BRANet: Graph-based Integration of Multi-omics Data with Biological a priori for Regulatory Network Inference

Surabhi Jagtap, Aurélie Pirayre, Frederique Bidard, Laurent Duval, Fragkiskos D. Malliaros

BRANET: Graph-based Integration of Multi-omics Data with Biological *a priori* for Regulatory Network Inference

Surabhi Jagtap[1,2], Aurélie Pirayre[2], Frederique Bidard[2], Laurent Duval[2],
and Fragkiskos D. Malliaros*[1]

[1] Université Paris-Saclay, CentraleSupélec, Inria, Gif-Sur-Yvette, France
Email: {surabhi.jagtap, fragkiskos.malliaros}@centralesupelec.fr
ORCID: 0000-0001-7599-5434 (SJ), 0000-0002-8770-3969 (FM)

[2] IFP Energies nouvelles, 1 et 4 avenue de Bois-Préau, 92852 Rueil-Malmaison, France
Email: {surabhi-vasantrao.jagtap, aurelie.chataignon, frederique.bidard-michelot, laurent.duval}@ifpen.fr
ORCID: 0000-0003-0112-3689 (AP), 0000-0002-7732-4666 (LD)

*corresponding author

Keywords: multi-omics data analysis, biological network integration, graph representation learning, regulatory network inference.

Abstract. Multilayer network embedding approaches are gaining large impetus to analyse omics data. Indeed, network integration approaches have demonstrated their efficiency for protein-protein interaction prediction, gene-regulatory network (GRN) inference, protein function prediction, and drug target identification. To our knowledge, very few network embedding methods have been specifically designed to handle heterogeneous multilayer networks. Moreover, in gene regulation studies MicroRNAs (miRNAs) are important non-coding RNAs and play key roles in tumorigenesis by targeting oncogenes or tumor suppressor genes. To promote the clinical application of miRNAs, the regulatory mechanism of a miRNA to mRNAs (genes) is very important. In this study, we propose BRANET, a novel multi-omics integration framework for multilayer heterogeneous networks. BRANET is an expressive and scalable method to learn node embeddings, leveraging random walk information within a matrix factorization framework. We evaluate BRANET on a TCGA pancreatic cancer dataset and demonstrate its efficiency for miRNA-mRNA regulatory network (MMRN) inference.

1 Scientific Background

Biological systems are composed of multiple interacting entities. Such entities can be genes, proteins, miRNAs, metabolites or epigenetic marks. It is a fundamental task to understand whether and how properties and activities of system entities interact. The number of high dimensional omics data measuring molecules (for instance, proteins, miRNA, mRNA) from biological samples have increased. Despite the wealth of numerous available omics datasets, there are some noticeable challenges regarding their acquisition, processing, efficient integration, and interpretation [1]. To this direction, networks are widely used to represent biological relationships (edges) between individual entities (nodes). The major challenge thus pertains to encode these networks in a way that they can effectively be used as input to machine learning models to perform downstream tasks such as bio-marker identification, drug-target prediction, disease-gene associations, GRN and MMRN inference [2]. In this study, we are inspired by graph representation learning (GRL) algorithms that allow us to encode the high-dimensional graph structure into compact embedding vectors [3]. As a target application, we focus on the task of MMRN inference, which will be described in detail in Sec. 2. In particular, we aim to embed graph-based heterogeneous multi-omic information in a lower-dimensional space towards inferring edges of MMRN.

In the related literature, a variety of algorithms and methodologies have been proposed for network integration that are based on random walks, matrix factorization, and neural networks [2]. However, the majority of them are applied for single-layer networks. Nevertheless, for biological networks there are only a few existing network integration strategies that leverage GRL. As one of the first proposed models, MASHUP [4] is a network integration framework based on matrix factorization that builds compact low-dimensional vector representations of proteins. MULTI-NET [5] is the extension of the SKIP-GRAM model to graphs that allows to perform random walks by defining paths to traverse the nodes. More recently, DEEPNF [6] is a network fusion method based on multimodal deep autoencoder (MDA) to integrate different heterogeneous networks. These approaches consider a multilayer network, learning vector representations for each node and indeed requires extensive parameter tuning. Besides, they are challenged when applied to omics data which demands comprehensive handling of data heterogeneity towards preserving biological relevance [1]. In such cases, it is necessary to obtain knowledge-based representations of the nodes, that can assist omics data analysis.

