POLISH ACADEMY OF SCIENCES
NENCKI INSTITUTE OF EXPERIMENTAL BIOLOGY
Pasteur 3, 02-093 Warsaw, Poland
Phone: (48-22) +5892209 Fax: (48-22) 822 53 42
http://www.nencki.gov.pl

Professor Bozena Kaminska, PhD
Laboratory of Molecular Neurobiology
email: b.kaminska@nencki.edu.pl

Warsaw, 15.08.2023

The review the PhD thesis by Shadi Darvish Shafigh entitled:
Probabilistic graphical models for mapping tumor clones in cancerous tissues and single cells

The thesis was supervised by: Dr hab. Ewa Szczurek, University of Warsaw, Faculty of Mathematics, Informatics and Mechanics and Prof. Alessandra Carbone, Sorbonne University, Faculty of Computer Science.

The doctoral dissertation of Ms Shadi Darvish Shafigh addresses an important and timely issue referring to the intratumoral heterogeneity and clonal evolution of tumor cells resulting in different genetic and phenotypic compositions of tumor cells in spatially distinct parts of the tumor. Tumor heterogeneity poses a challenge in patient treatment as it is impossible to eradicate all tumor clones/cells if the applied therapy targets a specific genetic vulnerability present in a subpopulation of cells. Elucidating tumor heterogeneity is impeded by difficulties in direct identification of single cells and the localization of different clones in the tumor tissue. Evolutionary modeling helps to understand how tumors arise and plays an increasingly important prognostic role in predicting disease progression and the outcome of medical interventions.

The PhD dissertation is composed of a short Introduction of topics, both biological and mathematical regarding various methods used in complex data analysis, forming a background for next chapters. Subsequent chapters represent the three manuscripts in which the PhD candidate is a leading author (one published in Genome Med 2021, one on BioRXv and one in preparation). Each of those chapters is composed of the manuscript organized in a classical ways with Introduction, Methods, Results and Discussion sections. As those manuscripts have multi-authors, it would be proper to indicate the exact contribution of the PhD candidate in the specific work.

In the Chapter 5 a probabilistic graphical model for integrating Clonal Architecture with genomic Clustering and Transcriptome profiling of single tUmor cellS (CACTUS) is proposed and used to analyze newly generated whole exome sequencing (WES), single-cell (sc)RNA-seq and scBCR (B cell receptor) sequencing data on malignant lymph nodes of two follicular lymphoma patients. The proposed model extends the cardelino, a Bayesian method for integrating somatic clonal and transcriptional heterogeneity within a population of cells (Nature Meth. 2020). The data on genomics of tumor clones, mutation transcript counts, and single-cell the BCR cluster information (clones with similarity of BCR heavy chain sequences) are combined in the CACTUS model to infer the clonal assignment of the clusters to cells and create the tumor evolutionary tree. Both the input clone genotypes and clustering are considered potentially imperfect and were corrected during the inference using all available data. CACTUS maps single cells to their clones based on comparing the allele specific transcript counts on mutated positions. A custom made script was used to generate single cell allelic transcript counts and identify reads intersecting with WES-based mutated positions.

Quantification of the agreement of the cell-to-clone assignment with gene expression profiles of the cells showed better values than the cardelino. The confidence of cell-to-clone assignment was measured as the concentration of the probability distribution of assigning a cell to clones, averaged across cells. For both subjects, the CACTUS mapped cells and BCR clusters with substantially higher confidence than the cardelino. One may wonder if data from 2 patient samples were sufficient to draw any conclusions about biological significance the data, particularly that only 1,500 cells per patient were sequenced. Cardelino was applied to a published cancer dataset and to newly generated matched scRNA-seq and exome-seq data from 32 human dermal fibroblast lines, so data were really strong. I wonder why the CACTUS was not compared or validated on the data from Okosun et al. Nat Genet. 2014 providing the information of genomics and transcriptomics of follicular lymphoma patients. Despite some limitations, CACTUS appears to be a step forward in establishing computational tools to resolve the tumor genetic heterogeneity.

