Università di Catania

Introduzione al Data Mining

BooleanNet on Breast Cancer

Autore:
Marco Ardizzone

Matricola:
X81001077

April 2021

Contents

1 Introduction

This project is meant to retrieve boolean implications from mRNA and proteins from Breast Cancer subtypes, in order to check whether there is evidence between a certain quantity of mRNA/proteins and another in case of a given Breast Cancer subtype. Data and source codes of StepMiner and BooleanNet was supplied by Professor Giovanni Micale [1]. This project's goal is to clean the dataset, split by cancer subtype and to check whether there is evidence of correlation between genes in the same subtype of cancer. Source code, Data and Plots for this project are available *here*.

2 Dataset Description

The Dataset comes from cBioPortal [2]. It is made up by 3 .txt files, *brca_clinical*, *brca_expression* and *brca_proteomics.*

- **brca_clinical**: contains clinical data from breast cancer patients, such as ID, Cancer Subtype and Stage.
- **brca_expression**: contains data about patients on the columns and data about the quantity of mRNA for each patients.
- **brca_proteomics**: contains data about patients on the columns and data about the quantity of proteines for each patients

3 StepMiner and BooleanNet

StepMiner [3] is an algorithm used to binarize matrixes of genes information, returning back matrixes having values $x_{i,j} \in \{-1, 0, 1\}$ based on the original values. Once the matrix is discretized, it is passed to BooleanNet.
BooleanNet [4] is an algorithm that, given a discretized matrix, returns a list of boolean relationships between genes.

