Briefings in Bioinformatics, 22(6), 2021, 1–15

https://doi.org/10.1093/bib/bbab270
Problem Solving Protocol

Elucidation of dynamic microRNA regulations in cancer progression using integrative machine learning

Haluk Dogan, Zeynep Hakguder, Roland Madadjim, Stephen Scott, Massimiliano Pierobon and Juan Cui

Corresponding author: Juan Cui, Department of Computer Science and Engineering, University of Nebraska-Lincoln, Lincoln, NE 68588-0115, USA. Tel: +1 402-472-5023; E-mail: jcui@unl.edu

Abstract

Motivation: Empowered by advanced genomics discovery tools, recent biomedical research has produced a massive amount of genomic data on (post-)transcriptional regulations related to transcription factors, microRNAs, long non-coding RNAs, epigenetic modifications and genetic variations. Computational modeling, as an essential research method, has generated promising testable quantitative models that represent complex interplay among different gene regulatory mechanisms based on these data in many biological systems. However, given the dynamic changes of interactome in chaotic systems such as cancers, and the dramatic growth of heterogeneous data on this topic, such promise has encountered unprecedented challenges in terms of model complexity and scalability. In this study, we introduce a new integrative machine learning approach that can infer multifaceted gene regulations in cancers with a particular focus on microRNA regulation. In addition to new strategies for data integration and graphical model fusion, a supervised deep learning model was integrated to identify conditional microRNA-mRNA interactions across different cancer stages. **Results:** In a case study of human breast cancer, we have identified distinct gene regulatory networks associated with four progressive stages. The subsequent functional analysis focusing on microRNA-mediated dysregulation across stages has revealed significant changes in major cancer hallmarks, as well as novel pathological signaling and metabolic processes, which shed light on microRNAs' regulatory roles in breast cancer progression. We believe this integrative model can be a robust and effective discovery tool to understand key regulatory characteristics in complex biological systems. **Availability:** http://sbbi-panda.unl.edu/pin/

Key words: gene regulatory network; miRNA binding; graphical models; Markov random field; Bayesian network; Gaussian process; functional analysis

Introduction

Dysregulation of gene expression in human disease represents a highly complex process involving various different mechanisms. In addition to transcription factors (TFs), microRNAs (miRNAs), a class of small non-coding RNAs, have been identified that can bind to the complementary sequences on their target mRNAs, act as post-transcriptional gene silencers through mRNA degradation and translation inhibition, and participate in numerous physiological processes including cancers [1, 2]. It has been recently revealed that miRNAs regulate target genes in a dynamic and conditional manner where the dramatic complexity can be explained by RNA competitive binding and

Haluk Dogan is a PhD student in the Department of Computer Science and Engineering (CSE) at the University of Nebraska- Lincoln (UNL). His research interests include artificial intelligence, machine learning and graphical models.
Zeynep Hakguder is a PhD student in the CSE department at UNL. Her research interest is in machine learning and artificial intelligence.
Roland Madadjim is an MS student in the CSE department at UNL, who has started his training in biomedical informatics.
Stephen Scott is an associate professor in the CSE department at UNL. His primary research focuses on artificial intelligence and machine learning.
Massimiliano Pierobon is an associate professor in the CSE department at UNL. His research area is molecular communication and bioinformatics.
Juan Cui is an associate professor in the CSE department at UNL. Her primary research interests include systems biology, biomedical information, and machine learning.
Submitted: 15 April 2021; **Received (in revised form):** 7 June 2021

multifaceted gene regulation [3–7]. One miRNA can simultaneously bind to various target mRNAs, long non-coding RNAs and circular RNAs, and meanwhile, one gene can be regulated by multiple miRNAs.

Competition takes place when different miRNAs target the same transcript at the same or close binding regions. Differently, cooperation exists when multiple miRNAs bind to different non-overlapping regions or different copies of the same target transcript [3–6]. Understanding the impact of miRNAs regulation in human disease is important but challenging as the dynamic nature of miRNA interactions, as well as the evolving functions, is largely undetermined. Moreover, other factors such as methylations, long non-coding RNAs (lncRNAs) and genetic mutations, and especially copy number variations (CNVs), also affect the equilibrium of gene expression in a certain biological condition. Each of the aforementioned mechanisms stresses a different dynamic property of gene regulation, underscoring the need of a causal model that can integrate all kinds of interactions.

Expression correlations between gene-gene and miRNA-gene were commonly used as strong indicators for interactions [8–11]. Prior studies have explored methods such as Bayesian [12], regression [13, 14] and machine learning [15] in studying gene regulations by both TFs and miRNAs. For example, we have previously built a meta-Lasso regression model based on a comprehensive set of genomic profiles, including both gene and miRNA expression in various cancer conditions, CNVs and DNA methylation, and sequencing-detected TF binding sites and miRNA-mRNA interactions [7]. The model considered the altered gene expressions as resulting from a combination of various regulations. Based on the conditional miRNA-gene interactions derived from each model, modularized miRNA regulation was assessed based on its involvement in fundamental human pathways [16]. By integrating cancer genomic data from The Cancer Genome Atlas (TCGA), we identified novel regulatory modules where participating miRNAs jointly bind to functionally related genes in different types of cancer [7]. This work demonstrated how modeling and information fusion can facilitate the discovery of miRNA competitive binding in human cancers, as well as modulated miRNA regulation. However, this previous model was focused on miRNA regulation at each individual gene, not accounting for gene-gene interactions, and therefore showed limitations in general functional analysis.

In this study, we explore integrated solutions to modeling miRNA-gene interactions by constructing comprehensive gene regulation networks (GRNs) because of two major reasons. First, sophisticated causal network models such as Bayesian Networks (BNs) enable us to use the power of causality and infer regulatory relationships between genes and miRNAs [1, 17–20]. Second, such probabilistic models are robust to noise in data, which makes this method appealing in biological data analysis where experimental and technical errors are inevitable. Despite these strengths, we are also aware of challenges in the following aspects. (i) Effective network fusion, referring to the integration of heterogeneous interaction networks inferred from different models. For example, authors of previous works have applied network analysis on the basis of generally predicted miRNA-gene networks to identify specific sub-networks associated with the condition of interest, e.g. human cancers [9, 21, 22], and none has fully addressed the multi-layer network fusion problem in a real biosystem. In BNs, inclusion of a prior derived from sequence-related features in the miRNA-gene regulation model may improve the structure learning stage but still fail to fully capture the complex dynamics throughout entire networks, where competition and cooperation are involved. Therefore, we need to design a more generalized framework for model integration while keeping the proper causality. (ii) Effective information fusion, referring to the integration of heterogeneous data analyses that reflect distinct regulatory mechanisms. For example, each type of high-throughput data such as microarray or RNA-seq-based expression profiles, CLIP- or CLASH-based miRNA-RNA interactions, and ChIP-seq TF binding profiles, as well as methylation and genetic profiles from DNA sequencing analysis can be used to infer a certain type of molecular interaction. Effective information fusion can transform a static interaction analysis into a semi-conditional interaction analysis, leading to more practically useful results. (iii) Computational feasibility. Learning BNs from data is an NP-hard problem; therefore, constraint- and heuristic-based structure learning algorithms should be considered to reduce the search space [23–25]. To this end, we can adopt a Markov Chain Monte Carlo-based structure learning algorithm (e.g., Madigan *et al.* [26]), which reproduces Markov chains on possible graphs by simulating adding, removing and reorienting edges to sample graphs preprocessing. For example, imbalanced sample size remains one of the major challenges in modeling complex cancer processes due to the difficulty in collecting large-scale samples, e.g. much fewer early stage tumor samples compared with available normal control samples. In most TCGA datasets, there are skewed numbers of samples in different cancer types or stage groups, which may cause biases when building and comparing models.

To summarize, we propose a new methodology based on mixed graphical models to infer novel regulatory mechanisms underlying cancer development and progression. We believe that a systematic understanding of the structural difference of GRNs across different phenotypes can shed light on cancer progression factors from regulatory perspectives. Our objectives is twofold: (i) we aim to model causal relations in GRNs by utilizing an information-theoretic approach, which prevents the learned cancer-related GRNs from deviating from the GRNs of normal samples; (ii) we aim to find confounding factors, identify indirect causal and evident effects, common causes and effects between the variables, as well as identifying their biological functions. We apply this analysis to breast cancer data to demonstrate the use and power of this new approach.

Materials and Methods

Datasets

RNA-Seq data on both gene and miRNA expression were collected on The Cancer Genome Atlas Breast Carcinoma by using GDCRNATools R package [27]. Common samples with both gene and miRNA data available were extracted, which involves 104 solid tissue normal and 1072 tumor samples (Stage 1: 179, Stage 2: 608, Stage 3: 242, Stage 4: 20). After Trimmed Mean of M (TMM) normalization (29) and Voom transformation [28], analysis of differentially expressed genes (DEGs) were performed by limma [29] and edgeR [30] methods. DEGs with fold-change less than 2 are filtered out and only the common DEGs (1218 up-regulated, 1236 down-regulated) found by the limma and edgeR were used for the downstream analysis.

The miRNA-mRNA interactome profiles obtained from starBase database [2] reveal more than 2,500,000 reported CLIP/-CLASH interactions where 863 066 interactions are identified in different breast cancer cell lines BT474, MCF7 and MDA-MB-231 [31–34].

