Decision on "Cancer phylogenetic tree inference at scale from 1000s of single cell genomes" by Solehi et al submitted to PCI Mathematical & Computational Biology

Amaury Lambert (recommender)

August 25, 2022

Brief foreword. Dear Sohrab Salehi,

I wish to apologize for the long delay you've been waiting to get feedback on your manuscript.

The report of Reviewer 1 arrived in January. Reviewer 2 then had us wait for a long time until s/he finally wrote a poorly informative report, as you will see. We asked her/him to make a more thorough review, which s/he accepted to do. We then waited for two more months until s/he finally said s/he wouldn't do it.

As a consequence, we had to find a third reviewer in emergency, who fortunately sent us her/his report early August.

I am really sorry for this unusually bad experience, and sincerely hope this delay does not put you in a too awkward situation for your publication schedule.

Recommender's summary. Let me first give my own, self-contained summary of the manuscript.

This paper presents and applies a new Bayesian inference method of phylogenetic reconstruction for multiple sequence alignments in the case of low sequencing coverage but diverse copy number aberrations (CNA), with applications to single cell sequencing of tumors. The idea is to take advantage of CNA to reconstruct the topology of the phylogenetic tree of sequenced cells in a first step (the 'sitka' method), and in a second step to assign single nucleotide variants (SNV) to tree edges (and then calibrate their lengths) (the 'sitka-snv' method).

The data are assumed to be in the form of an integer-valued $C \times L$ matrix A, where C is the number of cells, L is the number of loci (here, loci are segments of prescribed length called 'bins'), whose entry a_{ij} at row i and column j is the (preprocessed) number of copies, called 'copy number state' (CNS), of locus j in cell i.

These data are then summarized/simplified into a binary-valued $C \times L$ matrix Y, whose entry at row i and column j is $y_{ij} = 0$ if in cell i, the CNS at locus j and at locus $j + 1$ are equal (i.e., $a_{ij} = a_{i,j+1}$), and $y_{ij} = 1$ otherwise.

Biologically speaking, when $y_{ij} = 1$, in the ancestral lineage of cell i, at least one genomic rearrangement has occurred, and more specifically the gain or loss of a segment with at least one endpoint in locus j or in locus $j + 1$; this event is viewed as a 'mutation at marker j', where marker j is the point where loci j and $j + 1$ touch. The authors expect the infinite-allele assumption to approximately hold (i.e., that at most one mutation occurs at any given marker and that 0 is the ancestral state). They refer to this assumption as the 'perfect phylogeny assumption'. By only recording from CNA events the endpoints at which they occur, the authors lose the information on copy number state (and also forget the dependencies between these endpoints), but they gain the assumption of independence of the mutational processes occurring at different sites, which approximately holds for CNA endpoints but certainly not for CNS.

The goal of sitka is to produce a posterior distribution on phylogenetic trees conditional on the matrix Y, where here a phylogenetic tree is understood as containing the information on 1) the topology of the tree but not its edge lengths, and 2) for each edge, the identity of markers having undergone a mutation, in the sense of the previous paragraph.

For any given phylogenetic tree t (in the previous sense), we can define $x_{ij}(t) = 0$ or 1 according to whether cell i carries a mutation at marker j (i.e., descends from an edge carrying a mutation at j) or not, based on tree t. The posterior of a tree t is a measure of agreement of the matrices Y and $X(t)$. More specifically, it is the probability of Y under the assumption that conditional on t, the variables $y_{ij}(t)$ are independent and the law of $y_{ij}(t)$ conditional on $x_{ij}(t) = \varepsilon$ is Bernoulli with parameter p_ε, where $p_0 = r^{FP}$ is called a 'rate of false positive' and $1 - p_1 = r^{FN}$ is called a 'rate of false negative'.

The results of the method are tested against synthetic datasets simulated under various assumptions, including conditions violating the perfect phylogeny assumption and compared to results obtained under other baseline methods. The method is extended to assign SNV to edges of the tree inferred by sitka. It is also applied to real datasets of single cell genomes of tumors.

Main comments. I concur with the comments of Reviewers 1 and 3, in particular:

- It would be good to improve the structure of the paper and expand some bits in order to make it readable by a wider audience (cf. comment of Reviewer 3). For example, the authors might like to expand the Introduction in order to have the reader better understand the context (low sequencing coverage but additional information coming from CNA), the specificities of the method, its main aspects and its applications (subclonal structure?), similarly as in my personal summary.

  There are also several concepts and tools that should be defined more accurately (perfect phylogeny, overlapping/non-overlapping CNA, main principles of methods like UPGMA and the like, doublet/mouse cell/cycling cell, delta method, Sackin/Colless/Yule...).

  There are also Supplemental figures that could be included in the main text (e.g., Supplemental Figures 1 to 4).

- Following up on the previous point, a wealth of existing methods are cited in Introduction but their relation to sitka, as well as the difference/similarity with the benchmarked methods, should be better explained (maybe proposing a rough classification). Some other methods should be benchmarked, as mentioned by Reviewer 1.

