Causal Disturbance Analysis: A Novel Graph Centrality Based Method for Pathway Enrichment Analysis

Pourya Naderi Yeganeh and M. Taghi Mostafavi

Abstract—Pathway enrichment analysis models (PEM) are the premier methods for interpreting gene expression profiles from high-throughput experiments. PEM often use *a priori* background knowledge to infer the underlying biological functions and mechanisms. A shortcoming of standard PEM is their disregarding of gene interactions for mathematical simplicity, which potentially results in partial and inaccurate inference. In this study, we introduce a graph-based PEM, namely Causal Disturbance Analysis (CADIA), that leverages gene interactions to quantify the topological importance of perturbations in pathway organizations. In particular, CADIA uses a novel graph centrality model, namely Source/Sink, to measure the topological importance. Source/Sink Centrality quantifies a genes importance as a receiver and a sender of biological information, which allows for prioritizing the perturbations that are more likely to disturb a pathways functionality. CADIA infers an enrichment score for a pathway by deriving statistical evidence from Source/Sink centrality of the perturbed genes and combines it with classical over-representation analysis. Through real-world experimental and synthetic data evaluations, we show that CADIA can uniquely infer critical pathway enrichments that are not observable through other PEM. Our results indicate that CADIA is sensitive towards topologically central perturbations and provides a robust framework for interpreting high-throughput data.

Index Terms—Bioinformatics, Enrichment Analysis, Pathway Analysis, Graph Modeling, Genomics.

1 Introduction

A key objective of a typical high-throughput study is to identify the gene profile changes that associate with experimental conditions and phenotypic observations. Ultimately, an important task in this type of study is to interpret the underlying biological mechanisms to draw high-level insight. The most common interpretation approach is inferring the association of the perturbations with known biological processes [1], [2]. The known processes often come from *a priori* curated classes of genes and proteins – including cell functions, localization, disease drivers, and biological pathways [3], [4]. Pathway Enrichment Models (PEM) is an umbrella term for the inference methods that determine the prevalence of *a priori* classes based on the perturbation [1].

Over-representation analysis (ORA) and gene set analysis (GSA) are the two main standard PEM categories [5]. ORA generally evaluates the prevalence (enrichment) of perturbed genes in an *a priori* gene-function class by using cut-off methods for identifying the perturbed genes. ORA determines prevalence (enrichment) of an *a priori* class if the frequency of perturbations in the class is higher than the global frequency of perturbations. ORA uses straightforward statistical tests, such as hypergeometric and Chi-squared, for evaluating the incidence frequency. These tests require to assume that gene perturbations are independent events, which is one of the main limitations of ORA.

• *Authors are with the Department of Computer Science, College of Computing and Informatics, The University of North Carolina at Charlotte, NC, 28223.*
E-mail: pnaderiy@uncc.edu, taghi@uncc.edu

On the other hand, GSA determines a class's prevalence (enrichment) by evaluating the distribution of its members in a global sorted list of genes from the high-throughput experiment. The list is sorted based on some specific parameters such as perturbation intensity or phenotypic correlations. This approach enables GSA to avoid using cut-off parameters. GSA find a class as prevalent if its members are placed close together and tend to appear in the extreme ends of the list [1]. Although ORA and GSA are common in genomic studies, they have a limitation of being oblivious to the interactions between the genes [5]. Similar to ORA, GSA is not sensitive to the underlying topology of the pathways. For example, as long as the ranking is consistent in the sorted list, GSA is not sensitive whether the gene with the highest sorting criteria is topologically central or not.

Biological interactions are integral components of sustaining and proliferating cellular functions. Accordingly, a gene might be centrally involved in several biological mechanisms and interact with many other genes. In this case, the perturbation of the central gene may cause dysfunction in its associated functions. For example, TP53 gene regulates critical processes, and its dysfunction can disrupt DNA-repair and apoptosis in cancers [6]. A well-established body of literature shows that the position of genes/proteins in the biological networks may determine their importance to biological organisms [7], [8]. For instance, Jeong et al. showed that the number of interactions of a gene/protein (degree distribution) correlates with the probability of its removal being lethal [7]. The approach of ORA and GSA in disregarding the interactions results in evaluating the perturbed genes as having the same importance and, consequently, not being sensitive to their topological positions.

Given the existing wealth of information on the interactions, the approach of the standard PEM may deliver partial and inaccurate inference [5].

A new and emerging category of PEM, namely Network-based PEM, focuses on leveraging the topological properties of genes and proteins for identifying the enrichment of *a priori* functional classes. These models, such as Signalling Pathway Impact Analysis (SPIA) and EnrichNet, rely on the underlying networks to detect unique and critical pathway enrichments that are observable through standard PEM [5], [9], [10], [11]. While network-based PEMs have provided new perspectives for biological inference, their abstractions of pathway organizations do not necessarily capture key topological features of pathways. For instance, EnrichNet measures the importance of perturbation by evaluating their distance to different pathways using a global interaction network [11]. However, EnrichNet does not consider the direction of interactions. Likewise, SPIA uses the underlying graph of a pathway to quantify an accumulated perturbation that accounts for regulatory relationships [12]. SPIA dismisses downstream pathway genes as zero importance which results in neglecting some key pathway components from the analysis.

Biological pathways, particularly signaling pathways, often have an upstream-to-downstream organization. This organization indicates the interactions between the genes and proteins have a temporal and biochemical order. The downstream nodes, i.e. the nodes with no out-going interactions, can be critical to the function of pathways. For instance in ErbB signaling, a well-studied cancer pathway, there are several critical oncogenes/tumor suppressors in both of its upstream and downstream ends [13], [14] – including ELK, JUN, and ERBB2. Similarly, most of the signaling pathways contain genes/proteins in their downstream that are critical elements of cellular functions.

Our motivation is to incorporate the underlying structure of the pathways into PEM to obtain an informative inference method. In a recent work, we described an approach to use the number of associated chains of biochemical reactions for enrichment analysis [10]. That model showed power in detecting critical enrichments. However, the model had limitations as it required ignoring some interactions for obtaining directed acyclic graphs of pathways. That model considered all interaction chains as having the same importance and attributed no significance to the downstream nodes.

This study presents a novel graph centrality-based methodology that addresses existing issues of network-based PEM, namely Causal Disturbance Algorithm (CADIA). CADIA evaluates the importance of the upstream/downstream genes and interaction chains in signaling pathways. In particular, CADIA holds the following characteristics for topological evaluations of the genes within a pathway:

- It is sensitive to the direction of interactions, and it conserves the structure and order of interactions.
- Perturbation of a gene/protein relates to the activity of all of its downstream and upstream targets.
- Perturbation of a gene/protein has a stronger effect on its direct targets compared to the indirect ones.

To model these characteristics, we introduce a novel graph centrality model, namely Source/Sink. We then use Source/Sink centrality to construct the pathway enrichment analysis pipeline of CADIA for inferring high-throughput perturbation data. Our experimental and synthetic data evaluations show that CADIA can uniquely detect critical pathway enrichments in a variety of settings compared to standard and state-of-the-art PEM.

2 System and Methods

2.1 Graph Modeling of Pathways

This study uses directed graph representation for modeling the biological pathways. The nodes represent gene-encoded-elements. The edges represent interactions, immediate or mediated by some non-gene-encode elements. Formally, let $G = (V, E)$ represent the graph corresponding to a pathway. The set of vertices (nodes), $V(G) = \{v_1, v_2, \ldots, v_n\}$, represents n distinct elements. The set of edges (links), $E(G) = \{e_1, e_2, \ldots, e_m\}$, represents m distinct directed interactions between the nodes. Each edge, $e_m = (v_i, v_j)$, is an ordered pair that indicates a regulatory or causal relationship from gene-encoded element v_i to v_j.

Chains of regulatory and biochemical interactions can be expressed by using the concept of *walk*; a sequence of the graph nodes $(v_i, \ldots, v_k, v_{k+1}, \ldots, v_j)$ in which any two consecutive vertices are connected by a link, $(v_k, v_{k+1}) \in E$. An ij-walk of a graph is a walk that starts at node i and ends at node j. The size of a walk is the number of edges in a walk. For consistency, this article considers a single node as a walk of length zero. A *closed walk* is a walk in which the start point and the end-point correspond to the same vertex. The transpose of a graph, G^T, is a graph with reversed edge directions. In this case, $V(G^T) = V(G)$ and $E(G^T) = \{(u, v)|(v, u) \in E(G)\}$. We define the *walk-space* of a graph node, $\mathbf{W}_G(v)$, to be the set of all walks that start from the node, $\mathbf{W}_G(v) := \{w_i \mid w_i$: a vu-walk in G$\}$. Any graph with n vertices has an equivalent representation of a $n \times n$ square matrix form, also known as the adjacency matrix, A_G. Formally:

$$[A_G]_{ij} = \begin{cases} 1, & (v_i, v_j) \in E \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The adjacency matrix notation allows for straightforward computation of some graph properties. The adjacency matrix of a transposed graph is the transpose of the adjacency matrix, $A_{G^T} = {A_G}^T$. The total number of ij walks of length k in a graph is the ij^{th} element of the k^{th}power of the adjacency matrix, $[{A_G}^k]_{ij}$. Accordingly, the total number of all walks of length k that start from a node v_i in the graph can be expressed using the following formula:

$$\begin{aligned} \sum_{\substack{w_j \in \mathbf{W}_G(v_i), \\ |w_j| = k}} 1 &= \sum_j [{A_G}^k]_{ij} \\ &= \delta^T(v_i)[{A_G}^k]\mathbb{1} \end{aligned} \tag{2}$$

The above formulation denotes the sum of all elements in the i^{th} row of the adjacency matrix. $\delta^T(v_i)$ is the Kronecker delta which is a row vector of size n where i^{th} location is 1 and zero elsewhere. $\mathbb{1}$ is an $n \times 1$ column vector of size n with 1s for all elements.

