Using ontology embeddings for structural inductive bias in gene expression data analysis

Maja Trębacz, Zohreh Shams, Mateja Jamnik, Paul Scherer
Nikola Simidjievski, Helena Andres Terre, Pietro Liò

Department of Computer Science and Technology,
University of Cambridge

UNIVERSITY OF CAMBRIDGE
Department of Computer Science and Technology

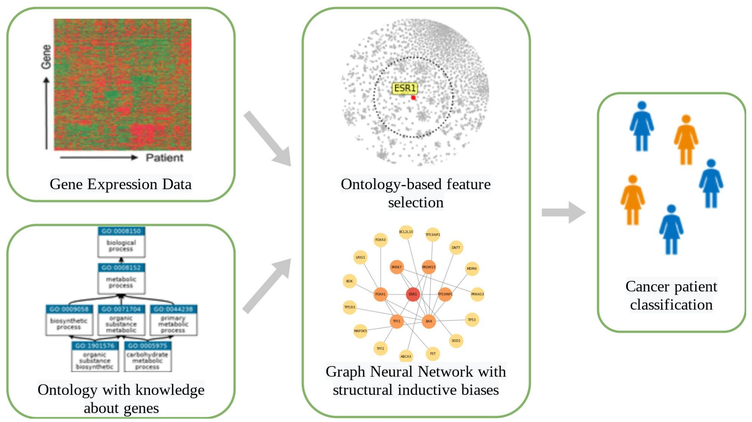
Overview

Stratifying cancer patients based on gene expressions

+ allows personalising diagnosis and treatment planning
– data is noisy and machine learning models struggle to identify true dependencies
– data is extremely highly dimensional as it contains expression values for over 20'000 genes per patient
– number of samples in the datasets is low

Ontology: structured representations of semantic knowledge commonly used to represent biological concepts

Key question: can biological knowledge from ontologies improve performance of ML techniques for stratifying cancer patients?



Contributions

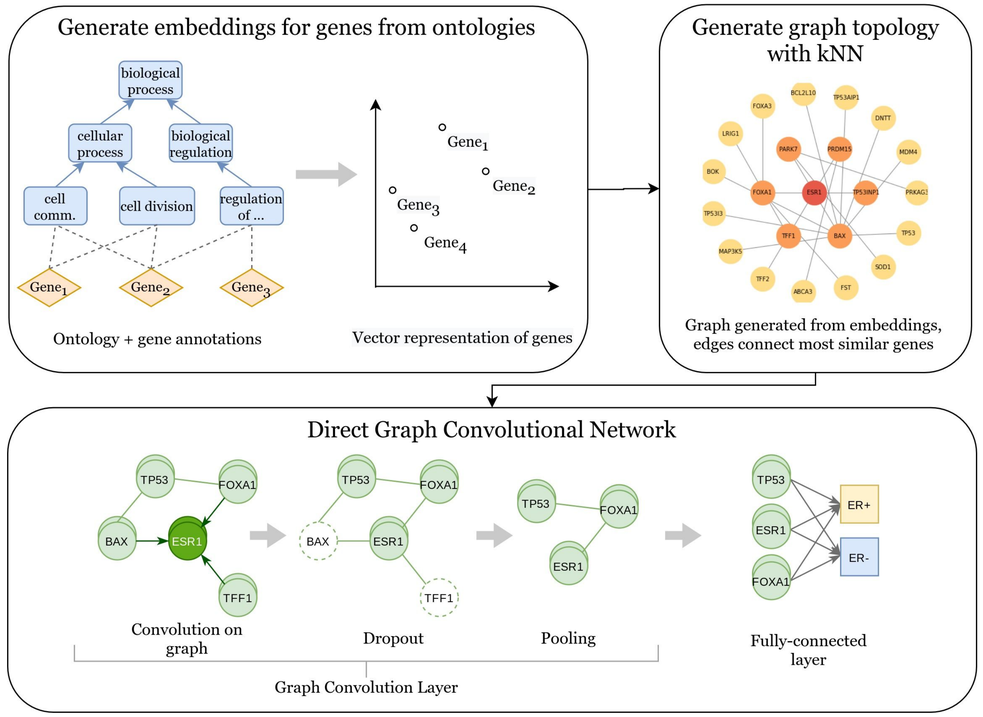
- Method, OntoGCN, using **ontology embeddings of genes** to impose structural inductive bias on the deep learning model by using **graph convolutional network (GCN)**
- Novel **ontology-based feature selection** pipeline
- Results showing improvements **for predicting clinical targets** from high-dimensional low-sample data

