D \cup E \cup F| \geq 2$, we have $|A|, |C|, |D|, |F| \geq 1$.

Tree rearrangement operations play an important role in phylogenetics. One reason for this is that they each induce a metric on the collection of binary phylogenetic X -trees and thus enable one to quantify the ‘‘closeness’’ of any pair of such trees. In particular, let T_1 and T_2 be two binary phylogenetic X -trees and let $\Theta \in \{\text{SPR}, \text{TBR}\}$. The Θ -distance between T_1 and T_2 is the minimum number of operations that is required to transform T_1 into T_2 . We denote this distance by $d_\Theta(T_1, T_2)$. It is well-known that, for each Θ , one can always get from T_1 to T_2 by such a sequence of operations and $d_{\text{TBR}}(T_1, T_2) \leq d_{\text{SPR}}(T_1, T_2) \leq 2d_{\text{TBR}}(T_1, T_2)$.

Theorem 2.1. *Let T_1 and T_2 be two distinct binary phylogenetic X -trees. Then there is a phylogenetic clock \mathcal{G} on X that displays $\{T_1, T_2\}$ if and only if $d_{\text{TBR}}(T_1, T_2) = 1$. Moreover, in that case, there are unique edges e_1 and e_2 such that, up to suppressing degree-two vertices, $\mathcal{G} \setminus e_1$ is isomorphic to T_2 and $\mathcal{G} \setminus e_2$ is isomorphic to T_1 .*

Proof. Suppose that there is a phylogenetic clock \mathcal{G} on X that displays both T_1 and T_2 . Then, as T_1 and T_2 are distinct, it follows by definition that there are two distinct edges e_1 and e_2 such that, up to suppressing degree-two vertices, $\mathcal{G} \setminus e_1$ and $\mathcal{G} \setminus e_2$ are isomorphic to T_1 and T_2 . This implies that, for each i , T_i can be obtained from $\mathcal{G} \setminus \{e_1, e_2\}$ by adding e_i in the appropriate way. By the definition of TBR, we deduce that $d_{\text{TBR}}(T_1, T_2) = 1$.

Now suppose that $d_{\text{TBR}}(T_1, T_2) = 1$. Then, up to suppressing degree-two vertices, T_2 can be obtained from T_1 by deleting an edge e_1 say in T_1 , and then joining the resulting components by a new edge e_2 say. Now let \mathcal{G} be the graph that is obtained from T_1 by adding e_2 so that $\mathcal{G} \setminus e_1$ is isomorphic to T_2 . Since adding e_2

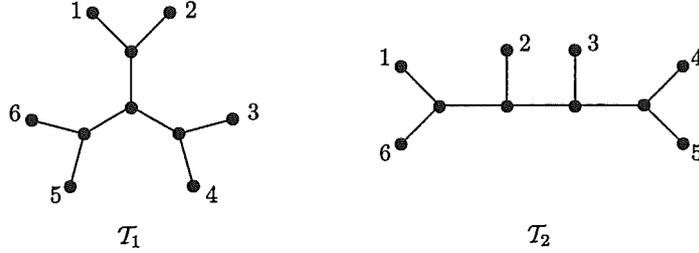


FIGURE 4. A counterexample to the converse of Proposition 2.2.

creates exactly one cycle, it follows that \mathcal{G} is a phylogenetic clock on X . Moreover, up to suppressing degree-two vertices, $\mathcal{G} \setminus e_1$ and $\mathcal{G} \setminus e_2$ are isomorphic to \mathcal{T}_1 and \mathcal{T}_2 , respectively. Thus \mathcal{G} displays \mathcal{T}_1 and \mathcal{T}_2 .