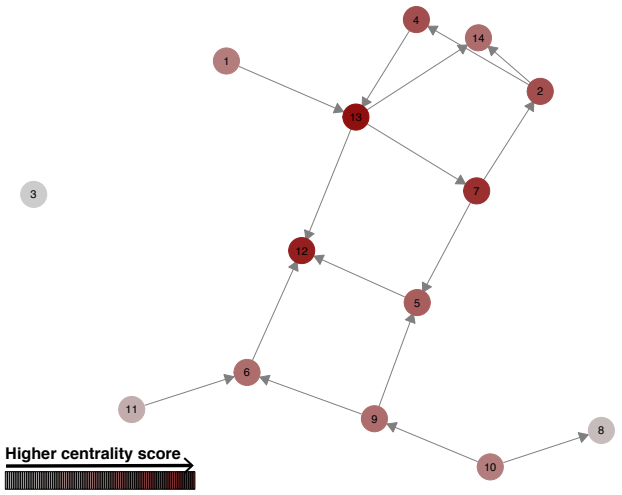

Fig. 1. The figure shows the ranking of the nodes of an example graph according to Source-Sink centrality which measures the importance of a node as a sender and a receiver of information. Here, the more intense colors indicate higher ranking. In comparison, common centrality models only focus on nodes as a sender of information of the nodes. Subsequently, in the standard models, the nodes at the receiving ends lose their importance. For example, node 12 in this figure would have minimal centrality value based on standard models because it has no outgoing edges. The figure is based on a random graph generated using the network package and visualized using "GGally" r package [16].

2.2 Source/Sink Centrality

A graph centrality is a function from $V(G)$ to real numbers for describing a topological ranking (importance) of the nodes in a network [15]. To calculate the centrality of a gene/protein in a biological pathway, we introduce a novel graph centrality model that is designed to quantify how the disturbance of a node can affect the activities of the downstream and upstream targets separately. In this context, the disturbance can be up-regulation, down-regulation, mutation, or any other gene-level perturbation. In our model, a node can be central as a sender (source) or a receiver (sink) of information. This concept considers a node as topologically central if it is the downstream target (receiver) of many chains of biochemical reactions, or many chains of biochemical reactions initiate from that node (sender).

In the example network illustrated in Figure 1, disturbance of node 13 can affect upstream and downstream interactors. In contrast, changes in node 3 do not affect any other nodes. With this condition, the importance of a node v derives from aggregating the number of possible walks that either start from v or end at v. The additive contribution of each walk to the centrality is dependent on its length. Two separate components quantify this concept; one is the centrality of a node as a source, and the other is as a sink. Formally:

$$C_{Source}(v) := \sum_{w_j:\ vu-walk\ of\ G} \alpha^{|w_j|} \tag{3}$$

Here, α is a dampening factor that ensures the walks of larger lengths to have smaller contributions to the Source centrality of node v. This parameter ensures two conditions. First, perturbation of a gene/protein has a greater effect on direct targets compared to indirect ones. Second, the effect of a perturbation is the same on all immediate gene/proteins. In the example of Figure 1, a perturbation of node 10 is more likely to cause a disturbance in the activities of nodes 8 and 9, compared to those of nodes 12 or 5. This formulation allows to handle an infinite number of walks as any graph with closed walks has an infinite number of walks. For example in Figure 1, the sequence $(4, 13, 7, 2, 4)$ forms a loop and it can create walks as large as any number. A small enough choice for α guarantees that Formula 3 converges. Formula 3 quantifies the importance of a node as a source (sender) for other nodes in G. Similarly, the definition of the centrality of a node as a sink (receiver) is:

$$\begin{aligned} C_{Sink}(v) :&= \sum_{w_j:\ uv-walk\ of\ G} \alpha^{|w_j|} \\ &= \sum_{w_j \in \mathbf{W}_{G^T}(v)} \alpha^{|w_j|} \end{aligned} \tag{4}$$

The above equality holds since any uv-walk of G is a vu-walk of G^T. In this context, the centrality of a node v is a weighted sum of the walks that end at v. This formulation assigns higher centrality values to the nodes that are the endpoint of many walks. For example in Figure 1, node 3 has no incoming edge and node 12 has three incoming edges. We then define the Source/Sink centrality as the sum of the two separate components from Formulas 3 and 4.

$$C_{ssc}(v) := C_{Source}(v) + \beta C_{Sink}(v) \tag{5}$$

Here, β is zero or a positive real number and denotes the relative importance of source component versus sink. When β takes small values, the Source/Sink centrality shifts towards the capacity of the nodes as sources. When β grows larger, the centrality shifts towards higher the sink capacity. For computing C_{ssc}, Equations 3,4, and 5 can be expressed using adjacency matrices. The following formulations calculate the Source centrality:

$$\begin{aligned} C_{Source}(v) &= \sum_k \sum_{\substack{w_j \in \mathbf{W}_G(v), \\ |w_j|=k}} \alpha^{|w_j|} \\ &= \sum_k \alpha^k \sum_{\substack{w_j \in \mathbf{W}_G(v), \\ |w_j|=k}} 1 \end{aligned} \tag{6}$$

Replacing the inner sum with Formula 2 gives:

$$\begin{aligned} C_{Source}(v) &= \sum_k \alpha^k \delta^T(v)[A^k]\mathbb{1} \\ &= \delta^T(v)\left[\sum_{k=0}^{\infty} \alpha^k A^k\right]\mathbb{1} \end{aligned} \tag{7}$$

A sufficient condition for the summation to be convergent is $\alpha \leq 1/\lambda_1$, where λ_1 is the largest positive eigenvalue of the adjacency matrix [17]. If $\alpha < 1/\lambda_1$, then $I - \alpha A$ is invertible. Having this condition, summarizing the portion of the equation in the bracket will give:

$$\begin{aligned} \sum_{k=0}^{\infty} \alpha^k A^k &= (I - \alpha A)^{-1}(I - \alpha A) \sum_{k=0}^{\infty} \alpha^k A^k \\ &= (I - \alpha A)^{-1} \end{aligned} \tag{8}$$

Sink centrality can be derived using the same procedure. The same α would also work for the transpose graph because the set of eigenvalues of a matrix and its transpose are equivalent. Subsequently:

$$C_{Source}(v_i) = \delta^T(v_i)(I - \alpha A)^{-1}\mathbb{1} \quad (9)$$

$$C_{Sink}(v_i) = \delta^T(v_i)(I - \alpha A^T)^{-1}\mathbb{1} \quad (10)$$

Plugging the above formulas into Formula 5 gives:

$$C_{ssc}(v_i) = \delta^T(v_i)\big[(I - \alpha A)^{-1} + \beta(I - \alpha A^T)^{-1}\big]\mathbb{1} \quad (11)$$

The above formulation allows for accounting for a node as both sender and receiver (Figure 1). In pathway annotations, the upstream genes are mainly senders (source) signaling. Likewise, the downstream genes/proteins are mainly receivers of the signaling. The Source-Sink Centrality allows attributing topological importance to both upstream and downstream nodes. Individual formulas of source and sink centrality are closely related to Katz-Bonacich centrality which is a standard method in the study of social networks [18]. In Katz-Bonacich centrality model, nodes with no out-degree would have the same centrality value of zero.

2.3 Constructing a PEM using Source/Sink Centrality

CADIA uses Source/Sink centrality to calculate an enrichment score for a set of perturbed genes. In particular, CADIA extracts a topological statistical evidence by measuring the aggregated importance of the perturbed genes. CADIA also uses an additional statistical evidence from ORA and combines it with Source-Sink Centrality to increase sensitivity, Similar to a methodology that was used in SPIA [12].

