Deep Pathway Analysis V2.0: A Pathway Analysis Framework Incorporating Multi-Dimensional Omics Data

Yue Zhao and Dong-Guk Shin

Abstract—Pathway analysis is essential in cancer research particularly when scientists attempt to derive interpretation from genome-wide high-throughput experimental data. If pathway information is organized into a network topology, its use in interpreting omics data can become very powerful. In this paper, we propose a topology-based pathway analysis method, called DPA V2.0, which can combine multiple heterogeneous omics data types in its analysis. In this method, each pathway route is encoded as a Bayesian network which is initialized with a sequence of conditional probabilities specifically designed to encode directionality of regulatory relationships defined in the pathway. Unlike other topology-based pathway tools, DPA is capable of identifying pathway routes as representatives of perturbed regulatory signals. We demonstrate the effectiveness of our model by applying it to two well-established TCGA data sets, namely, breast cancer study (BRCA) and ovarian cancer study (OV). The analysis combines mRNA-seq, mutation, copy number variation, and phosphorylation data publicly available for both TCGA data sets. We performed survival analysis and patient subtype analysis and the analysis outcomes revealed the anticipated strengths of our model. We hope that the availability of our model encourages wet lab scientists to generate extra data sets to reap the benefits of using multiple data types in pathway analysis. The majority of pathways distinguished can be confirmed by biological literature. Moreover, the proportion of correctly indentified pathways is 10 percent higher than previous work where only mRNA-seq and mutation data is incorporated for breast cancer patients. Consequently, such an in-depth pathway analysis incorporating more diverse data can give rise to the accuracy of perturbed pathway detection.

Index Terms—Pathway analysis, Bayesian network, data integration

1 Introduction

Pathway analysis is crucially important in many areas of biomedical research including cancer research. Molecular pathways succinctly record previously known gene and protein regulatory information into networks and high-throughput genomic data is interpreted over the networks. However, pathway analysis research is still rapidly evolving, particularly in the area where multiple genomic data sets are applied to the prior knowledge organized into networks. Previously, we presented a method in which the pathway route-as opposed to the entire pathway itself-is used as the unit of analysis [1], [2]. Nevertheless, this approach has many limitations:

- The model can handle only gene expression data and mutation data and it was not general enough to handle other omics data.
- Many types of regulation events such as phosphorylation and methylation in the pathway were not considered in the model.
- Handling of missing values in the given data set was not taken into account in the model.

The goal of this paper is to extend our previous pathway analysis system, namely, DPA V1.0, to handle additional data sets such as proteomics and CNV data along with dealing with additional regulation event types such as phosphorylation and methylation. The key tenet of our extended model remains the same. By combining even more diversified omics data types, our pathway analysis model should improve the accuracy of identifying perturbed pathway routes. As the unit of our analysis method is a pathway route, our method should help scientists gain a deeper insight into the biological phenomenon that they aim to obtain from their data generation experiments. The rest of the paper is organized as follows. Section 2 reviews existing pathway analysis methods. Section 3 describes the model settings and assumptions in detail. Section 4 presents computational results including a statistical significance study similar to that of PARADIGM [3]. It also includes outcomes of survival analysis and patient subtyping analysis using TCGA Breast Cancer Study (BRCA) and Ovarian Cancer Study (OV). Finally, Section 5 is the conclusion.

2 Related Work

Numerous efforts have been made to study how to use pathway information in analyzing genomics data. The first generation systems used gene set enrichment methods [4].

- *The authors are with the Department of Computer Science and Engineering, University of Connecticut, Storrs, CT 06269 USA. E-mail: {yue.2.zhao, dong.shin}@uconn.edu.*

Manuscript received 30 Oct. 2018; revised 24 Aug. 2019; accepted 24 Sept. 2019. Date of publication 7 Oct. 2019; date of current version 3 Feb. 2021. (Corresponding authors: Yue Zhao and Dong-Guk Shin.) Digital Object Identifier no. 10.1109/TCBB.2019.2945959

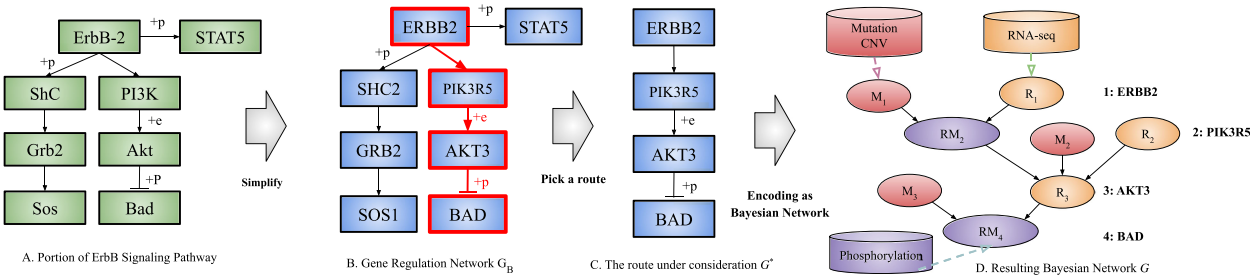

Fig. 1. Conversion Workflow. Part of the ErbB pathway shown in part A is simplified by keeping only specific interactions and genes, resulting in a gene regulation network G_B in Part B. A route G^* shown in part C, starting from ERBB2 to BAD, is extracted from G_B and converted to a Bayesian Network G in part D.

The second generation systems started to use pathways organized into networks.

SPIA [5] measured pathway significance by statistical testing against random permutation. The framework by Korucuoglu et al. [6], [7] encoded the pathway as a Bayesian network. After removing cycles in the graph, they trained the model with expression data. DRAGEN [8] aimed to detect differentially expressing genes by performing a hypothesis testing designed to compute if linear model has identical parameters. Regarding approaches to exploit multi-dimensional omics data, PARADIGM [3] presented a novel method modeling the pathway as a factor graph and introduced a method to perform inference. The approach by Verbeke et al. [9] ranked the pathways by p-value obtained by encoding pathway logic into a global network. The p-value was calculated based on a hypothesis test where the null hypothesis was that the concerned pathway can be picked randomly. Most recently, Altered Pathway Analysis tool (APA) [10] aimed to detect altered pathways by dynamically calculating pathway rewiring through computing correlation between genes, but this system did not use prior knowledge. In contrast, our previous work [2] dynamically encoded pathway routes as a Bayesian network and carried out pathway analysis by incorporating expression and mutation data over the organized network topology. In addition, to the best of our knowledge, this work was the first topology based pathway analysis system as far as identifying biological perturbation in terms of pathway routes.

3 Methods

3.1 Modeling Pathway Graph

A pathway is a graph with biological molecules as nodes and regulation interactions as edges. The edges in the pathway are categorized into two subtypes: 1. Protein activation and inhibition; 2. Gene expression and repression. We note that the terms activation and inhibition are used to capture protein interactions while the terms expression and repression are used to capture transcriptome regulation relationships. This graph modeling convention can be used in many existing pathway databases (e.g., KEGG, Reactome, etc.).

The edges representing protein activation and inhibition model protein interaction relationships typically captured via protein structure and function. We define them as (Protein's) *Functional Interactions*. These edges are further categorized into the following subtypes, namely, phosphorylation (+p), ubiquitination(+u), glycosylation(+g), methylation (+m), dephosphorylation(-p), deubiquitination(-u), deglycosylation(-g), or demethylation(-m), which are represented by tags in the parentheses in the KEGG database as shown in Fig. 1 A. These tags are defined as *Evidence Tag* to indicate the types of interactions occurring between proteins.

The edges representing gene expression and repression model the proteins DNA binding events and their end effects on the expression level change of the involved genes (e.g., a transcription factor binding on its target gene to either initiate or suppress the expression level of the target gene). We call this second type of interactions as *Expression Interactions*. Functional Interactions and Expression Interactions are treated differently in the model. Expression Interaction edges are further categorized into two types, '+e' to denote expression and as '-e' to denote repression.

3.2 From Pathway Route to Bayesian Network

Fig. 1 illustrates the steps for converting a pathway route into a Bayesian network. Fig. 1 A shows a small fraction of the example pathway, ErbB, which has been adapted from the KEGG pathway database [11]. The first step of the conversion process is to simplify the pathway by eliminating the parts that are included in the pathway but are irrelevant to the modeling (e.g, metabolites, small molecules, etc.). During the simplification, node names are standardized into gene names. Fig. 1 B presents the result of this process and we call the resulting graph a gene regulation network G_B. Unlike most existing approaches that merely keep activation and inhibition interactions after the simplification, the Evidence Tag is kept in G_B. The next step is to identify all possible "routes" available from the given G_B. As an example, Fig. 1 C shows a route, G^*, which starts from ERBB2 and ends at BAD. The selected route is then converted into a discrete Bayesian Network, G, as shown by Fig. 1 D. In the end, the Bayesian Network encoding the biological logic in pathway route G^* is integrated with its corresponding omics data to measure the perturbance of G^*.

This method treats each "route" as a unit of pathway analysis. The route-based modeling idea assumes that it is crucial to identify which portion(s) of the pathway is(are) either abnormally activated or suppressed. This route-based method is designed to isolate, for example, whether the effect of ERBB2 amplification is more prominent through ERBB2→SHC2 path or ERBB2→PIK3R5 path, or even through both.

Before introducing the method to integrate G and omics data, we further elaborate the process of converting a

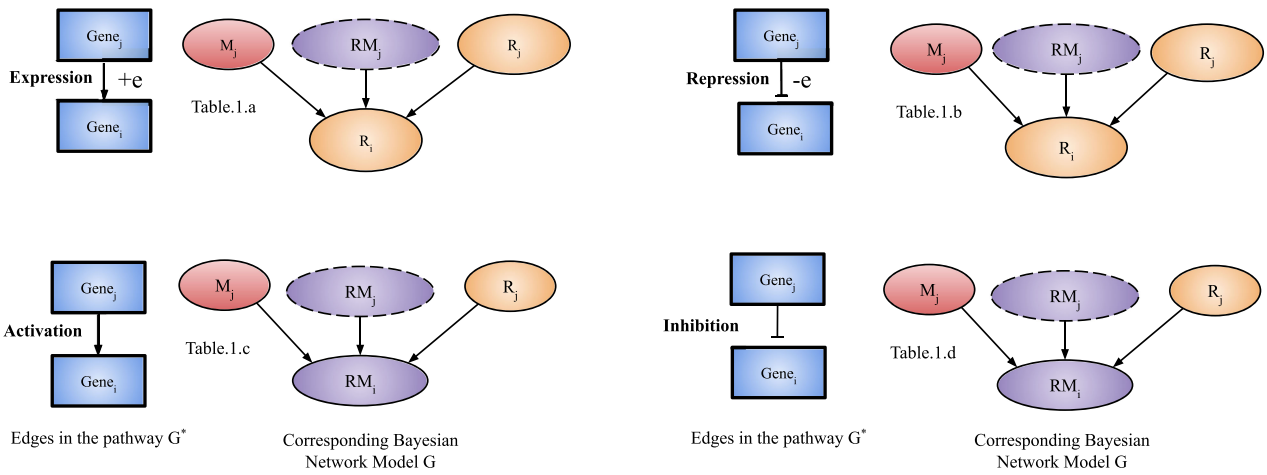

Fig. 2. Converting the edges in the pathway to Bayesian Network. The pathway edges in G^* on the left is converted to Bayesian Network G on the right. The RM_j have a dashed border because it will only exists when the edge regulating g_j is one of the Functional Interactions. In the route setting, we have $j = i - 1$.

pathway route G^* to a Bayesian network G. This process is illustrated in Fig. 2. For each gene, g_i, in the route, three types of nodes are created:

- R_i: a random variable representing expression level status on gene g_i
- M_i: a random variable representing DNA variation status (e.g., mutation) on gene g_i
- RM_i: a random variable representing protein functional status on gene g_i

Fig. 2 illustrates the four different cases, two for Expression Interactions and two for Functional Interactions. In each of four cases, $Gene_i$, three types of nodes M_i, RM_i and R_i are created in its corresponding Bayesian counterpart. Concrete examples for creating such corresponding Bayesian network nodes are illustrated in Fig. 1 D. For example, RM_2, M_2 and R_2 are created for PIK3R5 appearing in G^* of Fig. 1 C.

