**JOURNAL OF COMPUTATIONAL BIOLOGY**
**Volume 18, Number 3, 2011**
**© Mary Ann Liebert, Inc.**
**Pp. 507–522**
**DOI: 10.1089/cmb.2010.0265**

# Algorithms for Detecting Significantly Mutated Pathways in Cancer

FABIO VANDIN,[1] ELI UPFAL,[1] and BENJAMIN J. RAPHAEL [1,2]

## ABSTRACT

**Recent genome sequencing studies have shown that the somatic mutations that drive cancer development are distributed across a large number of genes. This mutational heterogeneity complicates efforts to distinguish functional mutations from sporadic, passenger mutations. Since cancer mutations are hypothesized to target a relatively small number of cellular signaling and regulatory pathways, a common practice is to assess whether known pathways are enriched for mutated genes. We introduce an alternative approach that examines mutated genes in the context of a genome-scale gene interaction network. We present a computationally efficient strategy for *de novo* identification of subnetworks in an interaction network that are mutated in a statistically significant number of patients. This framework includes two major components. First, we use a diffusion process on the interaction network to define a local neighborhood of "influence" for each mutated gene in the network. Second, we derive a two-stage multiple hypothesis test to bound the false discovery rate (FDR) associated with the identified subnetworks. We test these algorithms on a large human protein-protein interaction network using somatic mutation data from glioblastoma and lung adenocarcinoma samples. We successfully recover pathways that are known to be important in these cancers and also identify additional pathways that have been implicated in other cancers but not previously reported as mutated in these samples. We anticipate that our approach will find increasing use as cancer genome studies increase in size and scope.**

**Key words:** algorithms, pathways, statistical significance, cancer genomes.

## 1. INTRODUCTION

Cancer is a disease that is largely driven by somatic mutations that accumulate during the lifetime of an individual. Decades of experimental work have identified numerous cancer-promoting oncogenes and tumor suppressor genes that are mutated in many types of cancer. Recent cancer genome sequencing studies have dramatically expanded our knowledge about somatic mutations in cancer. For example, large projects such as The Cancer Genome Atlas (TCGA) (TCGA, 2008), the Tumor Sequencing Project (TSP) (Ding et al., 2008), and the Cancer Genome Anatomy Project (Greenman et al., 2007) have sequenced hundreds of protein coding genes in hundreds of patients with a variety of cancers. Other efforts

[1]Department of Computer Science and [2]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island.

have taken a global survey of approximately 20,000 genes in one to two dozen patients (Wood et al., 2007; Jones et al., 2008; Parsons et al., 2008). These studies have shown that: tumors harbor on average approximately 80 somatic mutations; two tumors rarely have the same complement of mutations; and thousands of genes are mutated in cancer (Wood et al., 2007). This mutational heterogeneity complicates efforts to distinguish functional *driver* mutations from sporadic, *passenger* mutations. One approach to identify genes with driver mutations is to find genes that are mutated at significant frequency in a collection of tumors from different patients. While some cancer genes are mutated at high frequency (e.g., well-known cancer genes such as TP53 or KRAS), most cancer genes are mutated at much lower frequencies. Thus, the observed frequency of mutation is an inadequate measure of the importance of a gene, particularly with the relatively modest number of samples that are tested in current cancer studies.

It is widely accepted that cancer is a disease of pathways, and it is hypothesized that somatic mutations target genes in a relatively small number of regulatory and signaling networks (Hahn and Weinberg, 2002; Vogelstein and Kinzler, 2004). Thus, mutational heterogeneity is explained by the fact that there are myriad combinations of mutations that cancer cells can employ to perturb the behavior of these key pathways. The unifying themes of cancer are thus not solely revealed by the individual mutated genes, but by the interactions between these genes. Standard practice in cancer sequencing studies is to assess whether genes that are mutated at sufficiently high frequency significantly overlap known cancer pathways (TCGA, 2008; Ding et al., 2008; Sjoblom et al., 2006; Wood et al., 2007; Parsons et al., 2008; Lin et al., 2007).

Finding significant overlap between mutated genes and genes that are members of known pathways is an important validation of existing knowledge. However, restricting attention to these known pathways does not allow one to detect novel group of genes that are members of less characterized pathways. Moreover, it is well known that there is crosstalk between different pathways (Vogelstein and Kinzler, 2004; McCormick, 1999), and dividing genes into discrete pathway groupings limits the ability to detect whether this crosstalk is itself a target of mutations. An additional source of information about gene and protein interactions is large-scale interaction networks, such as the Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009), STRING ( Jensen et al., 2009), and others (Bader et al., 2001; Salwinski et al., 2004). These resources incorporate both well-annotated pathways and interactions derived from high-throughput experiments, automated literature mining, cross-species comparisons, and other computational predictions. Many researchers have used these interaction networks to analyze gene expression data. For example, Ideker et al. (2002) introduced a method to discover subnetworks of differentially expressed genes, and this idea was later extended in different directions by others (Nacu et al., 2007; Liu et al., 2007; Ulitsky et al., 2008; Karni et al., 2009; Ma et al., 2007; Hescott et al., 2009; Chuang et al., 2007).