In contrast to the regular use of centrality models, where the individual centrality of the nodes is important, CADIA measures the centrality of the set of the perturbed genes. A notion of accumulated centrality for a subset of nodes quantifies this concept. Let $U = \{u_1, u_2, \ldots, u_m\}$ denote a subset of $V(G)$. We measure the aggregate centrality, namely causal disturbance, of U by using the following:

$$Agg(U) := \prod_{u_i \in U} C_{ssc}(u_i) \quad (12)$$

CADIA calculates the statistical significance of the aggregate score from an observed set of perturbations using a bootstrap sampling approach. Let $Agg(U_0)$ denote an observed causal disturbance of a pathway from m perturbed genes. The statistical significance of $Agg(U_0)$ is:

$$P_{ssc} = \mathbb{P}\Big\{Agg(U) > Agg(U_0) \;\Big|\; |U| = |U_0|\Big\} \quad (13)$$

P_{ssc} denotes the probability of observing a higher aggregate source/sink score (causal disturbance) when a subset U of $V(G)$ with size k is randomly selected. CADIA uses the probability density function (PDF) of $Agg(U)$ is to extract the P_{ssc} by calculating the right-hand side area under the PDF curve.

A property of P_{ssc} is that it remains invariant under a broad range of manipulations. For example, P_{ssc} is invariant to any positive scaling in the Formula 11. This allows for rearranging the definition of $C_{ssc}(\,)$ in a more symmetrical representation, the illustration and proof of this rearrangement is provided in the supplementary material. The definition P_{ssc} also allows for manipulation in the definitions of $Agg(\,)$ for facilitating implementation and computational purposes. Formally, Formula 12 can be re-written by taking logarithm from its right-hand side as follows:

$$\begin{aligned} Agg^*(U) &:= log\Big(\prod_{u_i \in U} C_{ssc}(u_i)\Big) \\ &:= \sum_{u_i \in U} log(C_{ssc}(u_i)) \end{aligned} \quad (14)$$

Since the objective is to evaluate the extremeness of an observed aggregate score, the logarithm operation does not change P_{ssc} on the condition that $log(C_{ssc}(v_i)) \geq 0$. This condition is satisfied because the minimum centrality of each node according to the definition of the Source/Sink centrality equals one. Formally:

$$\begin{aligned} P_{ssc} &= \mathbb{P}\Big\{Agg(U) > Agg(U_0) \;\Big|\; |U| = |U_0|\Big\} \\ &= \mathbb{P}\Big\{Agg^*(U) > Agg^*(U_0) \;\Big|\; |U| = |U_0|\Big\} \end{aligned} \quad (15)$$

The use of the product-based evaluation in P_{ssc} allows for creating sensitivity towards the experimental observations where the perturbed genes mainly have intermediate centrality values. The biological networks may contain hubs (nodes with extremely high centrality). A summation-based evaluation would dismiss the mainly-intermediate-centrality cases as non-significant in favor of the cases with few hubs and majority low-centrality. The product-based aggregate score (Formula 14) also allows increasing specificity. In particular, it smooths the aggregate score in the the instances where the perturbation set contains only a few hubs (e.g. a single hub), and the rest of the elements are unimportant nodes.

2.4 combining Source/Sink Centrality with ORA

To increase the sensitivity, CADIA uses additional statistical evidence from ORA. This use of additional evidence is similar to the approach of SPIA [12]. This study uses the hypergeometric test to calculate the p-values of over-representation analysis (P_{ora}). Formally, suppose that an experimental procedure evaluates individual perturbations for $B = \{b_1, b_2, \ldots, b_l\}$ genes. In practical cases, B is the set of all of the transcripts of the high-throughput machinery. Also, let $D = \{d_1, d_2, \ldots, d_k\}$ be the perturbed genes, $D \subset B$. Let $V = \{v_1, v_2, \ldots, v_n\}$ be the nodes of a pathway and m be the number of pathway genes that are perturbed, $|D \cap V| = m$. Let X be the random variable that denotes the number of perturbed genes in the pathway. The over-representation p-value of the pathway is the probability of observing m or more perturbations in the pathway, given an overall perturbation set of size k. Hypergeometric test statistics is one of the methods to calculate the over-representation p-values [5]. The p-value of over-representation, P_{ora}, is formally defined as:

$$P_{ora} := \mathbb{P}\Big\{X > m\Big\} \sim hyper(k, l, m, n) \quad (16)$$

P_{ora} and P_{ssc} are independent because given any m, the knowledge of P_{ora} does not add any information regarding P_{ssc}. A formal proof can be constructed by using the definition of Formula 13; P_{ssc} is independent from $\mathbb{P}\{X = |U_0|\}$ because the definition of P_{ssc} contains a condition of$\{|U| = |U_0|\}$. Similarly, P_{ssc} is independent from $\mathbb{P}\{X = |U_0| + i\}$ for all values of i. Also, the probabilities $\mathbb{P}\{X = |U_0|+i\}$ are mutually exclusive for all i's. Therefore, P_{ssc} and $\sum_i \mathbb{P}\{X = |U_0| + i\}$ are independent. Here, the summation of probability adds up to P_{ora}.

Given the independence of P_{ora} and P_{ssc}, it is possible to combine them into one test-statistic for producing higher statistical power. Fisher's method for meta-analysis uses Chi-square estimates to combine independent p-values [19]. In particular, let the random variable X indicate the product of P_{ora} and P_{ssc}. The chi-squared test calculates the probability of observing smaller values for the product. A Chi-squared test with four degrees of freedom [19] estimates this combined p-value as following:

$$\begin{aligned} P_{cadia} &= \mathbb{P}\Big\{X \leq P_{ora}.P_{ssc}\Big\} \\ &= -2[ln(P_{ssc}) + ln(P_{ORA})] \sim \chi^2_4 \end{aligned} \tag{17}$$

Here, P_{cadia} denotes the combined probability of the topological evidence (P_{ssc}) and the ORA evidence (P_{ora}). From a computational perspective, the time complexity of CADIA is similar to SPIA. For a pathway of size n with m rounds of sampling, the time complexity is of $O(n^3 + m.n)$. The n^3 component is for a one-time calculation of Source-Sink Centrality, which depends on a matrix inversion. The $m.n$ component depends on the number of sampling rounds for n random perturbed gene (n is the upper-bound).

3 Model Evaluation

We evaluate CADIA from two aspects. First, we determine whether CADIA can make unique and informative inferences by using multiple real-world experimental and comparison with other PEM. Second, to ensure that inference of CADIA is not a result of false-positives, we evaluate CADIA by testing against randomly generated inputs. PEMs are primarily exploratory tools with the purpose of detecting possible pathway associations of some experimental data. A direct proof for the superiority of one PEM versus others is often not feasible because of the exploratory objective and the complex nature of biological systems. For this reason, PEM are mainly evaluated by testing on real-world datasets and comparison with other existing models [1], [12]. This section outlines the experimental evaluation procedures of CADIA on three independent datasets. We also outline synthetic data evaluation procedures to verify the quality of CADIA's pipeline.

3.1 Experimental Data Evaluation

We used three real-world datasets to evaluate the inference of CADIA– all retrieved from the NCBI gene expression omnibus [20], [21]. For consistency, all three were from mRNA expression datasets from the common platforms. Also, the datasets were chosen from cancer-related datasets because of the abundance and depth of literature on signaling pathways in cancers, allowing to contrast CADIA results against existing evidence and other methods [6]. We compared CADIA to other PEMs including SPIA, ORA, GSA, and EnrichNet.

First, we used an ovarian cancer dataset by Bowtell and colleagues, which contained 60 High-grade serous ovarian cancer and 30 Low malignant potential tumors [20]. This data was retrieved using the accession code GSE12172. Second, we used a colorectal cancer gene expression dataset by Mogushi and colleagues, retrieved using the accession code GSE21510. We selected the subset of 25 normal colon tissues and 19 homogenized cancer tissues from this dataset for differential expression analysis [21]. Third, we used a dataset from gastric cancer patients by Bing Ya and colleagues that contained 21 normal samples and 111 cancer samples, retrieved using the accession code GSE54129.

For each dataset, the log of RMA normalized mRNA expressions was used to calculate differential expressions. Limma package was used to calculate the significance and fold-change of each differential expression [22]. The adjusted p-values for each differential expression were subjected to multiple hypothesis testing using the Benjamini-Hochberg False Discovery Rate (FDR) [23]. Each dataset was subjected to a specific fold-change (FC) and FDR criteria for gene selection to create differential expression sets from multiple settings and different sizes. In particular, GSE12172 was subjected to the filtering criteria $|FC| < 1$ and $FDR < 0.05$. GSE21510 was subjected to the filtering criteria $|FC| < 1$ and $FDR < 0.005$. GSE54129 was subjected to the filtering criteria $|FC| < 3$ and $FDR < 0.05$.

3.2 Background Pathways

All the PEMs in this study used the biological pathways from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [4]. SPIA and EnrichNet used internal lists of pathways. We used KEGGGraph package in R to parse the pathway for CADIA, ORA, and GSA [24]. All pathways were selected from KEGG classifications of *"Environmental Information Processing"*, *"Cellular Processes"*, *"Organismal System"*, *"Human Diseases"*, and *"Drug Development"*.