This setting is motivated from the central dogma of molecular biology as shown in Fig. 3. For a gene regulation network G_B, a route G^* is simply a subgraph of G_B, $G^* \subseteq G_B$, $G^* = (V^*, E^*)$ where $V^* = \{g_1, \ldots, g_{k_{G^*}}\}$, g_i represents the ith gene and k_{G^*} is the number of genes contained in the route G^*, $E^* = \{e_{ij} | 1 \leq i < k_{G^*} \text{ and } j = i+1\}$. For each edge in G^*, $e_{i-1,i}, 1 < i \leq k_{G^*}$, if $i < k_{G^*}$ and $e_{i-1,i}$ is one of the Functional Interactions in G^*, three nodes are created in the corresponding Bayesian Network G : R_i, M_i and RM_i for g_i. On the other hand, if $i < k_{G^*}$ and $e_{i-1,i}$ is one of the Expression Interactions, only two nodes R_i and M_i will be created. The first gene, g_1 will always have two nodes created: R_1 and M_1 while $g_{k_{G^*}}$ will only have one node, either $R_{k_{G^*}}$ (if $e_{k_{G^*}-1,k_{G^*}}$ is Expression Interaction) or $RM_{k_{G^*}}$ (if $e_{k_{G^*}-1,k_{G^*}}$ is Functional Interaction). In this way, there will usually be three nodes for target genes of Functional interactions and two nodes for that of Expression interactions. The two nodes, R and M, represent the expression and DNA variation status, respectively. These two nodes are essential for all genes. However, for Functional Interactions, the protein functional status is included in these two nodes, thus a third node RM is required.

After creating nodes for each gene in the route G^*, the edges in the Bayesian Network G is added dynamically according to the edges in the pathway route G^*. For $g_i \in V^*, 1 < i \leq k_{G^*}$, if $e_{i-1,i} \in E^*$ is one of the Functional Interactions, edges are created pointing from all the nodes for the parent gene g_{i-1} to the RM node for the child gene g_i. Namely, we add edges from $R_{i-1}, M_{i-1}, (RM_{i-1})$, to RM_i. On the other hand, if $e_{i-1,i}$ is an Expression Interaction, edges from the nodes of g_{i-1} $(R_{i-1}, M_{i-1}, (RM_{i-1}))$ to R_i is created instead. The whole process is introduced formally in Algorithm 1.

Algorithm 1. Converting Pathway Interaction G^* to Bayesian Network G

```
1: procedure ConvertRoute(G*)
2:   Create two nodes R_1 and M_1 in G.
3:   for 1 < i ≤ k_{G*} do
4:     if i ≠ k_{G*} then
5:       Create two nodes R_i and M_i in G.
6:     if e_{i-1,i} ∈ Functional Interactions then
7:       Create RM_i in G.
8:     end if
9:     else
10:    if e_{i-1,i} ∈ Functional Interactions then
11:      Create RM_i in G.
12:      else
13:      Create R_i in G.
14:    end if
15:  end if
16:  if e_{i-1,i} ∈ Functional Interactions then
17:      Create edges from R_{i-1}, M_{i-1}, (RM_{i-1}), to RM_i.
18:  else
19:      Create edges from R_{i-1}, M_{i-1}, (RM_{i-1}) to R_i.
20:  end if
21:  end for
22:  return G
23: end procedure
```

For the scenario illustrated in Fig. 1, since ERBB2 is activating (Functional Interaction) PIK3R5 in G^*, two nodes for ERBB2, R_1 and M_1, are created to point to RM_2, i.e., the RM node for PIK3R5. Once activated, PIK3R5's protein binds to AKT3 and that event increases the gene expression of AKT3 (Expression Interaction). In this case, all three types of nodes for PIK3R5 are created to point to R node of AKT3, namely, R_3 as illustrated in Fig. 1 D. In the figure, three distinct node

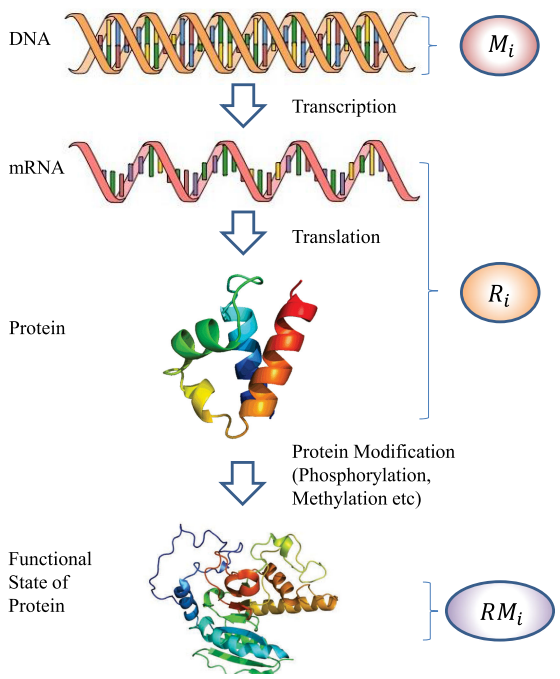

Fig. 3. Central dogma of molecular biology for gene g_i where different nodes for g_i corresponds to a certain stage.

colors are used to represent three different data sources, denoted as cylinders. The node values are imported from their respective data sources. For example, all R node values come from mRNA expression database (RNA-seq cylinder) for the same patient or sample under examiniation.

The conditional probability table corresponding to edges in Bayesian Network G is determined by the type of the edge in G^* as shown in Tables 1 a, 1 b, 1 c, and 1 d. The assumption is that, given the edge $e_{j,i}$, the expression level (R_i) (or the functional status RM_i), of the gene g_i is affected by its parent's expression status R_j, the DNA functional status M_j and the Protein functional status RM_j (if that exists).

3.3 Prior Probability Distribution Associated with the Bayesian Network

After conversion, the resulting Bayesian Network G is formally defined as follows: $G = (V, E)$, where $V = RR \bigcup MM \bigcup RMS$, $RR = \{R_i, i \in \{1, \ldots, k_{G^*}\}\}$. $MM = \{M_i, i \in \{1, \ldots, k_{G^*} - 1\}\}$. $RMS = \{RM_i, i \in \{j : e_{j-1,j}$ is one of Functional Interactions in $G^*\}\}$. M_i, R_i and RM_i are now defined in detail. Since the DNA information is not affected by any interactions in the pathway route and M_i does not have a parent node in G, the random variable M_i follows a discrete distribution as shown in (1). The random variable M_i has two possible values: $+1$ representing that g_i functions normally on DNA level, and -1 representing a function loss, i.e., g_i's DNA original biological function is disrupted. The probability distribution can be set to indicate that the prior has no specific preference on these two levels

$$P(M_i = +1) = P(M_i = -1) = 0.5. \tag{1}$$

Random variable R_i follows a different probability distribution based on the location of gene g_i in path G^*. Suppose g_i is the starting node in G^*, R_i's distribution is shown in (2)

TABLE 1
Conditional Table for Each Type of Edge

(a) THE REGULATION PROCESS e_{ji} IN G^* IS EXPRESSION

$(RM_j\&)M_j\&R_j$	$R_i = +1$	$R_i = 0$	$R_i = -1$
$+1$	$1-\epsilon-\tau$	τ	ϵ
-1	ϵ	τ	$1-\epsilon-\tau$

*$0 < \epsilon < \tau < 1 - \epsilon - \tau$

(b) THE REGULATION PROCESS e_{ji} IN G^* IS REPRESSION

$(RM_j\&)M_j\&R_j$	$R_i = +1$	$R_i = 0$	$R_i = -1$
$+1$	ϵ	τ	$1-\epsilon-\tau$
-1	$1-\epsilon-\tau$	τ	ϵ

(c) THE REGULATION PROCESS e_{ji} IN G^* IS FUNCTIONAL ACTIVATION

$(RM_j\&)M_j\&R_j$	$RM_i = +1$	$RM_i = 0$	$RM_i = -1$
$+1$	$1-\epsilon-\tau$	τ	ϵ
-1	ϵ	τ	$1-\epsilon-\tau$

(d) THE REGULATION PROCESS e_{ji} IN G^* IS FUNCTIONAL INHIBITION

$(RM_j\&)M_j\&R_j$	$RM_i = +1$	$RM_i = 0$	$RM_i = -1$
$+1$	ϵ	τ	$1-\epsilon-\tau$
-1	$1-\epsilon-\tau$	τ	ϵ

The conditional probability table for each of four types of edge e_{ji} is displayed. The first column represents the conditional value for parent gene $g_j, j = i - 1$, the other columns represent the values for child gene g_i. The operator & is defined in (3).

$$P(R_i = +1) = P(R_i = -1) = P(R_i = 0) = 1/3, \tag{2}$$

where +1 represents gene g_i up-regulated in expression level, -1 represents gene g_i down-regulated in expression level and $R_i = 0$ for neutral. However, if g_i ($i > 1$) has a parent, i.e., gene g_{i-1} in G^*, R_i follows the conditional probability table in Table 1 a (Table 1 b) if $e_{i-1,i}$ is expression(repression) in G^*. In order to illustrate our model more clearly, we define the following operator & in the conditional tables. It is similar to the AND operator, as shown in (3). For instance, $(RM_{i-1}\&)M_{i-1}\&R_{i-1}$ has the value of $+1$ if none of the three (or two) variables are -1. Otherwise, it has value of -1. The parentheses around '$RM_{i-1}\&$' is used to show that the RM node is optional. If RM node does not exist, the term $(RM_{i-1}\&)M_{i-1}\&R_{i-1}$ is reduced to $M_{i-1}\&R_{i-1}$.

$$A_1\&A_2\&, \ldots, A_{n-1}\&A_n = \begin{cases} -1 & \exists i \in [1, n] \ s.t. \ A_i = -1 \\ +1 & otherwise \end{cases}. \tag{3}$$

Next we show the biological insight behind the conditional probability table for R_i. Here we focus on the expression table (Table 1 a); the repression table (Table 1 b) is built in a similar way. If the parent gene of g_i, g_{i-1}, has no function loss in DNA, it is mapped to an overexpression and the functional status of g_{i-1}'s protein is fully activated, namely $M_{i-1}\&R_{i-1}\&RM_{i-1} = +1$, then the target g_i would be highly likely to overexpress, i.e., $R_i = +1$, given the edge between

them in G^* is 'expression'. If there is no Functional Interaction $e_{i-2,i-1}$ targeting at g_{i-1}, only R_{i-1} and M_{i-1} exist in the conditional table. Consequently,

$$P(R_i = +1 | M_{i-1} \& R_{i-1}(\& RM_{i-1}) = +1) = 1 - \epsilon - \tau,$$

while

$$P(R_i = -1 | M_{i-1} \& R_{i-1}(\& RM_{i-1}) = +1) = \epsilon,$$

where ϵ and τ are, respectively, the probability of observing $R_i = -1$ and $R_i = 0$, given $M_{i-1} \& R_{i-1}(\& RM_{i-1}) = +1$. The parameter are set to make $\varepsilon = 0.2$ and $\tau = 0.375$ in this work so that we can penalize the inconsistency more than the uncertainty. Similarly, if the parent gene of g_i has DNA function loss (e.g., function loss mutation), its expression level is down-regulated, or the protein of g_{i-1} is not activated ($M_{i-1} \& R_{i-1} \& RM_{i-1} = -1$), then the downstream regulation process towards g_i is not likely functioning. Therefore, g_i would tend to be down-regulated, making $R_i = -1$, and hence the corresponding probability would be flipped.