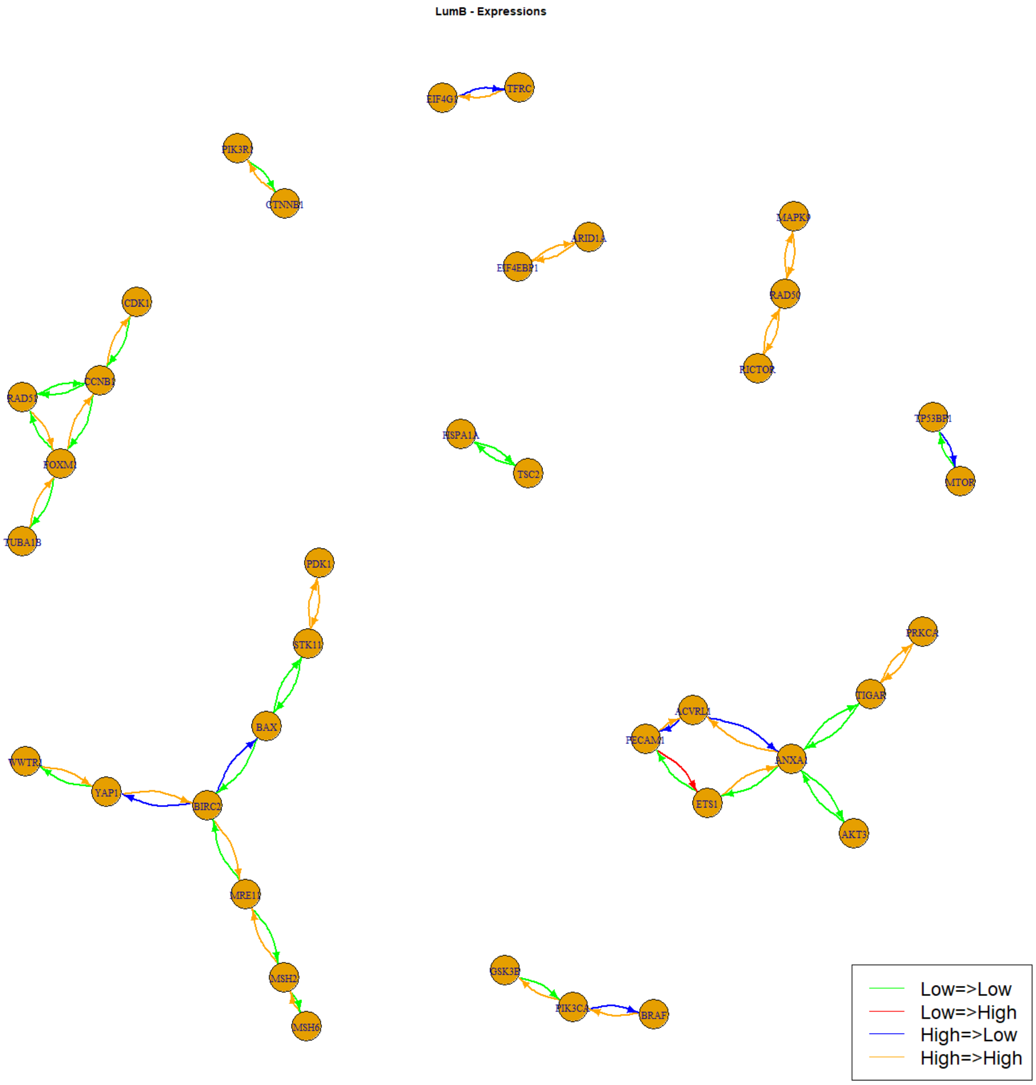
4 Data Cleaning

In order to perform StepMiner and BooleanNet, it is mandatory to clean the dataset and to split by cancer subtype.

1. Remove *NA* values from *brca_clinical*.
2. Split *brca_clinical* in: *brca_clinical.LumA*, *brca_clinical.LumB*, *brca_clinical.Basal*, namely 3 kinds of cancer subtype.
3. Introduce *brca_expression.LumA*, *brca_expression.LumB*, *brca_expression.Basal*, 3 tables which contains data from gene expressions of patients suffering from LumA, LumB and Basal cancer
4. Introduce *brca_proteomics.LumA*, *brca_proteomics.LumB*, *brca_proteomics.Basal*, 3 tables which contains data from proteins of patients suffering from LumA, LumB and Basal cancer

5 Networks

Once data cleaning is done, StepMiner and BooleanNet are applied on the data, in order to obtain 6 adjacency matrixes. Then, using *igraph* library, graphs are created and plotted.

