

---

---

# Protein datasets analysis using PASTA

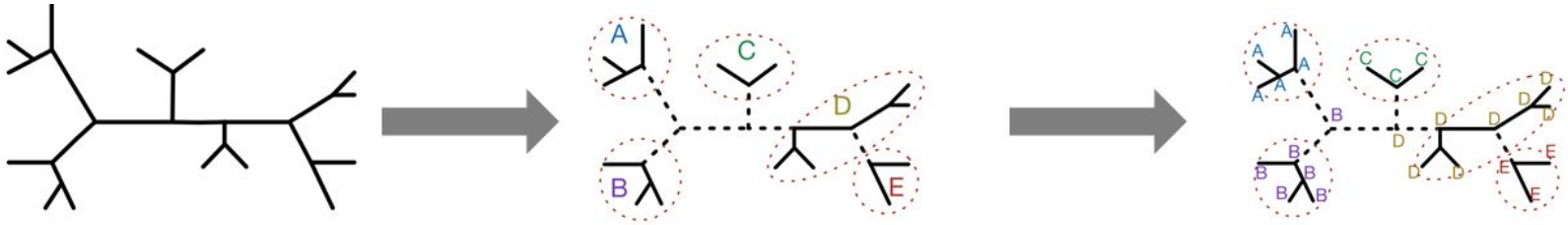
— Ashok Arjunakani & Lily Barghi —

---

---

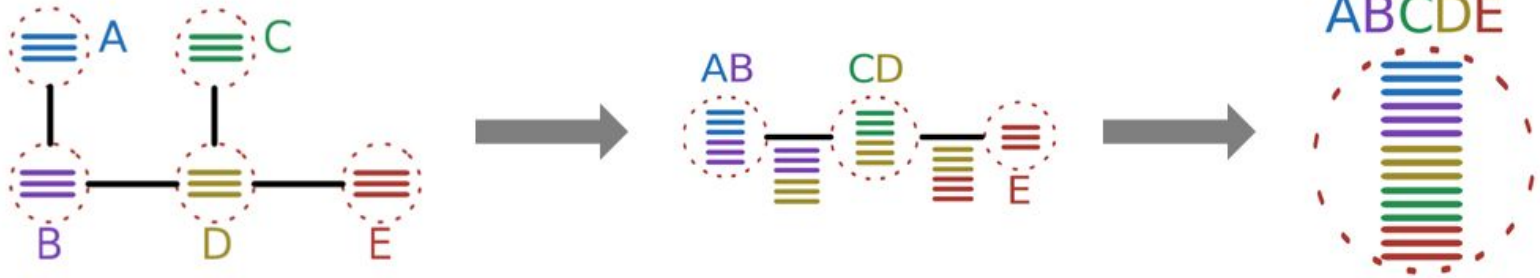
# PASTA Algorithm

1. Obtain initial alignment
2. Decompose Tree into subsets



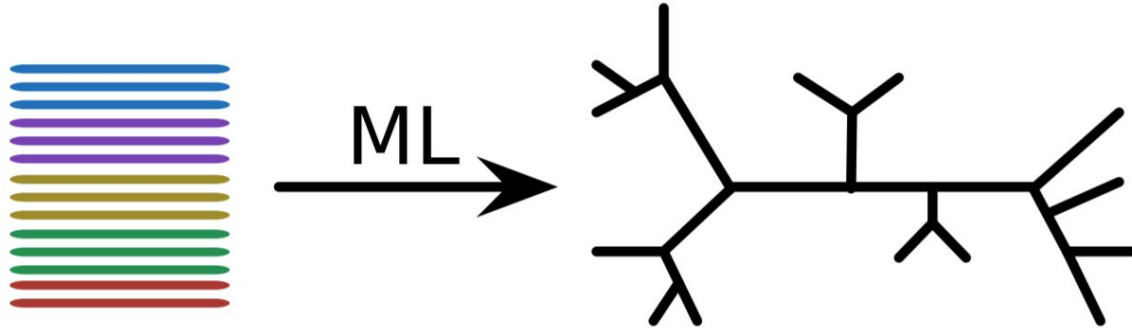
# PASTA Algorithm

- Subsets aligned and merged together to create multiple sequence alignment



# PASTA Algorithm

4. Use Maximum Likelihood tree to compute tree estimation



# Our Project

- Use the different multiple sequence alignment methods and tree estimation methods found in PASTA to assess which collection of methods causes the best accuracy and efficiency
- Using aligned Protein Sequences
- Generated alignment and tree will be compared to reference tree and alignment.
  - Reference tree is generated by using RAxML

# Protein Datasets

1. BlaR1 peptidase: 2,747 sequences
  - a. <http://merops.sanger.ac.uk/cgi-bin/merops.cgi?id=M56>
2. Ulp1 peptidase: 5,954 sequences
  - a. <http://merops.sanger.ac.uk/cgi-bin/famsum?family=C48>
3. Amidophosphoribosyltransferase precursor: 11,342 sequences
  - a. <http://merops.sanger.ac.uk/cgi-bin/famsum?family=C44>
4. Dipeptidyl-peptidase VI: 13,452 sequences
  - a. <http://merops.sanger.ac.uk/cgi-bin/famsum?family=C40>
5. Pepsin A: 16,022 sequences
  - a. <http://merops.sanger.ac.uk/cgi-bin/famsum?family=A1>

# Checking Accuracy of MSA & Trees

- FastSP: checks accuracy of multiple sequence alignment (MSA) based on SPFN
  - SPFN: comparison to reference alignment to see how many amino acids are missing in estimated sequence alignment
- Reference trees will be created using RAxML and compared using Robinson-Foulds (RF) distance from Dendropy
