SHARING INFORMATION TO RECONSTRUCT PATIENT-SPECIFIC PATHWAYS IN HETEROGENEOUS DISEASES

ANTHONY GITTER[1,2], ALFREDO BRAUNSTEIN[3,4], ANDREA PAGNANI[3,4], CARLO BALDASSI[3,4], CHRISTIAN BORGS[1], JENNIFER CHAYES[1], RICCARDO ZECCHINA[3,4], ERNEST FRAENKEL[2,*]

[1] *Microsoft Research, Cambridge, MA, USA*
[2] *Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*
[3] *DISAT and Center for Computational Sciences, Politecnico di Torino, Turin, Italy*
[4] *Human Genetics Foundation, Turin, Italy*
[*] *E-mail: fraenkel-admin@mit.edu*

Advances in experimental techniques resulted in abundant genomic, transcriptomic, epigenomic, and proteomic data that have the potential to reveal critical drivers of human diseases. Complementary algorithmic developments enable researchers to map these data onto protein-protein interaction networks and infer which signaling pathways are perturbed by a disease. Despite this progress, integrating data across different biological samples or patients remains a substantial challenge because samples from the same disease can be extremely heterogeneous. Somatic mutations in cancer are an infamous example of this heterogeneity. Although the same signaling pathways may be disrupted in a cancer patient cohort, the distribution of mutations is long-tailed, and many driver mutations may only be detected in a small fraction of patients. We developed a computational approach to account for heterogeneous data when inferring signaling pathways by sharing information across the samples. Our technique builds upon the prize-collecting Steiner forest problem, a network optimization algorithm that extracts pathways from a protein-protein interaction network. We recover signaling pathways that are similar across all samples yet still reflect the unique characteristics of each biological sample. Leveraging data from related tumors improves our ability to recover the disrupted pathways and reveals patient-specific pathway perturbations in breast cancer.

Keywords: Prize-collecting Steiner forest, breast cancer, protein-protein interactions

1. Introduction

Cancer is caused by mutations or other alterations that perturb normal biological processes in a manner that confers a selective growth advantage to the mutated cell. Massive efforts to sequence the DNA of thousands of tumors have detected hundreds of thousands of mutations [1]. However, due to the heterogeneity of tumors, very few genes are mutated frequently enough to be identified as driver genes [1] — those that yield a growth advantage — and generally the significantly mutated genes are already known cancer genes [2]. Fortunately, although even tumors within a specific subtype of cancer may be genetically diverse, the perturbed pathways are similar [1]. A promising direction is therefore combining genomic data with complementary data types to focus on these signaling pathways [2] and computationally searching for 'driver pathways' instead of individual driver genes.

Existing algorithms for analyzing cancer are unable to learn patient-specific driver pathways. Many algorithms find modules or subnetworks of altered genes [3–8] but produce a single set of modules for all tumors instead of tumor-specific predictions, limiting the potential for individualized therapies. PARADIGM [9] addresses this issue by combining multiple types of data to learn protein and pathway activity for each individual tumor. However, it relies on

fixed collections of pathways from pathway databases, which are inconsistent and incomplete even in model organisms like yeast [10] and can be altered by gain-of-function mutations [11].

De novo pathway discovery has been successful in other biological settings [10, 12–18], but previous approaches are not suitable for analyzing genomic alterations in cancer patients. Most pathway inference algorithms operate on a single set of input. In the cancer setting, this input is data from a single tumor, which makes it very difficult to determine which meaningful genes should compose the driver pathway amid the more numerous passenger mutations.

To overcome the noisiness of the input, we propose to discover tumor-specific driver pathways by leveraging the wealth of data that is available for other tumors of the same cancer subtype. Instead of learning pathways independently for all tumor samples we study all tumors simultaneously, constraining the predicted pathways to be similar. This idea is similar to what is known as multitask learning in other domains [19]. As we demonstrate in simulated settings and with real data from basal-like breast cancer tumors, such an approach can recover individualized driver pathways that contain common core elements that are relevant to the disease even though they may not be mutated in each tumor.

2. Methods

2.1. *Prize-collecting Steiner forest*

The prize-collecting Steiner forest (PCSF) algorithm [16] is a computational technique for *de novo* signaling pathway discovery. Given a biological network, such as a protein-protein interaction (PPI) network, and a set of proteins in the network that are believed to be relevant to a disease or condition of interest, PCSF returns a small subnetwork that connects a subset of the disease-related proteins with high-confidence paths. These paths typically reveal additional proteins termed 'Steiner nodes' that were not initially implicated as disease proteins but are useful in forming concise, trusted connections among the disease proteins. The discovered subnetwork is a forest, a collection of trees.