Data augmentation

In our dataset, each stage group has different numbers of samples. In order to reduce the effect of size difference and

conduct an unbiased experiment, we used Conditional Variational Autoencoder (CVAE) [35] to produce equal number of samples across groups. Similar to Variational Autoencoder (VAE) [36], CVAE as shown in Figure 1a is also a generative model but it supports to generate data label. The objective function of VAE is optimizing log likelihood of data $P(X)$ as

$$\log P(X) - D_{KL}\left[Q(z \mid X) \parallel P(z \mid X)\right] = E[\log P(X \mid z)] - D_{KL}\left[Q(z \mid X) \parallel P(z)\right]. \quad (1)$$

The VAE model has two parts: the encoder $Q(z \mid X)$ and the decoder $P(X \mid z)$, where z denotes latent variables. Since VAE directly models z both in encoder and decoder parts, it cannot generate data for a given class. However, CVAE takes the class variable into consideration when optimizing its objective function in

$$\log P(X \mid c) - D_{KL}\left[Q(z \mid X, c) \parallel P(z \mid X, c)\right] = E[\log P(X \mid z, c)] - D_{KL}[Q(z \mid X, c) \| P(z \mid c)], \quad (2)$$

where c denotes the class label, and both encoder and decoder now are conditioning on latent variables z and class labels c.

The detailed description about the robustness evaluation and the augmentation effect on network inference was provided in Supplementary Table 1, see Supplementary Data available online at http://bib.oxfordjournals.org/.

Gene regulation network reconstruction

Figure 1b depicts the overall workflow to construct a GRN that explains regulatory mechanisms based on expression and interactome profiles.

First, a Gaussian graphical model (GGM) was explored to explain the dependency relationship between genes, the variables in a continuous multivariate system. In order to learn the underlying relationships embedded under complex GRN, we assume that our data are sampled from the following multivariate Gaussian distribution:

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \quad (3)$$

where $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Sigma}$ is the covariance matrix. $\boldsymbol{\Sigma}$ is a square positive definite matrix and $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is called precision matrix. We can rewrite the formula in Equation 3 for $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Omega}$ as

$$p(x_1, x_2, \ldots, x_n \mid \boldsymbol{\mu} = 0, \boldsymbol{\Omega}) = \frac{|\boldsymbol{\Omega}|^{1/2}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\sum_i \omega_{ii}(x_i)^2 - \sum_{i<j}\omega_{ij}x_i x_j\right). \quad (4)$$

Equation 4 can be considered as a continuous Markov Random Field with potentials defined on every node and edge where $\omega_{ii}(x_i)^2$ is a node potential denoted as $\phi(x_i)$, and $\omega_{ij}x_i x_j$ is an edge potential denoted as $\phi(x_i, x_j)$. Given $n \times p$ data matrix $\boldsymbol{X}$, where n is the number samples, p is the number of genes, and observations $x_1, \ldots, x_n$ are independent and identically distributed (i.i.d) and sampled from $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is $p \times p$ positive definite matrix. Two variables p_i and p_j are conditionally independent if and only if $\boldsymbol{\Omega}[i, j] = 0$ [37]. The problem for learning the conditional independence relationship between the variables with the given data becomes now estimating the coefficients ω_{ii} and ω_{ij} shown in Equation 4.

The scaled log-likelihood of a sample $x \in \mathbb{R}^p$ in a GGM with mean $\boldsymbol{\mu}$ and precision matrix $\boldsymbol{\Omega}$ is, up to a constant given by

$$\mathcal{L}(\boldsymbol{\Omega}, \boldsymbol{x}) \equiv \log\det(\boldsymbol{\Omega}) - (\boldsymbol{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Omega}(\boldsymbol{x}-\boldsymbol{\mu}). \quad (5)$$

We define the average scaled log-likelihood of N samples $x^{(1)}, \ldots, x^{(n)}$, which depends only on sample covariance matrix $\widehat{\Sigma}$ as

$$\mathcal{L}(\boldsymbol{\Omega}, \widehat{\boldsymbol{\Sigma}}) \equiv \frac{1}{N}\sum_n \mathcal{L}\left(\boldsymbol{\Omega}, x^{(n)}\right) = \log\det(\boldsymbol{\Omega}) - \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}). \quad (6)$$

It is common to encounter sparse networks in real-world applications. We impose a sparsity assumption in our learning problem because of three reasons: (i) biological networks are often sparse [38]; (ii) computations on dense graphs require huge amount of resources and (iii) dense graphs are difficult to interpret. Banerjee *et al.* [39] showed that finding the sparse precision matrix which fits the best to a dataset is an NP-hard problem. Additionally, $p \times p$ covariance matrix Σ requires $O(p^2)$ parameters for accurate estimation; however, we often have $n \ll p$. Therefore, some form of regularization can be used to make the computation tractable. Structured sparsity can be obtained by regularizing with ℓ_1-norm. Our goal is to solve the following regularized maximum likelihood problem by minimizing regularized minus log-likelihood as follows:

$$\min_{\boldsymbol{\Omega}>0} \mathcal{L}(\boldsymbol{\Omega}) := \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}) - \log\det(\boldsymbol{\Omega}) + \lambda\|\boldsymbol{\Omega}\|_1. \quad (7)$$

Equation 7 is a convex optimization problem where regularization parameter $\lambda > 0$, and linear term $(\mathrm{tr}(\widehat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}))$, the negative log determinant function $(\log\det(\boldsymbol{\Omega}))$, the ℓ_1 penalty and the set of all positive definite matrices are convex. The solution to the convex optimization problem in Equation 7 is known as the graphical lasso [40]. Learning the structures using the observations in different groups separately does not take into consideration the similarities between their structures. In fact, the structure of a graphical model on a single sample group should not deviate much from the rest. Since differences between the graphical models are of interest, Danaher *et al.* [41] proposed a technique for jointly estimating multiple graphical models. They solved the following optimization problem subject to constraint that $\boldsymbol{\Omega}^1, \ldots, \boldsymbol{\Omega}^{(K)}$ are positive definite:

$$\min_{\{\boldsymbol{\Omega}>0\}} \mathcal{L}(\{\boldsymbol{\Omega}\}) := \sum_{k=1}^{K} \mathrm{tr}(\widehat{\boldsymbol{\Sigma}}^{(k)}\boldsymbol{\Omega}^{(k)}) - \log\det(\boldsymbol{\Omega}^{(k)}) + P(\{\boldsymbol{\Omega}\}), \quad (8)$$

where $P(\{\boldsymbol{\Omega}\})$ denotes a convex penalty function. They defined two regularization functions to foster the precision matrices to share certain characteristics. Their first proposed regularization function, fused graphical lasso as shown in Equation 9, applies ℓ_1 regularization for sparsity constraint, and the fused lasso [42] penalty regularization function to the differences between corresponding elements of each pair of precision matrices to

encourage similar edge values.

$$P(\{\mathbf{\Omega}\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{ij}^{(k)}| + \lambda_2 \sum_{k<k'} \sum_{i,j} |\omega_{ij}^{(k)} - \omega_{ij}^{(k')}|, \quad (9)$$

where λ_1 and λ_2 are non-negative tuning parameters. Second regularization function they proposed, group graphical lasso as shown in Equation 10, also applies ℓ_1 regularization for sparsity constraint, and the group lasso penalty [43] to the (i, j) element across all K precision matrices in order to have an identical pattern of non-zero elements in the precision matrices.

$$P(\{\mathbf{\Omega}\}) = \lambda_1 \sum_{k=1}^{K} \sum_{i \neq j} |\omega_{ij}^{(k)}| + \lambda_2 \sum_{i \neq j} \sqrt{\sum_{k=1}^{K} \omega_{ij}^{(k)^2}}. \quad (10)$$

We tune the regularization parameters λ_1 and λ_2 using an approximation of Akaike information criterion (AIC) defined as follows:

$$\mathrm{AIC}(\lambda_1, \lambda_2) = \sum_{k=1}^{K} \left[n_k \mathrm{tr}\left(\hat{\mathbf{\Sigma}}^{(k)} \hat{\mathbf{\Theta}}^{(k)}_{\lambda_1,\lambda_2} \right) - n_k \log \det(\hat{\mathbf{\Theta}}^{(k)}_{\lambda_1,\lambda_2}) + 2E_k \right], \quad (11)$$

where K is the number of class, E_k is the number of non-zero elements in $\hat{\mathbf{\Theta}}^{(k)}_{\lambda_1,\lambda_2}$ and n_k is the number of observations in class k. Grid search can be used to select λ_1 and λ_2 that minimizes $\mathrm{AIC}(\lambda_1, \lambda_2)$ score. When the number of variables p is too large, computing $\mathrm{AIC}(\lambda_1, \lambda_2)$ for large number of (λ_1, λ_2) pairs is a computationally intensive task. One way to remedy this problem is first searching λ_1, fixing its optimal value for a search over λ_2.

miRNA-gene binding network

Next, we learned a BN to represent the binding relationship between miRNAs and genes. All the reported miRNAs-gene interaction sites were collected from the starBase [2] database. Among all 2454 DEGs, starBase reported 166 669 interactions for 2030 genes that involve 617 unique miRNAs. We used these interactions to build the evidence matrix E, where $E[i, j] = 1$ if there is a reported interaction between gene i and miRNA j in starBase. We first used the Greedy Hill Climbing method to find initial directed acyclic graphs (DAGs), and then applied the Tabu search algorithm starting with those DAGs with tabu size 100 and a maximum of 2 changes that decreases the score of model. The Bayesian Dirichlet equivalence uniform (BDeu) [44] scoring implemented in aGrUM package [45] was used for this score-based learning process.

