Phylogenetic Closure Operations and Homoplasy-Free Evolution

Tobias Dezulian[1] and Mike Steel[2]

[1] University of Tübingen, Germany
dezulian@informatik.uni-tuebingen.de
[2] University of Canterbury, New Zealand
M.Steel@math.canterbury.ac.nz

Abstract: Phylogenetic closure operations on partial splits and quartet trees turn out to be both mathematically interesting and computationally useful. Although these operations were defined two decades ago, until recently little had been established concerning their properties. Here we present some further new results and links between these closure operations, and show how they can be applied in phylogeny reconstruction and enumeration. Using the operations we study how effectively one may be able to reconstruct phylogenies from evolved multi-state characters that take values in a large state space (such as may arise with certain genomic data).

1 Phylogenetic Closure Operations

Meacham (1983), building on the earlier work of Estabrook and McMorris (1977) and others, described two formal rules for deriving new phylogenetic relationships from pairs of compatible characters. These rules are particularly simple and appealing, and we describe them in the following subsections. There are interesting mathematical relationships between these rules, and we develop and those and indicate new results that support their use in modern phylogenetic modeling.

1.1 Rules for Partial Splits

In this paper a *partial X-split* refers to a partition of some subset of X into two disjoint nonempty subsets, say A and B, and we denote this by writing $A|B$ $(= B|A)$. Also we say that a phylogenetic X-tree $\mathcal{T}$ *displays* a partial X-split $A|B$ if there exists at least one edge of $\mathcal{T}$ whose deletion from $\mathcal{T}$ separates the leaves labeled by the species in A from the species in B. This concept is illustrated in Fig. 1.

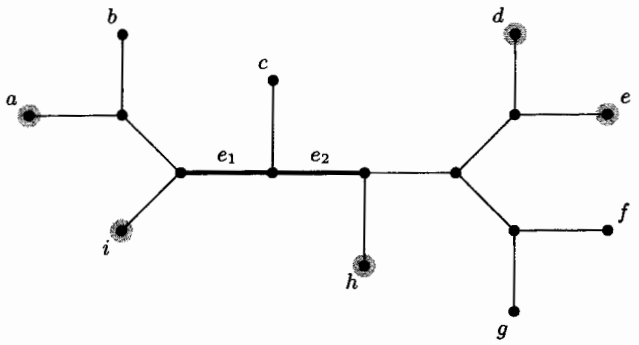


Fig. 1. The phylogenetic X–tree shown above displays the partial X–split $\{a,i\}|\{d,e,h\}$. Deleting edge e_1 or e_2 separates $\{a,i\}$ from $\{d,e,h\}$.

Given a collection Σ of partial X–splits and a partial X–split $A|B$, we write $\Sigma \vdash A|B$ if every phylogenetic X–tree that displays each partial X–split in Σ also displays $A|B$. Let $A_1|B_1$ and $A_2|B_2$ be two partial X-splits. Meacham's two rules can be stated as follows.

(**M1**): If $A_1 \cap A_2 \neq \emptyset$ and $B_1 \cap B_2 \neq \emptyset$ then

$$\{A_1|B_1, A_2|B_2\} \vdash A_1 \cap A_2|B_1 \cup B_2, A_1 \cup A_2|B_1 \cap B_2.$$

(**M2**): If $A_1 \cap A_2 \neq \emptyset$ and $B_1 \cap B_2 \neq \emptyset$ and $A_1 \cap B_2 \neq \emptyset$ then

$$\{A_1|B_1, A_2|B_2\} \vdash A_2|B_1 \cup B_2, A_1 \cup A_2|B_1.$$

The underlying theorem (Meacham, 1983) that justifies these two rules is the following:

> ANY PHYLOGENETIC X–TREE THAT DISPLAYS THE PARTIAL X–SPLITS ON THE LEFT OF (**M1**) OR (**M2**) ALSO DISPLAYS THE CORRESPONDING PARTIAL X–SPLITS ON THE RIGHT.

1.2 Rules for Quartet Trees

At around the same time as Meacham's paper, researchers in stemmatology in Holland, building on the earlier pioneering work of Colonius and Schulze (1981), described rules for combining quartet trees. In particular, Marcel Dekker in his MSc thesis (1986) investigated two 'dyadic' rules, which take as their input two quartet trees and produce one or more output quartet trees.

Standard terminology refers to a fully-resolved phylogenetic tree on four leaves as a *quartet tree* and we write it as $ab|cd$ if the interior edge separates the pair of leaves a, b from the pair of leaves c, d. Also we say that a phylogenetic X–tree $\mathcal{T}$ *displays* the quartet tree $ab|cd$ if there is at least one interior edge of

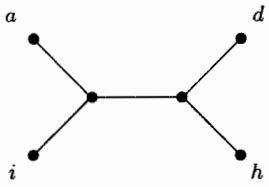


Fig. 2. A quartet tree $ai|dh$ that is displayed by the tree in Fig.1.

$\mathcal{T}$ that separates the pair a, b from the pair c, d. These concepts are illustrated in Fig.1 and Fig.2. For any phylogenetic X–tree $\mathcal{T}$ we let $\mathcal{Q}(\mathcal{T})$ denote the set of all quartet trees that are displayed by $\mathcal{T}$.

Returning now to quartet rules, for a set $\mathcal{Q}$ of quartet trees we write $\mathcal{Q} \vdash ab|cd$ precisely if every phylogenetic tree that displays $\mathcal{Q}$ also displays $ab|cd$. We call the statement $\mathcal{Q} \vdash ab|cd$ a *quartet rule*, and it is *dyadic* if $|\mathcal{Q}| = 2$. There are precisely two dyadic quartet rules:

(**Q1**): $\{ab|cd, ab|ce\} \vdash ab|de$ and

(**Q2**): $\{ab|cd, ac|de\} \vdash ab|ce, ab|de, bc|de$.

The underlying (and easily proved) theorem that justifies these two rules is the following:

> ANY PHYLOGENETIC X–TREE THAT DISPLAYS THE QUARTET TREES ON THE LEFT OF (**Q1**) OR (**Q2**) ALSO DISPLAYS THE CORRESPONDING QUARTET TREE(S) ON THE RIGHT.

Thus for a set $\mathcal{Q}$ of quartet trees we may form the *dyadic closure* of $\mathcal{Q}$ subject to either or both rules. More precisely, for $\theta \subseteq \{1, 2\}$, let qcl_θ denote the minimal set of quartet trees that contains $\mathcal{Q}$ and is closed under rule (**Qi**) for each $i \in \theta$. In practice $\mathrm{qcl}_\theta(\mathcal{Q})$ can be obtained from $\mathcal{Q}$ by constructing a sequence

$$\mathcal{Q} = \mathcal{Q}_1 \subseteq \mathcal{Q}_2 \subseteq \cdots$$

where $\mathcal{Q}_{i+1}$ consists of $\mathcal{Q}_i$ together with all additional quartets that can be obtained from a pair of quartets in $\mathcal{Q}_i$ by applying the rule(s) allowed by θ. Then $\mathrm{qcl}_\theta(\mathcal{Q})$ is just $\mathcal{Q}_i$ for the first index i for which $\mathcal{Q}_{i+1} = \mathcal{Q}_i$. Note that the sequence $\mathcal{Q}_i$ is uniquely determined by $\mathcal{Q}$ and $\mathrm{qcl}_\theta(\mathcal{Q})$ is the minimal subset of quartet trees containing $\mathcal{Q}$ that is closed under the rule(s) in θ.

The construction of $\mathrm{qcl}_\theta(\mathcal{Q})$ is useful for the following reasons. First, if $\mathcal{Q}$ is incompatible, we may discover this by finding a pair of contradictory quartets (of the form $ab|cd, ac|bd$ in $\mathrm{qcl}_\theta(\mathcal{Q})$). Second, we may find that $\mathrm{qcl}_\theta(\mathcal{Q})$ consists of all the quartets of a tree, from which we can thereby not only verify that $\mathcal{Q}$ is compatible, but easily construct a tree that displays $\mathcal{Q}$. This is precisely what Dekker found for some of his trees describing the copying history of manuscripts.

Dekker showed that there exist quartet 'rules' of order 3 that could not be reduced to repeated applications of the two dyadic rules. Dekker also found irreducible rules of order 4 and 5, leading to his conjecture that there exist irreducible quartet rules of arbitrarily large order, a result subsequently established in Bryant and Steel (1995). Similar phenomena occur with split closure rules, for which there also exist higher order rules for combining partial X–splits, and indeed a third order rule was described by Meacham (1983).

Since dyadic quartet rules are provably incomplete, why then should we bother with them? Two reasons seem compelling. First, in an area where most interesting questions are NP-complete (cf. Ng, Steel,and Wormald, 2000; Steel, 1992), computing dyadic quartet closure can, reassuringly, be carried out in polynomial-time. Second, there are now several sufficient conditions known where quartet closure will yield all the quartets of a tree. We shall describe one of these now, after recalling some terminology.