The goal of this work is to propose BRANET, a novel multi-omics integration framework inspired from graph embedding techniques, and to examine its application to the task of MMRN inference. Moreover, an expressive transition probability is considered to relate nodes within random walks, towards learning informative latent node representations. We leverage a properly chosen random walk matrix (PPMI matrix that we use), that allow us to capture relevant context around each node of interest. More precisely, we introduce network integration with the concept of multilayered heterogeneous graph embeddings, that perform matrix factorization by approximating the spectrum of a PPMI matrix. In our preliminary empirical analysis, we evaluate BRANET for the task of MMRN inference. We apply it over gene (mRNA) and miRNA expression datasets for pancreatic cancer [8] and compare its performance to baseline methods.

Source code and data availability: https://github.com/Surabhivj/BRANet.

2 Materials and Methods

Data acquisition: MicroRNAs (miRNAs) are small noncoding RNAs employed by the cells for gene (mRNA) regulation. A single miRNA ($\approx$ 22 nucleotides) can regulate the expression of numerous genes. Our dataset is comprised of $2,065$ samples ($1,836$: Tumor; 228: Normal) for $1,045$ miRNAs and $20,501$ mRNAs. We eventually perform differential expression analysis to obtain the list of omics features (mRNAs and miRNAs) that are differentially expressed in normal vs. tumor samples. miRNA dysregulation is known to be associated with cancer as they are actively involved in mechanisms like genomic instabilities, abnormal transcriptional control, altered epigenetic regulation, and biogenesis machinery defects. To study such mechanisms, we select upregulated miRNAs and downregulated mRNAs for the same experimental conditions to infer MMRN.

Workflow of BRANET: In Fig. 1, an illustration of the BRANET framework is presented. It consists of four main components to infer relationships from omics data. Given a set of p omics matrices $\mathbf{X}^{(i)}_{|n_i|\times|m_i|}, i = 1, \ldots, p$, where n_i is a set of omics features (for instance, genes, miRNAs, CpGs) and m_j represents the samples of test and control expression datasets, BRANET infers a network of $|N|$ nodes, where $N \subseteq \{n_1 \cup n_2 \cup \cdots \cup n_p\}$. We select N by performing differential expression analysis on each omics data independently. We first build a multilayer network G, represented

