BIOINFORMATICS ORIGINAL PAPER

Vol. 28 no. 21 2012, pages 2811–2818
doi:10.1093/bioinformatics/bts517

Systems biology

Advance Access publication August 24, 2012

Static network structure can be used to model the phenotypic effects of perturbations in regulatory networks

Ariel Feiglin[1], Adar Hacohen[1], Avital Sarusi[1], Jasmin Fisher[2], Ron Unger[1] and Yanay Ofran[1,*]

[1]The Goodman faculty of life sciences, Bar Ilan University, Ramat Gan 52900, Israel and [2]Microsoft Research Cambridge, 7 JJ Thomson Avenue, Cambridge CB3 0FB, UK

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Biological processes are dynamic, whereas the networks that depict them are typically static. Quantitative modeling using differential equations or logic-based functions can offer quantitative predictions of the behavior of biological systems, but they require detailed experimental characterization of interaction kinetics, which is typically unavailable. To determine to what extent complex biological processes can be modeled and analyzed using only the static structure of the network (i.e. the direction and sign of the edges), we attempt to predict the phenotypic effect of perturbations in biological networks from the static network structure.

Results: We analyzed three networks from different sources: The EGFR/MAPK and PI3K/AKT network from a detailed experimental study, the TNF regulatory network from the STRING database and a large network of all NCI-curated pathways from the Protein Interaction Database. Altogether, we predicted the effect of 39 perturbations (e.g. by one or two drugs) on 433 target proteins/genes. In up to 82% of the cases, an algorithm that used only the static structure of the network correctly predicted whether any given protein/gene is upregulated or downregulated as a result of perturbations of other proteins/genes.

Conclusion: While quantitative modeling requires detailed experimental data and heavy computations, which limit its scalability for large networks, a wiring-based approach can use available data from pathway and interaction databases and may be scalable. These results lay the foundations for a large-scale approach of predicting phenotypes based on the schematic structure of networks.

Contact: yanay@ofranlab.org

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 4, 2012; revised on July 31, 2012; accepted on August 14, 2012

1 INTRODUCTION

Analysis of regulatory networks has been used to explore the molecular mechanisms that underlie biological processes and to identify new regulatory modules (Hughey *et al.*, 2010; Przytycka *et al.*, 2010; Shamir and Karlebach, 2008). Network-based modeling tools are often used to account for the difference between healthy and diseased cells (Sorger *et al.*, 2011). They have also been used to predict the effect of perturbations, e.g. inhibiting one protein/gene or more, on other proteins/genes in the network (Li *et al.*, 2010; Maslov and Ispolatov, 2007; Mitsos *et al.*, 2009; Prill *et al.*, 2010; Wang *et al.*, 2009). For example, Ruths *et al.* successfully predicted changes in protein levels in the MAPK/AKT signaling network in breast tumor cells in response to perturbation of two genes (Ruths *et al.*, 2008). A wide variety of methods have been developed to analyze regulatory networks and provide predictions at different levels of detail (Fisher and Piterman, 2010; Shamir and Karlebach, 2008). The input for these methods ranges from quantitative kinetic parameters, such as reaction rates, to logical constraints defining the interactions (e.g. AND/OR relationships). Continuous methods such as ordinary differential equations (ODEs) model the rate of change of each component in the network and provide detailed quantitative information regarding the networks dynamics (Hughey *et al.*, 2010; Shamir and Karlebach, 2008). However, ODEs require comprehensive knowledge of kinetic parameters, which are unknown for most networks, and therefore their applicability is limited (Arisi *et al.*, 2006; Bailey, 2001; Papin *et al.*, 2005).

Alternatively, in discrete logic-based models, each component in the network has a discrete level, which is determined at every time step through a logical function. Two main methods in this category are Boolean networks (Glass and Kauffman, 1973; Thomas, 1973) and Petri nets (Reddy *et al.*, 1996). Both have been used to model regulatory networks and proven useful for gaining mechanistic insights and for predicting phenotypes (Chaouiya, 2007; Morris *et al.*, 2010). Such approaches do not depend on quantitative data but rather on the structure of the network along with a set of logical constraints. Although the low resolution of logic-based models imposes limitations on their predictive power, they are applicable to systems that could not be modeled with ODEs owing to lack of kinetic information (Ruths *et al.*, 2008).

High-throughput technologies are being used extensively to identify physical and functional relationships between proteins and genes on a large scale. These data are often deposited in interaction databases such as IntAct (Hermjakob *et al.*, 2010), Biogrid (Tyers *et al.*, 2011) and STRING (Jensen *et al.*, 2009). Pathway databases such as KEGG (Kanehisa *et al.*, 2010) and the Pathway Interaction Database (PID) (Schaefer *et al.*, 2009), which are manually curated, also rely on these data. They are typically represented by schematic diagrams, with no quantitative parameters or logical constraints. Consequently, continuous models or logic-based models are not applicable to most of the available interaction data and to most known networks.

*To whom correspondence should be addressed.

Despite this limitation, these data have previously been used to gain insight and suggest novel molecular mechanisms. For example, Ideker *et al.* used binary protein–protein and protein–DNA interactions to generate mechanistic hypotheses that explain experimental expression data (Ideker *et al.*, 2002). This work was later extended and used to infer regulatory pathways from subsets of candidate genes using a literature-derived network of biological relationships (Rajagopalan and Agarwal, 2005). More recently, an elegant method applying edge consistency in directed and signed networks was used to infer causal perturbations (Chindelevitch *et al.*, 2012; Enayetallah *et al.*, 2011).

