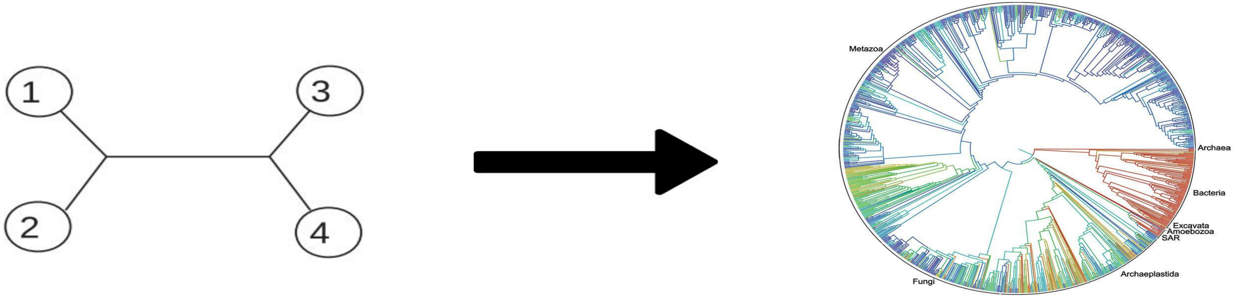
Fast and Curious: Amalgamating Quartet Trees Using MaxCut

Molecular Phylogenetics and Evolution 2012
Authors: Sagi Snir and Satish Rao
Journal Club Presentation by: Emma Hamel

Motivation

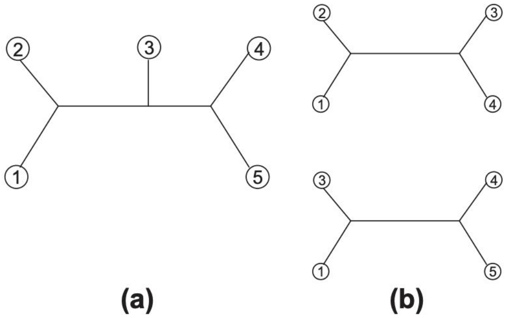
- Phylogenetic reconstruction methods are computationally intensive, thus limiting the amount of taxa that they can be run on
- One divide-and-conquer approach runs methods on smaller (overlapping) subsets of taxa and then combines the subset trees by estimating a **supertree**
    - Related the creation of a Tree of Life, a supertree that encompasses all known organisms
- **Quartet amalgamation** is one approach for estimating supertrees



Taken from wikipedia

Maximum Quartet Consistency (MQC) Problem

- Find the tree satisfying the largest number of input quartets
- NP-hard optimization problem
- Example: “A quartet tree ((a,b),(c,d)) is **satisfied** by a tree T if in T, there is an edge (or a path in general) separating a and b from c and d.”



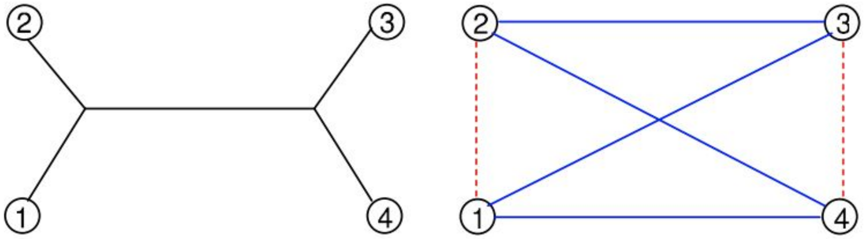
The two trees on the left were induced from the tree on the right

Quartet MaxCut: High Level

- **Input:** A set S of species and a set Q of quartet trees on S
- **Output:** A supertree T on S
- **Divide-and-conquer method that, at each recursive step, divides the input species set S in half, defining a bipartition in the output tree T**
    - Pick bipartition that maximizes the ratio between satisfied to violated quartets; to do this
        - Create quartet graph G(Q)
        - Find cut with maximal weight for G (NP-hard problem)

Quartet Graph G(Q)

- Vertices are defined by species set S
- Edges are defined by input quartets
- Each quartet adds 4 “good” edges and 2 “bad” edges
  - “Good” edges (red dashed) get a positive weight (ɑ) and “bad” edges (blue) get a negative weight (-ɑ)