We propose to identify ''significantly mutated subnetworks'' (connected subnetworks whose genes have more mutations than expected by chance) *de novo* in a large gene interaction network. This problem differs from the gene expression problem of Ideker et al. (2002) in that a relatively small number of genes might be measured, a small subset of genes in a pathway may be mutated, and a single mutated gene may be sufficient to perturb a pathway. The naive approach to *de novo* identification of mutated subnetworks is to examine mutations on all subnetworks or all subnetworks of a fixed size. This approach is problematic. First, the enumeration of all such subnetworks is prohibitive for subnetworks of a reasonable size. Second, the extremely large number of hypotheses that are tested makes it difficult to achieve statistical significance. Finally, biological interaction networks typically have small diameter due to the presence of ''hub'' genes of high degree. There are reports that cancer-associated genes have more interaction partners than non-cancer genes (Lin et al., 2007; Jonsson and Bates, 2006), and indeed highly mutated cancer genes like TP53 have high degree in most interaction networks (e.g., the degree of TP53 in HPRD is 236). Such correlations might lead to a large number of ''uninteresting'' subnetworks being deemed significant.

We propose a rigorous framework for *de novo* identification of significantly mutated subnetworks. Our approach employs three strategies to overcome the difficulties described above. First, we formulate an *influence* measure between pairs of genes in the network using a diffusion process defined on the graph. This quantity considers one gene to influence another gene if they are both close in distance on the graph *and* there are relatively few paths between them in the interaction network. We use this measure to build an *influence graph* that includes only the tested genes but encodes the neighborhood information from the larger network. Second, we identify subnetworks using either a combinatorial model that finds subnetworks mutated in large number of samples, or an enhanced influence model in which the influence between pairs of genes is weighted by the number of mutations observed on these genes. Finally, we derive a *two-stage multiple hypothesis test* that mitigates the testing of a large number of hypotheses in subnetwork discovery. The first stage of the test computes the significance of the *number* of discovered subnetworks of a given size (rather than each individual subnetwork), and the second stage bounds the false discovery rate (FDR) of the list of discovered subnetworks.

We tested our approach using somatic mutation data from two recently published studies: (i) 601 genes in 91 glioblastoma multiforme patients from The Cancer Genome Atlas (TCGA) project; (ii) 623 genes in 188 lung adenocarcinoma patients sequenced during the Tumor Sequencing Project (TSP). In both datasets, we identify statistically significant mutated subnetworks that are enriched for genes on pathways known to be important in these cancers. Our approach is the first, to our knowledge, to demonstrate a computationally efficient strategy for *de novo* identification of *statistically significant* mutated subnetworks. We anticipate that our approach will find increasing use as cancer genome studies increase in size and scope.

## 2. METHODS

We describe our approach for the identification of significantly mutated pathways in cancer. In Section 2.1, we define the model used for the data we consider. In Section 2.2, we define the *influence graph*, which encodes the information in the interaction network and is used as input by the two methods we design. In Section 2.3, we describe a combinatorial model to identify mutated pathways, show that the corresponding optimization problem is NP-hard, and give an approximation algorithm for the problem. In Section 2.4, we develop a computationally efficient enhanced influence model that combines the information in the influence graph and the mutation data to identify mutated subnetworks. Finally, in Section 2.5, we design a statistical test to assess the significance of the networks reported by our methods.

### 2.1. Mathematical model

We model the interaction network by a graph $G = (V, E)$, where the vertices in $V$ represent individual proteins (and their associated genes), and the edges in $E$ represent (pairwise) protein-protein or protein-DNA interactions. Let $\mathcal{T} \subseteq V$ be the subset of genes that have been tested, or assayed, for mutations in a set $\mathcal{S}$ of samples (patients). The size of $\mathcal{T}$ will vary by study; for example, some recent works resequenced hundreds of genes (TCGA, 2008; Ding et al., 2008), whereas others examine nearly all known protein-coding genes in the human genome (Wood et al., 2007; Jones et al., 2008; Parsons et al., 2008). We assume that each gene $g$ is assigned one of two labels, *mutated* or *normal*, in each sample. Let $M_i$ denote the subset of genes in $\mathcal{T}$ that are mutated in the $i$th sample, for $i = 1, \ldots, |\mathcal{S}|$. Let $\mathcal{S}_j$ be the samples in which gene $g_j \in \mathcal{T}$ is mutated, for $j = 1, \ldots, |\mathcal{T}|$, and let $m = \Sigma_i |M_i|$ be the total number of occurrences of mutated genes observed in all samples.

We define a *pathway* or *subnetwork* to be a connected subgraph of $G$. This definition matches the common biological usage of the term where pathways are not restricted to be linear chains of vertices. We generally do not know whether more than one gene must be mutated to perturb a pathway in a sample, and thus will assume that a pathway is mutated in a sample if *any* of the genes in the pathway are mutated.

### 2.2. Influence graph

Our goal is to identify subnetworks that are significant with respect to the set of mutated genes in the samples. The significance of a subnetwork is derived from (i) the number of samples that have mutations in the genes of the subnetwork, and (ii) the interactions between genes in the subnetwork in the context of the topology of the whole network. For example, consider two possible scenarios of mutated nodes (Fig. 1). In the first scenario, the two mutated nodes are part of a linear chain in the interaction network. In the second scenario, the two mutated nodes are connected through a high-degree node. In the first scenario, there is a single path joining the two mutated nodes, and thus we are more surprised by this local clustering of mutations than in the second scenario, where the two nodes are connected by a node that is present in a large number of possible paths.

