

AD \_\_\_\_\_

(Leave blank)

Award Number:  
W81XWH-05-1-0267

TITLE:  
Functional proteomic analysis of signaling networks and response to targeted therapy

PRINCIPAL INVESTIGATOR:  
Pralhad Ram, PhD

CONTRACTING ORGANIZATION:  
University of Texas MD Anderson Cancer Center  
Houston, TX 77030

REPORT DATE:  
March 2009

TYPE OF REPORT:  
Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command  
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: (Check one)

- Approved for public release; distribution unlimited
- Distribution limited to U.S. Government agencies only;  
report contains proprietary information

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

**REPORT DOCUMENTATION PAGE**Form Approved  
OMB No. 074-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503

<b>1. AGENCY USE ONLY (Leave blank)</b>		<b>2. REPORT DATE</b> 20-mar-2009	<b>3. REPORT TYPE AND DATES COVERED</b> Final 21 FEB 2005 - 20 FEB 2009	
<b>4. TITLE AND SUBTITLE</b> Functional proteomic analysis of signaling networks and response to targeted therapy			<b>5. FUNDING NUMBERS</b>	
<b>6. AUTHOR(S)</b> Prahlad T. Ram  Email: prahlad.ram@mssm.edu				
<b>7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)</b>  University of Texas MD Anderson Cancer Center 1515 Holcombe Blvd Houston, TX 77030  E-Mail: <a href="mailto:pram@mdanderson.org">pram@mdanderson.org</a>			<b>8. PERFORMING ORGANIZATION REPORT NUMBER</b>	
<b>9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)</b>  U.S. Army Medical Research and Materiel Command Fort Detrick, Maryland 21702-5012			<b>10. SPONSORING / MONITORING AGENCY REPORT NUMBER</b>	
<b>11. SUPPLEMENTARY NOTES</b>				
<b>12a. DISTRIBUTION / AVAILABILITY STATEMENT</b> Approved for Public Release; Distribution Unlimited			<b>12b. DISTRIBUTION CODE</b>	
<b>13. ABSTRACT (Maximum 200 Words)</b> The purpose of the research done has been to determine the regulation of the EGFR network and identify how manipulations of the network alter signal flow to bypass targeted inhibitions. The scope of the project is to understand the network and determine which molecules have to be targeted to inhibit tumor cell proliferation. We have completed the tasks of the grant proposal. In summary we have developed a quantitative phosphor-proteomic approach to measure changes in signaling in response to EGF in breast cancer cells. We have developed a quantitative model that incorporates the experimental data. Using this model we studied the signaling dynamics. We also manipulated specific nodes in the network and measured changes in the signal flow. Using this data in the model we determined feedback loops that are present in the cell. Using the developed model we predicted what combinations have to be targeted to overcome feedback bypass of targets. These predictions were tested experimentally. The results from our work has been published in different parts in 5 publications. The DOD proposal has also allowed us to obtain a NIH R01 grant.				
<b>14. SUBJECT TERMS</b>  None provided.			<b>15. NUMBER OF PAGES</b>	
			<b>16. PRICE CODE</b>	
<b>17. SECURITY CLASSIFICATION OF REPORT</b> Unclassified	<b>18. SECURITY CLASSIFICATION OF THIS PAGE</b> Unclassified	<b>19. SECURITY CLASSIFICATION OF ABSTRACT</b> Unclassified	<b>20. LIMITATION OF ABSTRACT</b> Unlimited	

## Table of Contents

Introduction.....	3
Body.....	3
Key Research Accomplishments.....	8
Reportable Outcomes.....	8
Conclusions.....	8
References.....	8
Appendices.....	9-48

Final Report

**Introduction**

The purpose of the research done has been to determine the regulation of the EGFR network and identify how manipulations of the network alter signal flow to bypass targeted inhibitions. The scope of the project is to understand the network and determine which molecules have to be targeted to inhibit tumor cell proliferation.

**Body**

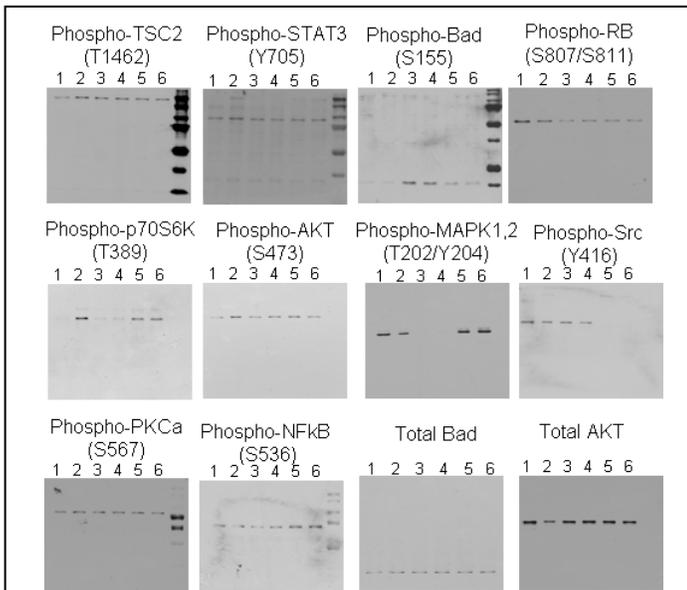
The aims of the proposal were as follows.

1. Determine the dynamics of the EGFR/MAPK/Stat3/PI3K signaling network in response to EGF and the drugs Iressa and Herceptin in breast cancer cells.
2. Determine how signals are integrated and routed within this EGF response network and identify important nodes that regulate the network.
3. Determine what combinations of targeted inhibitors effectively reduce proliferation for each of the breast cancer cell lines.

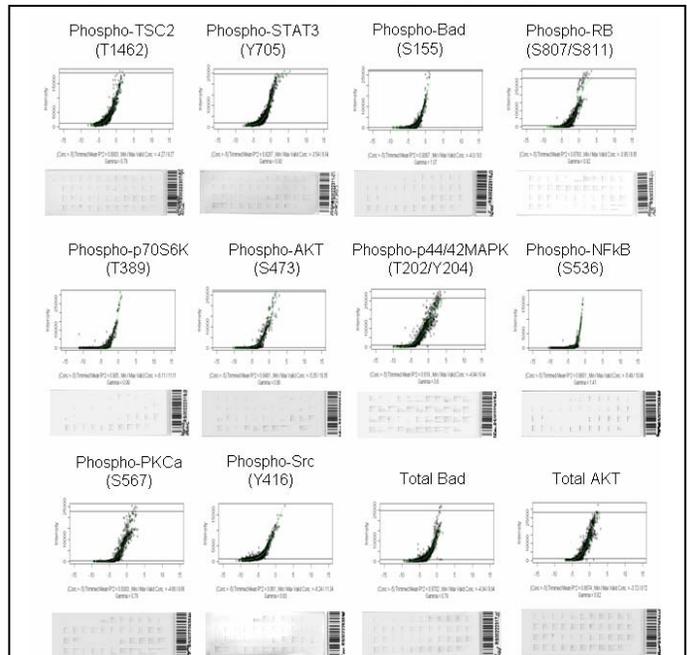
Since the final report guidelines state that “Journal publications **can be** substituted for detailed descriptions of specific aspects of the research” we have included 5 of our papers that describe results in detail.

**Aim 1** of the application was the development of quantitative reverse phase protein micro arrays to determine the changes in the signaling network. We have accomplished this task.

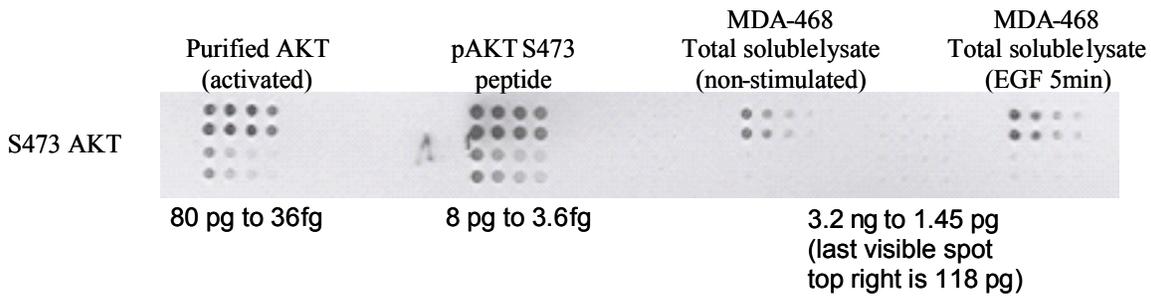
Figure 1 and 2 below shows development of the quantitative array to measure changes in phospho-protein



**Figure 1. Western blot of MCF10A cells.** MCF10A breast cells were treated with Lapatinib, Dasatinib or DMSO as control followed by stimulation with EGF. 1. Control, 2. EGF, 3. Lapatinib, 4. Lapatinib + EGF, 5. Dasatinib, 6. Dasatinib + EGF. Cell lysates were aliquoted and used on the RPPA and for Western blot analysis.

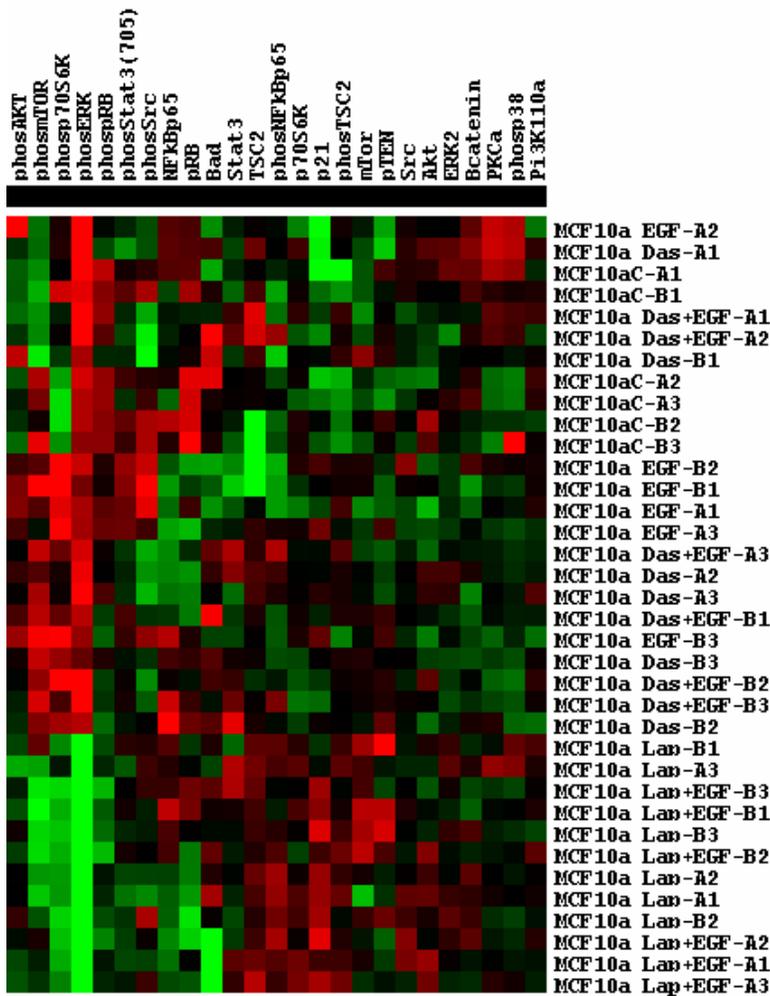


**Figure 2 RPPA slides and statistical data analysis of each slide.** An aliquot of the MCF10A lysate was spotted on RPPA slides and each slide was probed with the antibody listed. The same antibodies used for the Western blot in Figure 3 was used for the RPPA. The graphs show the supercurve analysis data fit, the R<sup>2</sup> values are between 0.82 and 0.96..



**Figure 3. Quantification of changes in activity of signaling molecules using RPPA.** Purified protein, EGF treated and non-stimulated MDA-468 cell lysates and phospho-AKT peptide were serially diluted and spotted on a RPPA slide. The slide was probed using phospho-S473 AKT antibody. The slide is shown along with the range of concentrations of the samples spotted on the array.

levels in cells upon treatment with EGF.

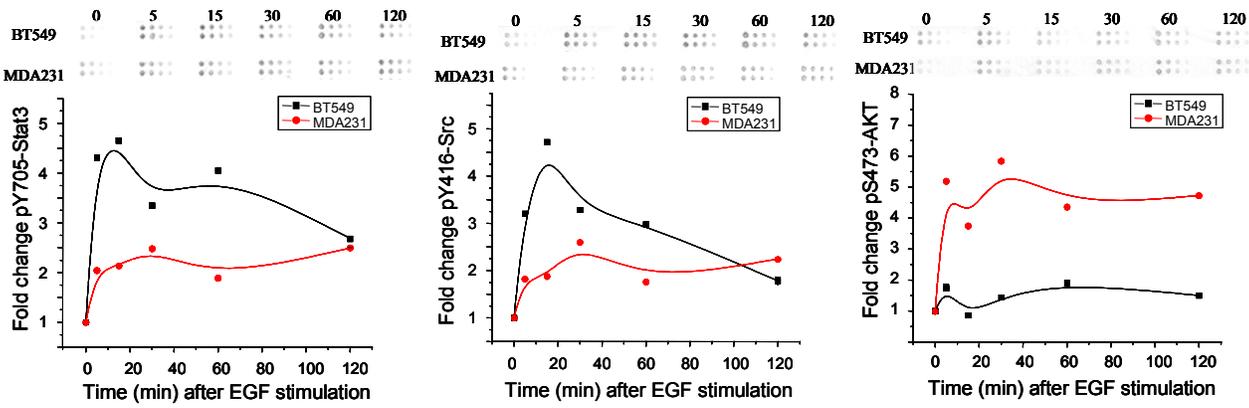


**Figure 4 RPPA from MCF10A breast cells.** MCF10A cells were incubated with 2 different pharmacological inhibitors for 4 hours or with DMSO as control. Cells were then stimulated with EGF or vehicle for 20 minutes. The cell lysates were spotted on the RPPA and probed using 30 different phospho and total antibodies. The data is mean centered for each antibody: **black** - mean intensity value, **red** - increased intensity of signal and **green** - decreased intensity of signal. Lap- Lapatinib, Das- Dasatinib, C-control, A&B 1,2,3 – experiment replicate numbers (n=6).

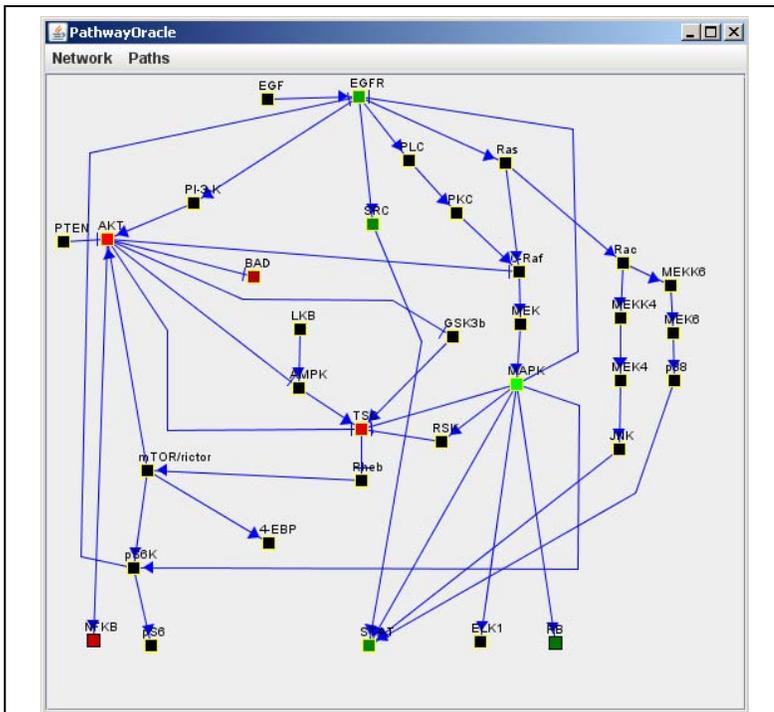
Figure 3 shows the quantitative aspect of the RPPA whereby we are able to detect picogram amounts of protein.

Figure 4 shows the RPPA data from the same lysates used in Fig 1 and 2. As can be seen in the figure 4 we are able to simultaneously measure protein levels of a large number of conditions.

The data in figure 5 shows the quantification of changes in the dynamic activity of signaling molecules upon stimulation with EGF. We have extensively investigated the dynamic changes in signaling of several members of the signaling network, including MAPK, AKT, Stat3, Src, S6K, p38, NFkB, and JNK (please see attached publications).



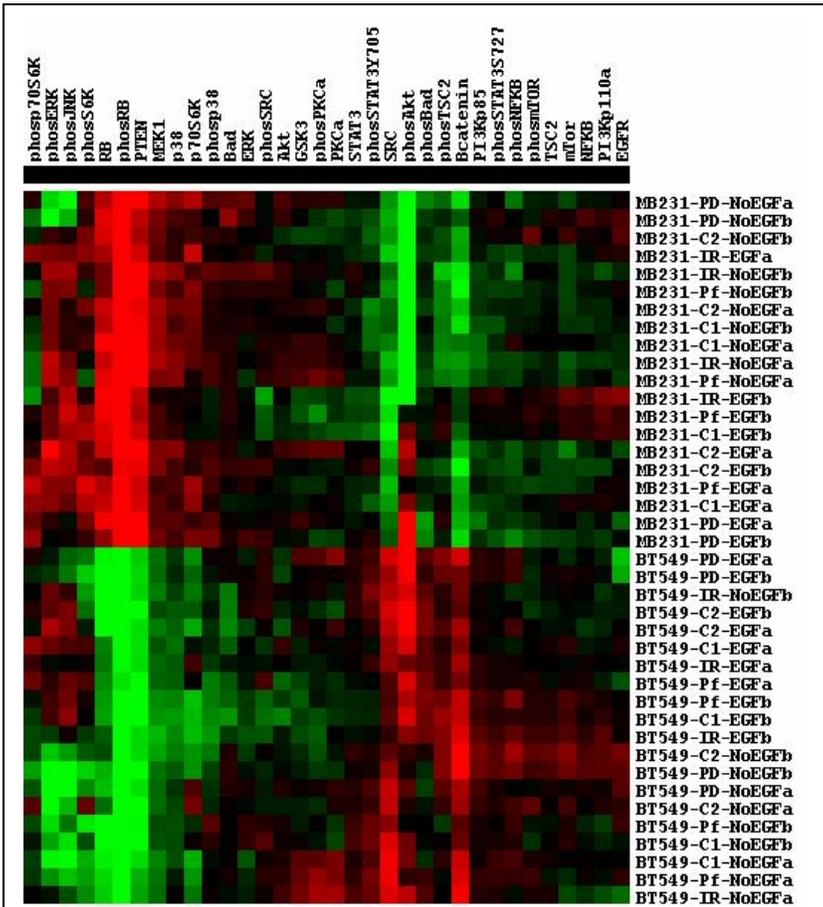
**Figure 5 Simultaneous measurements of activity states of molecules within the EGF signaling network.** Human breast cancer cells (MDA231 and BT549) were stimulated with 20 ng/ml EGF for the times indicated. The cells were lysed, quantified, and equal concentrations of lysates and their serial dilutions spotted on the array. The slides were then probed with phospho-antibodies to  $\gamma$ 705Stat3,  $\gamma$ 416Src, and s473AKT as well as antibodies to total Stat3, Src, and AKT. The slides were scanned and spot intensity quantified. The change in activity was measured for only those concentrations that were within the linear range. The activity was normalized to total protein levels and the data shown on the graph above.



**Figure 6 PathwayOracle model of Lapatinib signaling in MCF10A cells.** Red indicates increase in signal compared to control non treated cells, and green indicates a decrease in signal, black are no changes or non-measured nodes.

**Aim2** of the grant was to integrate the biological experimental data into a computational model to determine dynamic properties of the network.

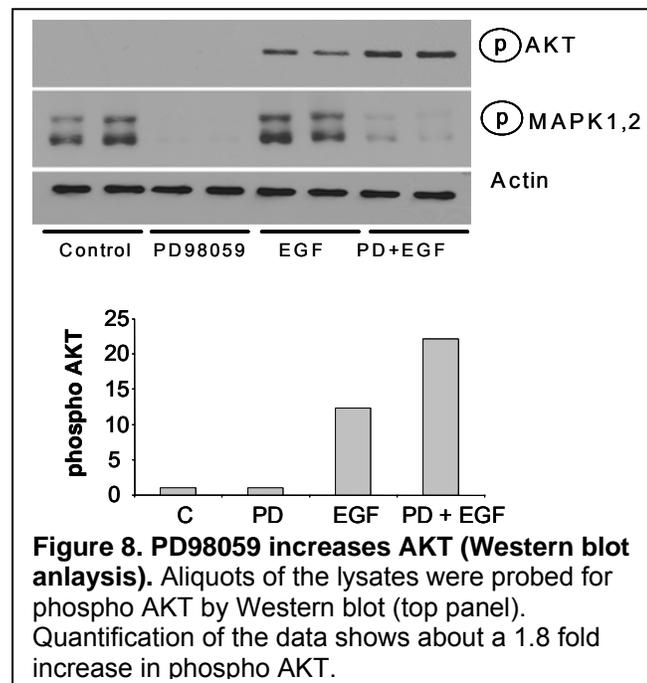
Figure 6 shows the computational analysis of information flow in the EGFR network from experimental data measured by RPPA. Please see attached publications Ruths et al 2008 and Iadevaia et al 2010)



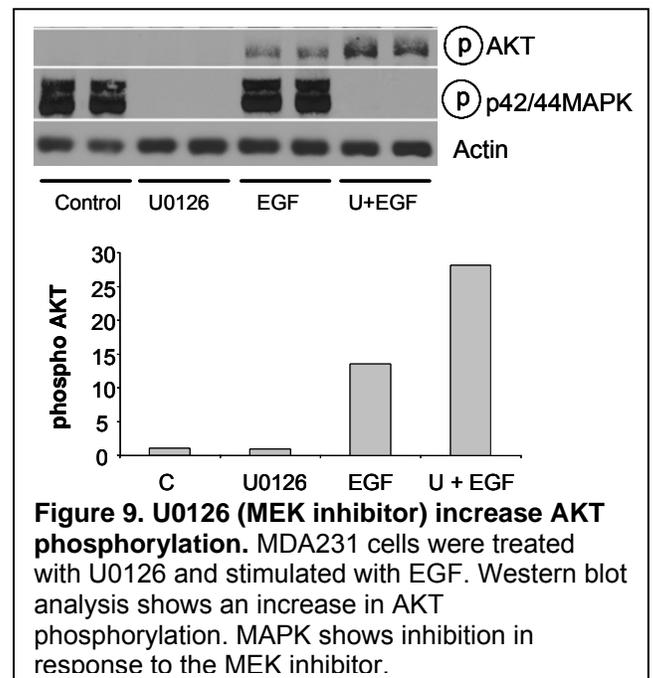
**Figure 7. RPPA analysis of BT549 and MDA231 cells.** The two cell lines were treated with MEK (PD), EGFR (IR), AKT (Pf) inhibitors, or control (C) and stimulated with EGF for 30 minutes (EGF), or non-EGF stimulated, a&b are two independent samples. Black is mean intensity, red increase, and green decrease in intensity.

We perturbed individual nodes in the network to determine how signaling is altered through the network. Figure 7 shows the RPPA data from one such experiment where three different drugs were used in two cell lines and the cells were stimulated with EGF. Similar experiments were done for the entire panel of cell lines.

Analysis of the data revealed several feedback loops that became activated when one node was inhibited. One such loop was the increase in pAKT when MEK was inhibited. This is seen in the RPPA as well as by western blot analysis (Figure 8&9). Details of these can be found in the attached publications.



**Figure 8. PD98059 increases AKT (Western blot analysis).** Aliquots of the lysates were probed for phospho AKT by Western blot (top panel). Quantification of the data shows about a 1.8 fold increase in phospho AKT.

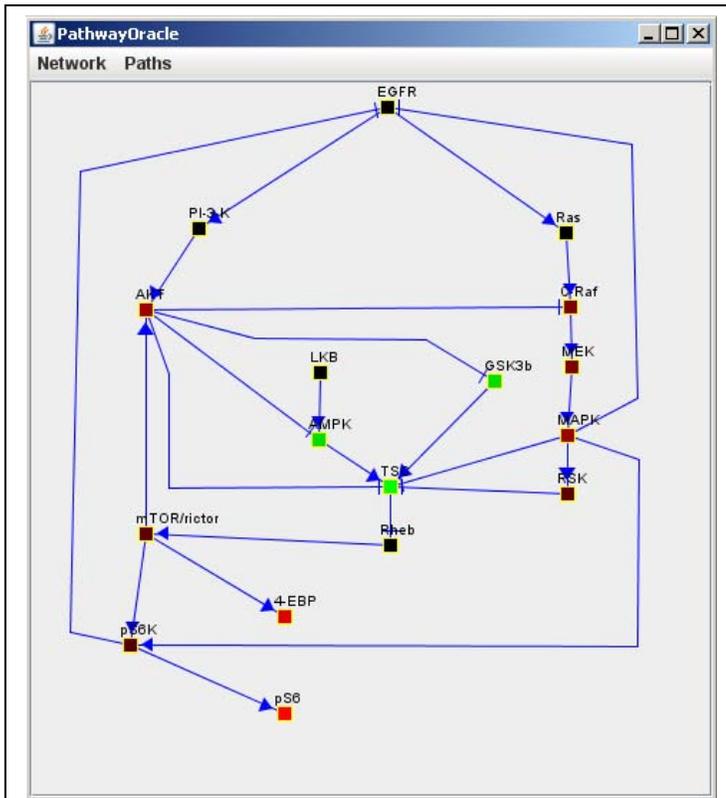


**Figure 9. U0126 (MEK inhibitor) increase AKT phosphorylation.** MDA231 cells were treated with U0126 and stimulated with EGF. Western blot analysis shows an increase in AKT phosphorylation. MAPK shows inhibition in response to the MEK inhibitor.

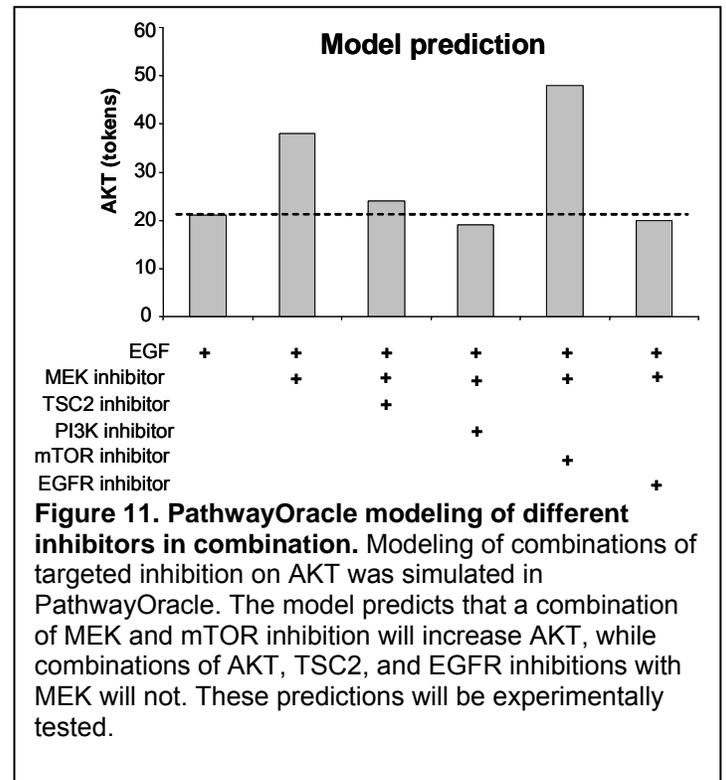
**Aim 3** of the grant was to use data and the model to determine optimal combinations for each cell line.

Based on the experimental data (Figure 7-9) and our developed model (Figure 10) we predicted different combinations of targets (Figure 11). We experimentally tested (Figures 12 & 13) the predictions and showed

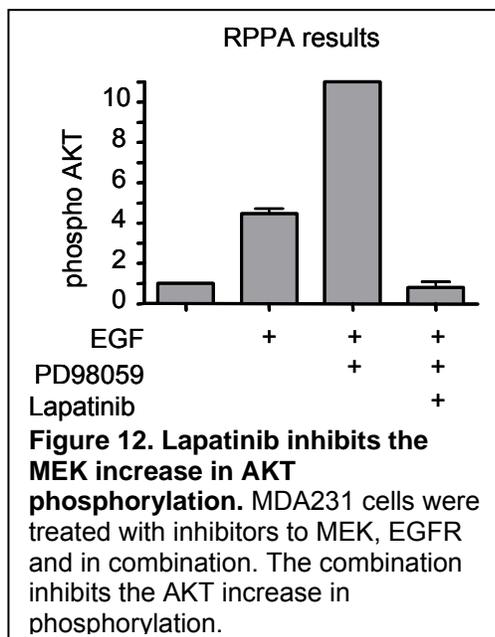
that infact combination targeting can overcome deficiencies of single targeted agents. These and additional data are seen in the publications Ruths et al 2008 and Iadevaia et al 2010.



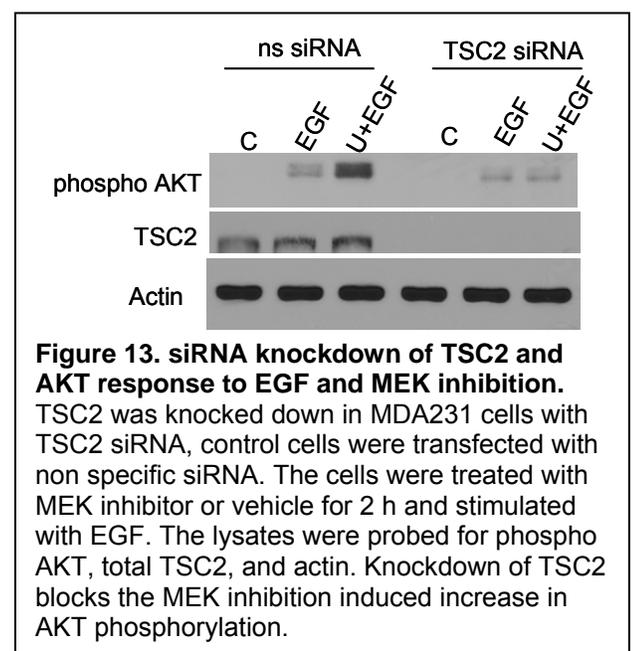
**Figure 10 Subnetwork showing the connectivity of the EGFR/MAPK/AKT subnetwork.** The colors indicate changes in phosphorylation in response to EGF compared to non stimulated control cells. Red is increase and green is decrease in phosphorylation.



**Figure 11. PathwayOracle modeling of different inhibitors in combination.** Modeling of combinations of targeted inhibition on AKT was simulated in PathwayOracle. The model predicts that a combination of MEK and mTOR inhibition will increase AKT, while combinations of AKT, TSC2, and EGFR inhibitions with MEK will not. These predictions will be experimentally tested.



**Figure 12. Lapatinib inhibits the MEK increase in AKT phosphorylation.** MDA231 cells were treated with inhibitors to MEK, EGFR and in combination. The combination inhibits the AKT increase in phosphorylation.



**Figure 13. siRNA knockdown of TSC2 and AKT response to EGF and MEK inhibition.** TSC2 was knocked down in MDA231 cells with TSC2 siRNA, control cells were transfected with non specific siRNA. The cells were treated with MEK inhibitor or vehicle for 2 h and stimulated with EGF. The lysates were probed for phospho AKT, total TSC2, and actin. Knockdown of TSC2 blocks the MEK inhibition induced increase in AKT phosphorylation.

## **Key research accomplishments**

Aim 1. We developed a quantitative RPPA to measure dynamics of signaling network in response to EGF

Aim2. We developed a computational model and integrated the experimental data.

Aim 3. We predicted and tested combinations of targets based on underlying experimental data.

## **Reportable outcomes**

The grant has help support research and personnel that has resulted in 5 publications (see references). The data from the grant has also allowed me to obtain a NIH-R01 grant (R01 CA125109). The results from the work were also given during a talk and poster presentation at the 2008 ERA of HOPE meeting in Baltimore, MD.

## **Conclusions**

We have successfully completed the objectives of the grant. We have learnt that targeted manipulation of the signaling network can lead to unforeseen changes elsewhere in the network, therefore we need to understand what these other changes are and determine optimal combinations to kill breast cancer cells.

## **Bibliography of all publications resulting from this grant**

1. Ruths D, Nakhleh L, Iyengar MS, Reddy S, Ram PT Graph-theoretic hypothesis generation in signaling networks *J Comp Biol* 2006 13:1546-1557
2. Ruths D, Muller M, Tseng J-T, Nakhleh L, Ram PT The signaling Petri Net-based simulator: A non-parametric strategy for characterizing the dynamics of cell-specific signaling networks *PLoS Comp Biol* 2008 Feb 29;4(2):e1000005
3. Ruths D, Nakhleh L, Ram PT Rapidly Exploring Structural and Dynamic Properties of Signaling Networks Using PathwayOracle *BMC Systems Biology* 2008 19 2:76
4. Komurov K, White M, Ram PT Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data *PLoS Comp Biol* 2010 In press
5. Iadevaia S, Lu Y, Morales F, Mills GB, Ram PT Identification of optimal drug combinations targeting cellular networks: integrating phospho-proteomics and computational network analysis *Cancer Research* 2010 In press

## **List of personnel supported by this grant**

Dr. Prahlad T. Ram PhD - PI

Dr. Melissa Muller PhD – Post-doctoral fellow

Dr. Kakajan Komurov PhD – Post-doctoral fellow

Jen-Te Tseng MS – Graduate Student

Bettye Cox BS – Research technician

## **Degrees obtained**

Jen-Te Tseng obtained his MS degree while being funded by this grant

## **Appendix**

The 5 papers listed above are attached as appendix

# Graph-theoretic Hypothesis Generation in Biological Signaling Networks

Derek A. Ruths<sup>1</sup> Luay Nakhleh<sup>1</sup> M. Sriram Iyengar<sup>2</sup>  
Shrikanth A. G. Reddy<sup>3</sup> Prahlad T. Ram<sup>3</sup>

<sup>1</sup> Department of Computer Science, Rice University, Houston, TX 77005, USA  
{druths, nakhleh}@rice.edu

<sup>2</sup> UT School of Health Information Sciences, Houston, TX 77030, USA  
msriram@uth.tmc.edu

<sup>3</sup> UT M.D. Anderson Cancer Center, Houston, TX 77030, USA  
{pram, sareddy}@mdanderson.org

**Abstract.** Biological signaling networks comprise the chemical processes by which cells detect and respond to changes in their environment. Such networks have been implicated in the regulation of important cellular activities including cellular reproduction, mobility, and death. Though technological and scientific advances have facilitated the rapid accumulation of information about signaling networks, utilizing these massive information resources has become infeasible except through computational methods and computer-based tools. To date, visualization and simulation tools have received significant emphasis. In this paper, we present a graph-theoretic formalization of biological signaling network models that are in wide but informal use, and formulate two problems on the graph: the *Constrained Downstream* and *Minimum Knockout* Problems. Solutions to these problems yield qualitative tools for generating hypotheses about the networks, which can then be experimentally tested in a laboratory setting. Using established graph algorithms, we provide a solution to the Constrained Downstream Problem. We also show that the Minimum Knockout Problem is NP-Hard, propose a heuristic, and assess its performance. In tests on the Epidermal Growth Factor Receptor (EGFR) network, we find that our heuristic reports the correct solution to the problem in seconds. Source code for the implementations of both solutions is available from the authors upon request.

## 1 Introduction

In this paper, we use the term *biological networks* to refer to cell signaling networks, chains of reactions involved in triggering, propagating, and processing signals within the cell. These networks regulate many cellular activities that are critical to the health of the cell and the larger systems to which it may belong. Altered biological networks have been implicated as the cause of many devastating diseases including cancer [15], heart disease [11], congenital abnormalities [9], metabolic disorders [15], and immunological abnormalities [15].

Significant research efforts to identify and map biological networks, aided by new technologies and scientific methods, have amassed vast databases of molecules and putative interactions among them. Given the immense scale of networks now in common use, computational techniques to filter, search, and reason about them have become indispensable.

Existing research on computational tools in this area has focused on two forms of analysis: visualization of the networks [17, 13, 22, 10], and detailed simulations of small subnetworks based on initial conditions, reaction rates, and other molecule and reaction-specific parameters [19, 12, 24, 21]. Because of the difficulty of determining these parameters, higher-level models are used whenever possible. Recent efforts have also developed hypothesis generation techniques that use model checking and formal verification in order to qualitatively reason about networks [6, 5, 26, 8, 7, 27]. Hypothesis testing tools establish the set of most-likely outcomes of an experiment, providing insights into experimental design, thereby reducing the investment of time and labor-intensive laboratory work. Existing hypothesis generation tools require statements about the properties of individual reactions in networks, details that are often unavailable for many networks. In this paper, we present a framework for computational hypothesis testing that only depends on the simplest property of a reaction - its reactants and products. Our framework combines currently-used graph-based network representations with graph algorithms. We also formalize two biologically significant problems useful for hypothesis testing.

The *Constrained Downstream Problem* seeks the set of reactions in a biological network that leads from one set of molecules to another, such that the set is constrained to include reactions from a given set and exclude reactions from another given set. This is a useful tool in the design of drugs to modify or inhibit certain biological functions while preserving others. At the signaling network level this would help to identify molecules or sets of molecules that have to be targeted to inhibit function of a sub-network while preserving signal flow to a different sub-network. A biological endpoint for this type of problem would be if one wanted to identify a molecule (or a set of molecules) to inhibit proliferation while at the same time preserving metabolic or secretory functions. We provide a polynomial-time algorithm for solving this problem.

The *Minimum Knockout Problem* seeks a minimum-size set of molecules whose removal (or *knocking out*) from the biological network makes the production of a set of molecules impossible given an initial set of molecules. The minimum knockout problem is very important in the identification of molecular targets for therapies, especially in cancer. Traditional chemotherapeutics function by killing rapidly dividing cells, the end result being both cancer cells as well as normal cells are killed, hence the hair loss

and gastro-intestinal side effects of these drugs. In the past few years there has been a great effort in developing drugs that specifically target signaling molecules that are aberrantly functioning in cancer cells. The clinical trials and data from these drugs show that they are limited in their ability, and function best in combination with other targeted drugs. Therefore, the biological problem here is to identify the optimal and minimal sets of molecules that have to be targeted to block network function. This will allow the development of therapeutics that can efficiently kill cancer cells while still preserving normal cells.

The rest of the paper is organized as follows. In Section 2 we introduce a graph formalization of biological networks. In Section 3 we formulate the Constrained Downstream problem, and present a polynomial-time solution of it. In Section 4 we formulate the Minimum Knockout problem, prove its NP-hardness, and devise a randomized heuristic for solving it. In Section 5 we analyze the accuracy and performance of the proposed heuristic for the Minimum Knockout Problem on a large biological network. In Section 6 we conclude and outline future research directions.

## 2 Biological Networks

Standard models of biological networks encompass various molecules and interactions among them that occur on and within the cell membrane. An example of such a model is given in Fig. 1. These models consist of instances of two fundamental components: (1) a *molecule*, either inorganic (such as oxygen,  $O_2$ ), or organic such as proteins, segments of DNA, RNA, or even *complexes* consisting of one or more molecules attached to one another; and (2) an *interaction*, which is a change that occurs to one or more molecules. A change to a molecule will either change a property of the molecule (activity and/or localization), bind one or more molecules together, or break one or more molecules apart.

A salient feature of biological networks is the common occurrence of *feedback loops* (Fig. 1) in which a molecule  $a$ , through a series of interactions, gives rise to another molecule  $b$  that directly changes properties of molecule  $a$  through interactions. A *negative feedback loop* is a chain of interactions which decreases the activity of molecule  $a$  in the network; a *positive feedback loop* increases the activity of molecule  $a$  in the network.

### 2.1 Existing Representations and Models of Biological Networks

Standard models of biological networks have served as the basis for a number of computational models allowing simulating and reasoning about cell processes. We briefly review some of these models in this section.

The Systems Biology Markup Language (SBML) is an XML schema for representing biological networks in addition to regulatory and metabolic networks [1]. The model uses chemical nomenclature: molecules are called *species* and interactions are called *reactions*. In the model, a reaction has reactant, product, and modifier species. Reactants and modifiers are both inputs to the reaction, differing only in that a modifier affects the reaction without undergoing any changes. The SBML file format and underlying network structure is well-accepted and supported by a large number of tools [1].

Petri Nets are considered hybrid models of networks because they model both the global topology of the network as well as some of the reaction-specific parameters that

determine the quantitative behavior of the system. They have been used with varying success to simulate biological networks [28]. In translating the biological network into a Petri Net, each molecule and each interaction is given its own node. Each molecule is initialized with some number of ‘tokens’ which are then iteratively reallocated by the interactions to which they are connected. For a more detailed discussion of PetriNets, see [23]; for an example application see [28].

Differential and algebraic models attempt to simulate the quantitative characteristics of a network using mathematical formulae as well as constants and parameters that have been determined for each reaction in the network [3, 4, 14, 25, 19, 12, 24, 21]. While undisputedly the most accurate of available techniques, these methods are not yet able to simulate large networks efficiently and accurately.

Model checking and formal verification techniques use logical models of networks to make qualitative assertions about their temporal properties (i.e., whether a certain reaction will ever take place under certain conditions). These tools have received significant attention due to their ability to support rapid qualitative hypothesis generation without requiring significant information specific to the networks of interest [6, 5, 26, 8, 7, 27]. In contrast to differential, algebraic, and Petri Net models which require numerical parameters, logical models require qualitative properties of individual reactions. Further, the temporal logic mechanisms that these approaches use are limited in their expressive power vis-à-vis general querying of biological networks.

The rapid pace of lab-based research on biological networks forces biologists to deal with large numbers of reactions and molecules. The lack of much quantitative or qualitative data for these reactions and molecules, coupled with the complexity of questions that biologists would ask about the networks, significantly hinder the applicability and appropriateness of the techniques described above. There is a significant need for tools that provide hypothesis generation capabilities in the absence of detailed network information. For many known networks, the only experimentally confirmed detail is the existence of reactions and the identities of their reactants and products. To the best of our knowledge, hypothesis generation tools that operate on this information alone are lacking. In this paper, we propose a model and approach for such tools and provide a working implementation of two problems useful for hypothesis generation.