Definitions: We say that a set $\mathcal{Q}$ of quartet trees *defines a phylogenetic X–tree* $\mathcal{T}$ precisely if $\mathcal{T}$ is the only phylogenetic X–tree that displays each quartet tree in $\mathcal{Q}$. In this case it is easily shown that $\mathcal{T}$ must be fully resolved, and $|\mathcal{Q}| - (|X| - 3) \geq 0$; if in addition we have $|\mathcal{Q}| - (|X| - 3) = 0$ then we say that $\mathcal{Q}$ is *excess-free*. □

Theorem 1. *Suppose $\mathcal{Q}$ is a set of quartet trees and $\mathcal{Q}$ contains an excess-free subset that defines a phylogenetic tree $\mathcal{T}$. Then* $\mathrm{qcl}_2(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$.

The only known proof of this result, in Böcker et al. (2000), is an easy consequence of the 'patchwork' theory in Böcker (1999) and Böcker, Dress and Steel (1999). This in turn is based on one of the most mysterious and apparently difficult results in phylogenetics, which concerns the proof an innocuous-looking yet powerful theorem: *Any set $\mathcal{Q}$ of two or more quartet trees that is excess-free and defines a phylogenetic tree is the disjoint union of two non-empty, excess-free subsets.*

In Section 2 we shall provide another sufficient condition under which dyadic quartet closure, in this case using rule (**Q1**), will yield all the quartets of any fully resolved tree.

1.3 The (Almost) Happy Marriage

The reader may now be wondering what, if any, connection exists between the rules that Meacham described and those of Dekker. It turns out there is a very close (but not exact) correspondence between these operations, and to explain this we introduce some further notation.

Definitions: We say that a partial split $A'|B'$ *refines* another partial split $A|B$ precisely if $A \subseteq A'$ and $B \subseteq B'$ (or $A \subseteq B'$ and $B \subseteq A'$) and we denote this by writing $A|B \preceq A'|B'$. Note that $\preceq$ is a partial order on the set $\Sigma(X)$ of all partial X–splits, and if $A|B \preceq A'|B'$ but $A|B \neq A'|B'$, then as is usual we shall write $A|B \prec A'|B'$. Given a set Σ of partial X–splits, we define the *reduction* of Σ to be the set $\rho(\Sigma)$ of partial X–splits defined by:

$$\rho(\Sigma) = \{A|B \in \Sigma : \text{ there is no } A'|B' \in \Sigma \text{ for which } A|B \prec A'|B'\}.$$

□

We should note that a set of partial X–splits Σ conveys no more phylogenetic information than does its reduction. This is due to the following, easily established result: The set of phylogenetic X–trees that displays each split in Σ is identical to the set of phylogenetic X–trees that displays each split in $\rho(\Sigma)$. Therefore it will be convenient for us to reduce whenever we feasible to ensure that the sets of partial X–splits do not become excessively large.

Definitions: For a set Σ of partial X–splits and $\theta \subseteq \{1, 2\}$ let

$$W_\theta(\Sigma) := \{\Sigma' \subseteq \Sigma(X) : \Sigma \subseteq \Sigma' \text{ and } \Sigma' \text{ is closed under } (\mathbf{Mi}) \text{ for all } i \in \theta\}.$$

Notice that $W_\theta(\Sigma) \neq \emptyset$ since $\Sigma(X) \in W_\theta(\Sigma)$, and if $\Sigma_1, \Sigma_2 \in W_\theta(\Sigma)$ then $\Sigma_1 \cap \Sigma_2 \in W_\theta(\Sigma)$. Thus the set $\cap W_\theta(\Sigma)$ $(= \cap\{\Sigma : \Sigma \in W_\theta(\Sigma)\}$ is well-defined, and it is the (unique) minimal set of partial X–splits that contains Σ and which is also closed under Meacham's rules (**Mi**) for all $i \in \theta$. Finally, let

$$\text{spcl}_\theta(\Sigma) := \rho(\cap W_\theta(\Sigma)).$$

Thus $\text{spcl}_\theta(\Sigma)$ is the reduction of the set $\cap W_\theta(\Sigma)$. □

We can construct $\text{spcl}_\theta(\Sigma)$ by repeatedly applying the rules allowed by θ to construct a sequence $\Sigma = \Sigma_1 \subseteq \Sigma_2 \subseteq \cdots$ until the sequence stabilizes. At this point (as well as at any point along the sequence that we wish) we can then apply reduction. This construction was suggested by Meacham (1983) and our aim here is to establish some of its useful features.

Example 1. Let

$$\Sigma = \{\{a, b\}|\{c, d\}, \{a, b\}|\{c, e\}, \{a, c\}|\{d, e\}, \{b, c\}|\{d, e\}\}.$$

Then $\text{spcl}_1(\Sigma) = \{\{a, b\}|\{c, d, e\}, \{a, b, c\}|\{d, e\}\}$. □

Example 2. Let

$$\Sigma = \{\{b, f\}|\{e, g\}, \{a, f\}|\{d, g\}, \{a, e\}|\{c, d\}, \{a, e\}|\{b, c\}, \{a, d\}|\{c, g\}\}.$$

Then $\text{spcl}_2(\Sigma) = \{\{c, g\}|\{a, b, d, e, f\}, \{b, f\}|\{a, c, d, e, g\}, \{a, e\}|\{b, c, d, f, g\}, \{a, b, e, f\}|\{c, d, g\}\}$. □

Notice that rule (**M2**) has the property that the derived partial splits refine the two input partial splits. Consequently

$$|\text{spcl}_2(\Sigma)| \leq |\Sigma|.$$

Rule (**M1**) does not have this property. Note also that $\text{spcl}_{1,2}(\Sigma)$ is not necessarily equal to $\text{spcl}_1(\text{spcl}_2(\Sigma))$ or to $\text{spcl}_2(\text{spcl}_1(\Sigma))$.

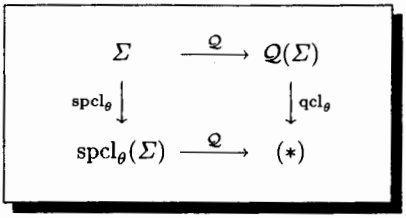
For a set Σ of partial X–splits, let $\mathcal{Q}(\Sigma)$ denote the set of *induced* quartet trees, defined by

$$\mathcal{Q}(\Sigma) := \{aa'|bb' : a, a' \in A, b, b' \in B, A|B \in \Sigma\}.$$

A phylogenetic X–tree displays the partial X–splits in Σ if and only if $\mathcal{T}$ displays the quartet trees in $\mathcal{Q}(\Sigma)$. Note that $\mathcal{Q}(\Sigma) = \mathcal{Q}(\rho(\Sigma))$ and so

$$\mathcal{Q}(\text{spcl}_\theta(\Sigma)) = \mathcal{Q}(\cap W_\theta(\Sigma)). \tag{1}$$

Given a set Σ of quartet splits we may construct the quartet closure (under rules in θ) of $\mathcal{Q}(\Sigma)$, or we may construct the induced quartets of the split closure (under rules in θ) of Σ. By either route we derive a collection of quartet trees from a collection of partial X–splits. A fundamental questions is: When are the resulting sets of quartet trees identical? In other words, when does the following diagram commute?



The answer to this question is given in the following result.

Theorem 2. *Let Σ be a collection of partial X–splits. Then*

$$\text{qcl}_\theta(\mathcal{Q}(\Sigma)) = \mathcal{Q}(\text{spcl}_\theta(\Sigma))$$

for $\theta = \{1\}$ and $\theta = \{1, 2\}$. For $\theta = \{2\}$ we have

$$\text{qcl}_\theta(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\text{spcl}_\theta(\Sigma))$$

and containment can be strict.

Theorem 2 ensures that $\mathcal{Q}(\text{spcl}_1(\Sigma))$ and $\mathcal{Q}(\text{spcl}_{1,2}(\Sigma))$ can both be computed in polynomial time, which is not obvious since $\text{spcl}_1(\Sigma)$ and $\text{spcl}_{1,2}(\Sigma)$ could presumably be very large.

Proof of Theorem 2: First note that

$$\text{qcl}_2(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\text{spcl}_2(\Sigma)) \tag{2}$$

by Semple and Steel (2001). And Example 2 shows that containment can be strict, for in this example $ab|cd \in \mathcal{Q}(\text{spcl}_2(\Sigma)) - \text{qcl}_2(\mathcal{Q}(\Sigma))$. Note that containment can be strict even when, as in Example 2, we have $\mathcal{Q}(\text{spcl}_2(\Sigma)) = \mathcal{Q}(\mathcal{T})$ for a tree $\mathcal{T}$ defined by Σ.

Next we show that

$$\mathcal{Q}(\mathrm{spcl}_1(\Sigma)) = \mathrm{qcl}_1(\mathcal{Q}(\Sigma)). \tag{3}$$

To do so, we first show that

$$\mathcal{Q}(\mathrm{spcl}_1(\Sigma)) \subseteq \mathrm{qcl}_1(\mathcal{Q}(\Sigma)). \tag{4}$$

Consider the sequence $\Sigma = \Sigma_0, \Sigma_1, \Sigma_2, \ldots, \Sigma_N = \mathrm{spcl}_1(\Sigma)$, where each Σ_{i+1} is obtained from Σ_i by using (**M1**)whenever applicable on any pair of splits in Σ_i.

