DISCOVERY OF MUTATED SUBNETWORKS ASSOCIATED WITH CLINICAL DATA IN CANCER

FABIO VANDIN, PATRICK CLAY, ELI UPFAL, BENJAMIN J. RAPHAEL

Department of Computer Science, and Center for Computational Molecular Biology, Brown University, Providence, RI 02912, U.S.A.

E-mail: {vandinfa,pclay,eli,braphael}@cs.brown.edu

A major goal of cancer sequencing projects is to identify genetic alterations that determine clinical phenotypes, such as survival time or drug response. Somatic mutations in cancer are typically very diverse, and are found in different sets of genes in different patients. This mutational heterogeneity complicates the discovery of associations between individual mutations and a clinical phenotype. This mutational heterogeneity is explained in part by the fact that driver mutations, the somatic mutations that drive cancer development, target genes in cellular pathways, and only a subset of pathway genes is mutated in a given patient. Thus, pathway-based analysis of associations between mutations and phenotype are warranted. Here, we introduce an algorithm to find groups of genes, or pathways, whose mutational status is associated to a clinical phenotype *without* prior definition of the pathways. Rather, we find subnetworks of genes in an gene interaction network with the property that the mutational status of the genes in the subnetwork are significantly associated with a clinical phenotype. This new algorithm is built upon HotNet, an algorithm that finds groups of mutated genes using a heat diffusion model and a two-stage statistical test. We focus here on discovery of statistically significant correlations between mutated subnetworks and patient survival data. A similar approach can be used for correlations with other types of clinical data, through use of an appropriate statistical test. We apply our method to simulated data as well as to mutation and survival data from ovarian cancer samples from The Cancer Genome Atlas. In the TCGA data, we discover nine subnetworks containing genes whose mutational status is correlated with survival. Genes in four of these subnetworks overlap known pathways, including the focal adhesion and cell adhesion pathways, while other subnetworks are novel.

1. Introduction

A major goal of cancer sequencing projects such as The Cancer Genome Atlas is to identify genetic and epigenetic alterations that determine clinical phenotypes, such as survival time or drug response. There are a number of reports of genes whose mutational status is associated with survival such as KRAS mutations,[1] EGFR amplifications,[2] and PTEN mutations.[3] Despite the rapid increase in catalogs of somatic mutations in cancer genomes,[4,5] progress in determining which mutations, or mutations in which genes, determine clinical phenotypes remains slow. The difficulty is due in part to the extensive *mutational heterogeneity* exhibited by cancer genomes, where the somatic mutations or mutated genes vary widely across patients. This mutational heterogeneity is a consequence of two features of the somatic mutation process in cancer. First, the somatic mutations in each cancer genome are a mixture of functional *driver* mutations responsible for cancer, and random *passenger* mutations that accumulate during tumor progression but are inconsequential for cancer. Second, driver mutations target not just single genes, but also groups of genes in signaling or regulatory pathways.[6,7] Thus, different patients may have different subsets of driver mutations in key pathways, and thus driver mutations are distributed over a large number of genomic locations in many different

genes. A natural approach for finding clinically relevant mutations is first to determine driver mutations, and then to test each of these for clinical association. However, the problem of distinguishing driver from passenger mutations is itself a challenge. Many cancer genome studies attempt to predict driver mutations (or mutated genes) by finding those with a statistical significant frequency of occurrence in a large cohort of patients. But the power of this approach is reduced by mutational heterogeneity. Moreover, it may be mutations within a pathway, and not a single gene, that determine a clinical phenotype in a heterogeneous cohort of cancer patients.

A more powerful approach to test associations between mutation and phenotype is to test these associations at the pathway level, rather than at the level of single mutation or single genes. Ideally, one would perform such a test using collection of all relevant biological pathways. No such comprehensive collection currently exists, although databases such as KEGG[8] , Reactome,[9,10] and others are important efforts in this direction. An alternative source of information is genome-scale interaction networks that record (binary) interactions between proteins. Examples of such data sources for human are STRING[11] and HPRD.[12] Interaction networks have proven to be a useful source of information for analyzing genomic data, particularly gene expression data. Chuang *et al.*[13] introduced a method to find subnetworks of interaction networks whose gene expression predicts progression to metastasis in breast cancer. This work extends an earlier approach to find differentially expressed subnetworks introduced in Ideker *et al.*[14] and extended in Dittrich *et al.*[15] and in Beisser *et al.*[16] Also, Vaske *et al.*[17] present a method to infer patient-specific genetic activities. The latter method relies on curated pathway interactions among genes and predict the degree of alteration of pathway's activities in the patient using probabilistic inference. More recently, methods to discovery mutated subnetworks have been introduced.[18,19]

