EpiTracer – an algorithm for identifying epicenters in condition-specific biological networks

Narmada Sambaturu
IISc Mathematics Initiative
Indian Institute of Science
Bangalore–560012, India
narmada.sambaturu@biochem.iisc.ernet.in

Madhulika Mishra
Department of Biochemistry
Indian Institute of Science
Bangalore–560012, India
madhulika@biochem.iisc.ernet.in

Nagasuma Chandra
IISc Mathematics Initiative
Department of Biochemistry
Indian Institute of Science
Bangalore–560012, India
nchandra@biochem.iisc.ernet.in

***Abstract*—Diseases in biological systems may result from small perturbations in a complex network of protein-protein interactions (PPIs). The perturbations typically affect a small set of proteins, which then go on to disturb a larger part of the network. Biological systems attempt to counteract these perturbations by launching a stress-response, resulting in a complex pattern of variations in the cell. We present an algorithm, EpiTracer which identifies the key proteins, termed epicenters, from which a large number of the changes in PPI networks ripple out. We propose a new centrality measure, ripple centrality, that measures how effectively a change at a particular protein can ripple across the network, by identifying condition specific highest activity paths obtained by mapping gene expression profiles to the PPI network.**

We perform a case study on a dataset (E-GEOD-61973) where the gene PARK2 was intentionally overexpressed in human glioma (U251) cell line and analyze the top 10 ranked epicenters. We find that EpiTracer identifies PARK2 as the most important epicenter in the perturbed condition. Analysis of the other top-ranked epicenters showed that all of them were involved in either supporting the activity of PARK2 or counteracting it, indicating that the cell had activated a stress-response. We also find that 5 of the identified epicenters did not have significant differential expression, proving that our method is capable of finding information that simple differential expression analysis cannot.

The source code is available at Github (https://github.com/narmada26/EpiTracer).

***Keywords*-network mining; influential nodes; ripple centrality; perturbation analysis; condition-specific network**

I. Introduction

Biological systems are the result of a complex set of interactions, captured effectively by networks. Diseases typically occur due to a perturbation at one or more locations, affecting the nature, abundance and interactions of certain proteins in the system [1] [2]. Although the perturbation usually affects only a small set of proteins, the effects of the perturbation ripple across a much larger portion of the network. The reasons for such ripple effects are many fold. The proteins in the immediate vicinity of the perturbation experience a change as a result of direct interactions with the affected proteins. Cascade effects also exist, which affect proteins quite distant from the source of the perturbation. Aside from this, the cell may also attempt to restore its equilibrium by launching a stress-response, which itself can be through multiple mechanisms [3]. Thus a complex picture of variations is presented by a cell in any given disease condition. Given a cell in the throes of this tug-of-war, it would be very interesting and useful to identify the proteins which are the key players in spreading the perturbation and/or reacting to it. These key players are referred to as the epicenters of that condition.

Most studies in biology, work with very simplistic models, or keep the field of view restricted to a small number of parameters. However the underlying biological system is actually a large and complex one, and the simplistic models lose out valuable information by abstracting out these details. Further, a lot of data is generated by microarray experiments, which are available publicly on various databases. For instance, over 50,000 datasets are available on Omnibus [4]. The full potential of this vast amount of data is far from being realized because of the lack of good analysis and interpretation pipelines. With algorithms such as EpiTracer, we hope to bridge this gap, making analysis possible for large scale and detailed models, thus more closely reflecting the intricate workings of living creatures. In this paper, we work with a network consisting of nearly half the complement of human genes.

In the current state-of-the-art, there are no well-established methods to identify the epicenter. Most methods require the existence of a causal network, with each edge being from a cause to an effect [5] [6]. However the current knowledge on causal relationships even for a pair of nodes is minimal, and is restricted to only a small set of processes. As a result, constructing large causal networks is not directly feasible. Network motif based approaches have been used to identify important nodes in directed biological networks [7]. However this method relies only on network information, and has no way of incorporating gene expression data. Other methods attempt to identify the node which, when intentionally perturbed, would spread the perturbation the fastest [8]. This is different from identifying the epicenter of a naturally occurring perturbation, which is a more nuanced

978-1-4673-6799-8/15/$31.00 ©2015 IEEE

and complicated as well as biologically important scenario.

