Diploma project

Graph Neural Networks for Identification of Robust Biomarkers from Multi-Omics Data

Lukas Ruzicka

January 22, 2023

State of Art

Biomarkers are nowadays considered to be any biological characteristics that can be objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacological responses to a therapeutic intervention [12]. By understanding how measurable biological processes relate to clinical outcomes it is possible to further improve disease treatments as well as deepen our knowledge of normal physiology. Studies of biomarkers should always use clinical outcomes as an ultimate measure, at least for retrospective analysis of biomarker correlation success, to avoid unnecessary health risks [17].

Nowadays, existence of high-throughput techniques on large samples and the need to understand underlying biochemical processes results in availability of multi-omics data [18]. Omics technologies are defined as high-throughput biochemical assays that measure comprehensively and simultaneously molecules of the same type from a biological sample, such as genomics (DNA/RNA), transcriptomics (transcripts), proteomics (protein quantity) etc.. It may quite often happen that multi-omics datasets are not fully complete, which can be caused by reasons such as budget limitations or experimental constraints [2]. Since many statistical methods require the datasets to be complete or balanced, several approaches on how to deal with such problem were already proposed [23] [5] [22].

Early attempts to classify multi-omics data were unsupervised methods, such as clustering [21] [15]. The simplest form of integrating the data is called early integration, which concatenates all the features for each omics type. [7] [15] both report better results when multiple omics data types were used compared to only single or pairwise data type integration. However, the effect of noise becomes greater when multiple omics data are being utilized and thus adding a feature selection step is mandatory [21]. Unfortunately this approach equally weighs each data type and suffers from a curse of dimensionality as the higher-dimensional features dominate the clustering. [20]. Several methods were developed to deal with such problems by assigning weights to features [19] or regularization while assuming latent lower dimensional distributions [13]. Late integration performs clustering for each omics type separately and the clustering solutions are integrated to obtain a single clustering solution. This approach however performs poorly when each of the omics type contains only a weak signal and isn't able to detect the correlations between different omics types [20]. At last, intermediate integration methods create models that incorporate all omics types, such as models based on similarity networks, e.g. [24] (builds similarity network with message passing techniques), dimension reduction models, e.g. [10] (assumes that the variation in each omics can be partitioned to a variation that is joint between all omics), [9], [16] (approaches using CCA) or model which construct statistical model of data, e.g. [15] (assumes low dimensional representation of the original data population).

As the current medicine rapidly advances, with new modern methods more and more curated datasets with detailed annotations are becoming available. This leads to increase of interest in supervised learning techniques [26]. Highly flexible and performing graph neural networks (GNN) are proving to be able to further improve the performance of learning/classification on multi-omics data, being able to exploit the interactions across different omics data types. Multiple GNN frameworks for working with multi-omics data have already been proposed and their performance tested on several distinct datasets. [26] uses mRNA expression data, DNA methylation data and microRNA expression data. Based on similarity weighted networks, GNNs for each omics type are built separately, which utilize features and correlations between samples. Then, cross-omics discovery tensor is assembled from GNN outputs and forwarded into view correlation discovery network, which further exploits cross-omics correlation. [25] utilizes single cell RNA sequences to infer cell-to-cell relationships. It is using three autoencoders in a loop. The feature autoencoder learns a dimensional representation of the input as embedding. Upon that, graph autoencoder learns the topology of a graph and the input expression matrix of the feature autoencoder is constructed by cluster autoencoder. All runs until convergence. [7] works with CNA, methylation, miRNA and gene expression data and uses a graph-based semi-supervised learning method. For each omics type, the graph was constructed, with edges representing similarities between cancer patients, extracted from different omics data types. [20] classifies kidney renal clear cell carcinoma patients through integrating pathways with help of graph kernels. Similarity between two patients (represented as nodes) is given by node attributes along the shortest path in pathway graph. Smoothing is performed so that attributes in proximity also contribute to similarity. [27] uses pathway graphs to map the input layer onto low-dimensional functional layer, which then serves as a modified input layer of typical deep neural network. Lastly, [1] assigns layer to each omics data type and then so called guidance graph, which incorporates prior knowledge about the regulatory interactions among features at distinct omics layers, is used to model the inter-omics interactions between layers.

