CS 581 COURSE PROJECT PRESENTATION
DACTAL V2: RE-IMPLEMENTATION OF DACTAL

SAYANTANI BASU

SUPERVISED BY PROFESSOR TANDY WARNOW

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

INTRODUCTION

- **DACTAL** – Nelesen et al. (2012) [1]
- **Input** – Set of unaligned sequences, **Output** – Tree on entire dataset
- Two-phase methods – Computationally intensive
- Alternative – Alignment-free methods
- ***Divide-and-Conquer strategy + Iteration***

MOTIVATION

- Interesting to **change certain modules** of DACTAL [1] pipeline
- **Impact of starting tree** on the pipeline
- Behavior of DACTAL [1] on **large datasets**
- Re-implemented pipeline – **DACTAL v2**

EXISTING DACTAL PIPELINE

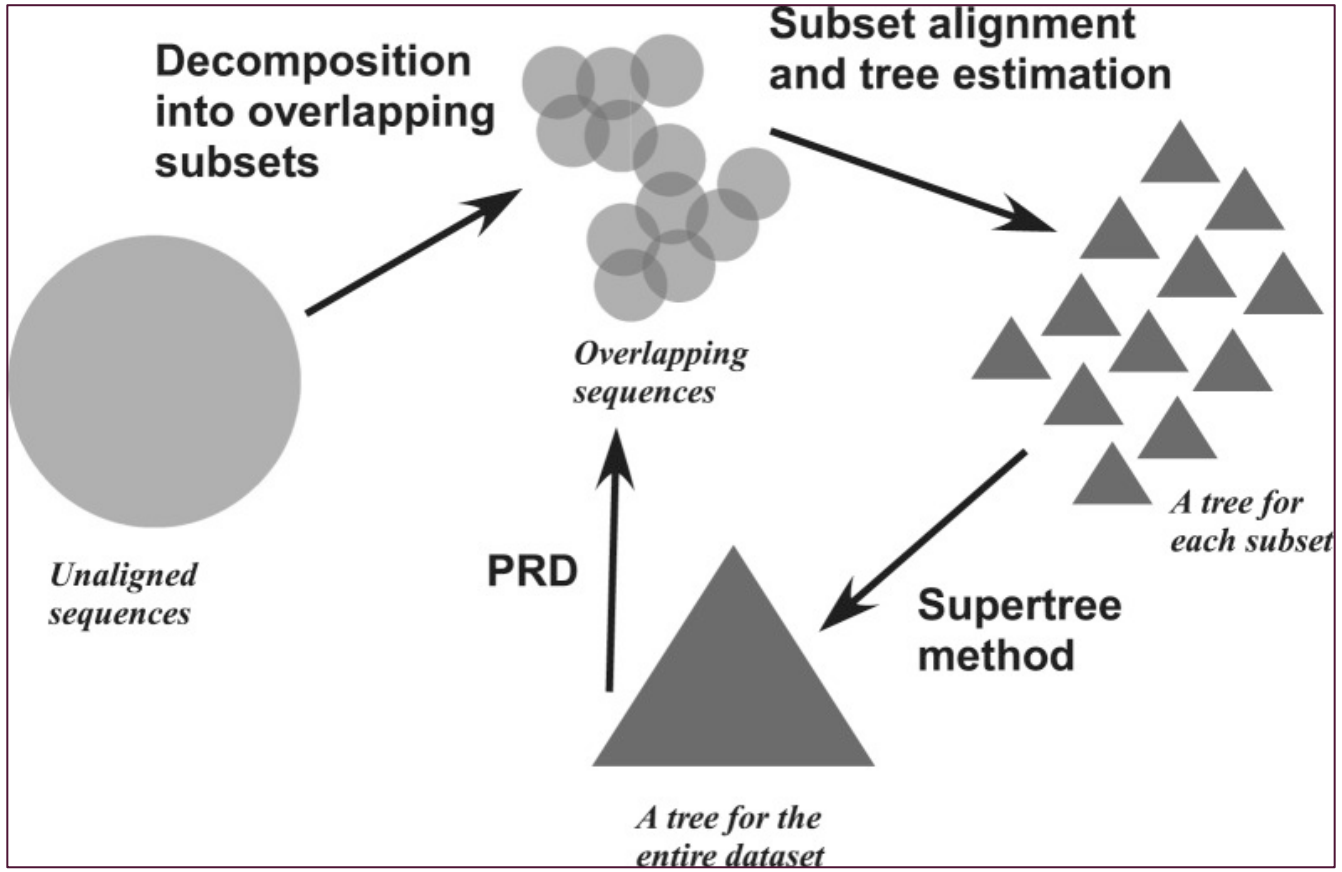
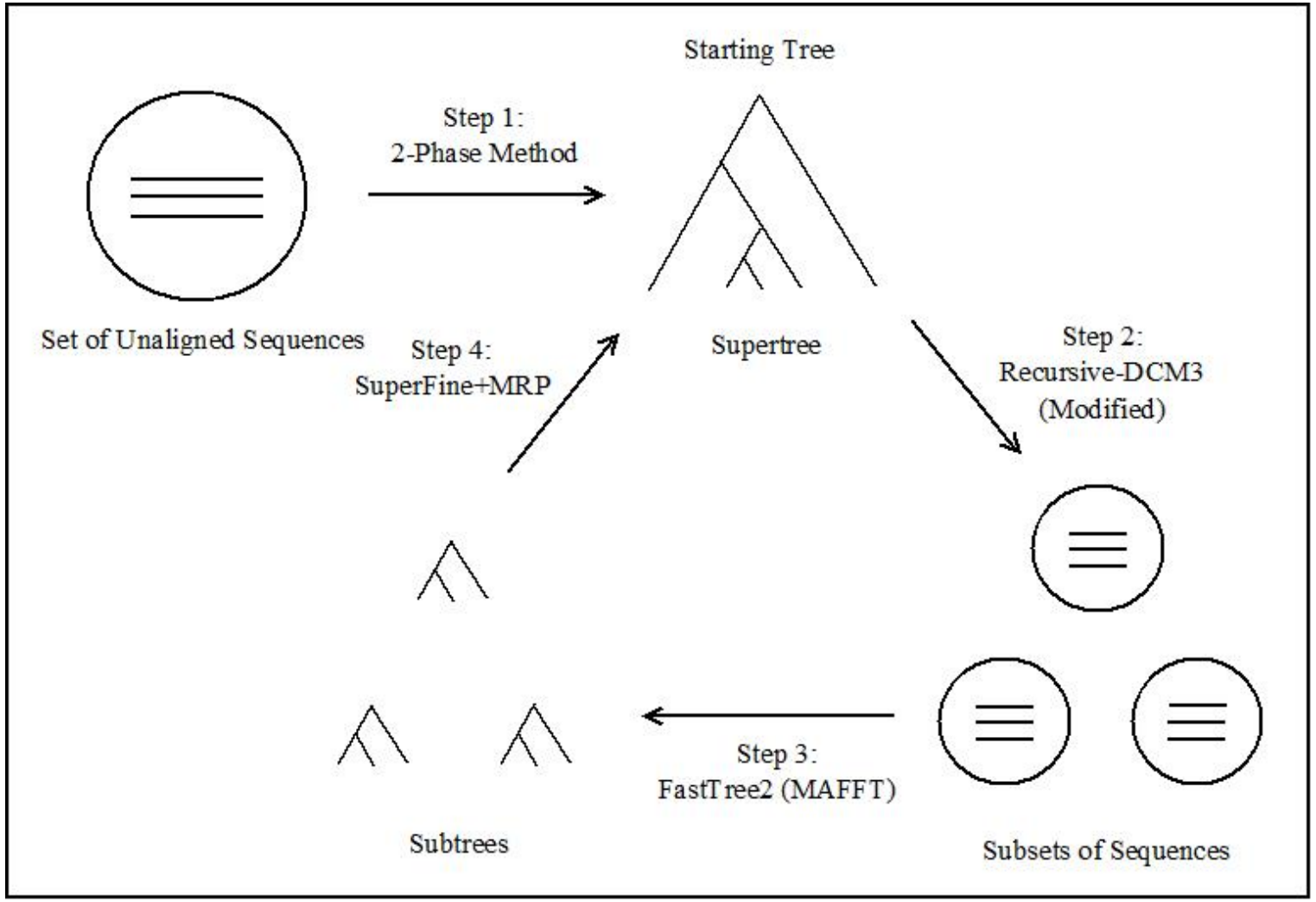


