Published online 24 November 2010

Nucleic Acids Research, 2011, Vol. 39, No. 4 ***e22***
doi:10.1093/nar/gkq1207

Discovering pathways by orienting edges in protein interaction networks

Anthony Gitter[1], Judith Klein-Seetharaman[2], Anupam Gupta[1] and Ziv Bar-Joseph[1,*]

[1]Computer Science Department, Carnegie Mellon University and [2]Department of Structural Biology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

Received August 13, 2010; Revised October 24, 2010; Accepted November 8, 2010

ABSTRACT

Modern experimental technology enables the identification of the sensory proteins that interact with the cells' environment or various pathogens. Expression and knockdown studies can determine the downstream effects of these interactions. However, when attempting to reconstruct the signaling networks and pathways between these sources and targets, one faces a substantial challenge. Although pathways are directed, high-throughput protein interaction data are undirected. In order to utilize the available data, we need methods that can orient protein interaction edges and discover high-confidence pathways that explain the observed experimental outcomes. We formalize the orientation problem in weighted protein interaction graphs as an optimization problem and present three approximation algorithms based on either weighted Boolean satisfiability solvers or probabilistic assignments. We use these algorithms to identify pathways in yeast. Our approach recovers twice as many known signaling cascades as a recent unoriented signaling pathway prediction technique and over 13 times as many as an existing network orientation algorithm. The discovered paths match several known signaling pathways and suggest new mechanisms that are not currently present in signaling databases. For some pathways, including the pheromone signaling pathway and the high-osmolarity glycerol pathway, our method suggests interesting and novel components that extend current annotations.

INTRODUCTION

Reconstructing interaction networks in the cell is one of the great challenges of computational biology. Work in this area using high-throughput data sets focused on the reconstruction of regulatory networks (1–3), the analysis of metabolic networks (4,5) and the discovery of signaling networks and pathways (6,7). However, while data about the directionality of an interaction are available when using high-throughput data to reconstruct and analyze regulatory and metabolic networks, this information is often missing for signaling networks. For example, ChIP-chip and ChIP-Seq studies (8,9) identify which transcription factors regulate genes, studies of microRNAs often look for targets (10) and motif studies are performed upstream of genes (11). Similarly, metabolic networks are often modeled using knowledge regarding the order of genes and enzymes (12). In contrast, even though signaling networks are directed, the available protein–protein interaction (PPI) data are almost always undirected (13,14). Thus, it is challenging to reconstruct these networks since it requires not only the right set of proteins and interactions but also the directionality for each edge when assembling pathways.

Recent proteomic studies have examined interactions between cellular proteins and the molecules and agents that affect them [e.g. host–pathogen interactions (15)]. In many cases, we can also determine the proteins that are impacted downstream of these initial interactions, either through expression or through knockdown studies (16–18). Thus, an important challenge is to determine the signaling networks or pathways that are used to transmit information from known sources to known targets. To reconstruct these networks we need to infer an orientation for undirected PPI networks in order to identify directed paths between sources and targets. This is a difficult problem because there are many paths that can link two proteins in the interaction network. Fortunately, we can rely on a few established assumptions to simplify the problem. First, it is likely that biological responses are controlled by reasonably short signaling cascades, so we can only search for length-bounded paths. Pathways in signaling databases such as KEGG (19) and the *Science Signaling* Database of Cell Signaling (http://stke.sciencemag.org/cm/) on average contain only five edges between a target and its closest

*To whom correspondence should be addressed. Tel: 412 268 8595; Fax: 412 268 3431; Email: zivbj@cs.cmu.edu

source (Supplementary Methods), and previous signaling pathway prediction methods have focused on pathway segments of only 3–4 edges (7). Second, we have varying degrees of confidence in the available interaction data [e.g. small-scale versus high-throughput experiments (20)] and, as we show, focusing on the more confident edges leads to better pathways. Finally, in many cases there are overlapping parallel pathways linking sources and targets (21–23) so selecting an orientation that generates multiple possible pathways may produce better reconstruction results.

Although much attention has been given to the signaling pathway prediction problem, nearly all previous work does not consider the orientation of the paths and simply selects subsets of edges, yielding undirected predictions. One of the earliest undirected pathway prediction algorithms was NetSearch (24). NetSearch enumerated linear pathways and ranked all putative pathways by clustering the gene expression profiles of pathway members and generating hypergeometric distribution-based scores. Since linear paths do not fully capture the complexity of signaling networks, Scott *et al.* (6) used a color-coding technique to search for paths and higher order structures (trees and parallel paths) in a weighted protein interaction graph. Lu *et al.* (25) presented a randomized divide-and-conquer algorithm that, like Scott *et al.*, supported complex non-linear pathways structures. PathFinder (7) integrated multiple data sources to extract association rules describing protein function in known signaling pathways and then used these rules, along with additional expression data, to detect new pathways of interest in the network. Whereas many previous methods searched for source–target pathways individually, Zhao *et al.* (26) formulated a linear program to identify a single global signaling subnetwork that satisfies various constraints. We refer to their technique as the unoriented edge selection algorithm. Recognizing the trade-offs between local and global search approaches, Yosef *et al.* (27) presented an algorithm that combined the two objectives and could be tuned to give preference to one or the other on a particular run. While all of these methods led to useful findings, none of them generates directed pathways. As we show in the 'Results' section, by ignoring the edge orientations these methods lose important information that improves pathway reconstruction and thus contain far fewer known signaling pathways in their predictions.

Relatively few methods have been developed to try to explicitly address the edge orientation problem. In (28), the authors defined the maximum tree orientation (MTO) problem, which focused on reachability. They considered a source–target pair to be satisfied as long as any single path of arbitrary length connected them. As a result, cycles in the PPI network could be contracted and the problem was equivalent to orienting a tree. While this variant of the edge orientation problem can be approximated well, such a structure cannot give preference to short paths or high-confidence edges and also ignores redundant pathways. Liu *et al.* (29) predicted directed signaling pathways in multiple species. However, because their method relies on specific protein domain interactions, it does not scale to the entire proteome. Indeed, as the authors noted, coverage, the fraction of interactions in the test set for which predictions could be made, was $<50\%$ at the thresholds they used. Probabilistic graphical models have also been used to orient edges when trying to explain knockout effects via a physical interaction network consisting of PPI and protein–DNA interactions (30). The Physical Network Models algorithm constructs a factor graph and applies belief propagation to infer both PPI directionality and regulatory effect (inhibition or activation). While this approach works well for relatively small networks and short pathways, as we show in the 'Results' section, it does not scale well. SPINE (31) adapts the Physical Network Models formulation but expresses the problem as an integer program. However, SPINE only focuses on identifying activation and repression regulatory effects of either proteins or edges and does not attempt to orient the network. Conversely, our goal is to determine directionality in PPI signaling networks where the positive and negative regulatory effects upon genes are not the primary concern.

In this article, we formalize the orientation problem for length-bounded pathways in weighted interaction networks and show that the problem is non-deterministic polynomial-time hard (NP-hard). We next present three algorithms for this problem and their approximation guarantees. Two of the algorithms use methods developed to solve weighted Boolean satisfiability (SAT) problems and the third is based on probabilistic selections. We applied all three algorithms to PPI networks using simulated and biologically derived sources and targets. As we show using the simulated sources and targets, the algorithms perform very well and in practice achieve very good solutions in reasonable time. Using real signaling networks, we show that our algorithms can recover many known pathways and improve upon prior methods for pathway discovery. We also analyzed pathways discovered by the algorithms that do not appear in current signaling databases. In many cases, these match known knowledge about the directionality of the interactions within pathways. Other predictions raise interesting biological hypotheses.

