OXFORD

An optimization framework for network annotation

Sushant Patkar[1] and Roded Sharan[2,*]

[1]Computer Science, University of Maryland, College Park, MD 20740, USA and [2]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv 69978, Israel

*To whom correspondence should be addressed.

Abstract

Motivation: A chief goal of systems biology is the reconstruction of large-scale executable models of cellular processes of interest. While accurate continuous models are still beyond reach, a powerful alternative is to learn a logical model of the processes under study, which predicts the logical state of any node of the model as a Boolean function of its incoming nodes. Key to learning such models is the functional annotation of the underlying physical interactions with activation/repression (sign) effects. Such annotations are pretty common for a few well-studied biological pathways.

Results: Here we present a novel optimization framework for large-scale sign annotation that employs different plausible models of signaling and combines them in a rigorous manner. We apply our framework to two large-scale knockout datasets in yeast and evaluate its different components as well as the combined model to predict signs of different subsets of physical interactions. Overall, we obtain an accurate predictor that outperforms previous work by a considerable margin.

Availability and implementation: The code is publicly available at https://github.com/spatkar94/NetworkAnnotation.git.

Contact: roded@post.tau.ac.il

1 Introduction

With increased mapping of physical interactions in living cells (Huttlin *et al.*, 2017), we now have a blueprint of the inner workings of the cell. However, the functional interpretation of this map to simulate the behavior of the cell under different genetic and environmental cues is still beyond reach. One fundamental piece of information that is often missing is the annotation of the network interactions with direction of signal flow and functional activation/repression (sign) effects. The interpretation of the latter effects depends on the type of the physical interaction being considered. For protein–DNA interactions (PDIs), a +/– sign describes a regulatory effect; for protein–protein interactions (PPIs), it represents a functional activation/repression effect. Currently, direction and sign information is available to only a few well-studied pathways (see Fig. 1 for an example), although a large fraction (40–70%) of the PPIs are expected to admit such an annotation (Silberberg *et al.*, 2014). The inference of such annotation information is a precondition to any logical model of a system under study [see, e.g. (Morris *et al.*, 2010)].

A pioneering work by Yeang *et al.* (2004) for network annotation suggested a simple yet effective logical model for signaling whereby physical interactions are directed and signed, and a signal can flow along a directed path with its effect being the aggregate effect of its member interactions, i.e. the product of their signs. To tackle the annotation challenge, Yeang *et al.* suggested a machine learning framework, but their method was limited to physical networks of small scale where it is possible to enumerate all paths. Subsequent work in this area adopted the signaling model suggested by Yeang *et al.* but employed (to the most part) combinatorial methods to learn the hidden annotations. In the most common scenario, one is given a partially annotated physical interaction network and a list of pairs of genes obtained from knockout experiments in which a knockout gene (cause, or source) affected the expression of some other gene (effect, or target). The goal is to annotate the remaining interactions in the physical network with directions and signs such that a maximum number of knockout pairs can be explained by the model.

The problem of inferring interaction directions so that a maximum number of pairs admit a directed path from the cause to the effect was shown to be non-deterministic polynomial time (NP)-hard and a sublogarithmic approximation algorithm was given for it (Blokh *et al.*, 2013). Using SAT solvers and integer linear programming (ILP) techniques, optimal algorithms were given for various variants of the problem (Gitter *et al.*, 2011; Silverbush *et al.*, 2011; Silverbush and Sharan, 2014), potentially restricting the length of the path connecting each cause-effect pair. In addition, a machine learning based inference method was suggested by Stelzl *et al.* (2005).

