CAPTURING A PHYLOGENETIC TREE WHEN THE NUMBER OF CHARACTER STATES VARIES WITH THE NUMBER OF LEAVES

MIKE STEEL

Abstract. We show that for any two values \(\alpha, \beta \in (0, 1)\) for which \(\alpha + \beta > 1\) there exists a value \(N\) so that for all \(n \geq N\), and any binary phylogenetic tree \(T\) on \(n\) leaves there exists a set of at most \(n^\alpha\) characters that capture \(T\), and for which each character has at most \(n^\beta\) states. Here ‘capture’ means that \(T\) is the only phylogenetic tree that is compatible with the characters. Our short proof of this combinatorial result is based on the probabilistic method.

Suppose that \(k\) characters on \(n\) taxa captures some phylogenetic tree \(T\) (i.e. the characters are compatible with \(T\) and no other tree; thus \(T\) is a binary unrooted tree with the \(n\) taxa as its leaves). If each of the \(k\) characters has at most \(r\) states, then a simple and known inequality states that \(k\) must be at least \((n - 3)/(r - 1)\) (Proposition 4.2 of [5]). Remarkably, this lower bound was recently shown [2] to be sharp for every value of \(r \geq 2\), provided that \(n \geq n_0\), where \(n_0\) is some (increasing) function of \(r\). In this note, we consider how small \(k\) can be when \(r\) is allowed to depend on \(n\). From [1, 3] it is known that \(k = 4\) holds for a certain series \(r_n\) for which \(n/r_n = O(1)\), so we focus on the setting where both \(r_n\) and \(n/r_n\) tend to infinity with increasing \(n\). We consider what happens when \(r_n\) grows as a sub-linear function of \(n\), by constraining \(r_n\) to be less or equal to an \(n^\alpha\) for \(\alpha \in (0, 1)\), in which case the inequality \(k \geq (n - 3)/(r - 1)\) implies that \(k\) must exceed \(n^\frac{\beta}{2}\) for \(\beta = 1 - \alpha\), for \(n\) sufficiently large. The following result is independent of the result from [2] mentioned above, in the sense that neither result directly implies the other. Our short proof involves a simple application of the probabilistic method, the Chernoff bound, and a property of the random cluster model on trees established in [4].

Theorem 1. For any two values \(\alpha, \beta \in (0, 1)\) for which \(\alpha + \beta > 1\) there exists a value \(N\) so that for all \(n \geq N\), and any unrooted binary phylogenetic tree \(T\) on \(n\) leaves there exists a set of at most \(\lfloor n^\alpha \rfloor\) characters that capture \(T\), and for which each character has at most \(r_n = \lfloor n^\beta \rfloor\) states.

Proof. Let \(X\) denote the leaf set of \(T\). Consider the random cluster model on \(T\) in which each edge of \(T\) is independently cut with probability \(p_n = r_n/4n\), or left intact with probability \(1 - p_n\). This leads to a partition of \(X\) corresponding to the equivalence relation that two leaves are related if and only if they lie in the same connected component of the resulting graph. We will regard such a partition as equivalent to a character (with the number of ‘states’ of the character being the number of blocks of the partition). Notice that \(\lim_{n \to \infty} p_n = 0\). Let \(Y\) denote the random number of edges of \(T\) that are cut. Then \(Y\) has a binomial distribution \(Y \sim Bin(2n - 3, p_n)\), which has mean \(\mu_n = (2n - 3)p_n = (\frac{1}{2} - o(1))n^\beta\). By a multiplicative form of the ‘Chernoff bound’ in probability theory, \(\Pr(Y \geq 2\mu_n) \leq \exp(-4\mu_n/3)\) and since \(r_n > 2\mu_n\) we obtain:

\[
\Pr(Y \geq r_n) \leq \exp(-4\mu_n/3). \tag{1}
\]

