A Polynomial Algorithm for the Minimum Quartet Inconsistency Problem with $O(n)$ Quartet Errors

Gang Wu, Jia-Huai You, Guohui Lin*
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
wgang,you,ghlin@cs.ualberta.ca

Abstract

We show that for the Minimum Quartet Inconsistency *problem, if the number of quartet errors is $O(n)$, where n is the number of taxa under consideration, then it can be solved in polynomial time. This improves the previously best algorithmic result saying that if the number of quartet errors is at most $(n-3)/2$ then the problem can be solved in polynomial time.*

1. Introduction

Phylogenetic analysis to characterize the evolutionary relationships among a set of taxa provides the fundamental understanding of the taxa on various aspects including their origin and functional relationships. It has become a routine work to perform such an analysis at the availability of genetic data. Traditional phylogenetic analysis faces the data disparity problem, which becomes severe at the availability of whole genomes. One way of resolving this issue is to use different but the most applicable phylogenetic analysis methods to handle different subsets of taxa, and then to assemble a global phylogenetic tree out of the achieved sub-trees for the subsets. The promise of this approach is that, since every piece of phylogenetic analysis on subsets is of high confidence, the global analytical results must also be of high confidence, though there might be needs to resolve potential conflicts among the sub-trees for subsets. The quartet-based phylogeny construction methods can be classified into such efforts to construct a (global) phylogeny from sub-trees on subsets of 4 taxa.

In our discussion of phylogeny construction, the phylogeny is an unrooted binary tree whose leaves bijectively map to the set of taxa and every internal node in the tree has degree 3. In quartet-based phylogeny construction methods, researchers try to build a phylogeny (called a *quartet topology*) for every subset (or most subsets) of 4 taxa (called a *quartet*), and then assemble a global phylogeny for the whole set of taxa to satisfy all the quartet topologies that have been built, or if not at all possible, to satisfy as many of them as possible. For a quartet, there are 3 possible phylogenies or topologies associated with it. For simplicity, we use $[s_1, s_2|s_3, s_4]$ to denote the quartet topology in which the path connecting s_1 and s_2 doesn't intersect the path connecting s_3 and s_4.

Given a phylogeny T on a set of taxa S, for every quartet X, we can derive a topology for X by computing the induced subtree of T on X. Such a set of $\binom{n}{4}$ induced quartet topologies is denoted as Q_T. The computationally interesting problem is in the other way: The input is a set Q of quartet topologies in which, for every quartet, there is at most one topology for it (i.e. no ambiguity). The question is whether or not there exists one global phylogeny T for S such that each quartet topology $q \in Q$ is the same as the one derived from T. If q is the same as the quartet topology derived from T for the quartet, then T *satisfies* q or q is *consistent* with T. If there exists one global phylogeny T satisfying all quartet topologies in Q, i.e. $Q \subseteq Q_T$, then Q is *compatible* and T is a phylogeny *associated* with Q. The above recognition problem is called the *Quartet Compatibility Problem* (QCP).

If the input quartet topology set Q contains exactly one topology for every quartet, then Q is *complete*; Otherwise, Q is *incomplete*. Subsequently, we have the complete QCP problem and the incomplete QCP problem. It has been known that the complete QCP problem can be answered in $O(n^5)$ time, where n is the size of taxa set S. Furthermore, if Q is compatible, then the associated phylogeny T is unique and can be constructed within the same time. The situation changes when Q becomes incomplete, where the recognition problem becomes NP-complete. If Q is not compatible, for a solution phylogeny T, we call those quartet topologies in $Q - Q_T$ *quartet errors* on T. The more interesting computational problem is the optimization problem where Q (either complete or incomplete) is not com-

*To whom correspondence should be addressed.

patible and the goal is to construct a phylogeny to satisfy as many quartet topologies as possible. This is the so-called *Maximum Quartet Consistency Problem* (MQC). A dual minimization problem to the MQC problem, the *Minimum Quartet Inconsistency Problem* (MQI), where the input is the same, is to construct a phylogeny to minimize the number of quartet errors. Despite the fact that the MQC and the MQI problems have the same optimal solution(s), their approximabilities differ a lot. The MQC problem is NP-hard. The complete MQC problem admits a *Polynomial Time Approximation Scheme* (PTAS); The incomplete MQC problem is MAX SNP-hard. The complete MQI problem is NP-hard and it admits $O(n^2)$-approximation algorithms. No algorithms with better approximation ratios are presently known, despite some tries [4]. The incomplete MQI problem is nonapproximable within any finite ratio.

