Applied Mathematics Letters 22 (2009) 1778–1780

Contents lists available at ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

An improved bound on the maximum agreement subtree problem

Mike Steel [a], László A. Székely [b,*]

[a] Biomathematics Research Centre, University of Canterbury, New Zealand

[b] Department of Mathematics, University of South Carolina, United States

ARTICLE INFO

Article history:
Received 19 March 2009
Received in revised form 22 June 2009
Accepted 22 June 2009

Keywords:
Phylogeny
Phylogenetic tree
Maximum agreement subtree
Binary tree
Caterpillar
Cyclic order

ABSTRACT

We improve the lower bound on the extremal version of the maximum agreement subtree problem. Namely we prove that two binary trees on the same n leaves have subtrees with the same $\geq c \log\log n$ leaves which are homeomorphic, such that homeomorphism is the identity on the leaves.

A *phylogenetic X-tree* is a binary tree in which the leaves are labelled bijectively with labels from a set X (usually $\{1, 2, \ldots, n\}$) and internal vertices are unlabelled. Two phylogenetic X-trees are considered the same if there is a label-preserving graph isomorphism between them.

If T is phylogenetic X-tree and $Y \subseteq X$ is a set of labels, then the *induced binary subtree* $T|_Y$ is defined as follows: (a) take the subtree induced by Y in T, and (b) substitute paths in which all internal vertices have degree 2 by edges. $T|_Y$ is a phylogenetic Y-tree (see Fig. 1).

If $|Y| = 4$, the induced binary subtree is often identified with an unordered partition of Y into two two-element sets, obtained by removing the (unique) internal edge of $T|_Y$. This partition is known as a *quartet split*. It is known that the $\binom{n}{4}$ quartet splits of a phylogenetic X-tree with $|X| = n$ determine the phylogenetic tree through a polynomial time algorithm. This was first observed in 1981 by Colonius and Schultze [1], in the context of stemmatology, and was developed further in 1986 by Bandelt and Dress [2].

An important algorithmic problem, known as the *maximum agreement subtree problem*, is the following: given two phylogenetic X-trees, find a common induced binary subtree of the largest possible size.

This problem has a history that spans more than 25 years, from papers in the early 1980s by Gordon [3], and Finden and Gordon [4], to its implementation in the late 1990s in the widely used phylogenetic software PAUP [5]. Somewhat surprisingly, this problem can be solved in polynomial time [6] (see also [7,8]).

Here we focus on the extremal version of the problem, where we ask how small a maximum agreement subtree can possibly be. This is motivated by a recent approach to tree comparison [9] based on the distribution of the maximum agreement subtree for pairs of randomly selected trees—thus we are interested here in the range of this distribution.

Let mast(n) denote the smallest order (number of leaves) of the maximum agreement subtree of two phylogenetic X-trees with $|X| = n$. In 1992, Kubicka, Kubicki, and McMorris [10] showed that $c_1(\log\log n)^{1/2} < \text{mast}(n) < c_2 \log n$ with some explicit constants.

* Corresponding author. Tel.: +1 803 749 4137.
E-mail addresses: M.Steel@math.canterbury.ac.nz (M. Steel), szekely@math.sc.edu (L.A. Székely).

0893-9659/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aml.2009.06.019

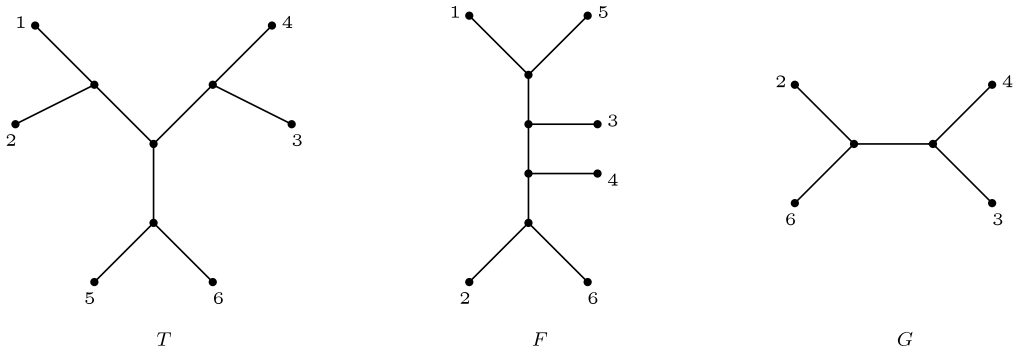

Fig. 1. For $X = \{1, 2, 3, 4, 5, 6\}$ and the two phylogenetic X-trees shown (T and F), a maximum agreement subtree is the phylogenetic tree $G = T|_Y = F|_Y$ shown, where $Y = \{2, 3, 4, 6\}$.

