Accounting for robustness in modeling signal transduction responses

Mohammad Omid Oftadeh and Sayed-Amir Marashi
department of Biotechnology, College of Science, University of Tehran, Tehran, Iran

ABSTRACT
A classical question in systems biology is to find a Boolean model which is able to predict the observed responses of a signaling network. It has been previously shown that such models can be tailored based on experimental data. While fitting a minimum-size network to the experimentally observed data is a natural assumption, it can potentially result in a network which is not so robust against the noises in the training dataset. Indeed, it is widely accepted now that biological systems are generally evolved to be very robust. Therefore, in the present work, we extended the classical formulation of Boolean network construction in order to put weight on the robustness of the created network. We show that our method results generally in more relevant networks. Consequently, considering robustness as a design principle of biological networks can result in more realistic models.

1. Introduction

Cellular response to external signals is often mediated through a complex protein signaling network. Such an intricate network determines the behavior of a cell type when a certain set of signals are encountered. A detailed understanding of cell behavior has a central role in cell biology and bioengineering. Although the outcome of applying the external stimuli can be determined by conducting experiments on a certain cell type, the number of experiments required to screen various combinations of the input signals tremendously rises when the number of the stimulators increases. In practice, it is necessary to construct a model to enable us predicting the cell responses by using a limited set of experimental data. Furthermore, such a model is inherently an enabling framework to elucidate the underlying mechanism of signal transduction.

The protein-protein interaction graphs are created either directly from mass spectrometry and other high-throughput methods or by inferring from the literature [1,2]. This interaction graph can be considered as a pre-knowledge network (PKN), which supposedly includes all of the signal transduction interactions. In other words, a signal transduction network model should be found as a ‘subnetwork’ of PKN.

It should be noted that interaction graphs created by both experiment- and literature-based methods suffer from shortcomings that prevent them from predicting the cellular response to a predefined stimuli properly [3]. Protein interaction networks which are obtained by experimental data often neglect the intrinsic and mechanistic properties of interactions. Literature-based networks, on the other hand, are reconstructed by collecting information from unrelated experiments performed under potentially uncomparable conditions. Consequently, such networks may not be relevant for modeling realistic conditions.

In practice, a subnetwork of PKN should be selected by maximizing the agreement between the predictions of the model and the experimental results. To this end, a simple and yet efficient approach is application of Boolean network models. In a Boolean network, it is assumed that each interaction is either present or absent in the final network. Similarly, each protein is either activate or inactivate. Logical operators (AND, OR, and NOT) are utilized to determine the relationship between different interactions.

Optimizing Boolean models against experimental data was originally proposed by Saez-Rodriguez et al. [3] using a genetic algorithm. Subsequently, Mitsos et al. utilized mixed integer linear programing (MILP) instead of genetic algorithm to establish their model [4]. The objective function of these methods is constituted of two terms: the first term takes the difference between experimental data and model predictions into account while the second term penalizes the size of the network. In this vein, the algorithm searches for a subnetwork of PKN with the best predictive ability while eliminating those interactions with subtle impact on the predictive power.

It has been suggested the biological networks, including signaling transduction, have evolved in a way that allows them to perform their task robustly [5,6]. Robustness in signaling networks can be defined as the reliable transmission of information through the network in the presence of internal perturbation [7]. In other words, the presence of alternative pathways between specified input and output can maintain information transmission when one of the pathways is inhibited [8]. In this work, we present an updated MILP formulation for pruning PKN while optimizing
the signaling robustness. We show that considering the defined robustness as a principle design to construct the signaling model has improved the predictive ability.

2. Materials and methods

2.1. Boolean model tailoring based on the experimental data

The experimental data and the PKN for seven different growth and inflammatory ligands and their underlying signaling networks in HepG2 cells (ExtLiverBMC) were obtained from CellNOpt webpage ([9], http://www.cellnopt.org). Briefly, the experiments include treatments of HepG2 cells utilizing various combinations of seven cytokines (IFNγ, IL6, LPS, TGFα, TNFα, IGF1, and IL1α) and seven inhibitors (i.e. inhibitors of GSK, IKK, mTOR, MEK12, p38, JNK12, and PI3K). For each treatment, the quantity of measured species had been determined at two different time points, i.e. 0 and 30min. Then, the previously proposed method [3] was used to normalize the measurements for each species to [0,1].

