BNrich: A Bayesian network approach to the pathway enrichment analysis

- Samaneh Maleknia[1], Ali Sharifi-Zarchi[2,3], Vahid Rezaei Tabar[4,5], Mohsen Namazi[6], Kaveh Kavousi[1,*]

- [1]Laboratory of Complex Biological Systems and Bioinformatics (CBB),Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.
- [2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.
- [3]Department of Stem Cells and Developmental Biology, Cell Science Research Center, Royan Institute for Stem Cell Biology and Technology, Tehran, Iran.
- [4]Department of Statistics, Allameh Tabataba'i University, Tehran, Iran.
- [5]School of Biological Sciences, Institute for Research in Fundamental Sciences, Tehran, Tehran, Iran.
- [6] Laboratory of Biomolecular Modeling (LBM), Department of Bioinformatics, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran.

*Assistant Professor of Systems Biology and Bioinformatics, Laboratory of Complex Biological Systems and Bioinformatics (CBB),Department of Bioinformatics, Institute of Biochemistry and Biophysics (IBB), University of Tehran

Abstract

Motivation: One of the most popular techniques in biological studies for analyzing high throughput data is pathway enrichment analysis (PEA). Many researchers apply the existing methods without considering the topology of pathways or at least they have overlooked a significant part of the structure, which may reduce the accuracy and generalizability of the results. Developing a new approach while considering gene expression data and topological features like causal relations regarding edge directions will help the investigators to achieve more accurate results.

Results: We proposed a new pathway enrichment analysis based on Bayesian network (BNrich) as an approach in PEA. To this end, the cycles were eliminated in 187 KEGG human signaling pathways concerning intuitive biological rules and the Bayesian network structures were constructed. The constructed networks were simplified by the Least Absolute Shrinkage Selector Operator (LASSO), and their parameters were estimated using the gene expression data. We finally prioritize the impacted pathways by Fisher's Exact Test on significant parameters. Our method integrates both edge and node related parameters to enrich modules in the affected signaling pathway network. In order to evaluate the proposed method, consistency, discrimination, false positive rate and empirical P-value criteria were calculated, and the results are compared to well-known enrichment methods such as signaling pathway impact analysis (SPIA), bi-level meta-analysis (BLMA) and topology-based pathway enrichment analysis (TPEA).

Availability: The R package is available on carn.

Contact: kkavousi@ut.ac.ir

1 Introduction

Over the past few decades, the fast advances of high-throughput technologies have assisted the investigators to comprehensively monitor complicated biological systems. Subsequently, the innovative methods for all-inclusive surveys have been developed [1] providing a summary of the most advantageous information, all the while preserving and enhancing comprehensiveness. The pathway enrichment analysis (PEA) is one of the most popular among these methods [2–5].

The outcomes of the PEA methods can be lead to a definite and limited set of genes and intergenic relations through a pool of biological data, in order to perform later phases of research. These methods can be divided into three major categories under the chronologically sorted list of events [6]. The first generation is over-representation analysis (ORA) which uses just a subset of data (e.g., DEGs); inasmuch as the cut-off threshold is defined by the user, varying results might be obtained via ORA [7,8]. The second generation, the functional class scoring (FCS)

methods, utilize all available expression data, even when information about the network structure of the pathways is missed [9–11]. The last generation, the topology-based pathway analysis (TPA), exploits biological (causal/non-causal) relation between pathway components [12–16]. In most cases, TPA methods provide a more accurate analysis [6]. However, a substantial proportion of these methods consider only the relationship between each gene and its neighbors [17–20] which ignores the complete topology of the network. Consequently, the methods are needed to deliberate the entire structure of the signaling pathways, as much as possible.

In this paper, we introduce BNrich, a new TPA method, based on the Bayesian Network (BN). The literature demonstrated that BN is a worthwhile technique for integrating and modeling biological data with causal relationships [21–25]. Actually, there are some pathway enrichment analysis methods using BN [26] or Dynamic BNs which are learned with time series gene expression data [27].

This study uses the BN to model variations in downstream components (children) as a consequence of change in upstream components (parents). The method employs signaling pathways as underlying structures for the inferred BNs [2,28,29] and gene expression data to estimate its parameters [26,30]. We eliminate cycles of inferred networks based on the biologically intuitive rules instead of using the computational algorithms [31]. In the BNrich, the edges that directly connect a gene product to itself or originated from the nucleus resulting in cycle formation, are removed. The inferred networks are simplified in two steps; unifying genes and LASSO. Likewise, we use originally continuous gene expression data rather than discretized data [26] to BN parameters learning. Given the methods of parameter learning in the BN, we estimate regression coefficients by continuous data [32,33]. The final impacted pathways are gained through Fisher's exact test. This method can represent effective genes and biological relations in the impacted pathways based on the significance level.

We evaluate the proposed method on 187 human non-metabolic KEGG pathways using eight various gene expression datasets. Eventually, we compare the proposed framework with the other pathway topology-based methods, namely, SPIA [12], BLAM [15,34] and TPEA [16].

2 Methods

BNs as probabilistic graphical models that provide arguments in uncertainty framework by a directed acyclic graph (DAG) represent relations among random variables [32,35]. The nodes and arrows in BNs represent random variables and conditional dependence relations, respectively. Here, a group of experts manually proved these causal relationships in the labs [36]. The ability to represent prior data and knowledge in updatable relationships is one of the most important advantages of BNs; therefore, many researchers studied biological data benefiting BNs' stated advantage [24,27,37].

Generally, reconstruction of BNs is implemented in two steps: structure learning and parameter learning. We first attained structures of BNs by KEGG pathways then simplified them. We estimated parameters using gene expression data via *maximum likelihood estimation* (MLE).

2.1 Reconstruction of BN Structures

The KEGG pathways are the collection of directed biological graphs whose biological causal relations have already been confirmed. Because they are not entirely acyclic, some of them require structural modification to eliminate the cycle. In this research these pathways become acyclic through a novel biological strategy. In the following, the simplification of the networks is performed with two steps of unifying genes and the LASSO regression to prepare the BN networks for estimating the parameters.

