CS 581, Fall 2020
intro to course projects

Baqiao

New algorithm for quartet amalgamation

- Develop new heuristic for the MQC (max quartet consistency) problem (NP-hard)
  - Given quartet trees on taxa set X. Find a tree T that satisfies the maximum number of input quartet trees
  - Constrained combinatorial optimization through DP?
- Compete with Quartet MaxCut
  - Divide-and-conquer heuristic that uses max-cut to determine the split
  - No theoretical guarantee, but good performance

Alignments Joining Under Sequence Length Heterogeneity

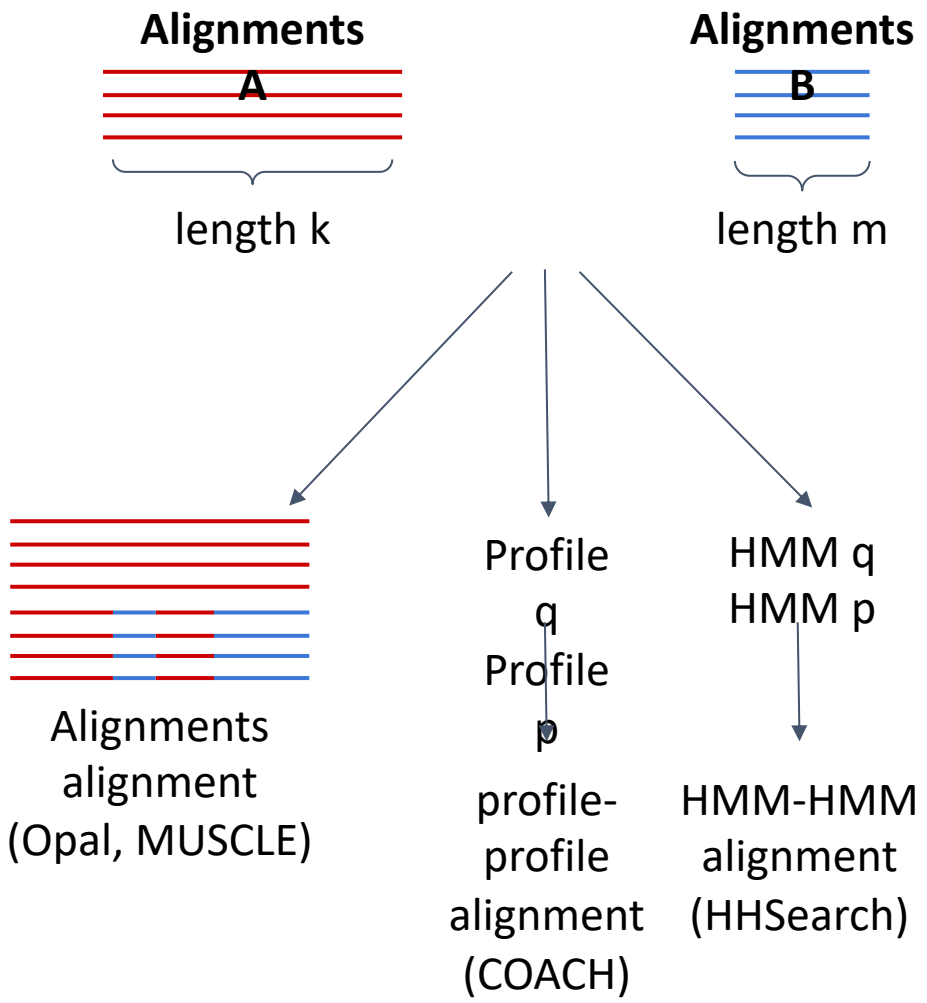
Chengze Shen, Qinghui Zhou

Motivations

- We have existing methods for aligning alignments, such as alignments alignment, profile-profile alignments and HMM-HMM alignments.
- However, there are few studies discussing the impact of sequence heterogeneity on merging alignments.

Goals

- Evaluate existing methods performances on two disjoint MSAs with sequence length heterogeneity.
- **Softwares**: Opal, MUSCLE, HHSearch, and other more recent softwares.
- **Data**: MSAs obtained from two disjoint subsets of sequences. The full dataset should have multiple peaks on the sequence length histogram (e.g. 16S.T)
- **Evaluation**: For each method, SPFN and SPFP comparing to true/estimated MSAs obtained from the full dataset.