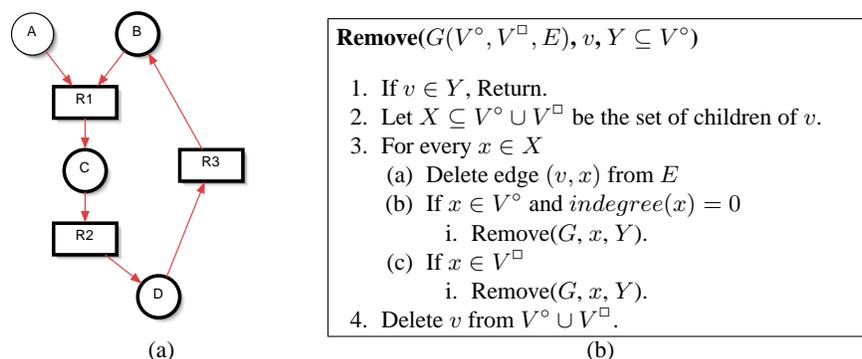
## 2.2 A Graph Formulation of Biological Networks

Numerous tools exist that visualize networks as graphs in which molecules and interactions appear as interconnected nodes [17, 13, 22, 10]. Here, we formalize this graph representation in order to provide a model on which hypothesis testing problems can be posed, analyzed and solved. In this section we introduce this graph-theoretic model, which we call the *Pathway Graph*.

**Definition 1.** A **pathway Graph** is a directed graph,  $G = (V^\circ, V^\square, E)$ , with two types of nodes: **molecule-nodes**,  $V^\circ$ , and **interaction-nodes**,  $V^\square$ , with the following properties:

1.  $V^\circ \cap V^\square = \emptyset$  and
2. For every  $(u, v) \in E$ ,  $u$  and  $v$  are not of the same type.

Property (1) in Definition 1 implies that each node in the graph is either a molecule-node or an interaction-node. Property (2) reflects that the fact that, biologically, a molecule



**Fig. 1.** (a) A diagram of a pathway graph model of a biological network. Circles represent molecules and rectangles represent interactions. (b) The *Remove* procedure for removing a node from a pathway graph and propagating its effect.

cannot directly produce another molecule except through an interaction, nor can a reaction lead to another reaction except through a molecule.

The effect of the “removal” of molecule-nodes from the pathway graph is of significant interest to researchers because it models the effect of drugs that inhibit sections of the network. In particular, they are interested in the connectivity of the graph resulting from the removal of those nodes. Biologically, the effect of removing a node  $v$  in a pathway graph usually propagates further to other nodes reachable from  $v$ : interactions involving the removed molecule can no longer occur, the products of those interactions are no longer produced, and so on. Procedure *Remove* in Fig. 1(b) is a formal description of the “propagation effect” of the removal of a node  $v$  in a pathway graph. An additional set of nodes,  $Y$ , is also specified to indicate the nodes at which to terminate the propagation. The *Remove* procedure can be extended to apply to a set of nodes in a straightforward manner: *Remove*( $G(V^\circ, V^\square, E), X, Y$ ). In this case, *Remove* is applied successively to the nodes in  $X$  (this application yields the same result regardless of the order of nodes).

### 3 The Constrained Downstream Problem

A critical piece of information necessary to predict the outcome of biological network experiments is the set of molecules and interactions dependent on a given set of molecules and/or interactions. In the pathway graph model, all elements in the graph that are dependent on (*downstream from*) a set of molecules and interactions are reachable from that set.

Because of feedback loops in the network, often the set of downstream nodes will contain a significant portion of the network, sometimes the entire network. In order to reduce the number of downstream nodes returned, a biologist may choose to apply certain constraints to the downstream node search. Constraints restrict the solution to the set of nodes belonging to a path from nodes in set  $S$  to nodes in set  $T$  that includes one or more nodes contained in set  $I$  and not containing any nodes in set  $X$ . These nodes belong to a subset of all possible downstream nodes. This is a useful tool in the design of drugs to modify or inhibit certain biological functions while preserving

others. At the signaling network level this would help to identify molecules or sets of molecules that have to be targeted to inhibit function of a sub-network while preserving signal flow to a different sub-network.

**Problem 1. THE CONSTRAINED DOWNSTREAM PROBLEM**

**Input:** Pathway graph  $G = (V^\circ, V^\square, E)$  and four sets  $S, T \subset V^\circ$  and  $I, X \subset (V^\circ \cup V^\square)$ .

**Output:** Subgraph  $G' = (U^\circ, U^\square, E')$  where

1.  $U^\circ \subseteq V^\circ, U^\square \subseteq V^\square$ , and  $E' \subseteq E$ ;
2.  $\forall u \in (U^\circ \cup U^\square), \exists [s \in S, t \in T]$  such that  $s \xrightarrow{G'} u$  and  $u \xrightarrow{G'} t$ <sup>4</sup>;
3.  $(U^\circ \cup U^\square) \cap X = \emptyset$ ;
4. every path from a node in  $S$  to a node in  $T$  passes through a node in  $I$ ; and
5.  $G'$  is the maximum subgraph that satisfies conditions 1–4.

$G' = \mathbf{FindDownstream}(G = (V^\circ, V^\square, E), S, T, I, X)$

1.  $\forall t \in T, Visited[t] = 1; \forall t \notin T, Visited[t] = 0$ ;
2.  $\forall t \in T, OnPath[t] = 1; \forall t \notin T, OnPath[t] = 0$ ;
3.  $\forall v \in V^\circ \cup V^\square, AboveInclude[v] = 0$ ;
4.  $G' = (\emptyset, \emptyset, \emptyset)$ ;
5. For every  $s \in S$   
     $CalcDownstream(G, s, I, \emptyset, G')$ .
6. Return  $G'$ ;

$\mathbf{CalcDownstream}(G, v, I, P, G')$

1. If  $Visited[v] == 0$ 
  - (a)  $Visited[v] = 1$ ;
  - (b) Let  $C$  be the children of  $v$ ;
  - (c) For every  $c \in C$   
     $CalcDownstream(G, c, (p_1, \dots, p_k, v))$ ;
2. Else
  - (a) If  $OnPath[v] == 0$   
    Return;
  - (b) If  $AboveInclude[v] == 1$ 
    - i.  $V_{G'} = V_{G'} \cup P$ ;
    - ii.  $E_{G'} = E_{G'} \cup \{(p_1, p_2), \dots, (p_{k-1}, p_k)\}$ ;
    - iii.  $\forall p \in P, AboveInclude[p] = 1$ ;
  - (c) Else If  $P \cap I \neq \emptyset$ 
    - i.  $V_{G'} = V_{G'} \cup P$ ;
    - ii.  $E_{G'} = E_{G'} \cup \{(p_1, p_2), \dots, (p_{k-1}, p_k)\}$ ;
    - iii.  $\forall p \in P, AboveInclude[p] = 1$ ;
  - (d) Else, Return;

**Fig. 2.** The algorithm for the Constrained Downstream Problem.

The algorithm in Fig. 2 solves the Constrained Downstream Problem with time complexity  $O(|V^\circ \cup V^\square|)$ . Analysis of the running time as well as proof of the correctness of the algorithms are omitted due to space constraints.

<sup>4</sup> We write  $x \xrightarrow{G} y$  to denote that node  $y$  is reachable from node  $x$  in graph  $G$ .

## 4 The Minimum Knockout Problem

A problem of significant interest to experimental biologists researching networks implicated in disease is the minimum knockout problem. In this problem, for a given pathway graph, a minimal set of nodes is sought such that the removal of these nodes disconnects a given set of (source) molecules,  $S \subset V^\circ$ , from another given set of (target) molecules,  $T \subset V^\circ$ . The minimum knockout problem is very important in the identification of molecular targets for therapies, especially in cancer. Traditional chemotherapeutics function by killing rapidly dividing cells, the end result being both cancer cells as well as normal cells are killed, hence the hair loss and gastro-intestinal side effects of these drugs. Therefore, the biological problem here is to identify the optimal and minimal sets of molecules that have to be targeted to block network function. Formally, we define the problem (decision version) as follows.

*Problem 2. THE MINIMUM KNOCKOUT PROBLEM (MKO)*

**Input:** Pathway graph  $G = (V^\circ, V^\square, E)$ , two sets of nodes  $S, T \subset V^\circ$ , and a positive integer  $Q$ .

**Question:** Does there exist a set  $U \subseteq ((V^\circ \cup V^\square) - (S \cup T))$  with  $|U| \leq Q$  such that  $\text{Remove}(G, U, S \cup T)$  yields graph  $G' = (V'^\circ, V'^\square, E')$  in which for every  $s \in S$  and  $t \in T$ ,  $s \not\rightsquigarrow t$ ?

We first prove that MKO is NP-Hard, and then present an efficient and accurate randomized heuristic for solving it. We prove the NP-hardness of the problem by a reduction from the Minimum Set Cover Problem [16].

*Problem 3. THE MINIMUM SET COVER PROBLEM (MSC)*

**Instance:** Collection  $C$  of subsets of a finite set  $B$  and a positive integer  $K \leq |C|$ .

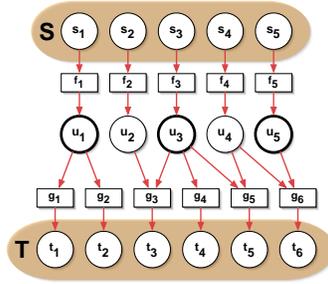
**Question:** Does  $C$  contain a cover for  $B$  of size  $K$  or less, i.e., a subset  $C' \subseteq C$  with  $|C'| \leq K$  such that every element of  $B$  belongs to at least one member of  $C'$ ?

**Theorem 1.** *MKO is NP-Hard.*

*Proof.* Given an instance  $\langle B = \{b_1, \dots, b_m\}, C = \{C_1, \dots, C_n\}, K \rangle$  of MSC, we construct an instance  $\langle G, S, T, Q \rangle$  of MKO as follows.

- Pathway graph  $G = (V^\circ, V^\square, E)$  where
  - $V^\circ = \{s_i, u_i : C_i \in C\} \cup \{t_i : b_i \in B\}$ .
  - $V^\square = \{f_i : C_i \in C\} \cup \{g_i : b_i \in B\}$ .
  - $E = \{(s_i, f_i) : 1 \leq i \leq n\} \cup \{(f_i, u_i) : 1 \leq i \leq n\} \cup \{(u_i, g_j) : b_j \in C_i\} \cup \{(g_i, t_i) : 1 \leq i \leq m\}$ .
- $S = \{s_i \in V^\circ\}$ .
- $T = \{t_i \in V^\circ\}$ .
- $Q = K$ .

Fig. 3 gives an example of the construction. The graph  $G$  constructed by the reduction satisfies the conditions of Definition 1, and hence it is a pathway graph. We now establish the validity of the reduction by showing that  $\langle B, C, K \rangle$  is a yes-instance of MSC if and only if  $\langle G, S, T, Q \rangle$  is a yes-instance of MKO.



**Fig. 3.** The  $G$ ,  $S$ , and  $T$  components of the MKO instance constructed by the reduction in the proof of Theorem 1 for the MSC instance with  $B = \{b_1, b_2, b_3, b_4, b_5, b_6\}$ ,  $C = \{\{b_1, b_2\}, \{b_3\}, \{b_3, b_4, b_5\}, \{b_5, b_6\}, \{b_6\}\}$ , and  $K = 3$ ;  $Q = 3$ .

$\Rightarrow$  Let  $C' \subseteq C$  with  $|C'| \leq K$  be a cover for  $B$ . Then, by construction, every node in the set  $Y = \{g_i \in V^\square\}$  has an incoming edge from a node in the set  $X = \{u_i : C_i \in C'\}$ . Since  $Y$  contains only interaction nodes, applying Remove (Fig. 1(b)) to all nodes in  $X$  will disconnect all paths from nodes in  $S$  to nodes in  $T$ . Since  $|X| = |C'| \leq K$  and  $Q = K$ , it follows that  $\langle G, S, T, Q \rangle$  is a yes-instance of MKO.

$\Leftarrow$  Assume there does not exist a cover of size  $K$  or less for  $B$ . Then, for every set  $C' \subseteq C$  with  $|C'| \leq K$ , there is at least one  $b' \in B$  such that  $b' \notin \cup_{c \in C'} c$ . By construction of  $G$ , it follows that for any subset of  $X = \{u_i : C_i \in C'\}$  of size  $Q$  or less, there exists at least one node in  $Y = \{g_i \in V^\square\}$  that is not a child of any node in  $X$ . Hence, removing all nodes in  $X$  will not disconnect all paths from  $S$  to  $T$ . Since every node in  $T$  has a unique parent in  $Y$ , it follows that there does not exist a set of nodes of size  $Q$  or less that disconnects all paths from nodes in  $S$  to nodes in  $T$ . Hence,  $\langle G, S, T, Q \rangle$  is not a yes-instance of MKO. This finishes the proof, thus establishing that MKO is NP-hard.

#### 4.1 An Efficient and Accurate Randomized Heuristic for MKO

We now give an efficient and accurate heuristic for solving MKO; the heuristic is an iterative randomized search, with running time  $O(nmk)$ , where  $n$  is the number of nodes in the input pathway graph,  $m$  is the number of nodes in the constrained downstream subgraph, and  $k$  is the number of iterations. In the worst-case scenario,  $m = n$ ; however, in our experiments on a large pathway graph, we found that  $k, m \ll n$ . The heuristic is outlined in Fig. 4, and makes use of the following lemma.

**Lemma 1.** *Let  $U$  be a minimum knockout set for  $S$  and  $T$  in graph  $G_d = \text{Downstream}(G, S, T)$ , where  $G_d = (V_d^\circ, V_d^\square, E_d)$ . Then, there exists a minimum knockout set  $U'$  such that*

1.  $|U'| = |U|$  and
2.  $U \subseteq (V_d^\circ \cup (\text{Children}(S) \cap V_d^\square))$ .

The formal proof is omitted due to space constraints. Intuitively, this lemma states that if node  $v \in V^\square$  is an element of a solution to MKO, then  $v$  can be replaced by some  $v' \in V^\circ$  where  $(v', v)$  is an edge in the graph. The validity of this lemma follows from the definition of the *Remove* procedure (Fig. 1). The only  $V^\square$  nodes that cannot

<p><b>MinKnockout</b>(<math>G = (V^\circ, V^\square, E), S, T, m</math>)</p> <ol style="list-style-type: none"> <li>1. <math>U = FindDownstream(G, S, T, \emptyset, \emptyset)</math></li> <li>2. <math>U^\circ = U \cap V^\circ</math></li> <li>3. <math>U^\square = U \cap V^\square</math></li> <li>4. <math>C = \{u \in U : u \in U^\circ \vee u \in (Children(S) \cap U^\square)\}</math></li> <li>5. For <math>i = 1</math> to <math>m</math> <ol style="list-style-type: none"> <li>(a) <math>G' = G</math></li> <li>(b) <math>S_i = \emptyset</math></li> <li>(c) While <math>S \stackrel{G'}{\rightsquigarrow} T</math> <ol style="list-style-type: none"> <li>i. <math>c \in (C - S_i)</math></li> <li>ii. <math>G' = Remove(G', c, S, T)</math></li> <li>iii. <math>S_i = S_i \cup \{c\}</math></li> </ol> </li> </ol> </li> <li>6. <math>j = argmax_i  S_i </math></li> <li>7. Return <math>S_j</math></li> </ol>
---

**Fig. 4.** An iterative and randomized heuristic for MKO.

be replaced in this manner are the children of  $S$  (since elements of  $S$  cannot appear in the solution).

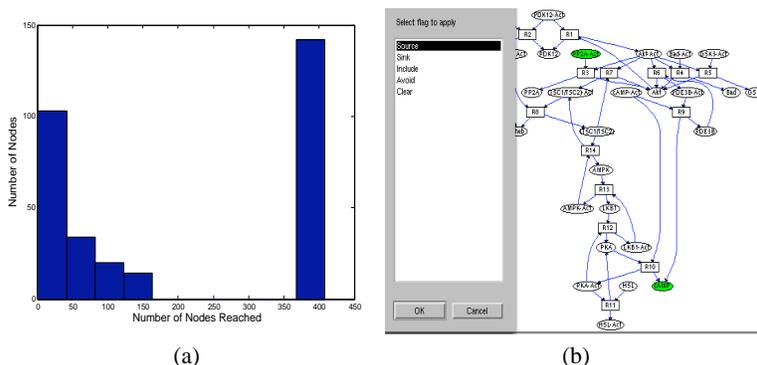
The intuition for the heuristic is to exhaustively search a small (relative to the size of the actual pathway graph) set of nodes for the smallest knockout set for  $S$  and  $T$ . By constructing the set of nodes so that it does contain a knockout set (though not necessarily a globally minimal knockout set), the algorithm is guaranteed to find a solution, though it may not be minimal.

Before we describe our heuristic, we review background material that will be used in the heuristic. Given a directed graph  $G = (V, E)$ , and two sets  $S, T \subseteq V$ , a path in  $G$  is an  $S$ – $T$  path if it runs from a node in  $S$  to a node in  $T$ . A set  $C \subseteq V$  is called  $S$ – $T$  disconnecting if  $C$  intersects each  $S$ – $T$  path ( $C$  may intersect  $S \cup T$ ).

**Theorem 2.** (Menger’s Theorem [20]) *Let  $G = (V, E)$  be a directed graph and let  $S, T \subseteq V$ . Then, the maximum number of node-disjoint  $S$ – $T$  paths is equal to the minimum size of an  $S$ – $T$  disconnecting node set.*

Now, we are in position to describe our heuristic. We construct the search set,  $C$  (line 4), to have the properties of containing a knockout set and being small relative to the number of nodes in the entire pathway graph as follows.

1.  $C = FindDownstream(G, S, T, \emptyset, \emptyset)$ . By Menger’s theorem, a knockout set is contained in the set of nodes that comprise all paths connecting  $S$  and  $T$  because one such knockout set is a node from each disjoint path connecting  $S$  to  $T$ . Since the Constrained Downstream Problem constructs this set,  $C$  contains a knockout set. In addition, the constrained set of downstream nodes for most choices of  $S$  and  $T$  will contain far fewer nodes than the entire pathway graph.
2.  $C = (C \cap V^\circ) \cup (C \cap Children(S))$ . Lemma 1 states that, except for the nodes with elements of  $S$  as inputs, any  $V^\square$  node that occurs in a minimum knockout set can be replaced by a single  $V^\circ$  node. Thus, by searching only the  $V^\circ$  nodes between  $S$  and  $T$ , the search set is further reduced in size and still contains a knockout set.



**Fig. 5.** (a) The distribution of nodes in the EGFR pathway graph by the total number of nodes in the pathway graph they can reach. (b) A screen shot from the implementation of the model and algorithms in this paper.

After constructing the search set  $C$  (lines 1–4), the algorithm performs  $m$  randomized searches over all nodes in set  $C$  for a knockout set. Within each search (loop on line 5c), a knockout set is iteratively constructed by removing a randomly selected node (line 5(c)i) from the graph until  $S$  and  $T$  are disconnected. Of the  $m$  knockout sets constructed, the knockout set with fewest members is returned.

While the heuristic does not guarantee an upper-bound on error, our experiments show that, for the Epidermal Growth Factor Receptor (EGFR) biological network [22], this heuristic finds a minimum-knockout set every time. We discuss the experiments and performance in more detail next in Section 5.

## 5 Experimental Results and Discussion

We studied the performance of the heuristic for the Minimum Knockout Problem on the biological data set published in [22]. In this work, the authors constructed a comprehensive epidermal growth factor receptor (EGFR) signaling network. This network is known to have a significant role in cancer development and proliferation. Given the amount of research currently focused on this network, benchmarks for our heuristic on this network will likely give a very accurate sense of how well the heuristic will perform.

The EGFR network contains 292 interactions involving 330 different molecules. The graph of this network is highly connected as shown in Fig. 5(b). Nearly half of the molecules reach between 350 and 400 nodes in the graph.

To test the heuristic, we manually selected 30 pairs of  $S$  and  $T$  node sets from the network. The only selection criteria applied was a rough attempt to choose  $S$  and  $T$  so that the nodes in opposite sets were far from one another, increasing the likelihood of non-trivial solutions. Beyond this, the nodes were selected at random. Sets varied in size from 1 to 10.

*Heuristic Accuracy:* The heuristic was set to run for 100 iterations on each of the  $(S, T)$  pairs. In every case, the heuristic reported a minimum knockout set of size 1. Since the smallest possible minimum knockout set has cardinality equal to 1, we con-

cluded that every time the heuristic correctly identified a minimum knockout set. This result is remarkable for two reasons.

(1) A minimum knockout set of 1 occurs with unexpected frequency. This result is best explained by the degree of connectivity in the network. Fig. 5(a) shows that over half of the nodes in the network have very extensive connectivity within the graph. This is consistent with other studies of connectivity within biological networks [18, 2, 29]. Because of the properties of the *Remove* operation, removing such a highly connected node from the graph will have global impact on the connectivity of other nodes.

(2) The heuristic correctly found a minimum knockout set every time. This is certainly a property of the network: if one chooses a molecule at random in the network, there is a 50% chance that it will connect to 400 other nodes in the network. Ultimately, while it is easy to envision cases that will be difficult for the heuristic to handle, our results indicate that the EGFR network, well-studied and important in research, has few difficult cases, if any.

We observed that, though correct, often the heuristic chose nodes which disconnected large sections of the graph from  $S$ . Biologically, it is favorable to target nodes that have the smallest global impact while still disconnecting  $S$  and  $T$ . While, preliminary analysis was not able to establish whether the disconnection of these nodes was necessary, we consider the problem of identifying the minimum knockout set that also minimizes the number of non- $T$  nodes disconnected from  $S$  to be an important extension to this problem.

*Heuristic Performance:* We implemented the heuristic shown in Fig. 4 in Java. A set of 100 iterations of the algorithm took approximately one second to complete on a Apple 1.33 GHz G4 laptop running Mac OS X.

A tool implementing the model and algorithms described in this paper is available for download. The model and algorithms are implemented in Matlab; for better performance, the interface is implemented in Java. Fig. 5(b) shows a screenshot of the tool in use on a small model of the AKT network. The tool can load networks stored in the SBML format, allowing biologists to import networks designed in CellDesigner and other biological network editors [13].

## 6 Conclusions and Future Work

In this paper we have presented a formal graph model that permits the use of graph theory to reason about the properties of biological networks. In addition, we have characterized two important research questions pertaining to biological networks, formulated them on our model, and provided efficient and accurate algorithms for solving them. To our knowledge, this is the first paper to formally define and propose a computational solution to the Minimum Knockout Problem. Despite being NP-Hard, our heuristic shows excellent performance on a large and important network in the research community.

Moving forward, we recognize that a useful addition to the current heuristic for the Minimum Knockout Problem is the ability to return a set of minimum knockout sets rather than just a single one. Furthermore, we intend to consider additional biological constraints, such as selecting the minimum knockout set with the least impact to global connectivity of the graph. There is also work to be done in studying other existing and new problems under the pathway graph model.

## References

- [1] Systems biology markup language webpage. <http://www.sbml.org>, 2005.
- [2] A. L. Barabasi and Z. N. Oltvai. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, 5(2):101–113, 2004.
- [3] U. S. Bhalla and R. Iyengar. Emergent properties of networks of biological signaling pathways. *Science*, 283(5400):381–387, 1999.
- [4] U. S. Bhalla, P. T. Ram, and R. Iyengar. MAP Kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science*, 297(5583):1018–1023, 2002.
- [5] N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages, and V. Schachter. Modeling and querying biomolecular interaction networks. *Theoretical Computer Science*, 325(1):25–44, 2004.
- [6] N. Chabrier-Rivier and F. Fages. Symbolic model checking of biochemical networks. In C. Priami, editor, *CMSB'03: Proc. 1st Workshop on Computational Methods in Systems Biology*, volume 2602 of *Lecture Notes in Computer Science*, pages 149–162. Springer, 2003.
- [7] N. Chabrier-Rivier, F. Fages, and S. Soliman. The biochemical abstract machine biocham. In V. Danos and V. Schachter, editors, *CMSB'04: Proc. 2nd Workshop on Computational Methods in Systems Biology*, Lecture Notes in Computer Science. Springer-Verlag, 2004.
- [8] S. Eker, M. Knapp, K. Laderoute, P. Lincoln, and C. Talcott. Pathway logic: Executable models of biological networks. *Theoretical Computer Science*, 71(eker-etal-02wrla), 2002.
- [9] E. Belloni et al. Identification of sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nature Genetics*, 14:353–356, 1996.
- [10] M. I. Aladjem et al. Molecular interaction maps – a diagrammatic graphical language for bioregulatory networks. *Sci. STKE*, pe8, 2004.
- [11] D. S. Feldman, C. A. Carnes, W. T. Abraham, and M. R. Bristow. Mechanisms of disease:  $\beta$ -adrenergic receptors alterations in signal transduction and pharmacogenomics in heart failure. *Nature Clinical Practice Cardiovascular Medicine*, 2:475–483, 2005.
- [12] C. Fu, Z. Qi, and J. You. A bioambients based framework for chain-structured biomolecules modelling. In J. Zhang, J. He, and Y. Fu, editors, *CIS'04: Proc. 1st International Symposium on Computational and Information Science*, volume 3314 of *Lecture Notes in Computer Science*. Springer-Verlag, 2004.
- [13] A. Funahashi, M. Morohashi, H. Kitano, and N. Tanimura. Celldesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1(5):159–162, 2003.
- [14] A. Hoffmann, A. Levchenko, M. L. Scott, and D. Baltimore. The ikappab-nf-kappab signaling module: temporal control and selective gene activation. *Science*, 298(5596):1241–1245, 2002.
- [15] T. Hunter. Signaling – 2000 and beyond. *Cell*, 100(1):113–127, 2000.
- [16] R. M. Karp. Reducibility among combinatorial problems. In R. E. Miller and J. W. Thatcher, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972.
- [17] H. Kitano, A. Funahashi, Y. Matsuoka, and K. Oda. Using process diagrams for the graphical representation of biological networks. *Nature Biotechnology*, 23(9):961–966, 2005.
- [18] A. Ma'ayan, S. L. Jenkins, S. Neves, A. Hasseldine, E. Grace, B. Dubin-Thaler, N. J. Eungdamrong, G. Weng, P. T. Ram, J. J. Rice, A. Kershenbaum, G. A. Stolovitzky, R. D. Blitzer, and R. Iyengar. Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science*, 309(5737):1078–1083, 2005.
- [19] T. C. Meng, S. Somani, and P. Dhar. Modeling and simulation of biological systems with stochasticity. *Silico Biology*, 4(2):293–309, 2004.

- [20] K. Menger. Zur allgemeinen kurventheorie. *Fundamenta Mathematicae*, 10:96–115, 1927.
- [21] M. Nagasaki, A. Doi, H. Matsuno, and S. Miyano. A versatile petri net based architecture for modeling and simulation of complex biological processes. *Genome Informatics*, 15(1):180–197, 2004.
- [22] K. Oda, Y. Matsuoka, A. Funahashi, and H. Kitano. A comprehensive pathway map of epidermal growth factor signaling. *Molecular Systems Biology*, msb41000014–E1–E17, 2005.
- [23] J. L. Peterson. *Petri Net Theory and the Modelling of Systems*. Prentice-Hall, 1981.
- [24] A. Phillips and L. Cardelli. A correct abstract machine for the stochastic pi-calculus. In *Transactions on Computational Systems Biology*, 2005.
- [25] A. E. Smith, B. M. Slepchenko, J. C. Schaff, L. M. Loew, and I. G. Macara. Systems analysis of ran transport. *Science*, 295(5554):488–491, 2002.
- [26] M. Sriram. Modelling protein functional domains in signal transduction using maude. *Briefings in Bioinformatics*, 4(3):236–245, 2003.
- [27] N. Tran, C. Baral, V. Nagaraj, and L. Joshi. Knowledge-based interactive framework for hypothesis formation in biochemical networks. In *Proceedings of the Workshop on Data Integration in the Life Sciences*, 2005.
- [28] D. Tsavachidou and M. Liebman. Modeling and simulation of pathways in menopause. *Journal of the American Medical Informatics Association*, 9(5):461–471, 2002.
- [29] S. Wuchty, Z. N. Oltvai, and A. L. Barabasi. Evolutionary conservations of motif constituents in the yeast protein interaction network. *Nat. Genet.*, 35(2):176–179, 2003.

# The Signaling Petri Net-Based Simulator: A Non-Parametric Strategy for Characterizing the Dynamics of Cell-Specific Signaling Networks

Derek Ruths<sup>1\*</sup>, Melissa Muller<sup>2</sup>, Jen-Te Tseng<sup>2</sup>, Luay Nakhleh<sup>1</sup>, Prahlad T. Ram<sup>2</sup>

<sup>1</sup> Department of Computer Science, Rice University, Houston, Texas, United States of America, <sup>2</sup> Department of Systems Biology, University of Texas M. D. Anderson Cancer Center, Houston, Texas, United States of America

## Abstract

Reconstructing cellular signaling networks and understanding how they work are major endeavors in cell biology. The scale and complexity of these networks, however, render their analysis using experimental biology approaches alone very challenging. As a result, computational methods have been developed and combined with experimental biology approaches, producing powerful tools for the analysis of these networks. These computational methods mostly fall on either end of a spectrum of model parameterization. On one end is a class of structural network analysis methods; these typically use the network connectivity alone to generate hypotheses about global properties. On the other end is a class of dynamic network analysis methods; these use, in addition to the connectivity, kinetic parameters of the biochemical reactions to predict the network's dynamic behavior. These predictions provide detailed insights into the properties that determine aspects of the network's structure and behavior. However, the difficulty of obtaining numerical values of kinetic parameters is widely recognized to limit the applicability of this latter class of methods. Several researchers have observed that the connectivity of a network alone can provide significant insights into its dynamics. Motivated by this fundamental observation, we present the signaling Petri net, a non-parametric model of cellular signaling networks, and the signaling Petri net-based simulator, a Petri net execution strategy for characterizing the dynamics of signal flow through a signaling network using token distribution and sampling. The result is a very fast method, which can analyze large-scale networks, and provide insights into the trends of molecules' activity-levels in response to an external stimulus, based solely on the network's connectivity. We have implemented the signaling Petri net-based simulator in the PathwayOracle toolkit, which is publicly available at <http://bioinfo.cs.rice.edu/pathwayoracle>. Using this method, we studied a MAPK1,2 and AKT signaling network downstream from EGFR in two breast tumor cell lines. We analyzed, both experimentally and computationally, the activity level of several molecules in response to a targeted manipulation of TSC2 and mTOR-Raptor. The results from our method agreed with experimental results in greater than 90% of the cases considered, and in those where they did not agree, our approach provided valuable insights into discrepancies between known network connectivities and experimental observations.

**Citation:** Ruths D, Muller M, Tseng J-T, Nakhleh L, Ram PT (2008) The Signaling Petri Net-Based Simulator: A Non-Parametric Strategy for Characterizing the Dynamics of Cell-Specific Signaling Networks. *PLoS Comput Biol* 4(2): e1000005. doi:10.1371/journal.pcbi.1000005

**Editor:** Satoru Miyano, The University of Tokyo, Japan

**Received:** September 18, 2007; **Accepted:** January 18, 2008; **Published:** February 29, 2008

**Copyright:** © 2008 Ruths et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** DR and LN were supported in part by a Seed Grant awarded to LN from the Gulf Coast Center for Computational Cancer Research, funded by John and Ann Doerr Fund for Computational Biomedicine. J-TT was supported in part by a training fellowship from the Pharmacoinformatics Training Program of the Keck Center of the Gulf Coast Consortia (NIH grant 5 T90 DK070109-03), and MM and PTR were supported in part by Department of Defense grant BC044268 to PTR.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: [druths@rice.edu](mailto:druths@rice.edu)

## Introduction

Signaling networks are complex, interdependent cascades of signals that process extracellular stimuli, received at the plasma membrane of a cell, and funnel them to the nucleus, where they enter the gene regulatory system. These signaling networks underlie how cells communicate with one another, and how they make decisions about their phenotypic changes, such as division, differentiation, and death. Further, malfunction of these networks may alter phenotypic changes that cells are supposed to undergo under normal conditions, and potentially lead to devastating consequences on the organism. For example, altered cellular signaling networks can give rise to the oncogenic properties of cancer cells [1,2], increase a person's susceptibility to heart disease [3], and have been shown to be responsible for many other

devastating diseases such as congenital abnormalities, metabolic disorders and immunological abnormalities [1,4].

In light of the crucial role signaling networks play in the proper functioning of cells and biological systems as a whole, and given the grave consequences their alterations may have on the behavior of cells, elucidating the connections in the networks, and understanding how they operate, are two central questions in cell biology. However, unlike the "pathway view" of signaling as linear cascades, signaling networks are highly interconnected, involve cross-talk among several pathways, and contain feedback and feed-forward loops [5]. Figure 1 illustrates this issue in a network of signaling cascades, which is stimulated by EGF and contains several players in cancer pathways. For example, multiple paths lead from EGFR to mTOR-Raptor, resulting in feed-forward loops. Some of these paths activate mTOR-Raptor,

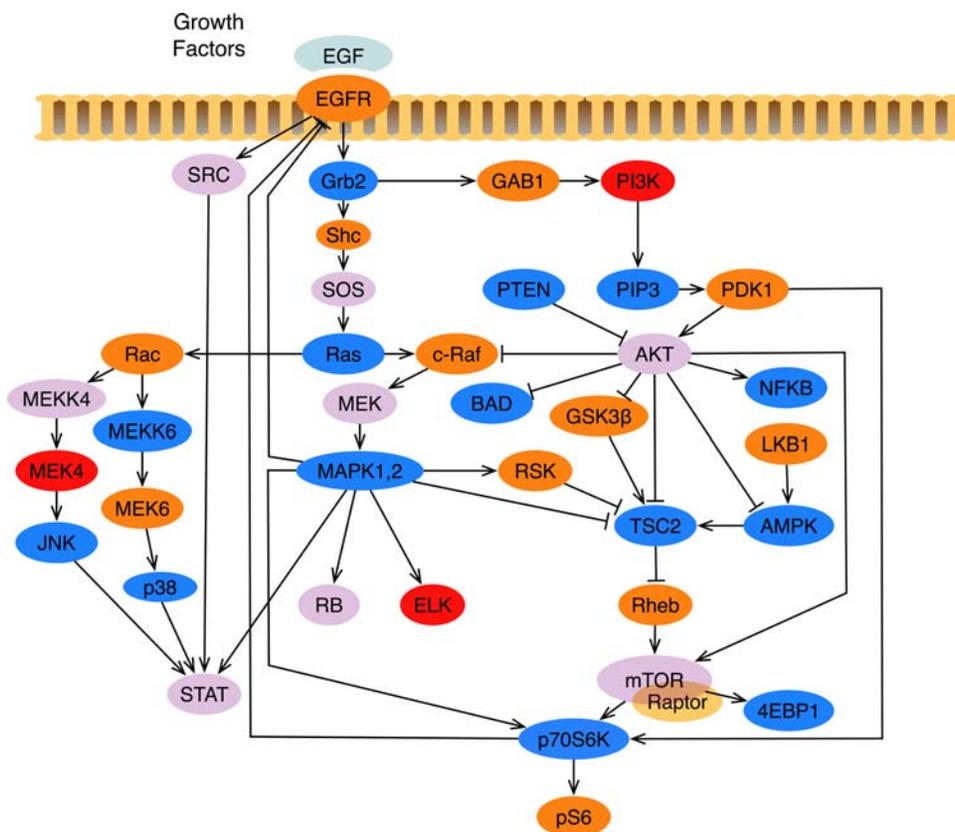
**Author Summary**

Many cellular behaviors including growth, differentiation, and movement are influenced by external stimuli. Such external stimuli are obtained, processed, and carried to the nucleus by the signaling network—a dense network of cellular biochemical reactions. Beyond being interesting for their role in directing cellular behavior, deleterious changes in a cell's signaling network can alter a cell's responses to external stimuli, giving rise to devastating diseases such as cancer. As a result, building accurate mathematical and computational models of cellular signaling networks is a major endeavor in biology. The scale and complexity of these networks render them difficult to analyze by experimental techniques alone, which has led to the development of computational analysis methods. In this paper, we present a novel computational simulation technique that can provide qualitatively accurate predictions of the behavior of a cellular signaling network without requiring detailed knowledge of the signaling network's parameters. Our approach makes use of recent discoveries that network structure alone can determine many aspects of a network's dynamics. When compared against experimental results, our method correctly predicted 90% of the cases considered. In those where it did not agree, our approach provided valuable insights into discrepancies between known network structure and experimental observations.

while others inhibit it. Further, the network contains two feedback loops, one from p70S6K to EGFR and another from MAPK1,2 to EGFR.

These and other complexities make it very difficult to analyze signaling networks by experimental biology approaches alone. As a result, computational methods have been developed and combined with experimental biology approaches, producing powerful tools for the analysis of these networks [6]. These computational methods produce hypotheses that guide the experimental design, leading to more informative experiments, while experimental results help refine the computational models, resulting in more accurate predictive tools.

In a recent survey, Papin et al. classified existing computational methods into two categories: *structural* and *dynamic* network analysis [6]. Structural network analysis is mainly based on the network's connectivity, which is typically readily available from numerous public signaling network databases (e.g., [7–9]), and makes inferences about global network properties as well as individual protein functions. This category can be further refined into two sub-categories, both of which are solely based on connectivity information, yet differ in the type of answers they provide. For example, the methods described in [10–13] infer “static” properties of the network, such as numbers of paths, reachability results, etc. In a series of papers, Palsson and co-workers [6,14–16] introduced extreme pathway analysis techniques, which are more appropriate for metabolic networks, yet have been applied to signaling networks to characterize various properties of networks, such as redundancy and cross-talk. Similar analyses have also been



**Figure 1. The Model Signaling Network.** A MAPK1,2 and AKT network downstream from EGFR, which we assembled from various sources, and used for the case study analysis in this work. An edge from u to v ending with an arrow indicates an activating reaction, while an edge ending with a plunger indicates an inhibiting reaction. With the exception of TSC2, all nodes have self-inhibitory edges, which were added to model the external cellular machinery that regulates the concentration of the active form of the proteins [36–43]. Colors were selected to enhance readability of the network. doi:10.1371/journal.pcbi.1000005.g001

formalized and conducted using the principles of S- and T-invariants in Petri Nets (e.g., [17–20]).

Methods for dynamic network analysis use, in addition to the network connectivity, the kinetic parameters of the biochemical reactions. The goal of these methods is to model the actual kinetics of the network and obtain through simulation the actual quantities of proteins involved in signal transduction. One of the most widely used techniques in this category is systems of ordinary differential equations (ODEs) (e.g., [21–25]). Within such a system, each reaction is modeled by a series of equations connecting reactant concentrations to product concentrations through differential relationships involving reaction rate constants. Given the difficulty of obtaining the numerical values of kinetic parameters [19,26] and standardization of the parameters and models [27], the applicability of these methods is limited in practice to small-scale networks [6,28].

Petri Nets have also been used for simulating the dynamics of signaling networks [29–31]. While such approaches somewhat relax the necessity for biologically exact kinetic parameters, current Petri Net-based approaches still require the selection of weights and/or probability distributions for individual interactions in the model. As a result, selecting the values for Petri Net parameters presents challenges similar to those encountered in ODE modeling.

Structural network analysis assumes mainly connectivity information about the model, and provides insights into global, static properties of the network. Dynamic analysis in general assumes numerical values of the kinetic parameters, and provides predictions of network dynamics by quantifying the change in concentration and activity-level (the concentration of the active form of a given protein) of the individual proteins and complexes in the network. To obtain a more detailed analysis one must either solve parameter optimization problems for a large number of molecules and interactions or conversely experimentally derive these values.

Given the difficulty of obtaining numerical values of kinetic parameters [19,26] and the implications this has on the applicability of dynamic analysis methods [6], it is imperative to develop innovative approaches that combine the attractive low requirements of structural network analysis techniques with the detailed answers provided by dynamic analysis techniques—specifically the response of individual proteins to signals which travel through the network.

Several recent efforts in this direction have produced encouraging results. An approach using a boolean network simulation method, based on work in the area of gene regulatory networks, successfully used only signaling network connectivity information to predict the speed of signal transduction through a stomata signaling network [32]. The use of piecewise linear systems of ODEs have also had success in analyzing some of the dynamics of gene regulatory and signaling networks without using exact kinetic parameters (e.g., [33–35]). The obstacle to extending the method in [32] to model individual protein responses to signal transduction is the boolean model used to discretize the signal as it propagates. In a boolean model, the signal is either present or absent at each node in the network. Such two-state models of signal transduction simplify the underlying biochemistry to the point where it is difficult to model changes in protein concentration more precisely than present or absent. Modeling such gradients of concentration changes and the effects of those changes may be important to predicting individual protein responses, motivating our effort to devise more fine-grained ways to model and simulate the dynamics of signaling networks. The challenges to using linear-piecewise ODEs to model a signaling network center around the issue of identifying all the ODEs required to model the underlying network as well as scalability issues involved in simulating large systems of ODEs.

In this paper, we extend the synchronized Petri net model and firing policy such that the resulting framework models cellular signaling processes. We call this extension the signaling Petri net (SPN). By coupling this with a novel strategy for Petri net execution and sampling, we obtain a method capable of characterizing some dynamics of signaling networks while using only connectivity information about these networks.

To validate our method, we studied the MAPK1,2 and AKT network shown in Figure 1 in two breast cancer cell lines. This network was chosen because the EGFR receptor and its downstream signaling network play a very important role in development, differentiation, and oncogenic transformation. Two very important signaling molecules within the cell are MAPK and AKT, both of which can be activated by EGFR, and contains several potential regulatory paths between them. We constructed a model network of EGF regulation of MAPK and AKT which includes several feedback and feed-forward loops all of which were constructed based on experimental findings from different laboratories around the world [36–43]. We analyzed, both experimentally and computationally, the change in activity-level of several proteins in response to targeted manipulation of TSC2 and mTOR-Raptor. Using the model network, the predictions from our method agreed with experimental results in over 90% of the cases, and in those where they did not agree, our method correctly identified discrepancies that could be traced back to incompleteness in the network connectivity model.

## Materials and Methods

Our approach combines elements of the boolean network simulator in [18] with a synchronized Petri net model [44]. In [18], Li et al. present a non-parametric approach that accurately predicts the speed of signal propagation through a network. However, as their method assumes a binary model of activation—every protein is either active (*true*) or inactive (*false*)—modeling a range of activity-levels is difficult. Petri nets, while able to model concentrations using tokens, require parameters describing the kinetic characteristics of the network, which are typically difficult to obtain.

