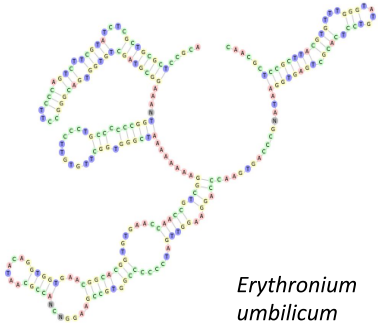
Incorporating Information About ITS2 RNA Secondary Structures Into *Erythronium* Phylogenies

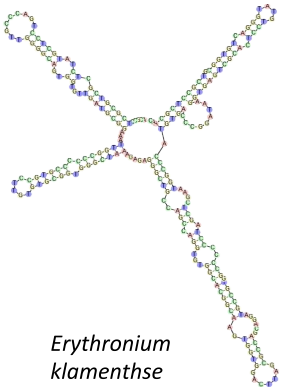
Sol Taylor-Brill

Introduction

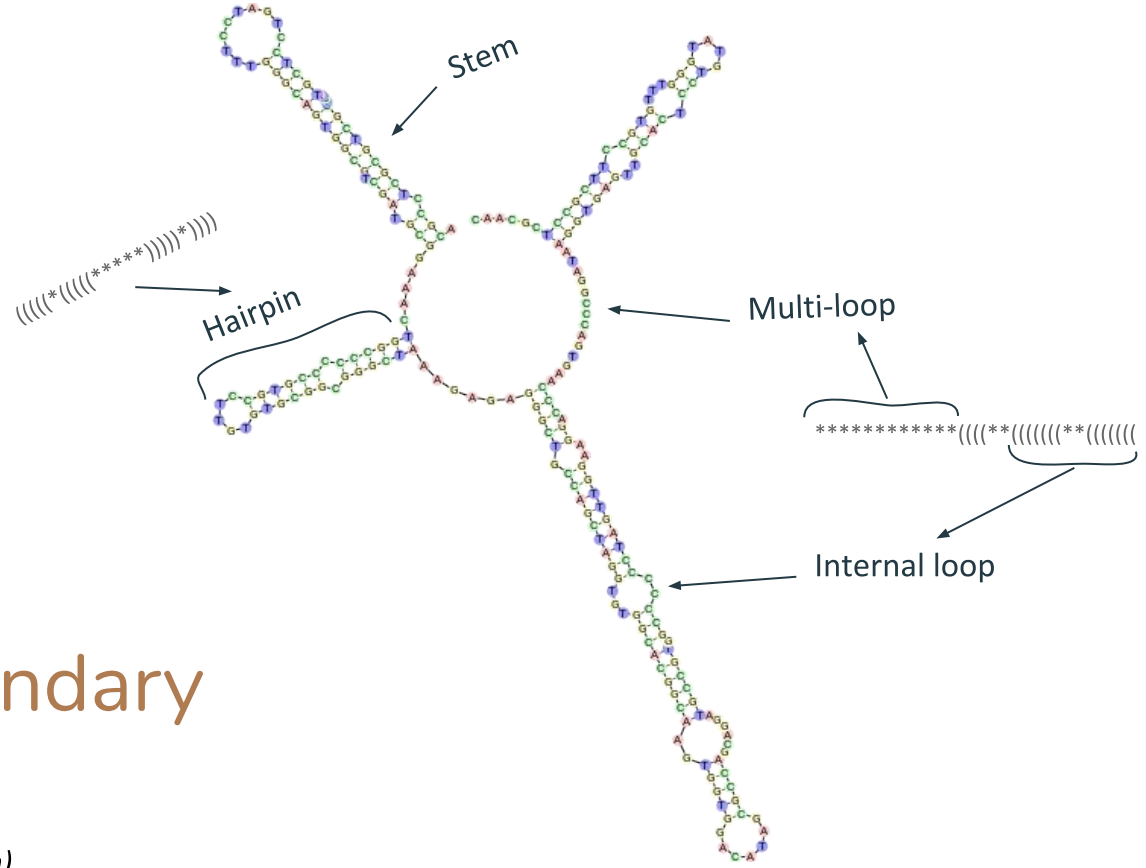
- The Internal Transcribed Spacer Region (ITS 1 and 2) of 18S-26S nuclear ribosomal DNA has long been used to determine phylogenetic relationships due to high copy number, ease of amplification and high level of sequence variation (Baldwin et al., 1995)
- More recent studies have shown that incorporating structural information can make these phylogenies more accurate and robust (Keller et al., 2010)



