



Sequence length bounds for resolving a deep phylogenetic divergence

Mareike Fischer, Mike Steel*

Allan Wilson Centre for Molecular Ecology and Evolution, Biomathematics Research Centre, University of Canterbury, Private Bag 4800, Christchurch, New Zealand

ARTICLE INFO

Article history:

Received 16 June 2008

Received in revised form

15 August 2008

Accepted 28 September 2008

Available online 11 October 2008

MSC:

05C05

92D15

Keywords:

Phylogenetic tree

DNA sequences

Markov process

Maximum parsimony

ABSTRACT

In evolutionary biology, genetic sequences carry with them a trace of the underlying tree that describes their evolution from a common ancestral sequence. The question of how many sequence sites are required to recover this evolutionary relationship accurately depends on the model of sequence evolution, the substitution rate, divergence times and the method used to infer phylogenetic history. A particularly challenging problem for phylogenetic methods arises when a rapid divergence event occurred in the distant past. We analyse an idealised form of this problem in which the terminal edges of a symmetric four-taxon tree are some factor (λ) times the length of the interior edge. We determine an order λ^2 lower bound on the growth rate for the sequence length required to resolve the tree (independent of any particular branch length). We also show that this rate of sequence length growth can be achieved by existing methods (including the simple ‘maximum parsimony’ method), and compare these order λ^2 bounds with an order λ growth rate for a model that describes low-homoplasy evolution. In the final section, we provide a generic bound on the sequence length requirement for a more general class of Markov processes.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

When sequence sites evolve independently under a Markov process along the branches of a tree \mathcal{T} , the sequences observed at the tips contain information concerning the underlying tree. This allows for the tree \mathcal{T} to be reconstructed accurately from sufficiently long sequences; this is the basis of modern molecular systematics (Felsenstein, 2003). The number of sites required to reconstruct \mathcal{T} accurately depends on how long the edges of the tree are. More precisely, it depends on the expected number of substitutions on each branch (edge) e of the tree—which we refer to as the *branch length* of e (this is the product of the temporal duration of the branch and the substitution rate).

A number of authors (e.g. Churchill et al., 1992; Lecointre et al., 1994; Saitou and Nei, 1986; Townsend, 2007; Wortley et al., 2005; Xia et al., 2003; Yang, 1998) have considered various ways to quantify the phylogenetic signal in aligned DNA sequences, and to estimate the sequence length required to reconstruct a phylogenetic tree. Most of these studies have involved simulation or heuristic approaches, although some analytical bounds have also been obtained (Mossel and Steel, 2005; Steel and Szekely, 2002). Typically, these bounds state that

if an interior branch length is very short, or if a terminal (external) branch length is long, then a large number of sites will be required.

In this paper we explore these results further by obtaining bounds that are expressed purely in terms of the relative sizes of the branch lengths, not their absolute values. One motivation for our approach is that different genes are known to evolve at different rates, so that any particular branch length will depend on which gene is considered; however, the ratios of the branch lengths will be unchanged if the gene-specific rate applies uniformly across the tree.

A particularly difficult tree reconstruction problem, requiring long sequences to resolve, arises when one has an interior edge with a short branch length incident with edges (or subtrees) having large branch lengths. Such a scenario occurs, for example, when speciation events in rapid succession (leading to short branch lengths) occurred in the distant past (leading to the large branch lengths for the incident edges). Several examples of this have been highlighted in the literature (Lockhart et al., 2006; Rokas and Carroll, 2006) and include the origin of metazoa and the origin of photosynthesis.

In this paper we analyse a scenario which, although somewhat idealised, nevertheless captures the essence of this problem—a four-taxon tree, where the terminal edges have equal branch lengths that are $\lambda > 1$ times the branch lengths of the interior edge, and a simple symmetric model of site evolution (specifically, we assume sites evolve independently according to a common two-state Markov process).

* Corresponding author. Tel.: +64 3 366 7001x7688; fax: +64 3 364 2587.

E-mail addresses: email@mareikefischer.de (M. Fischer), m.steel@math.canterbury.ac.nz (M. Steel).

We provide a mathematical analysis to the question of how many sites are required to resolve the tree correctly (from the three possible resolved topologies on four taxa). We are particularly interested in how the growth of the sequence length, k , depends on λ , independent of the absolute value of a particular edge length. We establish that k must grow at the rate λ^2 , which implies that regardless of how fast (or slow) any particular sequence is evolving, we can set explicit lower bounds on the length of sequences required to resolve the tree. We then show that for our setting, the growth in k need not be any worse than this quadratic growth in λ , because an existing method (namely, maximum parsimony) achieves this growth rate. This does not imply that maximum parsimony is the ‘best’ method for tree reconstruction; we chose it simply because we can analytically calculate tree reconstruction probabilities for this method. Our results complement an earlier simulation-based analysis (Yang, 1998). We contrast our results by considering a quite different model of site evolution (the infinite state model) and establishing that order λ growth in k can sometimes suffice for this model.