Our method models signal flow as the pattern of token accumulation and dissipation within places (proteins) over time in the Petri net. Transitions in the network represent directed protein interactions; each transition models the effect of a source protein on a target protein. Through transition firings, the source can influence the number of tokens assigned to the target, called the *token-count*, modeling the way that signals propagate through protein interactions in cellular signaling networks.

In order to overcome the issue of modeling reaction rates in the network, signaling dynamics are simulated by executing the signaling Petri net (SPN) for a set number of steps (called a *run*) multiple times, each time beginning at the same initial marking. For each run, the individual signaling rates are simulated via generation of random orders of transition firings (interaction occurrences). When the results of a large enough number of runs are averaged together, we find that the series of token-counts correlate with experimentally measured changes in the activity-levels of individual proteins in the underlying signaling network. In essence, the tokenized activity-levels computed by our method should be taken as abstract quantities whose changes over time correlate to changes that occur in the amounts of active proteins present in the cell. It is worth noting that some of the most widely used experimental techniques for protein quantification—western blots and microarrays—also yield results that are treated as indications, but not exact measurements, of protein activity-levels within the cell. Thus in some respects, the predictions returned by

our SPN-based simulator can be interpreted like the results of a western blot or microarray experiment looking at changes relative to “control”.

The key insight behind our approach is the assumption that, while all network parameters determine the actual signal propagation to some extent, the network connectivity is the most significant single determinant. While this is clearly a gross simplification, several researchers have observed that the connectivity of a biological network dictates, to a great extent, the network’s dynamics [18,45–47]. Some have conjectured that biological network connectivities have evolved to have a stabilizing effect on the overall network behavior, making the network more resilient to local fluctuations in other network parameters such as reaction rates and protein binding affinities [45,47]. Here we present the *signaling Petri net* (SPN) model and the signaling Petri net-based simulator whose designs collectively utilize this assumption and couple it with a Petri net tokenization scheme that quantifies the changes in protein activity-levels that occur as signals propagate through the network. In the following sections, we describe the synchronized Petri net, how we extended it to create the signaling Petri net, and a novel strategy for executing the signaling Petri net to simulate signaling network dynamics.

### Petri Nets

A Petri net is a graph that consists of two types of nodes, *places*, and *transitions* [44]. Edges in the graph, called *arcs*, are directed and connect places to transitions or transitions to places. Thus, the Petri net is a bipartite graph. Formally, a Petri net is a 4-tuple  $Q = \langle P, T, I, O \rangle$  where

- $P = \{p_1, p_2, \dots, p_m\}$  is the set of places,
- $T = \{t_1, t_2, \dots, t_n\}$  is the set of transitions,
- $I = \{i_1, i_2, \dots, i_k\}$  is the set of input arcs where for all  $(u, v) \in I$ ,  $u \in P$  and  $v \in T$ , and
- $O = \{o_1, o_2, \dots, o_l\}$  is the set of output arcs where for all  $(u, v) \in I$ ,  $u \in T$  and  $v \in P$ .

In order to simulate a dynamic process, a number of tokens is assigned to each place in order to indicate the presence of some quantitative property. This assignment of tokens to places encodes the state of the system and is called a marking, denoted  $\mathbf{m}$ . A *marked Petri net*,  $R = \langle Q, \mathbf{m}_0 \rangle$ , is a Petri net with a marking  $\mathbf{m}_0$ , called the initial marking. For the remainder of this paper, the term *Petri net* (PN) refers to a marked Petri net.

Changes in the state of the system are simulated by *executing* the Petri net—evaluating the effect of transitions on the marking of the network. These changes in marking are induced by sequential *firing* one or more transitions. When a transition fires, it removes a token from each place connected to it by input arcs and adds a token to each place connected to it by output arcs. The number of tokens removed from inputs and added to outputs can be specified by weighting the input arcs. However, as our extension does not use this weighting property, we do not consider this very common PN formulation here.

A transition can only fire when it is *enabled*, meaning that each of its input places has at least one token in the current marking. If a transition  $t$ , when fired on a marking  $\mathbf{m}_1$ , produces marking  $\mathbf{m}_2$ , then we write  $\mathbf{m}_1 \xrightarrow{t} \mathbf{m}_2$ .

This notation can be extended to represent the effect of firing a series of transitions. A *firing sequence*,  $\sigma = (t_1, t_2, \dots, t_j)$  is a sequence of transitions. The sequence’s cumulative effect on the system’s state is denoted  $\mathbf{m}_0 \mid \sigma \mathbf{m}_j$  where  $\mathbf{m}_0$  is the initial marking and  $\mathbf{m}_j$  is the marking produced by the firing of the sequence of transitions in the order specified in  $\sigma$ . In this paper, we write  $\mathbf{m}_j^\sigma$  to indicate the

marking produced by the first  $j$  transitions in  $\sigma$ . Therefore, in the above example,  $\mathbf{m}_0^\sigma = \mathbf{m}_0$  and  $\mathbf{m}_1^\sigma = \mathbf{m}_j$ .

For a more complete introduction to types of Petri nets and their properties, we refer the reader to [44].

**Synchronized Petri nets.** Synchronized Petri nets model systems in which the firing of a transition is triggered by a specific event that occurs in the environment. The marked Petri net is extended to include a set of these events and a mapping function that assigns an event to each transition. When transition  $t$ ’s assigned event occurs, transition  $t$  is fired. Formally, a synchronized Petri net is a 3-tuple  $\langle R, E, Sync \rangle$ , where [44]:

- $R$  is a marked Petri net,
- $E = \{e_1, e_2, \dots, e_s\}$  is a set of events, and

$Sync: T \rightarrow EU\{\mathbf{e}\}$  maps each transition in the Petri net to an event. Event  $\mathbf{e}$  is the *always occurring event*. Any transition associated with  $\mathbf{e}$  is always immediately fired upon becoming enabled.

When executing a synchronized Petri net, transition  $t$  is fired when its associated event  $e = Sync(t)$  occurs. The order in which events are generated depends upon the environment which generates them. Just as in the marked Petri net, when a transition fires, it removes one token from each place connected by input arcs and gives one token to each place connected by output arcs.

As will be discussed in the next sections, we extend the synchronized Petri net paradigm to model the dynamics of a signaling network. To our knowledge, ours is the first use of the synchronized Petri net to model biochemical systems. In principle it is well suited to signaling networks since places represent proteins, tokens represent concentrations, and transitions represent directed protein interactions. A model of signaling event occurrence can be used to generate events and fire transitions, providing a way of simulating the signaling network’s behavior. These and other design details will be discussed in the next section.

### The Signaling Petri Net-Based Simulator

A high-level sketch of our simulator is given in Figure 2. Details and rationale for specific design decisions will be discussed in subsequent sections.

During the simulation, the input signaling Petri net is executed multiple times on a firing sequence constructed by the signaling event generator. The signaling event generator imposes an ordering on transition firing such that it creates a two-time scale simulation. The smaller time scale is discretized as the firing of a single transition. This unit is referred to as the *firing* time scale. Firing steps are nested within a larger time scale, called time *blocks*, in which each transition is fired exactly once. Thus, there are  $|T|$  firings per block. Since the simulation is run for the specified number of time blocks,  $B$ , there are  $B|T|$  firing steps in the simulation.

The time structure for an example simulation is illustrated in Figure 3. This dual-time approach is necessitated by the rate parameter sampling strategy we employ. Since the rate parameters are not known, our method executes many simulation runs (Step 2 in Figure 2) in order to sample the space of possible rate parameters. The markings returned by these runs are then averaged (Step 3 in Figure 2). The only requirement placed on the different rate parameter values is that all events occur within the same larger time frame—the time block. Therefore, within every time block all edges are evaluated once, though not necessarily in the same order.

This idea of evaluating random event orderings within a two-time scale system has appeared before in the domain of transcriptional networks [48]. In that study, Chaves et al. employed a two-time scale formulation of network updates similar in concept to the one we describe here. In their work, they assumed a boolean model of regulation and characterized the

```

PROCEDURE SIMULATE( $S, B, r$ )
1. Set the initial marking of  $S$ 
2. For  $i = 1$  to  $r$ 
    (a) Generate a random sequence of the events in  $E$ 
    (b) Simulate the network by executing the transitions associated with the events in the generated sequence
    (c) Record the number of tokens at each node, for each time step
3. For each node, compute the number of tokens at each time unit  $t$ , averaged over  $r$ 
    
```

**Figure 2. A High-Level Outline of the Procedure for Simulating a Signaling Network.** The input to the procedure is a signaling Petri net,  $S$ , the number of time units to simulate the network for,  $B$ , and the number of runs for which to repeat the simulation,  $r$ . The random generation of event ordering is employed to simulate the stochasticity in reaction rates and the differing times of signal arrivals. doi:10.1371/journal.pcbi.1000005.g002

effect of different relative rates of transcription within the same network on the final steady state reached. In contrast, our method is designed to operate on tokenized models of signaling networks with the ultimate intent of predicting the activity-level changes of proteins in the underlying signaling network over time.

In the next sections, we discuss in greater detail the core design decisions underlying our method: the signaling Petri net, transition firing, signaling network event generator, constructing the initial marking for the model, and sampling signaling rates. We then discuss how our strategy can be used to predict the outcome of perturbation experiments.

**The Signaling Petri Net**

The goal of our method is to predict the signal flow through a cell-specific network under specific experimental conditions. As a result, the signaling Petri net model must characterize the connectivity of the signaling network, the connectivity-level network properties that are unique to the cell type and experimental conditions under which the network is being studied, and the signaling processes of activation and inhibition.

The signaling Petri net is a synchronized Petri net with: 1) a specific way of modeling activating and inhibiting interactions using places, transitions, and arcs; 2) a one-to-one correspondence between events and transitions such that every transition is associated with a unique event; 3) modified rules regarding how many tokens are moved in response to a transition firing; and 4) a signaling network event generator.

Places correspond to the activated forms of signaling proteins. The number of tokens assigned to place  $p$  in marking  $\mathbf{m}_s, \mathbf{m}_s(p)$ , abstractly represents the amount of active protein  $p$  present in that

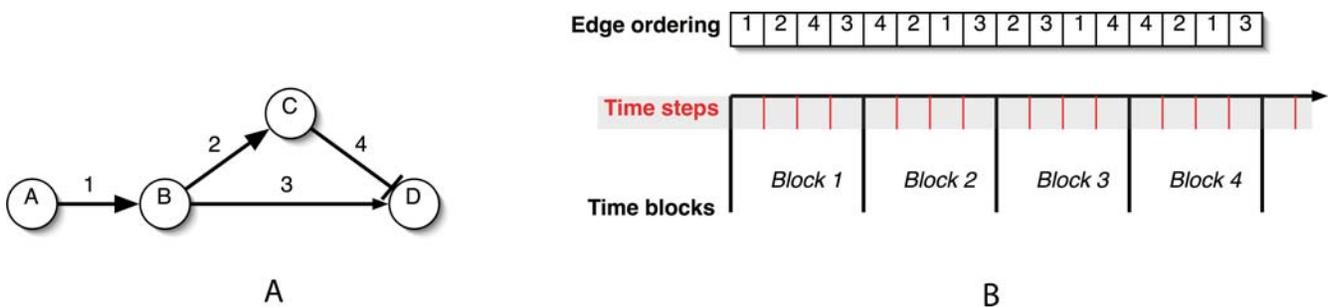
network state. Signaling interactions are modeled using transitions and their connected input and output arcs. Each transition,  $t$ , is associated with a unique signaling event,  $e$ , such that when  $e$  occurs, transition  $t$  fires. Figure 4 shows the equivalent signaling Petri net for a signaling network.

Formally, a signaling Petri net is a 3-tuple  $S = \langle R, E, Sync \rangle$ , where:  $R$  is a marked Petri net,  $E$  is a set of signaling events such that  $|E| = |T|$  and there is no *always occurring event*, and  $Sync: T \rightarrow E$  is a one-to-one mapping which assigns each transition a unique signaling event.

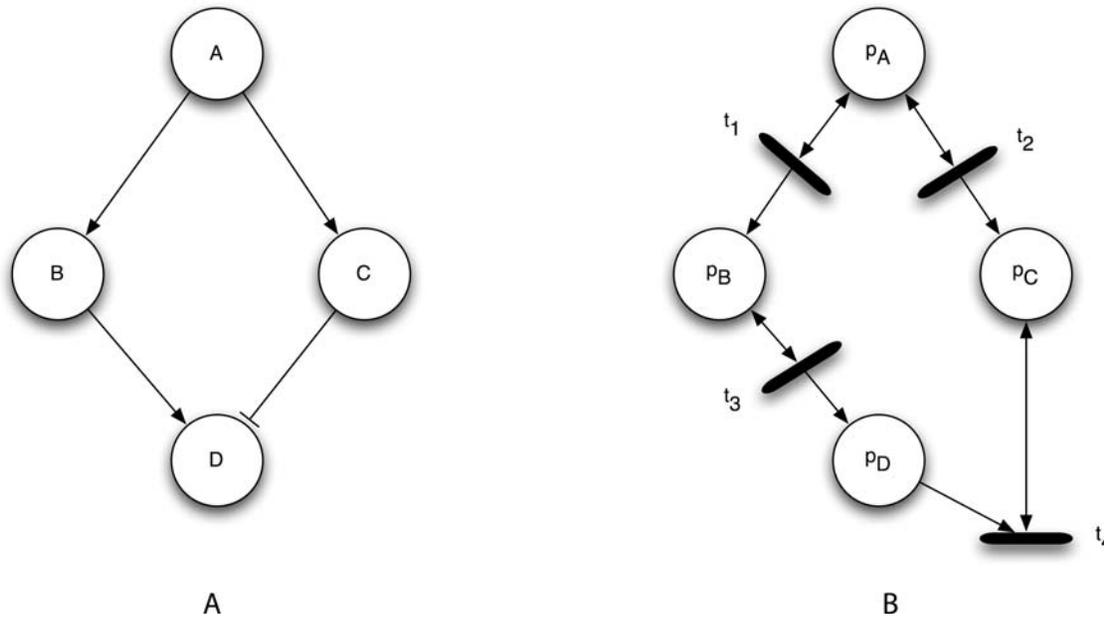
The initial marking of a signaling Petri net,  $\mathbf{m}_0$ , represents the state of rest from which the network is starting and being simulated. Proteins whose concentrations are known to be high can be given a large number of tokens, and those whose concentrations are known to be low can be assigned few or zero tokens. Attention to the initial marking is central to modeling cell-specific networks. In many cell lines, specific proteins are known to contain mutations that render them perpetually active or inactive [49]. Furthermore, experimental studies frequently involve the targeted manipulation of various proteins within the network. Both of these phenomena induce state changes in certain proteins at various time points that must be modeled. The way in which these are modeled will be discussed when the simulator design is explained.

**Transition Firing**

When a signaling interaction  $A \rightarrow B$  ( $A$  activates  $B$ ) or  $A \dashv B$  ( $A$  inhibits  $B$ ) occurs, it has the effect of changing the state of the system by modifying the activity-level of  $A$  and/or  $B$ . Thus, in the SPN used to model this network, the associated transition,  $t$ , will fire at



**Figure 3. The Effects of Reaction Rates on Signal Propagation.** (A) By changing the speed of signaling edge 3, the value of  $D$  at the end of a single simulation step can be reversed. If edge 3 is slower than the cascade  $B \rightarrow C \rightarrow D$ , then  $D$  will be active. If edge 3 is faster than the cascade, then  $D$  will be inactive. (B) An example of how the simulator might evaluate the individual edges during a run. In each time block, every edge is evaluated once. Each edge evaluation corresponds to one time step. Note that the order of the edge evaluation is shuffled during each time block in order to sample the space of possible relative signaling rates. doi:10.1371/journal.pcbi.1000005.g003



**Figure 4. An Example Signaling Network and Its Corresponding Petri Net.** An example signaling network (A) and its corresponding Petri net (B). Each signaling protein in the network, A, B, and C, are designated as places  $p_A$ ,  $p_B$ , and  $p_C$ . Signaling interactions become a transition node and its input and output arcs. Note that the connectivity for an activating edge differs from that of an inhibitory edge. doi:10.1371/journal.pcbi.1000005.g004

time  $\tau$  and produce marking  $\mathbf{m}_{\tau+1}$  from  $\mathbf{m}_{\tau}$ . The way in which  $\mathbf{m}_{\tau+1}$  is computed from  $\mathbf{m}_{\tau}$  depends on the set of input and output arcs attached to the transition as well as the number of tokens moved by the transition.

The combination of input and output arcs connected to a transition is determined exclusively by the type of interaction and the transition firing model. However, different topologies, combinations of input and output arcs, are needed to model the different biochemical processes that mediate protein-protein interactions in a signaling network. Here we examine four of the most common biochemical processes, identify the corresponding topological motifs, and ultimately devise a modeling policy best suited for non-parametric simulation of signal flow.

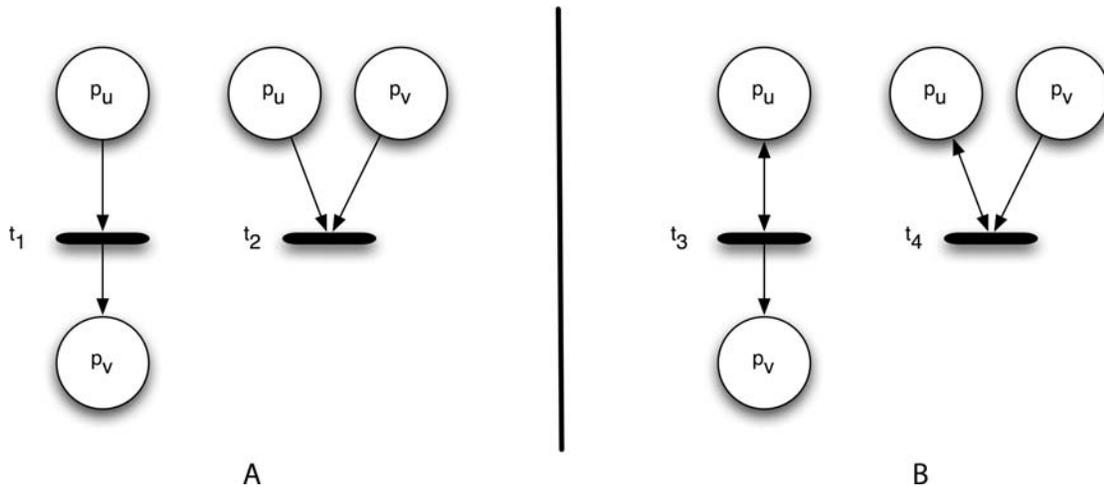
In *post-translational modification* (PTM), a protein mediates the addition or removal of a phospho group at a specific phosphorylation site on another protein. In *GTP/ATP binding*, a protein triggers the exchange of GDP (ADP) from GTP (ATP) on another protein. In a *recruitment* process, a protein mediates the relocalization of another protein to a different part of the cell. Finally, in a *complexing* process, a protein binds to another protein to create a complex, which can then participate in other reactions. In the first two processes, the mediating protein usually acts as an enzyme that participates in the reaction but is not consumed by the reaction. In the latter two processes, the participating protein often becomes unavailable to other reactions, transiently while the protein recruitment is taking place and for longer durations when complexing occurs. To model these two cases, we identified the two different token-passing policies implemented by the different topological motifs depicted in Figure 5.

**Token consumption.** In this policy,  $u \rightarrow v$  consumes tokens in  $u$  in order to generate new tokens for  $v$ . In order to model this,  $p_u$  is connected to transition  $t_1$  through an arc and  $p_v$  is connected to  $t_1$  through an output arc. When  $t_1$  fires, some number of tokens in  $p_u$  are moved into  $p_v$ . Similarly,  $u \rightarrow v$  consumes tokens in  $u$  in order to consume tokens in  $v$ . This is modeled by connecting  $p_u$  to  $t_2$  with an input arc and  $p_v$  to  $t_2$  with an input arc. When  $t_2$  fires, some number

of tokens are removed from both  $p_u$  and  $p_v$ . This policy models a recruitment or complexing event in which  $u$  binds to another molecule, thereby creating a molecule of type  $v$ . *A molecule of type  $u$  has been consumed in order to generate or deactivate a molecule of type  $v$ .*

**Token conservation.** In this policy,  $u \rightarrow v$  generates new tokens for  $v$  while conserving those in  $u$ . In order to model this,  $p_u$  is connected to transition  $t_3$  through a read arc. Node  $p_v$  is connected to  $t_3$  through an output arc. When  $t_3$  fires, some number of tokens in  $p_u$  is read (but not removed) and copied into  $p_v$ . Similarly,  $u \rightarrow v$  consumes tokens in  $u$  while conserving those in  $v$ . This is modeled by connecting  $p_u$  to  $t_4$  with a read arc and  $p_v$  to  $t_4$  with an input arc. When  $t_4$  fires, some number of tokens in  $p_u$  are read and removed from  $p_v$ . Enzymes will often behave in this way: inducing a change in a molecule ( $v$ ) without themselves undergoing any change. *A molecule of  $u$  has induced a change in a different molecule of type  $v$  without itself changing state.*

Ideally, for each interaction in the network, the associated transition could be embedded in the topology corresponding to the interaction's underlying biochemical mechanism. However, connectivity-level knowledge of the network does not provide this information for each interaction. In the absence of these details, we use one token-passing policy for all interactions in the network. We implemented and tested both the consuming and conserving policies and found that token conservation provides significantly more accurate results when compared to experimentally derived data. This is not surprising, as post-translational modification and GTP/ATP binding events are responsible for many activation state changes in signaling networks [1,50–52]. It is worth noting that our approach does not restrict the net structure to token conserving topologies. Thus, it is possible to use the token consumption topologies where such processes are known to occur. However, as our focus in this paper is designing a purely non-parametric simulation method, we consider the use of information regarding the biological mechanism of signaling as a potential way to further improve the accuracy of our method's predictions and identify this as a direction for future work.



**Figure 5. The Topological Motifs for Differing Signaling Processes.** (A) The token consumption motifs for complexing and recruitment. Transition  $t_1$  encodes activation of  $v$  by the binding or consumption of  $u$ . Transition  $t_2$  encodes deactivation of  $v$  by the binding or consumption of  $u$ . In both cases, the number of tokens of  $p_u$  decreases immediately after transitions  $t_1$  and  $t_2$  fire. (B) The token conserving motifs for PTM and GTP/ATP binding. Transition  $t_3$  encodes enzymatic activation of  $v$  by  $u$ . Transition  $t_4$  encodes enzymatic inhibition of  $v$  by  $u$ . In both cases, the number of tokens of  $p_u$  remains unchanged immediately after transitions  $t_3$  and  $t_4$  fire.  
doi:10.1371/journal.pcbi.1000005.g005

The transition topologies, as described above, do not designate how the number of tokens added to or removed from  $p_v$  is determined. However, we know that in biochemical signaling networks concentration has an effect on the strength of a signaling event [53–55]. Specifically, the higher  $u$ 's concentration, the stronger its effect on  $v$ —the more tokens that  $p_u$  has, the more tokens of  $p_v$  should be affected (generated or consumed).

However, because of the stochastic nature of the underlying biochemistry, it would be inaccurate to assume that *all* active  $u$  molecules will always participate in an interaction with  $v$ . In order to accommodate this observation, when transition  $t$  fires, we randomly select the number of  $p_u$ 's tokens to be involved in the subsequent evaluation of the transition, which we call a *signaling event*. Note that, according to our choice of topology,  $p_u$  can always be identified as the node connected to the transition by a read arc. In this paper, we assume a uniform distribution for selecting the number of tokens involved in a given signaling event, but acknowledge that other distributions may be more appropriate under certain circumstances and identify this as a topic deserving further consideration.

Let  $m_s(x)$  denote the number of tokens in node  $x$  at time  $s$ . For an interaction  $(u,v)$ , under the token conservation policy detailed above,  $u$ 's token-count remains unchanged after the firing of  $t$ , whereas  $v$ 's token-count is updated based on the following formula:

$$m_s(v) = \begin{cases} m_{s-1}(v) + \text{random}(0, m_{s-1}(u)) & \text{if } u \text{ activates } v \\ \max\{0, m_{s-1}(u) - \text{random}(0, m_{s-1}(u))\} & \text{if } u \text{ inhibits } v \end{cases}$$

where  $\text{random}(p,q)$  is a random integer drawn from a uniform distribution over the range  $[p,q]$ .

If we employ the policy of token passing with consumption, then after  $m_s(v)$  has been computed based on the formula above,  $m_s(u)$  is updated as:

$$m_s(u) = m_{s-1}(u) - \min\{m_{s-1}(u), |m_s(v) - m_{s-1}(v)|\}.$$

### Signaling Network Event Generator

The SPN topology and transition token-number selection policy alone do not specify the speed with which individual signaling interactions occur. However, such rates must be accounted for when simulating a signaling network. ODEs characteristically model such details as reaction rate constants; parameterized Petri nets specify these in a variety of ways including transition firing rates and firing probabilities [17,30]. In synchronized Petri nets, the environment controls the generation of events. Thus, the signaling network event generator is responsible for controlling the timing and ordering of signaling events. However, as our objective is a non-parametric simulation method, our approach must either estimate these parameters or operate without explicit knowledge of them.

Estimating reaction rates using only connectivity is currently beyond the predictive or inferential capabilities of computers. While there has been some work in the area of predicting reaction rates, all results of which we are aware require knowledge about the mechanism of signaling (e.g., [56]). As a result, without enriching the SPN model, it is doubtful that rate parameters can be accurately estimated.

For this reason, the signaling network event generator operates without explicit knowledge of the rate parameters. To compensate for this “missing” knowledge, we make use of an observation of signaling networks discussed earlier: a network's connectivity determines its dynamics. Several studies have found that the connectivity of biochemical networks desensitizes them to small fluctuations in the kinetic biochemical parameters [45–47]. Understood within the context of evolution – a stochastic process that tweaks signaling network parameters across generations – this is a highly desirable property as it ensures that an offspring remains viable despite fluctuations in the exact tuning of its cellular machinery. If this property holds, then small fluctuations in the rate parameters should have a marginal effect on the overall propagation of signal through the network. We can consider these small effects to be noise obscuring the underlying dynamics of the network connectivity. By taking many samples of the network dynamics under a variety of reaction rate assignments and then averaging these dynamics, we simultaneously reduce the noise

introduced by any one rate assignment and strengthen the underlying dynamic characteristics of the network's connectivity.

However, since reaction rate constants can vary by several orders of magnitude—from  $10^{-10}$  to  $10^3$ , the task of correctly selecting parameters *close* to the true parameters is non-trivial. In fact, without having some estimate of the actual rate parameters, it is unclear as to how to measure closeness at all. Clearly, these are among the issues that make parameter estimation so difficult for ODE and Petri net approaches. Since our comparisons will be relative and not absolute, we take a relative approach to modeling rate parameters. The space of possible rate values is *the space of possible signaling event orderings*.

This idea is illustrated in Figure 3A. Protein A affects the activity of protein D through two separate pathways. Assuming that A is active to begin with, the relative speed of these two pathways determines the final activity of D. If the pathway through C is faster than the pathway BFD, then D will be active. However, if the pathway speeds are reversed, then D will remain inactive. The overall outcome of this network can be represented without any use of numeric reaction rates by representing the reaction rates as an ordering over all the edges in the network. We can extend this idea to the SPN by observing that there exists a unique event for each signaling edge in the signaling network.

This sampling strategy is the motivation for the dual-time framework depicted in Figure 3B and implemented by the signaling network event generator shown in Figure 6. *Time blocks* are the larger time intervals during which every signaling event occurs exactly once. Since every transition in the SPN is associated with a unique event, each transition will fire exactly once in each time block. *Transition firings* are the smaller time units that impose a strict sequential order on the occurrence of signaling events. While this strict sequentiality of firing models relative reaction rates, it also discretizes the effect of signaling events. Though this is consistent with the definition of transition firing in discrete time Petri nets (only one transition is evaluated at a given point in time) [44], in biological signaling networks there is no such serial evaluation constraint. However, our validation with experimental data suggests that this discretization approximation does not affect the overall validity of the simulation results.

```

PROCEDURE GENERATESIGNALINGEVENTS( $E, n$ )
1.  $k = |E|$ 
2.  $\sigma$  an empty array of size  $(k \times n)$ 
3.  $i = 1$ 
4. for  $b = 1$  to  $n$ 
   (a)  $E' = E$ 
   (b) while  $E' \neq \emptyset$ 
     i.  $e$  = a random event from  $E'$ 
     ii.  $\sigma[i] = e$ 
     iii.  $E' = E' - \{e\}$ 
     iv.  $i = i + 1$ 
5. Return  $\sigma$ 

```

**Figure 6. The Algorithm That Implements the Signaling Network Event Generator.** This routine generates the time block/firing structure. Given a set of events,  $E$ , and the number of blocks for which the SPN will be executed,  $n$ , GENERATESIGNALINGEVENTS generates  $n$  blocks of events, each consisting of  $|E|$  events ordered randomly. In each block, every event in  $E$  occurs exactly once. doi:10.1371/journal.pcbi.1000005.g006

## Defining the Initial State

As mentioned previously, the initial state of the SPN is the initial marking,  $\mathbf{m}_0$ . As the SPN provides no explicit information on how this marking should be built, we propose three ways to construct the initial state: zero, basal, or experimentally derived. In a zero initial state, the simulator initializes all proteins to have zero tokens. The basal initial state is a random distribution of activation levels intended to model the cell when no impulses due directly to external stimuli are propagating through the signaling network. Though a basal network is considered at rest, in general it will not have a zero marking since signal flows are known to occur even in unstimulated signaling networks through autocrine and paracrine secretions by the cells. The experimentally derived initial state is based on knowledge about the activity levels of various proteins just prior to the addition of the external stimuli.

When accurate experimental data is available such as results from microarrays or western blots, the experimentally derived initial state may be the most accurate. A challenge in using experimental data, however, is determining how best to assign numbers of tokens based on the experimentally observed activity levels.

In the absence of reliable experimental data, the basal initial state seems more accurate than the zero initial state. However, it presents the challenge of properly selecting the basal activity-levels to assign to each protein in the model network. In [18], a basal initial state was constructed by activating a small number of randomly selected proteins in the signaling network. However, the work in [18] was done using a boolean model. Translating this approach into a tokenized model creates the additional complexity of determining how many tokens each basally active protein should receive. The correct values are likely to depend on the specific signaling network and experimental conditions.

We performed preliminary tests to compare the effect of using different basal versus zero markings on the outcome of the simulator. We found that the basal and zero states produced indistinguishable predictions so long as less than 30% of the proteins were activated and a small number of tokens ( $<5$ ) were used when constructing the basal marking. This is not as surprising as it may seem at first. Inhibitory edges will quickly consume a small number of tokens scattered throughout the network, effectively returning much of the network to the zero state before a stimulation event can propagate through.

Furthermore, while validating our method, we also compared the predictions produced by SPNs based on a zero initial state and experimentally derived initial state. These, too, did not produce noticeably different final results for similar reasons as discussed above. Details of these comparisons will be discussed further in the Results and Discussion sections.

However, since all three initial state construction strategies yield qualitatively identical predictions, using zero initial states has the advantage of invoking the fewest unnecessary assumptions about the network (as in the case of the basal initial state) and requiring the least experimental data (as in the case of the experimentally derived state). Nonetheless, in our implementation of the tool, we allow for using any one of these three initial state construction strategies.

## Modeling Cell-Specific Signaling Networks

Whereas consensus signaling networks typically represent the connectivity in normal cells, many experiments are conducted on abnormal cells in which oncogenic mutations, gene knockouts, and pharmacological inhibitors have altered the behavior of various signaling nodes in the network. In an SPN, these alterations to the signaling network can be modeled by adding/removing transitions

(and associated input/output arcs) and explicitly setting the token count for various proteins in the initial state.

The two network alterations which are commonly induced by oncogenic mutations, gene knockouts, or pharmacological inhibitors are constitutively high or low protein activity-levels, meaning that a protein is either unable to be inhibited or unable to be activated. The simulator allows for proteins to be specified as either fixed *High* or *Low*. Here we explain how these are modeled by changes to the SPN.

If protein *u* is fixed high, then this protein cannot be inhibited. Thus, all transitions that remove tokens from  $p_u$  are removed from the SPN. The fact that *u* is high, however, also suggests that it maintains a higher activity level in general. Therefore, in the initial state,  $\mathbf{m}_0(p_u) = H$ , where *H* is a non-zero number of tokens. Since all inhibiting transitions have been removed from the SPN, throughout any execution, place  $p_u$  will always have at least *H* tokens.

In experiments, we have observed that the choice of the value of *H* does not change the relative outcome of the simulations. While *H* will affect the actual number of tokens present in a given place as well as the number of time blocks required to observe certain activity-level changes, the relative changes in activity-level (number of tokens) among different proteins (places) does not change. As a result, one is free to select any reasonable value of *H* (for our experiments, we used *H*=10) as long as this *H* is held constant across all simulations whose results will be compared.

If protein *u* is fixed low, then this protein cannot be activated. Thus, all transitions that add tokens to  $p_u$  are removed from the SPN. The fact that *u* is low, however, also suggests that it maintains a constantly low activity level in general. Therefore, in the initial state,  $\mathbf{m}_0(p_u) = L$ , where *L* is a small number of tokens (in our simulations we use *L*=0). Since  $p_u$  is only inhibited, we observed that all constitutively low proteins quickly had their marking reduced to zero.

Unlike the value of *H*, extra caution must be taken when selecting values for representing *L*. A value of *L* that is too large can destabilize the early propagation of signal through the network. In our experiments, we obtained best results for values of *L* very close to or equal to zero ( $L \leq 2$ ). Beyond this, the final results obtained depended on other values in the network, the strength of the signal, and the duration of the simulation.

### Simulating a Signaling Network

Figure 7 provides more detailed versions of the simulation algorithm outlined in Figure 2. Steps 1 and 2 of the SIMULATE procedure constructs the initial marking and net topology to incorporate perpetually high proteins, *H*, and perpetually low proteins, *L*. In this paper, proteins that are assigned high activity-levels receive an initial token count of 10 in order to model a higher-than-average initial activity-level. As discussed earlier, using other values of *H* scale the activity-levels of all the proteins in the network, but will not qualitatively change their relative activities.

The loop in Step 3 runs *r* individual simulation runs. Each run receives a different event ordering,  $\sigma^e$ , thereby implementing the interaction rate sampling strategy. The time block/step structure is contained within the ordering  $\sigma^e$  (see Figure 6C). As a result, the SPN execution step simulates the events by firing their associated transition. Only those markings that correspond to time block boundaries are sampled.

After SIMULATE finishes collecting the time block markings from all the runs, Step 4 computes the average markings for each time block and Step 5 returns these averages.

```

PROCEDURE SIMULATE(S, H, L, B, r)
1. For each p ∈ H
   •  $m_0(p) = 10$ ;
   •  $I = I - \{(p, t) : t \in T \text{ and } (t, p) \notin O\}$ 
2. For each p ∈ L
   •  $m_0(p) = 0$ ;
   •  $I = I - \{(t, p) : t \in T\}$ ;
3. for i = 1 to r
   •  $\sigma^e = \text{GenerateSignalingEvents}(E, B)$ ;
   • Execute  $\mathbf{m}_0^i | \sigma | \mathbf{m}_{B|T}^i$ ;
4. For each p ∈ P and  $0 \leq b \leq B$ 
   •  $\bar{m}_b(p) = \frac{1}{r} \sum_{i=1}^r m_{b|T}^i(p)$ ;
5. Return  $(\bar{\mathbf{m}}_1, \bar{\mathbf{m}}_2, \dots, \bar{\mathbf{m}}_B)$ 
    
```

**Figure 7. The Procedure for Simulating a Signaling Petri Net.**

SIMULATE predicts the signal flow through the SPN *S*. The simulation is run for *B* time blocks; the results of *r* runs are averaged to produce the final result. Most of the work is done by the signaling Petri net execution procedure detailed in the preceding sections. This execution actually performs an individual run. This procedure takes the initial marking,  $\mathbf{m}_0$ , and applies the sequence of transitions triggered by the event sequence,  $\sigma^e$ . This ordering, generated by the algorithm in Figure 6, has the dual time structure in which each block of edges contains every event in *E* exactly once. Each firing evaluates the effect of one transition. The markings at the end of each time block are extracted in Step 5.

doi:10.1371/journal.pcbi.1000005.g007

### Simulating a Perturbation Experiment

We tested the accuracy and performance of our method by simulating the effect of two different targeted manipulations to a well-known signaling network. We compared these predictions to experimental results produced by performing the actual manipulations on two separate cancer cell lines.

The perturbations we considered in this study altered the constitutive activity-level of various proteins in the network (as opposed to affecting specific signaling interactions). Therefore, we modeled the perturbations as changes in the high and low proteins— $H^c$  and  $L^c$  for the control (unperturbed) network and  $H^p$  and  $L^p$  for the perturbed network.

A variant of the SIMULATE method was required to quantify how a perturbation changed the protein token-counts for each time block. Figure 8 shows the algorithm we used. In the procedure DIFFERENTIALSIMULATE, the input *S* provides the consensus SPN. Inputs  $H^c$  and  $L^c$  specify the control high and low proteins, the inputs  $H^p$  and  $L^p$  specify the perturbed high and low proteins. After Steps 1–5 construct two separate SPNs for the control and perturbed conditions, the loop in Step 6 performs *r* independent simulations over the control and perturbed models. Step 6d computes the difference between the markings at the end of each time block in the perturbed and control networks. The marking difference  $\mathbf{d}_j^i = \mathbf{m}_j^p - \mathbf{m}_j^c$  yields the marking  $\mathbf{d}_j^i$  where  $d_j^i(v) = m_j^p(v) - m_j^c(v)$  for each  $v \in P$ . Following the loop, the marking differences are averaged to obtain the time series  $(A_1, A_2, \dots, A_B)$  where  $A_b(v)$  is the average change in the token-count for protein *v* at time block *b*.

For values of  $|A_b| > 0$  for a given molecule *v*, we can conclude that the perturbation caused a change in the activity-level of *v* at time block *b* only if the difference observed is statistically

```

PROCEDURE DIFFERENTIALSIMULATE( $S, H^c, L^c, H^p, L^p, B, r$ )
1.  $S^c = S, S^p = S$ ;
2. For each  $p \in H^c$ 
   (a)  $m_0^c(p) = 10$  and  $I^c = I^c - \{(p, t) : t \in T^c \text{ and } (t, p) \notin O^c\}$ 
3. For each  $p \in L^c$ 
   (a)  $m_0^c(p) = 0$  and  $I^c = I^c - \{(t, p) : t \in T^c\}$ ;
4. For each  $v \in H^p$ 
   (a)  $m_0^p(v) = 10$  and  $I^p = I^p - \{(v, t) : t \in T^p \text{ and } (t, v) \notin O^p\}$ 
5. For each  $v \in L^p$ 
   (a)  $m_0^p(v) = 0$  and  $I^p = I^p - \{(t, v) : t \in T^p\}$ ;
6. for  $i = 1$  to  $r$ 
   (a)  $\sigma^c = \text{GenerateSignalingEvents}(E, T)$ ;
   (b) Execute  $\mathbf{m}_0^c(\sigma)\mathbf{m}_{B|T}^c$ ;
   (c) Execute  $\mathbf{m}_0^p(\sigma)\mathbf{m}_{B|T}^p$ ;
   (d) For  $j = 0$  to  $B$ 
       i.  $\mathbf{d}_j^i = \mathbf{m}_{j|T}^p - \mathbf{m}_{j|T}^c$ 
7. For each  $p \in P$  and  $0 \leq b \leq B$ 
   (a)  $\Delta_b(p) = \frac{1}{r} \sum_{i=1}^r \mathbf{d}_b^i(p)$ ;
8. Return  $(\Delta_1, \Delta_2, \dots, \Delta_B)$ ;
    
```

**Figure 8. The Algorithm for Predicting the Effect on Signal Propagation of a Targeted Manipulation.** The algorithm for predicting the effect on signal propagation of a targeted manipulation on signaling network with connectivity  $G$ . The ‘ $c$ ’ and ‘ $p$ ’ superscripts are used to denote parameters in the control and perturbed versions, respectively, of the SPN. doi:10.1371/journal.pcbi.1000005.g008

significant. We use a t-test to determine whether this change is statistically significant for protein  $v$  at time block  $b$ . Computing the t-test for two distributions (control and perturbation) requires knowledge of the mean ( $\mu_{c,b}$  and  $\mu_{p,b}$ ) as well as the variance ( $\sigma_c^2$  and  $\sigma_p^2$ ) for both distributions. In order to obtain these parameters for the control network, a large number,  $X$ , of independent simulations is run. Simulation  $i$  provides a single series of markings,  $(\bar{\mathbf{m}}_1^i, \bar{\mathbf{m}}_2^i, \dots, \bar{\mathbf{m}}_B^i)$ . The mean is then computed:

$$\mu_{c,b,v} = \frac{\sum_{i=1}^X \bar{m}_{b(v)}^i}{X}.$$

The variance is computed similarly:

$$\sigma_{c,b,v}^2 = \frac{\sum_{i=1}^X (\bar{m}_{b(v)}^i - \mu_{c,b,v})^2}{X - 1}.$$

The parameters  $\mu_{p,b,v}$  and  $\sigma_{p,b,v}^2$  for the perturbed network are computed as described above by substituting the perturbed network for the control network. Using these parameters, the t-value for molecule  $v$  at time block  $b$  can be computed from the formula

$$t - \text{value} = \frac{\mu_{c,b,v} - \mu_{p,b,v}}{\sqrt{\frac{\sigma_{c,b,v}^2}{X} + \frac{\sigma_{p,b,v}^2}{X}}}.$$

The statistical significance of the difference can then be obtained by comparing the t-value to the desired critical value.

Note that the DIFFERENTIALSIMULATE procedure and the associated significance test can predict the effect not only of

perturbations, but also of any two different experimental (or cellular) conditions imposed on the same signaling network. As a result, in addition to perturbation experiments, our method can also be used to study the effects of other phenomena that induce changes in the propagation of signal through a signaling network.

### Cell-Specific Signaling Network Models

Figure 1 shows the signaling network we analyzed. We obtained the core connectivity from a published literature survey on the EGFR network [57]. We added to this several other well-established interactions taken from literature [36–43]. The response of this network to various perturbations was measured and simulated in two separate breast cancer cell lines: MDA231 and BT549. The core signaling Petri net used,  $S^{\text{EGFR}}$ , is captured by the following signaling proteins and interactions: places (the set  $P$ ): VEGFR, VSRC, VRac, VMEKK4, VMEK4, VJNK, VMEKK6, VMEK6, VSTAT, VGrb2, VShc, VSOS, VRB, VELK, VBAD, VNFKB, VRAS, VGAB1, VPIP3, VPI3K, VPDK1, VPTEN, Vc-Raf, VAKT, VLKB1, VMEK, VGSK3 $\beta$ , VAMPK, VTSC2, VMAPK1,2, VRSK, VRheb, VmTOR-Raptor, V4EBP1, Vp70S6K, Vp38, and VpS6.

