The Journal of Open Source Software

swsc: A sitewise UCE partitioner

Ryan A. Hagenson[1]

1 Omaha's Henry Doorly Zoo and Aquarium

DOI: 10.21105/joss.01116

Software

- Review
- Repository
- Archive

Submitted: 04 December 2018
Published: 05 December 2018

Summary

Ultraconserved elements (UCEs) are regions of the genome that retain partial identity across a vast number of species. This identity retention makes UCEs especially useful for inferring otherwise intractable phylogenies. UCE partitioning acts to split a UCE into three parts: variable left flank, conserved core, and variable right flank. The heightened variation found in the flanks allows for phylogenetic inferences (Crawford et al., 2012,Baca, Alexander, Gustafson, & Short (2017),Blaimer, Lloyd, Guillory, & Brady (2016),Faircloth et al. (2012),Faircloth, Sorenson, Santini, & Alfaro (2013),McCormack et al. (2012),Smith, Harvey, Faircloth, Glenn, & Brumfield (2014),Harrington2016,Moyle2016).

Based on a method originally described by (Tagliacollo & Lanfear, 2018) as Sliding-Window Site Characteristics (SWSC), `swsc` partitions UCEs based on chosen sitewise metrics such as Shannon's entropy or GC percentage. Input is either a modified Nexus file or standard FASTA+CSV files, containing the concatenated UCE sequences of individuals under analysis along with the range of each UCE in the concatenation (see `example-data/` in code repository for example formats). Output is a single CSV containing the UCE partitions. Optionally, a configuration file for`PartitionFinder2` can be produced.

The original method by (Tagliacollo & Lanfear, 2018) used a sequential, brute-force approach, considering all potential core windows from the provided minimum window size up to 1/3 of each UCE's length. This is inefficient when small minimum window sizes combined with either many UCEs or large UCEs. The method herein uses a candidate windows plus extension procedure.

Overview of `swsc`'s candidate window plus extension procedure:

1. Generate candidate windows of size `--minWin` across the UCE (both from the start forward and the end backwards)
    - Example: a UCE of length 120 and `--minWin` of 50 has forward windows of $1-50$ and $51-100$, as well as backwards windows of $81-120$ and $31-80$.
2. Find the best C candidates
    - Fitness is determined by minimum sum of square errors of sitewise metrics, minimum variance of left flank, core, right flank lengths, and user preference for `--largeCore` or not, in that order.
    - Minimum sum of square error finds the best core windows, minimum variance acts to select more centered cores, while user preference allows flexibility in desired results
3. Extend the best C candidates by 1/2 of `--minWin` in both directions; as well, consider the single maximum window made by the lowest starting position and highest stopping position among best C candidates (not extended further)

4. Find the best window within the extended candidate windows set using the same criteria as before

All UCEs are processed concurrently using the gorountines afforded by the Go programming language – further speed up is bounded by Amdahl's law (Amdahl, 1967) as only IO is done sequentially.

Statement of Need

UCEs have the ability to resolve otherwise intractable phylogeny questions, but as these questions confront resolving more distant relationships or require longer UCEs to capture enough variation in the flanks to be useful, a more efficient method is required. Using a candidate window plus extension procedure cuts the combinatorial search from definite:

$\sum_{i=0}^{i=N} \frac{n_i(n_i+1)}{2}$

where N is the number of UCEs, $n_i = (UCE_{i,L} - m \times 3) + 1$, and $UCE_{i,L}$ is the length of the i-th UCE

down to an upward bound of:

$2\sum_{i=1}^{i=N} \lfloor \frac{UCE_{i,L}}{m} \rfloor + C\frac{m(m+1)}{2} + 1$

where N is the number of UCEs, $UCE_{i,L}$ is the length of the i-th UCE, m is the minimum window size, and C is the number of best candidates to consider.

In the initial project `swsc` was built for this change equated to roughly a 10^4 order of reduction in the search space.

Acknowledgements

I acknowledge Cynthia L. Frasier, Timothy M. Sefszek, and Melissa T. R. Hawkins for their input on decisions made during development.

References

Amdahl, G. M. (1967). Validity of the single processor approach to achieving large scale computing capabilities. In *AFIPS spring joint computer conference*.

Baca, S. M., Alexander, A., Gustafson, G. T., & Short, A. E. (2017). Ultraconserved elements show utility in phylogenetic inference of Adephaga (Coleoptera) and suggest paraphyly of 'Hydradephaga'. *Systematic Entomology*, *42*(4), 786–795. doi:10.1111/syen.12244

Blaimer, B. B., Lloyd, M. W., Guillory, W. X., & Brady, S. G. (2016). Sequence capture and phylogenetic utility of genomic ultraconserved elements obtained from pinned insect specimens. *PLoS ONE*, *11*(8), 1–20. doi:10.1371/journal.pone.0161531

Crawford, N. G., Faircloth, B. C., McCormack, J. E., Brumfield, R. T., Winker, K., & Glenn, T. C. (2012). More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters*, *8*(5), 783–786. doi:10.1098/rsbl.2012.0331

Faircloth, B. C., McCormack, J. E., Crawford, N. G., Harvey, M. G., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements anchor thousands of genetic

markers spanning multiple evolutionary timescales. *Systematic Biology*, *61*(5), 717–726. doi:10.1093/sysbio/sys004

Faircloth, B. C., Sorenson, L., Santini, F., & Alfaro, M. E. (2013). A Phylogenomic Perspective on the Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved Elements (UCEs). *PLoS ONE*, *8*(6). doi:10.1371/journal.pone.0065923

McCormack, J. E., Faircloth, B. C., Crawford, N. G., Gowaty, P. A., Brumfield, R. T., & Glenn, T. C. (2012). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Research*, *22*(4), 746–754. doi:10.1101/gr.125864.111

Smith, B. T., Harvey, M. G., Faircloth, B. C., Glenn, T. C., & Brumfield, R. T. (2014). Target capture and massively parallel sequencing of ultraconserved elements for comparative studies at shallow evolutionary time scales. *Systematic Biology*, *63*(1), 83–95. doi:10.1093/sysbio/syt061

Tagliacollo, V. A., & Lanfear, R. (2018). Estimating improved partitioning schemes for ultraconserved elements. *Molecular Biology and Evolution*, *35*(7), 1798–1811. doi:10.1093/molbev/msy069