2.1.1 Reconstruction of BNs' structures utilizing signaling pathways

All the pathways of this study have been extracted from the KEGG pathways database[1]. They include all human, non-metabolic pathways in which gene products are as nodes and their numbers of edges vary between 10 to 1500 [19]. Totally, pathways included in this study are 187 of which 69 contain cyclic pathways. The elimination of the

1- KEGG PATHWAY Database available in: https://www.kegg.jp/kegg/pathway.html (Release 90.0, April 1, 2019)

edges is accomplished through biological approach rather than computational methods [26]. The cyclic pathways have been converted to the acyclic graph while considering their biological perspective as follows:

- Edges that connect two nodes with the same label (gene product) were eliminated.
- According to the high importance of the cytoplasmic processes in this context, the feedbacks originating from nucleus resulting in cycle formation, were discarded.
- The main direction of signaling flux is from the cell membrane to the nucleus, in most of the cases. Therefore, in the remaining cycles, the edges which were opposite of the signaling flux direction and would lead to cycle configuration were removed.

The examples of these rules are displayed in supplementary (Figure S1). All 187 pathways were converted to a directed acyclic graph (DAG) after applying the above rules and eliminating 1,600 edges (less than 0.02 total edges); and we employed them as BNs structures.

2.1.2 BN simplification

Though we extracted the structure of BNs directly from biological pathways, these structures need to be moderated and simplified before estimating the parameters. Thus, while maintaining the ability to estimate the correct parameters, inferencing from the final networks will also be more tractable [38]. This simplification takes place in two stages. The first step is unifying which utilizes classical BN pruning by deleting nodes for which gene expression data is not accessible depending on the dataset platform [39,40]. Similarly, genes that are not in the pathways are also removed from gene expression data to prepare the dataset for the parameter learning. Thereafter, the networks which have at least 5 edges remain in the analysis.

The second step of simplification is driven by the parameter learning model. In BN, the number of outgoing and incoming edges of a node shows the number of children and number of parents, respectively, and the expression level of the node is a function of expression levels of its parents [32].

The parameter learning method in BN depends on the distribution type of the variables including discrete, continuous and hybrid (Scutari, 2010). Here, continuous gene expression data is provided at two conditions of the case (disease or exposed) and control to learning BNs parameters. Hence, the mean value of gene expression for each node can be modeled as a linear regression of its parents' gene expression. On the other hand, in many of the pathway databases including KEGG, a node is merely a position and can represent multiple gene products. Concurrently, many of these nodes are regarded as inputs of a single node such that some of them have up to 65 input nodes (parents). Given the type of estimation of parameters, this large number of parents means a large number of independent variables in the regression line and can lead to the problem of collinearity. To solve the problem, before the final estimation of the parameters, the LASSO regression is used to reduce the number of edges for each network [41,42]. LASSO as a variable selection and regularization method, minimizes the sum of squares of errors by an upper bound on the sum of absolute values of the model parameters as a solution to the l_1 optimization problem [41,42]. This optimization problem can be written as an estimate of the parameter:

$$\hat{\beta}(\lambda) = \operatorname{argmin}_{\beta}\left(\frac{\|\mathbf{Y}-\mathbf{X\beta}\|_2^2}{n} + \lambda\|\beta\|_1\right) \qquad (1)$$

Where $\|\mathbf{Y}-\mathbf{X\beta}\|_2^2 = \sum_{i=0}^{n}(Y_i - (X\beta)_i)^2$ and $\|\beta\|_1 = \sum_{j=1}^{k}|\beta_j|$ and $\lambda \geq 0$ is the parameter that controls the strength of the penalty, the larger the value of λ, the greater amount of shrinkage [41]. The best value for λ was obtained with minimum mean error by using the cross validation (CV) procedure.

The LASSO method was applied for learning BN structure in some studies[43,44]. In this study, the Lasso regression is implemented on each of the remained networks from the previous stage for each dataset. We run LASSO regression for each of the aforesaid network structures for both case and control data. Afterwards, all edges which have zero coefficient for both networks are excluded. Due to computational concerns, the sample size should be greater than 20 across case and control samples.

2.2 BN Parameter learning

In the BN simplification phase, we use LASSO to resolve the collinearity problem using CV to obtain minimum mean error. The LASSO can be applied just on the regression lines with more than one independent variable, while at least a quarter of the nodes in signaling pathway networks are either roots or have just one parent, which means that they have at most one independent variable and hence LASSO cannot be applied to estimate their parameters. To have a unified parameter estimation technique for all nodes and noticing the fact that continuous gene expression data are exploited in the BNrich method [33,45] the maximum likelihood estimation (MLE) technique is carried out for parameter estimation. This estimation is done at two states of the case and control data which share identical structures. For each node Y, $Y = \beta_0 + \sum_{i=1}^{p-1} \beta_i X_i + \varepsilon$, in which , $X_1 \ldots X_{p-1}$ are the parents of Y [32]. In the network, $\beta_1 \ldots \beta_{p-1}$, are interpreted as the weights of the edges and the amounts of biological functions based on the case and control data. The fixed value of β_0 is the intercept of the regression line. In other words, the expected mean value of Y in case that all X_i $(\mathrm{i} = 1, 2 \ldots \mathrm{p} - 1)$ are zero, is β_0. Indeed, β_0 is the amount of variable that is not modeled by its parents values.

2.3 Testing the equality of regression coefficients

Estimated Parameters of Bayesian Networks are derived from the following distribution [46]:

$$\hat{\beta} \approx N(\beta, \sigma^2 (X^T X)^{-1}) \qquad (2)$$

Note that this approximation can also be used for β_0 by adding a column of ones to the matrix X. Statistic σ in above distribution is "standard deviation of the residuals" and was estimated by $\hat{\sigma} = \sqrt{\frac{\sum_{\mathrm{i}=0}^{p-1} \varepsilon_i}{n-p}}$, in which n is the number of samples and p is the number of parameters (β_i), and ε_i is the ith residual $(i = 0, 1, 2 \ldots p - 1)$

Node Y has parents X_1, $X_2 \ldots X_{p-1}$ in the network. When $\hat{Y}_c$ and $\hat{Y}_h$ are the estimations of Y for the case and control datasets respectively which are modeled as follows:

$$\hat{Y}_c = \beta_{c0} + \beta_{c1}\mathrm{X}_1 + \beta_{c2}\mathrm{X}_2 + \cdots + \beta_{c(p-1)}\mathrm{X}_{p-1} + \varepsilon_c \qquad (3)$$

$$\hat{Y}_h = \beta_{h0} + \beta_{h1}\mathrm{X}_1 + \beta_{h2}\mathrm{X}_2 + \cdots + \beta_{h(p-1)}\mathrm{X}_{p-1} + \varepsilon_h$$

