Extending Similarity Network-Based Classifiers to the Non-Coding Genome and Deep Learning

Jie Yu[1], Duncan Forster[2], Shraddha Pai[3,4]

1. University of Waterloo, Waterloo, Canada. 2. Donnelly Centre for Cellular and Biomolecular Research, Toronto, Canada.
3. Ontario Institute for Cancer Research, Toronto, Canada. 4. Department of Medical Biophysics, University of Toronto.

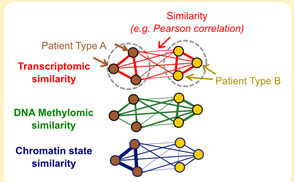
Contact: shraddha.pai@utoronto.ca

Abstract

Similarity networks provide a useful framework for multi-modal data integration, suitable for applications such as gene function prediction and patient classification[1]. We previously developed a supervised learning algorithm which converted heterogeneous patient data into the common space of patient similarity networks (PSN) and used these networks as input features[2] (netDx.org). In addition to excellent classification performance and handling missing data, netDx provides interpretability by allowing users to group genes into pathway-level features. However, the pathway-based grouping approach is of limited value for genomic data outside coding regions. Moreover, the current framework has limited scalability in the number of nodes and networks and does not take advantage of improved discriminability available in the deep learning framework.

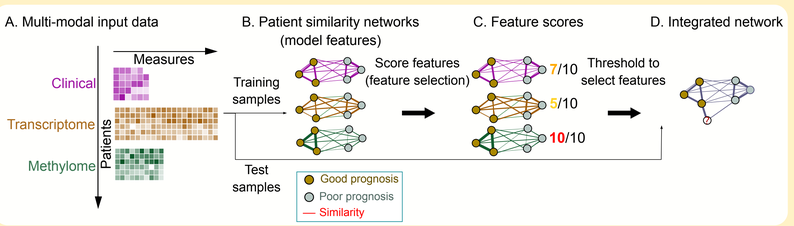
Here, we describe two recent areas of work addressing these limitations. In the first, we classify binary survival in PFA ependymomas using tumour DNA methylomes organized with prior knowledge of brain tissue- and cell-specific expression, transcription factor binding sites and chromatin state. In the second, we extend a recently developed framework from Forster *et al.*[3] for multiple network integration based on graph convolutional networks, to classification. Developing an approach to score features for interpretability remains an active area of research.

Background



Patient similarity networks (PSN) are networks where nodes are patients and edges are weighted by pairwise similarity for a data type. Similarity metric is user-defined (e.g. Pearson correlation for transcriptomic similarity).

The PSN framework allows building of classifiers that are accurate, generalizable, can integrate heterogenous data, and handle missing data[1].



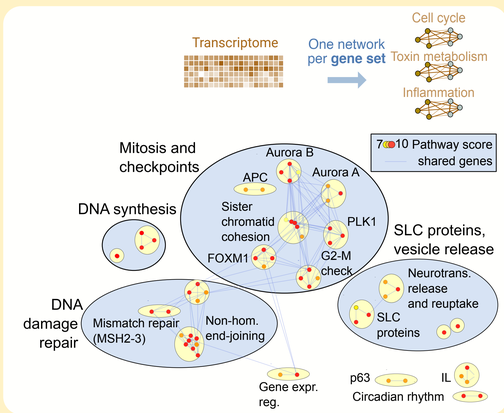
How netDx works. Each data layer (A) is converted to common PSN space (B). Each network is one feature that goes into the model. Feature selection is used to score predictive features (C), and top-scoring features are combined into a single, integrated patient network (D). Unknown patients (test samples) are classified by relative similarity to known examples (training samples).

Designing pathway features that may provide mechanistic insight.

Top: Transcriptomic measures can be grouped into sets that represent curated pathways to create pathway-level features.

Bottom: Example of a binary breast tumour classifier[2] (N=384 tumours) using pathway features.

Predictive pathway themes are consistent with known dysregulated signaling pathways.