Although the representation of biological networks as schematic diagrams is widespread, it is clearly a harsh oversimplification of reality and can only give a partial view of the real chain of events (Tyson *et al.*, 2001). However, as a large portion of physical and functional interaction data are available only at this level of abstraction, it is important to explore to what extent these networks capture the complexity of biological processes. To this end, we devised a simple algorithm for modeling regulatory networks, based on their schematic wiring alone. In particular, given a regulatory network that depicts known pairwise relationships within a biological process, this algorithm predicts, for each protein/gene, whether it will be upregulated, downregulated or will remain unchanged if one or more of the proteins/genes in the network is perturbed (e.g. inhibited by a drug). To determine the effect of a perturbation of one protein/gene on another protein/gene in the network, all the paths between them are assessed. We tested this approach on three different networks from three independent sources, for which experimental perturbation data were available. First, we attempted to predict the effect of drug perturbations on a relatively small but carefully constructed protein signaling network that originated in a detailed experimental study by Nelander *et al.* (Nelander *et al.*, 2008). This study also included a comprehensive set of drug perturbations and their experimental readouts that were used to validate our predictions. Next, we tested a small gene regulatory network derived from STRING (Jensen *et al.*, 2009). Finally, we predicted the effect of drug perturbations on a large network of hundreds of nodes, including all the National Cancer Institute (NCI)–curated pathways available in the Protein Interaction Database (Schaefer *et al.*, 2009). The experimental drug perturbation data used to evaluate our predictions for the two last networks were gleaned from microarray data from the connectivity map (CMAP) project (Golub *et al.*, 2006).

Our approach was able to correctly predict between 71% and 82% of experimental readouts, depending on the quality of the pairwise data that were used to construct the network. This shows that the network structure itself captures a significant part of the dynamic behavior of the network. We compare our approach with a method proposed by Yeang *et al.* (Yeang *et al.*, 2004) who used network structure to predict the effect of perturbations where the paths between the perturbed node and the readout are consistent. Our method provides correct predictions in many cases where the previous method could not be applied (e.g. when there are multiple contradicting paths, which in some networks constitute >50% of the cases).

Finally, our analysis demonstrates how it is possible to integrate data from a variety of databases and repositories, such as high-throughput microarray experiments, drug-target information and interaction databases, to create a useful platform to predict and evaluate the effect of perturbations in the cell.

2 METHODS

2.1 Algorithm

The input to our algorithm is a directed weighted network and a pair of nodes (x,y), where x is the target of a perturbation and y is the node to be evaluated. The algorithm predicts whether node y is upregulated, downregulated or unchanged as the result of inhibiting node x. The acyclic paths between node x and node y are computed through recursive exhaustive search.

The effect (E) of inhibiting node x, on the level of node y, is defined as the sum of individual effects of all paths from x to y, multiplied by -1:

$$E(x,y) = -\sum_{i=1}^{n} F(x,y)_i$$

where n is the number of acyclic paths from x to y, and $F(x,y)$ is the function that quantifies the effect of the i^{th} path. In turn, $F(x,y)_i$ is defined as the multiplication of weights along the i^{th} path:

$$F(x,y)_i = \prod_{j=1}^{l} w_j$$

where l is the length of the i^{th} path, and w_j is the weight of the j^{th} edge. The combined effect (CE) of multiple perturbations is defined as the sum of individual perturbations:

$$CE(x_1, x_2 \ldots x_p, y) = \sum_{k=1}^{p} E(x,y)_k$$

where p is the number of perturbations, and $E(x,y)_k$ is the effect of the k^{th} perturbation. As our algorithm aims to predict the direction of change (DOC) of the combined effect, we transform the data as follows:

$$DOC = \begin{cases} up-regulated & (\uparrow)\ if\ \mathrm{CE} > \mathrm{T} \\ down-regulated & (\downarrow)\ if\ \mathrm{CE} < \mathrm{T} \\ unchanged & (-)\ O.W. \end{cases}$$

where T is a predefined threshold. Here, we did not optimize the parameters, and, somewhat arbitrarily, set the threshold T to be zero and the weights (w_j) to be $+0.5$ and -0.5 for activating and inhibiting interactions, respectively. Using $\mathrm{T}=0$ dictates that in most cases, a perturbation would affect all genes. Choosing other values for T will allow one to ignore small changes. Obviously, the choice of optimal thresholds may change according to the system, the network and the desired specificity and sensitivity.

2.2 Control and background models

We created three null models as controls for assessing the significance of the predictions. Two of the null models are based on 1000 simulations in which the 'predictions' were made randomly maintaining either the ratio of upregulation/downregulation as in the experimental data (control 1) or that ratio as in results obtained from the algorithm (control 2). The third control was based on running the algorithms on 1000 randomized networks that preserve the in- and out-degree of each network node (control 3).