Here we introduce an approach to find subnetworks of genes in an interaction network with the property that mutations in the genes in the subnetwork are correlated with a clinical parameter. Specifically, for the clinical parameter of survival time, we identify subnetworks such that the survival time of patients with mutations in the subnetwork is significantly different from patients with no mutation in the subnetwork. To the best of our knowledge this is the first approach to find subnetworks whose mutations are correlated with survival. Previous methods[13–16] utilized gene expression data and incorporate a variety of different scoring methods to identify subnetworks/modules and to compute their statistical significance. We accomplish this goal by extending the HotNet algorithm previously introduced by some of us.[18] Our algorithm represents association scores for individual genes as sources of "heat" on the vertices (genes) of the interaction network, and uses a heat-diffusion model to propagate heat to surrounding vertices. We extract "significantly hot" subnetworks with a statistical test that rigorously bounds the false discovery rate (FDR) on the derived subnetworks. We apply our method to finding mutated subnetworks correlated with survival in simulated data and ovarian cancer data from The Cancer Genome Atlas. In the TCGA data, we discover nine subnetworks containing genes whose mutational status is correlated with survival. Genes in four of these subnetworks overlap known pathways, including the focal adhesion and cell adhesion pathways, while other subnetworks are novel.

2. Methods

2.1. *Generalized HotNet*

Our new method* builds on the HotNet algorithm.[18] A schematic of our method is given in Figure 1. Suppose we are given a gene/protein interaction network $G = (V, E)$ where the vertices represent genes and each edge represents a (binary) interaction between pair of genes, or their corresponding proteins. We are also given a score σ_g for each gene g. Our goal is to find subnetworks of G whose combined scores are statistically significant. There are two challenges that must be addressed. First, it is not feasible to evaluate all possible subnetworks: the large number of subnetworks of a reasonable size (e.g. containing at least 5 genes) implies that testing each incurs large computational burden, and more importantly would require a severe multiple hypotheses testing correction. Second, the topology of the interaction network means that the individual subnetworks cannot be treated as independent hypotheses. We first describe how our method addresses the topology of the network and then describe the statistical methodology.

The subnetworks of interest are determined both by the scores of their genes, and the interactions between the genes. However, human interaction networks are fairly irregular, and typically include "hub genes" with very high degree. Thus, it is important to account for the topology of the interaction network when evaluating subnetworks. For example, consider the following two scenarios. In the first scenario, two genes with high score are connected by a single vertex of degree 2 in the network. In the second scenario, the two genes with high score are connected by a vertex of very high degree. Because there are many such paths through the high degree vertex, it is more likely to see genes with high score connected through a high degree node (the second scenario) than a low degree one (the first scenario). This situation generalizes to more complicated scenarios where neighbors of neighboring vertices may have widely varying degrees. To formalize the intuition behind these scenarios, the HotNet algorithm considers the score of a gene as a quantity of "heat" placed on the gene, and allow heat to diffuse over the edges. Heat placed on a vertex of low degree will diffuse slowly through the graph and thus the neighbors of the vertex will remain hot for long periods of time. In contrast, heat placed on a vertex of high degree will diffuse quickly to the neighbors and none of these vertices will be hot.

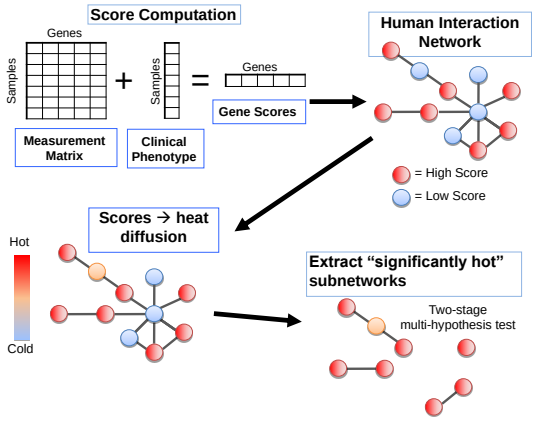


Fig. 1. Generalization of HotNet for clinical data.

HotNet uses a heat diffusion process and a statistical test to derive "significantly hot"

*An implementation of Generalized HotNet is available as a separate tool at: `http://cs.brown.edu/~braphael/software.html`

subnetworks. Significant subnetworks are thus determined by both the score of the genes in the subnetwork and the local topology of the subnetwork. HotNet consists of several steps. First, we use a heat diffusion process to derive a measure of influence between two genes in the interaction network. This measure depends only on the topology of the network. Next, the scores of genes are used to enhance the influence measure, defining the heat exchanged between any two pair of genes. Then, we remove cold edges (i.e., edges with low exchange of heat), dividing the network into subnetworks. Finally, we apply a two-stage statistical test to rigorously bound the false discovery rate (FDR) of the identified subnetworks. We briefly describe each of these steps. Additional details and analyses of the FDR bound are in Vandin *et al.*[18]

