MULTI-GNN FOR LUNG CANCER BIOMARKERS

Alessandro Artoni, Michael Bianco

1 Introduction

The term biomarker refers to any gene, protein, and methylation site whose presence in the pathological condition (eg lung cancer) is abnormal compared to the basal state. For example, a gene is a biomarker of lung cancer if it is mainly expressed in samples with lung cancer compared to healthy ones. To date, many methods have been used to identify biomarkers. However, the dependence between genes, proteins, and methylation sites must be thoroughly investigated.

2 Related Work

In the field of bioinformatics several remarkable techniques and models have been developed to address the challenges of finding biomarkers in lung cancer. Cheer et al. employed an intermediate fusion technique, integrating histopathology, clinical, and expression data through unsupervised encoders, culminating in joint representations for predicting patient survival across multiple cancer types. This multimodal fusion strategy accommodated potential absence of certain modalities, paving the way for enhanced interpretability and discovery of inter-modality associations. Li et al. introduced the "classified information index" evaluation standard, employing SVM-RFE to select five relevant features, achieving high accuracy on experimental datasets by mitigating redundancy. Similarly, Shipp et al. utilized the signal-to-noise ratio method, selecting 30 features for DLBCL classification, yielding a 91% recognition rate. However, reliance on classification and evaluation criteria may lead to redundancy issues due to genes' similar expression patterns. Park et al. proposed a multi-omics integration approach for Alzheimer's disease prediction, demonstrating superior accuracy compared to single data methods. Similarly, in lung adenocarcinoma (LUAD) prediction, a DNN model was employed to identify biomarkers utilizing integrated multi-omics data. By evaluating various feature selection techniques and prediction models, distinct biomarkers were identified, underscoring the impact of methodological choices on biomarker discovery. Our proposed method presents a novel multi-modal approach for lung cancer biomarker discovery, using different methodologies to identify common genes crucial for classification.

3 Data download and preprocessing

This work is deal with the study of lung tumor samples freely available in the Genomic Data Commons (GDC) database. We have downloaded for each samples methylation data and mRNA gene expression data, separating tumor samples from normal samples. To connect genes with each other we downloaded Protein-Protein Interaction table from STRING database and Illumina 27 annotations to connect genes with methylation sites. Then the following preprocessing has been performed:

- mRNA: 5000 genes x 1712 samples (1389 tumor samples + 323 normal samples) In principle data were 60000 features, then we discarded all non protein coding genes, applied $log_2(fpkm\_uq\_unstranded)$ and selected the 5000 genes with the highest variance between all samples.
- Meth data: 4235 sites x 1712 samples The original dataset contains over 480000 sites. We discarded sites not connected to protein coding genes, sites with all NaN values and sites not connected with the 5000 genes with highest variance.

4 Multi Omics Biomarker Identification

Since the aim of the project was to find lung cancer biomarkers, our proposed methods try to integrate methylation and gene expression data with a multi-modal approach that use an a-priori deterministic function and a learnable layer.

4.1 Integration with Learnable Layer

Our proposed method involves a model made of a convolutional layer followed by a linear layer. The input data is a tridimensional tensor of shape (1712, 5000, 2), where the first dimension represent patients, the second one genes and in the third we have gene expression values and the mean of the methylation sites for the specific gene and patient. We calculated the mean values of methylation data because each gene can be associated to multiple methylation sites and their values determine if the transcription of the linked gene is blocked or not. Methylation values are between [0, 1] and from a biological point of view we know that higher value of methylation data block the transcription of the relative gene (and viceversa for lower values), for this reason we take $1 - mean_meth$) as value for methylation of each gene and patient in the third dimension.

The idea is to use the convolutional layer to merge both methylation and genes channels to obtain an overall representation for each gene and use this to train a linear layer with a weight for each gene. The training task is a binary classification problem to separate tumor and non-tumor patients, and the model reaches an accuracy of 81%. After training we took 13 genes from the ones with the highest absolute weight value, such as CLCA2, GPR39, WNK2, FGF23,

```
NET(
  (conv1): Conv1d(2, 1, kernel_size=(3,), stride=(1,), padding=(1,))
  (fc1): Linear(in_features=5000, out_features=1, bias=True)
)
```

Figure 1: Network for gene and methylation data

known as overexpressed genes or biomarkers in lung cancer. We also passed the whole dataset through the convolutional layer to take its output as a learned combination of gene expression and methylation data.

4.2 A-Priori Deterministic Design Integration

We also proposed a deterministic method to compare the results obtained with the Learnable Layer method. This one takes the values of the methylation sites associated to each gene for each patient, computes the mean and applies the following formula:

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

Our proposal tries to increment the gene expression value in a percentage inversely proportional to the mean methylation value. For instance if a gene has two methylation sites connected with a mean value of 0.1, its gene expression value is incremented by 90% of its original value.

4.3 Comparison

In this section we discuss the results obtained with our proposed method on multi-modal data. For both we applied five evaluation techniques (Louvain, Lasso, SVM RFE, PCA, Select KBest) on the final data representation, where for each patient there's associated an integrated value of each gene, to find most relevant genes for the classification task.

