Theoretical Computer Science 646 (2016) 1–15

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Reduction rules for the maximum parsimony distance on phylogenetic trees

Steven Kelk [a,*], Mareike Fischer [b], Vincent Moulton [c], Taoyang Wu [c,*]

[a] Department of Knowledge Engineering, Maastricht University, P.O. Box 616, 6200 MD Maastricht, Netherlands
[b] Institut für Mathematik und Informatik, Walther-Rathenau-Straße 47, 17487 Greifswald, Germany
[c] School of Computing Sciences, University of East Anglia, Norwich, NR4 7TJ, United Kingdom

ARTICLE INFO

Article history:
Received 23 December 2015
Received in revised form 9 June 2016
Accepted 6 July 2016
Available online 16 July 2016
Communicated by A. Marchetti-Spaccamela

Keywords:
Phylogenetics
Parsimony
Fixed parameter tractability
Chain
Incongruence
Treewidth

ABSTRACT

In phylogenetics, distances are often used to measure the incongruence between a pair of phylogenetic trees that are reconstructed by different methods or using different regions of genome. Motivated by the maximum parsimony principle in tree inference, we recently introduced the maximum parsimony (MP) distance, which enjoys various attractive properties due to its connection with several other well-known tree distances, such as TBR and SPR. Here we show that computing the MP distance between two trees, a NP-hard problem in general, is fixed parameter tractable in terms of the TBR distance between the tree pair. Our approach is based on two reduction rules – the chain reduction and the subtree reduction – that are widely used in computing TBR and SPR distances. More precisely, we show that reducing chains to length 4 (but not shorter) preserves the MP distance. In addition, we describe a generalization of the subtree reduction which allows the pendant subtrees to be rooted in different places, and show that this still preserves the MP distance. On a slightly different note we also show that Monadic Second Order Logic (MSOL), posited over an auxiliary graph structure known as the display graph (obtained by merging the two trees at their leaves), can be used to obtain an alternative proof that computation of MP distance is fixed parameter tractable in terms of TBR-distance. We conclude with an extended discussion in which we focus on similarities and differences between MP distance and TBR distance and present a number of open problems. One particularly intriguing question, emerging from the MSOL formulation, is whether two trees with bounded MP distance induce display graphs of bounded treewidth.

1. Introduction

Finding an optimal tree explaining the relationships of a group of species based on datasets at the genomic level is one of the important challenges in modern phylogenetics. First, there are various methods to estimate the "best" tree subject to certain criteria, such as e.g. Maximum Parsimony or Maximum Likelihood. However, different methods often lead to different trees for the same dataset, or the same method leads to different trees when different parameter values are used. Second, the trees reconstructed from different regions of the genome might also be different, even when using the same criteria. In any case, when two (or more) trees for one particular set of species are given, the problem is to quantify how different the trees really are – are they entirely different or do they agree concerning the placement of most species?

* Corresponding authors.
E-mail addresses: steven.kelk@maastrichtuniversity.nl (S. Kelk), Taoyang.Wu@uea.ac.uk (T. Wu).

http://dx.doi.org/10.1016/j.tcs.2016.07.010
0304-3975/

In order to answer this problem, various distances have been proposed (see e.g. [24]). A relatively new one is the so-called Maximum Parsimony distance, or MP distance for short, which we denote d_{MP} [14,19,21]. This distance (which is a metric) is appealing in part due to the fact that it is closely related to the parsimony criterion for constructing phylogenetic trees, as well as to the Subtree Prune and Regraft (SPR) and Tree Bisection and Reconnection (TBR) distances. Indeed, it is shown in [21] that the unit neighborhood of the MP distance is larger than those of the SPR and TBR distances, implying that a hill-climbing heuristic search based on the MP distance will be less likely to be trapped in a local optimum than those based on the SPR or TBR distances. Recently, it has been shown that computing the MP distance is NP-hard [14,19] even for binary phylogenetic trees. For practical purposes it is therefore desirable to determine whether computation of d_{MP} is fixed parameter tractable (FPT). Informally, this asks whether d_{MP} can be computed efficiently when d_{MP} (or some other parameter of the input) is small, irrespective of the number of species in the input trees. We refer to standard texts such as [12] for more background on FPT. Such algorithms are used extensively in phylogenetics, see e.g. [26] for a recent example.