We note that while we focus on orienting undirected protein interaction networks, the algorithms we present are applicable to other biological network orientation problems as well and can also be used with mixed graphs, which consist of undirected PPI and directed interactions (e.g. protein–DNA binding).

MATERIALS AND METHODS

Our goal is to orient edges in the protein interaction network in order to extract the high-confidence signaling pathways activated as part of a response program. Below, we formally define the problem, prove that it is NP hard and then discuss several methods for approximating the optimal solution for this problem. Supplementary data and source code for the methods presented in the article are available from the supplementary web site: http://www.sb.cs.cmu.edu/OrientEdges.

Formulating the edge orientation optimization problem

We assume we are given a weighted undirected graph $G = (V,E)$ which represents our current knowledge of protein interactions. We are also given a maximum path length k and source–target pairs of the form $< s_i,t_i >$ such that $s_i \in S \subseteq V$ and $t_i \in T \subseteq V$. Our goal is to orient edges $e = (u,v) \in E$ from u to v or from v to u such that the weight of all satisfied paths between sources and targets with length at most k is maximized. Each simple path takes the form $p = (v_1,v_2),(v_2,v_3),...,(v_l,v_{l+1})$ where $v_1 = s_i$, $v_{l+1} = t_i$ and $l \leq k$ for some pair $< s_i,t_i >$. A path is satisfied in a given network orientation if and only if for every edge (v_j,v_{j+1}) along the path the edge is oriented from v_j to v_{j+1} in the network. Multiple paths may exist between a single source–target pair as long as paths with the same source and target have at least one disjoint edge (as mentioned above, parallel pathways are very common). After orientation there may be directed source–target paths in the graph that contain more than k edges, but they are not incorporated into the objective function.

All vertices and edges in the graph have real-valued weights denoted $w(v)$ and $w(e)$, respectively. While all vertices (proteins) have the same weight in our current implementation, allowing for varying protein weights is a useful feature in cases where some proteins are known to be involved in the response. The edge weights are assigned based on the confidence in each protein interaction, which in our implementation depends on the type of experimental support provided for that edge. Weights represent our confidence in the presence of the edge or in the involvement of a gene in the response, and the weight of an entire path p is $w(p) = \prod_{v \in p} w(v)^* \prod_{e \in p} w(e)$. Since we use weights in the range [0,1] to represent edge confidence, this definition of path weight causes long paths to have lower weights than short paths. Thus, the objective in the Maximum Edge Orientation (MEO) problem is to maximize the function:

$$\sum_{p \in P} I_S(p)^* w(p)$$

where P is the set of all unique paths between sources and targets with length at most k and $I_S(p)$ is an indicator function that has the value 1 if path p is satisfied.

Although we currently assume that edge weights are symmetric, one simple yet powerful generalization is to allow asymmetric edge weights when there is a prior belief that one orientation of an edge is more likely than the other. Incorporating such information involves using the appropriate direction-specific weight for each edge when calculating $w(p)$, but does not require any adjustments to the proposed MEO approximation algorithms.

MEO is NP hard

Similar to Medvedovsky *et al.* (28), we sketch a proof that MEO is NP hard for any $k \geq 2$ by reduction from Maximum Directed Cut (MAX-DI-CUT) (32). See the Supplementary Methods for an extended proof. Given a directed graph $G = (V,E)$, the objective of MAX-DI-CUT is to partition the vertices V into sets A and B, where $A \subseteq V$ and $B = V - A$, such that the number of directed edges that begin in A and end in B is maximized. To reduce a MAX-DI-CUT instance $G = (V,E)$ to MEO, we add a new node C and construct an undirected graph $H = (V',E')$, where $V' = V \cup \{C\}$ and $E' = (v',C)$ for all $v' \in V$ (Figure 1). All edges and vertices in H are given a weight of 1 so that for all p, $w(p) = 1$. For every directed edge (u, v) in the MAX-DI-CUT instance, we create a source–target pair $< u,v >$ in the MEO instance.

Observe that there is a one-to-one mapping between an orientation O of H and a cut $A \subseteq V$ of G. If an edge (v',C) in H is oriented toward C, then place the corresponding vertex v in the set A. For all edges (v',C) oriented away from C, include v in the set B. Thus, for any satisfied path $(v'_1,C),(C,v'_2)$ in H, the directed edge (v_1,v_2) will be across the cut in G (Figure 2). Since all paths have a weight of 1, the weight of all satisfied paths, which is the property MEO maximizes, is equal to the number of edges across the cut in G.

Note that the above reduction proves that the problem is hard even for $k = 2$. MAX-DI-CUT cannot be approximated within a factor of 12/13 (33), which implies MEO is inapproximable within 12/13 for $k \geq 2$. However, our problem is even harder for larger (yet still reasonable) values of k. Consequently, we can reduce MAX-3-SAT, which is harder to approximate (33), to MEO with $k \geq 5$ yielding the stronger inapproximability bound of 7/8 for this range of k (proof omitted).

Since MEO is NP hard even for small k, we describe approximation algorithms for orienting the graph with varying theoretical guarantees and running times. For an instance m of a maximization optimization problem, if the optimal value of the objective function is OPT(m) and an approximation algorithm guarantees a value of

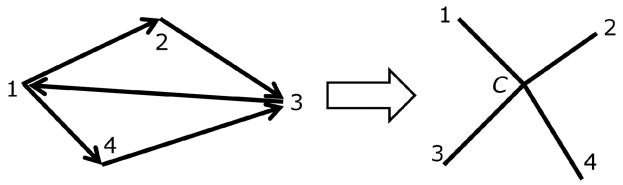

Figure 1. An example of the MAX-DI-CUT to MEO transformation. The MEO graph has the same vertices as the MAX-DI-CUT graph plus an additional center vertex, to which all other vertices are connected. The MAX-DI-CUT edges are used to define the MEO source–target pairs.

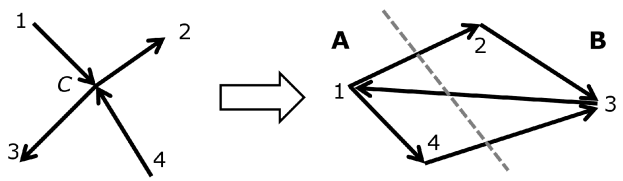

Figure 2. Mapping an orientation of the MEO instance back to a directed cut. An orientation in the MEO problem uniquely defines a cut in the MAX-DI-CUT instance. The number of satisfied paths in the MEO instance is identical to the number of directed edges from A to B.

at least APX(m), we say the algorithm guarantees an r-approximation where

$$r = \frac{\text{APX}(m)}{\text{OPT}(m)}$$

Random orientation

The simplest approximation algorithm randomly assigns an orientation to each edge in the graph. For a particular path, let the orientation an edge takes when the path is satisfied be the optimal orientation for that edge with respect to the path. After a random orientation, each edge in a particular path will be optimally oriented with probability 1/2. Since the path contains at most k edges and all edges are oriented independently, the probability that a given path is satisfied is:

$$P(I_S(p) = 1) = \prod_{e \in p} P(I_O(e, p) = 1) = \prod_{e \in p} \tfrac{1}{2} \geq \left(\tfrac{1}{2}\right)^k$$

where $I_O(e, p)$ is an indicator function that takes value 1 if the edge e is optimally oriented for path p. Thus, the expected value for a path is $E[p] \geq \left(\frac{1}{2}\right)^k * w(p)$ and by linearity of expectation the random orientation yields a $\frac{1}{2^k}$-approximation. In practice, we deterministically fix the orientation of any edges that are used in the same direction by all paths that contain them and only randomly orient the remaining edges. This can only improve the likelihood that a particular path is satisfied, thus the approximation guarantee is not affected.