Protein-protein interactions (PPIs) play a significant role in majority of biological processes. On top of physical bond structure, PPIs can also contain information about protein cooperation and mutual regulation [11]. PPI datasets are often visualized as graphs, where proteins are represented as nodes and interactions as connections between nodes [8]. Furthermore, the knowledge of PPIs has proven to be helpful in new drug discovery as well as the prevention and diagnosis of diseases [6]. There have already been some successful implementations of PPIs into GNN frameworks. [3] proposed a machine learning method based on GNNs and prior knowledge embedding in form of PPIs for single-cell RNA to protein prediction. [14] is another GNN framework for learning PPI patterns that constructs residue level graph and determines its structure based on spatial arrangement of proteins in 3D space.

Lastly, there are six major PPI databases , among others [8]. Since pairwise overlaps of those libraries are reaching only up to 75 %, comprehensive datasets should be then obtained by integration over multiple PPI databases. It's also worth noting that each PPI database uses different ontologies and thus identical PPIs may differ among different databases [8].

Methodology

Dataset descriptions

Data were split between 5 datasets, each containing data for a specific omic type. See Table 1 below for description of individual datasets. Also, a dataset containing labels was provided, which specified 3 classification tasks - binary "disease" classification task and "risk" and "mutation" classification tasks with 3 classes.

dataset	number of features	number of samples
allRNA	58216	86
circRNA	51018	86
miRNA	2553	105
piRNA	556	104
TE	687	112

Table 1: Overview of omic-specific datasets

Models used

For the purpose of having a reference to which GNN can be compared to, a set of baseline models was chosen and their performance tested on each classification task. These three classifiers were chosen: Multinomial Logistic Regression (MLR), Support Vector Machine with linear kernel (SVM) and Gradient Boosting Classifier (GBC).

MOGONET framework was used as a representative of GNNs.

Data preprocessing

Due to the fact that datasets were incomplete and imbalanced, preprocessing and feature selection was necessary before proceeding with classification tasks. Firstly, only samples present in all datasets were kept. This resulted in keeping only 66 samples in each dataset.

Different feature selection methods were used for different classifiers. In case of MLR, L1 regularization (Lasso) was used. In case of SVM, Recursive feature elimination was used and in case of GBC features were selected based on the Mean decrease of impurity. In case of multinomial logistic regression, parameter C was being tweaked to control the strength of regularization. For both the SVM and GBC, a number of selected features was specified prior to running the regularization. Overview of feature selection methods with their respective models and parameters is shown below in Table 2.

classifier	feature selection method	modified parameter	values
MLE	L1 regularization (Lasso)	C	0.42, 0.6, 0.8, 1.0
SVM	Recursive Feature Elimination	n_features	5, 10, 20, 100
GBC	Mean decrease of impurity	n_features	5, 10, 20, 100

Table 2: Overview of baseline classifiers and their respective feature selection methods

In case of MOGONET, Maximum Relevance Minimum Redundancy (MRMR) algorithm was used. Due to memory allocation problems with the first dataset ("allRNA"), the dataset was divided into two halves and from each a set of n features were extracted. Those two sets were then merged into a "secondary" dataset of $2n$ features and MRMR used again to extract the final amount of features. 200 features were selected from each omic-specific dataset. This number was reused from the original paper. [26].

Baseline evaluation

Each baseline model was trained using cross-validation, $k = 6$. To analyze contribution of each omic-type, performance of baseline models was first measured using only omic-specific datasets separately. After that, models were tested on the complete dataset including all omic types. For each baseline model, precision, recall, F-1 score and AUC were computed. Also, ROC curves were plotted. In order to plot ROC curves for multiclass classification tasks, the (1) and (2) labels were together considered as positive class.

MOGONET evaluation

For each classification task, MOGONET was run for different values of k parameter, which represents the average number of edges retained per sample in the similarity network constructed by network. Values 2, 3, 5 and 10 were tested. Number of training epochs was set to 1000. In case of binary classification, test accuracy, F-1 score and AUC were reported. In case of multi-class classification, test accuracy, weighted F-1 score and F-1 macro were reported. These values were average over 5 runs for each value of k Furthermore, 30 most influential biomarkers were also reported for each classification task.