In this paper, we have developed an algorithm called EpiTracer, which is capable of identifying the epicenter of changes in the network, and highlighting the paths through which the change propagates. An epicenter must be highly active in order to exert its influence, and must have good connectivity in order for the influence to spread. We define a new centrality measure called ripple centrality, which gives a combined measure of a node′s activity as well as its connectivity, thus giving us a handle on how well influence ripples out of that node. This can be used to rank nodes, with the top-ranked nodes qualifying as epicenters. The algorithm works by first narrowing the search space by identifying the sub-network which is most active, retaining only paths with high activity in the perturbed condition. We then calculate the ripple centrality of each node in the condition-specific highest activity network, identifying the top 10 as the most important epicenters. We demonstrate the efficacy of the algorithm by carrying out a case study on a dataset where PARK2 was intentionally overexpressed in human glioma (U251) cell line. The algorithm was able to identify the perturbed gene as the most important epicenter even though it was given no knowledge of the perturbation.

II. MATERIALS AND METHODS

The inputs used by the algorithm are (a) a high-density protein-protein interaction network, and (b) condition-specific gene expression profiles. The inputs are explained below with respect to the case study.

A. Network Reconstruction

A weighted directed human protein-protein interaction (PPI) network was reconstructed, with 10,306 nodes and 74,404 edges. The base network was taken from Khurana et al., 2013, which contains known and predicted protein-protein interactions, genetic interactions and regulatory networks with directions [9]. To this, metabolic interactions were added from KEGG [10].

B. Gene Expression Profiles

Gene-expression data (E–GEOD–61973) for PARK2 overexpression was taken from Array-express. The authors have deposited the transcriptome profile as a result of PARK2 overexpression in U251 cell line and control (GFP) U251 cell line. Normalization of microarray data was performed using GeneSpringX 12.6.1, with Robust Multichip Averaging (RMA) [11]. For differential gene analysis, a 1.5-fold cut-off was applied (P-value $\leq$ 0.05 by T-test with Benjamini-Hochberg false discovery rate correction).

C. Biological Analysis

Gene set enrichment was performed against KEGG [10] database using WebGestalt [12]. The statistical test used for analysis was a hypergeometric test with P-value of 0.05 with FDR correction. Cytoscape was used for network visualization, and the Cytoscape plugin ClueGO [13] was used for GO module enrichment.

D. Condition-specific Networks

Condition specific networks were constructed by setting node weight to be the normalized signal intensity for each gene in that condition. $w_i^x = SI^x$ where w_i^x is the weight of node i in condition x, and SI^x is the normalized signal intensity in condition x. The edge weight or edge cost was taken as a function of the abundance of the participating proteins, as

$$c_i^x = \frac{1}{\sqrt{w_u^x * w_v^x}}$$

where c_i^x is the cost of edge i in condition x, and w_u^x, w_v^x are the weights of the nodes comprising the edge. Taking the inverse makes sure that a highly active interaction has very low edge cost.

For a path with n edges, the cost of the path is given by the sum of costs of the edges involved in the path.

$$pathcost = \sum_{i=1}^{n} c_i^x$$

where c_i^x is the edge cost for each edge in the path, and n is the length of the path. A shortest path algorithm will preferentially choose edges with the least cost for a given source and destination, which in our formulation translates to identifying the highest activity path.

E. EpiTracer Algorithm – Rationale

The epicenter by definition, should be highly active and participate in high activity paths only in the perturbed condition. We extract condition-specific highest activity paths (CSHAP) by identifying highest activity paths (HAP) in each condition, and discarding common paths. These CSHAPs induce a network, referred to as the condition-specific highest activity network (CSHAN). An epicenter, by definition, should also be able to reach many nodes in the network and the paths to such nodes from the epicenter should be highly active. To capture this, we introduce a new parameter called *ripple centrality*, as explained below.