Similar to R_i, random variable RM_i has three possible values: $\{+1, 0, -1\}$, where +1 if gene g_i has its protein activated by its parent gene g_{i-1} through $e_{i-1,i}$, -1 if gene g_i is inhibited, and 0 otherwise. Recall that RM_i is introduced only when the interaction $e_{i-1,i}$ in G^* is a Functional Interaction between g_i and g_{i-1} in G^*. In this case, RM_i follows the conditional probability table from Tables 1 c to 1 d.

The biological insight behind the conditional probability table for RM_i is built based on the central dogma of molecular biology, as shown in Fig. 3. Here we focus on the functional activation table (Table 1 c); the functional inhibition table (Table 1 d) is built similarly. If the parent gene of g_i, g_{i-1}, has no function loss in DNA, it is overexpressed and g_{i-1}'s protein is successfully activated (if RM_{i-1} exists) ($M_{i-1} \& R_{i-1} \& RM_{i-1} = +1$), then interaction $e_{i-1,i}$ takes effect. The target g_i protein is highly likely to be up-regulated, i.e., $RM_i = +1$, given the edge between the two nodes in G^* is the functional activation. As a result,

$$P(RM_i = +1 | M_{i-1} \& R_{i-1} \& RM_{i-1} = +1) = 1 - \epsilon - \tau,$$

while

$$P(RM_i = -1 | M_{i-1} \& R_{i-1} \& RM_{i-1} = +1) = \epsilon,$$

where ϵ and τ are, respectively, the probability of observing $RM_i = -1$ and $RM_i = 0$. Similarly, if the parent gene of g_i has DNA function loss (e.g., due to mutation) or its expression level is down-regulated, or the protein of g_{i-1} is not activated successfully ($M_{i-1} \& R_{i-1} \& RM_{i-1} = -1$), then the downstream regulation process towards g_i is not likely to be functioning. Therefore, g_i would tend to be not activated, i.e., making $RM_i = -1$. In this case, the corresponding probability would be flipped.

3.4 Ranking the Route

3.4.1 A Score Based on Conditional Probability

Given $(\mathbf{r}, \mathbf{m}, \mathbf{rm})$ and a set of data observations for the random variables in Bayesian Network G from a specific patient s, we could rank the path G^* with the probability of observing $\mathbf{r}$, $\mathbf{m}$ and $\mathbf{rm}$ conditioning on the Bayesian network model G, $P(\mathbf{R} = \mathbf{r}, \mathbf{M} = \mathbf{m}, \mathbf{RM} = \mathbf{rm} | G)$. The larger the probability, the more likely the pathway route is perturbed since the observation is highly consistent with the biological regulation G^* encoded in G. One problem of using this probability as a measure is that the probability will be higher if fewer data is observed. Thus the score displayed in (4) given in [12] is used instead, where the conditional probability is normalized by $P(\mathbf{R}, \mathbf{M}, \mathbf{RM} \text{ are consistent} | G)$.

$$\begin{aligned} &Score_s(G^*, \mathbf{r}, \mathbf{m}, \mathbf{rm}) \\ &\quad = \frac{P(\mathbf{R} = \mathbf{r}, \mathbf{M} = \mathbf{m}, \mathbf{RM} = \mathbf{rm} \,|\, G)}{P(\mathbf{R}, \mathbf{M}, \mathbf{RM} \text{ are consistent} \,|\, G)} \\ &P(\mathbf{R} = \mathbf{r}, \mathbf{M} = \mathbf{m}, \mathbf{RM} = \mathbf{rm} \,|\, G) \\ &\quad = \sum_{\mathbf{R}=\mathbf{r}, \mathbf{M}=\mathbf{m}, \mathbf{RM}=\mathbf{rm}} P(\mathbf{R}, \mathbf{M}, \mathbf{RM}) \\ &\quad = \sum_{\mathbf{R}=\mathbf{r}, \mathbf{M}=\mathbf{m}, \mathbf{RM}=\mathbf{rm}} \prod_{Pa^G(R_i)=\emptyset} P(R_i) \prod_{1 \le i < k_{G^*}} P(M_i) \\ &\qquad \prod_{Pa^G(R_i)\neq\emptyset} P(R_i \,|\, Pa^G(R_i)) \\ &\qquad \prod_{Pa^G(RM_i)\neq\emptyset} P(RM_i \,|\, Pa^G(RM_i)), \end{aligned} \tag{4}$$

where $Pa^G(X)$ is the set containing parent nodes of the node X in Bayesian Network G. $P(\mathbf{R}, \mathbf{M}, \mathbf{RM} \text{ are consistent} | G)$ is the probability that the random variables with observations are fully consistent with the biological regulations encoded in the pathway route, given $R_1 = +1, M_1 = +1$. For instance, suppose the pathway route only contains two gene products g_1 and g_2 where g_1 activates g_2 by phosphorylation, i.e., $G^* : g_1 \xrightarrow{+p} g_2$. Then we have (5)

$$\begin{aligned} &P(\mathbf{R}, \mathbf{M}, \mathbf{RM} \text{ are consistent} \,|\, G) \\ &\qquad = P(R_1 = +1, M_1 = +1, RM_2 = +1 \,|\, G), \end{aligned} \tag{5}$$

since g_2 is the last node in the route and the interaction e_{12} is phosphorylation, then R_2 and M_2 are not included in the model.

A high score means that the path G^* is highly likely perturbed based on the given data. A path G^* could only get a high score if the observations are highly consistent with pathway information contained in the Bayesian Network G. Inconsistency between data and the model would lower the score greatly since the conditional probability will be ϵ instead of $1 - \epsilon - \tau$ during the calculation of the score. Advantages of this measure are

- The analysis could be done across pathways, i.e., once multiple pathways are merged in a biologically meaningful way, this measure could recognize a significantly meaningful route across different pathways.
- By decomposing the pathway graph as routes, the conflict that people encounter when treating a pathway as a graph is eliminated. And more complicated directional regulations like phosphorylation can be handled.
- Even though some observation values are flipped due to random errors from the genomic data (for example, it is observed to be -1 when it is actually $+1$), the

whole path would still have a high score if the other genes are assigned with consistent observations.

Let the data here come from one patient, s, indicating that the score is specifically tailored to the patient s.

The perturbed route could have two possible status, enhanced or suppressed, as we have defined in [1]. Here an extended definition is required: a pathway route G^* is defined as *enhanced* if the last gene's $R(RM$ if it exists) node is observed to be the same as the expected value. The expected value is calculated based on biological regulation encoded in G^* by supposing $R_1 = +1, M_1 = +1$. The route is defined to be *suppressed* if the observation of the last node is opposite to the expectation. The score is now revised to produce a new signed score, $sScore$, as shown in (6).

$$sScore_s(G^*, \mathbf{r}, \mathbf{m}) = \tilde{I}(o_{|G^*|}, \dot{o}_{|G^*|}) \cdot Score_s(G^*, \mathbf{r}, \mathbf{m}, \mathbf{rm}), \quad (6)$$

where o_{G^*} is the observation of the last gene in the route G^*. Function $\tilde{I}$ is defined in (7). The signed score varies from -1 (highly suppressed) to $+1$ (highly enhanced). If $e_{|G^*|-1,|G^*|}$ is an Expression Interaction, then $o_{|G^*|} = r_{|G^*|}$. Otherwise, $o_{|G^*|} = rm_{|G^*|}$ since $e_{|G^*|-1,|G^*|}$ is Functional interaction. $\dot{o}_{|G^*|}$ is the expected observation of the last gene in the route conditioning on $R_1 = +1, M_1 = +1$. For the same example $G^* : g_1 \xrightarrow{+p} g_2$, $\dot{o}_{|G^*|}$ becomes the expected value of RM_2. The expected value is $+1$ for RM_2 since g_2 is expected to be activated given $R_1 = +1, M_1 = +1$.

$$\tilde{I}(x, y) = \begin{cases} +1 & x = y \\ -1 & x \neq y \end{cases}. \quad (7)$$

Finally, we propose the measure for a whole pathway based on the route score. The pathway score for pathway G_B based on data from a group of subjects S, $pScore_S(G_B)$, is displayed in (8)

$$pScore_S(G_B) = \frac{1}{\sum_{G^* \in G_B} w_{G^*}} \sum_{G^* \in G_B} w_{G^*} I\left(\frac{1}{|S|} \sum_{s \in S} Score_s(G^*) \geq \beta\right). \quad (8)$$

The above equation is formulated for the following reasons. The pathway could be partitioned into several routes. We then simply measure the significance of this pathway, G_B, using the proportion of routes that have an average of all the patients' scores, calculated by (4), that is larger than threshold β. Moreover, we weight the route score by route length so that a score from a longer route can have a bigger impact.

3.5 Data Integration

The observations for each variable in the Bayesian Network G will come from multiple types of data, as shown in Fig. 1. D. The gene expression variable R_i value can be measured by many types of gene expression data, for instance, microarray, mRNA-seq, reverse phase protein array (RPPA), among others. Here mRNA-seq is chosen. R_i's observation r_i is generated with mRNA-seq RPKM. If both protein data and mRNA-seq data are available for the same gene of the same patient and these two data sets exhibit conflicting observations, then we use the protein data observation to overwrite the one from mRNA-seq data since mRNAs may be degraded while proteins are present for longer half-lives, thus making protein data much more reliable.

The data observed for random variable M_i, m_i, is the congenital functional status for gene g_i. Observation $m_i = -1$ if it can be observed from mutation or CNV data, i.e., g_i's DNA variation is a loss-of-function or a silent mutation. For instance, if we observe a loss-of-function mutation from the mutation database or a serious copy number loss from the CNV database, then $m_i = -1$. Otherwise, $m_i = +1$ standing for not observing functional deficiency in g_i DNA. We assume that mutation annotation tools is helpful in finding if a particular mutation is a function loss mutation or not.

