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 Abstract

Breast cancer still presents a significant global health challenge, requiring improved treatment options with minimized side effects. Since traditional drug development is costly, time-consuming, and risky, hindering progress. Drug repurposing offers a potential solution by finding new applications for approved drugs, therefore reducing costs and development time. This research project focuses on a computational approach, utilizing genetic associations to identify target genes for drug development. By analysing treatment response and genetic data, including gene expression profiles, the research project aims to identify promising gene candidates as potential drug targets. The evaluation includes assessing gene knockout scores and predicting drug effectiveness at different treatment doses. The goal is to contribute to personalized and efficient treatment options for breast cancer patients. It was shown that associations between genes and treatments could be made even though with only slight potential for drug repurposing use. However, predicting treatment response by drug dose is able to accurately help in evaluating potential drug repurposing candidates. In general, this research project highlights an approach in finding new gene targets and drugs applications by analysing genetic data and evaluating the potential usefulness of such repurposed drugs.

2 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.

3 Materials and Methods

3.1 Data

3.1.1 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

3.1.2 Cellline Datasets

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

The copy number data frame contains the copy number of each gene for each cell line used in the PRISM screen. Each row of the data frame represents a specific cell line, while the columns correspond

to different genes. The row names and column names of the data frame are assigned using the respective cell line IDs and gene designations. The copy number of a gene refers to the number of copies present in the genome of an individual cell. Theoretically, each gene has a copy number of two, because one gene copy of each parent exists in the genome. However, mutations can lead to variations in the copy numbers, resulting in either higher or lower values than the expected two copies.

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) Lastly, a data frame containing the gene knockout scores for the cancer cell lines. This means that each gene has been silenced via CRISPR or another method and the resulting effect on cell proliferation was tested. A low gene knockout score indicates shows that the cell proliferation has been severly inhibited. Therefore the importance of genes can be estimated by how low the gene knockout score is.

3.2 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.

3.3 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.

4 Results

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

4.1 Gene search engine

Analysing the main datasets led to many individual data formats such as data frames, lists etc.; which contain relevant information gathered by our code. 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.

4.2 List of inhibitory drugs

To compile a list of inhibitory drugs, the first step involved filtering of the prism dataset. This dataset consists of 1396 distinct drugs that were tested on their effect on cell growth at concentrations ranging from 0.000194147 nM to 12.655 nM. Further Analysis showed that every drug was tested at $4 - 8$ different doses. In the prism data set, the lowest values, indicate the highest effectiveness. To identify inhibitory drugs, we calculated the average prism score for every drug. Then we specifically chose the concentration that exhibited the lowest average prism score for every drug. To further understand the impact of the doses on the prism score, we compared the concentrations associated with the lowest and highest average prism score for every drug.

PLOT 1 and PLOT 2

Once the most effective drug concentration combinations were identified, we proceeded with an additional round of filtering. This time only the drugs resulting in an average prism score lower than -2 were selected. This threshold was chosen because its allowing the removal of 60 % of the provided drugs.

This resulted in a data frame containing 582 inhibitory drugs (columns) and the respective prism scores on 481 cell lines (rows).

4.3 UMAP

4.4 Gene analysis

4.4.1 Correlation analysis

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. This was important to see if the number of gene copies can be an indicator for gene expression. 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. Around 15.23% of the correlations are negative. 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.

4.4.2 Statistical testing of important 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.

4.4.3 Dataframe for targets involving genes

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 expression score, mean prism 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.

4.5 Linear regression

Linear regression is a statistical technique to model the relationship between a dependent variable and one or more independent variables and fitting a straight line to the data. We used linear regression models to predict the average prism score based on the concentration for every drug. We created a function that takes drug names as inputs and returns a scatter plot with a linear fit on top and additional information about how well the model performs. This includes the R2 value, and a Shapiro Wilk normality test performed on the residuals. The R2 value is an indicator of how well the straight line fights the data point and if there is a linear relation between the points. The Shapiro Wilk normality test is commonly used to assess if the residuals follow a normal distribution. This is important to evaluate the performance because deviations from normal distribution suggest an underlying true pattern, that is not captured by the model.

Next we created a function which takes drug names and concentration values as an Input and returns predicted mean cell growth values for the concentration values.

To further confirm the relationship between concentration and the mean cell growth, we calculated the Pearson correlation for each drug.

5 Discussion

gene search engine: An informative gene search engine is a promising advancement in the field of genomic data analysis. Regarding large screenings, utilizing a specific profile as an input to the engine, offers accurate results, enabling other researchers to identify genes and their possible association to another treatment. Respectively, one could have got a possible important gene through literature analysis or comparing every meaningful value in our big database manually, but it is time-consuming and one could lose the overview. The demonstrated scalability and function of our engine does not display the current standard of such engines with a wide range of abilities, being GEMINI and Sigcom LINCS. While further improvements may be considered, our gene search engine occurred to be a usable tool for exploring gene-associated data for an overview and contributes to helping researchers to find their specific gene-treatment interaction.