We compare corresponding β parameters between each pair of trained networks (for each pathway) in case and control states, by t-test [47] by degree of freedom as follows:

$$df = \frac{(V_c + V_h)^2}{\frac{(V_c)^2}{n_c - 1} + \frac{(V_h)^2}{n_h - 1}} \qquad (4)$$

When $V_h = \frac{S^2(\beta_{hi})}{n_h}$, $S^2(\beta_{hi})$ is estimated variance for i^{th} parameters of considering regression equation in control data, n_h is number of control individuals, $V_c = \frac{S^2(\beta_{ci})}{n_c}$, $S^2(\beta_{ci})$ is estimated variance for i^{th} parameters of considering regression equation in case (exposed or disease) data and n_c is number of case individuals. For the t-tests to yield comparable p-values, all variances and sample sizes are hypothesized to be different.

2.4 Prioritize the impacted pathways

We determine the enriched pathways by Fisher's exact test which was utilized by many researchers for implementing enrichment analysis techniques [48–50]. In this stage, all paired parameters in case and control states are tested and significant parameters in any network are determined. Then, the Fisher's exact test, as an overrepresentation test, is used for each network. The table of Fisher's exact test for kth BN is as follows:

	#Significant parameters	#Non significant parameters
#Parameters are in k^{th} BN	m_k	$n_k - m_k$
#Parameters are NOT in k^{th} BN	$M - m_k$	$N - M - n_k + m_k$

$k \in \{1,2,3, \dots, K\}$; K is the number of remained networks,
m_k= number of significant parameters in the kth BN,
n_k= number of all parameters in the kth BN,
M = number of all significant parameters in all BNs,
N = number of all parameters in all BNs.
All of the trained BNs are examined by Fisher's exact test, then significant pathways are determined as the results of the enrichment analysis. We calculate the adjusted P-value, by the Benjamini-Hochberg False Discovery Rate (FDR) correction method.

2.5 More insight into the impacted pathways

The impacted pathways can be visualized according to the significant threshold of the coefficients chosen by the user. The smaller threshold of the p-value (or FDR) results in the more compact pathway. The nodes' and edges' attributes in k^{th} pathway can be defined as below:

$$Att_{ij}^{k} = (\beta_{ijc}^{k} - \beta_{ijh}^{k}) \times \ln\left(-log10\left(FDR_{ij}^{k}\right)\right) \qquad (5)$$

Where (k =1,…,K) in which K shows the number of final networks and β_{ijh}^{k} is the jth parameter (coefficient) for i^{th} node in the healthy control data when $i \in \{1, 2 \dots m\}$ and $j \in \{0, 1, 2 \dots p-1\}$. The parameter β_{ijc}^{k} is the corresponding parameter (coefficient) of case data in the kth pathway. FDR_{ij}^{k} is also acquired from T-test between these two parameters.
As stated previously, the parameter $\beta_{i0(h/c)}$ shows the ith node attribute for (h)ealthy or (c)ase data. Subsequently, the attribute related to edge from node i to j is calculated by $\beta_{ij(h/c)}$ when $j \in \{1, 2 \dots p-1\}$.
We exemplify a part of Platelet activation pathway enriched in the HT29 (Table1), the human colon cancer cell line dataset, to represent the advantage of the BNrich result visualization (section 3.3).

2.6 Evaluation

2.6.1 Other methods for comparison

Three methods, i.e. TPEA, BLMA, and SPIA, were selected among the topology-based pathway analysis methods to be compared with the proposed method. The topology-based pathway enrichment analysis (TPEA) method, integrates topological properties and takes into account global upstream/downstream positions of genes in pathways [16]. BLMA [15,34] uses the gene expression data directly. It is worth mentioning that BLMA is able to analyze several datasets simultaneously. Finally, the signaling pathway impact analysis (SPIA) method combines the features of the classical enrichment analysis with measurements of the actual perturbation on a given pathway under a given condition [12];although this method is not new, it is still considered as a benchmark to examine new PEA methods [16,51,52].

2.6.2 Evaluation criteria

We evaluate our method based on the four criteria (elaborated on below), namely, consistency, discrimination, False positive rate and empirical P-value by permutation test.

- **Consistency** measures the number of overlapping nodes and edges among independent datasets under similar studying conditions, at a given rank threshold (top 10, 20… 70 significant pathways) [19,52]. We conduct consistency analysis on the last four datasets shown in Table1, related to colon cancer. This analysis has been performed in two phases: consistency analysis among four datasets and consistency analysis along dataset GSE44076 [52] by resampling 70% of data without replacement.

- **Discrimination** measures the difference between results obtained from datasets originating from sufficiently different experimental conditions. These criteria are calculated by 100 times resampling without replacement for any subset 20%, 35%, 50%, 65%, 80% and 95% of datasets [53]. We perform discrimination analysis by first five datasets in Table1.
- **False positive rate** is computed based on true target pathways in two cell line datasets. The cell line data is employed because their target pathways are more robust than the actual human data. Moreover, in contrast to the simulated data, the cell line data is biased towards neither the user preferences nor the evaluation method and leads to more generalized results.
- **Empirical P-values** are attained using the permutation test on cell line datasets A375 (Table1) by generating 1000 simulated datasets in which we randomly assigned samples of the base dataset to two groups. This approach actually gives the observed type I error rate [46,54]. The empirical p-value is calculated by the formula below:

$$\begin{cases} p_{epm}^{k} = \frac{\sum_{i=1}^{1000} \varphi\left(P_{obs_i}^{k} < \alpha_k\right)}{1000} \\ \varphi(x) = \begin{cases} 1; & if\ x\ is\ true \\ 0; & o.w. \end{cases} \end{cases} \quad (6)$$

where p_{epm}^{k} is empirical p-value based on α_k, the real p-value of pathway k achieved by the BNrich. P_{obs_i} is the observed p-value obtained from the ith permutation experiment for the kth pathway and function φ counts the number of observed p-values which are less than α_k in permuted datasets.