Some of the pathways potentially had incomplete information. Incomplete and partial information may cause inconsistency in a PEM [25]. Therefore, some pathway were excluded from the analysis to preserve consistency of graph analysis. The exclusion criteria were 1– pathways contained more than 50% abandoned nodes (without any edges), 2– their largest connected component was less than ten nodes, and 3– their edge count was less than 20. A total of 51 pathways exhibited these characteristics. Also, five pathways with the largest eigenvalue of more than 10 were excluded from analysis since they imposed too small values of α for CADIA. A final set of 143 pathways passed the analysis criteria for CADIA. The Supplementary Table provides a list of the pathways analyzed in CADIA.

3.3 Comparison of PEMs

We analyzed the differentially expressed genes from the three datasets by using CADIA, ORA, GSA, EnrichNet, and SPIA. The methods calculate enrichment p-values for each

pathway. For each dataset, there were multiple pathways analyzed by each model. To control the expected values of false positives, we used the Benjamini-Hochberg False Discovery Rate (FDR) criteria to correct the p-values. A pathway enrichment score was considered statistically significant if its respective FDR-corrected p-value was less than 0.05 ($FDR \leq 0.05$).

P-values of Source-Sink Centrality (P_{ssc}) in CADIA were calculated based on 10000 rounds of iteration for bootstrap sampling. A bootstrap sampling of 10^4 rounds allows computing p-values as small as 10^{-4}. The parameters $\alpha = 0.1$ and $\beta = 1$ were used for calculating Source-Sink Centrality in CADIA. The choice of $\beta = 1$ ensures that Source-Sink centrality is maximally distinct from the Source component and the Sink component (proof in the Supplementary Material). As for the parameter α, we are interested in having the largest possible values, while preserving a reasonable coverage of pathways. Source-Sink centrality is closely related to Katz-Bonacich model, and prior studies have shown that the choice of α in Katz model can strongly affect the centrality rankings [18]. For these reasons, $\alpha = 0.1$ is the maximal choice to ensure the pathway coverage and the convergence of Source-Sink calculation (only 5 pathways had to be excluded because they required a smaller choice of α, see Subsection 3.2).

Over-representation p-values (P_{ora}) were calculated using hypergeometric test. SPIA p-values were calculated using the SPIA R package [12]. GSA p-values were calculated using two available implementations GAGE and F-GSEA [26], [27]. EnrichNet pathway analysis was accessed through its online portal [11]. All the data analysis in this study were performed in R and related Bioconductor packages when possible [28].

3.4 Synthetic Data Evaluation

We tested CADIA on random inputs of different sizes to verify that the results were not outcomes of false positives. We also tested ORA on the same randomly generated data to contrast P_{ora} and P_{ssc}. Ideally, a PEM should not detect significant enrichments for a randomly selected input. In practice, a test of random data may generate false positive. Therefore, it is desired to measure the rates using a controlled false positive criterion. The synthetic evaluation procedure of CADIA to measure false positive rates was:

1) Set $n = 100$.
2) Select a random subset of n genes.
3) Calculate P_{ora}, P_{ssc} and P_{cadia}.
4) Evaluate the number of enriched pathways by each method ($FDR \leq 0.05$).
5) Repeat steps 2, 3, and 4 for 10 times and record the average number of false positives.
6) If $n \leq 5000$, do $n = n + 100$ and go to step 2.

CADIA and ORA parameters and background data are described in subsection 3.3. The 10 repeats at each input size allow more accurate estimates of false positives rates. The 100–5000 range provides a variety of reasonable input sizes for measuring CADIA and ORA. The Supplementary Material provides additional evaluation of the null distribution of P_{ssc} and P_{ora}.

In addition, we applied the Source/Sink centrality to the ErbB signaling pathway to showcase its ranking procedure. ErbB signaling is a suitable choice for in-depth analysis because of 1- the existence of extensive literature on its mechanisms, 2- being a suitable example of upstream/downstream mechanisms [13], 3- its small size for visualization purposes. We compared Source/Sink centrality to three other well-known centrality models; Degree centrality, Betweenness centrality, and Katz centrality. A comprehensive description of these models and their applications can be found in the reference [15]. To compare the models, we used the ranking of the nodes produced by each centrality method. A higher rank value indicates higher centrality. In the case of having the same values, the minimum rank was assigned to all ties.

4 Results and Discussion

CADIA detected critical pathway enrichments for ovarian cancer, colorectal cancer, and gastric cancer that were not observable by SPIA, ORA, GSA, and EnrichNet – supported by evidence from the literature. Also, CADIA dismisses some pathway enrichments from SPIA and ORA, many of which do not have any particular association with the experimental data. ORA and SPIA are the most relevant methods to CADIA, and their comparisons are provided in Tables 1–9. The Supplementary Material contains the evaluations of EnrichNet and two implementations of GSA, both of the methods fail to capture numerous significant pathways that are detected by CADIA. Our synthetic data evaluation provides insight regarding the performance of CADIA and reliability of its results. The synthetic data evaluation shows that CADIA is not prone to make false positives above the expected level. A case of example analysis provides insight regarding the performance of Source/Sink Centrality compared to standard centrality models.

4.1 Experimental Data Evaluation

Ovarian Cancer Data (GSE12172): Based on 1333 differentially expressed genes in the ovarian cancer dataset, CADIA uniquely identifies three pathways — *PI3K-AKT signaling*, *Focal Adhesion*, and *Ras signaling* pathways (Table 1). These are well-studied perturbed pathways in ovarian cancer [29], [30].

CADIA detects enrichment of PI3K-AKT signaling (FDR-corrected p-value $\leq 7.82 \times 10^{-3}$) by utilizing a P_{ssc} of $\leq 2.5 \times 10^{-3}$. PI3K-AKT is a cancer associated pathway that regulates many critical cellular mechanisms, including cellular proliferation, survival, and apoptosis [31], [32], [33], [34]. PI3K-AKT is activated in ovarian cancer and its utility for therapeutic approaches [29], [31], [32], [35]. Similarly, CADIA detects enrichment of Focal Adhesion pathway (FDR-corrected p-value $\leq 2.41 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 9.80 \times 10^{-3}$. Focal Adhesion is well studied in cancers – particularly ovarian cancer –and is associated with cellular migration, proliferation, and differentiation [36]. In addition, CADIA detects enrichment of Ras signaling pathway (FDR-corrected p-value $\leq 2.20 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 2.00 \times 10^{-4}$. Ras signaling activates cellular proliferation and growth and is associated cancers [31], [37].

CADIA discards some pathways with insignificant Source/Sink topological evidence, some of which not having clear connections to ovarian cancer. For example, SPIA detects *cytokine-cytokine receptor interactions* pathway which is not detected by ORA or CADIA. Our literature search failed to identify any established results for the association of this pathway with ovarian cancer. On the other hand, CADIA provides strong topological evidence for *"Pathways in cancer"* for which SPIA fails to provide topological evidence (More details in supplementary Table S.II). Also, SPIA uniquely detects *mineral absorption* pathway for which the literature search failed to identify any established results for its association ovarian cancer. *Mineral absorption* was among the pathways that did not pass the quality criteria and did not qualify for CADIA because of its incomplete information.

TABLE 1
Statistically significant pathway enrichments identified by CADIA from Ovarian Cancer Data (GSE12172)

Name§	ID	P_{ora}	P_{ssc}	**CADIA**†	FDR_{ora}†
MicroR...	05206	3.66e-08	2.65e-01	2.70e-05	5.23e-06
Oocyte ...	04114	3.13e-04	3.00e-04	1.09e-04	1.49e-02
p53 sig...	04115	2.83e-07	4.80e-01	1.09e-04	2.02e-05
***PI3K-...**	04151	7.34e-03	2.50e-03	7.82e-03	8.31e-02
***Ras si...**	04014	3.65e-01	2.00e-04	2.20e-02	9.97e-01
***Focal...**	04510	1.01e-02	9.80e-03	2.41e-02	9.02e-02
Proges...	04914	4.82e-04	3.51e-01	3.19e-02	1.72e-02
Pathwa...	05200	2.30e-03	8.08e-02	3.19e-02	4.71e-02

§ Names truncated for space limitation
† FDR corrected P_{cadia} and P_{ora}
* Unique to CADIA

TABLE 2
Statistically significant pathway enrichments identified by ORA from Ovarian Cancer Data (GSE12172)

Name§	ID	CADIA†	$\mathbf{FDR_{ora}}$†
MicroRNAs in cancer	05206	2.70e-05	5.23e-06
p53 signaling pathway	04115	1.09e-04	2.02e-05
Oocyte meiosis	04114	1.09e-04	1.49e-02
Progesterone-mediated oocyt...	04914	3.19e-02	1.72e-02
***Proteoglycans in cancer**	05205	7.59e-02	3.29e-02
ECM-receptor interaction	04512	6.85e-02	4.63e-02
Pathways in cancer	05200	3.19e-02	4.71e-02

§ Names truncated for space limitations
† FDR corrected P_{cadia} and P_{ora}
* Unique to ORA

TABLE 3
Statistically significant pathway enrichments identified by SPIA from Ovarian Cancer Data (GSE12172)