Erythronium umbilicum



Erythronium klamenthse



RNA Secondary Structures

(Erythronium mesochoreum)

Introduction

- *Erythronium* is a genus in the *Liliaceae* family which contains several species that are endemic to OR and CA.
- A number of possible phylogenies have been proposed for this genus. However, species-level distinctions are still debated. (Clennet et al, 2012)
- Goal of this project was to incorporate structure data into an Erythronium phylogeny for comparison against a phylogeny from the literature (based on sequence + morphology data)
- Additionally, I was inspired by a 2007 thesis by Ginger Jui to investigate whether the free energies of these structures is related to the climate of the plant's regions.

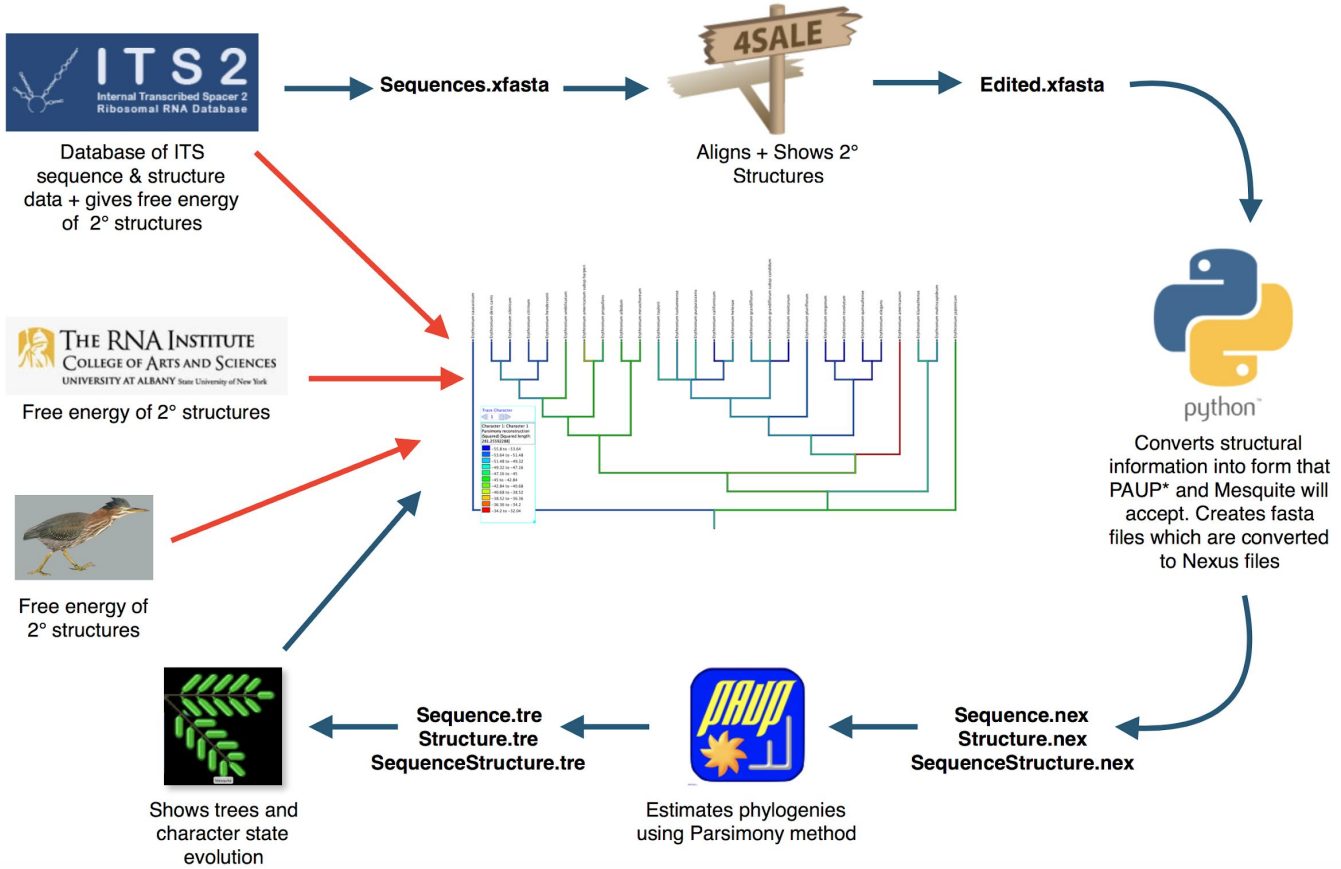
Erythronium dels-canis
Source: https://en.wikipedia.org/wiki/Erythronium