  The claim that sitka relaxes "the independence assumptions required by existing phylogenetic methods" is not sufficiently well explained. Indeed, the method does not assume that copy number states evolve independently at different sites, but it assumes that the endpoints of CNA events occur independently, which may approximately hold for say the left endpoint, but not when combining both endpoints (see comment of Reviewer 3). Worse than that, it assumes independence of false positive/negative processes between lineages of different cells. It would be good to emphasize these aspects, to discuss the advantages and shortcomings of these assumptions (in Discussion) and also, as asked by Reviewer 3, to test violation of within-site independence, for example by assuming in synthetic experiments that sizes of CNA events are not exponentially distributed but e.g., always have the same fixed value or follow a heavy-tailed distribution.

Could you maybe quantify the trade-off (mentioned line 58) between scalability/computational time and estimation accuracy ?

- Following up on the previous point (again), don't you think it would be more natural to model violation of 'perfect phylogeny assumption' by modeling directly the biological CNA process (gains and losses) as you do it verbally at the bottom of page 3 and in Supp Fig 3? In particular, I don't understand the IS violation procedure applied to the processed data (merging two columns): how do you do the merging and why does it mimic homoplasy?

  More generally, can you argue why you apply a lossy transformation to the data before analyzing it? At first sight, it looks like you lose a lot of information by replacing CNS by a binary variable telling whether contiguous bins have different CNS or not. In addition, the method assumes that the ancestral state of this binary variable is 0, which of course does not always hold in reality. Why can't you encode the data by the *difference between CNS at two contiguous bins*, so you don't lose the pseudo-independence of marker evolution at different marker sites but can keep the information on CNS and compute tree likelihood under the model of CNA evolution used in the simulations? Please discuss this.

Minor comments.

- General comment – isn't there sometimes a double meaning of the word 'locus'? (used both to denote a bin and a marker)
- line 49 – The fact that likelihood-based methods perform statistically better than e.g., distance-based ones, should be supported by examples or references.
- line 104 – Please add that this procedure consists in passing from a type I to a type II tree. Anyway, I'm not sure you really need to explain the reader the difference between type I and type II trees.
- line 130 – What do you mean by "sitka's performance *degrades gracefully* in the face of some of the key types of expected violation of the perfect phylogeny assumption"?
- line 222 – The Discussion section seriously needs to be fleshed out.
- line 322 – Remove "disjoint"
- line 339 – Did you test the robustness of the method related to the upper bound of the support of the prior of false positive/negative rates?

- line 359 – Please give a reference for the “ ’rich gets richer’ behaviour built-in into the prior, which is viewed as useful in many Bayesian non-parametric models” (see also comment by Reviewer 1)?

- line 391 – I like your definition of a Gibbs sampling algorithm (“an MCMC move with no rejection step”), but I am not sure it is very academic.

- line 429, eq (6) – Shouldn’t you have $\pi(t', d\theta)$ rather than $\pi(t, d\theta)$? Maybe give a name to this argmin for future reference (see comment about lines 472–474).

- line 452 – Here and at some other places, you normalize scores by the score of the worst performing method. It seems weird because in the presence of a very poor-performing method for a given dataset, this will tend to overrate all alternative methods.

- line 470 – Isn’t the “best possible tree” just the true tree?

- lines 472–74 – What is the difference between the “greedy estimator (GE) of Section 9.4.5” and the “trace search estimator (TSE) defined as a tree in the sampler trace that minimizes the sample L1 distance (Section 9.4.5)” ? After the latter definition, it seems to me that the TSE is given by Eq (6). Please give a mathematical formula for the estimator which is not defined by Eq (6) and specify which is which.

- lines 478–484 and lines 496–498 – Please specify what is measured (RF distances? Normalized? Confidence intervals?).

- line 478 – Can you explain in Discussion why the global model can outperform the local model?

- lines 509 – Add “of size s” (to “An island...”)

- line 512 – Is there a reason why the violation rate thus estimated has anything in common with the violation rate defined in the synthetic experiments?

- line 552 – From my personal experience, the range of β for which Beta-splitting trees are interesting and realistic is $(-2, 0)$ rather than $(-1, 10)$.

- line 566 – Why not follow the same procedure as previously? (Beta-splitting trees and, as in my last main comment, simulation of the biological CNA process)

- line 611–613 – Please give the mathematical formula defining $g_{\cdot,j}$.

- lines 622–624 – Please be more precise and elaborate notation to let the matrix o explicitly depend on its arguments (h, w, z?).

- line 669- – ‘loci’ should be ‘locus’ (twice).

- Supp Fig 2 – Please specify that the red nodes in (a) correspond from top to bottom to markers 2, 3, 1 in this order. Also if you feel it is important for the reader to understand the difference between type I and type II trees, it might be good to display a type I tree with more interior marker nodes on the same edge.

- Supp Fig 3 – “By the infinite site argument” is confusing (assumption vs approximation?)

- Supp Fig 10 – Please specify that only Sackin and Colless indices are normalized and have positive values indicating more imbalance (it is the contrary for β). Do you have a sense why Sackin and Colless indices always give very similar values and why β is consistently estimated by ≈ -1?