Learning Signature Exposures from Gene Expression at Single-Cell Resolution: Regular vs. Multitask Learning of Individual Regression Models

Ariel Potolski Eilat (A.PotolskiEilat@student.tudelft.nl) *Ivan Stresec, Sara Costa (Supervisors)* *Dr. Joana Gonçalves (Responsible Professor)*

01. Introduction

- **Mutational Signature**: Patterns of mutation in the RNA of single-cells
  - Can indicate the cause of the tumor (e.g. smoking)

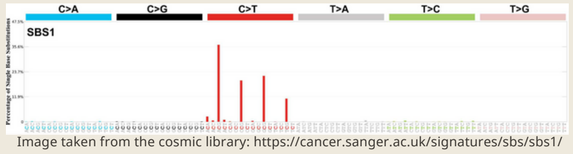


Image taken from the cosmic library: https://cancer.sanger.ac.uk/signatures/sbs/sbs1/

- **Gene Expression:** level of activity of a gene in a cell
- Previous work done in the area [2]:
  - It was found that the activity of certain mutational processes are associated with changes in gene expression.
  - Bulk data
  - Classification problem - presence or absence of mutational signatures using the gene expression data

02. Research Question

Are mutational signature exposures of single-cell data predictable from the cells' gene expression?
Sub-questions:

1. How does multitask learning compare to regular regression models for predicting mutational signatures from gene expression?
2. How well do these models predict mutational signature exposures when applied to unseen gene expression profile data?

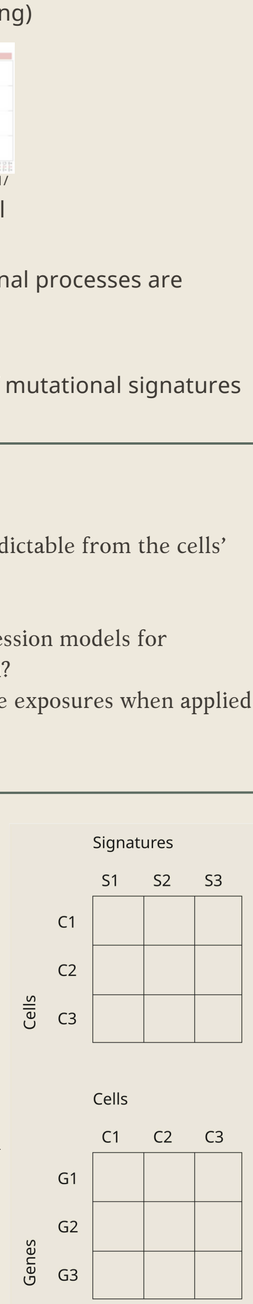
03. Methodology

Data

- 688 cells from one breast-cancer
- Mutational signature exposure matrix: number of mutations caused by a specific signature in a specific cell
- Gene expression matrix: number of times a gene is expressed in a cell

Preprocessing

- Filtered out signatures that have zero exposure values.
- Split data into train, validation, and test sets
  - Experiment 1: random split of the data with a percentage of 70, 15, and 15, respectively
  - Experiment 2: cluster-based split (PCA + k-means)
    - Distribution shift in the data



    - Clusters assigned to the sets, trying to maintain a 70%/15%/15% split.
- Applied normalisation using CPM + log1p [3] and standardisation to gene expression matrix

Models selected

- RidgeCV: Solve individually for each signature
  - Different set of genes, different regularisation parameters
- MultiTaskLassoCV: Multitask solution
  - Same set of genes, same regularisation parameter

Metrics chosen

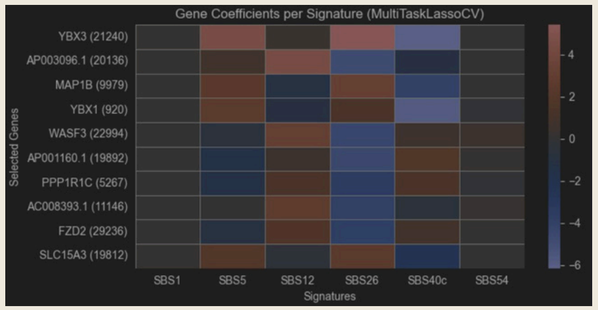
- R^2 and MSE

04. Results

Random Split

Model	R^2	MSE	$\frac{MSE}{Variance}$
RidgeCV	0.35	1155.77	0.171
MultiTaskLassoCV	0.32	1234.42	0.182

- RidgeCV's performance is slightly better, but not as much as expected.
  - Similar biological pathways between signatures, highly correlated genes, and/or high sparsity of single-cell data.
- MultiTaskLassoCV selected a set of 292 genes



Signature	Jaccard Coefficient	Overlapping Genes
SBS1	0.053	AP003096.1
SBS5	0.000	—
SBS12	0.111	AC008393.1, AP003096.1
SBS26	0.053	WASF3 (22994)
SBS40c	0.000	—
SBS54	0.000	—

- Shared genes might be involved in shared biological mechanisms and may be involved in core pathways
- Genes selected only by RidgeCV highlight the added interpretability of learning per-signature models.
  - May reflect signature specific regulatory mechanisms

Cluster-based split

- Silhouette value: 0.3045

Model	R^2	MSE	$\frac{MSE}{Variance}$
RidgeCV	-0.20	1561.60	0.601
MultiTaskLassoCV	-0.23	1635.07	0.630

- Significant drop in performance shows that generalisation is hard.
- Highly influenced by the clustering and data split
- Highlights the importance of training and evaluating models on more diverse cell populations before clinical use.
- Models are not powerful enough to learn signature exposures based on a non-representative sample

05. Conclusions

- Regular approach better reveals potential signature-specific influences.
- Multitask approach might be useful to find common underlying pathways, and when seeking a sparse set of predictors shared across mutational processes
- The significant drop in performance when applied to unseen data highlights the challenges of deploying these models in clinical settings.
- Signature exposures are hard to learn from a non-representative sample

06. Future Work

- Reconstruct mutational catalogues from predicted exposures
- Expand research to use more diverse and representative datasets
- Run experiments with nonlinear models
- Gene enrichment analysis

07. References

[1] L. B. Alexandrov, S. Nik-Zainal, D. C. Wedge, P. J. Campbell, and M. R. Stratton, "Deciphering Signatures of Mutational Processes Operative in Human Cancer," in Cell Reports, vol. 3, Elsevier Inc., 2013, pp. 246–259. [Online]. Available: https://www.cell.com/cell-reports/fulltext/S2211-1247(12)00433-0?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS2211124712004330%3Fshowall%3Dtrue

[2] L. Jiang, H. Yu, and Y. Guo, "Modeling the relationship between gene expression and mutational signature," in Quantitative Biology, vol. 11, 2023, pp. 31–43. Accessed: Apr. 22, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/epdf/10.15302/J-QB-022-0309

[3] https://www.sc-best-practices.org/preprocessing_visualization/normalization.html