Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mathematical Biosciences 187 (2004) 189–203
Mathematical Biosciences
www.elsevier.com/locate/mbs

A phase transition for a random cluster model on phylogenetic trees

Elchanan Mossel [a], Mike Steel [b,*]

[a] Computer Science and Statistics, University of California, Berkeley, CA, USA
[b] Biomathematics Research Centre, University of Canterbury, Room 623, Private Bag 4800, Christchurch, New Zealand

Received 11 April 2003; received in revised form 8 October 2003; accepted 8 October 2003

Abstract

We investigate a simple model that generates random partitions of the leaf set of a tree. Of particular interest is the reconstruction question: what number k of independent samples (partitions) are required to correctly reconstruct the underlying tree (with high probability)? We demonstrate a phase transition for k as a function of the mutation rate, from logarithmic to polynomial dependence on the size of the tree. We also describe a simple polynomial-time tree reconstruction algorithm that applies in the logarithmic region. This model and the associated reconstruction questions are motivated by a Markov model for genomic evolution in molecular biology.

Keywords: Phylogenetic tree; Phase transition; Random cluster model

1. Introduction

A central question in evolutionary biology is the following: how much information about historical relationships between species can be recovered from the genes they carry? In this paper we investigate an aspect of this question using a simple model, which we refer to as the *random cluster model*. This model is closely related to branching processes [1], and may be defined equivalently in terms of percolation, infinite state Potts models or random cluster models on trees. See e.g. [2–5] for basic background in statistical physics.

* Corresponding author. Tel.: +64-3 366 7001/7688/2600; fax: +64-3 364 2587.
E-mail addresses: mossel@stat.berkeley.edu (E. Mossel), m.steel@math.canterbury.ac.nz (M. Steel).

0025-5564/$ - see front matter
doi:10.1016/j.mbs.2003.10.004

The model may be viewed as an 'infinite state' Poisson process model where states are never re-visited in their evolution in a tree. For this model we are interested in how much data is required to reconstruct the underlying tree. We show that, provided the process is mildly conservative (namely, the probability of a state change on any edge is at most $\frac{1}{2}$) one requires just $O(\log(n))$ independent samples to reconstruct a tree with n leaves. Furthermore this bound is optimal, and there is a simple and fast algorithm for reconstructing this tree from the data. However when the process is less conservative, a phase transition occurs beyond which a polynomial number of samples is provably required for certain families of trees.

The structure of this paper is as follows. We begin by describing precisely the random cluster model, and we then summarize our main results in Theorem 1.1. We then describe the relevance of this model and our results to molecular systematics and to earlier results for other tree-based Markov models. In Section 2 we establish our result for the logarithmic region, and describe explicitly the constants involved, as well as providing a valid polynomial-time tree reconstruction method. In Section 3 we deal with the polynomial region.

1.1. The main result

Throughout this paper X is a finite set and we will let $n = |X|$. A *phylogenetic X-tree* (or more, briefly, a *phylogenetic tree*) is a tree $\mathscr{T}$ having leaf set X, and for which the interior vertices are unlabelled and of degree at least 3. If in addition each interior vertex has degree exactly 3 we say that $\mathscr{T}$ is *trivalent*. An example of a phylogenetic X-tree is shown in Fig. 1(a).

Two phylogenetic X-trees $\mathscr{T}$ and $\mathscr{T}'$ are regarded as equivalent if the identity map on X, regarded as a bijection from the set of leaves of $\mathscr{T}$ to the leaves of $\mathscr{T}'$ extends to a graph isomorphism between the two trees. Thus, for example, there are precisely three trivalent (and one non-trivalent) phylogenetic X-trees for any set X of size 4.

We now consider the following random process on a phylogenetic tree $\mathscr{T}$. 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 $V(\mathscr{T})$ of $\mathscr{T}$ into non-empty sets according to the equivalence relation that $u \sim v$ if v and v are in the same component of G. This model thus generates random partitions of $V(\mathscr{T})$, and thereby of X by connectivity, and we will refer to these partitions as $\overline{\chi}$ and χ, respectively. An example of such a partition χ is given in Fig. 1(b). For an element $x \in X$ we will let $\chi(x)$ denote the equivalence class containing x. We call the resulting probability distribution on partitions of X the *random cluster model* with parameters $(\mathscr{T}, p)$ where p is the map $e \mapsto p(e)$.

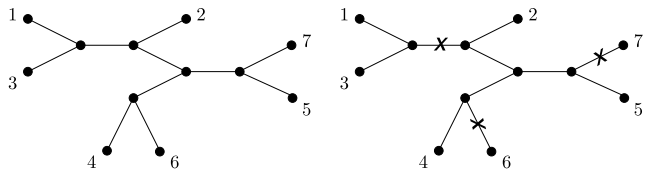

Fig. 1. (a) A trivalent phylogenetic X-tree $\mathscr{T}$ for $X = \{1, 2, \ldots, 7\}$. (b) For the random cluster model, cutting the edges of $\mathscr{T}$ that are marked by a cross induces the character χ on X given by $\chi = \{\{1, 3\}, \{2, 4, 5\}, \{6\}, \{7\}\}$.

In keeping with the biological setting we will call an arbitrary partition χ of X a *character* (on X). Let $\mathbb{P}(\chi|\mathscr{T},p)$ denote the probability of generating a character χ under the random cluster model with parameters $(\mathscr{T},p)$. We say a subset C of the set $E(\mathscr{T})$ of edges of $\mathscr{T}$ is a *cutset for χ on $\mathscr{T}$* if the partition χ of X equals that induced by the components of $(V(\mathscr{T}),E(\mathscr{T})-C)$. Then

$$\mathbb{P}(\chi|\mathscr{T},p)=\sum_{C}\prod_{e\in C}p(e)\prod_{e\in E(\mathscr{T})-C}(1-p(e)), \tag{1}$$

where the summation is over all cutsets C for χ on $\mathscr{T}$. Note that the number of terms in the summation described by Eq. (1) can be exponential with $|X|$. However by modifying the well-known dynamic programming approach for computing the probability of a character on a tree according to a *finite* state Markov process [6] one can compute $\mathbb{P}(\chi|T,p)$ in polynomial time in $|X|$.

Suppose we generate a sequence $\Pi=(\chi_1,\ldots,\chi_k)$ of k such independent characters on X where the generating pair (T,p) is unknown. We wish to reconstruct $\mathscr{T}$ with probability at least $1-\epsilon$ from Π.

The following theorem describes how the value of k is related to the size of $\mathscr{T}$ and properties of p.