An example of a quartet tree and its corresponding quartet graph

Definitions

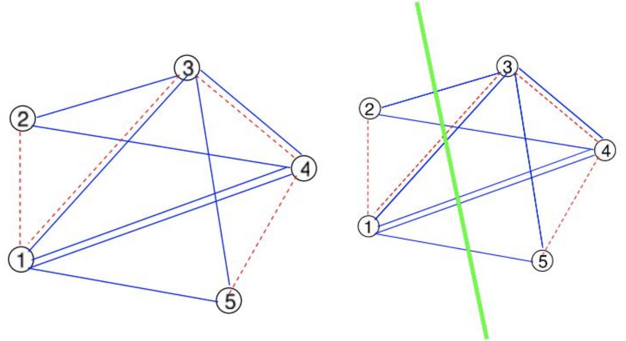
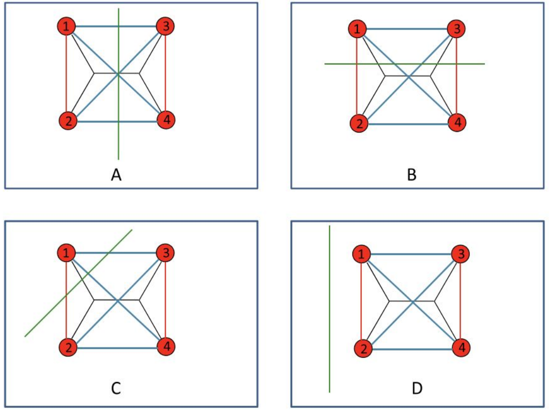


Fig. 4. Left: The graph $G(Q)$ induced by the quartets ((1,2), (3,4)) and ((1,3), (4,5)) from Fig. 3. Right: The maximum cut in $G(Q)$ separating {1,2} from {3,4,5}, therefore satisfies quartet ((1,2), (3,4)) but defers ((1,3), (4,5)) and hence contains six good and one bad edges.

- A cut is any bipartition A | B that divides the vertex set S of the graph
- An edge (x,y) is in the cut if its two vertices are on different sides of the bipartition
    - For example, x in A and y in B or vice versus
- The weight of a cut is the sum of the weights of the edges in the cut
    - Sum the weights for all possible edges (x,y) with x in A and y in B

More definitions

- **Unaffected Quartets:** All vertices in quartet are on the same side of the cut
- **Affected Quartets**: A cut separates some vertices in quartet
  - **Satisfied**: If one pair of sisters are in one part and the other pair is in the other part
  - **Violated**: Both pairs of sisters are separated
  - **Deferred**: One vertex is separated from the other three vertices by a cut



A = Satisfied, B = Violated, C = Deferred, D = Unaffected

Intuition

- Negative weights (-ɑ) are assigned to edges between sisters in each quartet
  - So putting sister vertices on the same side of the cut (bipartition) decreases the cut's weight
- Positive weights (ɑ) are assigned to edges between non-sisters in each quartet
  - So putting non-sisters on opposite sides of the cut (bipartition) increases the cut's weight

MaxCut Heuristic

“**Center of mass (COM)** point of a vertex is the closest point to all its neighbors, proportional to the edge weights of each neighbor”

1. Vertices randomly placed on the 3-dimensional sphere
2. Every vertex is moved towards its COM; repeat a constant number of times
3. Draw hyperplane through origin of sphere, dividing the vertex set into two sets A and B
4. Return bipartition A|B

The authors also introduce two improvements to prevent the heuristic from returning a trivial cut (bipartition), that is, when the cut is all taxa versus the empty set OR when the cut is one taxon versus all remaining taxa (singleton).