3 Results

3.1 Dataset Preparation

We applied eight microarray gene expression datasets of 1383 samples from the NCBI database (Table 1). These datasets include two sets of cell line data [55]. We described characteristics of selected individuals and the target pathways related to the perturbation compounds, in supplementary (T1, Tabel S1 and Tabel S2).

Table 1. Datasets for evaluation and comparison methods

Dataset ID	Study	Tissue type	Sample size (case/control)	Reference
N/A	A375	Tumor cell line skin	110/110	[55]
N/A	HT29	Tumor cell line large intestine	107/110	[55]
GSE93601[a]	Breast Cancer	Tumor breast	144/118	[56]
GSE63060[b]	Alzheimer	Blood	99/99	[57]
GSE44076	Colorectal Cancer	colon mucosa cells	98/148	[58]
GSE47756	Colorectal Cancer	peripheral blood monocytes	28/38	[59]
GSE28000[c]	Colorectal Cancer	colon	34/81	[60]
GSE23878	Colorectal Cancer	colorectal tissue	35/24	[61]

a) Age at diagnosis≥50; estrogen receptor status: Positive; alcohol intake: 0 g/day
b) Age > 60; Alzheimer patients matched by healthy controls in age and sex
c) GPL4133

3.2 BN structure simplification

As mentioned previously, simplifying BNs structures leads to a decrease in the number of edges and possibly networks. The number of edge reductions is different based on different dataset platforms according to the special

set of gene expression afforded by each dataset platform. As it is shown in supplementary (Tabel S3), BNrich resulted in a different number of edges and networks in diverse dataset platforms.

3.3 Insight to the impacted pathways by visualization

As an instance of BNrich output, we illustrated a part of Platelet activation pathway enriched in HT29 cell line dataset visualized by Cytoscape 3.6.1(Figure 1).

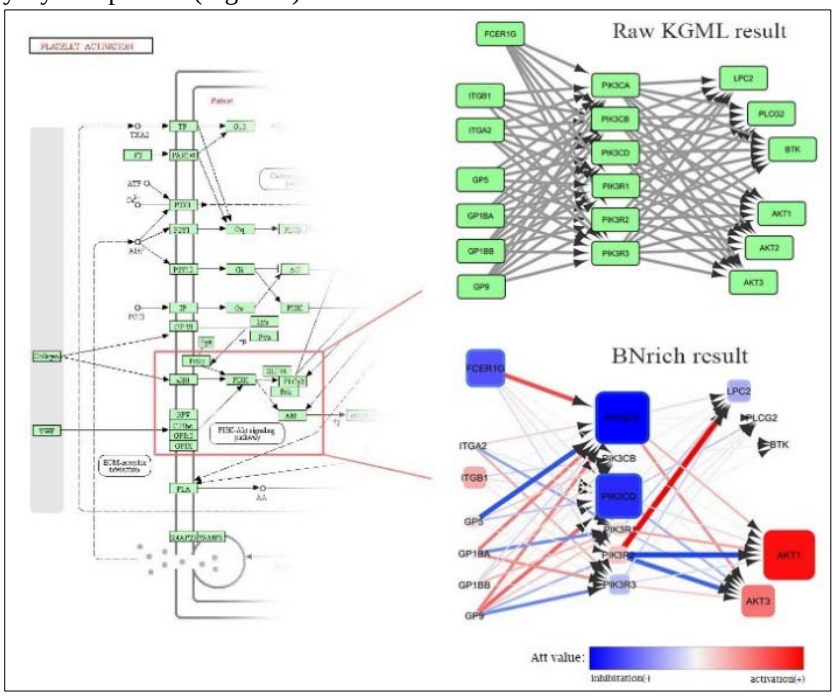

Fig1: The Platelet activation pathway enriched in HT29 dataset. A part of the pathway related to PIK3 gene as a BNrich output is displayed. The color and size of nodes and edges reflect the values and the absolute values of the attribute (Att) respectively.

In accordance with BNrich method stages, at first, Platelet activation pathway was imported as a graphNEL graph with 282 directed edges and 124 nodes. After simplification, a BN with 219 directed edges and 109 nodes was attained. BNrich could determine 204 significant directed edges and 103 significant nodes.

3.4 Evaluation

3.4.1 Consistency

We performed the methods of consistency on the four independent colorectal cancer datasets (the last four datasets in Table 1).

At the given number of top significant pathways, the results obtained by different methods on any pairwise datasets were shown in Figure 2. At the given FDR thresholds, the enriched pathways by different methods on any pairwise resampling datasets from GSE44076 were shown in supplementary (Figure S2).

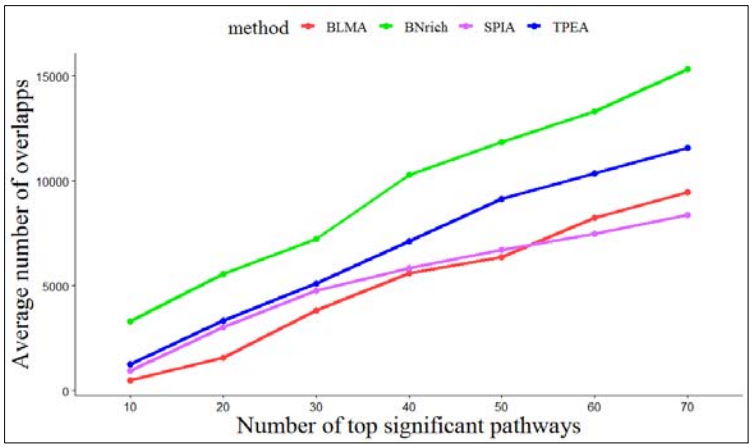

Fig2: Plot of the average number of overlaps among impacted pathways based on the different number of top enriched pathways determined in four colorectal cancer datasets.

3.4.2 Discriminant

The discriminant was computed by results of resampling datasets in four methods. Figure 3 shows the comparison of the discrimination values in GSE44076 vs HT29 datasets, among different methods. These results are achieved taking into account all the enriched pathways from different resampling levels of the first five datasets explained in Table1. All of the discriminant plots were shown in supplementary (Figure S3).