2.3 EGFR/MAPK and PI3K/AKT network

To test our algorithm, we used a model created by Nelander *et al.* (Nelander *et al.*, 2008), based on a computational and experimental platform regarding MCF7 human breast carcinoma cells. In this model, the

authors inhibited key components in the network and measured the resulting phenotypic effects. Additionally, they used these measurements to infer regulatory interactions and construct a regulatory network. The cells were treated with six inhibitors that were used to target EGFR (ZD1839), mTOR (rapamycin), MEK (PD0325901), PKC-delta (rottlerin), PI3 kinase (LY294002) and IGF1R (A12). Inhibitors were administered singly and in pairs, followed by stimulation with epidermal growth factor. As relevant readouts of phenotypic responses, phosphor-protein levels of seven regulators (p-AKT, p-ERK, p-MEK, p-eIF4E, p-RAF, p-P70S6K and pS6) were measured, as well as cell-cycle arrest and apoptosis. These readouts are given as quantitative levels of treated versus untreated cells standardized to an interval between -1 and $+1$. Using these readouts, the authors developed a computational strategy to infer the regulatory interactions between the network components. They inferred 23 interactions between 14 nodes, nine of which were designated as 'low significance' interactions. To create a network with a single connected component, we included all 23 interactions (Table 1 and Fig. 1A). Additionally, we binned the experimental readout values to either upregulated (value >0) or downregulated (value <0).

2.4 Using CMAP expression data

Normalized ratios of treated/untreated expression levels from the CMAP project (Golub *et al.*, 2006) were used. The original CMAP data set included 6100 instances, each representing the expression profile of cells treated with a small molecule at a specific concentration. Overall, 1310 small molecules were used at different concentrations and on different cell types. We mapped the small molecules used in CMAP to their known targets through the 'Drug Target' field in the DrugBank database (Knox *et al.*, 2011). In all, 2861 CMAP instances comprising 548 small molecules were found to have a known target (Supplementary Table S1). Overall, 552 gene targets were identified. In this study, we used only instances of MCF7 cells at their highest concentration of treatment.

Each expression profile in CMAP comprises normalized ratios of treated/untreated expression levels for 22 283 probes on Affymetrix gene chips. However, the expression level of a single gene may be represented by multiple, often contradictive, datapoints in CMAP. This can occur if more than one probe on the microarray is mapped to the same gene, or if the same experiment was duplicated at different (or the same) concentrations. Therefore, we defined the following policy to determine whether a gene is over- or underexpressed as a result of treatment: a probe on the microarray is considered regulated, only if the ratio of treated/untreated >1.25 (upregulated) or treated/untreated <0.8 (downregulated). A gene is considered regulated if 70% of its probes change in the same direction. Additionally, at least 50% of the probes must be either up- or downregulated. Using this policy, we tagged each gene as either upregulated, downregulated or as having no change. To further filter the data, a target gene was considered valid only if 50% of its direct neighbors in the regulatory network were regulated in the anticipated direction according to the sign of the edge. As this last filter dramatically reduces the amount of data, we used it only for the large NCI network, and not for the small TNF network.

Table 1. Network statistics

	Nodes	Interactions	Source
EGFR/MAPK	14	23	Nelander *et al.*
TNF	14	29	STRING
NCI	227	466	PID

2.5 TNF network

We identified the 25 proteins from the TGF-β receptor signaling pathway in Pathway Commons (Cerami *et al.*, 2011) that were targets of a drug in the CMAP project. The STRING server (Jensen *et al.*, 2009) was used to determine the regulatory interactions between this subset of proteins. First, we used the STRING web interface to get a full list of links between these components and then used the STRING data file (protein.actions.detailed.v9.0.txt) with information regarding interaction types to identify regulatory interactions. Removing isolated nodes left us with a regulatory network of 14 nodes and 29 edges (Table 1 and Fig. 1C). Merging this data with CMAP revealed a set of 16 perturbations over 13 proteins. Although the remaining (14th) protein was also targeted by a drug in CMAP, it is excluded because it did not cause a change in the expression of any of the other network nodes. Similarly, only 10 of the proteins are analyzed as readouts, as the remaining four did not show any change for any of the perturbations. We note that four interactions appeared twice with opposite signs. As both interactions may be true under different biological circumstances, we left both edges in the network. Many of the interactions in this network involve TNF, therefore we refer to this network as the TNF network.

2.6 NCI pathways network

The integrated XML file of all NCI pathways was downloaded from PID (Schaefer *et al.*, 2009). Generally, interactions in this file are described using the following edge types: input biomolecule (input), negative regulator (inhibitor), positive regulator (agent) and output. The input, agent and inhibitor edges point from a set of molecules to a specific process type, which in turn points with an output edge to the output molecule. To transform this network to a set of binary interactions, we connected each molecule from the source of the input agent and inhibitor edges to the target of the output edge. For interactions where no output edge exists, e.g. when the interaction triggers a biological process, edges were created from the molecules to the process. This procedure resulted in a network of 5916 nodes (many of which are complexes) and 13 050 interactions. For this network, we computed all paths of up to six nodes.

Considering only nodes that were either targets or readouts in CMAP, or part of the paths connecting them, resulted in a network of 227 nodes and 466 interactions (Table 1 and Results). Seven edges that were both activating and inhibiting were treated as described earlier. The network was rendered using Cytoscape with the force directed layout algorithm (Smoot *et al.*, 2011).

2.7 Comparing predictions with experimental results

When comparing our predictions with the experimental results, we only used experimental data points that were either upregulated or downregulated, and not points that had no change. Additionally, we excluded cases in which there was no path between the perturbed node and the readout node. Our assumption being that when experimental results imply a regulatory relationship between two nodes, but no pathway connects them, it proves that the pathway is not fully represented, and therefore our method cannot be tested on such a case.