Formally, the PPI network is represented as a weighted graph $G(V, E)$ where V is the set of proteins and E is the set of interactions between those proteins. A cost function assigns a cost $c(e) > 0 \quad \forall e \in E$ and a prize function P assigns prizes $p(v) \in \mathbb{R} \quad \forall\, v \in V$. Prizes are derived from biological data such as gene expression or quantitative proteomic data. $p(v) > 0$ indicates that the protein is biologically altered and should be included in the Steiner forest, if possible, with the magnitude indicating the degree of relevance to the disease or condition. $p(v) = 0$ denotes that there is no observed data for vertex v or no prior reason to believe it is relevant to the disease, and such vertices compose the potential Steiner nodes. The original PCSF optimization problem [16] is defined as $\underset{F}{\operatorname{argmin}}\, o(F)$ where

$$o(F) = \beta \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) + \omega\kappa \tag{1}$$

where V_F and E_F are the vertices and edges of the forest F and κ is the number of trees in the forest. β is a parameter that controls the tradeoff between including prizes and avoiding expensive edges, and ω is a parameter that controls how many distinct trees are in the forest.

A PCSF instance can be transformed into a prize-collecting Steiner tree (PCST) instance

by adding an artificial vertex v_0 that must be included in the Steiner tree and artificial edges $E_0 = V \times \{v_0\}$ with $c(e) = \omega \quad \forall e \in E_0$ [16]. Without loss of generality we can instead connect v_0 only to prize nodes, vertices for which $p(v) > 0$, because in an optimal solution any tree connected to v_0 must contain at least one prize. PCST is NP-hard so we recover an approximate solution using an efficient message-passing algorithm [13] that performs very well on benchmarks [20] and has been shown to be optimal in certain cases [20]. From the approximate PCST solution, we solve the original PCSF instance by deleting v_0 and its incident edges. In all analyses here, we set $\omega = 1.0$ to bias toward solutions with few connected components.

2.2. *Multi-sample prize-collecting Steiner forest*

The original PCSF formulation is designed for a single set of prizes from a single sample, condition, or patient. However, in many settings there are multiple samples that are expected to have some common properties even though the prizes may be very heterogeneous across the samples. This is particularly the case when the data are derived from patients who suffer from the same disease. In these cases, we would like to find a middle ground between two extremes. On the one hand, treating each patient in isolation ignores valuable data that can more accurately identify the common disease pathway. On the other, if we merge all the patient data, we miss patient-specific aspects of the disease. To address this challenge, we introduce the multi-sample prize-collecting Steiner forest (Multi-PCSF) problem.

We define 'artificial prizes' ϕ (described below) that are derived from the frequency at which a node is included in forests for all the samples. By adding ϕ to the sample-specific prizes, we introduce a link that constrains the forests to be similar but not identical. Below we introduce two alternative definitions for ϕ, one that tends to increase precision and one that promotes recall, and provide an algorithm to solve the Multi-PCSF problem.

Without loss of generality we assume that PCSF instances are transformed to PCST instances as described above. We further assume that β does not change during the execution of the algorithm, which allows us to redefine $p(v) = \beta\hat{p}(v)$ before execution, where $\hat{p}(v)$ are the original prizes from the biological data. We can then simplify Equation 1 to

$$o(F) = \sum_{v \notin V_F} p(v) + \sum_{e \in E_F} c(e) \tag{2}$$

which is a PCST instance whose solution can be transformed into a PCSF solution.

In the Multi-PCSF setting we have N samples and each sample $i \in \{1, \ldots, N\}$ has its own prize function P_i. The goal is to learn a collection of forests $\mathbf{F} = \{F_1, \ldots, F_N\}$ that are constrained to be similar to one another yet still reflect the diversity of the prizes in each sample. We expand the objective to create a joint objective function over the collection of forests $\mathbf{F}$ and solve $\underset{\mathbf{F}}{\operatorname{argmin}}\, o(\mathbf{F})$ where

$$o(\mathbf{F}) = \sum_{i=1}^{N} o(F_i) + \lambda \sum_{i=1}^{N} \sum_{v \notin V_{F_i}} \phi\left(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}\right) \tag{3}$$

The term $o(F_i)$ refers to the single forest objective function (Equation 2). The function ϕ is a new term that promotes similarity among all $F_i \in \mathbf{F}$ by introducing artificial prizes. The

parameter λ controls the tradeoff between F_i that is similar to the other forests versus F_i that concisely explains the observed data for tumor sample i. The role of λ is similar to how β controls the tradeoff between prizes and edge costs in the original PCST formulation.

The first of the two definitions of ϕ uses positive artificial prizes

$$\phi\left(\alpha, v, p(v), \mathbf{F}\right) = \begin{cases} \left(\frac{\sum_{i=1}^{|\mathbf{F}|} \mathbb{1}(v \in V_{F_i})}{|\mathbf{F}|}\right)^{\alpha}, & \text{if } p(v) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

The positive artificial prizes provide rewards for including nodes that are common to many other forests. $\mathbb{1}(v \in V_{F_i})$ is an indicator function that has the value 1 if forest F_i contains vertex v. The artificial prize on v is therefore determined by the fraction of other forests that contain v. The parameter α allows for non-linear relationships between the fraction and the artificial prize. As α grows, the vertices that are in many other forests will have larger artificial prizes relative to the vertices in few other forests.