2.2. Original MILP formulation

MILP formulation has been extensively exploited to construct optimal data-driven models in systems biology [10–12]. As mentioned above, this framework has previously been used for constructing optimal Boolean models based on a set of experimental data [4].

The constraints of the MILP formulation can be divided into two main groups. The first group consists of the previously proposed constraints utilized to construct the Boolean model of the signaling network [4]. For the detailed description of the first group of constraints see [4] and Text S1. On the other hand, we introduce further constraints in this work to take into account the optimality of the network robustness.

Let \( i = 1, \ldots, n_i, \ j = 1, \ldots, n_s, \ k = 1, \ldots, n_k \) represent the reactions, the molecular species, and the experiments, respectively. The following constraints were proposed to guarantee the consistency between the components of interactions:

\[
\begin{align*}
\text{(1)} & \quad z_{i,k} \leq y_i, \quad \forall i, \forall k \\
\text{(2)} & \quad z_{i,k} \leq x_{i,j,k}, \quad \forall i, \forall k, j \in R_i \\
\text{(3)} & \quad z_{i,k} \geq 1 - x_{i,j,k}, \quad \forall i, \forall k, j \in I_i \\
\text{(4)} & \quad z_{i,k} \geq x_{i,j,k}, \quad \forall i, \forall k, j \in P_i \\
\text{(5)} & \quad z_{i,k} \geq y_i + \sum_{j \in R_i} (1 - x_{i,j,k}) - \sum_{j \in I_i} x_{i,j,k}, \quad \forall i, \forall k \\
\text{(6)} & \quad \sum_{i,j \in P_i} x_{i,j,k} \geq x_{i,j,k}, \quad \forall j, \forall k \\
\text{(7)} & \quad x_{i,j,k} = 0, \quad \forall k, j \in M_{i,0} \\
\text{(8)} & \quad x_{i,j,k} = 1, \quad \forall k, j \in M_{i,1} \\
\text{(9)} & \quad a_{i,j} \leq b_i, \quad i = 1, \ldots, n_i \\
\text{(10)} & \quad x_{i,j,k} \in \{0,1\}, \quad y_i \in \{0,1\}, \quad z_{i,k} \in \{0,1\}, \quad \forall i, \forall j, \forall k 
\end{align*}
\]

Constraint (1) implies that only allowed reactions \( y_i = 1 \) can occur during each experiment. Constraints (2) and (3) guarantee that a reaction does not take place when some of its reagents are not present or one of its inhibitors are provided, respectively. Constraint (4) enforces the presence of a species when at least one of its producing reactions occurs. Constraint (5) verifies the occurrence of a reaction if all of its reagents and none of its inhibitors are present. Constraint (6) implies that a species is not formed when none of its producing reactions are active. Constraint (7) (resp. Constraint (8)) dictates the presence (resp. absence) of the input species when they are provided (resp. not provided) as the input. Constraint (9), which is called ‘the combinatorial constraint’, controls the relation among a set of logical operators which link two or more species.

Constraint (10) defines all variables as Boolean. Mitsos et al. have proved the above ILP can be reformulated as an equivalent MILP by relaxing Constraint (10), as follows [4]:

\[
\begin{align*}
\text{(11)} & \quad x_{i,j,k} \in \{0,1\}, \quad 0 \leq x_{i,j,k} \leq 1, \\
\text{(12)} & \quad y_i \in \{0,1\}, \quad 0 \leq y_i \leq 1, \\
\text{(13)} & \quad \forall i, j', \forall j \notin IN \iff \forall i, \forall k
\end{align*}
\]

where \( IN \) is the set of the input species.