Results

Shown below are results of baseline classifiers both for specific omic data and for combined omic data. Only the best performing instances are presented.

classifier	parameter	classification task	precision	recall	F-1	AUC
LR	C = 0.42	Disease	0.05	0.50	0.10	0.46
	C = 1.00	Risk	0.31	0.35	0.26	0.47
	C = 0.42	Mutation	0.21	0.33	0.26	0.48
SVM	n_features = 100	Disease	0.98	0.79	0.95	1.00
	n_features = 100	Risk	1.00	1.00	1.00	0.81
	n_features = 100	Mutation	0.98	0.92	0.95	0.50
GCB	n_features = 20	Disease	0.94	0.99	0.96	1.00
	n_features = 10	Risk	0.84	0.82	0.82	0.98
	n_features = 20	Mutation	0.98	0.92	0.95	0.64

Table 3: Classification results of baseline models on allRNA dataset

classifier	parameter	classification task	precision	recall	F-1	AUC
LR	C = 0.42	Disease	0.05	0.50	0.10	0.43
	C = 0.42	Risk	0.14	0.33	0.19	0.48
	C = 0.42	Mutation	0.21	0.33	0.26	0.48
SVM	n_features = 10	Disease	0.97	0.71	0.78	0.63
	n_features = 100	Risk	0.89	0.77	0.80	0.67
	n_features = 100	Mutation	0.92	0.75	0.78	0.51
GCB	n_features = 5	Disease	1.00	1.00	1.00	1.00
	n_features = 20	Risk	0.90	0.89	0.89	0.99
	n_features = 20	Mutation	0.91	0.85	0.88	0.51

Table 4: Classification results of baseline models on circRNA dataset

Below are shown results of classification with use of MOGONET. Each value in Table 9 is an average of 5 runs. Furthermore, in Table 10 are shown most important biomarkers for each classification task, ranked from top to bottom.

classifier	parameter	classification task	precision	recall	F-1	AUC
LR	C = 0.42	Disease	0.05	0.50	0.10	0.54
	C = 1.00	Risk	0.30	0.35	0.28	0.42
	C = 0.42	Mutation	0.21	0.33	0.26	0.48
SVM	n_features = 100	Disease	0.98	0.79	0.85	0.98
	n_features = 100	Risk	0.90	0.86	0.87	0.72
	n_features = 100	Mutation	0.91	0.60	0.65	0.44
GCB	n_features = 5	Disease	0.84	0.71	0.75	0.58
	n_features = 20	Risk	0.82	0.82	0.82	0.94
	n_features = 10	Mutation	0.70	0.65	0.67	0.74

Table 5: Classification results of baseline models on miRNA dataset

classifier	parameter	classification task	precision	recall	F-1	AUC
LR	C = 0.42	Disease	0.05	0.50	0.10	0.42
	C = 1.00	Risk	0.36	0.35	0.24	0.40
	C = 0.42	Mutation	0.21	0.33	0.26	0.47
SVM	n_features = 100	Disease	0.95	0.57	0.60	0.73
	n_features = 20	Risk	0.44	0.58	0.50	0.50
	n_features = 5	Mutation	0.21	0.33	0.26	0.41
GCB	n_features = 5	Disease	0.88	0.78	0.82	0.79
	n_features = 10	Risk	0.62	0.61	0.62	0.76
	n_features = 10	Mutation	0.55	0.55	0.55	0.45

Table 6: Classification results of baseline models on piRNA dataset

Discussion

One problem with the dataset, which is apparent right from the beginning is the unfortunate label imbalance in the ”disease” class, where only 7 out of 66 samples are labeled as 0, remaining as 1. This could result into inherent bias towards the positive label (1). Another potential mistake could have been made when it was necessary to divide one of the datasets into halves when performing feature selection using MRMR. Since at each iteration the best feature is chosen based on previously chosen best features [4], the final selection might be influenced by that.

Identified biomarkers could be furthermore verified in research papers or GO database. Baseline reported values should be modified (rather weighted values, also report those which are available for MOGONET). I'd also give a go to try another framework and try to compare performace, differences.

Github repository

Link for github repository: https://github.com/ruza22/DT

classifier	parameter	classification task	precision	recall	F-1	AUC
LR	C = 0.42	Disease	0.05	0.50	0.10	0.51
	C = 0.42	Risk	0.14	0.33	0.19	0.48
	C = 0.42	Mutation	0.21	0.33	0.26	0.47
SVM	n_features = 5	Disease	0.45	0.50	0.47	0.71
	n_features = 100	Risk	0.53	0.53	0.49	0.67
	n_features = 5	Mutation	0.21	0.33	0.26	0.37
GCB	n_features = 5	Disease	0.76	0.76	0.76	0.70
	n_features = 20	Risk	0.68	0.68	0.68	0.89
	n_features = 10	Mutation	0.73	0.64	0.67	0.51