To optimize Equation 3 we iteratively refine our estimates of the optimal forest for each sample given all other samples' current forests for a fixed number of iterations (five here) or until $\mathbf{F}$ converges. At the first iteration we set $\lambda = 0$ so that each optimal F_i is independent of $F_j \quad \forall i \neq j$ because there is no similarity constraint imposed. At all subsequent iterations, we update each F_i individually in a sequential random order using the fixed current estimate of all $\mathbf{F} \setminus \{F_i\}$. Below we show how to update F_i by formulating a new PCST instance with modified prizes. To derive the modified prizes we consider only the ith term of each summation in Equation 3 to approximately solve $\underset{F_i}{\operatorname{argmin}}\, o_i(\mathbf{F})$.

$$\begin{aligned} o_i(\mathbf{F}) &= o(F_i) + \lambda \sum_{v \notin V_{F_i}} \phi\left(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}\right) \\ &= \sum_{v \notin V_{F_i}} p_i(v) + \sum_{e \in E_{F_i}} c(e) + \lambda \sum_{v \notin V_{F_i}} \phi\left(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}\right) \\ &= \sum_{v \notin V_{F_i}} \left(p_i(v) + \lambda \phi\left(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}\right)\right) + \sum_{e \in E_{F_i}} c(e) \end{aligned} \tag{5}$$

By substituting the definition of $o(F_i)$ from Equation 2 into Equation 5 and rearranging the terms we can define a new prize function P_i' that adds artificial prizes to the original P_i

$$\begin{aligned} p_i'(v) &= p_i(v) + \lambda \phi\left(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}\right) \\ &= \begin{cases} \lambda \left(\frac{\sum_{i=1}^{|\mathbf{F} \setminus \{F_i\}|} \mathbb{1}(v \in V_{F_i})}{|\mathbf{F} \setminus \{F_i\}|}\right)^{\alpha}, & \text{if } p_i(v) = 0 \\ p_i(v), & \text{otherwise} \end{cases} \end{aligned} \tag{6}$$

We obtain the new PCST instance that can be solved as described in Section 2.1.

$$o_i(\mathbf{F}) = \sum_{v \notin V_F} p'(v) + \sum_{e \in E_F} c(e) \tag{7}$$

The alternative definition of ϕ uses negative artificial prizes, which encourage the algorithm to exclude potential Steiner nodes that appear in few other forests. We define

$$\phi\left(\alpha, v, p(v), \mathbf{F}\right) = \begin{cases} -\left(\frac{\sum_{i=1}^{|\mathbf{F}|} \mathbb{1}(v \notin V_{F_i})}{|\mathbf{F}|}\right)^{\alpha}, & \text{if } p(v) = 0 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

The algorithm is otherwise identical except the updated prize function P_i' becomes

$$\begin{aligned} p_i'(v) &= p_i(v) + \lambda\phi\left(\alpha, v, p_i(v), \mathbf{F} \setminus \{F_i\}\right) \\ &= \begin{cases} -\lambda\left(\frac{\sum_{i=1}^{|\mathbf{F}\setminus\{F_i\}|} \mathbb{1}(v \notin V_{F_i})}{|\mathbf{F} \setminus \{F_i\}|}\right)^{\alpha}, & \text{if } p_i(v) = 0 \\ p_i(v), & \text{otherwise} \end{cases} \end{aligned} \tag{9}$$

2.3. *Simulated data*

In our first analysis, we generated a synthetic scale-free PPI network using the Barabási-Albert preferential attachment model [21] with 1000 total nodes, 10 initial nodes, and 10 edges per new node attached (9900 total edges). We created artificial pathways by initiating a depth-first search from a randomly selected root node in the graph. The search visited at most two children per node up to a maximum depth of five. Given a pathway with m nodes and parameters f (pathway fraction) and n (noise level), we simulated patients by selecting $\lceil fm \rceil$ prizes from the pathway and $\lceil n\lceil fm \rceil \rceil$ noisy prizes (nodes that are not on the pathway) as mutated genes. For example, if we have a 1000 node network, $m = 25$, $f = 0.25$, and $n = 2.0$, we would randomly select 7 pathway members as true prizes and another 14 nodes from the 975 that are not pathway members as noisy prizes for each patient. All edges had a cost of 0.1, and we assigned a prize of 1.0 to all mutated genes.

We tested our Multi-PCSF algorithm under a variety of parameter configurations and for various f and n (Section 3.1). We varied one parameter at a time and set all others to their default value (Table 1). For all configurations we tested positive and negative artificial prizes. In each Multi-PCSF run, we simulated 25 patients per pathway and calculated the precision and recall (Equation 10) for each forest.