We prove by induction on i that $\mathcal{Q}(\Sigma_i) \subseteq \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$, which suffices to establish (4). For $i = 0$, we have $\mathcal{Q}(\Sigma_0) = \mathcal{Q}(\Sigma) \subseteq \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$. Now suppose that the induction hypothesis holds for i where $0 \leq i < N$. For the induction step let $q \in \mathcal{Q}(\Sigma_{i+1})$. If $q \in \mathcal{Q}(\Sigma_i)$, then $q \in \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$ as claimed. Thus we may suppose that $q \in \mathcal{Q}(\Sigma_{i+1}) - \mathcal{Q}(\Sigma_i)$. In this case, referring to (**M1**) we may suppose, without loss of generality, that q is induced by the split $A_1 \cup A_2 | B_1 \cap B_2$. In this case we may further suppose (since we assume $q \notin \mathcal{Q}(\Sigma_i)$) that $q = a_1a_2|b_1b_2$, where $a_1 \in A_1 - A_2$, $a_2 \in A_2 - A_1$, and $b_1, b_2 \in B_1 \cap B_2$. Furthermore, there exists an element $a \in A_1 \cap A_2$ for spcl_1 to be applicable and yield q, and thus $a_1a|b_1b_2 \in \mathcal{Q}(\Sigma_i)$ and $a_2a|b_1b_2 \in \mathcal{Q}(\Sigma_i)$. The induction hypothesis yields that $a_1a|b_1b_2 \in \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$ and $a_2a|b_1b_2 \in \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$. Application of (**Q1**) to the latter two quartets yields $q = a_1a_2|b_1b_2 \in \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$, which establishes the induction step and thereby the proof of (4).

We now establish the reverse inclusion, namely:

$$\mathrm{qcl}_1(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\mathrm{spcl}_1(\Sigma)). \tag{5}$$

Consider the sequence of quartet sets $\mathcal{Q}(\Sigma) = \mathcal{Q}_0, \mathcal{Q}_1, \mathcal{Q}_2, \ldots, \mathcal{Q}_N = \mathrm{qcl}_1(\mathcal{Q}(\Sigma))$, where each $\mathcal{Q}_{i+1}$ is obtained from $\mathcal{Q}_i$ by using (**Q1**) whenever applicable on any pair of quartets in $\mathcal{Q}_i$.

We prove by induction on i that $\mathcal{Q}_i \subseteq \mathcal{Q}(\mathrm{spcl}_1(\Sigma))$, thus establishing the theorem. For $i = 0$, we have $\mathcal{Q}_0 = \mathcal{Q}(\Sigma) \subseteq \mathcal{Q}(\mathrm{spcl}_1(\Sigma))$. Now suppose that the induction hypothesis holds for some i where $0 \leq i < N$. Let $q \in \mathcal{Q}_{i+1}$. If $q \in \mathcal{Q}_i$ then $q \in \mathcal{Q}(\mathrm{spcl}_1(\Sigma))$ as claimed, so we may assume that $q \in \mathcal{Q}_{i+1} - \mathcal{Q}_i$. In this case, without loss of generality, $q = ab|cd$ and there is some x and two quartet trees $q_1 = ab|cx$, $q_2 = ab|dx \in \mathcal{Q}_i$, so that one application of (**Q1**) yields q. By the induction hypothesis $q_1, q_2 \in \mathcal{Q}(\mathrm{spcl}_1(\Sigma))$. Let $\sigma_1 = \{a, b, a'_1, a'_2, \cdots, a'_j\} | \{c, x, b'_1, b'_2, \cdots, b'_l\} \in \mathrm{spcl}_1(\Sigma)$ be a split from which q_1 is derived, and let $\sigma_2 = \{a, b, a''_1, a''_2, \cdots, a''_m\} | \{d, x, b''_1, b''_2, \cdots, b''_n\} \in \mathrm{spcl}_1(\Sigma)$ be a split from which q_2 is derived. Note that $\sigma_1 \neq \sigma_2$ due to $q \in \mathcal{Q}_{i+1} - \mathcal{Q}_i$. Application of (**M1**) to σ_1 and σ_2 yields $\sigma_3 \in \mathrm{spcl}_1(\Sigma)$ where

$$\begin{aligned}\sigma_3 = \{a, b\} \cup (\{a'_1, a'_2, \cdots, a'_j\} \cap \{a''_1, a''_2, \cdots, a''_m\}) \\ | \{c, d, x, b'_1, b'_2, \cdots, b'_l, b''_1, b''_2, \cdots, b''_n\}.\end{aligned}$$

Since $q \in \mathcal{Q}(\{\sigma_3\}) \subseteq \mathcal{Q}(\mathrm{spcl}_1(\Sigma))$, this establishes (5). And combining (4) and (5) establishes (3).

We now show that

$$\mathcal{Q}(\text{spcl}_{1,2}(\Sigma)) = \text{qcl}_{1,2}(\mathcal{Q}(\Sigma)). \tag{6}$$

To do so we first establish that

$$\text{qcl}_{1,2}(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\text{spcl}_{1,2}(\Sigma)). \tag{7}$$

Construct a sequence of quartet sets $\mathcal{Q}(\Sigma) = \mathcal{Q}_0, \mathcal{Q}_1, \ldots, \mathcal{Q}_N = \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$ where $\mathcal{Q}_{i+1} = \text{qcl}_{p(i)}(\mathcal{Q}_i)$, and where

$$p(i) = \begin{cases} 1 & \text{if } i \text{ is odd;} \\ 2 & \text{if } i \text{ is even;} \end{cases}$$

We use induction to show that for each i we have $\text{qcl}_{p(i)}(\mathcal{Q}_i) \subseteq \mathcal{Q}(\text{spcl}_{1,2}(\Sigma))$. The case $i = 0$ is clear, and the inductive step for i odd follows the argument used to establish (5), while the inductive step for i even follows the argument used in Semple and Steel (2001) to establish (2).

It remains to establish the reverse inclusion, namely:

$$\mathcal{Q}(\text{spcl}_{1,2}(\Sigma)) \subseteq \text{qcl}_{1,2}(\mathcal{Q}(\Sigma)). \tag{8}$$

Consider the sequence $\Sigma = \Sigma_0, \Sigma_1, \Sigma_2, \ldots, \Sigma_N = \text{spcl}_{1,2}(\Sigma)$, where each Σ_{i+1} is obtained from Σ_i by using **(M1) and (M2)**whenever applicable on any pair of splits in Σ_i.

We use induction on i to show that $\mathcal{Q}(\Sigma_i) \subseteq \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$ holds for any i between 0 and N, thus establishing the claim. For $i = 0$, $\Sigma_0 = \Sigma$, and so $\mathcal{Q}(\Sigma_i) \subseteq \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$. Now suppose that the induction hypothesis holds for i where $0 \leq i < N$. For the induction step let $q \in \mathcal{Q}(\Sigma_{i+1})$. If $q \in \mathcal{Q}(\Sigma_i)$ then $q \in \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$, as claimed, so we may suppose that $q \in \mathcal{Q}(\Sigma_{i+1}) - \mathcal{Q}(\Sigma_i)$.

Now q can only have been derived by either using **(M1)**, in which case we refer to the argument used to establish (4) to show that $q \in \text{qcl}_1(\mathcal{Q}(\Sigma)) \subseteq \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$, or by using **(M2)**, which we consider now. For $\text{spcl}_{1,2}$ to be applicable on Σ_i and yield q, there exist two splits $\sigma_1 = A_1|B_1$ and $\sigma_2 = A_2|B_2$ with $\sigma_1, \sigma_2 \in \Sigma_i$ and elements x, y, z with $x \in A_1 \cap A_2$, $y \in A_2 \cap B_1$, $z \in B_1 \cap B_2$. Furthermore, without loss of generality, $q = a_1a_2|b_1b_2$, where $a_1 \in A_1 - A_2$ and $b_1, b_2 \in B_2$. Note that x, y, z are distinct, since otherwise σ_1 or σ_2 would contain an identical element on both sides of the split. Also note that $x \neq a_1$.

The following cases arise:

(I) $a_2 \in A_2$.

(a) $z \notin \{b_1, b_2\}$.

Consider the quartets $q_1 = a_1x|zy, q_2 = yx|zb_1, q_3 = yx|zb_2$ obtained from σ_1 and σ_2, and also the quartet $q_7 = a_2x|b_1b_2$. By the induction hypothesis, $\{q_1, q_2, q_3, q_7\} \subseteq \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$. Application of **(Q2)** on $\{q_1, q_3\}$ and $\{q_1, q_2\}$ produces $q_4 = a_1x|zb_2$ and $q_5 = a_1x|zb_1$, respectively, and one application of **(Q1)** on $\{q_4, q_5\}$ yields $q_6 = a_1x|b_1b_2$. Finally, an application of **(Q1)** on $\{q_6, q_7\}$ yields $q = a_1a_2|b_1b_2 \in \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$.

(b) $z = b_1$.
Proceed as in argument (a), to define and obtain q_1, q_3, q_4 and q_7. Since $q_4 = a_1x|b_1b_2$, application of (**Q2**) on $\{q_4, q_7\}$ yields $q = a_1a_2|b_1b_2 \in \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$.

(c) $z = b_2$.
Proceed symmetrically to argument (b).

(II)$a_2 \in A_1 - A_2$.
Proceed as in argument (I), and similarly define and obtain $q_1, \ldots, q_6 \in \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$ with $q_6 = a_1x|b_1b_2$. In contrast to argument (I), obtain $q_7 = a_2x|b_1b_2$ by the same line of argument used to obtain q_6, taking advantage of the symmetry of a_1 and a_2. As in (I), one application of (**Q1**) on $\{q_6, q_7\}$ yields $q = a_1a_2|b_1b_2 \in \text{qcl}_{1,2}(\mathcal{Q}(\Sigma))$.