Table 7: Classification results of baseline models on TE dataset

classifier	parameter	classification task	precision	recall	F-1	AUC
LR	C = 0.42	Disease	0.05	0.50	0.10	0.46
	C = 0.42	Risk	0.32	0.38	0.30	0.42
	C = 0.42	Mutation	0.21	0.33	0.26	0.55
SVM	n_features = 100	Disease	0.97	0.71	0.78	0.95
	n_features = 100	Risk	0.92	0.87	0.89	0.59
	n_features = 100	Mutation	0.92	0.68	0.75	0.40
GCB	n_features = 5	Disease	0.94	0.99	0.96	1.00
	n_features = 20	Risk	0.90	0.86	0.87	0.99
	n_features = 100	Mutation	0.93	0.86	0.89	0.65

Table 8: Classification results of baseline models on all omics dataset

Literature resources

[1] Zhi-Jie Cao and Ge Gao. “Multi-omics single-cell data integration and regulatory inference with graph-linked embedding”. In: *Nature Biotechnology* (2022), pp. 1–9.

[2] Ana Conesa and Stephan Beck. “Making multi-omics data accessible to researchers”. In: *Scientific data* 6.1 (2019), pp. 1–4.

[3] Xinnan Dai et al. “PIKE-R2P: Protein–protein interaction network-based knowledge embedding with graph neural network for single-cell RNA to protein prediction”. In: *BMC bioinformatics* 22.6 (2021), pp. 1–17.

[4] Chris Ding and Hanchuan Peng. “Minimum redundancy feature selection from microarray gene expression data”. In: *Journal of bioinformatics and computational biology* 3.02 (2005), pp. 185–205.

[5] François Husson and Julie Josse. “Handling missing values in multiple factor analysis”. In: *Food quality and preference* 30.2 (2013), pp. 77–85.

		k = 2	k = 3	k = 5	k = 10
Disease	Accuracy	0.875	0.888	0.938	0.938
	F-1	0.933	0.940	0.953	0.966
	AUC	1.00	1.00	1.00	1.00
Risk	Accuracy	0.938	0.925	0.938	0.888
	F-1 weighted	0.936	0.921	0.935	0.890
	F-1 macro	0.945	0.931	0.949	0.912
Mutation	Accuracy	0.900	0.875	0.875	0.888
	F-1 weighted	0.890	0.862	0.860	0.875
	F-1 macro	0.890	0.850	859	0.875

Table 9: Classification results of MOGONET

[6] Kanchan Jha, Sriparna Saha, and Hiteshi Singh. "Prediction of protein–protein interaction using graph neural networks". In: *Scientific Reports* 12.1 (2022), pp. 1–12.

[7] Dokyoon Kim et al. "Knowledge boosting: a graph-based integration approach with multi-omics data and genomic knowledge for cancer clinical outcome prediction". In: *Journal of the American Medical Informatics Association* 22.1 (2015), pp. 109–120.

[8] Benjamin Lehne and Thomas Schlitt. "Protein-protein interaction databases: keeping up with growing interactomes". In: *Human genomics* 3.3 (2009), pp. 1–7.

[9] Dongdong Lin et al. "Group sparse canonical correlation analysis for genomic data integration". In: *BMC bioinformatics* 14.1 (2013), pp. 1–16.

[10] Eric F Lock et al. "Joint and individual variation explained (JIVE) for integrated analysis of multiple data types". In: *The annals of applied statistics* 7.1 (2013), p. 523.

[11] Guofeng Lv et al. "Learning Unknown from Correlations: Graph Neural Network for Inter-novel-protein Interaction Prediction". In: *arXiv preprint arXiv:2105.06709* (2021).

[12] Richard Mayeux. "Biomarkers: potential uses and limitations". In: *NeuroRx* 1.2 (2004), pp. 182–188.

[13] Qianxing Mo et al. "Pattern discovery and cancer gene identification in integrated cancer genomic data". In: *Proceedings of the National Academy of Sciences* 110.11 (2013), pp. 4245–4250.

[14] Manon Réau et al. "DeepRank-GNN: a graph neural network framework to learn patterns in protein-protein interfaces". In: *bioRxiv* (2021).

[15] Ronglai Shen, Adam B Olshen, and Marc Ladanyi. "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis". In: *Bioinformatics* 25.22 (2009), pp. 2906–2912.

