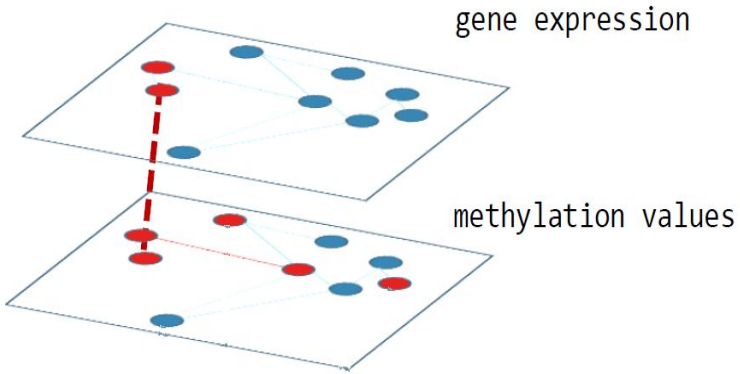
MULTI-GNN FOR LUNG CANCER BIOMARKERS

Michael Bianco, Alessandro Artoni

Introduction

The aim of the project is to identify biomarkers of lung cancer analyzing gene expression and methylation data.



Data download and preprocessing

GENE EXPRESSION

60000 GENES x 1712 SAMPLES

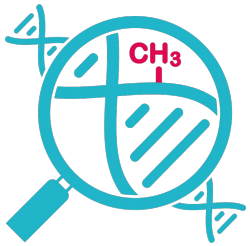
↓

20000 GENES x 1712 SAMPLES

↓

5000 GENES x 1712 SAMPLES

METHYLATION DATA

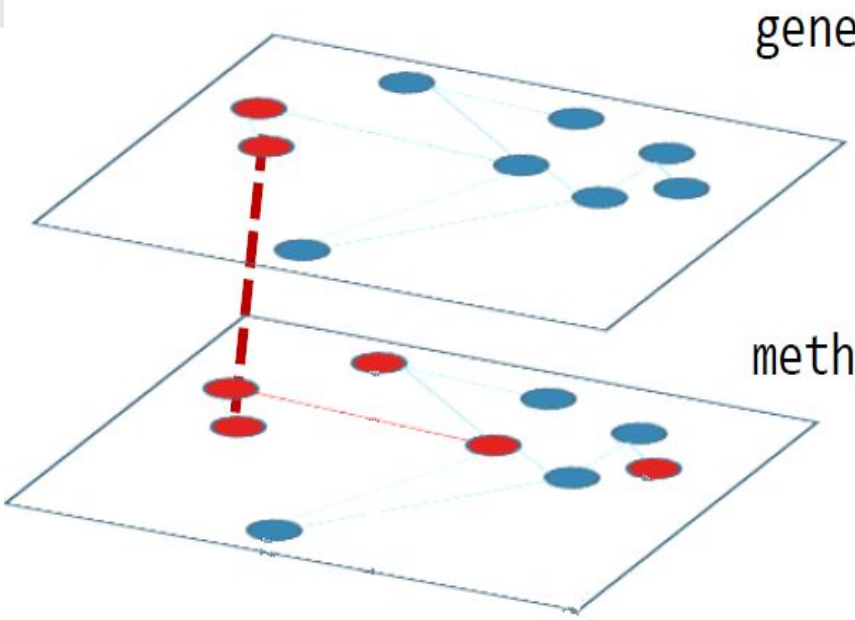


480000 SITES x 1712 SAMPLES

↓

4235 SITES x 1712 SAMPLES

Graph structure



Evaluation Techniques

Louvain: greedy algorithm that maximizes modularity of the graph

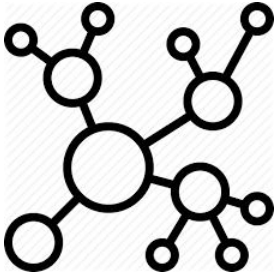
Lasso: minimizes the difference between the predicted values and the actual ones

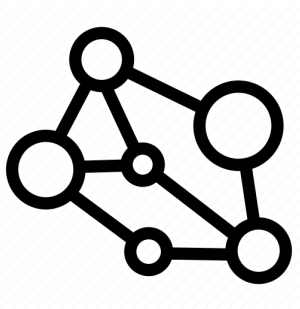
SVM RFE: feature selection technique used to identify the most important features for a predictive model

