SAND2022-12177C

Scalable Alignment and Tree Estimation on Large Protein Datasets

Chengze Shen

Baqiao Liu

Minhyuk Park

(plus Tandy Warnow and Kelly Williams)

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

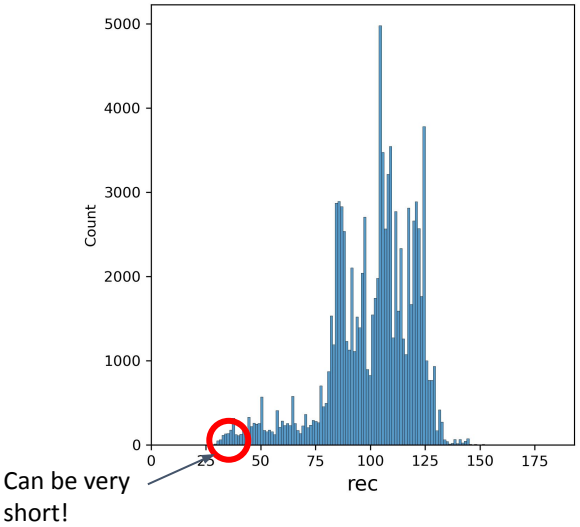
Motivation

- Bacterial genomes can collect mobile DNA elements, including prophages
  - They deliver genes modulating bacterial pathogenicity, metabolism, etc.
  - Site-specific chromosomal integration requires **integrase enzymes**
- Two main integrase protein families:
  - **Tyrosine integrase**
  - **Serine integrase** (with two main domains: **Res**olvase and **Rec**ombinase)
- The evolutionary history of the integrase proteins would help us better understand and interpret the integration process

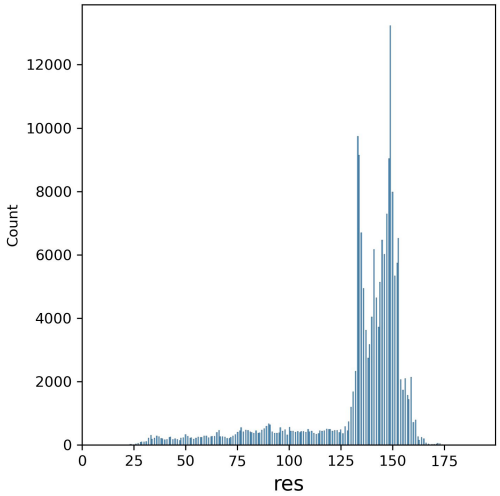
Algorithmic challenges for these datasets

- **Phylogenetic tree** estimation is challenging because
  a. It requires multiple sequence alignments
  b. Maximum Likelihood Tree estimation is NP-hard and the best heuristics (e.g., RAxML-ng and IQ-TREE 2) fail on ultra-large datasets
- **Multiple Sequence Alignment (MSA)** estimation is challenging because
  a. Most MSA methods do not run on large datasets
  b. Accuracy degrades with sequence length heterogeneity
- These datasets are very large (96,000 to > 700,000 sequences) and exhibit substantial sequence length heterogeneity

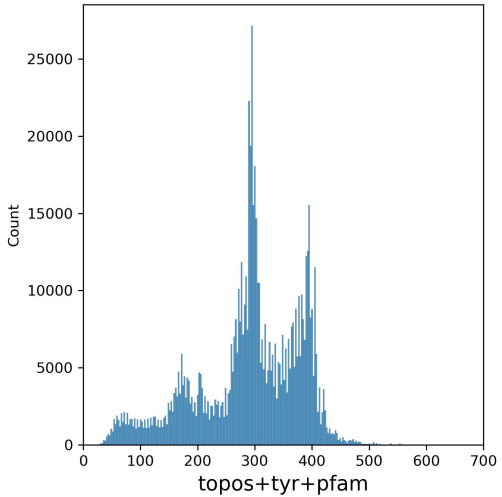
Dataset information



Rec: **96,773** sequences



Res: **186,802** sequences



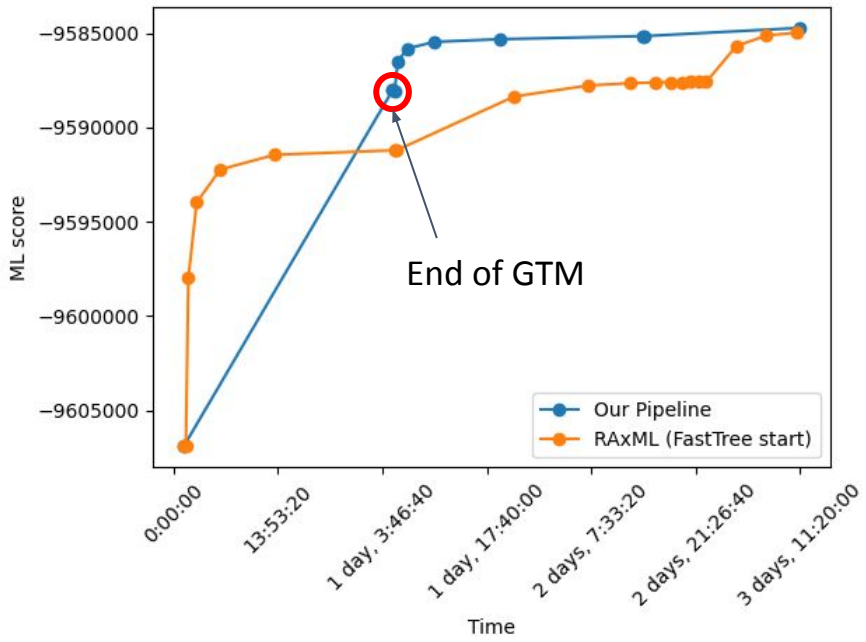
Tyr: **721,526** sequences

- Sequences are taken from 350,378 bacterial and archaeal genomes using Prodigal (Hyatt 2010)
- Specific domains (e.g., **Rec**ombinase, **Res**olvase or **Tyr**osine) are identified using corresponding Pfam Hidden Markov Models (Mistry 2021, Smyshlyaev 2021)
- Seed sequences from the HMMs are included