We also extend the approach to more general Markov processes on trees, obtaining exact, but less explicit lower bounds on k and which involve absolute (rather than relative) branch lengths. Our arguments are based on standard techniques from probability theory, such as central limit approximation, and information-theoretic arguments based on the properties of Hellinger distance.

2. Preliminaries

Consider an unrooted binary phylogenetic tree on four taxa, say 12|34, with branch length x for the interior edge e_5 and λx for the terminal edges e_1, \dots, e_4 , where $\lambda > 1$. This is illustrated in Fig. 1(a), and the topology of the tree is shown at the top of Fig. 1(b). The other two competing topologies (13|24 and 14|23) are also shown in Fig. 1(b). Here branch length refers to the expected number of substitutions under some continuous time substitution process.

Recall that a *binary character* or *site pattern* refers to an assignment to each taxon of a state from some two-element set, which we will denote through this paper as $\{\alpha, \beta\}$.

Suppose that a sequence of binary characters are generated independently and identically (i.i.d.) under a symmetric two-state model on the tree. This model is often called the CFN (Cavender–Farris–Neyman model) or more briefly the Neyman 2-state model (for more details see e.g. Semple and Steel, 2003). Although it is the simplest non-trivial Markov process on a tree, it allows for an exact analysis. Moreover, stochastic results for this model

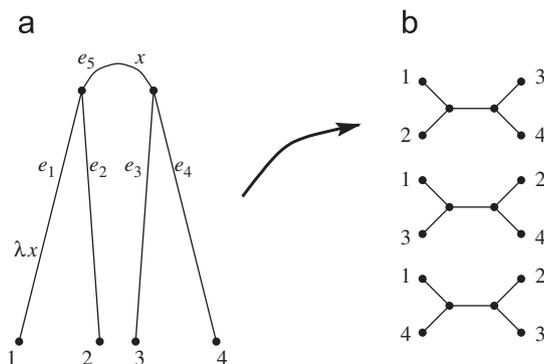


Fig. 1. (a) The generating tree with interior branch length x and all four terminal branch lengths equal to λx . (b) This tree has the topology 12|34, while the other two binary topologies are 13|24 and 14|23.

typically extend to more general finite-state models where an exact analysis is usually more complex (Mossel and Steel, 2005), and in Section 5 we show how some of our approaches extend to more general Markov processes.

If we denote the substitution probability on edge e_i by $P(e_i)$, then for each terminal edge we have $P(e_i) = \frac{1}{2}(1 - \exp(-2\lambda x))$ while for the central edge e_5 , we have $P(e_5) = \frac{1}{2}(1 - \exp(-2x))$. Let $\theta_i = 1 - 2P(e_i)$ for $i = 1, \dots, 5$. Then we can express these five θ_i values in terms of $\theta := e^{-2x}$ as follows:

$$\theta_i = \theta^{\lambda^2} \text{ for } i = 1, \dots, 4, \text{ and } \theta_5 = \theta.$$

Now, if we fix x and let λ grow, or, alternatively, if we fix λx and let x tend to zero, then it is easily shown that the sequence length k required to reconstruct the topology of the generating tree accurately tends to infinity. Informally, this is because under either of the two limiting situations described, the three trees in Fig. 1(b) will (in the limit) give the same probability distribution on site patterns, and so the three trees will describe any data equally well (a more formal proof can be provided by using Lemma 3.2). This holds for any tree reconstruction method that treats all three topologies fairly (if a method has an a priori preference for one topology, it will perform worse on an alternative topology). Moreover, if λx is fixed, then k grows at the rate $1/x^2$ as x tends to zero (by Theorem 4.1 of Steel and Szekely, 2002). However, if we do not fix x or λx in advance two fundamental questions arise: what is the slowest rate that k can possibly grow as a function of λ ? and (ii) does some value of x (dependent on λ) achieve this rate of growth for a certain tree reconstruction method? We will see that for the simple scenario described, the answers to these questions are (i) λ^2 and (ii) yes (up to a constant factor).

3. Lower bounds

The main result of this section is the following:

Theorem 3.1. *Suppose k sites evolve i.i.d. under a symmetric two-state model on some (unknown) four-taxon tree that has branch length x on the interior edge and λx on each terminal edge. Then any method that is able to correctly identify the underlying tree topology with probability at least $1 - \varepsilon$ requires*

$$k \geq c_\varepsilon \cdot \lambda^2$$

for any x , where $c_\varepsilon = \frac{1}{2}(1 - \frac{3}{2}\varepsilon)^2$.

To establish this result we require some preliminary results. We begin with a general information-theoretic bound on the number of i.i.d. observations required to reconstruct a discrete parameter in a general setting.

