PANI: A Novel Algorithm for Fast Discovery of Putative Target Nodes in Signaling Networks

Huey-Eng Chua[§] Sourav S Bhowmick[§] Lisa Tucker-Kellogg[‡]
Qing Zhao[§] C F Dewey, Jr[†] Hanry Yu[¶]

[§]School of Computer Engineering, Nanyang Technological University, Singapore
[‡]Mechanobiology Institute, National University of Singapore, Singapore
[¶]Department of Physiology, National University of Singapore, Singapore
[†]Division of Biological Engineering, Massachusetts Institute of Technology, USA
chua0530|assourav|zhaoqing@ntu.edu.sg, LisaTK|nmiyuh@nus.edu.sg, cfdewey@mit.edu

ABSTRACT

In biological network analysis, the goal of the *target identification problem* is to predict molecule to inhibit (or activate) to achieve optimum efficacy and safety for a disease treatment. A related problem is the *target prioritization problem* which predicts a subset of molecules in a given disease-related network which contains successful drug targets with highest probability. Sensitivity analysis prioritizes targets in a dynamic network model using principled criteria, but fails to penalize off-target effects, and does not scale for large networks. We describe PANI (**P**utative T**A**rget **N**odes Pr**I**oritization), a novel method that prunes and ranks the possible target nodes by exploiting concentration-time profiles and network structure (topological) information. PANI and two sensitivity analysis methods were applied to three signaling networks, MAPK-PI3K; myosin light chain (MLC) phosphorylation and sea urchin endomesoderm gene regulatory network which are implicated for example in ovarian cancer; atrial fibrillation and deformed embryos. Predicted targets were compared against the molecules known to be targeted by drugs in clinical use for the respective diseases. PANI is orders of magnitude faster and prioritizes the majority of known targets higher than both sensitivity methods. This highlights a potential disagreement between absolute mathematical sensitivity and our intuition of influence. We conclude that empirical, structural methods like PANI, which demand almost no run time, offer benefits not available from quantitative simulation and sensitivity analysis.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: [biology and genetics]

Keywords

drug target prioritization, profile shape similarity, target downstream effect, putative target score, algorithm

ACM-BCB '11, August 1-3, Chicago, IL, USA

1. INTRODUCTION

The emergence of new technologies facilitating integration of various drug development approaches led to an increase in customized [21] designs for high-throughput experiments, creating demand for computational automation to assist in the selection of molecule sets for multiplex assays. A *putative target* node in a signaling network is a protein that when perturbed is able to achieve desirable efficacy and safety in terms of regulation of a particular *output node*. Informally, an *output node* is a protein that is either involved in biological processes (e.g., proliferation) which may be deregulated, resulting in manifestation of a disease (e.g., cancer) or be of interest due to its physiological role in the disease. Regulation of the output node provides a means to restore normalcy to the diseased network [12]. *This paper proposes a novel approach to select molecules that have high probability as drug targets.* First, we formalize "target prioritization", the problem of choosing a set of *putative target* molecules for further study (Section 3). Next, in Section 4 we present a fast and novel algorithm called PANI (**P**utative T**A**rget **N**odes Pr**I**oritization), which uses network information and simple empirical scores to prioritize and rank biologically relevant target molecules in signaling networks.

PANI is a generic algorithm applicable to any biological signaling network. The algorithm PANI prunes the *candidate nodes* (nodes being considered for analysis) based on a *reachability rule* and prioritizes nodes using a score based on *profile shape similarity distance* (PSSD), *target downstream effect* (TDE) and *bridging centrality* (BC) [14]. Putative target nodes are nodes with high ranking score. In Section 5, we evaluate the performance of PANI by comparing it against two state-of-the-art global sensitivity analysis (GSA)-based techniques (multi-parametric sensitivity analysis (MPSA) [34] and SOBOL [27]) run on three signaling networks. Instead of defining success according to the internal logic of the original networks, the goal is to agree with empirical outcome: namely, to *predict* the set of molecules that is actually targeted by drugs given to human patients. Our study shows that PANI is orders of magnitude faster and can identify a majority of targets in these networks whereas many of these targets are ignored by MPSA [34] and SOBOL [27]. Finally, extrapolating trends from the results suggests some insights and possible reasons why empirical outcome of disease is not addressed well by sensitivity analysis.