Combining (7) and (8) establishes (6) and thereby the theorem. □

1.4 The Extended Family: Characters and Trees

The operations described above, on partial X–splits and on quartet trees, are not confined to these seemingly specialized inputs. Indeed they apply easily to more familiar phylogenetic objects—namely characters and trees. We pause to describe this connection here as it will be useful in Section 4.

Given a sequence $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ of partitions of X, which we shall refer to as (qualitative, unordered) *characters*, one can associate with $\mathcal{C}$ a set $\Sigma(\mathcal{C})$ of partial X–splits and a set $\mathcal{Q}(\mathcal{C})$ of quartet trees, defined as follows:

$$\Sigma(\mathcal{C}) = \{A|B : A, B \in \chi_i, \text{ for some } i \in \{1, \ldots k\}\}$$
$$\mathcal{Q}(\mathcal{C}) = \mathcal{Q}(\Sigma(\mathcal{C})).$$

Similarly, given a collection $\mathcal{P}$ of phylogenetic X–trees on overlapping leaf sets we may associate a set $\Sigma(\mathcal{P})$ of partial X–splits and a set $\mathcal{Q}(\mathcal{P})$ of quartet trees, defined as follows:

$$\Sigma(\mathcal{P}) = \cup_{\mathcal{T} \in \mathcal{P}} \Sigma(\mathcal{T}), \text{ and } \mathcal{Q}(\mathcal{P}) = \cup_{\mathcal{T} \in \mathcal{P}} \mathcal{Q}(\mathcal{T}).$$

The following result (Steel, 1992) shows that the phylogenetic compatibility of characters (or of trees with overlapping leaf sets) can be completely transformed into questions involving either partial X–splits or quartet trees.

Proposition 1. *Let $\mathcal{T}$ be a phylogenetic X–tree, $\mathcal{C}$ a collection of characters on X, and $\mathcal{P}$ a collection of phylogenetic trees whose leaf sets are subsets of X. The following are equivalent:*

- *$\mathcal{T}$ displays the characters in $\mathcal{C}$ (respectively the trees in $\mathcal{P}$).*
- *$\mathcal{T}$ displays the partial X–splits in $\Sigma(\mathcal{C})$ (respectively the partial X–splits in $\Sigma(\mathcal{P})$).*
- *$\mathcal{T}$ displays the quartet trees in $\mathcal{Q}(\mathcal{C})$ (respectively the quartet trees in $\mathcal{Q}(\mathcal{P})$).*

2 Meacham's First Rule Yields All Quartets from a Generous Cover

In this section we show that a sufficiently 'rich' subset of quartets from a fully resolved phylogenetic tree $\mathcal{T}$ suffices for the reconstruction of $\mathcal{T}$ by the quartet rule (**Q1**). We begin by introducing some terminology.

For any two vertices x, y of an X-tree $\mathcal{T}$, let *path set* $p(x,y) := \{x = v_1, v_2, .., v_i = y\}$ denote the set of all vertices traversed by the path from x to y in $\mathcal{T}$ and let *length* $l(x,y) := |p(x,y)| - 1$ denote the number of edges traversed by this path. The path between two inner vertices u, v of a phylogenetic tree $\mathcal{T}$, is said to be *distinguished* by a resolved quartet tree $ab|cd$ precisely if $p(a,b) \cap p(u,v) = \{u\}$ and $p(c,d) \cap p(u,v) = \{v\}$. A collection $\mathcal{Q}$ of quartet trees is a *generous cover* of $\mathcal{T}$ if $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ and if, for all pairs of interior vertices u, v of $\mathcal{T}$, there exists a quartet tree $ab|cd \in \mathcal{Q}$ that distinguishes the path uv.

In Mossel and Steel (2003) it was shown that if $\mathcal{Q}$ is a generous cover of $\mathcal{T}$ then $\mathcal{T}$ is the only tree that displays $\mathcal{T}$ and, furthermore, $\mathcal{T}$ can be reconstructed from $\mathcal{Q}$ by a polynomial-time algorithm. The aim of this section is to show a further result, namely that if $\mathcal{Q}$ is a generous cover of $\mathcal{T}$ then $\mathrm{qcl}_1(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$. First, however, we establish a definition and two lemmas.

Definition: An unordered pair $\{x, y\}$ of distinct leaves of a tree are said to form a *cherry* precisely if x and y are adjacent to a common vertex.

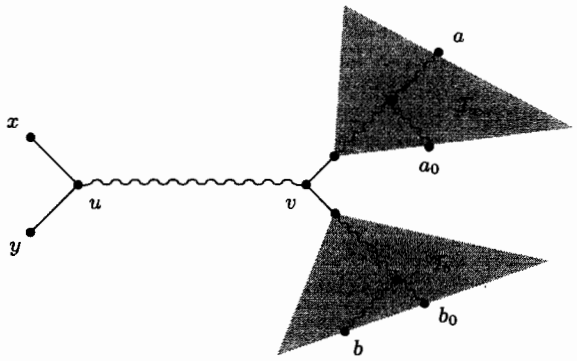


Fig. 3. Quartet $xy|a_0b_0$ distinguishes path uv.

Lemma 1. *Let $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ be a generous cover for a binary phylogenetic X-tree $\mathcal{T}$ and $\{x, y\}$ be a cherry of $\mathcal{T}$. Then $xy|ab \in \mathrm{qcl}_1(\mathcal{Q})$ for all $a, b \in X - \{x, y\}$.*

Proof. For any $a, b \in X - \{x, y\}$, consider the path uv uniquely distinguished by the quartet tree $xy|ab$ with vertex u adjacent to $\{x, y\}$ as in Figure 3.

Let the *distinguished path length* $dpl_{xy}(a,b) := l(u,v)$ for this path uv and let $dpl_{xyMAX} := max(\{dpl_{xy}(a,b) : a,b \in X - \{x,y\}\})$. Consider the three subtrees of $\mathcal{T}$ that would be derived by deleting vertex v. Let $\mathcal{T}_a$ be the subtree of $\mathcal{T}$ pendant to v containing a. Similarly, let $\mathcal{T}_b$ be the subtree of $\mathcal{T}$ pendant to v containing b.

We show that $ab|xy \in \mathrm{qcl}_1(\mathcal{Q})$ by applying induction on $\Delta(a,b) := dpl_{xyMAX} - dpl_{xy}(a,b)$. For $\Delta(a,b) = 0$, $\{a,b\}$ forms a cherry and due to the generous cover property of $\mathcal{Q}$: $xy|ab \in \mathcal{Q} \subseteq \mathrm{qcl}_1(\mathcal{Q})$. Suppose now that the result holds whenever $\Delta(a',b') = k$. Then, for $\Delta(a,b) = k+1$, since $\mathcal{Q}$ is a generous cover $\exists\, a_0 \in \mathcal{T}_A, b_0 \in \mathcal{T}_B : xy|a_0b_0 \in \mathcal{Q}$. We consider the following cases:

(i) $a = a_0, b = b_0$.
In this case $xy|ab \in \mathcal{Q} \subseteq \mathrm{qcl}_1(\mathcal{Q})$ as claimed.

(ii) $a = a_0, b \neq b_0$.
Since $\Delta(b,b_0) \leq k$, the induction hypothesis implies that $xy|bb_0 \in \mathrm{qcl}_1(\mathcal{Q})$. Furthermore $xy|a_0b_0 = xy|ab_0 \in \mathrm{qcl}_1(\mathcal{Q})$ and thus $xy|ab \in \mathrm{qcl}_1(\{xy|bb_0, xy|ab_0\}) \subseteq \mathrm{qcl}_1(\mathcal{Q})$.

(iii) $a \neq a_0, b = b_0$.
Symmetric argument to case (ii).

(iv) $a \neq a_0, b \neq b_0$.
As $\Delta(b,b_0) \leq k$, the induction hypothesis implies $xy|bb_0 \in \mathrm{qcl}_1(\mathcal{Q})$. Similarly, $xy|aa_0 \in \mathrm{qcl}_1(\mathcal{Q})$. So $xy|a_0b \in \mathrm{qcl}_1(\{xy|bb_0, xy|a_0b_0\})$ and thus $xy|ab \in \mathrm{qcl}_1(\{xy|aa_0, xy|a_0b\}) \subseteq \mathrm{qcl}_1(\mathcal{Q})$.

Thus, in all possible cases $xy|ab \in \mathrm{qcl}_1(\mathcal{Q})$, which completes the proof of Lemma 1.

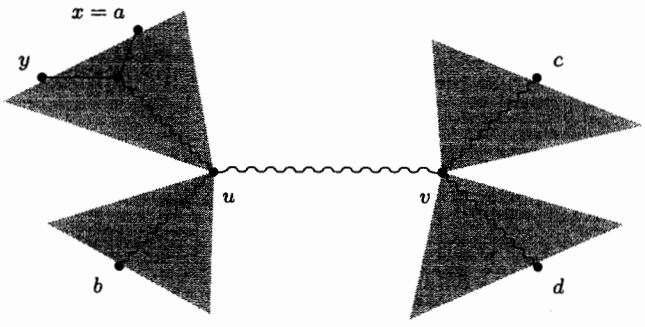


Fig. 4. Transition from $\mathcal{T}$ to $\mathcal{T}' = \mathcal{T}|X - \{x\}$ deletes x and suppresses z.