Protein interaction network motifs (the combination of arcs and transitions): VEGFR  $\rightarrow$  VGrb2, VGrb2  $\rightarrow$  VShc, VShc  $\rightarrow$  VSOS, VSOS  $\rightarrow$  VRAS, VGrb2  $\rightarrow$  VGAB1, VGAB1  $\rightarrow$  VPI3K, VEGFR  $\rightarrow$  VSRC, VSRC  $\rightarrow$  VSTAT, VPI3K  $\rightarrow$  VPIP3, VPIP3  $\rightarrow$  VPDK1, VRAS  $\rightarrow$  Vc-Raf, VPDK1  $\rightarrow$  VAKT, VRAS  $\rightarrow$  VRac, VRac  $\rightarrow$  VMEKK4, VMEKK4  $\rightarrow$  VMEK4, VMEK4  $\rightarrow$  VJNK, VJNK  $\rightarrow$  VSTAT, VRac  $\rightarrow$  VMEKK6, VMEKK6  $\rightarrow$  VMEK6, VMEK6  $\rightarrow$  Vp38, Vp38  $\rightarrow$  VSTAT, VPDK1  $\rightarrow$  Vp70S6K, VPTEN  $\uparrow$  VAKT, VAKT  $\uparrow$  Vc-Raf, VAKT  $\uparrow$  VGSK3 $\beta$ , VAKT  $\uparrow$  VTSC2, VAKT  $\uparrow$  VAMPK, VAKT  $\uparrow$  VBAD, VAKT  $\uparrow$  VNFKB, VAKT  $\rightarrow$  Vp70S6K, VLKB1  $\rightarrow$  VAMPK, VMEK  $\rightarrow$  VMAPK1,2, VMAPK1,2  $\rightarrow$  VRB, VMAPK1,2  $\rightarrow$  VELK, VMAPK1,2  $\rightarrow$  VSTAT, VGSK3 $\beta$   $\rightarrow$  VTSC2, VAMPK  $\rightarrow$  VTSC2, VMAPK1,2  $\rightarrow$  VEGFR, VMAPK1,2  $\uparrow$  VTSC2, VMAPK1,2  $\rightarrow$  Vp70S6K, VMAPK1,2  $\rightarrow$  VRSK, VRSK  $\uparrow$  VTSC2, VTSC2  $\uparrow$  VRheb, VRheb  $\rightarrow$  VmTOR-Raptor, VAKT  $\rightarrow$  VmTOR-Raptor, VmTOR-Raptor  $\rightarrow$  V4EBP1, VmTOR-Raptor  $\rightarrow$  Vp70S6K, Vp70S6K  $\uparrow$  VEGFR, VSRC  $\uparrow$  VSRC, VRac  $\uparrow$  VRac, VMEKK4  $\uparrow$  VMEKK4, VMEK4  $\uparrow$  VMEK4, VJNK  $\uparrow$  VJNK, VMEKK6  $\uparrow$  VMEKK6, VMEK6  $\uparrow$  VMEK6, VSTAT  $\uparrow$  VSTAT, VGrb2  $\uparrow$  VGrb2, VShc  $\uparrow$  VShc, VSOS  $\uparrow$  VSOS, VRAS  $\uparrow$  VRAS, Vc-Raf  $\uparrow$  Vc-Raf, VMEK  $\uparrow$  VMEK, VMAPK1,2  $\uparrow$  VMAPK1,2, VRB  $\uparrow$  VRB, VELK  $\uparrow$  VELK, VRSK  $\uparrow$  VRSK, VGAB1  $\uparrow$  VGAB1, VPIP3  $\uparrow$  VPIP3, Vp38  $\uparrow$  Vp38, VPI3K  $\uparrow$  VPI3K, VPDK1  $\uparrow$  VPDK1, VAKT  $\uparrow$  VAKT, VBAD  $\uparrow$  VBAD, VNFKB  $\uparrow$  VNFKB, VAMPK  $\uparrow$  VAMPK, VmTOR-Raptor  $\uparrow$  VmTOR-Raptor, Vp70S6K  $\uparrow$  Vp70S6K, VpS6  $\uparrow$  VpS6, V4EBP1  $\uparrow$  V4EBP1.

Notice that the last several edges are self-inhibitory loops (e.g.,  $v_{\text{Ras}} \uparrow v_{\text{Ras}}$ ). These loops are used to model regulatory mechanisms that are not present in the model network.

For molecules that do not have specific inhibitory edges modeled in the network, we use the self-inhibitory loop to prevent exponential increase in the token counts and to model inhibitory mechanisms beyond the scope of the network. For example, consider the molecule Ras in the network shown in Figure 1. In the model, this protein is not inhibited. However, biologically we know that Ras has intrinsic GTPase function which inactivate itself. In order to model this, we introduce a self-inhibitory loop.

The differences between the two cell-specific networks are captured by following activity assignments to various proteins in the SPN. In the MDA231 cell line,  $H^{\text{MB}} = \{v_{\text{Ras}}, v_{\text{EGF}}\}$  and  $L^{\text{MB}} = \emptyset$ . In the BT549 cell line,  $H^{\text{BT}} = \{v_{\text{EGF}}\}$  and  $L^{\text{BT}} = \{v_{\text{PTEN}}\}$ .

Of the two perturbations we considered, one significantly knocked down the activity-level of TSC2 and the other knocked down mTOR-Raptor. While the core SPN still modeled these networks, separate *perturbed* activity-assignments were required for each cell line-perturbation pairing:  $L^{\text{MB-TSC2}} = L^{\text{MB}} \cup \{v_{\text{TSC2}}\}$ ,

$$L^{MB-mTOR} = L^{MB} \cup \{v_{mTOR-Raptor}\}, L^{BT-TSC2} = L^{BT} \cup \{v_{TSC2}\} \text{ and } L^{BT-mTOR} = L^{BT} \cup \{v_{mTOR-Raptor}\}.$$

**Setup for Perturbation Experiments**

**Cell culture and stimulation.** Human MDA-MB-231 (MDA231) and BT549 breast cancer cells were routinely maintained in RPMI supplemented with 10% FBS. For signaling experiments, logarithmically growing cells were serum-starved for 16 hours and then subjected to treatments by epidermal growth factor (EGF) (20 ng/mL) (Cell Signaling Technology, Beverly, Massachusetts) for 30 minutes. Controls were incubated for corresponding times with DMSO. To knock down TSC2, cells were treated with short interfering RNA (siRNA) (Dharmacon, Lafayette, Colorado) for 72 hours prior to EGF stimulation. Control cells were transfected with non-targeting (N/T) siRNA (Dharmacon, Lafayette, Colorado) prior to EGF treatment.

**Antibodies.** The following antibodies were used for immunoblotting: anti-phospho-p44/42 MAPK, anti-phospho-GSK3β (S21/S9); anti-phospho-AKT(ser473); anti-phospho-TSC2(T1462); anti-phospho-mTOR(S2448); anti-phospho-P70S6K(T389) (Cell Signaling Technology, Boston, Massachusetts); and anti-β-Actin (Sigma-Aldrich, St. Louis, Missouri).

**SDS-PAGE and immunoblotting.** Cells were lysed by incubation on ice for 15 minutes in a sample lysis buffer (50 mM Hepes, 150 mM NaCl, 1 mM EGTA, 10 mM Sodium Pyrophosphate, pH 7.4, 100 nM NaF, 1.5 mM MgCl2, 10% glycerol, 1% Triton X-100 plus protease inhibitors; aprotinin, bestatin, leupeptin, E-64, and pepstatin A). Cell lysates were centrifuged at 15,000 g for 20 minutes at 4°C. The supernatant was frozen and stored at -20°C. Protein concentrations were determined using a protein-assay system (BCA, Bio-Rad, Hercules, California), with BSA as a standard. For immunoblotting, proteins (25 μg) were separated by SDS-PAGE and transferred to Hybond-C membrane (GE Healthcare, Piscataway, New Jersey). Blots were blocked for 60 minutes and incubated with primary antibodies overnight, followed by goat anti-mouse IgG-HRP (1:30,000; Cell Signaling Technology, Boston, Massachusetts) or goat anti-rabbit IgG-HRP (1:10,000; Cell Signaling Technology) for 1 hour. Secondary antibodies were detected by enhanced chemiluminescence (ECL) reagent (GE Healthcare, Piscataway, New Jersey). All experiments were repeated a minimum of three independent times.

**Setup for perturbation simulations.** To select the block duration parameter, B, we compared the experimentally derived fold change of AKT in the MDA231 cell line to the AKT fold changes predicted for B = 10, 20, 50, 100, and 1000. We found B = 20 to be the best fit and used this value for all simulations in this study.

We also experimented with input parameter r, the numbers of individual simulation runs averaged per simulation. We tried a range extending from r = 100 to r = 1000. We found that no observable changes occurred in trends for r ≥ 400. Therefore, r = 400 was used for all simulations in this study.

We considered both the zero and experimentally derived initial states as the initial markings for the TSC inhibition simulations. The experimental states for both cell lines were derived from western blots produced from cells that were incubated in DMSO and serum-starved for 16 hours. Unsampled molecules were assigned a marking of zero. The number of tokens assigned to each sampled molecule was directly proportional to the darkness of the line on the western blot. This assignment was done by hand, though devising automated and standardized methods for the construction of experimentally derived initial states is an important direction for future work. Since most of the molecules in the

network were not sampled, only mTOR-Raptor, TSC2, GSK3β, p70S6K, AKT, and MAPK were given non-zero markings. The initial markings used are shown in Table 1.

Since experimental results for the mTOR-Raptor inhibition were obtained from literature, we did not have experimental results for construction of experimentally derived initial states. Therefore, we used the zero initial states for the mTOR-Raptor inhibition simulations.

**Results**

In order to evaluate the accuracy of our simulation method, we tested its predictions of the effect of targeted manipulations on two cell-specific versions of the signaling network depicted in Figure 1. In each cell line, a TSC2-specific siRNA was applied and the concentration of several key proteins in the EGFR network were sampled 30 minutes after stimulation with EGF. This was repeated in the absence of the TSC2 siRNA in order to obtain the concentration in the control network. We also collected a corpus of literature detailing the response of signaling proteins activity-levels to the inhibition of mTOR-Raptor using Rapamycin [43,58]. Predictions were generated by our simulator for the TSC2 and mTOR-Raptor perturbations in both cell lines.

**Simulation**

To simulate a perturbation, we used two networks both based on the signaling network shown in Figure 1: the control network for the cell line and the perturbed network for the cell line. The control networks for the cell lines were different because it was important to model the cell-specific mutations. In the case of the BT549 cell line, there is a mutation that leads to the loss of PTEN, which makes AKT always active. In the MDA231 cell line, there is a mutation in Ras, which makes it always active. As shown in the formulation of the model, these are modeled using fixed activity assignments in the simulator.

The TSC2 (mTOR-Raptor) perturbed network for a cell line was created by taking the control network and fixing the activity-level of TSC2 (mTOR-Raptor) to zero for the duration of the simulation, effectively simulating the pharmacological inhibition of the protein. For each cell-line/perturbation pair, we ran the simulator on the control and perturbed networks using the DIFFERENTIALSIMULATE procedure in Figure 8 which computed the change in token-counts induced by the perturbation for all proteins in the model. These change plots are shown in Figure 9 for TSC2 and in Figure 10 for mTOR-Raptor. We ran the simulations using both experimentally derived initial states as well

**Table 1.** Experimentally Derived Initial Markings Used in the Simulations.

Molecule	MB231		BT549	
	Control	TSC2 Inhibited	Control	TSC2 Inhibited
mTOR-Raptor	0	1	5	5
TSC2	0	0	6	0
GSK3β	5	3	3	6
p70S6K	0	2	0	0
AKT	0	0	7	7
MAPK	2	6	1	2

doi:10.1371/journal.pcbi.1000005.t001

as zero initial states. The initial state used did not change the overall trends observed in the simulations.

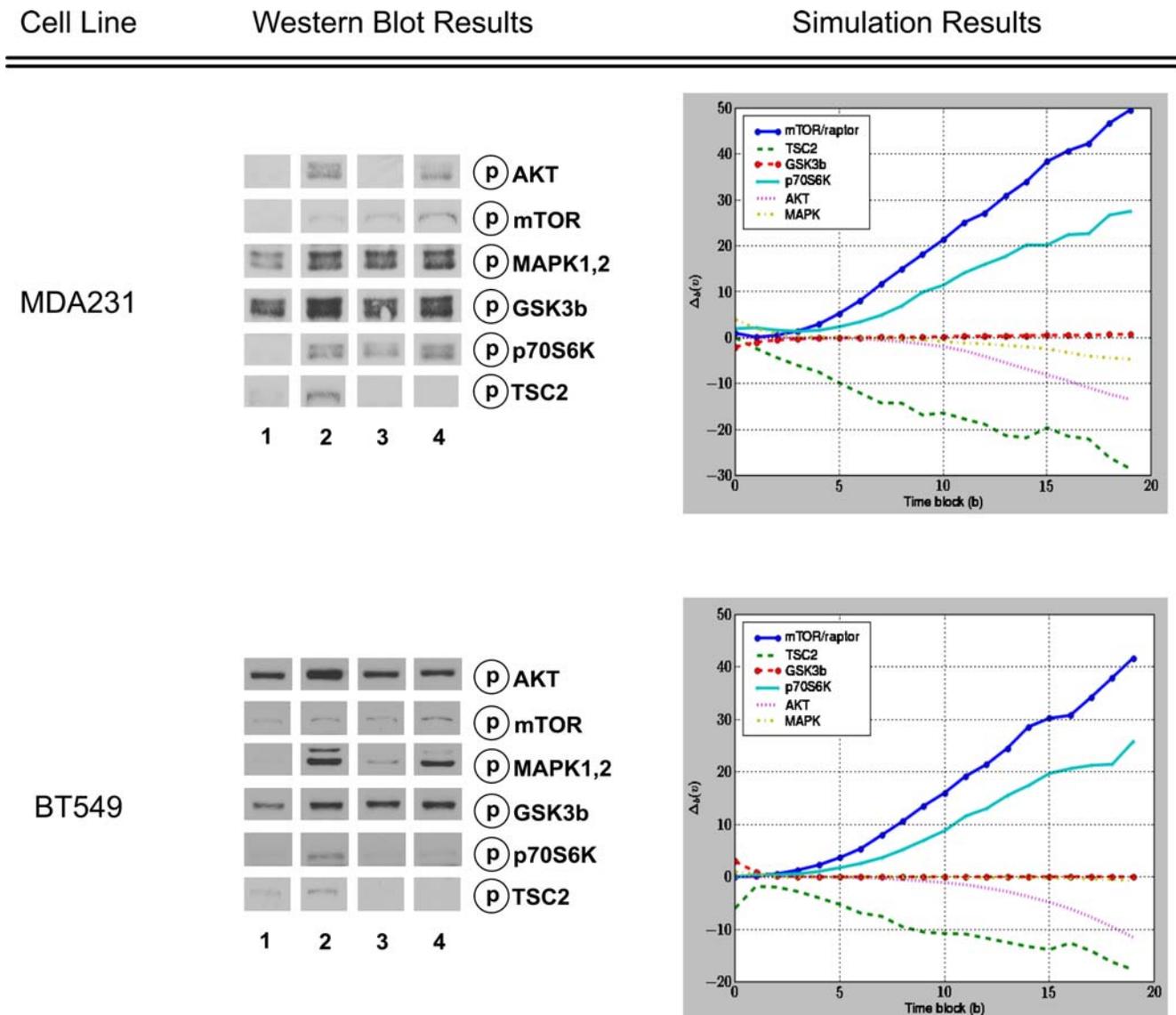
Using the t-test described in the Methods section, we also computed the statistical significance of the final time block ( $b = 20$ ) for each molecule considered. For each molecule considered, 400 runs, 20 time blocks, and 50 samples were used. With the exception of GSK3 $\beta$  which did not show a significant response to the perturbation, the changes of all other proteins sampled were beyond the 0.05 significance level (see Table 2). The statistical insignificance of the change in GSK3 $\beta$  is not surprising since, as shown in Figure 1, GSK3 $\beta$  is solely activated by LKB, a molecule fixed high in both cell lines. Thus, we should not expect either perturbation to have a significant effect on the activity of GSK3 $\beta$ , which is what the t-value indicates.

### Experimental Results

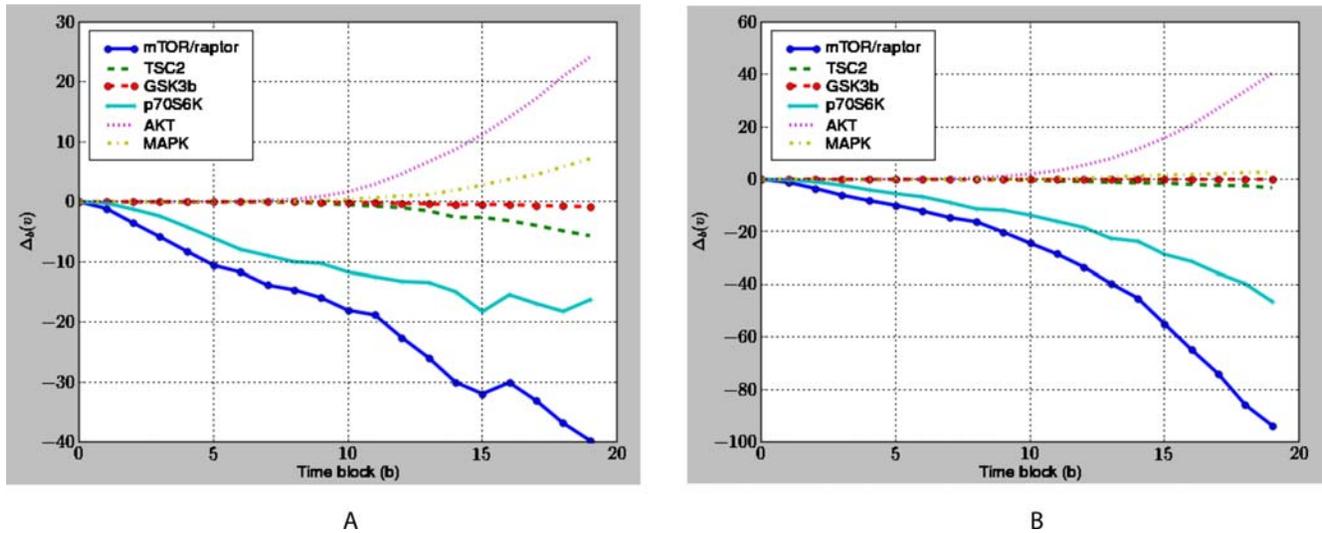
After the TSC2 perturbation was applied to a cell line, the protein concentrations were collected using western blots. Details are given in the Materials and Methods section. The western blot results are shown in Figure 9.

### Discussion

As can be seen in Table 3, our method correctly predicted the *relative* protein activity-level changes induced by the TSC2 perturbation in both cell lines, for most molecules sampled. Notice that *no change* (-) was reported for the predicted response of MAPK to the TSC2 perturbation despite the fact that a small change did occur in its marking during the simulation (see Figure 9)



**Figure 9. The Results of the TSC2 Perturbation Experiments and Simulations.** In the western blots, columns (or lanes) are as follows: (1) non-targeting (NT) control siRNA, (2) NT siRNA+EGF, (3) TSC2 siRNA, (4) TSC2 siRNA+EGF. The effect of the TSC2 siRNA on a given molecule can be assessed by comparing column 4 against column 2. For each molecule in the western blot, there is a corresponding simulation curve showing the predicted change in protein activity over time. For the purposes of this analysis, we compared the concentration change after 20 time steps (the left-most data points in the plots) for each molecule. Each simulation point corresponds to the average of 400 measurements that were computed using the procedure described in Figure 8. Experimentally derived initial states were used in the simulations. The results of both the experiments and simulations are qualitatively summarized in Table 3.  
doi:10.1371/journal.pcbi.1000005.g009



**Figure 10. The Predicted Response of the Network to an mTOR-Raptor Perturbation.** The predicted response of the network to a mTOR-Raptor perturbation in the (A) MDA231 and (B) BT549 cell lines. Our method predicts that the amount of available AKT increases in response to the perturbation, which is in agreement with results published in the literature [43,58]. Our method also predicts that the activity-level of p70S6K in the MDA231 cell line decreases in response to the perturbation, which has been observed experimentally [59]. Each point corresponds to the average of 400 measurements that were computed using the procedure described in Figure 8. doi:10.1371/journal.pcbi.1000005.g010

and the t-value for the change is significant (see Table 2). At first, interpreting this value as *no change* may seem misleading. However, one of the significant challenges in experimental perturbation experiments is separating true system responses from the background noise created by experimental variables that cannot be precisely controlled (among them cell population sizes, variability in microarray antibody binding effectiveness, and limited sensitivity of hardware and software used to quantify experimental results). As a result, a common practice is to only consider those substantial changes that are well beyond the background noise level. Our interpretation of the small predicted change in MAPK as *no change* reflects the fact that such small changes would not be detectable in microarray or western blot results. Thus, though such a small fluctuation might have occurred in the real data, it would not have been detected by the biologists and most likely would appear in the experimental data to have not changed.

Similar reasoning guided our decision to characterize the simulation (and experimental) results as either up (↑), down (↓), or no change (–) in general. Since the amount of protein

registered in a microarray or western blot is not always a reliable indicator of the exact amount of protein (or protein form) being measured, biologists are often reluctant to report degrees of increases or decreases—preferring binary observations such as *up* or *down* which are less subject to influence by extraneous experimental conditions. It is true that our simulation method produces precisely quantified increases or decreases which can be taken to indicate degrees of change in response to perturbations. However, as experimental techniques cannot reliably measure degrees of increase or decrease, we judged the binary (up or down) characterization to be a more reliable way of validating our method. Certainly, our method provides additional information of

**Table 2. The T-Values for the Molecules Sampled in the Microarray.**

Molecule	t-Value in MDA231	t-Value in BT549
mTOR-Raptor	41.72	30.53
TSC2	21.65	8.28
GSK3β	0.42	0.10
p70S6K	14.22	5.83
AKT	6.60	9.55
MAPK	16.35	18.93

The critical value for an alpha value of 0.05 with 50 samples is 2.0086. Note that the t-values for all molecules except for GSK3β are larger than this value, confirming that these changes are statistically significant. doi:10.1371/journal.pcbi.1000005.t002

**Table 3. Summary of the Effect of Perturbation Reported by Experimental and Simulated Methods.**

Molecule	MB231		BT549	
	Experiment	Simulation	Experiment	Simulation
mTOR-Raptor	↑	↑	↑ or –	↑
TSC2	↓	↓	↓	↓
GSK3β	–	–	–	–
p70S6K	↑	↑	↓	↑
AKT	↓ or –	↓	↓	↓
MAPK	–	–	–	–

The up arrow (↑) indicates that the perturbation caused a rise in the level of the phosphorylated protein; the straight line (–) indicates no change; and the down arrow (↓) indicates that a decrease occurred. Values in the *Experiment* column were estimated by comparing lanes 4 and 2 in Figure 9. We estimated the *Simulation* column by determining whether the top quartile of the distribution for the final time point was above, below, or at zero. In some cases it is difficult to judge for certain whether the total quantity of the phosphorylated protein changed or remained the same—both for the experimental and computational cases. In these situations, we indicated the uncertainty by listing the possible changes that the protein *could* have feasibly undergone. doi:10.1371/journal.pcbi.1000005.t003

degrees of change and we consider studying the accuracy of these degrees to be an important area for future work.

Our method also correctly predicted the activity-level change of AKT in response to mTOR-Raptor inhibition as reported by a number of studies [43,58]. Further, our method predicted that, when mTOR-Raptor is inhibited, the level of p70S6K in the MDA231 cell line decreased, which also had been observed experimentally [59].

The only incorrect prediction made by our method was the activity-level change of p70S6K in the BT549 cell line. However, BT549 cells contain an RB mutation [49] which could alter p70S6K phosphorylation [60]. It is a strength of our simulator that the discrepancy between our method's predictions and the experimental results identified a section of the model in which additional connectivity has been found which might account for the difference observed.

The predictions made by our simulator would be exceedingly difficult to derive by visual or manual inspection. Table 4 shows the number of paths between several pairs of compounds within the network. Where there is more than one path connecting two molecules, feed forward and feed backward loops are present. Attempting to determine, by hand, how these different loops will interact with one another is, by itself, a difficult endeavor even when not considering the additional task of deriving the rest of the network dynamics simultaneously. For the larger networks that are now becoming available, computational analysis becomes even more crucial to obtaining insights into the dynamic behavior of the network.

Despite the complexity of the network dynamics, it was straightforward to find and integrate the connectivity information used to build it. Most of the information sources [36–43] established the *existence* of various pathways and provided few or no biochemical or kinetic details. As a result, the literature we used would have provided little assistance in building a parameterized Petri net or ODE model. Due to the proliferation of curated signaling network repositories and searchable literature archives, connectivity information is relatively abundant which makes the ad hoc assembly of networks a relatively straightforward endeavor. This further underscores the advantage of using our method over ODEs or parameterized Petri nets to quickly model and characterize some of the dynamics of a signaling network.

For simulations that will be compared to experimental results, the time parameter must be selected carefully. The time parameter,  $B$ , indicates how many time blocks our method will simulate. The time block is an abstract unit of time. Therefore, before comparing experimental results and predictions, it is necessary to determine how many seconds, minutes, or hours correspond to a time block. This can be done by comparing a prediction of the simulator with the experimentally measured activity-level of one or two proteins at several time points in order to determine what time blocks correspond to the different sampled time points. In the present study, we calibrated our time blocks only once for two cell lines and six experimental conditions (two cell lines, with/without TSC2, with/without mTOR-Raptor). To select the time parameter we used the experimentally measured activity changes in two proteins at two time points. In contrast to other predictive dynamic analysis tools which require multiple time points and multiple protein samples in order to calibrate simulation and model parameters, our method has relatively low time and resource investment.

Besides the time parameter, the other component of our simulations which involved experimentally obtained knowledge was the initial states. The experimentally derived initial states require that some experimental data be available providing information on the initial concentrations of individual signaling proteins in the network prior to stimulation. However, in the

**Table 4.** Number of Paths Connecting Several Pairs of Compounds in the EGFR Model Used in Our Simulations

Source Protein	Destination Protein	Number of Paths
EGFR	TSC2	7
AKT	mTOR-Raptor	6
MEK	EGFR	4
AKT	p70S6K	8

The multiple paths connecting pairs of proteins highlight the complex interactions present within the network that give rise to its overall dynamic behavior.

doi:10.1371/journal.pcbi.1000005.t004

network that we considered here, the overall behavior of the network and of individual signaling proteins was resilient to changes in the initial states used. Zero and experimentally derived both produced the same overall change predictions. Thus, while experimentally derived initial states may be important for the simulation of some networks, it may well be the case that many networks (such as the one we considered in this paper) can be simulated without this knowledge—further reducing the experimental work that must be done prior to simulation.

The fact that our simulator produced accurate predictions for a variety of experimental conditions using the one core network model and set of simulation parameters also distinguishes our method from other predictive approaches. The only aspects of the model that were modified during the simulations were activity-levels reflecting the immediate effects of either the underlying tumor mutations (Ras and PTEN) or the perturbations (mTOR-Raptor and TSC2 targeted manipulation). In contrast, the accuracy of ODEs and Petri nets predictions are known to be sensitive to small changes to the model. For comparative studies such as the one conducted in this paper, an ODE or parameterized Petri net model might need to be re-constructed with different parameters for each experimental condition of interest. As a result, while it is possible to obtain our simulation results using these models, it remains beyond the capabilities of any existing ODE or parameterized Petri net system to provide insights into the effects of experimental conditions on the dynamic behavior of a signaling network with so little initial time and resource investment.

Though our method's predictions will not be as accurate as the results returned by a correctly parameterized ODE, biologists using our method can derive information about a network's dynamic behavior without having to conduct extensive experimentation and computationally expensive parameter estimation. This novel capability offers scientists the exciting prospect of being able to test hypotheses regarding signal propagation *in silico*. As a result, by using our method researchers can evaluate a wide array of network responses in order to determine the most promising experiments before even entering the laboratory.

## Acknowledgments

We gratefully acknowledge the three anonymous reviewers, whose comments and feedback improved the manuscript significantly.

## Author Contributions

Conceived and designed the experiments: DR MM JTT PTR. Performed the experiments: DR MM JTT. Analyzed the data: DR MM LN PTR. Contributed reagents/materials/analysis tools: DR MM LN PTR. Wrote the paper: DR MM LN PTR.

## References

- Hunter T (2000) Signaling-2000 and beyond. *Cell* 100: 113–127.
- Hanahan D, Weinberg RA (2000) The Hallmarks of Cancer. *Cell* 100: 57–70.
- Feldman DS, Cames CA, Abraham WT, Bristow MR (2005) Mechanisms of Disease: beta-adrenergic receptors alterations in signal transduction and pharmacogenomics in heart failure. *Nature Clinical Practice Cardiovascular Medicine* 2: 475–483.
- Belloni E, Muenke M, Roessier E, Traverse G, Siegel-Bartelt J, et al. (1996) Identification of Sonic hedgehog as a candidate gene responsible for holoprosencephaly. *Nat Genet* 14: 353–356.
- Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, et al. (2005) Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* 309: 1078.
- Papin JA, Hunter T, Palsson BO, Subramaniam S (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* 6: 99–111.
- The Cancer Cell Map (2006) <http://cancer.cellmap.org/cellmap/>.
- Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28: 27–30.
- Thomas PD, Campbell MJ, Kejariwal A, Mi J, Karlak B, et al. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res* 13: 2129–2141.
- Dasika MS, Burgard A, Maranas CD (2006) A computational framework for the topological analysis and targeted disruption of signal transduction networks. *Biophysical J* 91: 382–398.
- Eker S, Knapp M, Laderoute K, Lincoln P, Talcott C (2002) Pathway Logic: Executable Models of Biological Networks. *Electronic Notes Theoretical Computer Science* 71.
- Ruths D, Nakhleh L, Iyengar MS, Reddy SAG, Ram PT (2006) Graph-theoretic Hypothesis Generation in Biological Signaling Networks. *J Computational Biology* 13: 1546–1557.
- Schaub MA, Henzinger TA, Fisher J (2007) Qualitative networks: a symbolic approach to analyze biological signaling networks. *BMC Systems Biology* 1: 4.
- Papin JA, Palsson BO (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophysical J* 87: 37–46.
- Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO (2003) Metabolic pathways in the post-genomic era. *Trends Biochemical Sciences* 28: 250–258.
- Schilling CH, Letscher D, Palsson BO (2000) Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective. *J Theoretical Biology* 203: 229–248.
- Chaouiya C (2007) Petri net modelling of biological networks. *Briefings Bioinformatics* 8: 210–219.
- Li C, Suzuki S, Nakata M (2006) Structural Modeling and Analysis of Signaling Pathways Based on Petri Nets. *J Bioinformatics Computational Biology* 4: 1119–1140.
- Sackmann A, Heiner M, Koch I (2006) Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics* 7: 482–498.
- Steggles IJ, Banks R, Shaw O, Wipat A (2007) Qualitatively modelling and analysing gene regulatory networks: a Petri net approach. *Bioinformatics* 23: 336–343.
- Bhalla US, Ram PT, Iyengar R (2002) MAP kinase phosphatase as the locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* 297: 1018–1023.
- Bornheimer SJ, Maurya MR, Farquhar MG, Subramaniam S (2004) Computational modeling reveals how interplay between components of a GTPase-cycle module regulates signal transduction. *Proc Natl Acad Sci U S A* 101: 15899–15904.
- Hoffman A, Levchenko A, Scott ML, Baltimore D (2002) The I $\kappa$ B-NF- $\kappa$ B signaling module: temporal control and selective gene activation. *Science* 298: 1242–1245.
- Huang CY, Ferrell JE (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci U S A* 93: 10078–10083.
- Ferrell JE, Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* 280: 895–898.
- Bailey JE (2001) Complex biology with no parameters. *Nat Biotechnol* 19: 503–504.
- Novere NL, Finney A, Hucka M, Bhalla US, Campagne F, et al. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 23: 1509–1515.
- Arisi I, Cattaneo A, Rosato V (2006) Parameter estimate of signal transduction pathways. *BMC Neuroscience* 7: S6.
- Doi A, Fujita S, Matsuno H, Nagasaki M, Miyano S (2004) Constructing Biological Pathway Models with Hybrid Functional Petri Nets. In *Silico Biology* 4: 271–291.
- Hardy S, Robillard PN (2004) Modeling and Simulation of Molecular Biology Systems using Petri Nets: Modeling Goals of Various Approaches. *J Bioinformatics Computational Biology* 2: 619–637.
- Matsuno H, Tanaka Y, Aoshima H, Doi A, Matsui M, et al. (2003) Biopathways representation and simulation on hybrid functional Petri net. In *Silico Biology* 3: 389–404.
- Li S, Assmann SM, Albert R (2006) Predicting essential components of signal transduction networks: a dynamic model of guard cell abscisic acid signaling. *PLoS Biology* 4: e312–e328.
- Glass L, Kauffman A (1973) The logical analysis of continuous non-linear biochemical control networks. *J Theoretical Biology* 39: 103–129.
- Jong HD, Geiselman J, Batt G, Hernandez C, Page M (2004) Qualitative Simulation of the Initiation of Sporulation in *Bacillus subtilis*. *Bull Mathematical Biology* 66: 261–299.
- Muller M, Obeyesekere M, Mills GM, Ram PT (2007) Network topology determines dynamics of the mammalian MAPK1,2 signaling network: bi-fan motif regulation of C-Raf and B-Raf isoforms by FGFR and MC1R. *FASEB J*; In press.
- Avruch J, Hara K, Lin Y, Liu M, Long X, et al. (2006) Insulin and amino-acid regulation of mTOR signaling and kinase activity through the Rheb GTPase. *Oncogene* 25: 6361–6372.
- Inoki K, Ouyang H, Zhu T, Lindvall C, Wang Y, et al. (2006) TSC2 integrates Wnt and energy signals via a coordinated phosphorylation by AMPK and GSK3 to regulate cell growth. *Cell* 126: 955–968.
- Karbowiczek M, Cash T, Cheung M, Robertson GP, Astrinidis A, et al. (2004) Regulation of B-Raf kinase activity by tuberlin and Rheb is mTOR-independent. *J Biological Chemistry* 279: 29930–29937.
- Kwiatkowski DJ, Manning BD (2005) Tuberous sclerosis: a GAP at the crossroads of multiple signaling pathways. *Human Molecular Genetics* 14: R251–R258.
- Liang J, Shao SH, Xu ZX, Hennessy B, Ding Z, et al. (2007) The energy sensing LKB1-AMPK pathway regulated p27(kip1) phosphorylation mediating the decision to enter autophagy or apoptosis. *Nature Cell Biology* 9: 218–224.
- Ma L, Chen Z, Erdjument-Bromage H, Tempst P, Pandolfi PP (2005) Phosphorylation and functional inactivation of TSC2 by Erk implications for tuberous sclerosis and cancer pathogenesis. *Cell* 121: 179–193.
- Manning BD, Logsdon MN, Lipovsky AI, Abbott D, Kwiatkowski DJ, et al. (2005) Feedback inhibition of Akt signaling limits the growth of tumors lacking Tsc2. *Genes and Development* 19: 1773–1778.
- O'Reilly KE, Rojo F, She QB, Solit D, Mills GB, et al. (2006) mTOR inhibition induces upstream receptor tyrosine kinase signaling and activates Akt. *Cancer Research* 66: 1500–1508.
- David R, Alla H (2005) *Discrete, Continuous, and Hybrid Petri Nets*. Springer.
- Aldana M, Cluzel P (2003) A natural class of robust networks. *Proc Natl Acad Sci* 100: 8710–8714.
- Kauffman S, Peterson C, Samuelsson B, Troein C (2004) Genetic networks with canalizing Boolean rules are always stable. *Proc Natl Acad Sci* 101: 17102–17107.
- Klemm K, Bornholdt S (2005) Topology of biological networks and reliability of information processing. *Proc Natl Acad Sci U S A* 102: 18414–18419.
- Chaves M, Albert R, Sontag ED (2005) Robustness and fragility of Boolean models for genetic regulatory networks. *J Theoretical Biology* 235: 431–449.
- Neve RM, Chin K, Fridlyand J, Yeh J, Bachner FL, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10: 515–527.
- Bray D (1995) Protein molecules as computational elements in living cells. *Nature* 376: 307–312.
- Iyengar R, Birnbaumer L, editors (1989) *G Proteins*. New York: Academic Press.
- Jordan JD, Landau EM, Iyengar R (2000) Signaling networks: the origins of cellular multitasking. *Cell* 103: 193–200.
- Eungdamrong NJ, Iyengar R (2004) Modeling cell signaling networks. *Biology Cell* 96: 355–362.
- Eungdamrong NJ, Iyengar R (2004) Computational Approaches for modeling regulatory cellular networks. *Trends Cell Biology* 14: 661–669.
- Gianchandani EP, Brautigan DL, Papin JA (2006) Systems analyses characterize integrated functions of biochemical networks. *Trends Biochemical Sci* 31: 284–291.
- Blinov ML, Faeder JR, Goldstein B, Hlavacek WS (2006) A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. *BioSystems* 83: 136–151.
- Inoki K, Corradetti MN, Guan KL (2005) Dysregulation of the TSC-mTOR pathway in human disease. *Nat Genet* 37: 19–24.
- Sarbasov DD, Ali SM, Sabatini DM (2005) Growing roles for the mTOR pathway. *Curr Opin Cell Biol* 17: 596–603.
- Chen Y, Rodrik V, Foster DA (2005) Alternative phospholipase D/mTOR survival signal in human breast cancer cells. *Oncogene* 24: 672–679.
- Makris C, Voisin L, Giasson E, Tudan C, Kaplan DR, et al. (2002) The Rb-family protein p107 inhibits translation by a PDK1-dependent mechanism. *Oncogene* 21: 7891–7896.

Software

Open Access

## Rapidly exploring structural and dynamic properties of signaling networks using PathwayOracle

Derek Ruths\*<sup>1</sup>, Luay Nakhleh<sup>1</sup> and Prahlad T Ram<sup>2</sup>

Address: <sup>1</sup>Department of Computer Science, Rice University, Houston, Texas, USA and <sup>2</sup>Department of System Biology, University of Texas MD Anderson Cancer Center, Houston, Texas, USA

Email: Derek Ruths\* - druths@cs.rice.edu; Luay Nakhleh - nakhleh@cs.rice.edu; Prahlad T Ram - pram@mdanderson.org

\* Corresponding author

Published: 19 August 2008

Received: 18 December 2007

BMC Systems Biology 2008, 2:76 doi:10.1186/1752-0509-2-76

Accepted: 19 August 2008

This article is available from: <http://www.biomedcentral.com/1752-0509/2/76>

© 2008 Ruths et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** In systems biology the experimentalist is presented with a selection of software for analyzing dynamic properties of signaling networks. These tools either assume that the network is in steady-state or require highly parameterized models of the network of interest. For biologists interested in assessing how signal propagates through a network under specific conditions, the first class of methods does not provide sufficiently detailed results and the second class requires models which may not be easily and accurately constructed. A tool that is able to characterize the dynamics of a signaling network using an unparameterized model of the network would allow biologists to quickly obtain insights into a signaling network's behavior.

**Results:** We introduce *PathwayOracle*, an integrated suite of software tools for computationally inferring and analyzing structural and dynamic properties of a signaling network. The feature which differentiates *PathwayOracle* from other tools is a method that can predict the response of a signaling network to various experimental conditions and stimuli using only the connectivity of the signaling network. Thus signaling models are relatively easy to build. The method allows for tracking signal flow in a network and comparison of signal flows under different experimental conditions. In addition, *PathwayOracle* includes tools for the enumeration and visualization of coherent and incoherent signaling paths between proteins, and for experimental analysis – loading and superimposing experimental data, such as microarray intensities, on the network model.

**Conclusion:** *PathwayOracle* provides an integrated environment in which both structural and dynamic analysis of a signaling network can be quickly conducted and visualized along side experimental results. By using the signaling network connectivity, analyses and predictions can be performed quickly using relatively easily constructed signaling network models. The application has been developed in Python and is designed to be easily extensible by groups interested in adding new or extending existing features. *PathwayOracle* is freely available for download and use.

### Background

Reconstructing cellular signaling networks and understanding how they work are major endeavors in cell biology. The scale and complexity of these networks, however,

render their analysis using experimental biology approaches alone very challenging. As a result, computational methods have been developed and combined with experimental biology approaches, producing powerful

tools for the analysis of these networks. These tools aid biologists in interpreting existing experimental findings, evaluating hypotheses, enumerating possible biological behaviors, and, ultimately, in quickly designing experiments that maximize the amount of useful information gained. By assisting biologists in maximizing the amount of information obtained from their experiments through improved experimental design and more thorough analysis of results, computational tools increase the pace of scientific discovery.

Biological network analysis can generally be classified as either *structural* or *dynamic* [1]. Structural analysis provides insights into global properties of the network, among them decomposition of the network into functional modules (e.g., [2]), enumeration of signaling paths connecting arbitrary protein pairs (e.g., [3-5]), and the identification of key pathways that determine the behavior of the network (e.g., [2,6-10]). Dynamic methods, on the other hand, simulate the actual propagation of signals through a network by predicting the changes in the concentration of signaling proteins over time. These predictions will be of varying degrees of resolution and accuracy, depending largely on the accuracy and level of detail of the model from which they are produced.

The prevailing methods for dynamic analysis involve systems of ordinary differential equations (ODEs) [11,12]. These approaches require kinetic parameters for the individual biochemical reactions involved in the signaling process. This requirement often poses a significant hurdle for researchers as the numerical values of such parameters are difficult to obtain and may be the object of the researcher's project in the first place. In [13], we presented a novel signaling network simulation method which uses a non-parametric Petri net model of network to predict the signal flow under various experimental conditions. Our simulation method uses a novel technique to approximate the interaction speeds and predicts the qualitative behavior of the signaling network dynamics.

The advantage of our method over ODEs is the wide availability of connectivity-based models of signaling networks, and the relative speed with which they can be constructed. Numerous databases exist which catalog known signaling interactions (e.g., [14-16]). Thus, the existence and type (activating or inhibition) of an interaction can often be inferred directly from literature and/or these databases. This presents a stark contrast to the kinetic parameters required by ODEs, the numerical values for many of which must be determined experimentally for each experimental condition and cell line of interest [2].