When it comes to the observation of RM_i, rm_i, the data source becomes more complex. The data source needs to be determined by the specific subtype of *Functional Interaction* $e_{i-1,i}$, namely by the *Evidence Tag*. The general logic is summarized by the following equation in (9).

$$rm_i = Type_{i-1,i} * Tag_{i-1,i} * RawValue_i \quad (9)$$

where $Type_{i-1,i} = +1$ if $e_{i-1,i}$ is activation and $Type_{i-1,i} = -1$ if $e_{i-1,i}$ is inhibition; $Tag_{i-1,i} = +1$ if $e_{i-1,i}$ has a *Evidence Tag* sign of $(+)$, i.e., $+p, +m, +u$ or $+g$ while $Tag_{i-1,i} = -1$ if $e_{i-1,i}$ is with a *Evidence Tag* of $(-)$ sign, i.e., $-p, -m, -u$ or $-g$. $RawValue_i = +1$ if the database shows that the gene is Phosphorylated, Methylated, Ubiquitinated or Glycosylated and $RawValue_i = -1$ if the database shows that the gene is Dephosphorylated, Demethylated, Deubiquitinated or Deglycosylated. For instance, if the regulation process $e_{i-1,i}$ is inhibition ($Type_{i-1,i} = -1$) tagged with phosphorylation ($+p$, $Tag_{i-1,i} = +1$) and phosphorylation data for the patient shows that g_i is phosphorylated ($RawValue_i = +1$), then $rm_i = -1 * 1 * 1 = -1$ represents that g_i protein is inhibited. If the data shows that the gene is not phosphorylated ($RawValue_i = -1$), then $rm_i = -1 * 1 * (-1) = +1$ indicating that g_i protein is not inhibited successfully through $e_{i-1,i}$. If $e_{i-1,i}$ has *Evidence Tag* of demethylation ($-m$) instead, then the value of $RawValue_i$ is determined from methylation data. The same goes with the other possible interactions: Phosphorylation, Dephosphorylation, Ubiquitination, Glycosylation etc. For the Functional Interactions with no Evidence tag, we assume that the edge is always working with $r_{i-1} = +1$, $m_{i-1} = +1$ and the formula in (10) is used instead.

$$rm_i = Type_{i-1,i} * min(r_{i-1}, m_{i-1}). \quad (10)$$

This formula indicates that given no function loss observation in M_{i-1} and no low expression observation in R_{i-1}, $e_{i-1,i}$ works and determines rm_i.

4 Results

4.1 Significance Analysis

The bioinformatics field frequently uses TCGA Breast invasive carcinoma data to test newly developed analysis models. We choose the same TCGA cancer data set to validate our model. Another cancer data set, Ovarian serous cystadenocarcinoma, is also analyzed with the same methodology to demonstrate the generality of our method. Four types of data

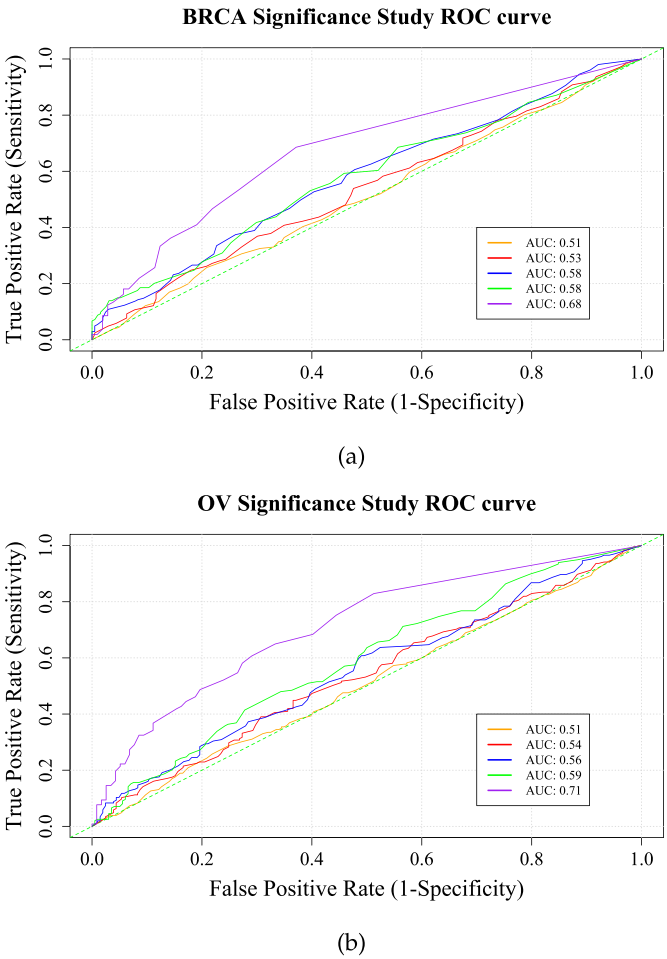

Fig. 4. (a) ROC curve for BRCA Significance study. (b) ROC curve for OV Significance study. The orange, red, blue, green, and purple curves correspond to $\beta = 0.2, 0.4, 0.6, 0.8, 1.0$ seperately. The threshold is picked from [0,1] with a step of 1/10000. We can see that by increasing the threshold β, the AUC is increased. The maximum AUC is achieved at $\beta = 1$ for both BRCA and OV study.

sets are used: mRNA-seq, mutation, and Copy Number Variation were downloaded from https://gdac.broadinstitute.org/ for both cancer studies and phosphorylation data were extracted from recent work [13] and [14], respectively.

The mRNA-seq data is processed as follows to obtain r_i, the observation for R_i: 0 RPKM is encoded as $r_i = -1$ and $r_i = +1$ if RPKM is positive, otherwise $r_i = 0$. This is due to the fact that BRCA and OV mRNA-seq data have no adequate paired normal samples. The value for each item is mapped to a node in the pathway by official gene symbol.

The mutation information is extracted from mutation assessor study [15], [16]. The mutation with a 'medium' or 'high' impact factor is encoded as a function loss mutation. Other mutations are encoded as no function loss mutation in the data. The value for each item is mapped to a node in the pathway using NCBI-protein ID.

Copy Number Variation (CNV) data is imported from GISTIC2 study [17], [18], where the copy number variation is quantified by integers varying from -2 to $+2$ and negative values are considered as copy number loss. CNV information determines the observation for M node, m, along with mutation information as we discussed in Section 3.5. The value for each item is mapped to a node in the pathway using its official gene symbol.

In the end, phosphorylation data is processed. The $RawValue_i$ (phosphorylation) $= +1$ if the same patient's phosphosite iTRAQ log2 ratio is positive for g_i and otherwise $RawValue_i$(phosphorylation)$= -1$. Missing values are encoded as 0. The value for each item is mapped to a node in the pathway using NCBI-protein ID. However, one challenge here is that values for different residues within the same protein may be inconsistent. KEGG pathway does not provide sufficient information on the specific residue involved for each phosphorylation. Consequently, only the consistent signals are used in the experiment.

Next, we perform a significance analysis similar to that of PARADIGM [3]. All KEGG Homo Sapien pathways are used in this study, and our entire pipeline is implemented in R using R package "KEGGgraph" [19] and "gRain" [20].

We produce decoy pathways by permuting the genes in the pathway while keeping the interactions unchanged. We generate one decoy pathway for each of 324 KEGG pathways. For each pathway, we extract all possible routes in it. Then for each route, we calculate the score for each pathway by pathway score formula in Deep Pathway Analysis (DPA) V1.0 [2] so that we can directly compare against the old version DPA. We rank the significant real pathways and their corresponding decoy pathways. A threshold is set to make a prediction, i.e., the cases having a score higher than the threshold are predicted to be the real pathway.

After obtaining False positive rate and True positive rate with various thresholds, the resulting Receiver operating characteristic (ROC) curve is computed as shown in Figs. 4a and 4b. The Area Under the Curve (AUC) gets 0.68(0.71) with the threshold $\beta = 1.00$ for BRCA(OV). This AUC value is reasonable since many real pathways are not differentially regulated in cancer data and thus may not be "recognizable". Similar AUCs were reported for SPIA (0.602) and PARADIGM (0.669)[3]. For the sake of comparison, we reran the same experiment with DPA V1.0 [2] which only utilizes mRNA-seq and mutation from the same BRCA data set. The AUC for DPA V1.0 is only 0.59 which is far below 0.68 from the new DPA V2.0. The new approach incorporating more types of omics data almost increases AUC by 0.1. This clearly shows the benefit of combining multiple types of omics data.

Lastly, we show in Tables 2 and 3, respectively, the top significantly perturbed pathways for Breast cancer and Ovarian Cancer, each filtered out from the 50 Homo Sapiens signaling pathways using the same $\beta = 0.99$. The second column corresponds to the pScore for the real pathway by (8). The third column is the statistical significance measure obtained by using 20 decoy pathways for each KEGG pathway. The pathways with 0 pScores are filtered out. The fourth column is the comparison with PARADIGM's analysis [3] indicating if PARADIGM calls the pathway as significant or not. Noticeable here is that many pathways picked by DPA V2.0 was not picked by PARADIGM. This difference in performance between the two systems is possibly due to (i) different probabilistic models used in two systems and (ii) in PARADIGM case only mRNA seq and copy number variation data from [21] was used.

TABLE 2
Significant Pathways for BRCA

Pathway	pScore	p-value	PARADIGM
Phosphatidylinositol signaling system	$9.900\,00 \cdot 10^{-03}$	0	No
Sphingolipid signaling pathway	$5.490\,00 \cdot 10^{-03}$	0.05	No
Apelin signaling pathway	$4.930\,00 \cdot 10^{-03}$	0.1	No
Phospholipase D signaling pathway	$4.560\,00 \cdot 10^{-03}$	0.15	Yes
Notch signaling pathway	$4.350\,00 \cdot 10^{-03}$	0.1	Yes
RIG I like receptor signaling pathway	$4.330\,00 \cdot 10^{-03}$	0	No
mTOR signaling pathway	$3.940\,00 \cdot 10^{-03}$	0	Yes
TNF signaling pathway	$3.070\,00 \cdot 10^{-03}$	0.05	Yes
Thyroid hormone signaling pathway	$3.020\,00 \cdot 10^{-03}$	0.1	No
Rap1 signaling pathway	$2.610\,00 \cdot 10^{-03}$	0	No
AGE RAGE signaling pathway	$2.370\,00 \cdot 10^{-03}$	0.1	No
VEGF signaling pathway	$2.350\,00 \cdot 10^{-03}$	0.15	Yes
NOD like receptor signaling pathway *	$2.250\,00 \cdot 10^{-03}$	0	No
MAPK signaling pathway	$2.170\,00 \cdot 10^{-03}$	0.05	Yes
Toll like receptor signaling pathway	$1.970\,00 \cdot 10^{-03}$	0	No
cAMP signaling pathway	$1.800\,00 \cdot 10^{-03}$	0.4	No
FoxO signaling pathway	$1.400\,00 \cdot 10^{-03}$	0.25	Yes
Calcium signaling pathway	$1.230\,00 \cdot 10^{-03}$	0.5	Yes
Wnt signaling pathway	$1.210\,00 \cdot 10^{-03}$	0	Yes
B cell receptor signaling pathway*	$1.210\,00 \cdot 10^{-03}$	0.05	No
Ras signaling pathway	$1.160\,00 \cdot 10^{-03}$	0.1	Yes
HIF 1 signaling pathway	$9.860\,00 \cdot 10^{-04}$	0.05	Yes
PI3K Akt signaling pathway	$8.930\,00 \cdot 10^{-04}$	0.4	Yes
Hippo signaling pathway	$6.990\,00 \cdot 10^{-04}$	0.05	No
Relaxin signaling pathway	$6.190\,00 \cdot 10^{-04}$	0.6	No
NF κ B signaling pathway	$4.020\,00 \cdot 10^{-04}$	0.15	Yes
p53 signaling pathway	$1.650\,00 \cdot 10^{-04}$	0.75	Yes
JakSTAT signaling pathway	$1.770\,00 \cdot 10^{-05}$	0.05	No

In this table, the columns corresponds to pathway names, the pathway score by (8), statistical p-value and indicator of discovery by PARADIGM [3].