Lemma 2. *Let $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ be a generous cover for a binary phylogenetic tree $\mathcal{T}$ and let $\{x, y\}$ be a cherry of $\mathcal{T}$. Then $\mathrm{qcl}_1(\mathcal{Q})$ is a generous cover for $\mathcal{T}' = \mathcal{T}|X - \{x\}$.*

Proof. Let u, v be any interior vertices of $\mathcal{T}'$. Since $\mathcal{Q}$ is a generous cover, there exists a quartet $ab|cd \in \mathcal{Q}$ that *distinguishes* the path uv (cf. Fig. 4). Now, either $x \notin \{a, b, c, d\}$, in which case we are done, or, without loss of generality, $x = a$. Note that then $y \notin \{b, c, d\}$. And since $xb|cd \in \mathcal{Q} \subseteq \mathrm{qcl}_1(\mathcal{Q})$ and, by Lemma 1, $xy|cd \in \mathrm{qcl}_1(\mathcal{Q})$, the quartet tree $yb|cd \in \mathrm{qcl}_1(\{xb|cd, xy|cd\}) \subseteq \mathrm{qcl}_1(\mathcal{Q})$ distinguishes uv independently of x, thus establishing Lemma 2.

Theorem 3. *Let $\mathcal{T}$ be a binary phylogenetic tree and suppose that quartet set $\mathcal{Q} \subseteq \mathcal{Q}(\mathcal{T})$ is a generous cover. Then $\mathrm{qcl}_1(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$.*

Proof. We use induction on $|X|$. For $|X| = 4$, $\mathrm{qcl}_1(\mathcal{Q}) = \mathcal{Q}(\mathcal{T})$ as needed. For $|X| = k + 1$ choose any cherry $\{x, y\}$ of $\mathcal{T}$. By Lemma 2, $\mathrm{qcl}_1(\mathcal{Q})$ contains a subset $\mathcal{Q}_x$ that is a generous cover of $T|X - \{x\}$. By the induction hypothesis therefore $\mathcal{Q}(\mathcal{T}|X - \{x\}) = \mathrm{qcl}_1(\mathcal{Q}_x)$. Similarly, $\mathrm{qcl}_1(\mathcal{Q})$ contains a subset $\mathcal{Q}_y$ that is a generous cover of $T|X - \{y\}$ and so $\mathcal{Q}(\mathcal{T}|X - \{y\}) = \mathrm{qcl}_1(\mathcal{Q}_y)$. Finally, Lemma 1 yields that $\mathcal{Q}_{xy} := \{xy|ab : a, b \in X - \{x, y\}\} \subseteq \mathrm{qcl}_1(\mathcal{Q})$.

Now,

$$\begin{aligned}
\mathcal{Q}(\mathcal{T}) &= \mathcal{Q}(\mathcal{T}|X - \{x\}) \cup \mathcal{Q}(\mathcal{T}|X - \{y\}) \cup \mathcal{Q}_{xy} \\
&\subseteq \mathrm{qcl}_1(\mathcal{Q}_x) \cup \mathrm{qcl}_1(\mathcal{Q}_y) \cup \mathrm{qcl}_1(\mathcal{Q}) \\
&\subseteq \mathrm{qcl}_1(\mathcal{Q}) \cup \mathrm{qcl}_1(\mathcal{Q}) \cup \mathrm{qcl}_1(\mathcal{Q}) \\
&= \mathrm{qcl}_1(\mathcal{Q}) \\
&\subseteq \mathcal{Q}(\mathcal{T}),
\end{aligned}$$

hence $\mathcal{Q}(\mathcal{T}) = \mathrm{qcl}_1(\mathcal{Q})$ as is required to establish the induction step.

Note that the converse of Theorem 3 is not true, as in Example 1 where $\mathcal{Q}(\Sigma)$ is not a generous cover of the tree $\mathcal{T}$ on five leaves defined by Σ, yet $\mathrm{qcl}_1(\mathcal{Q}(\Sigma)) = \mathcal{Q}(\mathcal{T})$.

3 An Application to Phylogenetic Enumeration

It is well known that a necessary, but not sufficient, condition for a set $\mathcal{Q}$ of quartet trees to define a fully resolved phylogenetic X–tree $\mathcal{T}$ is the following: each interior edge of $\mathcal{T}$ must be distinguished by at least one quartet. Consequently we must have $|\mathcal{Q}| \geq n - 3$, where $n = |X|$. Consider the case where we have a set $\mathcal{Q}$ of size $n - 3$. A natural question is how many fully resolved phylogenetic trees have the property that each of their interior edges is distinguished by $\mathcal{Q}$. When $n \leq 5$ this number can be at most 1, but when $n = 6$ this number can be 2 (an example is provided by the set $\mathcal{Q} = \{12|35, 34|26, 56|14\}$).

In the following result we use one of the dyadic quartet closure rules to establish an upper bound on this number in general. First we state a lemma (cf. Steel, 1992, or Theorem 6.8.8 of Semple and Steel, 2003) that is central to the proof.

Lemma 3. *Suppose $\mathcal{Q}$ is a collection of quartet trees that distinguishes every interior edge of some fully-resolved phylogenetic X–tree $\mathcal{T}$. If, in addition, there is some element $y \in X$ that is a leaf in each quartet tree in $\mathcal{Q}$, then $\mathcal{Q}$ defines $\mathcal{T}$.*

Theorem 4. *Suppose $\mathcal{Q}$ is a set of quartet trees of size $n-3$ where n is the size of the set X of leaf labels of $\mathcal{Q}$. Let $\mathrm{d}(\mathcal{Q})$ denote the set of fully-resolved phylogenetic X–trees for which each interior edge is distinguished by exactly one quartet tree in $\mathcal{Q}$. Then*

$$\log_2 |\mathrm{d}(\mathcal{Q})| \leq n - 3 - \lceil \frac{4(n-3)}{n} \rceil.$$

Proof. For each element $x \in X$ let $n(x)$ denote the number of quartet trees in $\mathcal{Q}$ that have leaf x. Consider the set of pairs (q, x) where $q \in \mathcal{Q}$ and x is a leaf of q. Counting this set in the two obvious ways, we obtain $\sum_{x \in X} n(x) = 4|\mathcal{Q}| = 4(n-3)$, and so the average value of $n(x)$ over all choices $x \in X$ is exactly $4(n-3)/n$. Consequently there exists an element $y \in X$ that lies in at least $r := \lceil 4(n-3)/n \rceil$ quartet trees from $\mathcal{Q}$. Now, for the set $\mathcal{Q}'$ consisting of the $n-3-r$ quartet trees in $\mathcal{Q}$ that do not contain leaf y, let us replace each quartet $q = ab|cd$ by either $S_1(q) := \{ab|cy, ab|dy\}$ or $S_2 := \{ay|cd, by|cd\}$. For each map $\pi : \mathcal{Q}' \to \{1, 2\}$ consider the collection

$$\mathcal{Q}[\pi] := (\mathcal{Q} - \mathcal{Q}') \cup (\cup_{q \in \mathcal{Q}'} S_{\pi(q)}(q)).$$

We claim that there exists a bijection

$$\beta : \mathrm{d}(\mathcal{Q}) \to \{\pi : \mathcal{Q}' \to \{1, 2\} : \mathcal{Q}[\pi] \text{ is compatible } \}.$$

This bijection associates to each phylogenetic X–tree $\mathcal{T}$ in $\mathrm{d}(\mathcal{Q})$ the map $\pi : \mathcal{Q}' \to \{1, 2\}$. Consider a quartet tree $ab|cd \in \mathcal{Q}'$. Since $ab|cd$ distinguishes some interior edge, say $e = \{u, v\}$, of $\mathcal{T}$, then leaf y is connected to precisely one of a, b, c, d in the forest obtained from $\mathcal{T}$ by deleting u and v. If y is connected to c or d in this forest then set $\pi(ab|cd) = 1$, otherwise set $\pi(ab|cd) = 2$. Note that $\mathcal{Q}[\pi]$ is compatible, since each quartet tree in $\mathcal{Q}(\pi)$ is displayed by $\mathcal{T}$, thus the function β that we have just described is well defined. To see that β is one-to-one, suppose that $\beta(\mathcal{T}) = \beta(\mathcal{T}') = \pi$. Then $\mathcal{T}$ and $\mathcal{T}'$ both display all the quartet trees in $\mathcal{Q}(\pi)$. Furthermore, each quartet tree in $\mathcal{Q}[\pi]$ contains leaf y, and $\mathcal{Q}[\pi]$ distinguishes every interior edge of $\mathcal{T}$. Thus, by Lemma 3, $\mathcal{T}' = \mathcal{T}$ and so β is one-to-one, as claimed. Finally, β is onto, since if $\pi : \mathcal{Q}' \to \{1, 2\}$ has the property that $\mathcal{Q}[\pi]$ is compatible, then if $\mathcal{T}$ is a tree that displays $\mathcal{Q}(\pi)$ then $\mathcal{T}$ displays each quartet tree in $\mathcal{Q}$; this follows by applying (**Q1**)—any tree that displays $S_1(q)$ (or $S_2(q)$) also displays q.