Eleanor and Kathie (Yirong)

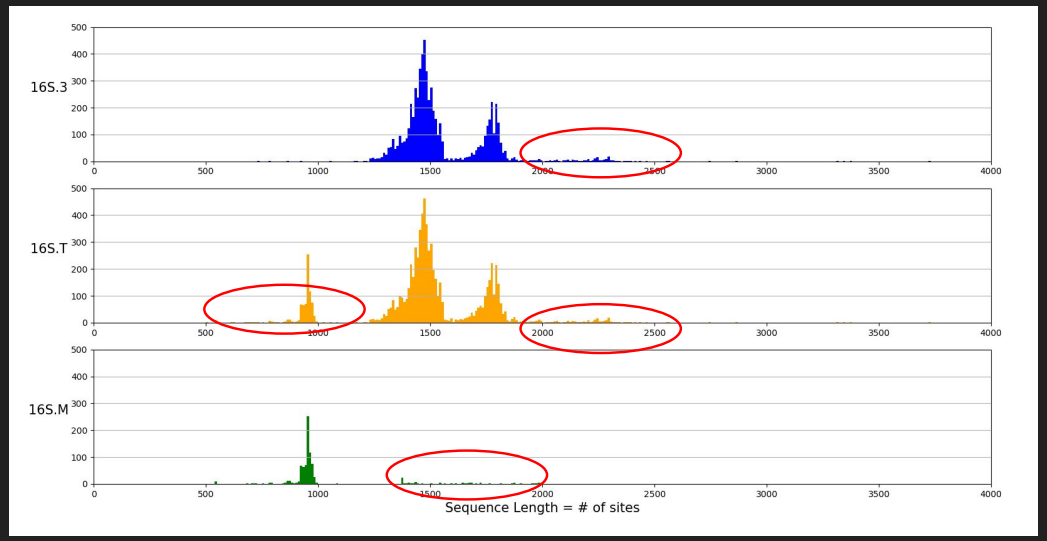
Objective: modify the phylogenetic placement method pplacer, so that it can work on backbone tree which has around 20,000 leaves

Idea:

- Find leaf y which is the closest leaf to query sequence x
- Find subtree T’ on backbone tree T based on y such that the size of T’ is close to but less than 5,000
- Use pplacer to find on placement of x on T’
- Based on the placement of x on T’, place x on tree T

Explore an improved mode of running UPP based on the effect of sequence length heterogeneity

Qikai



A challenge

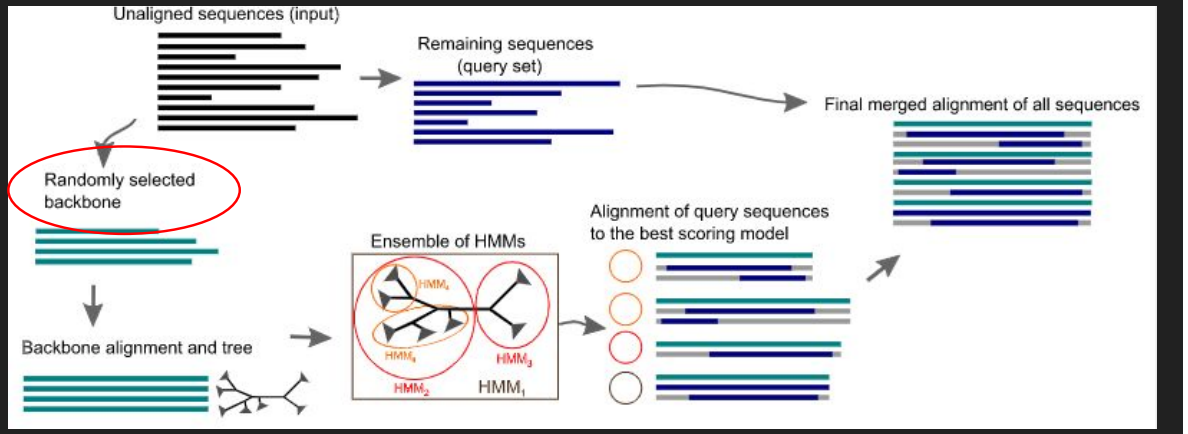


Figure source: Nguyen NP et al (2015)

By default, UPP will sample 1000 sequences whose length are within the interval [0.75 median, 1.25 median] for the backbone alignment.

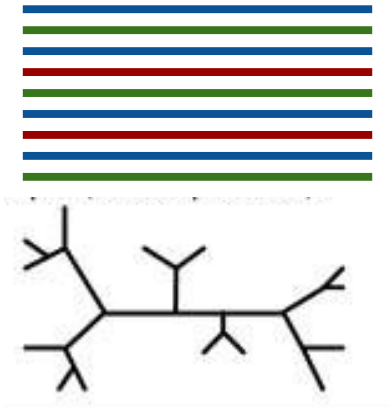
Is there a better way to sample the sequences of the backbone alignment?