Theorem 1.1. *Let $0<a\leqslant b<1$ and $0<\epsilon<1$ be fixed constants. Consider the random cluster model on any collection of the parameters $(\mathscr{T},p)$ where $\mathscr{T}$ is a trivalent phylogenetic tree, and $a\leqslant p(e)\leqslant b$ for all edges e of $\mathscr{T}$. Let k be the number of characters generated i.i.d. under this model, and $k_{\min}(\epsilon)$ be the minimal k such that the tree can be correctly reconstructed from the characters with probability at least $1-\epsilon$. Then, if n denotes the number of leaves of $\mathscr{T}$.*

(i) *$k_{\min}(\epsilon)$ grows logarithmically with n if $b<\frac{1}{2}$. In particular, if*

$$k\geqslant\frac{(1-b)^4}{a(1-2b)^4}\log\left(\frac{n^2}{\epsilon}\right),$$

then the tree can be reconstructed correctly with probability $1-\epsilon$. Furthermore, there is a polynomial-time (in n) algorithm for reconstructing $\mathscr{T}$ from the generated characters.

(ii) *$k_{\min}(\epsilon)$ can grow polynomially with n if $a>\frac{1}{2}$. In particular, for all h, if*

$$k\leqslant\frac{\epsilon(1-a)^h}{6}\left(\frac{n}{3}\right)^{-\log_2(2-2a)}, \tag{2}$$

then there exists a distribution on trivalent phylogenetic X-trees, such that if $\mathscr{T}$ is drawn according to the distribution, $p(e)=a$, for all edges of the trees, and characters are generated by $(\mathscr{T},p)$, then the probability of correctly reconstructing $\mathscr{T}$ given the k characters is bounded above by $\epsilon+3^{-3\times2^h}$.

Remark 1.2. We note that given a prior distribution $\mathbb{P}$ on the space of phylogenetic trees, the *probability of reconstruction* is well defined, once we assume that p is determined by $\mathscr{T}$. Indeed, given k characters $\chi_1,\ldots,\chi_k$ the best reconstruction algorithms will return the tree $(\mathscr{T},p)$ that maximizes

$$\mathbb{P}[(\mathscr{T},p)|\chi_1,\ldots,\chi_k] = \frac{\mathbb{P}[(\mathscr{T},p)]}{Z}\mathbb{P}[\chi_1,\ldots,\chi_k|(\mathscr{T},p)],$$

where Z is a constant independent of the tree. This follows from the Neyman–Pearson Lemma.

We now describe a motivation for this model from biology, and the relation of our results to some earlier work.

1.2. Relevance of the random cluster model to molecular systematics

In molecular phylogenetics the set X typically corresponds to the set of extant species (or genes) under study. Biologists seek to reconstruct a rooted tree that describes the evolution of these species from a common ancestor. The extant species under study are generally regarded as the leaves of the tree, and since speciation is usually regarded as a bifurcating process, the tree is viewed as a rooted binary tree with leaf set X. If we now suppress the root vertex of this tree we obtain a trivalent phylogenetic X-tree. This last step is not just a technical convenience – it is turns out that most models of genetic evolution (and consequently most tree reconstruction methods) can work just as naturally on unrooted trees as on rooted trees. Thus, a main goal of phylogenetic analysis is the reconstruction of trivalent phylogenetic X-tree by comparing genetic differences between the species in X.

Markov models are now standard for modeling the evolution of aligned genetic sequence data. Furthermore, these models are routinely used as the basis for phylogenetic tree reconstruction using techniques such as maximum likelihood [7]. In these models the state space (the set of possible values each character can take) is small – typically 4 for DNA sequence data (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 $\mathscr{T}$ that are assigned particular states do not generally form connected subtrees of $\mathscr{T}$ (in biological terminology this is because of 'homoplasy' – the evolution of the same state more than once in the tree). Consequently, the random cluster model is not an appropriate model for these characters.

However increasingly there is interest in genomic characters such as gene order where the underlying state space may be very large [8–11]. 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 – for example a re-shuffling 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 [9,12] 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 $|X|$ [13]. 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 [14].

This leads then to the following question, which is of both theoretical and practical interest: how many characters are required to reconstruct a phylogenetic tree correctly? More precisely, suppose the phylogenetic tree $\mathscr{T}$ is trivalent, and the probability of a net substitution on each

edge of the tree lies in the interval $[a,b]$ where $0 < a \leqslant b < 1$, and that we wish to reconstruct all such trees with probability at least $1-\epsilon$ for some $\epsilon > 0$. Let k be the required number of characters.

The random cluster model is the second model showing a phase-transition in this number k of characters needed for reconstruction. It is conjectured in [15] and proved in [16] that for tree-based Markov models based on the two state symmetric model

- k depends polynomially on $n = |X|$ for values of b above a certain critical value of $\frac{1}{2}\left(1-\frac{1}{\sqrt{2}}\right)$.
- Below that value and under some technical conditions, k depends logarithmically on n.

Theorem 1.1 shows that the situation with the random cluster model differs in two respects. Firstly, the critical value is 1/2 instead of $\frac{1}{2}\left(1-\frac{1}{\sqrt{2}}\right)$. This corresponds to the fact that in statistical physics models on the binary tree, the critical value for the extremality of the free measure or the Ising model is $\frac{1}{2}\left(1-\frac{1}{\sqrt{2}}\right)$, see [17–19], while the critical value for uniqueness of Gibbs measure, or the critical value for percolation is 1/2, see [4,5]. In [20] it is shown that for any Markov model, if the mutation rate is high then k depends polynomially on n.

The second respect in which the random cluster model differs from the symmetric two state model, is that for the random cluster model, the dependence of k on a has exponent -1 rather than -2, see [16,21,22].

2. The case $p_{\max} < \frac{1}{2}$

We begin this section by introducing some useful terminology.

We will mostly use $a, b, c, \ldots$ to denote vertices of the tree, $x, y, z, \ldots$ to denote leaves of the tree, and $u, v, w, \ldots$ to denote interior vertices of the tree.

A *quartet tree* is a trivalent phylogenetic X-tree for $|X| = 4$. We can represent any quartet tree by the notation $xy|wz$ where x, y are leaves that are adjacent to one interior vertex, while w, z are leaves that are adjacent to the other interior vertex.

For any trivalent phylogenetic X–tree, $\mathscr{T}$ let $\mathscr{Q}(\mathscr{T})$ denote the set of quartet trees induced by $\mathscr{T}$ by selecting subsets of X of size 4. It is a fundamental result that $\mathscr{T}$ is uniquely determined by $\mathscr{Q}(\mathscr{T})$ [23].

