Exploring the Effect of Iteration on MAGUS

Ian Chen

CS 581 (Spring 2025): Algorithmic Computational Genomics

May 6, 2025

Table of Contents

Overview

- Estimate initial tree T_0
- For $i = 1, 2, 3, 4$:
  - Split input into constraint subsets using T_{i-1}
  - Generate constraint alignments (MAFFT -L-ins-i)
  - Merge into alignment (Graph Clustering Merger)
  - Generate guide tree T_i

Overview

- ClustalOmega vs MAGUS initial tree
- Experiment One: Comparing MAGUS Pipelines
  - Subset size (10, 25, 50, 100, 200)
  - GuideTree method (FastTree2, FastTree -noml)
- Experiment Two: Comparative Study
  - MAGUS(clustalo), MAGUS(magus)
  - MAGUS (default)
  - PASTA
  - MUSCLE
  - MAFFT
  - ClustalOmega

Overview

- ClustalOmega vs MAGUS initial tree
- Experiment One: Comparing MAGUS Pipelines
  - Subset size (10, 25, 50, 100, 200)
  - GuideTree method (FastTree2, FastTree -noml)
- Experiment Two: Comparative Study
  - MAGUS(clustalo), MAGUS(magus)
  - MAGUS (default)
  - PASTA
  - MUSCLE
  - MAFFT
  - ClustalOmega

Overview

- ClustalOmega vs MAGUS initial tree
- Experiment One: Comparing MAGUS Pipelines
  - Subset size (10, 25, 50, 100, 200)
  - GuideTree method (FastTree2, FastTree -noml)
- Experiment Two: Comparative Study
  - MAGUS(clustalo), MAGUS(magus)
  - MAGUS (default)
  - PASTA
  - MUSCLE
  - MAFFT
  - ClustalOmega

Datasets

All data are publicly available at the Illinois Data Bank.

- ROSE
  - 1000M1-M3, 1000L1-L3
  - 10 replicates
- RNASim
  - 1K, 10K subsampled sequences
  - 10 replicates
- 16S.3, 16S.T, 16S.B.ALL
  - Filter sequences lengths within 0.8x and 1.2x median

Criteria

Experiment One:

- SPFN, SPFP score (alignment accuracy, lower is better)
- RF score (guide tree accuracy, lower is better)
- Runtime (wall-clock time)

Experiment Two:

- SPFN, SPFP score (alignment accuracy, lower is better)
- Runtime (wall-clock time)

Criteria

Experiment One:

- SPFN, SPFP score (alignment accuracy, lower is better)
- RF score (guide tree accuracy, lower is better)
- Runtime (wall-clock time)

Experiment Two:

- SPFN, SPFP score (alignment accuracy, lower is better)
- Runtime (wall-clock time)