Suppose one has a finite set A , and each element $a \in A$ has an associated probability distribution on a finite set U . Suppose we observe k observations from U that are generated independently by the same unknown element $a \in A$. Suppose, furthermore, that some method M estimates the element of A that generated our observations and does so correctly with probability at least $1 - \varepsilon$ (regardless of which element a actually generated the data). Then we can set a lower bound on k in terms of a stochastic distance between elements of A . Recall that the *Hellinger distance* of two elements $a, a' \in A$ is defined as follows. If p and q denote the probability distribution induced by a and a' , respectively, then let

$$d_H^2(a, a') := \sum_{u \in U} (\sqrt{p_u} - \sqrt{q_u})^2 = 2 \left(1 - \sum_{u \in U} \sqrt{p_u q_u} \right). \tag{1}$$

The latter equality holds as $\sum_{u \in U} p_u = \sum_{u \in U} q_u = 1$. The following result is from Steel and Szekely (2002) (Theorem 3.1 and (2.7)).

Lemma 3.2. *If there is a subset A' of A of size $m \geq 2$ for which $d_H(a, a') \leq d$ for all $a, a' \in A'$ and some method M correctly identifies each element of A' with probability at least $1 - \varepsilon$ from k independently-generated elements in some set U , then*

$$k \geq \frac{1}{4} \left(1 - \frac{m}{m-1} \varepsilon\right)^2 d^{-2}.$$

In our setting, A will consist of the three binary four-taxon trees on leaf set $\{1, 2, 3, 4\}$, U will consist of the assignment of states of the elements of this leaf set, and m will be 3 (in this section) or 2 (in Section 5).

Let S be the set of possible binary site patterns on $\{1, 2, 3, 4\}$. These consist of the site patterns $s_1 := \alpha\alpha\beta\beta, s_2 := \alpha\beta\alpha\beta$ and $s_3 := \alpha\beta\beta\alpha$, and five non-informative ones s_4, \dots, s_8 (note that pairs of complementary site patterns—for example $\alpha\alpha\beta\beta$ and $\beta\beta\alpha\alpha$ —are regarded as equivalent). For any site pattern $s \in S$, let $p_s = \mathbb{P}(s|\mathcal{T}_1)$ (respectively, $q_s = \mathbb{P}(s|\mathcal{T}_2)$) be the probability that the site pattern s is generated on \mathcal{T}_1 (respectively, \mathcal{T}_2). We can express the probabilities p_{s_1} and p_{s_2} in terms of $\theta = e^{-2x}$ by using the Hadamard representation of [Hendy \(1989\)](#) (see [Semple and Steel, 2003](#), Section 8.6). We have

$$p_{s_1} = \frac{1}{8} \cdot (1 + 2 \cdot \theta^{2\lambda} - 4 \cdot \theta^{2\lambda+1} + \theta^{4\lambda}), \tag{2}$$

and

$$p_{s_2} = \frac{1}{8} \cdot (1 - 2 \cdot \theta^{2\lambda} + \theta^{4\lambda}) = \frac{1}{8}(1 - \theta^{2\lambda})^2. \tag{3}$$

To obtain an upper bound on the Hellinger distance for our problem, we require a further technical lemma.

Lemma 3.3. *Let $\gamma > 1$ and let $h(x) = x^\gamma(1-x)/(1-x^\gamma)$. Then the supremum of $h(x)$ for x in the half-open interval $[0, 1)$ equals $1/\gamma$.*

Proof. Since $\gamma > 1$ it can be checked that $h'(x) > 0$ for all x in $(0, 1)$, and so $\sup_{x \in [0, 1)} h(x) = \lim_{x \uparrow 1} h(x)$. By L'Hôpital's rule, we have $\lim_{x \uparrow 1} h(x) = 1/\gamma$. \square

Proof of Theorem 3.1. If any method has a probability of at least $1 - \varepsilon$ of correctly reconstructing each of the three binary trees on four taxa from i.i.d. sequences of length k then, by Lemma 3.2 with $m = 3$ we have

$$k \geq \frac{(1 - \frac{3}{2}\varepsilon)^2}{4} \cdot d_H^{-2}, \tag{4}$$

where d_H is the maximum Hellinger distance between any two of the three trees. Now, if each of the three trees has the $x, \lambda x$ combination of branch lengths (for interior, terminal branches, respectively) then, by symmetry, all three of these pairwise Hellinger distances are equal. Moreover, we claim that

$$d_H^{-2} \geq 2\lambda^2 \tag{5}$$

which together with (4) requires $k \geq c_\varepsilon \lambda^2$ for the choice of c_ε described. Thus it remains to establish (5).