Let A be the adjacency matrix of the interaction network G. That is, let $A(i,j) = 1$ if genes g_i, g_j are connected by an edge in G, and let $A(i,j) = 0$ otherwise. Let D be a diagonal matrix with $D(i,j) = 0$ if $i \neq j$, and $D(i,i) = deg(i)$, where $deg(i)$ is the degree of g_i in the graph (i.e., the number of genes that interact with g_i in the network). The matrix $L = -A + D$ is called the Laplacian matrix of G and the matrix $H_t = e^{-Lt}$ is the *heat kernel* of the graph[20,21] for a real number $t \geq 0$. Here e^X denotes the matrix exponential: $e^X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$. The entry $H_t(i,j)$ represents the heat found on gene i when a unit of heat is placed on gene j and allowed to diffuse for time t. Note that a different diffusion model based on the equilibrium value of a diffusion with loss[22] was employed in the original HotNet algorithm.[18] Fix a value $t > 0$, and let heat diffuse on the interaction network for time t. We define an influence measure $i(u,v)$ between pairs of genes u and v in the interaction network as $i(u,v) = H_t(v,u)$. Thus, the influence $i(u,v)$ of gene u on gene v is the heat observed on v after a length of time t when there is only a unit source of heat on node u at time 0. Since this measure is asymmetric (i.e., for a general diffusion kernel $i(u,v) \neq i(v,u)$), we define a symmetrized influence as $\tilde{i}(u,v) = \min\{i(u,v), i(v,u)\}$. The influence is thus defined only by the topology of the network and the value of the parameter t, and is defined for all pairs of genes based on the heat observed on the vertices.

HotNet uses the score for each gene to enhance the influence measure and then breaks the graph into connected components based on this influence measure. In particular, let $\sigma(g)$ be the score for gene g. Then for each pair (u,v) of genes we define a weight $w(u,v) = \max\{\sigma(u), \sigma(v)\} \times \tilde{i}(u,v)$. Next, given a parameter δ, HotNet removes all the edges (u,v) with $w(u,v) < \delta$. HotNet returns as output the connected components identified after the removal of the edges.

HotNet returns a list of subnetworks, each containing at least s genes, and employs a two-stage statistical test to assess the statistical significance of the returned list of subnetworks. The first stage of the test computes a p-value for the *number* of subnetworks with at least s genes that are returned, for different values of s, under a suitable null hypothesis (see Section 2.2.3 for the description of the null hypothesis used for the results presented in this work). Since the statistic is the number of subnetworks with at least s genes, and the possible values of s (and thus the number of tested hypotheses) is bounded by the number of genes with a score, the number of hypothesis is bounded by a quantity that is much smaller than the number of possible subnetworks of the interaction network. We can thus determine an s such that the number of connected components of size $\geq s$ is significant, with a particular

p-value. This p-value measures the significance of the number of subnetworks of a minimum size s, but does not say which, if any, of the individual subnetworks is significant. The second stage of the test estimates the false discovery rate (FDR) for the subnetworks in the list. (See Ref. 18 for further details of the statistical test.) In summary, HotNet returns: (i) a list of subnetworks, each with at least s genes; (ii) a p-value for the observed number of subnetworks; (iii) an estimated FDR for the list.

2.2. *Adaptation to Clinical Data*

There are three major steps in adapting the HotNet algorithm to applications in clinical data analysis: (i) selection of scoring function $\sigma(g)$ for individual genes; (ii) selection of the parameters t and δ; (iii) determination of the null hypothesis distribution.

2.2.1. *Gene Scores*

The input to HotNet is a score $\sigma(g)$ for each gene g. In order to find associations between mutations in subnetworks and clinical phenotypes, we derive a score for each gene g from the p-value of a statistical test that measures the association between the mutational status of g and a clinical parameter of interest. The particular statistical test depends on the clinical parameter. For example, if the clinical parameter is categorical, (e.g., classes for "response to treatment") the χ^2 test for independence can be used. For survival analysis we use the logrank[23] test to assess the significance of the difference between the survival curve of the set $M(g)$ of patients where g is mutated and the survival curve of the set $\overline{M}(g)$ of patients with no mutations in g.

For a gene g, we define the score $\sigma(g) = -2\log_e p_g$, where p_g is the p-value of the statistical test. Note that with this choice of score, when $t = 0$ the heat on a subnetwork $S = \{g_1, \ldots, g_{|S|}\}$ is equal to $-2\sum_i \log_e p_{g_i}$. This sum correspond to the statistic of Fisher's Method for combining p-values for (independent) statistical tests[24] . When $t > 0$, the total heat on a subnetwork S will depend on the topology of S and the topology of the entire network. For example, if S does not contain any high degree node (e.g., it is a linear path) or if it is a dense subgraph (e.g., a clique), the heat on S will be close to the case $t = 0$, and thus to the statistic of Fisher's Method. However, when vertices in S are connected to many other vertices outside S (e.g. a linear path going through an high degree vertex), the heat on S will diffuse to many other vertices, and result in a reduction in the combined p-value of the genes in S. Note that the gene scores are not a function of the null hypothesis employed by HotNet to assess the significance of the discovered subnetworks, but depends on the nature of the clinical data analyzed.

2.2.2. *Selection of parameters t and δ*

As described in Section 2.1, the execution of HotNet depends on the choice of parameters: t, the length of time that heat diffuses, and δ, the threshold for removing cold edges; i.e. edges of weight less than δ. In this way HotNet divides the network into subnetworks.

The parameter t controls the distance at which the score of a gene will diffuse in the network. If $t = 0$ there is no diffusion, while for $t = +\infty$, the heat distribution reaches equilibrium where all genes have the same heat. Using simulations we studied how the influence $i(u, v)$

changes as the distance between u and v increases for different values of t (data not shown). We fixed $t = 0.1$ since with this choice nodes at different distances receives distinguished amounts of heat with the diffusion process.