Sample from higher cluster? Lower cluster? Or a mix? .....

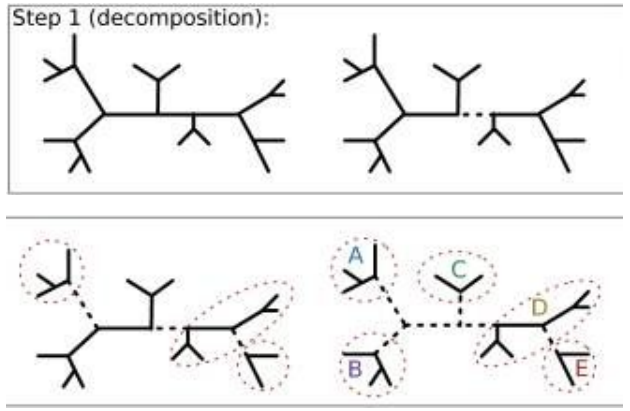
Divide-and-Conquer with pplacer
Placement of Query Sequences in Large Trees

Elizabeth and Malachi

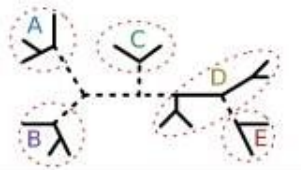
Input:
Aligned Sequences
+
Backbone Tree



Step 1:
Decompose backbone tree

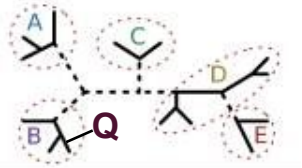


Step 2:
Run pplacer on subtrees
For each query sequence



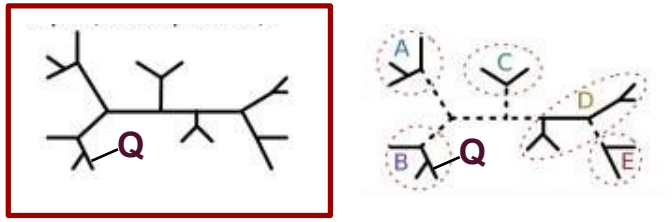
Score A, Score B, Score C,
Score D, Score E

Step 3:
Compare subtree placements
and select best placement



Step 4:
Return backbone tree with placement

Result



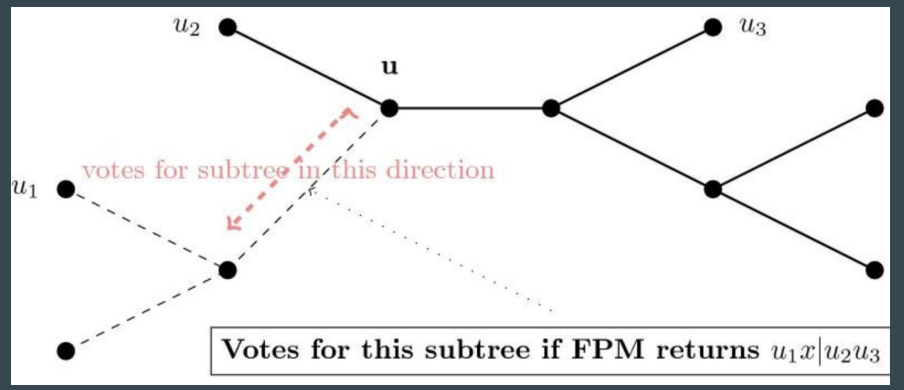
Tree images modified
from PASTA

Gillian

INC Improvements

Background:

- Incremental tree-building (INC)
- StitchINC as recursive SPRs



Goal: Improve INC's accuracy on large-scale phylogenies with stitchINC.

- Aim 1: Finish implementing and testing stitchINC.
- Aim 2: Benchmark stitchINC with INC:
  - Use subtrees from FastME as initial set decomposition
  - Run on sate-II datasets (1000M1)