The existence of the bijection β immediately implies that

$$\log_2 |\mathrm{d}(\mathcal{Q})| \leq \log_2 |\{\pi : \mathcal{Q}' \to \{1,2\}\}| = |\mathcal{Q}'| = n - 3 - r$$

which establishes the theorem.

4 Tree Construction In Homoplasy-Free Evolution

Markov models are now standard for modeling the evolution of aligned genetic sequence data. These models are routinely used as the basis for phylogenetic tree construction using techniques such as maximum likelihood (cf. Swofford et al., 1996). In these models the state space (the set of possible values each character can take) is typically samll; for DNA sequence data it is 4, but occasionally 2 for purine-pyrimidine data, or 20 for amino acid sequences. For such models the subsets of the vertices of a phylogenetic tree $\mathcal{T}$ that are assigned to particular states do not generally form connected subtrees of $\mathcal{T}$ (in biological terminology this is because of 'homoplasy,' which is the evolution of the same state more than once in the tree).

4.1 The Random Cluster Model

Recently, there is increasing interest in genomic characters such as gene order where the underlying state space may be very large (cf. Gallut and Barriel, 2002; Moret et al., 2001; Moret et al., 2002; and Rokas and Holland, 2000). For example, the order of k genes in a signed circular genome can take any of $2^k(k-1)!$ values. In these models whenever there is a change of state (e.g., a reshuffling of genes by a random inversion of a consecutive subsequence of genes), it is likely that the resulting state (gene arrangement) is a unique evolutionary event, arising for the first time in the evolution of the genes under study. Indeed Markov models for genome rearrangement such as the (generalized) Nadeau-Taylor model (cf. Moret et al., 2002; Nadeau and Taylor, 1984) confer a high probability that any given character generated is homoplasy-free on the underlying tree, provided the number of genes is sufficiently large relative to $|X|$ (Semple and Steel, 2002). In this setting the random cluster model is the appropriate (limiting case) model, and may be viewed as the phylogenetic analogue of what is known in population genetics as the 'infinite alleles model' of Kimura and Crow (1964).

We now consider the following random process on a phylogenetic tree $\mathcal{T}$. For each edge e let us independently either cut this edge with probability $p(e)$ or leave it intact. The resulting disconnected graph (forest) G partitions the vertex set $V(\mathcal{T})$ of $\mathcal{T}$ into non-empty sets according to the equivalence relation that $u \sim v$ if u and v are in the same component of G. This model thus generates random partitions of $V(\mathcal{T})$, and thereby of X by connectivity. In this way we can generate a character on X, as illustrated in Fig. 5, or

more generally, a sequence $\mathcal{C}$ of independently-generated characters. Following Mossel and Steel (2003), we call resulting probability distribution on partitions of X the *random cluster model* with parameters $(\mathcal{T}, p)$ where p is the map $e \mapsto p(e)$.

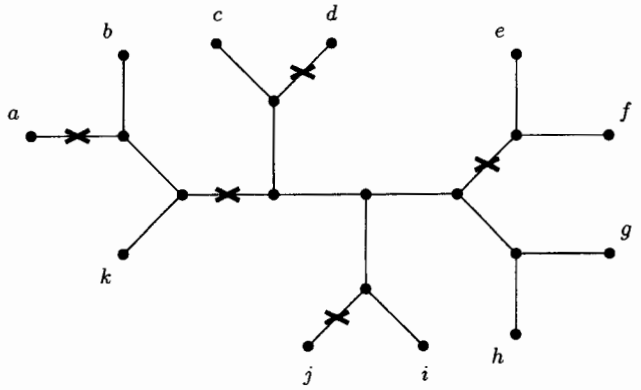


Fig. 5. Cutting the marked edges yields the character $\{a|bk|cghi|d|ef|j\}$.

Mossel and Steel (2003) show that the number of characters required to correctly reconstruct a fully resolved tree with n leaves grows at the rate $\log(n)$ provided the range of p is constrained to lie between any two fixed values that are less than 0.5. Here we use simulations to investigate how much phylogenetic information can be accessed using the various dyadic closure rules, when characters evolve according to this model.

4.2 Quantitative Closure Gains

For a compatible set $\mathcal{Q}$ of quartet trees, let

$$\mathrm{cl}(\mathcal{Q}) = \bigcap_{\mathcal{T} \in \mathrm{co}(\mathcal{Q})} \mathcal{Q}(\mathcal{T})$$

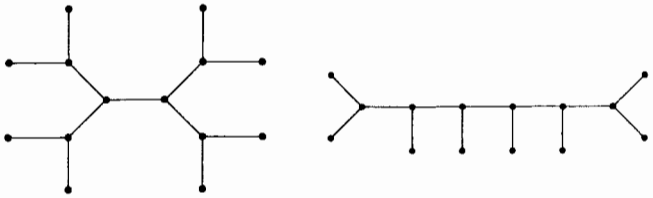
where $\mathrm{co}(\mathcal{Q})$ is the set of phylogenetic trees that display each of the trees in $\mathcal{Q}$. Thus $\mathrm{cl}(\mathcal{Q})$ consists of precisely those quartet trees that are displayed by every phylogenetic tree that displays $\mathcal{Q}$, and so $\mathrm{qcl}_1(\mathcal{Q}), \mathrm{qcl}_2(\mathcal{Q}), \mathrm{qcl}_{1,2}(\mathcal{Q})$ are all subsets of $\mathrm{cl}(\mathcal{Q})$.

This set $\mathrm{cl}(\mathcal{Q})$ is called the *closure* of $\mathcal{Q}$ and it has the property that $|\mathrm{cl}(\mathcal{Q})| = \binom{n}{4}$ precisely when $\mathcal{Q}$ defines $\mathcal{T}$. No polynomial-time algorithm is known for computing the closure of a set of quartets, whereas the dyadic closure can clearly be computed in polynomial time.

This situation poses interesting questions in the random cluster model that have practical consequences. For a sequence $\mathcal{C}$ of characters generated

independently under the random cluster model, consider the induced set $\mathcal{Q} = \mathcal{Q}(\mathcal{C})$ of quartet trees (as described in Section 1.4). How large are $\mathcal{Q} = \mathcal{Q}(\mathcal{C}), \mathrm{qcl}_1(\mathcal{Q}), \mathrm{qcl}_2(\mathcal{Q}), \mathrm{qcl}_{1,2}(\mathcal{Q})$ on average in relation to $\mathrm{cl}(\mathcal{Q})$; i.e., how many 'forced' quartet relationships can be derived using only quartet rules of order two? And, similarly, how does the size of these sets compare to the number of all quartets ($\binom{n}{4}$) derivable from the original tree?

To gain insight, the random cluster model was simulated on 8-leaf trees and sets of quartets $\mathcal{Q} = \mathcal{Q}(\mathcal{C})$ were derived from the resulting characters. The parameters used were tree shape, number of characters k, and the common value of edge-cutting probability ($p(e)$).



(a) A bushy tree shape (b) A caterpillar tree shape

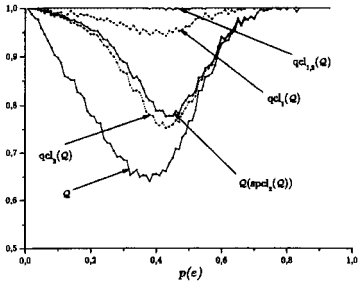
Fig. 6. Two extreme tree shapes on eight leaves used for random cluster model simulation: (a) shows a bushy tree shape, with minimal length for the longest inner path; (b) shows a caterpillar tree shape, with maximal length of the longest inner path.

We first simulated the model on trees having the bushy tree shape in Fig. 6(a). We ran averaged the results of 500 runs of the random cluster model for each set of parameter values. Each run gave a set of k characters C and the quartet sets $\mathcal{Q} = \mathcal{Q}(\mathcal{C})$, $\mathrm{qcl}_1(\mathcal{Q})$, $\mathrm{qcl}_2(\mathcal{Q})$, $\mathcal{Q}(\mathrm{spcl}_2(\mathcal{Q}))$, $\mathrm{qcl}_{1,2}(\mathcal{Q})$ and $\mathrm{cl}(\mathcal{Q})$, where $\mathrm{spcl}_2(\mathcal{Q}) = \mathrm{spcl}_2(\Sigma(\mathcal{Q}))$ with $\Sigma(\mathcal{Q}) = \{\{a,b\}|\{c,d\} : ab|cd \in \mathcal{Q}\}$ in line with Proposition 1. >From these quartet sets we computed the following values:

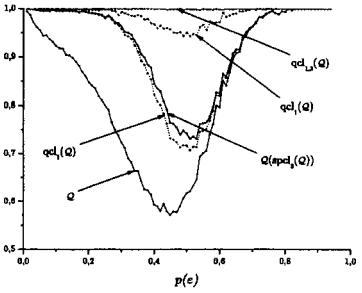
$$\frac{|\mathcal{Q}|}{|\mathrm{cl}(\mathcal{Q})|}, \frac{|\mathrm{qcl}_1(\mathcal{Q})|}{|\mathrm{cl}(\mathcal{Q})|}, \frac{|\mathrm{qcl}_2(\mathcal{Q})|}{|\mathrm{cl}(\mathcal{Q})|}, \frac{|\mathcal{Q}(\mathrm{spcl}_2(\mathcal{Q}))|}{|\mathrm{cl}(\mathcal{Q})|}, \frac{|\mathrm{qcl}_{1,2}(\mathcal{Q})|}{|\mathrm{cl}(\mathcal{Q})|}.$$

These values for the bushy tree shape on eight leaves, using 16 (respectively 32) characters, are shown graphically for varying $p(e)$ in the range $\{0, \ldots, .95\}$ (in steps of .01) in Fig. 7(a) (respectively Fig. 7(b)).