The Chapter 6 contains the work presented in the manuscript deposited at the BioRXv doi: https://doi.org/10.1101/2022.09.22.508914. The study addresses the important issue because with development and popularity of spatial transcriptomics technologies there is a great need for a good method of data deconvoluation and overcoming challenges posed by the presence of many cells in a single spot (i.e., 10X Genomics Visium). The PhD candidate proposes Tumoroscope, the probabilistic model that infers cancer clones and their high-resolution localization by integrating pathological images, WES and spatial transcriptomics data. The authors claim that in contrast to previous methods, Tumoroscope addresses the problem of deconvoluting the proportions of clones in spatial transcriptomics (ST) spots. Tumoroscope was applied to a reference prostate cancer dataset (1 patient) and a newly generated breast cancer dataset (1 patient, 5 regions) and revealed spatial patterns of clone colocalization and sub-regional mutual exclusion in the tumor tissue.

From a biologist point of view drawing conclusions with a single patient data is problematic and it would be good to see reproducibility of the method but as a proof of concept for a new computational approach it might be sufficient. As before, some aspects were compared to the cardelino performance that did not detect spatial pattern of domination of clones in sub-areas, as Tumoroscope did. The median correlation of clone proportions inferred by Tumoroscope between adjacent spots was significantly higher than the median correlation between distant spots, thus this result supports the correctness of the deconvolution of ST spots using Tumoroscope. The applicability and robustness the new method is discussed in the context of existing methodologies.

In the Chapter 7, ClonalGE, a novel statistical graphical model is presented that infers clone specific gene expression and the composition of clones present in localized spots within a tissue sample. It extends a previous model, Tumoroscope, by incorporating additional variables for gene expression in the spots. It explores WES data, histological images, sequencing data for somatic SNVs shared between WES and ST. ClonalGE leverages expression measurements of genes across the tissue sample to improve the accuracy of mapping clones to specific regions and allows the generation of distributions of clone-specific gene expression profiles, allowing subsequent differential expression analysis between pairs of clones. The method has been tested on 12 sections from one prostate cancer patient. The performed comparison show that ClonalGE reconstructs the gene expression profile of the tissue more accurately than Tumoroscope. The model attempts to expand understanding of the tumor heterogeneity, at the genetic and phenotypic levels.

The Chapter 8 recapitulates the content of the previous parts, the limitations of the technologies generating data and confines of the employed computational methods. The results are discussed in the context of their applicability and better performance over existing methods. The extension and future improvements the methods are discussed. This part is succinct, competent and show a maturity of the PhD candidate and her fluency in the scientific discourse.

Altogether, in this dissertation the PhD candidate presented the combination of new or improved mathematical methods and implemented those novel computational techniques to study the presence of genetically distinct sub-populations of cancer cells with different phenotypic behavior within a heterogeneous tumor. This has been done by integrating different types of cancer datasets. The presented analyses increased the reliability of applied methods in comparison to previously published methods. Each of the three described projects addressed the research problems in a specific manner and provided the refinement of the methods.

The dissertation is well written and illustrated with transparent schemes of the performed steps. I predict the results of the presented studies would be widely used and appreciated by the scientific community.

The minor weaknesses spotted:

1. The structure of the Introduction is peculiar and confusing. Just after a brief statement of about probabilistic graphical models (PGMs), the PhD candidate introduced a subchapter *1.1. Research topics covered in the thesis* in which the three project being a part of the dissertation are briefly summarized. In the next chapter the challenges of each projects are discussed, in particular related to deconvolution and separating individual components of a mixed sample in bulk sequencing data or feature allocation in the mixtures of aggregated reads coming from the cells belong to different clones with the same coordinates in the spatial transcriptomics spot. Next, she briefly describes how these stated limitations and biases have been overcame, how various solutions have been employed to increase accuracy and reliability.
2. I am confused about a meaning of the statement: *"Lack of ground truth"* and I think it is inaccurate in a scientific discourse. I think it was meant as uncertainty of predictions in regard to the real biological data and the candidate tries to provide some computational ways to estimate which classifier is more accurate. Such considerations are valid and interesting. However, verification of predictions with other biological approaches (in case of BCR clusters flow cytometry based quantification of clones or FISH in case of spatial transcriptomics) would be the best way to offer "ground truth". While such considerations are important, the entire fragment should find a place in the Discussion.
3. Some statements are trivial and inaccurate: i.e. p. 21 *"In this chapter, our aim is to establish a basic understanding of cancer and its properties"* . The Chapter is limited to the discussion of the cell cycle and there is much more mechanisms related to cancerogenesis than that. Saying" *There are some cells that do not want to divide again, or they divide very slowly"* is not accurate as cells are not autonomous to want to divide, cell division is a process very strictly controlled by numerous factors and the environment. *Inflammation is not a component of the immune system* but a response or reaction of the system to injury or infection. The PhD candidate uses sometimes an eccentric language to describe biological processes but in general does a fair job in presenting the hallmarks of cancer and features of selected cancers considering her education.
4. Regarding the comment in the Chapter 7: *"Previous studies did not decisively determine whether the genotypes of the clones determine their gene expression".* I think the motivation of those studies and the assumption was the genotypes of the cancer clones will determine their gene expression and of course nobody expects that the genotypes of the non-cancerous clones (that are the same) will determine differentially their gene expression.
5. There are several references in the text regarding "the lack of ground truth" and difficulties in estimation of the reliability of the methods. I wonder why the real data from those cancers acquired with cell sorting and genomics/transcriptomics (i.e., TCGA datasets or pathological datasets) had not been explored to find support for a validity of computational predictions.

The minor concerns raised above, do not diminish my high opinion about the dissertation. Taking into account the innovative and scientific value of the findings reported in the doctoral dissertation, the successful combination of carefully designed computational tools and fundamental research questions, I value this doctoral dissertation as an important contribution to computational sciences in the field of multi-omics and genomics of cancer.

In my opinion, this dissertation exceeds statutory requirements for doctoral dissertations, constitutes an original solution to an important scientific problem, demonstrates the excellent candidate's computational skills and good general biological knowledge in cancer, demonstrates the ability to independently conduct scientific work.

The reviewed dissertation meets all the conditions set out in Art. 187 of the Act of July 20, 2018 Law on Higher Education and Science (Journal of Laws of 2021, items 478, 619, 1630). Therefore, I am asking the RND Mathematics and Computer Sciences mgr. Shadi Darvish Shafighi to the next stages of doctoral proceedings.

In addition, considering the high substantive level of the dissertation, high level of originality and considerable achievements, transparent way of presenting the research topics, methodology and results, I would like to request that the dissertation be distinguished with an appropriate award.

Stwierdzam, że recenzowana rozprawa doktorska mgr. Shadi Darvish Shafighi spełnia wszystkie wymagania Ustawy z dnia 20 lipca 2018 r. – Prawo o szkolnictwie wyższym i nauce. Zwracam się do Rady Naukowej Dyscyplin Matematyka i Informatyka Uniwersytetu Warszawskiego o dopuszczenie mgr. Shadi Darvish Shafighi do dalszych etapów w postępowaniu w sprawie nadania stopnia doktora.

Ponadto, mając na uwadze wysoki poziom merytoryczny rozprawy, bardzo wysoki poziom oryginalności i znaczący dorobek, zwracam się z prośbą o wyróżnienie rozprawy stosowną nagrodą.

Z poważaniem

Podpis jest
prawidłowy
Dokument podpisany przez
Bożena Kamińska-
Kaczmarek
Data: 2023.08.13 09:46:27
CEST