1) Closeness centrality: Closeness centrality of a node u is defined as the reciprocal of the sum of shortest path costs from u to every reachable node v

$$C(u) = \frac{1}{\sum_v \sigma(u,v)}$$

where $\sigma(u,v)$ is the cost of the shortest path from u to v. Because of the way edge costs are formulated, a node u with highly active paths to a set of nodes v will have high closeness centrality (node Acl in Figure 1A).

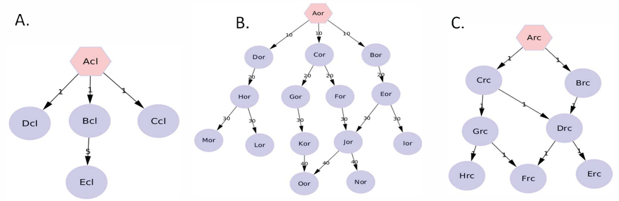


Figure 1. (A) Node Acl is the source of highly active paths, and has high closeness centrality. However it can only reach 4 nodes, and is not a good epicenter. (B) Node Aor can reach 14 nodes, but paths originating at Aor have low activity. Thus it is not a good epicenter. (C) Node Arc is the source of highly active paths and can reach a large number of nodes (7), making it the best candidate for an epicenter. The hexagon represents candidate epicenters

2) Outward Reachability: Given a node u, the number of nodes reachable from u is termed its outward reachability.

$$R_{out}(u) = |nodes\, reachable\, from\, u|$$

where $R_{out}(u)$ denotes outward reachability of u.

3) Ripple Centrality: Nodes such as node Acl (Figure 1A) can have extremely high-activity connections to very few nodes. Such a node would have high closeness centrality, but would make a poor candidate for an epicenter as any perturbation originating at this point would not propagate far. On the other hand, node Aor (Figure 1B) is able to reach a large number of nodes, but the paths originating at node Aor are of relatively lower activity. This node would have high outward reachability, but would make a poor candidate for an epicenter. Thus neither closeness centrality nor outward reachability are sufficient on their own. On the other hand, node Arc (Figure 1C) has highly active paths to a large number of nodes, and is the best candidate for an epicenter.

We formulate a new measure, *ripple centrality*, which serves as a logical AND between closeness centrality and outward reachability.

$$Ripple\, centrality(u) = C(u) * R_{out}(u)$$

Once the nodes have been ranked based on ripple centrality, we split the ranked nodes into two lists – (a) nodes which occur only in the perturbed CSHAN, and (b) nodes common to both CSHANs. The common nodes are key players both before and after the perturbation, and work as global epicenters. The nodes occurring only in the perturbed CSHAN are epicenters specific to the perturbation.

F. EpiTracer Algorithm

The EpiTracer algorithm consists of three modules (1) *highest_activity_paths* extracts the paths with cost inside a user-defined percentile threshold, (2) *condition_spec_han* uses *highest_activity_paths* to identify the highest activity network specific to each condition, and (3) the main module, *get_epicenters*, uses the above two modules to identify the top 10 epicenters in the perturbed condition, as well as the top 10 epicenters common to both conditions. The pseudocode for each module is provided in Algorithms 1, 2 and 3.

Algorithm 1: Function *highest_activity_paths*
input: network, percentile
output: highest activity paths

```
Calculate all-pairs-shortest-paths and path costs;
Discard paths with length 1;
sorted_paths = sort(paths, asc, path_cost);
return top percentile of sorted_paths;
```

Algorithm 2: Function *condition_spec_han*
input: network A, network B, percentile
output: condition specific han

```
A_hap = highest_activity_paths(A, percentile);
B_hap = highest_activity_paths(B, percentile);
common_paths = A_hap $\cap$ B_hap;
A_specific_hap = A_hap $-$ common_paths;
B_specific_hap = B_hap $-$ common_paths;
return (A_specific_hap.edges), (B_specific_hap.edges)
```