i502
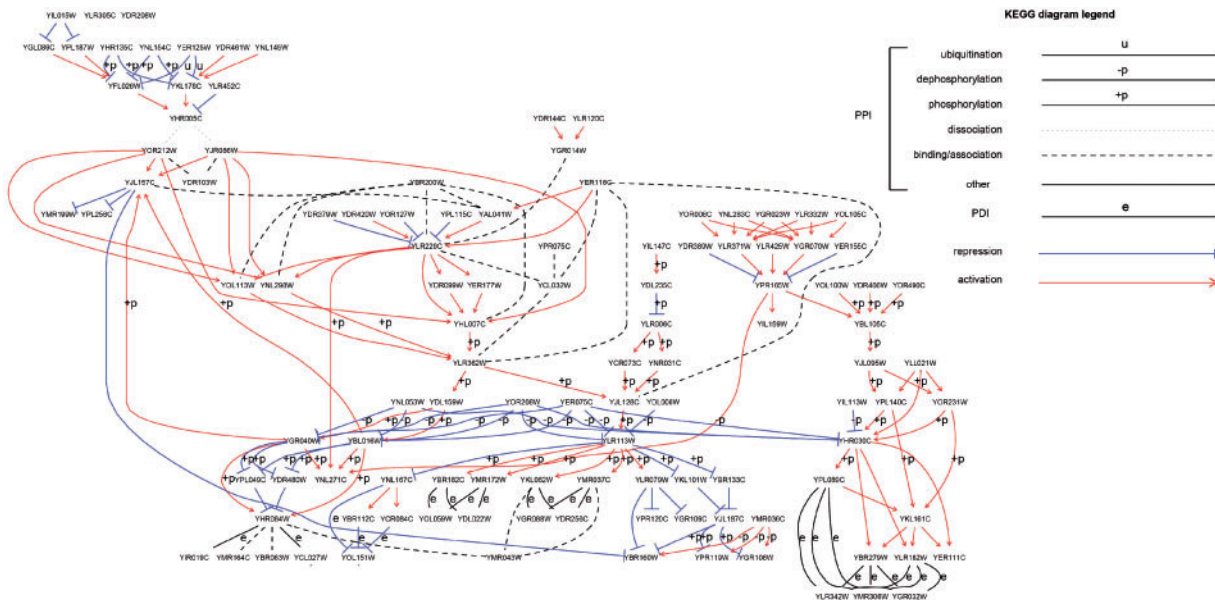
Fig. 1. The yeast signaling pathways from KEGG in one network depicting the organization of different types of physical interactions with their respective experimentally derived signs (activation/repression) and directions

In contrast, the problem of inferring interaction signs received far less attention. Ourfali *et al.* (2007) considered explanatory paths of very short length (3) and provided ILP formulations to maximize the expected number of pairs that can be explained in a probabilistic network. Peleg *et al.* (2010) showed that the sign assignment problem is NP-hard and developed network-free algorithms for predicting genome-wide effects of gene knockouts. A related approach using regression was adopted by Cosgrove *et al.* (2008) to distinguish direct and indirect targets of cell perturbation. Houri and Sharan (2012) were the first to tackle the problem of inferring physical interaction signs on a network while accounting for paths of any length. Specifically, they searched for an assignment that maximized the number of pairs that admit a path of the required sign. They provided network reduction techniques and an ILP formulation to solve this problem to optimality on current physical interaction networks. However, their algorithm could only account for a small fraction of physical interactions in the network (low coverage), as most were contracted in their network reduction step.

In this paper, we present novel network based ILP formulations for the purpose of predicting interaction signs in a physical network. The models we propose bypass the issue of network reduction and thus significantly improve the scale of predictions that can be made. In particular, we consider signaling models where a pair is explained by (i) a shortest path connecting the pair having a desired sign (ASP), (ii) a directed shortest path connecting its nodes having a pre-defined sign (AdirSP) and (iii) all shortest paths connecting its nodes having a desired sign (AllSP). We then evaluate the performance of each model in predicting physical interaction signs in yeast over two different gene expression datasets. We show that these models lead to ≈15-fold higher coverage and higher accuracy than the state-of-the-art method of Houri and Sharan (2012). Additionally, we propose a machine learning approach for predicting interaction signs that combines features from each of these models and show that it improves over any individual model in predicting signs of previously annotated interactions.

2 Materials and methods

2.1 An optimization framework for sign prediction

In this section we describe novel algorithms for inferring signs of physical interactions. We start by formally defining the problem and sketching the previous approach of (Houri and Sharan, 2012). Then, we study three variants of the original problem (each describing a hypothetical signaling model) and develop novel ILP formulations to solve them to optimality on current networks. We assume we are given a (potentially partially signed) physical interaction network along with a collection of cause-effect gene pairs, such as commonly obtained from knockout experiments. The maximum sign assignment (MSA) problem is to assign signs to the unsigned edges of the network in a way that best explains the given pairs. We say that a cause-effect pair (s, t) with sign δ_{st} ($+$ encoding down-regulation of t in response to the knockout of s, $-$ encoding up-regulation of t in response to the knockout of s) is *explained* or *satisfied* by a sign assignment, if there exists a path in the network from s to t whose aggregate sign (the product of the signs along its edges) is δ_{st}. Formally, MSA is defined as follows:

> Input. *A partially signed network G(V, E) and a set of k cause-effect pairs* $(s_1, t_1), \ldots, (s_k, t_k)$ *with signs* $\delta_{s_1 t_1}, \ldots, \delta_{s_k, t_k} \in \{+, -\}$

> Goal. *A sign assignment to the unsigned edges of the network such that a maximum number of input pairs are satisfied by the assignment.*

This problem focuses on the hypothetical A-path signaling model of Yeang *et al.* (2004). Houri and Sharan (2012) showed that due to the nature of the model, any unsigned edge that lies on a cycle in the network cannot be uniquely signed. They generalized this notion to any 2-connected component (or block) by determining if these components are *strongly signed*. They then proposed an approach to reduce the input network to an acyclic one by contracting all edges in these strongly signed components without affecting the maximum number of pairs that could be satisfied. In the reduced network, every pair is connected by a unique path, facilitating the formulation of an ILP to assign signs to the unsigned edges of this path such that

the number of satisfied pairs is maximized. A key drawback of this approach is that reducing the network to an acyclic one severely restricts the number of edges participating in the ILP (coverage) and, hence, restricts the number of interactions that can be uniquely signed. In subsequent paragraphs, we discuss three variants of MSA, each describing a different plausible signaling model, where edges lying on cycles may have unique signs and, hence, may no longer be contracted.