MIN-k-SAT and MAX-k-CSP approximation algorithms

Although the MEO problem is a maximization problem, an MEO instance can be transformed to a weighted MIN-k-SAT (34,35) instance. Weighted MIN-k-SAT is an optimization version of the traditional SAT problem in which weighted disjunctive clauses with at most k literals are given and the objective is to find the assignment to all variables that minimizes the sum of the weights of the satisfied clauses. In a similar manner, MEO can be reduced to MAX-k-CSP (constraint satisfaction problem), which maximizes conjunctive clauses instead of minimizing disjunctive clauses. MAX-k-CSP is more difficult to approximate than MIN-k-SAT, but the technique by Charikar *et al*. (36) can be used to obtain a $O\left(\frac{k}{2^k}\right)$-approximation ratio for MEO, improving upon the $\frac{1}{2^k}$-approximation guarantee obtained via random orientation. See the Supplementary Methods and Figures S1–S3 for further theoretical details, pseudocode, and a discussion of how we approximate the MIN-k-SAT and MAX-k-CSP instances in practice.

Improving approximations with local search

The solution returned by any of the algorithms described above can typically be improved by using it as the starting point for a local search instead of taking it directly as the final orientation. Specifically, local search in the MEO problem involves iteratively finding the edge that will yield the greatest improvement in the objective function if its orientation is changed and flipping that edge's direction. Complete pseudocode can be found in Supplementary Figure S4.

In practice, we have found that the local search procedure terminates quickly, but if worst case runtime is a concern, the number of iterations can be bounded by requiring that each edge flip improves the score by some fixed percentage of the score. While helpful in practice, local search does not improve the theoretical guarantees of any of the algorithms.

PPI databases and edge weight assignment

The unweighted PPI network consisted of the union of all yeast interactions in the BioGRID (37), IntAct (38) and MINT (39) databases. In the context of MEO, every edge and vertex in an unweighted network is given a default weight of 1 so that all paths have equal weight.

The first weighted network was constructed by taking the intersection of edges in the three PPI databases. Since the reliability of a reported PPI has been shown to increase with the number of observations of that interaction (40), we assigned a weight of 0.75 to interactions appearing in exactly two databases and a weight of 0.95 to edges in all three databases. Edges only present in a single database were discarded.

The second weighting scheme was based on the type of experiment(s) used to detect the interactions. These weights were calculated only for the BioGRID edges, because the calculation is dependent on the experimental types reported by BioGRID. Each PPI was assigned a probability that the reported pair of proteins truly physically interact, which was used as the weight of the edge between the two proteins in the MEO graph. As in the other weighting scheme, we place more trust in interactions that have been observed more frequently. Therefore, the probabilities were computed using both the confidence in the experimental systems used to detect the interaction and the number of separate publications that report the interaction. For each interaction between proteins $P1$ and $P2$, the probability their interaction is a true positive is given by the formula:

$$P(\text{interact}(P1, P2)) = 1 - \prod_{i \in I_{P1,P2}} (1 - c(i))$$

where i is a member of the set $I_{P1,P2}$, all of the distinct (based on experiment type and PubMed identifier) instances of that interaction in the PPI data set, and $c(i)$ is the confidence in the class of experiments to which i belongs. The confidence scores for the 15 types of experiments presently considered can be found in Supplementary Table S1, and Supplementary Table S4 compares the sizes of the different networks.

Selecting gold standard sources and targets

After identifying several signaling pathways in KEGG (19) and the *Science Signaling* Database of Cell Signaling (http://stke.sciencemag.org/cm/), we manually inspected the pathway diagrams to choose source and target proteins. Only proteins without parent nodes in the diagram were chosen as sources. Any protein that

was downstream of the sources was allowed to be a target, although preference was given to those proteins without children in the graph. We ensured that the set of sources and set of targets were disjoint. The Supplementary Methods describe the pathways used in the gold standard and Supplementary Table S3 provides a list of the sources and targets used in our evaluations.

Gold standard pathway evaluation

The results for random orientation with local search in Table 2 are based upon 20 random restarts. The MIN-SAT results are based on 20 initial orientations given by the MIN-SAT approximation algorithm. The MAX-CSP results are based on a single execution of its deterministic solver. MTO results are averaged over 20 runs, and random orientation results without local search (referred to as the oriented baseline) are averaged over 1000 runs. The unoriented edge selection algorithm is also deterministic so the results are based on a single execution. To evaluate its undirected predictions, every source–target path containing exactly six proteins was treated as a satisfied path.

Path ranking metrics

For our primary evaluation, we ranked all paths returned by the orientation algorithms by various criteria (Table 2) and calculated how many of the top 100 paths with five edges (containing exactly six proteins) are at least partially present in a gold standard pathway. Partially present means that at least four of the six proteins are found *consecutively* in both the gold standard and a satisfied path returned by the algorithm (see Supplementary Results for other path-matching criteria). Ranking paths by path weight is the most natural method, but we also explored ranking paths by the maximum, average or minimum value of different criteria that can be calculated for each edge or vertex on the path. The first such alterative ranking metric was the edge weight. The next was referred to as edge use, where the number of uses for a single edge is the number of times that edge is a member of satisfied paths. Although this metric does not directly incorporate the edge or path weights, they still influence the top-ranked paths when sorting by edge use because edge use is dependent on the network orientation, which is dependent on the path weights. The final ranking criterion was the vertex degree (the sum of the in and out degrees). Path weight was used to break ties when ranking by other metrics.

RESULTS

Experimental setup

We applied our orientation algorithms to random and real source–target pairs in protein interaction networks, where the real pairs were derived from known signaling pathways as described in the 'Materials and Methods' section. Since it is often not clear which sources are affecting which targets, in all tests we took the set of source–target pairs to be the Cartesian product of the set of sources and set of targets

Using real data, we compared our algorithms with two other methods for discovering pathways. The first is the reachability-focused MTO algorithm (28). The second is the undirected edge selection algorithm by Zhao *et al*. (26). Zhao *et al*. directly evaluated their technique against other notable undirected signaling pathway prediction algorithms (6,7,24) and showed that it compares favorably. Thus, we consider it to be representative of the general class of undirected methods. We selected these two methods for comparison because like MEO, they both can be expressed as integer programming problems and search for paths in an interaction network to explain source–target pairs. Although each of these algorithms makes different assumptions about the properties the discovered pathways should have, to our knowledge no existing work incorporates the combination of path constraints and objectives that our formulation does.

We also attempted to compare our methods with the Physical Network Models algorithm (30), however this algorithm was unable to scale to our test cases. After running for 10 days on a dual core 2.66 GHz machine with 2 GB of RAM, the algorithm was not close to termination on a test case involving 16 sources, 16 targets and paths of 5 edges in the yeast PPI network. In addition, when we ran a modified version of the algorithm that terminated after a fixed amount of time, it did not return any predicted pathways we could evaluate (see the Supplementary Methods for details).

Obtaining a protein interaction network for yeast

Our methods are applicable to any PPI data set. In addition, unlike some previous algorithms, they can utilize edge weight information. As we discussed in the 'Materials and Methods' section, weights can be derived from the source provided as evidence for the interaction (e.g. which type of experiment was used to identify the interaction) or by integrating multiple PPI databases (e.g. increasing the weight for those interactions that are supported by multiple databases). To test whether weighted edges help and if so which type of weight provides the most benefit, we downloaded all protein interactions for yeast from the BioGRID (37), IntAct (38) and MINT (39) PPI databases (Supplementary Table S4). We compared three types of networks: unweighted networks, a weighted network based on the intersection of these databases and a weighted network in which the weight depends on the type of experiment(s) that identified each edge as reported in BioGRID.

