Generating executable models from signaling network connectivity and semi-quantitative proteomic measurements

Derek Ruths*
School of Computer Science, McGill University, Quebec, Montreal Canada
**Email: derek.ruths@cs.mcgill.ca*

Luay Nakhleh
Department of Computer Science, Rice University, Houston, Texas USA
Email: nakhleh@cs.rice.edu

Executable biology is a discipline that is concerned with turning the specifications of a biological system into a computational model that can be simulated under different conditions to produce predictions about the behavior of the system. In this paper, we propose a computational framework consisting of a generalized execution strategy for signaling networks as well as a method for learning executable models from connectivity-maps and proteomic data pertaining to a specific signaling network. We call these data sources *semi-quantitative* because often they characterize the behavior of the system without providing reliably exact numerical measurements. To the best of our knowledge this is the first use of semi-quantitative data for building predictive models of biochemical systems.

Using our framework, we generate an executable model of a network of signaling pathways downstream of the epidermal growth factor receptor (EGFR) in the MCF-7 cell line. Using this executable model, we determine that our method performs as well as existing methods while using orders of magnitude less training data to achieve a comparable degree of accuracy.

1. Introduction

Within the domain of executable biology, an area of ongoing research and innovation concerns how to build executable models from biological data[1]. In this paper, we present a new modeling framework that uses connectivity maps and proteomic measurements to build executable models of signaling networks. We collectively call these data sources *semi-quantitative* because such data sets are often generated to the level of resolution that can identify trends, but not exact numerical quantities within the biological system (though it is important to note that, when desired, biologists can generate more precise measurements, though the effort can be significant and costly).

Our execution strategy features a simplified discrete-time representation of signal propagation in which each protein has a degradation rate parameter and each interaction has a weight parameter that abstractly models both strength and speed. The model for a specific signaling system is derived from an input connectivity map of protein interactions in the system. The model's parameter values are determined by solving an optimization problem in which values are indirectly constrained by semi-quantitative measurements taken from a set of perturbation experiments. This approach has several benefits.

Models for cell-specific signaling networks. Our method's use of perturbation experiments to learn parameter values makes it possible for biologists to build cell-specific executable models by providing perturbation experiments that characterize the behavior of a specific cell-line.

Large-scale model building and analysis. Because the discrete-time relationships can be derived directly from network connectivity and the non-linear optimization problem from raw experimental measurements, our framework can be automated (and, in fact, was automated when conducting all analyses discussed in this paper), making it well-suited to constructing executable models for preliminary and large-scale data sets. Since current high-throughput technologies can generate both—high-throughput microarrays often are used as a first-pass to identify potentially interesting features of a cell's signaling or gene regulatory network—our method provides researchers with the ability to analyze and use such vast amounts of data that are becoming available to them.

Modeling other cellular processes. Because the parameter learning process is independent of the specific execu-

*Corresponding author.

tion strategy we developed to capture signaling dynamics, it is possible to extend our framework to other biological processes beyond cellular signaling such as gene regulation and metabolism. Such an extension would involve developing a discrete-time representation of the cellular process and formalizing how experimental measurements of that system would correspond to entities within that representation.

We validate our method by building a predictive model of a network of signaling pathways downstream of EGFR in the MCF-7 cell-line using previously published experimental results[2]. The trained model correctly predicts the effect of a perturbation on a protein's activity-level 90% (63 out of 70) of the time. This high success-rate is particularly favorable when compared with the method in Ref. 2 which trained on 20 perturbation experiments from the same data set (as opposed to the 3 used by our method) in order to achieve the same level of accuracy. Note that training a model to have 100% accurate predictions is often complicated by the fact that a priori knowledge used to bootstrap the model may itself be incorrect or incomplete. This is the case with the EGFR network we consider in this paper: the 10 inconsistencies between the experiments and our model's predictions suggest that the dynamics of several components of the Ras pathway may be influenced by factors besides those present in the model.

Additionally, in a closer investigation of the model constructed by our method, we find that paths with the strongest weights correspond to interactions with known significance in the MCF-7 cell-line. This suggests that executable models constructed by our method can be not only predictive, but also descriptive of the underlying signaling mechanisms which can be useful to biologists in better understanding the structural and dynamic properties of a signaling network that determine aspects of its behavior.

2. Materials and Methods

2.1. A simplified model of signaling network dynamics

Dynamic models of biochemical systems fall into two classes: continuous-time and discrete-time. Continuous-time schemes typically model the behavior of the system as a first-order differential equation $\frac{dY}{dt} = f(Y(t))$, where $Y(t)$ is a vector containing the values of the state variables at time t. The trajectory that the system state vector follows at time t is determined by some function of the current state, $f(Y(t))$.