Name	ID	**SPIA**†	CADIA†
***Cell cycle**	04110	6.38e-09	NA
p53 signaling pathway	04115	1.11e-04	1.09e-04
***Chemokine signaling pathway**	04062	4.85e-04	3.17e-01
***Mineral absorption**	04978	1.48e-02	NA
Oocyte meiosis	04114	1.73e-02	1.09e-04
***Cytokine-cytokine receptor...**	04060	1.73e-02	4.76e-01
Progesterone-mediated oocyt...	04914	3.77e-02	3.19e-02

† FDR corrected p-values
* Unique to SPIA
NA: Not Analyzed by CADIA

TABLE 4
Statistically significant pathway enrichments identified by CADIA from Colorectal Cancer Data (GSE21510)

Name§	ID	P_{ora}	P_{ssc}	**CADIA**†	FDR_{ora}†
Oocyt...	04114	6.02e-05	1.10e-03	8.30e-05	1.72e-03
p53 s...	04115	9.48e-08	4.34e-01	8.30e-05	1.36e-05
Pathw...	05200	1.17e-06	9.45e-01	7.77e-04	8.39e-05
Micro...	05206	5.09e-06	4.22e-01	1.08e-03	2.43e-04
PPAR ...	03320	1.09e-05	3.24e-01	1.37e-03	3.89e-04
HTLV-...	05166	1.31e-04	4.95e-01	1.64e-02	3.11e-03
Proge...	04914	8.65e-04	1.20e-01	1.88e-02	1.55e-02
***Olfa...**	04740	9.97e-01	1.00e-04	1.88e-02	9.97e-01
***Hipp...**	04390	6.58e-03	5.64e-02	3.77e-02	6.27e-02
***Phos...**	04072	6.01e-02	7.00e-03	3.77e-02	1.95e-01
***Apop...**	04210	4.56e-02	8.00e-03	3.77e-02	1.64e-01
Chemo...	04062	2.89e-03	1.12e-01	3.77e-02	3.47e-02
***GnRH...**	04912	2.03e-02	1.81e-02	3.77e-02	1.02e-01
***Vasc...**	04270	2.15e-02	2.12e-02	3.77e-02	1.02e-01
Small...	05222	7.35e-04	5.88e-01	3.77e-02	1.50e-02
***Calc...**	04020	2.51e-02	2.50e-02	4.70e-02	1.16e-01

§ Names truncated for space limitations
† FDR corrected P_{cadia} and P_{ora}
* Unique to CADIA

Colorectal Cancer Data (GSE21510): Based on 2625 differentially expressed genes, CADIA uniquely detects six pathway enrichments in colorectal cancer data including *Apoptosis* and *Hippo Signaling* (Table 4).

CADIA detects enrichment of Apoptosis pathway (FDR-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 8.00 \times 10^{-3}$. Dysfunction of Apoptosis pathway – programmed cell death– is an important feature of cancers, and in particular colorectal cancer [6], [38]. Similarly, CADIA detects enrichment of *Hippo signaling* pathway (FDR-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 5.64 \times 10^{-2}$. Hippo signaling is well-studied in human neoplasms and control cellular proliferation and apoptosis [39]. CADIA detects enrichment of *GnRH signaling* pathway (FDR-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 1.81 \times 10^{-2}$. Literature evidence also show mechanisms in which GnRH signaling affects colorectal cancer [40]. Similarly, CADIA detects enrichment of *Phospholipase D signaling* pathway (fdr-corrected p-value $\leq 3.77 \times 10^{-2}$) by utilizing a P_{ssc} of $\leq 7.00 \times 10^{-3}$. Phospholipase D signaling is related to colorectal cancer through connections with Wnt signaling [41], [42].

ORA uniquely detects enrichment of Thyroid hormone signaling pathway for which the literature search did not find any results in supports of its association with colorectal cancer. Similarly, SPIA detects pathways that are not necessarily related to colorectal cancer such as Alzheimer's, Amoebiasis, Bile secretion, and Pancreatic Secretion (Table 6). CADIA did not calculate p-values for some of these unique SPIA pathways (NA entries in Table 6) because of their exclusion from analysis due to incomplete information. For example, SPIA detects a strong topological evidence for *Alzheimer's disease* pathway (pPERT = 5×10^{-6}, Supplementary Material Table SIII) which is not necessarily related to colorectal cancer. The detection of Alzheimer's is an instance where the incomplete information causes irrelevant outcomes for Network-based PEM by producing strong topological evidence.

TABLE 5
Statistically significant pathway enrichments identified by ORA from Colorectal Cancer data (GSE21510)

Name[§]	ID	CADIA[†]	$\mathbf{FDR_{ora}}$[†]
p53 signaling pathway	04115	8.30e-05	1.36e-05
Pathways in cancer	05200	7.77e-04	8.39e-05
MicroRNAs in cancer	05206	1.08e-03	2.43e-04
PPAR signaling pathway	03320	1.37e-03	3.89e-04
Oocyte meiosis	04114	8.30e-05	1.72e-03
HTLV-I infection	05166	1.64e-02	3.11e-03
Small cell lung cancer	05222	3.77e-02	1.50e-02
Progesterone-mediated oocyte...	04914	1.88e-02	1.55e-02
Chemical carcinogenesis	05204	5.92e-02	2.95e-02
***TGF-beta signaling pathway**	04350	8.87e-02	3.02e-02
Chemokine signaling pathway	04062	3.77e-02	3.47e-02
***Proteoglycans in cancer**	05205	5.92e-02	3.47e-02
***Thyroid hormone signaling...**	04919	1.37e-01	4.73e-02

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia} and P_{ora}
[*] Unique to ORA

TABLE 6
Statistically significant pathway enrichments identified by SPIA from Colorectal Cancer data (GSE21510)

Name	ID	**SPIA**[†]	CADIA[†]
***Cell cycle**	04110	4.86e-16	NA
p53 signaling pathway	04115	1.71e-05	8.30e-05
***RNA transport**	03013	1.27e-04	NA
PPAR signaling pathway	03320	3.54e-04	1.37e-03
***Mineral absorption**	04978	3.54e-04	NA
***Alzheimer's disease**	05010	9.23e-04	NA
HTLV-I infection	05166	1.46e-03	1.64e-02
***Amoebiasis**	05146	6.19e-03	NA
Oocyte meiosis	04114	7.68e-03	8.30e-05
***Bile secretion**	04976	9.12e-03	NA
Pathways in cancer	05200	9.12e-03	7.77e-04
***ECM-receptor interaction**	04512	1.21e-02	1.23e-01
Progesterone-mediated oocy...	04914	1.66e-02	1.88e-02
Small cell lung cancer	05222	1.91e-02	3.77e-02
Chemokine signaling pathway	04062	2.15e-02	3.77e-02
***Gap junction**	04540	2.76e-02	7.66e-02
***Transcriptional misregulat...**	05202	2.87e-02	NA
***Wnt signaling pathway**	04310	3.02e-02	8.87e-02
***Pancreatic secretion**	04972	4.99e-02	NA

[†] FDR corrected p-values
[*] Unique to SPIA
[NA:] Not Analyzed by CADIA

Gastric Cancer Data (GSE54129): Based on 133 differentially expressed genes, CADIA uniquely detects Wnt Signaling pathway in gastric cancer (Table 7). In the case of gastric cancer, CADIA detects Wnt signaling (FDR-corrected p-value $\leq 9.38 \times 10^{-3}$) by utilizing a P_{ssc} of $\leq 1.00 \times 10^{-4}$.

Wnt signaling is among the most well-studied cancer pathways, and there is a plethora of evidence for its activation in cancers including gastric [43]. This case indicates that the perturbed genes of the Wnt pathway are substantially important in the structure and makes the case of why a structural pathway analysis can detect unique discoveries. Compared to CADIA, ORA detects enrichment of Renin Secretion and Vascular muscle contractions, for which the literature suggests no particular relevance to the disease. Similarly, SPIA detects Ameobiasis and Malaria. The literature search failed to identify any established results for the association of these pathways with gastric cancer. These pathways were among the list that did not pass the quality criteria and did not qualify for CADIA because of their incomplete information.

The three presented experimental test cases indicate that the use of Source/Sink centrality in CADIA enables detection of critical pathway enrichment from biological data. Source/Sink centrality allows for attributing higher importance to the nodes that are missed by other network-based methods such as SPIA. Small p-values of Source/Sink centrality evidence indicates that CADIA is sensitive to perturbation of topologically central genes that are also important to a pathway's functionality. Although small p-values do not guarantee the dysfunction of any pathway, the support of literature for the experimental data shows the ability of CADIA in making an informative enrichments.