In the original MILP formulation, the following objective function was used to determine a minimal-sized Boolean network whose predictions agree best with the experimental results:

\[
\min \left( \sum_{j \in M_{i,2}} \sum_{k=1}^{n_k} (x_{i,j,k}^n + (1-2x_{i,j,k}^n)x_{i,j,k}) + \beta \sum_{j=1}^{n} r_j y_j \right)
\]

where variables \( x_{i,j,k}^n \) represent measured output species (the species that are not the reagent of any reaction in the network) normalized to the \([0,1]\) domain (for more details see Text S1).

2.3. Reformulating the MILP by considering network robustness

As discussed in the Introduction section, the robustness in a signaling network can be interpreted as the existence of alternative pathways between the input and output species (nodes). Such alternative pathways can ensure information transmissibility when the network is perturbed. This can be conceptualized as the resistance of the system to maintain the original relationship between inputs and outputs in various experiments.

For developing a formulation which considers the robustness of the network, we added a set of constraints to check the existence of pathways between two specified nodes. In addition to the set of input nodes, \( IN \), we define another set of nodes, \( OUT \), which represent the output nodes. For \( s \in IN \), Boolean variables \( u_{i,s} \) and \( v_{j,s} \) are defined as follows:

\[
\begin{align*}
\text{(14)} & \quad u_{i,s} = \begin{cases} 1 & \text{if } (j = s) \text{ or } (j \text{ is connected to } s) \text{ in experiment } k \\
0 & \text{otherwise} \end{cases} \\
\text{(15)} & \quad v_{j,s} = \begin{cases} 1 & \text{if } (j = s) \text{ or } (j \text{ is connected to } s) \text{ in experiment } k \\
0 & \text{otherwise} \end{cases}
\end{align*}
\]
The function is modified in the following way:

\[
u_{i,k}^k = \begin{cases} 
1 & \text{if } \left( v_{j,k}^j = 1 \text{ for at least one of } j \in R_i \right) \\
0 & \text{otherwise}
\end{cases}
\]

and (reaction $i$ is active) in experiment $k$

In other words, Boolean variable $v_{j,k}^j$ can be interpreted as an indicator attributed to node $s \in \text{IN}$ which can be transmitted to node $j$ iff a pathway exists from $s$ to $j$ in experiment $k$. Clearly, a pathway from node $s \in \text{IN}$ to node $t \in \text{OUT}$ exists in experiment condition $k$ iff $v_{s,k}^s = 1$.

Before describing the constraints, the above objective function is modified in the following way:

\[
\min \left( \sum_{j \in \text{MB,t}} \sum_{k=1}^{n_r} \left( x_{j,k}^m + (1-2x_{j,k}^m)x_{j,k} \right) \right) \\
+ \beta \sum_{i=1}^{n_r} r_{j} y_{i} + \alpha \sum_{s \in \text{IN}, t \in \text{OUT}} \left( \left| \sum_{k=1}^{n_r} v_{i,t}^k - n_{s,t} / 2 \right| \right)
\]

where $n_{s,t}$ is the number of experiments in which the input species $s$ is provided. The added summation will prioritize solutions in which the relation between each pair of input-output nodes remains preferably unchanged during the experiments (whether most of the time they are connected or most of the time they are not connected). Coefficients $\alpha$, $\beta$ control the tradeoff between the optimization of the robustness, the network size, and the predictive ability. It can be demonstrated that the absolute values can be converted to linear terms by defining variables $w_{s,t}$ and binary variables $b_{s,t}$ and adding these constraints to the formulation:

\[
\sum_{k=1}^{n_r} v_{i,t}^k - n_{s,t} / 2 \leq w_{s,t}, \quad s \in \text{IN}, t \in \text{OUT}, \forall k
\]

(11)

\[
- \sum_{k=1}^{n_r} v_{i,t}^k + n_{s,t} / 2 \leq w_{s,t}, \quad s \in \text{IN}, t \in \text{OUT}, \forall k
\]

(12)

\[
\sum_{k=1}^{n_r} v_{i,t}^k - n_{s,t} / 2 + M b_{s,t} \geq w_{s,t}, \quad s \in \text{IN}, t \in \text{OUT}, \forall k
\]