Hypotheses

1. The sequence + structure data will produce a phylogeny that's closer to results in the literature than the sequence or structure data alone.

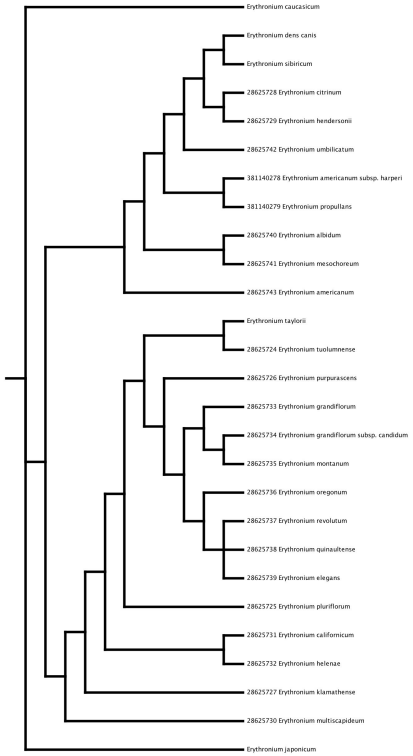
2. All three structure prediction algorithms will produce similar secondary structures for each species' RNA sequence.

3. RNA Secondary structure free energy will correspond to the climate of each species.



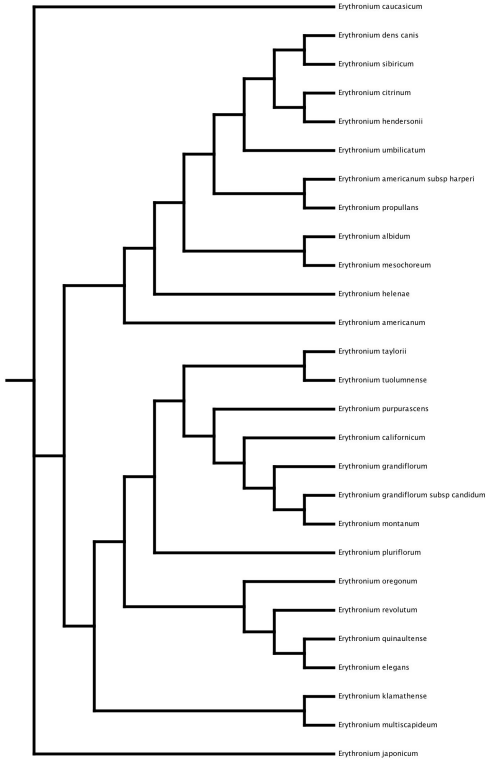
Data and python/R files available at: https://github.com/SolTB/Bio332

