Simultaneous Reconstruction of Multiple Signaling Pathways *via* the Prize-Collecting Steiner Forest Problem

Nurcan Tuncbag[1], Alfredo Braunstein[2,3], Andrea Pagnani[3], Shao-Shan Carol Huang[1], Jennifer Chayes[4], Christian Borgs[4], Riccardo Zecchina[2,3], and Ernest Fraenkel[1]

[1] Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
{ntuncbag,shhuang,fraenkel-admin}@mit.edu
[2] Department of Applied Science, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy
{alfredo.braunstein,riccardo.zecchina}@polito.it
[3] Human Genetics Foundation, Via Nizza 52, 10126 Torino, Italy
andrea.pagnani@hugef-torino.org
[4] Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142, USA
{jchayes,borgs}@microsoft.com

Abstract. Signaling networks are essential for cells to control processes such as growth and response to stimuli. Although many "omic" data sources are available to probe signaling pathways, these data are typically sparse and noisy. Thus, it has been difficult to use these data to discover the cause of the diseases. We overcome these problems and use "omic" data to simultaneously reconstruct multiple pathways that are altered in a particular condition by solving the prize-collecting Steiner forest problem. To evaluate this approach, we use the well-characterized yeast pheromone response. We then apply the method to human glioblastoma data, searching for a forest of trees each of which is rooted in a different cell surface receptor. This approach discovers both overlapping and independent signaling pathways that are enriched in functionally and clinically relevant proteins, which could provide the basis for new therapeutic strategies.

Keywords: Prize-collecting Steiner forest, signaling pathways, multiple network reconstruction.

1 Introduction

High-throughput technologies including mass spectrometry, chromatin immunoprecipitation followed by sequencing (CHIP-Seq), RNA sequencing (RNA-seq), microarray and screening methods have the potential to provide dramatically new insights into biological processes. By providing a relatively comprehensive view of the changes that occur for a specific type of molecule or perturbation, these approaches can uncover previously unrecognized processes in a system of interest. However, interpreting these data types together to provide a coherent view of the biological processes is still a challenging task. In order to discover how changes in

B. Chor (Ed.): RECOMB 2012, LNBI 7262, pp. 287–301, 2012.

different classes of molecules relate to each other, it is possible to map the data onto a network of known or predicted interactions. In the ideal case, the observed interactions would all lie near each other in a functionally coherent part of the interaction network (the interactome). However, due to false positives and false negatives in both the "omic" data and the interactome, the true situation is much more complex; advanced algorithms are needed to find meaningful connections among the data. Among the approaches that have been proposed to find these sub-networks from the interactome are network flow optimization [1, 2], network propagation [3], the Steiner tree approach [4-6], network inference from gene expression [7, 8], linear programming [9], maximum-likelihood [10], electric circuits [11-13], network alignment [14] and Bayesian networks [15].

In our previous work, we used the prize-collecting Steiner tree formalism to find an optimum tree composed of nodes detected in experiments (terminals) and nodes that were not detected (Steiner nodes). We assigned costs to each interaction reflecting our confidence that the reported interaction was real and assigned prizes for excluding any of the terminals from the tree based on confidence in the proteomic or transcriptional data. By minimizing the sum of the total cost of all edges in the tree and the total prize of all nodes not contained in the tree, we were able to obtain compact and biologically relevant networks [4, 6]. Despite the power of Steiner tree approach for identifying functionally coherent networks, it is restricted to discovering a connected subgraph, which may be an inadequate representation for many systems. In particular, we often expect there to be many simultaneously acting biological processes in the cell that may not be connected together by interactions in the currently known interactome. These processes may be unconnected either because they may involve essentially independent cell functions, or simply due to our imperfect knowledge of the interactome.

In this work, we formulate a forest (defined as a disjoint union of trees) approach to identify simultaneously acting pathways in biological networks using both proteomic and transcriptional data, We use a generalization of the message-passing algorithm for the Prize-collecting Steiner Tree (PCST) problem [4, 16]. We first demonstrate the forest approach by using it to integrate proteomic and transcriptional data in the yeast pheromone response, showing that the forest consists of trees enriched in specific and distinct biological processes. As an additional feature, directed edges, which are particularly useful for representing the effects of enzymes and transcriptional regulators on their targets, are also incorporated.

We reasoned that the Steiner forest approach could be utilized in modeling mammalian signaling where there are many more cell-surface receptors and downstream pathways than in yeast. In principle, the forest approach could uncover multiple, independent components of the biological response. Although the interactome data are much less complete for mammals than for yeast, we show that the same methods are applicable. We built prize-collecting Steiner forests derived from proteomic data from a model of glioblastoma multiforme (GBM) in which each tree was rooted in a different cell surface receptor representing independent signaling pathways and potential points of therapeutic intervention. The solution reveals several known pathways and some unexpected new ones that are altered in the disease and suggests potential therapeutic strategies. The modified algorithm can now be applied to a wide range of complex systems.

2 Methods

2.1 Datasets

Throughout this work, two different biological networks are used: the yeast interactome and the human interactome. We refer to nodes with prize values greater than zero as terminal nodes.

Yeast Dataset. The yeast interactome contains 34,712 protein-protein and transcription factor to target interactions between 5,957 nodes. The terminal node set contains 106 differentially phosphorylated proteins detected by mass spectrometry [17] and 118 differentially expressed genes [18] detected by microarray in response to the mating pheromone alpha factor. The node prizes are computed from the fold changes between treated and non-treated conditions. The edge costs are calculated by taking a negative log of the interaction probability. The details are available in [6]. In this study, we modified the transcription factor–DNA interactions to be directed edges. We also added to the interactome a set of directed edges that represent phosphorylation and dephosphorylation reactions between kinases, phosphatases and their substrates [19]. If these interactions are available in the original interactome, probabilities are retained. If they are not, the probabilities of these interactions are set uniformly to 0.8, based on the distribution of the probabilities in the original interactome. The final interactome contains 35,998 edges between 5,957 nodes. In both cases, the resulting interactomes are comprised of both undirected and directed edges.