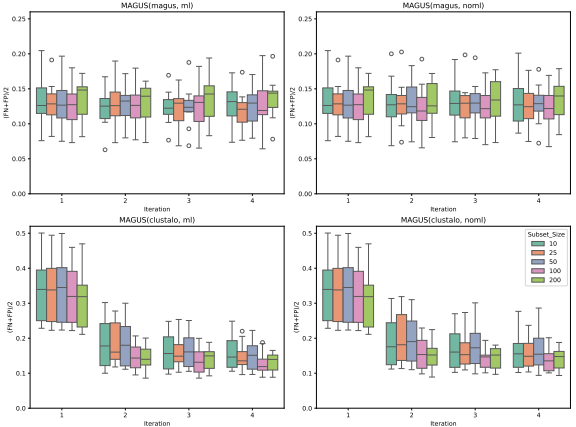
Computational Resources

All experiments are on the Illinois Campus Cluster

- 12 hour time limit
- 64 cores

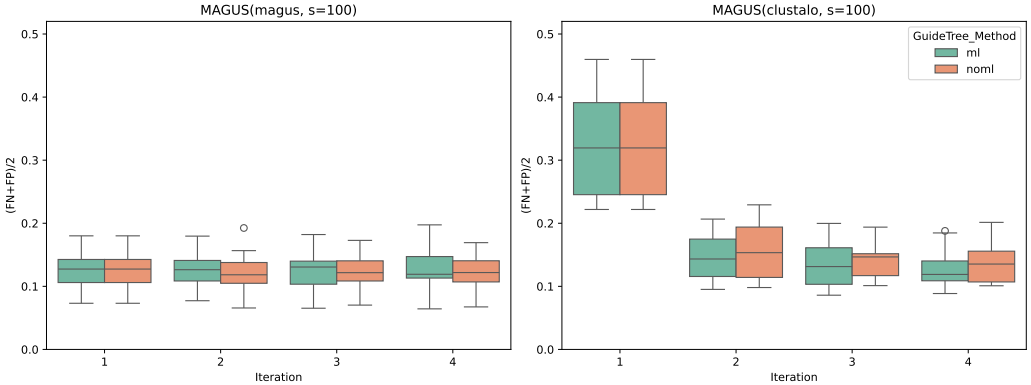
Experiment One: Comparing MAGUS Pipelines

Constraint subset size 100 is best



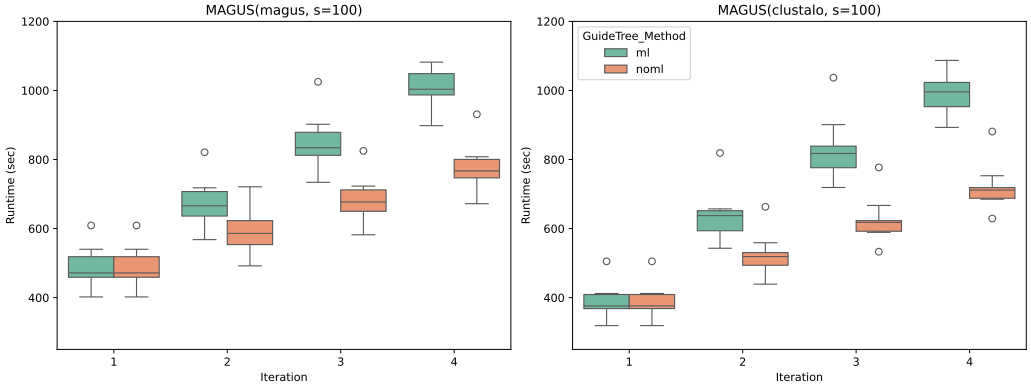
Experiment One: Comparing MAGUS Pipelines

Using ML heuristic is better accuracy . . .



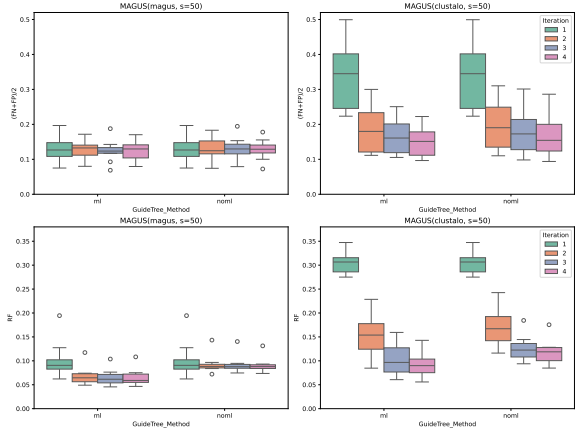
Experiment One: Comparing MAGUS Pipelines

Using ML heuristic is better accuracy . . . but it is also slower



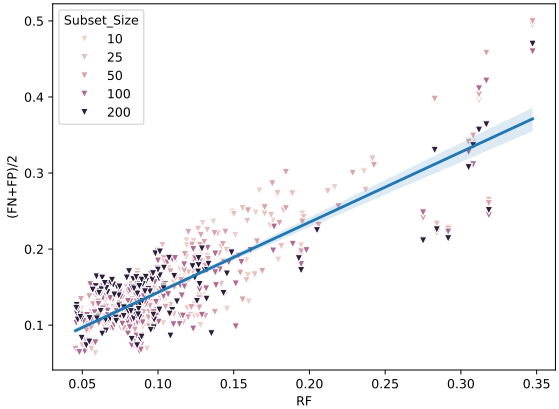
Experiment One: Comparing MAGUS Pipelines

