Machine Learning in Genomics

Biomarker discovery in Triple-Negative Breast Cancer

Team 2

Nathan Bigaud (IASD)

Bhargob Kakoty (MASH)

Anna Kanyuka (IMaLis)

Anna Kurowska (IMaLis)

Salome Papereux (IASD)

Matthieu Rolland (IASD)

08/04/2022

Introduction

Triple negative breast cancer (TNBC) is an aggressive form of breast cancer that lacks three receptors that are commonly found in breast cancer and are targeted with chemotherapy. It is also a highly heterogeneous form of cancer, with up to six identified subtypes [Kothari et al. 2020]. However, we observe differences in responses to chemotherapy among TNBC patients, which is the focus of our project.

In this project, we are given information on the response of a cohort of patients to three separate medications and we aim to identify biomarkers in gene expression data that could help predict a patient's response to these chemotherapy medications - here doxorubicine, cyclophosphamide, and paclitaxel. After two research studies on 450 and 130 people respectively, the overexpressing ABCD1 gene was identified as a potential biomarker. The data consists of bulk gene expressions, the treatment name and a binary variable indicating whether or not the patient responded to the treatment.

Our mission is to find biomarkers predictive of good treatment response among patients overexpressing the ABCD1 gene using machine learning (ML) techniques. The ML challenge stems from the shape of the data, which is high-dimensional with a limited number of samples (<800 patients for perhaps >10,,000 genes) - abbreviated below to HpLn for *High p Low n*. Our goal is to identify the subset of features in the data that allow a model capable of reliably predicting reaction to medication, modeled as a binary variable. We first look at how to use the data provided to the best of our ability (Plan A), and then dig into supplementary data sources (Plan B).

Plan A - Initial approach of the available gene expression data

Our first focus would be on making the most of the available data, assuming further data collection is a costly process. Our understanding of a biomarker in this case is the following; a gene whose expression is consistently different between ABCD1-overexpressing chemotherapy responsive and non-responsive groups.

1 - Data exploration and preparation

We would start by getting familiar with the data and performing a set of sense checks in order to verify the information given to us. This would include (i) investigating the data collection process to spot potential biases - for instance looking at sampling biases on ethnicities or age ; (ii) exploring the data's characteristics such as sparsity, presence of outliers, and visualization using dimensionality reduction techniques, to try and get modeling insights ; (iii) perform a rapid set of statistical tests (e.g. DGE) on the whole dataset to confirm the results of ABCD1 overexpression in well-responding patients.

After this step, we would focus on data preparation. We would start by isolating samples who overexpressed ABCD1 gene and prepare split into test/train/val splits using stratified sampling, bootstrapping or cross validation depending on model type.

An important choice to make is whether to treat the three drugs separately. In an ideal world, we would want to break the data into three datasets, as each drug has a different mode of action [Abu Samaan et al 2019]. We could then even use a multitask learning approach to jointly learn from those three related learning tasks. However, the sample size is very small, and is further shrunk down by our exclusive focus on patients overexpressing ABCD1, and we could end up working with very few samples per dataset. So as a first step, we would favor treating the three drugs together.

2 - First Model Approach

A first simple way to approach the problem is to perform univariate tests - e.g., Mann-Whitney U test - for each feature to find out statistically significant differences between the two groups. This approach is however limited, since it considers each feature in isolation. A good solution is to use more complex ML methods to select the most significant feature selection while training the model.

Model selection needs to prioritize interpretability and account for the HpLn setting. We are trying to predict a categorical variable in a biologically interpretable way. This means that in the trade-off between simple, interpretable models with moderate prediction accuracy (e.g. logistic regression) vs highly accurate but hardly interpretable models (e.g. DNN), we should favor the former. We are working with a low sample size, high-dimensional dataset. This implies avoiding models with a large number of parameters to avoid overfitting.