Lastly, suppose there is a phylogenetic clock \mathcal{G} on X that displays \mathcal{T}_1 and \mathcal{T}_2 . Since no two distinct edges f and f' of the cycle of \mathcal{G} have the property that $\mathcal{G} \setminus f$ is isomorphic to $\mathcal{G} \setminus f'$, it follows that the choice of e_1 and e_2 is unique. This completes the proof of the theorem. \square

Proposition 2.2. *Let \mathcal{T}_1 and \mathcal{T}_2 be two distinct binary phylogenetic X -trees. If $\{\mathcal{T}_1, \mathcal{T}_2\}$ is clock compatible, then $\Sigma(\mathcal{T}_1) \cup \Sigma(\mathcal{T}_2)$ is circular.*

Proof. Let \mathcal{G} be a phylogenetic clock on X that displays \mathcal{T}_1 and \mathcal{T}_2 . Let $x \in X$. Viewing \mathcal{G} drawn in the plane with its leaves on the outside of the cycle, trace around the outside of \mathcal{G} beginning at x , eventually returning to x . Let π be the cyclic permutation of X induced by the order in which the elements of X are met in this tracing. It is now easily checked that π is a circular ordering for both \mathcal{T}_1 and \mathcal{T}_2 , thus completing the proof of the proposition. \square

We remark here that the converse of Proposition 2.2 does not hold. For a counterexample, consider the pair of trees $\{\mathcal{T}_1, \mathcal{T}_2\}$ in Fig. 4. Then, with $\pi = (1, 2, \dots, 6)$, we have $\Sigma(\mathcal{T}_1) \cup \Sigma(\mathcal{T}_2) \subseteq \Sigma^\circ(\pi)$, and so $\Sigma(\mathcal{T}_1) \cup \Sigma(\mathcal{T}_2)$ is circular. However, $d_{\text{TBR}}(\mathcal{T}_1, \mathcal{T}_2) \geq 2$, and therefore, by Theorem 2.1, $\{\mathcal{T}_1, \mathcal{T}_2\}$ is not clock compatible.

We now consider the problem of determining precisely when an arbitrary collection of binary phylogenetic X -trees is clock compatible. To this end, we begin with the following proposition.

Proposition 2.3. *Let \mathcal{T}_1 and \mathcal{T}_2 be two binary phylogenetic trees on X , and suppose that $\{\mathcal{T}_1, \mathcal{T}_2\}$ is clock compatible. Then*

- (i) *If $d_{\text{TBR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$, but $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) \neq 1$, then there is exactly one phylogenetic clock on X that displays \mathcal{T}_1 and \mathcal{T}_2 .*
- (ii) *If $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$ and the pruned subtree consists of a single leaf, then there is exactly one phylogenetic clock on X that displays \mathcal{T}_1 and \mathcal{T}_2 .*

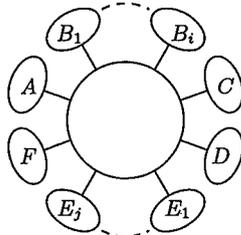


FIGURE 5. A schematic view of the phylogenetic clock described in (i) of Proposition 2.3.

- (iii) If $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$ and the pruned subtree has at least two leaves, then there are exactly three phylogenetic clocks on X that display \mathcal{T}_1 and \mathcal{T}_2 .

Proof. It follows by the definition of display that all phylogenetic clocks on X that display both \mathcal{T}_1 and \mathcal{T}_2 can be obtained by starting with \mathcal{T}_1 and adjoining a new edge e_2 . The edge e_2 is added in such a way that \mathcal{T}_2 can be obtained from the resulting phylogenetic clock on X by deleting an edge e_1 . By Theorem 2.1, there is exactly one choice for e_1 . Thus to prove the proposition, it suffices to consider the possible ways by which e_2 can be added to \mathcal{T}_1 . In establishing each of (i)–(iii), we make use of the schematic diagram of a TBR operation shown in Fig. 3. With regards to this diagram, it is clear that e_2 must join an edge of the minimal subtree of \mathcal{T}_1 that connects $A \cup B \cup C$ to an edge of the minimal subtree of \mathcal{T}_1 that connects $D \cup E \cup F$. Furthermore, as $d_{\text{TBR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$ or $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$, we have $|X| \geq 4$.