The analyis of the different doses revealed a trend, Figure 1 shows that higher doses result in a lower average prism score. Figue ?? also validates this discovery, 79 % of the treatments have a high correlation between concentration and average prism score. This can also be seen when looking at the distribution of R^2 values of the linear regression models (Figure ???) that were computed for every drug. The majority of drugs have a R^2 value above 0.5 and higher, which indicates a decent fit of the regression line. The Shapiro Wilk normality test performed on the residuals of the linear regression models showed that the residuals are normally distributed for 72 % of drugs. This is important because

it indicates that the model captures the underlying pattern of the data quite well. However, there are some drugs that have a low R^2 value and a p-value below 0.05 for the Shapiro Wilk normality test. This indicates that not all patterns in the data are captured by the model, this could be improved with for example multiple linear regression taking in accound more factors or a completly different approach.

The Pearson correlation analysis of the gene expression and the copy number validated our hypothesis partially. A trend of high positive correlation can be seen on the Fig. XX as shown in the results. This indicates that if there are more copies of a gene it is likely that the gene expression is upregulated for this gene as well. But as the graph shows, this is not always the case as 15,23% of the correlations are negative. This might be the case since gene regulation isn't as simple as only the gene copy. Other regulatory units that are on another part of the genome could be involved. It could also be that after some point the higher the copy number gets, the less effect it has on gene expression. Further research across cancer cell lines shows similar results (Shao *et al.*, 2019). They conclude that there is a mostly positive correlation between copy number and gene expression as well.

Correlation analysis of breast cancer concluded one gene that is particularly interesting, SDHC. The correlation score of SDHC is -0.751 which means that a high gene expression indicates low treatment score and therefore a good treatment response. The Wilcoxon-rank-sum test showed that the SDHC gene has a significantly lower mean knockout score compared to the other lineages. This suggest that SDHC has particular importance in the breast cancer cell lines in our research project. The gene SDHC codes for the succinate dehydrogenase complex subunit c, a subunit of the succinate dehydrogenase important for anchoring and stabilization. The succinate dehydrogenase is part of the respiratory chain in mitochondria. Potential risks of mutations are a disbalance in the respiratory chain and oxidative stress (Cerqua *et al.*, 2021). SDHC has been linked in the correlation analysis to only one treatment, known as tegafur. This drug targets the thymidylate synthase (Lee *et al.*, 2015). Using the prediction model from this research project shows that tegafur response in breast cancer has a high correlation with the dosage. The r-squared value is 0.672 and the residuals are normally distributed, which indicate that the model is a good fit. Literature research have shown that tegafur already has been used on some cases in breast cancer with positive results. Yet, it is unclear how effective the treatment is and if it opens a new potential for targeted application if a high gene expression in patients can be detected. As it has already been tested if tegafur has a potential effect in breast cancer, the objective of finding novel applications for treatments has not been met.

Our final dataframe revealed results regarding the importance of mTOR in breast cancer cell lines. A mean importance value of -1.268 emphasizes that the knock-out of mTOR could play a crucial role in breast cancer development. This matches with existing literature, which has shown that dysregulation of mTOR in the PAM pathway is associated with increased tumor growth in breast cancer (Zhu *et al.*, 2022). The mechanistic target of rapamycin (mTOR) encodes a protein kinase, functioning as a central regulator of essential cellular pathways involved in growth, proliferation and survival (Zhu *et al.*, 2022). The regulatory function of mTOR makes it a possible cause for various cancer types when mutated and consequently an attractive target for therapeutic interventions. Moreover, we obtained a mean expression level value of 4.736, which indicates a high expression as the average of all genes are 2.729. This supports the previous literature findings, which suggest that mTOR is commonly overexpressed

in breast cancer, further stressing its potential oncogenic involvement (Zhu *et al.*, 2022). In addition to that, we examined the copy number variation of the mTOR gene in breast cancer cell lines and the mean value of 0.912 could display genomic instability due to partial deletion of the gene as the copy number is lower than two. This discovery would further promote the role of mTOR in cancer development. Including the hypothesis, we made out the three values, possible treatments should be able to downregulate the function of the mTOR gene. In fact, the associated treatments presented in the final dataframe are mTOR inhibitors. Finally, our results reinforce the significance of mTOR in breast cancer. However, targeting mTOR in breast cancer should be investigated to develop safe treatment strategies. Nevertheless, our study about mTOR is similar to the groundwork findings, consequently suited for future research.

The basis of our gene analysis is the correlation analysis of treatment response and gene expression. The idea was that high absolute correlation could show genes that are associated with breast cancer. This means that a particular gene could be linked to a treatment. However, this does not imply causality meaning that the treatment could possibly not be affected directly by the treatment. The next selection process was on the importance of the genes via gene knockout score. The findings conclude genes and treatments that have been associated with cancer or further breast cancer but some only slightly. Novel opportunities for drug repurposing have not been found. In hindsight, this approach may not be optimal to further decrease the number of potential targets. As stated before, correlation of gene expression does not imply the direct causality between the treatment and the gene expression. So, looking at the importance of a gene via gene knockout score does not show that a gene could be a good target. With more time to analyse the data, we would want to test new ways in looking into the data and selecting genes. Possible approaches could be as following.

6 References

7 Appendix