Symbols	Description
V_{meta}	Set of meta nodes $\{v_{meta:1}, v_{meta:2}, \cdots, v_{meta:i}\}$ where $v_{meta:i}$ is the i^{th} strongly connected component (SCC).
ζ_u	Concentration-time profile $\{\varsigma_{u[1]}, \varsigma_{u[2]}, \cdots, \varsigma_{u[i]}\}$ of node u where $\varsigma_{u[i]}$ is the value at time point i.
DTW(ζ_u, ζ_v)	Dynamic time warping (DTW) distance between ζ_u and ζ_v.
$\rho_{u,v}$	Probability of perturbing node v when node u is perturbed.
θ_u	Degree of node u.
Φ_v	Set of profile shape similarity distances (PSSD) $\{\Phi_{(u_1,v)}, \Phi_{(u_2,v)}, \cdots, \Phi_{(u_i,v)}\}$ with respect to v where $\Phi_{(u_i,v)}$ is the PSSD value between ζ_{u_i} and ζ_v.
Υ	Set of target downstream effect (TDE) $\{\Upsilon_{u_1}, \Upsilon_{u_2}, \cdots, \Upsilon_{u_i}\}$ where Υ_{u_i} is the TDE value of node u_i.
Λ	Set of bridging centrality (BC) $\{\Lambda_{u_1}, \Lambda_{u_2}, \cdots, \Lambda_{u_i}\}$ where Λ_{u_i} is the BC value of node u_i.
Ψ_X	Ranked list $\{\psi_{X:u_1}, \psi_{X:u_2}, \cdots, \psi_{X:u_i}\}$ based on property X where $\psi_{X:u_i}$ is the rank of node u_i. Node u_1 will be assigned a higher rank than u_2 ($\psi_{X:u_1} < \psi_{X:u_2}$) if $X_{u_1} > X_{u_2}$.
ω_X	Scalar weight factor associated to property X.
T	Set of nodes $\{t_1, t_2, \cdots, t_i\}$ such that there exists a path from each node $t_i \in T$ to the output node.

Table 1: Notations.

2. RELATED WORK

Sensitivity analysis [13, 34] has been frequently proposed for target identification and its goal is to rank parameters according to the effect of a particular parameter perturbation (e.g., a kinetic rate constant change) on the output node. Since the parameter values of a real biological network vary depending on genetics, cellular environment and cell type, GSA-based methods are deemed more appropriate as they measure the effect on the output node when all parameters are varied simultaneously [34].

Although GSA-based methods can identify sensitive parameters, they have several limitations. They are computationally expensive, require large number of simulations; ignore off-target effects; and may miss "insensitive" nodes that may be important drug targets. GSA-based approaches typically create many sets of simulation data using some random samplers and then use some statistical measures on the simulation results to determine which parameters should be ranked higher. In contrast, PANI prunes "irrelevant" nodes to reduce computational cost, then ranks the nodes by computing an *aggregate score* that is based on certain structural *and* kinetic properties of the network, instead of using sensitivity and focussing *solely* on the kinetic aspect of the network.

3. TARGET PRIORITIZATION PROBLEM

In this section, we introduce the terminologies and problem that we address in this paper. The key notations used in this paper are summarized in Table 1. In order to validate our results, we choose signaling networks that have been well-studied for the roles their nodes play when targeted with relevant drugs for a specific disease. They are MAPK-PI3K [9], MLC phosphorylation [23], and endomesoderm [18] networks, which are implicated in ovarian cancer, atrial fibrillation, and gastrulation phase of embryonic development, respectively. In the sequel, we shall use the heregulin (HRG)-induced MAPK-PI3K signaling network in [9] as a running example. Phosphorylated ERK (ERKPP) is selected as the output node due to its role in ovarian cancer [29]. Details of the ordinary differential equation (ODE) model (BIOMD0000000146) in [9] can be found in Biomodels.net [19].

3.1 Profile Shape Similarity Distance (PSSD)

In signaling networks, signal responses to perturbation are typically measured in terms of phosphoprotein concentrations dynamics [17] represented as concentration-time profiles. In signaling networks, profiles with variable time delays are common since reactions occur at different and non-uniform rates [1]. Hence, compared to Euclidean distance measure, *dynamic time warping* (DTW) distance (non-linear measure), allows a more intuitive alignment between profiles [16] and is more suitable for biological time series data [1].