Without loss of generality, $\mathcal{T}_1 = 12|34$ and $\mathcal{T}_2 = 13|24$. Now, for all $i = 3, \dots, 8$, we have $p_{s_i} = q_{s_i}$. Furthermore, $p_{s_1} = q_{s_2}$ and $p_{s_2} = q_{s_1}$, as the given trees are identical except for their leaf labelling. Consequently, Eq. (1) can be simplified as follows:

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) = 2 \left(1 - \sum_{i=1}^8 \sqrt{p_{s_i} q_{s_i}}\right) = 2 \left(1 - \sum_{i=3}^8 p_{s_i} - 2\sqrt{p_{s_1} p_{s_2}}\right) \tag{6}$$

$$= 2(1 - (1 - p_{s_1} - p_{s_2}) - 2\sqrt{p_{s_1} p_{s_2}}) \tag{7}$$

$$= 2(p_{s_1} + p_{s_2} - 2\sqrt{p_{s_1} p_{s_2}}). \tag{8}$$

Let $\delta = \frac{1}{2}\theta^{2\lambda}(1 - \theta)$. Then $p_{s_1} = p_{s_2} + \delta$, and so Eq. (8) can be re-written as

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) = 4p_{s_2} \left(1 + \frac{\delta}{2p_{s_2}} - \sqrt{1 + \frac{\delta}{p_{s_2}}}\right). \tag{9}$$

Applying the inequality $\sqrt{1+y} \geq 1 + y/2 - y^2/4$, for any $y > 0$, to $y = \delta/p_{s_2}$ in (9), gives

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq \frac{\delta^2}{p_{s_2}} = 2 \left[\frac{\theta^{2\lambda}(1 - \theta)}{1 - \theta^{2\lambda}}\right]^2 \leq \frac{1}{2\lambda^2},$$

where the last inequality follows by invoking Lemma 3.3 with $\gamma = 2\lambda, x = \theta$. This establishes (5) and thereby completes the proof of the theorem. \square

4. An upper bound: the performance of maximum parsimony

We now show that the lower bound described above is essentially 'best possible' (up to a constant factor) for the given model, as it can be achieved for a certain choice of x by a simple tree reconstruction method, namely maximum parsimony (MP). This method selects the tree that requires the smallest number of substitutions to extend the sequences at the tips of the tree to (ancestral) sequences at all the interior vertices of the tree (for further background, the reader can consult, for example, [Felsenstein, 2003](#) or [Semple and Steel, 2003](#)).

The probability that MP correctly reconstructs the true tree 12|34 will be called the *MP reconstruction probability*. Let $f(\varepsilon)$ denote the one-sided ε -critical value for the standard normal distribution, defined by

$$f(\varepsilon) = z \Leftrightarrow \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt = \varepsilon.$$

Theorem 4.1. *Suppose k sites evolve i.i.d. under a symmetric two-state model on some (unknown) four-taxon tree that has branch length x on the interior edge and λx on each terminal edge. Then for a sequence c'_i with $\lim_{i \rightarrow \infty} c'_i = 4e^2$, the following holds: If $k \geq c'_i f(\varepsilon/2)^2 \cdot \lambda^2$, an interior branch length x exists for which the MP reconstruction probability is at least $1 - \varepsilon$.*

In order to prove this theorem, some preliminary work is required. Suppose we generate a sequence \mathcal{C} of k i.i.d. sites under the symmetric two-state model. Define the random variables X_i and Y_k as follows. Let

$$X_i = \begin{cases} 1 & \text{if } i\text{th character in } \mathcal{C} \text{ is of the kind } (\alpha, \alpha, \beta, \beta), \\ -1 & \text{if } i\text{th character in } \mathcal{C} \text{ is of the kind } (\alpha, \beta, \alpha, \beta), \\ 0 & \text{else} \end{cases}$$

and let

$$Y_k = \sum_{i=1}^k X_i.$$

The probability that MP will favour the tree 12|34 over 13|24 is then $\mathbb{P}(Y_k > 0)$. We will exploit the fact that the random variables X_i are i.i.d., and so Y_k can be approximated for large k by a normal distribution with a mean μ_k and a standard deviation σ_k . These two parameters can be easily described (just) in terms of θ, λ and k as follows.

Lemma 4.2.

1. $\mu_k = k \cdot \frac{1}{2}\theta^{2\lambda}(1 - \theta)$.
2. $\sigma_k^2 = k \cdot \frac{1}{4}(1 + 2\theta^{4\lambda+1} - 2\theta^{2\lambda+1} - \theta^{4\lambda+2})$.
3. $\mu_k/\sigma_k \geq \sqrt{k} \cdot \theta^{2\lambda}(1 - \theta)$.

Proof. Since X_1, \dots, X_k are independent and take values $+1, 0$ and -1 , we have

- (i) $\mu_k = k \cdot [\mathbb{P}(X_1 = 1) - \mathbb{P}(X_1 = -1)]$;
- (ii) $\sigma_k^2 = k \cdot [\mathbb{P}(X_1 = 1) + \mathbb{P}(X_1 = -1) - [\mathbb{P}(X_1 = 1) - \mathbb{P}(X_1 = -1)]^2]$.