First consider (i). Since $d_{\text{TBR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$, but $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) \neq 1$, we may assume that $|A|, |B|, |C|, |D|, |E|, |F| \geq 1$ in Fig. 3. By noting that $A|(X - A), C|(X - C), D|(X - D), F|(X - F)$ are all X -splits of \mathcal{T}_2 , this added edge cannot be joined to edges in any of the subtrees labelled A, C, D , and F . Furthermore, as $(A \cup B)|(X - (A \cup B))$ and $(E \cup F)|(X - (E \cup F))$ are both X -splits of \mathcal{T}_2 , this added edge cannot be joined to edges in B or E . It now follows that there is exactly one way in which e_2 can be appropriately added to \mathcal{T}_1 . Thus there is exactly one phylogenetic clock on X that displays both \mathcal{T}_1 and \mathcal{T}_2 . This phylogenetic clock is schematically shown in Fig. 5, where B_1, \dots, B_i are the subtrees of B attached to the path from e_1 to e_2 , and E_1, \dots, E_j are the subtrees of E attached to the path from e_2 to e_1 .

Now consider (ii). Without loss of generality, we may assume that, in Fig. 3, $|A| = 1$, and B and C are both empty. Using an approach similar to that in (i), it is easily seen that in this case there is also exactly one phylogenetic clock on X that displays both \mathcal{T}_1 and \mathcal{T}_2 .

Lastly, consider (iii). In this case, as $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$ and the pruned subtree has at least two leaves, precisely one of B or E is empty, and $|A|, |C|, |D|, |F| \geq 1$. Without loss of generality, we may assume that B is empty, in which case E is non-empty. Again using the approach used in (i), we deduce, in this case, that there are exactly three phylogenetic clocks on X that display both \mathcal{T}_1 and \mathcal{T}_2 . These three

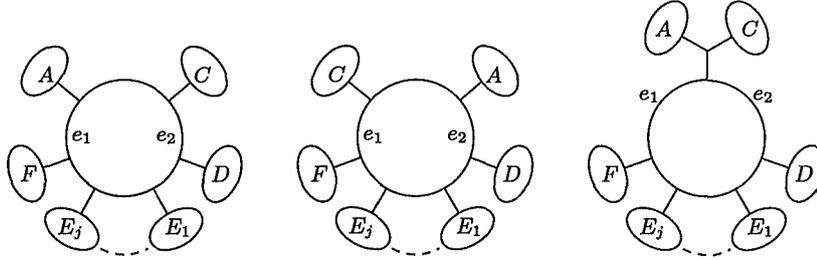


FIGURE 6. A schematic view of the phylogenetic clocks described in (iii) of Proposition 2.3.

phylogenetic clocks are schematically shown in Fig. 6. This completes the proof of the proposition. \square

Theorem 2.4. *Let \mathcal{P}' be a collection of binary phylogenetic trees on X with $|\mathcal{P}'| \geq 3$. Then \mathcal{P}' is clock compatible if and only if, for all subsets \mathcal{P} of size three, \mathcal{P} is clock compatible, in which case there is a unique phylogenetic clock on X that displays \mathcal{P}' .*

Proof. If there is a phylogenetic clock \mathcal{G} on X that displays \mathcal{P}' , then every 3-element subset of \mathcal{P}' is displayed by \mathcal{G} . This proves one direction of the theorem.

For the converse, suppose that \mathcal{P} is clock compatible for every 3-element subset \mathcal{P} of \mathcal{P}' . First assume that there is a pair \mathcal{T}_1 and \mathcal{T}_2 in \mathcal{P}' such that either the assumptions of (i) or (ii) in the statement of Proposition 2.3 hold. In either case, it follows by Proposition 2.3 that there is exactly one phylogenetic clock, \mathcal{G} say, on X that displays \mathcal{T}_1 and \mathcal{T}_2 . Since \mathcal{G} is unique and every 3-element subset of \mathcal{P}' is clock compatible, we now deduce that, for each $i \in \{3, 4, \dots, k\}$, there is exactly one phylogenetic clock that displays $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_i\}$ and that this tree is always \mathcal{G} . Hence, in this case, \mathcal{P}' is clock compatible and there is a unique phylogenetic clock on X that displays \mathcal{P}' .

