Literature review

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

Lukas Ruzicka

November 17, 2022

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 [11]. 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 [16].

Nowadays, existence of high-throughput techniques on large samples and the need to understand underlying biochemical processes results in availability of multi-omics data [17]. 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 [22] [4] [21].

Early attempts to classify multi-omics data were unsupervised methods, such as clustering [20] [14]. The simplest form of integrating the data is called early integration, which concatenates all the features for each omics type. [6] [14] 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 [20]. Unfortunately this approach equally weighs each data type and suffers from a curse of dimensionality as the higher-dimensional features dominate the clustering. [19]. Several methods were developed to deal with such problems by assigning weights to features [18] or regularization while assuming latent lower dimensional distributions [12]. 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 [19]. At last, intermediate integration methods create models that incorporate all omics types, such as models based on similarity networks, e.g. [23] (builds similarity network with message passing techniques), dimension reduction models, e.g. [9] (assumes that the variation in each omics can be partitioned to a variation that is joint between all omics), [8], [15] (approaches using CCA) or model which construct statistical model of data, e.g. [14] (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 [25]. 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. [25] 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. [24] 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. [6] 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. [19] 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. [26] 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 [10]. PPI datasets are often visualized as graphs, where proteins are represented as nodes and interactions as connections between nodes [7]. Furthermore, the knowledge of PPIs has proven to be helpful in new drug discovery as well as the prevention and diagnosis of diseases [5]. 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. [13] 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 [7]. 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 [7].

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

[5] 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.

[6] 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.

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

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

[9] 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.

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

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

[12] 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.

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

[14] 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.

[15] 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.

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

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

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

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

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

[21] 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.

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.