Linear regression models share a strong power of interpretability. Given that we have a binary response, a regularized version of logistic regression seems a perfect fit. When it comes to the choice of regularization, using the Lasso penalty helps obtain a sparse set of features, but we would favor an Elastic Net approach to select correlated features in a group. The regularization parameters can be found using a grid search or random search algorithm.

The expected result of the regression would enable us to rate feature importance based on the magnitude of the coefficients of the selected features and extract the top candidate biomarkers. Biomarkers' quality could be confirmed using the p-value or test statistic of a standard statistical test as mentioned above. As an overall predictive performance of the logistic regression model (and, also to facilitate comparison with other ML methods) we can use standard accuracy measures linclusing recall, sensitivity, F1, and AUROC.

3 - Other approaches available

While the linear regularized approach seen in class provides a scalable, interpretable model adapted to HpLn data, it cannot capture non-linear relationships between the features and the target variable. We could explore non-linear methods specifically adapted to feature selection and HpLn data.

A - Kernel methods

Our first focus would be on kernalized methods, as the most widely studied non-linear approach. Kernels are widely used in the GWAS community, in particular the Sequence Kernel Association Test (SKAT) [Wu et al. 2011]. For our problem, we would focus on generalizations of this approach, as outlined in [Azencott 2020]. Those use mutual information or the Hilbert-Schmidt Independence Criterion (HSIC) as association measures and require to be tweaked to avoid redundancy for our purpose.

We will focus on a SOTA kernel method, Block HSIC Lasso, proposed by [Climente-González et al. 2019]. Acknowledging the difficulty of estimating mutual information [Walter-Williams & Li, 2009], it focuses on Lasso HSIC [Yamada et al. 2014] and decreases the memory complexity by dealing with the data block-by-block. Under the hood, the algorithm optimizes a three part loss, focusing on (i) selecting features that are highly dependent on the phenotype, (i) penalizing selecting mutually dependent features, and (iii) Lasso regularization. It is worth noting that given the small sample size, we might even be able to use the original HSCI Lasso method.

Key issues in applying those methodology would be kernel selection and the interpretability of the feature selection process. Following [Climente-González et al. 2019], we would use the normalized Delta kernel as we are predicting a binary variable in a setting that closely fits the original paper's setting. Exploring further alternative kernel choices would require experimenting on our dataset, and perhaps looking more closely at more recent work described by [Azencott 2020], looking at kernel selection methods. Overall, we would lose part of our ability to explain the 'rationale' behind the feature selection process, due to the non-linear nature of the method.

B - Other non-linear approaches

As a second priority, we could explore adaptation of other common algorithms to the HpLn feature selection problem. In particular, we found two methods we would want to explore : RFE with random forest and FsNEt.

The Recursive Feature Elimination (RFE) algorithm searches for a subset of features by starting with all features in the training dataset and removing features until the desired number remains. In our case, we would use the Recursive Feature Elimination with Random Forest (RFE-RF) [Thalor et al. 2022], which applied to comparable data performs well and prevents overfitting. Random Forest algorithm is here used to help choosing features. RFE-RF allows elimination of the dataset's extraneous and redundant feature variables, highlighting promising biomarkers. The quality of predicting biomarkers would further be measured with the accuracy, precision-recall, and F1 score.

Finally, we would explore adaptations of deep learning methods to the HpLn setting, taking advantage of the ability of neural nets to discover complex nonlinear patterns. For instance, [Singh et al. 2020] introduced neural networks for features selection, through the neural network FsNet. FsNet addresses overfitting with two main approaches: it involves a selection layer that uses concrete random variables, and uses two tiny networks to predict the large weight matrices of the selection and reconstruction layers, this last one allowing to avoid overfitting. This DNN has proven to be robust on several high-dimensional biological datasets, and compares favorably to previous attempts at adapting DNN to the HpLn setting (e.g. Concrete Autoencoders). For our dataset, this method would come at the very end of the analysis, to confirm results from other non-linear classifiers and out of general curiosity.

