A Prize-Collecting Steiner Tree Approach for Transduction Network Inference

Marc Bailly-Bechet[1,2], Alfredo Braunstein[1], and Riccardo Zecchina[1]

[1] Microsoft TCI Research, Dipartimento di Fisica,
Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy
[2] Université de Lyon, F-69000, Lyon, CNRS, UMR5558,
Laboratoire de Biométrie et Biologie Evolutive, F-69622, Villeurbanne, France
mbailly@biomserv.univ-lyon1.fr, alfredo.braunstein@polito.it,
riccardo.zecchina@polito.it

Abstract. Into the cell, information from the environment is mainly propagated via signaling pathways which form a transduction network. Here we propose a new algorithm to infer transduction networks from heterogeneous data, using both the protein interaction network and expression datasets. We formulate the inference problem as an optimization task, and develop a message-passing, probabilistic and distributed formalism to solve it. We apply our algorithm to the pheromone response in the baker's yeast *S. cerevisiae*. We are able to find the backbone of the known structure of the MAPK cascade of pheromone response, validating our algorithm. More importantly, we make biological predictions about some proteins whose role could be at the interface between pheromone response and other cellular functions.

1 Introduction

Living cells need to react to a wide spectrum of changes –physical, chemical or biological– in their environment [1]. Conversely the cell reactions span from the activation of small-scale processes, *e.g.* synthesis of precise molecular components or excretion of others, to complex changes in the global cellular state, such as the diauxic shift or pheromone response and mating [2] in yeast. In order for the cell to survive, these changes must be tightly regulated. One type of regulation occurs through signaling cascades, which represents how the information propagates inside a cell, from receptor proteins to transcription factors and other effector proteins. At the molecular level, this information transits by activation or inactivation of specific signaling proteins. Activation mechanisms include a variety of protein-protein interactions such as conformation changes, or dimerization [3]; one of the most studied is the well-known phosphorylation-dephosphorylation system provided by kinases [4]. Known signaling cascades show desirable properties, from a system point of view: they act as low-pass filters, ensuring an adequate cell response only when external stimuli are above the molecular noise level [5], but they also provide signal amplification [6]. Recently, it was also shown that information could travel in both directions on signaling

P. Degano and R. Gorrieri (Eds.): CMSB 2009, LNBI 5688, pp. 83–95, 2009.

cascades, due to chemical equilibrium shifts in the cascades[7]. These properties can be used by the cell to tune the signal propagation and therefore the response to the environment.

The intersection of the signaling pathways forms the transduction network, whose nodes are proteins and whose edges represent protein interactions transmitting information. Due to the many interconnections between different signaling cascades in the transduction network, a precise regulation of the cross-talk between different pathways is necessary. One way to ensure pathway specificity in answer to a given signal is the usage of scaffold proteins which will specifically bind to other members of a given signaling cascade, increasing specificity of the response [8,9]. On the other hand, one way to diffuse signal to many pathways is to root them all to the same activator protein. The complexity of these cross-interactions has allowed evolution to shape these pathways so as to be very efficient in sensing and adjusting to the environment, but makes them very difficult to study independently and even to identify precisely. In this work we tackle the issue of transduction network inference from proteomics and transcriptomics data. Phosphoproteomics works have been led to reconstruct these cascades, but are still very expensive and time consuming. At the algorithmic level, this problem has been widely studied, mainly by inference of linear cascades. In this context Scott *et al.* [10] developed an algorithm based on color-coding to infer linear subparts of the transduction network, and found with good accuracy the known MAPK kinases cascades. White *et al.* [11] made a step forward by looking for transduction networks as a superposition of shortest paths on the protein interaction network (PIN). The focus of their method is to unbias the solution tree from the high connectivity bias, as often hubs of the PIN tend to be over-represented in the inferred networks, as a consequence of their high inbetweenness. Other works about the inference of transduction network include [12], who introduced a Steiner tree formalism to recover this network based on expression data and an existing PIN. This formalism states that the transduction network is a subtree of the global protein interaction network which contains all proteins of a given subset, named terminals, defined by the user. This subset is composed of proteins known to be part of the signaling network, or selected via another criterion such as expression level. The problem is then to reconstruct such a tree, respecting also other combinatorial constraints such as *e.g* small tree size or fixed tree depth. This last approach was recently developed by [13], who used an integer linear programming relaxation to find subnetworks involved in signal transduction, improving the algorithmic performance.