Suppose that $\mathscr{T}$ is a trivalent phylogenetic X–tree. We say that $\mathscr{T}$ *displays* a quartet tree $xy|wz$ (respectively, a set $\mathscr{Q}$ of quartet trees) if $xy|wz \in \mathscr{Q}(\mathscr{T})$ (respectively, if $\mathscr{Q} \subseteq \mathscr{Q}(\mathscr{T})$). For example the tree $\mathscr{T}$ in Fig. 1(a) displays the quartet tree 12|47.

For any three distinct vertices a, b, c of $\mathscr{T}$ let $\mathrm{med}(a,b,c)$ denote the *median vertex* of the triple a, b, and c; that is, the unique vertex of $\mathscr{T}$ that is shared by the paths connecting a and b, a and c and b and c.

A collection $\mathscr{Q}$ of quartet trees is a *generous cover* of $\mathscr{T}$ if $\mathscr{Q} \subseteq \mathscr{Q}(\mathscr{T})$ and if, for all pairs of interior vertices u, v there exists a quartet tree $xx'|yy' \in \mathscr{Q}$ for which $u = \mathrm{med}(x,x',v)$ and $v = \mathrm{med}(u,y,y')$. This concept is illustrated in Fig. 2. Note that if $\mathscr{Q}$ is a generous cover of a trivalent phylogenetic X-tree then $|\mathscr{Q}| \geqslant \binom{n-2}{2}$.

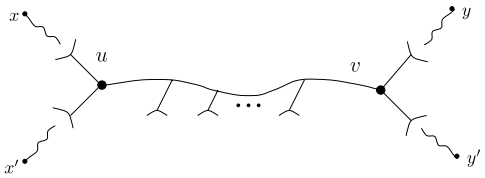

Fig. 2. A quartet $xx'|yy'$ for the pair $\{u, v\}$.

Given a sequence $\mathscr{C} = (\chi_1, \chi_2, \ldots, \chi_k)$ of characters on X, let

$$\mathscr{Q}(\mathscr{C}) = \{xx'|yy' : \exists i \in \{1, \ldots, k\} : \chi_i(x) = \chi_i(x') \neq \chi_i(y) = \chi_i(y')\}.$$

Consider a random cluster model on a phylogenetic X-tree $\mathscr{T}$. Recall that $p(e)$ denotes the probability that edge e is cut, and suppose that $p_{\min} := \min\{p(e)\} > 0$ and $p_{\max} := \max\{p(e)\} < \frac{1}{2}$. We write $q(e) = 1 - p(e)$, so $q_{\max} = \max\{q(e)\} < 1$ and $q_{\min} = \min\{q(e)\} > \frac{1}{2}$.

For any interior vertex v of $\mathscr{T}$, and any neighbour a of v, denote by $\alpha(v, a)$ the probability that there is a simple path $a_0 = v, a_1 = a, \ldots, a_n$, such that a_n is a leaf, and such that none of the edges on this path are cut.

Lemma 2.1. *Let v be an interior vertex and a a neighbour of v, then $\alpha(v, a) \geqslant g(q_{\min}) = \frac{2q_{\min}-1}{q_{\min}}$.*

Proof. The proof is by induction. If a is a leaf, then

$$\alpha(v, a) = q(\{v, a\}) \geqslant q_{\min} \geqslant g(q_{\min}),$$

as needed. Otherwise, let b and c be the neighbours of a different than v. Then

$$\alpha(v, a) = q(v, a)(1 - (1 - \alpha(a, b))(1 - \alpha(a, c))) \geqslant q_{\min}(\alpha(a, b) + \alpha(a, c) - \alpha(a, b)\alpha(a, c)).$$

The function $Z + Y - ZY$ is increasing in $Z, Y \in [0, 1]$. Therefore, by the induction hypothesis

$$\alpha(v, a) \geqslant q_{\min} g(q_{\min})(2 - g(q_{\min})) = \frac{2q_{\min} - 1}{q_{\min}} = g(q_{\min}),$$

as needed.

Finally, let $\beta := p_{\min} g(q_{\min})^4 > 0$. □

Lemma 2.2. *Consider the random cluster model on a trivalent phylogenetic X-tree $\mathscr{T}$, with associated parameter $\beta > 0$, and let $\mathscr{C}$ denote a sequence of k characters generated i.i.d. under this model. Provided*

$$k \geqslant \frac{1}{\beta} \log\left(\frac{n^2}{\epsilon}\right)$$

then, with probability at least $1 - \epsilon$, $\mathscr{Q}(\mathscr{C})$ is a generous cover of $\mathscr{T}$.

Proof. For any pair of interior vertices u, v, the probability that a character χ generated under this model satisfies $\chi_i(x) = \chi_i(x') \neq \chi_i(y) = \chi_i(y')$ for some $x, x', y, y' \in X$ with $u = \mathrm{med}(x, x', y)$ and $v = \mathrm{med}(x, y, y')$ is at least β. Consequently, the probability that $\mathscr{Q}(\mathscr{C})$ is not a generous cover for $\mathscr{T}$ is at most $n^2(1 - \beta)^k$ (since the number of pairs of interior vertices of $\mathscr{T}$ is $\binom{n-2}{2} < n^2$). The remainder now follows by standard algebra, together with the bound $\frac{1}{-\log(1-\beta)} \leqslant \frac{1}{\beta}$. □

Recall that a *cherry* of $\mathscr{T}$ is a pair of leaves that are adjacent to the same vertex.

Lemma 2.3. *Suppose that $\mathscr{Q}$ is a set of quartet trees on X, and that $\mathscr{T}$ is a trivalent phylogenetic X-tree. Suppose that $\{x,y\}$ is a cherry of $\mathscr{T}$. Construct a graph G_{xy} on vertex set $X-\{x,y\}$ with an edge between w, z precisely if $xy|wz \in \mathscr{Q}$. Then,*

(i) *If G_{xy} is connected, then any phylogenetic X-tree that displays $\mathscr{Q}$ has $\{x,y\}$ as a cherry.*
(ii) *If $\mathscr{Q}$ is a generous cover for $\mathscr{T}$, then G_{xy} is connected.*
(iii) *If $\mathscr{Q}$ is a generous cover for $\mathscr{T}$, then $\{x',y'\}$ is a cherry of $\mathscr{T}$ if and only if $\mathscr{Q}$ does not contain a quartet of the form $x'w|y'z$.*

Proof. For part (i), first note that any phylogenetic X-tree that displays both quartet trees $xx'|yy'$ and $xx'|yy''$ also displays $xx'|y'y''$ [23,24]. Thus, if G_{xy} is connected, and if a phylogenetic X-tree $\mathscr{T}'$ displays $\mathscr{Q}$, then $\mathscr{T}'$ also displays $\{xy|wz : w,z \in X-\{x,y\},\ w \neq z\}$. This implies that $\{x,y\}$ is a cherry of $\mathscr{T}'$.