The purpose of our note is to remove the square root sign from the lower bound. This is achieved by changing the order of two combinatorial steps, one resulting in taking the logarithm twice, the other taking a square root. Of course, the square root sign after the log log is no longer visible.

First we would like to exhibit a direct connection to Ramsey theory, which might explain the large gap between the lower and upper bounds for mast(n). Let $R_2^k(n, \ell)$ denote the smallest integer m such that for any coloration of the k-element subsets of any m-element set with colors red and blue, there exists an n-element subset of the m-element set such that every k-element subset of the n-element set is colored red, or there exists an ℓ-element subset of the m-element set such that every k-element subset of the ℓ-element set is colored blue (see Chapter 14 in [11]).

Claim. $\text{mast}[R_2^4(n, 6)] \geq n$.

Proof. We first recall an observation from [2] that for $|X| = 6$, any two phylogenetic X-trees share a quartet split. Given T and F arbitrary phylogenetic X-trees with $|X| = R_2^4(n, 6)$, color 4-subsets of X red if they define the same quartet split, otherwise blue. No six elements of X can have all 4-subsets blue by the previous reference, so there are n elements from X such that all their 4-subsets are colored red. As the binary tree is determined by its quartet splits, these n elements span a size n agreement subtree, thereby establishing the claim. □

This approach would give an explicit lower bound for mast(n) in the form of a multiply iterated logarithm, much weaker than $c_1(\log\log n)^{1/2}$.

Before proving our result, we quickly show $c_1(\log\log n)^{1/2} < \text{mast}(n)$ following the approach in the 1992 paper by Kubicka, Kubicki, and McMorris [10]. Recall that a *caterpillar* is a tree, which has a path such that every leaf has a neighbor on the path (for example, the tree F in Fig. 1). Let us be given two phylogenetic X-trees T and F with $|X| = n$. As our trees are binary, the diameter of T is at least $c_3 \log n$. Therefore T must have an induced binary caterpillar subtree with leaf set Y such that $|Y| \geq c_3 \log n$. Consider the induced binary subtree $F|_Y$, which must have diameter $\geq c_4 \log\log n$. Like we argued before, there should be a $Z \subseteq Y$ such that $F|_Z$ is a caterpillar and $|Z| \geq c_4 \log\log n$. Notice that $T|_Z = (T|_Y)|_Z$ is also a caterpillar. Recall the Erdős–Szekeres Theorem (Ex. 14.15 in [11]) for sequences: two sequences composed from the same $k^2 + 1$ items have either a common $k + 1$-length subsequence, or they have a common $k + 1$-length subsequence after reversing the order in one sequence. As caterpillar trees can be understood as sequences of their leaves, two caterpillar trees with the same $k^2 + 1$ leaves contain size $k + 1$ agreement subtrees. Apply this with the largest k such that $k^2 + 1 \leq c_4 \log\log n$.

Before turning to our main result, we need some definitions. We say that a phylogenetic X-tree T is *drawn on the plane* if it is drawn as a plane graph. The *circumference of a phylogenetic tree drawn on the plane* is the cyclic permutation of X, the leaf set, as we walk around T clockwise. This concept has been a useful combinatorial tool elsewhere (see, for example, [12]) and we illustrate it here in Fig. 1 by noting that the circumference of this drawing of T is the cyclic permutation (1, 4, 3, 6, 5, 2).

Note that for $Y \subseteq X$, the induced (by Y) binary subtree of T has a natural drawing following steps (a) and (b) by deleting edges and vertices from the plane drawing, and then removing the vertex designation of vertices of degree 2, but keeping the curve representing the path for representing the new edge. For this natural drawing of $T|_Y$, the circumference is the circumference of the drawing of T restricted to Y. For the tree T in Fig. 1 and the subset $Y = \{2, 3, 4, 6\}$, the circumference of the induced drawing of $T|_Y$ is the cyclic permutation (2, 4, 3, 6) (the same as the circumference of the given drawing of G), while the circumference of the induced drawing of $F|_Y$ is the cyclic permutation (2, 3, 4, 6).