Discrete-time models, in contrast, explicitly break time into a series of steps in which the behavior of the system is expressed as the inductive formula:

$$Y_{t+1} = f(Y_t) \tag{1}$$

where $f(x)$ is the transition function that evaluates to the next state visited after x. Often such discrete-time models are linear in the system state variables, in which case the state transition formula can be rewritten $Y_{t+1} = AY_t$, where A is the transition matrix. In models of metabolic networks, A corresponds to the stoichiometric matrix. This correlation does not extend to signaling systems, however, since the underlying biochemical reactions are rarely explicitly modeled. Regardless of the interpretation of A, a given state variable y^i_{t+1} is determined by

$$y^i_{t+1} = \sum_{1 \le j \le |Y|} a_{i,j} y^j_t$$

where $a_{i,j}$ is the element of A at row i, column j. Thus, the system's next state depends entirely upon the current state and the elements of A. These $a_{i,j}$ are the parameters of the system. Once the values of these have been determined and a starting condition, Y_0 has been specified, the model is complete.

Though continuous-time models seem to express the biochemical processes more accurately (the underlying system is spatially and temporally continuous in nature), discrete-time models enjoy a number of practical advantages over continuous-models that can make them the better suited for certain types of problems: (1) though the underlying biochemical systems may be continuous, time-series data is inherently discrete, representing one or more time points at which the state of the system was observed; (2) the inductive structure of Equation 1 makes it easy to derive the state space of the system; and (3) Equation 1 allows the explicit derivation of the finite sequence of states visited given a starting state and a number of time steps.

The third property is of particular interest to us here as we use the finiteness property of this sequence to efficiently find parameter values for a model that satisfy certain semi-quantitative properties. In order to take advantage of this finite state sequence property, we build a

discrete-time model of a signaling network with the form:

$$y^i_{t+1} = \max(\delta_i y^i_t + \sum_{j \in A_i} w_{j,i} y^j_t - \sum_{j \in H_i} w_{j,i} y^j_t, 0). \quad (2)$$

State variable i corresponds to the activity-level of a signaling protein, δ_i is the degradation rate of that protein, A_i are other proteins in the system that activate i, and H_i are other proteins in the system that inhibit i. Since A_i and H_i specify the proteins that interact directly with i, the A_i's and H_i's for all i's in the system constitute the *connectivity* of the system—the directed interactions that connect the proteins in the system together. The parameter $w_{j,i}$ denotes the strength of the effect that j has on i through the interaction that connects them. When the parameters δ_i and $w_{i,j}$ are specified and a starting point is selected, the resulting system can be simulated by iteratively evaluating the state equations for increasing values of time, t. Models similar to this have been used to capture transcriptional dynamics (e.g., Refs 3, 4).

Note that the model shown in Equation 2 is effectively a system of linear discrete mathematical formulae with discontinuities at zero. Within this project, we consider it an executable model for two reasons. First, and most fundamentally, we use the state equations to execute the model precisely as expressed using a computer (making this model one that is 'executed'). This is different from other mathematical models, such as ODEs, which are *simulated*, meaning that their behavior is approximated by computational evaluation. Second, as will be discussed in later sections, we select parameters values for the model by treating it as a computational model.

2.2. Semi-quantitative data from perturbation experiments

To determine values for δ_i and $w_{i,j}$, we require semi-quantitative data from perturbation experiments. A perturbation experiment activates or inhibits the function of one or more proteins (called *targets*) through the use of various mechanisms such as drugs, gene knockouts, or siRNA. These perturbations have varying effects on the response of other proteins and cell phenotypes to signaling events. For a given signaling protein, the perturbation's effect is measured by comparing the activity-level of that protein in an unperturbed cell to the activity-level of the same protein under the perturbed condition. Ordinarily the cell is stimulated prior to measuring the activity-levels in order to determine how the perturbed protein(s) influence the signal that reaches other proteins.

Given the unperturbed and perturbed activity-levels for proteins X, Y, and Z (X_u and X_p, Y_u and Y_p, Z_u and Z_p, respectively), we can make semi-quantitative assertions about the effect of the perturbation on the activity-level of each protein: $X_u < X_p$ if X increased in response to the perturbation, $Y_u > Y_p$ if Y decreased, and $Z_u = Z_p$ if Z exhibited no change.

It is possible to make many other kinds of semi-quantitative assertions about the experimental results. For example, the biologist may observe that the perturbed concentration of Z is greater than that of Y: $Z_p > Y_p$; or that the unperturbed value of Z appears to be two times that of X: $Z_u = 2X_u$. In fact, any observations taking such forms can be used to constrain the parameter values of the model. However, using such constraints must be done with great care since comparison across protein types and conditions may not be meaningful due to differing concentrations and measurement accuracy for various protein types.