The number of blocks of the partition of \(X\) induced by randomly cutting edges of \(T\) in this way is at most \(Y + 1\). Thus, the probability that a character, generated by the random cluster model with \(p_n\) value as specified, has strictly more than \(r_n\) states is at most \(\Pr(Y + 1 > r_n) = \Pr(Y \geq r_n) \leq \exp(-4\mu_n/3)\), by (1). Thus if we generate a set \(S_n\) of \(\lfloor n^\alpha \rfloor\) such characters the probability that at least one of these characters has more than \(r_n\) states is, by Boole’s inequality,
at most
\[ n^\alpha \exp(-4\mu_n/3) = n^\alpha \exp\left(-\frac{4}{3} - o(1)n^\beta\right) \to 0 \]
as \( n \to \infty \) (recall \( \beta > 0 \)). Thus, there exists some value \( N_1 \) for which, for any \( n \geq N_1 \), with probability at most \( 1/3 \), at least one character in \( S_n \) has more than \( r_n \) states.

Now, suppose that \( k \) i.i.d. characters are generated under the random cluster model on \( T \) with \( p_n \), as specified above. Then, from Lemma 2.2 and Theorem 2.4 of [4], this set of characters captures \( T \) with probability at least \( 1 - \varepsilon \) provided that
\[ k = \left\lceil \frac{1}{B} \log(n^2/\varepsilon) \right\rceil, \]
where \( B = p_n(2 - 1/p_n)^4 \sim p_n \). Now,
\[ \frac{1}{B} \sim \frac{1}{p_n} = \frac{4n}{r_n} \sim \frac{4n}{n^\beta} = 4n^{1-\beta}, \]
and since \( \alpha + \beta > 1 \) it follows that for any \( \varepsilon > 0 \):
\[ \frac{1}{B} \log(n^2/\varepsilon)/n^\alpha \sim 4n^{1-\beta} \log(n^2/\varepsilon)/n^\alpha \to 0 \quad \text{as} \quad n \to \infty. \]

Consequently, taking \( \varepsilon = 1/3 \), there is a value \( N_2 \) for which, for any \( n \geq N_2 \), we have \( \left\lceil \frac{1}{B} \log(n^2/\varepsilon) \right\rceil \leq \lceil n^\alpha \rceil \) for all \( n \geq N_2 \). Thus, if \( k = \lfloor n^\alpha \rfloor \) where \( n \geq N_2 \), the i.i.d. characters generated under the above process fails to capture \( T \) with probability at most \( 1/3 \).

Combining these two observations, if we set \( N = \max\{N_1, N_2\} \) then for all \( n \geq N \), a set \( S_n \) of \( \lfloor n^\alpha \rfloor \) randomly-generated characters fails to satisfy one or other (or both) of the following properties:

(i) \( S_n \) has no character with more than \( r_n \) states, and

(ii) \( S_n \) captures \( T \),

with probability at most \( 1/3 + 1/3 = 2/3 \), by Boole’s inequality. Thus there is a strictly positive probability that \( S_n \) will satisfy both of (i) and (ii), and so there must exist a set of at most \( \lfloor n^\alpha \rfloor \) characters, each having at most \( \lfloor n^\beta \rfloor \) states, which captures \( T \). This completes the proof. \( \square \)

Remark: Notice from the proof, that the condition \( \alpha + \beta > 1 \) can be replaced by \( |\alpha + \beta| = 1 \) if we allow \( \lfloor n^\alpha \rfloor \) characters to be replaced by \( \lfloor n^\alpha \rfloor \lfloor (8 + c) \log(n) \rfloor \), for any \( c > 0 \). However, Theorem 1 fails when \( \alpha + \beta = 1 \) without this adjustment to the number of characters, since \( (n - 3)/(r_n - 1) > n^\alpha \) for all sufficiently large \( n \), when \( r_n = \lfloor n^\beta \rfloor \), and \( \alpha + \beta = 1 \), and yet at least \( (n - 3)/(r_n - 1) \) characters are required to capture \( T \) by Proposition 4.2 of [5].

References


Biomathematics Research Centre, University of Canterbury, Christchurch, NZ

E-mail address: mike.steel@canterbury.ac.nz