Hubs present an extreme case of this phenomenon and result in many ‘‘uninteresting’’ subnetworks being deemed significant. Since some highly mutated cancer genes, like TP53, also have high degree in interaction networks it is not advisable to ignore these genes in the analysis of cancer mutation data. These examples show that both the number of samples that have mutations in the genes of the subnetwork and the interactions between genes in the subnetwork in the context of the whole network must be considered to

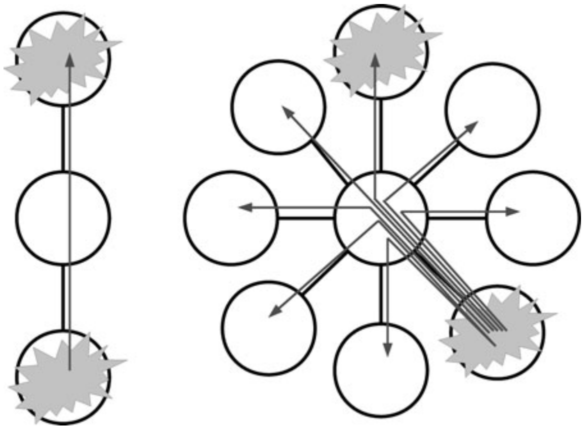

**FIG. 1.** Mutation on chain versus star graph.

derive the significance of a subnetwork. Considering only subnetworks of genes that are "close" in the network (i.e., with small shortest path distance) is not sufficient to overcome the problems highlighted above. Moreover, other graph mining approaches like dense subgraph identification (Feige et al., 1999) are also not appropriate, since not all subnetworks of interest (e.g., the chain in Fig. 1) are dense in edges.

We use a diffusion process on the interaction network to define a rigorous measure of *influence* between all pairs of nodes. To measure the influence of node $s$ on all the other nodes in the graph, consider the following process, described by Qi et al. (2008). Fluid is pumped into the source node $s$ at a constant rate, and fluid diffuses through the graph along the edges. Fluid is lost from each node at a constant first-order rate $\gamma$. Let $f_v^s(t)$ denote the amount of fluid at node $v$ at time $t$, and let $\mathbf{f}^s(t) = [f_1^s(t), \ldots, f_n^s(t)]^T$ be the column vector of fluid at all nodes. Let $L$ be the Laplacian matrix of the graph,[1] and let $L_\gamma = L + \gamma I$. Then the dynamics of this continuous-time process are governed by the vector equation $\frac{d\mathbf{f}^s(t)}{dt} = -L_\gamma \mathbf{f}^s(t) + \mathbf{b}^s u(t)$, where $\mathrm{b}^s$ is the elementary unit vector with 1 at the $s^{th}$ place and 0 otherwise, and $u(t)$ is the unit step function. As $t \to \infty$, the system reaches the steady state. The equilibrium distribution of fluid density on the graph is $\mathbf{f}^s = L_\gamma^{-1}\mathbf{b}^s$ (see (Qi et al., 2008). Note that this diffusion process is related to the diffusion kernel (Kondor and Lafferty, 2002) or heat kernel (Chung, 2007), which models the diffusion of heat on a graph, and these diffusion processes are in turn related to certain random walks on graphs (Doyle and Snell, 1984; Lovász, 1993). Diffusion processes and their related flow problems have been used in protein function prediction on interaction networks (Tsuda and Noble, 2004; Nabieva et al., 2005) and to define associations between gene expression and phenotype (Ma et al., 2007).

We interpret $f_i^s$ as the influence of gene $g_s$ on gene $g_i$. Computing the diffusion process for all tested genes gives us, for each pair of genes $g_j, g_k \in \mathcal{T}$, the influence $i(g_j, g_k)$ that gene $g_j$ has on gene $g_k$. Note that in general the influence is not symmetric; i.e. $i(g_j, g_k) \neq i(g_k, g_j)$. We define an *influence graph* $G_I = (\mathcal{T}, E_I)$ with the set of nodes corresponding to the set of tested genes, the weight of an edge $(g_j, g_k)$ is given by $w(g_j, g_k) = \min[i(g_k, g_j), i(g_j, g_k)]$, for all pairs of tested genes. If $n$ is the number of nodes in the interaction network, then the cost of computing $G_I$ is dominated by the complexity of inverting the $n \times n$ matrix $L_\gamma$.

## 2.3. *Discovering significant subnetworks: combinatorial model*

Given an influence measure between genes, the obvious first approach for discovering significant subnetworks is to identify sets of nodes in the influence graph $G_I$ that are (1) connected through edges with high influence; and (2) correspond to mutated genes in a significant number of samples. We fix a threshold $\delta$ and compute a *reduced influence graph* $G_I$ $(\delta)$ of $G_I$ by removing all edges with $w(g_i, g_j) < \delta$, and all nodes corresponding to genes with no mutations in the sample data. The computational problem is reduced to identifying the connected subgraphs of $G_I$ $(\delta)$ such that the corresponding sets of genes are altered in a significant number of patients.

The size of the connected subgraphs we discover is controlled by the threshold $\delta$. We choose sufficiently small $\delta$ such that, in the null hypothesis, in which the mutations are placed on nodes corresponding to tested

[1] $L = -A + D$, where $A$ is the adjacency matrix of the graph and $D$ is a diagonal matrix with $D_{i,i} = degree(v_i)$.

genes, it is unlikely that our procedure finds connected subgraphs with similar properties. Note that the value of $\delta$ depends only on the null hypothesis and not on the observed sample data (see Section 2.5 for details of the statistical analysis). Finding the connected subgraph of $k$ genes that is mutated in the largest number of samples is equivalent to the following problem, which we define as *connected maximum coverage* problem.

**Computational problem.** Given a graph $G$ defined on a set of $n$ vertices $V = \{v_1, \ldots, v_n\}$, a set $I$, a family of subsets $\mathcal{P} = \{P_1, \ldots, P_n\}$, with $P_i \in 2^I$ associated to $v_i \in V$, and a value $k$, find the connected subgraph $\mathcal{C}^* = \{v_i, \ldots, v_{i_k}\}$ with $k$ nodes in $G$ that maximize $|\cup_{j=1}^{k} P_{i_j}|$.