The previously cited approaches have been effective at finding already known signaling cascades, but made few predictions, mainly because of the time constraints of the available techniques in the field of combinatorial inference problems. Indeed the Steiner tree problem [14] is NP-hard and classical algorithms allowing to solve it at the probabilistic level are slow. Here we provide a new method from the class of message-passing algorithms to infer a Steiner tree from a weighted graph, which directly applies to infer transduction networks from a PIN and expression data. From a numerical point of view, message-passing algorithms are

probabilistic and distributed, allowing for a very fast resolution of inference problems [15], even for large networks. Moreover, our algorithm does not need *a priori* selected terminals (i.e proteins of interest), and compute the transduction network as a whole, instead of a sum of linear subparts, as was done in previous works. This results in a high exploratory power of combinatorial effects that could uncover biologically meaningful cross-talk. We apply it to the pheromone response in *S. cerevisiae*; results show that we are able to reconstruct accurately known pathways, to infer how the signal propagates in other signaling cascades of the cell, and to make functional predictions about a new group of genes implied in the pheromone response.

2 Material and Methods

The rationale of our model is that the transduction network is a subtree of the PIN, which should be composed with links of the PIN corresponding to real protein interactions, and proteins being of biological relevance for the biological process under study. Indeed, protein-protein interactions detected in proteomic assays contain a high fraction of false positives [16], creating the need to take into account in our model the statistical confidence we have for each link of the PIN. As proteomics data are still scarce, whether expression data are nowadays available in huge quantities, we hypothesized, as was previously done by [13,17,18,10,12], that genes being differentially expressed during the activation of the signaling pathway encode proteins being necessary for the signaling response itself, and employed expression data to measure the relative importance of each protein in a given environmental context. Therefore we could model the transduction network inference as an optimization problem, given weights for every edge of the PIN, to represent the propensity of the edge to be a false positive, and prizes for the nodes, proportional to the level of differential expression of the corresponding genes in the expression data relative to the phenomenon under study.

2.1 Inference of the Transduction Network

In general terms, we are interested in finding a "minimal" sub-network that is connected to a given protein node, known as the root. We will model this problem as a Prize-Collecting Steiner Tree on Graphs problem (see e.g. [19,20]). Given a network $G = (V, E)$ with positive (real) weights $\{w_l : l \in E\}$ on edges and $\{w_n : n \in V\}$ on vertices, we are interested in finding the connected sub-network that minimizes the following quantity:

$$C = \sum_{links} w_l - \lambda \sum_{nodes} w_n \tag{1}$$

It is easy to see that such network must be a tree (links closing cycles can be removed, lowering C). λ is a parameter regulating the balance between optimization of the two terms of the sum.

This problem is known to be NP-Hard, implying that is unlikely that an algorithm that can efficiently solve any instance of the problem exists. To solve it we will use a small variation of an extremely efficient heuristics based on belief propagation developed on [14] that is known to be exact on many classes of random networks [14,21]. The algorithm iterates the following set of equations for the quantities $\{\psi_{ij}\}_{(ij)\in E}$ (called "messages") to a fixed point:

$$\psi_{ji}^{t+1}(d_j, p_j) = -c_{jp_j} + \sum_{(kj)\in E\setminus(ij)} \max_{f(d_k,p_k,d_j,p_j)\neq 0} \psi_{kj}^{t}(d_k, p_k) \tag{2}$$

where $d_i \in \mathcal{D} = \{0, ..., D-1\}$, $p_i \in V(i)\cup\{\emptyset\}$, $\psi_{ji} : \mathcal{D}\times V(i)\cup\{\emptyset\} \to \mathbb{R}$ and f_{ij} is a characteristic function that ensures the condition $p_i = j \Rightarrow p_j \neq \emptyset, d_j = d_i - 1$ defined as follows:

$$\begin{aligned} f_{ij} &= g_{ij}g_{ji} \\ g_{ij} &= (1 - \delta_{p_j,i}(1 - \delta_{d_i,d_j-1}))(1 - \delta_{p_j,i}\delta_{p_i,\emptyset}) \end{aligned}$$

On a fixed point, the following quantities ("field") are computed:

$$\psi_j(d_j, p_j) = \sum_{(kj)\in E} \max_{f(d_k,p_k,d_j,p_j)\neq 0} \psi_{kj}^{t}(d_k, p_k)$$

Then a tree T^* is built from the parenthood relations defined as follows: define $d_j^*, p_j^* = \arg\max_{d_j,p_j} \psi_j(d_j, p_j)$. Then if $p_j^* \neq \emptyset$, define the parent of j as p_j. Otherwise, j does not belong to T^*. With a minimal non-degeneracy assumption on the initial fields, it is relatively straightforward to verify that with variables d_j^*, p_j^*, $f_{ij} = 1\ \forall(ij) \in E$ and this implies that T^* is indeed a tree. It can be proved in some limit cases that the algorithm is optimal, and verified experimentally that it generally gives an excellent approximation to the optimal. For more details see [21].