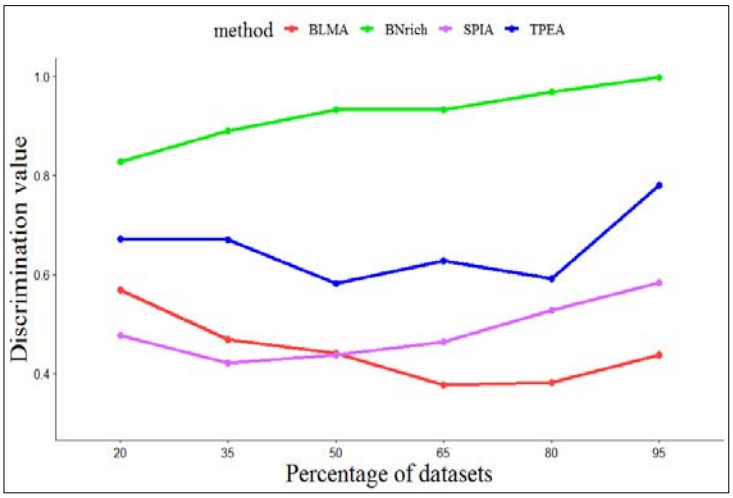

Fig3: The discriminant values of various percentages of sampling among GSE44076 vs HT29

3.4.3 False positive rate

As stated earlier, we calculated the false positive rate for two cell line datasets through determination of correct target pathways (Figure4).

3.4.4 Empirical P-value by permutation test

The empirical P-value was calculated for all pathways in dataset A375 is shown in Figure 5. According to formula 6, for $\boldsymbol{\alpha} < \mathbf{0.05}$, the 75 percent of empirical P-values are less than 0.05 and are demonstrated by the red circle. Indeed, 47 percent of the dots are above the green line.

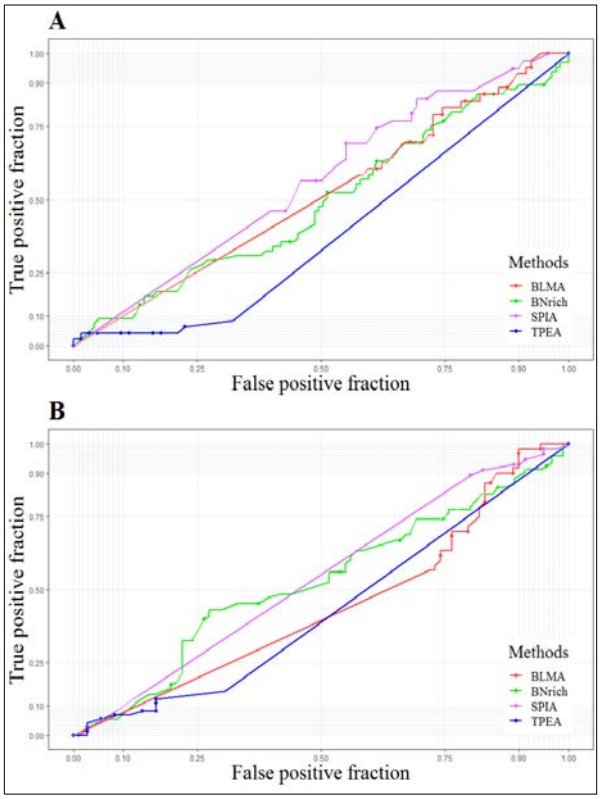

Fig4: The ROC curves according to correct pathways in HT29 cell line dataset (A) and in A375 cell line dataset (B)

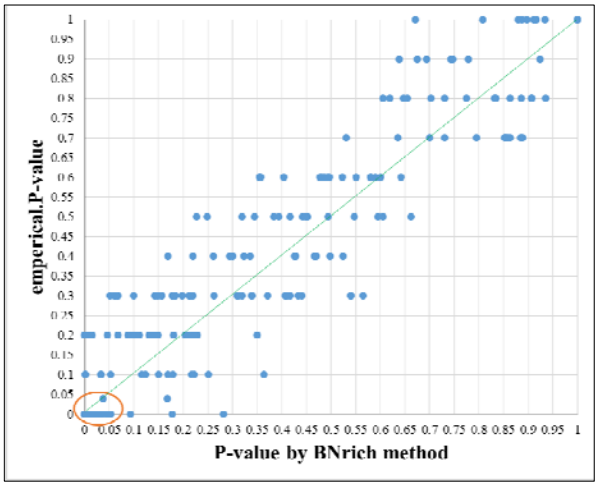

Fig5: Empirical P-value vs. P-value by BNrich method. Among P-values less than 0.05 by BNrich method, 75% of empirical P-values are less than 0.05 and demonstrated by the red circle. Also, 47% of dots are the top of the green line.

4 Discussion

PEA enables investigators to benefit a mechanistic intuition into gene lists produced by genome-scale studies. These methods recognize biological pathways enriched in a gene list rather than anticipating them by chance [62]. On the other hand, the upstream regulators are generally causal genes in complicated diseases and control some downstream gene products; therefore, knowing these causal relationships can detect genes which are more likely to perform a key function in a pathway [16,63]. Thus, among the PEA methods, those that incorporate causal relationships into their computations are biologically preferred[6] .

In the BNrich, the alteration of each component (gene product) is modeled by a linear regression equation with upstream components as regressors, in accordance with Bayesian network characteristics, while continuous gene expression data is applied to estimate parameters. Each gene expression variation is interpreted by two main factors: the level of changes in the gene's parents and its own changes. The impact of changes which are related to $p-1$ parents has been measured by $\beta_1 \ldots \beta_{p-1}$ parameters and illustrated as parameters of edges in BN. As previously stated, the presence of all edges (intergenic relations) has been already validated experimentally in vitro [36], and we only estimated their parameters in any gene expression dataset.

The effect of each individual component regardless of its parents has been shown by β_0 parameter, which represents a node of BN. Consequently, both node and edge variations have been considered for the analysis, relying on the BN hierarchical framework. Further, we simplified the networks and excluded collinearity; consequently, we could decrease spatial/memory complexity without missing vital information via LASSO regression analysis.

The BNrich differs fundamentally from earlier methods. Among the earliest methods, one of the most famous ones is an enrichment analysis method using BNs [26], which may initially appear to be similar to our approach; however, there are, at least, two major differences.