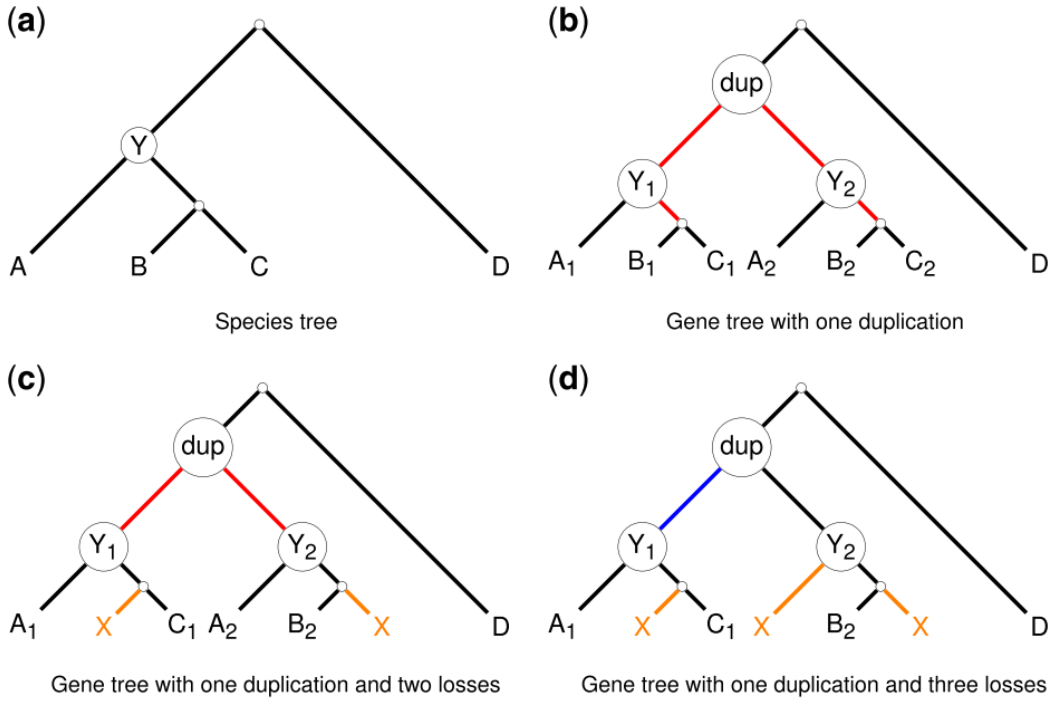
Zhang, Q., Rao, S., & Warnow, T. (2019). Constrained incremental tree building: new absolute fast converging phylogeny estimation methods with improved scalability and accuracy. *Algorithms for molecular biology : AMB*, *14*, 2. https://doi.org/10.1186/s13015-019-0136-9

FastMulRFS vs. ASTRAL-PRO

James and Mrinmoy

- **FastMulRFS:** modification (allowing for multi trees) of FastRFS, which uses a constraint set to solve the RF supertree problem.
- **ASTRAL-PRO:** modification of ASTRAL that solves a slightly different optimization problem comparing single copy to multi-copy trees.

Goal: Compare results of these two methods on both real and simulated GDL data under a variety of conditions in order to get an idea of what the relative strengths and weaknesses of the methods are.



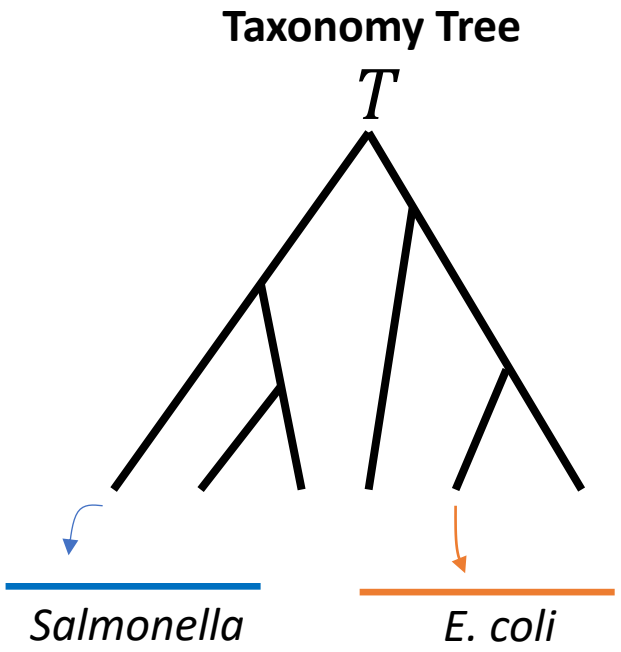
Example of Gene duplication and loss (specifically adversarial GDL)