Given many constraint-based, score-based and hybrid structure learning algorithms that have been proposed for BN structure learning, as reviewed and compared by Scutari *et al.* [46], we summarize here why we chose a score-based method in this study. In general, the advantage of score-based methods comes from the ability of formulating the learning problem as an optimization problem. The scoring approach is mainly implemented through two steps, one scoring the candidate structures with respect to given data while the other exploring the search space of structure. Local greedy and heuristics search methods are in necessity as the number of possible DAGs grows exponentially with the number of random variables. Scutari *et al.* [46] has conducted a performance analysis to compare algorithms in different categories. It was found that, based on both simulated and real-world data, constraint-based algorithms do not appear to be more efficient or more sensitive to errors than score-based algorithms and hybrid algorithms are not faster or more accurate than constraint-based algorithms. Tabu search generally outperforms the rest of the algorithms. BDe score was proposed which considers the likelihood equivalence but is still impractical to compute. BDeu overcomes both issues by defining a uninformative prior on model parameters [47]. When we trust our knowledge of the prior distribution, we can make the contribution of the prior to the posterior stronger. Since, in this study, we want the data to dominate the posterior, we set the equivalent sample size to 1 which is significantly smaller than our dataset. Our aim is to make discoveries from data without strong assumptions for which Bayesian scoring methods provide the means.

Once we obtained a DAG for binding network, we converted the DAG to its Markov equivalent undirected graphical model (moralized graph). In the equivalent undirected model, there is an undirected edge between two nodes if they share a directed edge in the original graph or they are parents of the same node. In the breast cancer case, the binding network in the DAG and undirected graphical model have 1678 and 3137 edges, respectively. In order to explain the impact of miRNA-mediated regulation in the gene interaction networks, we used the entropic Gromov-Wasserstein distance [48] to assess the similarity between two phenotypes by including only expressions of genes involved in direct interactions of miRNAs as well as the interactions of their dependencies. As shown in Figure 1C, calculating the distance of normal and cancer expression profiles to understand the impact of miRNA M1 involves both the expressions of direct interaction with G1 and all interactions of its dependencies G3. We only kept the distances if they are greater than the threshold, 0.0127, which is the distance between the normal and cancer profiles based on the entire DEGs. We then ranked the miRNA-mRNA interactions based on the distance with the reasoning that existence of top-ranked binding cases differentiates the regulatory mechanisms in cancer and normal more than other binding sites.

Conditional interactions identified using supervised neural network model

We explore a deep learning solution based on Convolutional Neural Network (CNNs) [49] to elucidate the conditional miRNA-mRNA interaction based on gene and miRNA expression profiles in this study. CNN is a special kind of deep neural network, designed to be spatially invariant and to recognize patterns directly from an input. It is composed of multiple building blocks such as convolution layers, pooling layers and fully connected layers. Early layers of CNN models learn low-level features, while deep layers learn high-level features which are composed of low-level features. We designed multiple layers of two-input 1D CNN model to discover spatial gene and miRNA features. In our CNN architecture as shown in Figure 1d, we added max pooling layers after convolution layers to reduce spatial dimensions which also helps to control overfitting. Additionally, we added dropout layers after dense layers to control overfitting. We truncated the CNN architecture at concatenation layer and saved the weights in the last dense layers before concatenation layers after training is complete.

We collected miRNA-gene binding interactions reported by CLASH, CLIP experiments for BT474, MCF7, MDA-MB-231 breast cancer cell lines limited to our DEG set from starBase. A total of

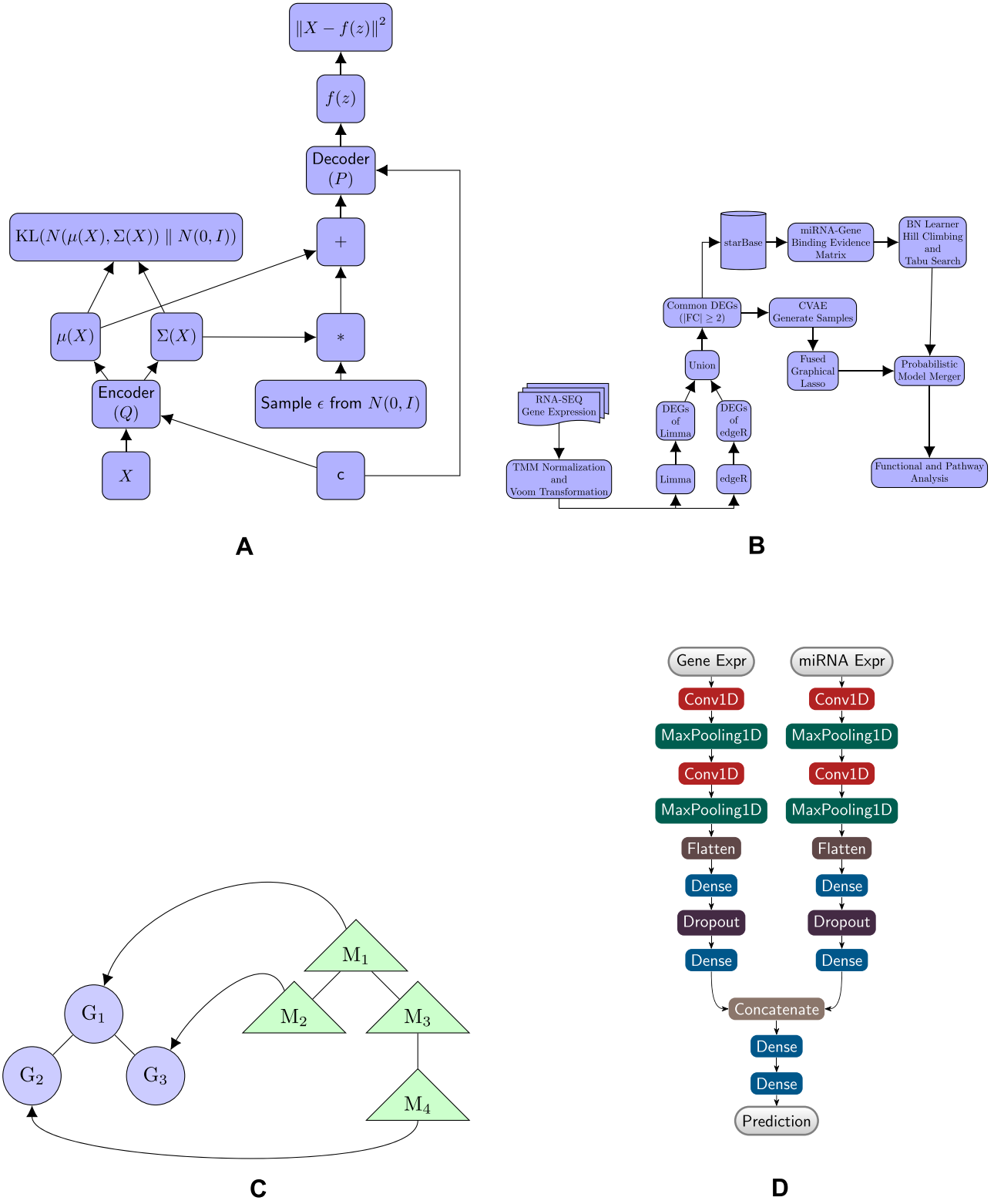

Figure 1. **(A)** The CVAE architecture. c denotes the class label and z denotes latent variables. Reparameterization trick allows fast sampling from distributions of normal and cancer samples learned from original data. This extension of VAE uses class information to learn class specific distributions. **(B)** The overall workflow. Following normalization and differential gene expression detection, our workflow supports generation of new samples to address the limitation of low number of observations compared with the number of variables. Our framework utilizes the miRNA-gene interactions as evidence in our probabilistic model. Inferences are strengthened by these evidences. Our workflow supports further functional analysis based on the differences in stage specific networks. **(C)** Illustration of connecting GRN and miRNA co-binding models. To remove all effects of a given miRNA, e.g. M1, we need to remove all its immediate targets, namely G1, and indirect targets through its dependencies, in the figure this is G3 due to M2. **(D)** Two input 1D CNN deep learning model. Each block shows the type of layer we used for creating architecture. Last dense layer is a classification layer to predict if there is a binding relationship between given miRNA and gene.