3 RESULTS

3.1 Predicting the effect of perturbations on a protein signaling pathway

We first predicted the effects of perturbations on the EGFR/MAPK and PI3K/AKT signaling pathway. To this end, we relied on the experiments of Nelander *et al.* (Nelander *et al.*, 2008) who used drugs to inhibit different proteins in these pathways, either one at a time or in pairs, in MCF7 breast cancer cells.

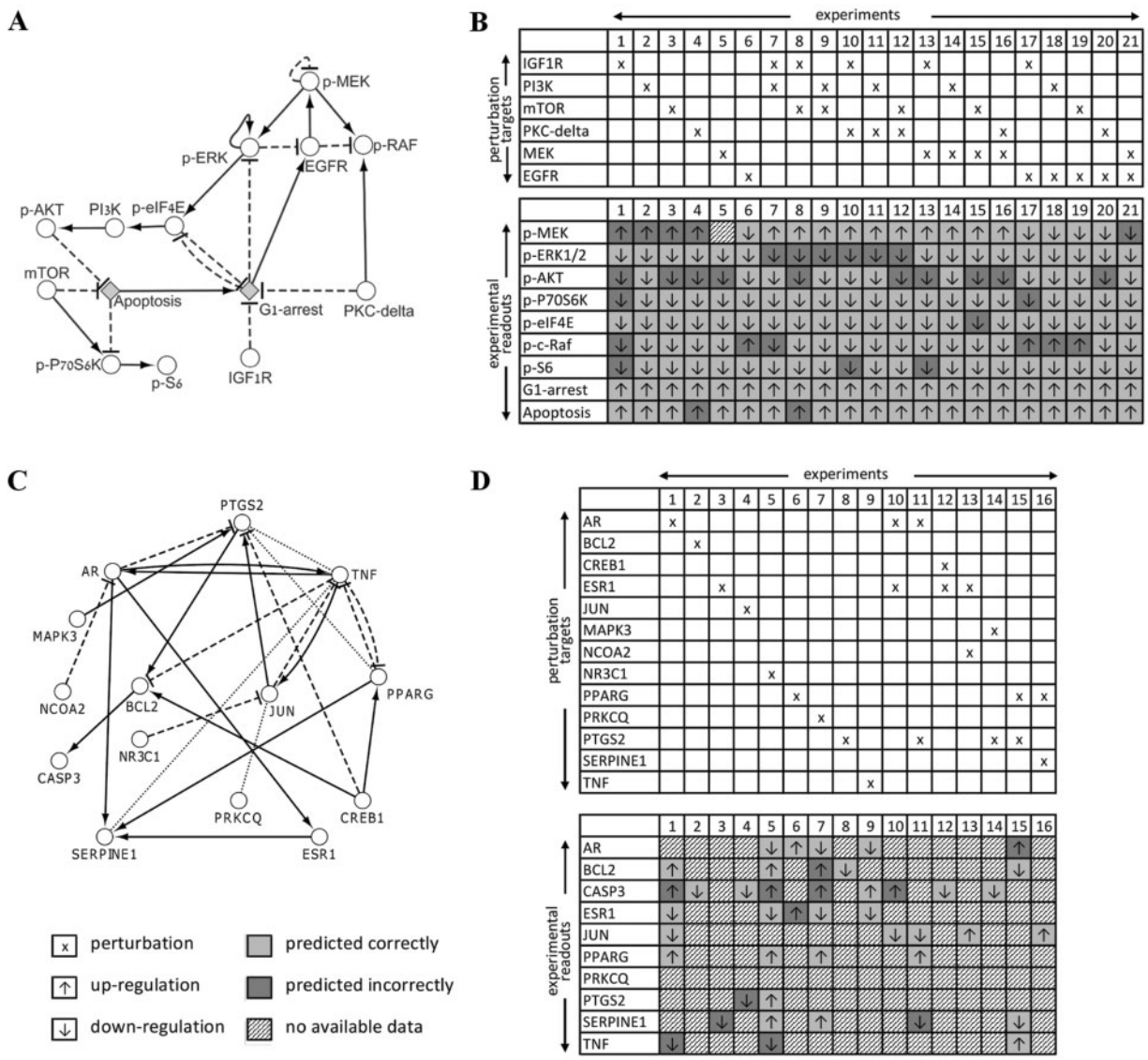

Fig. 1. Predicting perturbation effects on regulatory networks. Panels **A** and **C** depict the EGFR/MAPK and PI3K/AKT signaling network as derived from Nelander *et al.* (A), and the TNF network constructed using STRING (C). Solid arrowed edges denote activation, dashed blocked edges denote inhibition and faint dotted edges denote contradictive information regarding the sign of the interaction. Circles denote genes or proteins, whereas diamonds denote phenotypic processes. The prediction matrix of our method compared with the experimental readouts for the networks are shown in panels **B** and **D**. The experimental readouts for the EGFR/MAPK and PI3K/AKT signaling network were taken from Nelander *et al.* and discretized (B). The experimental readouts for the perturbations on the TNF network were extracted from expression data in the CMAP project (D). The top tables in these panels present the experimental layout, i.e. which proteins/genes were targeted by a drug (or a pair of drugs) in each experiment. The bottom tables present our predictions, where arrows indicate the direction of change (up- or downregulated) for each protein/gene, and the shade indicates whether the prediction agrees with the experimental readout (light grey) or not (dark grey). Results that could not be determined, as either no path exists between the perturbed node and the readout node or the expression data was inconclusive,are marked with diagonal stripes