We use the following procedure to choose δ. We generate 100 datasets with the distribution of the null hypothesis (see Section 2.2.3), and consider how the number of connected components of size at least s varies as a function of δ for small values of s (i.e., $s = 3, 4, 5$). Note that since we are considering datasets from the null distribution, the subnetworks observed should not be significant. We make a conservative choice, by choosing a first δ that gives the largest (average) number of subnetworks of size at least s. In our experiments we fix $s = 5$, but comparable values of δ are obtained for $s = 3, 4$. Since increasing the value δ corresponds to a more conservative test, we also consider values $\delta' \geq \delta$. In particular, we consider values $\delta' = 1.1\delta, 1.2\delta, \ldots, 2.0\delta$. Note that if we consider δ_1, δ_2 with $\delta_1 < \delta_2$, the subnetworks of size at least s found with δ_2 will all be part of the subnetworks of size at least s found with δ_1. Thus, increasing δ limits the output to subnetworks with larger enhanced influence.

2.2.3. *The Null Hypothesis Distribution*

The accuracy of HotNet depends on the appropriate choice of the null hypothesis distribution. Once this distribution is determined, we generate a large number of instances (1000 in our experiments) to estimate the p-values of the various events.

For survival data we use the following two null hypothesis:

(1) H_0^P: the mutation matrix is fixed and the survival data is permuted across the patients. That is, let c_i be the survival associated to patient i in the observed data, with $i \in \{1, \ldots, m\}$. In a dataset generated under the H_0^P null hypothesis the survival data for patient i is $c_{\pi(i)}$, where $\pi(i)$ is a permutation chosen uniformly at random among all permutations over the set $\{1, \ldots, m\}$.

(2) H_0^M: for each gene g, the set of patients in which g is mutated is chosen uniformly at random and independently of other mutations, preserving the frequency of mutation of the genes. That is, if g is mutated in f_g patients, in a dataset generated under the H_0^M null hypothesis g is mutated in a set of f_g patients chosen uniformly at random, independently of other mutations. (The survival data are not explicitly randomized, since randomizing mutations will already account for randomization of the survival data.)

The main difference between the two models is that H_0^M removes correlations between occurrences of mutations in the genes, which are preserved in H_0^P. Thus, H_0^M removes possible correlations between scores of genes in the network (since genes with correlated mutations have correlated scores). Note that the parameter δ depends on the particular null hypothesis chosen, since it is based on the distribution of the number of subnetworks of size at least s under the null hypothesis.

3. Results

We tested our algorithm on both simulated data and ovarian cancer data from The Cancer Genome Atlas (TCGA). For simulated data, we considered the patients as divided into two

classes, thus looking for groups of genes with different mutation status in the two classes. For cancer data we considered a test statistic for survival. Thus, we aim to find groups of genes whose mutation status was associated with patient survival.

3.1. *Simulated data*

We first assessed our method using simulated data. We simulated whole exome sequencing data coming from two classes C_1, C_2 containing 150 patients each. For example, C_1 is a class of patients that respond to a particular treatment, while C_2 is a class of patients that do not respond to the treatment. We assumed that a mutated subnetwork is only one possible cause of a patient being in class C_1. Thus, we *planted* non-random mutations into a subnetwork S consisting of the 4 genes JAG1, NOTCH1, MAML1, and CDK8 into 20% of the patients in C_1. In particular, for a subset P consisting of 20% of the patients of C_1 we mutated one of the genes in S. For all other patients and all other genes we generated mutations according to a background mutation rate of 1.7×10^{-6}, consistent with recent studies on cancer.[25] Note that the patients in P also contain random mutations (in all genes but the ones in S), and that the 4 genes in S were mutated at random in all patients but P.

For each gene g, we built a 2×2 contingency table in which the row variable is the mutation status (i.e., mutated or not) of g and the column variable is the class (i.e., C_1 or C_2) of a patient. We used a χ^2 test for the 2×2 contingency table to obtain the p-value of a gene. None of the four genes turned out to be significant after correction for multi-hypothesis testing. We then used the scores defined in Section 2.2.1 as input for the generalized version of HotNet. The only significant subnetwork reported by our algorithm corresponds to the set of 4 genes JAG1, NOTCH1, MAML1, CDK8 ($p \leq 0.01$, $FDR \leq 0.05$). This shows that our algorithm identifies a subnetwork associated with clinical data when none of the genes in the subnetwork were identified as significant when considered individually, and when the subnetwork is non-randomly mutated in only 20% of the patients of one class.

