UNIVERSITY OF PADUA

PREDICTIVE MODELING AND NETWORK ANALYSIS OF GENE EXPRESSION IN BREAST CANCER

Eleonora Mesaglio
Margherita Rigato
Francesco Lollato

TABLE OF CONTENTS

Overview

TOPIC

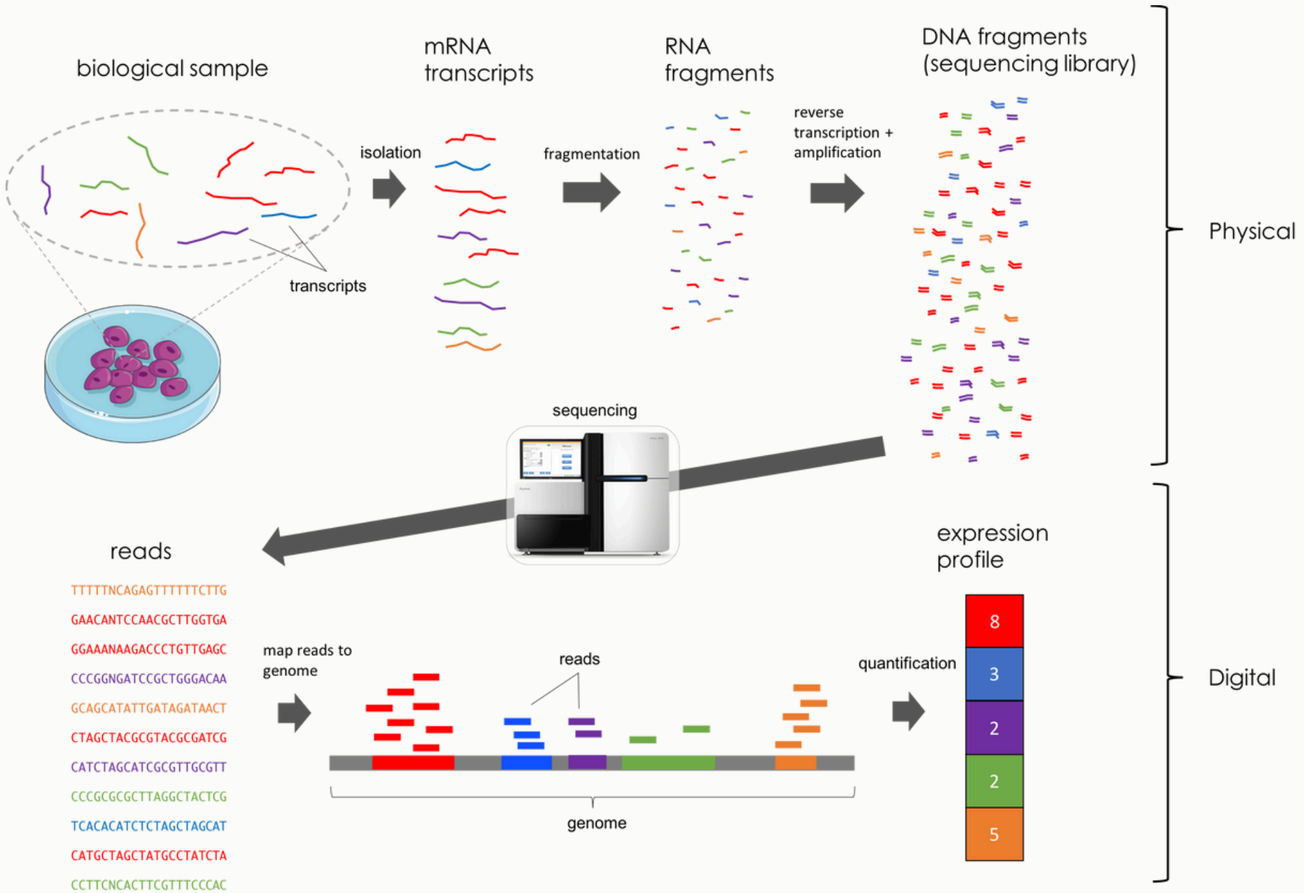
- Analysis of a gene expression dataset
- Data collected from **breast** tissues

GOALS

- **Classification of tissue type** (normal vs cancer) using gene expression values

- **Network analysis**: study of the relations between genes' levels of expression in normal tissues vs cancer tissues

Gene Expression Data Collection



More on Gene Expression

WHAT DO WE KNOW FROM GENE EXPRESSION LEVEL?