Image taken from Nelesen et al., 2012 [1]

PROPOSED DACTAL V2 PIPELINE



EXPERIMENTAL SETUP

- **Initial Tree** – used to get a starting tree for the pipeline
- **Parameters** – maximum subset size (s) and padding size (p),
  also number of iterations (maxiter)
- **Datasets** – Simulated ROSE Datasets [2]
- **Testing** – UIUC Campus Cluster
- **RF distance** – compare the ***error*** between the initial and final trees

$$RF\ Error\ Rate = \frac{FP + FN}{2n - 6}$$

where n= number of leaves

SOFTWARE USED

- Python 3
- MAFFT v7.407 [3]
- Clustal Omega v1.2.4 [4]
- FastTree2 v2.1.3 [5]
- RecDCM3 (Modified) [6]
- SuperFine [7]
- PAUP* [8]
- DendroPy 4 [9]

TESTING AND RESULTS TILL DATE (1)

Starting Tree	Dataset	No. of Iterations	Maximum subset size	Padding size	Initial Tree Error (RF)	Tree Error (RF) after Iteration(s)
FastTree2 on MAFFT	500L1	1	250	1	15.895%	16.398%
FastTree2 on MAFFT	500L1	1	250	2	15.895%	16.398%
FastTree2 on MAFFT	500L1	1	250	3	15.895%	16.398%

TESTING AND RESULTS TILL DATE (2)

Starting Tree	Dataset	No. of Iterations	Maximum subset size	Padding size	Initial Tree Error (RF)	Tree Error (RF) after Iteration(s)
FastTree2 on ClustalO	500L1	1	250	2	38.833%	30.785%
FastTree2 on ClustalO	1000M1	1	500	2	56.219%	50.853%
FastTree2 on ClustalO	1000M1	3	500	2	56.219%	30.943%
FastTree2 on ClustalO	1000M1	3	500	3	56.219%	31.494%

REFERENCES

[1] Nelesen, S., Liu, K., Wang, L.S., Linder, C.R. and Warnow, T., 2012. DACTAL: divide-and-conquer trees (almost) without alignments. *Bioinformatics*, *28*(12), pp.i274-i282.

[2] https://sites.google.com/eng.ucsd.edu/datasets/alignment/sate-i

[3] Katoh, K. and Standley, D.M., 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), pp.772-780.

[4] Sievers, F. and Higgins, D.G., 2014. Clustal Omega, accurate alignment of very large numbers of sequences. In *Multiple sequence alignment methods* (pp. 105-116). Humana Press, Totowa, NJ.

[5] Price, M.N., Dehal, P.S. and Arkin, A.P., 2010. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one*, *5*(3), p.e9490.

[6] Roshan, U., Moret, B.M., Williams, T.L. and Warnow, T., 2004. Rec-I-DCM3: A fast algorithmic technique for reconstructing large phylogenetic trees. In *Proc. 3rd IEEE Computational Systems Bioinformatics Conf. CSB" 04* (No. LCBB-CONF-2004-002, pp. 98-109). IEEE Press.

[7] Swenson, M.S., Suri, R., Linder, C.R. and Warnow, T., 2011. SuperFine: fast and accurate supertree estimation. *Systematic biology*, *61*(2), p.214.

[8] Swofford, D.L., 2001. PAUP*: Phylogenetic analysis using parsimony (and other methods) 4.0. B5.

[9] Sukumaran, J. and Holder, M.T., 2010. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, *26*(12), pp.1569-1571.