An obvious approach to address this question is to try to *kernelize* the problem. Roughly speaking, when given two trees, we seek to simplify them as much as possible without changing d_{MP} so that we can calculate the distance for the simpler trees rather than the original ones. Standard procedures that have been used to kernelize other phylogenetic tree distances are the so-called subtree and chain reductions (see, for example, [1,6,17]). In this paper we show that the chain reduction preserves d_{MP} and that chains can be reduced to length 4 (but not less). Moreover, we show that a certain generalized subtree reduction, namely one where the subtrees are allowed to have different root positions, also has this property, which extends a result in [21]. Both reductions can be applied in polynomial time.

These new results allow us to leverage the existing literature on TBR distance. Specifically, in [1] Allen and Steel showed that TBR distance, denoted d_{TBR}, is NP-hard to compute, by exploiting the essential equivalence of the problem with the Maximum Agreement Forest (MAF) problem: they differ by exactly 1. In the same article they showed (again utilizing the equivalence with MAF) that computation of d_{TBR} is FPT in parameter d_{TBR}. More specifically, it was shown that combining the subtree reduction with the chain reduction (where chains are reduced to length 3, rather than length 4 as we do here) is sufficient to obtain a reduced pair of trees where the number of species is at most a *linear* function of d_{TBR}. Careful reading of the analysis in [1] shows that a linear (albeit slightly larger) kernel is still obtained for d_{TBR} if chains are reduced to length 4 rather than 3. More recently, in [18] an exponential-time algorithm was described and implemented which computes d_{MP} in time $\Theta(\phi^n \cdot \text{poly}(n))$ where n is the number of species in the trees and $\phi \approx 1.618\ldots$ is the golden ratio. Combining the results of [1,18] with the main results of the current paper (i.e. Theorems 3.1 and 4.1) immediately yields the following theorem:

Theorem 1.1. *Let T_1 and T_2 be two unrooted binary trees on the same set of species X. Computation of $d_{\text{MP}}(T_1, T_2)$ is fixed parameter tractable in parameter $d_{\text{TBR}} = d_{\text{TBR}}(T_1, T_2)$. More specifically, $d_{\text{MP}}(T_1, T_2)$ can be computed in time $O(\phi^{c \cdot d_{\text{TBR}}} \cdot poly(|X|))$ where $\phi \approx 1.618\ldots$ is the golden ratio and $c \leq 112/3$.*

The constant 112/3 is obtained by multiplying the bound on the size of the kernel given in [1] ($28 \cdot d_{\text{TBR}}$) by a factor 4/3, which adjusts for the fact that here chains are reduced to length 4 rather than 3. Note also that Theorem 1.1 does not require us to apply the generalized subtree reduction: the traditional subtree reduction together with the chain reduction is sufficient.

We now summarize the rest of the paper. In the next section we collect some necessary definitions and notations, including a brief description of Fitch's algorithm which our proofs extensively use. Then in the following three sections we establish the two reductions for the MP distance, that is, the chain reduction and the subtree reduction, and remark that a theoretical variant of Theorem 1.1 could also be attained by leveraging Courcelle's Theorem [10,2], extending in a non-trivial way a technique introduced in [20]. Specifically, computation of $d_{\text{MP}}(T_1, T_2)$ can be formulated as a sentence of Monadic Second Order Logic (MSOL) posited over an auxiliary graph structure known as the display graph. The display graph is obtained by (informally) merging the two trees at their leaves. Crucially, the length of the sentence, and the treewidth of the display graph, are shown to be both bounded as a function of d_{TBR}.

We end with an extended discussion in which we focus on similarities and differences between MP distance and TBR distance. From a theoretical perspective the two distances sometimes behave rather differently but in practice d_{MP} and d_{TBR} are often very close indeed. The major open problem that remains is whether computation of d_{MP} is FPT when parameterized by itself. One possible route to this result is via a strengthened MSOL formulation, but this requires a number of challenging questions to be answered. In particular, can the treewidth of the display graph be bounded as a function of d_{MP} (rather than d_{TBR})? This in turn is likely to require new structural results on the interaction between (large grid) minors in the display graph and phylogenetic incongruency parameters.

2. Preliminaries

2.1. Basic definitions

An *unrooted binary phylogenetic tree* on a set of species (or, more abstractly, *taxa*) X is a connected, undirected tree in which all internal nodes have degree 3 and the leaves are bijectively labeled by X. For brevity we henceforth refer to these simply as *trees*, and we often use the elements of X to denote the leaves they label. In some cases, we have to consider