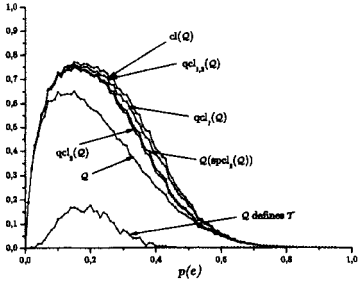
Furthermore, the number of quartets in each of the sets $\mathcal{Q}$, $\mathrm{qcl}_1(\mathcal{Q})$, $\mathrm{qcl}_2(\mathcal{Q})$, $\mathcal{Q}(\mathrm{spcl}_2(\mathcal{Q}))$, $\mathrm{qcl}_{1,2}(\mathcal{Q})$ and $\mathrm{cl}(\mathcal{Q})$ was put in relation to the total number of quartets obtainable from the original tree, yielding the values:



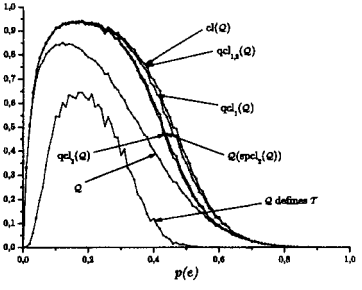
(a) The case $k = 16$. We plot the ratio of the size of the labeled quartet sets to the size of the closure.



(b) The case $k = 32$. We plot the ratio of the size of the labeled quartet sets to the size of the closure.



(c) The case $k = 16$. We plot the ratio of the size of the labeled quartet sets to the size of all quartets of the original tree. The average fraction of cases in which $\mathcal{Q}$ defines a tree is depicted and labeled "$\mathcal{Q}$ defines $\mathcal{T}$".



(d) The case $k = 32$. We plot the ratio of the size of the labeled quartet sets to the size of all quartets of the original tree. The average fraction of cases in which $\mathcal{Q}$ defines a tree is depicted and labeled "$\mathcal{Q}$ defines $\mathcal{T}$".

Fig. 7. Above Figures show the result of a random cluster model simulation on a bushy tree on 8 leaves yielding 16 (respectively 32) characters.

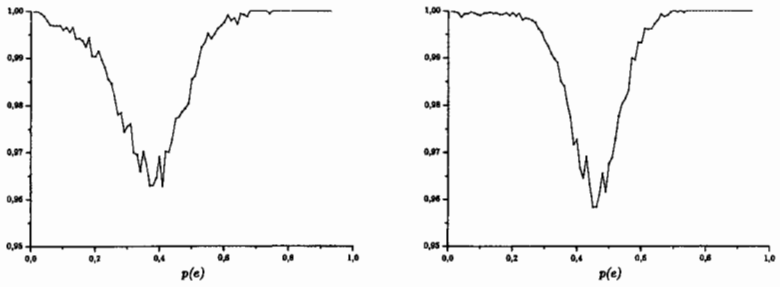
$$\frac{|\mathcal{Q}|}{\binom{n}{4}}, \frac{|\mathrm{qcl}_1(\mathcal{Q})|}{\binom{n}{4}}, \frac{|\mathrm{qcl}_2(\mathcal{Q})|}{\binom{n}{4}}, \frac{|\mathcal{Q}(\mathrm{spcl}_2(\mathcal{Q}))|}{\binom{n}{4}}, \frac{|\mathrm{qcl}_{1,2}(\mathcal{Q})|}{\binom{n}{4}}, \frac{|\mathrm{cl}(\mathcal{Q})|}{\binom{n}{4}}.$$

These values for the bushy tree shape on eight leaves, using 16 (respectively 32) characters are shown graphically for varying $p(e)$ in the range $\{0, \ldots, .95\}$ (in steps of .01) in Fig. 7(c) (respectively Fig. 7(d)). Interestingly, the resulting values do not depend very much on tree shape since similar simulations on the caterpillar type tree (Fig. 6(b)) on eight leaves and a completely random tree on eight leaves yielded a very similar result.

Quantifying the result from Theorem 2 that $\mathrm{qcl}_2(\mathcal{Q}(\Sigma)) \subseteq \mathcal{Q}(\mathrm{spcl}_2(\Sigma))$ and containment can be strict, the simulation shows that the value

$$\tau = |\mathrm{qcl}_2(\mathcal{Q})|/|\mathcal{Q}(\mathrm{spcl}_2(\mathcal{Q}))|$$

is larger than .95 on average in this setting. Its curve against $p(e)$ has a distinctive V-shape as shown in Fig. 8.



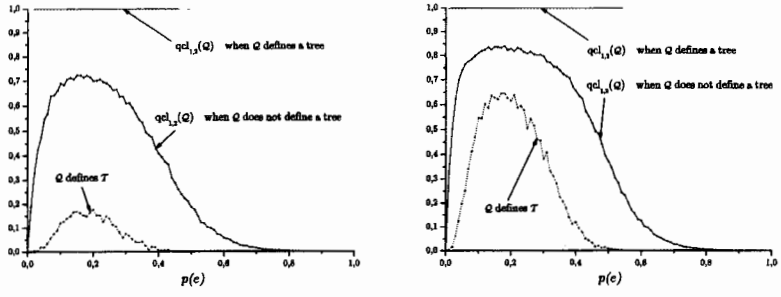
(a) The ratio τ for the case $k = 16$.

(b) The ratio τ for the case $k = 32$.

Fig. 8. The ratio τ is graphed against $p(e)$ in a random cluster model simulation with parameters: bushy tree shape, 8 leaves, quartets derived from k characters, 500 runs.

Note that in this simulation, the values $|\mathrm{qcl}_{1,2}(\mathcal{Q})|/\binom{n}{4}$ and $|\mathrm{cl}(\mathcal{Q})|/\binom{n}{4}$ are so close that they are indistinguishable in Fig. 7(c) and Fig. 7(d). This indicates that $|\mathrm{qcl}_{1,2}(\mathcal{Q})|/|\mathrm{cl}(\mathcal{Q})|$ is very close to 1 as is visible in Fig. 7(a) and Fig. 7(b), and thus the vast majority of quartets is already gained by $\mathrm{qcl}_{1,2}(\mathcal{Q})$. To explore this phenomenon further, we recorded the values of $|\mathrm{qcl}_{1,2}(\mathcal{Q})|/|\mathrm{cl}(\mathcal{Q})|$ separately, depending on whether $\mathcal{Q}$ defines a tree or not. These values are depicted in Fig. 9(a) (respectively Fig. 9(b)) for 16 (respectively 32) characters along with the average number of cases in which $\mathcal{Q}$ defines a tree. We found that in the majority of cases where Q did not define

a tree, $\mathrm{qcl}_{1,2}(\mathcal{Q}) = \mathrm{cl}(\mathcal{Q})$, but occasionally $\mathrm{qcl}_{1,2}(\mathcal{Q}) \neq \mathrm{cl}(\mathcal{Q})$, as expected by the comments in Section 1.2. However, an observation that was made consistently in all conducted simulations is that whenever Q defines a tree, $\mathrm{qcl}_{1,2}(\mathcal{Q})$ yields all the quartets of this tree; i.e., $\mathrm{qcl}_{1,2}(\mathcal{Q}) = \mathrm{cl}(\mathcal{Q})$. But in general this equality need not hold for a set $\mathcal{Q}$ that defines a tree (Steel, 1992). A further observation is that the average size of the quartet sets $\mathcal{Q}$ that define a tree decreases with $p(e)$ (results not shown).



(a) The case $k = 16$.

(b) The case $k = 32$.

Fig. 9. This plots the ratio of the size of the labeled quartet sets to the size of all quartets of the original tree. Furthermore, the average fraction of cases where $\mathcal{Q}$ defines a tree is depicted and labeled "$\mathcal{Q}$ defines $\mathcal{T}$".

4.3 Phylogenetic Information Content

Finally, we investigate a measure for the information content of characters described by Semple and Steel (2002). For a single character χ, a natural measure for the phylogenetic information content I of χ is the following:

$$I(\chi) = -\log(\mathrm{prop}(\chi))$$

where $\mathrm{prop}(\chi)$ is the proportion of fully-resolved phylogenetic X-trees for which χ is homoplasy-free. At the two extremes, when χ partitions X into singleton sets, or when χ has just one set (namely X), we have $I(\chi) = 0$, since then every tree is homoplasy-free under χ. Similarly, for a compatible set of characters $\mathcal{C} = \{\chi_1, \ldots, \chi_k\}$, we can define

$$I(\mathcal{C}) = -\log(\mathrm{prop}(\mathcal{C}))$$

where $\mathrm{prop}(\mathcal{C})$ is the proportion of fully-resolved phylogenetic X-trees $\mathcal{T}$ on which each $\chi \in \mathcal{C}$ is homoplasy-free.