The first variant we consider, A-shortest-path (ASP), considers a signaling model where the length of a satisfying path is always assumed to be the shortest possible. The shortest path assumption is motivated from the observation that signaling pathways tend to be of short length (Silverbush and Sharan, 2014). For each edge $(u,v) \in E$, let $x_{uv} = 1$ denote whether its sign is $-$ (0 if $+$). Similarly, we re-write the signs $\delta_{st} \in \{+,-\}$ as $\delta_{st} \in \{0,1\}$. Due to the nature of knockout experiments, there are usually much fewer sources compared to targets. Hence, for each source s, we construct a subnetwork $G_s(V_s, E_s)$ such that each edge in this subnetwork lies along a shortest path from s to one of its targets t. This is done by applying a breadth-first-search starting from each source and target (Silverbush and Sharan, 2014). Furthermore, we denote by $N_s(v)$ the set of neighbors of v in G_s and by d_{sv} the length of the shortest path from s to v. Additionally, for each pair (s,v) in G_s, we define auxiliary variables c_{sv}, r_{sv} where $c_{sv} = 0$ implies that under the selected sign assignment there exists a shortest path from s to v with aggregate sign r_{sv}, i.e. the node pair (s,v) is *satisfied* under the selected assignment. [Note, (s,s) is trivially assumed to be satisfied]. We also define E^+, E^- which represent subsets of edges in the ILP with known prior positive and negative signs, respectively. Then the following ILP formulation can be used to solve this variant of MSA:

$$\max \sum_{st} y_{st}$$

$$\text{s.t.} \quad 1 + \sum_{u \in \{N_s(v)|d_{sv}=d_{su}+1\}} (c_{su}-1) \leq c_{sv} \quad \forall s,v \in V_s \backslash s$$

$$r_{sv} = \text{XOR}(r_{su}, x_{uv}|c_{sv}=0) \quad \forall s,(u,v) \in \{E_s : d_{sv} = d_{su}+1\}$$

$$c_{st} + y_{st} \leq 1 \quad \forall(s,t)$$

$$r_{ss} = 0, c_{ss} = 0, r_{st} = \delta_{st} \quad \forall(s,t)$$

$$x_{uv} = 0 \quad \forall(u,v) \in E^+$$

$$x_{uv} = 1 \quad \forall(u,v) \in E^-$$

$$y_{st}, x_{uv}, r_{sv}, c_{sv} \in \{0,1\} \quad \forall s,t,u,v$$

The XOR relation between r_{sv}, r_{su} and x_{uv} is conditioned on the value of c_{sv}. That is, $r_{sv} = r_{su} \oplus x_{uv}$ only if $c_{sv} = 0$. It is linearized as follows:

$$r_{sv} - c_{sv} \leq 2 - x_{uv} - r_{su}$$

$$r_{sv} - c_{sv} \leq x_{uv} + r_{su}$$

$$r_{sv} + c_{sv} \geq x_{uv} - r_{su}$$

$$r_{sv} + c_{sv} \geq r_{su} - x_{uv}$$

Let l denote a layer of G_s such that all nodes belonging to this layer have $d_{sv} = l$. Given a feasible solution to the ILP, if $y_{st} = 1$ we can show that there exists a shortest path from s to t with aggregate sign δ_{st}. Indeed, if $y_{st} = 1$ then $c_{st} = 0$ by the third constraint. This implies that $\sum_{u \in N_s(t)|d_{st}=d_{su}+1} (c_{su}-1) < 0$. Thus, if t is in layer l of G_s, there must exist a neighbor u of t in layer $l-1$ such that $c_{su} = 0$. Furthermore, if $c_{st} = 0$, x_{ut} is bound by the XOR constraint to have a sign whose product with r_{su} is δ_{st}. Similarly, if $c_{su} = 0$, there must be a neighbor w in layer $l-2$ where $c_{sw} = 0$ and $r_{sw} \oplus x_{wu} \oplus x_{ut} = \delta_{st}$. By carefully investigating the constraints applicable to the subsequent

layers of G_s (i.e. $l-3,\ldots,0$) we find that there must exist a shortest path from s to t such that the product of signs along its edges is δ_{st}. The final two constraints incorporate prior knowledge of signs in the ILP.