For part (ii), we use induction on $|X|$. The result certainly holds for $|X| \leqslant 4$ so suppose that $|X| = k > 4$ and that $\{x,y\}$ is a cherry of $\mathscr{T}$. Since $\mathscr{T}$ is trivalent there exists another cherry of $\mathscr{T}$, say $\{w,w'\}$. Let $\mathscr{T}'$ be the trivalent phylogenetic tree obtained from $\mathscr{T}$ by deleting leaves w, w' and all edges adjacent to w, w' from $\mathscr{T}$. Let $\mathscr{Q}'$ be obtained from $\mathscr{Q}$ by deleting any quartet tree of the form $st|ww'$, and replacing all quartet trees of the form $tw|zz'$ and $tw'|zz'$ by $tu|zz'$ where $u \notin X$ denotes the unique neighbour of w and w'. Then $\mathscr{T}'$ still has the cherry $\{x,y\}$ and $\mathscr{Q}'$ is a generous cover for $\mathscr{T}'$ so by the induction hypothesis the associated graph G'_{xy} on the vertex set $(X-\{w,w'\})\cup\{u\}$ is connected. Now in G_{xy} there exists an edge connecting w and w' (since $xy|ww' \in \mathscr{Q}$, as $\mathscr{Q}$ is a generous cover for $\mathscr{T}$). Together with the connectivity of G'_{xy} it follows that G_{xy} is connected, as required.

For part (iii), the 'only if' direction follows immediately from the assumption that $\mathscr{T}$ displays $\mathscr{Q}$. For the 'if' direction, suppose that $\{x',y'\}$ is not a cherry of $\mathscr{T}$. Then if u and v denote the vertices of $\mathscr{T}$ that are adjacent to x' and y' we have $u \neq v$ and the assumption that $\mathscr{Q}$ is a generous cover for $\mathscr{T}$ implies the existence of a quartet tree $x'w|y'z \in \mathscr{Q}$, as required. □

Theorem 2.4. *Suppose that $\mathscr{Q}$ is a generous cover of a trivalent phylogenetic X-tree $\mathscr{T}$. Then $\mathscr{T}$ is the only phylogenetic X-tree that displays $\mathscr{Q}$.*

Proof. We use induction on $n = |X|$. The result certainly holds for $n = 4$ so suppose that it holds for $n = m \geqslant 4$ and that $|X| = m+1$. Select a cherry $\{x,y\}$, say, for $\mathscr{T}$. Suppose that $\mathscr{T}'$ is a phylogenetic X-tree that displays $\mathscr{Q}$. Combining parts (i) and (ii) of Lemma 2.3 $\{x,y\}$ is a cherry of $\mathscr{T}'$. Let $\mathscr{Q}'$ be obtained from $\mathscr{Q}$ by deleting any quartet tree of the form $xy|zz'$, and replacing all quartet trees of the form $yz|z'z''$ or $xz|z'z''$ by $uz|z'z''$, where $u \notin X$ denotes the unique neighbour of x and y.

Let $\widetilde{T}$ be the tree obtained from $\mathscr{T}$ by deleting x, y and all edges adjacent to x, y from $\mathscr{T}$. Define $\widetilde{T}'$ similarly. Then $\mathscr{Q}'$ is a generous cover of both $\widetilde{T}$ and $\widetilde{T}'$, so, by the induction hypothesis, $\widetilde{T} \cong \widetilde{T}'$. Now, since $\{x,y\}$ is a cherry of $\mathscr{T}$ and of $\mathscr{T}'$ this implies that $\mathscr{T}' \cong \mathscr{T}$, thereby establishing the induction step, and completing the proof. □

It follows from Theorem 2.4 and Lemma 2.2 that sequences of characters of length

$$k = \frac{(1-b)^4}{a(1-2b)^4} \log\left(\frac{n^2}{\epsilon}\right) = \mathrm{O}(\log n),$$

suffice to reconstruct the underlying tree with probability at least $1-\epsilon$. We thus obtain the first part of Theorem 1.1. Furthermore, this can be achieved by a polynomial-time algorithm. Indeed, there is a particularly simple recursive algorithm for reconstructing any trivalent phylogenetic X-tree $\mathscr{T}$ from any generous cover of $\mathscr{T}$ which goes as follows.

The first step is to find a cherry. This is done in the following manner. We find a pair of leaves x, y such that $\mathscr{Q}$ contains no quartet $xx'|yy'$. By Lemma 2.3(iii) such x, y is indeed a cherry. We now wish to reconstruct a tree on the set of leaves $X' = X - \{x, y\} \cup \{u\}$ where u denotes the unique neighbour of both x and y. This is done by replacing $\mathscr{Q}$ by $\mathscr{Q}'$ as in Theorem 2.4.

Remark 2.5. It is easy to generalize the above results to trees where all the interior degrees are at least 3. In the general case we define a generous cover as follows. We say that $\mathscr{Q}$ is a generous cover of a phylogenetic X-tree $\mathscr{T}$ if

- for all interior vertices u, v of $\mathscr{T}$ and
- all a, a' neighbours of u; b, b' neighbours of v, such that none of a, a', b, b' lies on the path connecting u and v,

there exists a quartet $xx'|yy' \in \mathscr{Q}$ such that

- x is at the end of a simple path $u, a, \ldots, x$,
- x' is at the end of a simple path $u, a', \ldots, x'$,
- y is at the end of a simple path $v, b, \ldots, y$,
- y' is at the end of a simple path $v, b', \ldots, y'$.

The proofs that a logarithmic number of samples suffices to obtain a generous cover with high probability and that a generous cover uniquely determine the tree are similar (the error bound $n^2(1-\beta)^k$ is now replaced by $n^4(1-\beta)^k$).

3. Lower bounds

In this section we establish lower bounds on the number k of characters needed for reconstruction. First, we prove a logarithmic lower bound which holds for all trees and all values of p bounded strictly between 0 and 1. This will establish that $k_{\min}(\epsilon)$ grows at least logarithmically in n in the first part of Theorem 1.1. Then in Section 3.2 we describe lower bounds that will establish the second part of Theorem 1.1.