TABLE 7
Statistically significant pathway enrichments identified by CADIA from Gastric Cancer data (GSE54129)

Name[§]	ID	P_{ora}	P_{ssc}	**CADIA**[†]	FDR_{ora}[†]
ECM-r...	04512	1.44e-07	6.23e-01	2.29e-04	2.13e-05
Focal...	04510	2.44e-06	4.60e-01	8.15e-04	1.21e-04
Gastr...	04971	1.07e-06	9.94e-01	8.15e-04	7.91e-05
Chemi...	05204	1.23e-03	7.10e-03	4.08e-03	2.76e-02
***Wnt ...**	04310	2.76e-01	1.00e-04	9.38e-03	9.88e-01
PI3K-...	04151	2.13e-04	3.24e-01	1.81e-02	7.89e-03

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia} and P_{ora}
[*] Unique to CADIA

TABLE 8
Statistically significant pathway enrichments identified by ORA from Gastric Cancer data (GSE54129)

Name[§]	ID	CADIA[†]	$\mathbf{FDR_{ora}}$[†]
ECM-receptor interaction	04512	2.29e-04	2.13e-05
Gastric acid secretion	04971	8.15e-04	7.91e-05
Focal adhesion	04510	8.15e-04	1.21e-04
PI3K-Akt signaling pathway	04151	1.81e-02	7.89e-03
***Renin secretion**	04924	1.35e-01	2.76e-02
Chemical carcinogenesis	05204	4.08e-03	2.76e-02
***AGE-RAGE signaling pat...**	04933	1.35e-01	2.76e-02
***Vascular smooth muscle...**	04270	1.26e-01	2.95e-02

[§] Names truncated for space limitations
[†] FDR corrected P_{cadia} and P_{ora}
[*] Unique to ORA

TABLE 9
Statistically significant pathway enrichments identified by SPIA from Gastric Cancer data (GSE54129)

Name	ID	**SPIA**[†]	$CADIA$[†]
ECM-receptor interaction	04512	2.02e-09	2.22e-04
Gastric acid secretion	04971	5.39e-06	7.97e-04
Focal adhesion	04510	1.69e-05	7.97e-04
***TGF-beta signaling pathway**	04350	4.54e-03	4.06e-01
***Malaria**	05144	4.54e-03	NA
Cytokine-cytokine receptor....	04060	4.39e-02	5.09e-01
***Amoebiasis**	05146	4.43e-02	NA
Vascular smooth muscle con...	04270	4.43e-02	1.23e-01

[†] FDR corrected p-values
[*] Unique to SPIA
[NA:] Not Analyzed by CADIA

The variety of differential expression set sizes in the experimental evaluation indicates the sensitivity of CADIA towards both small and large sets of perturbations. CADIA only requires a list of differentially expressed genes and a set of background pathways to produce the enrichment p-values (P_{cadia}). After the selection of differentially expressed genes, the method does not depend on a ranked list of genes nor their fold changes. Also, because of fewer limitations in preprocessing step, CADIA has broader coverage in pathway analysis and can infer the enrichment of several critical pathways that are not included in SPIA analysis, such as "Ras Signaling" and "PI3K-Akt Signaling".

The exclusion of incomplete pathways in CADIA allows avoiding the detection of inaccurate enrichments. With incomplete information, perturbation of any node with a non-minimal centrality score would produce small p-values for enrichment, and subsequently, produce false positive. Network-based enrichment analyses are prone to producing incorrect inferences when the pathway information [25]. The results produced by SPIA show instances where network-based PEM are prone to make irrelevant inferences. Although we took a filtering approach to disregard incomplete pathways (See Supplementary Material Table SI), using predicted interactions could benefit CADIA's enrichment analysis in future developments.

CADIA leverages two independent statistics of ORA (P_{ora}) and Source/Sink Centrality (P_{ssc}) to achieve increased sensitivity; an approach that was originally used in SPIA [12]. We have shown that in multiple instances, the significance of the topological evidence P_{ssc} allows compensating for the lack of strength in the over-representation evidence. Also, the lack of topological evidence allows dismissing marginal over-representation evidence that might be irrelevant to the experimental data. While it is also possible to Source-Sink Centrality in the GSA model through a methodology shown by Gu and colleagues [9], the choice of ORA allows to leverage two independent pieces of evidence simultaneously. Recent studies show that multi-evidence approaches for PEM can provide increased sensitivity and specificity [12], [44].

The Supplementary Material contains assessments of EnrichNet and GSA on the three test datasets. Those results show that GSA in some cases does not discover any significant pathways at the specified thresholds (not sensitive). In the other cases, GSA discovers numerous pathways and is oversensitive (as much as 94 out of 143, See Tables S.VIII-S.XIII in Supplementary Material). Also, EnrichNet fails to discover multiple critical pathways that are discovered by CADIA, SPIA, and ORA. For these reasons, this manuscript only displays the results of CADIA, ORA, and SPIA. Interested readers may refer to the Supplementary Material for a comparison of CADIA with GSA and EnrichNet.

4.2 synthetic data evaluation

The synthetic evaluation shows the number of false positives in CADIA is below the specified criterion. In particular, the repetitive testing of CADIA for different input sizes shows that the average of false positives is below FDR = 0.05 threshold. Figure 2 shows that the average false positive rates of the topological evidence (FDR-corrected P_{ssc}) is zero. When using ORA alone, the average false positive rate in some cases is not zero but is below the FDR = 0.05 threshold. Similarly, the combined evidence (FDR-corrected P_{cadia}) produces small averages of false positive rates. Figure 2 shows that the controlled false positive rate of CADIA is consistent across a wide range of random perturbation input size (100–5000). These results indicate the specificity of CADIA and ensure that the experimental data inferences are not results of false positive.

Additional synthetic data evaluation shows a uniform null distribution of P_{ssc} for 5000 random perturbation sets (results in Supplementary Material). The uniform distribution of P_{ssc} shows that this topological evidence is not biased towards making false-positives or false-negatives. The synthetic data evaluation does not find any correlation (correlation estimate = 0.005 and p-value = 0.2) between P_{ssc} and P_{ora} (See Supplementary Material).

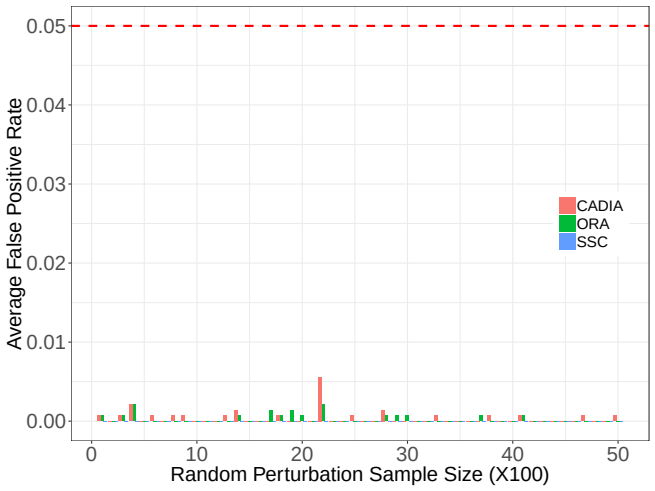

Fig. 2. The number of false positives of ORA, SSC, and CADIA for different sizes of the randomly sampled perturbed genes (10 repeats). The Y-axis is the average number of false positives from a $FDR \leq 0.05$ threshold. The red dashed-line is the FDR control threshold.

Figure 3 displays the bootstrap (5×10^5 rounds) sampling for the aggregate scores (Formula 14) of the 31 perturbed nodes from focal adhesion (from Table 1). The pattern of normal distribution, in this case, is explainable by the central limit theorem. The aggregate score estimates a multiple of the mean of log-centrality values (Formula 14). With sufficiently large perturbation set, the normal distribution can replace the empirical estimation of P_{ssc}. Also, the random aggregate scores of $C(v_i)$s are independent and identically distributed (iid), having the necessary conditions for the central limit theorem. This normal distribution can lead to another route for showing that P_{ssc} and P_{ora} are independent. If P_{ssc} follows a normal distribution, based on mean $C(v_i)$s, then it is independent of the outcomes of the hypergeometric distribution in Formula 16.

