Ruprecht-Karls University of Heidelberg
Faculty of Engineering Sciences
BSc Molecular Biotechnology

Drug viability screens for oncological and non-oncological treatments for breast cancer

Data Science Project Summer Semester 2023

Topic 5 Team 4

Submission date: 17.07.2023

Luis Herfurth, Aaron Eidenmüller, Sharujan Suthakaran, Simon Westermann

Abstract

Hier muss das Abstract eingefügt werden

Contents

Abbreviations

1 Introduction

As breast cancer is the most common type of cancer in women, accounting for approximately 2.3 million cases each year, treating breast cancer is still a major objective in global research (Łukasiewicz et al., 2021). "Even though a significant number of cancers do not always need to result in death, they significantly lower the quality of life and require larger costs in general." (Łukasiewicz et al., 2021) While therapies such as chemotherapy have significantly improved outcomes, they are often accompanied by notable side effects. Therefore, there is a critical need to discover new compounds that can enhance treatment outcomes for breast cancer while minimizing adverse effects. (Kumbhar et al., 2023) Nevertheless, the process of drug development is lengthy and expensive, typically taking around 12-15 years and costing billions of dollars to bring a new compound to market (Wouters et al., 2020). This challenge is further amplified in the field of cancer, as cancer cells exhibit a heightened resistance to treatment with estimations that only about 5% of potential drugs enter the clinical testing (Sleire et al., 2017). Consequently, researching and developing new cancer treatments pose substantial risks for pharmaceutical companies. Small companies and start-ups, in particular, require significant investments to undertake such endeavours. It is estimated that for every dollar a pharmaceutical company invests in drug development, less than a dollar is returned, resulting in minimal profitability and discouraging potential investors. This issue is particularly detrimental to the development of medications for rare diseases and those prevalent in developing countries, as the financial incentives are even more limited (Pushpakom et al., 2019). A potential solution to this problem is drug repurposing, which involves utilizing already approved drugs for the treatment of other diseases. By repurposing existing drugs, the cost can be significantly reduced to around 300 million dollars, and the development time can be greatly shortened, as these drugs have already undergone the necessary regulatory processes. (Nosengo, 2016) By addressing these challenges through drug repurposing, we can expedite the availability of new treatment options for breast cancer and other diseases, making substantial progress in improving patient outcomes while maximizing resource efficiency. Drug repurposing can be approached through various methods, including computational and experimental approaches. In this research project, the chosen approach was a computational one, specifically focusing on genetic associations. The objective was to identify genes associated with the disease that could potentially serve as targets for future drug development (Pushpakom et al., 2019). To achieve this, treatment response data for cancer cells obtained using the PRISM method, along with genetic data such as gene expression profiles of the cancer cell lines, were collected. These data were utilized to explore new opportunities for drug repurposing. In the beginning, the treatments in the data sets are categorized on their potential of inhibition. This data will then be used for analysis of similarity between treatments as well as being the basis for a genetic analysis of the data. Specifically, this research project emphasizes the use of gene expression as an indicator for identifying genes that could serve as new drug targets. With advancements in analysis techniques and the increasing affordability of gene expression analysis, it has become feasible to analyse gene expression for individual patients and identify potential drugs for targeted treatment (Chawla et al., 2022). The potential target genes are subsequently evaluated based on their gene knockout scores, which indicate the extent to which the knockout of the target gene inhibits cell growth significantly. Furthermore, predicting the effectiveness of the possibly repurposed drugs depending on the treatment

dose is another objective of our research. As during the preclinical and clinical testing phase drugs are evaluated on their effective dose and their maximum dose, it is important to be able to predict if a drug can have a beneficial effect in the permitted dose.
By leveraging the genetic association approach and focusing on gene expression analysis as well as analysing the inhibiting treatments, this research aims to identify promising candidates for drug repurposing and pave the way for the development of new treatment options tailored to specific patients.

2 Materials and Methods

2.1 Data

2.1.1 Prism Datasets

Prism: effect of the treatment (columns) on cell growth of the cell lines (rows); includes drug, dosage and assay used

Prism.treat: for each treatment (rows) further information on the treatment and drug

Prism.cl: contains information about the different celllines

2.1.2 Cellline Datasets

Prism.exp: contains levels of gene expression. Celllines (rows) and genes (columns)

Prism.cnv: contains copy number levels of genes. Normal is CN = 2. Gene names (rows) and celllines (columns)

Prism.snv: marks mutation in the different celllines als functional or nonfunctional to the cancer.

Prism.achilles: has information on how important a gene is for cell survival. Was generated using knockdown celllines. Gene names (rows) and celllines (colums)

2.2 Data clean up/Filtering

Show distributions after cleanup Abbildung für prism vorher nachher. Für andere maybe im Clean up

Before performing the cleanup process, a subset of the data frame was created specifically for the copy number data of the 22 breast cancer cell lines. This subset was obtained by extracting the ID codes of the breast cell lines from the comprehensive data frame that included information on all cell lines used in the PRISM screen. The extracted ID codes were saved in a vector, which was then utilized to construct a new data frame containing the copy number information solely for the 22 breast cell lines. Next, both copy number data frames were checked for missing values, and it was determined that no missing values were present. Consequently, the data frames were deemed suitable for further analysis and processing.