TABLE 3
Significant Pathways for OV

Pathway	pScore	p-value	PARADIGM
Phosphatidylinositol signaling system	$2.130\,00 \cdot 10^{-02}$	0	No
TNF signaling pathway	$1.250\,00 \cdot 10^{-02}$	0.05	Yes
Sphingolipid signaling pathway	$1.130\,00 \cdot 10^{-02}$	0	No
Apelin signaling pathway	$1.060\,00 \cdot 10^{-02}$	0	No
AGE RAGE signaling pathway	$9.400\,00 \cdot 10^{-03}$	0	No
RIG I like receptor signaling pathway	$9.300\,00 \cdot 10^{-03}$	0	No
Phospholipase D signaling pathway	$8.600\,00 \cdot 10^{-03}$	0.25	No
mTOR signaling pathway	$8.460\,00 \cdot 10^{-03}$	0	Yes
VEGF signaling pathway	$8.290\,00 \cdot 10^{-03}$	0.1	Yes
MAPK signaling pathway	$5.460\,00 \cdot 10^{-03}$	0	Yes
Rap1 signaling pathway	$5.350\,00 \cdot 10^{-03}$	0	No
Toll like receptor signaling pathway	$4.820\,00 \cdot 10^{-03}$	0.1	No
p53 signaling pathway	$4.750\,00 \cdot 10^{-03}$	0.05	Yes
NOD like receptor signaling pathway *	$4.080\,00 \cdot 10^{-03}$	0	No
Ras signaling pathway	$3.880\,00 \cdot 10^{-03}$	0	Yes
Wnt signaling pathway	$3.320\,00 \cdot 10^{-03}$	0	Yes
HIF 1 signaling pathway	$2.330\,00 \cdot 10^{-03}$	0.2	Yes
PI3K Akt signaling pathway	$2.100\,00 \cdot 10^{-03}$	0.25	Yes
Thyroid hormone signaling pathway	$1.940\,00 \cdot 10^{-03}$	0.35	No
cAMP signaling pathway	$1.590\,00 \cdot 10^{-03}$	0.5	Yes
Hippo signaling pathway	$1.420\,00 \cdot 10^{-03}$	0.2	No
Relaxin signaling pathway *	$1.330\,00 \cdot 10^{-03}$	0.3	No
Calcium signaling pathway	$1.270\,00 \cdot 10^{-03}$	0.6	Yes
Hedgehog signaling pathway	$1.170\,00 \cdot 10^{-03}$	0.05	Yes
AMPK signaling pathway	$1.020\,00 \cdot 10^{-03}$	0.3	No
NF κ B signaling pathway	$9.220\,00 \cdot 10^{-04}$	0.25	Yes
FoxO signaling pathway	$6.600\,00 \cdot 10^{-04}$	0.7	Yes
Estrogen signaling pathway	$4.630\,00 \cdot 10^{-04}$	0.05	Yes
JakSTAT signaling pathway	$3.810\,00 \cdot 10^{-05}$	0	No

In this table, the columns corresponds to pathway names, the pathway score by (8), statistical p-value and indicator of discovery by PARADIGM [3].

4.1.1 BRCA Pathways Identified

We examined how much of the pathways that DPA V2.0 predicted has already been known in the Breast cancer literature as part of validation effort for our method.

This newly imported KEGG pathway, *Phosphatidylinositol signaling*, appears to be highly relevant to the commonly disrupted *PI3K-Akt* pathway in breast cancer [22]. Oestrogen is known to trigger the *sphingolipid signaling* cascade in various tissues including breast cancer [23]. *Apelin*-13 induces MCF-7 cell proliferation and invasion via phosphorylation of ERK1/2 [24]. *Phospholipase D* overexpresses in breast cancer cell and its implication in survival signals as discussed in [25]. Activated *Notch signaling* and up-regulation of tumor-promoting Notch target genes have been observed in human breast cancer [26]. Therapeutically active *RIG-I* agonist induces immunogenic tumor cell killing in breast cancers [27]. Hynes and Boulay report the *mTOR* pathway as a potential mechanisms activating breast cancer [28]. According to Wang and Li, LOX-1 is up-regulated by *TNF* in endothelial cell, which promotes the adhesion and transendothelial migration of MDA-MB-231 breast cancer cells [29]. In human breast cancer cells *thyroid hormone signaling* overshadows estrogen signaling on SMP30 gene leading to induction of apoptosis [30]. *Rap1* GTPase is found to be driving breast cancer cell migration with β1-integrin as suggested in [31]. *RAGE* expression is upregulated widely in aggressive triplenegative breast cancer (TNBC) cells [32]. Vascular endothelial growth factor (*VEGF*) is the most prominent among the angiogenic cytokines and is believed to play a central role in the process of neovascularization, both in cancer as well as in other inflammatory diseases [33]. *MAPK* pathway involving ERK-1 and ERK-2 is known very relevant to breast cancer progression [34]. An excellent review summarizing the role of the *toll-like receptor* (TLR) signaling pathway on breast cancer risk, disease progression, survival, and disease recurrence is given in [35]. Involvement of the *cAMP*/protein kinase A pathway and of mitogen-activated protein kinase in the anti-proliferative effects of anandamide in human breast cancer cells is discussed in [36]. *FOXO* transcription factors both suppress and support breast cancer progression [37]. When it comes to the *Calcium signaling* pathway, specific Ca(2+) channels reportedly play important roles in the proliferation and invasiveness of breast cancer cells [38]. Blockade of *Wnt*/β-catenin signaling suppresses breast cancer metastasis [39]. *Ras* pathway activation in breast cancer was specifically discussed [40]. *HIF-1* plays a key role in regulating breast cancer progression and metastasis [41]. Activation of the *(PI3K)/Akt*/mammalian target of rapamycin (mTOR) pathway is commonly reported in breast cancer [42]. Dysfunction of *Hippo* pathway components is linked to breast cancer stem cell regulation and the connection between the disease and genetic variations in the pathway is reported in [43]. *Relaxin* enhances in-vitro invasiveness of breast cancer cell lines by upregulation of S100A4/MMPs signaling [44]. The critical role for *NF-κB signaling* pathway is discussed in [45]. More aggressive disease progression and poorer overall survival are suggested for *p53* mutation breast cancer patients [46]. The *JAK/STAT* pathway is active in breast cancer metastasis [47].

Overall, we found that 89 percent $(23/25)$ (which is 5 percent higher than the result previously reported in [2]) of the pathways identified in our anlaysis has some published facts implicated in breast cancer, suggesting that our analysis is producing meaningful outcomes. In contrast, only $14/25$ (56 percent) of these pathways are identified by PARADIGM, among which are well known cancer pathways such as TNF, mTOR, MAPK, Wnt, PI3K-Akt etc. PARADIGM also fails to discover well established cancer pathways such as Hippo, Toll-like, and JakSTAT. We also note that the known major cancer pathway PI3K-Akt is not ranked at the top in our approach. This is possibly due to the fact that the pathway score by (8) introduces a penalty on pathway size. The fraction of perturbed route gets

smaller if the pathway size being analyzed is larger. PI3K-Akt is one of those large pathways. Lastly, in Table 2 the pathways that we have not known reference in the literature are marked with $*$ and these may deserve extra attention by the wet bench scientists.

4.1.2 OV Pathways Identified

The ovarian cancer result is validated in a similar way. The result is also compared to the one from PARADIGM that used only mRNA seq and copy number variation data [48]. Only top pathways with score larger than 0 are displayed here due to space limit.

Like in BRCA, *Phosphatidylinositol signaling*, a pathway highly relevant to PI3K-Akt pathway is ranked top and the disruption of PI3K-Akt pathway is well established in ovarian cancer [22]. *TNF* secretion by ovarian cancer cells is known constitutively stimulating cytokines, chemokines, and angiogenic factors that in turn promote colonization of the peritoneum and neovascularization for developing tumor deposits [29]. The role of *sphingolipids* in various cancers including ovarian cancer is discussed in [49]. *Apelin*/APJ may play an important role in promoting angiogenesis and progression in ovarian cancer [50]. Receptor for Advanced Glycation Endproducts (*RAGE*) is expressed in ovarian tissue and associated with ovarian carcinoma [51]. Ovarian cancer cells stimulated via *RIG-I* may become apoptotic [52]. A novel role of *phospholipase D* in agonist-stimulated lysophosphatidic acid synthesis is identified in ovarian cancer cells [53]. The mammalian target of rapamycin is frequently activated in epithelial ovarian cancer and is regarded as an attractive therapeutic target for therapy [54]. *VEGF* has also been implicated in the pathogenesis of ovarian cancer according to [55]. For the *MAPK* pathway, MEK4 was shown to suppress metastasis based on its down-regulation in prostate and ovarian cancers with a high risk of metastasis [56]. *Rap1A* is known to promote ovarian cancer metastasis via activation of ERK/p38 and notch signaling [57]. *Toll-like receptor* signaling may play an important role in ovarian cancer (OC) progression [58]. The paradoxical roles of toll-like receptor signaling in the progression of ovarian cancer is discussed in [59]. The role of tumor suppressor *p53* and the Rb pathway in EOC with particular attention to association of p53 to high grade serous carcinomas (as opposed to low grade and benign tumors) is reviewed [60]. The *Ras-signaling* pathway has attracted considerable attention as a target for anticancer therapy because of its important role in carcinogenesis [61]. About 23.5 percent ovarian cancer cell lines having RAS/RAF pathway aberrations is reported [62]. *Wnt*/β-catenin pathway is implicated in epithelial ovarian cancer, specifically its role in chemoresistance and its potential role as a target for chemosensitization [63]. *HIF-1*α overexpression was correlated with apoptosis in most ovarian cancer tumours and the combination of apoptosis and HIF-1α overexpression was associated with increased patient survival [64]. According to [65], *PI3K-Akt* pathway was frequently altered in many cancers, including ovarian cancer. *Thyroid hormone signaling* may explain the suggested epidemiological links between hyperthyroidism and ovarian cancer [66]. In ovarian cancer cells, *cAMP* has been shown to mediate integrin-dependent adhesion of OVCAR-3 cells to fibronectin through the Epac [67]. *Hippo signaling* pathway may define an important pathway in progression of ovarian cancer in [68]. *Calcium signals* inhibition sensitizes ovarian carcinoma cells to anti-Bcl-xL strategies through Mcl-1 down-regulation [69]. The crucial role of *Hedgehog signaling* in the development and progression of ovarian cancer is highlighted [70]. *AMPK* activators may not only prevent cancer progression and metastasis but also can be applied as a supplement to enhance the efficacy of cisplatin-based chemotherapy in ovarian cancer patients [71]. *NF-*κ*B* may play a critical role in ovarian cancer [72]. Inactivation of *FOXO* or overexpression of FOXM1 is associated with tumorigenesis and cancer progression [73]. Active *estrogen* receptor-alpha signaling in ovarian cancer models was discussed in [74]. Inhibition of the *JAK2*/*STAT3* pathway in ovarian cancer resulted in the loss of cancer stem cell-like characteristics and a reduced tumor burden [75].

Overall, we found that 93 percent $(27/29)$ of the pathways in Table 3 have some published facts implicating them in ovarian cancer etiology, suggesting again that our analysis is producing meaningful outcomes. PARADIGM reported $15/29$ of these pathways including the well-known ones such as TNF, mTOR, MAPK, Wnt, PI3K-Akt, etc. PARADIGM does not recognize the established cancer pathways like Rap1, Toll-like, Hippo and JakSTAT pathways. We observe a similar limitation of our approach which ranks many major cancer pathways at medium level, similarly as in the BRCA case.