The second variant we study, 'A-directed-shortest-path' (AdirSP), additionally assumes each shortest path explaining a pair to be directed from the cause to the effect. It is worth noting that one cannot adapt existing ILP solutions to the orientation and sign assignment problems, as both rely on reducing the input graph into an acyclic one. This reduction does not work when simultaneously optimizing both. Instead, we simply adapt the ASP formulation above to simultaneously find sign and direction assignments to the network. Specifically, we consider a pair (s,t) to be satisfied by a sign and direction assignment over the network if a directed shortest path from s to t in this assignment has aggregate sign δ_{st}. We call this variant of MSA the 'A-directed-shortest-path' (AdirSP). Let $o_{uv} = 1$ denote whether an edge (u,v) is directed from u to v (0 if from v to u) and let the flow variables f_{uv}^s indicate the existence of a flow from u to v. The flow variables allow computing pair reachability in a directed network. The new ILP is:

$$\max \sum_{st} y_{st}$$

$$\text{s.t.} \quad o_{uv} + o_{vu} = 1 \quad \forall(u,v) \in E$$

$$f_{uv}^s \leq \sum_{w \in N_s(u)\backslash v} f_{wu}^s \quad \forall s,(u,v) \in \{E_s :$$

$$d_{sv} = d_{su}+1, d_{su} \geq 1\}$$

$$f_{uv}^s \leq o_{uv} \quad \forall s,(u,v) \in E_s$$

$$a_{uv}^s = (1 - f_{uv}^s) \text{ OR } c_{su} \quad \forall s,(u,v) \in \{E_s : d_{sv} = d_{su}+1\}$$

$$1 + \sum_{u \in N_s(v)|d_{sv}=d_{su}+1} (a_{uv}^s - 1) \leq c_{sv} \quad \forall s,v \in V_s \backslash s$$

$$r_{sv} = \text{XOR}(r_{su}, x_{uv}|c_{sv}=0, f_{uv}^s=1) \quad \forall s,(u,v) \in \{E_s : d_{sv} = d_{su}+1\}$$

$$c_{st} + y_{st} \leq 1 \quad \forall(s,t)$$

$$r_{ss} = 0, c_{ss} = 0, r_{st} = \delta_{st} \quad \forall(s,t)$$

$$x_{uv} = 0 \quad \forall(u,v) \in E^+$$

$$x_{uv} = 1 \quad \forall(u,v) \in E^-$$

$$y_{st}, x_{uv}, o_{uv}, a_{uv}^s, r_{sv}, c_{sv}, f_{uv}^s \in \{0,1\} \quad \forall s,t,u,v$$

The first constraint ensures that each edge has a unique orientation. In some feasible solution, if $f_{uv}^s = 1$, then the second and third constraint ensure that a directed path exists from s to v containing edge (u,v).

Note that the XOR relation that helps determine the sign of an edge now additionally depends on the existence of a flow in that edge. The constraint is linearized as follows:

$$r_{sv} - c_{sv} - 1 + f_{uv}^s \leq 2 - x_{uv} - r_{su}$$

$$r_{sv} - c_{sv} - 1 + f_{uv}^s \leq x_{uv} + r_{su}$$

$$r_{sv} + c_{sv} + 1 - f_{uv}^s \geq x_{uv} - r_{su}$$

$$r_{sv} + c_{sv} + 1 - f_{uv}^s \geq r_{su} - x_{uv}$$

Another change from the previous formulation is the definition of auxiliary variables a_{uv}^s for each edge participating the ILP. Their value depends on the flow in edge (u,v) originating from s and on c_{su}. The OR relation between these variables is linearized as follows.

$$a_{uv}^s \leq (1 - f_{uv}^s) + c_{su}$$

$$a_{uv}^s \geq 1 - f_{uv}^s$$

$$a_{uv}^s \geq c_{su}$$

Given a feasible solution in which $y_{st} = 1$, we show that there exists a shortest path oriented from s to t such that its aggregate sign is δ_{st}. Let t be in layer l of the shortest path graph G_s. If $y_{st} = 1$, then by the seventh constraint $c_{st} = 0$. It follows that $\sum_{u \in N_s(t)|d_{st}=d_{su}+1} (a_{ut}^s - 1) < 0$ (by constraint 5), which implies that there exists a neighbor u in layer $l - 1$ where $a_{ut}^s = 0$. This implies $f_{ut}^s = 1$, $c_{su} = 0$ (constraint 4) and δ_{st} must be the product of the signs given by x_{ut} and r_{su} (constraint 6). Additionally, $c_{su} = 0$ implies there exists a neighbor w in layer $l - 2$ where $a_{wu}^s = 0$ (constraint 5). This implies $f_{wu}^s = 1$, $c_{sw} = 0$ and $r_{su} \oplus x_{wu} \oplus x_{ut} = \delta_{st}$. In this manner after carefully investigating the constraints through subsequent layers of G_s (i.e. $l - 3, l - 4, \ldots, 0$) we can find a directed shortest path from s to t such that the product of signs along its edges is δ_{st}. The last two constraints account for signs that are already known.