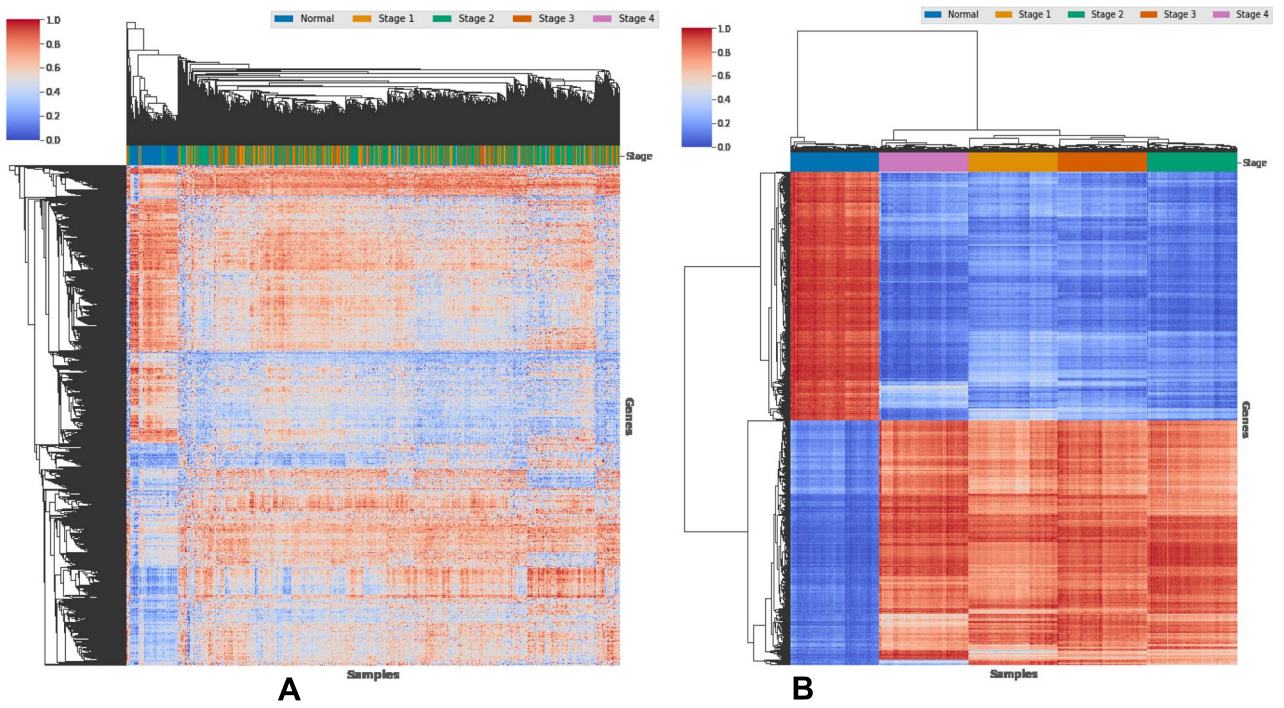

Figure 2. Hierarchical clustering view of (**A**) the original and (**B**) after data augmentation gene expression profiles in the TCGA-BRCA dataset across normal and four stages.

80 199 miRNA-gene interactions reported by at least three different experiments between 596 unique miRNAs and 2454 DEGs. These reported interactions serve as positive instances in our classifier. One branch of the architecture gets gene expression values, and the other branch gets miRNA expression values and they are concatenated down the line for performing binary classification task to predict binding relationships from expression values. We performed batch training with batch size 16 for 100 epochs. As the number of negative labeled pairs is much larger than positives, we sub-sampled equal number of pairs for training 10 different times. We used repeated 5-fold cross validation (CV) where repeat number is 10 for model selection. In addition to interaction prediction, one main objective of creating this CNN model was to demystify perplexing disposition of genes and miRNAs seen in a multidimensional space in a projected learned manifold. We used t-distributed stochastic neighbor embedding (t-SNE) manifold learning algorithm [50] to visualize learned features that are associated with interacted genes and miRNAs in Euclidean space.

Cross-stage functional analysis

We built networks for each progressive stage following the aforementioned steps and then investigated the structural differences of the learned models. Specifically, we focus on gene-gene or miRNA-gene interactions that are newly introduced to each stage and absent in the preceding stage, which are defined as stage-specific interactions. For instance, if there exists an interaction between gene A and gene B in stage 1, and interaction between gene B and gene C in stage 2, then we only consider gene A for stage 1, and gene C for stage 2 for functional analysis. We excluded the common gene B in these interactions from set enrichment analysis because including such a common gene may undermine differences between pairing patterns specific to stages. Based on this information, functional roles enriched in each stage was analyzed through the following approaches. First, genes that interact and function together would more likely be in the same cluster, along with their miRNA regulators. Gaussian Mixture Model (GMM), an unsupervised parametric statistical model, was used for clustering. Specifically, K-component GMM, $\mathcal{G}$, is defined as

$$\mathcal{G} = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{12}$$

where the parameters for each component $\theta_k = \{\pi_k, \mu_k, \Sigma_k\}$ are mixing coefficients, mean and covariance. The covariance matrix of a Gaussian distribution determines volume, shape and orientation of the clusters. Specifying covariance matrix type provides different models that may better explain the structure of the data. Approaching clustering problem from a probabilistic viewpoint, reduces the problem into inferring θ_k. Our goal is to model given data $\mathbf{X}$ with linear superpositions of multiple Gaussians. In order to estimate θ_k, we need to maximize the log-likelihood of the *GMM* given by

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{n=1}^{N} \log \left\{ \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}. \tag{13}$$

Non-linear optimization of the likelihood function is required for high-dimensional datasets. Since there are no closed-form solutions to $\frac{\partial}{\partial\theta} \log p(\mathbf{X}|\boldsymbol{\theta}) = 0$, we used Expectation-Maximization algorithm [51] to find a maximum likelihood solution for the GMM model. Since there is no prior knowledge for optimum number of clusters and the covariance type, we create a model for each combination of covariance type (spherical, tied, diagonal

Table 1. Significantly enriched pathways identified by SPIA as activated

Pathway Name	Status	Cluster IDs
Amphetamine addiction	Activated	9
Bacterial invasion of epithelial cells	Activated	6
Chemokine signaling pathway	Activated	11
Complement and coagulation cascades	Activated	15
Cytokine-cytokine receptor interaction	Activated	11
Dilated cardiomyopathy	Activated	10
ECM-receptor interaction	Activated	4
Fanconi anemia pathway	Activated	8
Focal adhesion	Activated	4
HTLV-I infection	Activated	9
Insulin signaling pathway	Activated	16
Oocyte meiosis	Activated	8
p53 signaling pathway	Activated	8
Salmonella infection	Activated	9
Serotonergic synapse	Activated	6

Table 2. Significantly enriched pathways identified by SPIA as inhibited

Pathway Name	Status	Cluster IDs
Adipocytokine signaling pathway	Inhibited	16
Alcoholism	Inhibited	7
Amoebiasis	Inhibited	4
Cell cycle	Inhibited	8
Fc gamma R-mediated phagocytosis	Inhibited	11
Focal adhesion	Inhibited	6,15
HTLV-I infection	Inhibited	8
Influenza A	Inhibited	11
Malaria	Inhibited	10
Measles	Inhibited	11
Pancreatic cancer	Inhibited	8
Pathways in cancer	Inhibited	8,10,13
PPAR signaling pathway	Inhibited	16
Progesterone-mediated oocyte maturation	Inhibited	8
Systemic lupus erythematosus	Inhibited	7
Tight junction	Inhibited	6

and full) with the number of clusters ranged from 1 to 50. After estimating the parameters for each model, we used Bayesian information criterion (BIC) [48] to select the model that explains data the best.

Given a list of genes in the clusters, the Gene Ontology (GO) Enrichment Analysis was performed to assess their biological significance by testing over-representation of GO terms. We used Fisher's exact test to calculate a p-value determining the probability of identifying that many genes for a given term by chance alone. As we test multiple GO terms simultaneously and these tests are highly correlated, individual p-value of each test is not a good indicator that a term is enriched. Therefore, we used a Benjamini–Hochberg multiple-testing correction with a p-value < 0.05.

Cells undergo aberrant regulation of signaling pathways during the process of cancer development. Further pathway analysis was conducted using Signaling Pathway Impact Analysis (SPIA) [52] to interpret the functional changes of cell signaling by using the topological information of signaling pathways. Based on the altered gene expression, SPIA measures the perturbation in pathways. It outputs two probabilities based on the over-representation of DEGs in a given pathway and the perturbation of the pathway reflected by the gene expression changes propagated along the pathway topology. The first probability represents the significance of the given pathway using over-representation evidence. Modeling the distribution of number of DEGs in a pathway with hypergeometric distribution, SPIA calculates the probability of having DEGs at least as many as the number observed in a particular pathway. The second probability uses the perturbation amount information and represents the probability of having a total perturbation greater than the one observed in the given pathway. A global probability value is obtained by combining these two probabilities and used to rank the pathways to significance test the perturbation.

Results

Stage-specific regulatory networks in breast cancer

In the breast cancer dataset, there are five class labels: normal, stage 1, stage 2, stage 3 and stage 4. We split our dataset to 80 and 20% for training and testing, respectively. In our CVAE architecture, the input, hidden and latent dimensions are 2454, 400 and 50, respectively. We do batch training with the batch size 64 for 1000 epochs and we set learning rate to 0.001. After training is complete, we generate 500 samples for each of the five groups. As we can see in Figure 2a, original expression profile of DEGs does not show clearly different patterns across five groups, which is partially due to the fact that the number of samples in each group is largely imbalanced and insufficient for a decent separation. On the other hand, with generated samples by CVAE as shown in, Figure 2b, it achieves to differentiate samples in different groups as well as improving the distinction of profile for up and down regulated genes. Additionally, clustering analysis shows that normal and cancer samples fall into distinct clusters. Note that the heatmap is used to help demonstrate our model performance visually on generated new data. The idea is not improving the clustering algorithm performance, but to show that model generates meaningful data. Within the cancer cluster, cancer stages 2 and 3 are the closest, while stage 4 is the most distant.

In order to demonstrate that network analysis benefits from this data augmentation process without generating artifacts, we have included an analysis based on simulated data in the supplementary materials. We used AIC criterion for model selection, and found that $\ell_1 = 0.7$ and $\ell_2 = 0.025$ gives the best fitted model to our data. We investigated the structural differences of the learned models for each cancer stage and explained the functional roles by enrichment analysis. The GRN models for each cancer stage are learned together with normal samples. We hypothesized that GRN model for different cancer stages should not deviate much from the GRN for normal condition. Figure 3a shows the numbers of edges in each subset and their intersections. The large number of common edges in the intersection of all GRNs verifies our hypothesis.