For each such perturbation or a combination thereof, they measured phenotypic readouts of phosphor-protein levels for seven regulators in these pathways. They also measured the effect of the perturbations on cell-cycle arrest and apoptosis. The authors then developed an algorithm that searched for a set of ODEs that will explain the measurements. Their best model comprised 23 interactions amongst 14 components of the EGFR/MAPK and PI3K/AKT signaling pathways (see Methods). Table 1 summarizes the network characteristics. We used this network to predict the phenotypic effects that were measured in the experiment. The results of our predictions are presented in Table 2 and illustrated in Figure 1. Our method correctly predicted the experimental results (i.e. if a protein is upregulated or downregulated) for 81% (153/188) of the readouts (Table 2 and Fig. 1A and B). To assess the significance of these predictions, we compared them to three background null models (see Methods). Briefly, one null model was based on randomly designating the effect for each node (e.g. upregulation or downregulation), preserving the ratio of

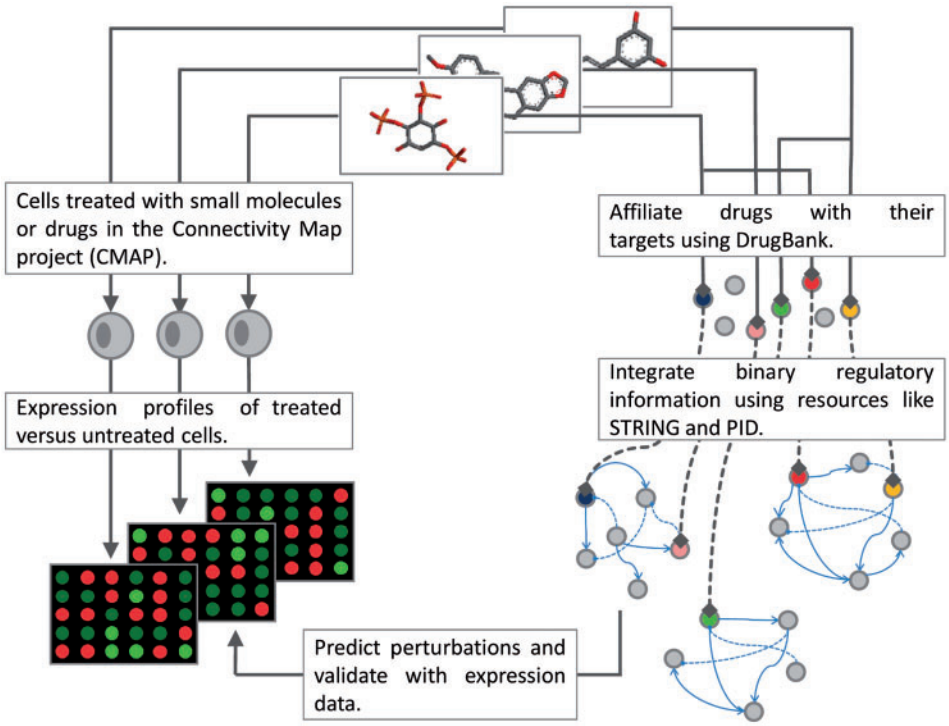

Fig. 2. Integrating data from different bioinformatics resources. In the Connectivity Map (CMAP) Project, cells were treated (perturbed) with hundreds of small molecules, some of which are drugs, and expression profiles were measured. The ratios of treated/untreated profiles are used to assign each protein/gene as upregulated, downregulated or unchanged in each experiment (Left). The small molecules used in CMAP can be mapped to their protein/gene target through DrugBank. In turn, this information can be integrated with regulatory networks through pathway and interaction databases e.g. STRING and PID Right. Finally, predictions can be made regarding the effect of perturbations on the components of the regulatory networks and then validated with the expression data in CMAP

Table 2. Prediction results

	Regulatory networks		Control 1[a]		Control 2[b]		Control 3[c]	
	Precision	%	%	*P*-value	%	*P*-value	%	*P*-value
EGFR/MAPK	153/188	81.4	52.0 (±0.04)	<0.001	56.5 (±0.03)	<0.001	48.7 (±0.08)	<0.001
TNF	31/43	72.1	50.0 (±0.08)	0.014	50.2 (±0.08)	0.014	48.4 (±0.07)	0.002
NCI	143/202	70.8	54.0 (±0.03)	<0.001	53.2 (±0.03)	<0.001	46.0 (±0.09)	<0.001

[a]Randomly designating the direction of change with the ratio of experimental data.[b]Randomly designating the direction of change with the ratio of prediction data.[c]Predictions for randomized networks, preserving in- and out-degree.

upregulation/downregulation as in the experimental data (control 1). The second null model was based on random designation maintaining the ratio of upregulation/downregulation as in the predictions (control 2). The third control executed the prediction algorithm on a set of 1000 randomized networks that preserve the in-degree and out-degree of each component (control 3). As seen in Table 2, the prediction was highly significant with respect to all three background models, ($P < 0.001$ in multiple simulation tests for all three cases).

