Politechnika
Śląska

SILESIAN UNIVERSITY OF TECHNOLOGY

Faculty of Automatic Control, Electronics and Computer Science

Institute of Automatic Control

Integrative data analysis methods in multi-omics molecular biology studies for disease of affluence biomarker research

Doctoral Dissertation

by

Anna Papież

Supervisor

prof. dr hab. inż. Joanna Polańska

2019

Gliwice, POLAND

Abstract

The need for transforming large amounts of data in the life sciences drives the development of statistical and data mining algorithms for merging and validation of biomedical experiments. Although this issue has been previously commonly acknowledged in the scientific community, the constantly increasing amounts of data require continuous efforts towards the optimization of data analysis pipelines. Therefore, the aim of this thesis is to investigate diverse approaches for high-throughput molecular biology integrative data analysis to enable the discovery of disease of affluence biomarkers. The work consists of a detailed overview of existing advancements in high-throughput molecular biology techniques data integration, followed by the demonstration of novel algorithms for combined analysis of data derived from multi-platform and multi-domain experiments.

Initially, an original batch effect identification algorithm based on dynamic programming is presented, as correcting for these effects constitutes a part of the intra-experiment data integration pipeline. Its performance on identifying batch structure is proven to be highly efficient, and moreover, batch effect preprocessing entails potential new knowledge discovery in studied diseases and conditions.

Subsequently, two microarray data sets obtained using different platforms for biomarker research in breast cancer patients are analyzed to highlight the potential of measurement transformation to achieve computational and biological consistency. The statistical and data mining integrative approaches with functional validation and profile modeling provides a comprehensive solution for elucidating dose response mechanisms and potential biomarker signatures. Moreover, custom statistical integrative methods applied to a transcriptomics and proteomics data set on ischemic heart disease plutonium mine workers enabled discrimination of dose dependent protein expression changes from the age dependent changes and validation of pathways identified previously in the proteomic data. Another approach to data integration, which enabled the identification of factors playing a key role in differentiation of irradiated samples, was conducted on multi-tissue exosome proteomics data.