The underlying assumption in both signaling models above is that a single path is sufficient to force a pre-defined effect. However, due to the inherent stochasticity in signaling, this might not always be the case (Ladbury and Arold, 2012). Moreover, on careful examination of signed interactions of Figure 1, we find that for any node pair connected by more than one completely signed path, the product of signs on each path is the same. Hence, we strengthen the pair satisfaction assumption in the ASP model to require that a pair (s, t) is satisfied if all shortest paths connecting s to t admit the same aggregate sign δ_{st}. We call this variant 'All-shortest-paths' (AllSP) and solve for it using the following formulation:

$$\max \sum_{st} y_{st}$$

$$\text{s.t.} \quad c_{su} \leq c_{sv} \qquad \forall s, (u, v) \in \{E_s : d_{sv} = d_{su} + 1\}$$

$$r_{sv} = \text{XOR}(r_{su}, x_{uv}|c_{sv} = 0) \quad \forall s, (u, v) \in \{E_s : d_{sv} = d_{su} + 1\}$$

$$c_{st} + y_{st} \leq 1 \qquad \forall(s, t)$$

$$r_{ss} = 0, c_{ss} = 0, r_{st} = \delta_{st} \qquad \forall(s, t)$$

$$x_{uv} = 0 \qquad \forall(u, v) \in E^+$$

$$x_{uv} = 1 \qquad \forall(u, v) \in E^-$$

$$y_{st}, x_{uv}, r_{sv}, c_{sv} \in \{0, 1\} \qquad \forall s, t, u, v$$

As above, let t belong to layer l of G_s. Given a feasible solution to this new formulation, if $y_{st} = 1$, c_{st} must be 0 (from third constraint). Hence, for every neighbor u of t that lies in layer $l - 1$ of G_s, $c_{su} = 0$ (from first constraint). This in turn constrains the sign assignment of the respective edges (i.e. $r_{su} \oplus x_{ut} = \delta_{st}$, for all neighbors u in layer $l - 1$). By carefully investigating the constraints through subsequent layers of G_s (i.e. $l - 2, l - 3, \ldots, 0$), it becomes apparent that for any node v in G_s, all shortest paths from s to v must admit the same aggregate sign (r_{sv}). Hence, all shortest paths from s to t must have an aggregate sign δ_{st}.

Notably, the models discussed above permit mathematically efficient formulations. Specifically, if p is the number of sources ($p \ll k$), then each formulation contains $O(k + p|V| + |E|)$ variables and $O(k + p(|V| + |E|))$ constraints.

2.2 Quantifying the activation/repression potential of a physical interaction

Each of the above models may admit multiple sign assignments with optimal or near optimal scores. Hence, it is necessary to quantify the robustness of a sign assignment to an edge. To this end, we solve each ILP repeatedly n times; each time adding a small Gaussian noise of mean 0 and variance 0.01 to the objective function as shown below. This stochastic approach, motivated by Hazan and Jaakkola (2012), effectively results in a random sampling of

different likely solutions that exist nearby in the optimum solution space, thereby allowing us to assess the robustness of the sign on each edge. The procedure is as follows:

```
1: procedure GetScores(ILP, n)
2:     scores_uv ← 0, ∀(u, v) ∈ E that are in ILP
3:     for i = 1: n
4:         set objective : ∑_st(1 + ε_st)y_st, where ε_st ∼ N(0, 0.01)
5:         x* ← solve(ILP)
6:         scores_uv = scores_uv + x*_uv/n, ∀(u, v) ∈ E that are in ILP
7:     return scores
```

An edge score close to 1 implies that the sign is negative with high confidence, a score close to 0 implies a positive sign with high confidence and a score close to 0.5 implies that the sign on that edge cannot be uniquely determined (possibly implicating the absence of an activation/repression effect). For efficiency, we use $n = 10$ throughout. Our results remain qualitatively the same for larger values of n.

3 Results

3.1 Input data

We focused our analysis on budding yeast (*Saccharomyces cerevisiae*). We obtained 4095 PDIs spanning 2079 proteins (conserved across at least two other yeast species) from MacIsaac et al. (2006). We additionally downloaded 2930 high-quality experimentally verified PPIs from Yu et al. (2008), 1361 kinase–substrate/phosphatase–substrate interactions (KPIs) among 802 proteins from Breitkreutz et al. (2010), and 189 physical interactions from signaling pathways of yeast in Kyoto Encyclopedia of Genes and Genomes (KEGG). We merged these sets into a *unified* yeast network of 8268 unique physical interactions among 3695 proteins.