It quantifies the **level of activity** of a gene.

ALTERATION (wrt normal) IN GENE EXPRESSION LEVELS:

- **Warning sign** that something is not going the way it should
- Can be the **starting point** in the developement of a disease.
- Can lead to MALFUNCTIONS

In-depth knowledge of how gene expression changes from normal tissues to cancer ones is very important for:

- **Early diagnosis** (thanks to the discovery of new biomarkers)
- more efficient treatments

Dataset

Gene / Sample_id	gene A	gene B	gene C	...
sample 1				
sample 2				
...				
sample 289	...			

Expression level of gene A in sample 289

SOURCE: Gene Expression Omnibus (GEO) platform

DIMENSION:

289 samples

- 143 samples of breast adenocarcinoma (cancer)
- 146 samples of normal breast tissue (normal)

(BALANCED DATASET)

35981

genes considered (columns)

Preprocessing

- **MISSING VALUES DETECTION**: none.
- **QUANTILE NORMALIZATION**: we found in the documentation that it was **already done**. It is necessary to **make comparisons between samples**.

SPLITTING OF THE DATA INTO TRAINING AND TEST SETS

- Training set: 75% of the original dataset, still balanced.
  217 statistical units
- Test set: 25% of the original dataset
  72 statistical units

Can we directly apply Logistic regression? No

In the high-dimensional setting, in which the **number of features p is larger than the sample size**, Logistic regression cannot be used without modification. When p > N, any linear model is **over-parametrized**, and **regularization is needed** to achieve a stable fit.

- Hastie T., Tibshirani R., Wainwright M. (2015) Statistical Learning with Sparsity The Lasso and Generalizations, CRC Press Taylor & Francis Group

Logistic Regression with LASSO penalty

Model

Logistic Lasso: combine logistic regression with a L1 penalization.

Parameter λ:
We applied **cross validation** on the training set to obtain the best value of λ.
Best λ: 0.01217137

Performance

	TN	TP
PN	32	8
PP	0	32

- #selected variables : **72**

- Test error: **11.11%**
- Sensitivity: **80%**
- Specificity: **100%**

L1-regularized linear SVM with hinge loss

Model

- SVM: aims to find the optimal decision boundary that separates classes in the feature space, by **maximizing the margin.**

$$\min_{\beta_0,\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} [1 - y_i f(x_i; \beta_0, \beta)]_+ + \lambda \|\beta\|_1 \right\}$$

penalization for the hinge loss

hinge loss

Parameter λ:

- We applied **cross validation** on the training set to obtain the best value of λ.
- Best λ: 0.1124

Performance

	TN	TP
PN	32	7
PP	0	33

- #selected variables : **117**

- Test error: **9.72%**
- Sensitivity: **82.5%**
- Specificity: **100%**

L1-regularized linear SVM with squared hinge loss

Model

- squared hinge loss: in contrast to hinge loss, it is differentiable everywhere.

$$\min_{\beta_0,\beta}\left\{\frac{1}{n}\sum_{i=1}^{n}[1-y_i f(x_i;\beta_0,\beta)]_+^2+\lambda\|\beta\|_1\right\}$$

penalization for the squared hinge loss

squared hinge loss

Parameter λ:

- We applied **cross validation** on the training set to obtain the best value of λ.
- Best λ: 0.0586

Performance

	TN	TP
PN	32	8
PP	0	32

- #selected variables : **84**

- Test error: **11.11%**
- Sensitivity: **80%**
- Specificity: **100%**

Lasso-Least Squares Hybrid Estimator

LASSO:

- very effective for variable selection
- it is a biased estimator

Another approach: **Lasso-Least Squares Hybrid Estimator**

- variables selected by Lasso were used in a Logistic Regression model
- goal: obtain unbiased coefficient estimates

Problems faced:

- strong collinearity among LASSO-selected variables, preventing algorithm convergence

Elastic Net Logistic Regression

Model

Why Elastic Net?