2.3 Dimension reduction

Uniform Manifold Approximation and Projection (UMAP) was utilized as a method for reducing the dimensionality of data in this study. UMAP is a non-linear technique designed to retain local structure and capture intricate relationships within high-dimensional datasets. The 'umap' function from the 'umap' package in R was employed to apply UMAP. The method involves several steps: initially, pairwise distances are computed between data points using a selected distance metric. Subsequently, a neighborhood graph is constructed by connecting each point to its closest neighbors. UMAP then estimates a fuzzy-topological representation of the data based on this graph, capturing the connectivity strength via a fuzzy simplicial set. To find a low-dimensional representation that minimizes the inconsistency between pairwise similarities in the original high-dimensional space and the reduced space, an optimization process driven by stochastic gradient descent is conducted. This iterative optimization accounts for attractive and repulsive forces, ultimately yielding a lower-dimensional embedding that represents the data. The 'n_neighbors' parameter specifies the number of nearest neighbors considered during graph construction, and the 'min_dist' parameter governs the minimum distance between points in the low-dimensional embedding. The resulting UMAP embedding maintains local structure and provides a visually interpretable representation, facilitating data exploration and analysis.

3 Results

First include positiv results; if space is left include negativ results: UMAP, K means clustering, promoting drugs describe goal, describe process, describe outcome

3.1 Gene search engine

Goal: Arbeitsvereinfachung; Outcome: Overview over data Für Präsentation als Visualisierungstool pitchen Maybe Website, ***Discussion***

Analysing the main datasets led to many individual data formats such as data frames, lists etc.; which contain relevant information gathered by our code. ***In general, one could get a possible important gene through literature analysis or comparing every meaningful value in our database manually, which is time-consuming and one could lose track.*** The aim of the function of the search engine was to quickly search for a gene of interest and display its attached values out of our main cell line datasets (prism.exp, prism.achilles, prism.cnv) for a more simple process of gene analysis. Based on this code, applications like looking for suitable treatments in prism.treat, loops which undergo printing every gene and their attached values in breast cancer celllines and the first approach for a relevant final data frame were realized. The conclusive outcome of this engine development was a final modified search engine, in which one can type in the gene of interest and it prints every relevant value or information to get a decisive overview for drug repurposing applications.

3.2 List of inhibitory drugs

Results von Data clean up und filtering. Goal: List of Inhibitory Drugs; Outcome: List of Inhibitory Drugs Bilder vergleich liste vergleich ohne threshold und mit threshold Maybe oncological drugs rein screenen

3.3 UMAP

3.4 Gene analysis

3.4.1 Correlation analysis

treatment response / gene expression; Goal: finding relevant genes; Outcome: giant data frame -> used for further work

copy number / gene expression Goal: looking if hypothesis correct; Outcome: Histogram of correlations

The initial step in gene analysis was examining the correlation between different variables to determine if there were any connections between them. This approach aimed to identify genes that could potentially serve as indicators of a drug's effectiveness. To begin, a correlation analysis between the gene expression data in prism.exp and the gene copy number data in prism.cnv was conducted. The method used was the Pearson correlation. Our hypothesis was that the copy number would correlate highly with gene expression, as higher copy numbers typically lead to increased transcription (Shao et al., 2019). Figure XX presents the histogram of the correlation calculations, which demonstrates that, for the majority of genes, a positive correlation indeed exists. The mean correlation was approximately 0.325, with a median of 0.345. Yet, the correlation is not as high as was expected. In the subsequent step, we constructed a correlation matrix using Pearson correlation to examine the relationship between gene expression data from prism.exp and treatment response data from prism.treat. This correlation matrix allowed us to further refine our understanding of which genes were associated with specific treatments. The resulting matrix contains correlations between 18,805 genes and 1,395 treatments. The data was then used in the further gene analysis.

3.4.2 Statistical testing of important genes

Test wich of the found genes are for breast cancer of interest Goal: find out which one are negativ, which ones are lower than other lineages; Outcome: 2 genes

A threshold of absolute correlation greater than 0.75 was applied to select genes for subsequent analysis. This threshold was selected as it is the mean of the highest absolute correlation for each treatment. The filtering resulted in selection of 3925 genes out of the total 18119 genes in the prism.achilles dataset. These genes were then sorted based on their prism.achilles scores being lower than 0 and then a one-sided Wilcoxon rank sum test with significance level 0.05 was performed to assess if their mean scores were significantly lower than those of other lineages. As previous Shapiro-Wilk-Test showed the data is not normally distributed and therefore a non-parametric test was needed. The p-values were

adjusted using the False Discovery Rate (FDR) correction. One gene showed a significant difference, SDHC with a p-value of 0.025.

3.4.3 Dataframe for targets involving genes

mean of data frames. Threshold for what genes are relevant. Used indings from correlation tests Goal: finding interesting genes; Outcome: Data frame with many genes -> 48 genes data set with filtering after gene knockout score

Moreover, the 3925 preselected genes were compiled into a data frame for further research. Genes with a mean achilles score below -1 for breast cell lines were included in the final data frame. 108 preselected genes fulfilled this criterion. This data frame includes information such as the frequency of high absolute correlation with a treatment, mean prism.exp score, mean prism.achilles score and the p-value from the Wilcoxon rank sum test. Furthermore, it provides additional details on gene mutations in breast cancer cell lines, treatments targeting the gene, and its association with any cancer hallmarks. This data frame concludes our gene analysis as it combines our gene selection process with general information we gather about the genes along the way. Two genes showed a high frequency amongst treatments, mTOR and AURKA. These genes were associated with more than 20 treatments during the analysis.

3.5 Linear regression

Perform drug by drug to avoid weird plot; For every drug one linear regression, R^2 Value and with those showing, that many of them are very good. Prediction model for concentration and drug name. Plot um das gnaze zu veranschaulichen; Goal: Regression/Prediction model; Outcome: Regression/Prediction model

4 Discussion

Search for papers mentioning certain genes found in the targets of the inhibitory drugs or

5 References

6 Appendix