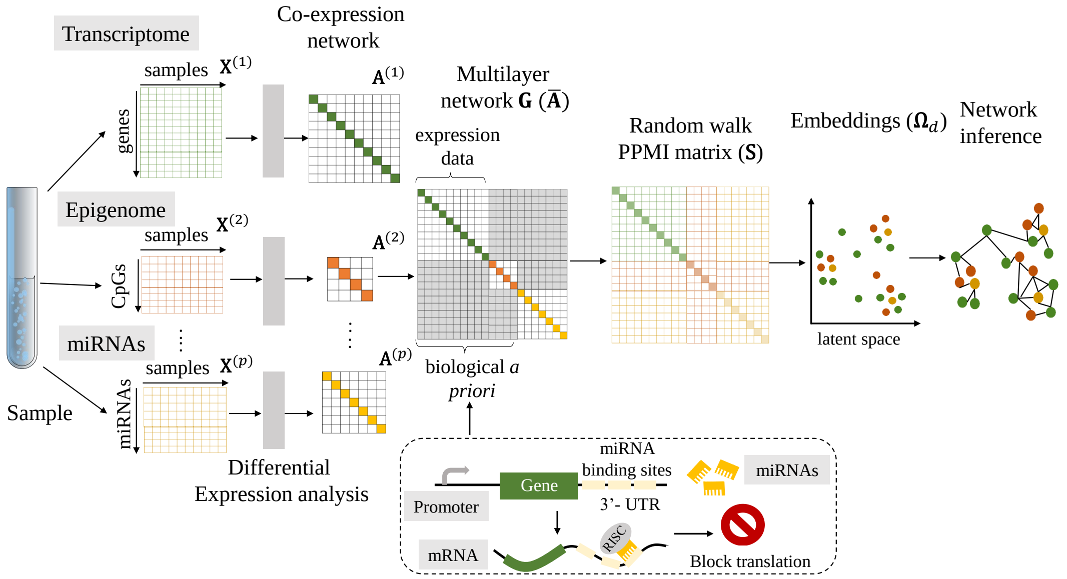


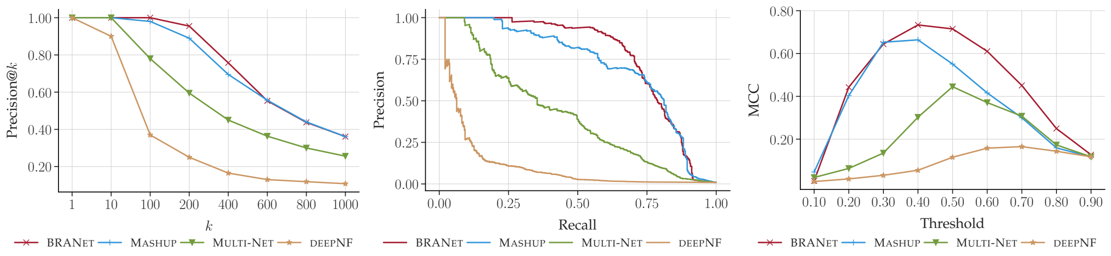
Figure 1: Overview of the BRANET model.

by its supra-adjacency matrix $\bar{\mathbf{A}}$, enhanced with biological a-priori information. Then, for this graph we obtain a random walk-based positive pointwise mutual information (PPMI) matrix (**S**) via a closed-form solution. Matrix **S** is then factorized using Singular Value Decomposition (SVD), and the d-dimensional embedding vectors $\mathbf{\Omega}_d \in \mathbb{R}^{|N| \times d}$ ($d \ll |N|$) are given by its top-d singular vectors. Next, we describe the workflow of BRANET in detail.

Differential expression analysis: We identify $|N|$ important omics features for the selected p datasets. To do this, we measure the level of expression, also known as differential expression analysis of each element in n_i for "control" and "test" samples. Features are selected based on the empirical fold change (FC) threshold that defines up-regulation ($\log_2(FC) \geq 2$) and down-regulation ($\log_2(FC) \leq -2$). For the selected $|N|$ features, co-expression networks are constructed. Intra-omics relationship is defined based on the Pearson correlation coefficient (ρ) for $|\rho| > 0.8$. It provides a metric of similarity among feature's expression level across different samples and conditions.

Construction of a multilayer network: The networks obtained from the above step correspond to intra-omics feature similarity matrices and the biological *a priori* knowledge is the known information about these features. This, for example, could be the binding sites for miRNAs in the $3'$–UTR (untranslated region) of their target mRNAs (genes/transcripts) or the presence of epigenetic mark in the promoter region of a gene. A multilayer network (G) is built using these matrices and biological *a priori* knowledge, that is given by the supra-adjacency matrix $\bar{\mathbf{A}}$ defined as: $\bar{\mathbf{A}} = \bigoplus_p \mathbf{A}^{(p)} + \mathbf{C}$, where $\bigoplus_p \mathbf{A}^{(p)}$ is the intra-layer adjacency matrix and **C** is a block matrix with zero diagonal blocks that stores *a priori* knowledge of inter-layer connections.