Algorithm 3: Function *get_epicenters*
input: network A, network B, percentile
output: top 10 epicenters (B only, common)

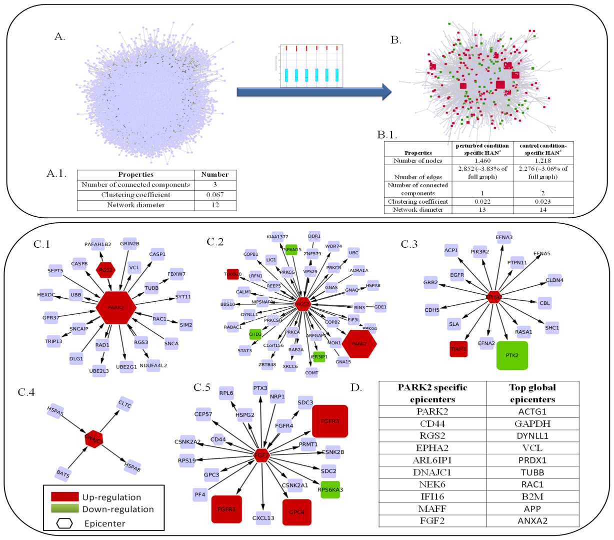
```
A_shan, B_shan=condition_spec_han(A, B, percentile);
common = A_shan.nodes $\cap$ B_shan.nodes;
B_only_nodes = B_shan.nodes $-$ common;
for all node $\in$ B_shan.nodes do
    C(node) = closeness centrality of node;
    $R_{out}$(node) = outward reachability of node;
    Ripple centrality(node) = C(node) $*$ $R_{out}$(node);
end for
ranked = sort(B_shan.nodes, desc, Ripple centrality);
ranked_B_only = ranked $\cap$ B_only_nodes;
ranked_common = ranked $\cap$ common;
return top 10 in (ranked_B_only, ranked_common);
```

III. Results

The algorithm was implemented in Python 2.7, using Networkx1.7. Dijkstra′s algorithm [14] was used for shortest path computation. The code was run on a xeon server having 16 cores, and was able to complete analysis of a network with 10,306 nodes and 74,404 edges in less than 30 minutes.

A. System Description

A summary of network properties is shown in Figure 2A. The gene expression profile used is for the overexpression of PARK2. PARK2 (PARKIN) is an E3 ubiquitin ligase whose dysfunction has been associated with the progression of Parkinsonism and human malignancies. However its function in cancer remains unclear. In the dataset, microarrays were



A.1.

Properties	Number
Number of connected components	3
Clustering coefficient	0.067
Network diameter	12

B.1.

Properties	perturbed condition specific HAN*	control condition-specific HAN*
Number of nodes	1,460	1,218
Number of edges	2,852 (~3.83% of full graph)	2,276 (~3.06% of full graph)
Number of connected components	1	2
Clustering coefficient	0.022	0.023
Network diameter	13	14

D.

PARK2 specific epicenters	Top global epicenters
PARK2	ACTG1
CD44	GAPDH
RGS2	DYNLL1
EPHA2	VCL
ARL6IP1	PRDX1
DNAJC1	TUBB
NEK6	RAC1
IFI16	B2M
MAFF	APP
FGF2	ANXA2

Figure 2. (A) Human PPI network comprising of 10,306 nodes and 74,404 edges. Color of the node is based on differential expression of the gene, red color indicates upregulation and green indicates downregulation. Hexagon shape represents epicenters. (A.1) Table of network properties for human PPI network (B) PARK2 overexpressed condition specific highest activity network (CSHAN); network properties in table B.1. (C) Modules of differentially expressed epicenters, with immediate neighbours being in the same module. (D) List of top most epicenters for global as well as PARK2 specific HAN.

performed to understand the global reprogramming of gene expression upon PARK2 overexpression in human glioma (U251) cell line. After normalization and filtration, 605 genes were found to be down-regulated and 1,089 genes were upregulated in response to PARK2 overexpression (fold change $\geq$ 1.5). In general, genes associated with cell cycle, ubiquitin mediated proteolysis, ErbB signaling pathway, MAPK, JAK-STAT signaling, WNT signaling, Hedgehog signaling pathway and pathways related to lipid metabolism were differentially expressed.