3.2. *Ovarian TCGA data*

We next considered mutation data from ovarian cancer patients from The Cancer Genome Atlas. For each of these patients, we marked a gene as mutated if a somatic point mutation (or small indel) was present (as measured by exome resequencing) or if a focal copy number aberration (CNA) was present (as measured by array copy number data). Thus, the somatic mutation data was reduced to an $m \times n$ binary mutation matrix measuring the mutation status of $m = 316$ patients for $n = 17301$ genes. Genes not mutated in any patient were removed. Moreover, we removed CNAs for which the sign of the aberration was not the same in at least 90% of the patients with the aberration. This data was the same used in the HotNet analysis of TCGA ovarian publication.[25] The clinical data we used is preliminary data for 266 of these 316 patients, and thus we restricted our attention to these 266 patients. We considered the overall survival (in months) that measures the time from the patient's first surgery to their last followup or death in months, and vital status values (LIVING or DECEASED) that describe if the data is censored or not. For the H_0^P null hypothesis, the overall survival and vital status were treated as combined survival unit.

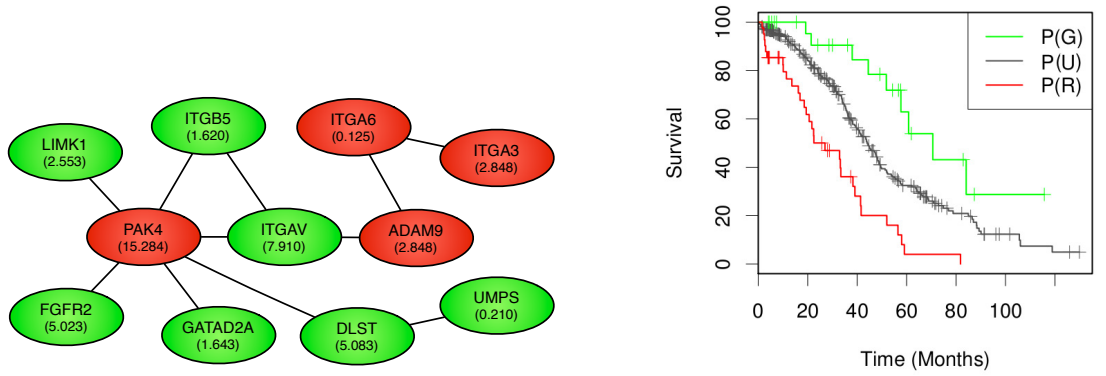
For each gene g, we considered the set $M(g)$ of patients in which g is mutated and the set $\overline{M}(g)$ of patients in which it is not mutated. We compared the Kaplan-Meier survival curve obtained for patients in $M(g)$ with the survival curve for patients in $\overline{M}(g)$. (We used the R package `survival` to compute the survival curves.) For each gene, we used the *logrank* statistical test to test the hypothesis that there is no difference between the two survival curves. In particular, we used the `survdiff` function of `survival` package in R to compute the p-value for each gene. To focus on genes having a non negligible effect on the set of patients analyzed, we removed genes that were mutated in fewer than five patients. For each p-value we derived a score as described in Section 2.2.1. We also removed outliers: five genes (C2orf65, DOK1, DQX1, LOXL3, and SEMA4F) with p-value less than 10^{-10}. The remaining scores constitute the input to HotNet. Using the procedure of Section 2.2.2 we computed a threshold δ for null hypothesis H_0^P and a threshold δ for null hypothesis H_0^M. In both cases we obtained the threshold $\delta = 0.11$. We then ran our algorithm for thresholds $\delta' = 1.0\delta, 1.1\delta, 1.2\delta, \ldots, 2.0\delta$. The best results were obtained using the threshold $\delta' = 0.22$, and are reported below.

Using this approach HotNet identifies 12 candidate subnetworks containing at least 10 genes: the p-value for observing 12 subnetworks containing at least 10 genes is ≤ 0.05 under H_0^P and < 0.008 under H_0^M, and the FDR for the set of 12 subnetworks is ≤ 0.57 under H_0^P and ≤ 0.43 under H_0^M. The FDR is a conservative estimate of the ratio of false positives among all subnetworks reported by our algorithm, and implies that approximately 5 of the subnetworks reported by HotNet are significant. Since we included CNAs in our analysis, our results may contain potential artifacts resulting from functionally related genes that are both neighbors on the interaction network and close enough on the genome that they are affected by the same CNA. To reduce such artifacts, we applied two heuristics. First, we removed candidate subnetworks returned by HotNet that contain 3 or more genes in the same focal CNA in more than 1% of the patients. Second, for subnetworks with 2 genes g_1, g_2 in the same focal CNA in more than 1% of the patients, we removed the genes that are not found in the subnetwork when alterations in either g_1 or g_2 are removed. Of the 12 subnetworks identified by HotNet, 9 remain after these CNA filtering heuristics. To gain additional support for individual subnetworks and to focus attention on subnetworks with known biological function, we computed the overlap between the genes in candidate subnetworks and known pathways from the KEGG database.[8] For subnetworks that are enriched for at least one KEGG pathway, the best enrichments are reported in Table 1. Those 4 subnetworks are reported in Figs. 2–5. We note that none of the genes in those 4 subnetworks would be flagged as significant using a log rank test on single genes (with FDR ≤ 0.2).