What we did

- (Alignment) We developed new MSA methods that produce more accurate MSAs on large datasets with sequence length heterogeneity than current methods – Chengze Shen and Baqiao Liu
- (Tree estimation) We developed a new approach to large-scale maximum likelihood phylogeny estimation that is able to produce better maximum likelihood scores than leading heuristics (RAxML, IQ-TREE 2, and FastTree) – Minhyuk Park

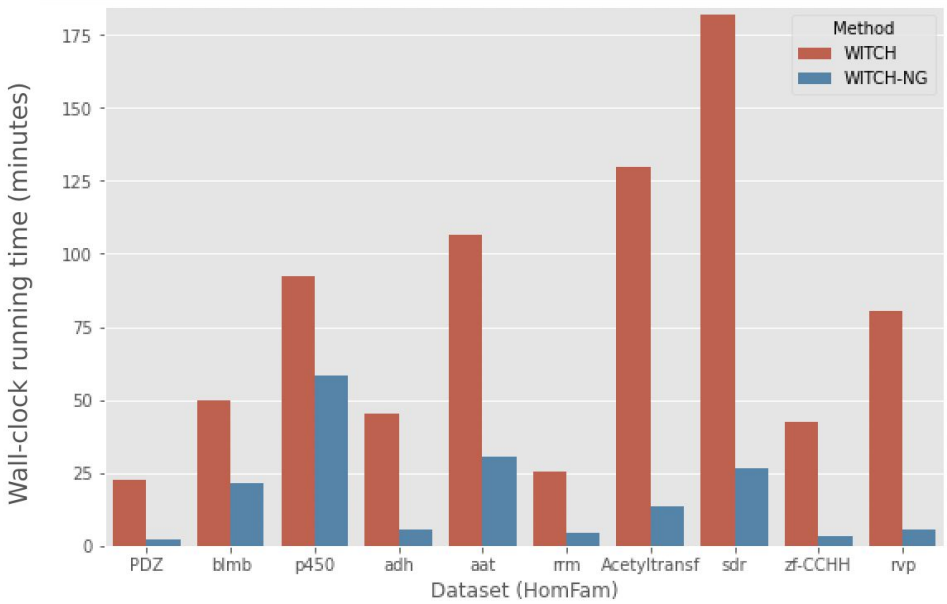
GTM - large scale tree estimation on 78k-sequence dataset



- Guide Tree Merger (Smirnov and Warnow 2021):
  a. Divides input sequences into subsets
  b. Uses RAxML on subsets
  c. Combines trees
- Our pipeline: given starting tree, run GTM and continue with RAxML
- Observation: We find better ML scores than RAxML with the same starting tree

Figure caption: Our pipeline and RAxML (FastTree start) begin with FastTree's topology and ML score. Our pipeline uses a divide-and-conquer pipeline followed by RAxML for further optimization.

WITCH-ng - improved runtime for WITCH alignment



Running time comparison in phase 3 (sequence adding) between WITCH-ng and WITCH (lower is better) on large Homfam data (n >= 14950).

- Large scale data – running time matters
- WITCH – accurate for sequence length heterogeneity, but can be slow
- WITCH-ng: makes WITCH faster
  a. Same except phase 3
  b. Simpler algorithmic design
  c. Better algorithmic engineering

SALMA - Scalable Alignment using MAFFT-Add

- A divide-and-conquer method improving on UPP (Nguyen et al. 2015)
- Designed to align hundreds of thousands of sequences with sequence length heterogeneity

What it does:

1. Select subset *S'* (full-length sequences) from *S* (all input sequences)
2. Align and build tree *T* on *S'*
3. Decompose tree *T* into small subsets
4. Assign *S* \ *S'* to their best subset using HMMs
5. Use MAFFT to add *S* \ *S'* into assigned subset alignments
6. Merge extended subset alignments

Accuracy on Tyrosine (>700k sequences)

	Within-Pfam	Cross-Pfam
SALMA	**96.2%**	**62.1%**
WITCH-ng	92.6%	55.5%
UPP	82.4%	59.7%
FAMSA	87.0%	9.8%

Table 1. Normalized homology scores (sensitivity according to Pfam seed sequences) for alignments on Tyrosine (>700k sequences).

- SALMA results are promising.
- Many standard methods failed to run (e.g., Clustal-Omega, MAFFT, PASTA).

Conclusion and Future Work

- We are using SALMA for alignments
- Next steps
  - Improve our ML tree heuristics - Minhyuk Park
  - Align and construct trees for both integrase protein families - Baqiao Liu, Chengze Shen, Minhyuk Park, and Tandy Warnow
  - Interpret trees and biological discovery - Kelly Williams

Contact all of us!

chengze5@illinois.edu, baqiaol2@illinois.edu, minhyuk2@illinois.edu

warnow@illinois.edu, kpwilli@sandia.gov

Funding: (a) LDRD from Sandia Lab to KW, (b) NSF grant #2006069 to TW