B. Highest Activity Paths (HAPs)

Shortest paths and path costs between all pairs of nodes were identified. It was observed that the number of paths in the top 0.2 percentile was twice the number of paths in the top 0.1 percentile. Thus the conservative threshold of 0.1 percentile was chosen. Since the node and edge weights are different in the two conditions, the same cut-off results in a different set of HAPs in each condition. Interestingly, the edges involved in the HAPs themselves form a well-connected network.

C. Condition-Specific Highest Activity Network (CSHAN)

Among the HAPs, some paths were found to be common to both conditions. Such paths are always extremely active, and give us no information relevant to the perturbation. Such paths were removed from both networks, giving us the condition-specific highest activity paths (CSHAPs), and the networks induced by them (CSHANs). In the data under study, 48,949 HAPs were common to both conditions, and were removed. This left us with 9,621 HAPs specific to the control setting, and 18,779 paths specific to the perturbed setting. The properties of these networks are given in Figure 2B.

The perturbed CSHAN has 1,756 genes (Figure 2B), of which 75 were down-regulated and 130 up-regulated. These belonged to the functional categories of cell cycle, MAPK, ErbB, p53 and mTOR signaling pathway, ubiquitin mediated proteolysis, regulation of actin cytoskeleton and oocyte meiosis.

D. Tracing the Epicenter

After extracting the CSHANs, the nodes were ranked in descending order of their ripple centrality. The ranked list was then separated into two nodes occurring only in the perturbed CSHAN, and nodes occurring in both CSHANs. Although common paths are removed, common nodes can still remain. The nodes common to both CSHANs are referred to as global epicenters, and the top 10 are illustrated in Figure 2D. PARK2 was identified as the highest ranked epicenter among the nodes which are highly active only in the perturbed condition. The perturbation in the dataset under study was the overexpression of the PARK2 gene. Although this knowledge was not used to guide the algorithm in any way, the algorithm was able to identify PARK2 as the highest ranked epicenter of the perturbation.

1) Biological interpretation: Top ranked epicenters common to both conditions were found to belong to highly conserved and ubiquitously expressed proteins such as TUBB, GAPDH, VCL, ACTG1, DYNLL1 and ANXA2. In glioma cells, RAC1 promotes cell migration and invasion. APP gene is also associated with neurite growth, neuronal adhesion and axonogenesis [15]. PRDX1 gene is involved in redox regulation of the cell. B2M gene is associated with MHC class I antigen presentation.

Further, biological function of top ranked epicenters active only in the PARK2 overexpression condition were revisited to understand their significance in this scenario (Figure 2C). Out of the ten genes being examined, five genes were found to be differentially expressed, namely PARK2, RGS2, EPHA2, DNAJC1 and FGF2. PARK2 was highlighted as the most important epicenter in the PARK2 overexpressed condition. PARK2 is an E3 ubiquitin ligase which negatively regulates cell cycle by degrading Cyclin E and D. RGS2, the second most important epicenter in the PARK2 overexpression condition, is involved in G0 to G1 transition [15]. Inhibition of EPHA2 gene leads to stalling of cells in G0/G1 phase [16]. FGF2 blocks cell proliferation and causes a G2/M arrest [17]. When considered together, our analysis revealed that most of the top ranked genes were associated with cell cycle regulation.

2) Immediate influence zone of top-ranked epicenter: In order to understand the cellular response to the top-ranked epicenter specific to the perturbed condition (PARK2 in this case), the influence zone around it was analysed. For this, the subgraph induced by PARK2 and the nodes up-to two hops up/down-stream of PARK2 were considered (Figure 3A), and GO enrichment was performed. Since PARK2 is a cell cycle regulator, enrichment was performed specifically for cell cycle regulation. Interestingly, it was found that the PARK2 influence zone was highly enriched for cell cycle regulation (Figure 3B), including G2/M transition and G1/S transition of mitotic cell cycle, mitotic cell cycle, positive and negative regulation of cell cycle.