2.2 Data Source and Definition of the Weights

The yeast protein interaction network (PIN) was built by combining data from two databases : DIP [22] and MIPS [23]. The combined network has 5217 nodes and 22637 edges. To define their weights, edges were divided in two categories: a high confidence one, containing links extracted from small-scale experiments or found many times; and a low confidence one, containing links found only once in a large-scale experiment. We defined the two corresponding weights so as to maximize the correlation of our weight set and the one of [24], giving a weight $w_l = 1$ for high confidence edges (24.9% of the PIN) and a weight $w_l = 1.74$ for low confidence edges. The choice of this weight set as a reference is based on the observation that it is one of the most reliable [25], and does not derive the weights from expression data.

We analyzed 56 expression datasets from [26]. We computed node prizes for each dataset in a classical way by taking $w_n = -\log(p_n)$, where p_n is the p-value of differential expression of node n in the corresponding microarray. Though, a high prize was attributed to genes having a significant p-value in the expression data.

The 56 datasets were analyzed independently with values of λ ranging from 0.05 to 0.9. The chosen root was the receptor protein STE2 in datasets comparing cells submitted or not to pheromone α action. In datasets with an artificially overexpressed gene under GAL4 promoter control, this gene was chosen as the root. If the strain used contained deletions, the corresponding genes were removed from the PIN prior to inference. In datasets comparing deleted strains to wild type strains without exposition to pheromone, deleted genes were selected as roots.

2.3 Statistical Analyses

Functional homogeneity of the trees inferred for each expression microarray and each value of λ was assessed by comparing the number of GO Slim annotations [27] shared by interacting proteins in the inferred Steiner trees, and random trees with same root and size, with edges probabilistically weighted as in the real data or not. Random trees were generated 50 times and results were averaged.

Steiner proteins were defined as proteins present in the Steiner tree with $w_n < \frac{1}{\lambda}$: such proteins have a local cost to be added in the tree, which has to be compensated. Enrichment of the inferred trees in proteins of interest was estimated by comparison with random trees generated with permuted expression data, for $\lambda = 0.2$ (30 iterations).

3 Results

Our algorithm infers an organism transduction network, using as a support the PIN and expression data to find a Steiner tree maximizing the level of differential expression of its nodes (genes) *and* built preferentially with edges of high confidence. The free parameter λ (see Mat. Meth.) regulates the balance between optimization on the edges and on the nodes, and therefore regulates the tree size. For each microarray given as input, the Steiner tree found is a representation of the transduction network activated in the corresponding condition. The Steiner trees representing transduction networks were inferred in 56 expression datasets from a study about pheromone response [26], with 7 different values of λ. A statistical description of the trees found is provided in Table 1. As expected, both the frequency of high-cost links selected and the average tree size increase with λ.

As an integrity check we analyzed the correlation between the tree size and the average prize w_n of the nodes in the datasets, which is a direct measure of the numbers of genes differentially expressed on the microarray (Fig 1). As expected, the average tree size increases both with λ and average node prize; indeed, this second dependence even seems linear, a property that could be interesting to detect anomalies in the inferred trees.

An averaged representation of the trees found for $\lambda = 0.2$ is given in Fig 2. Proteins usually found as members of the pheromone response pathway are present, such as FUS3, GPA1 or SST2; some missing intermediates appear for higher

Table 1. Statistical properties of the trees inferred. One can see the evolution of the average tree properties with increasing values of the parameter λ, notably the increase in average tree size and decrease in the fraction of Steiner proteins found. The global fraction of high-cost links in the PIN is 75.1%, notably higher than the fraction present in the inferred trees.