Now in the two-state symmetric model and the generating tree in Fig. 1(a), we have

$$\mathbb{P}(X_1 = 1) = p_{s_1} \quad \text{and} \quad \mathbb{P}(X_1 = -1) = p_{s_2},$$

where p_{s_1}, p_{s_2} were given above in Eqs. (2) and (3), respectively. Parts (1) and (2) of the lemma now follow by substitution of the expressions for p_{s_1}, p_{s_2} into (i) and (ii), respectively. For Part (3), note that Parts (1) and (2) imply that

$$\frac{\mu_k}{\sigma_k} = \sqrt{k} \cdot \frac{N_\theta}{D_\theta}, \tag{10}$$

where $N_\theta = \theta^{2\lambda}(1 - \theta)$; $D_\theta = \sqrt{1 + 2\theta^{4\lambda+1} - 2\theta^{2\lambda+1} - \theta^{4\lambda+2}}$. We now show that $D_\theta \leq 1$. We have $1 + 0.5\theta^{2\lambda+1} \geq \theta^{2\lambda}$ and so $2\theta^{2\lambda+1}(1 - \theta^{2\lambda} + 0.5\theta^{2\lambda+1}) \geq 0$. Thus $1 - 2\theta^{2\lambda+1}(1 - \theta^{2\lambda} + 0.5\theta^{2\lambda+1}) \leq 1$, which implies that $D_\theta^2 \leq 1$. Part (3) now follows from (10) by the inequality $D_\theta \leq 1$. \square

Proof of Theorem 4.1. Note that the MP reconstruction probability is the probability that MP will favour the true tree 12|34 over both alternative trees on four taxa, namely 13|24 and 14|23. Recall that the event of the tree 12|34 being favoured over 13|24 can be expressed as $\mathbb{P}(Y_k > 0)$. The event of 12|34 being favoured over 14|23 can be expressed similarly by defining the random variables \tilde{X}_i and \tilde{Y}_k which are analogous to X_i and Y_k , using the character $(\alpha, \beta, \beta, \alpha)$ instead of $(\alpha, \beta, \alpha, \beta)$. Then, the MP reconstruction probability can be written as $\mathbb{P}((Y_k > 0) \cap (\tilde{Y}_k > 0))$. Let

$$Z_k = \frac{Y_k - \mu_k}{\sigma_k}.$$

Thus, Z_k is the normalised difference of the parsimony score between tree 13|24 and 12|34 for a k i.i.d. characters generated by the tree in Fig. 1(a). By Lemma 4.2(3) we have

$$\mathbb{P}(Y_k \leq 0) = \mathbb{P}\left(Z_k \leq -\frac{\mu_k}{\sigma_k}\right) \leq \mathbb{P}(Z_k \leq -\sqrt{k}\theta^{2\lambda}(1 - \theta)). \tag{11}$$

Now, by symmetry of the branch length of the generating tree in Fig. 1(a), we have $\mathbb{P}(Y_k \leq 0) = \mathbb{P}(\tilde{Y}_k \leq 0)$. Moreover, by Boole's inequality

$$\mathbb{P}((Y_k > 0) \cap (\tilde{Y}_k > 0)) \geq 1 - \mathbb{P}(Y_k \leq 0) - \mathbb{P}(\tilde{Y}_k \leq 0),$$

which, combined with (11), furnishes the following inequality for the MP reconstruction probability:

$$\mathbb{P}((Y_k > 0) \cap (\tilde{Y}_k > 0)) \geq 1 - 2\mathbb{P}(Y_k \leq 0) \geq 1 - 2\mathbb{P}(Z_k \leq -\sqrt{k}\theta^{2\lambda}(1 - \theta)). \tag{12}$$

Now, $\theta^{2\lambda} \cdot (1 - \theta)$ has a unique local maximum in $[0, 1]$, namely at $\theta' := 1 - (1/2\lambda + 1)$, at which it takes the value α_λ/λ , where $\alpha_\lambda = (1 - 1/(1 + 2\lambda))^{2\lambda} \cdot \lambda/(1 + 2\lambda) \rightarrow \frac{1}{2}e^{-1}$ as $\lambda \rightarrow \infty$. Moreover, the difference between the distribution of Z_k and a standard normal distribution tends uniformly to zero as λ (and hence k) grows. This follows by applying standard bounds on the central limit theorem approximation (see, for example, Zahl, 1966; one cannot directly apply the usual form of the central limit theorem as the distribution of the X_i 's is changing with increasing λ). Thus we have $\mathbb{P}(Z_k \leq -\sqrt{k}\alpha_\lambda/\lambda) \leq \varepsilon/2$ provided that k grows at the rate $c'_\lambda \lambda^2 f(\varepsilon/2)^2$ for a sequence $c'_\lambda \rightarrow 4e^2$ as $\lambda \rightarrow \infty$.