In our case, we have $G = G_I(\delta)$, $V$ is the subset of genes in $\mathcal{T}$ mutated in at least one sample, and for each $g_i \in V$ the associated set is $\mathcal{S}_i$. The connected maximum coverage problem is related to the maximum coverage problem (Hochbaum, 1997), where, given a set $I$ of elements, a family of subsets $F \subset 2^I$, and a value $k$, one needs to find a collection of $k$ sets in $F$ that covers the maximum number of elements in $I$. This problem is NP-hard as set cover is reducible to it.

If the graph $G$ is a complete graph, the connected maximum coverage problem is the same as the maximum coverage problem. Thus, the connected maximum coverage problem is NP-hard for a general graph. Moreover, we prove that the problem is still hard even on simple graphs such as the star graph. (Shuai and Hu, 2006, give a similar result for the connected set cover problem.)

**Theorem 1.** *The connected maximum coverage problem on star graphs is NP-hard.*

The proof is in the appendix. Since the connected maximum coverage problem is NP-hard even for simple graphs, we turn to approximate solutions. It is not hard to construct a polynomial time $1 - \frac{1}{e}$ approximation algorithm for spider graphs (analogous to the result in Shuai and Hu [2006] for the connected set cover problem). Since this algorithm cannot be applied to the network here, we construct an alternative polynomial time algorithm that gives $O(1/r)$ approximation when the radius of the optimal solution $\mathcal{C}^*$ is $r$. The pseudocode is shown in Figure 2.

Our algorithm obtains a solution $\mathcal{C}_v$ (thus, a connected subgraph) starting from each node $v \in V$, and then returns the best solution found. To obtain $\mathcal{C}_v$, our algorithm executes an *exploration phase*, i.e., for each node $u \in V$ it finds a shortest path $p_v(u)$ from $v$ to $u$. Let $\ell_v(u)$ be the set of nodes in $p_v(u)$, and $P_v(u)$ the elements of $I$ that they cover. After this *exploration phase*, the algorithm builds a connected subgraph $\mathcal{C}_v$ starting from $v$. At the beginning we have $\mathcal{C}_v = \{v\}$. $P_{\mathcal{C}_v}$ is the set of elements covered by the current connected subgraph $\mathcal{C}_v$. Then, while $|\mathcal{C}_v| < k$, the algorithm chooses the node $u \notin \mathcal{C}_v$ such that: $u = \arg\max_{u \in V}\left\{\frac{|P_v(u) \backslash P_{\mathcal{C}_v}|}{|\ell_v(u) \backslash \mathcal{C}_v|}\right\}$ and $|\ell_v(u) \cup \mathcal{C}_v| \leq k$; the new solution is then $\ell_v(u) \cup \mathcal{C}_v$. The main computational cost of our algorithm is due to the exploration phase, that can be performed in polynomial time. We have the following (proof in the Appendix):

**Theorem 2.** *The combinatorial algorithm gives a $\frac{1}{cr}$-approximation for the connected maximum coverage problem on G, where $c = \frac{2e-1}{e-1}$ and r is the radius of optimal solution in G.*

**Combinatorial Algorithm**

```
Input: Influence graph $G_I$ and parameters $\delta$ and $k$
Output: Connected subgraph $\mathcal{C}$ of $G_I(\delta)$ with $k$ vertices
Construct $G_I(\delta)$ by removing from $G_I$ all edges with weight $< \delta$;
$\mathcal{C} \leftarrow \emptyset$;
for each node $v \in V$ do
    $\mathcal{C}_v \leftarrow \{v\}$;
    for each $u \in V \backslash \{v\}$ do $p_v(u) \leftarrow$ shortest path from $v$ to $u$ in $G_I(\delta)$;
    while $|\mathcal{C}_v| < k$ do
        //$\ell_v(u) =$ set of nodes in $p_v(u)$; $P_v(u) =$ elements of $I$ covered by
          $\ell_v(u)$; $P_{\mathcal{C}_v} =$ elements covered by $\mathcal{C}_v$; $P_{\mathcal{C}} =$ elements covered by $\mathcal{C}$
        $u \leftarrow \arg\max_{u \in V \backslash \mathcal{C}_v : |\ell_v(u) \cup \mathcal{C}_v| \leq k}\left\{\frac{|P_v(u) \backslash P_{\mathcal{C}_v}|}{|\ell_v(u) \backslash \mathcal{C}_v|}\right\}$;
        $\mathcal{C}_v \leftarrow \ell_v(u) \cup \mathcal{C}_v$;
    if $|P_{\mathcal{C}_v}| > |P_{\mathcal{C}}|$ then $\mathcal{C} \leftarrow \mathcal{C}_v$;
return $\mathcal{C}$;
```

**FIG. 2.** Pseudocode of the algorithm for the combinatorial model.

For our experiments, we implemented a variation of this algorithm that for each pair of nodes ($u$, $v$) considers all the shortest paths between $u$ and $v$, and then keeps the one that maximizes $\frac{|P_v(u)|}{|\ell_v(u)|}$ to build the solution $\mathcal{C}_v$. With this modification, the algorithm is not guaranteed to run in polynomial time in the worst-case, but ran efficiently for all our experiments.

### 2.4. Discovering significant subnetworks: the enhanced influence model

We developed an alternative, computationally efficient, approach for identifying subnetworks that are significant with respect to the gene mutation data. The *Enhanced Influence Model* is based on the idea of enhancing the influence measure between genes by the number of mutations observed in each of these genes, and then decomposing an associated *enhanced influence graph* into connected components.