Human Dataset. Protein-protein interactions in the STRING database (version 8.3) are used as the data source for the human interactome [20]. Here, the probabilities from experiments and database evidence channels are combined to obtain the final probability of the interactions. Interactions with a combined probability greater than 0.8 are included in the interactome. The receptor molecules are collected from the Human Plasma Membrane Database [21] where 331 receptors are available in the interactome derived from STRING. The phosphoproteomics data in [22] is combined with the interactome in humans for the GBM test case. From this dataset, 72 proteins containing phophorylated tyrosine peptides are present in our human interactome.

2.2 Prize-collecting Steiner Tree Problem

For a given, directed or undirected network $G(V, E, c(e), p(v))$ of node set V and edge set E, where a $p(v) \geq 0$ assigns a prize to each node $v \in V$ and $c(e) \geq 0$ assigns a cost to each edge e ∈ E. The aim is to find a tree $T(V_T,E_T)$, by minimizing the objective function:

$$f(T) = \beta \sum_{v \notin V_T} p(v) + \sum_{e \in E_T} c(e) \tag{1}$$

where the first term is β times the sum of the node prizes not included in the tree T and the second part is the sum of the edge costs of T. Note that

$$\sum_{v \notin V_T} p(v) = -\sum_{v \in V_T} p(v) + const \tag{2}$$

so that minimizing *f(T)* amounts to collecting the largest set of high prize vertices while minimizing the set of large cost edges in a trade-off tuned by β. As a starting point, we consider the message-passing algorithm for the PCST problem introduced in [4]. The message-passing algorithm converts the global problem of finding the optimal tree into a set of local problems that can be solved efficiently. These equations are solved iteratively in a computationally efficient way. Here we present a generalization of the message passing algorithm designed to solve the PCST problem on directed networks (*i.e.* where in general c(e{i,j}) might be different from c(e{j,i})). In this variant, the optimization will be done on directed rooted trees, where choice of the root (which will be part of the candidate tree) is an external parameter of the algorithm.

2.3 Prize-collecting Steiner Forest (PCSF) Problem

A type of PCSF has already been considered in [23, 24]. In these works penalties are assigned to each pair of nodes either directly connected in the tree (*i.e.* edges belonging to the forest), or completely disconnected (*i.e.* in different forest components). Here we consider a different PCSF construction for a given, directed or undirected network *G(V, E, c(e), p(v))* of node set *V* and edge set *E*, where a $p(v) \geq 0$ assigns a prize to each node $v \in V$ and $c(e) \geq 0$ assigns a cost to each edge e ∈ F. The aim is to find a forest $F(V_F, E_F)$ that minimizes the objective function:

$$f'(F) = \beta \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) + \omega \cdot \kappa \qquad (3)$$

where κ is the number of trees in the forest and ω is new tuning parameter explained below. A practical way of minimizing *f'* consists in casting the PCSF into a PCST on a slightly modified graph. The idea is to introduce an extra root node v_0 into the network connected to each node $v \in V$ by an edge (v, v_0) with cost ω [25]. The PCST algorithm is employed on the resulting graph $H(V \cup \{v_0\}, E \cup Vx\{v_0\})$ and the solution will be called *T*. We define the forest *F* as *T* with all edges that point to the root removed. It is straightforward to see that the tree *T* is minimal for *f* if and only if the forest *F* is minimal for *f'*. Typically, the algorithm is run for different values of β and ω.

We used the previously published message-passing approach as the underlying implementation for this forest search [4], as many of our networks exceeded the capacity of the linear programming approaches. The message-passing approach is computationally fast and robust to the noise in the network as well. Although this algorithm is not guarranteed to find the optimal solution, in practice the networks it discovers are very similar to the exact solution. Introducing the artificial edges allows the algorithm to identify one or more trees that are only connected to the artificial node and not to each other. Although this modification seems algorithmically straightforward, its biological implications are very important. The concept is illustrated in Figure 1. In that example, two distinct pathways are connected only through spurious edges. The main difference between the tree formalism and the forest formalism is that the former one that connects as many of the experimental data as it can in a single network. As a result, it will either have to exclude some of the data that relate to distinct biological processes or add spurious edges to force these data to connect to the tree while the latter

one allows the corresponding nodes to be included in distinct trees. However, the forest formalism is able to locate distinct biological processes into different sub-trees through the artificial node. The artifical node and edges give the flexibility of generating several sub-trees without paying any penalty.

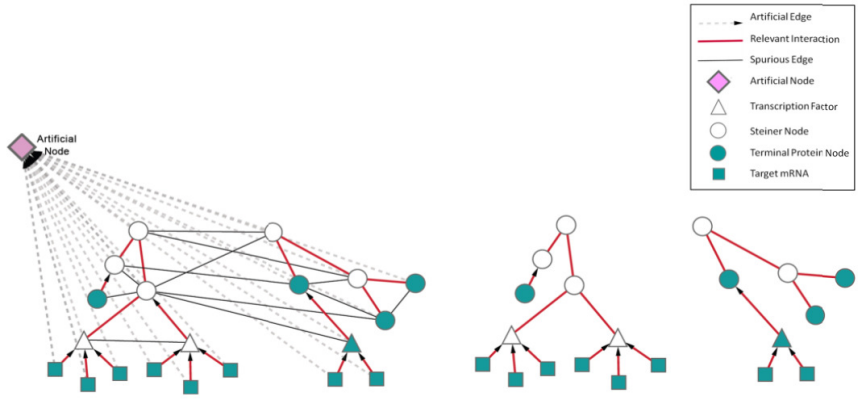

Fig. 1. Conceptual illustration of the PCSF algorithm. The left panel shows an interactome and the right panel shows the Steiner forest constructed from that interactome. The direction of transcription factor to target and kinase/phosphatase to substrate interactions are pointing towards the root node (opposite to the biological direction). In this scenario, there are spurious edges between these two pathways in the interactome. The PCSF algorithm provides the advantage to connect these distinct pathways artificially.