4 - Functional analysis

After retrieving the list of candidate biomarkers, we will focus on investigating the biological properties of our biomarkers. We will perform a co-expression network analysis on the expression values of all genes in order to compute whether any two genes are connected in a network. We are interested in gene networks because genes with similar expression patterns are likely involved in common biological processes, regulated together or have similar functions [Nisar et al., 2021].

First of all, we will choose an appropriate co-expression measure and selection threshold method - for instance Pearson's correlation coefficient and weighted gene co-expression networks method - in order to create a graph displaying the gene networks. Subsequently, we will complete a gene set (each network corresponds to a gene set) enrichment analysis on expression values of these biomarkers. Our goal is to identify pathways that have relevance to the TNBC, and select them for further analysis. We will use the KEGG pathway database in order to identify a list of pathways that are significantly (p-value <0.05) over-represented in the gene sets.

Potential pathways of interest could be related to cell proliferation, apoptosis and other pathways that are known to be involved in cancer physiology and development [the Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium, 2015]. Once relevant pathways are identified, we will select the biomarkers that are found in such pathways for further analysis.

Plan B - Further data augmentation, cross-validation and collection

One we have made the most out of the available data, we will focus on three types of additional data, by increasing order of costliness. We will first look at how to augment the data we currently have with basic patient data and existing markers of TNBC cancer types, then at public omics dataset for cross-validation, and finally additional experiments to be performed on our patients' data.

1 - Data Augmentation

In order to improve our models, we will request more information on the patient data from the research group. Information that will be helpful to know include a detailed questionnaire about the patients and the treatment response. The questionnaire will allow us to learn the demographic characteristics, medical history, lifestyle characteristics, and risk factors for cancer and other chronic diseases. This information will be useful in data integration as well as the conduction of our own experiments.

Also, we will ask the researchers if they have any information on the specific TNBC subtype for each patient. Since there are multiple subtypes of TNBC, it will be interesting to know whether the biomarker ABCD1, as well as any other potential biomarkers that may be discovered, are overexpressed in all or only some subtypes of TNBC. If the biomarker is overexpressed in only some subtypes, this will mean that it is not a universal biomarker of TNBC, and this is interesting. Alternatively, we could use existing research and our gene expression data to infer the subtypes, based on known biomarkers - see for instance [Bissanum et al., 2021]. This is however a more labor-intensive and less precise approach.

More extensive background information on patient characteristics could be used in two distinct ways. First, reassess the risk of sampling biases by looking at summary statistics, and if relevant applying the most suitable bias correction technique - see for instance [Cortes et al. 2008]. Second, adding these as features for inference, with the proper normalization depending on the model - being particularly aware of the risk that some indicators, such as age, may 'overshadow' the inference power of our biomarkers.

2 - Cross-validation

The next step in our analysis is to assess the quality of the biomarkers identified by cross-validating our results with publicly available transcriptomic datasets. In order to render the cross-validation informative, the chosen datasets need to share as many similarities with our expression dataset as possible. Given the triple mutation characteristic of TNBC cancer, it is crucial that the comparison expression data used comes from experiments on TNBC cell lines or from TNBC patients undergoing chemotherapy. It is essential that the datasets contain expression measurements for ABCD1 and that for all samples this gene is overexpressed. Ideally, all samples should derive from the treatment group that was treated with one of the three drugs present in our analysis. In practice, we will look for data from experiments testing similar drug classes. Just as it was done in the analysis of our expression data, the validation will compare ABCD1-positive chemotherapy responsive to non-responsive samples.

We will compare our data to expression data collected with microarray and RNA-seq technologies in order to analyze as many relevant datasets as possible, given the specific requirements of our data. We will download raw expression datasets from Gene Expression Omnibus (GEO) and The Cancer Gene Atlas (TGCA). Prior to the analysis, we will pre-process gene expression data according to the technology it was measured with. For each analyzed dataset, we need to verify the quality, as well as, perform background correction and normalization steps. It is important to state, comparison can only be made between genes that are included in all datasets.