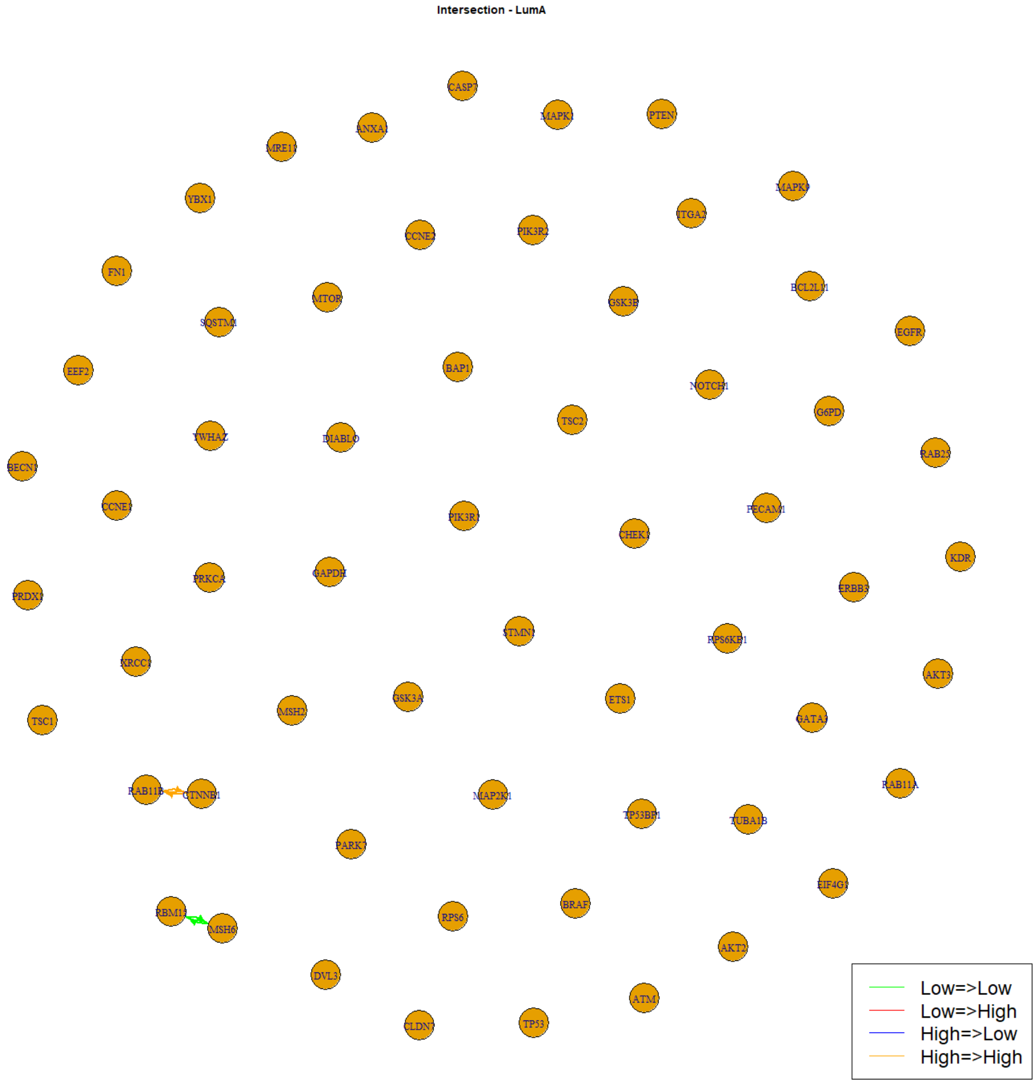


Relationships between mRNA in LumB Breast Cancer.

6 Intersections

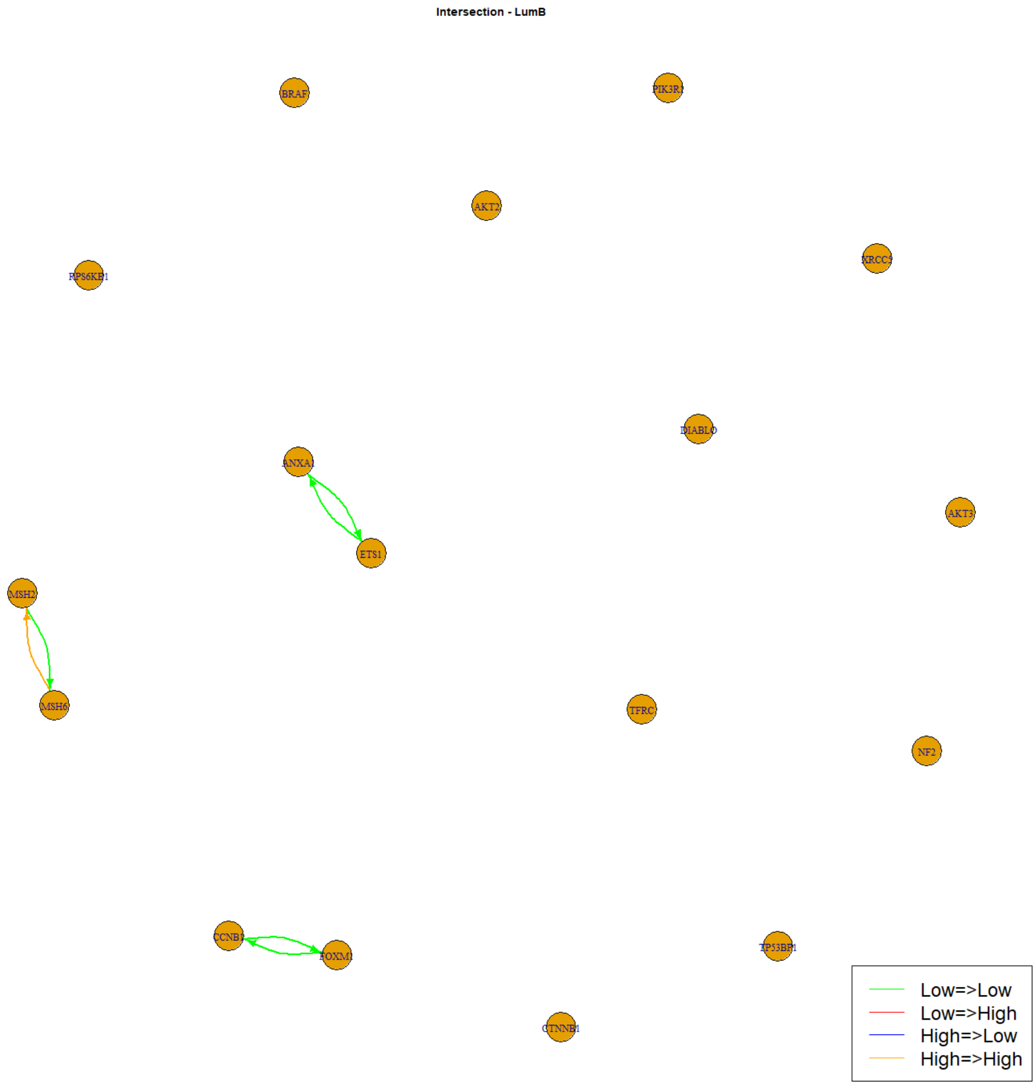
It is interesting to intersect graphs, in order to check which genes are present in both mRNA and protein expressions.