3.2 Automated merging of regulatory networks and drug perturbation data

To further explore the extent to which the wiring of the static network allows the modeling of biological processes, we attempted to automatically construct regulatory networks, based on data-mining approaches, and to map experimental perturbations to these networks. Figure 2 illustrates this data mining and integration process. Briefly, CMAP (Golub *et al.*, 2006) is a collection of genome-wide transcriptional expression data from cultured human cells treated with bioactive small molecules. Thus, CMAP provides a large set of perturbations and their experimental readouts. DrugBank (Wishart *et al.*, 2006), identifies biologically active molecules and their target proteins. Thus, we mapped the small molecules used in CMAP to hundreds of protein targets. STRING (Jensen *et al.*, 2009) identifies pairwise relationships in high-throughput data. As an example, we examined the TNF signaling pathway taken from the Pathway Commons database (Cerami *et al.*, 2011). We chose this

Fig. 3. NCI network. The combined network composed of all pathways in NCI was intersected with CMAP data, creating a subnetwork of 227 nodes and 466 edges [rendered using Cytoscape's force directed layout (Smoot *et al*., 2011)]. Solid nodes represent genes whose expression was changed in CMAP as a result of a perturbation and for which we provide predictions. This illustration emphasizes the difficulty of making meaningful predictions on such a complex network. Nevertheless, the performance on this large, automatically generated network was high

pathway, as many of its nodes were targets of a drug in CMAP. Using DrugBank, we identified the 25 proteins in this pathway that were targets of drugs in CMAP. We then used STRING to identify regulatory interactions between these targets. Excluding isolated nodes results in a network of 14 proteins/genes and 29 interactions (Table 1 and Fig. 1C). Merging this network with the CMAP data resulted with 16 perturbation combinations (nine single perturbations and seven dual ones, i.e. one drug on two proteins) on 13 of the proteins/genes in the network (see Methods, Fig. 1D).

3.3 Predicting the effect of perturbations on the TNF regulatory network

Using our method, we attempted to reproduce the experimental readouts of CMAP perturbation experiments on the TNF regulatory network, and determine whether each protein/gene was upregulated or downregulated by each perturbation or combinations thereof. Table 2 and Figure 1C and D show the performance of our approach on this network. The static wiring successfully predicted the experimentally measured effect of perturbations for 72.1% (31/43) of the available experimental data. We used the same three background models described earlier as controls. Again, our results were significantly better than all control models (Table 2 lists the *P*-values for each model based on multiple simulation tests).

3.4 Predicting the effect of perturbations on a large network of integrated pathways

Next, we show that our algorithm performs well on a large network as well. We constructed a network based on all NCI-curated pathways available in the PID (Schaefer *et al*., 2009) (see Methods). Intersecting this network with CMAP resulted with a network of 227 nodes and 466 interactions, and 202 experimental readouts on two perturbed nodes (Table 1). Figure 3 illustrates this network and demonstrates the difficulty of making meaningful predictions at such a high level of complexity. Nevertheless, our suggested approach was able to correctly predict whether a gene was up- or downregulated as a result of a drug perturbation in 71% (143/202) of the cases. Again, this result is highly significant compared with the three control models described earlier ($P<0.001$ for all three control models, Table 2).

3.5 Relevant comparisons

A previous method by Yeang *et al*. (Yeang *et al*., 2004) predicts the effect of a perturbation as the common effect of the connecting paths and predicts the effect of a single path as the product of $+1/-1$ effects of its member interactions. Although similar to ours, this method is only applicable if no contradicting paths exist. There are three possible scenarios for the paths between a perturbed node and readout node: (1) a single path, (2) multiple consistent paths and (3) multiple contradicting paths. Table 3 shows the number of occurrences and successful predictions for each of these scenarios in our networks. For example, in the EGFR/MAPK pathway, out of a total of 188 relevant experiments, 108 (57%) have contradicting paths. Therefore, the ability of our method to handle such rampant cases is essential.

Another difference between our method and the one of Yeang *et al*. is the use of $+0.5/-0.5$ weights. Although we did not optimize the algorithm to these (or any other) specific weights, our main point is to demonstrate the effect of using weights that are <1. This causes the overall effect to decay exponentially along the path and thus gives more influence to nearby nodes. This stipulation is based on the biological intuition that distant nodes, which require relaying the signal through more nodes, are more likely to have a smaller effect than closer ones. Considering only the sign of the path (i.e. assigning weights of $+1/-1$ to the edges) results in predictions poorer than the ones we have shown. This is demonstrated for all three networks (Table 3). For single pathways or multiple consistent pathway scenarios, using $+0.5/-0.5$ is the same as using the sign alone. The difference is only in cases where multiple contradicting paths exist and must be ranked. For example, the precision for multiple contradicting paths in the EGFR/MAPK pathway was 73% with $+0.5/-0.5$ weights and dropped to 62% using the sign alone.

Our method could also be assessed in comparison with an intuitive approach that is based only on the direct interactions in a regulatory network. Although such a method would have a high precision, we would pay a steep price in recall. For example, using only the direct interactions in the EGFR/MAPK network would increase our precision from 81.4 to 85.7%, but will enable only 56 predictions, instead of 188.