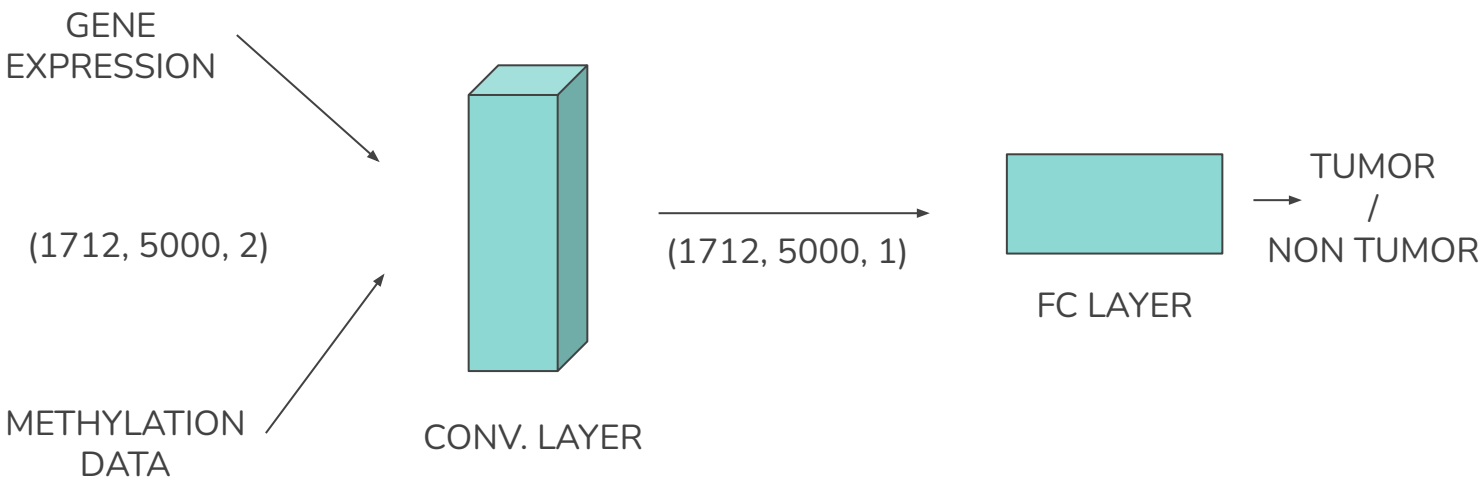
PCA: dimensionality reduction technique used to identify the principal components that captures the majority of the information

Select KBest: selects the K most important features from a given dataset

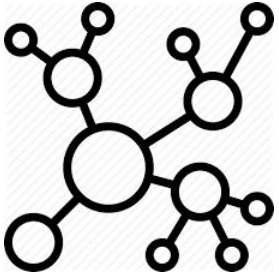
Methylation Data Integration
Learnable Layer Integration







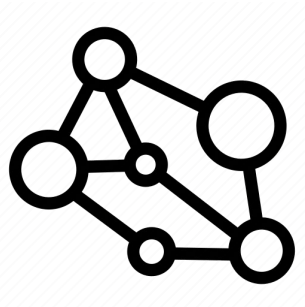
Methylation Data Integration
A-Priori Deterministic Design Integration



GENE
EXPRESSION

$$gene\_expression\_value * (1 + (1 - mean\_meth\_values))$$

WGCNA



METHYLATION
DATA

$$10^{0.5+0.5*corr(i,j)}$$

Overall Results

With our analysis we have found the following known biomarkers for lung cancer

- Gene + Meth learnable layer integration: CPXCR1 (tumor suppr.), TRIM55, TMEM171 (colon cancer), GABRG1, ZBBX (nasoph. cancer), ACP7 (brain tumor)
- Gene + Meth a-priori integration: DEFA3, SPP1, CXCL13, AKR1B10, LIM2, IGSF9, ADCY8, PLAC1, CHRNB4, FAM83A, KRT16, LEP.
- Gene expression: PITX2, BARX1 (over expr.), CST1, PRAME (over expr.).

Biomarkers Found

	Biomarkers for LC	Tumor Suppressor	Biomarkers for other tumors
Learnable layer method	TRIM55, GABRG1	CPXCR1	TMEM171, ZBBX, ACP7
A-priori integration	DEFA3, SPP1, CXCL13, AKR1B10, LIM2, IGSF9, ADCY8, PLAC1, CHRNB4, FAM83A, KRT16, LEP	X	X
Gene expression	PITX2, CST1, BARX1, PRAME	X	X

Methylation Data Integration
Learnable Layer Integration

- ✅ Self-learned representation of the overall data (gene + meth)
- ✅ One weight for each gene
- ✅ Weights are updated taking into account the whole dataset every epoch

- ❌ Finds more tumor suppressor genes with respect to known biomarkers

Methylation Data Integration
A-Priori Deterministic Design Integration

- ✅ Finds more known biomarkers for lung cancer
- ✅ Based on a formula following knowledge about biological transcriptome process

- 🟨 Increments every gene expression value, using a fixed deterministic function

Gene Expression Classification

Network

```
GCN(
  (conv1): GCNConv(5000, 128)
  (conv2): GCNConv(128, 64)
  (conv3): GCNConv(64, 2)
  (classifier): Linear(in_features=2, out_features=2, bias=True)
)
```

Results

SVM	RF	DTC	KN	GCN
99.2%	98.7%	97.2%	98%	98.2%

Graph Convolutional Network

$$\mathbf{h}_v^{(l)} = \sigma\left(\sum_{u \in N(v)} \mathbf{W}^{(l)} \frac{\mathbf{h}_u^{(l-1)}}{|N(v)|}\right)$$

GCN layer performs 2 phases

- Message passing: multiply node features with weight matrix W and normalize by node degree.
- Message aggregation: sum over messages from neighbors, then apply sigmoid activation function.

Methylation Data Classification

Network

```
GCN(
  (conv1): GCNConv(4235, 128)
  (conv2): GCNConv(128, 64)
  (conv3): GCNConv(64, 2)
  (classifier): Linear(in_features=2, out_features=2, bias=True)
)
```

Results

SVM	RF	DTC	KN	GCN
99.1%	98.2%	96.7%	96.2%	99.1%

Conclusion

Thanks to our methods we found that a multi modal approach is better than working on a single modality, because the intersection between our five methods in the multi omics approach found one gene, meanwhile the single omics found zero genes. In addition, the multi omics approach found much more biomarkers than the single omics one. Further analysis can be performed to validate or check our results from a biological point of view, maybe adding some a-priori knowledge on gene expression transcription and methylation.

THANKS FOR THE ATTENTION

Michael Bianco, Alessandro Artoni