6.1 mRNA and proteins in LumA Cancer



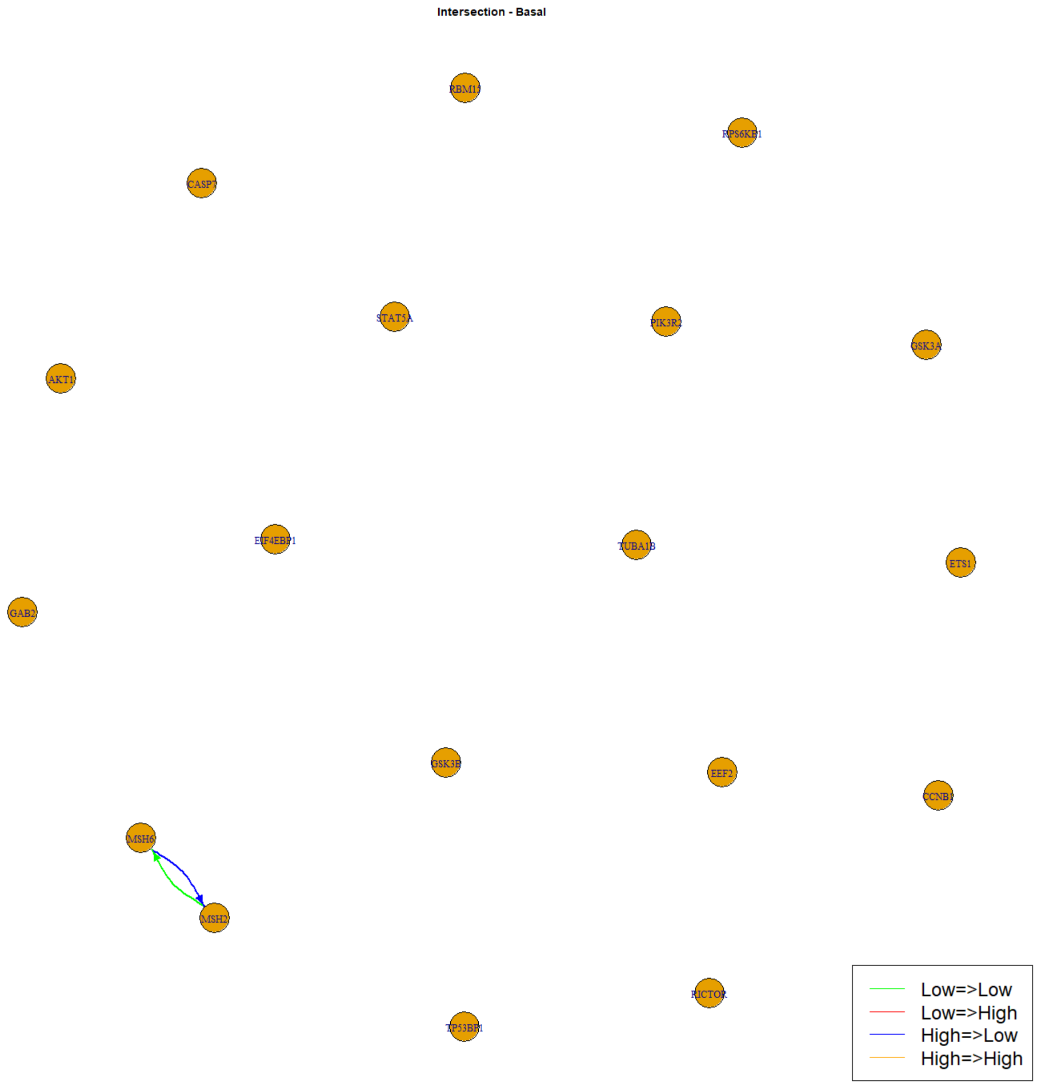
Relationships between mRNA and proteins in LumA Breast Cancer.