Table 1. Multi-PCSF parameters

Parameter	Values tested	Default
α	$1, 2, 3$	2
β	$0.25, 0.5, 1.0$	0.5
λ	$0.5, 1.0, 2.0$	1.0
f	$0.1, 0.25, 0.5, 1.0$	0.25
n	$0, 0.5, 1.0, 2.0$	0.5

$$\text{precision} = \frac{\text{correct predictions}}{\text{total predictions}} \qquad \text{recall} = \frac{\text{correct predictions}}{\text{pathway members}} \tag{10}$$

2.4. *Human data*

We evaluated Multi-PCSF using two types of human data: canonical pathways and breast cancer data from 98 patients. For both human analyses we used physical PPI from STRING (version 9.0) [22]. Using the edge scores $s(e)$ from STRING, we removed low confidence interactions with $s(e) < 0.5$ and defined edge costs as $\max(0.01, 1 - s(e))$. We downloaded the 'Epidermal Growth Factor Receptor Pathway' (EGFR) from the *Science Signaling* Database of Cell Signaling [23], translating all pathway node names into gene symbols. Three non-protein nodes could not be mapped and retained their original names. We selected only a single gene symbol per gene family to maintain the original pathway topology. We downloaded National Cancer Institute-Nature Pathway Interaction Database (PID) pathways [24] and mapped UniProt ids to gene symbols. To calculate P-values for PID pathway enrichment, we used the right-tailed Fisher's exact test. All P-values were corrected for multiple hypothesis testing by multiplying them by the number of hypotheses tested (Bonferroni correction).

We obtained The Cancer Genome Atlas (TCGA) breast cancer data from the Broad Institute's Genome Data Analysis Center Firehose (April 21, 2013 analysis run). We considered only the 98 basal-like tumors defined in [25]. For each tumor i, we defined the prize on a gene to be $p_i(g) = p_i^m(g) + p_i^p(g)$ where $p_i^m(g)$ is the number of non-silent mutations or indels in gene g and $p_i^p(g)$ denotes proteomic changes in the reverse phase protein array data. If an antibody exhibited a $\log_2$ scale fold change with magnitude of at least 1.0, we set $p_i^p(g)$ to be that magnitude and took the maximum magnitude when multiple antibodies mapped to a single gene. To simulate 100 patients in the EGFR pathway, we set $f = 0.25$ and $n = 10.0$ and generated noisy prizes as described above. We used $\alpha = 2$, $\beta = 1.0$, and $\lambda \in \{0.5, 1.0, 2.0, 5.0\}$. For the breast cancer analysis we set $\alpha = 2$, $\beta = 0.5$, and $\lambda = 1.0$.

2.5. *HotNet analysis*

We ran generalized HotNet (version 1.0.0) [5, 26], which takes a gene-gene influence matrix and a score on genes as input. We used the influence matrix packaged with HotNet, which is derived from the Human Protein Reference Database (HPRD) PPI network [27], and set the gene score to be $\sum_{i=1}^{N} p_i(g)$ where N is the number of basal-like breast cancer tumors. We allowed HotNet to choose the optimal δ parameter, which it selected as $\delta = 0.05$, and used all other default parameters (1000 permutations, smin of three, and smax of ten). We defined 'HotNet PID pathways' as the five PID pathways that most significantly overlapped a HotNet subnetwork, which happened to be the same 864-gene HotNet subnetwork for all five.

3. Results

We tested Multi-PCSF in three increasingly challenging settings to demonstrate how sharing information across samples improves pathway recovery for each individual sample. In the first two test cases, we generated prizes from a known reference pathway and quantified how well

the pathway was recovered. In the third, we analyzed data from 98 patients with basal-like breast cancer tumors and showed that Multi-PCSF produces individualized representations of the signaling pathways that are perturbed in this breast cancer subtype.

3.1. *Recovering simulated pathways*

In order to quantitatively evaluate whether Multi-PCSF improves pathway recovery, we first simulated prizes for cancer samples with a common driver pathway. We simulated a 1000 node scale-free network, which reflects the topology of real PPI networks [28] and allowed us to run Multi-PCSF under a wide range of parameter configurations (solving 32500 PCST instances) to ensure its advantages are not limited to specific settings. We generated a driver pathway that would be altered in each tumor. We then randomly assigned prizes in each synthetic tumor sample to a fraction of the pathway members as well as a fraction of other proteins that are not on the pathway, which represent noisy passenger mutations. We ran baseline PCSF (which does not share information across samples) and Multi-PCSF and calculated precision and recall (Equation 10) for the nodes and edges of each forest. We assessed the average performance over ten synthetic pathways (Figure 1).