However, for the remainder of this paper, we consider (without loss of generality) the three fundamental assertions: $Y_u < Y_p$, $Y_u > Y_p$, and $Y_u = Y_p$ as the types of semi-quantitative data that constrain the training process.

2.3. Training a model using semi-quantitative data

Given the connectivity for a signaling network of interest—the sets A_i and H_i for all proteins i in the system—we designed a training method that takes a set of semi-quantitative data from perturbation experiments and infers values for the parameters δ_i and $w_{i,j}$ that make the resulting model reproduce the maximum number of semi-quantitative behaviors specified possible (when the appropriate perturbed conditions are simulated).

Our method works by converting the model and the semi-quantitative data into a series of constraints for a non-linear optimization problem. The optimization algorithm is directed to find values for all δ_i and $w_{i,j}$ such that the model's behavior satisfies as many semi-quantitative data constraints as possible.

2.3.1. *Modeling perturbation experiments*

Note that a perturbation experiment can be characterized as the set of inhibited proteins, $P \subseteq \mathbb{P}$. The perturbed sig-

naling network is structurally the same as the unperturbed network except where the perturbation has its effect. As a result, the state equations of the perturbed network, S^P, are largely the same as those in the unperturbed network, S^0:

$$S^P[i] := \begin{cases} S^0[i] & \text{if } i \notin P \\ y^i_{t+1} = 0 & \text{if } i \in P \end{cases}$$

where $S^X[i]$ is the state equation for protein i under condition X (the set of inhibited proteins).

Given the state equations for a perturbation experiment, S^P, and the unperturbed signaling network, S^0, we can compute the semi-quantitative change in protein i's activity-level due to the perturbation by simulating both networks from some initial state Y_0. The predicted semi-quantitative change in protein i is:

$$\hat{q}^P_i = \begin{cases} < \text{ if } \Delta^P_i < -\epsilon \\ > \text{ if } \Delta^P_i > \epsilon \\ = \text{ if } -\epsilon \leq \Delta^P_i \leq \epsilon \end{cases}$$

where $\Delta^P_i = \sum_{0 \leq t \leq T}(S^0[i,t] - S^P[i,t])$ is the difference in the activity-level of protein i over the time of the simulation between the unperturbed (S^0) and perturbed (S^P) conditions ($S^X[i,t]$ denotes the value of the state equation for protein i at time t under condition X). The ϵ parameter is incorporated into the definition in order to desensitize the measure to extremely small, probably insignificant, changes (e.g., $\Delta^P_i = 10^{-12}$ most likely does not indicate a change of any significance).

2.3.2. *Training a model using semi-quantitative data from perturbation experiments*

To train a model, S^0, a set of perturbation experiments, $\mathbf{P} = \{P_1, P_2, ..., P_n\}$ and semi-quantitative perturbation experiment observation, $Q = \{(x_1, p_1, q_1), (x_2, p_2, q_2), ..., (x_R, p_R, q_R)\}$, are provided. A semi-quantitative perturbation experiment observation, $(x, p, q) \in Q$ specifies the response of protein p to perturbation P_x: $q \in \{<, >, =\}$ indicates the way that the activity-level of protein p changed in response to perturbation P_x with respect to the unperturbed system S^0.

The objective of the training procedure is to select an initial condition, Y_0, degradation rates, δ_i, and interaction weights, $w_{i,j}$, such that when the original and perturbed systems are simulated (S^0 and $S^{P_1}, S^{P_2}, ..., S^{P_n}$, respectively), $\hat{q}^{P_x}_p = q$ is true for as many semi-quantitative results, $(x, p, q) \in Q$, as possible.

As with most training procedures, ours is a search for parameter values that cause the model to which they belong to behave in a certain way. We formalize the parameter search as a non-linear optimization problem in which the parameters are free variables constrained by (1) the state equations in S^0 and S^P, (2) the semi-quantitative behavioral assertions, Q, and (3) a set of logical constraints: $0 \leq \delta_i \leq 1$ (the activity-level of a protein can never fall below zero), and $w_{i,j} \geq 0$ (the effect of a protein can not be negative)[a].