4.2 Survival Analysis

The goal of survival analysis is to identify which pathway route could highly affect the mortality of the OV and BRCA patients. After processing score calculation for all of the 324 Homo Sapiens pathways, all the route scores are saved into a data matrix where columns correspond to the routes and rows correspond to the patients. This data matrix includes more than 20,000 features after removing routes with only two genes (which is too small a pathway route) or routes having same score for all patients (not contributing to classification). We apply Lasso Cox proportional hazards model to derive the relationship between predictor variables (routes) and survival time [76]. The Lasso regression methods generate sparse features, setting all the unimportant feature coefficients to zero. The implementation is done using the R-package glmnet in [77].

Next we use default values for most of the required parameters of the package except λ which chooses proper limit of total parameters' $L-1$ Norm. We have chosen 0.1 (0.3) for λ and this value controls the 10-fold cross validation error to minimal in BRCA(OV) study. Among the 20,000 pathway routes (features), only 4(14) of them are recognized by the model for BRCA(OV). The Kaplan-Meier (KM) plots for these routes are shown, respectively, in Figs. 7 and 8. These figures show that all identified routes have a significant p-value ($p < 0.05$) in the hypothesis test.

The four routes identified for breast cancer are from MAPK, Hippo and Cancer pathways. All three pathways are already known to be highly involved in breast cancer [34], [43]. On the other hand, the 14 routes for ovarian cancer are from Retinol metabolism, ErbB, Rap1, Calcium, Chemokine, Gap junction, and Th1 and Th2 cell differentiation pathways of which only five are well annotated with ovarian cancer in the literature [31], [38], [78], [79], [80].

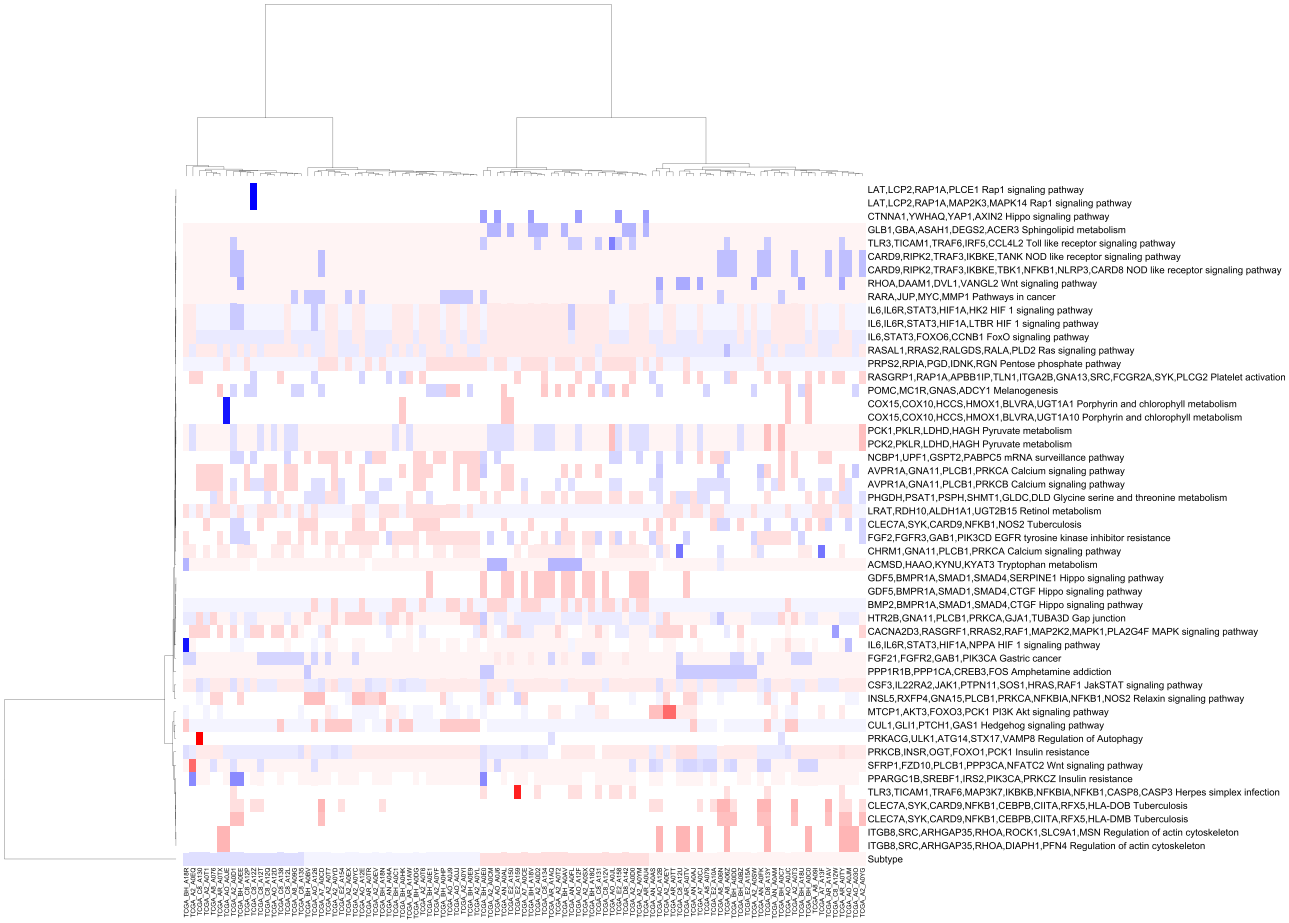

Fig. 5. The heatmap for BRCA subtypes: the columns corresponds to patients and the rows corresponds to pathway routes. Red indicates enhancement while blue indicates suppressiveness. Four subtypes in the last row (from left to right) corresponds to Her2-enriched, Lumina A, Basal-like, and Lumina B, respectively. Each subtypes can be associated with several route markers.

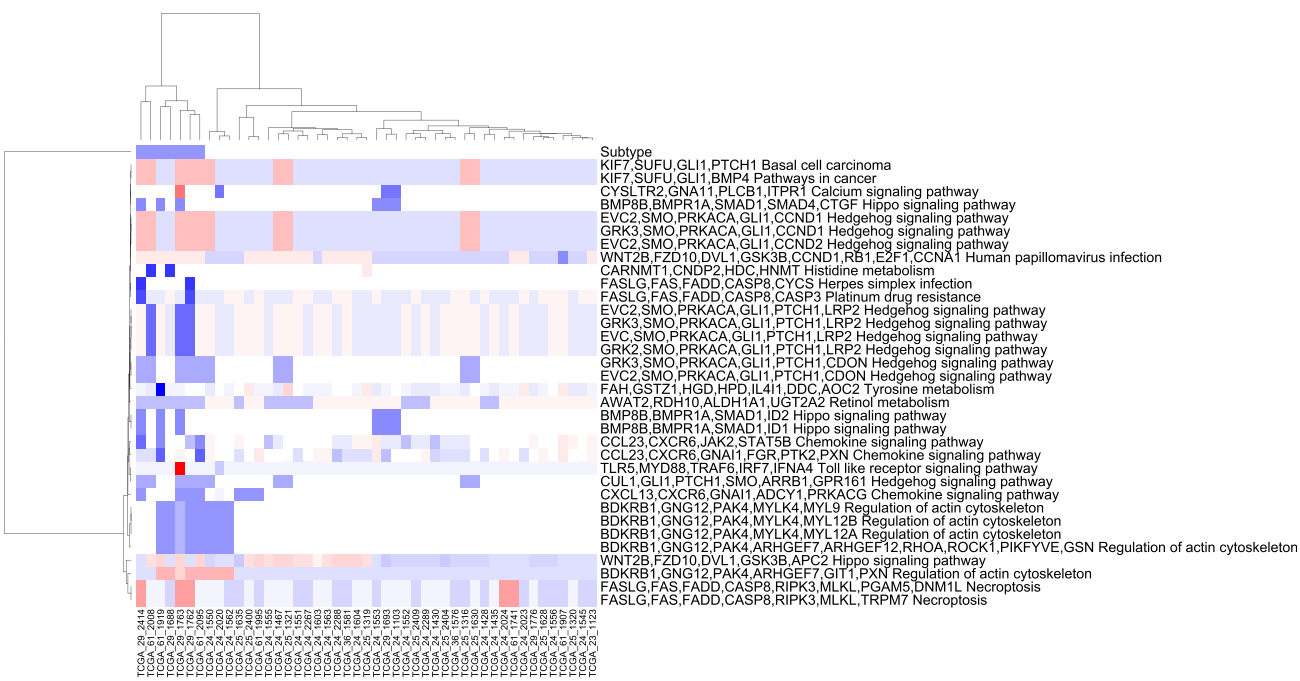

Fig. 6. The heatmap for OV subtypes: the columns corresponds to patients and the rows corresponds to pathway routes. Red indicates enhancement while blue indicates suppressiveness. Two subtypes in the first row corresponds to G2 (blue) and G3 (white), respectively. One can clearly tell the distinct pattern between the routes scores of two subtypes: G2 patients have quite many pathway routes highly suppressed.

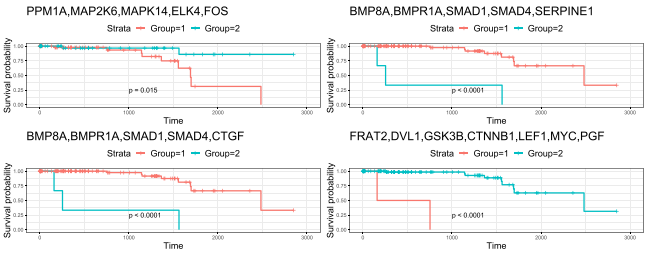

Fig. 7. BRCA Kaplan Meier plots for the selected routes by the model with p-value attached. Patients in Group 1 and Group 2, respectively, are higher or lower than the mean score for the same route. Routes are shown in the title above each plots.

4.3 Patient Subtyping with Route Scores

In this section, we discuss if pathway route scores can be used to identify cancer subtypes, specifically, in this case subtypes of BRCA and OV. BRCA PAM50 subtypes and OV Historic code are considered. BRCA PAM50 subtype contains Her2-enriched, Luminal A, Luminal B and Basal-like. OV historic code includes G2 and G3 in the OV data. These information were extracted from cBioPortal [81] and mapped by TCGA patient barcode.

All routes from all the 324 homo pathways were aggregated after removing routes with only two genes and constant score routes. A multinomial regression LASSO model [82] is fitted to predict each subtype given the route scores of each patient. By setting the penalty coefficient of 0.15, we computed the top features for each subtype by ordering the features by absolute value of coefficient estimate as LASSO is known to eliminate unimportant features. We extracted the top 50(34) routes with largest absolute values of coefficient estimates and visualized how their routes and patient subtype from BRCA(OV) are clustered in Fig. 5 (Fig. 6).