λ	Tree size (# prot)	Fraction of high-cost edges	Fraction of Steiner proteins
0.05	1.5±1.1	0.034	0.471
0.1	9.7±15.8	0.058	0.295
0.2	85.0±123.1	0.273	0.248
0.3	173.2±222.8	0.345	0.233
0.5	337.3±363.7	0.389	0.213
0.7	478.5±450.6	0.404	0.198
0.9	612.7±516.2	0.407	0.188

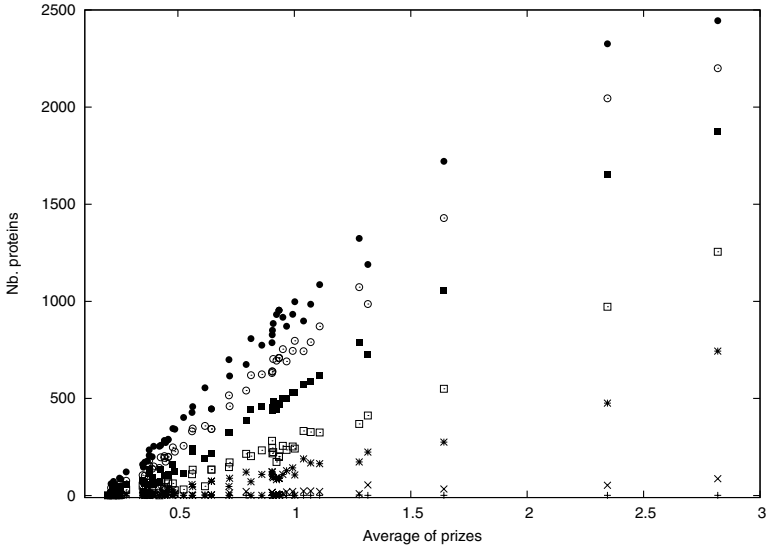

Fig. 1. This figure shows the strong correlation between the average node prize in each dataset (x-axis) and the number of proteins found in the inferred tree (y-axis). Different types of points correspond to different values of λ: vertical crosses $\lambda = 0.05$, diagonal crosses $\lambda = 0.1$, stars $\lambda = 0.2$, empty boxes $\lambda = 0.3$, filled boxes $\lambda = 0.5$, empty circles $\lambda = 0.7$, filled circles $\lambda = 0.9$. Note the linearity of the relation for each given value of λ.

values of λ. To assess the quality of the trees found, we computed the average number of shared GO Slim annotations between neighbors, and compared it to random trees, either weighted or not (Fig 3). the average number of common annotations is higher for low values of λ (Fig 3), showing a clear functional enrichment of the Steiner trees. Topology and PIN weights only account for a part

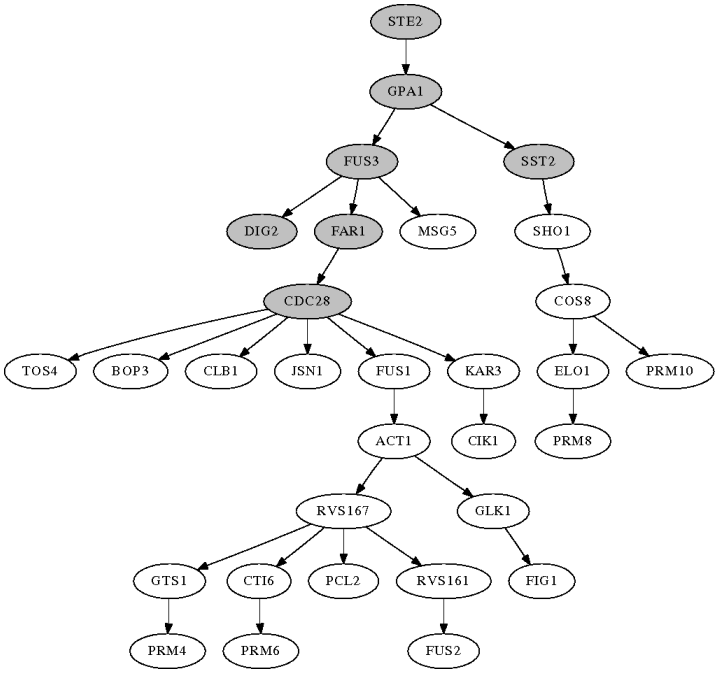

Fig. 2. This tree is formed by the superposition of all 56 Steiner trees found for $\lambda = 0.2$. Link intensity is proportional to the number of times the link was found, either in one sense or another; in case of links inferred in different directions, the orientation represented is the one mostly found. Links found in less than 30% of the trees are not shown for clarity. Grey nodes represent the proteins involved in the known pheromone pathway.

of this enrichment, shown by the simulations with random weighted trees, the rest being a combined consequence of both the proteins and the paths selected in the tree, which can thus be considered to represent biologically meaningful transduction networks. For high values of λ, this enrichment is not visible, and we will therefore focus on results at low λ.