In order to build the non-linear optimization problem, a simulation time, T, must be specified. Optionally, a set of weights for individual constraints can be specified $\Omega = \{\omega_1, ..., \omega_{|Q|}\}$. Conceptually, these weights can be used to make the optimizer favor satisfying certain constraints over others. If Ω is not specified, all constraints are assumed to be equally important (i.e. $\omega_i = 1$ for all $1 \leq i \leq |Q|$)[b]. The problem is then constructed as follows:

- **Free variables**
  - $S^0[i,t]$ - the activity-levels for each protein, $1 \leq i \leq N$, for each time step, $t \in \{0, 1, ..., T\}$, in the original network
  - $S^{P_k}[i,t]$ - these are the activity-levels for each protein, $1 \leq i \leq N$, for each time step, $t \in \{0, 1, ..., T\}$, in the network corresponding to the perturbation experiment, P_k
  - $0 \leq \delta_i \leq 1$ - the degradation rate of each protein
  - $w_{i,j} \geq 0$ for all interactions - the interaction weight of each edge in the network
  - $X[r] \in \{0, 1\}$ for all semi-quantitative data constraints $1 \leq r \leq R$.
- **Constraints**
  - State equations for the unperturbed and

[a] It is worth noting that, because this non-linearity takes such a regular form, we suspect that there may be more optimal search strategies than a general non-linear optimization algorithm. We identify this as a topic for future work.

[b] We have not used these weights in our formulation (i.e., $\omega_i = 1$ for all ω_i). These weights are included in this formulation since they may be useful for researchers using the method on data for which some observations are more certain or important than others.

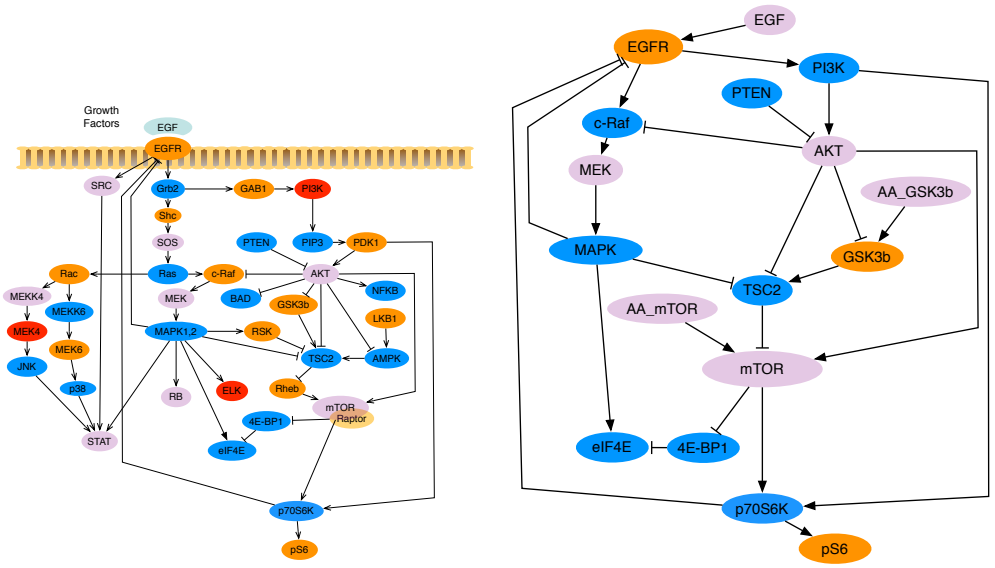


Fig. 1. (a) A detailed diagram of the EGFR signaling network. (b) The EGFR signaling network largely restricted to the proteins inhibited or measured in the experiments reported in *Nelander et al.*.

perturbed networks: S^0 and S^{P_i} for $P_i \in \mathbf{P}$

- Semi-quantitative changes due to perturbations as characterized in Q:
  - The following rule is produced for all rules $(x_r, p_r, \text{`<'})$: (proteins that increased in response to the perturbation)

$$X[r]\sum_{t=0}^{T}(S^0[p_r,t] - S^{P_{x_r}}[p_r,t]) < X[r](-\epsilon_r)$$

  - The following rule is produced for all rules $(x_r, p_r, \text{`>'})$: (proteins that decreased in response to the perturbation)

$$X[r]\sum_{t=0}^{T}(S^0[p_r,t] - S^{P_{x_r}}[p_r,t]) > X[r]\epsilon_r$$

  - The following rule is produced for all rules $(x_r, p_r, \text{`='})$: (proteins that did not change in response to the perturbation)

$$X[r]\left|\sum_{t=0}^{T}(S^0[p_r,t] - S^{P_{x_r}}[p_r,t])\right| \leq X[r]\epsilon_r$$

- **Objective function:** maximize $\sum_{r=1}^{R}\omega_r X[r]$

The choice to use ϵ_r rather than a strict inequality was based on the need to ensure that the optimization algorithm did not satisfy the condition using a trivial difference (e.g., 10^{-20}) and the desire to incorporate support for changing the difference thresholds that signaled a semi-quantitative change (recall the use of a similar ϵ parameter earlier in the definition of Δ_i). At present, the value of ϵ is manually selected by the biologist in order to influence what changes are considered significant. Better characterizing choices of ϵ and the effects these have on the outcome of the training procedure are important directions for future work.