Iteration improves accuracy when the accuracy is poor ($> 15\%$ error)



Experiment One: Comparing MAGUS Pipelines

The accuracy of guide tree matters ($R^2 = 0.775$)



Experiment One: Comparing MAGUS Pipelines

MAGUS(clustalo)

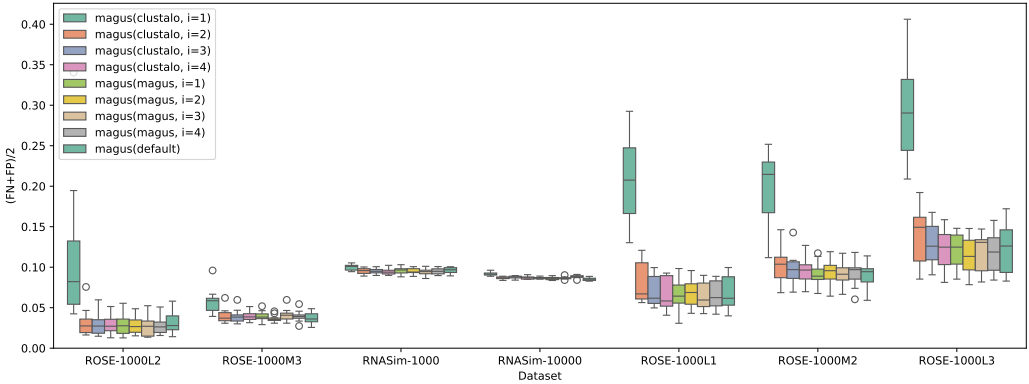
- $s = 100$
- FastTree -noml

MAGUS(magus)

- $s = 100$
- FastTree -noml

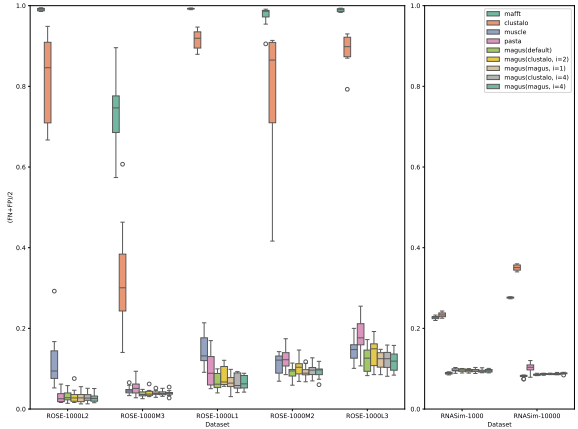
Experiment Two: Comparative Study

1000L3 and 1000M2 are the hardest conditions . . . it takes 3 iterations for MAGUS(clustalo) to match accuracy of default MAGUS but 2 iterations on everything else



