Università degli Studi di Milano– Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione

Milan, ITALY, 2023-07-26

Marco Antoniotti
Dipartimento di Informatica, Sistemistica e Comunicazione
Università degli Studi di Milano Bicocca U14
Viale Sarca 336
I-20126 Milan, MI, ITALY

tel.: +39 02 6448 7901
web site: http://dcb.disco.unimib.it
email: marco.antoniotti@unimib.it

REPORT ON THE DOCTORAL THESIS PRESENTED BY SHADI DARVISH SHAFIGI:
"Probabilistic graphical models for mapping tumor clones in cancerous tissues and single cells"

The PhD thesis submitted by Shadi Darvish Shafighi discusses the application of Probabilistic Graphical Models (PGMs) to the overall problem of describing the clonal structure of a tumour, as it progresses from early, to advanced, and eventually to metastatic stages. This is a very important problem to tackle, as its understanding will have immediate therapy consequences impacting both the life expectancy and the well-being of a patient. The thesis presents work that leverages the current availability of *bulk* and *single-cell* sequencing data in the form of DNAseq and RNAseq data; the thesis also presents an innovative take on the novel field of *spatial transcriptomics*, which is the latest type of biotechnology analysis techniques used to elucidate the behaviour of biological systems, of course including cancer.

The thesis starts with introductory material and then it presents three "projects" which are used to analyse different aspects of the clonal reconstruction problem. Each project is centred on a particular application of PGMs to a given type of data, reified in a specific tool constructed (and distributed). The three tools are CACTUS, Tumoroscope and clonalGE, covering different aspects of the analysis tasks as well depicted in Figure 1.1; the three tools have been published either on relevant journals or at conferences in the field and are available on the usual platforms.

The introductory material of the thesis starts with a description of the relevant data analysis problems deriving from the nature of the available measurements, followed by an exposition of much biological background including the revised "Hallmark of Cancer" put forth by Hanahan and Weinberg and the overall notion of cancer heterogeneity in its declensions as treated in the thesis. The three tumour types analysed in the three projects – prostate, breast, and follicular lymphoma – are also introduced. Next there is a quick and to the point description of the kind of data produced by sequencing techniques and by Hematoxylin and Eosin (H&E) stained images.

The next chapter of the thesis, the fourth, is a synthetic introduction to the statistical methods used in the three projects. PGMs are introduces alongside the usual fundamental concepts of likelihood, posterior and posterior maximization. All of this is an introduction to the next three chapters describing the three projects.

The three chapter on CACTUS, Tumoroscope and clonalGE, have a similar structure, with their core dedicated to the definition and fine tuning of a specialized PGM that will solve the problem at hand. A discussion of the actual use of the model presented concludes each chapter.

The chapter on CACTUS describes how the method is applied to the analysis of Follicular Lymphoma. The method combines a preliminary phylogeny reconstructed from whole exome sequencing (WES), read counts

from Single Cells sequencing experiments and cell clusters defined on the basis of identical B-cell receptor (BCR) heavy chains sequences. The overall goal of CACTUS is to reconcile the clones initially obtained by the phylogeny reconstruction procedure (which uses off-the shelf tools: FALCON-X, GATK and Canopy) with the BCR clusters. The chapter proceeds with the description of the development of the CACTUS PGM that is used to estimate the probabilities of assignments of a cluster to a clone; each step is clearly described and justified. Next, the chapter proceeds by describing how the model is used to produce results by means of a standard Gibbs sampler; again, the details of each sampling step are painstakingly described and justified. The chapter continues by clearly describing the validation steps and comparing the results with those of a previous method (cardelino). As usual, with these kinds of validations, several different measure must be employed to justify the improvements claimed; the chapter how several standard measures are used to assess assignments based on gene expression (exome sequencing): Root mean standard deviation (RMSSTD), connectivity, Dunn index and Calinski-Harabasz (CH) index. Gini Index and Entropy measures were used to assess the assignments of cells to clones.
All in all, CACTUS utilizes shared BCR sequences defining clusters of single cells to assign cells to clones, effectively handling errors and dropouts in single cell RNA sequencing and addressing the challenge of inferring accurate clonal structures. This contribution represents progress in developing computational tools to resolve tumour heterogeneity and understand the functional diversification of tumour cell subpopulations by integrating genotype with gene expression profiles.

The chapter on Tumoroscope expands on the theme of defining PGMs to analyse data that has become very important in the past four or five years: i.e., spatial tissue derived transcriptomic data (*spatial transcriptomic* – ST – data for short). ST aims to identify spatial heterogeneity in a (tumorous) tissue slice. The analyses reported in the thesis pertain Breast and Prostate cancer data, while most of the validation is done by extensive simulation studies. The method reconstructs the clonal architecture of a tumour tissue stating from H&E images and it can correctly deconvolve the signal in each image spot.
Again, the presentation of the method is accompanied by a careful development of the PGM and of the relative Gibbs sampler used to estimate the posterior probabilities of the spot vs transcriptome assignments.

The final chapter of the thesis, prior to the conclusions, describes the clonalGE method that relies and integrates the results of Tumoroscope to associate gene expression profiles to spatially heterogenous clones. clonalGE is an innovative statistical graphical model designed to infer clone-specific gene expression and the composition of clones in localized spots within tissue samples. clonalGE utilizes a combination of WES data, H&E images, and read counts for somatic SNVs shared between WES and ST. Additionally, it leverages gene expression measurements across the tissue sample to enhance the precision of clone mapping to specific regions. Moreover, clonalGE enables the generation of distributions for clone-specific gene expression profiles, facilitating subsequent differential expression analysis between pairs of clones. This model advances the comprehension of tumor heterogeneity, encompassing not only genetic but also phenotypic aspects.
As with the previous chapters, the steps of the development of the PGM and of the Gibbs sampler used to estimate the outcome probabilities are very well detailed and justified; also, in this case, a study of hyper-parameter initialization is performed, thus lending more robustness to the results.

The thesis work of Ms. Shafighi's is of extremely high quality especially in the consistent application of a methodology that, starting from the understanding of the current cancer data available, constructs PGMs and Gibbs samplers that allow to find appropriate posterior probabilities of various assignments of inferred clonal architectures.

The work is very well organized and readable, with an extensive and precise bibliography. The written language demonstrates an exceptionally high standard. The extensive range of questions tackled, and the achieved results attest to Ms. Shafighi's diverse and proficient computational skills. In conclusion, based on

Università degli Studi di Milano– Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione

the impressive quality and quantity of the work undertaken, the candidate's remarkable productivity, and the excellence of the already published articles, I wholeheartedly authorize the defense of Ms. Shafighi's thesis. It has been a pleasure to serve as a reviewer of this work.

Milan, 2023-07-26

Marco Antoniotti

U14, Viale Sarca 336 – 20126 Milano, ITALIA – Tel. +39 02 6448 7901 – Fax. +39 02 6448 7809