*Definition 1. Given two discrete time series ζ_u and ζ_v, the **dynamic time warping distance** between them is defined recursively as:*

$$\text{DTW}(\zeta_u, \zeta_v) = \xi(First(\zeta_u), First(\zeta_v)) + Min \begin{cases} \text{DTW}(\zeta_u, Rest(\zeta_v)) \\ \text{DTW}(Rest(\zeta_u), \zeta_v) \\ \text{DTW}(Rest(\zeta_u), Rest(\zeta_v)) \end{cases}$$

where $First(\zeta_u) = \{\varsigma_{u[1]}\}$, $Rest(\zeta_u) = \{\varsigma_{u[2]}, \varsigma_{u[3]}, \cdots, \varsigma_{u[n]}\}$, $\xi(\varsigma_{u[i]}, \varsigma_{v[j]}) = (\varsigma_{u[i]} - \varsigma_{v[j]})^2$ and $\varsigma_{u[i]}$ is the value of ζ_u at time point i [16].

Although DTW distance is robust to time warping, it can miss similar profiles that have undergone y-axis warping [16] due to signals amplification or attenuation [2] and inversely similar profiles which are common for inhibitors [28] in signaling networks. In order to address these limitations, the profiles are Z-normalized to minimize the effects of y-axis warping; and DTW distances are computed for both the original profile (ζ_u) and the inversely similar profile (ζ'_u), of which the smaller distance is selected as the PSSD ($\Phi_{(u,v)}$).

*Definition 2. Given a concentration-time profile ζ_u having n time points, denoted as $\zeta_u = \{\varsigma_{u[0]}, \cdots, \varsigma_{u[n]}\}$, let m be the median value of ζ_u. The corresponding **inverted profile** is denoted as $\zeta'_u = \{\varsigma'_{u[0]}, \cdots, \varsigma'_{u[n]}\}$ where $\varsigma'_{u[i]} = 2 \times m - \varsigma_{u[i]}$.*

*Definition 3. Given a signaling network $H = (V_H, E_H)$, let ζ_u, ζ_v be the Z-normalized concentration-time profiles of $u, v \in V_H$. The **profile shape similarity distance** of u with respect to v is defined as:*

$$\Phi_{(u,v)} = Min(\text{DTW}(\zeta_u, \zeta_v), \text{DTW}(\zeta'_u, \zeta_v))$$

3.2 Target Downstream Effect (TDE)

Perturbations of nodes *downstream* of the target node is one of the contributing factors of off-target effects for drugs [20]. The *target downstream effect* of a node v assesses this risk based on the probability of perturbing a *downstream node* w and the likelihood of w causing off-target effects. Node w is *downstream* of v if there exists a path from v to w. The probability of perturbing a downstream node depends on the likelihood of the existence of a path from v to w (path probability). Hence, it can be calculated by assigning suitable edge weights using edge confidence score in protein-protein interaction (PPI) databases [30] and then multiplying the weights of all edges in the path. If there are multiple paths from v to w, the overall probability can be computed as the maximum of all paths' probabilities. The likelihood of a *downstream node* causing off-target effect is dependent on the degree of the node since high degree nodes are more likely to be involved in essential PPIs [11].

*Definition 4. Given $H = (V_H, E_H)$, let W be the set of downstream nodes of $v \in V_H \setminus W$. Let $\rho_{v,w}$ be the probability of perturbing $w \in W$ when target node v is perturbed and θ_w be the degree of w. The **target downstream effect** of v is defined as $\Upsilon_v = \sum_{w \in W} (\rho_{v,w} \times \theta_w)$.*

3.3 Bridging Centrality (BC)

The *bridging centrality* identifies *bridging nodes* (nodes with high *bridging centrality* value) which are located between functional modules in the signaling network and mediate signal flow between the modules [14]. Compared to hub nodes (nodes with high degree), bridging nodes are more effective drug targets with fewer off-target effects [14]. The *bridging centrality* of a node is the product of two ranks, namely, the inverses of *betweenness centrality* [3] and *bridging coefficient* [14], since bridging nodes have higher *betweenness centrality* and *bridging coefficient* than other nodes [14] and the ranking function used in this paper assigns higher rank to larger value. The *betweenness centrality* of a node v is $\Omega_v = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$ where σ_{st} is the number of shortest paths from node s to node t and $\sigma_{st}(v)$ is the number of shortest paths from s to t passing through v [3]. The *bridging coefficient* of a node v is $\Gamma_v = \frac{1}{\theta_v} \sum_{i \in N_v, \theta_i > 1} \frac{\eta_i}{\theta_i - 1}$ where θ_v is the degree of v, N_v is the set of neighbors of v, and η_i is the number of outgoing edges of node $i \in N_v$ [14].