With very few exceptions, Multi-PCSF improves both the precision and recall under all tested parameter configurations. The improvements in recall, how much of the reference path-

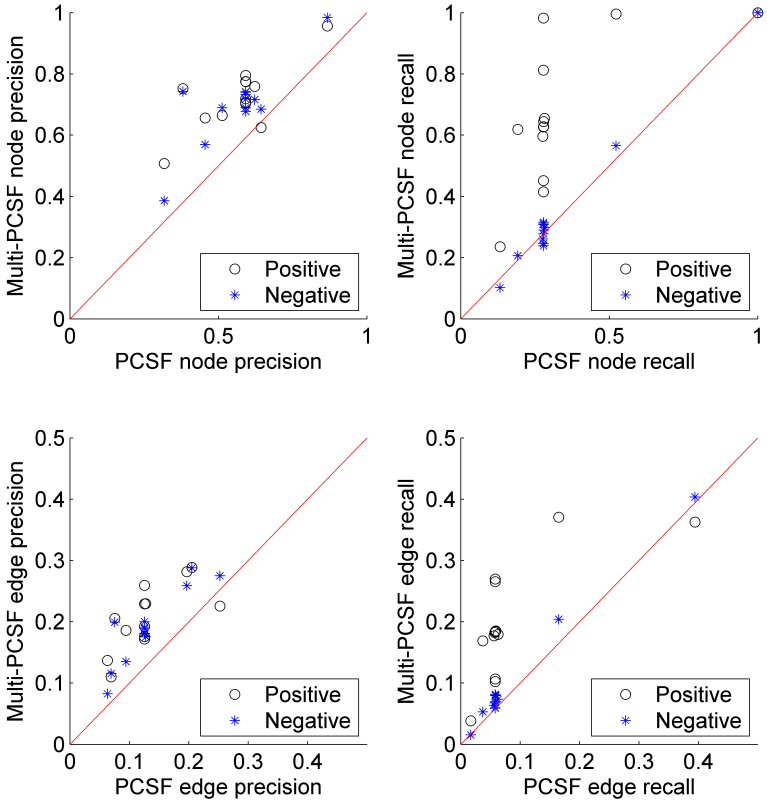

Fig. 1. Node and edge precision and recall for Multi-PCSF versus PCSF on simulated pathways. Positive and negative refer to the Multi-PCSF artificial prizes. Points above the red diagonal indicate instances where Multi-PCSF outperforms PCSF.

way is recovered, are especially notable. In the best case Multi-PCSF node recall is 3.5 times greater than PCSF and edge recall is 4.6 times greater. On this instance PCSF node recall is 0.28 signifying that for most synthetic tumors the prize nodes are the only pathway members that could be recovered. Multi-PCSF node recall is 0.98 — in most cases the entire pathway could be recovered. Positive artificial prizes yield greater improvements in recall than negative artificial prizes. With positive prizes, Multi-PCSF includes proteins that are shared by many other forests even if they are not needed to connect additional prize nodes. Conversely, with negative prizes Multi-PCSF is more likely to use such nodes as Steiner nodes but will not include them in a forest unless they help connect prize nodes.

3.2. *Recovering the EGFR signaling pathway*

Having established that Multi-PCSF can substantially improve pathway recovery in a simulated setting, we assessed its performance in a human PPI network. We selected the human EGFR pathway as the hypothetical driver pathway that was perturbed in a cohort of simulated tumors and applied both Steiner forest algorithms. The randomly generated prizes in this setting were much noisier than in the simulated pathway setting to better reflect the large number of passenger alterations per driver mutation in real cancer datasets. For every functional prize selected from the EGFR pathway, we added ten noisy prizes from elsewhere in the PPI network. We simulated 100 tumor samples, ran PCSF and Multi-PCSF, and calculated precision and recall (Figure 2). For Multi-PCSF we varied λ, which controls the strength of the constraint that requires forests to be similar to one another.

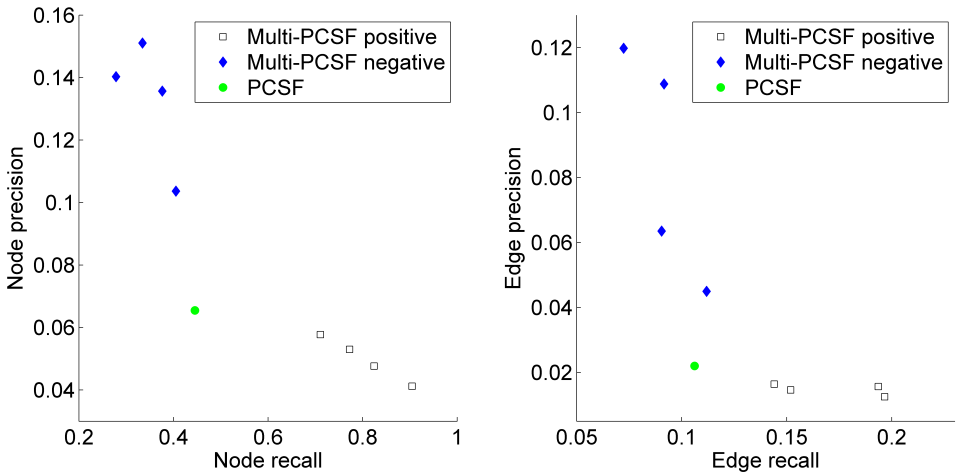

Fig. 2. Precision-recall graphs for Multi-PCSF with positive and negative artificial prizes and baseline PCSF on the human EGFR pathway. The four Multi-PCSF points correspond to different values of λ.