References:
1. Pai S and GD Bader (2018). J Mol Biol. 430.
2. Pai S, Hui S, Isserlin R, Shah MA, Kaka H and GD Bader (2019). Mol Sys Biol.
3. Forster D, Boone C, Bader GD and B Wang. (2021). bioRxiv preprint.

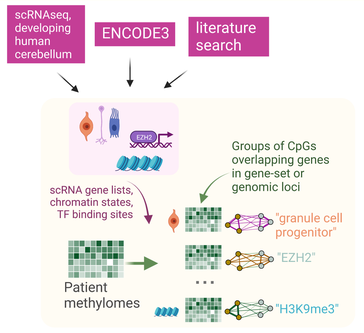
Designing Interpretable Features for Patient DNA methylomes

Motivation

One area of research is extending feature design to the non-coding genome, of relevance to classification tasks that use epigenetic measures such as DNA methylation, but also genetic data (e.g. SNPs, CNVs). Here we focus on modelling DNA methylation, which is the most commonly used 'omic data type for brain tumour diagnostics[1], predicts treatment resistance in glioblastoma[2], and which is dysregulated in some groups of pediatric brain tumours, such as Group 3 and 4 medulloblastoma and PFA ependymoma[3,4].

Strategy

Just as we grouped gene-level transcriptomic measures into pathways ("Background"), here we group base-level CpG methylation - the unit measured in methylomic assays - by transcription factor binding sites, chromatin states, and marker genes for cell populations of disease relevance (e.g. those in developing human cerebellum for pediatric brain tumours).



Application

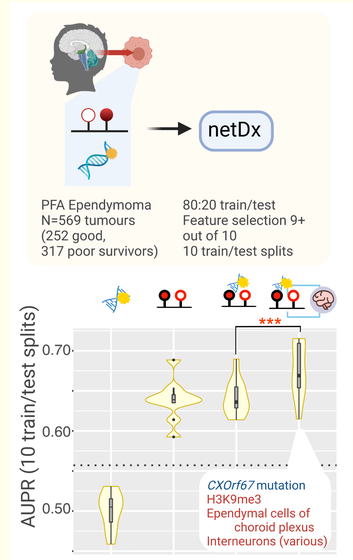
Methods: We predicted binarized survival in PFA-Ependymoma using 569 brain tumour methylomes[4] (Illumina 450K). Using netDx, we evaluated a model that grouped CpG-level methylation into sets reflecting 25 cell types from the developing human cerebellum[5], EZH2 binding sites, 3 chromatin states (ENCODE), and brain super enhancers[6] (80:20 train/test split, feature selection >=8/10; 10 splits). This model was compared to a baseline lacking that did not use prior knowledge (single feature for all methylome).

Preliminary results: Organizing methylomes by prior knowledge significantly improves prognostic prediction (AUPR: 67.5+/- 3.5, mean +/- SD; baseline model: 64.2 +/- 2.3; $p < 5x10^{-3}$, one-sided WMW). Features that predict prognosis are consistent with known dysregulation in PFA ependymomas[4], including *CXOrf67* mutation, methylation in H3K9me3 sites, marker genes for ependymal cells of choroid plexus and for multiple interneuron classes.

Conclusion & Future Directions

Interim Conclusion: Prior knowledge can improve survival prediction in PFA-EP and identify features reflecting tumour biology.

Current Directions: We are validating this finding in an independent dataset and extending this approach to Group 3/4 medulloblastoma, pediatric hindbrain tumours with uncharacterized epigenetic components.



Application to binary survival prediction in PFA ependymomas. Top: Predictor design, integrates *CXOrf67* mutation and methylomes. Bottom: Performance of models with (L-R) mutation only; methylome, baseline; mutation + baseline methylome; mutation + methylome with prior knowledge.