Note that a logarithmic lower bound on k for tree-based Markov models on a fixed state space is guaranteed by trivial counting arguments [25]. However these arguments do not apply when the size of state space is infinite, or finite but variable with $|X|$. Indeed it has recently been shown that for any trivalent phylogenetic X-tree $\mathscr{T}$ there is an associated set of four characters $\mathscr{C}_{\mathscr{T}}$ for which

$\mathscr{T}$ is the only phylogenetic X-tree that can generate $\mathscr{C}_{\mathscr{T}}$ with positive probability under a random cluster model [13,26]. Thus it is reasonable to ask whether O(1) characters might suffice to reconstruct $\mathscr{T}$ under the random cluster model, at least for certain restrictions on p in the parameter pair $(\mathscr{T}, p)$. Proposition 3.4 excludes this possibility.

In this section we also prove a polynomial lower bounds for 'deep' trees, where $p_{\min} > 1/2$ – thus establishing the second part of Theorem 1.1. The proof follows a general principle already demonstrated in [16,20]. It says that if the mutual information between the root and level n decays exponentially, then for 'deep' trees, a polynomial number of samples are needed for reconstruction.

We will also encapsulate the following idea from [20].

Lemma 3.1. *Let $(\mathscr{T}_1, p_1), \ldots, (\mathscr{T}_m, p_m)$ be a sequence of phylogenetic X-trees. Consider the following model for generating k characters. Choose one of the trees $(\mathscr{T}_i, p_i)$ uniformly at random with probability $\frac{1}{m}$ and then generate all k characters according to the random cluster model on $(\mathscr{T}_i, p_i)$.*

Assume that k partitions $\overline{\mathscr{C}'} = (\overline{\chi}_1, \ldots, \overline{\chi}_k)$ are generated via the random cluster model on one of the above trees $(\mathscr{T}', p)$ and let $\mathscr{C}' = (\chi_1, \ldots, \chi_k)$ denote the induced characters.

Let $Z = Z(\overline{\mathscr{C}'}, \mathscr{T}', p)$ be a random variable defined in terms of $(\mathscr{T}', p)$ and the partitions $\overline{\mathscr{C}'}$ and assume that there exists a subset $I = I(Z(\overline{\mathscr{C}'}, \mathscr{T}', p)) \subset \{1, 2, \ldots, m\}$ such that

$$\mathbb{P}[(\mathscr{T}_i, p_i)|Z, \mathscr{C}'] = \begin{cases} \frac{1}{|I|} & \text{if } i \in I, \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Then the probability of reconstructing the tree given the k characters is at most $\mathbb{E}\left[\frac{1}{|I|}\right]$.

Proof. Assume that in the reconstruction process in addition to $\mathscr{C}'$ one is also given the value of the random variable Z. Clearly, this does not decrease the reconstruction probability.

From (3) it follows that conditioned on Z, all the trees (T_i, p_i) for $i \in I$ are equally likely. It now follows that the reconstruction probability is at most $\mathbb{E}\left[\frac{1}{|I|}\right]$. □

We now turn to some definitions. Given a phylogenetic X-tree $\mathscr{T}$ and a set of edges $F \subset E$ of $\mathscr{T}$, we write $\mathscr{T}/F$ for the tree obtained from $\mathscr{T}$ by contracting all the edges in F. We call $\mathscr{T}/F$ a *factor tree* of $\mathscr{T}$. Similarly, we write $(\mathscr{T}, p)/F$ for the tree $(\mathscr{T}/F, p')$, where $p'(e) = p(e)$ for all edges e of $\mathscr{T}/F$.

We let E^o be the set of interior edges of $\mathscr{T}$, i.e. the set of edges $e = (u, v)$ where neither u or v are leaves of $\mathscr{T}$. Finally, we say that a phylogenetic tree $\mathscr{T}$ *displays* a sequence $\mathscr{C} = (\chi_1, \chi_2, \ldots, \chi_k)$ of characters if $\mathscr{T}$ displays (as defined earlier) the quartet trees $\mathscr{Q}(\mathscr{C})$. This is equivalent to requiring that, for each character $\chi \in \mathscr{C}$, the subtrees of $\mathscr{T}$ connecting $\{x \in X : \chi(x) = \alpha\}$ are vertex disjoint across all $\alpha \in \chi(X)$.

Proposition 3.2. *Let $\mathscr{T}$ be a trivalent phylogenetic X-tree and $E \subset E^o$. The number of trivalent phylogenetic X-trees which have $\mathscr{T}/E$ as a factor is at least $3^{|E|}$.*

Proof. For $k \geqslant 2$ let $B(k) = \prod_{i=1}^{k-2}(2i-1)$, the number of trivalent phylogenetic X-trees for a given set X of size k [27]. It is easy to see that if $\mathscr{T}$ is a phylogenetic X-tree where all the interior degrees are at least 3, then the number of trivalent phylogenetic X-trees which have $\mathscr{T}$ as a factor, $f(\mathscr{T})$, is

exactly $\prod_w B(d(w))$, where the product is taken over all interior vertices of w of $\mathscr{T}$ and $d(w)$ is the degree of w in $\mathscr{T}$.

We also note that if $d_1 \geqslant 3$ and $d_2 \geqslant 3$, then $B(d_1 + d_2 - 2) \geqslant 3B(d_1)B(d_2)$.

We now prove the claim by induction on $|E|$. The claim is trivial when $|E| = 0$ or 1. For the induction step, assume that the claim holds for sets up to size $|E| - 1$ and let $e \in E$. By the induction hypothesis, $f(\mathscr{T}/(E - \{e\})) \geqslant 3^{|E|-1}$. Let u and v be the endpoints of e in the tree $\mathscr{T}/(E - \{e\})$, and $d(u)$, $d(v)$ their degrees in that tree. By contracting e, the vertices u, v are replaced by a single vertex of degree $d(u) + d(v) - 2$. Therefore

$$f(\mathscr{T}/E) = f(\mathscr{T}/(E - \{e\}))\frac{B(d(u) + d(v) - 2)}{B(d(u))B(d(v))} \geqslant 3f(\mathscr{T}/(E - \{e\})) \geqslant 3^{|E|},$$

as needed. □

3.1. Logarithmic lower bounds

We will shortly present a result (Proposition 3.4) which implies that $k_{\min}(\epsilon)$ grows at least logarithmically in the first part of Theorem 1.1, thereby completing the proof of the first part of that theorem. First we state the following lemma.

Lemma 3.3. *Consider a phylogenetic X-tree $\mathscr{T}$. Given k partitions $\overline{\mathscr{C}} = (\overline{\chi}_1, \ldots, \overline{\chi}_k)$ of $V(\mathscr{T})$, let $\mathscr{C} = (\chi_1, \ldots, \chi_k)$ be the induced characters at the leaves and*

$$F = \{e = \{u, v\} \in E^o : \forall 1 \leqslant j \leqslant k,\ \overline{\chi}_j(u) = \overline{\chi}_j(v)\} \tag{4}$$

Then any phylogenetic X-tree $\mathscr{T}'$ that has $\mathscr{T}/F$ as a factor, displays $\mathscr{C}$.