- LASSO struggles with correlated predictors

Elastic Net:

- combines L1 and L2 penalties
- allows selection of correlated variables together
- balances regularization and feature selection

Parameters:

- α: balances Lasso and Ridge penalties. cv: α=0.744
- λ: controls penalty strength. cv: λ=0.0112

Performance

	TN	TP
PN	32	7
PP	0	33

- #selected variables: **104**

- Test error: **9,72%**
- Sensitivity: **82,5%**
- Specificity: **100%**

Random Forest

Model

Random Forest:

- ensemble method combining multiple decision trees
- non-linear model

Parameters:

- grid search: *mtry, num.trees, min.node.size*
- each combination evaluated using Out-Of-Bag (OOB) error
- *mtry*=180, *num.trees*=400, *min.node.size*=10

Performance

	TN	TP
PN	31	5
PP	1	35

- Test error: **8.33%**
- Sensitivity: **87,5%**
- Specificity: **96,88%**

Models comparison

Accuracy:

- RF is more accurate than the other models
- all models are competitive

Sensitivity and Specificity:

- RF excelles at identifying positive cases (highest sensitivity)
- RF has slightly lower specificity (more false positives)
- linear models (SVM, Elastic Net) are slightly lower in sensitivity
- linear models show perfect specificity (no false positives)

Observations for cancer detection:

- maximizing sensitivity is critical to avoid treatment delays due to false negatives
- RF is the most suitable choice, balancing strong sensitivity and good specificity
- linear models may be preferred if avoiding false positives is the highest priority

	Test Error	Sensitivity	Specificity
Lasso SVM hinge loss	9.72%	82.5%	100%
Logistic Reg. Elastic Net	9.72%	82.5%	100%
Random Forest	8.33%	87.5%	96.88%

Selected Genes

Which genes have the greatest influence on the presence or absence of cancer?

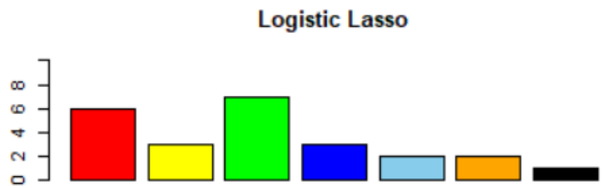
We selected the top **20 most important genes** for each model:

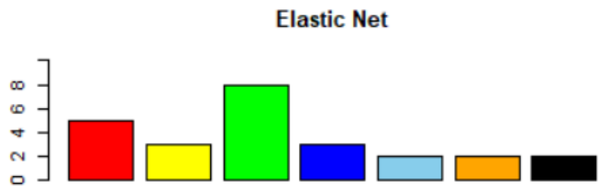
- coefficients with strongest magnitude

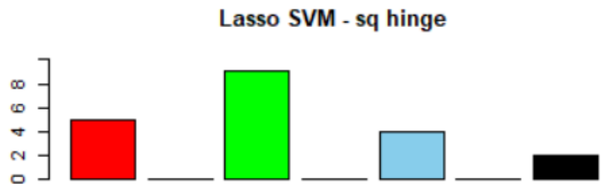
Common genes:

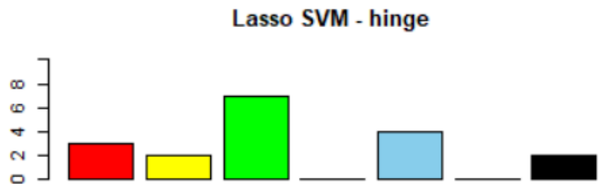
- lincRNA.chr9.34741650.34747975_F
- lincRNA.chr12.76652565.76706444_F
- lincRNA.chr2.102597068.102606493_R
- NM_001001971
- A_33_P3296313
- A_33_P3355694