The analysis was further focused onto the nodes downstream of PARK2 in order to gain better insight into the influence exerted by PARK2 itself. (Figure 3C). Mapping of nodes was done on the basis of differential expression, edge weight and epicenter ranking. It was found that many downstream genes such as MDM2, CHK1, SQSTM1 and DUSP1 were involved in cell cycle regulation.

Since overexpression of PARK2 inhibits the progression of cell cycle, the expected response from the cell would be to modify other regulatory mechanisms of cell cycle progression to counteract this arrest. Examination of the nodes downstream of the top-ranked epicenter (PARK2) showed that this was indeed the case (Figure 3D). Major remodeling can be inferred from the G0/G1 and G1/S transition. SQSTM1 (P63) is involved in exiting of the cell from the M phase in the cell cycle. CD44, EPAH2, RGS2 and ARL6IP1 are positive regulators for G0/G1 transition. MDM2 is an activator of G1/S transition as it inhibits P53 and Rb proteins. However, CHEK1 and DUSP1 are repressors of G1/S phase transition. CHEK1 acts as a Cyclin E repressor by inhibiting Cdc at the DNA-repair check-point. DUSP1 is a repressor of the MAPK pathway [18]. FGF2 and NEK6 are repressors of G2/M phase transition [19].

IV. Discussion

The epicenters identified by EpiTracer are the nodes from which highly active paths originate and connect to a large part of the network. These epicenters constitute the nodes from which most of the influence ripples out in the specific condition. Epicenters are not necessarily the source of the perturbation. However, we can expect the source of the perturbation to be topologically close to the top-ranked epicenters. In the case study, the top-ranked epicenter coincided with the source of the perturbation.

In our study, it was observed that the largest strongly connected component (LSCC), which is the largest subgraph such that there exists a directed path from node u to node v and also from node v to node u for every pair of nodes u, v, played an important role in the spread of the perturbation. The epicenter was found to be a part of the LSCC in the CSHAN.