Proof. The proof follows from the fact that $\mathscr{T}/F$ displays $\mathscr{C}$. □

Proposition 3.4

(i) *Consider the random cluster model on a trivalent phylogenetic X-tree $(\mathscr{T}, p)$. Then for all positive integer t, k and $1 > \epsilon > 0$, if*

$$\sum_{e \in E^o} q(e)^k > 2\log(t/\epsilon), \tag{5}$$

then with probability at least $1 - \epsilon$, given k generated characters, $(\chi_1, \ldots, \chi_k)$, there are t distinct trivalent phylogenetic X-trees that display $(\chi_1, \ldots, \chi_k)$.

(ii) *Condition* (5) *holds whenever*

$$k \leqslant \frac{\log(n-3) - \log(2\log(t/\epsilon))}{-\log q_{\min}}, \tag{6}$$

(iii) *For all $\epsilon > 0$ and all t positive integer, if the prior probability on trees is the uniform distribution over all trivalent phylogenetic X-trees, where $p(e) = p$ for all edges, and if k satisfies* (6) *then the probability of correctly reconstructing the tree is bounded above by $\epsilon + \frac{1}{t}$.*

Proof. Part (i): We will show that with probability at least $1-\epsilon$, the set F in (4) is of size at least $\log t \geqslant \log t/\log 3$. This will imply the first claim by Lemma 3.3 and Proposition 3.2.

Note that the size of F can be written as $\sum_{e\in E^o} Y_e$, where $\mathbb{P}[Y_e = 1] = q(e)^k$ and where Y_e are independent 0/1 random variables. Letting $Y = \sum_{e\in E^o} Y_e$, we see that

$$\mathbb{E}[Y] = \sum_{e\in E^o} q(e)^k. \tag{7}$$

By standard large deviation results (see, for example [28] [Theorem A.1.13]), for $s > 0$,

$$\mathbb{P}[Y \leqslant \mathbb{E}[Y] - s] \leqslant \exp\left(-\frac{s^2}{2\mathbb{E}[Y]}\right).$$

Therefore, if $s^2 \geqslant -2\mathbb{E}[Y]\log\epsilon$, then $\mathbb{P}[Y \leqslant \mathbb{E}[Y] - s] \leqslant \epsilon$. In order to obtain $|F| \geqslant \log t$, with probability at least $1-\epsilon$, it suffices that

$$\mathbb{E}[Y] - \sqrt{-2(\log\epsilon)\mathbb{E}[Y]} \geqslant \log t. \tag{8}$$

This is equivalent to

$$\left(\sqrt{\mathbb{E}[Y]} - \sqrt{\frac{-\log\epsilon}{2}}\right)^2 \geqslant \log t - \frac{\log\epsilon}{2},$$

which holds if

$$\mathbb{E}[Y] \geqslant \left(\sqrt{\frac{-\log\epsilon}{2}} + \sqrt{\log t - \frac{\log\epsilon}{2}}\right)^2. \tag{9}$$

From the inequality $(x+y)^2 \leqslant 2(x^2+y^2)$, for $x, y > 0$, it is easy to see that (9) and thereby (8) holds whenever

$$\mathbb{E}[Y] \geqslant 2\left(\frac{-\log\epsilon}{2} + \log t - \frac{\log\epsilon}{2}\right) = 2\log t - 2\log\epsilon.$$

The sufficiency of (5) to establish the claim of part (i) now follows, in view of (7).

Part (ii): Condition (6) implies (5), since for any trivalent tree on n leaves, $|E^o| = n-3$.

Part (iii): We use Lemma 3.1. Given the k partitions $\overline{\mathscr{C}} = (\overline{\chi}_1, \ldots, \overline{\chi}_k)$ of $V(\mathscr{T})$, we define the random variable Z as $\mathscr{T}/F$. Let $m \geqslant 1$ be the number of trivalent phylogenetic X-trees that have the factor $\mathscr{T}/F$.

Note that if $\mathscr{T}/F$ is not a factor of $\mathscr{T}'$ then $\mathbb{P}[(\mathscr{T}', p)|Z, \mathscr{C}] = 0$, while if $\mathscr{T}/F$ is a factor of $\mathscr{T}'$, then

$$\mathbb{P}[(\mathscr{T}', p)|Z, \mathscr{C}] = \frac{\mathbb{P}[(\mathscr{T}', p)]\mathbb{P}[Z, \mathscr{C}|(\mathscr{T}', p)]}{\mathbb{P}[Z, \mathscr{C}]} = \frac{\mathbb{P}[(\mathscr{T}', p)](1-p)^{k|F|}\mathbb{P}[\mathscr{C}|(\mathscr{T}, p)/F]}{\mathbb{P}[Z, \mathscr{C}]} = \frac{1}{m}. \tag{10}$$

(The last equality follows from the fact that the penultimate quantity in (10) is the same for all trees $\mathscr{T}'$ having $\mathscr{T}/F$ as a factor.)

By Lemma 3.1 the reconstruction probability is at most $\mathbb{E}[1/m]$. By the parts (i) and (ii) of Proposition 3.4 it follows that

$$\mathbb{E}[1/m] \leqslant \frac{\mathbb{P}[m \geqslant t]}{t} + \mathbb{P}[m < t] \leqslant \frac{1}{t} + \epsilon,$$

as required. □

3.2. Polynomial lower bounds

We will shortly present a result (Proposition 3.7) that establishes the second part of Theorem 1.1, after we have introduced some further lemmas. We will again use the notion of factor tree. We now assume that the tree is trivalent and that $p_{\min} > \frac{1}{2}$.

Lemma 3.5. *Consider a phylogenetic X-tree $\mathscr{T}$. Given k partitions $\overline{\mathscr{C}} = (\overline{\chi}_1, \ldots, \overline{\chi}_k)$ of $V(\mathscr{T})$, let*

$$G = \{e = \{u, v\} \in E^o : \forall 1 \leqslant j \leqslant k, \forall x \text{ a leaf}, \ \overline{\chi}_j(x) \neq \overline{\chi}_j(u) \text{ and } \overline{\chi}_j(x) \neq \overline{\chi}_j(v)\} \tag{11}$$

Then any phylogenetic X-tree $\mathscr{T}'$ that has the factor $\mathscr{T}/G'$ for $G' \subset G$, displays the sequence $\mathscr{C}$ of induced characters.