The first difference is related to the modification of the BNs structure. Although in both methods, the BN structures are derived from signaling pathways, the afore-mentioned earlier method removed the cycles of these structures through a completely computational method introduced in 1995 [31] without any biological interpretation, while our method excludes some edges to eliminate cycles in the acquired networks through biological concepts.

Secondly, for learning BN structures, the BNrich, (i.e. our adopted method) utilizes continuous gene expression data, while it is the discretized fold-change data which is exploited in the competing earlier method. As a result, BNrich is more likely to potentially utilize transcriptome profiles. Certainly, another study is required to examine the extent to which the precision and accuracy of final results are affected by the information loss occurring during discretization.

The users can have a survey on enriched pathways in terms of both nodes and edges. For example, in Figure 1, we demonstrated a part of the Platelet activation pathway, which is enriched in HT29 cell line dataset. This cell line was exposed to perturbation compound, Dasatinib, an outcome of which is inhibiting PIK3. PIK3 gene has six paralogs which have been affected differentially by the compound. The colors of the nodes and edges are related to the values of the attribute (Att), while their size is related to absolute values of Att. Thence, among these paralogs, PIK3CA and PIK3CD were the most inhibited paralogs, while PIK3R2 had the strongest biological relations with the next genes in this pathway. It is worth mentioning that these effects are just related to this pathway. Likewise, the researchers need to examine the genes and their relationships in all the enriched pathways in order to gain more accurate insights.

The method of interest was evaluated through a variety of approaches. Figure 2 demonstrated that BNrich is more consistent compared to other methods, and it can identify similar pathways from same clinical condition data in different platforms of datasets. In Figure S2, the BNrich, after TPEA method, is also more consistent than the other two methods. Of course, it should be noted that the number of overlapping nodes and edges is directly related to the number of impacted pathways; as a result, inasmuch as the TPEA method produces many false positives in each run especially in a large dataset, this output is not unexpected.

Considering the second criteria, namely discrimination, BNrich method could better discriminate among specific experimental states in diverse studies; furthermore, with reference to false positive rate, our method results were quite comparable with those of previous reliable studies.

The results of the permutation test revealed that the accuracy of P-value in the introduced method is satisfactory. The mean of empirical P-value in each level (0.01, 0.05, 0.1 . . .) is at least equal to the P-value of the method. The equality of these two values means that BNrich performs as accurately as an exact test. Knowing that the last stage of the method was conducted using Fisher's exact test, this conclusion was expected. Generally, accuracy of BNrich is not be any means less than traditional and acceptable tests.

4.1 Concluding Remarks

The main advantages of the BNrich, as a biologically intuitive method, include: exploiting the structural information acquired from signaling pathways, such as causal relationships between genes, to achieve more precise pathway enrichment in comparison with current well-known methods; controlling the false positive rate in networks with the LASSO regression [64]; applying a bi-stage simplification method for inferring the final structure of networks; and Using an integrative approach which considers both edge and node related parameters to enrich modules in the affected signaling networks.

4.2 Suggestions for Further Research

One way to upgrade the desired algorithm is applying additional information related to phenotype, with the purpose of assigning weigh to edges of the signaling pathways. It can lead to an increase in the speed of execution and a rise in the accuracy of the results while implementing LASSO regression.

Acknowledgements

The authors would like to thank Dr. Marco Scutari from Institute Dalle Molle of Studies on Arti-ficial Intelligence (IDSIA) and Mao Tanabe from Institute of Chemical Research, Kyoto Univer-sity, for their expert insights and suggestions shared via several emails contributing greatly to the quality of this research.

Funding

This work has been supported by Institute of Biochemistry and Biophysics, University of Tehran, Tehran.
Conflict of Interest: none declared.

References

1. Ma, J.; Shojaie, A.; Michailidis, G. Network-based pathway enrichment analysis with incomplete network information. *Bioinformatics* **2016**, *32*, 3165–3174.
2. Lou, S.; Ren, L.; Xiao, J.; Ding, Q.; Zhang, W. Expression profiling based graph-clustering approach to determine renal carcinoma related pathway in response to kidney cancer. *Eur. Rev. Med. Pharmacol. Sci.* **2012**, *16*, 775–780.
3. Guo, Q.; Zhong, M.; Xu, H.; Mao, X.; Zhang, Y.; Lin, N. A Systems Biology Perspective on the Molecular Mechanisms Underlying the Therapeutic Effects of Buyang Huanwu Decoction on Ischemic Stroke. *Rejuvenation Res.* **2015**, *18*, 313–325.
4. Ye, Q.; Lei, L.; Aili, A.X. Identification of potential targets for ovarian cancer treatment by systematic bioinformatics analysis. *Eur. J. Gynaecol. Oncol.* **2015**, *36*, 283–289.
5. Zhang, G.; Yang, P. Bioinformatics Genes and Pathway Analysis for Chronic Neuropathic Pain after Spinal Cord Injury. *Biomed Res. Int.* **2017**, *2017*.
6. García-Campos, M.A.; Espinal-Enríquez, J.; Hernández-Lemus, E. Pathway Analysis: State of the Art. *Front. Physiol.* **2015**, *6*, 383.
7. Khatri, P.; Draghici, S.; Ostermeier, G.C.; Krawetz, S.A. Profiling Gene Expression Using Onto-Express. *Genomics* **2002**, *79*, 266–270.
8. Pavlidis, P.; Qin, J.; Arango, V.; Mann, J.J.; Sibille, E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochem. Res.* **2004**, *29*, 1213–22.
9. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15545–50.
10. Xiong, H. Non-linear tests for identifying differentially expressed genes or genetic networks. *Bioinformatics* **2006**, *22*, 919–923.
11. Jiang, Z.; Gentleman, R. Extensions to gene set enrichment. *Bioinformatics* **2007**, *23*, 306–313.
12. Tarca, A.L.; Draghici, S.; Khatri, P.; Hassan, S.S.; Mittal, P.; Kim, J.S.; Kim, C.J.; Kusanovic, J.P.; Romero, R. A novel signaling pathway impact analysis. *Bioinformatics* **2009**, *25*, 75–82.
13. Gu, Z.; Liu, J.; Cao, K.; Zhang, J.; Wang, J. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Syst. Biol.* **2012**, *6*.
14. Martini, P.; Sales, G.; Massa, M.S.; Chiogna, M.; Romualdi, C. Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic Acids Res.* **2013**, *41*, e19–e19.
15. Nguyen, T.; Tagett, R.; Donato, M.; Mitrea, C.; Draghici, S. A novel bi-level meta-analysis approach: Applied to biological pathway analysis. *Bioinformatics* **2016**, *32*, 409–416.
16. Yang, Q.; Wang, S.; Dai, E.; Zhou, S.; Liu, D.; Liu, H.; Meng, Q.; Jiang, B.; Jiang, W. Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Brief. Bioinform.* **2019**, *20*, 168–177.
17. Zhang, J.; Li, J.; Deng, H.W. Identifying Gene Interaction Enrichment for Gene Expression Data. *PLoS One* **2009**, *4*.
18. Liu, Y.; Koyutürk, M.; Barnholtz-Sloan, J.S.; Chance, M.R. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases. *BMC Syst. Biol.* **2012**, *6*.
19. Han, J.; Shi, X.; Zhang, Y.; Xu, Y.; Jiang, Y.; Zhang, C.; Feng, L.; Yang, H.; Shang, D.; Sun, Z.; et al. ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. *Sci. Rep.* **2015**, *5*.
20. Liu, L.; Wei, J.; Ruan, J. Pathway enrichment analysis with networks. *Genes (Basel).* **2017**, *8*.
21. Yu, J.; Smith, V.A.; Wang, P.P.; Hartemink, A.J.; Jarvis, E.D. Advances to Bayesian network inference for generating causal networks from observational biological data. *Bioinformatics* **2004**, *20*, 3594–3603.