We define the *enhanced influence* graph $H$. The set $V_H$ of vertices of $H$ is given by all genes $g_j$ with at least one mutation in the data. The weight of edge ($g_j$, $g_k$) in $H$ is given by the enhanced influence

$$h(g_j, g_k) = w(g_j, g_k) \times \max\{|\mathcal{S}_j|, |\mathcal{S}_k|\}, \qquad (1)$$

for each pair of genes $g_j$, $g_k \in V_H$. Recall that $\mathcal{S}_j$ is the set of samples in which $g_j$ is altered, and $h$ is thus defined by the observed mutation data. Thus, the strength of connection between two nodes in the enhanced influence graph is a function of both the influence between the nodes in the interaction network and the number of mutations observed in their corresponding genes. Next we remove all edges with weight smaller than a threshold $\delta$ to obtain a graph $H(\delta)$. We return the connected components in $H(\delta)$ as the significant subnetworks with respect to the mutation data and the threshold $\delta$. The pseudocode is shown in Figure 3.

The computational cost is the complexity of computing all connected components in a graph with $S$ nodes, where $S$ is the number of mutated genes, which is linear in the size of the graph. The significance of the discovered subnetworks depends on the choice of $\delta$. We choose sufficiently small $\delta$ such that in the null hypothesis, in which the mutations are randomly placed in nodes corresponding to tested genes according to an appropriate null distribution, it is unlikely that our procedure finds connected components of similar size (see Section 2.5).

### 2.5. Statistical analysis

In this section, we describe a statistical test to assess the significance of our discoveries. The test we design can be used to assess the significance with respect to any null hypothesis on the distributions of mutations among the genes. In particular, we consider null hypothesis distributions in which the mutated genes are randomly allocated in the network, i.e., when the occurrence of mutations are independent of the network topology. Other distributions in which the occurrence of mutations *are not* independent of the network topology could be considered. For example, it has been previously reported that there is a correlation between the degree of a gene in a interaction network and its number of mutations (Cui et al., 2007). The data we analyze below did not display such a strong correlation, and so we do not consider this type of distribution in our experiments.

We employ two null hypothesis distributions: in $H_0^{\text{sample}}$ a total of $m = \Sigma_i |M_i|$ mutations are placed uniformly at random in the nodes corresponding to the $|\mathcal{T}|$ tested genes. While easier to analyze, this model does not account for the fact that in the observed data a large number of mutations are concentrated in a few genes (e.g., TP53). Thus, we also use a second null hypothesis distribution, $H_0^{\text{gene}}$, generated by permuting the identities of the tested genes in the network. That is, we select a random permutation $\sigma$ of the set

**Enhanced Influence Algorithm**

```
  Input: Influence graph $G_I$ and parameter $\delta$
  Output: Connected components of $H(\delta)$
$V_H \leftarrow \{g_j : \mathcal{S}_j \neq \emptyset\}$;
$E \leftarrow \{g_j, g_k : g_j, g_k \in V_H, g_j \neq g_k\}$;
$H \leftarrow (V_H, E, h)$;
$E(\delta) \leftarrow \{(g_j, g_k) \in E : h(g_j, g_k) \geq \delta\}$;
$H(\delta) \leftarrow (V_H, E(\delta))$;
return connected components of $H(\delta)$;
```

**FIG. 3.** Pseudocode of the algorithm for the enhanced influence model.

$\{1, \ldots, |\mathcal{T}|\}$, and we assign gene $g_j$, which was mutated in the set of samples $\mathcal{S}_j \subseteq \mathcal{S}$, to the location of gene $g_{\sigma(j)}$ in the original network. Note that, in a random mutation dataset, the set $\mathcal{S}_j$ of samples in which gene $g_j$ is altered is given by the null hypothesis distribution when computing the enhanced influence (1) or the combinatorial model. In contrast, the influence graph is fixed and given by the interaction network and the set of tested genes.

*2.5.1. A two-stage multi-hypothesis test.* A major difficulty in assessing the statistical significance of the discovered subnetworks is that we test simultaneously for a large number of hypotheses; each connected subnetwork in the interaction graph with at least one tested gene is a possible significant subnetwork and thus an hypothesis. The strict measure of significance level in multi-hypothesis testing is the *Family Wise Error Rate (FWER)*, the probability of incurring at least one Type I error in any of the individual tests. An alternative, less conservative approach to control errors in multiple tests is the the *False Discovery Rate (FDR)* (Benjamini and Hochberg, 1995). Let $V$ be the number of Type I errors in the individual tests, and let $R$ be the total number of null hypotheses rejected by the multiple test. We define $FDR = E[V/R]$ to be the expected ratio of erroneous rejections among all rejections (with $V/R = 0$ when $R = 0$). Let $t$ be the total number of hypothesis tested. Applying either measure to our problem, a discovery would be flagged as statistically significant only if its $p$-value is $O\ (1/t)$, which is impractical in the size of our problem. Instead, building on an idea presented in Kirsch et al. (2009), we develop a two-stage test for our problem that allows us to flag a number of subnetworks in our data as statistically significant with small FDR values.