Previous to analyses, a technical bias has to be accounted for. Due to differences in in-betweenness – or connectivity, see [11] –, certain proteins occur more or less often in the Steiner trees. Indeed, proteins with a high in-betweenness in the PIN tend to be frequently present in the Steiner trees, even if they are attributed a low prize. From a probabilistic point of view, including these proteins in the Steiner tree allows to gain access to proteins with a positive contribution to the global tree cost, enough to compensate for their own relative costs. One can see this trend in Fig 4: proteins selected more often have a high in-betweenness. Still, this correlation is only partial ($R^2 = 0.37$), and let ample space for other factors to explain presence of certain proteins in the final trees.

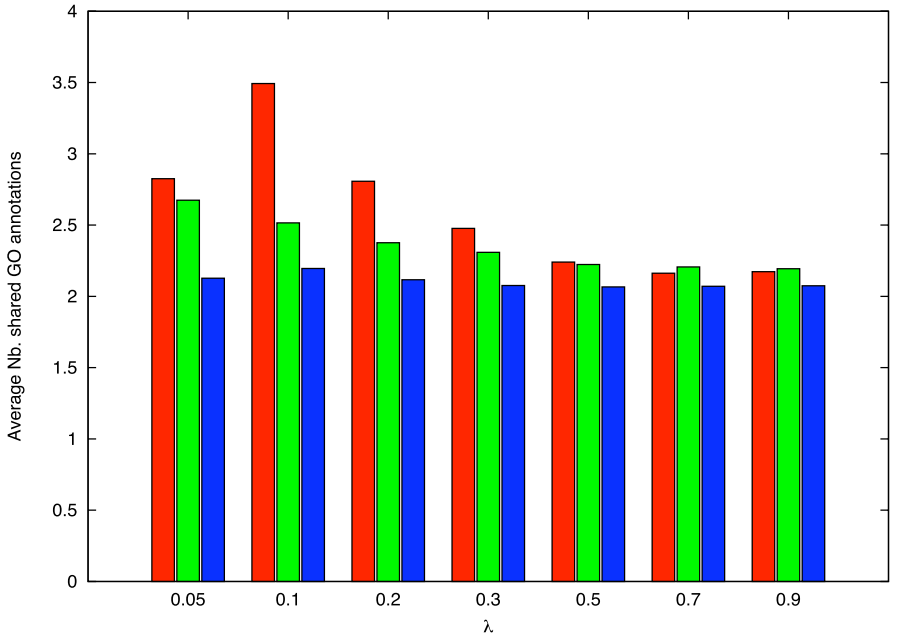

Fig. 3. Histogram of the average number of shared GO Slim annotations per link, on average on all 56 inferred trees at a given value of λ. First histogram represents the real values, second random weighted trees, and third random unweighted trees (See Mat. Meth.). Note the high differences for low values of λ.

Table 2. Properties of the 11 putative Steiner proteins found the most frequently in $\lambda = 0.2$ datasets. k stands for connectivity and "In-bet." for in-betweenness.

Gene name	Protein name	Frac. found (real data)	Frac. found (random data)	Ratio	k	In-bet. (x10^5)
YBR160W	CDC28	0.66	0.57	1.2	227	15
YDR388W	RVS167	0.52	0.31	1.7	121	4.9
YHL048W	COS8	0.45	0.09	5.0	46	0.63
YFL039C	ACT1	0.45	0.17	2.7	47	1.1
YER118C	SHO1	0.43	0.06	7.0	42	1.5
YJR091C	JSN1	0.43	0.44	1.0	293	25
YCL040W	GLK1	0.43	0.003	144	6	0.25
YBR159W	IFA38	0.41	0.09	4.8	101	1.9
YGL181W	GTS1	0.41	0.09	4.5	43	1.5
YPL181W	CTI6	0.41	0.02	20.3	26	0.26
YMR059W	SEN15	0.34	0.007	47.5	57	1.4

