Reading comprehension questions for week 4 (viruses)

Reading Group Information for Fall 2021

Last updated: 2021-11-16 09:39:14 -05:00

Questions for Group A – (and letters)

1. Outline the methods. (Look at website for information on alignment, etc: https://www.fluxus-engineering.com/index.htm)
2. What is a median joining network (MJN)?
   - What is the input and output?
   - What do the leaves represent and how are they labeled? What does their size (diameter) mean?
   - What do the internal nodes in the output represent? Are they labeled and if so how? What does their size mean?
   - What do the edges in the output represent? Are they directed or undirected? Are they labeled / annotated with any information?
   - What is a "reticulation" in an MJN?
   - Give a 1–2 sentence description of these methods (see letter by Chookajorn but you will likely need to look at abstracts for the cited papers).
3. Does an MJN represents an evolutionary history? Does it assume a particular model of evolution? What assumptions does it make about evolution? Do you think these assumptions are realistic for virus genomes?
4. Given what you know about MJN, how would you interpret Figure 1 (what do the node colors mean)? What do the authors say about the evolution of these three groups? Does this figure support what they say? More broadly, does their analysis support what they say?
5. The figure 1 caption notes that the "reticulations are caused by recurrent mutations at np11083"? What does this mean?
6. How could recombination and lateral/horizontal gene transfer impact phylogenetic inference or the MJN?
7. What is a founder effect and how does it relate to MJN?
8. What does it mean to use an outgroup to polarize character transformations?
9. How would you redesign this study?
10. How did this paper get published in PNAS? Also note the date of this study was March 30, 2020 (sent for review March 17, 2020).

Questions for Group B – Accuracy in Near-Perfect Virus Phylogenies

1. What is a perfect phylogeny? What is a near perfect phylogeny (from the probabilistic perspective)?
2. When will the optimal solution to maximum likelihood, maximum parsimony, and maximum compatibility problems be the same?
3. Why might virus genomes sampled/assembled during outbreaks be modeled with a near-perfect phylogeny?
4. Using Figure 4, which characters (1, 2, 3, 4) evolve perfectly in the true tree and which do not? Why does the estimated tree place A and B as siblings? Why does the estimated tree have a polytomy? Calculate the FNR and FPR for these trees.
5. Using Figure 1, what would happen if a fifth perfect character changed state on one of two the dotted edges. Draw the resulting maximum parsimony or maximum compatibility tree.
6. Give an example similar to Figure 1 for a balanced tree. How does tree shape impact the probability of a false positives and why?
7. The authors provide some theory regarding the expected false positive rate in terms of tree shape, the tree length per site, the number of leaves (n), and the number of characters (m). What are some of the (simplifying) assumptions they make?
8. Outline the simulation study for producing Figure 2A and 2B.
9. Discuss the differences between Figure 2A and 2B at when the tree length per site is less than 1 and greater than 1. How do you explain these trends given the differences between the simulation conditions?
10. Based on the section "extension to ML inference," would you use MP or ML to estimate a virus phylogeny? Do you recommend any analyses to determine which method to select?
11. How do you calculate bootstrap support? Why could bootstrap support be problematic when $m \ll n$? How do the authors suggest support should be estimated instead?
12. What are the implications of this study for recent analyses of SARS-CoV-2 genomes?

Questions for Group C – Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic

1. Describe recombination e.g. What is the result? When does it occur?
2. What is a recombination break point? What is a break point free region (BFR)?
3. Why do the authors decide to identify BFRs prior to their phylogenetic analyses? What would happen if they instead estimated one tree for the entire genome under the GTR+GAMMA model (e.g. using RAxML or IQTree)?
4. What is phylogenetic incongruence (PI)? Why is it a signal of recombination?
5. Why did the authors use multiple methods to detect recombination break points? List the methods used to detect break points. If you have time, give a one sentence description of each.
6. Does recombination occur uniformly across the genome? Do recombination break points correspond to functional genes?
7. Give an example where homoplasy can be explained via recombination.
8. Do you agree or disagree with the following methodological choice "BFRs were concatenated if no phylogenetic incongruence signal could be identified between them."
9. Explain figure 3. How can both evolutionary trees be correct?
10. How would you calculate root to tip divergence in figure 4? How do the authors explain differences between Figure 4a and Figure 4b.
11. In Figure 5, why is the uncertainty greater for the timing of deeper nodes (i.e. nodes further back in time) in the tree?
12. Even if we account for recombination, what are other sources of model violation?
13. Why were geographic analyses important to the authors when considering recombination?
14. Did the authors conclude that transmission from bats to pangolins was important for SARS-CoV-2 evolution (i.e. adaption so that it could be transmitted to humans)? Why or why not?

Questions for Group D – Phylogenetic supertree reveals detailed evolution of SARS-CoV-2

1. What do the authors say is their motivation for using a supertree methods?
2. Outline their data analyses steps.
3. What is the maximum size of the MRP matrix? How come the MRP matrix is smaller?
4. Give a list of the supertree methods used. If you have time, give a 1–2 sentence description of each method. Note that the first supertree approach described should have been called MRL analysis (https://doi.org/10.1186/1748-7188-7-3).
5. What does a bootstrap support value of 55 mean for the gene trees in Figure 3? What about the supertrees? Do you think this is "high" support? What issues could make computing bootstrap support challenging?
6. Compare and contrast MRL versus MRP supertree methods as well as the consensus tree approach. Would you have predicted MRP to be the best method?
7. Looking at Figure 3, list some differences between the branches of different trees. Consider which trees place RaTG13 (BatMN996532) as sibling of SARS-CoV-2 and which don't. What processes could explain the differences between the phylogenetic trees estimated on each gene?
8. Why is the ORF1ab gene important?
9. What is recombination and why could it impact phylogeny estimation? Do you agree with the authors statement that "this problem could be avoided by constructing a supertree based on protein sequence."
10. The authors say that there analyses "the MRP pseudo-sequence supertree analysis firmly disputes bat coronavirus RaTG13 be the last common ancestor of SARS-CoV-2." Do you think their analyses are "firmly dispute" this finding?
11. Given their motivation for using a supertree method, do you agree with their choice of methods? Do you have any questions or comments about their analyses? What changes would you make to their analyses if any?

Presentation Order

- A
- B
- C
- D