In this paper, we present the software tool *PathwayOracle*, an integrated environment for connectivity-based structural and dynamic analysis of signaling networks, supporting

- visualization of signaling network connectivity;
- two versions of the simulation method described in [13] where
  - the first allows prediction of signal flow through a given network for a specific experimental condition, and
  - the second predicts the difference in signal flow through a given network induced by two different experimental conditions;
- enumeration of the paths connecting arbitrary pairs of nodes in the network; and
- visualization of experimental concentration data on the signaling network display.

In future releases we plan on expanding capabilities in all three areas of analysis – dynamic, structural, and experimental – with a focus on providing effective ways of integrating results from each together.

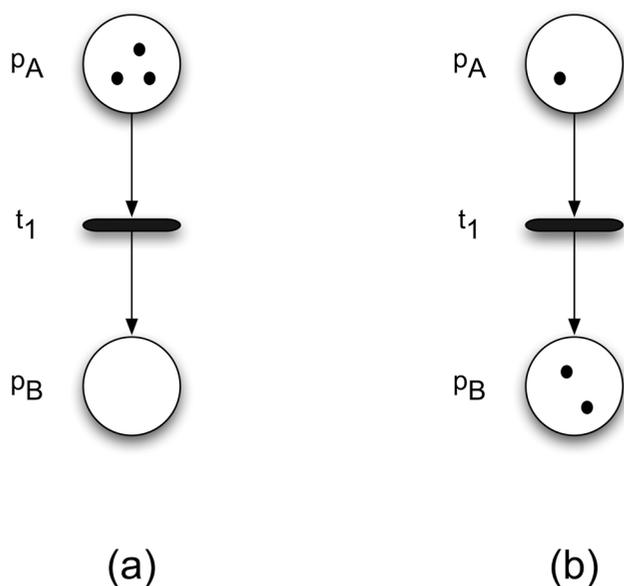
*PathwayOracle* has been designed in a modular fashion in order to facilitate extension of existing capabilities and the addition of new features.

Since *PathwayOracle*'s most distinctive analytical capability involves the signaling Petri net simulator, a new dynamic analysis technique for signaling networks, we first provide an overview of the signaling Petri net modeling approach. Then in subsequent sections, we focus on *PathwayOracle* and explain the architecture and core concepts underlying the tool and then examine the individual features, how they can be used, and how they compare to existing tools.

### **The Signaling Petri Net Simulator**

Petri nets provide a graphical and executable model of processes in which information or material flows among a series of places or entities [17]. A Petri net consists of places, transitions, and tokens (see Figure 1). Quantities of tokens are assigned to individual places. This assignment is called a *marking*. As Figure 1 illustrates, the network flow is modeled by the reassignment of tokens to individual places in the Petri net in response to transition firings.

A signaling Petri net is an extension of the Petri net formalism to model a signaling network. Places are signaling



**Figure 1**  
**An example of how tokens move among places.** In a Petri net, quantities of tokens are assigned to places. In (a), three tokens are assigned to place  $p_A$  and zero tokens are assigned to place  $p_B$ . The two places are connected by a transition,  $t_1$ . The arcs in and out of  $t_1$  indicate the direction in which tokens move. When  $t_1$  fires, it moves some number of tokens from  $p_A$  and puts them in  $p_B$ . In (b), transition  $t_1$  has fired and moved two tokens from  $p_A$  to  $p_B$ .

proteins and transitions implement directed protein interactions; each transition models the effect of a source protein on a target protein. The marking of (number of tokens in) protein  $p$  at time  $t$  is interpreted as the activity-level of that protein – the number of activated molecules of that type. Figure 2 shows the correspondence between a signaling network and a signaling Petri net model.

The signaling Petri net simulator models signal flow as the pattern of token accumulation and dissipation within proteins over time in the Petri net. Through transition firings, the source can influence the marking of (the number of tokens assigned to) the target, modeling the way that signals propagate through protein interactions in cellular signaling networks.

In order to overcome the issue of modeling reaction rates in the network, signaling dynamics are simulated by executing the signaling Petri net (SPN) for a set number of steps (called a *run*) multiple times, each time beginning at the same initial marking. For each run, the individual signaling rates are simulated via generation of random orders of transition firings (interaction occurrences). When the results of a large enough number of runs are averaged together, we find that the change in distribution

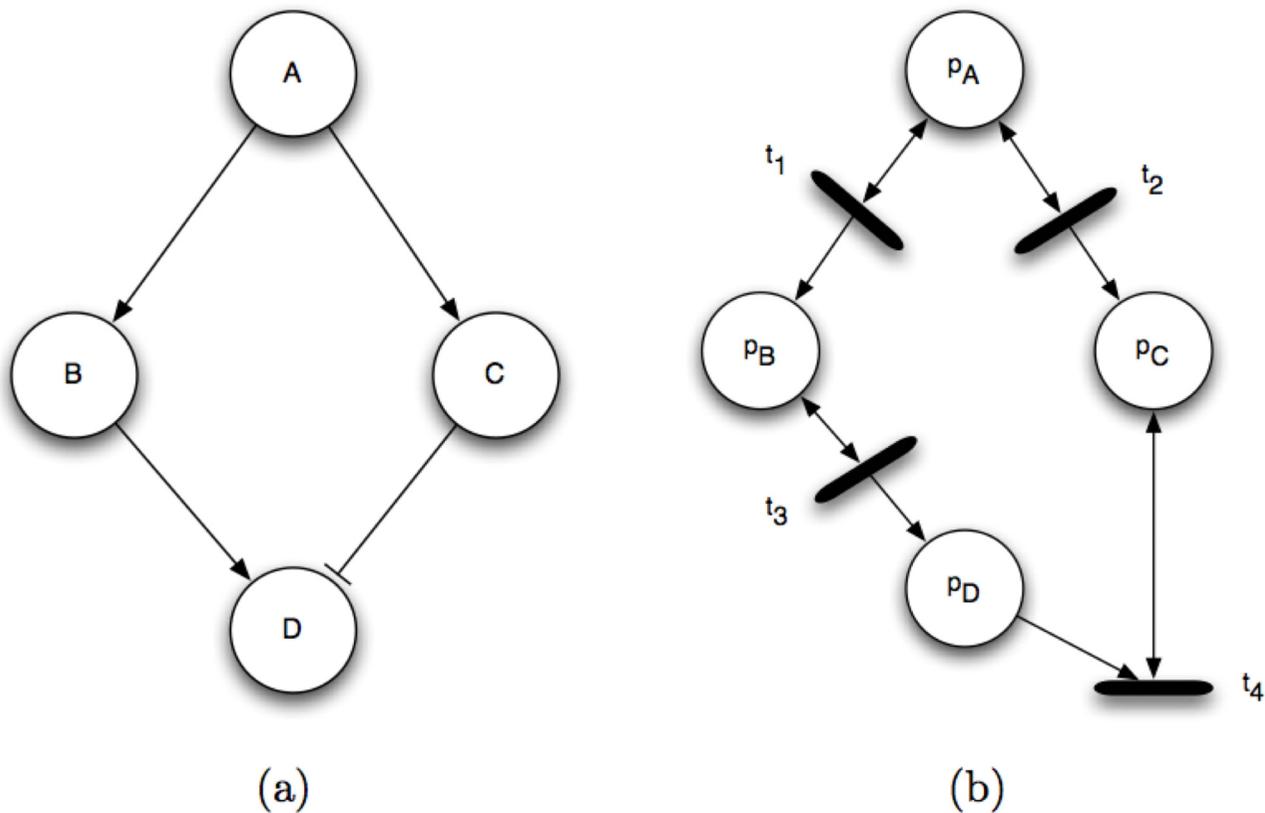
of tokens in the network correlate with experimentally measured changes in the activity-levels of individual proteins in the underlying signaling network. In essence, the tokenized activity-levels computed by our method should be taken as abstract quantities whose changes over time correlate to changes that occur in the amounts of active proteins present in the cell. It is worth noting that some of the most widely used experimental techniques for protein quantification – western blots and microarrays – also yield results that are treated as indications, but not exact measurements, of protein activity-levels within the cell. Thus in some respects, the predictions returned by our SPN-based simulator can be interpreted like the results of a western blot or microarray experiment looking at changes relative to "control".

During a simulation run, the simulator imposes a strict ordering on transition firing such that it creates a two-time scale simulation. The smaller time scale is discretized as the firing of a single transition. This unit is referred to as the *firing* time scale. Firing steps are nested within a larger time scale, called time *blocks*, within which each transition is fired exactly once. The values returned by the simulator are the averaged token-counts for each protein at each time-block (across all runs).

Figure 3 provides a small example of a simulation run whose duration is two time blocks. As mentioned previously, within a given time block, each transition fires exactly once. Thus, in the table (Figure 3(c)), there is one column for each transition in each time block. The ordering of the transitions is shuffled in each time block in order to sample a different set of signaling rates within the networks.

In the first time block, transition  $t_2$  fires first: it reads 2 tokens out of *Grb2* and places 2 additional tokens in *Ras*. Transition  $t_1$  fires second, reading 3 tokens out of *Grb2*. Transition  $t_3$  is evaluated last. The final marking for the network, highlighted as the red column in block 1 is used by the simulator as the marking for that block when averaging across runs.

At the conclusion of block 2, compare the values highlighted in red in the Initial column and at the end of both blocks. Note how the distribution of tokens have changed over the course of the simulation. *Grb2* has the same number of tokens, implying that its activity-level has remained unchanged – this is consistent with the signaling network since no activating or inhibiting edges affect it in the model. *AKT*'s token-count has risen, consistent with the fact that it is only activated in the signaling network. *Ras*'s token-count has fallen which is one plausible behavior of the system since it is activated by *Grb2*, but inhibited by *AKT*.

**Figure 2**

**An example signaling network and its corresponding Petri net.** An example signaling network (a) and its corresponding Petri net (b). Each signaling protein in the network, A, B, and C, is designated as a place  $p_A$ ,  $p_B$ , and  $p_C$ . A signaling interaction becomes a transition node and its input and output arcs. Note that the connectivity for an activating edge differs from that of an inhibitory edge.

### Implementation

*PathwayOracle* is written in Python [18]. The user experience is oriented around visualization of and interaction with three main types of data: the signaling network, markings, and paths. At any given time, one signaling network is open, which is the basis for all analyses. Any simulation or concentration data is loaded and inspected as markings. Currently all static analyses revolve around paths, which are the third data type. In the following subsections, these individual data types and the user interfaces to them are discussed in more detail.

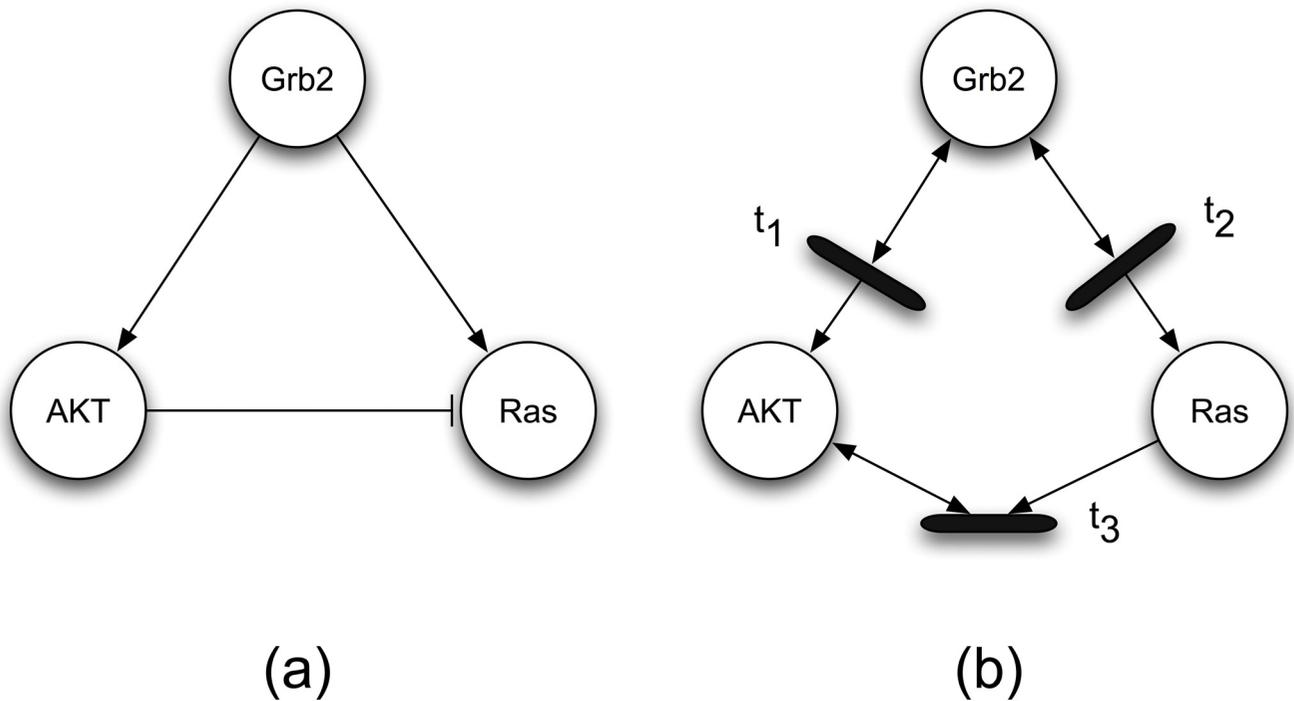
### The Signaling Network Model

While the implementation of our methods use the signaling Petri net model discussed in an earlier section of this paper, we provide a simpler and more convenient representation of the network to the user which omits the internal topology of the transitions and allows the user to specify interactions simply as either activating or inhibit-

ing. Thus, for the remainder of this paper we use the following definition of the signaling network which is consistent with the experience the user will have when working with *PathwayOracle*. The signaling network connectivity is a directed graph  $G = (V, E)$  where

- $V$  is the set of nodes, which are signaling proteins and complexes (hereafter referred to collectively as *signaling nodes*) and
- $E$  is the set of edges, which are signaling interactions. Each edge is of one of two types:  $u \rightarrow v$  for activation and  $u \vdash v$  for inhibition.

Within *PathwayOracle*, each signaling node has a name, unique within the network. A signaling edge has no properties besides its type and is only defined by its *source* and *target*.



Proteins	Initial	<i>Block 1</i>			<i>Block 2</i>		
		t <sub>2</sub>	t <sub>1</sub>	t <sub>3</sub>	t <sub>1</sub>	t <sub>3</sub>	t <sub>2</sub>
<i>Grb2</i>	5	5	5	5	5	5	5
<i>AKT</i>	3	3	6	6	10	10	10
<i>Ras</i>	6	8	8	4	4	0	2

(c)

**Figure 3**

**An example signaling Petri net simulation.** (a) is the signaling network being simulated. (b) is the signaling Petri net that models that signaling Petri net. The table in (c) provides the markings for the Petri net over the course of a simulation run whose duration is two time blocks. The proteins are given the initial marking shown in the *Initial* column. Each subsequent column corresponds to a single time step during which one transition fired, producing a new marking of the network. The bold number in each column indicates which protein's marking was affected by the transition that fired in that time step. The red columns – always the last time step in the block – highlight the markings whose values would be averaged and used as part of the final result. These red columns are the sources of the markings that *PathwayOracle* reports.

In order to facilitate the rapid construction of such signaling network models, we devised a file format called the *Connectivity Format*. It is capable of expressing both general networks as well as paths. When representing a network in the format, as shown in the example in Figure 4(b), one signaling interaction is written on a line with the format

$$u \rightarrow v \text{ or } u - | v$$

where  $u$  is the name of the source signaling node and  $v$  is the name of the target signaling node. Each node is taken to represent the active form of the protein it is named for. Thus, from the example above, the interaction PI-3-K $\rightarrow$ AKT means that the active form of PI-3-K increases the activity-level of AKT whereas the interaction PTENAKT means that the active form of PTEN decreases the activity-level of AKT. While these types of unparameterized relationships can be represented in SBML, SBML was designed for encoding much more information than just connectivity [19]. As a result, we deemed it appropriate to design a more concise format for our purposes. However, in a future release, *PathwayOracle* will support loading and saving in the SBML format.

At a given point in time, only one signaling network can be open in *PathwayOracle*. The main window displays a

graphical representation of the network. The layout of the network can be modified by dragging nodes or by *shift*-clicking on edges to create, remove, or move waypoints. These layouts can be saved with the network and loaded again.

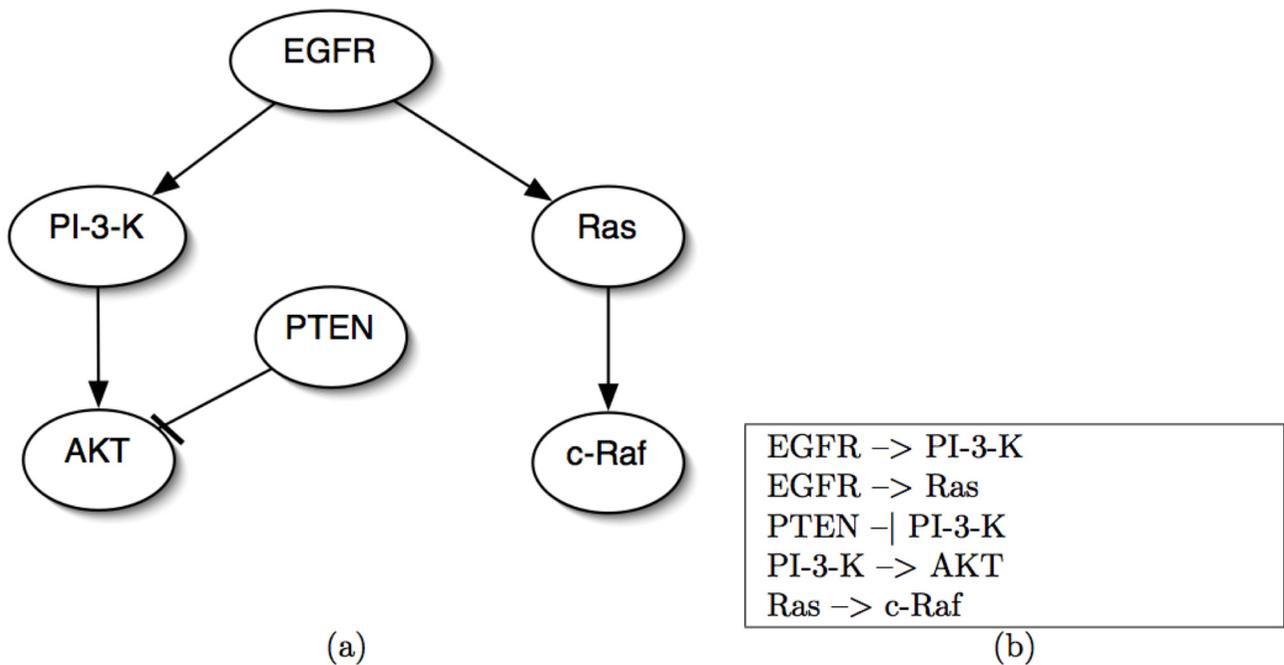
**Signaling Network Markings**

In signaling networks, signal flow is measured and quantified as the fluctuation of concentrations of various forms of signaling proteins over time. In *PathwayOracle*, we model concentrations using the concept of a network *marking*, which was adapted from Petri nets in which it was first used [9].

**Markings**

In *PathwayOracle*, a marking,  $\mu$  is an assignment of real values to the nodes of a signaling network such that every signaling node receives a value. Earlier, the concept of a marking was introduced as the assignment of tokens to protein places in the signaling Petri net. In a signaling Petri net, tokens are discrete. In *PathwayOracle*, a marking is an average of the markings from many independent simulation runs, which gives rise to the real, rather than integral values, assigned by the marking.

As discussed earlier, the value of the marking of a signaling node,  $\mu(v)$ , can be interpreted as an estimate of the



**Figure 4**  
**An example of a Network in the Connectivity Format.** (a) A graphical representation of a signaling network's connectivity. (b) The signaling network in (a) written in the *Network Connectivity Format*.

concentration or change in concentration of the active form of the signaling protein  $v$  (we call the amount of the active form of the signaling protein its *activity-level*). The two different versions of the simulator generate markings with these different meanings. The first simulator predicts the signal flow due to an experimental condition and generates markings whose values are taken to represent the actual activity-level of signaling protein present over the assumed basal levels. The second version of the simulator predicts the difference in signaling due to changing experimental conditions. The values assigned by markings produced by this simulator correspond to the *change* in the activity-level of the protein induced by the change in experimental condition. This will be discussed further in the Results and Discussion section.

**Marking Series**

In order to model signal *flow*, a single marking is not enough since it only provides a single snapshot of concentrations throughout the network. A *marking series* is an sequence of markings,  $(\mu_1, \mu_2, \dots, \mu_T)$  in which the marking  $\mu_i$  is a snapshot of the concentration distribution at time step  $t$ . Thus, it is possible to see how the activity-level of protein  $v$  changed by plotting the values  $\mu_1(v), \mu_2(v), \dots, \mu_T(v)$ . *PathwayOracle* provides the ability to do this.

*PathwayOracle* supports loading a marking series dataset from *comma-separated value (.csv)* files. As shown in Figure 5(a), the file has a header row which specifies, for each column, the name of the molecule whose concentration values will appear in that column. Each subsequent row contains the value assignments for a marking: the second row contains the marking for time step 1, the third row contains the marking for time step 2, and so on.

**Marking Groups**

In many experiments, the activity-level of various proteins are sampled at different time points and under different

experimental conditions. Since the *marking series* is not able to represent changes due to different experimental conditions, we introduced the more general concept of a *marking group* in which each marking can correspond to an arbitrary activity-level distribution. Each marking is given a descriptive label that can be used to identify the conditions under which the activity-level was sampled.

Like the marking series, a marking group is loaded from a *.csv* file. However, unlike the marking series in which each row corresponds to a time step, in the marking group, each row corresponds to an independent marking (experimental condition). As shown in Figure 5(b), the first row is a header row specifying the molecule names for each column, the first column specifies the names for the individual markings (experimental conditions).

**The Marking Manager**

*PathwayOracle* includes a specific user-interface, the *Marking Manager*, designed to manage the three different types of markings. The Marking Manager provides a central interface within which it is possible to view all markings loaded and inspect them in ways that are relevant to their type (marking, marking series, or marking group). The specific ways in which markings can be inspected will be discussed further in the *Results* section.

**Signaling Paths**

The current structural analysis capabilities available in *PathwayOracle* allow inspection of signaling paths within the network. A signaling path  $p$  is a sequence of nodes,  $(v_1, v_2, \dots, v_k)$  where  $v_i \in V \forall 1 \leq i \leq k$ , and  $(v_i, v_{i+1}) \in E \forall 1 \leq i < k$ . In this case, we say that node  $v_1$  is the source of path  $p$ , and node  $v_k$  is the *target* of  $p$ . Given a path, a variety of statistics may be of interest to the user. Additionally, it may be useful to view the path within the larger network. *PathwayOracle* provides these capabilities which will be discussed in the Results and Discussion section.

mTOR/raptor,	AKT,	EGFR,	RSK
0.3,	0.2,	0.1,	1.1
...	...	...	...
2.1,	0.001,	0.1,	1.5

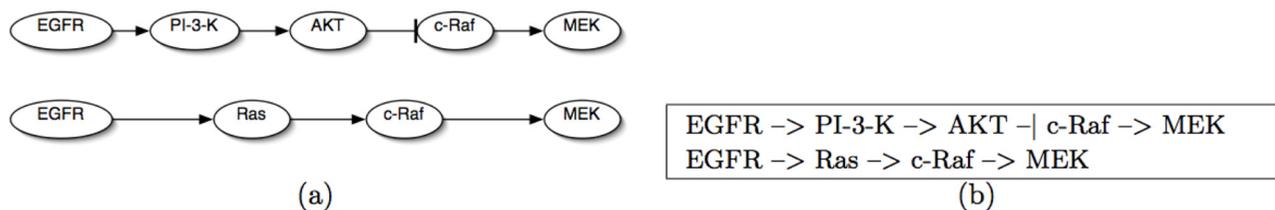
(a)

,	mTOR/raptor,	AKT,	EGFR,	RSK
DMSO,	0.3,	0.2,	0.1,	1.1
...	...	...	...	...
EGF_30min,	2.1,	0.001,	0.1,	1.5

(b)

**Figure 5**

**Examples of marking series and group file formats.** (a) An example marking series dataset in the *comma-separated value* file format. The first row specifies the signaling proteins whose concentrations were measured. Each row thereafter specifies the concentration for a given time step: row  $i$  specifies the concentrations for each signaling protein at time step  $i - 1$ . (b) An example marking group dataset in the *comma-separated value* file format. The first row specifies the signaling proteins whose concentrations were measured. The first column specifies the names for each marking in the group dataset. The numbers in each row specify the concentration measured for each signaling protein in that marking.

**Figure 6**

**An example of a Path in the Connectivity Format.** (a) A graphical representation of two signaling paths. (b) The signaling paths in (a) represented in the *Connectivity Format*. Each line corresponds to a single signaling path.

Sets of paths can be saved to a file and loaded back into a session. Like networks, paths are also stored in the Connectivity Format. When representing a set of paths, as shown in Figure 6, the full node names and the edge types are written so that all path information is directly available within the file itself. One line contains one path.

## Results

*PathwayOracle* provides a variety of tools for analyzing the structural and dynamic properties of a signaling network based on its connectivity. While its main differentiating feature is the ability to predict signal flow through a network using only the connectivity of the signaling network, *PathwayOracle* also provides the ability to visualize the network, analyze its connectivity, and inspect concentration-based experimental data.

With the exception of the signaling Petri net simulator, *PathwayOracle*'s features can be found in various combinations in other tools. Figure 7 provides a matrix of the features and capabilities of several tools most commonly used for signaling network analysis. While other tools support a variety of simulation techniques, *PathwayOracle*, alone, provides non-parameterized simulation capabilities. It is worth noting that the commercial software package CellIllustrator [20] provides Petri net-based simulation capabilities. The difference between CellIllustrator and *PathwayOracle* Petri net approaches is the extensive set of kinetic parameters required by CellIllustrator in order to simulate a biological system. In this regard, hybrid functional Petri nets, the underlying technology used by CellIllustrator, are not significantly different from ODEs.

Another important distinguishing characteristic of *PathwayOracle* is the combination of features that it supports. Biological network analysis is a multi-faceted process that may involve structural, dynamic, and data analysis in parallel. Whereas other tools tend to focus on one or two of these general areas of analysis, we considered it important for *PathwayOracle* to incorporate all three in order to pro-

vide the researcher a single environment in which all their analysis could be done. In future releases we plan to increase *PathwayOracle*'s support for all three of these directions of investigation: structural, dynamic, and data analysis.

In the remainder of this section, we discuss the features currently available in *PathwayOracle*.

### Network Visualization

As in many other computational analysis tools for signaling networks (e.g., [20,21]), an interactive graphical representation of the signaling network connectivity is at the center of the *PathwayOracle* interface. The main window provides a visualization of the signaling network connectivity. This visualization interface allows the user to edit the layout of the network by clicking on and dragging nodes and by *shift*-clicking on edges to create, remove, or move waypoints. Waypoints are points that lie on an edge. Holding down *shift* will display all edge waypoints. Existing waypoints can be dragged to change the path that an edge follows. Right-clicking on a waypoint will remove it. Left-clicking on a straight segment of the edge will create a new waypoint.

The network visualization also provides a view onto which path and experimental data analysis may be mapped. As will be discussed in subsequent sections, selected paths may be highlighted in this view and markings from experiments can set the colorings of individual nodes.

### Network Signal Flow Simulation

The main feature differentiating *PathwayOracle* from other tools, such as CellDesigner [20] and COPASI [22], is its ability to simulate signal flow using an unparameterized signaling network model. Simulations can be performed in two different ways. In the first (*Single Simulation*), the simulator predicts the signal flow through the network for a specific experimental condition. In the second (*Differential Simulation*), the simulator predicts the difference in

Analysis Type	Features	CellDesigner	CellIllustrator	CellNetAnalyzer	COPASI	Cytoscape	Matlab SB Toolkit	PathwayOracle
	Open Source	-	-	✓	✓	✓	-	✓
	Visual Network Editor	✓	✓	✓	-	✓	-	✓
Experimental Data Analysis	Microarray Visualization	-	-	-	-	✓	-	✓
	Microarray Analysis	-	-	-	-	✓	-	-
Structural Analysis	Structural Statistics	-	-	✓	-	✓	-	-
	Path Finding	-	-	✓	-	✓	-	-
	+/- Path Finding	-	-	✓	-	-	-	✓
	Flux Analysis	-	-	✓	✓	-	-	-
	Boolean Analysis	-	-	✓	-	-	-	-
Dynamic Analysis	ODE Simulation	✓	-	-	✓	-	✓	-
	Hybrid PN Simulation	-	✓	-	-	-	-	-
	Signaling PN Simulation	-	-	-	-	-	-	✓

**Figure 7**

**A comparison of features supported by tools commonly used for signaling network analysis.** The table shows the features and analytical capabilities supported by different tools commonly used for the analysis of signaling networks. Tools included in the comparison are: CellDesigner [20], CellIllustrator [24], CellNetAnalyze [25], COPASI [22], Cytoscape [21], the System Biology Toolkit for Matlab [26], and PathwayOracle.

signal flow due to two different experimental conditions on the same network. These simulation methods themselves are described in [13]. Here we focus on how simulations are configured, run, and analyzed.

Whereas the consensus networks typically represent the connectivity in normal cells, many experiments are conducted on abnormal cells in which oncogenic mutations, gene knockouts, and pharmacological inhibitors have altered the behavior of various signaling nodes in the network. In *PathwayOracle* users can model these cell- and experiment-specific conditions by specifying each signaling node as either *High*, *Low*, or *Free*. The *High* state models any condition under which a protein's activity-level is held high for the duration of the experiment. This may be due to external stimulation or a known mutation in the protein that makes it constitutively active, for example. Similarly, a *Low* state models any phenomenon that forces a protein to have a persistently suppressed activity-level. This may be due to mutations that render the protein inactive, gene knockouts, or pharmacological inhibitors that force the activity-level of the protein low. In general, most signaling nodes will be *Free*, which means that their activity-level is unconstrained throughout the simulation. Only those nodes designated as *High* or *Low* will have their activity-level fixed for the duration of the simulation.

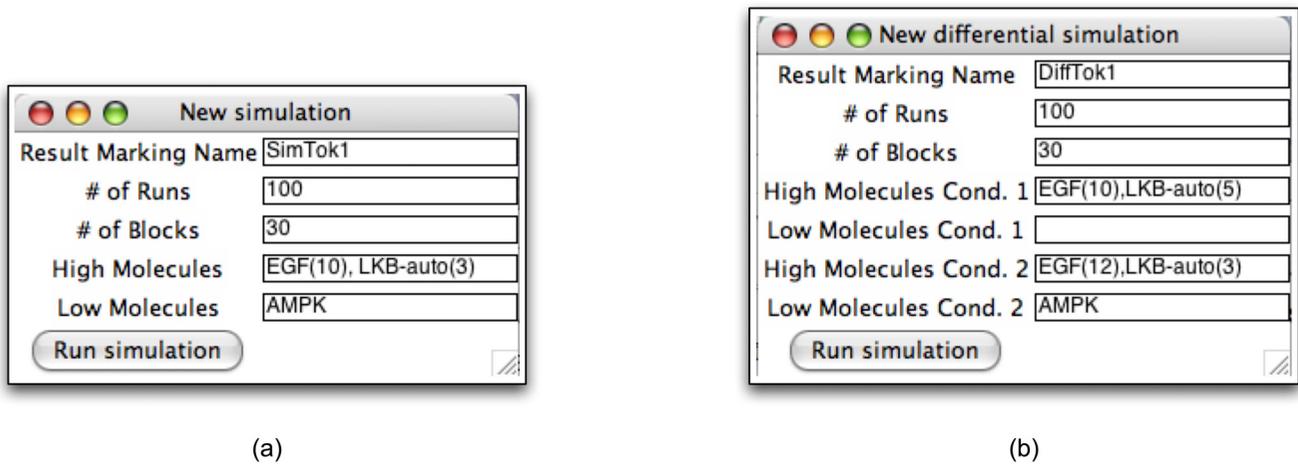
In order for a protein to be held high during the simulation, it is necessary to indicate the initial activity-level that the protein will be elevated to. This is done by specifying the number of tokens that the protein will receive. Since a protein with a *High* state cannot be inhibited (even if inhibitory edges target it in the actual network), the pro-

tein's activity level will never fall below this initial value. The initial value for a High protein is indicated by placing it in parentheses next to the protein's name, as shown in Figure 8. Two other parameters that must be specified for a simulation are:

- the number of simulation runs to perform and
- the number of time blocks

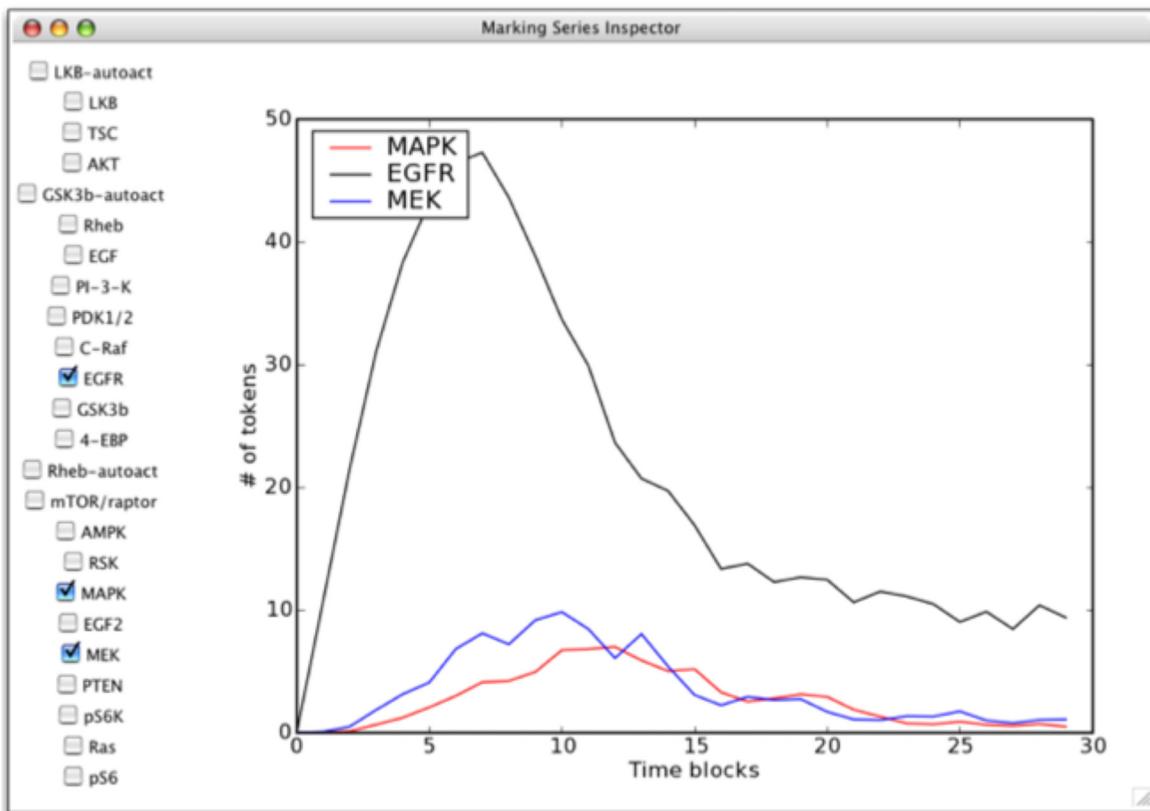
The number of runs sets the number of independent simulations whose time block markings are averaged together to yield the overall simulation markings. In general, using more runs is a tradeoff between reliability of the results and simulation speed. In practice, the number of runs needed is dependent on the signaling network model and should be selected by observing the reproducibility of the simulation results. An appropriate number of iterations will be large enough so that for a given experimental condition, the results are very similar across multiple simulations.

The time block, as discussed earlier, is a fundamental unit of time in the simulator. The appropriate number of time blocks for which to simulate will vary depending on the size of the signaling network and the scale of the network behavior of interest. Generally it should be selected by running simulations for a variety of time block values and determining which yields the most biologically reasonable activity-level changes for a known protein. While this is a manual process in the current version of *PathwayOracle*, we are investigating automated methods for estimat-



(a)

(b)



(c)

**Figure 8**

**The tokenized simulator user interface.** (a) The setup window for the tokenized simulator. The simulation is being configured to have two High nodes, EGF and LKB-auto. EGF will be initialized with a token-count of 10, LKB-auto with a token-count of 3. The token-count of AMPK will be zero for the duration of the simulation. (b) The setup window for the differential simulator. Two different scenarios are being compared through simulation: different token assignments are being tried with EGF and LKB-auto, with and without AMPK being fixed low. (c) The plot window for the marking series generated by a simulation. Observe that the signaling nodes whose activity-levels are plotted correspond to those selected in the checklist directly to the left of the plot.

ing the number of time blocks by training against experimental time series data.

In *PathwayOracle*, the setup window for the *Single Simulation* (see Figure 8(a)) prompts the user for a single experimental condition. The setup window for the *Differential Simulation* (see Figure 8(b)) prompts the user for two experimental conditions. Both simulators produce a marking series. The tokenized simulation marking series corresponds to the activity-level time series predicted for the specified experimental condition. The differential simulation marking series corresponds to the change in activity-levels over time produced by switching from experimental condition 2 to experimental condition 1. The marking series produced by a simulation can be accessed through the Marking Manager. Choosing to *inspect* a marking series will present the user with a blank plot. By selecting signaling nodes, the plot is populated by the marking series values for individual nodes over time, as shown in Figure 8(c).

While this plot generation capability exists in many other dynamic simulation tools, the simplicity of the model used for simulation and the speed with which a simulation runs set *PathwayOracle* apart from other tools which require specification of the numerical values of kinetic parameters for each reaction in the network of interest (e.g., [20,22]). *PathwayOracle*, because of its novel approach, does not have such requirements. It is worth noting, however, where *PathwayOracle* provides approximations of signal flow, an ODE generates the actual concentration changes using extremely detailed and accurate models of the underlying biochemistry. The simulators in *PathwayOracle* provide an attractive, time- and resource-saving alternative to this more exhaustively parameterized techniques. In particular, *PathwayOracle's* features will benefit researchers interested in quickly assessing characteristics of signal flow in their network.

For some networks, biologists will have partial knowledge of kinetic parameters or of other biological details which the signaling Petri net model does not, at present, consider. By integrating this knowledge into the simulator, it may be possible to improve the simulator's predictions. We identify this as a direction for future investigation. As the signaling Petri net simulator is extended, these new capabilities will be incorporated in future releases of *PathwayOracle*.

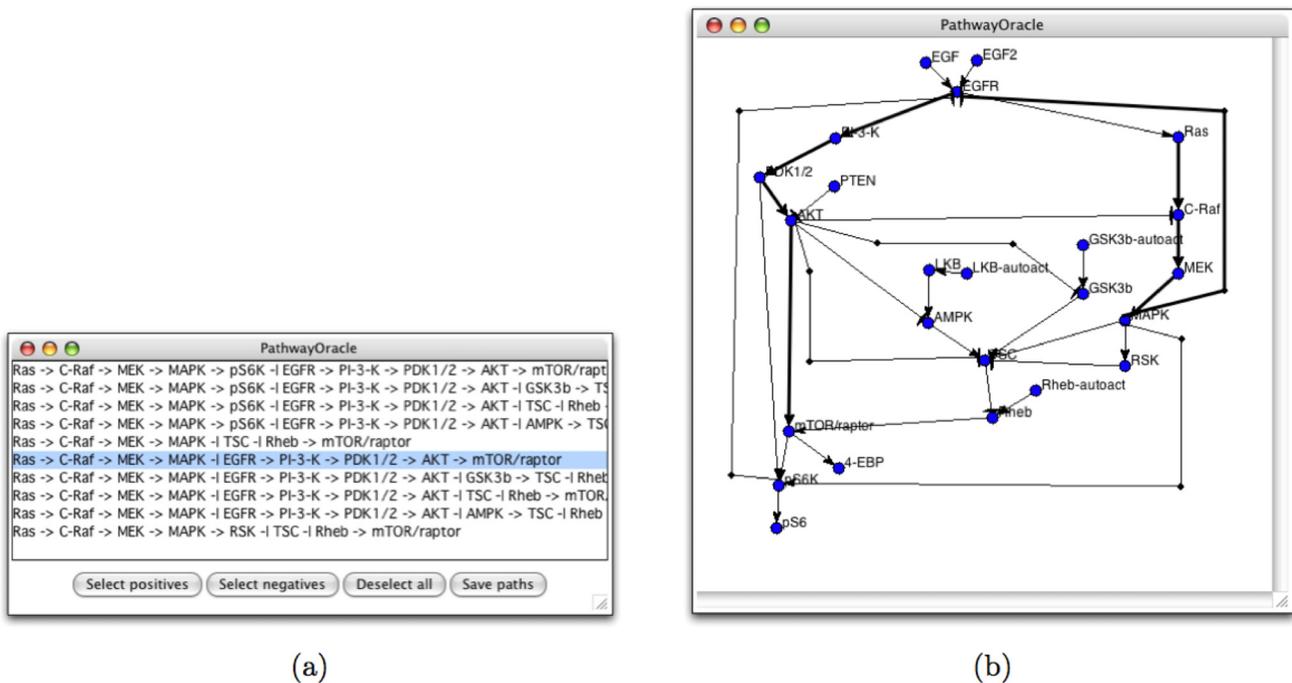
### Signaling Path Analysis

The use of the simulators and plotting tools allows the user to observe trends in the activity-level of individual signaling nodes over time. Since the activity-level of a node is determined by the activity-level of other nodes in the network, the activity-level time series of a node may be

explained by changes in the activity-level history of nodes upstream of it. In order to investigate such indirect interactions, it is useful to enumerate all the paths leading from a specific protein to the protein of interest. *PathwayOracle* provides this capability. Additionally, it provides various statistics on the set of paths linking two signaling nodes as well as a classification of the effect of each path as either *coherent* or *incoherent* (e.g. [23]). A coherent path is a directed series of interactions that leads from  $x$  to  $y$  such that an increase in the activity-level of  $x$  causes an increase in the activity of  $y$  and a decrease in the activity-level of  $x$  causes a decrease in the activity-level of  $y$ . An incoherent path is a directed series of interactions leading from  $x$  to  $y$  such that an increase in the activity-level of  $x$  causes a decrease in the activity-level of  $y$  and a decrease in the activity-level of  $x$  causes an increase in the activity-level of  $y$ . It is possible to classify a path  $p$  as either coherent or incoherent by counting the number of inhibitory edges along  $p$ . A path with an even number of inhibitory edges is coherent; a path with an odd number of inhibitory edges is incoherent [5]. This logic is assumed in *PathwayOracle*. All simple paths (paths without loops) connecting two specified signaling nodes are enumerated by an exhaustive depth-first search. These paths then are classified as either coherent or incoherent, and presented to the user for further inspection in a window similar to the one shown in Figure 9(a). When a path is selected in the results window, it is highlighted in the main window, allowing the user to evaluate it within the context of the complete network (see Figure 9(b)).

