Molecular supermatrix construction

Sequence retrieval

- Sequences were downloaded with restez (Bennett et al, 2018)
    - GenBank ()
- Sequence clusters were identified using phylotaR (Bennett et al. 2018)
    - This approach uses BLAST (BLAST ref)
    - phylotaR was run for each of the family taxonomic IDs from NCBI, see table.
- Very lenient sequence selection: minimum 5 species and a minimum MAD score – a measure of sequence length disparity – of 0.4 (Sanderson et al. 2008)
- Species names not conforming to genus species form (i.e. *Genus species*), were dropped.
- In total 33 suitable sequence clusters were identified.

Alignment

- Pasta (https://github.com/smirarab/pasta) (Mirarab et al 2014; Mirarab et al 2015; Balaban et al. 2019)
    - We used the default settings, which makes use of mafft ().
- Alignments were reduced in size and gaps were dropped using TRIMAL (Capella-Gutiérrez et al. 2009)

Supermatrices

- Supermatrices were compiled from the alignments using functions from the R package gaius (Bennett et al. 2019)
- Three supermatrices were constructed each weighting number of species and proportion of gaps in the supermatrix differently.
    - “Low” strictness matrix that dropped no species or columns: 155 species, 82% gaps, 25143 bps and 33 clusters
    - “Mid” strictness matrix that dropped species and columns: 46 species, 58% gaps, 17286 bps and 23 clusters.
    - “High” strictness matrix that dropped species and columns: 11 species, 22% gaps, 7494 bps and 10 clusters.
- The best phylogenetic partitioning scheme and mutation model was determined for each supermatrix using PartitionFinder2 (Lanfear et al 2016) using the Greedy Algorithm (Lanfear et al 2012) and PhyML (Guindon et al 2010). (See the partition finder config file.)
- For the “low” strictness matrix where no columns were dropped, the wobble base pair was considered in the partitioning scheme.
- A guide tree of just the molecular supermatrices was then estimated using a fast RAxML run ().

Details

- The pipeline for running the analysis is reproducible.
- The code was run in R (v3) and all external programs were called using the outsider package (Bennett et al. 2019)
- Code, data and key results are available via GitHub.

References

- Bennett et al. (2018). restez: Create and Query a Local Copy of GenBank in R. Journal of Open Source Software, 3(31), 1102. https://doi.org/10.21105/joss.01102
- Bennett, D., Hettling, H., Silvestro, D., Zizka, A., Bacon, C., Faurby, S., … Antonelli, A. (2018). phylotaR: An Automated Pipeline for Retrieving Orthologous DNA Sequences from GenBank in R. Life, 8(2), 20. DOI:10.3390/life8020020
- BLAST
- Sanderson, M. J., Boss, D., Chen, D., Cranston, K. A., & Wehe, A. (2008). The PhyLoTA Browser: Processing GenBank for molecular phylogenetics research. *Systematic Biology*, **57**(3), 335–346. DOI:10.1080/10635150802158688
- Mirarab S, Nguyen N, Warnow T. PASTA: ultra-large multiple sequence alignment. Sharan R, ed. Res Comput Mol Biol. 2014:177-191.
- Mirarab S, Nguyen N, Guo S, Wang L-S, Kim J, Warnow T. PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. J Comput Biol. 2015;22(5):377-386. doi:10.1089/cmb.2014.0156.
- Balaban, Metin, Niema Moshiri, Uyen Mai, and Siavash Mirarab. “TreeCluster : Clustering Biological Sequences Using Phylogenetic Trees.” BioRxiv, 2019, 591388. doi:10.1101/591388.
- Mafft
- Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics (Oxford, England), 25(15), 1972–1973. doi:10.1093/bioinformatics/btp348
- Bennett gaius
- Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., Calcott, B. (2016) PartitionFinder 2: new methods for selecting partitioned models of evolution formolecular and morphological phylogenetic analyses. Molecular biology and evolution. DOI: dx.doi.org/10.1093/molbev/msw260
- Lanfear, R., Calcott, B., Ho, S. Y., & Guindon, S. (2012). PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular biology and evolution, 29(6), 1695-1701.
- Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology, 59(3), 307-321.
- Bennett et al. (2019). Install and run programs, outside of R, inside of R. Journal of Open Source Software, *In Review*.
- RAxML