Quartet MaxCut Algorithm

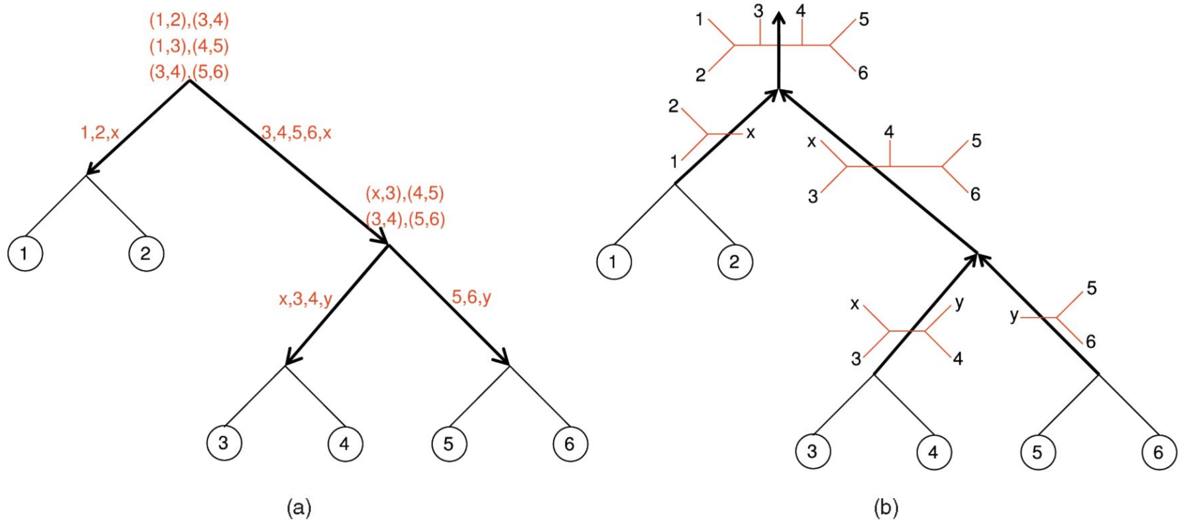


Fig. 8. The operation of the high-level QMC algorithm on the quartets $\{((1,2),(3,4)),((1,3),(4,5)),((3,4,),(5,6))\}$. (a) The recursive calls with the input quartets at every call (node) and the splitted taxa along the edges. (b) The rollback from the recursion with the returned trees along the edges.

Taken from Snir and Rao, 2010.

Quartet MaxCut Algorithm

Function: QuartetMaxCut(Q, S)

1. If the set **Q** of quartets is empty, return the tree T on the species set **S**
2. Construct the quartet graph **G(Q)**
3. Use heuristic to find the “maximum” cut **A|B** for **G(Q)**
4. Add an artificial taxon to both species sets **A** and **B**
5. Create a set **QA** of quartets for leaf set **A**
    a. Add quartets from **Q** with 4 leaves in **A** (unaffected quartets)
    b. Add quartets from **Q** with 3 leaves in **A** (deferred quartets), labeling the taxon in **B** as the artificial taxon
6. **TA = QuartetMaxCut(QA, A)** [Recurse on leaf set A]
7. Repeat steps 5 and 6 for leaf set **B**
8. Join trees **TA** and **TB** at the artificial taxon to get a tree **T** on species set **S**
9. Return **T**

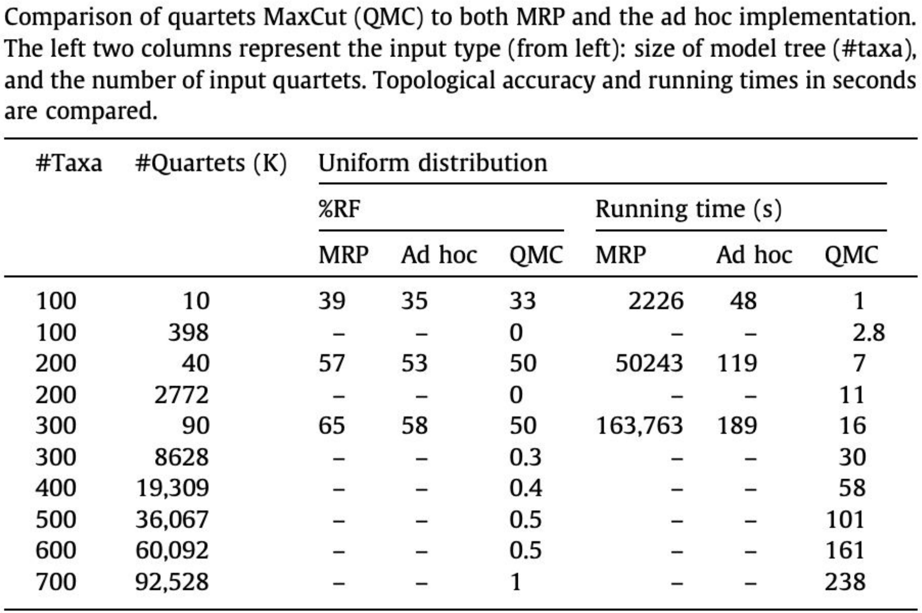
Evaluation Overview

1. Generate a set Q of quartet trees from a model tree
2. Estimate a tree from Q using supertree methods
   - Paup*'s implementation of Matrix Representation with Parsimony (**MRP**)
   - Old version of QuartetMaxCut (**Ad Hoc**)
   - New and improved of Quartet MaxCut (**QMC**)
3. Compare estimated tree to model tree using Robinson-Foulds (RF) distance