6.2 mRNA and proteins in LumB Cancer



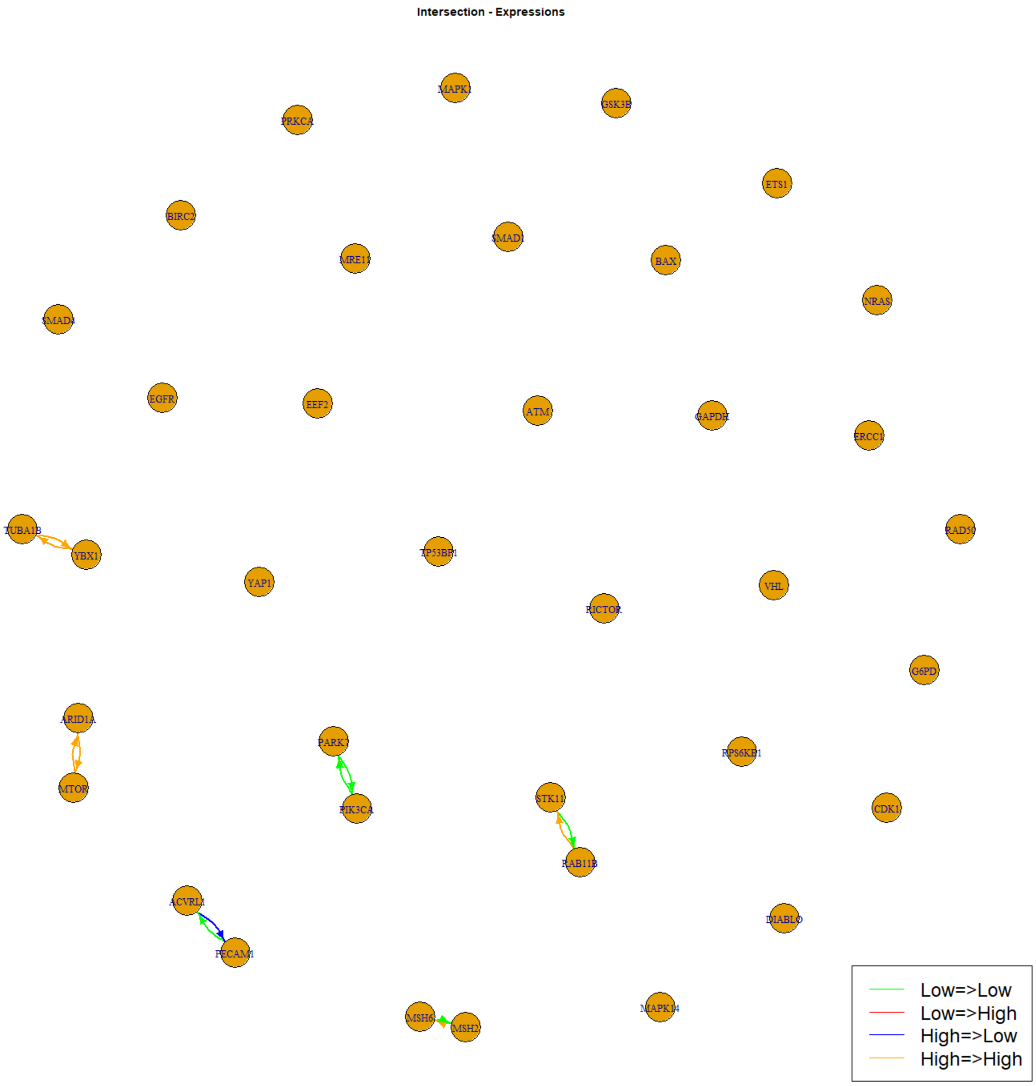
Relationships between mRNA and proteins in LumB Breast Cancer.