We demonstrate our method through the analysis of the Enhanced Influence model. A similar technique was applied to the Combinatorial model. Let $C_1, \ldots, C_\ell$ be the set of connected components found in the enhanced influence graph $H(\delta)$. Testing for the significance of these discoveries is equivalent to simultaneously testing for $2^{|\mathcal{T}|}$ hypotheses. To reduce the number of hypotheses, we focus on an alternative statistic: the *number* of discoveries of a given size. Let $\tilde{r}_s$ be the number of connected components of size $\geq s$ found in the graph $H(\delta)$, and let $r_s$ be the corresponding random variable under the null hypothesis ($H_0^{\text{sample}}$ or $H_0^{\text{gene}}$). We are testing now for just up to $\mathcal{K} = |\mathcal{T}|$ simple hypotheses, for $s = 1, \ldots, \mathcal{K} : E_s \equiv$"$\tilde{r}_s$ conforms with the distribution of $r_s$". Testing each hypothesis with confidence level $\alpha/\mathcal{K}$, the first stage of our test identifies the smallest size $s$ such that with confidence level $\alpha$ we can reject the null hypothesis that $\tilde{r}_s$ conforms with the distribution of $r_s$.

The fact that the number of connected components of size at least $s$ is statistically significant does not imply necessarily that each of the connected components is significant. We now add a second condition to the test that guarantees an upper bound on the FDR:

**Theorem 3.** *Fix $\beta_1, \beta_2, \ldots, \beta_{\mathcal{K}}$ such that $\sum_{i=1}^{\mathcal{K}} \beta_i = \beta$. Let $s$* be the first $s$ such that $\tilde{r}_s \geq \frac{\mathbf{E}[r_s]}{\beta_s}$. If we return as significant all connected components of size $\geq s^*$, then the FDR of the test is bounded by $\beta$.*

The proof is in the Appendix. In our tests, we have used $\beta_i = \frac{\beta}{2^i}$ for the $i^{th}$ largest $s$ tested (with $\beta_s = \beta - \sum_i \beta_i$ for the smallest $s$), since we are more interested in finding large connected components.

*2.5.2. Estimating the distribution of the null hypothesis.* The null hypothesis distributions can be estimated by either a Monte-Carlo simulation ("permutation test") or through analytical bounds.

Using Monte-Carlo simulation, two features of our method significantly reduce the cost of the estimates. First, the Influence Graph $G_I$ is created *without* observing the sample data. The mutation data and $G_I$ are then combined to create the sample dependent graphs $G_I\ (\delta)$ and $H(\delta)$. Thus, the Monte Carlo simulation needs to run on the graph $G_I$, which is significantly smaller than the original interaction network (in our data the original interaction network had 18796 nodes while the influence graph had only about 600 nodes). Second, our statistical test does not use the $p$-values of individual connected subgraphs/components but the $p$-value of the distribution of the number of connected subgraphs/components of a given size. Thus, for this test it is sufficient to estimate $p$-values that are a magnitude larger, and therefore require significantly fewer rounds of simulations. These features allowed us to compute the null distributions through Monte-Carlo simulations for the size of our data with no significant computational cost.

For a larger number of tested genes, we can estimate the null hypothesis through analytical bounds. Consider for example the Enhanced Influence model, and assume that the $|\mathcal{T}|$ tested genes are randomly permuted among the $|\mathcal{T}|$ nodes of the graph $G_I$ to generate a random graph $\bar{H}(\delta)$. Let $M$ be the number of genes with observed mutations, and let $s_{max}$ be the maximum number of mutations of any gene. Since we are interested in $\delta$ that partitions the graph into a number of connected components, we choose the maximum $\delta$ such that for any node $g_i$ in $G_I$ no more than $\alpha M/|\mathcal{T}|$ of the adjacent edges have weights that satisfy $s_{max}w(g_i,\ g_j) \ge \delta$, for some fixed $\alpha < 1$. For this choice of $\delta$, the expected number of connected components of size $k$ in $\bar{H}(\delta)$ is bounded by $\binom{|\mathcal{T}|}{k} k^{k-2}\alpha^{k-1} \le \frac{M}{k^2}\alpha^{k-1}$. Since connected components are disjoint, their occurrences are negatively correlated, and we can stochastically bound the distribution of $r_s$ with a binomial distribution with the above expectation. A similar bound can be computed for the other models and null hypothesis distributions, and for (somewhat) less restrictive conditions on $\delta$.

## 3. EXPERIMENTAL RESULTS

We applied our approach to analyze somatic mutation data from two recent studies. The first dataset is a collection of 453 validated nonsynonymous somatic mutations identified in 601 tested genes from 91 glioblastoma multiforme (GBM) samples from The Cancer Genome Atlas (TCGA, 2008). In total, 223 genes were reported mutated in at least one sample. The second dataset is a collection of 1013 validated nonsynonymous somatic mutations identified in 623 tested genes from 188 lung adenocarcinoma samples from the Tumor Sequencing Project (Ding et al., 2008). In total, 356 genes were reported mutated in at least one sample. For the Enhanced Influence model, we also considered simulated data.

We use the protein interaction network from the Human Protein Reference Database (June 2008 version) (Keshava Prasad et al., 2009), which consists of 18796 vertices and 37107 edges. We derive the influence graph for each dataset by directly computing the inverse[2] of $L_\gamma$. The results presented below are obtained by fixing the parameter $\gamma = 8$, which is approximately the average degree of a node in HPRD (after the removal of disconnected nodes). Similar results were obtained with $\gamma = 1$ or $\gamma = 30$.

The resulting influence graphs have weights $i(g_j,\ g_k) \neq 0$ for almost all pairs $(g_j,\ g_k)$ of tested genes: less than 2% of the weights are zero in the GBM graph, while all weights in the lung adenocarcinoma graph are positive.