We found that both the presence of edge weights and the manner in which they are derived greatly impact the quality of the predicted pathways. In both weighted networks, our orientation algorithm recovered known signaling pathways even though none of its highest confidence predictions in the corresponding unweighted networks aligned with any gold standard paths (Supplementary Table S5). In addition, experiment type-based weights led to many more valid predictions than weights derived from database intersection so we

focused on that weighting scheme for the remainder of our evaluations. Supplementary Table S6 shows that this phenomenon is not algorithm dependent—all algorithms we examined benefit from the weighted network.

Algorithm runtimes

Scalability is an important issue for methods analyzing high-throughput data sets especially because current data are incomplete and networks for other organisms may be larger than those for yeast. We thus used the yeast network to examine the runtimes of our orientation algorithms and MTO for various combinations of maximum path length and source–target pairs. Runtimes for our algorithms include the time to enumerate paths and run local search, which composes nearly the entire runtime of our random orientation algorithm. Table 1 presents the runtimes of the algorithms for various combinations of sources, targets and maximum path length (k) using a dual core 2.66 GHz machine with 2 GB of RAM. Times for MTO and the random orientation algorithm are averaged over 50 runs per instance.

For smaller instances, all algorithms are very fast, terminating in less than a second. As expected, the randomized algorithm scales very well even for paths with six proteins ($k = 5$) making it practical for large networks with many sources and targets, and even the MIN-SAT-based algorithm executes in less than an hour on the largest instance. MTO's runtime is primarily affected by the network size and not the number of sources and targets, and it is completely independent of k.

The number of possible paths in the network grows by roughly one order of magnitude for every additional node allowed in a path. Thus, we did not measure the runtimes for cases where there are seven or more nodes in the pathway ($k \geq 6$). See Supplementary Table S7 for details.

Algorithms outperform approximation guarantees on simulated source–target pairs

To evaluate our orientation algorithms from a theoretical perspective, we examined the objective function values achieved in practice with respect to the approximation guarantees by using the real interaction network and simulated source–target pairs. We set the maximum path length to 5 (allowing for six proteins in each pathway), which is longer than the 3–4 edges preferred by previous pathway prediction algorithms (7,31). We randomly selected five unique sources and 10 unique, distinct targets for each test case leading to 50 source–target pairs per instance. We computed an upper bound on the optimal score for each instance (Supplementary Methods). This upper bound can be used to obtain a lower bound on the performance of the algorithms since the ratio of their objective value achieved to the upper bound could be even larger if the actual optimal score replaced the upper bound in the ratio. Recall that larger ratios correspond to better approximations.

Table 1. Algorithm runtimes in seconds

Sources	Targets	k	MTO	Random	MIN-SAT	MAX-CSP
4	4	4	2.0	0.1	0.1	0.3
8	8	4	2.0	0.2	0.5	0.4
16	16	4	2.0	0.4	11.0	2.3
4	4	5	2.0	0.8	3.7	3.8
8	8	5	2.0	1.8	65.7	10 802.7
16	16	5	2.0	16.2	2742.5	10 806.7

Local search was included for all three of our algorithms. See Supplementary Table S3 for the sources and targets used.

Figure 3 shows the fraction of the upper bound achieved by the algorithms on instances with simulated sources and targets (MTO and the unoriented edge selection algorithm do not use the MEO objective function and are therefore not included in this evaluation). Note that even for a fixed number of sources and targets, the number of possible paths in the network varies greatly due to network topology.

We observe that for those instances that yield fewer paths, the best approximation algorithm either achieves the optimal value or finds an orientation with value >99% of the upper bound. Even in the worst case we encountered, the best ratio achieved is >0.7, which is far better than the $\frac{k}{2^k} = \frac{5}{32} \approx 0.16$ best known theoretical guarantee of the MAX-k-CSP algorithm.

The benefit of local search varies greatly by algorithm and by the number of paths. As expected, the random orientations without local search perform much worse than the orientations after search. For the smaller instances and one larger instance with roughly 50 000 paths, the MIN-SAT algorithm obtains an excellent orientation without search. However, in the worst instance local search improves the MIN-SAT score nearly 2-fold. Of all three algorithms, MAX-CSP is the top performer without local search, and search does little to improve its orientations. This is not surprising because its underlying solver already uses an internal search-based strategy.

Interestingly, all three algorithms achieve quite similar ratios after local search across all instances we tested, to the extent that their respective points on the plot oftentimes overlap. This suggests that in practice the local search itself is more important when finding an optimal orientation than the actual algorithm used to obtain the starting point for the local search.

Evaluating algorithms using gold standard pathways

To confirm that the orientations produced by our algorithms not only achieve good approximation ratios but also produce biologically meaningful results, we compared the oriented networks with all yeast signaling pathways from KEGG and the *Science Signaling* Database of Cell Signaling. In Supplementary Methods, we describe the individual pathways in our gold standard set, the overlap between the gold standard and PPI network (Supplementary Table S2) and the sources and targets we selected (Supplementary Table S3). The gold standard network is small compared with the complete interaction network, containing 188 proteins and 286 interactions.

We considered using biochemical and metabolic pathways from *Saccharomyces* Genome Database (SGD) (http://www.yeastgenome.org/biocyc/) as well. However,

we found that these data are not appropriate for evaluating our pathway predictions because <2% of consecutive pairs of proteins in these metabolic pathways interact in the high-confidence BioGRID PPI network. Consequently, even an optimal orientation of the network cannot recover the vast majority of the SGD pathways.

Since the oriented networks can contain thousands of paths connecting the source–target pairs (Supplementary Table S7), we needed a method for identifying which paths are most likely to be biologically meaningful. We tested several such methods including path weight; min, max and average edge weight; min, max and average edge usage; and min, max and average node degree. See 'Materials and Methods' section for a description of each ranking method. Table 2 summarizes the results of this evaluation. Forty percent of the top-ranked paths discovered by the local search algorithm (following random orientation) are partially present in the gold standard when sorting by minimum edge use. Note that since the pathway databases are incomplete, the number of biologically valid pathways discovered is even larger (see 'Discussion' section).

We found that path weight, average and minimum edge weight and minimum edge use are useful criteria for ranking pathways for most algorithms whereas vertex degree is a poor ranking criterion. Of the three edge use-based metrics, the minimum edge use is consistently

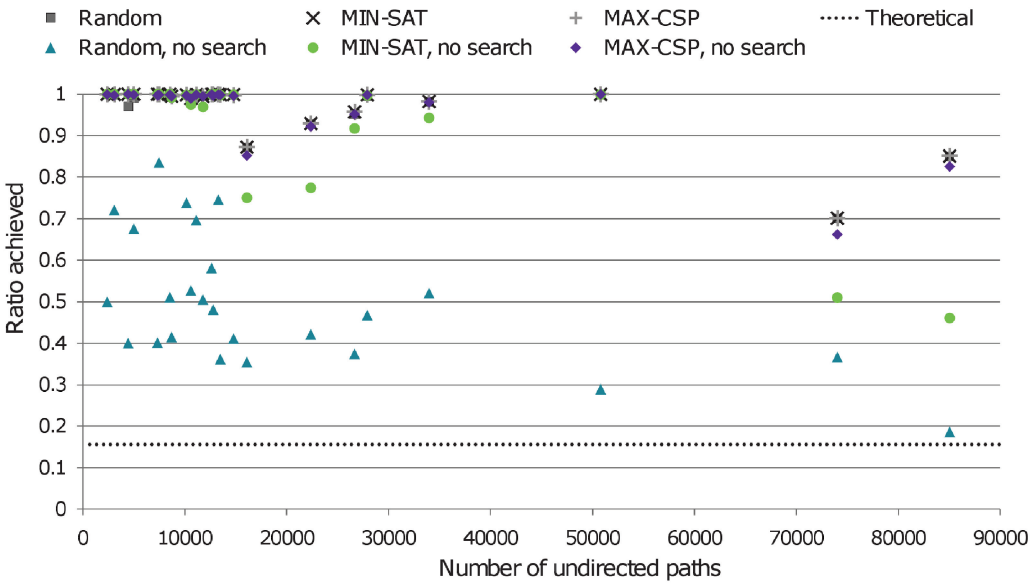

Figure 3. Fraction of the objective function upper bound achieved on instances with simulated sources and targets. After local search, all approximation algorithms perform much better than the MAX-k-CSP theoretical guarantee on instances with simulated source–target pairs and find orientations whose objective function values are virtually indistinguishable. The number of undirected paths includes all paths from a source to a target before the network is oriented. The *y*-axis plots the ratio achieved by each algorithm, which is the score of the orientation returned by the algorithm divided by the upper bound on the optimal objective function value. For each instance, there are six points (one for each algorithm with and without local search) that have the same x-coordinate, the number of undirected paths, and different y-coordinates, the ratios achieved. Instances have been ordered along the *x*-axis by the number of distinct source–target paths in the network before orientation, which is a coarse indication of the difficulty of the instance.