6.3 mRNA and proteins in Basal Cancer



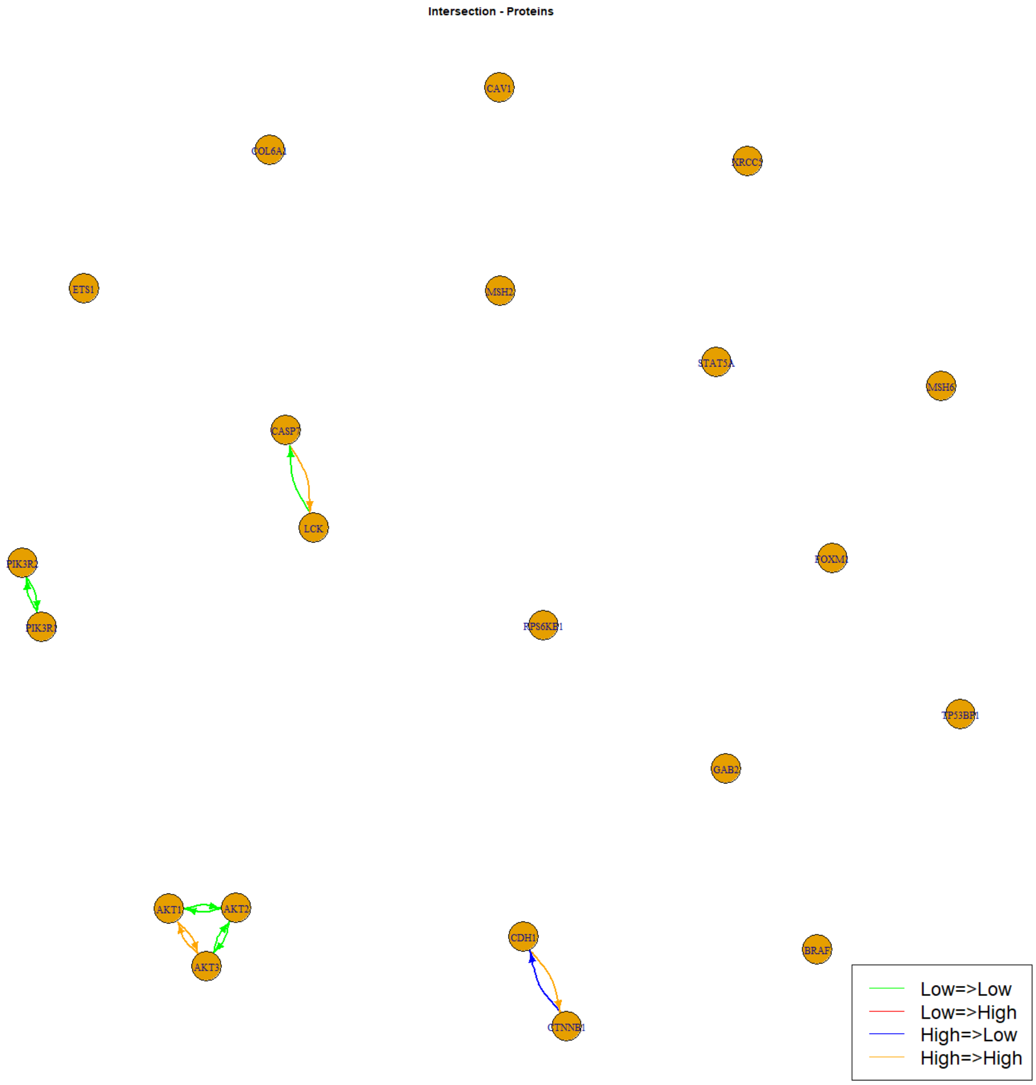
Relationships between mRNA and proteins in Basal Breast Cancer.

