SaffronTree: Fast, reference-free pseudo-phylogenomic trees from reads or contigs.

Andrew J. Page[1], **Martin Hunt**[1], **Torsten Seemann**[2], **and Jacqueline A. Keane**[1]

1 Pathogen Informatics, Wellcome Trust Sanger Institute **2** University of Melbourne

DOI: 10.21105/joss.00243

Software

- Review
- Repository
- Archive

Licence
Authors of JOSS papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Summary

When defining bacterial populations through whole genome sequencing (WGS) the samples often have unknown evolutionary histories. With the increased use of next generation WGS in routine diagnostics, surveillance and epidemiology a vast amount of short read data is available, with phylogenetic trees (dendograms) used to visualise the relationships and similarities between samples. Standard reference and assembly based methods can take substantial amounts of time to generate these phylogenetic relationships, with the computation time often exceeding the time to sequence the samples in the first place. Faster methods (Ondov et al. 2016; Wood and Salzberg 2014) can loosely classify samples into known taxonomic categories, however the loss of granularity means the relationships between samples is reduced. This can be the difference between ruling a sample in or out of an outbreak, which is a clinically important finding for genomic epidemiologists. Other methods (Boratyn et al. 2014) are closed source which prevents independent scrutiny. SaffronTree utilises the k-mer profiles between samples to rapidly construct a tree, directly from raw reads in FASTQ format or contigs in FASTA format. It supports NGS data (such as Illumina), 3rd generation long read data (Pacbio/Nanopore) and assembled sequences (FASTA). Firstly, a k-mer count database is constructed for each sample using KMC (Kokot, Długosz, and Deorowicz 2017). Next, the intersection of the k-mer databases is found for each pair of samples, with the number of k-mers in common recorded in a distance matrix. Finally, the distance matrix is used to construct a UPGMA tree (Sokal and Michener 1958) in Newick format. This tree method was chosen as it is fast, however the final result is lower quality than slower methods which perform ancestral sequence reconstructions (Stamatakis 2014). The computational complexity of the algorithm is O(N^2), so is best suited to datasets of less than 50 samples. This can give rapid insights into small datasets in minutes, rather than hours. SaffonTree provides better granularity than MLST as it uses more of the underlying genome, can operate at low depth of coverage, is reference free, species agnostic, and has a low memory requirement.

References

Boratyn, Greg, Christiam Camacho, Scott Federhen, Yuri Merezhuk, Tom Madden, Conrad Schoch, and Irena Zaretskaya. 2014. "MOLE-Blast a New Tool to Search and Classify Multiple Sequences." ftp://ftp.ncbi.nlm.nih.gov/blast/documents/moleblast_poster2014.pdf.

Kokot, M., M. Długosz, and S. Deorowicz. 2017. "KMC 3: counting and manipulating k-mer statistics." *ArXiv E-Prints*, January.

Ondov, Brian D, Todd J Treangen, Páll Melsted, Adam B Mallonee, Nicholas H Bergman, Sergey Koren, and Adam M Phillippy. 2016. "Mash: fast genome and metagenome distance estimation using MinHash." *Genome Biology* 17 (1): 1–14. doi:10.1186/s13059-016-0997-x.

Sokal, Robert R., and Charles D. Michener. 1958. "A Statistical Method for Evaluating Systematic Relationships." *The University of Kansas Science Bulletin* 38: 1409–38. http://ci.nii.ac.jp/naid/10011579647/en/.

Stamatakis, Alexandros. 2014. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." *Bioinformatics (Oxford, England)* 30 (9): 1312–3. doi:10.1093/bioinformatics/btu033.

Wood, Derrick E, and Steven L Salzberg. 2014. "Kraken: ultrafast metagenomic sequence classification using exact alignments." *Genome Biology* 15 (3): R46. doi:10.1186/gb-2014-15-3-r46.