Regarding data acquired from microarrays, if it was generated with the same chip as our data (Affymetrix U133A Plus 2.0 array), it can be processed according to the protocol used in our analysis and be compared to our data directly. Data acquired with different microarray chips or RNA-seq will require different quality control, background correction and normalization methods [Jaksik et al, 2018; Evans et al., 2018]. If genes have different amounts of corresponding probes between microarray chips, expression for each gene should be averaged from the corresponding probes [Pozhitkov et al., 2014].

We will perform the cross-validation in two ways. First we will repeat our analysis from section A on publically available datasets in order to verify whether using our method we obtain the same biomarkers from other datasets. Secondly, we will compare expression profiles of the biomarkers identified in Plan A between treatment responsive and non-responsive conditions. We want to observe whether our biomarker's expression has the same profile across the two conditions in other datasets, which would support the biomarkers' usefulness in predicting response to the chemotherapy.

The reliability of the biomarkers from section A analysis will increase with 1) ML analysis of public datasets identifying the same biomarkers and 2) the expression profiles of biomarkers (between chemotherapy responsive and non-responsive ABCD1-expressing samples) being similar across experiments. The level of the expected result similarity will depend on, and be adjusted according to the characteristics of the comparison datasets (e.g. Is expression data available for all identified biomarkers? Do the same drugs are used?). Genes that are the least identified as biomarkers or whose expression profile is not similar across most of the datasets will be removed from further analysis. The exact cut-off value and criteria of rejection will only be decided upon once the public datasets are retrieved.

3 - Further experiments

In order to further corroborate the validity, and understand the genetics and biological significance of the biomarkers of interest, we need to conduct additional experiments on the remaining genetic material from the original cohort of 450 patients. Since this material was used for a microarray study, we make an assumption that the remaining genetic material comprises both DNA and RNA in bulk. We assume that we also have data on protein purification from our 450 patients.

A multi-omics approach could help us consolidate our existing knowledge on biomarkers. Specifically, genomics and proteomics will be useful.

- **We will start with genomics**: using the genetic material of the original cohort of 450 patients, we can sequence it, to get the genotype of each patient. We can then look for a difference in genotypes between the two groups expressing our known biomarker: group that is responsive to treatment (80%) and group that is not responsive to treatment (20%). We will hope to find a difference in genotypes between the two groups, assuming that even one differing SNP can cause this difference. If there is a difference in genotypes between the two groups and there is one or multiple SNPs, this genotype will code for a difference in phenotypes as well (which is expressed as different responses to treatments).

- **Following this logic, we can validate our hypothesis with proteomics**. We hypothesize that this SNP leads to a different structure of protein in question, which means that the drug molecule can no longer bind it (lock and key model) and this results in non-responsiveness to treatment in the 20% of patients expressing the biomarker. Alternatively, this SNP can also lead to differences in protein abundance between the two groups which can similarly affect the responsiveness of the two groups to chemotherapy (in the 20% non-responsive group there are not enough proteins for the drug molecules to bind). Finally, it would also be useful to cross check our data with databases, as the data needs to meet similar conditions that are described in the section above.

Finally, while this may be outside of the scope of our analysis, the usefulness of the selected biomarkers in predicting the response to chemotherapy must be verified in large clinical studies. This is needed on account of the commonly reported discrepancy between usefulness of biomarkers in small research studies and clinical runs [Strimbu and Tavel, 2010]. This could be done in large clinical tests with TNBC ABCD1-positive patients where the expression of the selected biomarkers would be measured during the duration of the treatment with doxorubicine, cyclophosphamide, and paclitaxel separately. Testing the drugs separately is important in order to select the most suitable biomarker to each chemotherapy type. Hopefully, a proportion of the biomarkers identified from our analysis will show similar expression patterns that are predictive of chemotherapy success. Finally, repeating all the analyses on single cell omic data could allow us to acquire a more accurate understanding of the nature of, and hence assess the quality of the selected biomarkers.