Experiment 1

- A set Q of quartets were created by sampling uniformly at random from model trees with n = 100 to 700 taxa
- # of quartets was either $|Q| = n^2$ or $n^{2.8}$
- 10% of quartets in Q were made to disagree with the model tree

Experiment 1: Results



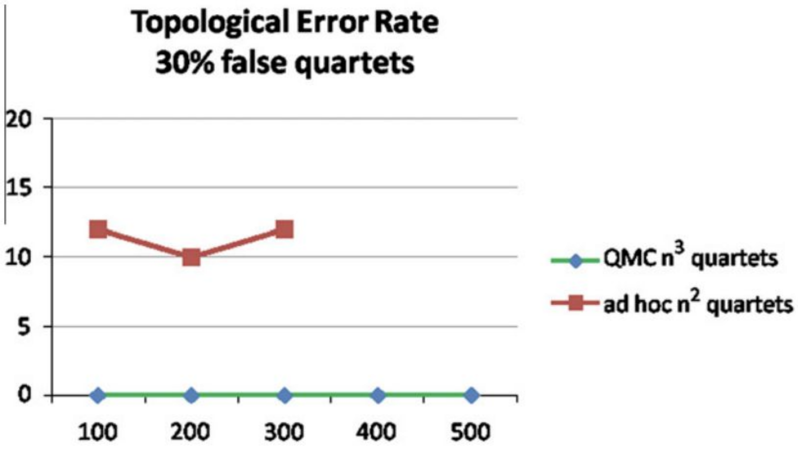
Comparison of quartets MaxCut (QMC) to both MRP and the ad hoc implementation. The left two columns represent the input type (from left): size of model tree (#taxa), and the number of input quartets. Topological accuracy and running times in seconds are compared.

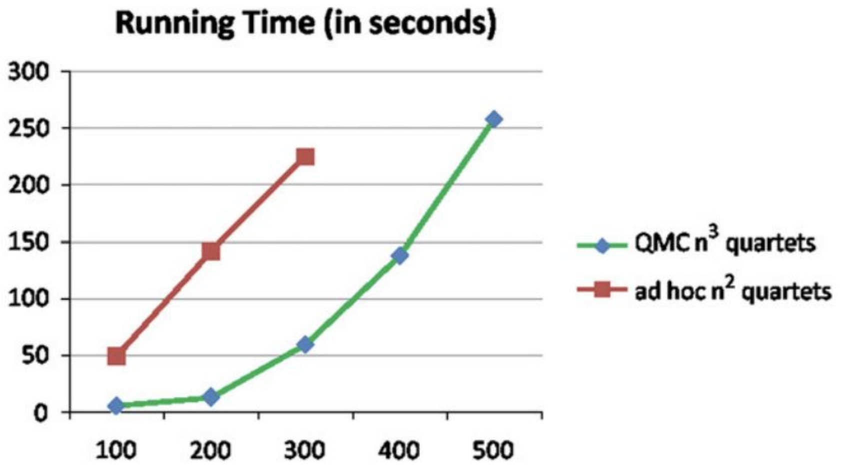
#Taxa	#Quartets (K)	Uniform distribution					
		%RF			Running time (s)		
		MRP	Ad hoc	QMC	MRP	Ad hoc	QMC
100	10	39	35	33	2226	48	1
100	398	–	–	0	–	–	2.8
200	40	57	53	50	50243	119	7
200	2772	–	–	0	–	–	11
300	90	65	58	50	163,763	189	16
300	8628	–	–	0.3	–	–	30
400	19,309	–	–	0.4	–	–	58
500	36,067	–	–	0.5	–	–	101
600	60,092	–	–	0.5	–	–	161
700	92,528	–	–	1	–	–	238

Experiment 2

This experiment is exactly like experiment 1 except 30% (instead of 10%) of the quartet trees disagree with the model tree.

