Reading comprehension + discussion questions for week 5 (presenter's choice)

Reading Group Information for Fall 2021

Last updated: 2021-11-28 23:18:50 -05:00

Questions for Group A – Integrative inference of subclonal tumour evolution from single-cell and bulk sequencing data

1. Define variant allele frequency (VAF)?
2. What is a clonal tree? What is a cell lineage tree? What is a mutation tree? Specifically, define what the leaves, internal nodes, and edges represent. (hint: see figures and also the section called "Tree models of tumor evolution").
3. What do the above definitions tell us about the model assumptions for tumor evolution?
4. What types of evolutionary histories (or evolutionary information) are bulk sequencing data useful for reconstructing? What are some of the challenges that arise when building a clone tree from bulk sequencing data?
5. Repeat #4 for single cell sequencing data.
6. Give a high level description of the B-SCITE method.
7. What is the relationship between a tumor phylogeny (shown in Figure 1c) and a mutation tree (shown in Figure 1a)? How do you transform a clonal tree into a mutation tree? Given a mutation tree, give an algorithm to determine all of the "allowed" clones?
8. What is a doublet? Given an example how a doublet could mislead phylogenetic analyses (in the simplified case where you seek a perfect phylogeny to explain your error-free single cell sequencing data)?
9. Repeat #8 for other common issues like sequencing error and CNAs.
10. How were the methods compared given that B-SCITE and SCITE return a mutation tree, OncoNEM returns return clonal trees, and ddClone returns mutational clusters?
11. Define V-measure and co-clustering accuracy.
12. Based on the simulation study, under which conditions would you expect B-SCITE to have an advantage? Would you expect this based on your high level understanding of the methods? Why or why not?
13. Overall, are you satisfied with the method evaluation or what other evaluation metrics would you be interested in?

Questions for Group B – Rapid Neighbour-Joining

1. Give a brief overview of the traditional NJ algorithm.
2. How does QuickTree attempt to improve upon NJ?
3. How do QuickJoin and RapidNJ attempt to improve upon NJ?
4. What is the average running time of QuickJoin? What is the worst-case running time of QuickJoin? Why does QuickJoin fail to outperform QuickTree (according to the authors)? Do you agree with this assessment based on the data (Figures 1–5)?
5. Describe the RapidNJ algorithm. Consider working out an example to illustrate the use of the I and S data structures S and I, as well as how they get updated. A good case study could be when RapidNJ is given a matrix that is additive for tree $((((1,2),3),4),5)$; with all branches having unit length.
6. How does the variable q_{min} impact the running time of RapidNJ? (note: the similarities to branch-and-bound approaches)
7. Discuss trade-offs in how data structures are updated from the theoretical and practical perspective.
8. Based on the experimental study, what is the next best method compared to RapidNJ? What is it's worst-case running time?
9. Why do you think the results look better behaved on the simulated data (Figure 5) compared to the biological data (e.g. Figure 3)?
10. What other experiments or results would you want to see to evaluate the utility of RapidNJ compared to other methods?
11. What do you think will be a more significant challenge in practice – the running time or the storage?
12. In high performance computing, it is common to profile code to determine which functions the code spends most of its time (so that those functions can be targeted for optimization). Now consider that NJ is just one function used in a phylogenetics pipeline (e.g. it is used after MSA estimation or even after gene tree estimation when used in the context of species tree estimation). Do you think that NJ-like methods will be the bottleneck in such pipelines? Why or why not?

Questions for Group C – Alignment- and reference-free phylogenomics with colored de Bruijn graphs

1. List some of the challenges to traditional phylogeny estimation pipelines (i.e. MSA + ML), noted by the authors?
2. List some of the challenges to alignment-free pipelines, noted by the authors?
3. Define a colored de Bruijn graphs. What do the vertices and edges represented? How is the graph colored?
4. Give an overview of the authors approach. How are (ordered) splits defined from a colored de Bruijn graph and weighted? How does this relate to evolutionary events like SNVs and indels? How are these ordered splits assembled into a tree? or split-network?
5. Define a split-network. What do the vertices and edges represent? What does it mean when there are cycles in the graph? To what extent does a split network represent an evolutionary history? Do you find them helpful to look at?
6. How does the authors' approach, SANS, compare to other alignment-free methods in terms of accuracy, running time, and storage?
7. For each of the biological data sets, how accurately does SANS reconstruct reference phylogenies? (note: either give the RF metric or give a brief description e.g. "recovers major clades")
8. How would you expect this method to perform if data evolve under an infinite sites model, down the reference phylogeny? What about if data evolve under the GTR model? What if different regions of the genome have different evolutionary histories due to population-level processes?
9. What are the biggest strengths/weaknesses of this approach? Overall, do you think this is a promising? Why or why not?

Questions for Group D – SAQ: semi-algebraic quartet reconstruction method

1. What is the input and output of SAQ?
2. A GM model tree (T, Θ) defines a probability distribution p on site patterns. How many possible site patterns are there if the sequences are on alphabet space $\{A, C, G, T\}$? Recall:
   - T denotes the rooted tree topology, T^u denotes the unrooted version of T, and Θ denotes the numeric parameters. What are the numeric parameters associated with the GM model? (described in main text and appendix). How does this differ from the number of parameters in the GTR model? Or the Jukes-Cantor model?
   - We can think of sites evolving *i.i.d.* down the model tree.
   - Alternatively, we can think of each site pattern as being generated by a roll of a die defined by p.
   - In practice, p is estimated from a multiple sequence alignment (MSA).
3. How is the flattening matrix computed? How does $\tilde{p}$ relate to F_{T^u}? (recall that T^u is typically denoted as a quartet, for example 12|34).
4. Ignoring the "12—34 leaf-transformations" for now, describe the steps in the SAQ algorithm and why they work (note: this is essentially the Erik+2 method).
5. Why are the "12—34 leaf-transformations" important at the high level? (if time permits at the end, go back and discuss this part of the paper)
6. Based on Table 1, rank the methods from best to worst for the different model conditions. How does method performance change based on the model conditions?
7. How are data generated under the mixture model at the high level? How does the length of the internal branch in the model tree impact absolute and relative method performance for mixture data?
8. What methods evaluated are statistically consistent under the (hom)GTR model? under the GM model? under the mixture model?
9. How well does theory (consistency) and method performance relate?
10. Why do others sometimes use MP or NJ (with log-det also called paralinear distances)? Was NJ displayed for mixture models? Was MP displayed for non-mixture models?