Approximating Minimum Quartet Inconsistency (*Abstract*)

Gianluca Della Vedova * Tao Jiang † Jing Li‡ Jianjun Wen §

A fundamental problem in computational biology which has been widely studied in the last decades is the reconstruction of evolutionary trees from biological data. Unfortunately, almost all its known formulations are NP-hard. The compelling need for having efficient computational tools to solve this biological problem has brought a lot of attention to the analysis of the *quartet paradigm* for inferring evolutionary trees [1, 2, 8]. Given a quartet of taxa $\{a, b, c, d\}$, there are 3 possible degree-3 trees connecting the taxa as terminals. Each such tree is called a *quartet topology*. The quartet methods proceed by first estimating the topology of each quartet of taxa and then recombining the inferred quartet topologies into an evolutionary tree. A major difficulty in this approach derives from the fact that quartet topology inference methods often make mistakes, and thus may result in a set Q of quartet topologies that is inconsistent with any evolutionary tree. Therefore, the problem of recombining the quartet topologies of Q to form an estimate of the correct evolutionary tree is naturally formulated as an optimization problem that looks for a tree T maximizing the number of *consistent* quartets (*i.e.* $|Q \cap Q_T|$), or equivalently minimizing the number of *inconsistent* quartets (*i.e.* $|Q \quad Q_T|$), where Q_T denotes the unique set of quartet topologies induced by T. The above (complementary) problems are referred to as *Maximum Quartet Consistency* (MQC) and *Minimum Quartet Inconsistency* (MQI) problems. In a recent paper [4], it has been shown that MQC is NP-hard, but it admits a PTAS, using the technique of *smooth integer polynomial programming* and exploiting the natural *denseness* of the set Q (*i.e.* $|Q| = \binom{n}{4} = \theta(n^4)$, where n is the number of taxa). However, the PTAS only guarantees an evolutionary tree that may deviate from Q by ϵn^4 quartet topologies, for any small constant $\epsilon > 0$. Some methods, such as maximum likelihood, gives good results in practice (*i.e.* the quartet errors are $\ll O(n^4)$). In this case, the trees computed by the PTAS are not very useful since they may actually contain many more errors than those in the input set Q. Hence, it is logical to look for efficient approximation algorithms that aim at minimizing the number of inconsistent quartet topologies directly. In fact, some heuristics for this problem have been proposed in [1, 8, 9]; but they do not have guaranteed performance.

In this abstract, we report our partial progress on approximating MQI. It is known that MQI is approximable within ratio $O(n^2)$ [5]. We show that the *local cleaning* algorithm introduced in [3] achieves a better ratio. For any input set Q of quartet topologies, let $E_{opt}(Q)$ denote the optimum of the MQI problem on Q.

Theorem. The local cleaning algorithm approximates MQI with ratio $\min\{O(n^2), E_{opt}(Q)\}$.

A stronger result can be proved for the *hypercleaning* algorithm [3] algorithm. Namely, setting $m = \Theta(\frac{\log n}{\log \log n})$, if hypercleaning returns a unique tree, then such a tree approximates $E_{opt}(Q)$ with ratio $\min\{O(n^2/m), E_{opt}(Q)/m\}$.

We also propose a new greedy algorithm that iteratively identifies pairs of (original or derived) taxa that should be joined together. Consider a set of weighted taxa, where each taxon a has a nonnegative integer weight $w(a)$. For any four taxa a, b, c, d, a *mixed topology* $q(a, b, c, d)$ of these taxa is a multiset of w_1 $ab|cd$'s, w_2 $ac|bd$'s and w_3 $ad|bc$'s, where w_1, w_2, w_3 are nonnegative integers and $w_1 + w_2 + w_3 = w(a)w(b)w(c)w(d)$. The *Generalized Minimum Quartet Inconsistency* (GMQI) problem is defined as follows: Given a set S of weighted taxa and Q containing mixed topologies for each quartet of taxa, find an evolutionary tree T minimizing $|Q \quad Q_T|$, where Q_T is the set of mixed quartet topologies induced by T. Let $a, b \in S$, $\{c, d\} \subseteq S$, if $q(a, b, c, d) = \{w_1(ab|cd), w_2(ac|bd), w_3(ad|bc)\}$, define $B_{ab}(c, d) = \{w_2(ac|bd), w_3(ad|bc)\}$, $Q^{a*b} = \bigcup_{\{c,d\} \subseteq S} B_{ab}(c, d)$. Given $a, b, c, d, e \in S$, let $q(a, c, d, e) = \{w_1(ac|de), w_2(ad|ce), w_3(ae|cd)\}$ and $q(b, c, d, e) = \{w'_1(bc|de), w'_2(bd|ce), w'_3(be|cd)\}$. Let $w_i + w'_i = \max\{w_j + w'_j, 1 \leq j \leq 3\}$. W.l.o.g., assume $i = 1$ in the above maximization, and define $P_{ab}(c, d, e) = \{w_2(ad|ce), w_3(ae|cd), w'_2(bd|ce), w'_3(be|cd)\}$, $Q^{a+b} = \bigcup_{\{c,d,e\} \subseteq S} P_{ab}(c, d, e)$, $q(a, b) = |Q^{a*b} \cup Q^{a+b}|$.

We are now ready to describe our (recursive) greedy

*Dipartimento di Statistica, Università degli Studi di Milano-Bicocca. Email:gianluca.dellavedova@unimib.it.

†Dept. of Comp. Sci., UC Riverside. jiang@cs.ucr.edu.