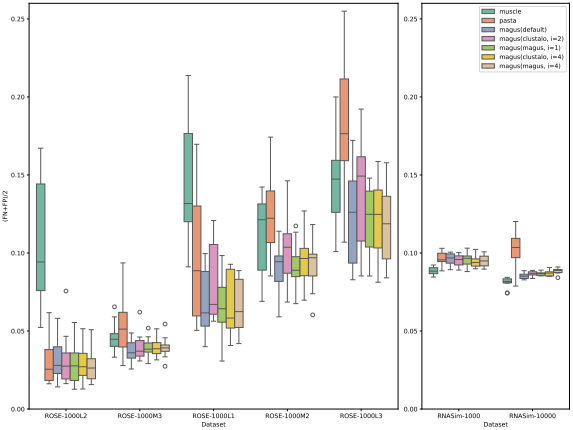
Experiment Two: Comparative Study

ClustalOmega and MAFFT are performing poorly



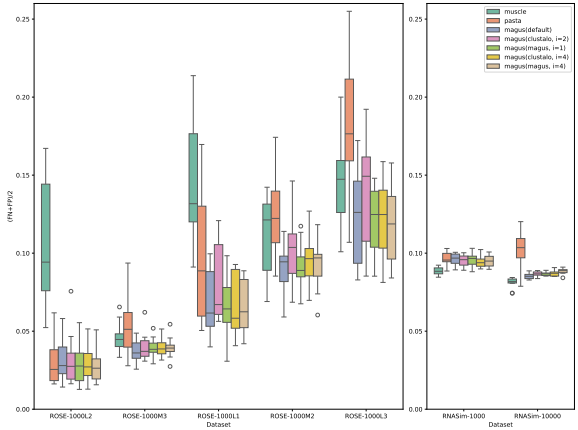
Experiment Two: Comparative Study

All MAGUS pipelines are better than PASTA



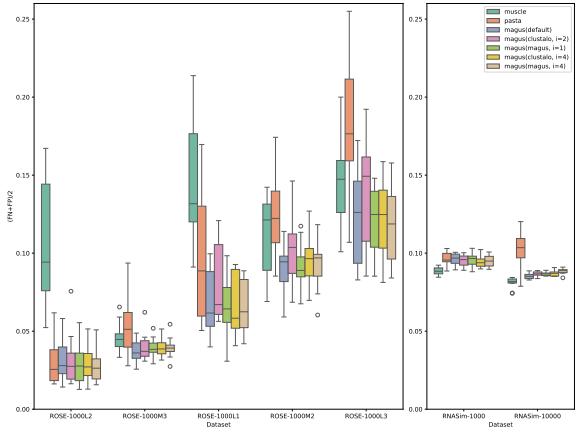
Experiment Two: Comparative Study

MUSCLE is worse than MAGUS on ROSE conditions . . .



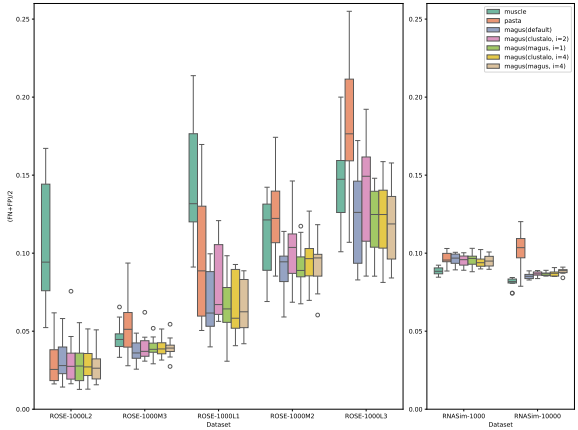
Experiment Two: Comparative Study

MUSCLE is worse than MAGUS on ROSE conditions . . . but better on RNASim



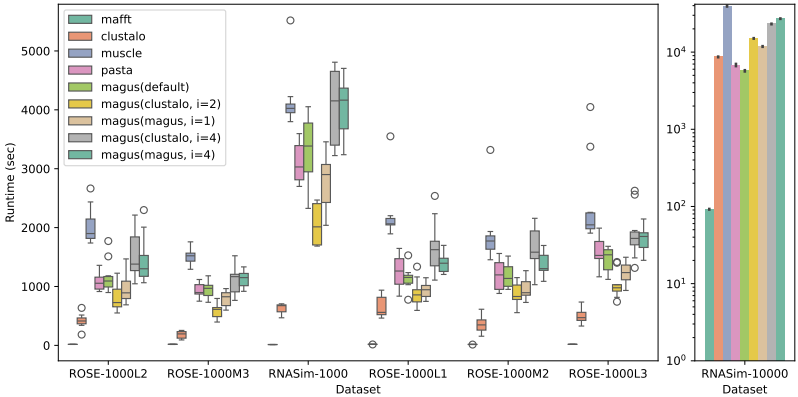
Experiment Two: Comparative Study

1000L3 and 1000M2 are the hardest conditions . . .