Tuning the Parameters. The parameters to be tuned in this problem are ω and β. The number of components of the solution (κ) depends strongly on the parameter ω, but it also depends on the β value: e.g. for $\beta = 0$ the optimal solution is the empty forest for all values of other parameters. For other values of β, while some sub-trees are composed of a single node, some others are composed of large number of nodes in the resulting forest. A forest with many very small trees (a single node each) would be obtained with very small artificial edges cost. The limiting case in the other direction is a single tree resulting from very large artificial edge cost. Therefore, there is a non-trivial interaction between the two parameters (β and ω). In principle, this two dimensional (β, ω) space of parameters should be explored. In this way, we get many possible types of forest: many small trees, many large trees, few small trees, few large trees. The effect of ω and β intervals highly depends on the distribution of edge costs and node prizes in the targeted interactome, so these parameters will be different for different datasets. For the yeast dataset, ω values are tuned between [0.005, 0.1] and β values are tuned between [1, 20].

Functional Annotation. For functional enrichment analysis, the BINGO plug-in [26] of Cytoscape [27] was used. The p-value significance threshold of 0.05 was used, which is corrected for multiple hypothesis testing, and all yeast proteins were used as the background set for the yeast dataset. For the human dataset, the functional

enrichment is performed by using all human proteins as background set. All network visualizations were performed in Cytoscape [27].

3 Results

3.1 The PCSF Approach Reveals Parallel Working Pathways in Addition to Hidden Individual Proteins or Genes in Yeast Pheromone Response

High-throughput experimental methods like mass-spectrometry are capable of simultaneously detecting changes in many distinct biological processes that will not be connected by physical interactions. However, the PCST approach searches for a tree structure in the interactome that connects as many of the experimental data as it can. As a result, it will either have to exclude some of the data that relate to distinct biological process or add spurious edges to force these data to connect to the tree. The main advantage of PCSF approach over PCST is that PCSF does not force the system to be connected in a single network, and it can automatically separate multiple pathways.

We tested the PCSF algorithm using data from the yeast pheromone response, which we had previously analyzed using the prize-collecting Steiner tree approach. The data consist of phosphoproteomic and transcriptional changes induced by mating pheromone, and the network is enriched with directed transcription factor-target and kinase/phosphatase-substrate reactions. The edge costs of the interactome were computed as the negative log of the interaction probabilities, and node prizes were obtained from the scheme detailed in [6]. To explore the space of solutions, we tuned the ω and β parameters between [0.005, 0.1] and [1, 20], respectively. The minimum, maximum and average size and number of trees in the constructed PCSFs are extracted for each (ω, β) pair and the distribution of these values along ω parameter is plotted. We looked for a solution in a region where the number of trees and average size of the trees in the forest are closest to each other. By these criteria, the best solution is found when $\omega = 0.025$ and $\beta = 13$. We note that in order to explore these parameters, we constructed 400 solutions to the PCST problem. This number of calculations is only practical using the message-passing algorithm, but not with the integer linear programming based approaches.

The solution to PCSF problem places distinct functional classes in seperate sub-trees. In this solution, there are six trees, each containing more than 10 nodes. In Figure 2, each tree is labeled with its corresponding pathway. Small sub-trees such as $T_{3\text{-}6}$ are enriched in specific biological processes including the PKC pathway, actin organization, protein folding and kinetochore, and DNA and chromatin pathways, while larger trees contain multiple processes. For example, the largest subtree, T_1, contains the pheromone core MAPK pathway with CDC28 related proteins and the second largest one, T_3, contains transcription and transport processes (see Figure 2). There are two different yeast MAPK pathways; the pheromone-induced MAPK and the protein kinase C (PKC) pathways [28, 29]. The PCSF algorithm correctly separates these two pathways into different trees. The largest tree in size is T_1 contains pheromone-induced MAPK pathway but the PKC pathway is located in T_3. While the former one induces cells to differentiate and be prepared for mating, the latter one is involved in cell integrity and new cell wall synthesis.