22. Gendelman, R.; Xing, H.; Mirzoeva, O.K.; Sarde, P.; Curtis, C.; Feiler, H.S.; McDonagh, P.; Gray, J.W.; Khalil, I.; Korn, W.M. Bayesian network inference modeling identifies TRIB1 as a novel regulator of cell-cycle progression and survival in cancer cells. *Cancer Res.* **2017**, *77*, 1575–1585.
23. Luo, Y.; El Naqa, I.; McShan, D.L.; Ray, D.; Lohse, I.; Matuszak, M.M.; Owen, D.; Jolly, S.; Lawrence, T.S.; Kong, F.-M. (Spring); et al. Unraveling biophysical interactions of radiation pneumonitis in non-small-cell lung cancer via Bayesian network analysis. *Radiother. Oncol.* **2017**, *123*, 85–92.
24. Agrahari, R.; Foroushani, A.; Docking, T.R.; Chang, L.; Duns, G.; Hudoba, M.; Karsan, A.; Zare, H. Applications of Bayesian network models in predicting types of hematological malignancies. *Sci. Rep.* **2018**, *8*, 6951.
25. Ramos, J.; Das, J.; Felty, Q.; Yoo, C.; Poppiti, R.; Murrell, D.; Foster, P.J.; Roy, D. NRF1 motif sequence-enriched genes involved in ER/PR −ve HER2 +ve breast cancer signaling pathways. *Breast Cancer Res. Treat.* **2018**, *172*, 469–485.
26. Isci, S.; Ozturk, C.; Jones, J.; Otu, H.H. Pathway analysis of high-throughput biological data within a Bayesian network framework. **2011**, *27*, 1667–1674.
27. Kourou, K.; Papaloukas, C.; Fotiadis, D.I. Integration of Pathway Knowledge and Dynamic Bayesian Networks for the Prediction of Oral Cancer Recurrence. *IEEE J. Biomed. Heal. Informatics* **2017**, *21*, 320–327.
28. Zhi-wei, J.; Zhen-lei, Y.; Cai-xiu, Z.; Li-ying, W.; Jun, L.; Hong-li, W.; Zhan-jun, Z.; Man, L.; Peng-tao, L.; Jie, Z.; et al. Comparison of the Network Structural Characteristics of Calcium Signaling Pathway in Cerebral Ischemia after Intervention by Different Components of Chinese Medicine. *J. Tradit. Chinese Med.* **2011**, *31*, 251–255.
29. Fu, C.; Deng, S.; Jin, G.; Wang, X.; Yu, Z.G. Bayesian network model for identification of pathways by integrating protein interaction with genetic interaction data. *BMC Syst. Biol.* **2017**, *11*.
30. Korucuoglu, M.; Isci, S.; Ozgur, A.; Otu, H.H. Bayesian pathway analysis of cancer microarray data. *PLoS One* **2014**, *9*, 1–8.
31. Spirtes, P.; Richardson, T. Directed Cyclic Graphical Representations of Feedback Models. *Proc. Elev. Conf. Uncertain. Artif. Intell.* **1995**, 1–37.
32. Neapolitan, R.E. *Learning Bayesian networks*; first.; Pearson Prentice Hall: Chicago, 2004; ISBN 0130125342.
33. Scutari, M. Learning Bayesian Networks with the bnlearn R Package. *J. Stat. Softw.* **2010**, *35*, 1–22.
34. Nguyen, T.; Mitrea, C.; Draghici, S. Network-Based Approaches for Pathway Level Analysis. *Curr. Protoc. Bioinforma.* **2018**, *61*, 8.25.1-8.25.24.
35. Pearl, J. *Probabilistic reasoning in intelligent systems: networks of plausible inference*; Morgan Kaufmann, 2014; ISBN 0080514898.
36. Kanehisa, M.; Furumichi, M.; Tanabe, M.; Sato, Y.; Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **2017**, *45*, D353–D361.
37. Gorenshteyn, D.; Zaslavsky, E.; Fribourg, M.; Park, C.Y.; Wong, A.K.; Tadych, A.; Hartmann, B.M.; Albrecht, R.A.; García-Sastre, A.; Kleinstein, S.H.; et al. Interactive Big Data Resource to Elucidate Human Immune Pathways and Diseases. *Immunity* **2015**, *43*, 605–614.
38. Allen, D.; Darwiche, A. RC_Link: Genetic linkage analysis using Bayesian networks. *Int. J. Approx. Reason.* **2008**, *48*, 499–525.
39. Shachter, R.D. Evidence absorption and propagation through evidence reversals. *Mach. Intell. Pattern Recognit.* **1990**, *10*, 173–190.
40. Shachter, R.D. Evaluating Influence Diagrams. *Oper. Res.* **2008**, *34*, 871–882.
41. Tibshirani, R. The lasso method for variable selection in the cox model. *Stat. Med.* **1997**, *16*, 385–395.
42. Bühlmann, P.; Geer, S. van de *Statistics for high-dimensional data: Methods, Theory and Applications*; 2011; ISBN 9783642201929.
43. Fu, F.; Zhou, Q. Learning sparse causal Gaussian networks with experimental intervention: Regularization and coordinate descent. *J. Am. Stat. Assoc.* **2013**, *108*, 288–300.
44. Han, S.W.; Chen, G.; Cheon, M.-S.S.; Zhong, H. Estimation of Directed Acyclic Graphs Through Two-Stage Adaptive Lasso for Gene Network Inference. *J. Am. Stat. Assoc.* **2016**, *111*, 1004–1019.
45. Nagarajan, R.; Scutari, M.; Lèbre, S. *Bayesian Networks in R*; Springer New York: New York, NY, 2013; ISBN 978-1-4614-6445-7.
46. Wasserman, L. All of Statistics: A Concise Course in Statistical Inference. In *The American Statistician*; 2004; Vol. 59, pp. 161–163&216–217 ISBN 0387402721.
47. Andrade, J.M.; Estévez-Pérez, M.G. Statistical comparison of the slopes of two regression lines: A tutorial. *Anal. Chim. Acta* **2014**, *838*, 1–12.
48. Kuleshov, M. V.; Jones, M.R.; Rouillard, A.D.; Fernandez, N.F.; Duan, Q.; Wang, Z.; Koplev, S.; Jenkins, S.L.;