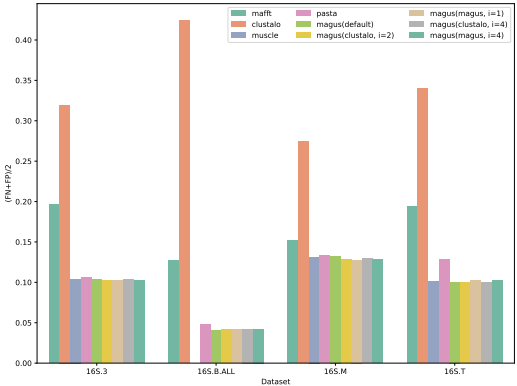
Experiment Two: Comparative Study

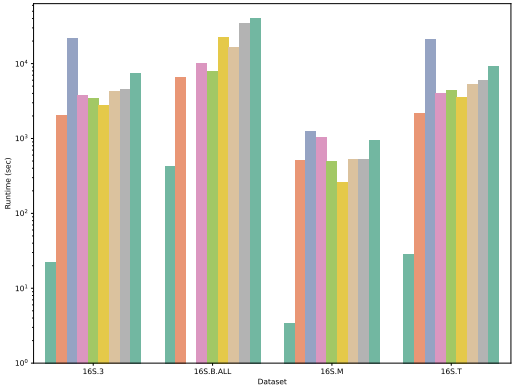
MUSCLE is slowest, MAFFT is fastest. MAGUS is slow



Experiment Two: Comparative Study

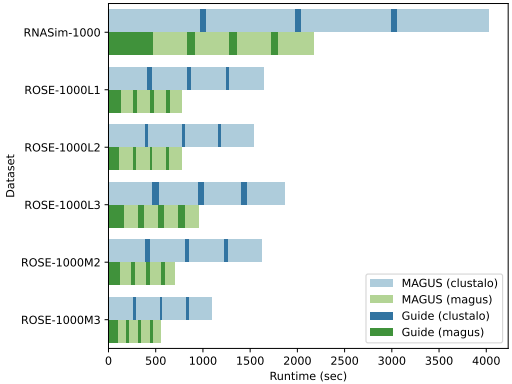
16S is too "easy" to distinguish across methods

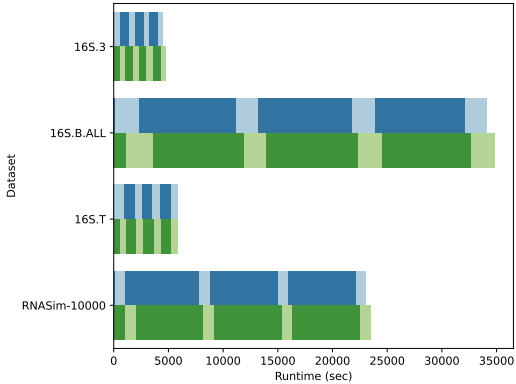




Experiment Two: Comparative Study

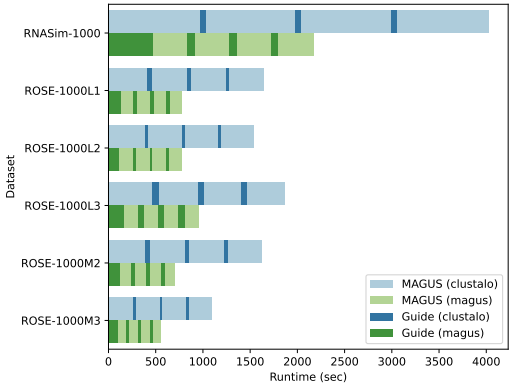
On 1000 sequence datasets, most of runtime is in estimating backbone alignments ...

