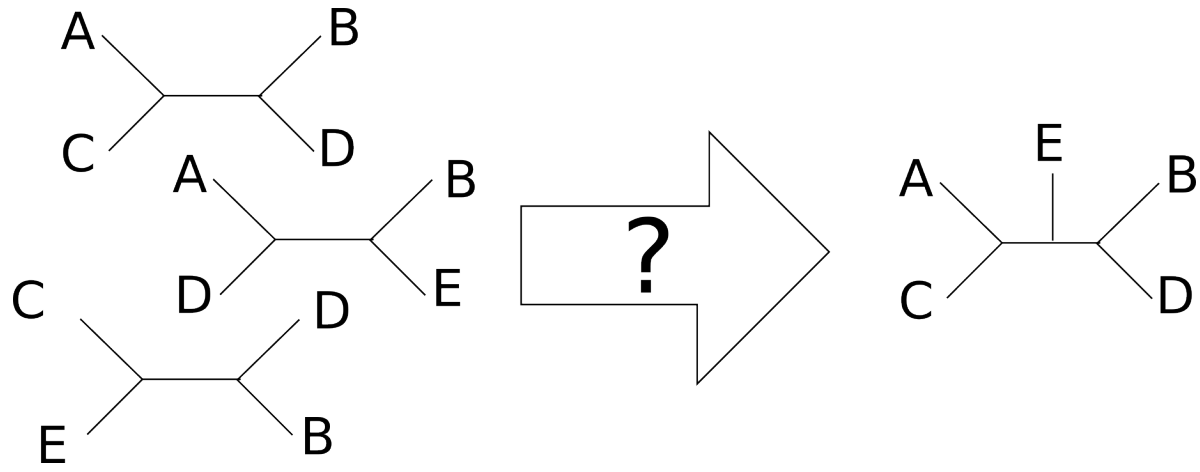
A Polynomial-Time Approximation Scheme for Maximum Quartet Compatibility

Pranjal Vachaspati

UIUC - CS598AGB

Incomplete Maximum Quartet Consistency [I-MQC]

Given quartet set Q over taxon set X and some integer k, is there some tree T that induces at least k of the quartets in Q?



- Shown to be NP-Hard (reduction to BETWEENNESS) by (Steel, 1992)
- Also Max SNP-hard - only constant-factor approximations exist

Maximum Quartet Consistency [MQC]

Given quartet set Q over every four-taxon subset of taxon set X and some integer k, is there a tree T that induces at least k of the quartets in Q?

- This is still NP-hard
- But, we have a polynomial-time approximation scheme

Approximating NP-Hard Problems

Inapproximable	Approximation factor is a function of n	Max-Clique: $O(n^{1-\epsilon})$ Set Cover: $O(\log n)$
APX/Max-SNP	Constant-factor approximation in $p(n)$ time	Traveling salesman Max-Parsimony
PTAS	$(1 \pm \epsilon)$ approximation in $f(1/\epsilon)p(n)$ time	Euclidean traveling salesman Maximum quartet consistency
FPTAS	$(1 \pm \epsilon)$ approximation in $p(1/\epsilon)p(n)$ time	Knapsack Problem

Polynomial Time Approximation Scheme

- Given complete quartet set Q (of size $\binom{n}{4}$), there is some tree *TOPT* that maximizes $|Q_{TOPT} \cap Q|$
- Find *TAPX* in polynomial time such that

$$|Q_{TAPX} \cap Q| \geq (1 - \epsilon)|Q_{TOPT} \cap Q|$$

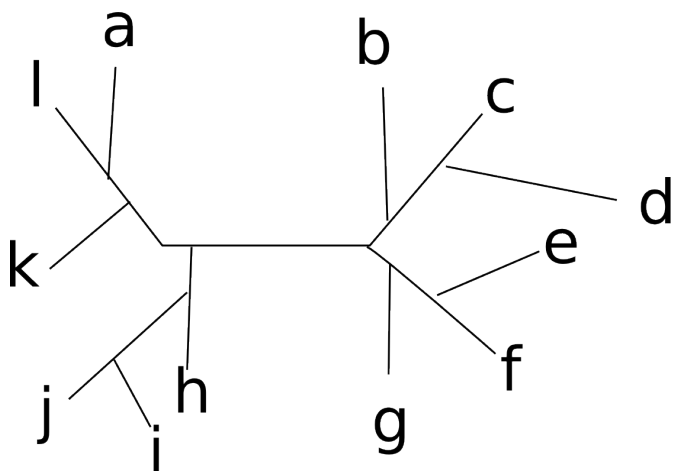
- By choosing a random tree, $|Q_{TOPT} \cap Q| \geq \frac{1}{3}\binom{n}{4}$
- Then for some c, our desired *TAPX* has the property

