Silesian University of Technology in Gliwice, Poland
Automatic Control, Electronics, and Computer Science Department

Title: Classification of white blood cells based on single-cell sequencing data for biodosimetry purposes

Author: Katarzyna Sieradzka

Supervisor: prof. dr hab. inż. Joanna Polańska
Advisor: dr Christophe Badie

Acknowledgments: This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029.

Abstract

Single-cell RNA-sequencing (scRNA-seq) is an increasingly widely used technology to analyze the transcriptome of many single cells. By sequencing the genome of single cells, it is possible to avoid the data generalization problem in sequencing technologies that do not focus on individual cells. As a result of utilizing this technology, high-dimensional data is generated, which requires more and more computing resources to make the proper analysis. The scRNA-seq technology is essential for investigating cell-to-cell heterogeneity in analyzing the impact of specific factors, such as a cellular response to ionizing radiation. Unconscious or conscious exposure to radiation induces changes in cell responses caused by modifications in the expression of many genes regulating cell lives. The analysis of such modifications can reveal genes that are most involved in the radiation response. Such analysis can also demonstrate gene communication pathways that could give insight into what changes occur throughout a complex cellular system. By combining the knowledge about the level of radiation-induced changes and the available sequencing technologies, we can perform appropriate analysis steps that will allow us to learn about the genes that respond to radiation.

This work has two main goals. One is purely biological, while the other is related to engineering. The biological aim of this study is to search for known and new genes of radiation response based on the data from single-cell RNA-sequencing techniques. This way, the differences in the gene signature of normal cells and those subjected to ionizing radiation will be determined. A fundamental goal in engineering is to create an appropriate bioinformatic analysis workflow to partially automate the consecutive steps of working with high-dimensional data from single-cell sequencing experiments. The main aspects included in the proposed method are based on feature selection procedures and the problem of cell classification itself. It is a considerable challenge, especially considering the very high complexity and dimensionality of the analyzed data, but also the expectations of achieving satisfactory results regarding the quality of the classification of irradiated cells. The expected outcome of the created tool is primarily related to the biological purpose of the research, i.e., to the recognition of the complete genetic profile of cells irradiated in an ex vivo environment.

The first work stage focuses on data quality control. For this purpose, two ex vivo samples, technical repetitions of the same experiment, were tested. Several statistical and visualization paths were developed to allow detailed analysis of the quality of both genes and cells. The methodology used, especially the unsupervised classification approach utilized for visualization, allows for drawing an unambiguous conclusion about the significant heterogeneity of the studied data sets. Therefore, attempts were made to determine the cause of such cell heterogeneity using public and own-developed mathematical and statistical methods. Moreover, a list of subpopulation-specific marker genes was also used to designate white blood cell subpopulations. It was proved that the chosen research path determined the cause of the internal heterogeneity in complex data sets related to the occurrence

of highly-differentiated cell subtypes. Moreover, as a result of a series of analyses, it was possible to detect frequently occurring subpopulations of this fraction in the quantitative context and rare and small subpopulations of white blood cells. The works' main stage aims to build a classifier based on logistic regression methods. The purpose of the classifier is to distinguish control and ionizing radiation-subjected cells. At this stage of the work, only the T-cells subpopulation was considered as it constituted most of the selected white blood cell subtypes. What is essential, the applied procedure made it possible to remove the substantial heterogeneity of the data set. Next, to standardize the structure of the analyzed data set, there was also performed the data normalization procedure. A feature selection procedure was based on cells and genes prepared this way. For this purpose, an own-implemented workflow was developed, enabling the classification of normal and irradiated cells with adequate measures of classification quality. As a result of using the implemented workflow, a radiation response genes panel was finally obtained. Interestingly, a significant majority of found genes correspond to current literature reports. While reducing the impact of the heterogeneity in the data set allowed to improve the classification quality to obtain very satisfactory results with the value of the weighted accuracy, based on the hold-out test data set, at above 93%. Additionally, a detailed analysis was made using the neural networks approach to compare logistic regression-based workflow with another well-known method. Another machine-learning analysis workflow was created that is compatible with the stated goal to recognize irradiated cells set out in the dissertation. This approach was primarily aimed at checking and comparing the quality of classifications resulting from using two different feature selection techniques. Using neural networks made it possible to obtain promising results, with a classification quality value of almost 91%. Moreover, such results were achieved in a much shorter time frame, comparing neural networks with a logistic regression-based approach. On the other hand, what is even more critical in undertaking analysis this way, it was also possible to compare the genetic profiles of irradiated cells resulting from the logistic regression and neural networks-based approaches. It occurred that 8 out of 10 genes creating the neural networks-based model are familiar with the logistic regression-based procedure. These well-known genes of radiation response include RPS19P1, BAX, DDB2, RPS27L, PHPT1, CCNG1, TNFSF8, and AEN.

This doctoral dissertation shows that using data derived from a precise and detailed technology, such as scRNA-seq, it is possible to determine the specific gene structure of cells subjected to ionizing radiation. This work also made it possible to compare two machine-learning techniques: logistic regression and neural networks-based approaches. Several bioinformatics methods and different workflows developed can be used in the future as support in medicine, science, and engineering. The developed method for feature selection and irradiated cell classification met the challenges posed in the dissertation with very high efficiency. This research describes exactly the workflow of high-dimensional data analysis from single-cell sequencing experiments, such as the extended quality control, through the recognition of radiation response genes, the determination of the irradiated cells gene signature, classification of white blood cells along with the subpopulations recognition, the comparison of machine learning procedures in terms of high dimensional data analysis and observations' classification, and also the biological interpretation of the results. Therefore this work covers, with a detailed description of the proposed analysis steps and the effects in the form of results, all aspects necessary to achieve the assumed goals, combining them into a logical workflow with appropriate comments and inferences from both the technical and engineering side, and supports these aspects in the form of a biological interpretation.