References

[1] Dutil et al. Towards gene expression convolutions using gene interaction graphs. In International Conference on Machine Learning (ICML) Workshop on Computational Biology (WCB), 2018.
[2] Chen et al. Predicting Candidate Genes From Phenotypes, Functions, And Anatomical Site Of Expression. Bioinformatics, 2020
[3] Hashir et al. Is graph biased feature selection of genes better than random? Machine Learning in Computational Biology (MLCB) meeting, 2019.
[4] Crawford and Greene. Graph biased feature selection of genes is better than random for many genes. BioRxiv, 2020.
[5] Curtis et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature, 486(7403):346–352, 2012.
[6] Warde-Farley et al. The genemaniaprediction server: biological network integration for gene prioritization and predicting gene function. Nucleic acids research, 38(suppl_2):W214–W220, 2010
[7] Szklarczyk et al.String v11: protein–protein association networks with increased coverage, supporting functionaldiscovery in genome-wide experimental datasets. Nucleic acids research, 47(D1):D607–D613,2019.

OntoGCN: Graph Convolutional Network for processing gene expressions directed by ontology embeddings

- OntoGCN enforces convolutions on the genes related by similarity and thus captures localised patterns of data, similarly as convolutional neural networks capture spatial relationships of pixels in the images [1].
- The gene embeddings are generated with DL2vec [2] graph-based method learning over three biomedical ontologies (GO, UBERON, and MP).
- One can freely control the sparsity of the network or the number of neighbours for each of the nodes. This overcomes problems with using curated networks (e.g. PPI) [3,4].



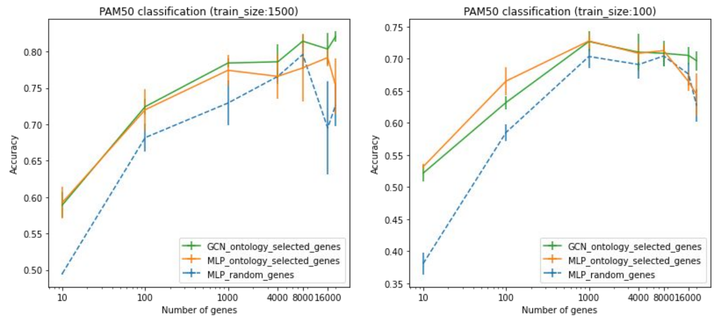
Experiments and Results

We evaluate via classification of breast cancer patients from their genomic data collected by METABRIC [5], aiming at targets:

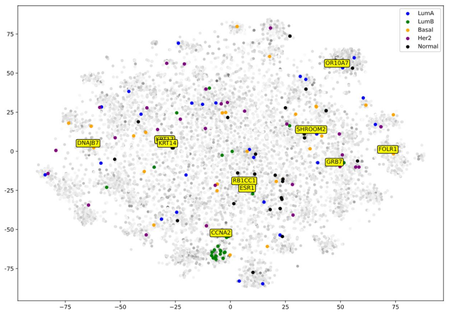
- PAM50 (5-class molecular cancer subtypes)
- ER (2 immunohistochemistry subtypes)
- iC10 (11 IntegrativeCluster subtypes)

Table 1: Performance comparison of the methods on the PAM50, ER and iC10 patient classification tasks (accuracy±std), using all genes as input features. Bold font marks the cases when the model obtained statistically significantly better results than the baselines.

	PAM50		ER		IC10	
train size	100	1500	100	1500	100	1500
OntoGCN w/o node embeddings	**72.3±2.9**	**81.2±0.8**	**91.4±1.0**	93.7±1.0	48.2±3.1	**74.3±2.3**
OntoGCN w/ node embeddings	**72.7±3.7**	**81.6±2.2**	90.8±0.5	93.8±1.2	50.4±2.3	73.7±3.3
Random Forest	69.3±1.1	78.4±1.0	88.7±1.2	93.1±1.2	66.7±0.9	71.3±2.1
Multi-Layer Perceptron	60.1±5.0	77.9±2.5	88.8±1.8	90.9±5.0	40.4±4.5	68.9±1.6



Accuracy on the PAM50 classification varied over the number of genes used



t-SNE visualisation of the ontology embedding space showing high-weight genes for PAM50

	PAM 50	
	train size=100	train size=1500
OntoGCN	**72.3±2.9**	**81.2±0.8**
Random graph	68.9±3.9	78.0±4.7
GeneMANIA graph [6]	71.7±2.6	80.1±2.1
STRINGdb graph [7]	71.2±2.5	80.7±1.2
Random Forest	69.3±1.1	78.4±1.0
Multi-Layer Perceptron	60.1±5.0	77.9±2.5

Performance comparison of the graphs directing GCN structure (accuracy ± std on PAM50)