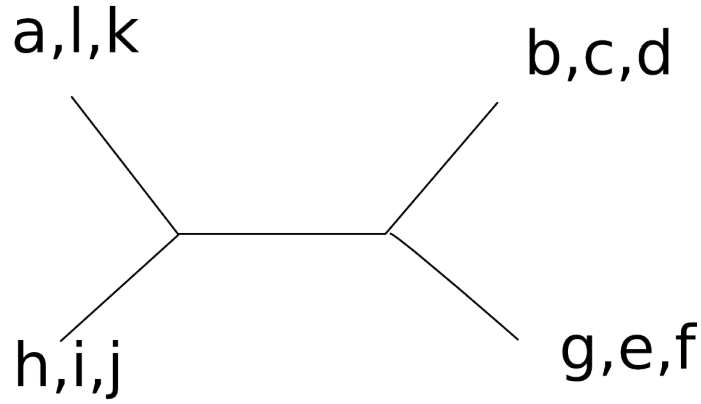
$$|Q_{TAPX} \cap Q| \geq |Q_{TOPT} \cap Q| - cn^4$$

k-bin decomposition

- For all T, Q, k, there exists a tree T_k with k leaves and multiple taxa at each leaf that satisfies

$$|Q_{T_k} \cap Q| \geq |Q_T \cap Q| - (c'/k)n^4$$

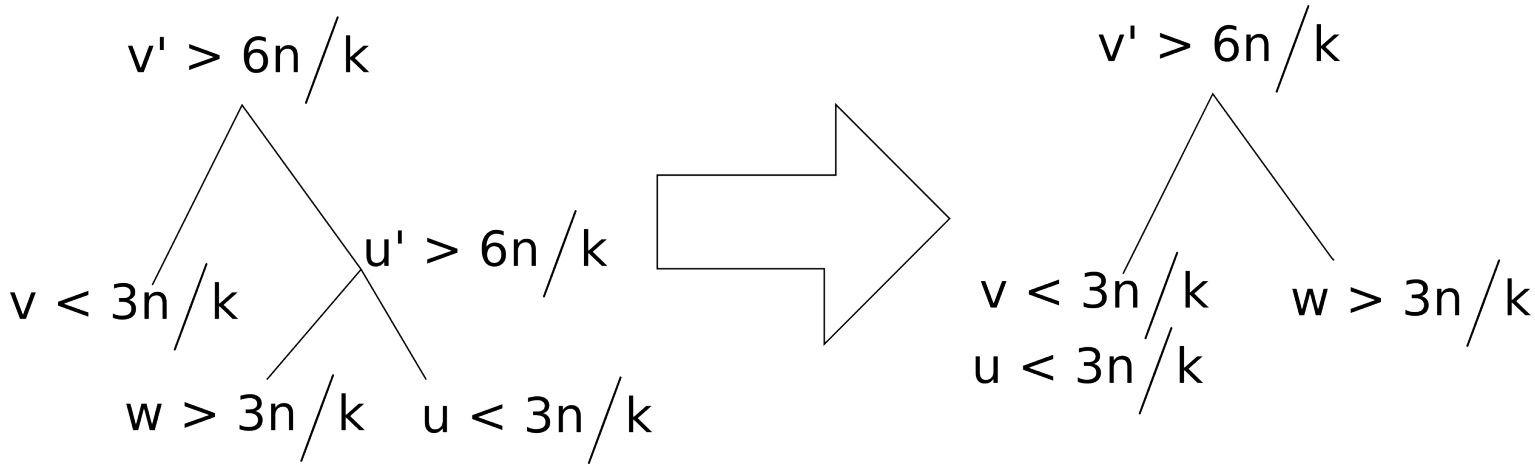




- How do we generate this?