When all constraint weights are equal (e.g., $\omega_r = 1$ for all r), then the objective function forces the optimization algorithm to find parameter values that satisfy the maximum number of semi-quantitative constraints. Giving the optimization algorithm the flexibility to ignore specific constraints is important since certain network structures might make satisfying some semi-quantitative constraints impossible. In these cases, rather than failing outright, the optimization algorithm simply satisfies all other semi-quantitative constraints.

The constraint weights, $\Omega = \{\omega_1, ..., \omega_{|Q|}\}$, are used to bias the optimizer towards satisfying certain perturbation constraints over others. This is useful when some experimental results have higher confidence than others.

In such cases, the more highly supported experimental result constraints can be given larger weights in order to cause the optimizer to favor satisfying them over other results in which the researcher has less confidence.

The resulting non-linear optimization problem generated as described can be solved by a variety of off-the-shelf optimization algorithms. All results generated in this paper were produced using the BONMIN software package[5].

A web-based interface for this method is available at `http://www.ruthsresearch.org/monarch`.

3. Results and Discussion

We evaluated our method's performance on a series of perturbation experiments conducted on the MCF-7 cell-line and published in Ref. 2. In these experiments, a series of proteins were targeted: EGFR (ZD1839), mTOR (rapamycin), MEK (PD0325901), PKC-δ (rottlerin), PI3-kinase (LY294002), and IGF1R (A12 anti-IGF1R inhibitory antibody). In total, 21 different perturbation experiments were conducted. In each, one or two of these molecules were inhibited, after which EGF stimulation was applied. Phospho-levels for several proteins were measured at the end of each experiment: p-AKT-S473, p-ERK-T202/Y204, p-MEK-S217/S221, p-eIF4E-S209, p-c-RAF-S289/S296/S301, p-P70S6K-S371, and pS6-S235/S236. The effect of these perturbations on two phenotypic processes, cell cycle arrest and apoptosis, were also measured.

For our analysis, we considered a subset of molecules involved in signaling directly downstream of EGFR; this network (hereafter, the *EGFR network*) is shown in Figure 1(a). Because several of its members have known oncogenic properties, this network is of significant interest to the biomedical research community. Based on this subset, we considered all protein targets except IGF1R and PKC-δ—both of which are not recognized members of EGFR signaling[6, 7]. This provided a set of 10 perturbation experiments (out of the 21 in Ref. 2). We included phospho-levels for all proteins measured. Since our current methods are focused on signaling processes, we did not consider the two phenotypic processes since these are the result of a combination of signaling, transcriptional, and metabolic processes.

The network in Figure 1(a) was reduced in order to minimize the number of proteins and interactions in the model for which measurement information was not available. The motivation for this is to limit the number of parameters whose values are unconstrained by observations, which otherwise makes the parameter space much larger. Clearly, however, it is desirable to support such unmeasured proteins in a predictive model. We identify the problem of extending our methods to handle such unconstrained signaling members as a direction for future work.

The connectivity for the network induced by the measured molecules is shown in Figure 1(b). This reduced form of the EGFR network was obtained by keeping only proteins that either (1) were targets, (2) were measured, or (3) were required to maintain connectivity among targets and measured proteins in a non-trivial way. GSK3b was retained in order to ensure that TSC2 had at least one activating input. The molecules AA_mTOR and AA_GSK3b were added in order to model significant sources of activity that reside outside of the EGFR network (GSK3b activity is largely determined by environmental factors and mTOR is activated by Rheb which maintains a high basal activity-level).

To test the predictive ability of our method, we performed a cross-validation procedure in which the model parameters were trained using semi-quantitative data from three experiments. The resulting model was then used to simulate the remaining seven perturbation conditions.[c] The predicted activity-levels from these simulations were interpreted as semi-quantitative observations (e.g., the perturbation caused an increase/decrease/no-change in p-AKT). These predicted observations were then compared to the true semi-quantitative changes in the data. The correctness of the trained model was taken to be the percent of predictions that agreed with the semi-quantitative experimental results.

[c]This training-to-testing proportion was selected in order to emphasize the smaller data-requirements of our method, which will be discussed in the following section.