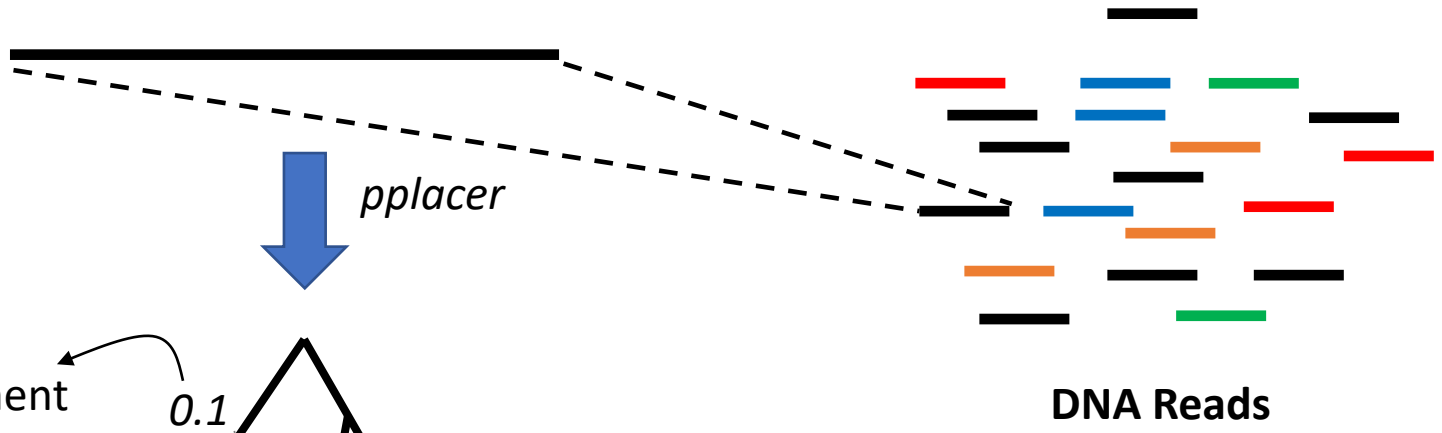
Picture from:

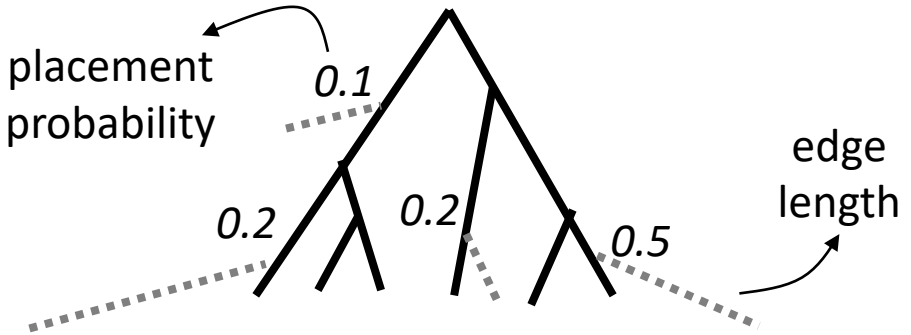
Erin K Molloy and Tandy Warnow. FastMulRFS: fast and accurate species tree estimation under generic gene duplication and loss models. *Bioinformatics*, 36(Supplement$_1$) : $i57 - -i65, 072020$.

Novel Species Discovery with *Tipp*

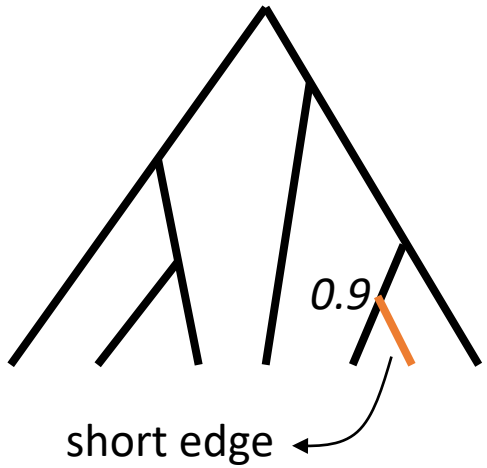
Grant and Vishal



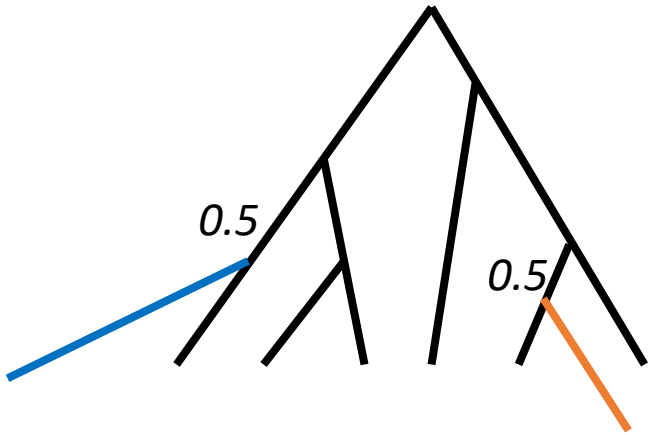




Known species (*E. coli*)



Novel species (*Colmonella*)



Novel species (*B. coli*)