6.4 mRNA expression in all three Cancers



Relationships between mRNA of different Breast Cancer subtypes.

6.5 Proteins in all three Cancers



Relationships between proteins of different Breast Cancer subtypes.

7 Conclusion

It was possible to find relationships between genes in the same subtype of cancer and also to intersect them. In particular:

- **LumA Cancer**:
  - $RBM15Low \implies MSH6Low$ and vice versa
  - $RAB11BHigh \implies CTNNB1High$ and vice versa
- **LumB Cancer**:
  - $CCNB1Low \implies AOXM1Low$ and vice versa
  - $ANXA1Low \implies ETS1Low$ and vice versa
  - $MSH6High \implies MSH2High$
  - $MSH2Low \implies MSH6Low$
- **Basal Cancer**:
  - $MSH2Low \implies MSH6Low$
  - $MSH6High \implies MSH2Low$
- **mRNA all three cancers**:
  - $XUBA1BHigh \implies YBX1High$ and vice versa
  - $ARID1AHigh \implies MTORHigh$ and vice versa
  - $PARK7Low \implies PIK3CALow$ and vice versa
  - $ACVRL1High \implies PECAM1Low$
  - $PECAM1Low \implies ACVRL1Low$
  - $STK11Low \implies RAB11BLow$
  - $RAB11BHigh \implies STK11High$
  - $MSH6Low \implies MSH2Low$
  - $MSH2High \implies MSH6High$

- **proteins all three cancers**:
  - $PIK3R2Low \implies PIK3R1Low$ and vice versa
  - $CASP7High \implies LCKHigh$
  - $LCKLow \implies CASP7Low$
  - $CDH1High \implies CTNNB1High$
  - $CTNNB1High \implies CDH1Low$
  - $AKT1Low \implies AKT2Low$ and vice versa
  - $AKT2Low \implies AKT3Low$ and vice versa
  - $AKT3High \implies AKT1High$ and vice versa

References

[1] Giovanni Micale : http://www.medclin.unict.it/docenti/giovanni.micale

[2] cBioPortal : https://www.cbioportal.org
`cBioPortal provides visualization, analysis and download of large-scale cancer genomics data sets.`

[3] StepMiner : Extracting binary signals from microarray time-course data.
`D. Sahoo, D. L. Dill, R. Tibshirani, S. K. Plevritis: Nucleic Acids Research, 2007, Vol. 35, No. 11, pp. 3705-3712`

[4] BooleanNet : Boolean implication networks derived from large scale, whole genome microarray datasets.
`D. Sahoo, D. L. Dill, A. J. Gentles, R. Tibshirani, S. K. Plevritis: Genome Biology, 2008, 9:R 157`