Proof. The claim will follow once we show that $\mathscr{T}/G'$ displays $\mathscr{C}$. We may couple the process on $\mathscr{T}$ and on $\mathscr{T}/G'$ in such a way that for all partitions and all edges e which belong both to $\mathscr{T}$ and $\mathscr{T}/G'$ an edge is cut in $\mathscr{T}$ if and only if it is cut in $\mathscr{T}/G'$. Note that using this coupling, the characters on X that are induced by the partitions of $V(\mathscr{T})$ by restriction to the leaves of $\mathscr{T}$ and $\mathscr{T}/G'$ coincide. □

Lemma 3.6. *Let $G' \subset E$ be a set of edges of a trivalent phylogenetic X-tree $\mathscr{T}$. Given k characters, the probability that $G' \subset G$ is at least*

$$1 - 2k \sum_{e \in G'} (2q_{\max})^{d(e)}, \tag{12}$$

where $d(e)$ is the distance of e to the set of leaves of $\mathscr{T}$, i.e., for $e = \{u, v\}$, $d(e)$ is the minimum of the distance of u to the set of leaves of $\mathscr{T}$ and the distance of v to the set of leaves of $\mathscr{T}$.

Proof. We assume that $q_{\max} < 1/2$, as bound (12) is trivial if $q_{\max} \geqslant 1/2$.

Given a fixed edge $e = \{u, v\}$, let L_u be the set of leaves that are connected to u via a path not containing v. Define L_v similarly.

The expected number of leaves x such that $\overline{\chi}(u) = \overline{\chi}(x)$ or $\overline{\chi}(v) = \overline{\chi}(x)$ in a partition $\overline{\overline{\chi}}$ of $V(\mathscr{T})$ is given by

$$\sum_{x \in L_u} \mathbb{P}[\overline{\chi}(u) = \overline{\chi}(x)] + \sum_{x \in L_v} \mathbb{P}[\overline{\chi}(v) = \overline{\chi}(x)] \leqslant \sum_{x \in L_u} q_{\max}^{d(x,u)} + \sum_{x \in L_v} q_{\max}^{d(x,v)}, \tag{13}$$

where $d(x, x')$ is the graph metric distance between x and x'. Since $2q_{\max} < 1$, the sum $\sum_{x \in L_u} q_{\max}^{d(x,u)}$ becomes smaller if an element $x \in L_u$ is replaced by two elements x_1, x_2 with $d(x_1, u) = d(x_2, u) = d(x, u) + 1$. A similar statement holds for L_v.

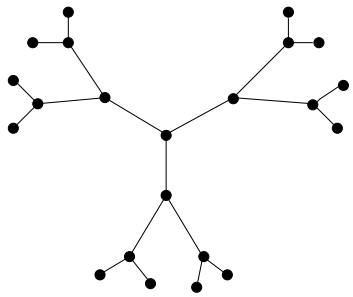

Fig. 3. A 3-level 3-regular tree.

It therefore follows, that both sums in the right hand side of (13) are maximized when $|L_u| = |L_v| = 2^{d(e)}$ and for all $x \in L_u$ ($x \in L_v$) it holds that $d(x,u) = d(e)$ ($d(x,v) = d(e)$).

From (13) we now obtain that the expected number of leaves x such that $\overline{\chi}(u) = \overline{\chi}(x)$ or $\overline{\chi}(v) = \overline{\chi}(x)$ is bounded by $2(2q_{\max})^{d(e)}$.

Therefore if we let s denote the size of the set

$$\{(i, e = \{u,v\}, x) : 1 \leqslant i \leqslant k,\ e \in G',\ x \text{ a leaf and } \overline{\chi}_i(u) = \overline{\chi}_i(x) \text{ or } \overline{\chi}_i(v) = \overline{\chi}_i(x)\}$$

then the expected value of s is at most $2k \sum_{e \in G'} (2q_{\max})^{d(e)}$. In particular, with probability at least that given by (12), s is zero, in which case $G' \subset G$. □

Recall that an r-level 3-regular tree is a tree $\mathscr{T}_r$ having an interior vertex that is separated from each leaf by exactly r edges. Fig. 3 shows an example of such a tree for $r = 3$. Note that, for an r-level 3-regular tree has $n = 3 \times 2^{r-1}$ leaves. We let $G_{r,h}$ be the set of edges of $\mathscr{T}_r$ at distance at least $r - h - 1$ from the set of leaves. Finally, we let $\mathscr{T}_{r,h} = \mathscr{T}_r / G_{r,h}$.

Proposition 3.7

(i) *Let $r \geqslant 1, h \geqslant 1$, and $\mathscr{T}$ a trivalent phylogenetic X-tree, which has the factor $\mathscr{T}_{r,h}$. Suppose we generate a sequence $\mathscr{C}$ of k characters using $\mathscr{T}$ under the random cluster model. Then with probability at least $1 - \epsilon$, there exists at least $3^{3 \times 2^h}$ distinct trivalent phylogenetic X-trees that display $\mathscr{C}$, where*

$$\epsilon = 3 \times 2^{h+1} k (2q_{\max})^{r-h-1} \tag{14}$$

(ii) *If we assume furthermore that the prior probability on trees is the uniform distribution over all X-trees that have $T_{r,h}$ as a factor and that $p(e) = p$ for all edges, then the probability of correctly reconstructing the tree is bounded by*

$$\frac{6k}{(1-p)^h} \left(\frac{n}{3}\right)^{\log_2(1-p)+1} + 3^{-3 \times 2^h}.$$

Note that part (ii) of Proposition 3.7 implies the second part of Theorem 1.1.

Proof. Part (i): Let G' be the set of edges of $\mathscr{T}$ at distance at least $r - h - 1$ from the set of leaves. By Lemma 3.6, $G' \subset G$ with probability at least $1 - \epsilon$, where G is defined in (11). Since the size of

G' is 3×2^h, it follows from Proposition 3.2, that there exist at least $3^{3\times 2^h}$ distinct trivalent phylogenetic X-trees which have the factor $\mathscr{T}/G' = \mathscr{T}_{r,h}$. By Lemma 3.5, all those trees display $\mathscr{C}$. This completes the proof of part (i).

Part (ii): Let A be the following event defined on the space of trees which have $\mathscr{T}_{r,h}$ as a factor, where trees are chosen uniformly at random.

- For all leaves x, all edges $e = \{u, w\}$ of $\mathscr{T}$ which are not edges of $\mathscr{T}_{r,h}$ and all characters $\overline{\chi}_j$ for $1 \leqslant j \leqslant k$, it holds that $\overline{\chi}_j(x) \neq \overline{\chi}_j(u)$ and $\overline{\chi}_j(x) \neq \overline{\chi}_j(w)$.