Bibliography

Abu Samaan, T. M., Samec, M., Liskova, A., Kubatka, P., & Büsselberg, D. (2019). Paclitaxel's mechanistic and clinical effects on breast cancer. *Biomolecules*, *9*(12), 789.

Azencott, C. A. (2020). Machine learning tools for biomarker discovery (Doctoral dissertation, Sorbonne Université, UPMC).

Bissanum, R., Chaichulee, S., Kamolphiwong, R., Navakanitworakul, R., & Kanokwiroon, K. (2021). Molecular Classification Models for Triple Negative Breast Cancer Subtype Using Machine Learning. *Journal of personalized medicine*, *11*(9), 881.

Climente-González, H., Azencott, C. A., Kaski, S., & Yamada, M. (2019). Block HSIC Lasso: model-free biomarker detection for ultra-high dimensional data. Bioinformatics, 35(14), i427-i435.

Cortes, C., Mohri, M., Riley, M., & Rostamizadeh, A. (2008, October). Sample selection bias correction theory. In *International conference on algorithmic learning theory* (pp. 38-53). Springer, Berlin, Heidelberg.

Evans, C., Hardin, J., & Stoebel, D. M. (2017). Selecting between-sample RNA-seq normalization methods from the perspective of their assumptions. *Briefings in Bioinformatics*, *19*(5), 776–792. https://doi.org/10.1093/bib/bbx008

Jaksik, R., Iwanaszko, M., Rzeszowska-Wolny, J., & Kimmel, M. (2015). Microarray experiments and factors which affect their reliability. *Biology Direct*, *10*(1). https://doi.org/10.1186/s13062-015-0077-2

Kothari, C., Osseni, M. A., Agbo, L., Ouellette, G., Déraspe, M., Laviolette, F., ... & Durocher, F. (2020). Machine learning analysis identifies genes differentiating triple negative breast cancers. Scientific reports, 10(1), 1-15.

Nisar, M., Paracha, R. Z., Arshad, I., Adil, S., Zeb, S., Hanif, R., Rafiq, M., & Hussain, Z. (2021). Integrated analysis of Microarray and RNA-seq data for the identification of hub genes and networks involved in the pancreatic cancer. *Frontiers in Genetics*, 12. https://doi.org/10.3389/fgene.2021.663787

Pozhitkov, A. E., Noble, P. A., Bryk, J., & Tautz, D. (2014). A revised design for microarray experiments to account for experimental noise and uncertainty of probe response. *PLoS ONE*, *9*(3). https://doi.org/10.1371/journal.pone.0091295

Singh, D., Climente-González, H., Petrovich, M., Kawakami, E., & Yamada, M. (2020). Fsnet: Feature selection network on high-dimensional biological data. *arXiv preprint arXiv:2001.08322*.

Strimbu, K., & Tavel, J. A. (2010). What are biomarkers? *Current Opinion in HIV and AIDS*, *5*(6), 463–466. https://doi.org/10.1097/coh.0b013e32833ed177

Thalor, A., Joon, H. K., Singh, G., Roy, S., & Gupta, D. (2022). Machine learning assisted analysis of breast cancer gene expression profiles reveals novel potential prognostic biomarkers for triple-negative breast cancer. *Computational and Structural Biotechnology Journal*.

The Mutation Consequences and Pathway Analysis working group of the International Cancer Genome Consortium. (2015). Pathway and network analysis of cancer genomes. *Nature Methods*, *12*(7), 615–621. https://doi.org/10.1038/nmeth.3440

Walters-Williams, J., & Li, Y. (2009, July). Estimation of mutual information: A survey. In International Conference on Rough Sets and Knowledge Technology (pp. 389-396). Springer, Berlin, Heidelberg.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. The American Journal of Human Genetics, 89(1), 82-93.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., & Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. Neural computation, 26(1), 185-207