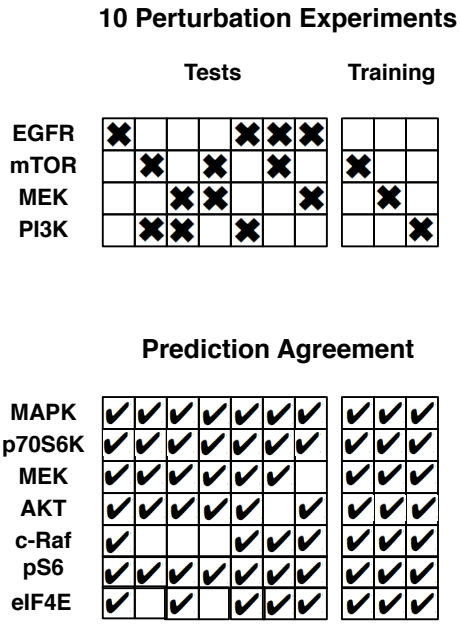


Fig. 2. The agreement of one of the best trained model's predictions with perturbation experiments reported in *Nelander et al*. Columns are the individual experiments, rows correspond to molecules. The columns set apart to the far right constitute the three experiments used to train the model. In the perturbation experiments matrix, a bold "**x**" indicates inhibited molecules. In the prediction agreement matrix, a "✓" square indicates that our method's prediction for that molecule in that condition agreed with the experimental measurement. Our method correctly predicted 90% (63 out of 70) of the test experiment measurements.

Each different triplet of perturbation experiments yielded a different parameterized model (the full set of training triplets and their predictive accuracy is provided in the Supporting Information). Most triplet training sets yielded models with $> 70\%$ accuracy. The best trained models obtained had 90%, 63 out of 70, predictive accuracy (70 data points = [7 observations per experiment] $\times$ [10 different experimental conditions]). One of these models was selected for further analysis and is shown in Figure 2. As a point of comparison, the predictive model reported in Ref. 2 was trained and tested on this same data set. Though they trained their method on 20 of the 21 experiments, their method's ability to recall the correct semi-quantitative change for a given molecule in a specific perturbation experiment was also 90% (63 out of 70). Thus, despite using much less and only semi-quantitative interpretations of the experimental data, our method was able to predict the behavior of individual molecules with a comparable degree of accuracy.

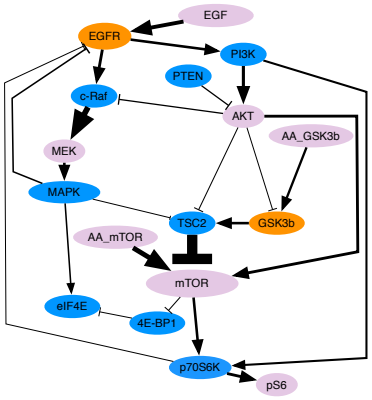
The 10 disagreements between our method's predictions and the experimental data may be due to cell-specific signaling properties, some of which are suggested in Ref. 2. It is worth noting that the majority of errors occur along the c-Raf pathway (i.e., c-Raf, MEK, and eIF4E). Far from being a random distribution of discrepancies throughout the network, the concentration of inconsistencies in this pathway suggests that this part of the model is incomplete. Several discrepancies arise for c-Raf under three different perturbations. c-Raf is known to be activated by Ras and by various isoforms of PKC, none of which is PKC-δ[6, 7]. Nonetheless, Nelander *et al.* detect a significant interaction between PKC-δ and c-Raf suggesting that, in the MCF-7 cell-line, this isoform may have some interaction with c-Raf. The absence of such a signaling mechanism in our model could well account for the inconsistencies concerning c-Raf.

The discrepancies in the dynamics of eIF4E under the MEK/mTOR perturbation may be related to the complicated mechanisms actually governing eIF4E. Experimental results report eIF4E *increasing* in response to this perturbation. Regardless of parameter values, the connectivity of our model cannot explain this since MAPK and mTOR are the only activators of eIF4E activity. This suggests that the increase in eIF4E activity in response to this perturbation is either the result of an entirely different mechanism or experimental error.

Both our method and the method in Ref. 2 generated discrepancies when predicting the response of AKT to the EGFR/mTOR perturbation. Under the perturbation, AKT is reported to have shown no change (0.0 fold increase). While it is certainly possible for AKT to have not changed, it is also possible that the change (up or down) was sufficiently small as to not register as a change during analysis: note that in Ref. 2, the AKT blots are quite dark and cover much of the channel, factors that make discerning small fold changes more difficult. It is also possible that AKT signaling occurs differently in the MCF-7 cell-line due to a known mutation in PIK3CA (the catalytic subunit of PI3K) which causes MCF-7 cells to have higher basal levels of AKT phosphorylation than normal cells[7].

Like AKT, the MEK activity inconsistencies under the EGFR/MEK perturbation may be the result of the existence of some mechanism not present in our model. Typically, MEK is activated through the pathway

EGFR $\rightsquigarrow$ c-Raf $\rightsquigarrow$ MEK. However, MEK is observed to increase while c-Raf activity drops, which cannot be explained by interactions in the model. Thus, other cell-specific signaling pathways may dominate MEK's activity under this perturbation. Close inspection of the western blots for c-Raf in Ref. 2 also raise the possibility that the reported changes are simply artifacts of the western blots themselves.