Example 3. Assume

$$\chi_1 = \{\{1,2\},\{4,5\},\{3\},\{6\}\}$$
$$\chi_2 = \{\{3,4\},\{1,6\},\{2\},\{5\}\}$$
$$\chi_3 = \{\{5,6\},\{2,3\},\{1\},\{4\}\}$$
$$\chi_4 = \{\{5,6\},\{1,3\},\{2\},\{4\}\}$$

with $\mathcal{C}_1 = \{\chi_1, \chi_2, \chi_3\}$ and $\mathcal{C}_2 = \{\chi_1, \chi_2, \chi_4\}$. Then $\text{prop}(\chi_i) = 35/105$ and $I(\chi_i) = -\log(35/105)$ for all $i \in \{1,2,3,4\}$. On the other hand, $\text{prop}(\mathcal{C}_1) = 2/105$ and $I(\mathcal{C}_1) = -\log(2/105)$, whereas $\text{prop}(\mathcal{C}_2) = 1/105$ and $I(\mathcal{C}_2) = -\log(1/105) \neq I(\mathcal{C}_1)$.

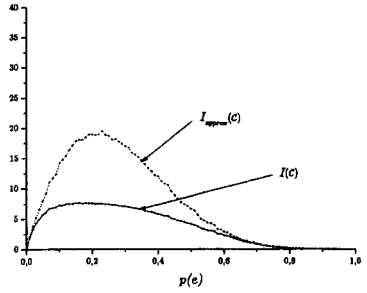
Fortunately, for a single character, $I(\chi)$ can easily be computed directly (without having to enumerate all trees to obtain the proportion) in polynomial time (Theorem 2 of Carter et al., 1990; see also Theorem 4.3 of Semple and Steel, 2002). But for a set $\mathcal{C}$ of characters, determining $\text{prop}(\mathcal{C})$ is at least as hard as determining whether a set of quartets defines a phylogenetic tree. This is illustrated in Example 3, where $\mathcal{C}_1$ and $\mathcal{C}_2$ both contain characters equivalent to three quartet trees, yet only $\mathcal{C}_2$ defines a phylogenetic tree and thus $I(\mathcal{C}_1) \neq I(\mathcal{C}_2)$. The computational complexity of determining whether a compatible set of quartet trees defines a phylogenetic tree is still open (cf. Semple and Steel, 2003).

For $\mathcal{C} = \{\chi_1, \ldots, \chi_k\}$, consider $I_{approx}(\mathcal{C}) = \sum_{1 \leq i \leq k} I(\chi_i)$, which can easily be computed in polynomial time. For a fully resolved phylogenetic X-tree $\mathcal{T}$, selected uniformly at random from the set of all such trees, if the events (for $i = 1, \ldots k$) defined by $E_i := \{\mathcal{T} \text{ displays } \chi_i\}$ were independent (in the statistical sense), then $I(\mathcal{C}) = I_{approx}(\mathcal{C})$. In general, however, $I(\mathcal{C}) \neq I_{approx}(\mathcal{C})$ and it is of interest to compare $I(\mathcal{C})$ with $I_{approx}(\mathcal{C})$ when $\mathcal{C}$ is generated under the random cluster model.

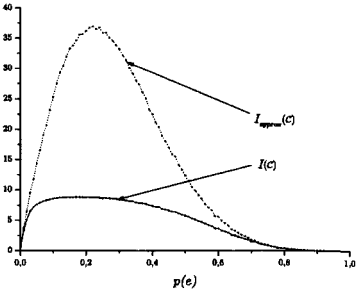
Figures 10(a) and 10(b) depict the values of $I(\mathcal{C})$ and $I_{approx}(\mathcal{C})$ in same random cluster model setting described in the previous section and shown in Fig. 7 for 16 (respectively 32) characters. The results suggest that $I_{approx}(\mathcal{C})$ tends to overestimate $I(\mathcal{C})$ (i.e, the events E_i are positively correlated), and that $I_{approx}(\mathcal{C})$ is only close to $I(\mathcal{C})$ when $p(e)$ is close to either 0 or 1.

5 Acknowledgments

We thank the New Zealand Institute for Mathematics and its Applications (NZIMA) for support under the *Phylogenetic Genomics* programme. We also thank two referees for some helpful comments on an earlier version of this manuscript.



(a) The case $k = 16$.



(b) The case $k = 32$.

Fig. 10. Figures (a) and (b) depict the values of $I(C)$ and $I_{approx}(C)$ as discussed in Section 4.3.

References

1. Aho, A. V., Sagiv, Y., Szymanski, T. G., and Ullman, J. D. (1981). "Inferring a Tree from Lowest Common Ancestors with an Application to the Optimization of Relational Expressions," *SIAM Journal on Computing*, **10**, 405–421.
2. Bandelt, H.-J., and Dress, A.W.M. (1986). "Reconstructing the Shape of a Tree from Observed Dissimilarity Data," *Advances in Applied Mathematics*, **7**, 309–343.
3. Bryant, D. and Steel, M. (1995). "Extension Operations on Sets of Leaf-Labelled Trees," *Advances in Applied Mathematics*, **16**, 425–453.
4. Böcker, S. (1999). *From Subtrees to Supertrees.* Unpublished PhD thesis. Fakultät für Mathematik, Universität Bielefeld, Bielefeld.
5. Böcker, S., Dress, A. W. M., and Steel, M. (1999). "Patching Up X-Trees," *Annals of Combinatorics*, **3**, 1–12.
6. Böcker, S., Bryant, D., Dress, A. W. M., and Steel, M. A. (2000). "Algorithmic Aspects of Tree Amalgamation," *Journal of Algorithms*, **37**, 522–537.
7. Carter, M., Hendy, M. D., Penny, D., Székely, L. A., and Wormald, N. C. (1990). "On the Distribution of Lengths of Evolutionary Trees," *SIAM Journal on Discrete Mathematics*, **3**, 38–47.
8. Colonius, H. and Schulze, H. H. (1981). "Tree Structures for Proximity Data," *British Journal of Mathematical and Statistical Psychology*, **34**, 167–180.
9. Dekker, M. C. H. (1986). *Reconstruction Methods for Derivation Trees.* Unpublished Masters thesis. Vrije Universiteit, Amsterdam, Netherlands.
10. Estabrook, G. F. and McMorris, F. R. (1977). "When Are Two Qualitative Taxonomic Characters Compatible?" *Journal of Mathematical Biology*, **4**, 195–200.
11. Gallut, C. and Barriel, V. (2002). "Cladistic Coding of Genomic Maps," *Cladistics*, **18**, 526–536.

12. Huber, K.T., Moulton, V. and Steel, M. (2002). "Four Characters Suffice to Convexly Define a Phylogenetic Tree," Research Report UCDMS2002/12, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand.
13. Kimura, M. and Crow, J. (1964). "The Number of Alleles That Can Be Maintained in a Finite Population," *Genetics*, **49**, 725–738.
14. Meacham, C. A. (1983). "Theoretical and Computational Considerations of the Compatibility of Qualitative Taxonomic Characters," in *Numerical taxonomy*, NATO ASI Series, Vol. G1, ed. J. Felsenstein, Berlin: Springer-Verlag, pp.304–314.
15. Moret, B.M.E. Tang,J. Wand, L.S. and Warnow, T. (2002). "Steps Toward Accurate Reconstruction of Phylogenies from Gene-Order Data," *Journal of Computer and System Sciences*, **65**, 508–525.
16. Moret, B.M.E., Wang, L.S., Warnow, T. and Wyman, S. (2001). "New Approaches for Reconstructing Phylogenies Based on Gene Order," Proceedings of the 9th International Conference on Intelligent Systems for Molecular Biology ISMB-2001, *Bioinformatics*, **17**, S165–S173.
17. Nadeau, J.J. and Taylor, B.A. (1984). "Lengths of Chromosome Segments Conserved Since Divergence of Man and Mouse," *Proceedings of the National Academy of Sciences USA*, **81**, 814–818.
18. Ng, M. P., Steel, M., and Wormald, N. C. (2000). "The Difficulty of Constructing a Leaf-Labelled Tree Including or Avoiding a Given Subtree," *Discrete Applied Mathematics*, **98**, 227–235.
19. Rokas, A., Holland P.W.H. (2000). "Rare Genomic Changes as a Tool for Phylogenetics," *Trends in Ecology and Evolution*, **15**, 454-459.
20. Semple, C. and Steel, M. (2001). "Tree Reconstruction Via a Closure Operation on Partial Splits," in *Proceedings of Journées Ouvertes: Biologie, Informatique et Mathématique*, Lecture Notes in Computer Science, eds. O. Gascuel and M.-F. Sagot, Berlin: Springer-Verlag, pp.126–134.
21. Semple, C. and Steel, M. (2002). "Tree Reconstruction from Multi-State Characters," *Advances in Applied Mathematics*, **28**, 169–184.
22. Semple, C. and Steel, M. (2003). *Phylogenetics*, Oxford, U.K.: Oxford University Press.
23. Steel, M. (1992). "The Complexity of Reconstructing Trees from Qualitative Characters and Subtrees," *Journal of Classification*, **9**, 91–116.
24. Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). "Phylogenetic Inference," in *Molecular Systematics* (2nd edn.), eds. D. M. Hillis, C. Moritz, B. K. Marble, Sunderland U.S.A.: Sinauer, pp. 407–514.
25. Mossel, E. and Steel, M. (2003). "A Phase Transition for a Random Cluster Model on Phylogenetic Trees," *Mathematical Biosciences* (in press).