Complex biological networks exhibit some non-trivial statistical properties. In the undirected graphs, the degree of a vertex $v, d(v)$, shows the number of neighbors of vertex v. A degree distribution is homogeneous if most of the degree values are close to the average such as Gaussian distribution. On the contrary, a degree distribution is heterogeneous, if most of the nodes have a low degree and a few of the nodes have a very high degree such as power law or exponential distribution.

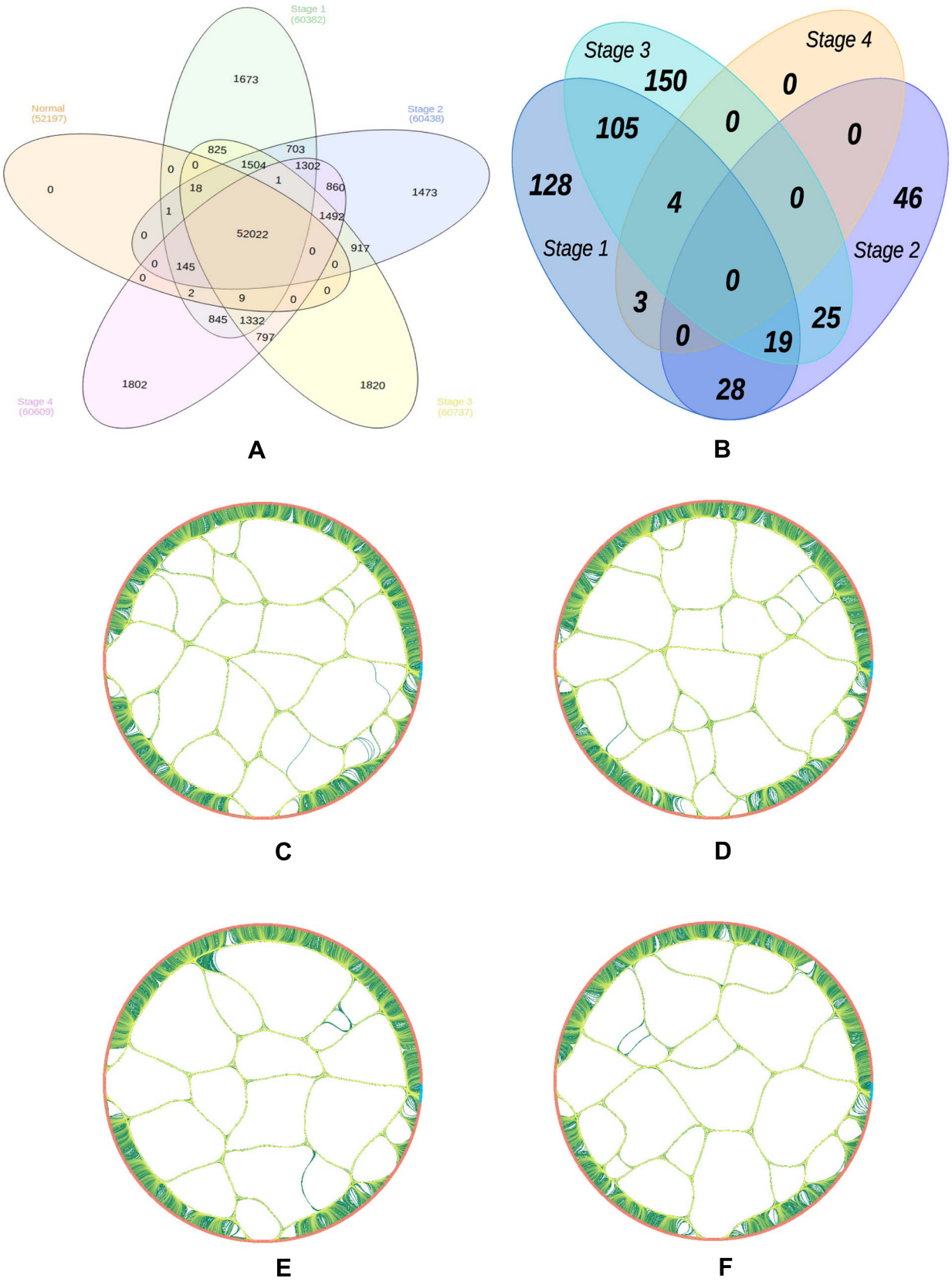

Figure 3. (**A**) Venn Diagram of the numbers of edges detected in each respective graphical models across normal and cancer stages 1–4. (**B**) Venn Diagram of the stage-specific miRNAs across four cancer stages. The networks show interactions that are specific to (**C**) Stage 1, (**D**) Stage 2, (**E**) Stage 3 and (**F**) Stage 4. The red nodes in the network represent genes and the blue nodes represent miRNAs. The nodes are arranged in the same order in each network.

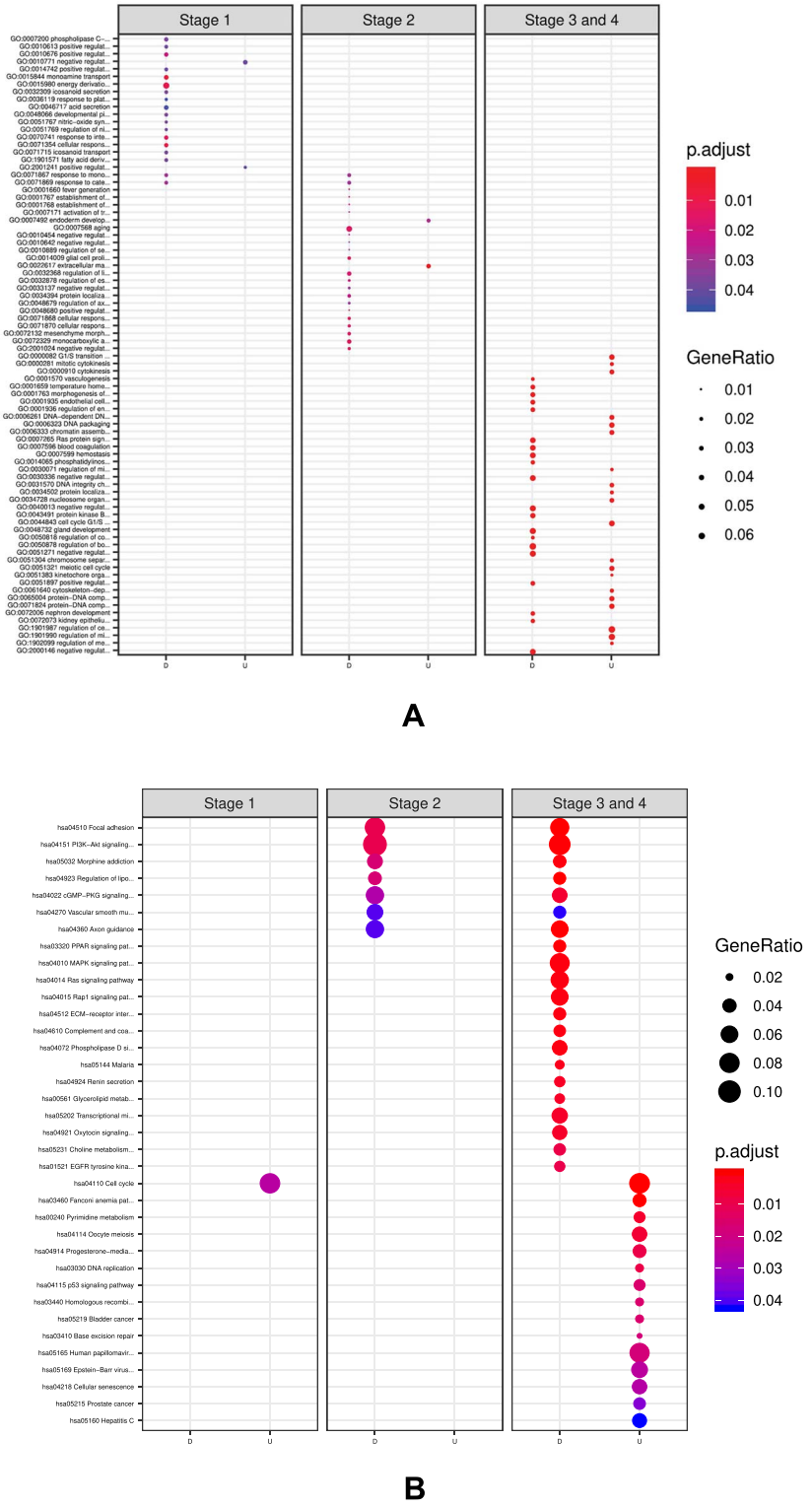

Figure 4. Enriched functional groups in different cancer stages based on enrichment analysis (**A**) on GO Biological Process, (**B**) and KEGG pathways. D: down-regulated; U:up-regulated.