Now assume that, for every pair of trees in \mathcal{P}' , the assumptions of (iii) in Proposition 2.3 hold. Let \mathcal{T}_1 and \mathcal{T}_2 be a pair of trees in \mathcal{P}' . Then, by Proposition 2.3, there are exactly three phylogenetic clocks, $\mathcal{G}_1, \mathcal{G}_2$, and \mathcal{G}_3 say, on X that display \mathcal{T}_1 and \mathcal{T}_2 . Now consider $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_i\}$, where $\mathcal{T}_i \notin \{\mathcal{T}_1, \mathcal{T}_2\}$. By assumption, there is a phylogenetic clock on X that displays $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_i\}$. Moreover, this tree must be one of the three phylogenetic clocks that display \mathcal{T}_1 and \mathcal{T}_2 . For each $j \in \{1, 2, 3\}$, it follows by Theorem 2.1 that, up to degree-two vertices, there is a unique pair of edges in \mathcal{G}_j such that the deletion of one results in \mathcal{T}_1 and the deletion of the other results in \mathcal{T}_2 . By considering the remaining edges of the cycles of $\mathcal{G}_1, \mathcal{G}_2$, and \mathcal{G}_3 , it is straightforward to deduce that the binary phylogenetic X -trees that result by deleting such an edge are distinct. This implies that there is exactly one phylogenetic clock on X that displays $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_i\}$. If, for all i , the phylogenetic clock displaying $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_i\}$ is the same, then \mathcal{P}' is clock compatible and this phylogenetic clock on X is the only such clock. Therefore assume that for some distinct

i and j , the phylogenetic clock that displays $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_i\}$ is not isomorphic to the phylogenetic clock that displays $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_j\}$. We may also assume that the former clock is \mathcal{G}_1 and the latter clock is \mathcal{G}_2 . By an argument similar to that used earlier in this paragraph, there is a unique phylogenetic clock that displays $\{\mathcal{T}_1, \mathcal{T}_i, \mathcal{T}_j\}$. Since \mathcal{G}_1 displays $\{\mathcal{T}_1, \mathcal{T}_i\}$, we deduce that it is \mathcal{G}_1 . But \mathcal{G}_1 does not display \mathcal{T}_j ; a contradiction. This completes the proof of the theorem. \square

The sufficient part of the hypothesis in Theorem 2.4 is sharp in the sense that it is not sufficient for \mathcal{P}' to be clock compatible if every subset of \mathcal{P}' of size two is clock compatible. To see this, take \mathcal{P}' to be the collection consisting of all three binary phylogenetic X -trees, where $|X| = 4$. Then it is easily checked that each of the three 2-element subsets of \mathcal{P}' are clock compatible. However, the union of the X -splits of the trees in \mathcal{P}' is not circular and so, by the contrapositive of Proposition 2.2, \mathcal{P}' is not clock compatible.

Theorem 2.1, Proposition 2.3, and Theorem 2.4 provide the basis and validity for the following polynomial-time algorithm for determining the clock compatibility of a collection of binary phylogenetic X -trees. We leave the formal details to the reader.

Algorithm: CLOCKCOMPATIBILITY(\mathcal{P}, \mathcal{G})

Input: A collection \mathcal{P} of binary phylogenetic X -trees.

Output: A phylogenetic clock \mathcal{G} on X that displays \mathcal{P} or the statement \mathcal{P} is not clock compatible.