We now apply Lemma 3.1 with the random variable Z where Z takes the value 1 if A holds and the value $\mathscr{T}$ if A does not hold (where $\mathscr{T}$ is the tree that generated the sequence). We let I have the same meaning as in Lemma 3.1.

Note that $\mathbb{P}[(\mathscr{T}', p)|Z, \mathscr{C}, A^c] = 1$ if $\mathscr{T}' = \mathscr{T}$. Moreover, by coupling the cut events on the edges of $\mathscr{T}_{r,h}$ for all trees $\mathscr{T}'$ which have $\mathscr{T}_{r,h}$ as a factor, it follows that $\mathbb{P}[(\mathscr{T}', p)|Z, \mathscr{C}, A]$ has the same value for all trees that have $T_{r,h}$ as a factor. Moreover there are at least $3^{3\times 2^h}$ such trees.

We therefore conclude from Lemma 3.1, that the probability of reconstruction is bounded by

$$\mathbb{E}\left[\frac{1}{|I|}\right] = \mathbb{P}[A^c]\mathbb{E}\left[\frac{1}{|I|}|A^c\right] + \mathbb{P}[A]\mathbb{E}\left[\frac{1}{|I|}|A\right] \leqslant \mathbb{P}[A^c] + 3^{-3\times 2^h},$$

which by part (i) of Proposition 3.7 is bounded by

$$3 \times 2^{h+1} k (2-2p)^{r-h-1} + 3^{-3\times 2^h} = \frac{6k}{(1-p)^h}\left(\frac{n}{3}\right)^{\log_2(1-p)+1} + 3^{-3\times 2^h},$$

as needed. □

Remark. It can be shown that, in the case $p_{\min} > \frac{1}{2}$ polynomial dependence of k on n is is not just necessary, but also sufficient for the correct reconstruction of trivalent phylogenetic trees (with high probability).

Acknowledgements

We thank an anonymous referee for helpful comments that have improved the paper. Some of the research reported here was conducted when the first author was a PostDoc at the theory group at Microsoft Research. The first author was also supported by a Miller fellowship at U.C. Berkeley. The second author thanks the *New Zealand Institute for Mathematics and its Applications* (Phylogenetic Genomics Programme).

References

[1] K.B. Athreya, P.E. Ney, Branching Processes, Springer, 1972.
[2] H.O. Georgii, Gibbs measures and phase transitions, de Gruyter Studies in Mathematics, Walter de Gruyter and Co., Berlin, vol. 9, 1988.
[3] H.O. Georgii, O. Häggström, C. Maes, The random geometry of equilibrium phases, in: C. Domb, J.L. Lebowitz (Eds.), Phase Transitions and Critical Phenomena, Academic Press, London, vol. 18, 2001, p. 1.

[4] G. Grimmett, Percolation, 2nd Ed., Springer, Berlin, 1999.
[5] Y. Peres, Probability on Trees: An Introductory Climb. Springer Lecture Notes in Math, vol. 1717, 1999, p. 193.
[6] J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, J. Mol. Evol. 17 (1981) 368.
[7] D.L. Swofford, G.J. Olsen, P.J. Waddell, D.M. Hillis, Phylogenetic inference, in: D.M. Hillis, C. Moritz, B.K. Marble (Eds.), Molecular Systematics, 2nd Ed., Sinauer, Sunderland, USA, 1996, p. 407.
[8] C. Gallut, V. Barriel, Cladistic coding of genomic maps, Cladistics 18 (2002) 526.
[9] B.M.E. Moret, J. Tang, L.S. Wang, T. Warnow, Steps toward accurate reconstruction of phylogenies from gene-order data, J. Comput. Syst. Sci. 65 (3) (2002) 508.
[10] B.M.E. Moret, L.S. Wang, T. Warnow, S. Wyman, New approaches for reconstructing phylogenies based on gene order, in: Proc. 9th Int. Conf. on Intelligent Systems for Molecular Biology ISMB-2001, Bioinformatics, vol. 17, 2001, p. S165.
[11] A. Rokas, P.W.H. Holland, Rare genomic changes as a tool for phylogenetics, Trends Ecol. Evol. 15 (2000) 454.
[12] J.J. Nadeau, B.A. Taylor, Lengths of chromosome segments conserved since divergence of man and mouse, Proc. Nat. Acad. Sci. USA 81 (1984) 814.
[13] C. Semple, M. Steel, Tree reconstruction from multi-state characters, Adv. Appl. Math. 28 (2002) 169.
[14] M. Kimura, J. Crow, The number of alleles that can be maintained in a finite population, Genetics 49 (1964) 725.
[15] M. Steel, My favourite conjecture, http://www.math.canterbury.ac.nz/~mathmas/conjecture.pdf, 2001.
[16] E. Mossel, Phase transitions in phylogeny, Trans. Am. Math. Soc. (in press).
[17] P.M. Bleher, J. Ruiz, V.A. Zagrebnov, On the purity of limiting gibbs state for the Ising model on the bethe lattice, J. Stat. Phys. 79 (1995) 473.
[18] W. Evans, C. Kenyon, Y. Peres, L.J. Schulman, Broadcasting on trees and the Ising model, Ann. Appl. Prob. 10 (2) (2000) 410.
[19] E. Mossel, Recursive reconstruction on periodic trees, Random Struct. Algor. 13 (1998) 81.
[20] E. Mossel, On the impossibility of reconstructing ancestral data and phylogenies, J. Comput. Biol. 10 (2003) 669.
[21] P.L. Erdös, L.A. Székely, M. Steel, T. Warnow, A few logs suffice to build (almost) all trees (I), Random Struct. Algor. 14 (1999) 153.
[22] M. Steel, L.A. Székely, Inverting random functions (II): explicit bounds for discrete maximum likelihood estimation, with applications, SIAM J. Discr. Math. 15 (2002) 562.
[23] H. Colonius, H.H. Schulze, Tree structures for proximity data, Brit. J. Math. Stat. Psychol. 34 (1981) 167.
[24] H.-J. Bandelt, A.W.M. Dress, Reconstructing the shape of a tree from observed dissimilarity data, Adv. Appl. Math. 7 (1986) 309.
[25] M. Steel, L.A. Székely, Inverting random functions (I), Ann. Comb. 3 (1999) 103.
[26] K.T. Huber, V. Moulton, M. Steel. Four characters suffice to convexly define a phylogenetic tree. Research Report UCDMA2002/12, Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand, 2002.
[27] Schröder, E. Vier, Combinatorische probleme, Z. Math. Phys. 15 (1870) 361.
[28] N. Alon, J.H. Spencer, The Probabilistic Method, 2nd Ed., John Wiley, 2000.