Table 3. Path scenarios

	Single path		Multiple consistent paths		Contradicting paths		Contradicting paths—Sign only	
	Precision	%	Precision	%	Precision	%	Precision	%
EGFR/MAPK	21/23	91.3	53/57	93.0	79/108	73.1	67/108	62.0
TNF	20/25	80.0	1/1	100.0	10/17	58.8	7/17	41.2
NCI	64/94	68.0	55/71	77.0	23/37	62.0	20/37	54.0

4 DISCUSSION

Static network representations have been criticized as too simplistic, as every network model 'implies a set of dynamical relationships among its components and, therefore, demands to be converted into a set of mathematical equations that describe the temporal and spatial evolution' (Tyson *et al.*, 2001). The need to complement static networks with logical constraints or other mathematical and computational means that may capture the complexity of biological processes is still frequently mentioned as a major challenge in systems biology (Auffray *et al.*, 2010; Fisher and Piterman, 2010). Indeed, integrative models that added ODEs as an additional layer to network models have provided novel insights (Alberghina *et al.*, 2009; Hughey *et al.*, 2010; Shamir and Karlebach, 2008). In this study, we attempted to determine to what extent the static, schematic network structure captures the complexity of dynamic biological processes. Our results show that the schematic structure of regulatory networks captures most of the complexity of the biological process and enables the prediction of most of the effects of perturbations on proteins/genes in the network, without requiring dynamic and kinetic data.

Notably, we applied our suggested simple approach to three networks that represent different biological realities and different kinds of pairwise relationships. In the EGFR/MAPK and PI3K/AKT network, the edges represent protein signaling and depict a cascade of direct interactions between proteins, where activation usually refers to phosphorylation, and inhibition usually refers to dephosphorylation. In contrast, the TNF network is a gene regulatory network that represents the direct or indirect relationship (e.g. gene A upregulates gene B) between the expression of genes. The NCI-based network is essentially an ensemble of pathways representing catalytic and regulatory relationships. Although distinguishing between network types is generally important, in our analysis, the same simple principle of summation of the effect of perturbations was seamlessly applicable for all networks, as we mainly consider the direction and the sign of the interactions. As our algorithm assumes that the effect of multiple paths is additive, microarray data should be adjusted when analyzed in this context. For example, using log transform of microarray expression levels is more appropriate than the original normalized values.

Another important difference between the networks we analyze is the quality of the data and the method of curation. In the EGFR/MAPK and PI3K/AKT, we relied on a carefully constructed network that is based on a wealth of experimental data and on sophisticated ODE models. In a sense, what we attempted to do in this case is to validate the static network wiring suggested by the authors, using a much simpler computational apparatus. We have shown that even when we forgo the ODEs and rely only on the static structure of the network, we manage to correctly predict upregulation and downregulation that is consistent with experimental measurements.

The two other processes we analyzed, namely the TNF and the NCI networks, show that the wiring is surprisingly robust. We constructed the TNF network based on an automatic data-mining approach from electronic resources that are based mostly on high-throughput data. Existing interaction/co-expression data are notoriously noisy, partial and inaccurate. Indeed, the experimental measurements we used (taken from the high-throughput CMAP experiment) suffer from the same problems. The fact that we were still able to make correct predictions based on these networks and these experimental perturbations suggests that even when the pairwise relationships that underlie the network are dubious and partial, the wiring can still provide meaningful predictions of the phenotype. The NCI example shows that the method we suggest is both robust and scalable. This is a large network with thousands of paths. Such networks are prone to amplify even small errors and lead to meaningless results. Moreover, networks of this magnitude are much harder to handle computationally. However, the analysis of the network using our algorithm took seconds and the performance remained high.

Our finding regarding the power of static wiring to account for network-wide phenomena is consistent with previous studies that have shown that the intramodular wiring is highly conserved evolutionarily across different species (Ideker *et al.*, 2008; Zinman *et al.*, 2011). One may hypothesize that the kinetic details may be less conserved, yet the wiring itself is maintained.

There is no doubt that when quantitative approaches such as ODEs or logic-based models are applicable, they can provide more accurate predictions of the effect of perturbations. Sophisticated regulatory circuits, which depend on a delicate balance of concentrations and precise timing, play an important role in regulation (Kholodenko, 2006). For example, the EGFR/MAPK signaling cascade is known to be controlled by a negative feedback loop (Keyse, 2000). Although this negative regulation is distinctly represented in the network we used, as seen in Figure 1A (inhibition of ERK on EGFR and the self-inhibition of MEK), our method failed to effectively model it. This may explain why our predictions of the readouts for phosphorylated MEK were somewhat less precise than the predictions for the other proteins. However, in most cases quantitative approaches

are not applicable owing to lack of logical and kinetic information. Moreover, these modeling approaches typically require heavy computations and are thus not trivially scalable for larger networks.

Existing databases have tremendous amounts of data that may enable the construction of reliable networks. But to construct these networks, one needs to data mine and integrate data from many sources. Clever methods for integrating pairwise information can lead to more accurate and comprehensive networks. Once such networks are obtained, our results suggest they will enable some predictions without requiring kinetic and logical information. Moreover, we demonstrate that even automatically constructed networks may provide valuable predictions of phenotype.

This study indicates that despite the importance of elaborate control mechanisms, a large portion of the control over the system lies solely in its schematic wiring. This important characteristic enables the use of basic modeling methods like ours to get an indication of a systems response to a set of perturbations for the wide range of interaction data usually available only at a low level of description.

ACKNOWLEDGEMENTS

We thank Guy Nimrod for useful discussion and Inbal Sela for critical reading of the manuscript. We thank those who deposit their experimental results in publically available databases and to those who maintain these databases.

Funding: This work is supported by Microsoft Research.

Conflict of Interest: none declared.

REFERENCES

Alberghina,L. *et al.* (2009) Systems biology of the cell cycle of Saccharomyces cerevisiae: from network mining to system-level properties. *Biotechnol. Adv.*, **27**, 960–978.