Representation learning: To embed nodes from different omics modalities into a common latent space towards capturing relevant information of inter- and intra-omics relationships, we construct a PPMI matrix **S** for graph G. The PPMI matrix is defined by the random walk transition probabilities to traverse nodes within and across layers. Starting from node n in G, a random walk traverses the multilayer graph, moving across neighborhood nodes chosen uniformly at random. This process repeats for a predefined number of walks per node. Nevertheless, for large networks, simulating random walks is computationally expensive and therefore it is not a recommended approach. To address this limitation, we leverage the relationship between random walk-based GRL algorithms that rely on the SKIP-GRAM model (for instance, DEEPWALK) and matrix factorization [7]. In particular, a multilayer random walk matrix (**S**) is defined by com-



(a) Precision@k

(b) Precision-Recall curve

(c) Mathews correlation coefficient

Figure 2: (a) Precision@k for top $1,000$ edges compared to baseline methods. (b) Precision-Recall curve for all inferred edges compared to baseline methods.

puting the closed-form of a properly normalized random walk transition matrix. For any graph G, $\mathbf{S}$ is given by:

$$\mathbf{S} = \log\left\{ \frac{\text{vol}(G)}{bT} \left[\frac{1}{T} \sum_{r=1}^{T} \mathbf{P}^{r} \right] \mathbf{D}^{-1} \right\}, \tag{1}$$

where $\bar{\mathbf{A}}$ and $\mathbf{D}$ are the adjacency and degree matrices of the graph G respectively, $\mathbf{P}$ is the "power" matrix defined by $\mathbf{D}^{-1}\bar{\mathbf{A}}$, and vol$(G)$ is the sum of the node degrees of G. T corresponds to the window size and b is number of negative samples [7]. In order to obtain node embeddings from matrix $\mathbf{S}$, we perform spectral decomposition using SVD, given by, $\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$. Since $\mathbf{S}$ is a real and symmetric matrix, $\mathbf{U}$ and $\mathbf{V}$ correspond to the real eigenvector matrices and $\mathbf{\Sigma}$ is the diagonal eigenvalue matrix. The integrated embedding matrix $\mathbf{\Omega}_d$ of dimension $|N| \times d$ is given by the first d eigenvectors of $\mathbf{S}$, appropriately weighted by the square root of $\mathbf{\Sigma}_d$ as: $\mathbf{\Omega}_d = \mathbf{U}_d\sqrt{\mathbf{\Sigma}_d}$.

Network inference: The similarity score for each omics feature in $\mathbf{\Omega}_d$ is defined by computing the scalar product for each inter- and intra-omics interaction. The integrated network is inferred by selecting the top edges of the nodes that have high *FC* value.

3 Results

Experimental setup: We substantiate BRANET for omics inference. After performing differential expression analysis, $1,070$ omics features are selected ((as mentioned in above section). We compute co-expression networks (> 0.8) for 192 miRNAs and 878 mRNAs. The biological *a priori* is given by the presence of miRNA binding site in the $3'$–UTR region [9]. As mentioned in the section above, a multilayer network G is constructed, whose adjacency is given by $\bar{\mathbf{A}}$. We learn node embeddings using the methodology described in Sec. 2 (Fig. 1). The parameters for constructing the PPMI random walk matrix, such as the window size T and number of negative samples b are set to 3 and 1 respectively, whereas the embedding dimension d is 128. The same pipeline and embedding size are used for the baselines MASHUP, DEEPNF, and MULTI-NET. To infer regulatory interaction from the learned embeddings, we define the similarity between the embedding vectors by computing the scalar product for each miRNA-mRNA interaction.

Evaluation: We investigate the ability of our model to infer regulatory interactions by reconstructing the MMRN for pancreatic cancer datasets. As a ground truth for the evaluation, we use a MMRN available in public databases and demonstrate its effectiveness by comparing to other network integration approaches. In practice, biological networks show small-world topological properties, where nodes are linked by a short chain of acquaintances. These properties could be extracted by focusing on important edges in the graph. In our context of binary edge inference, the precision metric computes the accuracy to retrieve correctly inferred edges. Therefore, to evaluate the performance of graph inference and to retrieve such relevant information, we measure the precision at the top k inferred edges (Precision@k), that corresponds to the number of correctly inferred edges among the top k ones. We choose to study the top $1,000$ edges of the inferred MMRN (Fig. 2a). We compare the performance of our model to the performance of the baseline methods used. As shown in Fig. 2a, our method is able to outperform DEEPNF, MULTI-NET and MASHUP by correctly inferring top 100 edges. Looking at the performance for the top $1,000$ edges, MASHUP proves to perform equally well as the proposed BRANET model.