Definition 5. Given the inverses of betweenness centrality rank ($\Psi_{\frac{1}{\Gamma:v}}$) and bridging coefficient rank ($\Psi_{\frac{1}{\Omega:v}}$) of node v, its ***bridging centrality*** *is defined as* $\Lambda_v = \Psi_{\frac{1}{\Gamma:v}} \times \Psi_{\frac{1}{\Omega:v}}$.

3.4 Putative Target Prioritization

The above three properties are used to determine if a node is a *putative target node*. A *putative target node* must promise better output node regulation (better efficacy) and reduced off-target effects than other nodes, which means smaller PSSD, smaller TDE, and larger BC values. Hence, *putative targets* prioritization is equivalent to a rank aggregation problem [24] with nodes ranked based on each property and the rankings aggregated into a combined score (*putative target score*). Nodes having top scores are called *putative target nodes* and prioritized over other nodes. We use the weighted-sum approach to aggregate the rank. This allows poor performance in one criterion to be compensated by good performance in other criteria, resulting in approximate ranking and hence, approximate prioritization. As we shall see in Section 5, this approximate prioritization is good enough as it can prioritize majority of the known drug targets over other nodes.

Definition 6. Given a signaling network $H = (V_H, E_H)$ and an output node $v_o \in V_H$, let Φ_{v_o} be the PSSD *property evaluated with respect to v_o, Υ and Λ be the* TDE *and* BC *properties, respectively. Let ω_c be the weight associated with property $c \in C = \{\Phi_{v_o}, \Upsilon, \frac{1}{\Lambda}\}$ and $\Psi_{c:v}$ be the rank of node $v \in V_H$, based on property c and normalized to a range of [0 1]. Then, the* ***putative target score*** *of a node v is defined as* $score_{v,C} = \sum_{c \in C} (\omega_c \times \Psi_{c:v})$ *where* $\sum_{c \in C} \omega_c = 1$.

Definition 7. Given a signaling network $H = (V_H, E_H)$ and an output node $v_o \in V_H$, the goal of the ***putative targets prioritization problem*** *is to rank the nodes using the putative target score (Ψ_{score}) such that top ranking nodes are prioritized as putative targets.*

The weights $\omega_{\Phi_{v_o}}$, ω_Υ and $\omega_{\frac{1}{\Lambda}}$ affect the putative target score and hence the decision of whether a node is a putative target. Interestingly, among the entire range of weights (ω_c) we tested, the minimum number of top ranking targets (*MinNode*) required to identify at least 75% of the relevant known drug targets in [26] is 19 and 72 for the MAPK-PI3K and MLC phosphorylation networks, respectively. The size of the networks are 36 and 105, respectively. Further, we note that the impact of ω_c on the ranking result reduces with increasing network size. When ω_c varies in the range [0.1 – 0.9], the Spearman ranking coefficients are in the ranges [0.45 – 1] and [0.8 – 1], respectively for the MAPK-PI3K and the endomesoderm network (622 nodes). Due to space constraints, the effects of using different weights are described in [5]. In the sequel, we assign $\omega_{\Phi_{v_o}} = 0.4$, $\omega_\Upsilon = 0.3$ and $\omega_{\frac{1}{\Lambda}} = 0.3$.

4. THE ALGORITHM PANI

The algorithm consists of two phases which we shall elaborate in turn. Due to space constraints, the pseudocode and the time and space complexities are given in [5].

Phase 1: Target Pruning. There are four subphases: *bipartite graph conversion*, *directed acyclic graph* (DAG) *conversion*, DAG *indexing* and *reachability-based pruning*. The first two subphases preprocess the input hypergraph into a DAG which has a consistent topological ordering, making indexing of the DAG easier subsequently. The hypergraph is converted into its corresponding bipartite graph using [7]. Then, the SCCs are identified in the graph using [31] and replaced with corresponding meta nodes $v_{meta:i}$ using [32] to form the DAG. We index the DAG using [4] to facilitate efficient evaluation of node reachability (*reachability-based pruning*) which reduces target search space.