Figure 4, illustrates an example of the ability of Source-Sink Centrality in attributing importance to both upstream and downstream nodes. In the case of the ErbB pathway, the signal receptors associated genes EGFR and ERBB2 are critical sources that initialize activities (upstream), while the genes MYC, JUN, and ELK are critical endpoint receivers (downstream) [13]. Figure 4 shows the relative importance

TABLE 10
Centrality scores of different algorithms for ErbB Signalling pathway

Gene	SSC	Bet	Deg	Katz	Gene	SSC	Bet	Deg	Katz	Gene	SSC	Bet	Deg	Katz
EGFR	88	87	88	88	CDKN1A	56	1	1	1	JUN	28	1	1	1
ERBB4	87	86	87	87	CDKN1B	56	1	1	1	CRK	26	41	36	31
ERBB3	86	66	85	85	GSK3B	56	1	1	1	CRKL	26	41	36	31
ERBB2	85	1	85	85	MAP2K1	54	74	36	38	STAT5A	24	1	1	1
AKT1	82	79	77	72	MAP2K2	54	74	36	38	STAT5B	24	1	1	1
AKT2	82	79	77	72	SOS1	52	83	60	61	AREG	21	1	26	53
AKT3	82	79	77	72	SOS2	52	83	60	61	EGF	21	1	26	53
GRB2	81	88	60	71	ARAF	49	69	36	40	TGFA	21	1	26	53
GAB1	80	85	84	84	BRAF	49	69	36	40	NRG3	19	1	26	50
MAP2K4	78	67	60	56	RAF1	49	69	36	40	NRG4	19	1	26	50
MAP2K7	78	67	60	56	HRAS	46	76	60	58	SRC	18	41	26	26
NCK1	76	72	80	82	KRAS	46	76	60	58	MYC	17	1	1	1
NCK2	76	72	80	82	NRAS	46	76	60	58	ABL1	8	1	1	1
PIK3CA	68	58	60	63	BTC	43	1	36	77	ABL2	8	1	1	1
PIK3CB	68	58	60	63	EREG	43	1	36	77	CAMK2A	8	1	1	1
PIK3CD	68	58	60	63	HBEGF	43	1	36	77	CAMK2B	8	1	1	1
PIK3CG	68	58	60	63	NRG1	41	1	36	75	CAMK2D	8	1	1	1
PIK3R1	68	58	60	63	NRG2	41	1	36	75	CAMK2G	8	1	1	1
PIK3R2	68	58	60	63	MAPK10	38	51	36	31	PRKCA	8	1	1	1
PIK3R3	68	58	60	63	MAPK8	38	51	36	31	PRKCB	8	1	1	1
PIK3R5	68	58	60	63	MAPK9	38	51	36	31	PRKCG	8	1	1	1
MTOR	67	82	60	52	PAK6	31	44	36	43	EIF4EBP1	5	1	1	1
PLCG1	65	54	80	80	PAK1	31	44	36	43	RPS6KB1	5	1	1	1
PLCG2	65	54	80	80	PAK2	31	44	36	43	RPS6KB2	5	1	1	1
SHC1	61	1	26	27	PAK3	31	44	36	43	CBL	2	1	1	1
SHC2	61	1	26	27	PAK4	31	44	36	43	CBLB	2	1	1	1
SHC3	61	1	26	27	PAK5	31	44	36	43	CBLC	2	1	1	1
SHC4	61	1	26	27	PAK6	31	44	36	43	PTK2	1	1	1	1
ELK1	60	1	1	1	MAPK1	29	56	36	31					
BAD	56	1	1	1	MAPK3	29	56	36	31					

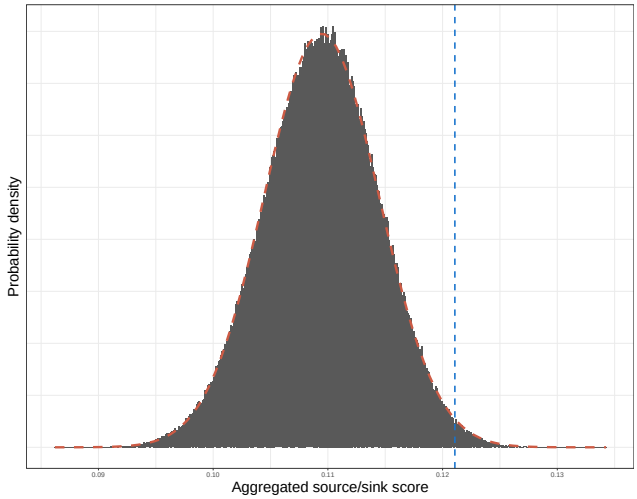

Fig. 3. Null distribution of aggregate centrality score for calculating P_{ssc}. The figure is generated based on 31 perturbation in Focal adhesion pathway from ovarian cancer data (Table 1). The X-axis denotes the aggregate centrality score from Formula 14. The red dashed line indicates the normal distribution fit based on the observed mean and standard deviation of the null aggregate scores. The blue line the is experimental observation and its right-hand side area under the curve is P_{ssc}.

scores of source/sink centrality for each gene in the ErbB pathway. Genes at the upstream the pathway, including EGFR, ERBB1, and ERBB2, are recognized by Source/Sink as high centrality (Table 10). Also, Source/Sink centrality distinguished between the downstream nodes such as MYC, JUN, and ELK1. Other standard centrality measures assign low importance to terminal nodes of pathways. For example, ELK1, BAD, PTK2, MYC, and JUN would have the same centrality score regardless of their underlying biological functions and topological position in the graph (Figure 4 and Table 10).

A general centrality measure may fail to capture the downstream importance and assign low centrality values. This observation extends to the definition of topological importance in other network-based PEM. In SPIA for example, the downstream nodes will have the lowest importance because they have zero (or low) out-degree. A possible alternative solution is to sacrifice the network directions, like in that of EnrichNet [11]. The undirected graph approach will potentially deliver incomplete results because the topological features of the graph rely on the directions of the nodes. Evident by our experimental validation, addressing the issues with common centrality models in CADIA enables to detect unique pathway enrichments while delivering consistent results with a low false positive rate.

5 Conclusion

In this study, we have shown that Causal Disturbance Analysis (CADIA) is a network-Based model that provides a unique and robust perspective for pathway enrichment analysis. CADIA takes into account the position of a gene in the pathway by devising a novel centrality model, Source-Sink. CADIA takes an unordered list of differentially expressed genes and produces a p-value that indicates enrichment of a pathway. Our evaluations show that CADIA can infer the critical pathway associations from experimental data that are not observed by standard and state-of-the-art enrichment analysis models. Source/Sink centrality allows

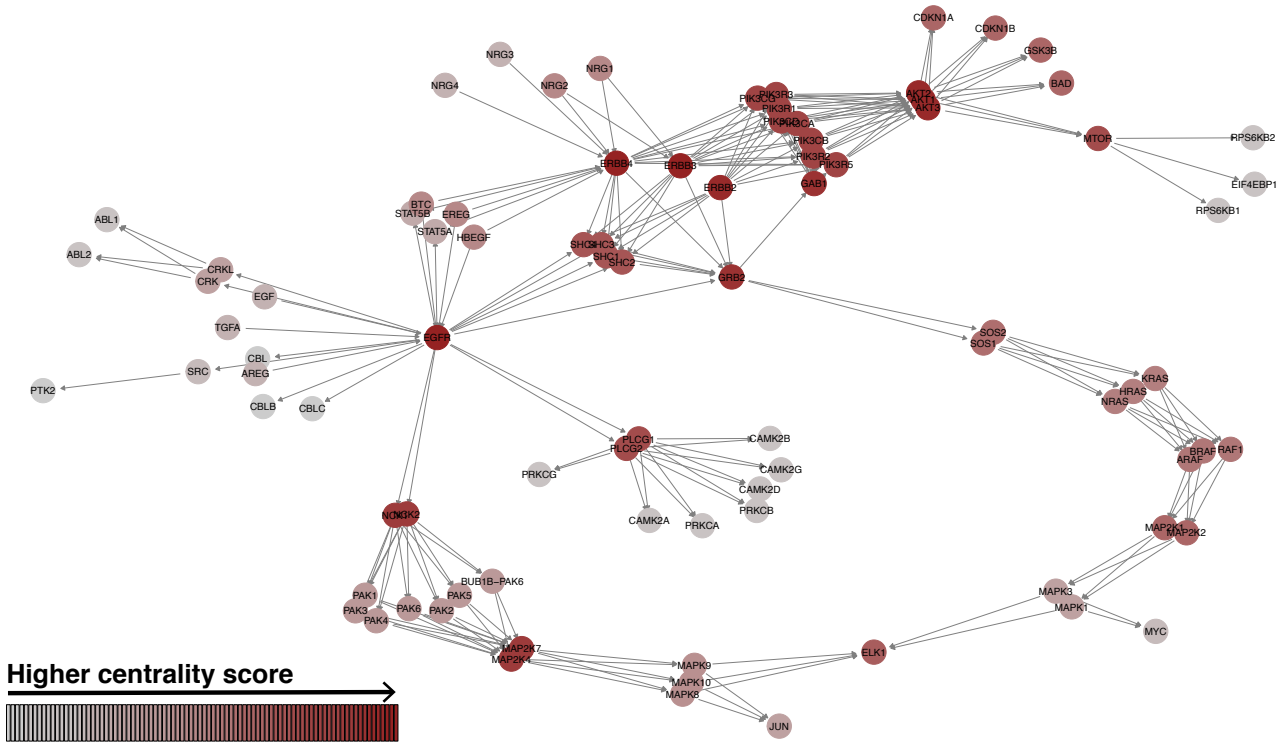

Fig. 4. Application of Source/Sink centrality to ErbB signaling pathway. The color intensity indicates the ranking assigned by source/sink centrality. This figure shows the ability of source/sink centrality to the terminal nodes of the pathways such as ELK1, JUN, and BAD. A standard centrality score for directed graphs might assign zero importance to terminal nodes (See Table 10 for more details).

attributing higher importance to the genes that are topologically important in both upstream and downstream of the pathways. Our results show that this topological-based evidence can provide unique and informative observations for inference when used in combination with frequency-based evidence.