Table 2. Number of top-ranked predicted paths that correspond to known signaling pathways

Algorithm	Path weight	Max. edge weight	Avg. edge weight	Min. edge weight	Max. edge use	Avg. edge use	Min. edge use	Max. degree	Avg. degree	Min. degree
Random + search	37	11	36	34	0	0	**40**	10	0	0
MIN-SAT	**2**	0	**2**	1	0	0	0	1	0	0
MIN-SAT + search	33	9	32	28	0	0	**40**	10	0	0
MAX-CSP	14	7	14	**16**	0	0	**16**	3	0	0
MAX-CSP + search	7	5	6	7	0	0	**16**	3	0	0
MTO	**3.2**	**3.2**	**3.2**	**3.2**	3.0	3.0	3.0	3.0	2.8	**3.2**
Unoriented edge selection	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**	**20**
Oriented baseline	9.5	4.3	**9.8**	7.5	0.4	0.2	3.2	4.6	0	0

For each of the algorithms, all satisfied paths with exactly five edges (six proteins) were ranked by various criteria. The table shows the number of the top 100 ranked paths that partially matched gold standard pathways. Bold text denotes the highest scoring ranking metric(s) for each algorithm. Sixteen sources and 16 targets derived from the gold standard signaling pathways were used (Supplementary Table S3). Supplementary Figures S5–S8 show receiver operating characteristic curves comparing random + search, MTO and unoriented edge selection in greater detail. Supplementary Tables S8 and S9 contain additional results for smaller sets of sources and targets.

the most informative. This demonstrates that predicted pathways that contain only edges that are critical to a large number of other satisfied paths correspond to the gold standard better than pathways that contain some edges that belong to many other paths and some edges that are isolated. The average and minimum vertex degree criteria yield top-ranked paths that generally do not match known signaling pathways because they consist only of paths that contain the highest degree protein, Hek2, which is not known to be involved in our gold standard signaling pathways.

The finding that three of the four most informative ranking metrics are dependent on the edge weights provides further evidence that our edge weight assignments play an important role in identifying signaling paths. As seen in Supplementary Table S6, when the PPI network is unweighted it is not possible to sort paths using these criteria. Almost all of the remaining criteria are unable to rank predicted paths as well as the weight-based metrics, thus edge weights are crucial for selecting a high-confidence subset of paths from all predictions.

Orientation improves pathway identification

Surprisingly, although all three algorithms achieved similar fractions of the upper bound on simulated instances, the fastest method we presented, random orientation followed by local search, is able to recover a far greater number of gold standard pathways in its top-ranked paths than the CSP-based algorithm for all criteria used and performs as good as or better than MIN-SAT with search in all cases. Therefore, even though the MIN-SAT and MAX-CSP algorithms are interesting from a theoretical perspective, there is little reason to prefer them in practice over the random orientation with local search, which is much faster and can handle larger values of k. The benefits of local search are highlighted by the MIN-SAT algorithm, which performs drastically better when local search is applied. Unlike our algorithms, MTO and unoriented edge selection do not produce more biologically meaningful results after local search (Supplementary Table S10).

On average MTO finds only three pathways that partially match the gold standard no matter what ranking criteria is used. This reflects the different objective of MTO. Since it attempts to connect source–target pairs with paths of arbitrary length, very few of the resulting paths are reasonably short. In fact, in many runs we found that the MTO-oriented network did not even contain 100 source–target paths with exactly six proteins, whereas our algorithms find thousands of such paths. For the minimum edge use ranking criteria, our random orientation with search discovers 13 times as many known pathways as MTO.

Our evaluation also highlights the weaknesses of the undirected edge selection algorithm, which can only identify 20 paths in the gold standard regardless of the ranking criteria used. This is only half of what our random orientation with search discovers when ranking by minimum edge use and demonstrates that crucial network edges can be overlooked when subnetworks are selected without regard to edge orientation. In fact, the unoriented edge selection method discarded so many of these relevant edges that it found less than 100 source–target paths containing at most six proteins, which is why its evaluation was not affected by the ranking criteria used. These results strongly indicate that the unique edge orientation constraint utilized by our algorithms helps improve the quality of the pathways these methods recover.

As a control, we also calculated how many gold standard pathways could be recovered by random orientations *without* local search, which we refer to as 'Oriented baseline' in Table 2. We found that on average $<10\%$ of the top-ranked pathways were present in a gold standard pathway for any of the ranking criteria, which is much lower than the results when random orientations are followed by local search.

DISCUSSION

Modern experimental techniques provide information about proteins that directly interact with environmental factors and about the downstream effects of these interactions. In this article, we presented algorithms for reconstructing the pathways activated during such responses. Such pathways are important for understanding how signals are transmitted in the cell and often lie upstream of the regulatory networks that are activated in these responses.

These pathways primarily consist of interacting proteins so a natural way of searching for them is to use large-scale protein interaction databases. However, this is challenging for several reasons. First, protein interaction data are noisy. In addition, there are often several paths that can link sources and targets. Finally, the protein interaction data are undirected, whereas pathways are typically a chain of directed events.

To solve these problems we presented several algorithms for orienting protein interaction edges. Our algorithms rely on a number of reasonable biological assumptions including limiting the path length, using the confidence in the interaction edges and allowing for parallel pathways between sources and targets. The algorithms perform very well in practice, and notably the simplest algorithm consistently achieves better orientations than the more sophisticated and time-consuming methods. As we showed using known pathways, our orientation algorithms substantially improve upon previous methods for discovering pathways in protein interaction data sets. Furthermore, the preference for multiple parallel pathways implicit in our objective function was shown to be beneficial in our evaluation.

Analyzing pathways identified by our methods

Given the success of the methods in recovering known pathways, we asked whether the novel pathways that ranked highly according to our criteria may also be correct and represent information that is missing from current databases. We divided the pathways predicted by our random orientation with local search algorithm into

three groups and analyzed the top 20 pathways in each group using the path weight for ranking. The first (Figure 4A) contains pathways of five or six proteins that were present, in their entirety, in the signaling databases. The second (Figure 4B) are pathways predicted by our method that consist of exactly six proteins and partially overlap a known pathway. For these we asked whether the additional interactions may represent known or sensible extensions to the pathway that were not previously known or were not recorded in the databases. The third (Figure 4C) are pathways discovered by our method that do not match any known pathways in the databases. For these we asked whether they represent known pathways not in the databases or novel hypotheses that make sense biologically.

In all three figures, we merged overlapping linear paths discovered by our algorithm. Our algorithm's predictions can be easily merged in this manner to form larger signaling networks because each edge is oriented uniquely in all paths. This feature of our orientation algorithm demonstrates its advantages over undirected methods. In undirected approaches, although edges in a single predicted path have an implicit orientation because information is known to flow from source to target, these local orientations are not globally consistent across all predictions. Thus, the predictions may either be considered in isolation or merged into less informative undirected networks [e.g. the pheromone response predictions by Scott *et al.* (6)].