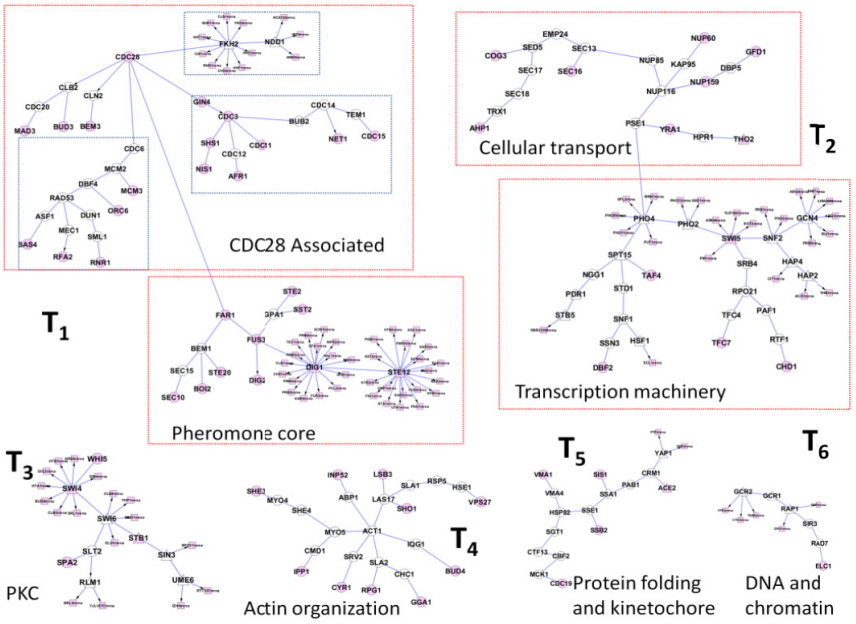

Fig. 2. Prize-collecting Steiner Forest (PCSF) of the Yeast Pheromone Response Network. Functional groups annotated by Gene Ontology (GO) are tagged with red boxes. In this PCSF, the rectangular nodes are DNA, triangular nodes are transcription factors and the circular nodes are proteins. Terminal nodes are colored red.

The core pheromone response pathway component in T_1 includes the STE2 receptor. In this sub-tree, the STE2-GPA1-FUS3 interaction is in the core of pheromone response. In addition, the MAP kinase FUS3 activates several transcription factors such as, STE12, DIG1, DIG2 for the expression of mating related genes. T_1 contains DNA replication proteins and cell cycle proteins associated with CDC28 as well. Here, the connection between the MAP kinase pathway in the pheromone core and the CDC28 associated sub-network is constructed through the interaction between FAR1 and CDC28. FAR1 is a direct inhibitor of CDC28/CLN2 complex and functions in orienting cell polarization. This association blocks the cell cycle progression. FUS3 phosphorylates FAR1, and only phosphorylated FAR1 can associate with CDC28/CLN2 complex. All these interactions and these connected pathways are correctly located into the same sub-tree.

The algorithm correctly identifies SLT2, which was not detected in the phosphoproteomic data, as a key node in regulating new cell wall synthesis. SLT2 is a serine/threonine MAP kinase activated in a cascade starting with PKC. The phosphoproteomic data are not sufficient for the algorithm to pick up the upstream pathway. However, in T_3, the algorithm links SLT2 to several transcription factors that mostly function in cell wall integrity and biosynthesis. SLT2 activates RLM1 [30], SWI4 [31] and SWI6 transcription factors. RLM1 functions in the maintenance

of cell integrity. SWI4/SWI6 regulates the expression of genes functioning in cell wall synthesis and G1/S transition of the cell cycle.

Transcriptional machinery and transport proteins are located in T_2, seperate from other trees. The connection between transcriptional machinery and cellular transport part is achieved by the interaction between PHO4 and PSE1. Although these two proteins are experimentally undetected, the PCSF algorithm locates them in the same sub-tree. Direct association of PSE1 to PHO4 is required for the import of PHO4 into the nucleus [32]. Nuclear pore components (NUP60, NUP85, NUP116, NUP159) are located in T_2 because nuclear transport is achieved through the nuclear pore [32]. In this sub-tree, the transcription factor PHO2 functions in a combinatorial manner with PHO4 and SWI5 [33].

Table 1. GO enrichments of the sub-trees in the PCSF illustrated in **Fig. 2**

Subtree Name	GO Enrichment - Biological Process	Corr p-value
T_1	regulation of cell cycle	1.97×10^{-17}
	cell division	2.60×10^{-17}
	cell cycle	3.02×10^{-17}
T_2	transcription	7.07×10^{-13}
	regulation of nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	2.36×10^{-12}
	nuclear transport	7.30×10^{-8}
T_3	positive regulation of gene-specific transcription	2.20×10^{-5}
	regulation of gene-specific transcription	7.75×10^{-5}
	positive regulation of transcription, DNA-dependent	9.11×10^{-5}
T_4	actin filament-based process	1.51×10^{-9}
	endocytosis	4.42×10^{-9}
	actin cytoskeleton organization	9.32×10^{-9}
T_5	protein folding	1.60×10^{-3}
	protein refolding	1.60×10^{-3}
	kinetochore assembly	4.25×10^{-3}
T_6	positive regulation of glycolysis	2.54×10^{-4}
	regulation of glycolysis	2.54×10^{-4}
	positive regulation of transcription	2.54×10^{-4}

In addition to the pathway analysis, we utilized GO biological process annotations to find the specific biological processes enriched in these trees. In **Table 1**, the top three annotations for each tree are tabulated along with their corrected p-values. These results show that this method effectively locates different biological processes into different trees. Instead of forcing all nodes to be connected in a single network, this "forest" representation composed of multiple sub-trees is more useful for distinguishing distinct pathways. The forest solution retains enrichment for the expected biological process, such as response to stress, cell cycle, signaling and transport. Further, by adding directions between transcription factors to targets and enzyme to substrate interactions, we are able to obtain condition-specific transcription factors and compact networks.

3.2 The PCSF Algorithm Reveals Coordinately Acting Receptor Molecules Functioning in Human GBM by Integrating Receptome, Interactome and Proteomics Data

Having demonstrated that the PCSF algorithm can successfully distinguish parallel-working pathways in yeast, we used it to identify cell surface receptors associated with signaling pathways altered in disease. Cell surface receptors are an interesting class of molecules to study, as they may be particularly easy to target with therapeutic agents. There is increasing evidence that some proteins are "undruggable," in other words hard or impossible to target with small molecule-based therapies because their three-dimensional shape does not have any appropriate concave sites to which these proteins can bind. In contrast, cell surface receptors can either be targeted with their natural ligand, modified forms of the natural ligand, small molecules that insert into the naturally occurring binding pocket or antibodies.