Phase 2: Target Prioritization. In this phase, the set of pruned nodes T (filtered from Phase 1) is prioritized based on their PSSD, TDE and BC. The concentration-time profiles used to compute the PSSD may be obtained from experiments or *in silico* simulations of biological models. In this paper, we use the latter approach. For instance, the MAPK-PI3K ODE model [9] was simulated in *Copasi* using parameters: {*duration*=1800 seconds, *intervals*=6 seconds}[1]. Table 2 reports the normalized ranks of nodes for PSSD, TDE and the inverse of BC, denoted as $\Psi_{\Phi_{\text{ERKPP}}}$, Ψ_Υ and $\Psi_{\frac{1}{\Lambda}}$, respectively.

5. EXPERIMENTAL RESULTS

PANI is implemented in Java JDK 1.6. In this section, we present the experiments conducted to evaluate its performance and report some of the results obtained. More detailed results are available in [5]. We compare PANI against MPSA [34] and SOBOL [27]. The SBML-SAT tool [35] is used to perform MPSA and SOBOL analysis[2]. We use three real-world signaling networks as our dataset, namely MAPK-PI3K [9], MLC phosphorylation [23], and endomesoderm [18]. Due to space constraints, we will focus mainly on the results related to MAPK-PI3K network here. We run all experiments on an Intel 1.86GHz dual core processor machine with 2GB RAM, running Microsoft Windows XP. For the MAPK-PI3K network, we set $|\zeta| = 300$, $\omega_{\Phi_{v_o}} = 0.4$, $\omega_\Upsilon = 0.3$ and $\omega_{\frac{1}{\Lambda}} = 0.3$. We set $\rho_{v,w} = 1$ since the network is very well-studied.

Execution times. Table 3 reports the execution times of the three methods on three networks of increasing size. The

[1]The actual CPU time required for the simulation is about 1 second using a 32-bit operating system with 2GB RAM and a dual core processor at 1.86GHz. The simulation time is unrelated to the duration parameter which intuitively, corresponds to the range of ζ and is related to $|\zeta|$ ($\frac{duration}{interval} = |\zeta|$).

[2]SBML-SAT was obtained from http://sysbio.molgen.mpg.de/SBML-SAT/ and the default number of simulations set to 2000 and 10000 for MPSA and SOBOL, respectively.

Ψ_P	Node	$\Psi_{\Phi_{v_o}}$	Ψ_Υ	$\Psi_{\frac{1}{\Lambda}}$	Ψ_M	Ψ_S
Kinase						
1	ERKPP$^\flat$	34	16	21	32	32
2	AktPIP$\dagger^\flat$	28	13	24	35	35
4	ERKP$^\flat$	21	17	19	33	33
5	RP$\dagger^\flat$	33	5	26	27	27
6	RHRG2$^\flat$	32	4	27	25	25
8	AktPIP3$\dagger^\flat$	16	12	28	36	36
9	AktPIPP$\dagger^\flat$	27	13	9	34	34
10	Raf$^\star\dagger^\flat$	13	11	30	19	19
11	MEKPP$\dagger^\flat$	17	14	16	30	30
12	PI3K$^\star\dagger^\flat$	22	3	29	29	29
13	MEK$^\flat$	20	10	18	3	3
14	MEKP$^\flat$	19	13	10	31	31
15	ERK$^\flat$	18	16	5	1	1

Ψ_P	Node	$\Psi_{\Phi_{v_o}}$	Ψ_Υ	$\Psi_{\frac{1}{\Lambda}}$	Ψ_M	Ψ_S
Kinase						
16	Raf$\dagger^\flat$	14	9	22	9	9
17	PI3K$^\flat$	25	2	21	13	12
18	RPI3K$^{\star\flat}$	31	3	11	23	23
19	Akt	29	9	2	14	11
20	RHRG	26	3	17	26	26
23	RPI3K	24	3	8	24	24
27	E	12	8	3	7	6
35	R	3	1	2	11	5
Phospholipid						
3	PIP3$^\flat$	23	11	31	28	28
7	PI$^\flat$	30	9	20	6	14
GTPase						
21	RasGTP	5	7	32	18	18
25	RasGDP	6	6	15	8	13

Ψ_P	Node	$\Psi_{\Phi_{v_o}}$	Ψ_Υ	$\Psi_{\frac{1}{\Lambda}}$	Ψ_M	Ψ_S
Phosphastase						
22	MKP3	12	15	4	10	8
29	PP2A	12	8	1	12	7
Adaptor molecule						
24	ShGS	7	4	25	17	17
28	Shc	4	2	23	2	2
32	GS	9	3	7	4	4
34	ShP	1	2	12	16	16
Tyrosine kinase receptor:adaptor molecule complex						
26	RShGS	11	3	14	22	22
30	RShc	8	3	13	20	20
31	RShP	10	3	6	21	21
Others						
33	internalization	15	1	3	15	15
36	HRG	2	1	2	5	10

Table 2: Prioritization result with ERKPP as output node (v_o). Nodes marked with †and $^\flat$ are known ovarian cancer drug targets in [26] and PANI-identified putative target nodes, respectively. Ψ_P, Ψ_M and Ψ_S denote the rankings based on PANI, MPSA and SOBOL, respectively.