1. Choose any two trees \mathcal{T}_1 and \mathcal{T}_2 in \mathcal{P} .
2. Decide whether or not $d_{\text{TBR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$.
 - (a) If no, then halt and return \mathcal{P} is not clock compatible.
 - (b) If yes, then construct a phylogenetic clock \mathcal{G} on X that displays \mathcal{T}_1 and \mathcal{T}_2 . In the case $d_{\text{SPR}}(\mathcal{T}_1, \mathcal{T}_2) = 1$ and the pruned subtree has at least two leaves, construct all three phylogenetic clocks \mathcal{G}_1 , \mathcal{G}_2 , and \mathcal{G}_3 on X that display \mathcal{T}_1 and \mathcal{T}_2 .
3. Select another tree $\mathcal{T}_3 \in \mathcal{P}$.
 - (a) If exactly one phylogenetic clock is constructed in the previous step, then check to see whether or not \mathcal{G} displays \mathcal{T}_3 . If not, then halt and return \mathcal{P} is not clock compatible.
 - (b) If three phylogenetic clocks are constructed in the previous step, then check to see whether or not \mathcal{G}_1 , \mathcal{G}_2 , or \mathcal{G}_3 displays \mathcal{T}_3 . (At most one such tree has this property.) If not, then halt and return \mathcal{P} is not clock compatible.
4. Let \mathcal{G} denote the phylogenetic clock that displays $\{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$. For each $\mathcal{T}_i \in \mathcal{P} - \{\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3\}$, check to see whether or not \mathcal{G} displays \mathcal{T}_i . If not, then halt and return \mathcal{P} is not clock compatible. Otherwise return \mathcal{G} .

3. COUNTING PHYLOGENETIC CLOCKS

In this section, we use generating functions to derive the following exact expressions for the number of distinct phylogenetic clocks on a fixed set X .

Theorem 3.1. *Let X be a finite set of size $n \geq 3$.*

(i) *Let $c(n)$ denote the number of phylogenetic clocks on X . Then*

$$c(n) = (n-1)!2^{n-2} - \frac{(2n-2)!}{(n-1)!2^{n-1}}.$$

(ii) *For each $k \geq 3$, let $c(n, k)$ denote the number of phylogenetic clocks on X whose unique cycle is of length k . Then*

$$c(n, k) = \frac{(2n-k-1)!}{(n-k)!2^{n-k+1}}.$$

In proving Theorem 3.1, we make use of the following notation: for a power series $f(x)$, we let $[x^n]f(x)$ denote the coefficient of x^n in $f(x)$.

For $|X| \geq 2$, a *rooted binary phylogenetic X -tree* is a rooted tree whose root has degree two and every other interior vertex has degree three, and whose leaf set is X . If $|X| = 1$, then the tree consisting of a single-root vertex labelled by the element in X is a rooted binary phylogenetic X -tree. For all $n \geq 1$, let $r(n)$ denote the number of rooted binary phylogenetic trees on a set X of size n . For each $n \geq 2$, the number $r(n)$ is given by

$$(1) \quad r(n) = \frac{(2n-2)!}{(n-1)!2^{n-1}} = 1 \times 3 \times \cdots \times (2n-3),$$

a well-known result that dates back to 1870 [7].

For establishing Theorem 3.1, it will be convenient for us to consider one particular way in which $r(n)$ can be derived. Let

$$R(x) = \sum_{n \geq 1} r(n) \frac{x^n}{n!}$$

be the exponential generating function for $r(n)$. Now notice that if we delete the root of a binary phylogenetic tree that has $n \geq 2$ leaves, we obtain an unordered pair of rooted phylogenetic binary trees for which the numbers of labelled leaves in the resulting pair of trees sum to n . Since the labels can be distributed freely between these two trees, it follows that, for all $n \geq 2$,

$$r(n) = \frac{1}{2} \sum_{i=1}^{n-1} \binom{n}{i} r(i)r(n-i).$$

This expression for $r(n)$ translates into the more succinct equation

$$(2) \quad R(x) = \frac{1}{2}R(x)^2 + x.$$

The term “ $+x$ ” in (2) accounts for the case where we have just a single isolated root vertex. If we regard (2) as a quadratic equation (in $R(x)$), and choose the root whose power series has non-negative coefficients, we get

$$(3) \quad R(x) = 1 - \sqrt{1-2x}.$$

Now, for all $n \geq 1$,

$$[x^n](1 - \sqrt{1-2x}) = \frac{r(n)}{n!}.$$

Therefore, as $r(n) = n![x^n]R(x)$, we obtain (1).