We modified our approach to identify cell-surface receptors associated with phosphoproteomic changes that occur in a model of glioblastoma. We use the artificial node to represent external stimuli (including autocrine loops) that potentially activates multiple receptor molecules, by connecting this node only to cell surface receptors, of which 331 are present in our human interactome. After running the prize-collecting Steiner tree algorithm and removing the artificial node, each sub-tree will contain one receptor as the starting node. The receptors selected in the solution of PCSF represent those most closely connected to the measured phosphoproteomic data and are therefore likely to be main contributors of the disease.

We applied this approach to phosphotyrosine data for a model of human GBM [22] representing phosphorylation differences between cells expressing an oncogenic mutation in the EGFR protein and cells with an inactive form of this receptor tyrosine kinase. The result is a set of eleven compact trees each rooted in one of the 331 potential receptors. The selected receptors in order of their tree sizes are EGFR, ERBB2, CD36, IGF1R, PTCH1, A2MR, SDC2, MET, ITGB3, NPR1 and EPHA2 (see Fig. 3). Although the algorithm had no direct knowledge that the data represented the results of mutation in EGFR, it selected this as the root of the largest tree. In fact, each of the four top receptors has a known link to cancer. EGFR and ERBB2 are EGF-family receptors, and it is known that EGFR is mutated in more than 50% GBM cases [34]. IGF1R is overexpressed in many tumors and mediates proliferation and resistance to apoptosis, and it is currently an anti-cancer treatment target [35]. Because IGF1R is also abnormally active in GBM, its inhibition is presented as a potential therapy to arrest the tumor growth [36]. It has been previously shown that the EGF and IGF pathways cross-talk [37], and IGF1R mediates resistance to anti-EGFR therapy in glioma cells [38]. Although CD36 functions in brain specific angiogenic regulation [39] and the interactions between CD36-Fyn-Yes lead to calcium and neurotransmitter release [40], its relation to GBM has not been studied in detail.

Although the algorithm is constrained to identify independent trees, we can observe the potential for cross-talk between different receptors by adding back all the edges among the selected nodes. We noticed two receptors selected by the algorithm, namely MET and ITGB3 (integrin-β3), are also very important, despite the fact that their corresponding sub-trees each contain only two nodes. When all edges are put

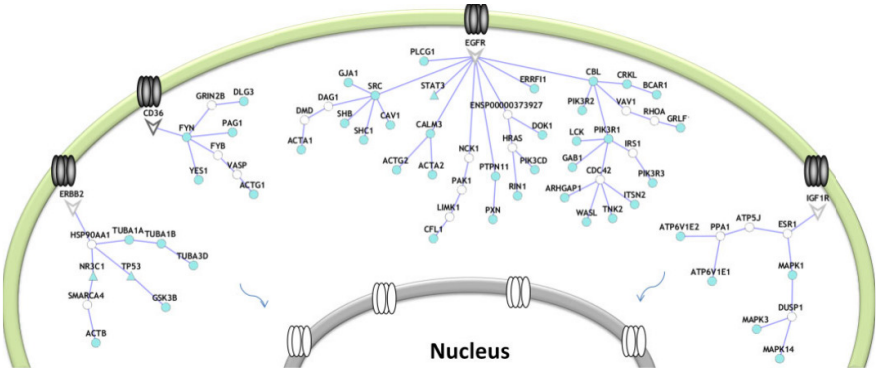

Fig. 3. Network representation of the PCSF for human glioblastoma dataset. Each tree is rooted from a cell surface receptor. The receptor molecules are represented by the arrowheads, transcription factors in triangles and other proteins as circles. Terminal nodes are colored in cyan.

back in the Steiner forest we observe extensive links between these two receptors and the EGFR sub-tree. MET has links to seven proteins out of nine first neighbors of EGFR and to 16 proteins in total of the EGFR rooted sub-tree, and ITGB3 has links to seventeen members of the EGFR sub-tree. By contrast, IGF1R has links only to three proteins and ERBB2 has link only to two proteins in the first neighbors of EGFR. Although this information was not provided to the algorithm, MET is detected as differentially phosphorylated in the original data, and a MET inhibitor synergizes with an EGFR inhibitor [22].

Mammalian signaling systems frequently demonstrate a high degree of cross-talk. If two receptors share many common downstream components, the algorithm need only choose one of these as a root node to explain all the terminal nodes. We, therefore, introduce a perturbation-based approach to improve the sensitivity of the algorithm in identifying receptors that share many downstream components with the selected root nodes. In this analysis, selected receptor molecules (the root of the largest tree in the forest and other receptors in its corresponding sub-family available in the forest) and all their interactions are removed from the interactome and PCSF algorithm is applied to the remaining network. Through this *in silico* knock-out experiment, we can find the other receptor molecules whose role may be masked in the presence of the receptors in the initial forest.

We first knocked-out two EGF/ERBB sub-family receptors, EGFR and ERBB2, from the network and re-generated the PCSF. In the resulting network, 37 out of 42 nodes in the down-stream of EGFR in the original tree are connected to other receptors. PDGFR (Platelet-derived growth factor receptor) is the root of the largest tree, covering 23 nodes linked to EGFR in the original network. This observation suggests that PDGFR may have many overlapping functions with EGFR. In fact, several studies have shown that PDGFR is critical in brain tumorogenesis [41, 42]; mutation of PDGFR causes alteration in the intracellular signaling [43] and it is a therapeutic target in GBM [44]. Another 14 nodes down-stream of EGFR are shared by MET (seven nodes), IGF1R (five nodes) and ITGB3 (two nodes). Although these

four receptors (PDGFR, MET, IGF1R and ITGB3) capture many of the nodes that were down-stream of EGFR, five nodes are not captured by any other receptors. These may represent signaling that is uniquely downstream of EGFR.