Network ($H = (V_H, E_H)$)	$\lvert V_H\rvert$	Execution Time			$\frac{\lvert\tau_{\text{MPSA}}\rvert}{\lvert\tau_{\text{PANI}}\rvert}$	$\frac{\lvert\tau_{\text{SOBOL}}\rvert}{\lvert\tau_{\text{PANI}}\rvert}$
		τ_{PANI}	τ_{MPSA}	τ_{SOBOL}		
MAPK-PI3K network [9]	36	∼6sec	∼18min	∼3hrs	180	1800
MLC phosphorylation network [23]	105	∼11sec	∼2hr	∼21hrs	654.55	6872.73
Endomesoderm network [18]	622	∼251sec	-	-	-	-

Table 3: Execution times of various approaches.

SBML-SAT tool encounters segmentation violation error for the endomesoderm network whereas PANI takes around 251 sec. Observe that PANI *is at least two orders of magnitude faster than* MPSA *and* SOBOL.

Quality and relevance of results. We validate the quality of the results by the minimum number of top scoring nodes needed for identifying all the relevant known drug targets in [26] (*MinNode*); the top-3 drug target classes in terms of their average putative target score; and biological relevance of potential drug targets identified by PANI, MPSA and SOBOL. The quality of predicted targets can be evaluated empirically by comparing them against the known targets of drugs that are chosen for trials in human. We define the reference set of "good quality" targets as ovarian cancer drugs in [26] whose therapeutic effect is associated to regulation of ERK since ERKPP upregulation has been implicated in cancer [29] and [9] is based on Chinese hamster ovary cell. Details of the curation steps and drug targets are in [5].

The *MinNode* is 16, 36 and 36 for PANI, MPSA, and SOBOL (Table 2), respectively. Paired t-test analysis on the ranks of known drug targets shows that PANI ranks drug targets higher than MPSA and SOBOL ($p < 0.01$). Receiver Operating Characteristic (ROC) analysis reveals that PANI (area under the curve (AUC)=0.853) identifies known drug targets better than SOBOL (AUC=0.246) and MPSA (AUC=0.246). In 68 out of 100 random prioritization trials, PANI also ranks drug targets higher based on paired t-test statistics ($p < 0.05$). Hence, PANI *is able to identify known ovarian cancer drug targets using much fewer top scoring nodes and tends to prioritize known drug targets better than* MPSA, SOBOL *and random prioritization.* From Table 2, we note that PANI's top-3 drug target classes are phospholipids, kinases and GTPases while MPSA's and SOBOL's are adaptor molecules, phosphatases and other classes. Protein kinases are recognized as important drug targets [6]. Although phospholipids, phosphatases and GTPases are generally not considered good drug targets, they have been seriously considered recently [10, 22, 25]. PANI *identifies classes with relevance as drug targets which are distinct from* MPSA *and* SOBOL. When we examine the biological relevance of nodes having significantly different ranks (rank difference of $\frac{|T|}{2}$ or more) based on PANI, MPSA and SOBOL, we find that they have high biological relevance as potential ovarian cancer drug targets [8, 33]. We also note that the set of nodes ranked high in PANI, but low in MPSA or SOBOL contains mainly known drug targets while the set of nodes ranked high in MPSA or SOBOL, but low in PANI contains some targets whose efficacy may be dependent on the expression level of other proteins [33].

Remark. We note that GSA-based approaches tend to prioritize nodes with long pathways to the output (such as receptors) whereas PANI tends to choose targets with short pathways. The high sensitivity of concentration levels of downstream protein to that of upstream initiating proteins is mathematically correct due to amplifying effect of the signal transduction cascade provided that there is no influences on molecules in the cascade which would disrupt signal propagation [15]. Hence, PANI may have done well because of its bias towards short signaling distances which relies on fewer biochemical/pathway assumptions and creates less vulnerability to a larger number of unforeseen effects. Although signaling distance can prioritize known drug targets (experimental results in [5]), its low granularity (assigning multiple nodes to the same rank) limits its usage in networks with large SCC.