Phylogenies



Sequence Only

Treelength = 755, CI = 0.708, RI = 0.817



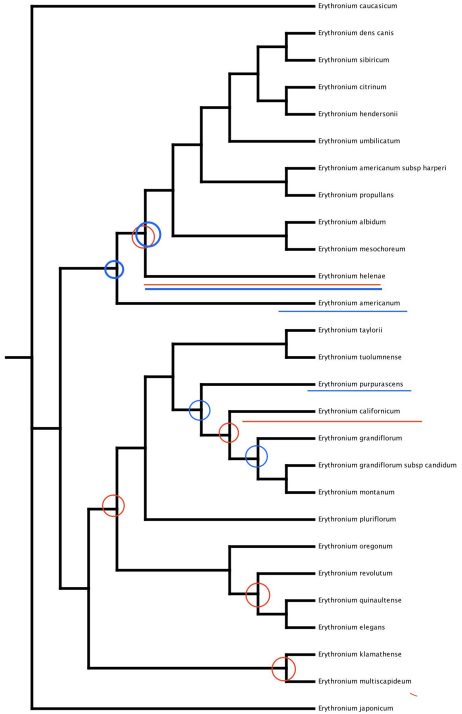
Sequence + Structure

Treelength = 1333, CI = 0.613, RI = 0.769

Structure Only

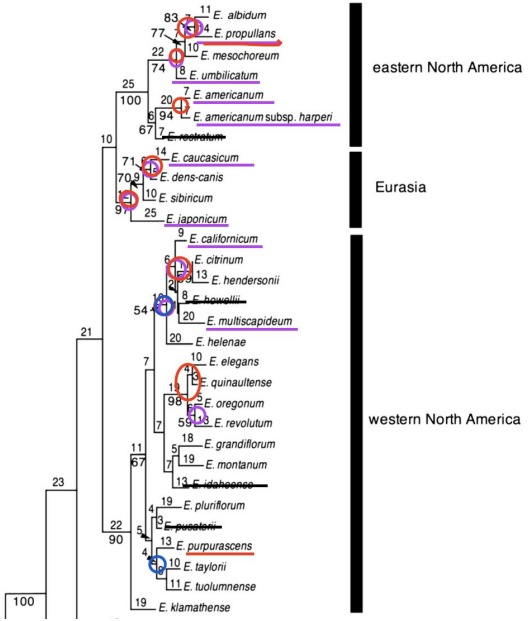
Treelength = 526, CI = 0.536, RI =0.75

Results: Differences Between Trees



- The “Sequence Only” tree minimizes homoplasy (CI = 0.708, RI = 0.817); “structure only” tree maximizes homoplasy (CI = 0.536, RI = 0.75)
- The “Sequence only” and “Sequence + structure” trees differed by:
  - *Californicum* and *helenae* are on one branch in “seq only” and ended up very distant in “S+S”
  - *Revolutum*, *quinaultense*, and *elegans* form a triad in “Seq only” but are differentiated in “S+S”
  - *Klamenthse* and *multiscapideum* branch off from a common ancestor in “S+S” but split off from different more distant ancestors with an extra step in “seq only”
- “Structure only” and “Sequence + structure” trees differed by:
  - *Helenae* branches off of *californicum* in “Structure only” but is very distant from *californicum* in “S+S”
  - *Americanum* branches off beneath *helenae* in “S+S” but branched off early in the bottom part of the phylogeny in “structure only”
  - *Grandiflorum*, *grandiflorum subsp. candidum* and *montanum* are a triad in “structure only”; extra step in “S+S”
  - *Taylorii*, *tuolumnense*, and *purpurascens* are a triad in “structure only”; Extra step.

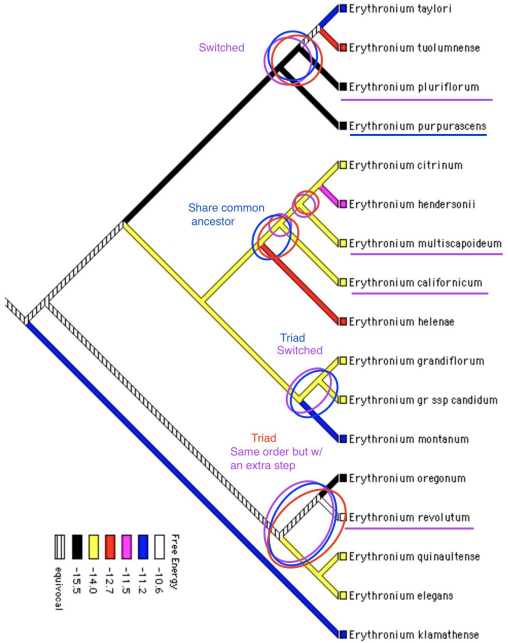
Results: Comparison to the Literature



(Clennett, Chase, Forest, Maurin, & Wilkin, 2012)

- All 3 of my phylogenies differed greatly from the phylogeny constructed from morphological and *ITS*, *matK* and *rps16* sequencing data created by Clennett et al (2012).
- While there were certain differences between how the phylogenies differed from Clennett et al.'s they all did about equally poorly.
- Interesting that the area that all three of my phylogenies had highly conserved was the place that had the most differences from Clennett et al.'s.