We extracted all 110 487 knockout pairs spanning 6228 proteins from Reimand et al. (2010) and additionally 699 771 pairs spanning 6110 proteins from Kemmeren et al. (2014). A pair was assigned a positive sign if the target gene was repressed in response to knockout of the source, and a negative sign if the target gene was activated/up-regulated. We restricted ourselves to knockout pairs such that the absolute log fold change in expression of the target gene is >2 and FDR < 0.001. This leaves us with 1756 significant knockout pairs from Reimand et al. (2010), referred to here as the *Reimand set*, and 3524 significant knockout pairs from Kemmeren et al. (2014), referred to here as the *Kemmeren set*. The above choice of thresholds was made while taking into consideration the inherent computational complexity of the problem.

3.2 Validation data

For a systematic validation of our sign prediction models we collected sign information as follows. Only 147 of 192 physical interactions in yeast had an experimentally confirmed sign from KEGG (See Fig. 1).

In addition, following Houri and Sharan (2012), we extracted gene ontology (GO) molecular function annotations related to transcriptional activators (GO: 0045893) and transcriptional repressors (GO: 0045892). PDIs originating from transcriptional activators were given a positive sign whereas PDIs originating from transcriptional repressors were given a negative sign. Finally, we also

extracted information on protein kinases (GO: 0004672) and protein phosphatases (GO: 0004721). We reasoned that since there are roughly three times as many confirmed functionally activating phosphorylation sites compared to repressive ones (PhosphoNET database, www.phosphonet.ca), and that 71% of phosphorylation interactions of yeast in KEGG are annotated as activating and 81% of de-phosphorylation interactions of yeast are annotated as repressing, kinase–substrate interactions tend to be activating while phosphatase–substrate interactions tend to be repressing. Thus, physical interactions linking a GO annotated kinase and a substrate were given a positive sign whereas interactions linking a GO annotated phosphatase to a substrate were given a negative sign. Any interaction in the unified network that had conflicting signs was left unsigned (unless it had sign information from KEGG, in which case this latter information was used).

In summary, the validation set consists of three groups of signed interactions in the network: (i) 2014 (1131+, 883−) signed PDIs, (ii) 1044 (872+, 172−) signed kinase/phosphatase–substrate interactions and (iii) 147 (96+, 51−) signed KEGG interactions.

3.3 Performance evaluation of individual models

We evaluated each of the four models presented above (A-path/ASP/AdirSP/AllSP) in a 5-fold cross-validation setting on the unified yeast network, focusing on the interactions covered by each model, i.e. participating in the corresponding ILP. To this end, we randomly divided all signed and covered interactions into five equal parts. Using each model, we predicted the activation/repression potential of the interactions in each part (see Section 2.2) while constraining the signs of interactions in the remaining parts. Then we measured the performance of the activation/repression scores of a given model across the five parts for different subsets of signed interactions covered by the model. For each subset, we denote its set of covered positive and negative interactions by E^+ and E^-, respectively.

As a benchmark, we discuss the performance of the previous A-path model. Recall that in this model we should contract all interactions that lie in a strongly signed block of size ≥ 3 (see Section 2.1). Since all blocks were strongly signed, this resulted in an acyclic network with 77% of the interactions contracted. When working with knockout pairs from the Reimand set, we observe that only 1% of all the network interactions participate in the ILP constraints due to network reduction, and 25 of them belong to the validation set. Due to low coverage over the validation set, we instead evaluated this framework using knockout pairs from the Kemmeren set. Overall, 4% of network interactions are covered in this instance and 73 interactions from the validation set were part of the ILP formulation, yielding an AUC of 0.66. Since there were only 73 interactions to validate our predictions, we could not evaluate the performance on individual subsets.

Next, we evaluated the ASP, AdirSP and AllSP models over the unified network. Tables 1 and 2 summarize the performance over the validation PDIs, KPIs and the KEGG interactions. We find that our new formulations lead to sign assignments on 35% of network interactions when working with the Reimand set and 59% of network interactions when working with the Kemmeren set; ≈15-fold coverage increase compared to previous work (A-path).

In order to directly compare the performance of the A-path model to our suggested alternative models, we evaluated them on the restricted validation set of 73 interactions covered by the A-path model. On this set ($|E^+| = 49$, $|E^-| = 24$) the performance of AdirSP was lower to A-path (AUC of 0.64), while ASP and AllSP had better performance (AUCs of 0.73 and 0.68, respectively).