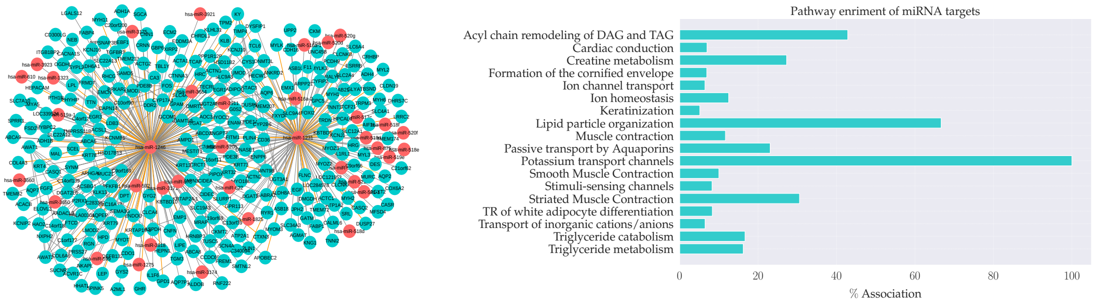
Although we are mainly interested to study the most important edges, we withdraw the network size bias and further measure the performance of our model for all edges. To do this, we have computed the area under the Precision-Recall curve (AUPR) for the whole inferred network (Fig. 2b). More formally, Precision $= \frac{TP}{TP+FP}$; Recall = $\frac{TP}{TP+FN}$ where, *TP*: true positives; *FP*: false positives; and *FN*: false negatives. The Precision-Recall curve shows the trade-off between Precision (result relevancy) and Recall (measure of how many truly relevant results are returned) for different thresholds. High area under the curve (AUPR) represents high precision and recall, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. In the obtained results, we observe that the AUPR of BRANET (0.76) is higher than MASHUP (0.74), DEEPNF (0.110) and MULTI-NET (0.42). Moreover, since biological datasets are usually unbalanced (the number of true edges is much smaller than the number of all possible edges), accuracy can provide overoptimistic estimation of the classifier's ability on the class with large number of samples. Therefore, to avoid this bias we compute Matthews correlation coefficient (MCC). The results show that MCC of BRANET for different threshold of edge scores is higher than MASHUP, DEEPNF and MULTI-NET (Fig. 2c). Overall we also observed that, to obtain a well trained model, methods like DEEPNF require large training data involving tuning of numerous parameters. This may face overfitting/underfitting problems. Moreover, the best performance of such baseline models can be achieved by extensive hyper-parameter tuning.

In Fig. 3a, we show the inferred network for the top 500 edges of important nodes (high *FC*) for the selected pancreatic cancer datasets. Nodes in orange and turquoise correspond to miRNAs and mRNA respectively. The edges in yellow are the truly inferred edges that were not present in the input graphs besides are the result of integration. From Fig. 3a, we can observe that this network is driven by two miRNAs (hsa-miR-1246 and hsa-miR-1231) that are well studied miRNA biomarkers for pancreatic cancer research. In addition to this, Fig. 3b shows the pathway enrichment of miRNA targets in Fig. 3a. Ion channels have been well studied and are associated with the malignant phenotype of cancer cells and contribute to all basic cellular processes such as proliferation, differentiation, and apoptosis. Potassium transport channels show the highest association, also known to play role in pancreatic duct adenocarcinoma [10].