In Figure 4A, the path Sln1→Ypd1→Ssk1→Ssk22→Pbs2 is a component of the high-osmolarity glycerol (HOG) pathway. The filamentous growth pathway contains the cascade Msb2→Cdc42→Ste20→Ste11→Ste7. The remaining paths that begin at Rga1 or Ste50 and extend to Dig1, Dig2, Fus3, Ste7 and Ste12 are members of the pheromone signaling pathway.

Partial match pathways

For the partial match pathways (Figure 4B), we found evidence that many of their edges missing from the

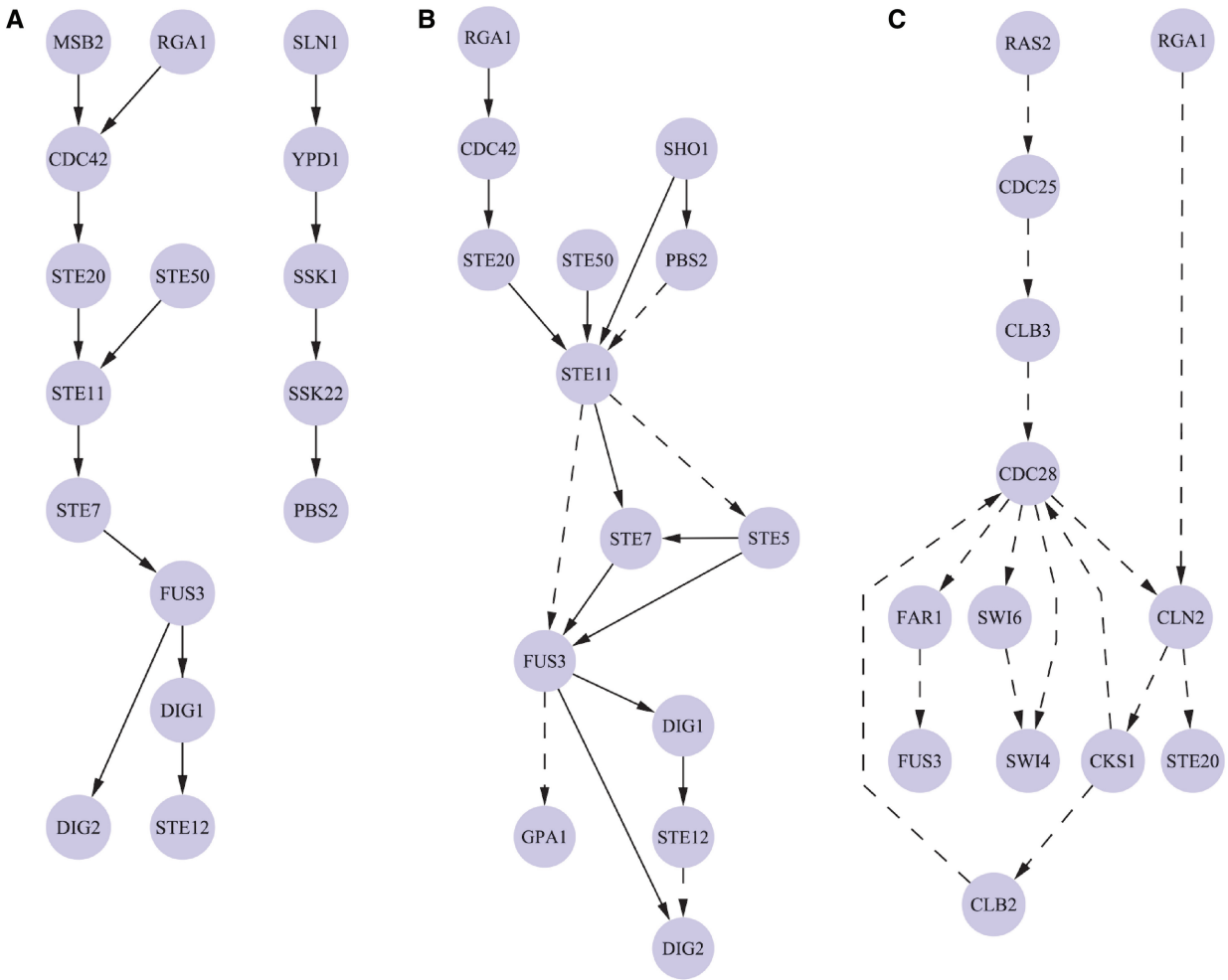

Figure 4. The top-ranked pathways discovered by the random orientation plus local search algorithm. Solid edges were present in the gold standard and dashed edges were absent or oriented in the opposite direction. (**A**) Pathways that are completely contained within a known gold standard pathway. (**B**) Pathways that partially overlap a gold standard path but contain new edges as well. (**C**) Pathways that do not have any edges in common with our set of gold standard pathways. Images were generated with Cytoscape (http://www.cytoscape.org/) and do not contain all of the top-ranked paths per category but rather a highly overlapping subset.

databases are in fact valid and that our algorithm discovered previously unknown variants of common signaling pathways. Some of these paths in the pheromone signaling pathway contain the edge Ste11→Ste5. In the evaluation summarized in Table 2, this edge was considered a mistake since in the gold standard it was oriented in the opposite direction. That orientation is based on a model in which Ste5, after being recruited by Ste4, mediates Ste20 phosphorylation of Ste11 by facilitating the complex formation via its scaffolding function. However, it was shown recently that Ste5 and Ste11 already form a tight complex in the cytosol, in fact with the highest affinity (50 nM) as compared with all other pairwise interactions between Ste5, Ste7, Ste11 and Fus3 (41). Thus, our predicted Ste11→Ste5 edge is also valid. This interaction is included in a number of paths because there is redundancy in the function of some components downstream of this edge so several of the partial matches are in fact complete matches.

Another predicted interaction that disagrees with the direction in the gold standard database is Pbs2→Ste11. However, Pbs2 is a scaffold protein that simultaneously binds the osmosensor receptor Sho1, the upstream MAPKKK Ste11 and the downstream MAPK Hog1 (42). Thus, even though Ste11 acts on Pbs2, its scaffolding function makes the edge direction ambiguous because formation of the signaling complex at Sho1 is required, and Sho1 and Pbs2 have therefore been termed 'coscaffolds'. Thus, drawing the edge in both directions is reasonable.

A particularly interesting prediction is the edge Fus3→GPA1, which was not found in the gold standard database. GPA1 is the Gα protein that is activated by pheromone stimulation of the membrane receptors which are G protein-coupled receptors. Thus, GPA1 is located close to the top input level of the pathway and is a critical step in mediating the sequence of six consecutive intracellular events leading to Ste12 activation. Recently, it was found that there is a feedback loop from Fus3 (the kinase that directly activates Ste12) to GPA1 to Ste4 (another subunit in the heterotrimeric G protein complex), which is phosphorylated by Fus3 and negatively regulates the pathway (43). Thus, the predicted Fus3→GPA1 edge is supported by this experimentally demonstrated feedback loop.

While most of the orientation results either agree with the gold standard orientation or with recent studies, we found two cases where the orientation determined by the algorithm is likely wrong. The first is the Ste11→Fus3 edge where both partners are part of the same macromolecular complex but the logic progression of the signal requires another partner in the complex. The second is the Ste12→Dig2 edge where again a third protein is involved in the communication of signal. Thus, in both cases the complex membership may confuse the algorithm by creating 'shortcuts' that are not biologically meaningful. The Supplementary Results contain further discussion of the partial match pathways.