To further explore the network, we removed PDGFR in addition to EGFR and ERBB2. In the new network, the MET receptor partially replaces PDGFR. It has been shown that the MET receptor is activated in GBM and it might be a therapeutic target [45]. Similar to the MET receptor tree, the sub-tree containing ITGB3 receptor also collects several of the nodes previously associated with EGFR in its corresponding sub-tree. It is interesting to note that integrins function as both upstream and downstream effectors of growth factor receptors, such as EGFR, IGF1R, PDGFR, MET [46]. Integrins and their relation to GBM have not been studied in detail, which may have clinical importance in GBM.

During all these leave-one-receptor-out tests, IGF1R is present in the resulting PCSF, and it retains all proteins in the original network. The downstream network of IGF1R starts with the estrogen receptor (ESR1) interaction and it contains several MAPKs. It has been shown that ESR1 and IGF1R are cross-regulated in the brain and activate the MAPK/ERK pathway. This system of interactions results in some neural functional regulations in the brain; such as, synaptic plasticity, neurotic growth, and neuronal survival [47]. The size of the trees corresponding to the down-stream of MET and ITGB3 receptors increases at each knock-out. Also, the FYN related downstream pathway of CD36 is swapped to be downstream of MET receptor, although CD36 is not knocked-out. This result implies that FYN-related pathway may be activated by several receptors.

To further validate the relevancy of these receptor molecules (EGFR, ERBB2, IGF1R, CD36, PDGFR, MET and ITGB3), we used the TCGA GBM Gene Ranker (http://cbio.mskcc.org/tcga-generanker/). This server combines available literature information and TCGA data for individual genes to score them. All selected receptors are among highly ranked genes (genes having a score greater than 2.0) in GBM (calculated scores are as follows: EGFR: 15.75, MET: 11.75, ERBB2: 9.25, PDGFRB: 7.25, IGF1R: 4.0, ITGB3: 3.75, CD36: 2.0), with EGFR, MET and ERBB2 having the highest rank in the database.

We performed randomization tests to check the reliability of the output of the algorithm. Here, terminal nodes, their prizes and the parameter set are kept same with the original PCSF analysis of GBM. In addition, number of nodes, and edges and edge costs are the same as in the original interactome. Only the edges are re-shuffled randomly within the network. The randomization test is repeated ten times on different interactomes. These characteristics show that random PCSFs contain many more sub-trees when compared to the original PCSF and these sub-trees are not structured like the original trees; most of the trees in the random forests are 'stringy', composed of nine proteins at most. Further, random trees are not enriched for a specific biological process and none of the receptors found in the original PCSF are selected in the random PCSFs. The algorithm uses substantially more Steiner nodes to connect terminal nodes in random case. We performed another randomization test by reshuffling the nodes in the original interactome. In this way, the degree distribution is retained. In these randomizations we retain the same terminal nodes, prizes and the parameter set are kept same with the original PCSF analysis, but these proteins have now been randomly mapped to other nodes. The results show that the characteristics

of the sub-trees in the random PCSFs are similar with the previous random case; they are 'stringy', not structured and not enriched for functions. However, this time the total number of nodes included in the PCSF is not as large as in the previous random case. These results show that the real PCSF solution is significantly different than the random solutions. It is particularly important that the receptors found by the algorithm run on the GBM data are not selected in the randomizations, supporting the hypothesis that these receptors are biologically relevant.

4 Discussion

We present a method for simultaneously discovery of multiple pathways by searching for "forests" consisting of multiple trees. We are able to solve this problem efficiently, even for large human networks by a simple modification of the previously published message-passing solution for the Steiner tree problem. When applied to the pheromone response data on the directed yeast interactome, the PCSF approach reveals several parallel pathways affected by yeast pheromone. Some of these parallel pathways contain multiple, coherently acting processes, such as pheromone response and the CDC28 associated pathway, or transport and transcriptional activity of PHO4. Others contain only one process, such as actin organization or protein folding.

The advantages of the forest approach are most apparent when used to study mammalian cells, which respond to a large number of hormones, growth factors and cytokines. Applying this approach to proteomic data from a model of GBM results in a forest composed of several sub-trees, each of which is rooted from a receptor molecule. The PCSF algorithm is able to select receptors relevant to GBM from hundreds of molecules in the human receptome. The solution reveals several known pathways and some unexpected new ones. EGFR, ERBB2, IGF1R and CD36 are starting nodes of the largest sub-trees in the PCSF. This set of receptor molecules was selected by the algorithm among hundreds of receptors, and the literature search shows that each of the selected receptors is clinically relevant to GBM. To find additional receptors whose downstream signaling pathways overlap with the selected receptors, we used an iterative approach that can be thought of as an *in silico* knock-out experiment. In this analysis, a selected receptor molecule and all its interactions are removed from the interactome and PCSF algorithm is applied to the remaining network. These calculations revealed the roles of PDGFR, MET and ITGB3 all of which have been previously linked to GBM.

Our method can be efficienly utilized to reconstruct networks that are enriched in functionally and clinically relevant proteins. Further, the algorithm is flexible, and can be modified for other types of data such as protein-small molecule inhibitor interactions and protein-metabolite interactions.

Acknowledgements. We thank Dr. Sara Gosline from MIT and Dr. Oznur Tastan from Microsoft Research for their critical reading and fruitful comments. This work is supported by NIH grants U54CA112967, R01GM089903 and Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the US Army Research Office (the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred) and used computing resources funded by the National Science Foundation under Award No. DB1-0821391. EF receives support from the Eugene Bell Career Development Chair. RZ acknowledges the ERC grant OPTINF 267915. The support from the EC grant STAMINA 265496 is also acknowledged by AB and RZ.