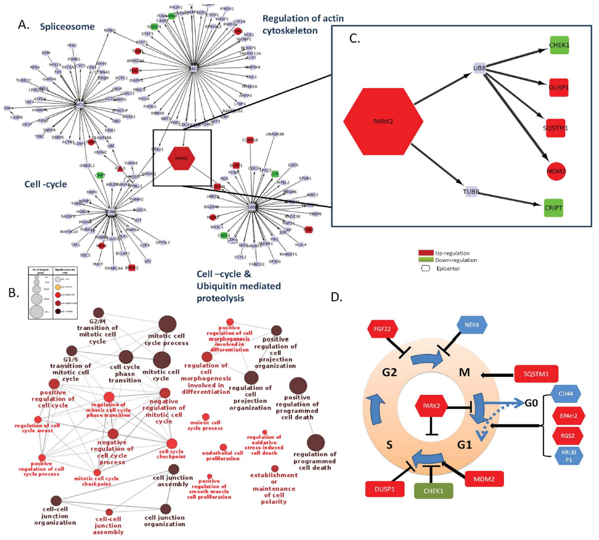

A.
Spliceosome
Regulation of actin cytoskeleton
Cell -cycle
Cell –cycle & Ubiquitin mediated proteolysis
C.
PARK2
UBB
TUBB
CHEK1
DUSP1
SQSTM1
MDM2
CRIPT
Up-regulation
Down-regulation
Epicenter
B.
G2/M transition of mitotic cell cycle
mitotic cell cycle process
positive regulation of cell morphogenesis involved in differentiation
positive regulation of cell projection organization
G1/S transition of mitotic cell cycle
cell cycle phase transition
mitotic cell cycle
regulation of cell morphogenesis involved in differentiation
regulation of cell projection organization
positive regulation of programmed cell death
positive regulation of cell cycle
regulation of mitotic cell cycle phase transition
negative regulation of mitotic cell cycle
meiotic cell cycle process
regulation of cell cycle arrest
negative regulation of cell cycle process
cell cycle checkpoint
endothelial cell proliferation
regulation of oxidative stress-induced cell death
regulation of programmed cell death
positive regulation of cell cycle process
mitotic cell cycle checkpoint
cell junction assembly
positive regulation of smooth muscle cell proliferation
establishment or maintenance of cell polarity
cell-cell junction organization
cell-cell junction assembly
cell junction organization
D.
FGF22
NEK6
G2
M
SQSTM1
PARK2
G0
CD44
EPAH2
RGS2
ARL6IP1
S
G1
DUSP1
CHEK1
MDM2

Figure 3. Detailed biological interpretation of PARK2 influence zone. (A) PARK2 influence zone consists of 118 nodes and 119 interactions. Color of node varies based on differential expression of the gene - red color indicates upregulation and green color indicates downregulation. Epicenters are represented by hexagons. (B) GO enrichment analysis of genes in the PARK2 influence zone showed that most of the influenced genes are involved in cell cycle regulation. (C) Focusing on the network two hops downstream of PARK2 showed that most of the proteins were involved in cell cycle regulation. (D) Mechanistic insight of cell cycle deregulation upon PARK2 overexpression.

It was also observed that the LSCC in the CSHAN was a subgraph of the LSCC in the parent graph. If the LSCC comprises a big enough percentage of the graph, we believe it might be possible to narrow the search space of epicenters even further by restricting the search to the LSCC alone, thus speeding up the algorithm.

V. Conclusion

We propose a new method, EpiTracer, to trace the epicenter of perturbations in a condition-specific biological network. Using only a network with static topology and microarray experiments in the relevant conditions, the algorithm can identify the condition-specific highest activity network (CSHAN), and associate each node in this CSHAN with a new measure called ripple centrality. The ripple centrality value of a node gives an indication of how influence of that node can ripple outwards into the rest of the network. The ripple centrality was used to identify the top candidates for epicenters specific to the perturbed condition, as well as epicenters common to both conditions. A case study was carried out on a dataset where the gene PARK2, an E3 ubiquitin ligase which regulates G1 to S phase transition by degrading Cyclin E and D, was intentionally overexpressed in human glioma (U251) cell line. EpiTracer was able to identify PARK2 as the most important epicenter specific to the perturbed condition. Biological analysis of the other top-ranked epicenters showed that all of them had functions relevant to cell cycle progression, and highlighted a scenario where the most important epicenters were involved in either spreading the influence of PARK2 or working to counteract its effect. Of the top 10 epicenters in the perturbed condition, 5 did not show significant differential expression, and yet were found to be biologically meaningful epicenters. This shows that our analysis can highlight more than what a simple differential expression analysis can.

The algorithm has been tested only on one dataset because of the complexity of biological interpretation. However, we expect that it will have general applicability, and will facilitate the understanding of cell behavior in response to perturbation. The algorithm highlights the key players or epicenters, which spread the perturbation and/or respond to the perturbation. The paths along which the influence of these epicenters ripples out is highlighted by the condition-specific highest activity network. These results can be used to gain a better understanding of the disease phenotype, and how the organism responds to it.

Acknowledgment

We thank Dr. Arun Konagurthu, Monash University, Australia for the useful discussions. We thank the Department of Biotechnology (DBT, India) for financial support.

References

[1] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabasi, “The human disease network,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 21, pp. 8685–8690, 05 2007.

[2] J. Padiadpu, R. Vashisht, and N. Chandra, “Protein–protein interaction networks suggest different targets have different propensities for triggering drug resistance,” *Systems and synthetic biology*, vol. 4, no. 4, pp. 311–322, 12 2010.

[3] M. A. Rowland, W. Fontana, and E. J. Deeds, “Crosstalk and competition in signaling networks,” *Biophysical journal*, vol. 103, no. 11, pp. 2389–2398, 12 2012.