[16] Amrit Singh et al. "DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays". In: *Bioinformatics* 35.17 (2019), pp. 3055–3062.

[17] Kyle Strimbu and Jorge A Tavel. "What are biomarkers?" In: *Current Opinion in HIV and AIDS* 5.6 (2010), p. 463.

[18] Indhupriya Subramanian et al. "Multi-omics data integration, interpretation, and its application". In: *Bioinformatics and biology insights* 14 (2020), p. 1177932219899051.

[19] Yijun Sun et al. "Improved breast cancer prognosis through the combination of clinical and genetic markers". In: *Bioinformatics* 23.1 (2007), pp. 30–37.

[20] Yasin Ilkagan Tepeli et al. "PAMOGK: a pathway graph kernel-based multiomics approach for patient clustering". In: *Bioinformatics* 36.21 (2020), pp. 5237–5246.

[21] Giulia Tini et al. "Multi-omics integration—a comparison of unsupervised clustering methodologies". In: *Briefings in bioinformatics* 20.4 (2019), pp. 1269–1279.

[22] Michel van de Velden and Tammo HA Bijmolt. "Generalized canonical correlation analysis of matrices with missing rows: a simulation study". In: *Psychometrika* 71.2 (2006), pp. 323–331.

[23] Valentin Voillet et al. "Handling missing rows in multi-omics data integration: multiple imputation in multiple factor analysis framework". In: *BMC bioinformatics* 17.1 (2016), pp. 1–16.

[24] Bo Wang et al. "Similarity network fusion for aggregating data types on a genomic scale". In: *Nature methods* 11.3 (2014), pp. 333–337.

[25] Juexin Wang et al. "scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses". In: *Nature communications* 12.1 (2021), pp. 1–11.

[26] Tongxin Wang et al. "MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification". In: *Nature Communications* 12.1 (2021), pp. 1–13.

[27] Lianhe Zhao et al. "DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis". In: *Computational and structural biotechnology journal* 19 (2021), pp. 2719–2725.

Disease	Risk	Mutation
AC004893.10	AC002553.4	ABCA1
hsa-miR-134-5p	USP48	hsa-miR-1268a
hsa-miR-128-2-5p	hsa-miR-1231	hsa-miR-1202
hsa-miR-128-3p	hsa-miR-1247-5p	hsa-miR-1234-3p
hsa-miR-129-5p	hsa-miR-127-5p	hsa-miR-1243
hsa-miR-130b-3p	hsa-miR-1273e	hsa-miR-125a-5p
hsa-miR-132-5p	hsa-miR-1275	hsa-miR-126-3p
hsa-miR-133a-5p	hsa-miR-128-2-5p	hsa-miR-126-5p
hsa-miR-136-3p	hsa-miR-129-2-3p	hsa-miR-127-3p
hsa-miR-1275	hsa-miR-1290	hsa-miR-1178-5p
hsa-miR-144-3p	hsa-miR-134-5p	hsa-miR-127-5p
hsa-miR-144-5p	hsa-miR-135b-5p	hsa-miR-1276
hsa-miR-1469	hsa-miR-136-3p	hsa-miR-128-1-5p
hsa-miR-152-3p	hsa-miR-138-5p	hsa-miR-1286
hsa-miR-153-3p	hsa-miR-144-3p	hsa-miR-1292-5p
hsa-miR-153-5p	hsa-miR-144-5p	hsa-miR-1296-3p
hsa-miR-1278	hsa-miR-1468-5p	hsa-miR-1185-2-3p
hsa-miR-127-5p	hsa-miR-1228-5p	hsa-miR-100-5p
hsa-miR-1537-5p	hsa-miR-1228-3p	hsa-miR-133a-5p
WDR19	hsa-miR-1178-3p	ZNF124
VASN	ZCCHC8	XRN1
VAV1	VPS13D	XXbac-BPG170G13.32
VIM	VPS53	XXbac-BPG254B15.10
VPS13D	VPS8	ZEB1
VPS8	VWA8	ZMYND8
VRK2	WDR43	ZNF121
WDR37	WRNIP1	ZNF354B
hsa-miR-127-3p	ZDHHC20	hsa-let-7e-5p
WWOX	hsa-miR-10b-5p	ZNF44
XXbac-BPG27H4.7	ZMYM2	ZNF541

Table 10: 30 most important biomarkes for each classificatino task, first ranked is at the top of the table