  - Script from Erin Molloy's presentation in CS466
  - RF distance: number of edges two trees do not share with each other

# Bibliography

- [1] Mirarab, Siavash, Nam Nguyen, Sheng Guo, Li-San Wang, Junhyong Kim, and Tandy Warnow. "PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences." *Journal of Computational Biology* 22.5 (2015): 377-86. Web. Apr. 2017.
- [2] Walle, I. Van, I. Lasters, and L. Wyns. "SABmark--a benchmark for sequence alignment that covers the entire known fold space." *Bioinformatics* 21.7 (2004): 1267-268. Web. Apr. 2017.
- [3] Edgar, R. C. "MSA benchmark collection." *Drive5*. N.p., n.d. Web. Apr. 2017.
- [4] Thompson, Julie D., Patrice Koehl, Raymond Ripp, and Olivier Poch. "BALiBASE 3.0: Latest developments of the multiple sequence alignment benchmark." *Proteins: Structure, Function, and Bioinformatics* 61.1 (2005): 127-36. Web. Apr. 2017.



# Bibliography continued

- [5] Le, Si Quang, Nicolas Lartillot, and Olivier Gascuel. "Phylogenetic Mixture Models for Proteins." *Philosophical Transactions of the Royal Society B: Biological Sciences* 363.1512 (2008): 3965–3976. *PMC*. Web. Apr. 2017.
- [6] Mirarab, S., and T. Warnow. "FASTSP: linear time calculation of alignment accuracy." *Bioinformatics* 27.23 (2011): 3250-258. Web. Apr. 2017.
- [7] Warnow, Tandy. Computational Phylogenetics An introduction to designing methods for phylogeny estimation. N.p.: n.p., n.d. Web
- [8] Molloy, Erin. "Compare\_trees.py." *GitHub*. CS466, n.d. Web. Apr. 2017.
- [9] Rawlings, Neil D., Alan J. Barrett, and Robert Finn. "Twenty years of the MEROPS database of proteolytic enzymes, their substrates and inhibitors." *Nucleic Acids Research* 44.D1 (2015): n. pag. Web. 13 Apr. 2017.