Jan Mielniczuk
Institute of Computer Science, PAS
& Faculty of Mathematics and Information Sciences, WUT

26.II.2024

Review of Ph.D. thesis by Senbai Kang, M.Sc., entitled
Probabilistic graphical models for inferring tumor phylogeny and genomic variants from single cell DNA genomic data

The following is a report on Ph.D. thesis by Senbai Kang, M.Sc., prepared at University of Warsaw and supervised by prof. Ewa Szczurek. The report has been made at the request of University of Warsaw.

Thematic content of the thesis. Advent of single-cell DNA sequencing methods provides new opportunities of studying occurrence and dynamics of genes' mutations. This is specially relevant in cancer therapies which are hindered by common occurrence of intra-tumor heterogeneity. The discussed work is a contribution to this area, in which modelling of mutations present at considered candidate loci is investigated in conjunction with modelling their history described by phylogenetic tree.

The main challenge here is to distinguish between mutations and errors due to noisy data. The model should take into account several factors, including over/under-representation of genes at certain regions due to whole-genome amplification, readout errors (ADO events) and deletion of genes. Moreover, in view of the fact that in diploid case at a given locus both genes are prone to mutation, it is worthwhile to distinguish between heterozygous and homozygous double mutations. The present work addresses these issues.

The first model, SIEVE, is a joint probabilistic graphical model of phylogenetic tree and read counts of nucleotides at candidate locations which takes into account possible inexactitudes mentioned above: non-uniform representation of all genomic regions in single-cell whole-genome amplification and ADO events. Moreover, considered phylogenetic tree allows for multiple mutations of the considered nucleotide. Its extension, DelSIEVE, incorporates deletions into the probabilistic graphical model. The main novel ingredients of the proposed models is a specific topology of phylogenetic tree and allowing for 12 types of genomic transitions at each considered locus due to distinguishing between two alternative

nucleotides in the case of SIEVE, what results in 9 types of observable mutation events. In the case of DelSIEVE, which allows additionally for deletion of genes, the model considers 17 types of observable mutation events stemming from 28 possible transitions.The Bayesian graphical model is fitted by assuming certain prior distributions of its parameters and using Monte Carlo Markov Chain (MCMC) methodology to obtain the solution.
The models and their performance are described in two multi-author papers (one published in Genome Biology (2022) and the second deposited on bioRxiv in 2023) for which S. Kang is the first author.

Evaluation of the thesis It is obvious for me that the thesis which concerns inference from single cell DNA genomic data, addresses very important subject in bioinfromatics. I will focus on modelling and data-analytic aspects of the study as they fall into the realm of my expertise.
Ihe presented models are based on plausible assumptions and use commonly accepted distributions as priors for count data models. They certainly contribute to more realistic modelling of noisy genomic data. DelSIEVE is an extension of SIEVE which takes into account possible deletions as well as double ADO events and thus improves on SIEVE. It is the only method available which allows for double deletions. The strength of both models is shown in action for synthetic data by comparison with competing approaches (CellPhy, SiFit, Monovar, SciΦ in the case of SIEVE and Monovar, SCIPHIN and SIEVE in the case of DelSIEVE). The simulations studies are carefully designed, incorporating different scenarios (quality of data, coverage quality and changing Copy Number Abberations). The results show that SIEVE outperforms all other methods but SciΦ in terms of standard metrics of variant calling (recall, precision, F1) and phylogeny tree recovery (branch score BS and normalised Robinson-Foulds distance RF). It also systematically outperforms SciΦ, but the advantage is small and likely not confirmed by statistical tests (which are missing) apart from case of high mutation rate (Figure 2.2). Analysis of real data sets using SIEVE shows in particular for TNBC16 data set several cases of heterozygous mutations and several violations of Finite Site Assumption (FSA). DelSIEVE additionally reveals here multiple somatic deletions.
It is worth underlining that DelSIEVE derives from careful analysis of weaknesses of SIEVE which is prone to explain somatic deletions as result of ADO events and as the result

inflates the amount of detected mutations.

The thesis is written in a very good English, and its editing is practically spotless - I have found only one typo (p. 60) and one misleading reference[1] which for the work of this size (150 pages) is quite an achievement. At some places, the Author is too laconic in explaining his approach and its rationale (see e.g. comments 1 and 2 below). Sometimes important piece of information is missing (e.g. prior distribution of M in (2.20)). Overall, however, I consider two presented models, SIEVE and DelSIEVE, as worthwhile and original contributions to modelling of mutations for genomic data.

That being said, there are some model building and data-analytic aspects relevant to the thesis' content, which in my view should be discussed, and in some cases investigated, more thoroughly.

1. The effort to build a model which realistically reflects data sampling problems in single cell DNA-sequencing necessarily results in a complicated construct with many parameters. The (natural) question which should be asked in such a case is whether the obtained graphical model is identifiable, that is whether the probabilistic structure is described by *the unique* set of parameters. In the case of high-dimensional models it is perennially hard task, however, it may be partially addressed by checking whether MCMC procedure yields approximately the same parameters when started from different initial points. I have not found discussion of this problem in the thesis;

2. The fact that the point raised in comment **1** is not mere mathematical subtlety, is evident in the case of DelSIEVE, where this model should differentiate between two events which has the same effect on data: ADOs and deletions. The Author is obviously aware of this problem (c.f. p. 62) and the analysis of synthetic data shows that the proposed model disentangles both phenomena reasonably well, but the reason why it happens deserves closer scrutiny;

3. Candidate site identification (section 2.2.1). This is likely to be established procedure not devised by the Author, but it is necessary to note that selection of candidate sites based on rejection of a test which ensures control of type one error applied in case of one site, will result of spurious rejections when applied to many sites;

4. Evaluation of variant calling (p. 43): it seems that it would be beneficial, for the sake of comparison of SIEVE with methods which do not account for the difference between

[1] Ch. Bishop's 'Pattern Recognition and Machine Learning' Springer, 2006 had one author only

one and two mutations, to consider the case when negative class corresponds to wildtype genotype and positive is a sum of single and double mutations' genotypes. Also, standard approach would be to consider three-class classification scenario, although this will be highly unbalanced case and would require careful treatment (this has been clarified by the Author in an e-mail exchange);

5. Run time analysis (section 2.2.6) The discussion lacks the important comment that CellPhy was run on 4 to 7 threads whereas the number of threads in the case of SIEVE was approximately 5 times more. Despite that the computing times were comparable.

I believe that the points 1-3 above are important and deserve discussion during a defence. My overall opinion is quite clear-cut, however.

General appraisal In view of my evaluation above, I am convinced that Ph.D. thesis prepared by Senbai Kang, M.Sc., is an original and successful attempt to jointly model tumor phylogeny and genomic variants. The thesis unequivocally testifies to his thorough knowledge of genomics, in particular of methods of single cell DNA sequencing as well as Bayesian graphical models modelling and statistical data analysis.

Conclusion

In my opinion the thesis 'Probabilistic graphical models for inferring tumor phylogeny and genomic variants from single cell DNA genomic data' by Senbai Kang, M.Sc., fulfils the formal Polish requirements concerning Ph.D. theses in the field of exact and natural sciences, discipline computer science, and I am in favour of him defending it.

Jan Michniczuk