References

1. Lan, A., Smoly, I.Y., Rapaport, G., Lindquist, S., Fraenkel, E., Yeger-Lotem, E.: ResponseNet: Revealing signaling and regulatory networks linking genetic and transcriptomic screening data. Nucleic Acids Res. (2011)
2. Yeger-Lotem, E., Riva, L., Su, L.J., Gitler, A.D., Cashikar, A.G., King, O.D., Auluck, P.K., Geddie, M.L., Valastyan, J.S., Karger, D.R., et al.: Bridging high-throughput genetic and transcriptional data reveals cellular responses to alpha-synuclein toxicity. Nat. Genet. 41(3), 316–323 (2009)
3. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. 6(1), e1000641 (2010)
4. Bailly-Bechet, M., Borgs, C., Braunstein, A., Chayes, J., Dagkessamanskaia, A., Francois, J.M., Zecchina, R.: Finding undetected protein associations in cell signaling by belief propagation. Proc. Natl. Acad. Sci. U.S.A. 108(2), 882–887 (2010)
5. Dittrich, M.T., Klau, G.W., Rosenwald, A., Dandekar, T., Muller, T.: Identifying functional modules in protein-protein interaction networks: an integrated exact approach. Bioinformatics 24(13), i223–i231 (2008)
6. Huang, S.S., Fraenkel, E.: Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks. Sci. Signal 2(81), ra40 (2009)
7. Friedman, N.: Inferring cellular networks using probabilistic graphical models. Science 303(5659), 799–805 (2004)
8. Bailly-Bechet, M., Braunstein, A., Pagnani, A., Weigt, M., Zecchina, R.: Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. BMC Bioinformatics 11, 355 (2010)
9. Ourfali, O., Shlomi, T., Ideker, T., Ruppin, E., Sharan, R.: SPINE: a framework for signaling-regulatory pathway inference from cause-effect experiments. Bioinformatics 23(13), i359–i366 (2007)
10. Yeang, C.H., Ideker, T., Jaakkola, T.: Physical network models. J. Comput. Biol. 11(2-3), 243–262 (2004)
11. Kim, Y.A., Wuchty, S., Przytycka, T.M.: Identifying causal genes and dysregulated pathways in complex diseases. PLoS Comput. Biol. 7(3), e1001095 (2011)
12. Missiuro, P.V., Liu, K., Zou, L., Ross, B.C., Zhao, G., Liu, J.S., Ge, H.: Information flow analysis of interactome networks. PLoS Comput. Biol. 5(4), e1000350 (2009)
13. Suthram, S., Beyer, A., Karp, R.M., Eldar, Y., Ideker, T.: eQED: an efficient method for interpreting eQTL associations using protein networks. Mol. Syst. Biol. 4, 162 (2008)
14. Sharan, R., Ideker, T.: Modeling cellular machinery through biological network comparison. Nat. Biotechnol. 24(4), 427–433 (2006)
15. Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., Causton, H.C., Pochanard, P., Mozes, E., Garraway, L.A., Pe'er, D.: An integrated approach to uncover drivers of cancer. Cell 143(6), 1005–1017 (2010)
16. Bayati, M., Borgs, C., Braunstein, A., Chayes, J., Ramezanpour, A., Zecchina, R.: Statistical mechanics of steiner trees. Phys. Rev. Lett. 101(3), 037208 (2008)
17. Gruhler, A., Olsen, J.V., Mohammed, S., Mortensen, P., Faergeman, N.J., Mann, M., Jensen, O.N.: Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway. Mol. Cell Proteomics 4(3), 310–327 (2005)