In summary, by (12), a value for θ exists, namely $\theta' = 1 - (1/(1 + 2\lambda))$, and thus a value for $P(e_5) = \frac{1}{2}(1 - \theta') = 1/2(1 + 2\lambda) \sim 1/4\lambda$ also exists, for which the MP reconstruction probability is at least $1 - \varepsilon$. This completes the proof. \square

4.1. Remarks

- Regarding Theorem 4.1, other tree reconstruction methods have a similar performance to MP when k grows at the rate λ^2 . Indeed it is possible that such methods will require shorter sequences, and better statistical properties on trees with different tree shapes (as MP is statistically inconsistent under some combinations of branch lengths that lie outside those considered in the scenario of Fig. 1). We have chosen to consider MP here, because the analysis is relatively straightforward and it suffices to prove the matching lower λ^2 bound.
- One can also derive a (non-asymptotic) form of Theorem 4.1 using Azuma's inequality (Alon and Spencer, 2000); however, the term in place of c'_λ is larger by a factor of 32.
- The optimal choice of x of (approximately) $1/4\lambda$ for MP has been observed in a slightly different setting by Townsend (2007).
- One can ask whether similar λ^2 bounds on k as given by Theorem 3.1 will apply for more complex models. We conjecture that for stationary, reversible, finite-state Markov processes, the results will be essentially the same for our tree in Fig. 1, up to a possibly different constant term in place of c_ε .
- For Markov processes in which the state space is countably infinite—and where a substitution is always to a new state (the 'random cluster model' for homoplasy-free evolution, described in Mossel and Steel (2004))—the situation regarding sequence length requirements is quite different. In this case, the required sequence length need only grow at the rate λ (not λ^2), as the following result shows.

Proposition 4.3. *Suppose k sites evolve i.i.d. under a random cluster model on some (unknown) four-taxon tree that has branch length x on the interior edge and λx on each terminal edge. Then for a constant C'_ε which depends just on ε , the following holds: If $k \geq C'_\varepsilon \cdot \lambda$, an x exists for which the MP reconstruction probability is at least $1 - \varepsilon$.*

Proof. In the random cluster model, the probability of a substitution event on an edge e can be written as $P(e) = 1 - \exp(-l)$ where l is the expected number of changes on the edge (the branch length). Now, the random cluster model only generates characters that are homoplasy-free on the generating tree; thus, MP will return the generating tree from a sequence of characters, provided this tree is the only one on which those characters are homoplasy-free. For a tree with topology 12|34, this will occur precisely if at least one of the k characters generated assigns taxa 1, 2 a shared state, and taxa 3, 4 a second shared state that is different to that assigned to 1, 2. The probability Q that any given character generated by the tree in Fig. 1(a) has this property is given by

$$Q = P(e_5) \prod_{i=1}^4 (1 - P(e_i)) = (1 - e^{-x})e^{-4\lambda x}.$$

Moreover, if $k \geq \log(\frac{1}{\varepsilon})/Q$ then $1 - (1 - Q)^k \geq 1 - \varepsilon$ (using the inequality $-\log(1 - Q) \geq Q$). Consequently, MP will correctly reconstruct the generating tree with probability at least $1 - \varepsilon$ provided that

$$k \geq \log(\varepsilon^{-1}) \cdot (1 - e^{-x})^{-1} e^{4\lambda x}. \tag{13}$$

Taking $x = 1/4\lambda$ we have $(1 - e^{-x})^{-1} e^{4\lambda x} \sim 4e\lambda$, which, in view of (13), establishes the result. \square

5. Lower bounds for more general models

In this section we derive a lower bound on the sequence length required for tree reconstruction, for a much wider range of

Markov processes. However, unlike the previous sections our bound is expressed in terms of the absolute branch lengths (or bounds on these) rather than in terms of ratios, and it involves constants that depend on the details of the model.

We first derive a general lemma. Consider any continuous-time, stationary and reversible Markov process. Let \mathcal{S} denote its state space, and in keeping with earlier terminology let $S = \mathcal{S}^4$ (thus in previous sections $\mathcal{S} = \{\alpha, \beta\}$). Let \mathcal{T}_1 and \mathcal{T}_2 be two topologically distinct four-taxon trees. Suppose that the branch lengths of \mathcal{T}_1 are arbitrary, and that each edge of \mathcal{T}_2 has the corresponding interior or pendant branch length specified by \mathcal{T}_1 (where the pendant edge incident with leaf i in \mathcal{T}_1 corresponds to the pendant edge incident with leaf i in \mathcal{T}_2). For $s = (s_1, s_2, s_3, s_4) \in S$, let p_s (respectively, q_s) denote the probability of generating s at the tips of \mathcal{T}_1 (respectively, \mathcal{T}_2). Let p'_s (respectively, q'_s) denote the conditional probability of generating s at the tips of \mathcal{T}_1 (respectively, \mathcal{T}_2) given that a substitution has occurred on the central edge of \mathcal{T}_1 (respectively, \mathcal{T}_2), and let $D_s := q'_s - p'_s$. Then we have the following result.

Lemma 5.1.

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq l^2 \cdot \sum_{s \in S} \frac{D_s^2}{p_s}$$

where l denotes the branch length of the interior edge of \mathcal{T}_1 .