Summary of other experimental results. First, aggregate ranking tends to perform better than individual ranking. For MAPK-PI3K network, the AUC are 0.853, 0.701, 0.763 and 0.833 for PANI, PSSD, TDE, and BC, respectively. Second, the execution time of PANI increases with profile length $|\zeta|$ when $|\zeta|$ varies in the range [10–1000]. We observe that $|\zeta| = 100$ is sufficient for reliable analysis since the correlation cofficients of the ranks at $|\zeta| \geq 100$ are $\sim 100\%$. Third, as mentioned in Section 3.4, varying the weights ω_c for the properties in the range [0.1 – 0.9] did not affect the ranking results significantly. Finally, selecting output nodes with no outlinks results in larger $|T|$ which varies linearly with execution time; and selecting output nodes in the same SCC produces closer rank correlation coefficient and more similar prioritization results. The choice of output nodes affects the decision of whether a candidate is pruned and the candidate node's PSSD value. Hence, output nodes in the same SCC have similar reachability property and are likely to share similar PSSD, suggesting that protein post-translational profiles in the same SCC for signaling networks are highly correlated. Interested readers may refer to [5] for details.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we propose PANI, a novel algorithm for selecting a set of putative target nodes from a signaling network and validation on several networks reveals it to be faster and more effective than GSA-based methods. Poor GSA performance may be because the sensitivity criterion seeks high-magnitude correlations, while PANI seeks robust correlations with few off-target effects. We note the unexpected trend of PANI ranking more actual clinical targets higher than mathematically rigorous GSA-based approaches. Hence, this work constitutes anecdotal evidence that heuristic common sense is still needed and useful for bridging the gap between the analysis of quantitative models, and the medical reasons why we build these models.

Future extension of this work includes producing more curated datasets of drug targets of additional diseases to facilitate improved statistical evaluation of target prioritization methods; extending PANI to handle multiple output nodes since a disease may be due to dysfunctions in various points instead of a single point of the networks; extending the dimensionality of the trigger, the concentration-time profiles and the shape similarity distance measure to handle models built around a variable dose input (e.g., bistable models); integrating heuristics from PANI such as filtering out molecules that are very poor by off-target criteria into GSA-based approaches. A potentially useful application of PANI is analysis of incomplete signaling networks with missing rate constants which GSA-based methods cannot perform. Proteomic methods such as SILAC are now providing an explosive increase of concentration-time profile for this purpose, but in the event that no rate parameters and incomplete concentrations are available, PANI can perform a partial analysis with concentration-time profiles of a partial set of nodes, and identify putative target nodes from within this partial set.

7. ACKNOWLEDGMENTS

The authors are supported by grant from the Singapore-MIT Alliance Programme in Computational and Systems Biology.

8. REFERENCES

[1] J. Aach et al. Aligning gene expression time series with time warping algorithms. *Bioinformatics*, 17(6):495–508, Jun 2001.

[2] A. Asthagiri et al. A computational study of feedback effects on signal dynamics in a mitogen-activated protein kinase (MAPK) pathway model. *Biotechnol Prog*, 17(2):227–239, 2001.

[3] U. Brandes. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, 25:163–177, 2001.

[4] L. Chen et al. Stack-based algorithms for pattern matching on DAGs. In *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, pages 493–504. VLDB Endowment, 2005.

[5] H. Chua et al. PANI: A novel algorithm for fast discovery of putative target nodes in signaling networks. www.cais.ntu.edu.sg/∼assourav/TechReports/PANI-TR.pdf, 2010.

[6] P. Cohen. Protein kinases–the major drug targets of the twenty-first century? *Nat Rev Drug Discov*, 1(4):309–315, Apr 2002.

[7] J. Engelfiet et al. A comparison of boundary graph grammars and context-free hypergraph grammars. *Information and Computation*, 84(2):163 – 206, 1990.

[8] M. S. Gordon et al. Clinical Activity of Pertuzumab (rhuMAb 2C4), a HER Dimerization Inhibitor, in Advanced Ovarian Cancer: Potential Predictive Relationship With Tumor HER2 Activation Status. *J Clin Oncol*, 24(26):4324–4332, 2006.