As the pathways data collections grow, the pathway coverage of CADIA will grow, and it will be able to produce more precise and valid inferences. The presented methodology can contribute to the applications of drug target discovery and biomarker discovery as it concerns pathway analysis based on the underlying topology.

Availability of Data

All data used in this study were retrieved from publicly available databases including KEGG and NCBI GEO. The source code and the documentation of CADIA is available through this link: https://github.com/pouryany/CADIA

Acknowledgments

The authors would like to thank the Department of Computer Science, College of Computing and Informatics at UNC Charlotte for their financial and other supports.

References

[1] A. Subramanian *et al.*, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, vol. 102, no. 43, pp. 15 545–15 550, 2005.

[2] D. K. Slonim and I. Yanai, "Getting started in gene expression microarray analysis," *PLoS computational biology*, vol. 5, no. 10, p. e1000543, 2009.

[3] M. Ashburner *et al.*, "Gene ontology: tool for the unification of biology," *Nature genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[4] M. Kanehisa *et al.*, "Kegg: new perspectives on genomes, pathways, diseases and drugs," *Nucleic Acids Research*, vol. 45, no. D1, pp. D353–D361, 2017.

[5] P. Khatri, M. Sirota, and A. J. Butte, "Ten years of pathway analysis: current approaches and outstanding challenges," *PLoS computational biology*, vol. 8, no. 2, p. e1002375, 2012.

[6] D. Hanahan and R. A. Weinberg, "Hallmarks of cancer: the next generation," *cell*, vol. 144, no. 5, pp. 646–674, 2011.

[7] H. Jeong *et al.*, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[8] U. Alon, "Network motifs: theory and experimental approaches," *Nature Reviews Genetics*, vol. 8, no. 6, p. 450, 2007.

[9] Z. Gu *et al.*, "Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes," *BMC systems biology*, vol. 6, no. 1, p. 56, 2012.

[10] P. Naderi Yeganeh and M. T. Mostafavi, "Use of structural properties of underlying graphs in pathway enrichment analysis of genomic data," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. ACM, 2017, pp. 279–284.

[11] E. Glaab *et al.*, "Enrichnet: network-based gene set enrichment analysis," *Bioinformatics*, vol. 28, no. 18, pp. i451–i457, 2012.

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

[13] Y. Yarden and M. X. Sliwkowski, "Untangling the erbb signalling network," *Nature reviews Molecular cell biology*, vol. 2, no. 2, p. 127, 2001.

[14] M. Volm, W. Rittgen, and P. Drings, "Prognostic value of erbb-1, vegf, cyclin a, fos, jun and myc in patients with squamous cell lung carcinomas," *British journal of cancer*, vol. 77, no. 4, p. 663, 1998.

[15] M. Newman, *Networks: An Introduction*. New York, NY, USA: Oxford University Press, Inc., 2010.

[16] B. Schloerke *et al.*, *GGally: Extension to 'ggplot2'*, 2018, r package version 1.4.0. [Online]. Available: https://CRAN.R-project.org/package=GGally

[17] C. D. Meyer, *Matrix analysis and applied linear algebra*. Siam, 2000, vol. 71.

[18] P. Bonacich and P. Lloyd, "Eigenvector-like measures of centrality for asymmetric relations," *Social networks*, vol. 23, no. 3, pp. 191–201, 2001.

[19] A. Birnbaum, "Combining independent tests of significance," *Journal of the American Statistical Association*, vol. 49, no. 267, pp. 559–574, 1954.

[20] M. S. Anglesio *et al.*, "Mutation of erbb2 provides a novel alternative mechanism for the ubiquitous activation of ras-mapk in ovarian serous low malignant potential tumors," *Molecular cancer research*, vol. 6, no. 11, pp. 1678–1690, 2008.

[21] S. Tsukamoto *et al.*, "Clinical significance of osteoprotegerin expression in human colorectal cancer," *Clinical cancer research*, vol. 17, no. 8, pp. 2444–2450, 2011.

[22] M. E. Ritchie *et al.*, "limma powers differential expression analyses for rna-sequencing and microarray studies," *Nucleic acids research*, vol. 43, no. 7, pp. e47–e47, 2015.

[23] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300, 1995.

[24] J. D. Zhang and S. Wiemann, "Kegggraph: a graph approach to kegg pathway in r and bioconductor," *Bioinformatics*, vol. 25, no. 11, pp. 1470–1471, 2009.

[25] J. Ma, A. Shojaie, and G. Michailidis, "Network-based pathway enrichment analysis with incomplete network information," *Bioinformatics*, vol. 32, no. 20, pp. 3165–3174, 2016.

[26] W. Luo *et al.*, "Gage: generally applicable gene set enrichment for pathway analysis," *BMC bioinformatics*, vol. 10, no. 1, p. 161, 2009.

[27] A. Sergushichev, "An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation," *BioRxiv*, p. 060012, 2016.

[28] R. C. Gentleman *et al.*, "Bioconductor: open software development for computational biology and bioinformatics," *Genome biology*, vol. 5, no. 10, p. R80, 2004.

[29] P. N. Yeganeh *et al.*, "Dysregulation of akt3 along with a small panel of mrnas stratifies high-grade serous ovarian cancer from both normal epithelia and benign tumor tissues," *Genes & cancer*, vol. 8, no. 11-12, pp. 784–798, 2017.

[30] J. Downward, "Targeting ras signalling pathways in cancer therapy," *Nature Reviews Cancer*, vol. 3, no. 1, p. 11, 2003.

[31] R. C. Bast Jr, B. Hennessy, and G. B. Mills, "The biology of ovarian cancer: new opportunities for translation," *Nature Reviews Cancer*, vol. 9, no. 6, p. 415, 2009.

[32] J. Luo, B. D. Manning, and L. C. Cantley, "Targeting the pi3k-akt pathway in human cancer," *Cancer cell*, vol. 4, no. 4, pp. 257–262, 2003.

[33] K. D. Courtney, R. B. Corcoran, and J. A. Engelman, "The pi3k pathway as drug target in human cancer," *Journal of clinical oncology*, vol. 28, no. 6, p. 1075, 2010.

[34] A. De Luca *et al.*, "The ras/raf/mek/erk and the pi3k/akt signalling pathways: role in cancer pathogenesis and implications for therapeutic approaches," *Expert opinion on therapeutic targets*, vol. 16, no. sup2, pp. S17–S27, 2012.

[35] S. Mabuchi *et al.*, "The pi3k/akt/mtor pathway as a therapeutic target in ovarian cancer," *Gynecologic oncology*, vol. 137, no. 1, pp. 173–179, 2015.

[36] A. K. Sood *et al.*, "Biological significance of focal adhesion kinase in ovarian cancer: role in migration and invasion," *The American journal of pathology*, vol. 165, no. 4, pp. 1087–1095, 2004.

[37] A. A. Adjei, "Blocking oncogenic ras signaling for cancer therapy," *Journal of the National Cancer Institute*, vol. 93, no. 14, pp. 1062–1074, 2001.

[38] B. Vogelstein and K. W. Kinzler, "Cancer genes and the pathways they control," *Nature medicine*, vol. 10, no. 8, p. 789, 2004.

[39] D. Pan, "The hippo signaling pathway in development and cancer," *Developmental cell*, vol. 19, no. 4, pp. 491–505, 2010.

[40] E. Carlsson *et al.*, "Potential role of a navigator gene nav3 in colorectal cancer," *British journal of cancer*, vol. 106, no. 3, p. 517, 2012.

[41] D. W. Kang, K.-Y. Choi *et al.*, "Phospholipase d meets wnt signaling: a new target for cancer therapy," *Cancer research*, vol. 71, no. 2, pp. 293–297, 2011.

[42] D. W. Kang *et al.*, "Phospholipase d1 inhibition linked to upregulation of icat blocks colorectal cancer growth hyperactivated by wnt/β-catenin and pi3k/akt signaling," *Clinical Cancer Research*, vol. 23, no. 23, pp. 7340–7350, 2017.

[43] J. Mao *et al.*, "Roles of wnt/β-catenin signaling in the gastric cancer stem cells proliferation and salinomycin treatment," *Cell death & disease*, vol. 5, no. 1, p. e1039, 2015.

[44] M. Alhamdoosh *et al.*, "Combining multiple tools outperforms individual methods in gene set enrichment analyses," *Bioinformatics*, vol. 33, no. 3, pp. 414–424, 2017.

Pourya Naderi Yeganeh Obtained his Bachelors degree in Computer Science from Sharif Univerity of Technology in 2013. He then joined UNC Charlotte where he currently is a Ph.D. Candidate in Computing and Informatics. His interests include discovery of diagnostic biomarkers for ovarian cancer. In addition, he researches network-based models for biological biological inference.

M. Taghi Mostafavi, Ph.D., is a faculty member and the director of Biomedical Instrumentation Laboratory at the Department of Computer Science, College of Computing and Informatics at UNC Charlotte. His current research interest is Biomedical Information Processing and Systems.