Proof. Let τ denote the probability that at least one substitution occurs on the interior edge of \mathcal{T}_1 , and let p_s^0 (respectively, q_s^0) denote the conditional probability of generating s on \mathcal{T}_1 (respectively, \mathcal{T}_2) given that no substitution occurs on the interior edge of \mathcal{T}_1 (respectively, \mathcal{T}_2). By the law of total probability we have

$$p_s = (1 - \tau) \cdot p_s^0 + \tau \cdot p'_s$$

and

$$q_s = (1 - \tau) \cdot q_s^0 + \tau \cdot q'_s.$$

Moreover, the assumptions on the correspondence between branch lengths of \mathcal{T}_1 and \mathcal{T}_2 imply that $p_s^0 = q_s^0$ for all $s \in S$ and so

$$q_s - p_s = \tau(q'_s - p'_s) = \tau D_s.$$

Now,

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) = 2 \left(1 - \sum_{s \in S} \sqrt{p_s q_s} \right) = 2 \left(1 - \sum_{s \in S} p_s \sqrt{1 + \frac{\tau D_s}{p_s}} \right).$$

Applying the inequality $\sqrt{1+y} \geq 1 + y/2 - y^2/2$ (for all $y \geq -1$) to $y = \tau D_s/p_s$ (and observing that $y \geq -1$ since $q_s \geq 0$), we obtain

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq 2 \left(1 - \sum_s p_s \left(1 + \tau \frac{D_s}{2p_s} - \tau^2 \frac{D_s^2}{2p_s} \right) \right).$$

Now, $\sum_s p_s = 1$, and $\sum_s D_s = 0$ (since $\sum_s q'_s = \sum_s p'_s = 1$) and so this last inequality reduces to

$$d_H^2(\mathcal{T}_1, \mathcal{T}_2) \leq \tau^2 \cdot \sum_{s \in S} \frac{D_s^2}{p_s}. \tag{14}$$

Furthermore, $\tau = \mathbb{P}(N > 0)$, where N is the number of substitutions occurring on the interior edge of \mathcal{T}_1 . However, $\mathbb{P}(N > 0) \leq \mathbb{E}(N)$; that is, $\tau \leq l$, which, together with (14), provides the inequality stated in the lemma. \square

We now apply this lemma to a slightly more restricted class of Markov processes to obtain the main result of this section.

Theorem 5.2. Suppose k sites evolve i.i.d. under a finite-state, stationary and reversible continuous-time Markov process in which

each state is accessible from any other state. Let l_0 be any strictly positive value. Consider this process on some (unknown) four-taxon tree that has branch length at most l on the interior edge and at least $L \geq l_0$ on each terminal edge. Then any method that is able to correctly identify with probability at least $1 - \varepsilon$ the underlying tree topology given these restrictions requires

$$k \geq \frac{C}{4} (1 - 2\varepsilon)^2 \cdot \frac{e^{cL}}{l^2},$$

where c and C are positive constants that depend only on R (the rate matrix for the process) and l_0 .

Proof. We exploit the fact that any Markov process of the type described converges to its unique stationary distribution at an exponential rate (see, for example, Theorem 8.3 of Rozanov, 1969). Let $\pi(s)$ denote the stationary probability of s under the model. For $j = 1, \dots, 4$, let $p(j) \in \{u, v\}$ be the end of the interior edge uv of \mathcal{T}_1 that is adjacent to leaf j (we may assume $p(1) = p(2) = u; p(3) = p(4) = v$), and let $S_{p(j)}$ denote the random state present at that vertex under the model. Then for any $s_j, s'_j \in \mathcal{S}$ there exist positive constants A, a (dependent on R) for which

$$|\mathbb{P}(S_j = s_j | S_{p(j)} = s'_j) - \pi(s_j)| \leq Ae^{-al_j} \tag{15}$$

(Rozanov, 1969, Theorem 8.3), where l_j denotes the branch length of the edge incident with leaf j . For $s = (s_1, s_2, s_3, s_4) \in S = \mathcal{S}^4$, let

$$\pi_s = \prod_{j=1}^4 \pi(s_j).$$

For $s', s'' \in \mathcal{S}$ let $p'(s', s'')$ denote the probability of generating state s' at u and the state s'' at v given that at least one substitution occurs on the edge uv . Then, by the Markov assumption, and recalling the definition of p'_s from Lemma 5.1, we have

$$p'_s = \sum_{(s', s'') \in \mathcal{S}^2} p'(s', s'') \cdot \prod_{j=1}^2 \mathbb{P}(S_j = s_j | S_u = s') \cdot \prod_{j=3}^4 \mathbb{P}(S_j = s_j | S_v = s''). \tag{16}$$