Fig. 5 clearly reveals a pattern associating some signature routes with prior encoded subtypes. A MAPK pathway route 'CACNA2D3, RASGRF1, RRAS2, RAF1, MAP2K2, MAPK1, PLA2G4F' is highly recognized for Her2-enriched patients. Gap junction route 'HTR2B, GNA11, PLCB1, PRKCA, GJA1, TUBA3D' seems to be activated for Lumina A subtype. Three Hippo signaling routes, 'CTNNA1, YWHAQ, YAP1, AXIN2', 'GDF5, BMPR1A, SMAD1, SMAD4, CTGF' and 'BMP2, BMPR1A, SMAD1, SMAD4, CTGF' are identified as the signature routes for Basal-like breast cancer patients. Route 'ITGB8, SRC, ARHGAP35, RHOA, DIAPH1, PFN4' from Regulation of actin cytoskeleton pathway is likely to be representative of Luminal B patients. For the OV heatmap in Fig. 6, a clear suppression pattern on many cancer pathway routes is observed for subtype G2, which confirms the fact that G2 ovarian cancer cells may undergo more differentiation [83]. Interestingly, routes starting from KIF7 to BMP4 in the cancer pathway appear to be more activated among G2 patients.

5 Conclusion

We presented an extended version of our Deep Pathway Analysis system which aims to uncover pathway routes (as opposed to the whole pathway) as the unit of analysis to pinpoint perturbed signals from various omics data sets. The key contribution of this extension is the Bayesian Network framework that can now include in its analysis multiple omics data types beyond transcriptome, CNV and mutation

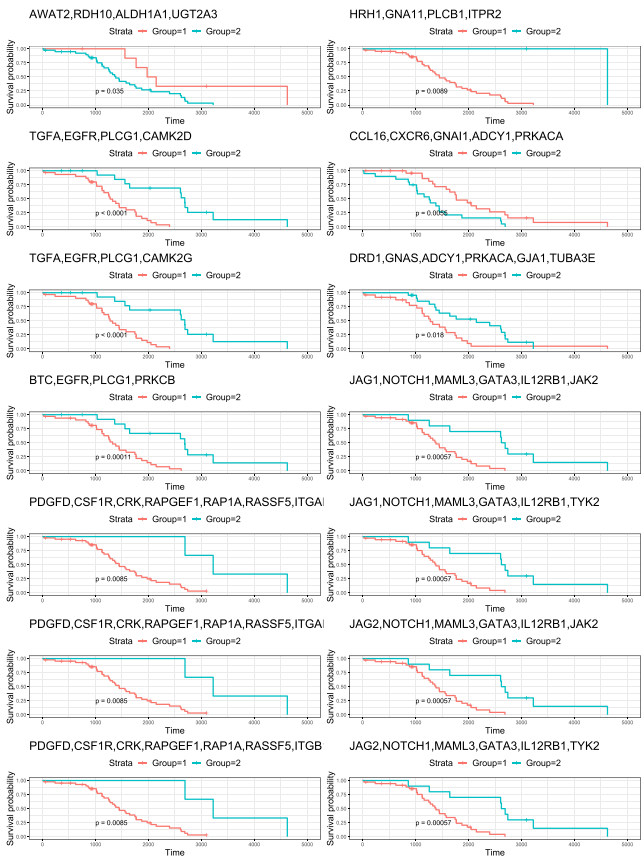

Fig. 8. OV Kaplan Meier plots for the selected routes by the model with p-value attached. Patients in Group 1 and Group 2, respectively, are higher or lower than the mean score for the same route. Routes are shown in the title above each plots.

such as proteomics, enabling the modeling of various additional molecular regulatory events such as phosphorylation, dephosphorylation, methylation demethylation and so on.

We demonstrated the effectiveness of our model through a significance study with two widely used TCGA cancer data sets, namely, BRCA and OV. The AUC for our extended model improved almost by 0.1 when compared to the result on BRCA by our old version DPA [1], [2], [84]. In comparison with PARADIGM, our system was able to identify multiple pathway routes that PARADIGM did not report but appear to be relevant to breast cancer progression from the literature survey. In terms of AUC our method reported comparable results similar to what PARADIGM reports for BRCA. This lack of significant improvement is possibly due to problem of missing data (e.g., protein residue information missing in the KEGG database we used). However, our framework is theoretically capable of analyzing a large number of heterogeneous omics data types concurrently.

We also performed survival analysis and patient subtype analysis with the pathway route scores produced by DPA. The results are promising as some routes appear to discern patient mortality through KM-analysis with a significance p-value far smaller than 0.05. The outcomes of the patient subtype analysis are also consistent with what have already been known in the literature. These two experiments support our hypothesis that route-based perturbed pathway identification is as good as the conventional methods and use of multiple omics data types is as good as using only a few but

offers the potential for improvement as long as proper omics data sets that can be analyzed together become available. Regarding future work, we are continuously looking for opportunities to apply our method to integrate additional omics data types. We also hope that the availability of our model encourages wet lab scientists to generate data sets that can be combined to derive more accurate interpretation from their newly generated data sets.

ACKNOWLEDGMENTS

The authors would like to thank the five anonymous reviewers for taking the time and effort which greatly helped them improve the manuscript. Yue Zhao's work was supported in part by the Pre-doctoral Fellowships from the Department of Computer Science and Engineering of the University of Connecticut. Dong-Guk Shin acknowledges his colleagues at The Jackson Laboratory for Genomic Medicine, Farmington Connecticut, Drs. Charles Lee, Jeffrey Chuang, and Edison Liu, for helping him learn and appreciate pressing computational issues in cancer genomics. Both authors also acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used in this research.

REFERENCES

[1] Y. Zhao, T. H. Hoang, P. Joshi, S.-H. Hong, and D.-G. Shin, "Deep pathway analysis incorporating mutation information and gene expression data," in *Proc. IEEE Int. Conf. Bioinf. Biomedicine*, 2016, pp. 260–265.

[2] Y. Zhao, T. H. Hoang, P. Joshi, S.-H. Hong, C. Giardina, and D.-G. Shin, "A route-based pathway analysis framework integrating mutation information and gene expression data," *Methods*, vol. 124, pp. 3–12, 2017.

[3] C. J. Vaske, et al., "Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm," *Bioinf.*, vol. 26, no. 12, pp. i237–i245, 2010.

[4] A. Subramanian, et al., "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Nat. Acad. Sci. United States America*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[5] A. L. Tarca, et al., "A novel signaling pathway impact analysis," *Bioinf.*, vol. 25, no. 1, pp. 75–82, 2009.

[6] M. Korucuoglu, et al., "Bayesian pathway analysis of cancer microarray data," *PloS One*, vol. 9, no. 7, 2014, Art. no. e102803.

[7] S. Isci, et al., "Pathway analysis of high-throughput biological data within a Bayesian network framework," *Bioinf.*, vol. 27, no. 12, pp. 1667–1674, 2011.

[8] S. Ma, T. Jiang, and R. Jiang, "Differential regulation enrichment analysis via the integration of transcriptional regulatory network and gene expression data," *Bioinf.*, vol. 31, no. 4, pp. 563–571, 2014.

[9] L. P. Verbeke, et al., "Pathway relevance ranking for tumor samples through network-based data integration," *PloS One*, vol. 10, no. 7, 2015, Art. no. e0133503.

[10] A. Kaushik, S. Ali, and D. Gupta, "Altered pathway analyzer: A gene expression dataset analysis tool for identification and prioritization of differentially regulated and network rewired pathways," *Scientific Reports*, vol. 7, 2017, Art. no. 40450.

[11] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, 2000.

[12] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA, USA: MIT Press, 2009.

[13] P. Mertins, et al., "Proteogenomics connects somatic mutations to signalling in breast cancer," *Nature*, vol. 534, no. 7605, pp. 55–62, 2016.

[14] H. Zhang, et al., "Integrated proteogenomic characterization of human high-grade serous ovarian cancer," *Cell*, vol. 166, no. 3, pp. 755–765, 2016.

[15] B. I. T. G. D. A. Center, "Mutation assessor," 2016. [Online]. Available: https://doi.org/10.7908/C1F18Z2Z

[16] B. I. T. G. D. A. Center, "Mutation assessor," 2016. [Online]. Available: https://doi.org/10.7908/C1TT4QCX

[17] B. I. T. G. D. A. Center, "SNP6 copy number analysis (GISTIC2)," 2016. [Online]. Available: https://doi.org/10.7908/C1NP23RQ

[18] B. I. T. G. D. A. Center, "SNP6 copy number analysis (GISTIC2)," 2016. [Online]. Available: https://doi.org/10.7908/C1P84B9Q

[19] J. D. Zhang and S. Wiemann, "KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor," *Bioinf.*, vol. 25, no. 11, pp. 1470–1471, 2009.

[20] S. Højsgaard, "Graphical independence networks with the gRain package for R," *J. Statistical Softw.*, vol. 46, no. 10, pp. 1–26, 2012. [Online]. Available: http://www.jstatsoft.org/v46/i10/

[21] Broad Institute TCGA Genome Data Analysis Center, "Paradigm pathway analysis of mRNA expression and copy number data," 2016. [Online]. Available: http://gdac.broadinstitute.org/runs/analyses__2016_01_28/reports/cancer /BRCA-TP/Pathway_Paradigm_mRNA_And_Copy_Number/nozzle.html

[22] T. Sasaki, et al., "Mammalian phosphoinositide kinases and phosphatases," *Prog. Lipid Res.*, vol. 48, no. 6, pp. 307–343, 2009.

[23] O. Sukocheva and C. Wadham, "Role of sphingolipids in oestrogen signalling in breast cancer cells: An update," *J. Endocrinology*, vol. 220, no. 3, pp. R25–R35, 2014.

[24] X. Peng, et al., "Apelin-13 induces MCF-7 cell proliferation and invasion via phosphorylation of ERK1/2," *Int. J. Mol. Medicine*, vol. 36, no. 3, pp. 733–738, 2015.

[25] J. Gomez-Cambronero, "Phosphatidic acid, phospholipase D and tumorigenesis," *Advances Biol. Regulation*, vol. 54, pp. 197–206, 2014.

[26] M. Reedijk, "Notch signaling and breast cancer," in *Notch Signaling in Embryology and Cancer*. Berlin, Germany: Springer, 2012, pp. 241–257.

[27] D. L. Elion, et al., "Therapeutically active RIG-I agonist induces immunogenic tumor cell killing in breast cancers," *Cancer Res.*, vol. 78, pp. 6183–6195, 2018.

[28] N. E. Hynes and A. Boulay, "The mTOR pathway in breast cancer," *J. Mammary Gland Biol. Neoplasia*, vol. 11, no. 1, pp. 53–61, 2006.

[29] X. Wang and Y. Lin, "Tumor necrosis factor and cancer, buddies or foes? 1," *Acta Pharmacologica Sinica*, vol. 29, no. 11, pp. 1275–1288, 2008.

[30] S. K. Mishra, et al., "In human breast cancer cells thyroid hormone signaling overshadows estrogen signaling on SMP30 gene leading to induction of apoptosis," AACR, 2012. [Online]. Available: https://cancerres.aacrjournals.org/content/72/8_Supplement/4961

[31] E. A. McSherry, et al., "Breast cancer cell migration is regulated through junctional adhesion molecule-A-mediated activation of rap1 GTPase," *Breast Cancer Res.*, vol. 13, no. 2, 2011, Art. no. R31.

[32] M. W. Nasser, et al., "RAGE mediates S100A7-induced breast cancer growth and metastasis by modulating the tumor microenvironment," *Cancer Res.*, vol. 75, pp. 974–985, 2015.

[33] M. W. Kieran, et al., "The VEGF pathway in cancer and disease: responses, resistance, and the path forward," *Cold Spring Harbor Perspectives Med.*, vol. 2, no. 12, 2012, Art. no. a006593.