‡Dept. of Comp. Sci., UC Riverside. jili@cs.ucr.edu.

§Dept. of Comp. Sci., UC Riverside. wjianju@cs.ucr.edu.

algorithm for GMQI. Consider a set S of weighted taxa and a set Q of mixed topologies for each quartet of taxa. The recursive step is (i) finding a pair of taxa a, b with the minimum $q(a, b)$, (ii) merging such taxa into a new (artificial) taxon with weight $w(a) + w(b)$, and then (iii) modifying the set Q of mixed topologies in a straightforward way to obtain a reduced instance of GMQI. Observing that an instance of MQI can be viewed as an instance of GMQI by assigning a unit weight to each taxon, we have the following conjecture:

Conjecture: the above is an $O(n)$-approximation algorithm for MQI.

We in fact suspect that the algorithm achieves $O(n)$-approximation for GMQI. We have so far obtained several preliminary results supporting the conjecture, and are working towards its complete proof.

We have compared the performances of the well-known Neighbor-Joining (NJ) algorithm, local cleaning, hypercleaning with $m = 2$ and $m = 3$, and the above greedy algorithm for GMQI on simulated DNA sequence data. Evolutionary trees were generated randomly using a birth-death process with ratio 2:1 and 1 out of 20 sampling [6]. On each tree, leaf sequences were generated using program Seq-Gen based on the Kimura 2-parameter (K2P) model with site-to-site rate variance [6]. The edge lengths of the tree represent the expected number of mutations per site along that edge. In the preliminary experiment, the transition vs transversion ratio was fixed as 1.6 to create some realistic bias for transition. Rate heterogeneity across sites are also accommodated by assuming a continuous gamma distribution [6] with parameters $\alpha = 10$ and $\beta = 1/\alpha$. Distances between sequences were calculated using program dnadist, assuming the K2P model with the correct transition vs transvertion ratio, and quartet topologies were inferred using the NJ algorithm. We have also considered the combination of local cleaning and the greedy algorithm in the experiment.

We ran the comparative test for 5, 10, 20 and 30 taxa, mutation rates .025, .05, .25 and 1 and sequence length 500, 1000, 5000. Due to space constraint, we can report here only some results for sequence length 5000, in terms of the average percentage of correct edges over 100 runs. It is interesting to see that the greedy GMQI algorithm has a comparable performance as the hypercleaning algorithm, although the latter has a much higher time complexity. Also, combining local cleaning with the greedy algorithm yields an improved performance that is not far from the performance of NJ. In general, the quartet methods performed in this experiment better than they did in [7], where the sequence data were simulated using a simpler model.

Num Taxa	Mutation rates 1	0.25	0.05	0.025
5	91,91,92	98,98,98	96,96,97	96,96,96
10	93,92,93	98,98,98	96,96,97	95,95,96
20	87,86,93	95,95,97	95,95,96	95,95,96
30	79,78,94	91,91,97	92,92,97	91,91,96

Table 1: Percentages of correct edges inferred by our GMQI algorithm with locally cleaned quartet topologies, the same with uncleaned quartet data, and NJ.

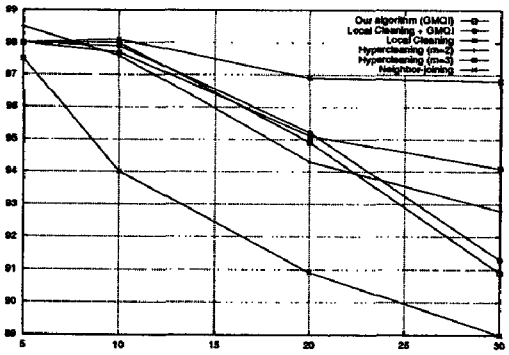

Figure 1: Comparison of the algorithms at mutation rate 0.25.

References

[1] A. Ben-Dor, B. Chor, D. Graur, R. Ophir and D. Pellep, Constructing Phylogenies from Quartets: Elucidation of Eutherian Superordinal Relationships. *Journal of Comp. Biol.* 5(3): 377–390, 1998.

[2] V. Berry and O. Gascuel, Inferring evolutionary trees with strong combinatorial evidence, *Theoretical Computer Science*, 240(2): 271–298, 2001.

[3] V. Berry, D. Bryant, T. Jiang, P. Kearney, M. Li, T. Wareham, and H. Zhang, A practical algorithm for recovering the best suppported edges in an evolutionary tree, *Proc. ACM-SIAM SODA2000*, 287–296, 2000.

[4] T. Jiang, P. Kearney, and M. Li. A Polynomial Time Approximation Scheme for Inferring Evolutionary Trees from Quartet Topologies and Its Application. *SIAM J. Comput.* 30(6): 1942-1961 (2000)

[5] T. Jiang, P. Kearney and M. Li, Open problems in computational molecular biology, *Journal of Algorithms* 34(1): 194–201, 2000.

[6] W.H. Li, Molecular Evolution, Sinauer Assoc., 1997

[7] K. St. John, T. Warnow, B.M.E. Moret, and L. Vawter, Performance study of phylogenetic methods: (unweighted) quartet methods and neighbor-joining, *Proc. 12th ACM-SIAM SODA*, 196-206, 2001.

[8] K. Strimmer and A. von Haeseler, Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies, *Molecular Biology and Evolution* 13(7), 964–969, 1996.

[9] S.J. Willson, Measuring inconsistency in phylogenetic trees, *Journal of Theoretical Biology* 190: 15–36, 1998.