All the subnetworks in Table 1 are enriched for at least one pathway reported in Crijns *et al.*[26] as containing more genes whose expression is correlated with overall survival than expected by chance. One of the subnetworks we found significantly overlaps genes in the focal adhesion pathway, that have been shown to be associated with survival. In particular, increased expression of the focal adhesion kinase (FAK) has been associated with shorter survival[27,28] in ovarian cancer, and the overexpression of ADAM9 has been correlated with brain metastasis in non-small cell lung cancer.[29] Moreover the depletion of MAP1S, part of another subnetwork S_2 we identify, has been associated with reduced survival.[30] To investigate the impact on survival

Table 1. Significant subnetworks identified by our method and significantly overlapping KEGG pathways. For each subnetwork, the genes in the subnetwork, the most significant KEGG pathways given by hypergeometric enrichment and the corresponding p-values are shown.

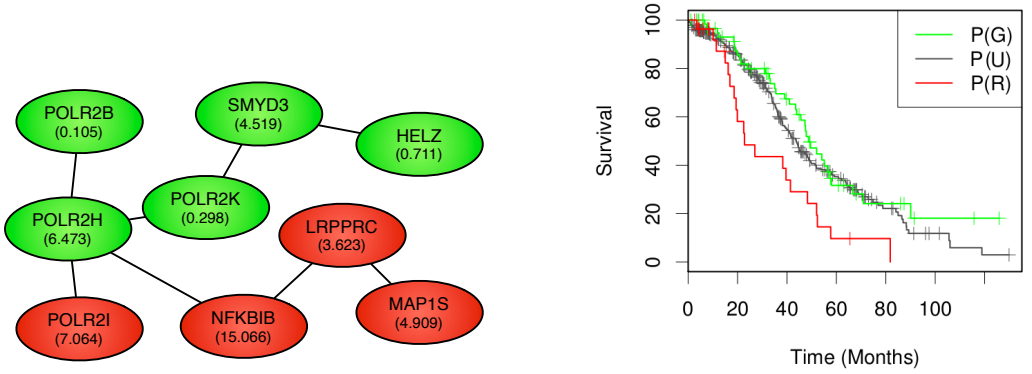
subnetwork	genes	pathway	enrichment p-val
S_1	ADAM9 ITGAV ITGA6	Regulation of actin cytoskeleton	1.1×10^{-9}
	ITGA3 ITGB5 LIMK1 FGFR2	Hypertrophic cardiomyopathy (HCM)	2.0×10^{-5}
	DLST UMPS PAK4 GATAD2A	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	1.2×10^{-5}
		Dilated cardiomyopathy	2.0×10^{-5}
		Focal adhesion	2.2×10^{-5}
S_2	POLR2I POLR2H POLR2K	RNA polymerase	3.1×10^{-8}
	HELZ MAP1S SMYD3	Pyrimidine metabolism	5.2×10^{-6}
	LRPPRC NFKBIB POLR2B	Purine metabolism	3.6×10^{-5}
S_3	MTF1 PVR ATP5J2	Lysosome	2.3×10^{-4}
	CD96 AP1G1 AP1M1 SHPRH	Cell adhesion molecules (CAMs)	4.8×10^{-2}
	LDLR PIGR LPP SCRIB GAK		
	AP1M2 PVRL3 PVRL2 LAMP1		
S_4	RAB8B OPTN RAB3IP TUBB	Calcium signaling pathway	7.4×10^{-4}
	RYR1 BIRC6 RYR2 CACNA1C	Cardiac muscle contraction	3.0×10^{-3}
	RABIF GDI2 TUBA4A RAB8A		
	HOMER3 CACNA1S RIMS2		



(a) (b)

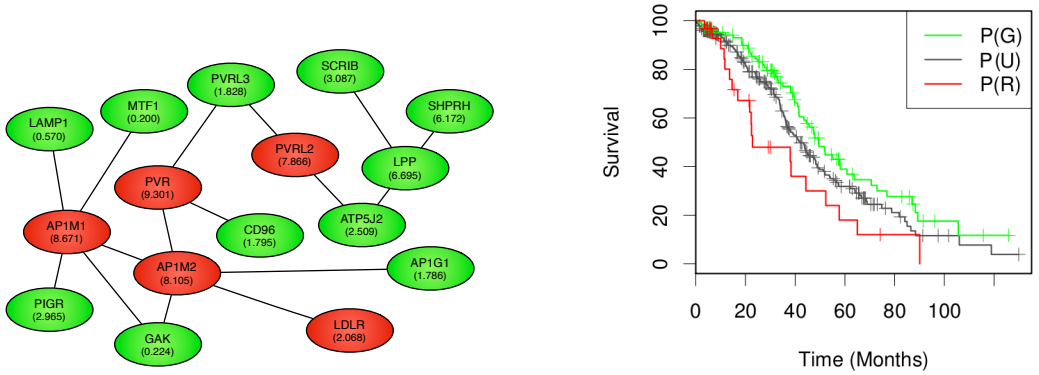
Fig. 2. Subnetwork S_1 identified by HotNet, significantly overlapping the focal adhesion pathway. (a) Interactions among genes as reported in HPRD. Genes are colored by their effect on the survival of patients: gene g is green if the median survival of patients where g is mutated exceeded that of patients where g is not mutated; gene g is red otherwise. The median survival is determined by the Kaplan-Meier estimator of the R function `survfit`. Numbers are the gene scores ($-2 \log_e p_g$). (b) Survival curves for: patients $P(G)$ with mutations only in green genes (green curve), patients $P(R)$ with mutations only in red genes (red curve), and patients $P(U)$ with mutations in neither red nor green genes (gray curve).