Table 1. Performance evaluation using the Reimand set (coverage of 35%)

Interaction	$|E^+|, |E^-|$	AUC (ASP)	AUC (AdirSP)	AUC (AllSP)
PDI	435, 458	0.75	0.63	0.84
KPI	205, 20	0.83	0.56	0.72
KEGG	40, 27	0.56	0.52	0.65

Table 2. Performance evaluation using the Kemmeren set (coverage of 59%)

Interaction	$|E^+|, |E^-|$	AUC (ASP)	AUC (AdirSP)	AUC (AllSP)
PDI	744, 653	0.63	0.59	0.83
KPI	522, 98	0.61	0.51	0.77
KEGG	46, 32	0.58	0.54	0.71

Table 3. Performance evaluation of the random forest classifier using the Reimand set

Interaction	$|E^+|, |E^-|$	AUC (classifier)
PDI	435, 458	0.86
KPI	205, 20	0.85
KEGG	40, 27	0.77

### 3.4 Performance evaluation of the combined model

Previous work as well as our models above vary in the assumptions they make on the way a knockout effect is explained, going all the way from requiring a single path of any length to requiring all paths of shortest length. Note that we adopt these models partly because they are grounded in our very own observations of cellular signaling pathways (see Section 2.1) and because they permit an efficient mathematical formulation. These descriptions are not perfect. In turn, the solution of each model allows different degrees of freedom on the signs of underlying interactions. To make the best inference possible for each physical interaction given the complex nature of cellular signaling, we integrate the predictions of each model in an ensemble. That is, using the sign scores from solutions to ASP, AdirSP and AllSP as features, we train a hybrid model, specifically a random forest classifier, that makes an overall prediction of the sign of an interaction (A-path was excluded due to low coverage).

The ensemble model is evaluated via nested cross-validation. In detail, the validation set is divided into the same five parts as above. Four of the parts are used for training the individual models to score the fifth part. Next, we perform a 5-fold cross-validation on the fifth part to train and test the classifier. Finally, using the cross-validated predictions across all parts, we report the mean classifier performance (AUC) against the signs of different validation subsets. Tables 3 and 4 summarize the performance of the random forest classifier on the different knockout sets and validation subsets.

The performances of the classifier and the individual models are depicted in Figures 2 and 3. Importantly, we observe that the classifier outperforms all individual models on the set of curated interactions from KEGG. It also outperforms the different models with respect to PDIs and KPIs on the Reimand set. The lower performance of the classifier on the KPI set (compared with the AllSP model) when working with the Kemmeren set is likely an artifact resulting from the skewed distribution of class labels. Such a skew may

Table 4. Performance evaluation of the random forest classifier using the Kemmeren set

Interaction	$|E^+|, |E^-|$	AUC (classifier)
PDI	744, 653	0.80
KPI	522, 98	0.67
KEGG	46, 32	0.81
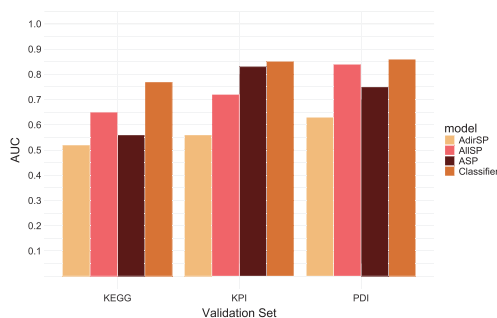
Fig. 2. Performance evaluation of all models using the Reimand set
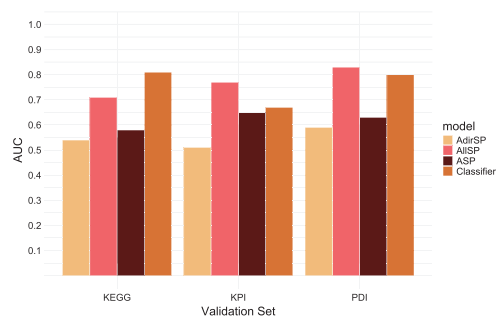
Fig. 3. Performance evaluation of all models using the Kemmeren set

influence ensemble classifier performance on unseen data. (Galar *et al.*, 2012).

4 Conclusions

In summary, we developed novel ILP formulations for predicting signs of physical interactions under different signaling models. We discussed the underlying assumptions guiding the predictions of each model and its advantages in terms of coverage relative to prior work by Houri and Sharan (2012). We then measured the cross-validation accuracy of our models in predicting signs across two knockout datasets to find that our models lead to improvement in accuracy and coverage over the previous state-of-the art method by Houri and Sharan (2012). Finally, we derive a hybrid signaling model based classifier that uses the sign assignment confidence scores of each model studied for predicting interaction signs. This was partly motivated by the fact that the three models presented in the paper, although mathematically efficient to represent, may be insufficient to capture the complex nature of cell signaling. Furthermore, this warrants the exploration of other plausible models that could be potentially integrated into the classifier to improve its predictions. For instance, one could additionally formulate an ILP that considers All-Directed Shortest paths (All-dirSP) as a plausible signaling mechanism to explain some of the knockout effects.