Identified pathways that do not match any database pathway

For pathways in Figure 4C, which do not overlap with any of the pathways in the databases we used, we found many edges that are either known or raise interesting biological hypotheses. The figure depicts nine of these paths that are cell-cycle related. For example, three of the pathways originate at Rga1, a regulatory protein important for cytokinesis (end of M) and bud site formation. It is known to interact with Cln2 (44). Cks1 activates Cdc28 (45) and sends the M cyclin Clb2 to degradation (46). Cks1, Cln2 and Cdc28 form a complex (47), and Cdc28 complexes phosphorylate many proteins in the G1/S transition, including Swi4 (48) and Swi6 (49) in a regular cell cycle, Ste20 (50) in a mating response and during filamentous growth and Far1 (51) in response to alpha factor. Another cascade starts with Ras2 and Cdc25 instead of Rga1. These proteins work together and are important for the exit from a G0 state (52). Along with Cdc28 they allow the G1/S transition by increasing Cln2 levels (53). Both Clb2 and Clb3 regulate Cdc28 activity and are expressed in the G2 and late S phases, respectively (54). The Supplementary Results and Table S11 contain additional analysis of these paths.

Motivation for orienting all PPIs

In some cases it may be ideal to leave certain PPI in the network undirected. However, in practice, orienting the entire network does not affect our ability to correctly discover signaling pathways due to the nature of the interaction data sets we use. In general, when a complex interacts with some external protein, all (or most) members of the complex are shown as interacting with that protein in PPI databases. This is a consequence of the high-throughput studies (e.g. pull-down assays) that often cannot distinguish between direct and indirect interactions. Thus, any orientation of the internal edges between complex members is appropriate because external proteins that interact with the complex are connected to both endpoints of the internal edges.

Several of our cell-cycle paths in Figure 4C demonstrate how our orientation of edges in a complex can correspond to the biological truth. Clb2 and Clb3 each form a complex with Cdc28, yet orienting the edges Clb2→Cdc28 and Clb3→Cdc28 is justified because these two proteins are also reported to activate Cdc28. In addition, the edges Cdc28→Cln2→Ste20 represent the Cdc28–Cln2 complex-mediating Ste20 even though this predicted path does not contain a direct Cdc28→Ste20 edge.

In fact, allowing our algorithms to leave certain edges undirected is not a viable option given our problem formulation. Orienting an undirected edge can only reduce the number of satisfied paths in the network and correspondingly lower the objective function. Thus, the optimal solution for all instances would be to not orient any edges. One reasonable way to overcome this would be to penalize solutions for including undirected edges, but setting the parameter that controls the trade-off between the objective function's penalty term and path weight term would

require much more training data (i.e. known signaling pathways) than is currently available.

Extensions to other species

While the work so far has focused on yeast, our methods can be directly applied to other species including humans. Large high-throughput interaction data sets for human are rapidly becoming available (14). Interactions of human proteins with proteins of infecting pathogens have also been cataloged recently (15). Other studies provide information about downstream genes (17,18). Combining these data sets provides both a network and a set of sources and targets that can be used by our method to infer pathways that are activated following infection. We are also interested in linking our orientation algorithms to methods that attempt to reconstruct regulatory networks (55). The combination of these two techniques would provide a connection between the signaling networks and the regulatory networks that are activated in response to environmental cues.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to express our gratitude to Itamar Simon for his assistance in analyzing the cell-cycle predictions.

FUNDING

This work was supported by National Institutes of Health (grant numbers 1RO1 GM085022, NO1 AI-5001 to Z.B.J); National Science Foundation (CAREER award 0448453 to Z.B.J); National Science Foundation Graduate Research Fellowship (to A.G.). Funding for open access charge: National Institutes of Health grant NO1 AI-5001.

Conflict of interest statement. None declared.

REFERENCES

1. Bar-Joseph,Z., Gerber,G.K., Lee,T.I., Rinaldi,N.J., Yoo,J.Y., Robert,F., Gordon,D.B., Fraenkel,E., Jaakkola,T.S., Young,R.A. *et al*. (2003) Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol*., **21**, 1337–1342.
2. Segal,E., Shapira,M., Regev,A., Pe'er,D., Botstein,D., Koller,D. and Friedman,N. (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet*., **34**, 166–176.
3. Margolin,A.A., Nemenman,I., Basso,K., Wiggins,C., Stolovitzky,G., Dalla Favera,R. and Califano,A. (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **7**, S7.
4. Covert,M.W., Knight,E.M., Reed,J.L., Herrgard,M.J. and Palsson,B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, **429**, 92–96.
5. Fischer,E. and Sauer,U. (2005) Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. *Nat. Genet*., **37**, 636–640.
6. Scott,J., Ideker,T., Karp,R.M. and Sharan,R. (2006) Efficient algorithms for detecting signaling pathways in protein interaction networks. *J. Comput. Biol*., **13**, 133–144.
7. Bebek,G. and Yang,J. (2007) PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. *BMC Bioinformatics*, **8**, 335.
8. Harbison,C.T., Gordon,D.B., Lee,T.I., Rinaldi,N.J., Macisaac,K.D., Danford,T.W., Hannett,N.M., Tagne,J., Reynolds,D.B., Yoo,J. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.
9. Mikkelsen,T.S., Ku,M., Jaffe,D.B., Issac,B., Lieberman,E., Giannoukos,G., Alvarez,P., Brockman,W., Kim,T., Koche,R.P. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
10. Lewis,B.P., Burge,C.B. and Bartel,D.P. (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are MicroRNA targets. *Cell*, **120**, 15–20.
11. Xie,X., Lu,J., Kulbokas,E.J., Golub,T.R., Mootha,V., Lindblad-Toh,K., Lander,E.S. and Kellis,M. (2005) Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature*, **434**, 338–345.
12. Cox,S.J., Levanon,S.S., Bennett,G.N. and San,K. (2005) Genetically constrained metabolic flux analysis. *Metab. Eng*., **7**, 445–456.
13. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al*. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.
14. Ewing,R.M., Chu,P., Elisma,F., Li,H., Taylor,P., Climie,S., McBroom-Cerajewski,L., Robinson,M.D., O'Connor,L., Li,M. *et al*. (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol*., **3**, Article no. 89.
15. Fu,W., Sanders-Beer,B.E., Katz,K.S., Maglott,D.R., Pruitt,K.D. and Ptak,R.G. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. *Nucleic Acids Res*., **37**, D417–D422.
16. Corbeil,J., Sheeter,D., Genini,D., Rought,S., Leoni,L., Du,P., Ferguson,M., Masys,D.R., Welsh,J.B., Fink,J.L. *et al*. (2001) Temporal gene regulation during HIV-1 infection of human CD4+ T cells. *Genome Res*., **11**, 1198–1204.
17. Brass,A.L., Dykxhoorn,D.M., Benita,Y., Yan,N., Engelman,A., Xavier,R.J., Lieberman,J. and Elledge,S.J. (2008) Identification of host proteins required for HIV infection through a functional genomic screen. *Science*, **319**, 921–926.
18. König,R., Zhou,Y., Elleder,D., Diamond,T.L., Bonamy,G.M., Irelan,J.T., Chiang,C., Tu,B.P., Jesus,P.D.D., Lilley,C.E. *et al*. (2008) Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. *Cell*, **135**, 49–60.
19. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.
20. von Mering,C., Krause,R., Snel,B., Cornell,M., Oliver,S.G., Fields,S. and Bork,P. (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.
21. Cobb,M.H. and Goldsmith,E.J. (1995) How MAP kinases are regulated. *J. Biol. Chem*., **270**, 14843–14846.
22. Schlaepfer,D.D., Jones,K.C. and Hunter,T. (1998) Multiple Grb2-mediated integrin-stimulated signaling pathways to ERK2/mitogen-activated protein kinase: summation of both c-Src- and focal adhesion kinase-initiated tyrosine phosphorylation events. *Mol. Cell. Biol*., **18**, 2571–2585.
23. Piloto,O., Wright,M., Brown,P., Kim,K., Levis,M. and Small,D. (2007) Prolonged exposure to FLT3 inhibitors leads to resistance via activation of parallel signaling pathways. *Blood*, **109**, 1643–1652.
24. Steffen,M., Petti,A., Aach,J., D'haeseleer,P. and Church,G. (2002) Automated modelling of signal transduction networks. *BMC Bioinformatics*, **3**, 34.
25. Lu,S., Zhang,F., Chen,J. and Sze,S. (2007) Finding pathway structures in protein interaction networks. *Algorithmica*, **48**, 363–374.