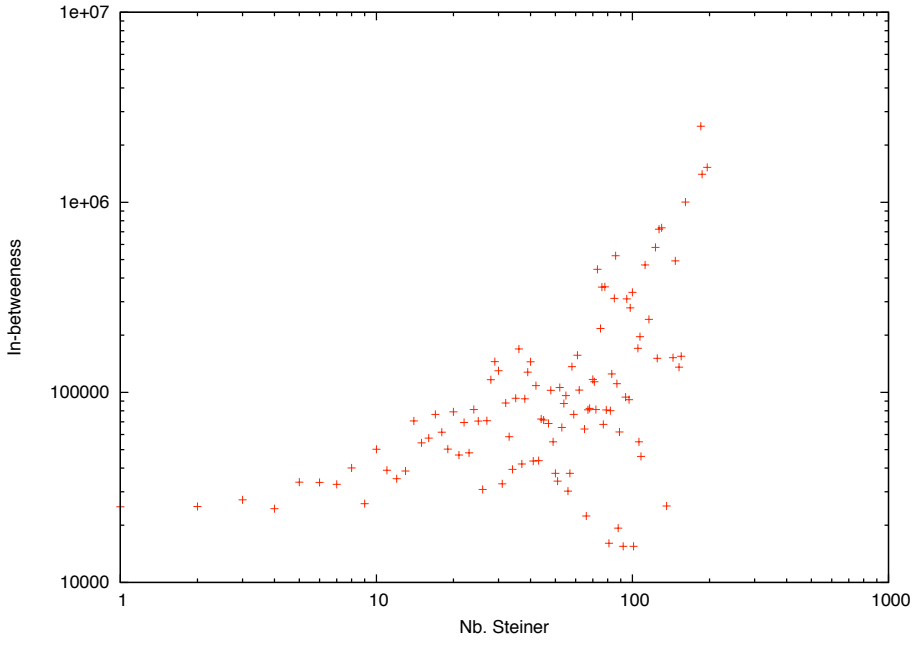

Fig. 4. Number of times a protein appears as Steiner (in all trees inferred) *vs* in-betweenness of the protein in the PIN. Note the correlation between them.

An interesting feature of our formalism is the definition of *Steiner proteins*, *i.e* proteins present in the Steiner trees without being highly differentially expressed. These proteins form bridges between groups of proteins with a positive contribution to the optimization criterion, and they could not be discovered by analyzing only the expression levels in the microarray, as they do not differ significantly from the background. Is is the combination of information from the PIN structure and expression data that unveil them. In the following analyses we focus on the Steiner proteins that appear at low values of λ, *i.e* those less distinguishable from the background expression.

In order to quantitatively measure the significance of Steiner proteins, we did a bootstrap experiment by generating Steiner trees for random expression data, obtained by permutations of the real datasets. Then, we compared the frequency of occurrence of proteins found very often as Steiner proteins in the real data to their frequency of occurrence in this randomized data; the ratio of these quantities was then used to assess the biological significance of the putative Steiner proteins (see Table 2), a high ratio meaning biologically meaningful inference and a low one typical of an artifact due to PIN topology and high in-betweenness bias.

Proteins with such a high ratio have an average in-betweenness (see Table 2). Using this table, one can easily see that the proteins CDC28, JSN1 and RVS167 should not be accounted as Steiner proteins, based on the ratio value and their very high in-betweenness. To get better insights about these proteins and their

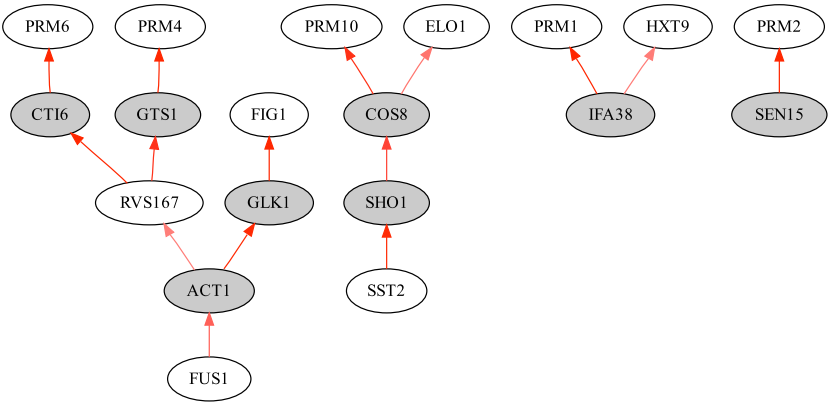

Fig. 5. Main first-order interactions of the proteins identified as *Steiner proteins*, $\lambda = 0.2$. Steiner proteins are shown in grey. Link intensity is proportional to the number of times the link is found when protein is considered as Steiner. Only links found in more than 50% runs are shown.

implication in the pheromone response as Steiner proteins, we looked which partners they interact with. The partners found in more than half of the trees are represented in Fig 5. Many interactants are membrane proteins, in particular PRM proteins [28]. One interesting feature is that the Steiner proteins COS8 and SHO1 seem to be strongly interacting, as ACT1 and GLK1 do either. We detail these two cases in the following paragraphs.

GLK1 give access to the FIG1 protein, a membrane protein which has already been implicated in the pheromone response and in particular in cell fusion [29]. Another protein implicated in the glucose metabolism, GTS1, is inferred as a Steiner protein. As both these proteins are interacting with the actin protein ACT1, one could hypothesize a cross-talk between pheromone response, glucose metabolism and cytoskeleton structure. If role of the cytoskeleton in mating is quite well-known, implication of the glucose metabolism is not, but could be a sign of a global regulation of the cellular state previous to mating, as GTS1 is also a known regulator of transcription.

