Advisor Diploma Thesis Review:
Haplotype Estimation using Sequencing Reads

by Bc. Anastasia Lebedeva

The student set off to improve the state-of-the-art genotype phasing algorithms (one of the last stages stage of a genome sequencing pipeline) by enhancing an existing algorithm and its model with additional data from an earlier stage. The student has proposed a modification of an existing state-of-the-art algorithm for a multi data source model, designed and implemented an algorithm in an existing program, implemented a preprocessing step to speed up the processing and tested the resulting algorithm on reasonably simulated data, obtaining results competing with the state-of-the-art.

Thesis overview

The thesis is well structured, with an illustrated introduction into the problem area, data formats, the problem itself and an overview of the prior work. In chapters 3 and 4, the author briefly describes the population-based phasing algorithm first employed in the Eagle 2 software and the extension to read-based phasing. The original algorithm, based on the PBWT transform, is not described in detail but that is understandable due to its complexity.

The joint mathematical model is described concisely but correctly, as well as the algorithm extensions to incorporate the read-based information into the model. While the full exposition of the algorithm would be more complete, the modularity of the HMM solving algorithm does not require that for a clear presentation.

The implementation introduces an intermediate format and a preprocessing step, both well-designed and described in chapter 5, including the asymptotic complexity and a practical discussion of the program performance and improvement potential. Chapters 6 and 7 describe the testing setup, choice of the testing dataset, rationale for the choice of testing parameters and the obtained results. These are presented with clear charts and discussed. The results include statistics and discussion of the types of resulting errors, one of the several points the thesis goes beyond the original requirements.

The language of the thesis is direct, concise and clear, on a level well suitable for a scientific publication. The bibliography is complete although the citations mostly lack issue numbers and publication years, a rather serious and needles omission.

Results overview

The presented results indicate that the algorithm outperforms both population-based and read-based methods, while the joint model based software (SHAPEIT) is much slower. This would need more testing and comparisons for a clear SOTA statement but that turns out to be rather difficult (partly due to low data availability and different conditions other tools use for benchmarking). The lack of a direct comparison is one of the potential improvements but indirect comparison is made and still indicates a favorable result. I believe that with some additional endeavor this line of work would be suitable for an impacted scientific publication and an application in the tools used in genomic pipelines.

Student contribution

The student surveyed the area and existing algorithms as well as their performance. She adapted and reformulated the SHAPEIT probabilistic model and augmented it for the Eagle 2 HMM state space. While the idea of incorporating the read information was a part of the thesis proposal, the algorithm itself was mostly a work of the student. The student implemented the algorithm on a loose reimplementation of the Eagle 2 algorithm developed by the advisor (me). While this code-base is developed for a different purpose (namely genotype refinement) and differs from Eagle 2 in several aspects, it is a good and comparable basis for experimenting with the model and performs similarly on genotype phasing itself. The student implemented both the algorithm and HMM state extensions on her own, as well as the pre-processing step and the intermediate format. The student also generated the simulated read data and performed all the tests.

Conclusion

The work and communication with the student was productive, with regular meetings to jointly plan the work schedule and track progress. She worked on the items effectively and actively proposed next steps, pointed out problems and was finding solutions to them, taking the ownership and responsibility for the thesis. Overall, it was a joy to work with her.

As stated at the start of the review, the thesis achieved its goals, showed not only the feasibility of joint-model phasing but also demonstrating its performance, opening the potential for further development. The student proved her ability to work on a state-of-the-art research project and present the results.

Recommended classification: **A**

July 12, 2018, Prague

Mgr. Tomáš Gavenčiak, PhD (advisor)
Dept. of Applied Mathematics
Charles University in Prague