### 3.1. Combinatorial model

We used the combinatorial model to extract a subnetwork of $k$ mutated genes that is mutated in the highest number of samples from GBM and lung adenocarcinoma, for $k = 10$ and $k = 20$. For both datasets, we used the procedure described in Section 2.3 to derive the threshold $\delta = 10^{-4}$ for the reduced influence graph $G_I$ ($\delta$). Table 1 shows that we find statistically significant subnetworks under both the $H_0^{\text{gene}}$ and $H_0^{\text{sample}}$ null hypotheses ($p$-values for $H_0^{\text{sample}}$ are computed without Monte-Carlo simulation). The genes in each subnetwork are reported in Table 2.

To assess the biological significance of our findings in GBM, we compared the genes in each subnetwork to the genes in pathways that were previously implicated in GBM and used as a benchmark in the TCGA

Table 1. Results of the Combinatorial Model

| | | | p-*value* | | *Pathway enrichment* p-*value* | | |
|---|---|---|---|---|---|---|---|
| *Dataset* | k | *Samples* | $H_0^{\text{sample}}$ | $H_0^{\text{gene}}$ | *All* | *RTK/RAS/PI(3)K* | *p53* |
| GBM | 10 | 67 | $<10^{-10}$ | $4\times10^{-3}$ | $3\times10^{-4}$ | $8\times10^{-4}$ | 0.19 |
| | 20 | 78 | $<10^{-10}$ | $<10^{-3}$ | $10^{-5}$ | $8\times10^{-5}$ | 0.05 |
| Lung | 10 | 140 | $<10^{-10}$ | 0.02 | $8\times10^{-6}$ | / | |
| | 20 | 151 | $<10^{-10}$ | 0.03 | $3\times10^{-3}$ | / | |

$k$ is the number of genes in the subnetwork. *Samples* is the number of samples in which the subnetwork is mutated. *p-value* is the probability of observing a connected subgraph of size $k$ mutated in a number of samples $\ge$ *samples* under the random model $H_0^{\text{sample}}$ or $H_0^{\text{gene}}$. *enrichment p-value* is the $p$-value of the hypergeometric test for overlap between genes in the identified subgraph and genes reported significant pathways in TCGA (2008) or Ding et al. (2008). For GBM, *enrichment p-value* is the $p$-value of the hypergeometric test for RTK/RAS/PI(3)K and p53 pathways.

[2] In contrast, Qi et al. (2008) derive a power series approximation to $L_\gamma^{-1}$ whose convergence depends on the choice of $\gamma$.

Table 2. Results of Combinatorial Model

| *Dataset* | k | *Samples* | *Genes* |
|---|---|---|---|
| GBM | 10 | 67 | INSR BCR TP53 PTEN EGFR<br>ERBB2 DST PIK3R1 PIK3CA SERPINA3 |
| | 20 | 78 | MDM2 FGFR1 BRCA2 CHEK1 COL1A2<br>ITGB3 TNK2 INSR BCR TP53<br>PTEN EGFR ERBB2 DST PIK3R1<br>PIK3CA NF1 SPARC PDGFRA SERPINA3 |
| Lung | 10 | 140 | CDC25A CHEK1 TP53 STK11 HRAS<br>KRAS ERBB4 EGFR NF1 PTEN |
| | 20 | 150 | MAPK8 PRKDC TP53 STK11 HRAS<br>KRAS EGFR PRKD1 NF1 ABL1<br>ERBB4 PTEN HD PRKCE SMAD2<br>TGFBR1 BAX RAPGEF1 PIK3CG ACVR1B |

Genes in the connected component of size *k* that covers the maximum number of samples in GBM and lung adenocarcinoma, as reported by our combinatorial algorithin.

publication (TCGA, 2008) (Fig. 4a). We find that our subnetworks are enriched for genes in the RTK/RAS/PI(3)K pathway and to a lesser extent, the p53 pathway. For the lung adenocarcinoma samples, we find that the subnetworks share significant overlap with the pathways reported in the original publication (Ding et al., 2008). These results demonstrate that the combinatorial model is effective in recovering genes known to be important in each of these cancers.

### 3.2. Enhanced influence model

**Simulated data.** We tested the ability of our enhanced influence model to recover significantly mutated pathways in simulated data. We extracted a well-curated network of 258 genes called "Pathways in cancer (hsa05200)" from the KEGG database (Kanehisa and Goto, 2000). We augmented this network with additional random edges so that 20% of the edges of the resulting network were random. We assigned mutations to a well-known cancer signaling pathway, PKC–RAF–MEK–ERK, a linear chain $\mathcal{P}$ of 4 genes,