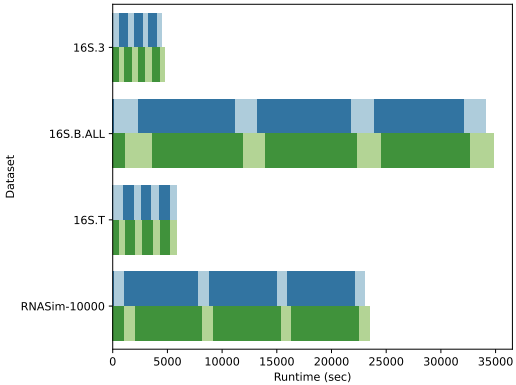




Experiment Two: Comparative Study

On 1000 sequence datasets, most of runtime is in estimating backbone alignments . . . but with larger datasets, most of runtime is in estimating guide tree





Takeaways

- The guide tree matters for MAGUS
- Iteration can overcome a bad initial tree
- Iteration more important on harder datasets

Takeaways

- Plan out how long the experiment will take
- Identify runtime bottlenecks
  - FastTree is largely not parallelizable
  - Should have run in parallel
- Account for mistakes
  - Ran out of storage (from not deleting temporary files)
  - Accidentally deleting data
  - Accidentally overriding data

Future Work

- Use an even better guide tree method?
  - For example, GTM pipelines?
- Change the base method
- Change the parameters in GCM
  - Change number of backbone alignments
  - Change the clustering method

Thank You

Thanks to Professor Warnow and TA Eleanor for the guidance throughout the semester!

Bibliography I

Jamie J Cannone, Sankar Subramanian, Murray N Schnare, James R Collett, Lisa M D'Souza, Yushi Du, Brian Feng, Nan Lin, Lakshmi V Madabusi, Kirsten M Müller, Nupur Pande, Zhidi Shang, Nan Yu, and Robin R Gutell. 2002.
The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron , and other RNAs.
BMC Bioinformatics 3, 1 (Jan. 2002), 2.
https://doi.org/10.1186/1471-2105-3-2

Robert C. Edgar. 2004.
MUSCLE: a multiple sequence alignment method with reduced time and space complexity.
BMC Bioinformatics 5, 1 (19 Aug 2004), 113.
https://doi.org/10.1186/1471-2105-5-113

Bibliography II

Kazutaka Katoh and Daron M. Standley. 2013.
MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.
Molecular Biology and Evolution 30, 4 (01 2013), 772–780.
https://doi.org/10.1093/molbev/mst010
https://academic.oup.com/mbe/article-pdf/30/4/772/6420419/mst010.pdf.

Kevin Liu, Sindhu Raghavan, Serita Nelesen, C. Randal Linder, and Tandy Warnow. 2009.
Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees.
Science 324, 5934 (2009), 1561–1564.
https://doi.org/10.1126/science.1171243
https://www.science.org/doi/pdf/10.1126/science.1171243.

Bibliography III

Siavash Mirarab, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. 2014.
PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences.
J Comput Biol 22, 5 (Dec. 2014), 377–386.
https://doi.org/10.1089/cmb.2014.0156

Siavash Mirarab and Tandy Warnow. 2011.
FASTSP: linear time calculation of alignment accuracy.
Bioinformatics 27, 23 (10 2011), 3250–3258.
https://doi.org/10.1093/bioinformatics/btr553
https://academic.oup.com/bioinformatics/article-pdf/27/23/3250/48862947/bioinformatics_27_23_3250.pdf.

Bibliography IV

Fabian Sievers and Desmond G Higgins. 2017.
Clustal Omega for making accurate alignments of many protein sequences.
Protein Sci 27, 1 (Oct. 2017), 135–145.
https://doi.org/10.1002/pro.3290

Vladimir Smirnov and Tandy Warnow. 2020.
MAGUS: Multiple sequence Alignment using Graph clUStering.
Bioinformatics 37, 12 (2020), 1666–1672.
https://doi.org/10.1093/bioinformatics/btaa992
https://academic.oup.com/bioinformatics/article-pdf/37/12/1666/39119282/btaa992.pdf.