We now introduce two further exponential generating functions. Let

$$C(x) = \sum_{n \geq 3} c(n) \frac{x^n}{n!}$$

and, for all $k \geq 3$, let

$$C_k(x) = \sum_{n \geq 3} c(n, k) \frac{x^n}{n!}$$

denote the exponential generating functions for $c(n)$ and $c(n, k)$, respectively, where $n \geq 3$. Both these generating functions are closely related to $R(x)$. In particular,

$$(4) \quad c(n, k) = \frac{1}{2k} \sum_{(n_1, \dots, n_k): n_1 + \dots + n_k = n} \frac{n!}{n_1! \dots n_k!} \prod_{i=1}^k r(n_i).$$

To justify the right-hand side of (4), first note that the term $\frac{n!}{n_1! \dots n_k!}$ counts the number of ways of assigning the n elements of X into k sets of size n_1, \dots, n_k , and the term $\prod_{i=1}^k r(n_i)$ is the number of choices of rooted binary phylogenetic trees that have specified leaf sets of sizes n_1, \dots, n_k where, for each i , $n_i \geq 1$. However, each phylogenetic clock with cycle length k generates exactly $2k$ such k -tuples of rooted binary phylogenetic trees, since we have k choices for which tree starts the cycle, and there are two directions that the cycle can be traversed. Equation 4 means that we may write $C_k(x)$ much more elegantly as

$$(5) \quad 2C_k(x) = \frac{1}{k} R(x)^k.$$

Since $C(x) = \sum_{k \geq 3} C_k(x)$, it follows by (5) that the following relationship between $C(x)$ and $R(x)$ holds:

$$(6) \quad 2C(x) = \frac{1}{3} R(x)^3 + \frac{1}{4} R(x)^4 + \dots$$

Using the identity

$$-\log(1-t) = t + \frac{1}{2}t^2 + \frac{1}{3}t^3 + \dots,$$

we can rewrite (6) as

$$(7) \quad C(x) = \frac{1}{2} \left(-R(x) - \frac{1}{2}R(x)^2 - \log(1 - R(x)) \right).$$

Replacing the term $\log(1 - R(x))$ in (7) by $\log(\sqrt{1 - 2x}) (= \frac{1}{2} \log(1 - 2x))$ as allowed by (3), and then the remaining term in (7), namely $-R(x) - \frac{1}{2}R(x)^2$, by $x - 2R(x)$ as allowed by (2), we get

$$C(x) = \frac{1}{2}x - R(x) - \frac{1}{4} \log(1 - 2x).$$

The expression for $c(n)$ in the statement of Theorem 3.1 now follows by routine manipulation. This establishes part (i).

To prove part (ii), we first evaluate $[x^n]R(x)^k$. Notice that one can write $R(x) = x\phi(R(x))$ for the function $\phi(x) = (1 - \frac{1}{2}x)^{-1}$. In such a situation, there is a

convenient tool for extracting $[x^n]R(x)^k$ known as the *Lagrange inversion formula* (see [3]). Applying this formula here (as was similarly applied in [1]), we obtain

$$[x^n]R(x)^k = \frac{1}{n}[\lambda^{n-1}]k\lambda^{k-1}\phi(\lambda)^n = \frac{k}{n}[\lambda^{n-k}](1 - \frac{1}{2}\lambda)^{-n} = \frac{k}{n} \binom{2n-k-1}{n-k} 2^{k-n}.$$

Therefore, by (5),

$$c(n, k) = n! \cdot \frac{1}{2k} [x^n]R(x)^k = \frac{(2n-k-1)!}{(n-k)!2^{n-k+1}}.$$

This establishes part (ii).