k-bin decomposition

1. Collapse all clades with fewer than $6n/k$ children
2. Then do this:

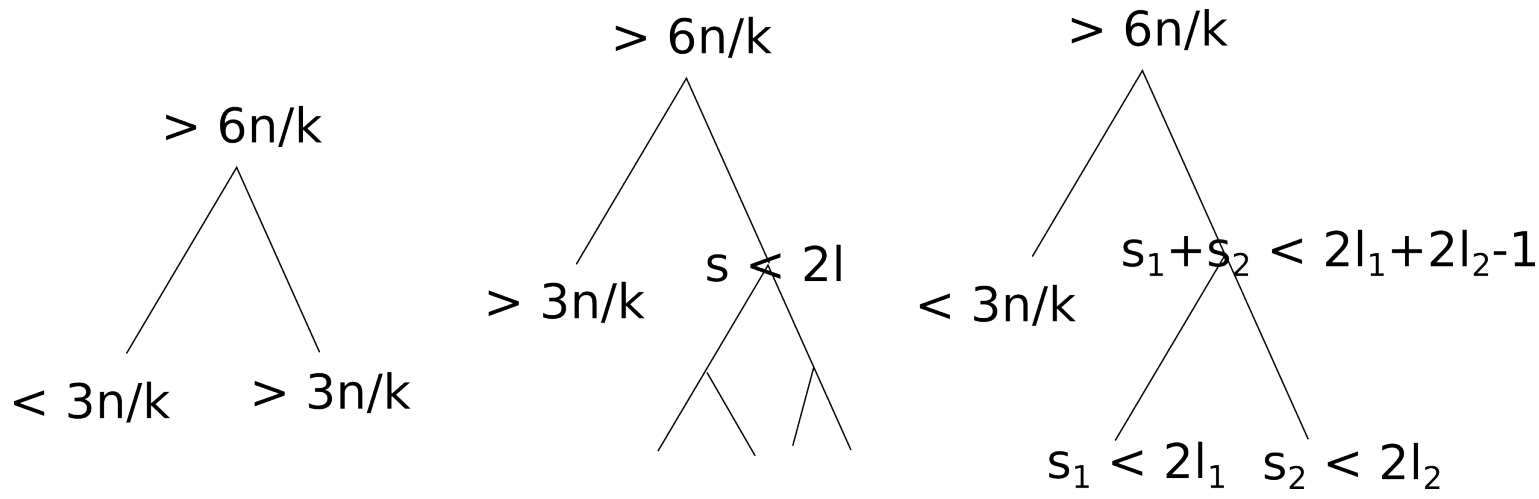


Observe that this still preserves quartets

k-bin decomposition

T_K has at most k bins:

- Lemma: We have at most twice as many small bins as large bins ($s < 2l$)



- Each large bin has at least $3n/k$ taxa
- There are at most $l = k/3$ large bins
- There are at most $3l = k$ bins

k-bin decomposition

$$|Q_{T_k} \cap Q| \geq |Q_T \cap Q| - (c'/k)n^4$$

- Every quartet on a, b, c, d with all taxa in different bins will agree
- At most $k(6n/k)^2 n^2 = 36n^4/k$ quartets with 2 taxa in the same bin
- At most $k(6n/k)^3 n = 216n^4/k^2 \leq 36n^4/k$ quartets with 3 taxa in the same bin
- At most $k(6n/k)^4 = 1296n^4/k^3 \leq 36n^4/k$ quartets with 4 taxa in the same bin
- In total, at most $\frac{108}{k}n^4$ missed quartets

- There are only a *constant* number (parameterized in n) of tree topologies over k leaves!
- We can try each of these topologies and pick the best one.
- All that remains is to assign labels to a tree topology.

Label-Bin Assignment

- Create nk $0-1$ variables x_{sb}, set to 1 if label s is assigned to bin b
- For each quartet $ab|cd$ in Q, the polynomial

$$p_{ab|cd}(x) = \sum_{ij|kl \in Q_{T_k}} x_{ai} x_{bj} x_{ck} x_{cl}$$

  is 1 iff the quartet exists in the labeled T_k
- So we want to maximize

$$p(x) = \sum_q p_q(x)$$

- subject to constraints

$$\forall s \in \mathit{labels}, \sum_{b \in \mathit{bins}} x_{bs} = 1$$

$$\forall b \in \mathit{bins}, \sum_{s \in \mathit{labels}} x_{bs} \leq 6n/k$$

- This is a smooth integer polynomial program, which has a randomized PTAS

Algorithm

Given a quartet set Q and a tolerance ϵ

1. Pick k, ϵ_1 such that

$$\epsilon \leq c'/(ck) + \epsilon_1/c$$

   where c is the fraction of quartets in Q induced by *TOPT* and c' is the constant from the k-bin decomposition analysis
2. For each of the $O(k!)$ k-tree topologies, find a ϵ_1 approximation to the optimal label-bin assignment
3. Arbitrarily resolve the best LBA for the best k-bin decomposition

Analysis

- The best k-bin decomposition misses $\frac{c'}{k} n^4$ quartets
- The best approximation to the best k-bin decomposition misses a further $\epsilon_1 n^4$ quartets
- Overall, we have a total of $|Q_{TOPT} \cap Q| - \left(\frac{c'}{k} + \epsilon_1\right) n^4$ correct quartets
- If $|Q_{TOPT} \cap Q| = cn^4$, we get $\left(1 - \frac{c'}{ck} - \frac{\epsilon_1}{c}\right) |Q_{TOPT} \cap Q|$ correct quartets

This is not a practical algorithm

- Suppose we want 1% error

$$\epsilon = 0.01 \leq c'/(ck) + \epsilon_1/c$$

- $c' \approx 100$ and $c \approx 1$
- Even if we can solve the LBA problem exactly
- $k \approx 10000$
- (this is an upper bound)

Related Problems

- Quartet Cleaning - a different application of the PTAS to eliminate bad quartets
- NP-hardness proof for MQC
- Open problems:
  - Is there a practical verison of this algorithm?
  - Is the algorithm still NP-hard if the input quartet set comes from gene trees?