It has been shown that metabolic, protein and gene interaction networks exhibit characteristics of scale-free networks that have heterogeneous degree distribution [53]. The ubiquity myth about scale-free networks still remain controversial. Therefore, we also tested the degree distribution of our final model against those similar to power-law distribution. We showed the degree distribution of our model restricted to vertices of lncRNA, miRNA and protein coding genes in Supplementary Figure 2, see Supplementary Data available online at http://bib.oxfordjournals.org/. In order to compare two distributions against

Figure 5. Enriched functional pathways that are mediated by stage-specific significant miRNAs across different cancer stages.

one another, we compute the likelihood of given data under the two competing distributions. If the likelihood-ratio test is positive, the first distribution is a better fit, otherwise the second. In order to check test significance, we used Vuong's method which gives a *p*-value that shows if the test conclusion is significant or not [54]. Based on the likelihood-ratio test, we found that degree distribution of lncRNAs and protein coding genes follows a stretched exponential distribution, and the degree distribution of miRNAs follows a truncated power-law distribution. This suggests that scale-free networks

are a consequence of preferential attachment process which induces a rich-gets-richer phenomenon. Furthermore, the power-law distribution is plausible for miRNAs due to the fact that miRNAs can bind to several genes. The degree distribution related to lncRNA can be biased because only a limited number of lncRNAs have been curated in the literature and included in this study. It is interesting to see when we performed a goodness-of-fit test based on a list of 1518 interactions reported in lncrna2target [55], the degree distribution also follows stretched exponential distribution.

Functional changes revealed by gene expression and gene-gene interactions across different cancer stages

Functional enrichment and ontology analysis are useful to explain important functions of genes of interest. Genes that have similar expression patterns are more likely to play a role in the same functions. We applied GMM clustering to get similar gene expression profiles in our clusters. A grid search of different priors and number of clusters is performed to find best clustering scheme for the data. The lowest BIC score was obtained for diagonal covariance matrix and 17 clusters. Disruptions in cancerous cells cause over proliferation and failure to control cell growth, division and migration. Given many of these disruptions are associated with cell signaling pathways, we applied the SPIA analysis into clusters of gene expressions in order to get a bird's-eye view of enriched signaling pathways. From the results of SPIA analysis, we deduced important activated and inhibited signaling pathways listed in Tables 1 and 2.

Next, we applied Kyoto Encyclopedia of Genes and Genomes (KEGG) and GO enrichment analysis to discover structural differences across cancer stages. Genes that were common in interactions across stages were omitted from gene list for the enrichment analysis. For example, in the case that gene A and gene B interact only in stage 1, gene A and gene C interact only in stage 2, we include gene B for stage 1 analysis and gene C for stage 2 analysis, omitting gene A from both. In Figure 4, we showed significant functions enriched in more advanced stage during cancer progression, reflected by the structural alteration of GRNs by using dotplot functionality of clusterProfiler R package [56]. For example, KEGG-annotated cell cycle pathway (hsa04110) is the most highly activated process in the stage 1 cancer group. When entering into stage 2, a few signalling pathways such as PI3K-Akt (hsa04151) and cGMP-PKG (hsa04022), and process related to focal adhesion (hsa04510), vascular smooth muscle contraction (hsa04270) and lipolysis regulation in adipocytes (hsa04923) appear to be suppressed. In the advanced stages of cancer, stage 3 and 4, more signaling and metabolic processes were altered with significantly activated cell cycle, oocyte meiosis, DNA replication and p53 signaling. In the meantime, the complemented GO-enrichment analysis provides additional biological processes altered during cancer progression from a slightly different perspective. This analysis provides us a better view of the functional transition when breast cancer progresses from early to more advanced stages.

Functional long non-coding RNAs

With the growth of interactome data derived by emerging sequencing technologies, it is highly compelling to integrate new types of molecular interactions, e.g. between miRNAs and non-coding RNAs such as circular RNAs, and mRNAs and lncRNAs, into the network through a robust system. In this study, we demonstrated how our model can facilitate new discoveries of new functions of those interactions. For example, the lncRNAs, a type of non-coding RNAs whose length are longer than 200 nucleotides, are known to play an important role in human complex diseases [57], but the annotation about disease associations is far from complete [58]. Unknown disease associations of lncRNAs can be uncovered by functional analysis of dependency relations of a given lncRNA. We verified this method on the example of TPT1-AS1. A known disease association of TPT1-AS1 is malignant glioma [59]. The cancerous breast tissue is known to be one of the primary origins of glioma [60]. A widely used practice to identify risk variants is to look in the genomic proximity of known factors, which, however, is argued to be misleading [61]. We used the reconstructed dependency relationships in our network model to find associated diseases for understudied molecules such as TPT1-AS1. Our analysis suggested the enrichment of glioblastoma multiforme which is a subtype of malignant glioma. Particularly, we used Schriml *et al.* [62]'s disease ontology method to provide associations between biomedical data and human diseases. The DOSE package was used to perform disease enrichment analysis for lncRNAs [63]. The significance was evaluated based on hypergeometric test and the expected false positives in a multiple hypothesis setting were adjusted using Benjamini–Hochberg method. Following the same methodology, we found the following enriched disease-lncRNA pairs: GATA3-AS1 and (DOID:2449) acromegaly; PVT1 and (DOID:299) adenocarcinoma; LINC00987 and (DOID:3355) fibrosarcoma, (DOID:8791) breast carcinoma *in situ* and (DOID:8719) *in situ* carcinoma. Note that with particular interest in lncRNA, one can follow the same BN based binding network analysis presented in this study by taking into consideration of experimental lncRNA-mRNAs interactions as priors to infer lncRNA interactions and assess the functions, which might generate more interesting result.

Functional changes revealed by miRNA-gene interactions across different cancer stages

Additionally, we found 617 miRNAs that have have interactions with identified DEGs according to starBase [2]; 20 441 out of 166 669 reported binding interactions make significant group difference by using entropic Gromov–Wasserstein probabilistic distance metric. The average degree of significant miRNAs is 217. The degree distribution of miRNAs marked as significant is shown in Supplementary Figure 3, see Supplementary Data available online at http://bib.oxfordjournals.org/. We consider the miRNAs with smaller degrees (less than 50) to be particularly important. This is because when an miRNA with a high degree is excluded from the dataset, many potentially important dependents are also taken off from the analysis which leads to a greater difference from the baseline. miRNAs with degrees in top-20% are: hsa-miR-129-5p, miR-140-3p, miR-146b-5p, miR-188-5p, miR-193a-5p, miR-28, miR-346, miR-3605-3p, miR-361, miR-455-5p, miR-671-3p, miR-320b, miR-193a-3p, miR-326, miR-330 and miR-501-3p.

We assessed miRNA functional roles as cancer progresses by increasing the miRNA analysis resolution to stage specific level. Stage-specific miRNA-gene interactions are examined in cancer versus normal samples to find sets of miRNAs that significantly change gene expression probability distributions from the baseline normal. In Figure 3b, we show the number of miRNAs that belongs to each subset. It is observed that when the cancer progressed to stage 4, stage specific miRNA activity has diminished which conforms to our beliefs that miRNAs play important roles in key steps towards cancer development; till a very advanced stage, cancer may have evolved by gaining new emergent regulatory mechanisms. We observe the most stage