Arisi,I. *et al.* (2006) Parameter estimate of signal transduction pathways. *BMC Neurosci.*, **7** (Suppl 1), S6.

Auffray,C. *et al.* (2010) The hallmarks of cancer revisited through systems biology and network modelling. In *Cancer Syst. Biol. Bioinform. Med.,* Springer Netherlands, pp. 245–266.

Bailey,J.E. (2001) Complex biology with no parameters. *Nature Biotechnol.*, **19**, 503–504.

Cerami,E.G. *et al.* (2011) Pathway commons, a web resource for biological pathway data. *Nucleic Acids Res.*, **39**, D685–D690.

Chaouiya,C. (2007) Petri net modelling of biological networks. *Brief. Bioinform.*, **8**, 210–219.

Chindelevitch,L. *et al.* (2012) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*, **28**, 1114–1121.

Enayetallah,A.E. *et al.* (2011) Modeling the mechanism of action of a DGAT1 inhibitor using a causal reasoning platform. *PLoS One*, **6**, e27009.

Fisher,J. and Piterman,N. (2010) The executable pathway to biological networks. *Brief Funct Genomics*, **9**, 79–92.

Glass,L. and Kauffman,S.A. (1973) The logical analysis of continuous, non-linear biochemical control networks. *J. Theor. Biol.*, **39**, 103–129.

Golub,T.R. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Hermjakob,H. *et al.* (2010) The intact molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

Hughey,J.J. *et al.* (2010) Computational modeling of mammalian signaling networks. *Wiley Interdiscip. Rev. Syst. Biol. Med.*, **2**, 194–209.

Ideker,T. *et al.* (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18** (**Suppl. 1**), S233–240.

Ideker,T. *et al.* (2008) Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science*, **322**, 405–410.

Jensen,L.J. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Keyse,S.M. (2000) Protein phosphatases and the regulation of mitogen-activated protein kinase signalling. *Curr. Opin. Cell. Biol.*, **12**, 186–192.

Kholodenko,B.N. (2006) Cell-signalling dynamics in time and space. *Nat. Rev. Mol. Cell. Biol.*, **7**, 165–176.

Knox,C. *et al.* (2011) DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs. *Nucleic Acids Res.*, **39**, D1035–D1041.

Li,F. *et al.* (2010) PerturbationAnalyzer: a tool for investigating the effects of concentration perturbation on protein interaction networks. *Bioinformatics*, **26**, 275–277.

Maslov,S. and Ispolatov,I. (2007) Propagation of large concentration changes in reversible protein-binding networks. *Proc. Natl. Acad. Sci. U S A.*, **104**, 13655–13660.

Mitsos,A. *et al.* (2009) Identifying drug effects via pathway alterations using an integer linear programming optimization formulation on phosphoproteomic data. *PLoS Comput. Biol.*, **5**, e1000591.

Morris,M.K. *et al.* (2010) Logic-based models for the analysis of cell signaling networks. *Biochemistry*, **49**, 3216–3224.

Nelander,S. *et al.* (2008) Models from experiments: combinatorial drug perturbations of cancer cells. *Mol. Syst. Biol.*, **4**, 216.

Papin,J.A. *et al.* (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell Biol.*, **6**, 99–111.

Prill,R.J. *et al.* (2010) Towards a rigorous assessment of systems biology models: the DREAM3 challenges. *PLoS One*, **5**, e9202.

Przytycka,T.M. *et al.* (2010) Toward the dynamic interactome: it's about time. *Brief. Bioinform.*, **11**, 15–29.

Rajagopalan,D. and Agarwal,P. (2005) Inferring pathways from gene lists using a literature-derived network of biological relationships. *Bioinformatics*, **21**, 788–793.

Reddy,V.N. *et al.* (1996) Qualitative analysis of biochemical reaction systems. *Comput. Biol. Med.*, **26**, 9–24.

Ruths,D. *et al.* (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.

Schaefer,C.F. *et al.* (2009) PID: the pathway interaction database. *Nucleic Acids Res.*, **37**, D674–D679.

Shamir,R. and Karlebach,G. (2008) Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.*, **9**, 770–780.

Smoot,M.E. *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics*, **27**, 431–432.

Sorger,P.K. *et al.* (2011) Comparing signaling networks between normal and transformed hepatocytes using discrete logical models. *Cancer Res.*, **71**, 5400–5411.

Thomas,R. (1973) Boolean formalization of genetic-control circuits. *J. Theor. Biol.*, **42**, 563–585.

Tyers,M. *et al.* (2011) The BioGRID interaction database: 2011 update. *Nucleic Acids Res.*, **39**, D698–D704.

Tyson,J.J. *et al.* (2001) Network dynamics and cell physiology. *Nat. Rev. Mol. Cell Biol.*, **2**, 908–916.

Wang,D.Y. *et al.* (2009) Computational modeling of the EGFR network elucidates control mechanisms regulating signal dynamics. *BMC Syst. Biol.*, **3**, 118.

Wishart,D.S. *et al.* (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.*, **34**, D668–D672.

Yeang,C.H. *et al.* (2004) Physical network models. *J. Comput. Biol.*, **11**, 243–262.

Zinman,G.E. *et al.* (2011) Biological interaction networks are conserved at the module level, *BMC Syst. Biol.*, **5**, 134.