References:
1. Capper *et al.* (2018). Nature. 555.
2. Qian XC *et al.* (1997) Cancer Res.
3. Northcott PA *et al.* (2014) Nature. 511.
4. Pajtler *et al.* (2018) Acta Neuropathol. 133.
5. Aldinger *et al.* (2021) Nat Neurosci. 24.
6. Jiang Y *et al.* (2019). NAR 47

Taking Similarity Network-Based Classification to Deep Learning

Motivation

The current algorithm for netDx is limited in tuneability, does not use GPUs for high-throughput training, and lacks a framework whereby trained models can be used for transfer learning (*i.e.* as a starting point for a related classification task). These problems with netDx would be resolved by adapting the algorithm to the deep learning framework, a fast evolving space in artificial intelligence research and a class of models with high discriminability[1]. Here we describe our efforts to adapt a recent graph convolutional network-based approach for multi-modal data integration into a classifier (BIONIC)[2].

How the algorithm works:

1.Pre-BIONIC: Patient data is converted into similarity networks using Pearson correlation as similarity metric. Optional: To build interpretability, prior knowledge of pathway definitions can be used to create pathway-level PSNs, similar to netDx.

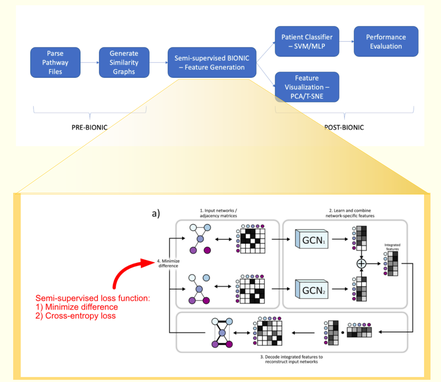
2. Modified BIONIC: This step uses both training and test samples. Test samples are masked during the training task.

a. Each adjacency matrix is put through a graph convolutional layer to generate features for each node in the network.

b. Network specific node features are summed to create integrated node features.

c. A semi-supervised approach has been added to BIONIC to incorporate labelled data. The loss function uses two components: 1) minimizing the mean-squared distance between a network reconstruction (produced from the integrated features) and all input networks (unsupervised), 2) cross-entropy loss using labelled data (semi-supervised). This step creates a single embedded PSN. This step uses 5-fold cross validation.

Workflow for GCN-based classifier

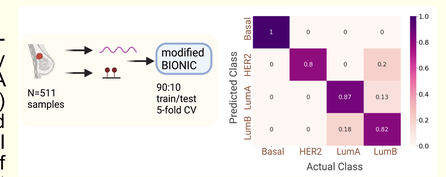


Overview of multi-modal data integration by BIONIC. The red arrow indicates the change made to the training loss function to create a semi-supervised classifier. *Adapted from Forster et al. with permission.*

3. Post-BIONIC: The embedding now contains training and test samples. It can be used with any standard classifier algorithm to classify test samples (e.g. SVM, MLP).

Application

We applied the above algorithm for 4-way classification of breast tumours by integrating gene expression and DNA methylation (N=511 tumours, TCGA[3]) using 90:10 train/test split and 5-fold cross validation. The model demonstrated an average F1-score of 0.88 across all 4 classes on the test set (N=52 samples, test accuracy 0.88).



Application to 4-way breast tumour classification by integrating transcriptome and DNA methylome (left). Results (right).

Interpretability? A work of active research is developing interpretability in this framework[4], similar to feature selection in the current netDx method. Current limitations include:

Perturbation-based approaches may not scale to designs with thousands of input features (e.g. pathway-based features)

To our knowledge, gradient-based feature importance scores (e.g. saliency maps) are not able to show the network-based importance; instead, they only provide information about the importance of each individual node/edge feature.

Preliminary exploration of attention coefficients and integration scaling factors suggests that these are uninformative to discriminate between network-based features.

References: 1. LeCun Y, Bengio Y, Hinton G. (2015). Nature. pp 436-444. 2. Forster DT, Boone C, Bader GD, Wang B. (2021). bioRxiv 3. TCGA (2012). Nature. 490. 4. Azodi CB, Tang J, Shiu S-H. (2020) TiGs 36.