### Experimental Data Analysis

A model of the connectivity of a signaling network makes it possible to identify components of the model that are inconsistent with experimental data or visa versa. *PathwayOracle* enables this kind of analysis by allowing users to load experimental concentration data and visualize it both as a heatmap (see Figure 10(a)) or superimposed on the network view (see Figure 10(b)). Several other software tools provide similar capabilities (e.g., [21]). In *PathwayOracle*, experimental concentration data is loaded as a marking group in which a single marking corresponds to a condition for which concentrations were sampled. Figure 10(a) shows a marking group with 24 conditions (rows). The concentration of seven signaling proteins were sampled for each condition. This is the heatmap view for the marking group. When a specific marking in the group is selected, the colors for that marking are applied to the network view. This is particularly useful when assessing whether the experimental data is consistent with the interactions in the model. In Figure 10, the MDA231-B-DMSO2 marking has been superimposed on the network. We can see that RSK has a relatively low concentration despite the high concentration of MAPK. Given that, in the model, RSK is activated by MAPK, this combi-

**Figure 9**

**The path interrogation user interface.** (a) The result window enumerating the set of all paths between *Ras* and *mTOR/raptor*. (b) The main network view showing the selected path highlighted.

nation of activity-levels seems unlikely to occur. Such an inconsistency suggests that there may be other signaling interactions contributing to the overall activity-level of RSK. Such an insight can help a researcher quickly identify areas where the model or experimental results need to be re-evaluated or improved.

#### Future Directions

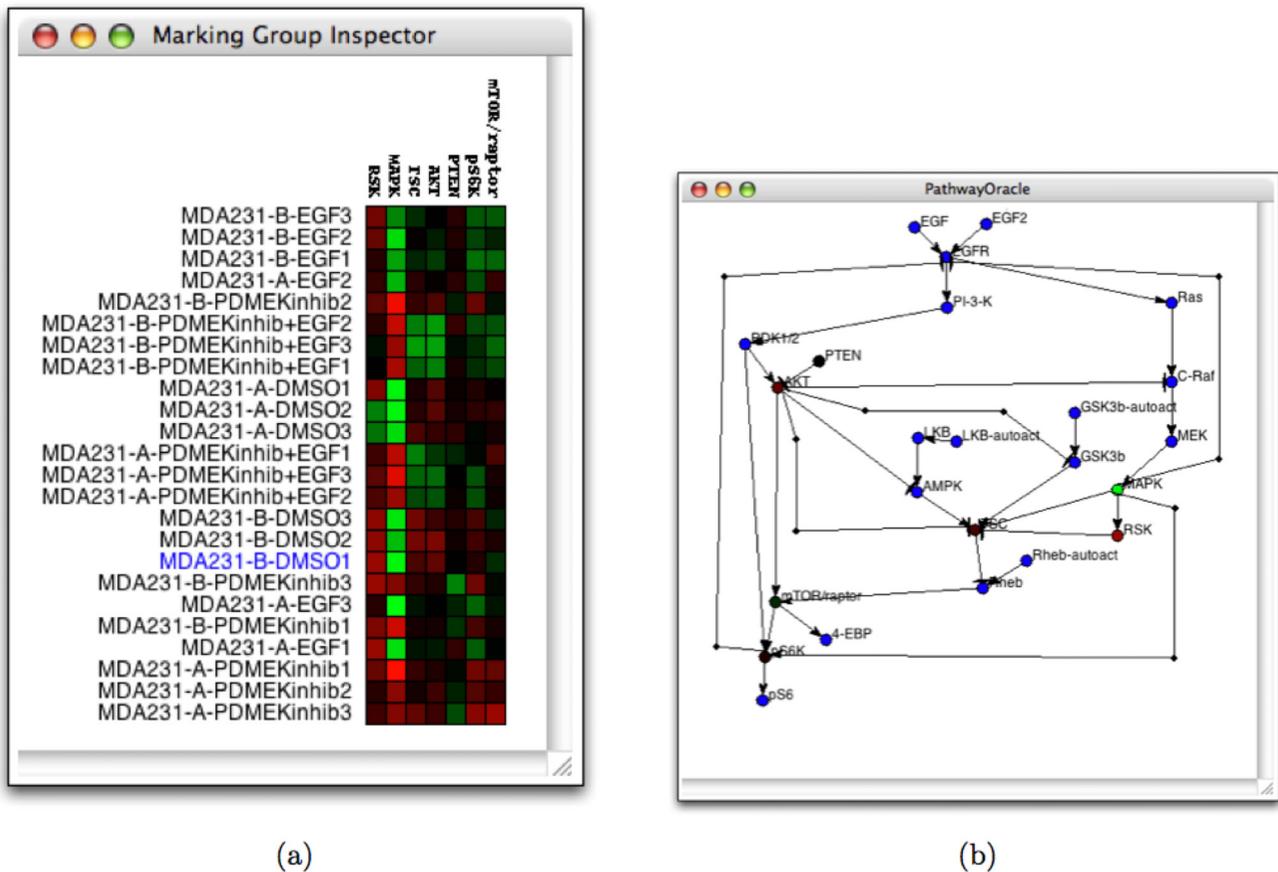
Our goal is to develop *PathwayOracle* into an integrated and expansive suite of tools that allow the biologist to extract as much information as possible from models of signaling network connectivity and experimental data relating to those models. We consider future directions for *PathwayOracle* to fall into several categories: network construction, network augmentation, experimental and computational analysis integration, and architecture.

One of the benefits of working with connectivity models of signaling networks is the abundance of databases and other online resources that publish connectivity-level data. Future versions of *PathwayOracle* will have support for querying such databases for connectivity components and, ultimately, for automated connectivity construction based on a set of signaling nodes specified by the user.

Analysis of network connectivity and topology is increasingly relevant to biological research. We intend to expand *PathwayOracle*'s structural analysis features to include the ability to search for and identify motifs in the signaling networks.

Network connectivity can also be inferred from experimental data, which provides another direction for research and development. By using experimental results to identify inconsistencies between experimental results and the current network model, it may be possible for *PathwayOracle* to augment the network with new connectivity based on hints supplied by experimental results. At present only experimental concentration data is supported. However, as experiments produce more information beyond concentration profiles of signaling nodes, we plan to expand the experimental data that *PathwayOracle* can load, visualize, and use as part of network analyses.

Experimental results can also provide computational analysis methods information that can improve their final predictions or decompositions. Taking advantage of the additional, potentially obfuscated, information present in experimental results to improve the results returned by computational tools is a major goal for future versions of *PathwayOracle*.

**Figure 10**

**The marking group user interface.** (a) The heat map visualization of a marking group. The selected marking, MDA231-B-DMSO1, is highlighted in blue. (b) The color distribution for the selected marking in the group is applied to the network view in the main window. Note that signaling nodes for which values were not given are not assigned a color on the valid red to green spectrum.

A longer term direction for *PathwayOracle* is the integration of transcriptional and metabolic network analysis. In the biological systems of interest, the behavior of any one of these networks is dependent on the characteristics of the other two. As a result, developing a complete understanding of signaling, transcriptional regulation, or metabolism depends in part on integrating knowledge from the others. Finally, an ongoing priority in the design of *PathwayOracle* is its role as an open platform for the development and deployment of new analytical capabilities by other groups. Currently *PathwayOracle* employs a modular architecture that facilitates easy integration of new functionality. However, in future releases we plan to expose a plugin interface which will make it easier to developers and researchers to develop and deploy tools within *PathwayOracle*.

## Conclusion

*PathwayOracle* is an integrated software environment in which biologists may conduct structural and dynamic analysis of signaling networks of interest. *PathwayOracle* is distinguished from other tools in the field of systems biology by its ability to predict the signal flow through a network using a simplified, connectivity-based model of the signaling network. Simulations are fast and, based on a published study, predictors of signal propagation. This novel simulation capability, combined with support for structural analysis of connectivity between pairs of proteins and for analysis of certain kinds of experimental data make *PathwayOracle* a powerful asset in the experimentalist's endeavor to gain a more complete understanding of the cellular signaling network.

## Availability and requirements

Project name: PathwayOracle

Project home page: <http://bioinfo.cs.rice.edu/pathwayoracle>

Operating system(s): Platform independent

Programming language: Python

Other requirements: Python 2.4 or higher

License: GNU GPL

Any restrictions to use by non-academics: None

## Authors' contributions

DR designed and developed the PathwayOracle application, participated in evaluating features for inclusion, and drafted the manuscript. LN participated in application design and feature selection. PTR contributed biological case studies and data for PathwayOracle feature design. All authors read and approved the final manuscript.

## Acknowledgements

DR and LN are supported in part by a Seed Grant awarded to LN from the Gulf Coast Center for Computational Cancer Research, funded by John and Ann Doerr Fund for Computational Biomedicine. PTR is supported in part by a Department of Defense grant BC044268.

## References

- Papin JA, Hunter T, Palsson BO, Subramaniam S: **Reconstruction of cellular signalling networks and analysis of their properties.** *Nature Reviews Molecular Cell Biology* 2005, **6**:99-111.
- Sackmann A, Heiner M, Koch I: **Application of Petri net based analysis techniques to signal transduction pathways.** *BMC Bioinformatics* 2006, **7**:482.
- Ruths D, Tseng JT, Nakhleh L, Ram PT: **De novo Signaling Pathway Predictions based on Protein-Protein Interaction, Targeted Therapy, and Protein Microarray Analysis.** *Proceedings of the RECOMB Satellite Workshop on Systems Biology and Proteomics, Lecture Notes in Bioinformatics (LNBI #4466)* 2007:62-72.
- Ruths D, Nakhleh L, Iyengar MS, Reddy SAG, Ram PT: **Graph-theoretic Hypothesis Generation in Biological Signaling Networks.** *Journal of Computational Biology* 2006, **13**(9):1546-1557.
- Klamt S, Saez-Rodriguez J, Lindquist JA, Simeoni L, Gilles ED: **A methodology for the structural and functional analysis of signaling and regulatory networks.** *BMC Bioinformatics* 2006, **6**:56.
- Papin JA, Palsson BO: **The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis.** *Biophysical Journal* 2004, **87**:37-46.
- Papin JA, Price ND, Wiback SJ, Fell DA, Palsson BO: **Metabolic pathways in the post-genomic era.** *TRENDS in Biochemical Sciences* 2003, **28**(5):250-258.
- Schilling CH, Letscher D, Palsson BO: **Theory for the Systemic Definition of Metabolic Pathways and their use in Interpreting Metabolic Function from a Pathway-Oriented Perspective.** *Journal of Theoretical Biology* 2000, **203**:229-248.
- Peleg M, Rubinfeld D, Altman RB: **Using Petri Net Tools to Study Properties and Dynamics of Biological Systems.** *Journal of the American Medical Informatics Association* 2005, **12**(2):181-199.
- Chaouiya C: **Petri net modelling of biological networks.** *Briefings in Bioinformatics* 2007, **8**(4):210-219.
- Eungdamrong NJ, Iyengar R: **Modeling cell signaling networks.** *Biology of the Cell* 2004, **96**(5):355-362.
- Eungdamrong NJ, Iyengar R: **Computational Approaches for modeling regulatory cellular networks.** *Trends Cell Biology* 2004, **14**(12):661-669.
- Ruths D, Muller M, Tseng JT, Nakhleh L, Ram PT: **The Signaling Petri Net-based Simulator: A Non-Parametric Strategy for Characterizing the Dynamics of Cell-Specific Signaling Networks.** *PLoS Computational Biology* 2008, **4**(2):e1000005.
- Kanehisa M, Goto S: **KEGG: Kyoto Encyclopedia of Genes and Genomes.** *Nucleic Acids Research* 2000, **28**(15):27-30.
- The Cancer Cell Map** [<http://cancer.cellmap.org>]
- Thomas PD, Campbell MJ, Kejariwal A, Mi J, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A: **PANTHER: a library of protein families and subfamilies indexed by function.** *Genome Research* 2003, **13**:2129-2141.
- David R, Alla H: *Discrete, Continuous, and Hybrid Petri Nets* Springer; 2005.
- Official Website for Python Programming Language** [<http://www.python.org>]
- Hucka M, Finney A, Sauro HM, H B, Doyle J, Kitano H, Arkin A, Bornstein B, Bray D, Cornish-Bowden A, Cuellar A, Dronov S, Gilles E, Ginkel M, Gor V, Goryanin I, Hedley W, Hodgman T, Hofmeyr J, Hunter P, Juty N, Kasberger J, Kremling A, Kummer U, Le Novere N, Loew L, Lucio D, Mendes P, Minch E, Mjolsness E, Nakayama Y, Nelson M, Nielsen P, Sakurada T, Schaff J, Shapiro B, Shimizu T, Spence H, Stelling J, Takahashi K, Tomita M, Wagner J, Wang J: **The Systems Biology Markup Language (SBML): A medium for representation and exchange of biochemical network models.** *Bioinformatics* 2003, **19**(4):524-531.
- Funahashi A, Tanimura N, Morohashi M, Kitano H: **CellDesigner: a process diagram editor for gene-regulatory and biochemical networks.** *BIOFILICO* 2003, **1**:159-162.
- Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Research* 2003, **13**(11):2498-504.
- Hoops S, Sahle S, Gauges R, Lee C, Pahle J, Simus N, Singhal M, Xu L, Mendes P, Kummer U: **COPASI – a Complex PATHway Simulator.** *Bioinformatics* 2006, **22**:3067-74.
- Alon U: *An Introduction to Systems Biology: Design Principles of Biological Circuits.* *Mathematical and Computational Biology Series* Chapman & Hall/CRC; 2007.
- Nagasaki M, Doi A, Matsuno H, Miyano S: **Genomic Object Net: I. A platform for modelling and simulating biopathways.** *Applied Bioinformatics* 2003, **2**(3):181-184.
- Klamt S, Saez-Rodriguez J, Gilles ED: **Structural and functional analysis of cellular networks with CellNetAnalyzer.** *BMC Systems Biology* 2007, **1**:2.
- Schmidt H, Jirstrand M: **Systems Biology Toolbox for MATLAB: A computational platform for research in Systems Biology.** *Bioinformatics* 2006, **22**(4):514-515.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)



# Use of Data-Biased Random Walks on Graphs for the Retrieval of Context-Specific Networks from Genomic Data

Kakajan Komurov<sup>1\*</sup>, Michael A. White<sup>2</sup>, Prahlad T. Ram<sup>1</sup>

<sup>1</sup> Department of Systems Biology, University of Texas M.D. Anderson Cancer Center, Houston, Texas, United States of America, <sup>2</sup> Department of Cell Biology, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America

## Abstract

Extracting network-based functional relationships within genomic datasets is an important challenge in the computational analysis of large-scale data. Although many methods, both public and commercial, have been developed, the problem of identifying networks of interactions that are most relevant to the given input data still remains an open issue. Here, we have leveraged the method of random walks on graphs as a powerful platform for scoring network components based on simultaneous assessment of the experimental data as well as local network connectivity. Using this method, NetWalk, we can calculate distribution of Edge Flux values associated with each interaction in the network, which reflects the relevance of interactions based on the experimental data. We show that network-based analyses of genomic data are simpler and more accurate using NetWalk than with some of the currently employed methods. We also present NetWalk analysis of microarray gene expression data from MCF7 cells exposed to different doses of doxorubicin, which reveals a switch-like pattern in the p53 regulated network in cell cycle arrest and apoptosis. Our analyses demonstrate the use of NetWalk as a valuable tool in generating high-confidence hypotheses from high-content genomic data.

**Citation:** Komurov K, White MA, Ram PT (2010) Use of Data-Biased Random Walks on Graphs for the Retrieval of Context-Specific Networks from Genomic Data. *PLoS Comput Biol* 6(8): e1000889. doi:10.1371/journal.pcbi.1000889

**Editor:** Nathan D. Price, University of Illinois at Urbana-Champaign, United States of America

**Received:** February 26, 2010; **Accepted:** July 15, 2010; **Published:** August 19, 2010

**Copyright:** © 2010 Komurov et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Funding:** This work was supported in part by NIH (CA71443) and the Robert Welch Foundation (I-1414) to MAW and DOD (BC044268) and NIH (R01CA125109) to PTR. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** The authors have declared that no competing interests exist.

\* E-mail: kkomurov@mdanderson.org

## Introduction

An important challenge in the analyses of high throughput datasets is integration of the data with prior knowledge interactions of the measured molecules for the retrieval of most relevant biomolecular networks [1–7]. This approach facilitates interpretation of the data within the context of known functional interactions between biological molecules and subsequently leads to high-confidence hypothesis generation. Typically, this procedure would entail identification of genes with highest or lowest data values, which is then followed by identification of associated networks. However, retrieval of most relevant biological networks/pathways associated with the upper or lower end of the data distribution is not a trivial task, mainly because members of a biological pathway do not usually have similar data values (e.g. gene expression change), which necessitates the use of various computational algorithms for finding such networks of genes [1,2,4,5,8–11]. One class of methods for finding relevant networks utilize optimization procedures for finding highest-scoring subnetworks/pathways of genes based on the data values of genes [2,8]. Although this approach is likely to result in highly relevant networks, it is computationally expensive and inefficient, and is therefore not suitable for routine analyses of functional genomics data in the lab. The most popular of the existing methods of extraction of relevant networks from genomic data, however, usually involve a network building strategy using a pre-defined focus gene set, which is typically a set of genes with most significant

data values (e.g. most over-expressed genes) [1,7]. The network is built by “filling in” other nodes from the network either based on the enrichment of interactions for the focus set (IPA -Ingenuity Pathway Analysis) [1], or based on the analysis of shortest paths between the focus genes (MetaCore) [7,12]. Both methods aim at identifying genes in the network that are most central to connecting the focus genes to each other. Problems associated with these methods have been outlined previously [7]. However perhaps most importantly, the central genes identified by these methods may have incoherent data values with the focus genes (e.g. the central genes may have reduced expression while the focus genes may have increased expression), as data values of nodes are not accounted for during the network construction process using the seed gene list. This may result in uninformative networks that are not representative of the networks most significantly represented in the genomic data (see Results). In addition, these methods do not account for genes with more subtle data values that collectively may be more important than those with more obvious data values [13]. Although powerful data analysis methods for finding sets of genes with significant, albeit subtle, expression changes have been developed (e.g. GSEA [13], Molecular concept maps [14], GenMAPP [15]), such an approach has not been incorporated into methods for extracting interaction networks that are most highlighted by the data.

In order to overcome these problems, we have employed the method of random walks in graphs for scoring the relevance of

## Author Summary

Analysis of high-content genomic data within the context of known networks of interactions of genes can lead to a better understanding of the underlying biological processes. However, finding the networks of interactions that are most relevant to the given data is a challenging task. We present a random walk-based algorithm, NetWalk, which integrates genomic data with networks of interactions between genes to score the relevance of each interaction based on both the data values of the genes as well as their local network connectivity. This results in a distribution of Edge Flux values, which can be used for dynamic reconstruction of user-defined networks. Edge Flux values can be further subjected to statistical analyses such as clustering, allowing for direct numerical comparisons of context-specific networks between different conditions. To test NetWalk performance, we carried out microarray gene expression analysis of MCF7 cells subjected to lethal and sublethal doses of a DNA damaging agent. We compared NetWalk to other network-based analysis methods and found that NetWalk was superior in identifying coherently altered sub-networks from the genomic data. Using NetWalk, we further identified p53-regulated networks that are differentially involved in cell cycle arrest and apoptosis, which we experimentally tested.

interactions in the network to the data. The method of random walks has been well-established for structural analyses of networks, as it can fully account for local as well as global topological structure within the network [16,17] and it is very useful for identifying most important/central nodes [16–18]. Here, instead of working with a pre-defined set of focus genes, we overlay the entire data distribution onto the network, and bias the random walk probabilities based on the data values associated with nodes. This method, NetWalk, generates a distribution of Edge Flux values for each interaction in the network, which then can be used for dynamical network building or further statistical analyses. Here, we describe the concept of NetWalk, demonstrate its usefulness in extracting relevant networks compared to Ingenuity Pathway Analysis, and show the use of NetWalk results in comparative analyses of highlighted networks between different conditions.

We tested NetWalk on experimentally derived genomic data from breast cancer cells treated with different concentrations of doxorubicin, a clinically used chemotherapeutic agent. Using NetWalk, we identify several previously unreported network processes involved in doxorubicin-induced cell death. From these studies we propose that NetWalk is a valuable network based analysis tool that integrates biological high throughput data with prior knowledge networks to define sub-networks of genes that are modulated in a biologically meaningful way. Use of NetWalk will greatly facilitate analysis of genomic data.

## Methods

### Calculating node probabilities using data

Integration of genomic data represented by a vector  $\mathbf{w}$  with the network data of interactions between genes (nodes) is performed by representing each interaction (edge) in the network in the form of a transition probability based on the data values (e.g. mRNA expression change, phenotype score from a genetic screen) of nodes within the immediate neighborhood:

$$p_{ij} = \frac{w_j}{\sum_{k \in \mathcal{N}_i} w_k} \quad (1)$$

where  $p_{ij}$  is the transition probability from node  $i$  to node  $j$ ,  $w_j$  is the experimental value for node  $j$ , and  $\mathcal{N}_i$  is the set of immediate downstream neighbors (undirected edges are considered bidirectional) of node  $i$ . If there are no downstream nodes of the node  $i$  ( $|\mathcal{N}_i| = 0$ ),  $p_{ij}$  is set to  $p_{ij} = 1/|n|$  for all  $j \in n \in n$ , where  $n$  is the set of all nodes in the network. The relevance score of each node in the network is defined by the probability of its visitation by the random walker, which is a function of both the local network connectivity as well the data values of nodes. So at any step  $k$  of this “random walk” process, the probability of a node being visited by the random walker is

$$g_i^k = \sum_{j \in n} g_j^{k-1} p_{ji} \quad (2)$$

where  $g_i^k$  is the probability of node  $i$  at step  $k$ ,  $p_{ji}$  is the transition probability from node  $j$  to node  $i$  and  $\mathcal{N}$  is the set of interacting neighbors of node  $i$ . This can be represented in a matrix form

$$\mathbf{g}^k = \mathbf{g}^{k-1} \cdot \mathbf{P} \quad (3)$$

where  $\mathbf{g}^k$  is the vector of probability values for all nodes in the network at step  $k$ , and  $\mathbf{P}$  is the transition probability matrix of the network. Obviously, since a “walk” can only be performed over adjacent nodes,  $p_{ij} > 0$  only if nodes  $i$  and  $j$  directly interact. The expression above can also be written as

$$\mathbf{g}^k = \mathbf{g}^0 \cdot \mathbf{P}^k \quad (4)$$

where  $\mathbf{P}^k$  is the transition probability matrix raised to the power  $k$ , and  $\mathbf{g}^0$  is the initial probability distribution over nodes (all  $1/|n|$ ). By the Perron-Frobenius theorem for stochastic matrices, as  $k \rightarrow \infty$   $k \rightarrow \infty$  (infinite random walk), the expression above converges to

$$\mathbf{g} = \mathbf{g} \cdot \mathbf{P} \quad (5)$$

where  $\mathbf{g}$  is the left eigenvector of  $\mathbf{P}$  associated with eigenvalue 1 and contains the final visitation probability values of nodes.

The final visitation probabilities of nodes depend on their data values, data values of their neighbors, as well as the local network connectivity. In order to further bias the random walk towards the input data values, we assigned a small probability  $q$  that the random walker will return to its starting node. Therefore, the expression for random walk with restart is given by

$$\mathbf{g} = \mathbf{g} \left( (1-q)\mathbf{P} + \frac{1}{|n|} \mathbf{q} \times \mathbf{1}^T \right) \quad (6)$$

where  $\mathbf{q}$  is a vector of all  $q$  of length  $|n|$  and  $\mathbf{1}$  is a vector of all 1: so that the restart probability is uniform among all nodes. However, we bias the restart probabilities to the data values of nodes, so that the random walker is more likely to return to its initial node if the data value of that node is high.

$$\mathbf{g} = \mathbf{g} \left( (1-q)\mathbf{P} + \frac{1}{\sum_{k \in n} w_k} \mathbf{q} \times \mathbf{w}^T \right) \quad (7)$$

In this way, the probability that the random walker will restart at another node  $i$  is directly proportional on the data value of node  $i$ , thereby even more biasing the process of random walk to the biological data.

### Calculating node probabilities for transcription factors

In the case of transcription factor - target gene interactions, these were reversed in the network so that the node values of target genes would contribute to the probabilities of the transcription factors, rather than the other way around. This is because the data values of target genes (i.e. mRNA expression change) are more informative of identifying regulation by transcription factors.

### Calculating edge flux values

To find networks of interactions between genes represented in the data, we scored each interaction in the network by

$$e_{ij} = \mathbf{g}_i \mathbf{P}_{ij} \quad (8)$$

where  $e_{ij}$  is the flux through edge  $ij$  and represents the score of *importance* of the interaction based on the data.

### Controlling for topological bias in the network

The node visitation frequencies in a random walk directly reflect the relative centralities of nodes in the network, and therefore are highly biased towards the local network topology. Although biasing the random walk to data values skews the visitation frequencies towards the supplied data values, there is still a significantly high correlation with node connectivity values (Figure S1), which suggests that the random walk process is highly biased to the highly connected hubs in the network. Therefore, it is important to control for topological bias in the network that stems either from its scale-free nature or the historical bias of highly studied genes. In order to control for topological biases in the network, we also calculated background visitation frequencies

$$\mathbf{g}_r = \mathbf{g}_r \left( (1-q) \mathbf{P}_r + \frac{1}{n} \mathbf{q} \times \mathbf{1}^T \right) \quad (9)$$

which is the same expression as in equation (7), with the exception that  $e_{r_{ij}} = \mathbf{g}_{r_i} \mathbf{P}_{r_{ij}}$ .  $\mathbf{P}_r$  is a transition probability matrix formed by letting  $w_i = 1$  for all  $i$ . Since  $\mathbf{g}_r$  is calculated without considering the data values of genes, it contains all the topological bias in the network. Therefore, to obtain relative visitation frequencies of genes ( $\mathbf{g}'$ ), we normalize values in  $\mathbf{g}$  by those in  $\mathbf{g}_r$ .

$$g'_i = \frac{g_i}{g_{r_i}} \quad (10)$$

Relative visitation frequency values in  $\mathbf{g}'$  have minimal correlation with node centralities, and have a high correlation with the supplied gene expression measurements (Figure S2), which indicates that relative visitation frequencies of nodes are highly biased towards the data.

Normalization of edge flux values is done by first calculating

$$e_{r_{ij}} = \mathbf{g}_{r_i} \mathbf{P}_{r_{ij}} \quad (11)$$

where  $\mathbf{e}_r$  is the edge score distribution vector calculated by letting  $w_i = 1$  for all  $i$ . Then, we normalize the data-biased edge flux values to  $\mathbf{e}_r$  to obtain normalized Edge Flux of interaction  $ij$  ( $EF_{ij}$ )

$$EF_{ij} = \log \left( \frac{e_{ij}}{e_{r_{ij}}} \right) \quad (12)$$

which gives the final normalized score distribution of edges, which reflects edge fluxes of nodes *relative* to what would be expected by topology alone in the given network.

### Data format and missing values

Because of the nature of random walks described above, the input values must be positive, possibly representing ratio of a test versus control sample (e.g. ratio of mRNA expression levels of treated to untreated samples). Missing values in the network are then assigned a value of 1, which represents a *no change* case in ratio values. Accordingly, the values of  $s$  are centered around 0, with higher values meaning higher probability relative to what would be expected by chance in the given network (i.e. networks of high data value nodes, e.g. increased gene expression), and lower values meaning lower visitation probability (i.e. networks with low data values, e.g. reduced gene expression) (see below).

### Effect of data distribution on Edge Flux values

In order to prevent disproportionate skewing of the node probabilities with extreme outliers in the data, the input data is normalized so that all  $w > k_{0.999}$  are assigned  $k_{0.999}$ , where  $k_{0.999}$  is the 99.9th percentile value of  $w$ . Similarly, all  $w < k_{0.001}$  are assigned  $k_{0.001}$ . With this procedure, the final normalized visitation frequencies of nodes are highly robust to differences in data distributions and ranges (see Figure S3).

### Network construction

We compiled protein-protein interactions from online databases HPRD [19] BIND [20], HomoMINT [21], Gene [22] and IntAct [23]. For directed interactions, we compiled signaling interactions from KEGG [24], BioCarta (<http://pid.nci.nih.gov/>) and TRANSPATH [25], as well as through manual curation of the undirected interactions based on published literature. Transcription factor-target interactions were obtained from ORegAnno [26] and TRANSFAC [27] databases. This resulted in a network of 10,473 genes connected by ~65,000 interactions.

In network-based analyses of genomic data, the analyses and therefore resultant hypotheses are limited by the gene coverage of the network. Therefore, it is crucial that the interaction network has as much gene coverage as possible. Since our main goal of network-based analyses is identification of relevant biological processes, the interactions represented in the network need not be direct physical interactions. For example, a concordant increase in the expression of genes involved in glucose metabolism will not be captured in network-based analyses of direct physical interactions, as metabolic enzymes within the same pathway rarely engage in direct physical interactions (with the exception of multifunctional complexes). Therefore, inclusion of *indirect* functional interactions in the network may help identify relevant biological processes that are not captured by direct interactions (see network plots below). In order to increase the coverage of our network, we added functional similarity interactions between genes, where an interaction means that the genes are involved in similar functional processes, such as a metabolic pathway (e.g. glycolysis) or a specific enzymatic reaction (e.g. oxidation/reduction). Functional similarity interactions were constructed using Gene Ontology (GO) annotations [28] as defined in the Entrez Gene database, and also metabolic pathway annotations in the KEGG database. Any two genes sharing a metabolic pathway annotation (but not signaling

pathways as they are already represented in protein-protein interactions) from KEGG were assigned an interaction. In the case of GO annotations, two genes were assigned an interaction if the overlap of their GO annotations was significant compared to the rest of the genes:

$$s_{ij} = \frac{|\bigcap_{k \in N} G_k|}{n}$$

where  $s_{ij}$  is the significance of overlap between genes  $i$  and  $j$ ;  $G_k$  is the set of genes that have the GO term  $k$ ;  $N$  is the set of GO terms common to genes  $i$  and  $j$ , and  $n$  is the total number of genes. If  $s_{ij} < 0.001$ , genes  $i$  and  $j$  were assigned an interaction.

Our final network contains 14,506 genes connected by 189,901 interactions. Gene coverage of our network of genes in our doxorubicin dataset is comparable to that in the Ingenuity Pathway Analysis (13,329 in our network versus 13,880 in IPA).

### Microarray analyses

MCF7 cells were grown in DMEM (Invitrogen) supplemented with 10% FBS (Gemini) to near confluency and treated with 1 or 10  $\mu\text{M}$  Doxorubicin (Sigma). Cells were collected at 0, 6, 12 and 24 hours post-treatment. Cell lysis and RNA extraction was done using Mirvana miRNA isolation kit (Ambion) and amplification using Illumina TotalPrep RNA amplification kit (Ambion). Equal amount of RNA from each sample was hybridized to Illumina HT12 BeadChip (Illumina). All procedures were performed exactly as described in the respective manuals. The experiments were repeated in triplicate.

### Analyses with IPA

Networks in IPA were generated using Core analysis with indicated data cutoffs for upregulated genes and using direct interactions with the cutoff for network size to be 70. Highest scoring 5 networks were merged and exported as text files.

### Network plotting

All network plottings were done using the *gplot* function in the *sna* package for R (<http://erzuli.ss.uci.edu/R.stuff/>).

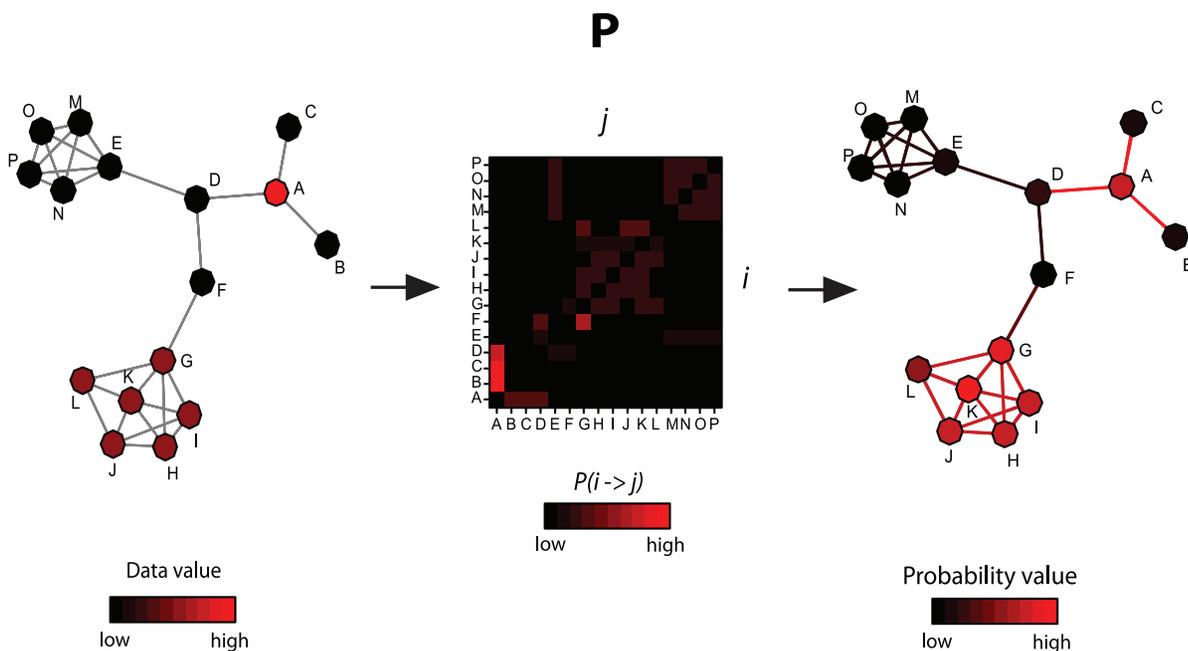
### Western blotting

Cells were treated as indicated and lysed in a sample lysis buffer (50 mM Hepes, 150 mM NaCl, 1mM EGTA, 10 mM Sodium Pyrophosphate, pH 7.4, 100 nM NaF, 1.5 mM  $\text{MgCl}_2$ , 10% glycerol, 1% Triton X-100 plus protease inhibitors; aprotinin, bestatin, leupeptin, E-64, and pepstatin A). Blotting was done using antibodies against p53 (Cell Signaling), p21 (Cell Signaling) and Actin (Sigma). The experiment was done in triplicate.

### Apoptosis assays

FACS: Cells were treated as indicated and after 24 hours trypsinized, fixed with 70% ethanol at  $-20^\circ\text{C}$  for 10 minutes and resuspended in Propidium Iodide solution. FACS analysis was performed in the Flow Cytometry core facility of M.D. Anderson Cancer Center.

Rhodamine 123 assay: Rhodamine 123 staining was performed as described [29]. Briefly, cells were treated as indicated and after 24 hours, trypsinized, spun down and resuspended in 10  $\mu\text{M}$  Rhodamine 123 (Invitrogen) in PBS for 30 minutes. Cells were washed in PBS and analyzed by FACS for Rhodamine 123 intensity (green).



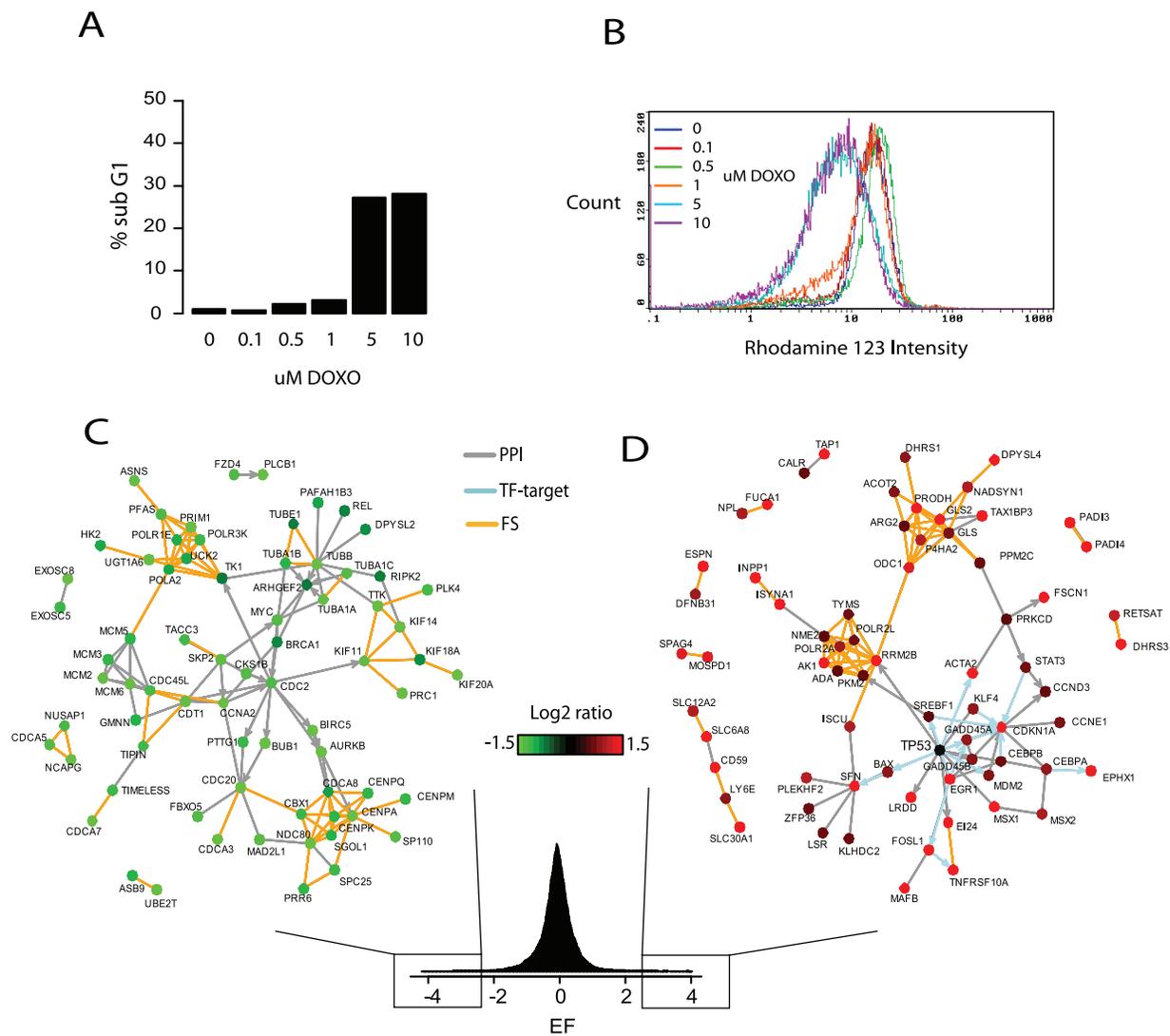
**Figure 1. General concept of NetWalk.** An imaginary network with artificial experimental data values is shown (e.g. relative gene expression values) on the left. Node A was assigned a value of 5, nodes G, H, I, J, K and L were assigned 2, and all the other nodes were assigned 1. A transition probability matrix **P** was constructed using the input data values and the network, with transition probabilities between adjacent nodes reflecting their data values (colors in the matrix reflect transition probabilities  $P(i \rightarrow j)$  according to the color key). Final visitation and flux values reflect the level of coherence between the experimental data of genes and their relative positioning within the network. Note that node colorings in the network on the right reflect relative visitation probabilities of nodes, and line colors of edges reflect the flux values according to the same color scale. doi:10.1371/journal.pcbi.1000889.g001

**Results**

**Generating networks with NetWalk**

Identifying common biological roles of genes whose expression are altered in a microarray experiment is one of the most frequently used strategies to understand the underlying biological processes and derive hypotheses [6,13–15,30]. This strategy is also implicit in NetWalk (Figure 1), as node visitation frequency values (hence *EF* values) calculated by NetWalk are based on 1) data values of nodes, 2) data values of their network neighbors and 3) the network connectivity among neighbors. Therefore, a node with a high data value that interacts with other nodes with high data values in the network will receive the highest node visitation and *EF* scores. Similarly, a node with a low data value that interacts with other nodes with low data values in the network will receive the lowest node visitation and *EF* scores.