Jagodnik, K.M.; Lachmann, A.; et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **2016**, *44*, W90–W97.

49. Hur, J.; Danes, L.; Krout, D.; Hsieh, J.-H.; Auerbach, S. Tox21 Enricher: Web-based chemical and functional enrichment analysis tool for Tox21 toxicity screening platform. *FASEB J.* **2017**, *31*, 613.1-613.1.
50. Mi, H.; Muruganujan, A.; Ebert, D.; Huang, X.; Thomas, P.D. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **2019**, *47*, D419–D426.
51. Zhang, J.-G.; Xu, C.; He, H.; Tian, Q.; Zhou, Y.; Qiu, C.; Deng, H.-W.; Tan, L.-J.; Chen, X.-D. Integrative analysis of transcriptomic and epigenomic data to reveal regulation patterns for BMD variation. *PLoS One* **2015**, *10*.
52. Mansoori, F.; Rahgozar, M.; Kavousi, K. FoPA: Identifying perturbed signaling pathways in clinical conditions using formal methods. *BMC Bioinformatics* **2019**, *20*.
53. Yu, C.; Woo, H.J.; Yu, X.; Oyama, T.; Wallqvist, A.; Reifman, J. A strategy for evaluating pathway analysis methods. *BMC Bioinformatics* **2017**, *18*.
54. Ihnatova, I.; Popovici, V.; Budinska, E. A critical comparison of topology-based pathway analysis methods. *PLoS One* **2018**, *13*.
55. Subramanian, A.; Narayan, R.; Corsello, S.M.; Peck, D.D.; Natoli, T.E.; Lu, X.; Gould, J.; Davis, J.F.; Tubelli, A.A.; Asiedu, J.K.; et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171*, 1437-1452.e17.
56. Wang, J.; Zhang, X.; Beck, A.H.; Collins, L.C.; Chen, W.Y.; Tamimi, R.M.; Hazra, A.; Brown, M.; Rosner, B.; Hankinson, S.E. Alcohol Consumption and Risk of Breast Cancer by Tumor Receptor Expression. *Horm. Cancer* **2015**, *6*, 237–246.
57. Sood, S.; Gallagher, I.J.; Lunnon, K.; Rullman, E.; Keohane, A.; Crossland, H.; Phillips, B.E.; Cederholm, T.; Jensen, T.; van Loon, L.J.C.; et al. A novel multi-tissue RNA diagnostic of healthy ageing relates to cognitive health status. *Genome Biol.* **2015**, *16*.
58. Solé, X.; Crous-Bou, M.; Cordero, D.; Olivares, D.; Guinó, E.; Sanz-Pamplona, R.; Rodriguez-Moranta, F.; Sanjuan, X.; De Oca, J.; Salazar, R.; et al. Discovery and validation of new potential biomarkers for early detection of colon cancer. *PLoS One* **2014**, *9*.
59. Hamm, A.; Prenen, H.; Van Delm, W.; Di Matteo, M.; Wenes, M.; Delamarre, E.; Schmidt, T.; Weitz, J.; Sarmiento, R.; Dezi, A.; et al. Tumour-educated circulating monocytes are powerful candidate biomarkers for diagnosis and disease follow-up of colorectal cancer. *Gut* **2016**, *65*, 990–1000.
60. B., J.; F., A.-P.; C.S., S.; J.K., S.; A.N., M.; J.J., Y.; T., K. Differential gene expression between African American and European American colorectal cancer patients. *PLoS One* **2012**, *7*.
61. Uddin, S.; Ahmed, M.; Hussain, A.; Abubaker, J.; Al-Sanea, N.; AbdulJabbar, A.; Ashari, L.H.; Alhomoud, S.; Al-Dayel, F.; Jehan, Z.; et al. Genome-wide expression analysis of Middle Eastern colorectal cancer reveals FOXM1 as a novel target for cancer therapy. *Am. J. Pathol.* **2011**, *178*, 537–547.
62. Müller, A.C.; Giambruno, R.; Weißer, J.; Májek, P.; Hofer, A.; Bigenzahn, J.W.; Superti-Furga, G.; Jessen, H.J.; Bennett, K.L.; Matsushima, Y.; et al. Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nat. Protoc.* **2019**, *22*, 924–934.
63. Carter, S.L.; Brechbühler, C.M.; Griffin, M.; Bond, A.T. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics* **2004**, *20*, 2242–2250.
64. Wang, L.; Michoel, T. Controlling false discoveries in Bayesian gene networks with lasso regression p-values. *bioRxiv* **2017**, 288217.