COS8 is found at the end of the SST2-SHO1 cascade. SST2 is the regulator of desensitization of the pheromone pathway, while SHO1 is the main cell osmosensor and initiates various signaling pathways. The subtree found behind COS8 is composed of membrane proteins and the fatty-acid elongase ELO1. Moreover, COS8 interacts often with proteins involved in sphingolipid synthesis, such as LAC1 and AUR1 (not shown in Fig 5 because they occur less than 50% of the time). Finally, the main cascade leading to IFA38, a beta-keto reductase implicated in fatty acid metabolism and also found as Steiner protein, is indeed passing by COS8. The multiple interactions of COS8 with these proteins, either membrane-spanning or located in the ER, allows to hypothesize that COS8

plays a role in the secretory pathway – probably in relation with sphingolipid synthesis– during pheromone response. Interestingly, COS8 is one but a member of a very conserved gene family [30], and finding the function of COS8 could help to understand the role of the entire family.

4 Discussion

In this work we presented an efficient strategy to infer structure relations from sparse gene expression information and protein-protein interaction probabilities. This approach is based on statistical physics principles, is scalable (completely parallelizable) and is expected to be well-suited for large networks. The scheme is highly efficient (the computation time scales as $D|E|$ where $|E|$ is the number of edges of the protein network, and it normally suffices to take $D = O(logN)$ to achieve optimal values). This property allowed us to explore values of the parameter λ and a large number of pathways very quickly, much faster than it would have taken with complete algorithms and other available heuristics.

The main drawback of the approach resides in its input limitations, that is, it cannot infer new interactions between proteins and must follow the structure of the PIN given as input. This makes it difficult to apply our method to organisms where the PIN is unknown or poorly described, which is the case for many organisms, such as human. However, this issue is becoming obsolete with the rise of new experimental techniques in the proteomic fields. Moreover, there are bioinformatic solutions that could be used in order not to be limited by this problem. First, one could add in the PIN very high cost edges on a set of putative interactions. Analysis of the Steiner trees with increasing values of λ, may allow to see, among the added edges, which are selected more frequently by the algorithm, and thereby discriminate between them. Second, as the methods to infer protein-protein interactions based solely on sequence become more efficient (see *e.g.* [31]), it should be possible to develop an integrated framework where protein interactions are inferred numerically before applying our methodology.

Our methodology, while using state-of-the-art computational techniques, is able to infer quantitatively which Steiner proteins could play a role in a given context, as represented by expression data. The network representation allows a clear interpretation: specific interactions are predicted in defined conditions. This type of predictions is easy to confirm experimentally by double-hybrid and genetic experiments, making our methodology an invaluable input for wet labs. Collaborations have indeed been started to experimentally test our predictions. Moreover, our algorithm could be made still more efficient, by including genetic or regulatory interactions in the base network or searching for protein complexes instead of protein interactions. Developments in this sense, coupled with experimental validations of our predictions, will finally allow the development an integrated message-passing framework for systems biology, in direct contact with experimental data and labs.

Acknowledgements

We thank S. Fortunato for the computation of the in-betweenness, and J.M. François for help in the biological analyses. M.B.B and A.B acknowledge a fellowship from Politecnico di Torino and a combined Microsoft research/Politecnico di Torino funding.