of mutations in the subnetworks identified by our algorithm we compared the survival of patients not having mutations in a particular subnetwork, with the survival of patients having a mutation in the subnetwork. Let $M(g)$ be the set of patients in which a gene g is mutated, and let $\overline{M}(g)$ be the set of patients in which g is not mutated. For a subnetwork S, let $S(G)$ be the set of genes in S for which the the mean survival of the patients in $M(g)$ is larger than the mean survival of the patients in $\overline{M}(g)$. Let $S(R)$ be the subset of S for which the



(a) (b)

Fig. 3. Subnetwork S_2 identified by HotNet, significantly overlapping the RNA polymerase pathway. (a) Interactions among genes as reported in HPRD. Genes are colored as in Figure 2 (b) Survival curves for: patients $P(G)$ with mutations only in green genes (green curve), patients $P(R)$ with mutations only in red genes (red curve), and patients $P(U)$ with mutations in neither red nor green genes (gray curve).

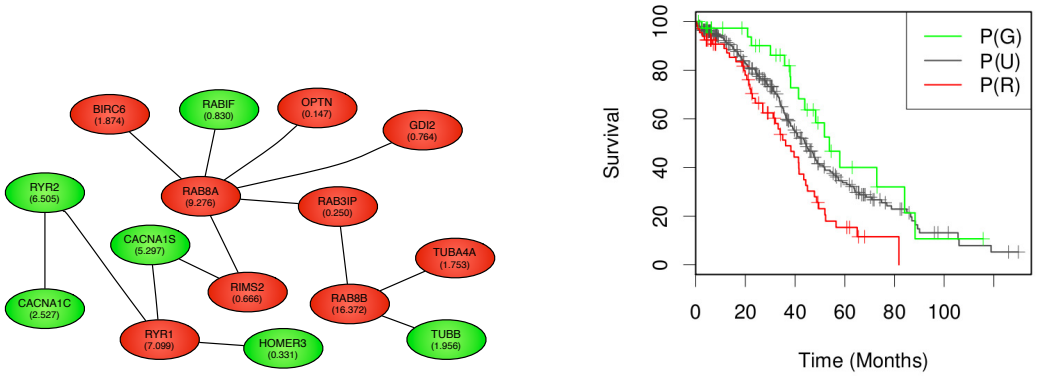


(a) (b)

Fig. 4. Subnetwork S_3 identified by HotNet, significantly overlapping the cell adhesion molecules pathway. (a) Interactions among genes as reported in HPRD. Genes are colored as in Figure 2 (b) Survival curves for: patients $P(G)$ with mutations only in green genes (green curve), patients $P(R)$ with mutations only in red genes (red curve), and patients $P(U)$ with mutations in neither red nor green genes (gray curve).

opposite holds. For each of the four subnetworks identified by our algorithm, we considered (i) patients $P(R)$ with mutations in genes in $S(R)$ but not in genes in $S(G)$; and (ii) patients $P(G)$ with mutations in genes in $S(G)$ but not in genes in $S(R)$. We also considered the set $P(U)$ of patients not having mutations in the subnetwork S. We then compared the survival curves of $P(R)$ and $P(U)$, and the survival curves of $P(G)$ and $P(U)$. The survival curves are reported in Figs 2–5, and the p-values (from logrank) for the tests are reported in Table 2. (In Table 2, A vs B denotes the comparison of survival for patients in A and of survival for patients in B.) For three of the subnetworks we found that at least one of the two tests has p-value $< 6 \times 10^{-3}$, and for the forth subnetwork both p-values are ≈ 0.06. This demonstrates that our method identifies subnetworks whose mutations are associated with survival.

We also used our method separating the analysis for genes with high survival and low



(a) (b)

Fig. 5. Subnetwork S_4 identified by HotNet, significantly overlapping the calcium signaling pathway. (a) Interactions among genes as reported in HPRD. Genes are colored as in Figure 2 (b) Survival curves for: patients $P(G)$ with mutations only in green genes (green curve), patients $P(R)$ with mutations only in red genes (red curve), and patients $P(U)$ with mutations in neither red nor green genes (gray curve).

Table 2. p-values from logrank test for the subnetworks of Table 1 enriched in KEGG pathways.