(13)

\[
- \sum_{k=1}^{n_r} v_{i,t}^k + n_{s,t} / 2 + M(1-b_{s,t}) \geq w_{s,t}, \quad s \in \text{IN}, t \in \text{OUT}, \forall k
\]

(14)

where $M$ is a sufficiently large number. Then, the objective function is modified as follows:

\[
\min \left( \sum_{j \in \text{MB,t}} \sum_{k=1}^{n_r} \left( x_{j,k}^m + (1-2x_{j,k}^m)x_{j,k} \right) \right) \\
+ \beta \sum_{i=1}^{n_r} r_{j} y_{i} + \alpha \sum_{s \in \text{IN}, t \in \text{OUT}} \left( \left| \sum_{k=1}^{n_r} v_{i,t}^k - n_{s,t} / 2 \right| \right)
\]

In addition to the previously introduced constraints, the following sets of constraints are also considered:

\[
u_{s,t}^k \geq z_{i,k} + v_{j,s}^j - 1, \quad \forall i, \forall k, s \in \text{IN}, j \in R_i
\]

(15)

\[
u_{s,t}^k \leq z_{i,k}, \quad \forall i, \forall k, s \in \text{IN}
\]

(16)

\[
u_{s,t}^k \leq \sum_{j \in R_i} v_{j,s}^j, \quad \forall i, \forall k, s \in \text{IN}
\]

(17)

\[
\sum_{\forall i,j \in P_k} u_{i,j} \geq v_{j,s}^j, \quad \forall j, \forall k, s \in \text{IN}
\]

(18)

where the binary variables $z_{i,k}$ represent whether the $i$th reaction is active or not. In the above formulation, Constraint (15) implies that when the $i$th reaction occurs in the $k$th experiment and the attributed indicator to at least one of its predecessor nodes is 1, its corresponding binary variable $u_{i,t}^k$ is forced to be equal to 1. On the other hand, Constraints (16) and (17) ensure, respectively, that when the $i$th reaction does not occur in the $k$th experiment or no indicator is attributed to any of its predecessor nodes, the corresponding variable $u_{s,t}^k$ is forced to be zero. Constraint (18) enforces that no indicator is attributed to a node if the corresponding variables $u_{i,t}^k$ are zero for all of its producing reactions, whereas Constraint (19) implies that if $u_{i,t}^k$ is equal to 1, indicator will be attributed to all of its predecessor nodes (products). Constraint (18) ensures that when an input node is specified, its indicator is set to 1 and Constraint (21) enforces that no indicator is attributed to the other input nodes. Finally, Constraint (22) states $u_{i,t}^k$, $v_{j,s}^j$, $b_{s,t}$ are binary variables.

### 2.4. Validation

To assess the predictive abilities of models, an independent set of experiments [13] was used as the validation dataset. Relevance of using these data as the validation dataset has been previously demonstrated [3]. The experiments include various combinations of four ligands (TGF$\alpha$, IL6, IGF1, and IL1$\alpha$) and four inhibitors (i.e. the inhibitors of IKK, MEK12, p38, and PI3K). The measurements had been performed at two time points, namely 0 and 30 min. Similar to the Section 2.1., the data were normalized to [0,1]. Model predictions were compared to the experimental data. False positive rates (FPR = 1-specificity) and true positive rates (TPR = sensitivity) were determined accordingly. More precisely, a small difference (less than 0.5) between model prediction and experimental data was assumed to represent an accurate prediction. Otherwise, the prediction was considered as incorrect.

### 3. Results and discussion

#### 3.1. Toy model

As an illustrative example, we present a small artificial toy model to compare performance of the two algorithms. The PKN for the toy network is shown in Figure 1(a) which constitutes eleven possible reactions among 10 species. The setups and results of the eight hypothetical experiments, used to develop the Boolean model, are provided in Table 1. Numbers designated by * and # represent the inconsistencies between experimental data and predictions of the models, constructed either by the previous method (minimizing
the size of the network, S-model) or by our method (maximizing the robustness of the network as well, SR-model), respectively. The diagrams of both of the constructed models are depicted in Figure 1(b,c).