In the EGFR setting, PCSF node precision is only 0.065 and edge precision is 0.022 because even the noisy prizes could often be connected to the EGFR pathway members. By sharing information across samples, Multi-PCSF is better able to discern which prizes are spurious and which potential Steiner nodes are preferable because they are perturbed in other samples. With positive artificial prizes, proteins that are members of other forests (as either prize nodes

or Steiner nodes) are introduced as Steiner nodes. This enhances recall, which increases with λ, culminating in a 2.0 times improvement in node recall and 1.9-fold edge recall improvement when $\lambda = 5.0$. The maximum node recall attained is 0.90. Even in this difficult setting, nearly all proteins on the pathway can be extracted from the PPI network at the expense of a decrease in precision. Parallel paths in the EGFR pathway cannot be captured by our inferred forests, which suggests that edge recall could potentially be further improved by applying perturbation techniques that merge multiple forests and produce more general topologies [16].

With negative artificial prizes, Multi-PCSF excludes proteins that are not useful in other forests, which boosts precision. When $\lambda = 5.0$ and negative prizes are used, Multi-PCSF node precision is 2.1 times greater than PCSF and edge precision is 5.4 times greater. In addition, when using a weaker similarity constraint ($\lambda = 0.5$), Multi-PCSF exhibits superior precision as well as a small improvement in edge recall.

3.3. *Pathways in breast cancer*

To assess Multi-PCSF's ability to interpret and suggest mechanistic hypotheses about real clinical data we applied it to TCGA breast cancer data [25], inferring the pathways perturbed in these tumors and their common and unique components. Because cancer subtypes defined by mRNA expression similarity are likely to share common driver pathways, we focus on only the basal-like breast cancer subtype (98 tumors). We calculated prizes using the TCGA non-silent mutations and proteomic data. Other data types such as copy number alterations can easily be integrated into our analysis, and we have previously shown how to combine epigenomic features and mRNA expression to place prizes on transcription factors [17]. Some of the tumors had sparse prizes so we used positive artificial prizes in Multi-PCSF to leverage its ability to construct more complete pathways based on alterations in other tumors.

Multi-PCSF achieves our goal of discovering pathways that have a common core structure and many individual characteristics connected to the core that reflect the diverse manners in which the driver pathways are affected in each tumor (Figure 3). The shared core is composed of 198 nodes (8.30% of all nodes appearing in any forest) that are present in all 98 forests. This core likely contains pathways that are altered in all patients despite their heterogeneous mutations. For example, we recover basal-like breast cancer-related proteins such as ATM, BRCA1, BRCA2, MYC, RB1, and TP53 [25]. In addition, we find HIF1A in the common core, consistent with the fact that high HIF1A pathway activity is a key feature of basal-like breast cancers [25]. By jointly analyzing all patients we find potential therapeutic targets that would have been missed in individual analyses. Two genes, ARHGDIA and SMAD2, do not appear in any forests when PCSF is run independently on each sample but appear in the Multi-PCSF common core. ARHGDIA encodes the protein RhoGDI-1, which is overexpressed in breast cancer and blocks chemotherapy drug-induced apoptosis in cancer cells [29]. SMAD2 knockdowns in breast cancer cells suggest it is a tumor suppressor [30].

Although many nodes are identical across the forests, the edges used to connect those nodes to each other vary as only 39 edges (1.36%) are common to all forests. Beyond the shared core, 1411 nodes (59.14%) and 1712 edges (59.55%) appear in only one forest. 917 nodes are Steiner nodes in at least one forest, including all nodes in the common core and 435 nodes

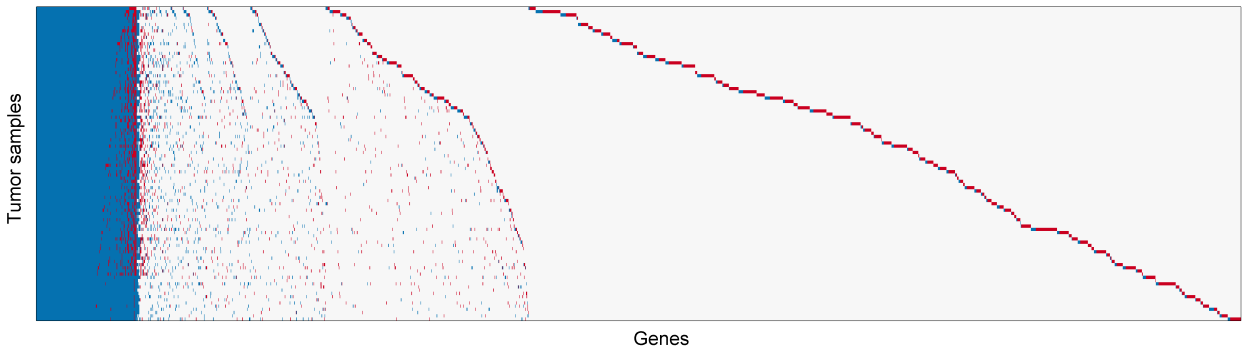

Fig. 3. The heat map summarizes all 98 Multi-PCSF forests. Each row represents the forest for a particular tumor sample and depicts which nodes are collected prizes (red), Steiner nodes (blue), and absent (white).

that are present in multiple but not all forests. The variation among the forests demonstrates that even tumors within a single subtype cannot be represented by a single pathway structure.

