UNIVERSITÀ
DEGLI STUDI
DI TRIESTE

Dept. of Mathematics and Geosciences
University of Trieste,
Via A. Valerio 12/1,
34127, Trieste, Italy.

phone +39 040 5582635
email gcaravagna@units.it
www www.caravagnalab.org

Trieste, 28th July 2023

To the PhD examination committee

REPORT ON THE DOCTORAL THESIS PRESENTED BY Shadi Darvish Shafigh

"Probabilistic graphical models for mapping tumor clones in cancerous tissues and single cells"

The PhD thesis submitted by Shadi Darvish Shafigh describes original work in the area of probabilistic graphical models for cancer integrative data analysis. The modern revolution of next generation sequencing has made a number of technologies accessible that can read out the molecular content of a tissue at various degrees of resolutions. In the context of a complex disease like cancer, where no single-gene can cause the neoplastic transformation of healthy cells, this technology has revolutionised the approach to study disease configuration in space and time, with and without treatment. This technology however poses analysis challenges, as it can measure thousands-dimensional data points, under variable and unknown error sources. For this reason, the adoption of advanced machine learning models has become necessary to make sense of modern sequencing datasets. While this has attracted a number of important machine learning researchers to the field of computational biology, many data analysis problems remain open and only partially approached. The continuous and fast technological turnover also demands for new solutions to be continuously developed.

In this framing, the thesis approaches some very foundational work on the identification of tumour clones (i.e., special population of interest in the context of cancer study) in tissues and single-cells, integrating data at different resolutions using the common ground of Bayesian probabilistic graphical models. One core aspect of the work is therefore the development of Markov Chain Monte Carlo samplers for the proposed models, and the creation of new machine learning tools that can be used to carry out advanced integrative analysis. Overall, this work approaches some core computational biology problems revolving around the biological question *"how do we map cancer genotypes into phenotypes"*, which comes forward in every chapter as the leitmotif of this work. Through 8 chapters (especially Chapters 5-7) sequencing data of different complexity are integrated, proposing solutions with increasing complexity. For every proposed model, a comparison with alternative methods in the field is presented, as well as some simple application to real data available in the public domain and from collaborators. For most parts, the main chapters (5 to 7) of the thesis are already published or about to be published in collaborative works. This is certainly an impressive amount of work.

Chapters 1 to 3 introduce the necessary background for an interdisciplinary thesis, e.g., basic cell biology concepts, cancer genomics and hallmarks and discusses the main contexts of application of this thesis. The basic types of sequencing data that is elaborated in this work is also discussed (whole-exome DNA, spatial RNA, immunohistochemistry and single-cell

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dept. of Mathematics and Geosciences
University of Trieste,
Via A. Valerio 12/1,
34127, Trieste, Italy.

phone +39 040 5582635
email gcaravagna@units.it
www www.caravagnalab.org

RNA), along with high-level bioinformatics pipelines used to preprocess these data before applying advanced machine learning algorithms.

Chapter 4 is a concise overview of probabilistic graphical models and sampling methods, refreshing basic concepts of Bayesian Network, Markov blanket, independence and Bayesian inference. In this chapter, more space is given to explain the Metropolis-Hastings and Metropolis-Within-Gibbs algorithms to sample posteriors, as well as clustering metrics that will be used in the successive chapters.

Chapter 5 presents CACTUS, a method to integrate the tumour clonal architecture with genomic clustering and transcriptome profiling of single cells, with an application to follicular lymphoma. CACTUS extends an earlier model to map single cells to clones based on allele-specific transcript counts on mutated positions, leveraging a Binomial likelihood with Beta priors, and a specific error model. The model parameters are sampled via Gibbs, and a performance better than competitors is observed on real and simulated data.

Chapter 6 presents TUMOROSCOPE, a method to integrate the clonal architecture over spatial transcriptomics data. As far as I am aware this is one of the first methods (if not the first) to approach this on spatial assays. The main idea is to combine a number of information (from imaging to spot-based RNA measurements) and carry out a clone deconvolution at the resolution of the spots in the assay (using similar Bayesian methods as in Chapter 5). A spot is the maximum resolution that can be achieved by these assays, and it is therefore necessary to regress spots-content against clonal structures to go beyond the resolution of the spot. This model has also a novel function due to a post-hoc regression procedure to infer RNA profiles of the clones, which overall allows mapping clones and their phenotypes in space. Even in this case simulations are used to measure the model performance, and breast and prostate cancer datasets are used as applications to real data.

Chapter 8 presents ClonalGE, an extension of TUMOROSCOPE that unifies the inference of the proportions of each clone (genotype/phenotype) per spot, with the estimation of a clonal gene expression profile. Essentially, it avoids the post-processing step of TUMOROSCOPE, and allows for explicit differential gene expression testing. This model uses Gamma and Poissons for clone counts as TUMOROSCOPE, and a reparameterized Negative Binomial for cluster expression profiles. Inferences are again performed via Metropolis-Hasting inside Gibbs sampling methodologies, and simulations confirm a better precision than TUMOROSCOPE, as well as higher-resolution deconvolution on the prostate cancer dataset previously analysed with TUMOROSCOPE.

Overall, the thesis of Shadi Darvish Shafigh is well structured and contains an accurate bibliography of the different areas of research addressed in her work. The technical machine learning contribution is very good, as most problems addressed were challenging, and developing sound Bayesian models was certainly non-trivial. This shows a very good machine learning skill set for the candidate. As mentioned above, each chapter corresponds to a published or submitted article, the material is well organised and I am confident that papers under review will be published soon.

Overall, there are two minor points that could be discussed with the candidate (and that do not require me to re-examine the thesis). One at the level of the implementation, another more at the level of the clones given in input to each method (more philosophical).

UNIVERSITÀ DEGLI STUDI DI TRIESTE

Dept. of Mathematics and Geosciences
University of Trieste,
Via A. Valerio 12/1,
34127, Trieste, Italy.

phone +39 040 5582635
email gcaravagna@units.it
www www.caravagnalab.org

Regarding the former, the author might make the implementation and release of each tool more structured. Sometimes I found them arranged at the level of scripts rather than tools with clear manuals and vignettes, which makes it hard to use these models for non-experts. Also, examples were mostly at the level of demos, whereas real-data analysis helps users better understand the models. This code might be available somewhere else linked to the publications, but I could not find it on GitHub. Regarding the latter point, instead, I understand that many tools rely on a priori known "clones". The field, in my opinion, has not yet converged to a stable definition of a "clone" as we can determine from bulk sequencing. A number of analyses here assume these inputs to be correct, but I am wondering if the integration of phenotype-level information has somehow the potential of resolving such conundrums. For example, in some real data the identified clones differ by a few differentially-expressed genes, are genetically similar (i.e., they share most mutations) and they are pretty much co-localised: *should this be an indication of the fact that those genetically–distinct input clones, at the end of the day, are the same clone?*

Summarising, considering the quality and quantity of the work done, the remarkable productivity of the candidate, and the quality of the articles already published, it is without any reservation that I deem the thesis as sufficient to grant a PhD, and I look forward to participating in the discussion.

Prof. Giulio Caravagna,

Associate professor of Computer Science
Head of the *Cancer Data Science Laboratory,*

Department of Mathematics and Geosciences,
University of Trieste,
Italy