Results: Comparison to the Literature

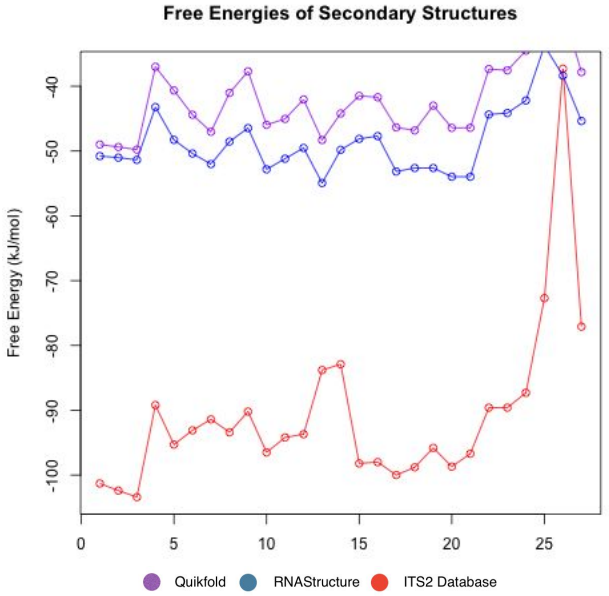


(Jui, 2007)

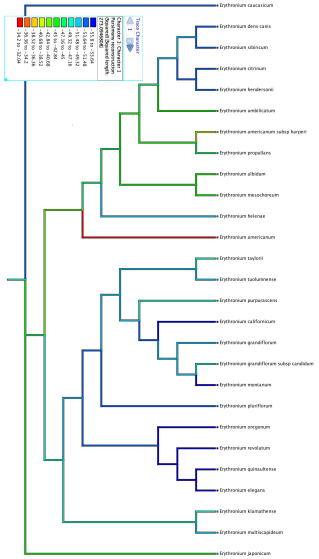
- My phylogenies were much more similar to Jui (2004) who used ITS sequence data and parsimony settings (confirming an earlier phylogeny by Allen et al. (2003)).
- The differences between my phylogenies and Jui's occurred mostly at the taxon level and were small rearrangements
- These differences were generally consistent across all three of my phylogenies.

Evolution of Free Energies

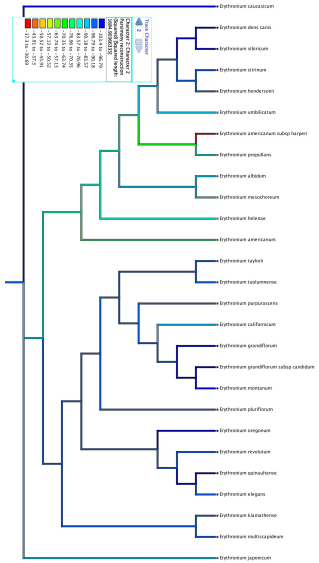
Results: Reliability of Algorithms



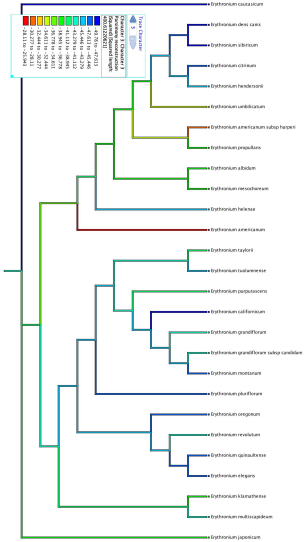
- The free energies given by *RNAstructure*, *Quikfold*, and the *ITS2 database* were significantly different ($t_{R,I;\,34}$ = -15.768, $t_{R,Q;\,51}$ = 4.322, $t_{Q,I;36}$ =17.765, p = <0.001)
- 95% CI for difference in means:
  - *RNAStructure - ITS2*: 36.78 - 47.66
  - *Quickfold - RNAStructure*: 3.44 - 9.40
  - *Quickfold - ITS2*: 43.09 - 54.19
- All three algorithms give significantly different free energies from the same sequence data.
- **However** All three are significantly correlated. Especially *RNAStructure* and and *Quikfold* (r= 0.938, F-stat = 184.9, p <0.001)