[9] M. Hatakeyama et al. A computational model on the modulation of mitogen-activated protein kinase (MAPK) and Akt pathways in heregulin-induced ErbB signalling. *Biochem J*, 373(Pt 2):451–463, Jul 2003.

[10] J. He et al. Antiphosphatidylserine antibody combined with irradiation damages tumor blood vessels and induces tumor immunity in a rat model of glioblastoma. *Clin Cancer Res*, 15(22):6871–6880, Nov 2009.

[11] X. He et al. Why do hubs tend to be essential in protein networks? *PLoS Genetics*, 2(6):e88, 06 Jun 2006.

[12] L. Hood et al. Systems biology and new technologies enable predictive and preventative medicine. *Science*, 306(5696):640–643, Oct 2004.

[13] D. Hu et al. Time-dependent sensitivity analysis of biological networks: Coupled MAPK and PI3K signal transduction pathways. *The J of Phy Chem A*, 110(16):5361–5370, 2006.

[14] W.-C. Hwang et al. Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery. *Clin Pharmacol Ther*, 84(5):563–572, Nov 2008.

[15] Y. Ishikawa et al. Cardiac Myosin Light Chain Kinase: A New Player in the Regulation of Myosin Light Chain in the Heart. *Circ Res*, 102(5):516–518, 2008.

[16] E. Keogh et al. Derivative dynamic time warping. In *In First SIAM International Conference on Data Mining (SDMŠ2001*, 2001.

[17] C. Kreutz et al. An error model for protein quantification. *Bioinformatics*, 23(20):2747–2753, Oct 2007.

[18] C. Kuhn et al. Monte carlo analysis of an ode model of the sea urchin endomesoderm network. *BMC Systems Biology*, 3(1):83, 2009.

[19] N. Le Novère et al. Biomodels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Res*, 34(Database issue):D689–D691, Jan 2006.

[20] D. Liebler et al. Elucidating mechanisms of drug-induced toxicity. *Nat Rev Drug Discov*, 4(5):410–420, May 2005.

[21] J. Low et al. Phenotypic fingerprinting of small molecule cell cycle kinase inhibitors for drug discovery. *Curr Chem Genomics*, 3:13–21, 2009.

[22] Q. Lu et al. Signaling through rho gtpase pathway as viable drug target. *Current Medicinal Chemistry*, 16:1355–1365(11), April 2009.

[23] A. Maeda et al. Ca2+-independent phospholipase a2-dependent sustained rho-kinase activation exhibits all-or-none response. *Genes to Cells*, 11:1071–1083, 2006.

[24] N. Mamoulis et al. Efficient top-k aggregation of ranked inputs. *ACM Trans. Database Syst.*, 32(3):19, 2007.

[25] J. McConnell et al. Targeting protein serine/threonine phosphatases for drug development. *Molecular Pharmacology*, 75(6):1249–1261, 2009.

[26] NIH. Clinicaltrials.gov, 2009. Accessed 2 Feb 2010.

[27] I. Sobolá. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.*, 55(1-3):271–280, 2001.

[28] R. Steinman et al. Activation of Stat3 by cell confluence reveals negative regulation of Stat3 by cdk2. *Oncogene*, 22(23):3608–3615, Jun 2003.

[29] R. Steinmetz et al. Mechanisms Regulating the Constitutive Activation of the Extracellular Signal-Regulated Kinase (ERK) Signaling Pathway in Ovarian Cancer and the Effect of Ribonucleic Acid Interference for ERK1/2 on Cancer Cell Proliferation. *Mol Endocrinol*, 18(10):2570–2582, 2004.

[30] D. Szklarczyk et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, 39(suppl 1):D561–D568, 2011.

[31] R. Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972.

[32] S. Trißl et al. Fast and practical indexing and querying of very large graphs. In *SIGMOD '07: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 845–856, New York, NY, USA, 2007. ACM.

[33] N. Yumoto et al. Expression of the erbb4 receptor causes reversal regulation of pp2a in the shc signal transduction pathway in human cancer cells. *Mol Cell Biochem*, 285(1-2):165–171, Apr 2006.

[34] Z. Zi et al. In silico identification of the key components and steps in IFN-gamma induced JAK-STAT signaling pathway. *FEBS Lett*, 579(5):1101–1108, Feb 2005.

[35] Z. Zi et al. SBML-SAT: a systems biology markup language (SBML) based sensitivity analysis tool. *BMC Bioinformatics*, 9:342, 2008.