Combining (15) and (16), there exist positive constants B, b (dependent only on R) such that

$$|p'_s - \pi_s| \leq Be^{-bl} \tag{17}$$

for all $s \in S$ (recall that $L \leq l_j$ for all j). Now, consider tree \mathcal{T}_2 which has branch lengths that correspond to those in \mathcal{T}_1 (as in Lemma 5.1). Then we also have

$$|q'_s - \pi_s| \leq Be^{-bl} \tag{18}$$

for all $s \in S$. Combining (17) and (18) using the triangle inequality gives

$$|D_s| = |q'_s - p'_s| \leq 2Be^{-bl}. \tag{19}$$

Moreover, since $l_j \geq l_0$ (for all j) and each state is accessible from any other state, we have $p_s \geq \delta$ (for some $\delta > 0$ dependent only on R and l_0). Combining this with (19) gives the following inequality, for all $s \in S$:

$$\frac{D_s^2}{p_s} \leq (4B^2/\delta)e^{-2bl}. \tag{20}$$

The theorem now follows from Lemmas 5.1 and 3.2 (with $m = 2$). \square

6. Concluding remarks

In this paper we have provided precise results for a specific and simple model (the two-state symmetric process), along with less explicit results for more general Markov processes (and phrased in

terms of absolute rather than relative branch lengths). The aim is to determine rigorous bounds on the sequence length required for resolving a deep divergence, which may shed light on debates as to whether some early radiations might be fundamentally unresolvable on the basis of current models and data.

Of course, in applications, other phenomena may further impede phylogenetic reconstruction (including substitution model mis-specification, lineage sorting and alignment artifacts, Philippe et al., 2005), however, these errors are unlikely to help tree reconstruction if our bound shows it is impossible even when the ideal model assumptions hold. We have seen that some models require significantly fewer characters for resolving a tree—in particular this holds for the random cluster model, and it is possible that new types of genomic data (involving rare genomic events where homoplasy is unlikely) can be described by these and related processes that preserve more phylogenetic signal regarding distant evolutionary divergences.

One limitation concerning our bounds is that they apply to pure Markov processes, in which each character evolves according to the same process. In molecular biology a common assumption is that there is a distribution of rates across sites, in which each site evolves at a rate (selected i.i.d. from some distribution) that acts as a multiplier for all the branch lengths in the tree (see e.g. Felsenstein, 2003; Semple and Steel, 2003). It would be interesting to extend the analysis in the last section to these models to obtain a lower bound on k analogous to Theorem 5.2.

Acknowledgements

We thank the *Allan Wilson Centre for Molecular Ecology and Evolution* for funding this work. We also thank the two anonymous referees for their helpful comments.

References

- Alon, N., Spencer, J.H., 2000. *The Probabilistic Method*. Wiley, New York.
- Churchill, G., von Haeseler, A., Navidi, W., 1992. Sample size for a phylogenetic inference. *Mol. Biol. Evol.* 9 (4), 753–769.
- Felsenstein, J., 2003. *Inferring Phylogenies*. Sinauer Press.
- Hendy, M.D., 1989. The relationship between simple evolutionary tree models and observable sequence data. *Syst. Zool.* 38, 310–321.
- Lecointre, G., Philippe, H., Van Le, H.L., Le Guyader, H., 1994. How many nucleotides are required to resolve a phylogenetic problem? The use of a new statistical method applicable to available sequences. *Mol. Phyl. Evol.* 3 (4), 292–309.
- Lockhart, P.J., Novis, P., Milligan, B.G., Riden, J., Rambaut, A., Larkum, T., 2006. Heterotachy and tree building: a case study with plastids and eubacteria. *Mol. Biol. Evol.* 23 (1), 40–45.
- Mossel, E., Steel, M., 2004. A phase transition for a random cluster model on phylogenetic trees. *Math. Biosci.* 187, 189–203.
- Mossel, E., Steel, M., 2005. How much can evolved characters tell us about the tree that generated them? In: Olivier Gascuel (Ed.), *Mathematics of Evolution and Phylogeny*. Oxford University Press, Oxford, pp. 384–412.
- Philippe, H., Delsuc, F., Brinkmann, H., Lartillot, N., 2005. *Phylogenomics*. *Annu. Rev. Ecol. Evol. Syst.* 36, 541–562.
- Rokas, A., Carroll, S.B., 2006. Bushes in the tree of life. *PLoS Biol.* 4 (11), e352.
- Rozanov, Y.A., 1969. *Probability Theory: A Concise Course*. Dover Publications, New York.
- Saitou, N., Nei, M., 1986. The number of nucleotides required to determine the branching order of three species, with special reference to the human-chimpanzee-gorilla divergence. *J. Mol. Evol.* 24, 189–204.
- Semple, C., Steel, M., 2003. *Phylogenetics*. Oxford Lecture Series in Mathematics and its Applications. Oxford University Press, Oxford.
- Steel, M., Székely, L., 2002. Inverting random functions II: explicit bounds for discrete maximum likelihood estimation, with applications. *SIAM J. Discrete Math.* 15 (4), 562–575.
- Townsend, J., 2007. Profiling phylogenetic informativeness. *Syst. Biol.* 56 (2), 222–231.
- Wortley, A.H., Rudall, P.J., Harris, D.J., Scotland, R.W., 2005. How much data are needed to resolve a difficult phylogeny? Case study in Lamiales. *Syst. Biol.* 54 (5), 696–709.
- Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its applications. *Mol. Phyl. Evol.* 26, 1–7.
- Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.* 47 (1), 125–133.
- Zahl, S., 1966. Bounds for the central limit theorem error. *SIAM J. Appl. Math.* 14 (6), 1225–1245.