Notably, the AdirSP model is the first to combine both direction and sign prediction within the same optimization framework and may be of independent interest to readers to expand on for the general problem of annotating a network with directions and signs.

A potential limitation of our approach is the computational cost involved in solving the models (especially if the network is sparsely annotated with signs). In this work, running the AllSP model on the Kemmeren set took the most time (up to 3 h to obtain a single solution). In all other cases it took on average 3 min to obtain a single solution (note that these values may differ based on the solver and the computer used). A theoretical comparison of efficiency can be done by analytically reasoning about the size of the search space for each model. The ASP and AdirSP models require $O(d)$ XOR constraints to be satisfied to explain a single knockout pair (s, t) (where d denotes the shortest path distance from s to t). In contrast, the AllSP model requires a much larger number of $O(|E|)$ XOR constraints to be satisfied to explain a single knockout effect (where $|E|$ denotes the number of edges in the network). Each constraint introduces at least one new Boolean variable and hence the search space for AllSP is much larger than that for ASP or AdirSP. This might pose a major limitation when it comes to predicting physical interaction signs over much larger networks, e.g. the human network, which still remains mostly unsigned. To enhance the scalability of our methods, one can pre-process the network and either contract interactions involved in protein complexes or pre-annotate them with positive signs (which are logically consistent with their role as a means to propagate signal forward without influencing the overall effect of the path taken). This vastly cuts down the size of the solution space; by half per interaction. The different models we proposed represent a trade-off between model coverage (of networks edges) and complexity (of its solution).

Another limitation of our results is that they are based on two approximate sources of sign information, namely KPI signs derived from whether the source protein is a kinase or a phosphatase and PDI signs derived from whether the involved transcription factor is an activator or a repressor. We expect our models to yield more accurate results as better quality sign information becomes available.

Acknowledgements

Funding

References

Blokh,D. *et al.* (2013) The approximability of shortest path-based graph orientations of protein–protein interaction networks. *J. Comput. Biol.*, **20**, 945–957.

Breitkreutz,A. *et al.* (2010) A global protein kinase and phosphatase interaction network in yeast. *Science*, **328**, 1043–1046.

Cosgrove,E.J. *et al.* (2008) Predicting gene targets of perturbations via network-based filtering of mrna expression compendia. *Bioinformatics*, **24**, 2482–2490.

Galar,M. *et al.* (2012) A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Trans. Syst. Man Cybern. C Appl. Rev.*, **42**, 463–484.

Gitter,A. *et al.* (2011) Discovering pathways by orienting edges in protein interaction networks. *Nucleic Acids Res.*, **39**, e22.

Hazan,T. and Jaakkola,T. (2012) On the partition function and random maximum a-posteriori perturbations. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pp. 991–998.

Houri,S. and Sharan,R. (2012) Sign assignment problems on protein networks. In: *WABI*. Springer, Berlin, pp. 338–345.

Huttlin,E.L. *et al*. (2017) Architecture of the human interactome defines protein communities and disease networks. *Nature*, **545**, 505–509.

Kemmeren,P. *et al*. (2014) Large-scale genetic perturbations reveal regulatory networks and an abundance of gene-specific repressors. *Cell*, **157**, 740–752.

Ladbury,J.E. and Arold,S.T. (2012) Noise in cellular signaling pathways: causes and effects. *Trends Biochem. Sci.*, **37**, 173–178.

MacIsaac,K.D. *et al*. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics*, **7**, 113.

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

Ourfali,O. *et al*. (2007) Spine: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, **23**, i359–i366.

Peleg,T. *et al*. (2010) Network-free inference of knockout effects in yeast. *PLoS Comput. Biol.*, **6**, e1000635.

Reimand,J. *et al*. (2010) Comprehensive reanalysis of transcription factor knockout expression data in *Saccharomyces cerevisiae* reveals many new targets. *Nucleic Acids Res.*, **38**, 4768–4777.

Silberberg,Y. *et al*. (2014) A method for predicting protein-protein interaction types. *PLoS One*, **9**, e90904.

Silverbush,D. and Sharan,R. (2014) Network orientation via shortest paths. *Bioinformatics*, **30**, 1449–1455.

Silverbush,D. *et al*. (2011) Optimally orienting physical networks. *J. Comput. Biol.*, **18**, 1437–1448.

Stelzl,U. *et al*. (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

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

Yu,H. *et al*. (2008) High-quality binary protein interaction map of the yeast interactome network. *Science*, **322**, 104–110.