26. Zhao,X., Wang,R., Chen,L. and Aihara,K. (2008) Uncovering signal transduction networks from high-throughput data by integer linear programming. *Nucleic Acids Res.*, **36**, e48.
27. Yosef,N., Ungar,L., Zalckvar,E., Kimchi,A., Kupiec,M., Ruppin,E. and Sharan,R. (2009) Toward accurate reconstruction of functional protein networks. *Mol. Syst. Biol.*, **5**, Article no. 248.
28. Medvedovsky,A., Bafna,V., Zwick,U. and Sharan,R. (2008) An algorithm for orienting graphs based on cause-effect pairs and its applications to orienting protein networks. In *Proceedings of the 8th international workshop on Algorithms in Bioinformatics*. Karlsruhe, Germany, pp. 222–232.
29. Liu,W., Li,D., Wang,J., Xie,H., Zhu,Y. and He,F. (2009) Proteome-wide prediction of signal flow direction in protein interaction networks based on interacting domains. *Mol. Cell Proteomics*, **8**, 2063–2070.
30. Yeang,C., Ideker,T. and Jaakkola,T. (2004) Physical Network Models. *J. Comput. Biol.*, **11**, 243–262.
31. Ourfali,O., Shlomi,T., Ideker,T., Ruppin,E. and Sharan,R. (2007) SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. *Bioinformatics*, **23**, i359–i366.
32. Halperin,E. and Zwick,U. (2001) Combinatorial approximation algorithms for the maximum directed cut problem. In *Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. Washington, D.C., USA, pp. 1–7.
33. Håstad,J. (2001) Some optimal inapproximability results. *JACM*, **48**, 798–859.
34. Bertsimas,D., Teo,C. and Vohra,R. (1999) On dependent randomized rounding algorithms. *Oper. Res. Lett.*, **24**, 105–114.
35. Kohli,R., Krishnamurti,R. and Mirchandani,P. (1994) The minimum satisfiability problem. *SIAM J. Discret. Math.*, **7**, 275–283.
36. Charikar,M., Makarychev,K. and Makarychev,Y. (2009) Near-optimal algorithms for maximum constraint satisfaction problems. *ACM Trans. Alg.*, **5**, 1–14.
37. Stark,C., Breitkreutz,B., Reguly,T., Boucher,L., Breitkreutz,A. and Tyers,M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.*, **34**, D535–D539.
38. Aranda,B., Achuthan,P., Alam-Faruque,Y., Armean,I., Bridge,A., Derow,C., Feuermann,M., Ghanbarian,A.T., Kerrien,S., Khadake,J. *et al.* (2009) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38(Suppl 1)**, D525–D531.
39. Chatr-aryamontri,A., Ceol,A., Palazzi,L.M., Nardelli,G., Schneider,M.V., Castagnoli,L. and Cesareni,G. (2007) MINT: the Molecular INTeraction database. *Nucleic Acids Res.*, **35**, D572–D574.
40. Deng,M., Sun,F. and Chen,T. (2003) Assessment of the reliability of protein-protein interactions and protein function prediction. In *Proceedings of the 8th Pacific Symposium on Biocomputing*. Kauai, Hawaii, pp. 140–151.
41. Maeder,C.I., Hink,M.A., Kinkhabwala,A., Mayr,R., Bastiaens,P.I.H. and Knop,M. (2007) Spatial regulation of Fus3 MAP kinase activity through a reaction-diffusion mechanism in yeast pheromone signalling. *Nat. Cell Biol.*, **9**, 1319–1326.
42. Zarrinpar,A., Bhattacharyya,R.P., Nittler,M. and Lim,W.A. (2004) Sho1 and Pbs2 act as coscaffolds linking components in the yeast high osmolarity MAP kinase pathway. *Mol. Cell*, **14**, 825–832.
43. Metodiev,M.V., Matheos,D., Rose,M.D. and Stone,D.E. (2002) Regulation of MAPK function by direct interaction with the mating-specific galpha in yeast. *Science*, **296**, 1483–1486.
44. Archambault,V., Chang,E.J., Drapkin,B.J., Cross,F.R., Chait,B.T. and Rout,M.P. (2004) Targeted proteomic study of the cyclin-Cdk module. *Mol. Cell*, **14**, 699–711.
45. Tang,Y. and Reed,S.I. (1993) The Cdk-associated protein Cks1 functions both in G1 and G2 in Saccharomyces cerevisiae. *Genes Dev.*, **7**, 822–832.
46. Kaiser,P., Moncollin,V., Clarke,D.J., Watson,M.H., Bertolaet,B.L., Reed,S.I. and Bailly,E. (1999) Cyclin-dependent kinase and Cks/Suc1 interact with the proteasome in yeast to control proteolysis of M-phase targets. *Genes Dev.*, **13**, 1190–1202.
47. Gavin,A., Bösche,M., Krause,R., Grandi,P., Marzioch,M., Bauer,A., Schultz,J., Rick,J.M., Michon,A., Cruciat,C. *et al.* (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.
48. Amon,A., Tyers,M., Futcher,B. and Nasmyth,K. (1993) Mechanisms that help the yeast cell cycle clock tick: G2 cyclins transcriptionally activate G2 cyclins and repress G1 cyclins. *Cell*, **74**, 993–1007.
49. Geymonat,M., Spanos,A., Wells,G.P., Smerdon,S.J. and Sedgwick,S.G. (2004) Clb6/Cdc28 and Cdc14 regulate phosphorylation status and cellular localization of Swi6. *Mol. Cell. Biol.*, **24**, 2277–2285.
50. Wu,C., Leeuw,T., Leberer,E., Thomas,D.Y. and Whiteway,M. (1998) Cell cycle- and Cln2p-Cdc28p-dependent phosphorylation of the yeast Ste20p protein kinase. *J. Biol. Chem.*, **273**, 28107–28115.
51. Blondel,M., Galan,J.M., Chi,Y., Lafourcade,C., Longaretti,C., Deshaies,R.J. and Peter,M. (2000) Nuclear-specific degradation of Far1 is controlled by the localization of the F-box protein Cdc4. *EMBO J.*, **19**, 6085–6097.
52. Folch-Mallol,J.L., Martinez,L.M., Casas,S.J., Yang,R., Martinez-Anaya,C., Lopez,L., Hernandez,A. and Nieto-Sotelo,J. (2004) New roles for CDC25 in growth control, galactose regulation and cellular differentiation in Saccharomyces cerevisiae. *Microbiology*, **150**, 2865–2879.
53. Dirick,L. and Nasmyth,K. (1991) Positive feedback in the activation of G1 cyclins in yeast. *Nature*, **351**, 754–757.
54. Hu,F., Gan,Y. and Aparicio,O.M. (2008) Identification of Clb2 residues required for Swe1 regulation of Clb2-Cdc28 in Saccharomyces cerevisiae. *Genetics*, **179**, 863–874.
55. Ernst,J., Vainas,O., Harbison,C.T., Simon,I. and Bar-Joseph,Z. (2007) Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.*, **3**, Article no. 74.