Theorem 1. *For a constant $c > 0$, we have*

$$\text{mast}(n) > c \log\log n.$$

Proof. Take two arbitrary phylogenetic X-trees, T and F, with $|X| = n$ and draw them in the plane. Cut the resulting circumferences anywhere to obtain two (linear) permutations of X. By the Erdős–Szekeres Theorem, there is subset $U \subseteq X$ such that the two permutations put U either into the same linear order, or into the opposite linear order, and $|U| \geq c_5 n^{1/2}$.

Like in the proof explained before the theorem, $T|_U$ has diameter $\geq c_3 \log |U| \geq c_6 \log n$. Therefore $T|_U$ has an induced binary subtree which is a caterpillar with leaf set V such that $|V| \geq c_6 \log n$. Consider now the induced binary subtree $F|_V$. The diameter of $F|_V$ is at least $c_3 \log |V| \geq c_7 \log \log n$, and therefore there should be a $Z \subseteq V$ such that $F|_Z = (F|_V)|_Z$ is a caterpillar and $|Z| \geq c_7 \log \log n$. Both $T|_Z$ and $F|_Z$ are caterpillars. By the choice of U, these two caterpillars have the same or mirror image circumferences. In the second case, starting this proof with the mirror image of the drawing of F, we can make sure that the caterpillars $T|_Z$ and $F|_Z$ have identical circumferences. Taking the longest path from $T|_Z$ (resp. $F|_Z$), this path partitions the $|Z| - 2$ non-endpoint leaves of $T|_Z$ (resp. $F|_Z$) into two classes corresponding to the two sides. We have two 2-partitions of $|Z| - 4$ or more elements into twoclasses - it is easy to see that some partition classes must have at least $(|Z| - 4)/4$ elements in common, say W. Now $T|_W = (T|_Z)|_W$ and $F|_W = (F|_Z)|_W$ are the common induced binary subtrees of T and F, and $|W| \geq c_8 \log \log n$. □

Remark. It would be interesting to see whether Theorem 1 can be tightened. In particular, it is conceivable that the much stronger bound $c' \log(n) < \text{mast}(n)$ holds, which would be best possible, up to the constant factor. Also, with slightly more care Theorem 1 can be expressed in a more explicit form: $\text{mast}(n) > \frac{1}{4} \log_2 \log_2(n - 1)$.

Acknowledgements

The first author was supported in part by the New Zealand Marsden Fund and the Allan Wilson Centre for Molecular Ecology and Evolution. The second author was supported in part by the NIH NIGMS contract 1 R01 GM078991-01, by the NSF DMS contract 0701111, by a Marie Curie Fellowship HUBI MTKD-CT-2006-042794, and by the 2007 Phylogeny programme of the Isaac Newton Institute, Cambridge, where this work started.

References

[1] H. Colonius, H.H. Schulze, Tree structures for proximity data, British Journal of Mathematical and Statistical Psychology 34 (1981) 167–180.
[2] H.-J. Bandelt, A.W.M. Dress, Reconstructing the shape of a tree from observed dissimilarity data, Advances in Applied Mathematics 7 (1986) 309–343.
[3] A.D. Gordon, Consensus supertrees: The synthesis of rooted trees containing overlapping sets of labelled leaves, Journal of Classification 3 (1986) 335–348.
[4] C.R. Finden, A.D. Gordon, Obtaining common pruned trees, Journal of Classification 2 (1985) 255–116.
[5] D.L. Swofford, PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4, Sinauer Associates, Sunderland, Massachusetts, 2003.
[6] M. Steel, T. Warnow, Kaikoura tree theorems: Computing the maximum agreement subtree, Information Processing Letters 48 (1993) 77–82.
[7] W. Goddard, E. Kubicka, G. Kubicki, F.R. McMorris, The agreement metric for labeled binary trees, Mathematical Biosciences 123 (1994) 215–226.
[8] E. Kubicka, G. Kubicki, F.R. McMorris, An algorithm to find agreement subtrees, Journal of Classification 12 (1995) 91–99.
[9] D.M. de Vienne, et al., A congruence test for testing topological similarity between trees, Bioinformatics 23 (2007) 3119–3124.
[10] E. Kubicka, G. Kubicki, F.R. McMorris, On agreement subtrees of two binary trees, Congressus Numerantium 88 (1992) 217–224.
[11] L. Lovász, Combinatorial Problems and Exercises, 2nd ed., North-Holland, 1993.
[12] C. Semple, M. Steel, Cyclic permutations and evolutionary trees, Advances in Applied Mathematics 32 (4) (2004) 669–680.