For the duration of the paper the MQC/MQI problem is assumed to mean the *complete* MQC/MQI problem, unless otherwise specified. There are several (exponential time) exact algorithms proposed for the MQC/MQI problem, as well as (polynomial time) heuristics. Among the heuristics, there is a class of *quartet cleaning* methods, one of which, the global edge cleaning algorithm, turns out to solve the MQC/MQI problem optimally when the number of quartet errors is no larger than $(n-3)/2$ [2, 5].

2. A Polynomial Algorithm for MQI with $O(n)$ Quartet Errors

Gramm and Niedermeier [5] presented a branch-and-bound algorithm for solving the MQI problem when the number of quartet errors is known exactly ahead of time. The interested readers may refer to [5] for more details. The most important idea in the branch-and-bound algorithm is to resolve global quartet conflicts through resolving local quartet conflicts [1, 3], referred to as *local conflicts*. A subset of 3 quartet topologies that involve exactly 5 taxa is called a *local subset*. If the quartet topologies in a local subset are incompatible, then the local subset becomes a *local conflict*. For example, $\{[a,b|c,d],[a,c|b,e],[a,c|d,e]\}$ is a local conflict.

Theorem 2.1 [5] *Given a complete quartet topology set Q over a taxon set S and a taxon $e \in S$, Q is compatible iff there exists no local conflict whose taxon set includes e.*

Note that there are $\binom{n-1}{3}$ quartet topologies in Q each of which involves taxon e and there are $\binom{n-1}{4}$ quartet topologies in Q none of which involves taxon e, where n is the number of taxa.

Lemma 2.2 *Let E denote the set of quartet errors in an optimal solution to the MQI problem. There exists a taxon e such that the number of quartet topologies in E involving e is less than or equal to $4|E|/n$.*

Lemma 2.3 *In the MQI problem, if there is no quartet error involving taxon e, then the problem can be solved in $O(n^4)$ time.*

PROOF. We build all the local conflicts whose taxon sets include taxon e. Since there is no quartet error involving taxon e, every such local conflict must contain exactly 2 quartet topologies involving e and the third quartet topology, denoted by q, should not involve e. Moreover, q must be changed to resolve the local conflict and q is included in at most 6 distinct local conflicts. By Theorem 2.1, we can determine all the quartet errors in $O(n^4)$ time. □

Theorem 2.4 *There is an $O(n^5 + 2^{4c}n^{12c+2})$-time algorithm that solves the MQI problem when the number of quartet errors is at most cn, where c is a positive constant and n is the number of taxa.*

PROOF. The key idea in the algorithm is that when there are at most cn quartet errors, there must exist a taxon e that is involved in at most $4c$ quartet errors (Lemma 2.2). Therefore, the algorithm tries every combination of k ($k \leq 4c$) quartet topologies involving taxon e, changes their topologies, and then applies Lemma 2.3 to solve the remaining problem. We remark that through some careful analyses, the running time of the algorithm can be decreased a bit. □

Acknowledgments: GW's research is supported by NSERC and CFI. JY's research is supported by NSERC. GL's research is supported by NSERC, CFI, and NNSF Grant 60373012.

References

[1] H. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advance in Applied Mathematics*, 7:309–343, 1986.

[2] V. Berry, T. Jiang, P. E. Kearney, M. Li, and H. T. Wareham. Quartet cleaning: Improved algorithms and simulations. In *Proceedings of the 7th Annual European Symposium on Algorithms (ESA'99)*, LNCS 1643, pages 313–324, 1999.

[3] H. Colonius and H. H. Schultze. Tree structure for proximity data. *British Journal of Mathematical and Statistical Psychology*, 34:167–180, 1981.

[4] G. Della Vedova, T. Jiang, J. Li, and J. J. Wen. Approximating minimum quartet inconsistency (abstract). In *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 894–895, 2002.

[5] J. Gramm and R. Niedermeier. A fixed-parameter algorithm for minimum quartet inconsistency. *Journal of Computer and System Science*, 67:723–741, 2003.