To summarize the empirical analysis, the performance of BRANET (Fig. 2a, 2b and 3a) is especially appealing mainly because of three reasons. First, our approach integrates experimental data with biological *a priori* knowledge which facilitates the inference of inter-omics relationships. Second, it can generate meaningful embeddings by preserving the inter- and intra-omics interactions. Third, its objective function is independent of the downstream task (MMRN inference), thereby it is adaptable to various omics data inference tasks.

4 Conclusion

Recent wide application of high-throughput experimental techniques has provided complex high-dimensional protein association data; in turn, the wide availability of these omics data have driven the need for the development of methods that can take advantage of this heterogeneous data. We have presented, an integrative analysis of



(a) Visualization of inferred MMRN

(b) Pathway enrichment of miRNA targets

Figure 3: (a) MMRN for the top $5,000$ edges. Nodes in orange correspond to miRNAs, while mRNAs are indicated with turquoise. Edges in yellow and grey represent truly inferred edges that are novel (not present in the input network). (b) Pathway enrichment of miRNA targets that are down-regulated in pancreatic cancer omics datasets.

multilayer heterogeneous networks for learning low-dimensional omics feature representations from different data types. BRANET relies on a GRL technique to learn embeddings that can capture relevant omics features from complex networks. Besides, we have presented a preliminary performance analysis comparing our approach with state-of-the-art integration methods.

In the future work, we intend to look at the added value of biologial *apriori* by solely considering expression datasets. Besides, we also intend to explore integration on other data types, such as epigenetic marks, protein sequences, and structures. Represented as similarity networks, these data types can aid researchers to give direction towards more accurate identification of biomarkers and insilico drug discovery.

Acknowledgements. Supported in part by ANR (French National Research Agency) under the JCJC project GraphIA (ANR-20-CE23-0009-01).

References

[1] I. Subramanian, S. Verma,S. Kumar, A. Jere, K. Anamika. “multi-omics data integration, interpretation, and its application”. *Bioinformatics and Biology Insights*, 14, 2020

[2] X. Yue, Z. Wang, J. Huang, S. Parthasarathy, S. Moosavinasab, Y. Huang, S. M. Lin, W. Zhang, P. Zhang, H. Sun. “graph embedding on biomedical networks: methods, applications and evaluations”. *Bioinformatics*, 36(4), pp. 1241–1251, 2019

[3] W. L. Hamilton, R Ying, J. Leskovec. “representation learning on graphs: methods and applications”. *IEEE Data Eng. Bull*, vol. 40, no. 3, pp. 52-74, 2017

[4] H. Cho, B. Berger, J. Peng. “compact integration of multi-network topology for functional analysis of genes”. *Cell Systems*, vol. 3, no. 6, pp.540-548, 2016

[5] A. Bagavathi, S. Krishnan. “multi-net: a scalable multiplex network embedding framework”. *Complex Networks and their Applications, Springer, Cham*, pp. 119-131, 2018

[6] V. Gligorijević, M. Barot, R. Bonneau. “deepNF: deep network fusion for protein function prediction”. *Bioinformatics*, 34(22), pp.3873-3881, 2018

[7] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, J. Tang. “network embedding as matrix factorization: Unifying DeepWalk, LINE, PTE, and node2vec”. *WSDM*, pp. 459-467, 2018

[8] T.D. Le, J. Zhang, L. Liu, H. Liu, J. Li. “miRLAB: an R based dry lab for exploring miRNA-mRNA regulatory relationships”. *PloS One*, 10(12), p.e0145386, 2015

[9] S. Griffiths-Jones, R.J. Grocock , S. V. Dongen, A. Bateman, A.J. Enright. “miRBase: microRNA sequences, targets and gene nomenclature”. *Nucleic Acids Research*, 34(suppl$_1$), pp.D140-D144, 2006

[10] V. Hofschröer, K. Najder, M. Rugi, R. Bouazzi, M. Cozzolino, A. Arcangeli, G. Panyi, A. Schwab. “ion channels orchestrate pancreatic ductal adenocarcinoma progression and therapy”. *Frontiers In Pharmacology*, 11, p.1979, 2020