In order to test the dependency of NetWalk output on the provided data, we performed deletions of portions of data and compared the resultant visitation frequencies to those of the original dataset. Correlation of node visitation frequencies to those of the full dataset closely followed the input data, suggesting that NetWalk output is highly dependent on the supplied data (Figure S4). However, this may also suggest that NetWalk output is mostly independent of the network connectivity. In order to test the dependence of NetWalk output on the network connectivity, we removed parts of the network and performed NetWalk analysis on the perturbed networks. The resultant node visitation frequencies correlate relatively poorly with those of the original network (Figure S5), indicating that the network connectivity substantially contributes to node visitation frequency values. We also performed a similar analysis with random deletions and additions of edges, rather than nodes, in the network, and found a similar dependence



**Figure 2. NetWalk analysis of low and high-dose doxorubicin response in MCF7 cells.** A) Apoptosis levels in MCF7 cells after 24 hours of stimulation with indicated doses of doxorubicin as measured by FACS analysis of DNA content (see Methods). B) FACS analysis of viable cells as indicated by loss of Rhodamine 123 staining(see Methods). C–D) Plots of interactions with lowest(B) and highest (C) EF values in samples treated with 1  $\mu$ M doxorubicin for 24 hours relative to control. Nodes are colored according to their gene expression change relative to control according to the color key. Edge coloring reflects type of interaction, PPI: protein-protein interaction, TF-target: gene regulation, FS: functional similarity. The distribution plot of all EF values is shown at the bottom. doi:10.1371/journal.pcbi.1000889.g002

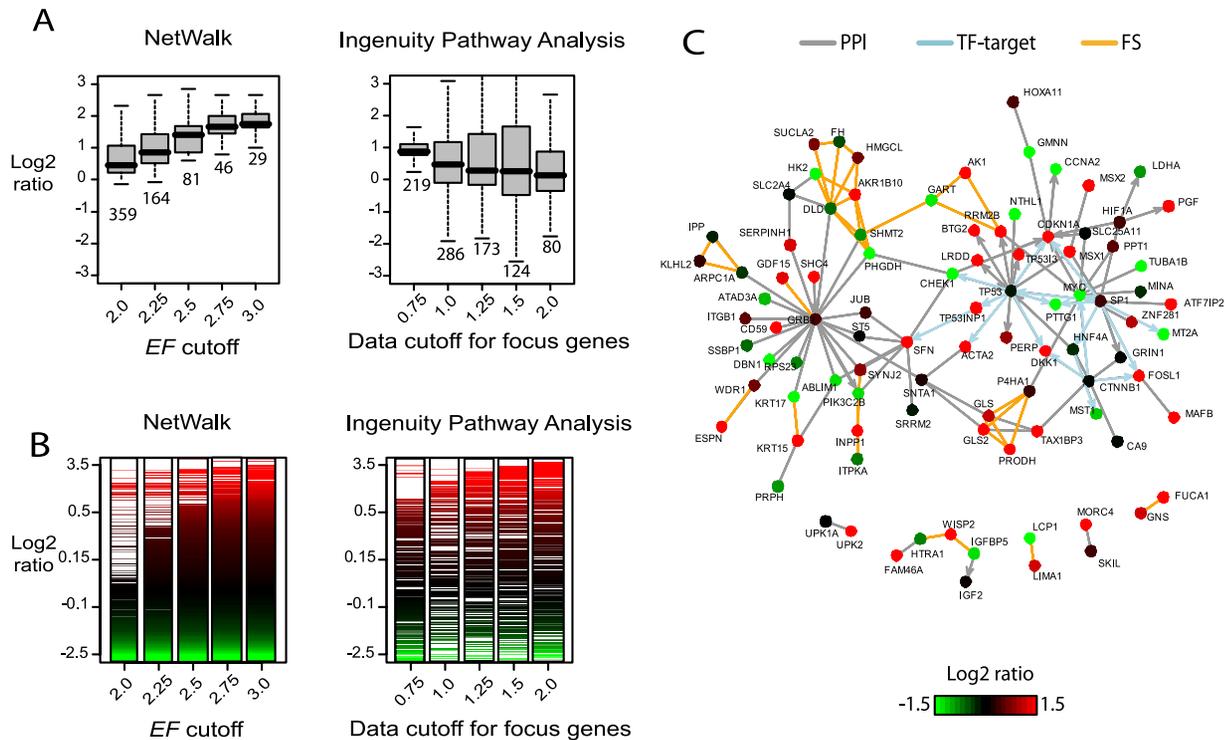
of the NetWalk output on the network connectivity (Figure S7). These analyses demonstrate that NetWalk output is highly dependent on both the supplied data as well as the network information.

To demonstrate the use of NetWalk in the extraction of relevant networks out of microarray gene expression data, we studied gene expression profiles of MCF7 cells subjected to sub-lethal and lethal doses of doxorubicin. We performed microarray gene expression analysis of MCF7 cells before and after treatment with 1 or 10  $\mu\text{M}$  doxorubicin for 6, 12 and 24 hours. In these cells, 1  $\mu\text{M}$  doxorubicin causes a cell cycle arrest in S-phase, while a 10  $\mu\text{M}$  dose induces cell death (Figure 2A–B). A NetWalk analysis of the ratio values (treated/untreated) for 1  $\mu\text{M}$  treatment was performed using  $q = 0.01$  (see Methods). The resulting distribution of edge flux values, and plots of edges with 100 highest and lowest *EF* values can be seen in Figure 2C–D. *EF* values are strictly biased towards the data, as the high and low-end networks are entirely composed of genes with, respectively, increased and reduced expression levels. In the Figure 2D, interactions in the cluster made of GLS, GLS2, P4HA2, ODC1 and PRODH genes (arginine and proline metabolism) have the highest *EF* scores due to both their high data values and tight interconnections with each other. Similarly, in the low-score network in Figure 2C, interactions in the cluster containing NDC80, CENPK, CBX1, CENPA and SGOL1 (centriole components) have the lowest *EF* scores. Nodes with moderate values that are in close proximity to other high value nodes within a tightly connected neighborhood will also get high scores, as is seen with TP53 in Figure 2B.

In order to demonstrate that the p53 network extracted by NetWalk is not an artifact of highly connected subnetworks, we performed a NetWalk analysis of baseline expression profile of MCF7 cells relative to other breast cancer cells as reported by Neve *et al* [31]. The most significantly upregulated networks in MCF7 cells relative to the rest of 53 breast cancer cells are those involved in the Estrogen Receptor signaling (Figure S6), a well-characterized dominant pathway in the estrogen receptor positive MCF7 cells. This analysis shows that NetWalk output does indeed reflect accurate quantification of highly biologically relevant networks based on the supplied data.

**EF scores are highly coherent with data values**

Contrary to the seed-based network building methods, NetWalk works with the whole data distribution and so does not require assignment of pre-defined cutoffs or focus gene sets. NetWalk procedure simply translates the gene-centric data values to corresponding interaction scores based on the coherence of the gene values with those in the local network neighborhood as well as the local interaction pattern in the network. Therefore, the results can be viewed at any user defined cutoff value for flexible generation of networks with highly coherent node values. The distribution of input node values and sample networks with different *EF* cutoffs shows that the node values within networks are highly coherent across a wide range of *EF* score cutoffs, which allows for high-confidence hypothesis generation about activated and inactivated network processes in response to DNA damage (Figure 3A–B). In comparison, the distribution of data values of nodes in the networks



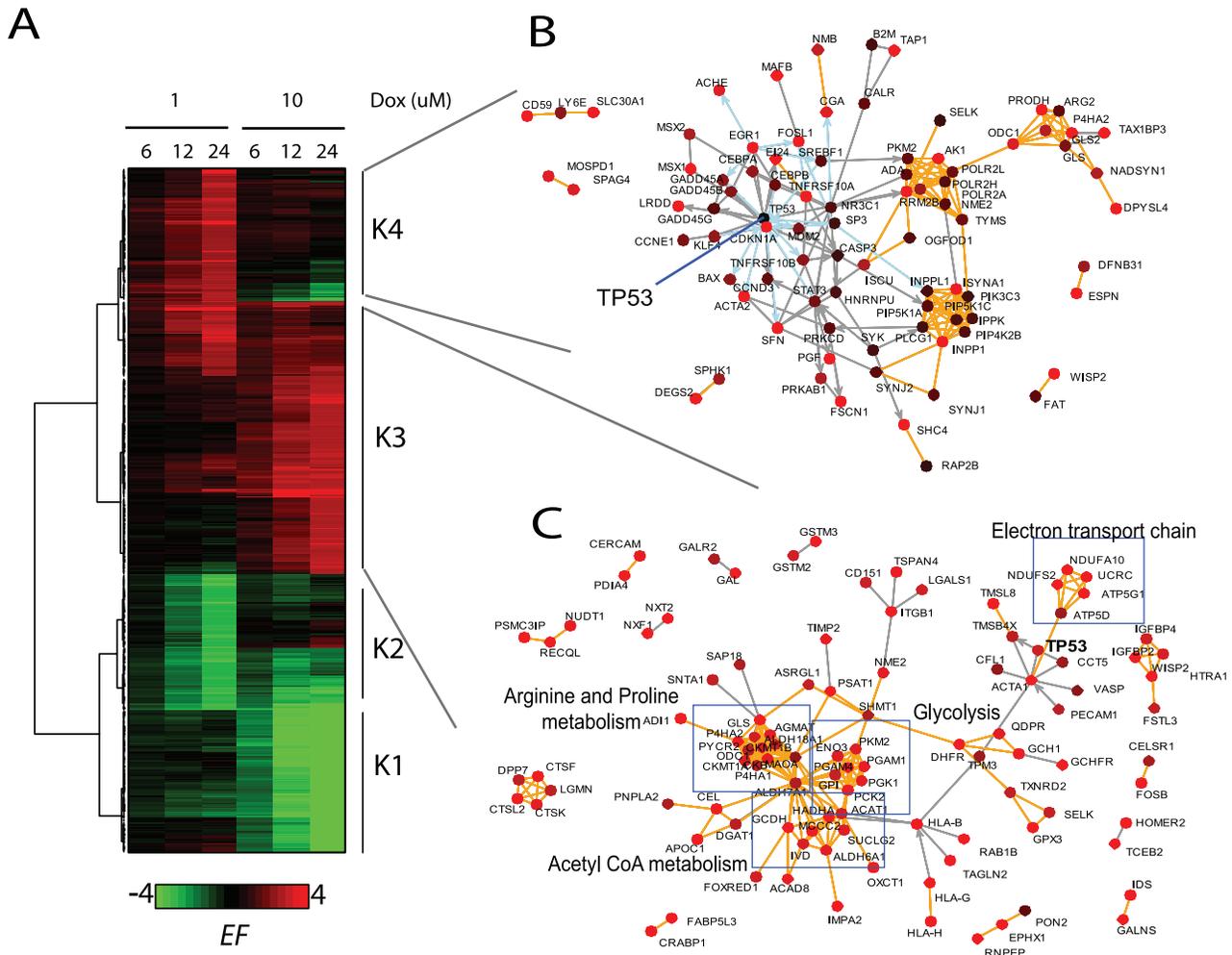
**Figure 3. Comparison of coherence of node values in highest scoring networks.** A) Boxplots of gene expression change values (1  $\mu\text{M}$  DOX, 24 hours relative to control) of nodes in networks generated by different cutoffs of *EF* values, or in networks generated by Ingenuity Pathway Analysis software using different data gene expression value cutoffs for the focus gene set (see Methods). B) Heatmaps showing position of genes in the networks in A in the whole data distribution. Positions of genes in the respective networks are indicated by a white line. C) A network of nodes generated by Ingenuity Pathway Analysis software with focus gene set using 1.5 as cutoff. Since original network plots in IPA lack node colorings for intermediate genes (non-focus genes), we extracted all nodes in the IPA-generated network and re-plotted them using our network, where we colored all nodes by their gene expression change. doi:10.1371/journal.pcbi.1000889.g003

generated by Ingenuity Pathway Analysis, which takes a focus gene list as input to build relevant networks, includes nodes with incoherent data values (see Figure 3A–C), which reduces confidence in the relevance of the generated networks to the data. The network of 124 genes retrieved by IPA using a cutoff of  $>1.5$  (60 focus genes) contains many genes with reduced expression values (Figure 3C), which were included in the network by the virtue of their connectivities but not data values. Consequently, the resulting network is not entirely representative of upregulated network processes in response to doxorubicin. Moreover, none of the networks identified by IPA contain all the genes involved in arginine-proline metabolism (compare Figures 2D and 3C) or any genes involved in the nucleotide metabolism that were retrieved by NetWalk (see cluster in Figure 2D containing RRM2B, AK1, POLR2A and NME2; compare with Figure 3C), demonstrating inability of seed-based methods to identify subnetworks with more subtle yet coherent gene expression values.

### Statistical analyses using NetWalk output to elucidate p53-mediated response to DNA damage

As stated earlier, an important feature of NetWalk is that the result is not a single or a collection of static networks, but a whole distribution

of numerical edge scores. In addition to their use for dynamical network construction of different sizes based on the user preference, these can be further subjected to standard statistical tests for a more detailed analysis. The heatmap of interactions with highest and lowest *EF* scores in each condition in our microarray dataset is shown in Figure 4A. As opposed to clustering with traditional heatmaps of gene expression values where cluster membership of genes is exclusive, here, a gene can appear in several different clusters but all with different interactions. So, analysis of expression with *EF* scores enables studying specific functions (i.e. interactions) of genes rather than their individual expression values. The heatmap shows that the activation and/or inactivation of several networks is specific to low- or high-dose doxorubicin treatment. The cluster K3, for example, is activated in response to high-dose doxorubicin, while K4 is more specifically activated in response low-dose doxorubicin. A plot of interactions in K3 reveals several metabolic pathways specifically activated in the high-dose treatment, including glycolysis, acetyl coenzyme A synthesis, arginine/proline metabolism and the mitochondrial electron transport chain (Figure 4C). There is also a p53-centered subnetwork containing several previously identified p53 target genes. The plot of interactions in K4 shows an extensive p53-centered network composed mostly of cell cycle regulatory proteins (e.g. CDKN1A (p21CIP) and several



**Figure 4. Clustering analysis of EF values in each condition.** A) Heatmap of highest and lowest EF values in each condition. Clustering was done using Ward’s method in R. B–C) Networks corresponding to K3 (B) and K4 (C). Node colorings are according to 24h of 1 and 10  $\mu$ M DOX treatments, respectively. Edge colorings are as in Figure 2C. doi:10.1371/journal.pcbi.1000889.g004

GADD45 genes) (Figure 4B). Interestingly, although p53 appears in both K3 and K4, its functions seem to be completely different in the low and high dose treatments. In response to low-dose doxorubicin, p53 is involved in the activation of cell cycle regulatory proteins, while under high-dose, it activates other targets, such as TMSB4X. Moreover, p53-target genes in cell cycle regulation in K3 are inactivated in high-dose doxorubicin (Figure 5A–B), which we confirmed by western blotting (Figure 5C), suggesting that p53 may act as a transcriptional activator of these genes during cell cycle arrest but as a repressor during apoptosis. This trend suggests not only that p53 may engage different targets during cell cycle arrest and apoptosis, but also shows dual behavior of p53 under these conditions. In addition, this analysis shows that energy and amino acid metabolisms may play an important role in doxorubicin-induced cell death. Here, clustering analysis using NetWalk results facilitated comparison of networks, rather than genes, between different conditions, leading to the identification of differential activities of p53 under low and high-dose doxorubicin treatment.

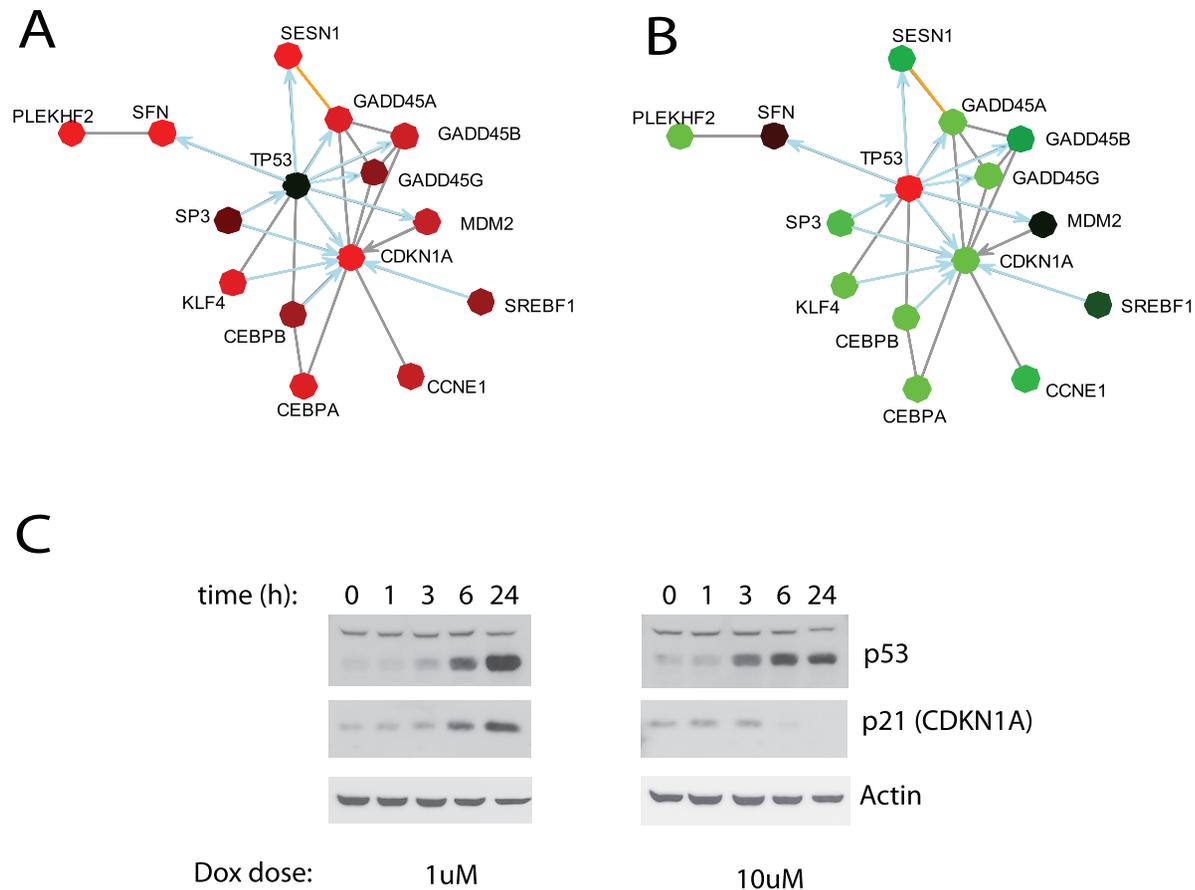
## Discussion

### NetWalk algorithm

Analyses of high content data within the context of biological interactions allow for high confidence hypothesis generation about mechanisms involved in the studied process. While some work has

been done on inferring novel causal interactions out of data [32–34], the most popular method is integration of data with prior knowledge on interactions to extract most relevant networks highlighted by the data. Most of the methods for extracting relevant networks rely on finding genes in the network that are most central to connecting the genes of interest identified from the data. The random walk process in NetWalk also scores most central genes in the network. However, rather than working on a small set of focus genes, NetWalk scores centralities of all genes in the network based on the whole data distribution. This is achieved by biasing the random walk transition probabilities between genes to their corresponding data values, which allows for higher visitation probabilities of nodes with high data values and lower probabilities of nodes with low data values. Since visitation probabilities of nodes in a random walk are also dependent on the visitation probabilities of their network neighbors, nodes with relatively moderate data values associated with those with higher values have the potential of high visitation by the random walk. Therefore, NetWalk scores nodes based on their data values, data values of their neighbors and local network connectivity.

Unlike most of the existing methods for network extraction, which typically give a set of networks as outputs [1,9], NetWalk gives a distribution of *EF* values that allows for flexibility in network construction using different *EF* cutoffs. In addition, *EF* scores can be subjected to further statistical tests for comparative



**Figure 5. p53-target cell cycle regulatory genes are specifically repressed during apoptosis.** A–B) Network plot of interactions in K3 (see Figure 4) related to cell cycle regulation. Nodes colored according to gene expression changes at 10 (A) or 1  $\mu$ M (B) doxorubicin treatment. C) Western blots of p53, p21 (CDKN1A gene product) protein levels over a time course after 1 and 10  $\mu$ M doxorubicin treatment. Actin levels shown as control.

doi:10.1371/journal.pcbi.1000889.g005

studies, allowing for network-based comparisons of multiple conditions.

Another important feature of NetWalk is its computational efficiency. We implemented a sparse matrix representation and multiplication, which allows for NetWalk to be run on a standard PC equipped with 1 gigabytes of memory. In our case (PC with Intel Xeon Quad processor), NetWalk run of a single dataset in our network (14,506 nodes and ~190,000 interactions) took about 2–3 seconds.

NetWalk analysis of the experimental data revealed a significant activation of networks involved in energy metabolism, including the glycolytic and mitochondrial electron transport chain components. At least one member of the electron transport chain, SCO2A, has been previously shown to be a p53 target [35], suggesting that some, if not most, of the metabolic genes activated in response to 10  $\mu$ M doxorubicin may be p53 target genes. A specific and extensive activation of the energy metabolism during p53-mediated apoptosis has not been previously reported, and therefore it is a novel finding facilitated by NetWalk analysis. Network analysis of experimental data using NetWalk revealed dual behavior of p53 under sublethal and lethal doses of DNA damage. In response to sublethal doses of DNA damaging agents, p53 activates a cell cycle arrest program centered around CDK inhibitors p21 (CDKN1A) and GADD45, as well as several pro-apoptotic genes, such as BAX and APAF1. However under lethal doses, p53 represses the cell cycle arrest machinery and activates an entirely different program. Use of NetWalk analysis allows network based analysis of genomic data as well as high confidence hypothesis generation and is a valuable tool in post-genomic analysis.

## Supporting Information

**Figure S1** Correlation of node visitation frequencies with node connectivities (left) and original data values (right) before normalization for network topology (see Text). R2 values show squared Spearman's rank correlation coefficients.  
Found at: doi:10.1371/journal.pcbi.1000889.s001 (0.08 MB PDF)

**Figure S2** Same as in Figure S1, but after normalization for network topological bias (see Text).  
Found at: doi:10.1371/journal.pcbi.1000889.s002 (0.09 MB PDF)

**Figure S3** Effect of data range on NetWalk output. Original mRNA expression changes in response to 1  $\mu$ M doxorubicin (ratio) were log<sub>2</sub>-transformed ( $d_i$ ), and then transformed back by taking exponential with different expansion factors  $f$ ,  $\sigma_i = f^{d_i}$  where  $\sigma_i$  is the transformed value of gene  $i$ ,  $d_i$  is the log<sub>2</sub>-transformed original ratio value of gene  $i$  and  $f$  is the expansion factor. Distributions of the transformed data with different expansion factors are shown in A. Numbers above each distribution chart

shows the expansion factor. Expansion factor of 2 corresponds to the original distribution. B) Correlation of visitation frequencies corresponding to each transformed dataset with the original visitation frequency values (i.e.  $f=2$ ). C) Correlation of visitation frequency values for each expansion factor with the supplied transformed data values. D–E) Highest scoring interactions calculated using transformed datasets with expansion factor D) 1.25 and E) 5. Note that the two networks are highly similar (~95% same node composition).

Found at: doi:10.1371/journal.pcbi.1000889.s003 (0.23 MB PDF)

**Figure S4** Effect of data deletions on NetWalk output. Portions of data were deleted and node visitation frequencies were calculated by NetWalk. Shown are the correlations of each deletion with the original node visitation frequency values (i.e. 0% deletion).  
Found at: doi:10.1371/journal.pcbi.1000889.s004 (0.03 MB PDF)

**Figure S5** Effect of network deletions on NetWalk output. A network corresponding to 690 nodes (highest scoring interactions in 1  $\mu$ M doxorubicin dataset) was selected and nodes were deleted at random. Correlation of resulting node visitation frequency values with the original unperturbed network of 690 nodes is shown (black). In addition, corresponding correlations with the node degrees in each network are also shown. Note that although total number of interactions are relatively similar in each deletion, the NetWalk output changes substantially due to changes in the local network connectivities.  
Found at: doi:10.1371/journal.pcbi.1000889.s005 (0.07 MB PDF)

**Figure S6** Highest scoring networks corresponding to estrogen receptor positive MCF7 cells relative to 58 other breast cancer cell lines. ESR1 (estrogen receptor gene) is highlighted.  
Found at: doi:10.1371/journal.pcbi.1000889.s006 (0.26 MB PDF)

**Figure S7** Effect of edge perturbations on NetWalk output. A random network corresponding to 755 nodes was selected out of the whole network (3721 interactions). A) Edges were deleted at random and correlation of the resultant node visitation frequencies were compared to that of unperturbed network. B) To the network in A where 50% of all edges were removed, we added random interactions between random pairs of nodes and compared the resultant NetWalk output with the initial NetWalk output at 50% deleted network.  
Found at: doi:10.1371/journal.pcbi.1000889.s007 (0.09 MB JPG)

## Author Contributions

Conceived and designed the experiments: KK PTR. Performed the experiments: KK. Analyzed the data: KK. Contributed reagents/materials/analysis tools: MAW. Wrote the paper: KK PTR.

## References

- Calvano SE, Xiao W, Richards DR, Felciano RM, Baker HV, et al. (2005) A network-based analysis of systemic inflammation in humans. *Nature* 437: 1032–1037.
- Chuang HY, Lee E, Liu YT, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. *Mol Syst Biol* 3: 140.
- Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, et al. (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* 2: 2366–2382.
- Ekins S, Nikolsky Y, Bugrim A, Kirillov E, Nikolskaya T (2007) Pathway mapping tools for analysis of high content data. *Methods Mol Biol* 356: 319–350.
- Ganter B, Zidek N, Hewitt PR, Muller D, Vladimirova A (2008) Pathway analysis tools and toxicogenomics reference databases for risk assessment. *Pharmacogenomics* 9: 35–54.
- Kaplow IM, Singh R, Friedman A, Bakal C, Perrimon N, et al. (2009) RNAiCut: automated detection of significant genes from functional genomic screens. *Nat Methods* 6: 476–477.
- Dezso Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, et al. (2009) Identifying disease-specific genes based on their topological significance in protein networks. *BMC Syst Biol* 3: 36.
- Lee E, Chuang HY, Kim JW, Ideker T, Lee D (2008) Inferring pathway activity toward precise disease classification. *PLoS Comput Biol* 4: e1000217.
- Nikolsky Y, Nikolskaya T, Bugrim A (2005) Biological networks and analysis of experimental data in drug discovery. *Drug Discov Today* 10: 653–662.
- Muller FJ, Laurent LC, Kostka D, Ulitsky I, Williams R, et al. (2008) Regulatory networks define phenotypic classes of human stem cell lines. *Nature* 455: 401–405.
- Ulitsky I, Shamir R (2009) Identifying functional modules using expression profiles and confidence-scored protein interactions. *Bioinformatics* 25: 1158–1164.
- Bugrim A, Nikolskaya T, Nikolsky Y (2004) Early prediction of drug metabolism and toxicity: systems biology approach and modeling. *Drug Discov Today* 9: 127–135.

13. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 102: 15545–15550.
14. Tomlins SA, Mehra R, Rhodes DR, Cao X, Wang L, et al. (2007) Integrative molecular concept modeling of prostate cancer progression. *Nat Genet* 39: 41–51.
15. Dahlquist KD (2004) Using GenMAPP and MAPPFinder to view microarray data on biological pathways and identify global trends in the data. *Curr Protoc Bioinformatics Chapter 7: Unit 7 5*.
16. Aldous D, Fill J Reversible Markov Chains and Random Walks on Graphs.
17. Lovasz L (1993) Random Walks on Graphs: A Survey. *Bolyai Society Mathematical Studies* 2: 1–46.
18. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A* 105: 1118–1123.
19. Mishra GR, Suresh M, Kumaran K, Kannabiran N, Suresh S, et al. (2006) Human protein reference database–2006 update. *Nucleic Acids Res* 34: D411–414.
20. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, et al. (2001) BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res* 29: 242–245.
21. Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res* 35: D572–574.
22. Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res* 35: D26–31.
23. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, et al. (2007) IntAct—open source resource for molecular interaction data. *Nucleic Acids Res* 35: D561–565.
24. Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28: 27–30.
25. Choi C, Krull M, Kel A, Kel-Margoulis O, Pistor S, et al. (2004) TRANSPATH-A High Quality Database Focused on Signal Transduction. *Comp Funct Genomics* 5: 163–168.
26. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, et al. (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. *Nucleic Acids Res* 36: D107–113.
27. Wingender E, Chen X, Hehl R, Karas H, Liebich I, et al. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28: 316–319.
28. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29.
29. Ferlini C, Scambia G (2007) Assay for apoptosis using the mitochondrial probes, Rhodamine123 and 10-N-nonyl acridine orange. *Nat Protoc* 2: 3111–3114.
30. Barrett T, Edgar R (2006) Mining microarray data at NCBI's Gene Expression Omnibus (GEO)\*. *Methods Mol Biol* 338: 175–190.
31. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* 10: 515–527.
32. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, et al. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34: 166–176.
33. Sachs K, Perez O, Pe'er D, Lauffenburger DA, Nolan GP (2005) Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308: 523–529.
34. Nelander S, Wang W, Nilsson B, She QB, Pratilas C, et al. (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol Syst Biol* 4: 216.
35. Matoba S, Kang JG, Patino WD, Wragg A, Boehm M, et al. (2006) p53 regulates mitochondrial respiration. *Science* 312: 1650–1653.

# Identification of Optimal Drug Combinations Targeting Cellular Networks: Integrating Phospho-Proteomics and Computational Network Analysis

Sergio Iadevaia, Yiling Lu, Fabiana C. Morales, Gordon B. Mills, and Prahlad T. Ram

## Abstract

Targeted therapeutics hold tremendous promise in inhibiting cancer cell proliferation. However, targeting proteins individually can be compensated for by bypass mechanisms and activation of regulatory loops. Designing optimal therapeutic combinations must therefore take into consideration the complex dynamic networks in the cell. In this study, we analyzed the insulin-like growth factor (IGF-1) signaling network in the MDA-MB231 breast cancer cell line. We used reverse-phase protein array to measure the transient changes in the phosphorylation of proteins after IGF-1 stimulation. We developed a computational procedure that integrated mass action modeling with particle swarm optimization to train the model against the experimental data and infer the unknown model parameters. The trained model was used to predict how targeting individual signaling proteins altered the rest of the network and identify drug combinations that minimally increased phosphorylation of other proteins elsewhere in the network. Experimental testing of the modeling predictions showed that optimal drug combinations inhibited cell signaling and proliferation, whereas nonoptimal combination of inhibitors increased phosphorylation of nontargeted proteins and rescued cells from cell death. The integrative approach described here is useful for generating experimental intervention strategies that could optimize drug combinations and discover novel pharmacologic targets for cancer therapy. *Cancer Res*; 70(17); OF1-11. ©2010 AACR.

## Major Findings

Simple and reliable strategies are needed to identify optimal combinations of molecular targeted drugs to treat individual cancer patients, to realize the fullest potential of a targeted therapeutic approach.

## Introduction

Cell signaling networks are complex systems that integrate information from the cellular environment (1–5). Maps of complex networks were derived by interconnecting the individual pathways obtained from experimental data (6, 7). These studies revealed that signaling networks contain numerous features, such as feedback and feedforward loops (8, 9), which render it virtually impossible for the human mind to decipher how signals are integrated within the path-

ways. Thus, computational approaches are needed to elucidate the regulatory properties of signaling networks (10–12).

Several groups have used ordinary differential equations (ODE) to analyze the dynamics of signaling networks and generate experimentally testable predictions (6, 13–17). The use of mass action ODE modeling, however, is impaired because of incomplete knowledge about the concentrations and kinetics of signaling intermediates.

Inferring the parameters for mass action modeling in signaling networks is challenging. The most common approach is to obtain parameters from the literature and fit the models to the experimental data to infer those that remain unknown (6, 13, 18–24). Unfortunately, the kinetic parameters reported in the literature may differ by orders of magnitude, depending on experimental conditions. Thus, it is difficult to determine whether discrepancies between computational and experimental data are due to inaccurate measures or incomplete modeling. Parameter estimation can be effectively accomplished using optimization methods, which enable quantitative model fitting to experimental data (25–31). However, the experimental techniques used to measure the activity of signaling proteins mainly provide qualitative or semiquantitative data. Optimization strategies are thus needed to identify sets of model parameters that equally fit the qualitative experimental data.

Another challenge in the analysis of signaling networks is the identification of optimal target combinations. The most common methods of computational target identification are based on formulating mathematical models and designing

**Authors' Affiliation:** Department of Systems Biology, The University of Texas M.D. Anderson Cancer Center, Houston, Texas

**Note:** Supplementary data for this article are available at Cancer Research Online (<http://cancerres.aacrjournals.org/>).

**Corresponding Authors:** Sergio Iadevaia or Prahlad T. Ram, M.D. Anderson Cancer Center, 7435 Fannin Street, Unit 950, P.O. Box 301429, Houston, TX. Phone: 713-563-2848; Fax: 713-563-4235; E-mail: siadevai@mdanderson.org or pram@mdanderson.org.

doi: 10.1158/0008-5472.CAN-10-0460

©2010 American Association for Cancer Research.

## Quick Guide To Equations And Assumptions

### Mass action modeling

The dynamics of the IGFR network in MDA-MB231 cells were described using a mass action model of ODEs formulated as follows:

Step 1: The pathways comprising the IGFR network were reconstructed into a set of chemical reactions that described the simplified mechanisms of activation and inhibition of relevant proteins. For example, mitogen-activated protein kinase (MAPK) phosphorylation was assumed to be catalyzed by MAPK kinase [MAP/ extracellular signal-regulated kinase kinase (MEK)/MAPKK] and occurred through an enzymatic reaction:



In equation 1,  $k_1^f$ ,  $k_1^r$ , and  $K_2$  are the forward, reverse, and dissociation kinetic rate constants, respectively.

Step 2: The set of chemical reactions was transformed into a system of coupled ODEs by assuming that the dynamics of the IGFR network obeyed the law of mass action. Specifically, the accumulation rate of the concentration of the  $i$ th signaling intermediate was expressed as the difference between its net rates of production ( $r_{p,i}$ ) and consumption ( $r_{c,i}$ ). Thus, the accumulation rate of the concentration of  $MEK^*$  was expressed as follows:

$$\frac{dMEK^*}{dt} = \sum r_{p,MEK^*} - \sum r_{c,MEK^*} = -k_1^f [MAPK][MEK^*] + (k_1^r + K_2)[C] \quad (2)$$

In equation 2,  $[MAPK]$ ,  $[MEK^*]$ , and  $[C]$  denote the concentration of MAPK,  $MEK^*$  and C, respectively.

The list of chemical reactions that described the consensus activation and inhibition mechanisms of proteins involved in the IGFR network and the corresponding system of ODEs are listed in Supplementary Material S1. To implement mass action modeling, it was necessary to infer the unknown model parameters, which are the kinetic rate constants and the initial concentrations of the proteins. In this regard, we trained the mass action model against transient data measured by RPPA using PSO. We selected PSO because of its superior ability to converge to more optimal solutions compared with other optimization algorithms (see Discussion).

### Particle swarm optimization

PSO is a stochastic algorithm that mimics the behavior of swarms of animals that search for food (36). Particles in the swarm have a position  $x_{ij}$ , a velocity  $v_{ij}$ , and a fitness  $f_i$  in which  $i$  and  $j$  represent the number of particles and the dimension of the space solution, respectively. Each particle remembers its best position  $x_{ij}^L$  locally and the best position  $x_j^G$  globally reached by the entire swarm. During the iterative search for food, particles update their position and velocities to improve their fitness according to the following rules:

$$\begin{aligned} v_{ij}(t+1) &= \omega v_{ij}(t) + c_1 r_1 [x_{ij}^L(t) - x_{ij}(t)] + c_2 r_2 [x_j^G(t) - x_{ij}(t)] \\ x_{ij}(t+1) &= x_{ij}(t) + v_{ij}(t) \end{aligned} \quad (3)$$

In equations 3,  $\omega$  is the inertia factor;  $r_1$  and  $r_2$  are two random numbers uniformly distributed in the interval  $[0,1]$ ; and  $c_1$  and  $c_2$  are the coefficients of self-recognition and social component (see ref. 37 and Supplementary Material S2 for details on parameters in equations 3).

In our settings, the particle positions represented the unknown parameter values used in the mass action model to generate computationally the time courses of proteins that are measured by RPPA; the particle velocities denoted the extent to which the parameter values were iteratively changed; and the particle fitness was defined as the distance between the time courses of proteins experimentally and computationally measured. Model parameters were randomly initialized and iteratively changed according to equation 3 until the distance between the time courses of the measured and predicted proteins was minimal (i.e., optimal fitness). The distance between computed and measured time courses was evaluated using the SD-weighted square error:

$$SqE = \sum_{j=1}^r \sum_{i=1}^s \frac{[\tilde{y}_{ij}^m - y_{ij}^c]^2}{\sigma(y_{ij}^m)} \quad (4)$$

In equation 4,  $\tilde{\gamma}_{ij}^m$  and  $\sigma(\gamma_{ij}^m)$  represent the mean and SD, respectively, of the proteins measured by RPPA, whereas  $y_{ij}^c$  denotes the protein levels computed using the mass action model. Moreover,  $s$  represents the total number of data points comprising a single time course, and  $r$  is the total number of time courses. PSO was implemented to minimize the SD-weighted square error and train the mass action model against RPPA data to estimate the unknown model parameters.

intervention strategies through environmental, genetic, and signaling perturbations (32–34). This approach can predict the effect of available drugs on signaling network dynamics, but it does not facilitate the search for drug combinations that would optimally inhibit aberrant signaling. Another strategy is to integrate mass action modeling with simulated annealing into a multiple-target optimal intervention (35). Because this approach is computationally expensive, alternative procedures are needed to enable the rapid search for targets in disease-related networks.

In this study, we used reverse-phase protein array (RPPA) to measure the transient response of the MDA-MB231 breast cancer cell line after stimulation by insulin-like growth factor (IGF-1). The reason for choosing the IGF receptor (IGFR) network is 2-fold: There is a large amount of experimental data and biological resources allowing us to build a consensus network and experimentally test it; components of this network are being targeted in several clinical trials for cancer therapy, thus having clinical applicability. We developed a computational procedure that integrated mass action modeling with particle swarm optimization (PSO) to train the model against normalized time courses of phosphorylated proteins in MDA-MB231 cells and infer sets of unknown model parameters that equally fit the measured data. The trained mass action model was used to predict the effect of a targeted perturbation and tested using experimental data. The trained and tested mass action model was then used to identify the most influential molecules responsible for aberrant cell signaling and determine the optimal combinations of inhibitors and small interfering RNAs for inhibiting abnormal signaling in MDA-MB231 cells. Immunoblotting and cell viability assay were then used to test and validate the effect of drug combinations predicted by the mass action model. Our integrative approach is useful for generating experimental intervention strategies that could optimize drug combinations and discovering novel pharmacologic targets for cancer therapy.

## Materials and Methods

### Cell culture and stimulation

The human MDA-MB231 breast cancer cell line (K-Ras and B-Raf mutants) was purchased from the America Type Culture Collection (ATCC). The cell line was validated by STR DNA fingerprinting using the AmpFISTR Identifier kit (Applied Biosystems). The STR profiles were compared with known ATCC fingerprints (<http://www.atcc.org/>) and to the Cell Line Integrated Molecular Authentication database version 0.1.200808 (38). Cells were cultured in RPMI supplemented with 5% fetal bovine serum. Cells were serum starved overnight and then subjected to treatment with 75 ng/mL IGF-1 (Cell Signaling Technology). For RPPA, cells were pretreated with 10  $\mu\text{mol/L}$  U0126 (Promega) for 4 hours, followed by IGF-1 stimulation for 5, 15, 30, 60, 90, or 120 minutes. For immunoblotting, cells were pretreated with 10  $\mu\text{mol/L}$  U0126, 50  $\mu\text{mol/L}$  LY294002 (Calbiochem-Nova-Biochem Corp.) and 50 nmol/L rapamycin (Calbiochem-

Nova-Biochem Corp.), individually or combined, for 1 hour, followed by IGF-1 stimulation for 5 or 60 minutes. For RPPA and immunoblotting, controls were incubated for the corresponding times with DMSO.

### Antibodies

The following antibodies were used for RPPA and immunoblotting: anti-phospho-MAPK (T202/Y204), anti-phospho-GSK3 (S21/S9), anti-phospho-AKT (ser473), anti-phospho-TSC2 (T1462), anti-phospho-mammalian target of rapamycin (mTOR; S2448), anti-phospho-P70S6K (T389), anti-MAPK (p44/42), anti-AKT, anti-TSC2 (28A7), anti-mTOR, anti-P70S6K, and anti-actin were from Cell Signaling Technology; and anti-GSK3 was from Santa Cruz Biotechnology, Inc.

### Immunoblotting

Immunoblotting was performed using standard procedures.

### Reverse-phase protein array

Serial diluted lysates were arrayed on nitrocellulose-coated FAST slides (Whatman) using the Aushon 2470 Arrayer (Aushon Biosystems). Each slide was probed with a primary antibody plus a biotin-conjugated secondary antibody. The signal was amplified using the DakoCytomation-catalyzed system (DAKO) and visualized using a 3,3'-diaminobenzidine colorimetric reaction. The slides were scanned, analyzed, and quantified using the customized Microvigene software (VigeneTech, Inc.) to measure spot intensity. Each dilution curve was fitted with the logistic model "Supercurve Fitting" (39). The mean values of the protein levels in the nonstimulated cells were used to normalize the time courses of the phosphorylated proteins measured in IGF-1-stimulated cells.

### Crystal violet cell viability assay

Viability assay was performed using standard procedures. Cells were treated for 3 days with the following: U0126 (concentration of 0.1–100  $\mu\text{mol/L}$ ), LY294002 (concentration of 0.1–100  $\mu\text{mol/L}$ ), or rapamycin (concentration of 0.1–100 nmol/L); combination of U0126 (concentration of 0.5–50  $\mu\text{mol/L}$ ) and LY294002 [fixed at its quarter maximal effective concentration ( $\text{EC}_{25}$ ) value of 3.8  $\mu\text{mol/L}$ ] or rapamycin (fixed at its  $\text{EC}_{25}$  value of 0.1 nmol/L); and combination of U0126 (fixed at its  $\text{EC}_{25}$  value of 3.5  $\mu\text{mol/L}$ ) and rapamycin (concentration of 0.5 to 50 nmol/L). Corresponding controls were incubated with DMSO. The  $\text{EC}_{25}$  of each inhibitor was estimated (Supplementary Material S3) using Microsoft GraphPad Prism.