HotNet [5, 26] is an algorithm for discovering PPI subnetworks that are significantly affected in a cancer dataset. We applied generalized HotNet to the basal-like tumor data, providing HotNet's HPRD-derived gene-gene influence matrix and the same mutation- and proteomic-based prizes as input. HotNet returned 109 subnetworks. One large subnetwork contained 864 proteins and the other subnetworks had two to seven members. HotNet's subnetworks significantly overlap PID pathways (Table 2), which we refer to as HotNet PID pathways (Section 2.5), demonstrating that HotNet can reveal which reference pathways are relevant in a cancer subtype. However, because it produces a single list of subnetworks for the entire subtype and does not reveal hidden pathway members (the equivalent of Steiner nodes in PCSF), it is difficult to use HotNet to generate mechanistic hypotheses or guide individualized treatment. Although HotNet would produce different results if we tune its parameters to generate smaller subnetworks or use an influence matrix derived from the STRING PPI network, these fundamental differences between HotNet and Multi-PCSF would remain.

Multi-PCSF not only recovers forests that capture the same annotated pathways as HotNet, but it also presents custom versions of those pathways for each tumor, which better

Table 2. HotNet PID pathways and whether they significantly overlap Multi-PCSF forests, PCSF forests, or both (corrected $P \leq 0.05$). If both, the table shows whether the overlap is better or worse for Multi-PCSF.

HotNet PID pathway	HotNet subnetwork overlap corrected P	Only Multi-PCSF	Better Multi-PCSF	Worse Multi-PCSF	Only PCSF
SHP2 signaling	9.36 E-10	65	33	0	0
IL2-mediated signaling events	2.97 E-9	36	62	0	0
Signaling events mediated by Stem cell factor receptor (c-Kit)	3.08 E-9	29	69	0	0
Integrins in angiogenesis	7.80 E-9	60	38	0	0
GMCSF-mediated signaling events	4.23 E-8	45	53	0	0

enables follow-up biological analysis. In many cases standard PCSF does not recover the reference pathways affected in the basal-like subtype because it does not leverage data from related tumors. For all tumors where the PCSF forest is significantly enriched with a PID pathway, the enrichment is stronger after sharing information with Multi-PCSF (Table 2). Individualized representations of the PID pathways, such as 'Signaling events mediated by Stem cell factor receptor (c-Kit)', could potentially lead to new therapeutic strategies for subsets of the basal-like breast cancer cases. KIT abnormalities have been implicated in several other cancers [31], and KIT-positive gastrointestinal stromal tumors have been approved for Gleevec (imatinib) treatment [32]. Post-processing procedures for prioritizing Steiner tree members have shown that highly-ranked Steiner nodes validate *in vitro* [17] and can be applied here to guide subsequent analysis of the individual pathway predictions.

4. Discussion

The prize-collecting Steiner forest algorithm is a powerful approach for integrating genomic, proteomic, transcriptional, and epigenomic data to reconstruct signaling pathways. Our multi-sample extension enables PCSF to analyze heterogeneous data, where prizes vary greatly across a collection of samples, and to exploit information from related samples despite the prize-level dissimilarities. Multi-PCSF is an especially pertinent tool for large-scale cancer profiling studies because the most frequently recurring alterations have already been identified (leaving the non-recurrent abnormalities for further interpretation) and we seek to understand the unique causes of oncogenesis in each tumor. The artificial prizes introduced in Multi-PCSF facilitate constructing accurate patient-specific driver pathways despite the presence of numerous passenger mutations by promoting genes that are driver pathway members in other tumors. Algorithms like HotNet can reveal which processes are affected in a patient cohort but do not guide individualized treatment (although recent diffusion-based algorithms [33] aim to lift this limitation). Multi-PCSF is also widely applicable beyond cancer and can model data from noisy biological replicates without initially aggregating all replicates, study responses to a collection of stimuli [34], or compare the immune responses to related viruses [15].

Acknowledgements

We thank Nurcan Tuncbag and Fabrizio Altarelli for discussions about Steiner forests as well as Anthony Soltis and Sara Gosline for preparing network data. This work was supported in part by the 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), by NIH grant U54-CA112967, and by European Grants FET Open No. 265496 and ERC No. 267915, as well as computing resources funded by the National Science Foundation under Award No. DB1-0821391.