References

1. Elston, T.C.: Probing pathways periodically. Sci. Signal 1(42) (2008); pe47
2. Dohlman, H.G., Slessareva, J.E.: Pheromone signaling pathways in yeast. Sci. STKE 2006(364) (December 2006); cm6
3. Luttrel, L.: Transmembrane signalling by g protein couple receptors. Methods Mol. Biol. 332, 3–49 (2006)
4. Chen, R.E., Thorner, J.: Function and regulation in MAPK signaling pathways: lessons learned from the yeast Saccharomyces cerevisiae. Biochem. Biophys. Acta 1773(8), 1311–1340 (2007)
5. Thattai, M., van Oudenaarden, A.: Attenuation of noise in ultrasensitive signaling cascades. Biophys. J. 82(6), 2943–2950 (2002)
6. Kholodenko, B.N.: Cell-signalling dynamics in time and space. Nat. Rev. Mol. Cell Biol. 7(3), 165–176 (2006)
7. Ventura, A.C., Sepulchre, J.A., Merajver, S.D.: A hidden feedback in signaling cascades is revealed. PLoS Comput. Biol. 4(3), e1000041 (2008)
8. Locasale, J.W., Chakraborty, A.K.: Regulation of signal duration and the statistical dynamics of kinase activation by scaffold proteins. PLoS Comput. Biol. 4(6), e1000099 (2008)
9. Bashor, C.J., Helman, N.C., Yan, S., Lim, W.A.: Using engineered scaffold interactions to reshape MAP kinase pathway signaling dynamics. Science 319(5869), 1539–1543 (2008)
10. Scott, J., Ideker, T., Karp, R.M., Sharan, R.: Efficient algorithms for detecting signaling pathways in protein interaction networks. J. Comput. Biol. 13(2), 133–144 (2006)
11. White, A., Ma'yan, A.: Connecting seed lists of mammalian proteins using steiner trees. Nature Precedings (2008)
12. Scott, M.S., Perkins, T., Bunnell, S., Pepin, F., Thomas, D.Y., Hallett, M.: Identifying regulatory subnetworks for a set of genes. Mol. Cell Proteomics 4(5), 683–692 (2005)
13. Zhao, X.M., Wang, R.S., Chen, L., Aihara, K.: Uncovering signal transduction networks from high-throughput data by integer linear programming. Nucleic Acids Res. 36(9), e48 (2008)
14. Bayati, M., Borgs, C., Braunstein, A., Chayes, J., Ramezanpour, A., Zecchina, R.: Statistical mechanics of steiner trees. Phys. Rev. Lett. 101(3), 037208 (2008)
15. Mézard, M., Parisi, G., Zecchina, R.: Analytic and algorithmic solution of random satisfiability problems. Science 297(5582), 812–815 (2002)
16. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M.,

Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868), 141–147 (2002)
17. Chang, W.C., Li, C.W., Chen, B.S.: Quantitative inference of dynamic regulatory pathways via microarray data. BMC Bioinformatics 6, 44 (2005)
18. Steffen, M., Petti, A., Aach, J., D'haeseleer, P., Church, G.: Automated modelling of signal transduction networks. BMC Bioinformatics 3, 34 (2002)
19. Johnson, D., Minkoff, M., Phillips, S.: The prize collecting steiner tree problem: theory and practice. In: Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 760–769 (2000)
20. Lucena, A., Resende, M.G.C.: Strong lower bounds for the prize collecting Steiner problem in graphs. Discrete Applied Mathematics 141(1-3), 277–294 (2004)
21. Bayati, M., Braunstein, A., Zecchina, R.: A rigorous analysis of the cavity equations for the minimum spanning tree. Journal of Mathematical Physics 49(12), 125206 (2008)
22. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D.: DIP: the database of interacting proteins. Nucleic Acids Res. 28(1), 289–291 (2000)
23. Güldener, U., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, H.W., Stümpflen: Mpact: the MIPS protein interaction resource on yeast. Nucleic Acids Res. 34(Database issue), D436–D441 (2006)
24. Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J.: Gaining confidence in high-throughput protein interaction networks. Nat. Biotechnol. 22(1), 78–85 (2004)
25. Suthram, S., Shlomi, T., Ruppin, E., Sharan, R., Ideker, T.: A direct comparison of protein interaction confidence assignment schemes. BMC Bioinformatics 7, 360 (2006)
26. Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y.D., Dai, H., Walker, W.L., Hughes, T.R., Tyers, M., Boone, C., Friend, S.H.: Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. Science 287(5454), 873–880 (2000)
27. The Gene Ontology Consortium: Gene ontology: tool for the unification of biology. Nat. Genet. 25, 25–29 (2000)
28. Heiman, M.G., Walter, P.: Prm1p, a pheromone-regulated multispanning membrane protein, facilitates plasma membrane fusion during yeast mating. J. Cell Biol. 151(3), 719–730 (2000)
29. Aguilar, P.S., Engel, A., Walter, P.: The plasma membrane proteins prm1 and fig1 ascertain fidelity of membrane fusion during yeast mating. Mol. Biol. Cell 18(2), 547–556 (2007)
30. Despons, L., Wirth, B., Louis, V.L., Potier, S., Souciet, J.L.: An evolutionary scenario for one of the largest yeast gene families. Trends Genet. 22(1), 10–15 (2006)
31. Weigt, M., White, R.A., Szurmant, H., Hoch, J.A., Hwa, T.: Identification of direct residue contacts in protein-protein interaction by message passing. Proc. Natl. Acad. Sci. U S A 106(1), 67–72 (2009)