Subnetwork S	$P(R)$ vs. $P(U)$	$P(G)$ vs. $P(U)$
S_1	4.9×10^{-6}	1.5×10^{-2}
S_2	4.3×10^{-3}	2.7×10^{-1}
S_3	5.5×10^{-2}	5.9×10^{-2}
S_4	5.8×10^{-3}	1.9×10^{-1}

survival phenotypes. In particular, we defined a gene to have high survival if the median survival of the patients with a mutation in the gene was higher than the median survival of the patients without mutation in the gene. Otherwise, we defined the gene to have low survival. We then ran HotNet twice: once considering only genes with high survival, and then only genes with low survival, applying the same filtering steps described above. For the high survival genes, our method identified 7 subnetworks of at least 10 genes (p-value ≤ 0.01, FDR ≤ 0.51 under H_0^M, and p-value ≤ 0.3, FDR ≤ 0.8 under H_0^P). Only 3 of those subnetworks remains after the CNA filtering heuristics, and they are not enriched for known pathways. For the low survival genes, our method identifies 5 subnetworks of at least 10 genes (p-value ≤ 0.05, FDR ≤ 0.63 under H_0^M, and p-value ≤ 0.37, FDR ≤ 0.82 under H_0^P). Only 1 of those subnetworks remains after the CNA filtering heuristics, and it is not enriched for known pathways.

4. Discussion

We described an extension of our HotNet algorithm to finding subnetworks of genes whose mutational status is associated with a phenotype of interest. We applied our algorithm to ovarian cancer patients from The Cancer Genome Atlas and find 9 subnetworks associated with survival, 4 of which significantly overlap well-known pathways. Although we presented results using either a χ^2 test or logrank test, the extensions to HotNet described here apply to

any gene score, thus allowing for a variety of statistical tests to be used for testing associations to clinical phenotypes.

5. Acknowledgements

This work is supported by NSF grants IIS-1016648 and CCF-1023160. BJR is supported by a Career Award from the Scientific Interface from the Burroughs Wellcome Fund and an Alfred P. Sloan Research Fellowship. The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at `http://cancergenome.nih.gov/`.

References

1. A. Lievre *et al.*, *Cancer Res.* **66**, 3992 (2006).
2. K. Iida *et al.*, *Br J Cancer* (2011).
3. J. S. Smith *et al.*, *J. Natl. Cancer Inst.* **93**, 1246 (2001).
4. M. Meyerson, S. Gabriel and G. Getz, *Nat. Rev. Genet.* **11**, 685 (2010).
5. L. Ding, M. C. Wendl, D. C. Koboldt and E. R. Mardis, *Hum. Mol. Genet.* **19**, R188 (2010).
6. B. Vogelstein and K. W. Kinzler, *Nat. Med.* **10**, 789 (2004).
7. W. C. Hahn and R. A. Weinberg, *Nat. Rev. Cancer* **2**, 331 (2002).
8. M. Kanehisa and S. Goto, *Nucleic Acids Res.* **28**, 27 (2000).
9. L. Matthews *et al.*, *Nucleic Acids Res.* **37**, D619 (2009).
10. D. Croft *et al.*, *Nucleic Acids Res.* **39**, D691 (2011).
11. D. Szklarczyk *et al.*, *Nucleic Acids Res.* **39**, D561 (2011).
12. T. S. Keshava Prasad *et al.*, *Nucleic Acids Res.* **37**, D767 (2009).
13. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee and T. Ideker, *Mol. Syst. Biol.* **3**, p. 140 (2007).
14. T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, *Bioinformatics* **18 Suppl 1**, S233 (2002).
15. M. T. Dittrich, G. W. Klau, A. Rosenwald, T. Dandekar and T. Muller, *Bioinformatics* **24**, i223 (2008).
16. D. Beisser, G. W. Klau, T. Dandekar, T. Muller and M. T. Dittrich, *Bioinformatics* **26**, 1129 (2010).
17. C. J. Vaske, S. C. Benz, J. Z. Sanborn, D. Earl, C. Szeto, J. Zhu, D. Haussler and J. M. Stuart, *Bioinformatics* **26**, i237 (2010).
18. F. Vandin, E. Upfal and B. J. Raphael, *J. Comput. Biol.* **18**, 507 (2011).
19. E. Cerami, E. Demir, N. Schultz, B. S. Taylor and C. Sander, *PLoS ONE* **5**, p. e8918 (2010).
20. R. I. Kondor and J. Lafferty, Diffusion kernels on graphs and other discrete structures, in *International Conference on Machine Learning*, 2002.
21. F. Chung, *Proceedings of The National Academy of Sciences* **104**, 19735 (2007).
22. Y. Qi, Y. Suhail, Y. Y. Lin, J. D. Boeke and J. S. Bader, *Genome Res.* **18**, 1991 (2008).
23. N. Mantel, *Cancer Chemother Rep* **50**, 163 (1966).
24. R. A. Fisher, *Statistical Methods for Research Workers* (Oliver and Boyd, 1938).
25. The Cancer Genome Atlas Research Network, *Nature* **474**, 609 (2011).
26. A. P. Crijns *et al.*, *PLoS Med.* **6**, p. e24 (2009).
27. A. K. Sood *et al.*, *Am. J. Pathol.* **165**, 1087 (2004).
28. M. K. Siu *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18622 (2010).
29. Y. Shintani, S. Higashiyama, M. Ohta, H. Hirabayashi, S. Yamamoto, T. Yoshimasu, H. Matsuda and N. Matsuura, *Cancer Res.* **64**, 4190 (2004).
30. R. Xie, S. Nguyen, K. McKeehan, F. Wang, W. L. McKeehan and L. Liu, *J. Biol. Chem.* **286**, 10367 (2011).