References

[1] B. Vogelstein, N. Papadopoulos, V. E. Velculescu, S. Zhou, L. A. Diaz and K. W. Kinzler, *Science* **339**, 1546 (2013).

[2] M. B. Yaffe, *Sci Signal* **6**, pe13 (2013).
[3] U. D. Akavia, O. Litvin, J. Kim, F. Sanchez-Garcia, D. Kotliar, H. C. Causton, P. Pochanard, E. Mozes, L. A. Garraway and D. Pe'er, *Cell* **143**, 1005 (2010).
[4] E. Cerami, E. Demir, N. Schultz, B. S. Taylor and C. Sander, *PLoS ONE* **5**, e8918 (2010).
[5] F. Vandin, E. Upfal and B. J. Raphael, *J Comput Biol* **18**, 507 (2011).
[6] G. Ciriello, E. Cerami, C. Sander and N. Schultz, *Genome Res* **22**, 398 (2012).
[7] J. Zhao, S. Zhang, L.-Y. Wu and X.-S. Zhang, *Bioinformatics* **28**, 2940 (2012).
[8] M. D. M. Leiserson, D. Blokh, R. Sharan and B. J. Raphael, *PLoS Comput Biol* **9**, e1003054 (2013).
[9] A. J. Sedgewick, S. C. Benz, S. Rabizadeh, P. Soon-Shiong and C. J. Vaske, *Bioinformatics* **29**, i62 (2013).
[10] A. Gitter, M. Carmi, N. Barkai and Z. Bar-Joseph, *Genome Res* **23**, 365 (2013).
[11] R. Brosh and V. Rotter, *Nat Rev Cancer* **9**, 701 (2009).
[12] C.-H. Yeang, T. Ideker and T. Jaakkola, *J Comput Biol* **11**, 243 (2004).
[13] M. Bailly-Bechet, C. Borgs, A. Braunstein, J. Chayes, A. Dagkessamanskaia, J.-M. François and R. Zecchina, *Proc Natl Acad Sci* **108**, 882 (2011).
[14] Y.-A. Kim, S. Wuchty and T. M. Przytycka, *PLoS Comput Biol* **7**, e1001095 (2011).
[15] A. Gitter and Z. Bar-Joseph, *Bioinformatics* **29**, i227 (2013).
[16] N. Tuncbag, A. Braunstein, A. Pagnani, S.-S. C. Huang, J. Chayes, C. Borgs, R. Zecchina and E. Fraenkel, *J Comput Biol* **20**, 124 (2013).
[17] S.-s. C. Huang, D. C. Clarke, S. J. C. Gosline, A. Labadorf, C. R. Chouinard, W. Gordon, D. A. Lauffenburger and E. Fraenkel, *PLoS Comput Biol* **9**, e1002887 (2013).
[18] N. Atias and R. Sharan, *Mol BioSyst* **9**, 1662 (2013).
[19] S. J. Pan and Q. Yang, *IEEE Trans Knowl Data Eng* **22**, 1345 (2010).
[20] I. Biazzo, A. Braunstein and R. Zecchina, *Phys Rev E* **86**, 026706 (2012).
[21] A.-L. Barabási and R. Albert, *Science* **286**, 509 (1999).
[22] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen and C. v. Mering, *Nucleic Acids Res* **39**, D561 (2011).
[23] N. R. Gough, *Ann NY Acad Sci* **971**, 585 (2002).
[24] C. F. Schaefer, K. Anthony, S. Krupa, J. Buchoff, M. Day, T. Hannay and K. H. Buetow, *Nucleic Acids Res* **37**, D674 (2009).
[25] The Cancer Genome Atlas Network, *Nature* **490**, 61 (2012).
[26] F. Vandin, P. Clay, E. Upfal and B. J. Raphael, *Pac Symp Biocomput* , 55 (2012).
[27] T. S. K. Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. H. Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady and A. Pandey, *Nucleic Acids Res* **37**, D767 (2009).
[28] H. Jeong, S. P. Mason, A. L. Barabási and Z. N. Oltvai, *Nature* **411**, 41 (2001).
[29] B. Zhang, Y. Zhang, M.-C. Dagher and E. Shacter, *Cancer Res* **65**, 6054 (2005).
[30] M. Petersen, E. Pardali, G. van der Horst, H. Cheung, C. van den Hoogen, G. van der Pluijm and P. ten Dijke, *Oncogene* **29**, 1351 (2010).
[31] J. Lennartsson and L. Rönnstrand, *Physiol Rev* **92**, 1619 (2012).
[32] H. Joensuu, *Nat Rev Clin Oncol* **9**, 351 (2012).
[33] E. O. Paull, D. E. Carlin, M. Niepel, P. K. Sorger, D. Haussler and J. M. Stuart, *Bioinformatics* (2013).
[34] S. J. C. Gosline, S. J. Spencer, O. Ursu and E. Fraenkel, *Integr Biol* **4**, 1415 (2012).