Interaction	Weight
4E-BP1 ⊣ eIF4E	0.158
AA_GSK3b → GSK3b	0.462
AA_mTOR → mTOR	0.939
AKT ⊣ c-Raf	0.271
AKT ⊣ GSK3b	0.162
AKT → mTOR	0.569
AKT ⊣ TSC2	0.214
c-Raf → MEK	1.130
EGF → EGFR	0.762
EGFR → c-Raf	0.554
EGFR → PI3K	0.491
GSK3b → TSC2	0.481
mTOR ⊣ 4E-BP1	0.199
mTOR → p70S6K	0.522
MAPK → eIF4E	0.321
MAPK ⊣ EGFR	0.299
MAPK ⊣ TSC2	0.055
MEK → MAPK	0.476
p70S6K ⊣ EGFR	0.157
p70S6K → pS6	0.546
PI3K → p70S6K	0.389
PI3K → AKT	0.616
PTEN ⊣ AKT	0.271
TSC2 ⊣ mTOR	2.127

Fig. 3. The EGFR signaling network model with relative interaction weights depicted by the width of arrows.

3.1. Interpretation of Interaction Weights

In addition to predictive capabilities, our method produces a model whose parameters have been derived from experimental data. There are several aspects of the interaction weights (shown in Figure 3) inferred for the EGFR network in the MCF-7 cell-line that offer insights into cell-specific signaling properties. The four heaviest pathways in the network are:

- EGFR $\rightarrow$ c-Raf $\rightarrow$ MEK $\rightarrow$ MAPK,
- EGFR $\rightarrow$ PI3K $\rightarrow$ AKT,
- EGFR $\rightarrow$ PI3K $\rightarrow$ p70S6K $\rightarrow$ pS6, and
- AA_GSK3b $\rightarrow$ GSK3b $\rightarrow$ TSC2 $\dashv$ mTOR.

Notice that the first three constitute the three ways in which EGF signal enters the network through the receptor. The interaction weights suggest a relative ordering in the strength of these different signaling paths (listed by signaling endpoint): pS6 $<$ AKT $<$ MAPK.

Cell-specific behavior of AKT. Our model suggests that the EGFR $\rightsquigarrow$ AKT pathway is much less significant than the c-Raf pathway. This is a surprising result when the general significance of the PI3K pathway is considered. Our method appears to have identified a cell-specific attribute, since MCF-7 has a PI3K mutation that induces the constitutive overexpression of AKT[8]. Additional evidence in support of this hypothesis is that, in our model, AKT was given a degradation rate slower than the network average degradation rate (approximately one standard deviation higher than the network-wide average degradation rate of 0.47, see Supporting Information) which will cause AKT to maintain its activity-level for longer than other members of the network.

Also notice that the relative strengths of $EGFR \rightsquigarrow MAPK$ and $EGFR \rightsquigarrow AKT \rightarrow mTOR$ suggest a relative ordering of the negative feedback loops that regulate $EGFR$. Because the $MAPK \dashv EGFR$ interaction receives stronger signal than the $p70S6K \dashv EGFR$ interaction, it is likely the case that in the MCF-7 cell-line, $MAPK$ is the stronger negative regulator of EGFR. This coincides with the results in Ref. 2 in which they found significant evidence of negative regulation of EGFR by MAPK, but no indication for that of p70S6K.

Tumor cell use of GSK3b. GSK3b participates in regulating a number of important cellular processes including cell cycle and energy metabolism[9]. A mounting body of experimental evidence also suggests that it may be a mechanism by which cancer cells satisfy their significant energy demands. The strong activation of GSK3b (and the very strong inhibition of mTOR) in our model may be an indication that MCF-7, a breast cancer cell-line, belongs to the class of tumor cells that up-regulates certain cell processes partially through increased GSK3b activity.

The presence of these pathways in our model as strong chains of interactions both provides additional evidence for the predictive capabilities of our method and demonstrates how the parameters of the models can be used to gain insights into the system being studied. These results also support the more general idea that semi-quantitative data alone is sufficient to gain insights into the relative importance and strength of interactions in a signaling network.

3.2. The Importance of Connectivity and Parameters

PROCEDURE RANDOMIZE(V,E)

(1) $D[v] = \text{degree}(v, E)$ for all $v \in V$

(2) $E' = \emptyset$

(3) For each $e \in E$

- $u, v, t = e$
- Choose $x \in V$ s.t. $D[x] > 0$
- $D[x] = D[x] - 1$
- Choose $y \in V$ s.t. $D[y] > 0$
- $D[y] = D[y] - 1$
- $E' = E' \cup \{(x, y, t)\}$

(4) Return $G' = (V, E')$

Fig. 4. The algorithm used to randomize the connectivity of a network $G = (V, E)$.