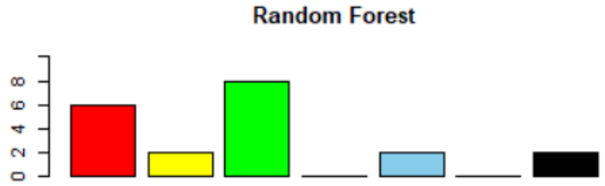
Biological Functions

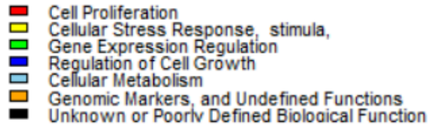








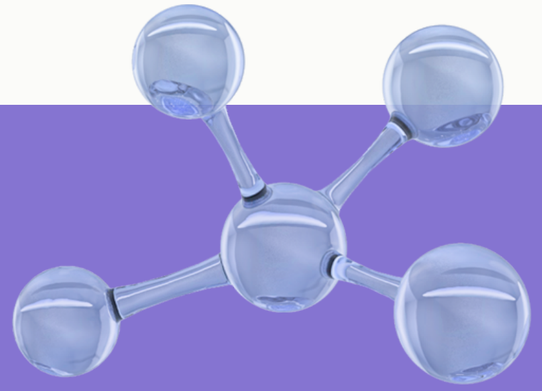




Predominance of

- **Gene Expression Regulation** (green)
- **Cellular Proliferation** (red)
- **Cellular Metabolism** (light blue)

Correlation Analysis



Goals

We analyzed the correlations between genes to:

- Find **patterns** in cancer and normal conditions
- Highlight **differences** in gene interactions in cancer and normal conditions

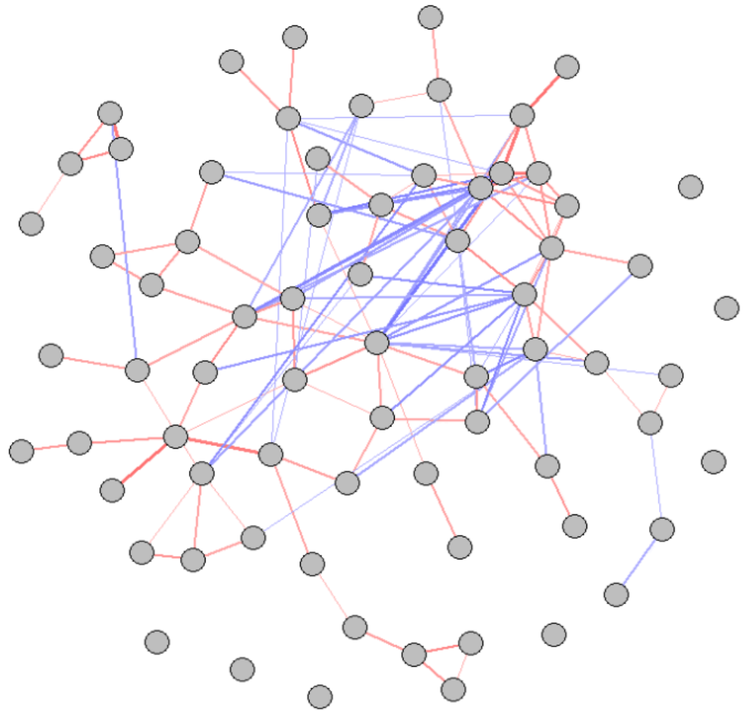
Preprocessing

- High dimensionality -> work just with the **72 variables** selected by Logistic Lasso
- **Split** the training set: **cancer vs normal** patients