Experiment 2: Results





Experiment 3

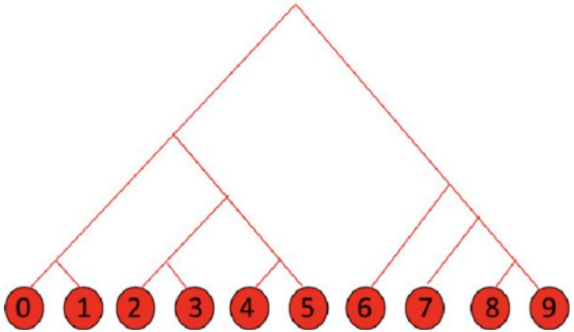
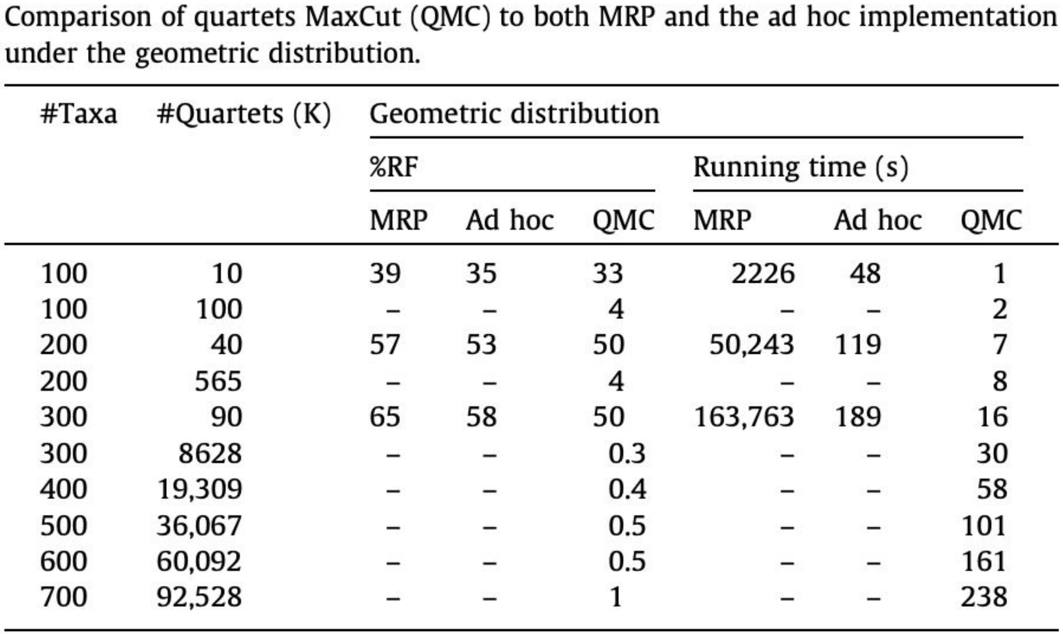


Fig. 8. A tree over 10 taxa. The quartet for taxa $\{1,6,7,9\}$ has topology $q = 1,6|7,9$ and $diam(q) = 7$ obtained by the pair 1,9. Under the geometric distribution, this quartet is chosen with probability 1/9.

Note that there is a typo in Fig. 8 caption. The quartet should be chosen with probability 1/7.

- A set Q of quartets was generated from a model tree and then sampled with probability of 1/(diameter of the quartet)
- For the diameter, the maximum number of edges between any pair of taxa in the quartet
- This procedure favors shorter quartets over longer quartets

Experiment 3: Results



Comparison of quartets MaxCut (QMC) to both MRP and the ad hoc implementation under the geometric distribution.

#Taxa	#Quartets (K)	Geometric distribution					
		%RF			Running time (s)		
		MRP	Ad hoc	QMC	MRP	Ad hoc	QMC
100	10	39	35	33	2226	48	1
100	100	–	–	4	–	–	2
200	40	57	53	50	50,243	119	7
200	565	–	–	4	–	–	8
300	90	65	58	50	163,763	189	16
300	8628	–	–	0.3	–	–	30
400	19,309	–	–	0.4	–	–	58
500	36,067	–	–	0.5	–	–	101
600	60,092	–	–	0.5	–	–	161
700	92,528	–	–	1	–	–	238

Experiment 4

- Biological dataset from Zhaxybayeva et al. 2006
  - 11 species
  - 1,128 gene trees
  - 214,729 quartets induced by gene trees, removing low-confidence quartets
    - Low-confidence was based on ratio between central edge to four external edges
- QMC applied to quartets recovered same species tree as Zhaxybayeva et al. 2006, showing strong evidence of horizontal gene transfer (HGT) highways

Conclusions

- Quartet MaxCut (QMC) is faster and more accurate than MRP or the older version of QMC
- Sampling larger numbers of quartets improves accuracy
- Quartet methods, such as QMC, may be useful for estimating species trees on datasets with horizontal gene transfer (HGT), for example, bacterial datasets