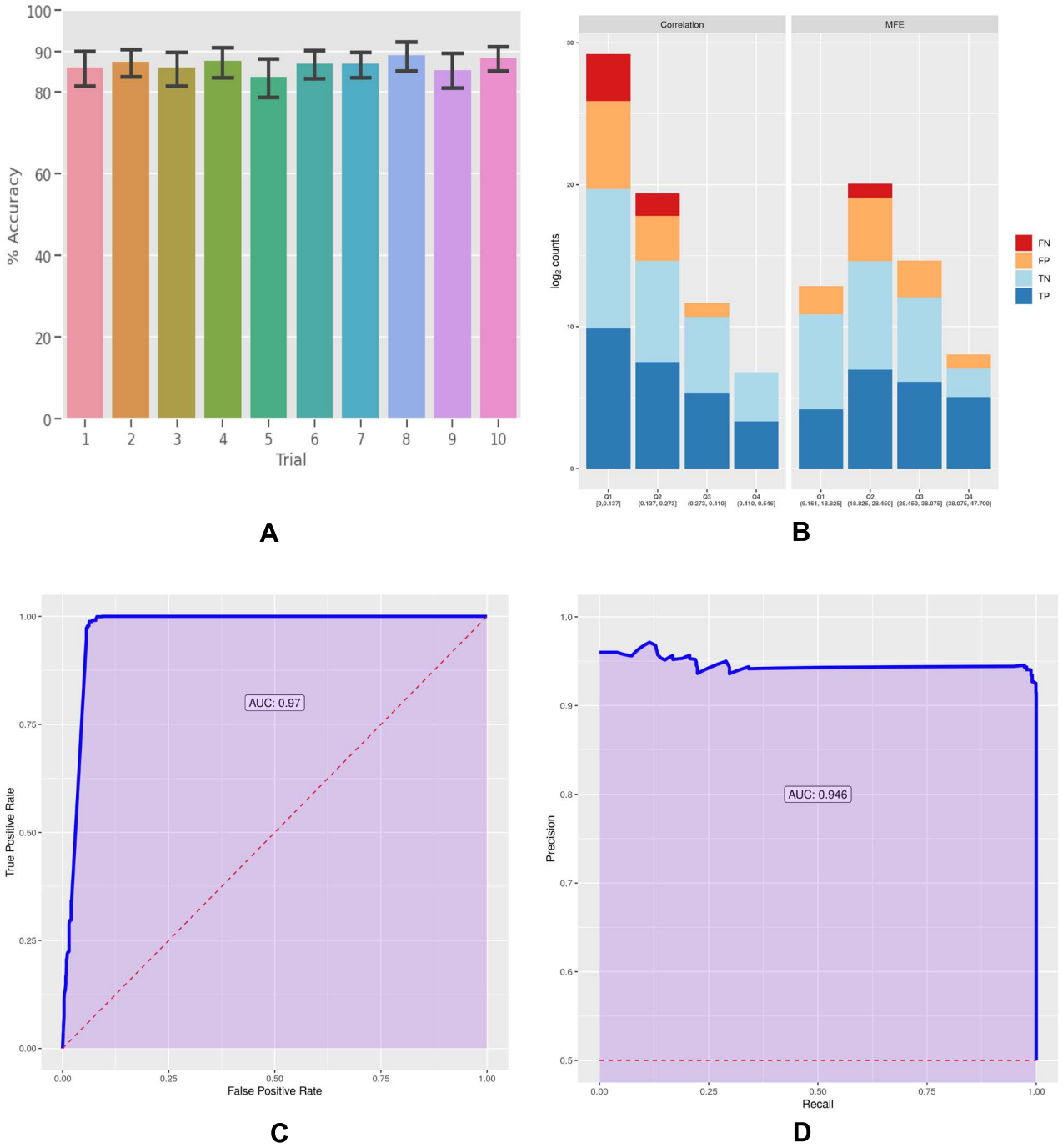

Figure 6. (**A**) The performance on 10-time repeated 5-fold cross validation for 10 different trials of two input 1D CNN model. For each trial, we feed different (gene, miRNA) pairs to model as input. The error bars show the 95% CI. (**B**) Performance conformance analysis shown between model performance and distribution of correlation and MFE values. The absolute value of quantization is used. Depth axis shows the counts in $\log_2$. The correlation does not directly indicate the binding mechanism but the complexes with lower energies are often more stable. The performance on (**C**) ROC. (**D**) Precision-recall (PR) curves for held out test set, where the red dashed lines show the expected performance if there is no learning achieved in the model, while the blue lines follow the performance of ROC and PR curves. The filled with violet area is an area under the curve.

specific miRNA activities in stages 1 and 3 which may lead to transitioning into the next phase.

We further perform a functional analysis on stage-specific miRNAs. In Figure 5, we show the significant functional groups while providing the detailed interactions on http://sbbi-panda.unl.edu/pin/pages-output/mirna/.

In stage 1, we see enrichment in vascular smooth muscle contraction, signaling pathways regulating pluripotency of stem cells, progesterone-mediated oocyte maturation, longevity regulating pathway, growth hormone synthesis, secretion and action, fatty acid degradation/biosynthesis, choline metabolism in cancer, calcium signaling pathway and apoptosis which are

Table 3. Performance of our supervised DL method for predicting significant miRNA-gene binding interactions found by our unsupervised network-based method for all cancer stages

Stage	Number of Significant Interactions	Supervised Method Accuracy
Stage 1	10 897	0.841
Stage 2	7267	0.925
Stage 3	12 421	0.815
Stage 4	58	0

enriched in stage 1 only. These pathways are highly related to the development and initiation. In stage 2, transcriptional misregulation in cancer, Th1 and Th2 cell differentiation, Ras signaling pathway, parathyroid hormone synthesis, secretion and action, PPAR signaling pathway, JAK-STAT signaling pathway, Hedgehog signaling pathway, HIF-1 signaling pathway and ECM-receptor interaction are enriched. These enriched pathways are related to mis-regulation and important known cancer signaling processes. In stage 3, Wnt signaling pathway, phosphatidylinositol signaling system and FoxO signaling pathway are enriched. Wnt and FoxO signaling pathways are shown to be associated with metastasis [64, 65].

Stage-associated miRNA-gene interactions inferred by CNN model

Based on the cross validation test, the CNN model has demonstrated promising prediction power on conditional miRNA-mRNA interactions. In Figure 6a, we show average test accuracies for 10 different trials. Error bars show 95% confidence interval (CI) over 1000 bootstraps. Accuracies fall into range between 80 and 90% in all 10 trials.

Previously, Yuan and Bar-Joseph [66] proposed a 2D CNN model based on 2D histograms. 2D histograms can be used to compare multidimensional factors in multidimensional tensors. We followed their architecture and training procedure, and got $\sim$ 53% testing accuracy. Our proposed 1D CNN model outperformed Yuan and Bar-Joseph [66]'s model by a wide margin. Additionally, we test our 1D CNN model with a held out test set to further assess its performance. We achieved $\sim$ 95.795% accuracy with 1089 true negative, 11 false positive, 88 false negative and 1166 true positive. In Figure 6C and D, we plot receiver operating characteristic (ROC) and precision-recall curves for this test set.

We conducted performance and conformance analysis to investigate the vaguely known role of correlation and minimum free energy (MFE) on gene and miRNA binding mechanism. We used absolute value of quantization to discretize the correlation and MFE values to quartiles, and we checked the overlap of prediction result sets with each quartiles. As we can see in Figure 6b, low correlation and unstable energy values fall into lower halves of the quartiles, and there is no linear relationship between correlation and binding relationship, but the lower the energy values, the more stable the complex.

We also compare the predictions of our unsupervised method based on a probabilistic distance metric and supervised method based on a 1D CNN model for all cancer stages. In Table 3, we showed number of significant miRNA-gene binding interactions reported in the literature by predictions algorithms and CLIP/-CLASH experiments by our unsupervised method, and accuracy of our supervised method for predicting these interactions for stages 1–3.

Furthermore, we projected learned features of genes and miRNAs in lower dimensional space with t-SNE method to interpret molecular structural characteristics of genes and miRNAs seen in our CNN-based model shown in Supplementary Figure 4, see Supplementary Data available online at http://bib.oxfordjournals.org/. We observed that genes' features show similar traits; however, miRNAs are highly dispersed and erratic. This may indicate miRNAs tend to have different properties in terms of expression and binding interactions in general, while genes that appear close together may have similar those properties in breast caner that are partially resulted by regulations among genes.

Conclusion

In this study, we present a systematic solution that can effectively identify miRNA binding by genomics-based modeling on context-dependent GRNs. In addition to common issues on this topic such as data fusion, we also design new strategies for addressing multiple other challenges such as model fusion and functional network comparison. We believe that it is the refinement on each of those key components that eventually leads to a better model. Particularly, gene and miRNA binding networks were inferred based on sequencing-derived expression data and interaction information. Data augmentation was performed to reduce statistical bias for our learner given the imbalanced numbers of samples in normal and different cancer stages. Specifically, we used the entropic Gromov–Wasserstein probabilistic distance metric to measure the effects of each miRNA-gene binding interaction and observed about 1/8 of the reported interactions contribute to significant difference in breast cancer versus control. Functional analysis based on obtained graphic models reveals important signaling processes involving miRNAs and other types of non-coding RNAs during cancer progression. Last, we proposed an unsupervised learning model to identify conditional miRNA-gene binding relationship, which has obtained a good performance. Learned features of genes and miRNAs from our model are visualized to interpret their characteristics. The major contribution of this study is the presentation of an integrative network learner that can merge continuous and discrete data models, and supports queries on variables of interest for interaction predictions, which can be generalized for similar applications in biomedical research.

Key Points

- Computational modeling of gene regulation networks remains a challenging task because of the complex interplay among different gene regulatory mechanisms, as well as the dynamic nature of interactions required in chaotic systems like cancers.
- Given the current challenges in handling network complexity and scalability, we have presented a new integrative learning framework that addresses the model fusion and heterogeneous data integration.
- As one major contribution of this study, the newly presented network learner can merge continuous and discrete data models and supports queries on variables of interest for interaction predictions, which can be generalized for similar applications in biomedical research.
- With a particular focus on microRNA regulation, this study has identified microRNA mediated regulations

associated with different progressive stages of cancer, which can provide new insights in cancer biology and potentially provide new targets for cancer management.
- Additionally, deep learning has shown promising performance in identifying context-dependent interactions based on both gene and microRNA profiles, as demonstrated by a CNN-based deep learning model developed in this study.

Supplementary data

Supplementary data are available online at *Briefings in Bioinformatics*.

Author's contributions

H.D.: Design and implementation of the algorithms, evaluation of the results and drafting the manuscript. Z.H.: Evaluation of the results and drafting the manuscript. R.M. and M.P.: Drafting the manuscript. S.S.: Design of the algorithms, evaluation of the results and drafting the manuscript. J.C.: Design of the algorithms, evaluation of the results and drafting the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank all SBBI members who have been involved in this work for providing helpful discussions and technical assistance. We also appreciate the UNL Holland Computing Center and the Open Science Grid for providing the computational facility.