In our toy model, experimental results and predictions of either of the models are inconsistent in three predictions each (see Table 1). Obviously, this implies that the first term in the objective function is of the same weight for both models. However, solving the optimization problem utilizing the Saez-Rodriguez et al. algorithm (S-model) yields the diagram in Figure 1(b), as this method prefers smaller sizes. On the other hand, the diagram in Figure 1(c) is more robust, since the relationships between input-output nodes are more resistant to perturbations. More specifically, consider the relationships between N2 (input) and N9 or N10 (outputs). In case of N2 and N9, they are disconnected in five experiments and connected in the other three experiments in S-model. This means that, in the best case, the relationship between N2 and N9 remains unchanged in only five of the experiments. On the other hand, in SR-model, an alternative pathway is created by selecting further interactions from the PKN. Existence of alternative pathways makes the relationship between the nodes more resistant to changes, i.e. the nodes are connected in six experiments and disconnected in the other two experiments.

Now, consider the relationship between N2 and N10 into account. In the S-model, these nodes are disconnected in five experiments and connected in other 3 experiments. In the SR-model, the interaction between N5 and N10 is eliminated, which results in a significant enhancement of network robustness, at the cost of only one more inconsistent prediction. This can also be interpreted in another way: our method determines that the interaction between N5 and N10 cannot be verified (at least) using this set of experiments and hence, neglects it.

3.2. Model validation

The predictive ability of S-model and SR-model was compared by plotting TPR vs. FPR (Figure 2) for different values of $\alpha$ and $\beta$ coefficients. Since $\alpha$ and $\beta$ can vary independently, various combinations of them were tested. Based on the results, the best performance was obtained when $\alpha$ is equal to $\beta$. Similar to the previous observations [3], FPR was decreased by increasing the value of $\beta$, i.e. decreasing the size of the network, at the cost of less sensitivity.

The effect of $\alpha$ alteration might be more complicated as increasing the robustness might result in both adding several new interactions and eliminating several others from the model. However, in the present model, increase in the robustness more often caused adding new interactions. In this vein, increase in the robustness hinders the decrement of TPR caused by decreasing the size and consequently yields solutions with better predictive ability.
Figure 3. F-score of SR-model predictions versus S-model using (a) 10%, (b) 20%, (c) 30%, (d) 40%, and (e) 50% noise. Matthews correlation coefficient (MCC) of SR-model predictions versus S-model using (f) 10%, (g) 20%, (h) 30%, (i) 40%, and (j) 50% noise.
3.3. Evaluating the effect of noise

An important concern in using a discrete abstraction to model a continuous system is that the presence of noise, even in small amplitude, might disrupt the model due to rounding to an erroneous level. To evaluate the effect of noise on the model predictions, in each case, we randomly removed a certain ratio of the experimental data and replaced them with random numbers chosen from domain [0, 1]. To make the experimental data more vulnerable to noises, half of the experiments were randomly eliminated. Subsequently, 10%, 20%, 30%, 40%, and 50% of experimental data were replaced by noised data. Each simulation was repeated 10 times. Using the validation dataset, the specificity and sensitivity for each model were calculated (Figures S1 and S2). Based on the results, accounting for robustness in the objective function improved considerably sensitivity of the model at the cost of marginal decrement of specificity. To assess and compare the quality of each model predictions, F-score and Matthews correlation coefficient (MCC) were calculated and plotted (Figure 3). While no significant difference between model predictions in the presence of 10% noise was observed, optimizing robustness of the network improved the quality of predictions remarkably when 20% and higher noises were added to the dataset. This indicates that predictive ability of SR-models is generally more resistant to the noise effect than S-models.

4. Conclusion

Robustness has been previously proposed as an important evolutionary design principle in biological networks modeling [14,15]. However, robustness can be defined in various biological networks distinctively, due to their different structure and task [5]. In this work, we formulated robustness in signaling networks by finding and selecting alternative pathways between connected input/output nodes from PKN. We indicated optimizing robustness alongside the size of the network can improve the predictive ability of the model. Besides, further analysis demonstrated that our formulation is more immune to the presence of noise in experimental data.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Sayed-Amir Marashi  http://orcid.org/0000-0001-9801-7449

References