18. Issel-Tarver, L., Christie, K.R., Dolinski, K., Andrada, R., Balakrishnan, R., Ball, C.A., Binkley, G., Dong, S., Dwight, S.S., Fisk, D.G., et al.: Saccharomyces Genome Database. Methods Enzymol. 350, 329–346 (2002)
19. Breitkreutz, A., Choi, H., Sharom, J.R., Boucher, L., Neduva, V., Larsen, B., Lin, Z.Y., Breitkreutz, B.J., Stark, C., Liu, G., et al.: A global protein kinase and phosphatase interaction network in yeast. Science 328(5981), 1043–1046 (2010)
20. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., et al.: STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res. 37(database issue), 412–416 (2009)
21. Ben-Shlomo, I., Yu Hsu, S., Rauch, R., Kowalski, H.W., Hsueh, A.J.: Signaling receptome: A genomic and evolutionary perspective of plasma membrane receptors involved in signal transduction. Sci. STKE 2003(187), RE9 (2003)
22. Huang, P.H., Mukasa, A., Bonavia, R., Flynn, R.A., Brewer, Z.E., Cavenee, W.K., Furnari, F.B., White, F.M.: Quantitative analysis of EGFRvIII cellular signaling networks reveals a combinatorial therapeutic strategy for glioblastoma. Proc Natl. Acad. Sci. U.S.A. 104(31), 12867–12872 (2007)
23. Chekuri, C., Ene, A., Korula, N.: Prize-Collecting Steiner Tree and Forest in Planar Graphs. Data Structures and Algorithms (2010)
24. Gupta, A., Konemann, J., Leonardi, S., Ravi, R., Schaefer, G.: An efficient cost-sharing mechanism for the prize-collecting Steiner forest problem. In: SODA 2007 Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (2007)
25. Bailly-Bechet, M., Bradde, S., Braunstein, A., Flaxman, A., Foini, F., Zecchina, R.: Clustering with shallow trees. J Stat Mech., 12010 (2009)
26. Maere, S., Heymans, K., Kuiper, M.: BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21(16), 3448–3449 (2005)
27. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 13(11), 2498–2504 (2003)
28. Buehrer, B.M., Errede, B.: Coordination of the mating and cell integrity mitogen-activated protein kinase pathways in Saccharomyces cerevisiae. Mol. Cell Biol. 17(11), 6517–6525 (1997)
29. Zarzov, P., Mazzoni, C., Mann, C.: The SLT2(MPK1) MAP kinase is activated during periods of polarized cell growth in yeast. EMBO J. 15(1), 83–91 (1996)
30. Garcia, R., Bermejo, C., Grau, C., Perez, R., Rodriguez-Pena, J.M., Francois, J., Nombela, C., Arroyo, J.: The global transcriptional response to transient cell wall damage in Saccharomyces cerevisiae and its regulation by the cell integrity signaling pathway. J. Biol. Chem. 279(15), 15183–15195 (2004)
31. Baetz, K., Moffat, J., Haynes, J., Chang, M., Andrews, B.: Transcriptional coregulation by the cell integrity mitogen-activated protein kinase Slt2 and the cell cycle regulator Swi4. Mol. Cell Biol. 21(19), 6515–6528 (2001)
32. Kaffman, A., Rank, N.M., O'Shea, E.K.: Phosphorylation regulates association of the transcription factor Pho4 with its import receptor Pse1/Kap121. Genes Dev. 12(17), 2673–2683 (1998)
33. Bhoite, L.T., Allen, J.M., Garcia, E., Thomas, L.R., Gregory, I.D., Voth, W.P., Whelihan, K., Rolfes, R.J., Stillman, D.J.: Mutations in the Pho2 (Bas2) transcription factor that differentially affect activation with its partner proteins Bas1, Pho4, and Swi5. J. Biol. Chem. 277(40), 37612–37618 (2002)

34. Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455(7216), 1061–1068 (2008)
35. Macaulay, V.M.: The IGF receptor as anticancer treatment target. In: Novartis Found Symp., 262:235-243; discussion 243-236, 265-238 (2004)
36. Kiaris, H., Schally, A.V., Varga, J.L.: Antagonists of growth hormone-releasing hormone inhibit the growth of U-87MG human glioblastoma in nude mice. Neoplasia 2(3), 242–250 (2000)
37. Adams, T.E., McKern, N.M., Ward, C.W.: Signalling by the type 1 insulin-like growth factor receptor: interplay with the epidermal growth factor receptor. Growth Factors 22(2), 89–95 (2004)
38. Chakravarti, A., Loeffler, J.S., Dyson, N.J.: Insulin-like growth factor receptor I mediates resistance to anti-epidermal growth factor receptor therapy in primary human glioblastoma cells through continued activation of phosphoinositide 3-kinase signaling. Cancer Res. 62(1), 200–207 (2002)
39. Kaur, B., Cork, S.M., Sandberg, E.M., Devi, N.S., Zhang, Z., Klenotic, P.A., Febbraio, M., Shim, H., Mao, H., Tucker-Burden, C., et al.: Vasculostatin inhibits intracranial glioma growth and negatively regulates in vivo angiogenesis through a CD36-dependent mechanism. Cancer Res. 69(3), 1212–1220 (2009)
40. Silverstein, R.L., Febbraio, M.: CD36, a scavenger receptor involved in immunity, metabolism, angiogenesis, and behavior. Sci. Signal 2(72), re3 (2009)
41. Dai, C., Celestino, J.C., Okada, Y., Louis, D.N., Fuller, G.N., Holland, E.C.: PDGF autocrine stimulation dedifferentiates cultured astrocytes and induces oligodendrogliomas and oligoastrocytomas from neural progenitors and astrocytes in vivo. Genes Dev. 15(15), 1913–1925 (2001)
42. Uhrbom, L., Hesselager, G., Nister, M., Westermark, B.: Induction of brain tumors in mice using a recombinant platelet-derived growth factor B-chain retrovirus. Cancer Res. 58(23), 5275–5279 (1998)
43. Clarke, I.D., Dirks, P.B.: A human brain tumor-derived PDGFR-alpha deletion mutant is transforming. Oncogene 22(5), 722–733 (2003)
44. Ziegler, D.S., Wright, R.D., Kesari, S., Lemieux, M.E., Tran, M.A., Jain, M., Zawel, L., Kung, A.L.: Resistance of human glioblastoma multiforme cells to growth factor inhibitors is overcome by blockade of inhibitor of apoptosis proteins. J. Clin. Invest. 118(9), 3109–3122 (2008)
45. Li, Y., Li, A., Glas, M., Lal, B., Ying, M., Sang, Y., Xia, S., Trageser, D., Guerrero-Cazares, H., Eberhart, C.G., et al.: c-Met signaling induces a reprogramming network and supports the glioblastoma stem-like phenotype. Proc. Natl. Acad. Sci. U.S.A. 108(24), 9951–9956 (2011)
46. Alam, N., Goel, H.L., Zarif, M.J., Butterfield, J.E., Perkins, H.M., Sansoucy, B.G., Sawyer, T.K., Languino, L.R.: The integrin-growth factor receptor duet. J. Cell Physiol. 213(3), 649–653 (2007)
47. Cardona-Gomez, G.P., Mendez, P., DonCarlos, L.L., Azcoitia, I., Garcia-Segura, L.M.: Interactions of estrogen and insulin-like growth factor-I in the brain: molecular mechanisms and functional implications. J. Steroid Biochem. Mol. Biol. 83(1-5), 211–217 (2002)