For the Learnable Layer Method we intersected the results of all the above evaluation methods and we found only one gene (CPXCR1), known in literature as a tumor suppressor gene. If we discarded Select KBest we got 30 genes, among which TRIM55 (tumor suppressor), TMEM171 (prognostic biomarker in colon cancer), GABRG1 (biomarker), ZBBX (downregulated in nasopharyngeal carcinoma), ACP7 (prognostic gene in brain tumor).

Then we re-applied the previous five methods on the output of the A-Priori Deterministic Method, obtaining 3873 candidate genes from Louvain, 401 with LASSO, 1000 from SVM RFE, 1000 from Select KBest and 500 from PCA. From the new intersection we obtained one gene (ADCY8), which is known as a potential biomarker for lung cancer. If we did not intersect PCA we got 20 candidate genes, most of them biomarkers in literature (DEFA3, SPP1, CXCL13, AKR1B10, LIM2, IGSF9, ADCY8, PLAC1, CHRNB4, FAM83A, KRT16, LEP).

Comparing the results of the two proposed methods we noticed that the A-Priori Deterministic method found more known biomarkers for lung cancer with respect to the Learnable Layer method, that found mostly prognostic biomarkers or tumor suppressors for other types of tumor.

5 Single Omics

We also considered the problem as a single omic analysis, working only on gene expression data. In this section we describe our methods and results.

5.1 Biomarker Selection

We selected the common genes between the output of Louvain algorithm, LASSO, Select KBest, SVM RFE and found 40 candidate biomarkers, among which PITX2, BARX1, CST1, PRAME. We searched these genes in literature, found that PITX2 and CST1 are known biomarkers for lung cancer, while BARX1 and PRAME are known as highly expressed genes in different types of human cancer, including lung cancer.

6 Classification Methods

In addition we considered the classification task to distinguish between tumor and non-tumor patients using gene expression and methylation data separately.

6.1 Gene Expression Classification

We performed patients binary classification (tumor and non-tumor) with various methods, among which Support Vector Machine (SVM), Random Forest (RF), Decision Tree Classifier (DTC), K-Neighbors Classifier (KN) and our implementation of a Graph Convolutional Network (GCN). Table 1 shows our results:

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

Table 1: Accuracy Table on gene expression

Nodes of the GCN are patients and edges are computed with Pearson Correlation coefficient with a threshold of 0.8. Input features are represented by gene expression values for each patient. The underlying figure shows the structure of the network, composed by three Graph Convolutional layers and one Linear Classifier.
GCN layer performs the two phases of message and aggregation with the following formula:

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

Figure 2: GCN Network for gene expression data

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

Figure 3: GCN message and aggregation

For each neighbor of the current node computes the message, multiplies the learnable weights and the previous message normalized by node degree. Then sum over messages from neighbors in performed and finally sigmoid activation function is applied.

6.2 Methylation Data Classification

Then we applied the previous methods to methylation data and Table 2 shows our results:

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

Table 2: Accuracy Table on methylation data

Nodes of the GCN are patients and edges are computed with Pearson Correlation coefficient with a threshold of 0.98. Input features are represented by methylation sites for each patient. The underlying figure shows the structure of the network, composed by three Graph Convolutional layers and one Linear Classifier.

7 Evaluation Metrics

7.1 Louvain

We developed a Graph in which nodes are represented by gene expression data and edges connect genes based on PPI.

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

Figure 4: GCN Network for methylation data

We divided tumor patients and non-tumor ones, associated weights on each edge based on Weighted Correlation Network Analysis (WCGNA) method, using the following formula:

$$10^{0.5+0.5*corr(i,j)} \tag{2}$$

This method performs Pearson Correlation coefficient on gene expression value and scales it in [1, 10], instead of [-1, 1], to weight more connections between genes.

Louvain is a greedy algorithm that maximizes modularity of the graph and operates in two phases:

1. Each node is assigned to its own community, next for each node, the change in modularity is calculated removing the node from its own community and moving it into the community of each neighbor.

2. Then each community is reduced to a single node and its connected edges are reduced to a single weighted edge. Once the new graph is created the first phase can be re-applied.

We applied this method on normal and tumor graph separately, to find communities between genes. Then we exploited communities to find genes that change communities and took them as candidate biomarkers.

7.2 Lasso

Least Absolute Shrinkage and Selection Operator (LASSO) is a regularization technique that helps feature selection. It minimizes the difference between the predicted values and the actual ones of the target variable, imposing a penalty on the absolute values of one coefficient for each input feature and trying to shrink the coefficients towards zero. The output is a value of the coefficient for each feature, some of them are exactly zero (so not relevant features), and others have a greater value.

7.3 RFE

Recursive Feature Elimination (RFE) is a feature selection technique used to identify the most important feature for a predictive model, applied to Support

Vector Machine. It recursively trains the model on the entire set of features, provides a score for each feature based on its importance in the classification, deletes the least important features and repeats the process until the desired number of features is reached. In our case we selected the 1000 most important features for the binary tumor/non-tumor classification.

7.4 PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique that projects the original data in a new reduced space in which the axis are represented by the top N eigenvectors with highest eigenvalues, retaining the most significant variance in the data and ensuring that the few principal components capture the majority of the information. For our purpose we selected the 500 most important genes.

7.5 Select KBEST

Select KBEST is a feature selection algorithm based on statistical tests. It selects the K most important features from a given dataset by training a classifier, assigns a score to each input feature to represent their importance and selects the ones with the highest scores. In our case we selected the 1000 most important features.

8 Conclusions

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.