RNAStructure



ITS2 Database

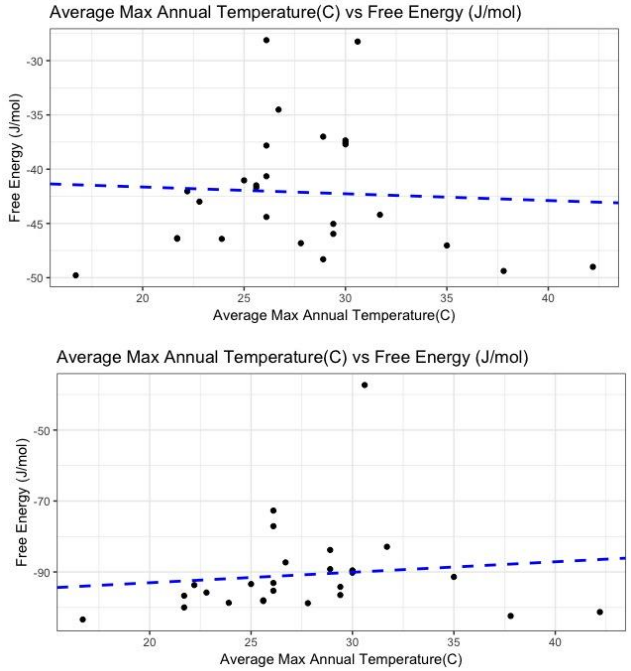


Quickfold

Results

- ITS2 Database resulted in a squared length of 1684.90
  - ITS2 free energies are equal to the calculated free energy of the top/only structure in the database
- RNAStructure Algorithm had a squared length of 273.09
  - RNAStructure free energies are equal to the mean of the top 5 predicted structures (mean sd = 0.57)
- Quickfold resulted in a squared length of 400.61

Results: Climate Data



- Top: *Quikfold* Free Energy Data vs. Average Max Annual Temperature Between 1985-2017 for Native Region.
  - **r = -0.0553**
  - $\hat{Y}$ = -40.3957 + -0.0623x
  - Adjusted R-Squared = -0.0368
  - ***p-value* = 0.7838 (no evidence of correlation)**
- Bottom: *ITS2 Database* Free Energy Data vs. Average Max Annual Temperature Between 1985-2017 for Native Region.
  - **r = +0.1184**
  - $\hat{Y}$ = -98.9438 + 0.2953x
  - Adjusted R-Squared = -0.0254
  - ***p-value* = 0.5563 (no evidence of correlation)**

Conclusion

- As compared to literature phylogenies (Jui, 2004 & Clennett et al., 2012), neither the “sequence only” nor the “sequence + structure” phylogeny were conclusively more accurate.
- The three algorithms used to predict RNA secondary structure and free energy produced significantly different results, implying that these programs are not completely accurate and that secondary structure information needs to be confirmed experimentally.
- There was no correlation between max temperature of native climate and secondary structure Free Energy
    - Jui (2007) found the same thing.
    - However, this is complicated by the wide range of temperatures across regions.

Future Directions

- Is there a better way to encode RNA structural data?
  - Counting hairpins seems overly simplistic and doesn't capture all the information
  - Coding nucleotide-by-nucleotide contains a lot of information but loses track of larger structural components which affect function
  - Perhaps some combination of these two methods or a morphometric method would be more effective at capturing this information.
  - Including CBC (Compensatory Base Change) information might help.
- Standardizing algorithms and confirming structures experimentally
- A pipeline that reduces the high amount of manual labor needed to perform these analyses.

References

Adebowale, A., Lamb, J., Nicholas, A., & Naidoo, Y. (2016). ITS2 secondary structure for species circumscription: case study in southern African Strychnos L. (Loganiaceae). *Genetica*, *144*(6), 639–650. https://doi.org/10.1007/s10709-016-9931-0

Arnaoudova, E., Haws, D., Huggins, P., Jaromczyk, J. W., Moore, N., Schardl, C., & Yoshida, R. (2010). Statistical Phylogenetic Tree Analysis Using Differences of Means. *ArXiv:1004.2101 [q-Bio]*. Retrieved from http://arxiv.org/abs/1004.2101