Funding

This research is funded by the NIH (1P20GM104320), NIH(1R01 DK107264)/NIFA (2016-67001-06314) and UNL Layman seed grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Su W-L, Kleinhanz RR, Schadt EE. Characterizing the role of miRNAs within gene regulatory networks using integrative genomics techniques. *Mol Syst Biol* 2011;**7**:490.
2. Li J-H, Liu S, Zhou H, *et al*. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2014;**42**:D92–7.
3. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism. *and function Cell* 2004;**116**:281–97.
4. Krek A, Grün D, Poy MN, *et al*. Combinatorial microRNA target predictions. *Nat Genet* 2005;**37**:495–500.
5. Cannell IG, Kong YW, Bushell M. How do microRNAs regulate gene expression? *Biochem Soc Trans* 2008;**36**(6):1224–31.
6. Seitz H. Redefining microRNA targets. *Curr Biol* 2009;**19**: 870–3.
7. Shu J, Silva BVRE, Gao T, *et al*. Dynamic and modularized MicroRNA regulation and its implication in human cancers. *Sci Rep* 2017;**7**:13356.
8. Liu B, Li J, Tsykin A. Discovery of functional miRNA-mRNA regulatory modules with computational methods. *J Biomed Inform* 2009;**42**:685–91.
9. Suzuki HI, Young RA, Sharp PA. Super-enhancer-mediated RNA processing revealed by integrative MicroRNA network analysis. *Cell* 2017;**168**:1000–1014.e15.
10. Ding M, Li J, Yu Y, *et al*. Integrated analysis of miRNA, gene, and pathway regulatory networks in hepatic cancer stem cells. *J Transl Med* 2015;**13**:259.
11. Quitadamo A, Tian L, Hall B, *et al*. An integrated network of microRNA and gene expression in ovarian cancer. *BMC Bioinformatics* 2015;**16**(Suppl 5):S5.
12. Chen X, Slack FJ, Zhao H. Joint analysis of expression profiles from multiple cancers improves the identification of microRNA-gene interactions. *Bioinformatics* 2013;**29**:2137–45.
13. Jacobsen A, Silber J, Harinath G, *et al*. Analysis of microRNA-target interactions across diverse cancer types. *Nat Struct Mol Biol* 2013;**20**:1325–32.
14. Balwierz PJ, Pachkov M, Arnold P, *et al*. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res* 2014;**24**:869–84.
15. Kazan H. Modeling gene regulation in liver hepatocellular carcinoma with random forests. *Biomed Res Int* 2016;**2016**:1035945.
16. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
17. Friedman N, Linial M, Nachman I, *et al*. Using Bayesian networks to analyze expression data. *J Comput Biol* 2000;**7**:601–20.
18. Guelzim N, Bottani S, Bourgine P, *et al*. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 2002;**31**:60–3.
19. Imoto S, Goto T, Miyano S. Estimation of genetic networks and functional structures between genes by using Bayesian networks and nonparametric regression. *Pac Symp Biocomput* 2002:175–86. PMID: 11928473.
20. Gendelman R, Xing H, Mirzoeva OK, *et al*. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells. *Cancer Res* 2017;**77**:1575–85.
21. Sadeghi M, Ranjbar B, Ganjalikhany MR, *et al*. MicroRNA and transcription factor gene regulatory network analysis reveals key regulatory elements associated with prostate cancer progression. *PLoS One* 2016;**11**:e0168760.
22. Sumazin P, Yang X, Chiu H-S, *et al*. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 2011;**147**:370–81.
23. Spirtes P, Glymour C. An algorithm for fast recovery of sparse causal graphs. *Soc Sci Comput Rev* 1991;**9**(1):62–72.
24. Chickering DM. Learning bayesian networks is NP-complete. In: *Learning from Data*. New York: Springer, 1996, 121–30.
25. Friedman N, Nachman I, Pe'er D. Learning Bayesian network structure from massive datasets: the "Sparse Candidate" Algorithm. In: Laskey KB, Prade H (eds). *UAI 99: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Sweden, July 30–August 1, 1999*. Stockholm, Sweden: Morgan Kaufmann, 1999, 206–15.
26. Madigan D, York J, Allard D. Bayesian Graphical Models for Discrete Data. *International Statistical Review / Revue Internationale de Statistique* 1995;**63**(2):215.

27. Li R, Qu H, Wang S, *et al*. GDCRNATools: an r/bioconductor package for integrative analysis of lncRNA, miRNA and mRNA data in GDC. *Bioinformatics* 2018;**34**(14):2515–7.
28. Law CW, Chen Y, Shi W, *et al*. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29.
29. Ritchie ME, Phipson B, Wu D, *et al*. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 2015;**43**:e47.
30. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 2010;**26**:139–40.
31. Pillai MM, Gillen AE, Yamamoto TM, *et al*. HITS-CLIP reveals key regulators of nuclear receptor signaling in breast cancer. *Breast Cancer Res Treat* 2014;**146**(1):85–97.
32. Milek M, Imami K, Mukherjee N, *et al*. DDX54 regulates transcriptome dynamics during DNA damage response. *Genome Res* 2017;**27**(8):1344–59.
33. Vanharanta S, Marney CB, Shu W, *et al*. Loss of the multifunctional RNA-binding protein RBM47 as a source of selectable metastatic traits in breast cancer. *Elife* 2014;**3**: e02734.
34. Fish L, Pencheva N, Goodarzi H, *et al*. Muscleblind-like 1 suppresses breast cancer metastatic colonization and stabilizes metastasis suppressor transcripts. *Genes Dev* 2016;**30**(4):386–98.
35. Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models. In: Cortes C, Lawrence ND, Lee DD *et al*. (eds). *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7–12, 2015, Montreal*. Canada: Quebec, 2015, 3483–91.
36. Kingma DP, Welling M. Stochastic gradient VB and the variational auto-encoder. In: *2nd International Conference on Learning Representations, ICLR*, 2014, Banff, AB, Canada.
37. Lauritzen SL. *Graphical models*. New York: Clarendon Press, 1996.
38. Hayden DP, Chang YH, Goncalves J, *et al*. Sparse network identifiability via Compressed Sensing. *Automatica* 2016;**68**:9–17.
39. Banerjee O, El Ghaoui L, d'Aspremont A. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J Mach Learn Res* 2008;**9**:485–516.
40. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 2008;**9**:432–41.
41. Danaher P, Wang P, Witten DM. The joint graphical lasso for inverse covariance estimation across multiple classes. *J R Stat Soc Series B Stat Methodology* 2014;**76**:373–97.
42. Hoefling H. A path algorithm for the fused Lasso signal approximator. *J Comput Graph Stat* 2010;**19**(4):984–1006.
43. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *J R Stat Soc Series B Stat Methodology* 2006;**68**(1):49–67.
44. Buntine WL. Theory refinement on Bayesian networks. In: D'Ambrosio B, Smets P (eds). *UAI 91: Proceedings of the Seventh Annual Conference on Uncertainty in Artificial Intelligence, University of California at Los Angeles, Los Angeles, CA, USA, July 13–15, 1991*. Los Angeles, CA: Morgan Kaufmann, 1991, 52–60.
45. Gonzales C, Torti L, Wuillemin P-H. aGrUM: A graphical universal model framework. In: *Advances in Artificial Intelligence: From Theory to Practice*. Basel, Switzerland: Springer International Publishing, 2017, 171–7.
46. Scutari M, Graafland CE, Gutiérrez JM. Who learns better bayesian network structures: constraint-based, score-based or hybrid algorithms? In: *International Conference on Probabilistic Graphical Models*. Prague, Czech Republic: PMLR, 2018, 416–27.
47. Heckerman D, Geiger D, Chickering DM. Learning bayesian networks: the combination of knowledge and statistical data. *Mach Learn* 1995;**20**(3):197–243.
48. Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;**6**(2):461–4.
49. LeCun Y, Bengio Y. *Convolutional Networks for Images, Speech, and Time Series, page* 255–258. Cambridge, MA, USA: MIT Press, 1998.
50. van der Maaten L, Hinton G. Visualizing data using t-sne. *J Mach Learn Res* 2008;**9**(86):2579–605.
51. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B Methodol* 1977;**39**(1):1–38.
52. Tarca AL, Draghici S, Khatri P, *et al*. A novel signaling pathway impact analysis. *Bioinformatics (Oxford, England)* 2009;**25**:75–82.
53. Broido AD, Clauset A. Scale-free networks are rare. *Nat Commun* 2019;**10**:1017.
54. Vuong QH. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica* 1989;**57**:307.
55. Cheng L, Wang P, Tian R, *et al*. LncRNA2target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res* 2018;**47**(D1):D140–4.
56. Yu G, Wang L-G, Han Y, *et al*. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* 2012;**16**(5):284–7.
57. Fan X-N, Zhang S-W, Zhang S-Y, *et al*. Prediction of lncRNA-disease associations by integrating diverse heterogeneous information sources with RWR algorithm and positive pointwise mutual information. *BMC Bioinformatics* 2019;**20**. https://doi.org/10.1186/s12859-019-2675-y.
58. Ding L, Wang M, Sun D, *et al*. TPGLDA: novel prediction of associations between lncRNAs and diseases via lncRNA-disease-gene tripartite graph. *Sci Rep* 2018;**8**(1). https://doi.org/10.1038/s41598-018-19357-3.
59. Wang W, Yang F, Zhang L, *et al*. LncRNA profile study reveals four-lncRNA signature associated with the prognosis of patients with anaplastic gliomas. *Oncotarget* 2016;**7**(47):77225–36.
60. Piccirilli M, Salvati M, Bistazzoni S, *et al*. Glioblastoma multiforme and breast cancer: Report on 11 cases and clinicopathological remarks. *Tumori J* 2005;**91**(3):256–60.
61. Rice T, Lachance DH, Molinaro AM, *et al*. Understanding inherited genetic risk of adult glioma - a review. *Neuro-Oncol Pract* 2015;**3**(1):10–6.
62. Schriml LM, Arze C, Nadendla S, *et al*. Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res* 2011;**40**(D1):D940–6.
63. Yu G, Wang L-G, Yan G-R, *et al*. DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 2014;**31**(4):608–9.
64. Zhan T, Rindtorff N, Boutros M. Wnt signaling in cancer. *Oncogene* 2016;**36**(11):1461–73.
65. Hornsveld M, Smits LMM, Meerlo M, *et al*. FOXO transcription factors both suppress and support breast cancer progression. *Cancer Res* 2018;**78**(9):2356–69.
66. Yuan Y, Bar-Joseph Z. Deep learning for inferring gene relationships from single-cell expression data. *Proc Natl Acad Sci* 2019;**116**(52):27151–8.