[4] R. Edgar, M. Domrachev, and A. E. Lash, “Gene expression omnibus: Ncbi gene expression and hybridization array data repository,” *Nucleic acids research*, vol. 30, no. 1, pp. 207–210, 01 2002.

[5] F. Martin, A. Sewer, M. Talikka, Y. Xiang, J. Hoeng, and M. C. Peitsch, “Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models,” *BMC bioinformatics*, vol. 15, no. 1, p. 238, 07 2014.

[6] A. Krämer, J. Green, J. Pollard, and S. Tugendreich, “Causal analysis approaches in ingenuity pathway analysis (ipa),” *Bioinformatics*, vol. 30, no. 4, pp. 523–530, 02 2013.

[7] P. Wang, J. Lü, and X. Yu, “Identification of important nodes in directed biological networks: A network motif approach,” *PLoS ONE*, vol. 9, no. 8, p. e106132, 08 2014.

[8] M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, “Identification of influential spreaders in complex networks,” *Nature Physics*, vol. 6, no. 11, pp. 888–893, 08 2010.

[9] E. Khurana, Y. Fu, J. Chen, and M. Gerstein, “Interpretation of genomic variants using a unified biological network approach,” *PLoS Comput Biol*, vol. 9, no. 3, p. e1002886, 03 2013.

[10] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “Data, information, knowledge and principle: back to metabolism in kegg,” *Nucleic acids research*, vol. 42, no. D1, pp. D199–D205, 11 2014.

[11] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, T. P. Speed *et al.*, “Exploration, normalization, and summaries of high density oligonucleotide array probe level data,” *Biostatistics*, vol. 4, no. 2, pp. 249–264, 04 2003.

[12] J. Wang, D. Duncan, Z. Shi, and B. Zhang, “Web-based gene set analysis toolkit (webgestalt): Update 2013,” *Nucleic acids research*, vol. 41, no. W1, pp. W77–W83, 05 2013.

[13] G. Bindea, B. Mlecnik, H. Hackl, P. Charoentong, M. Tosolini, A. Kirilovsky, W.-H. Fridman, F. Pagès, Z. Trajanoski, and J. Galon, “Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks,” *Bioinformatics*, vol. 25, no. 8, pp. 1091–1093, 04 2009.

[14] E. W. Dijkstra, “A note on two problems in connexion with graphs,” *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[15] M. Safran, I. Dalah, J. Alexander, N. Rosen, T. I. Stein, M. Shmoish, N. Nativ, I. Bahir, T. Doniger, H. Krug *et al.*, “Genecards version 3: the human gene integrator,” *Database*, vol. 2010, p. baq020, 08 2010.

[16] W. Yuan, Z. Chen, S. Wu, J. Guo, J. Ge, P. Yang, and J. Huang, “Silencing of epha2 inhibits invasion of human gastric cancer sgc-7901 cells in vitro and in vivo.” *Neoplasma*, vol. 59, no. 1, pp. 105–113, 2011.

[17] J. Salotti, M. H. Dias, M. M. Koga, and H. A. Armelin, “Fibroblast growth factor 2 causes g2/m cell cycle arrest in ras-driven tumor cells through a src-dependent pathway,” *PloS one*, vol. 8, no. 8, p. e72582, 08 2013.

[18] S. Shah, E. M. King, A. Chandrasekhar, and R. Newton, “Roles for the mitogen-activated protein kinase (mapk) phosphatase, dusp1, in feedback control of inflammatory gene expression and repression by dexamethasone,” *Journal of Biological Chemistry*, vol. 289, no. 19, pp. 13 667–13 679, 05 2014.

[19] M.-Y. Lee, H.-J. Kim, M.-A. Kim, H. J. Jee, A. J. Kim, Y.-S. Bae, J.-I. Park, J. H. Chung, and J. Yun, “Nek6 is involved in g2/m phase cell cycle arrest through dna damage-induced phosphorylation,” *Cell Cycle*, vol. 7, no. 17, pp. 2705–2709, 09 2008.