Within the context of work such as Ref. 10 which made predictions using only network connectivity (no parameters), an important question to answer is how much the presence of well-trained parameters contribute to the accuracy of this method. In order to understand the contribution of parameters and connectivity in this regard, we evaluated the accuracy achieved by a model (1) with the correct connectivity, but random parameter values, (2) random connectivity with trained parameter values, and (3) random connectivity and random parameter values. Correct connectivity corresponded to the connectivity in Figure 1; random connectivity corresponds to a network with all the nodes and edges in the correct network, connected in a randomized pattern (with only node degree preserved). The algorithm used to randomize a network's connectivity is shown in Figure 4. Trained parameters refers to using the optimal training data set to select good parameter values; random parameters refers to using parameter values selected within a range of 0 to 1 for retention parameters and 0 to 15 for interaction weights (note that various ranges for parameter values were tested with no change in the overall results we report next). For each scenario considered, 1000 networks were constructed and their accuracy tested against the 7 remaining data sets. Table 1 shows the outcome of the results.

Table 1. The contribution that correct connectivity and trained parameters make to overall model accuracy for the EGFR network.

Connectivity	Parameters	Accuracy
Correct	Trained	90% (63/70)
Correct	Random	59.3% (approx. 40/70)
Random	Trained	21.2% (approx. 15/70)
Random	Random	0.4% (approx. 3/70)

The results of these experiments indicate that connectivity is, by far, the most significant contributor to the accuracy of the model's predictions. Even when random parameters are used, predictions are correct nearly 60% of the time. Having trained parameters, however, does have an impact on accuracy: evidenced by the fact that trained parameters increase accuracy by another 25%.

What these results also show is that training parameters is not always susceptible to the issue of overfitting. While there is always concern that a sufficiently complicated system can always be parameterized to produce certain behavior, for the EGFR network considered here, the degree of connective complexity could only be fit to 21% (approximately 15 out of 70 data points) of the experimental data through training of parameter values.

4. Conclusions

The abundance of semi-quantitative experimental data (raw measurements that have been distilled into high-level behavioral trends or classes) both online and in individual labs as well as databases of network topology (e.g., KEGG[11]) makes such data an appealing source of information from which to build predictive models of biochemical networks. In this paper, we have presented a novel computational method for building executable models of signaling networks. Furthermore, we have shown that the models produced from semi-quantitative data have descriptive capabilities: the parameters derived from semi-quantitative experimental data can provide insights into the underlying signaling mechanisms. Using our method, we have provided further evidence that network connectivity (one kind of semi-quantitative data) is a strong determinant of network dynamics. Taken as a whole, the work presented in this paper suggests that models built from such data can provide profitable predictions. As a result, this line of inquiry deserves further exploration and development.

5. Acknowledgements

We are grateful to Prahlad T. Ram for providing information on the MCF-7 cell-line and to the authors of Ref. 2 for their explanation of the mechanism of MEK inhibitor PD0325901.

This work was supported by a Seed Grant awarded to Luay Nakhleh from the Gulf Coast Center for Computational Cancer Research, funded by John and Ann Doerr Fund for Computational Biomedicine, as well as by the National Cancer Institute [grant number R01CA125109]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Cancer Institute or the National Institutes of Health.

References

1. Fisher J, Henzinger TA (2007) Executable cell biology. Nat Biotechnol 25: 1239–1249.
2. 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: 11.
3. Ciliberti S, Martin O, Wagner A (2007) Robustness can evolve gradually in complex regulatory gene networks with varying topology. PLoS Comput Biol 3: e15.
4. Kwon YK, Cho KH (2008) Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. Bioinformatics 24: 987–994.
5. Bonami P, Biegler L, Conn A, Cornuejols G, Grossmann I, et al. (2008) An algorithmic framework for convex mixed integer nonlinear programs. Discrete Optimization 5: 196–204.
6. Corbit K, Foster D, Rosner M (1999) Protein Kinase C delta mediates neurogenic but not mitogenic activation of mitogen-activated protein kinase in neuronal cells. Molecular and Cellular Biology 19: 4209–4218.
7. Sozeri O, Vollmer K, Liyanage M, Firth D, Kour G, et al. (1992) Activation of the c-Raf protein kinase by protein kinase C phosphorylation. Oncogene 7: 2259–2262.
8. She QB, Chandarlapaty S, Ye Q, Lobo J, Haskell KM, et al. (2008) Breast tumor cells with PI3K mutation or HER2 amplification are selectively addicted to AKT signaling. PLoS ONE 3: e3065.
9. Martinez A (2008) Preclinical efficacy on GSK-3 inhibitors: Towards a future generation of powerful drugs. Med Res Rev 28: 773–796.
10. Ruths D, Muller M, Tseng JT, 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: e1000005.
11. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research 28: 27–30.