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., & Donoghue, M. J. (1995). The its Region of Nuclear Ribosomal DNA: A Valuable Source of Evidence on Angiosperm Phylogeny. *Annals of the Missouri Botanical Garden*, *82*(2), 247–277. https://doi.org/10.2307/2399880

Clennett, J. C. B., Chase, M. W., Forest, F., Maurin, O., & Wilkin, P. (2012). Phylogenetic systematics of Erythronium (Liliaceae): morphological and molecular analyses. *Botanical Journal of the Linnean Society*, *170*(4), 504–528. https://doi.org/10.1111/j.1095-8339.2012.01302.x

References

Edger, P. P., Tang, M., Bird, K. A., Mayfield, D. R., Conant, G., Mummenhoff, K., … Pires, J. C. (2014). Secondary Structure Analyses of the Nuclear rRNA Internal Transcribed Spacers and Assessment of Its Phylogenetic Utility across the Brassicaceae (Mustards). *PLOS ONE*, *9*(7), e101341. https://doi.org/10.1371/journal.pone.0101341

Gulko, B., & Haussler, D. (n.d.). *Using Multiple Alignments and Phylogenetic Trees to Detect RNA Secondary Structure*. 18.

ITS so much more | Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/j.tig.2015.02.005

Jui, G. C.-C. (2007). *Comparative Phylogenetic Analysis of 5.8S rRNA Hairpin Variation*. 77.

Keller, A., Förster, F., Müller, T., Dandekar, T., Schultz, J., & Wolf, M. (2010). Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. *Biology Direct*, *5*, 4. https://doi.org/10.1186/1745-6150-5-4

Maddison, W. P. (1991). Squared-Change Parsimony Reconstructions of Ancestral States for Continuous-Valued Characters on a Phylogenetic Tree. *Systematic Biology*, *40*(3), 304–314. https://doi.org/10.1093/sysbio/40.3.304

References

Pacific Bulb Society | Erythronium Three. (n.d.). Retrieved April 30, 2019, from https://www.pacificbulbsociety.org/pbswiki/index.php/ErythroniumThree

Plants Profile for Erythronium mesochoreum (midland fawnlily). (n.d.). Retrieved April 30, 2019, from https://plants.usda.gov/core/profile?symbol=ERME15

PRISM Climate Group, Oregon State U. (n.d.). Retrieved April 30, 2019, from http://www.prism.oregonstate.edu/explorer/

Särkinen, T., Bohs, L., Olmstead, R. G., & Knapp, S. (2013). A phylogenetic framework for evolutionary study of the nightshades (Solanaceae): a dated 1000-tip tree. *BMC Evolutionary Biology*, *13*(1), 214. https://doi.org/10.1186/1471-2148-13-214

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., … Consortium, F. B. (2012). Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, *109*(16), 6241–6246. https://doi.org/10.1073/pnas.1117018109

References

Seibel, P. N., Müller, T., Dandekar, T., Schultz, J., & Wolf, M. (2006). [No title found]. *BMC Bioinformatics*, *7*(1), 498. https://doi.org/10.1186/1471-2105-7-498

Seibel, P. N., Müller, T., Dandekar, T., & Wolf, M. (2008). Synchronous visual analysis and editing of RNA sequence and secondary structure alignments using 4SALE. *BMC Research Notes*, *1*(1), 91. https://doi.org/10.1186/1756-0500-1-91

Shaw, J., Lickey, E. B., Beck, J. T., Farmer, S. B., Liu, W., Miller, J., … Small, R. L. (2005). The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis. *American Journal of Botany*, *92*(1), 142–166. https://doi.org/10.3732/ajb.92.1.142

Shevock, J. R., Bartel, J. A., & Allen, G. A. (1990). DISTRIBUTION, ECOLOGY, AND TAXONOMY OF ERYTHRONIUM (LILIACEAE) IN THE SIERRA NEVADA OF CALIFORNIA. *Madroño*, *37*(4), 261–273.

Sjölander, K. (2010). Getting Started in Structural Phylogenomics. *PLOS Computational Biology*, *6*(1), e1000621. https://doi.org/10.1371/journal.pcbi.1000621

The ITS2 Database. (n.d.). Retrieved April 15, 2019, from http://its2.bioapps.biozentrum.uni-wuerzburg.de/