Correlation Graphs

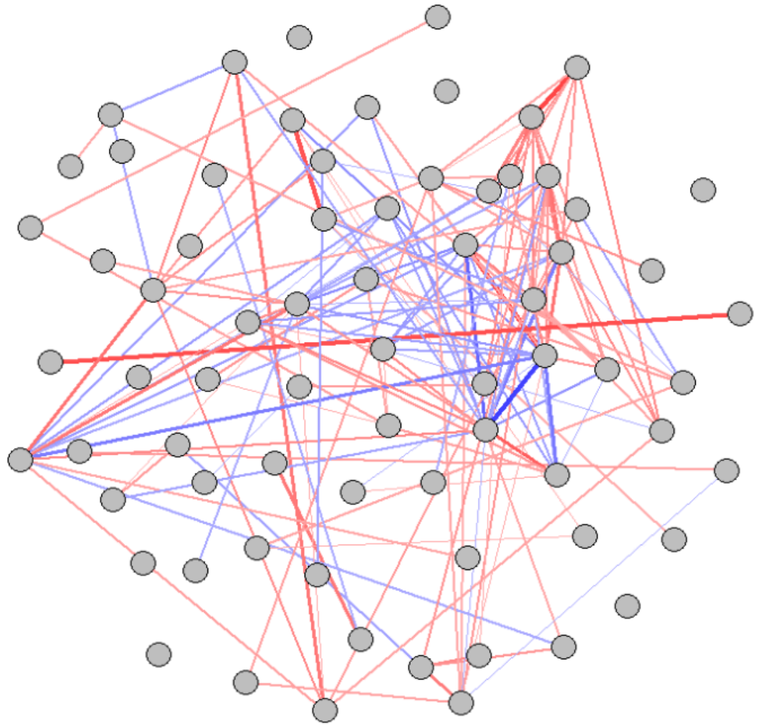
- Node -> gene
- Edge -> significant correlation between the two nodes
- Only correlation values greater in magnitude than 0.3 are represented

Graph for Normal Patients (Direct Correlation Colors)



- sparser
- weaker correlations
- mainly negative correlations

Graph for Cancer Patients (Direct Correlation Colors)



- denser
- intense correlations
- presence of positive correlations

Graphical Lasso

The correlation graphs do not distinguish between **direct vs indirect relationships**

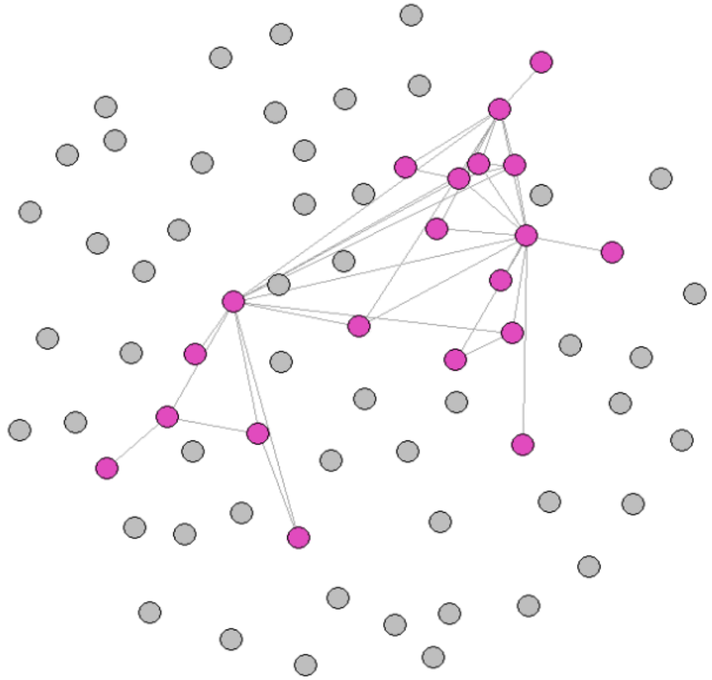
Graphical Lasso

Regularization parameter controlling the sparsity of the graph:

- **λ = 0.2**

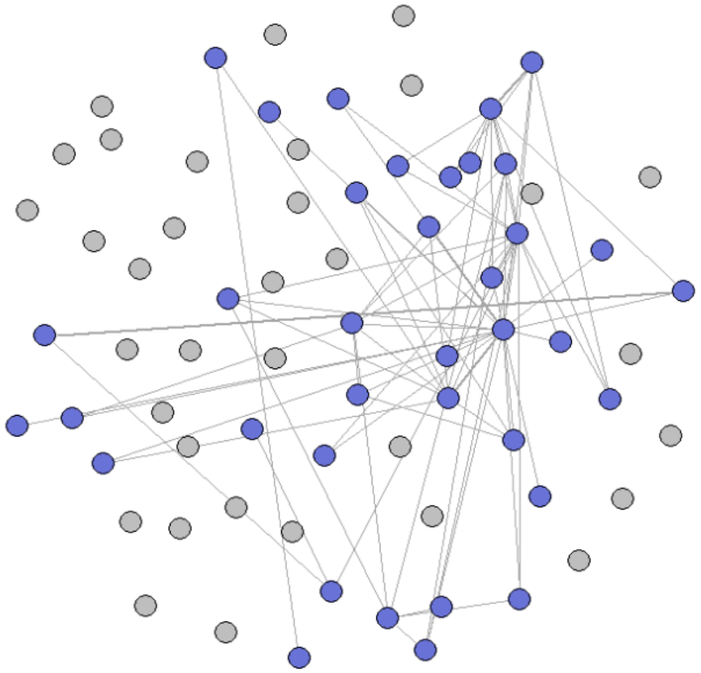
Balance between interpretability and relevance

Graphical Lasso for Normal Patients



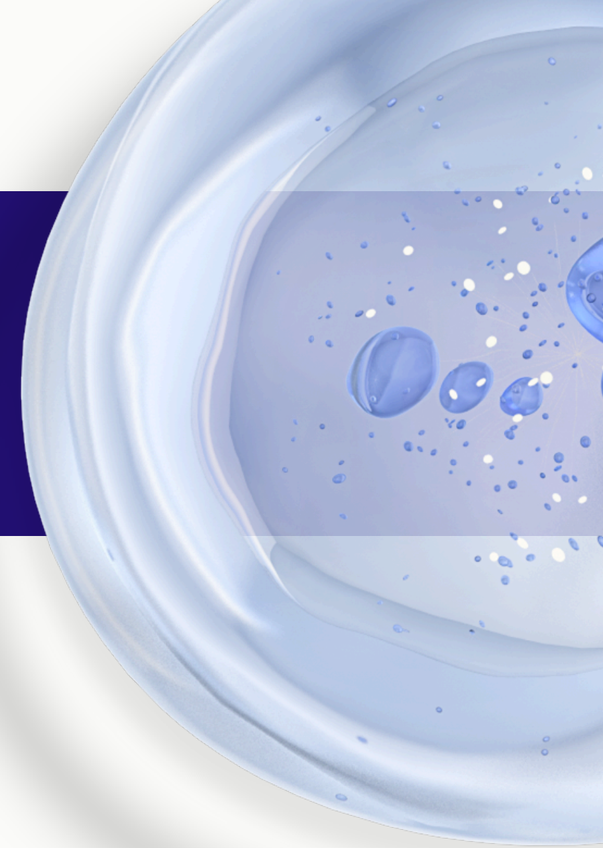
- sparser
- small clusters
- no connection to outer nodes

Graphical Lasso for Cancer Patients



- more interconnected
- bigger clusters
- connection to outer nodes

Conclusions



In cancer conditions gene interactions are **stronger** and more **widespread** -> less independence

The genes driving the most **important dependencies** are **different** in normal vs cancer conditions

Different biological processes at play, especially regarding **gene expression regulation, cellular proliferation** and **cellular metabolism**

Thank you for your attention!