MSA Benchmarking

Daniel Yuan and Stanley Liu

Intro

- Benchmarking 6 MSA software
  - 3 progressive methods
    - T-Coffee 11.00.8cbe486
    - MAFFT 7
    - PSAlign
  - 3 iterative methods
    - PRRN 4.1.0
    - DIALIGN 2.2.1
    - Muscle 3.8.31

- Progressive
  - Heuristics
  - Sequences -> Guide Tree -> MSA
- Iterative
  - Initial alignment -> Iteratively Realign -> MSA

Datasets

- 100M1 [7]
  - Simulated dataset
  - Dataset with 100 taxa with medium gap lengths
  - Using 10 replicates

Criteria

- Time
  - Amount of time it takes each software to run
- Accuracy
  - SP-score of estimated aligned sequence compared to the original true alignment from dataset
  - FastSP
    - Memory-efficient java app that can score alignments against a reference
- Efficiency
  - Normalized accuracy/time

References

[1] T. Warnow, Computational Phylogenetics

[2] Notredame, C. (n.d.). *T-Coffee Home Page*. [online] Tcoffee.org. Available at: http://www.tcoffee.org/Projects/tcoffee/#DOCUMENTATION [Accessed 9 Apr. 2017].

[3] Katoh, K. (2013). *MAFFT - a multiple sequence alignment program*. [online] Mafft.cbrc.jp. Available at: http://mafft.cbrc.jp/alignment/software/ [Accessed 9 Apr. 2017].

[4] En.wikipedia.org. (2017). *Multiple sequence alignment*. [online] Available at: https://en.wikipedia.org/wiki/Multiple_sequence_alignment#Iterative_methods [Accessed 3 Apr. 2017].

[5] Gotoh, O. (1997). *PRRN information*. [online] Genome.ist.i.kyoto-u.ac.jp. Available at: http://www.genome.ist.i.kyoto-u.ac.jp/~aln_user/prrn/index.html [Accessed 9 Apr. 2017].

[6] Morgenstern, B. and Abbedaim, S. (1999). *DIALIGN 2.2.1 User Guide*. [online] Hpcwebapps.cit.nih.gov. Available at: https://hpcwebapps.cit.nih.gov/multi-align/man/dialign.1.html [Accessed 9 Apr. 2017].

[7] Edgar, R. (n.d.). *MUSCLE documentation*. [online] Drive5.com. Available at: http://www.drive5.com/muscle/manual/ [Accessed 9 Apr. 2017].

[8] Linder CR, Suri R, Liu K, Warnow T. Benchmark datasets and software for developing and testing methods for large-scale multiple sequence alignment and phylogenetic inference. PLOS Currents Tree of Life. 2010 Nov 18 . Edition 1. doi: 10.1371/currents.RRN1195.

[9] Liu, K., S. Raghavan, S. Nelesen, C. R. Linder, T. Warnow. 2009. Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. Science 324:1561-1564.

[10] Siavash Mirarab, Tandy Warnow; FASTSP: linear time calculation of alignment accuracy. *Bioinformatics* 2011; 27 (23): 3250-3258. doi: 10.1093/bioinformatics/btr553