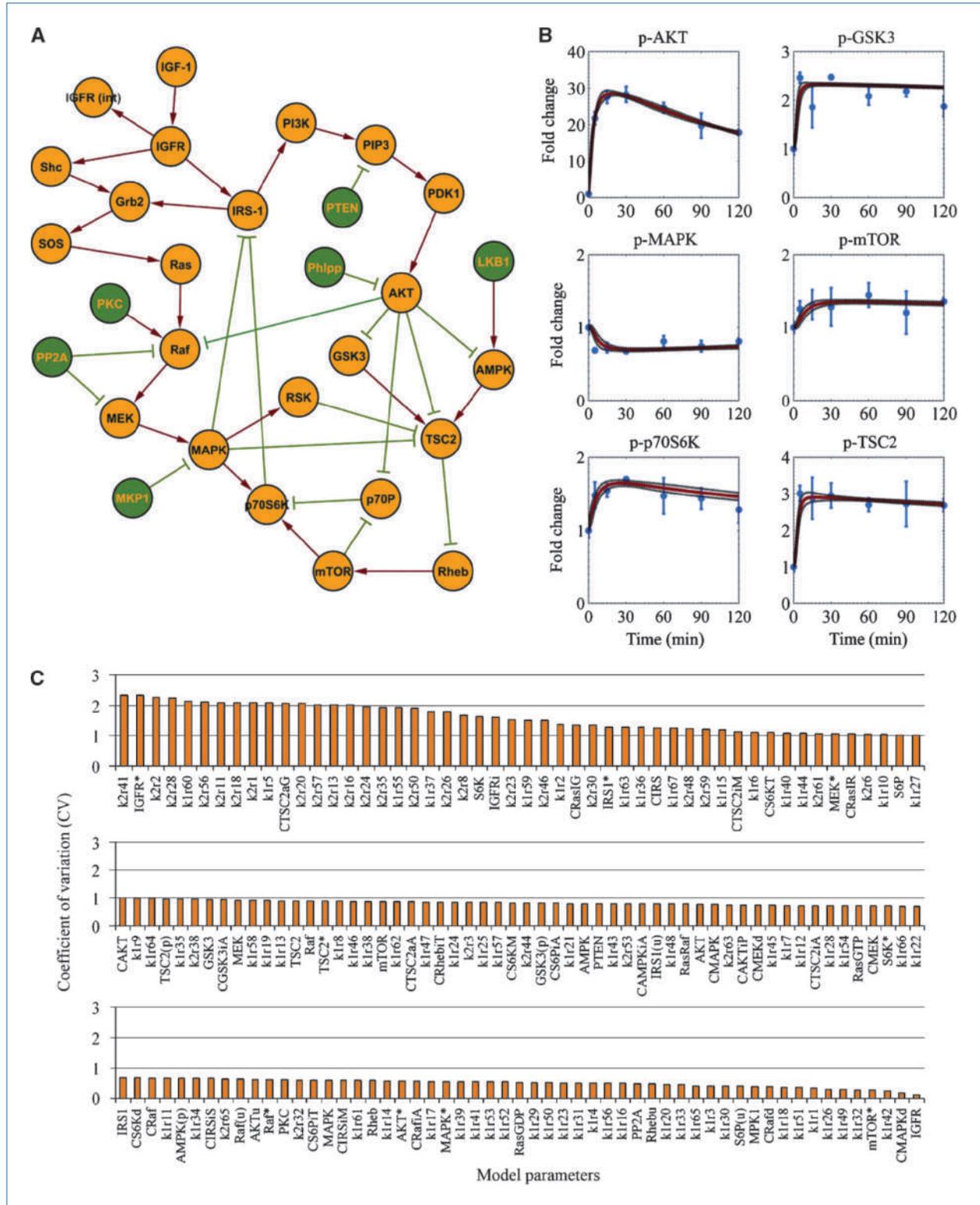
### Computational procedures

Computational procedures are described in Supplementary Material S2.

## Results

### IGFR-1 signaling detection by RPPA

Figure 1A shows the IGFR signaling network in the MDA-MB231 cell line. Signal transduction is originated when IGF-1

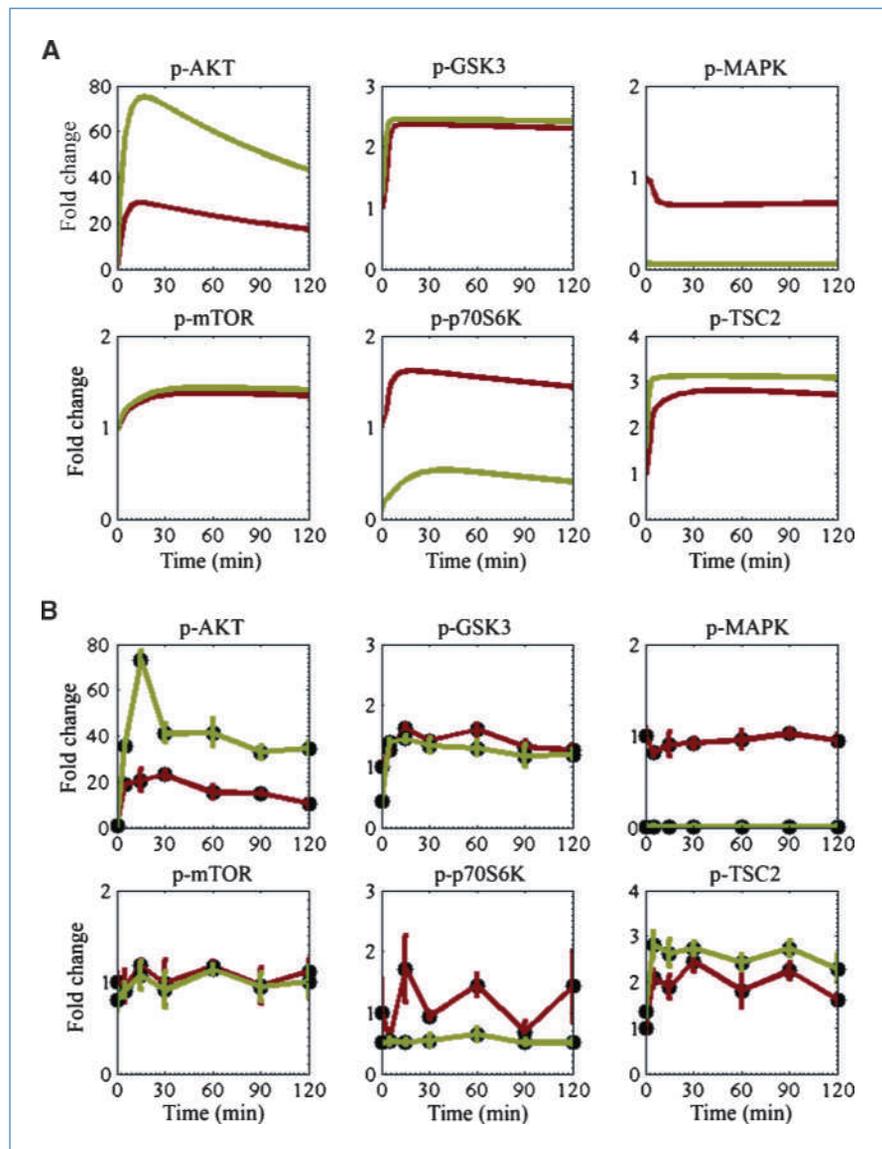


**Figure 1.** A, IGF1R signaling network topology in the MDA-MB231 cell line. Nodes, proteins; edges, protein interactions; red arrows, protein activation; and green plungers, protein inactivation. B, protein profiles were measured on RPPA in triplicate. The mean protein profiles of non-IGF-1-stimulated cells were used as controls for normalization. Circles, mean of the normalized protein profiles; bars, SD. Normalized time courses were computationally evaluated using the trained mass action model. Solid red lines, the mean time courses of the trajectories that equally fit the experimental data; dashed black lines, the fitting variability. C, histogram of model parameter regimens clustered according to the coefficient of variation ( $CV = SD/mean$ ).

complexes with IGFR and triggers IGFR autophosphorylation (40). Phosphorylated IGFR propagates the signal downstream through the MAPK and phosphoinositide-3-kinase (PI3K) pathways, and leads to MAPK and protein kinase B (PKB/AKT) phosphorylation (4, 5). The signals from the MAPK and PI3K cascades are routed to the mTOR pathway through tuberous sclerosis (TSC2) inactivation (1). Phosphorylated mTOR activates protein S6 kinase of 70 kDa (p70S6K), which inactivates the insulin receptor substrate (IRS-1) through a negative feedback loop (41). A detailed description of the network topology is provided in Supplementary Material S1. We used RPPA (42–46) to determine the changes in the phosphorylation of proteins in the IGFR network after IGF-1 stimulation. To account for the intrinsic variability of these assays, all experiments were performed in three independent

repeats. Figure 1B shows the time courses of the measured phosphorylated proteins; the curves show the protein fold change over the corresponding controls (Materials and Methods). After IGF-1 stimulation, the level of phospho-AKT peaked at 30 minutes (28-fold increase) and then settled toward a lower level at 120 minutes (18-fold increase). In contrast, signal transduction across the MAPK cascade remained essentially unchanged likely as a result of MAPK constitutive activation driven by *K-Ras* and *B-Raf* mutations in MDA-MB231 cells. AKT activation triggered glycogen synthase kinase (GSK3) and TSC2 downregulation through phosphorylation, and TSC2 inactivation facilitated phospho-mTOR and phospho-p70S6K upregulation. Thus, the levels of p-GSK3, p-TSC2, p-mTOR, and p-p70S6K initially increased and then adjusted to stationary levels.

**Figure 2.** The effect of MEK inhibition on IGFR network dynamics. A, protein profiles of IGF-1-stimulated MDA-MB231 cells predicted by the trained mass action model. Solid red lines, the protein time courses of the noninhibited cells; solid green lines, the protein profiles of MEK-inhibited cells. B, protein phosphorylation in MDA-MB231 cell lysates after stimulation with 75 ng/mL IGF-1 detected in triplicate by RPPA. Solid circles and red lines, the protein time courses of noninhibited cells; solid circles and green lines, the protein profiles of cells inhibited with MEK inhibitor for 4 h.



### Complementing mass action modeling with PSO

**Mass action modeling and model reduction.** To predict the dynamics of the IGFR network after IGF-1 stimulation in MDA-MB231 cells, we developed a mass action ODE model. Our formulation was based on a set of 77 chemical reactions that described the consensus activation and inhibition mechanisms of proteins involved in the IGFR network. The resulting mass action model was structured into 127 ODEs and 313 unknown parameters. To decrease the complexity of the model, we developed a reduced version of the original model. The 77 chemical reactions were reduced to a subset of 41 reactions to describe the simplified interaction mechanisms of the most relevant species in the IGFR network, and the original model was reduced to 65 ODEs and 161 model unknowns (Supplementary Material S4). We tested and validated the ability of the reduced model to adequately describe IGFR dynamics by showing that the protein profiles predicted by the reduced model matched those generated by the original model for randomly selected sets of parameters (Supplementary Material S5). Therefore, throughout the article, we exclusively used the reduced model to predict the dynamics of IGFR signaling network.

**Model training.** The measured time course data of proteins in MDA-MB231 cells contain relevant information about the regulatory loops comprising the IGFR network. To exploit this information to optimally inhibit aberrant pathways, we used PSO to fit the model to the time courses of p-AKT, p-MAPK, p-GSK3, p-mTOR, p-p70S6K, and p-TSC2 proteins and infer the 161 unknown parameters. Studies published in the literature typically use only two or three “readout” molecules to fit ODE models to experimental data and infer unknown parameters (14, 15, 26). In our study, we trained our model using six readout proteins and 126 experimental data points combined into a scalar fitness.

Because of the substantial degree of uncertainty in parameter estimation, fitting mass action models to the qualitative data measured on RPPA required the identification of multiple trajectories that equally resembled the measured protein profiles. Using the integrative mass action modeling PSO procedure, we identified 10 sets of model parameters that equally fit the measured data (Supplementary File S1). We characterized the parameter regimens by ranking the parameters according to their coefficient of variation (CV) and found that 69% of them had a CV smaller than 1 (Fig. 1C). We calculated the means and SD of the identified trajectories to represent the entire set and the fitting variability. Figure 1B shows the mean trajectories and the fitting variability identified by the mass action model, which had been trained using PSO against normalized protein profiles measured on RPPA after IGF-1 stimulation of MDA-MB231 cells. The simulation results indicated that the integrative procedure adequately fit the time courses of all measured proteins.

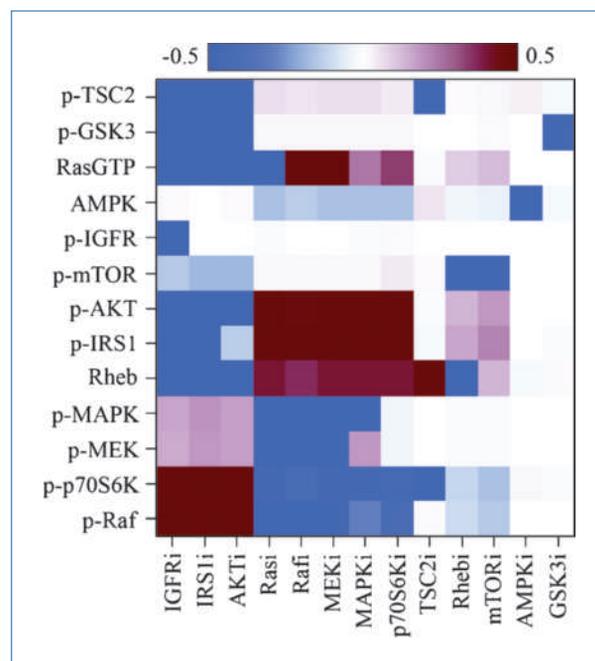
**Model testing.** To determine the ability of the trained model to correctly generate responses to perturbations that have not been explicitly included in the training data set, we used it to predict the dynamics of the IGFR network after inhibition of MEK. Figure 2A shows the transient IGFR signaling response to targeted MEK inhibition, as predicted by

the trained mass action model. MEK inhibition led to significant downregulation of its immediate downstream effector, p-MAPK. Inhibition of p-MAPK attenuated the inhibition of IRS-1 through direct interaction and through the p70S6K feedback loop. Consequently, p-AKT was upregulated. Activation of p-AKT increased the level of p-TSC2 but did not affect the level of p-mTOR or GSK3. Signals from the MAPK and mTOR cascades were integrated into the p70S6K pathway and led to p-p70S6K downregulation.

The computational results were experimentally tested using an independent set of 252 data points measured by RPPA. Figure 2B shows the levels of p-AKT, p-MAPK, p-GSK3, p-mTOR, p-p70S6K, and p-TSC2 detected in triplicate in IGF-1-stimulated MDA-MB231 cells in the absence of the MEK inhibitor and after 4 hours of incubation with the MEK inhibitor. The experimental data indicated that MEK inhibition increased p-AKT and p-TSC2 levels, decreased p-MAPK and p-p70S6K levels, and slightly decreased p-GSK3 levels but had no significant effect on p-mTOR levels. Despite the limited discrepancy between the computed and measured profiles of p-GSK3, the experimental results adequately matched those predicted by the model. Therefore, the trained mass action model correctly predicted the effect of MEK inhibition on IGFR dynamics.

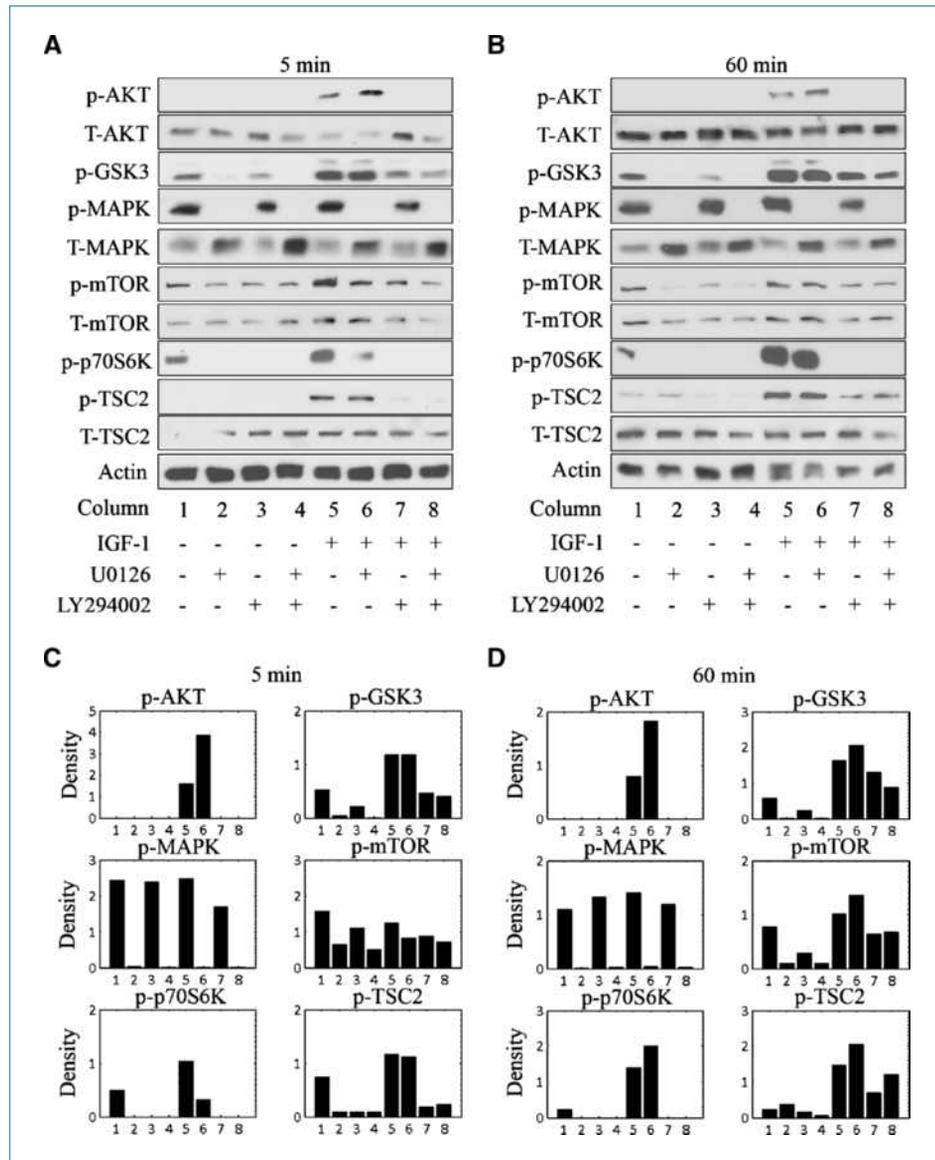
### Predicting inhibition of targeted molecules

**Individual inhibition of targeted molecules.** To determine how to select drugs with the ability to inhibit the



**Figure 3.** The effect of single-molecule inhibition on IGFR signaling in MDA-MB231 cells. Differential levels of proteins in single molecule-inhibited versus noninhibited MDA-MB231 cells after 2 h of stimulation with 75 ng/mL IGF-1 were predicted using the trained mass action model. Numerical values were converted to  $\log_{10}$ . Blue, inhibition; white, no variation; red, activation.

**Figure 4.** Combined inhibition of MEK and PI3K in MDA-MB231 cells. A and B, protein levels were detected in unstimulated cells in the absence of inhibition (columns 1) and after 1 h of incubation with the MEK and/or PI3K inhibitors (columns 2–4), and in cells stimulated with IGF-1 in the absence of inhibition (column 5) and after 1 h of incubation with the MEK and/or PI3K inhibitors (columns 6–8). Cells were stimulated with 75 ng/mL IGF-1 for 5 min (A) or 60 min (B). C and D, density of the bands after normalization with respect to actin.



pathways measured in MDA-MB231 cells, we used the trained and tested mass action model to predict the response of the IGF1R network after molecules in the network had been individually inhibited (Supplementary Material S6). Figure 3 and Supplementary Material S6 show the differential levels of proteins in inhibited versus noninhibited MDA-MB231 cells after IGF-1 stimulation, as predicted by the mass action model for 3 of the 10 sets of parameters inferred using PSO (Supplementary File S2). These three sets were randomly selected to repeat the computational analysis in triplicate. The modeling results suggested that targeting one molecule at a time may activate nontargeted molecules, likely as a result of feedback loop compensation. Thus, targeting a single molecule may not be sufficient to adequately inhibit aberrant signaling. Although inhibition of molecules in the MAPK pathway

was predicted to activate the PI3K/AKT pathway, inhibition of intermediates comprising the PI3K/AKT pathway was predicted to activate the MAPK pathway. Because these pathways are often upregulated in many tumors (47), the combined inhibition of the MAPK and PI3K/AKT cascades emerged as a candidate strategy to inhibit aberrant signaling in MDA-MB231 cells.

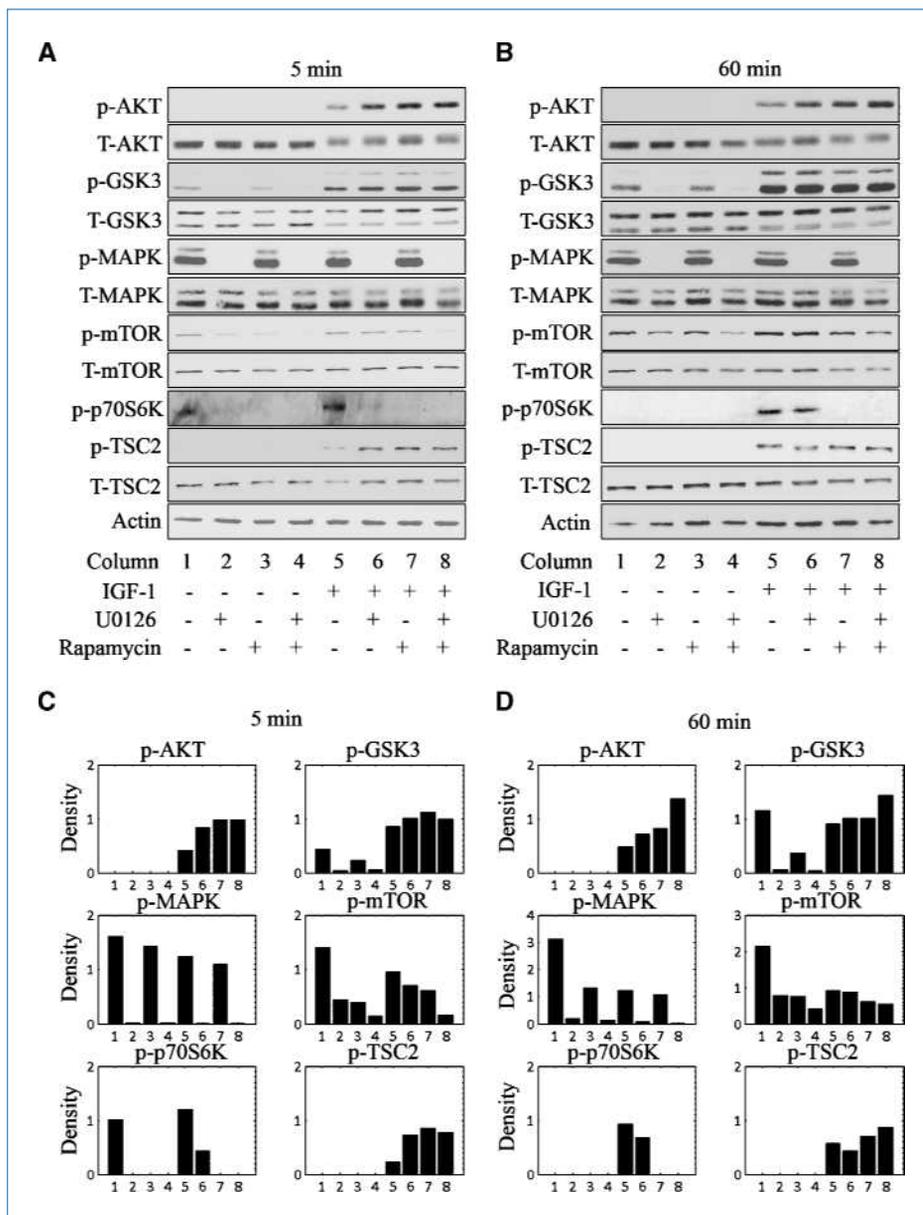
**Combined inhibition of targeted molecules.** Predicting the response of IGF1R networks to the inhibition of individual molecules may not necessarily identify optimal drug combinations for pharmacologic intervention. In contrast, perturbing all molecules in the network simultaneously would identify optimal combinations needed to inhibit aberrant signaling. From a computational point of view, the effect of small interfering RNAs can be mimicked by varying the initial concentration of signaling proteins. The effect of the drug

inhibitor can be approximately simulated by varying the values of rescaled kinetic rate constants, such as in the classic example of competitive inhibition (48). Integrating mass action modeling with a random sampling of kinetic constants (inhibitors) and initial protein concentrations (small interfering RNAs) within predefined intervals of values would thus provide an unbiased, unsupervised means of computationally predicting the effect of simultaneously perturbing all molecules in the IGFR network (computational procedures).

Combinations of signaling targets were identified by comparing the model parameters inferred using PSO from data measured in MDA-MB231 cells with randomly sampled model parameters that could restore user-defined signaling output. We defined the signaling network characterized by the

measured time courses of p-AKT, p-MAPK, and p-p70S6K as aberrant, as these proteins are often upregulated in many tumors (47). We defined the state at which p-AKT, p-MAPK, and p-p70S6K levels were inhibited by at least 5-fold as user-defined signaling output. To obtain reliable results, we identified 200 collections of combined perturbations (Supplementary File S3) that restored the user-defined signaling changes in MDA-MB231 cells for the sets of parameters listed in Supplementary File S2.

The most influential targets were scored according to the absolute value of the median deviation to the SD ratio (Supplementary Table S1). The computational analysis was repeated in triplicate using the same three randomly selected sets of parameters used to predict individual



**Figure 5.** Combined inhibition of MEK and mTOR in MDA-MB231 cells. A and B, protein levels were detected in unstimulated cells in the absence of inhibition (column 1) and after 1 h of incubation with the MEK and/or mTOR inhibitors (columns 2–4), and in cells stimulated with IGF-1 in the absence of inhibition (column 5) and after 1 h of incubation with the MEK and/or mTOR inhibitors (columns 6–8). Cells were stimulated with 75 ng/mL IGF-1 for 5 min (A) or 60 min (B). C and D, density of the bands after normalization with respect to actin.

inhibition of targeted molecules. Despite being ranked in a different order, the top five targets were the same for the three sets. All targets that scored as influential were characterized by a positive median deviation, which indicated activation of the reactions leading to inhibition of phosphorylated protein. Because p-IGFR, p-IRS-1, p-MEK, p-MAPK, and p-AKT were scored as the most influential targets, combined inhibition of the PI3K/AKT and MAPK pathways was predicted to optimally facilitate disruption of the loops responsible for aberrant signaling in MDA-MB231 cells.

### Experimental validation of modeling predictions

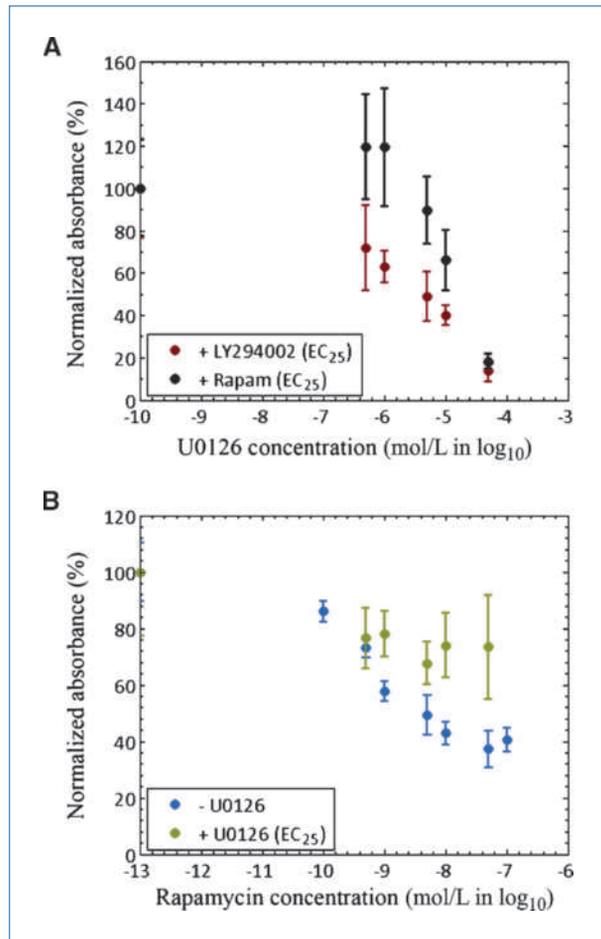
**The effect of drug combinations on the IGF network in the MDA-MB231 cell line.** We used immunoblotting to determine whether the combined inhibition of the MAPK and PI3K/AKT pathways would decrease the levels of p-AKT, p-MAPK, and p-p70S6K, and minimize changes in phosphorylation of other signaling proteins in the network. We also tested the combination of MEK and mTOR inhibitors to determine whether targeting pathways that differ from the predicted optimal combination would restore user-defined signaling changes in the MDA-MB231 cells.

Figures 4A and B show the levels of p-AKT, p-MAPK, p-GSK3, p-mTOR, p-p70S6K, and p-TSC2 detected in the absence of inhibition, and after 1 hour of incubation with the MEK and/or PI3K inhibitors in unstimulated cells and IGF-stimulated cells. The experimental data indicated that, in IGF-stimulated cells inhibited with MEK and PI3K inhibitors (column 8), the levels of all phosphorylated proteins were significantly decreased compared with those of the corresponding proteins detected in noninhibited, IGF-stimulated cells (column 5).

Figure 5A and B show the levels of p-AKT, p-MAPK, p-GSK3, p-mTOR, p-p70S6K, and p-TSC2 measured in the absence of inhibition, and after 1 hour of incubation with the MEK and/or mTOR inhibitors in unstimulated cells and IGF-stimulated cells. The experimental data indicated that p-MAPK, p-mTOR, and p-p70S6K levels in IGF-stimulated cells inhibited with MEK and mTOR inhibitors (column 8) were decreased compared with those of the corresponding proteins in noninhibited, IGF-1-stimulated cells (column 5). However, the p-AKT, p-GSK3, and p-TSC2 levels were increased.

Supplementary Table S2 shows the qualitative comparison between the measured and predicted differential levels of phosphorylated proteins in IGF-stimulated cells in the absence of inhibition, and after 1 hour of incubation with MEK and PI3K inhibitors or MEK and mTOR inhibitors. Note that the experimental results agree with the modeling predictions for both drug combinations. Therefore, as predicted by the mass action model, combined inhibition of the MAPK and PI3K/AKT pathways optimally inhibited aberrant networks, but combinations of MEK and mTOR inhibitors did not decrease the levels of p-AKT, p-MAPK, and p-p70S6K, and increased phosphorylation of nontargeted protein.

**Cell sensitivity to drug combinations.** Optimal inhibition of abnormal signaling networks must inhibit regulatory loops and redundant bypass to ultimately overcome the mechanism of feedback compensation that ensures cancer cell



**Figure 6.** Response curve of MDA-MB231 cells to dose concentration of drug inhibitors. A, cells were left nontreated as a control and incubated with LY294002 at its EC<sub>25</sub> (3.8  $\mu$ mol/L) or rapamycin at its EC<sub>25</sub> (0.1 nmol/L) in combination with U0126 at a concentration of 0.5 to 50  $\mu$ mol/L. B, cells were left nontreated as a control and incubated only with rapamycin at concentrations of 0.1 to 100 nmol/L or with a combination of U0126 at its EC<sub>25</sub> (3.5  $\mu$ mol/L) and rapamycin at a concentration of 0.5 to 50  $\mu$ mol/L. Absorbance was normalized with respect to the value detected for the controls and was expressed as a percentage. Solid circles, mean of normalized absorbance; bars, SD.

viability. To quantify the sensitivity of the MDA-MB231 cell line to MEK and PI3K, or MEK and mTOR inhibition, we used cell viability assays.

Figure 6A shows the effect of drug combinations on the viability of MDA-MB231 cells, which was measured as the normalized absorbance of viable cells as a function of increasing MEK inhibition (U0126) in combination with PI3K inhibitor (LY294002) or mTOR inhibitor (rapamycin). The experimental results indicated that inhibiting cells with a combination of LY294002 and increasing concentrations of U0126 resulted in a dose-dependent decrease in cell viability. In contrast, cells treated with a combination of rapamycin and increasing concentrations of U0126 showed no change in cell proliferation up to 1  $\mu$ mol/L of U0126 followed by partial rescue of cells with rapamycin. Combined MEK-PI3K

inhibition monotonically decreased cell viability, likely as a result of the optimal inhibition of the signaling pathways that led to inactivation of phosphorylated proteins (Fig. 4). In contrast, combined MEK-mTOR inhibition increased cell viability at low concentrations of the U0126 inhibitor, likely as a result of the nonoptimal inhibition of the signaling network that led to activation of p-AKT (Fig. 5). At high concentrations of the U0126 inhibitor, cell viability was significantly decreased for both drug combinations, likely as a result of U0126 inhibitor toxicity (Supplementary Material S3).

To test whether the combination of MEK and mTOR inhibitors rescued cell proliferation by activation of p-AKT, we performed cell viability assays with rapamycin alone or with rapamycin in combination with U0126. The experimental results shown in Fig. 6B indicated that addition of a MEK inhibitor to cells treated with rapamycin increased cell viability from 40% to 73% and rescued cells from cell death. Therefore, the experimental results suggested that optimal inhibition of aberrant signaling through combined inhibition of the MAPK and PI3K pathways was correlated with decreased cell viability. In contrast, nonoptimal combined targeted inhibition led to inadequate inhibition of the signaling networks and increased cell viability.

## Discussion

Integrating mass action modeling with optimization schemes is a quantitative approach to train ODE models using experimental data and identify optimal drug combinations that can inhibit signaling networks. PSO converged to more optimal solutions than did other optimization algorithms, including simulated annealing and genetic algorithms. Supplementary Table S3 summarizes the performance of the three algorithms in training the reduced mass action model against time courses of proteins (Supplementary File S4). Each simulation was repeated three times with different random seeds of the reduced model unknown parameters.

The most simple and intuitive strategy to inhibit aberrant networks consists of inhibiting the input sources that trigger signal transduction. Thus, individual inhibition of IGFR could

restore user-defined pathways in MDA-231 cells. However, constitutive p-MAPK activation driven by K-Ras and B-Raf mutations impairs this approach. The experimental results shown in Figs. 4 and 5, and the computational results shown in Fig. 3 also suggest that individual inhibition of targeted molecules frequently does not optimally inhibit cell signaling. A more effective inhibition of aberrant signaling is accomplished through multiple combined inhibitions of targeted molecules. The experimental results shown in Figs. 4 to 6 indicate that combined inhibition of the MAPK and PI3K/AKT pathways optimally inhibited the signaling networks and decreased cell viability. In contrast, combined inhibition of the MAPK and mTOR cascades led to significant activation of p-AKT and increased cell viability. Although several other kinases and pathways may potentially regulate the viability of the MDA-231 cells, the experimental results indicated that simultaneous inhibition of the MAPK and PI3K/AKT pathways was sufficient to significantly reduce cell proliferation.

In conclusion, we propose a computational procedure that can be used to rapidly generate experimentally testable intervention strategies that may lead to an optimal use of available drugs and the discovery of novel signaling targets. The procedure is currently being used to identify and validate drug combinations that can inhibit aberrant networks in a panel of human cancer cell lines.

## Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

## Grant Support

S. Iadevaia was supported in part by a training fellowship from the Pharmacoinformatics Training Program of the Keck Center of the Gulf Coast Consortia (NIH grant 5 T90 DK070109-04). This study was supported by the Kleberg Center for Molecular Markers, NIH P01CA099031, and The Komen Foundation (G.B. Mills), and DOD BC044268 and NIH R01CA125109 (P.T. Ram).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 02/23/2010; revised 07/02/2010; accepted 07/05/2010; published OnlineFirst 07/19/2010.

## References

- Manning BD, Cantley LC. AKT/PKB signaling: navigating downstream. *Cell* 2007;129:1261–74.
- Wullschlegel S, Loewith R, Hall MN. TOR signaling in growth and metabolism. *Cell* 2006;124:471–84.
- Seger R, Krebs EG. The MAPK signaling cascade. *FASEB J* 1995;9:726–35.
- Schlessinger J. Cell signaling by receptor tyrosine kinases. *Cell* 2000;103:211–25.
- Ullrich A, Schlessinger J. Signal transduction by receptors with tyrosine kinase activity. *Cell* 1990;61:203–12.
- Bhalla US, Iyengar R. Emergent properties of networks of biological signaling pathways. *Science* 1999;283:381–7.
- Weng G, Bhalla US, Iyengar R. Complexity in biological signaling systems. *Science* 1999;284:92–6.
- Ma'ayan A, Jenkins SL, Neves S, et al. Formation of regulatory patterns during signal propagation in a Mammalian cellular network. *Science* 2005;309:1078–83.
- Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet* 2007;8:450–61.
- Bhalla US. Understanding complex signaling networks through models and metaphors. *Prog Biophys Mol Biol* 2003;81:45–65.
- Justman QA, Serber Z, Ferrell JE, Jr., El-Samad H, Shokat KM. Tuning the activation threshold of a kinase network by nested feedback loops. *Science* 2009;324:509–12.
- Aldridge BB, Burke JM, Lauffenburger DA, Sorger PK. Physicochemical modelling of cell signalling pathways. *Nat Cell Biol* 2006;8:1195–203.
- Bhalla US, Ram PT, Iyengar R. MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network. *Science* 2002;297:1018–23.
- Birtwistle MR, Hatakeyama M, Yumoto N, Ogunnaike BA, Hoek JB, Kholodenko BN. Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses. *Mol Syst Biol* 2007;3:144.

15. Muller M, Obeyesekere M, Mills GB, Ram PT. Network topology determines dynamics of the mammalian MAPK1,2 signaling network: bifan motif regulation of C-Raf and B-Raf isoforms by FGFR and MC1R. *FASEB J* 2008;22:1393–403.
16. Tyson JJ, Chen K, Novak B. Network dynamics and cell physiology. *Nat Rev Mol Cell Biol* 2001;2:908–16.
17. Tyson JJ, Csikasz-Nagy A, Novak B. The dynamics of cell cycle regulation. *Bioessays* 2002;24:1095–109.
18. Alon U, Surette MG, Barkai N, Leibler S. Robustness in bacterial chemotaxis. *Nature* 1999;397:168–71.
19. Barkai N, Leibler S. Robustness in simple biochemical networks. *Nature* 1997;387:913–7.
20. Boman BM, Fields JZ, Bonham-Carter O, Runquist OA. Computer modeling implicates stem cell overproduction in colon cancer initiation. *Cancer Res* 2001;61:8408–11.
21. Edwards JS, Ibarra RU, Palsson BO. In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001;19:125–30.
22. Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators. *Nature* 2000;403:335–8.
23. Fussenegger M, Bailey JE, Varner J. A mathematical model of caspase function in apoptosis. *Nat Biotechnol* 2000;18:768–74.
24. Gardner TS, Cantor CR, Collins JJ. Construction of a genetic toggle switch in *Escherichia coli*. *Nature* 2000;403:339–42.
25. Balsa-Canto E, Peifer M, Banga JR, Timmer J, Fleck C. Hybrid optimization method with general switching strategy for parameter estimation. *BMC Syst Biol* 2008;2:26.
26. Chen WW, Schoeberl B, Jasper PJ, et al. Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol Syst Biol* 2009;5:239.
27. Hegger R, Kantz H, Schmuser F, Diestelhorst M, Kapsch RP, Beige H. Dynamical properties of a ferroelectric capacitor observed through nonlinear time series analysis. *Chaos* 1998;8:727–36.
28. Moles CG, Mendes P, Banga JR. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res* 2003;13:2467–74.
29. Rodriguez-Fernandez M, Mendes P, Banga JR. A hybrid approach for efficient and robust parameter estimation in biochemical pathways. *Biosystems* 2006;83:248–65.
30. Swameye I, Muller TG, Timmer J, Sandra O, Klingmuller U. Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc Natl Acad Sci U S A* 2003;100:1028–33.
31. Wang CC, Cirit M, Haugh JM. PI3K-dependent cross-talk interactions converge with Ras as quantifiable inputs integrated by Erk. *Mol Syst Biol* 2009;5:246.
32. Araujo RP, Liotta LA, Petricoin EF. Proteins, drug targets and the mechanisms they control: the simple truth about complex networks. *Nat Rev Drug Discov* 2007;6:871–80.
33. Rajasethupathy P, Vayttaden SJ, Bhalla US. Systems modeling: a pathway to drug discovery. *Curr Opin Chem Biol* 2005;9:400–6.
34. Schoeberl B, Eichler-Jonsson C, Gilles ED, Muller G. Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors. *Nat Biotechnol* 2002;20:370–5.
35. Yang K, Bai H, Ouyang Q, Lai L, Tang C. Finding multiple target optimal intervention in disease-related molecular network. *Mol Syst Biol* 2008;4:228.
36. Kennedy J, Eberhart R. Particle swarm optimization. Proceedings of the fourth IEEE International Conference on Neural Networks. Perth, Australia: IEEE Service Center; 1995, p. 1942–8.
37. Abraham A, Guo H, Liu H. Swarm intelligence: foundations, perspectives and applications. In: Nedjah N, de Macedo Mourelle L, editors. *Studies in Computational Intelligence*. Springer; 2006, p. 2–25.
38. Romano P, Manniello A, Aresu O, Armento M, Cesaro M, Parodi B. Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines. *Nucleic Acids Res* 2009;37:D925–32.
39. mdanderson.org. Houston: The University of Texas MD Anderson Cancer Center. [cited 2010 Feb 23]. Available from: <http://bioinformatics.mdanderson.org/OOMPA/>.
40. Vastrik I, D'Eustachio P, Schmidt E, et al. Reactome: a knowledge base of biologic pathways and processes. *Genome Biol* 2007;8:R39.
41. Easton JB, Kurmasheva RT, Houghton PJ. IRS-1: auditing the effectiveness of mTOR inhibitors. *Cancer Cell* 2006;9:153–5.
42. Sheehan KM, Calvert VS, Kay EW, et al. Use of reverse phase protein microarrays and reference standard development for molecular network analysis of metastatic ovarian carcinoma. *Mol Cell Proteomics* 2005;4:346–55.
43. Tibes R, Qiu Y, Lu Y, et al. Reverse phase protein array: validation of a novel proteomic technology and utility for analysis of primary leukemia specimens and hematopoietic stem cells. *Mol Cancer Ther* 2006;5:2512–21.
44. Mirzoeva OK, Das D, Heiser LM, et al. Basal subtype and MAPK/ERK kinase (MEK)-phosphoinositide 3-kinase feedback signaling determine susceptibility of breast cancer cells to MEK inhibition. *Cancer Res* 2009;69:565–72.
45. Stemke-Hale K, Gonzalez-Angulo AM, Lluch A, et al. An integrative genomic and proteomic analysis of PIK3CA, PTEN, and AKT mutations in breast cancer. *Cancer Res* 2008;68:6084–91.
46. Gonzalez-Angulo AM, Stemke-Hale K, Palla SL, et al. Androgen receptor levels and association with PIK3CA mutations and prognosis in breast cancer. *Clin Cancer Res* 2009;15:2472–8.
47. Hennessy BT, Smith DL, Ram PT, Lu Y, Mills GB. Exploiting the PI3K/AKT pathway for cancer drug discovery. *Nat Rev Drug Discov* 2005;4:988–1004.
48. Nelson DL, Cox MM, editors. *Lehninger principles of biochemistry*. New York: Worth Publishers; 2000.