[34] R. J. Santen, et al., "The role of mitogen-activated protein (MAP) kinase in breast cancer," *The J. Steroid Biochemistry Mol. Biol.*, vol. 80, no. 2, pp. 239–256, 2002.

[35] R. K. La Creis, et al., "Contribution of toll-like receptor signaling pathways to breast tumorigenesis and treatment," *Breast Cancer*, vol. 5, 2013, Art. no. 43.

[36] D. Melck, et al., "Involvement of the camp/protein kinase a pathway and of mitogen-activated protein kinase in the antiproliferative effects of anandamide in human breast cancer cells," *FEBS Lett.*, vol. 463, no. 3, pp. 235–240, 1999.

[37] M. Hornsveld, et al., "FOXO transcription factors both suppress and support breast cancer progression," *Cancer Res.*, vol. 78, pp. 2356–2369, 2018.

[38] I. Azimi, et al., "Calcium influx pathways in breast cancer: Opportunities for pharmacological intervention," *Brit. J. Pharmacology*, vol. 171, no. 4, pp. 945–960, 2014.

[39] G.-B. Jang, et al., "Blockade of wnt/β-catenin signaling suppresses breast cancer metastasis by inhibiting CSC-like phenotype," *Scientific Reports*, vol. 5, 2015, Art. no. 12465.

[40] E. Niemitz, "RAS pathway activation in breast cancer," *Nature Genetics*, vol. 45, no. 11, pp. 1273–1273, 2013.

[41] Z.-J. Liu, et al., "Hypoxia-inducible factor 1 and breast cancer metastasis," *J. Zhejiang University Sci. B*, vol. 16, no. 1, pp. 32–43, 2015.

[42] J. J. Lee, et al., "PI3K/Akt/mTOR inhibitors in breast cancer," *Cancer Biol. Med.*, vol. 12, no. 4, pp. 342–354, 2015.
[43] J. Zhang, et al., "Genetic variations in the hippo signaling pathway and breast cancer risk in african american women in the AMBER consortium," *Carcinogenesis*, vol. 37, pp. 951–956, 2016.
[44] W. Cao, et al., "Relaxin enhances in-vitro invasiveness of breast cancer cell lines by upregulation of S100A4/MMPs signaling," *Eur. Rev. Med. Pharmacological Sci.*, vol. 17, no. 5, pp. 609–617, 2013.
[45] K. Shostak and A. Chariot, "Nf-κb, stem cells and breast cancer: The links get stronger," *Breast Cancer Res.*, vol. 13, no. 4, 2011, Art. no. 214.
[46] M. Gasco, et al., "The p53 pathway in breast cancer," *Breast Cancer Res.*, vol. 4, no. 2, 2002, Art. no. 70.
[47] A. Bottos, et al., "Decreased NK-cell tumour immunosurveillance consequent to JAK inhibition enhances metastasis in breast cancer models," *Nature Commun.*, vol. 7, 2016, Art. no. 12258.
[48] Broad Institute TCGA Genome Data Analysis Center, "PARADIGM pathway analysis of mRNA expression and copy number data," 2016. [Online]. Available: http://gdac.broadinstitute.org/runs/analyses__2016_01_28/reports/cancer /OV-TP/Pathway_Paradigm_mRNA_And_Copy_Number/nozzle.html
[49] S. Ponnusamy, et al., "Sphingolipids and cancer: Ceramide and sphingosine-1-phosphate in the regulation of cell death and drug resistance," *Future Oncology*, vol. 6, no. 10, pp. 1603–1624, 2010.
[50] B. K. Devapatla, et al., "Apelin/apj pathway for targeting ovarian tumor microenvironment," AACR, 2016. [Online]. Available: https://cancerres.aacrjournals.org/content/76/14_Supplement/1272.short
[51] Y. Tekabe, et al., "Targeting rage expression in human ovarian cancer," *Clin. Oncology*, vol. 1, 2016, Art. no. 1055.
[52] K. Kübler, et al., "Targeted activation of RNA helicase retinoic acid–inducible gene-I induces proimmunogenic apoptosis of human ovarian cancer cells," *Cancer Res.*, vol. 70, pp. 5293–5304, 2010.
[53] C. Luquain, et al., "Role of phospholipase D in agonist-stimulated lysophosphatidic acid synthesis by ovarian cancer cells," *J. Lipid Res.*, vol. 44, no. 10, pp. 1963–1975, 2003.
[54] S. Mabuchi, et al., "Targeting mTOR signaling pathway in ovarian cancer," *Current Medicinal Chemistry*, vol. 18, no. 19, pp. 2960–2968, 2011.
[55] S. M. Moghaddam, et al., "Significance of vascular endothelial growth factor in growth and peritoneal dissemination of ovarian cancer," *Cancer Metastasis Rev.*, vol. 31, no. 1/2, pp. 143–162, 2012.
[56] A. S. Dhillon, et al., "Map kinase signalling pathways in cancer," *Oncogene*, vol. 26, no. 22, pp. 3279–3290, 2007.
[57] L. Lu, et al., "Rap1A promotes ovarian cancer metastasis via activation of ERK/p38 and notch signaling," *Cancer Medicine*, vol. 5, no. 12, pp. 3544–3554, 2016.
[58] M. Muccioli and F. Benencia, "Toll-like receptors in ovarian cancer as targets for immunotherapies," *Frontiers Immunology*, vol. 5, 2014, Art. no. 341.
[59] M. Muccioli, et al., "Toll-like receptors as novel therapeutic targets for ovarian cancer," *ISRN Oncology*, vol. 2012, 2012, Art. no. 642141.
[60] D. C. Corney, et al., "Role of p53 and rb in ovarian cancer," in *Proc. Ovarian Cancer*, 2008, pp. 99–117.
[61] A. A. Adjei, "Blocking oncogenic ras signaling for cancer therapy," *J. Nat. Cancer Inst.*, vol. 93, no. 14, pp. 1062–1074, 2001.
[62] A. J. Hanrahan, et al., "Genomic complexity and AKT dependence in serous ovarian cancer," *Cancer Discovery*, vol. 2, no. 1, pp. 56–67, 2012.
[63] R. C. Arend, et al., "The wnt/β-catenin pathway in ovarian cancer: A review," *Gynecologic Oncology*, vol. 131, no. 3, pp. 772–779, 2013.
[64] G. L. Semenza, "Targeting HIF-1 for cancer therapy," *Nature Rev. Cancer*, vol. 3, no. 10, pp. 721–732, 2003.
[65] B. Cheaib, et al., "The PI3K/Akt/mTOR pathway in ovarian cancer: Therapeutic opportunities and challenges," *Chin. J. Cancer*, vol. 34, no. 1, 2015, Art. no. 4.
[66] M. Rae, et al., "Thyroid hormone signaling in human ovarian surface epithelial cells," *The J. Clinical Endocrinology Metabolism*, vol. 92, no. 1, pp. 322–327, 2007.
[67] S. Rangarajan, et al., "Cyclic AMP induces integrin-mediated cell adhesion through EPAC and RAP1 upon stimulation of the β2-adrenergic receptor," *J. Cell Biol.*, vol. 160, no. 4, pp. 487–493, 2003.
[68] C. A. Hall, et al., "Hippo pathway effector yap is an ovarian cancer oncogene," *Cancer Res.*, vol. 70, no. 21, pp. 8517–8525, 2010.
[69] M.-L. Bonnefond, et al., "Calcium signals inhibition sensitizes ovarian carcinoma cells to anti-bcl-xl strategies through mcl-1 down-regulation," *Apoptosis*, vol. 20, no. 4, pp. 535–550, 2015.
[70] J. Szkandera, et al., "Hedgehog signaling pathway in ovarian cancer," *Int. J. Mol. Sci.*, vol. 14, no. 1, pp. 1179–1196, 2013.
[71] M. M. Yung, et al., "Targeting AMPK signaling in combating ovarian cancers: Opportunities and challenges," *Acta Biochimica Biophysica Sinica*, vol. 48, no. 4, pp. 301–317, 2016.
[72] K. L. White, et al., "Genomics of the NF-κb signaling pathway: Hypothesized role in ovarian cancer," *Cancer Causes Control*, vol. 22, no. 5, pp. 785–801, 2011.
[73] M. SC Wilson, et al., "FOXO and FOXM1 in cancer: The FOXO-FOXM1 axis shapes the outcome of cancer chemotherapy," *Current Drug Targets*, vol. 12, no. 9, pp. 1256–1266, 2011.
[74] C. L. Andersen, et al., "Active estrogen receptor-alpha signaling in ovarian cancer models and clinical specimens," *Clinical Cancer Res.*, vol. 23, pp. 3802–3812, 2017.
[75] K. Abubaker, et al., "Inhibition of the JAK2/STAT3 pathway in ovarian cancer results in the loss of cancer stem cell-like characteristics and a reduced tumor burden," *BMC Cancer*, vol. 14, no. 1, 2014, Art. no. 317.
[76] R. Tibshirani, et al., "The lasso method for variable selection in the cox model," *Statist. Medicine*, vol. 16, no. 4, pp. 385–395, 1997.
[77] N. Simon, et al., "Regularization paths for cox's proportional hazards model via coordinate descent," *J. Statistical Softw.*, vol. 39, no. 5, pp. 1–13, 2011. [Online]. Available: http://www.jstatsoft.org/v39/i05/
[78] Y. Yarden and M. X. Sliwkowski, "Untangling the ErbB signalling network," *Nature Rev. Mol. Cell Biol.*, vol. 2, no. 2, pp. 127–137, 2001.
[79] G. Brummer, et al., "Chemokine signaling facilitates early-stage breast cancer survival and invasion through fibroblast-dependent mechanisms," *Mol. Cancer Res.*, vol. 16, no. 2, pp. 296–308, 2018.
[80] F. Gava, et al., "Gap junctions contribute to anchorage-independent clustering of breast cancer cells," *BMC Cancer*, vol. 18, no. 1, 2018, Art. no. 221.
[81] E. Cerami, et al., "The CBIO cancer genomics portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discovery*, vol. 2, no. 5, pp. 401–404, 2012.
[82] J. Friedman, et al., "Regularization paths for generalized linear models via coordinate descent," *J. Statistical Softw.*, vol. 33, no. 1, 2010, Art. no. 1.
[83] V. W. Chen, et al., "Pathology and classification of ovarian tumors," *Cancer: Interdisciplinary Int. J. Amer. Cancer Soc.*, vol. 97, no. S10, pp. 2631–2642, 2003.
[84] Y. Zhao, S. Piekos, T. Hoang, and D.-G. Shin, "A framework using topological pathways for deeper analysis of transcriptome data," in *Proc. 14th Int. Symp. Bioinf. Res. Appl.*, 2018, p. XVIII.

Yue Zhao received the BS degree in applied mathematics from Shandong University, China, in 2012, and the MS degree in statistics from the University of Connecticut, in 2014. He is currently working toward the PhD degree in the Department of Computer Science and Engineering, University of Connecticut. His research interests include machine learning, statistics, and computational biology.

Dong-Guk Shin received the PhD degree in computer science and engineering, University of Michigan, in 1985. He joined the CSE Department at the University of Connecticut (UConn), in 1986 where he currently holds the rank of professor. He established the Bioinformatics and Bio-Computing Institute (BIBCI) at UConn with the NIH/NIGMS planning grant, in 2003. His current research interests include biological data mining and knowledge discovery from multi-omics datasets, specifically with emphasis on discovering gene regulatory pathway targets and creating context-specific gene regulatory pathways.