We end this section with the following consequence of Theorem 3.1 for which we recall the definition of a circular ordering of a phylogenetic clock from the introduction.

Corollary 3.2. *Let X be a finite set of size $n \geq 3$.*

- (i) *Let \mathcal{G} be a phylogenetic clock on X whose unique cycle has length k . Then the number of distinct circular orderings for \mathcal{G} is 2^{n-k+1} .*
- (ii) *Let π be a cyclic permutation of X . Then the number of phylogenetic clocks on X whose cycle has length k and for which π is a circular ordering is*

$$\binom{2n-k-1}{n-1}$$

Proof. To prove (i), we first note that a binary phylogenetic tree with m leaves, where $m \geq 3$, has precisely 2^{m-2} circular orderings (see, for example, [8]). Now let m_1, m_2, \dots, m_k denote the number of elements of X that appear (as leaves) on the k subtrees that are incident with the k vertices of the cycle in the phylogenetic clock \mathcal{G} . Then, as the cycle of \mathcal{G} can be traversed in two directions, it is now straightforward to see that the number of circular orderings for \mathcal{G} is

$$2 \prod_{i=1}^k 2^{(m_i+1)-2} = 2^{n-k+1}.$$

This establishes (i).

For the proof of (ii), let $c(n, k, \pi)$ denote the number of phylogenetic clocks on X whose unique cycles each have length k and for which π is a circular ordering. To evaluate $c(n, k, \pi)$, we will count the number of ordered pairs (\mathcal{G}, π) , where \mathcal{G} is phylogenetic clock on X whose unique cycle has length k and π is a circular ordering for \mathcal{G} . We do this count in two ways. Firstly, by Theorem 3.1(ii), there are

$$\frac{(2n-k-1)!}{(n-k)!2^{n-k+1}}$$

phylogenetic clocks whose unique cycle has length k . Furthermore, for each such clock, there are precisely 2^{n-k+1} circular orderings, by part (i). Hence the number of ordered pairs (\mathcal{G}, π) is

$$\frac{(2n-k-1)!}{(n-k)!}$$

Alternatively, we can calculate this number by noting that the number of cyclic permutations on X is $(n-1)!$ and, for every such cyclic permutation π , the number of phylogenetic clocks on X whose unique cycle has length k and for which π is a circular ordering is $c(n, k, \pi)$. Equating these two counts, we deduce (ii). \square

REFERENCES

- [1] M. Carter, M. D. Hendy, D. Penny, L. A. Székely, and N. C. Wormald, On the distribution of lengths of evolutionary trees, *SIAM J. Discrete Math.* 3 (1990) 38–47.
- [2] J. Felsenstein, *Inferring Phylogenies*, Sinauer Press 2004.
- [3] I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, John Wiley and Sons, New York 1983.
- [4] P. Legendre, Biological Applications of Reticulate Analysis, *J. Classification* 17 (2000) 191–195.
- [5] L. Nakhleh, T. Warnow, and C. Randal Linder, Reconstructing reticulate evolution in species - theory and practice, in: *Proceedings of the 8th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 2004, in press.
- [6] F. J. Rohlf, Phylogenetic models and reticulations, *J. of Classification* 17 (2000) 185–189.
- [7] E. Schröder, Vier combinatorische probleme, *Zeitschrift für Mathematik und Physik* 15 (1870) 361–376.
- [8] C. Semple and M. Steel, *Phylogenetics*, Oxford Uni. Press, 2003.
- [9] C. Semple and M. Steel, Cyclic permutations and evolutionary trees, *Adv. in Appl. Math.* 32 (2004) 669–680.

BIOMATHEMATICS RESEARCH CENTRE, DEPARTMENT OF MATHEMATICS AND STATISTICS, UNIVERSITY OF CANTERBURY, CHRISTCHURCH, NEW ZEALAND

E-mail address: c.semple@math.canterbury.ac.nz, m.steel@math.canterbury.ac.nz