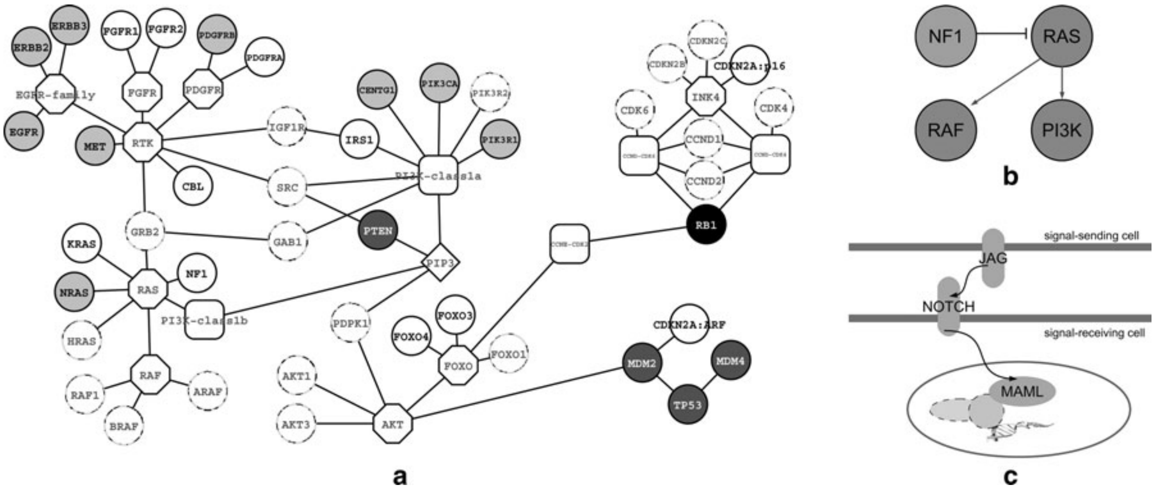

**FIG. 4.** **(a)** Overlap between subnetworks found by the enhanced influence model and significant pathways reported in TCGA (2008). Each circle is a gene; other shapes represent protein families or complexes, or small molecules. For each protein family and complex, tested genes are shown. "Dashed" nodes are tested genes that were not mutated in GBM, and thus cannot be returned as significant. Light gray nodes are found in the c.c. of size 22, dark gray nodes in the c.c. of size 18, and the black node (RB1) in a c.c. of size 2. **(b)** Pathway corresponding to one of the connected components extracted with enhanced influence model in lung. **(c)** Notch signaling pathway identified in the lung dataset.

Table 3. Results of the Enhanced Influence Model on GBM Samples

| | | $H_0^{\text{sample}}$ | | $H_0^{\text{gene}}$ | | *Enrichment* p-*value* | |
|---|---|---|---|---|---|---|---|
| *s* | *No. of c.c.* $\geq s$ | $\mu$ | p-*Value* | $\mu$ | p-*value* | *RTK/RAS/PI(3)K* | *p53* |
| 2 | 15 | 22.18 | 0.97 | 13.63 | 0.38 | / | / |
| 3 | 3 | 6.37 | 0.98 | 4.38 | 0.6 | / | / |
| 19 | 2 | $<10^{-3}$ | $<10^{-3}$ | 0.07 | $<10^{-3}$ | 0.9 | $4\times10^{-3}$ |
| 22 | 1 | $<10^{-3}$ | $<10^{-3}$ | 0.05 | 0.05 | $4\times10^{-6}$ | — |

*s* is the size of connected components (c.c.) found with our method. No. of *c.c.* $\geq s$ is the number of c.c. with *at least s* nodes. $\mu$ is the expected number of c.c. with $\geq s$ nodes under random models $H_0^{\text{gene}}$, $H_0^{\text{sample}}$. *p-value* is the probability of observing *at least* No. of *c.c.* $\geq s$ with at least *s* nodes in a random dataset. The last two columns show, for c.c. with $s > 3$, the result of the hypergeometric test for enrichment for RTK/RAS/PI(3)K, and p53 pathways, respectively.

so that at least one gene is mutated in *x*% of samples, for different *x*. We then randomly assigned mutations to all the genes in the network matching the observed values (e.g., number of samples, ratio between number of tested genes, and number of genes in the network) in GBM. We correctly identify $\mathcal{P}$ as significantly mutated ($P < 10^{-2}$, FDR $< 10^{-2}$) even when each gene in $\mathcal{P}$ is altered in $\leq$5% of the samples, but $\mathcal{P}$ is altered in 17% of the samples. Note that genes mutated in 5% of the samples were not reported as significantly mutated in TCGA (2008), demonstrating that our method correctly identifies a mutated path even when the individual genes in the path are not mutated in a significant number of samples. Moreover, $\mathcal{P}$ is the *only* significant pathway reported by our method. To verify that our influence measure takes into account the topology of the network, we added a number of edges to the RAF gene in $\mathcal{P}$, giving it high degree in the network. As expected, $\mathcal{P}$ is no longer identified as significant in the modified network.

**Cancer data.** We applied the enhanced influence model to the GBM and lung adenocarcinoma datasets. Following the procedure described in Section 2.4, we first computed the enhanced influence graph, using a threshold of $\delta = 0.003$ for the GBM data and $\delta = 0.01$ for the lung adenocarcinoma data. Table 3 shows the number and sizes of the connected components identified in the GBM data, and the associated *p*-values, the latter obtained using the method described in Section 2.5. Table 4 reports the genes in the connected components of size >3.

We identify two significant connected components with more than 19 genes (FDR $\leq$0.14). We find significant overlap ($P < 10^{-2}$ by hypergeometric test) between the 68 genes in our connected components and the set of all mutated genes in the same RTK/RAS/PI(3)K, p53, and RB pathways examined in the TCGA study (TCGA, 2008). The second largest connected component with 19 genes has significant overlap to the p53 pathway, while the largest connected component with 22 genes has significant overlap with the RTK/RAS/PI(3)K signaling pathway. In contrast to the combinatorial model, the enhanced influence model separates these two pathways into different connected components. Figure 4a illustrates the overlap between the mutated genes in connected components returned by our method and genes in the pathways reported in TCGA (2008).

For the lung data, Table 5 shows the sizes of connected components returned by the enhanced influence model and the *p*-values associated with each. Table 6 lists the genes in each connected component of size >5.

Table 4. Genes in Connected Components Obtained for GBM the Diffusion Model with $\delta = 0.003$

| *Size* | *Genes* |
|---|---|
| 22 | MSH2 ATM MSH6 PRKDC ATR BCR KLF6 GLI3 KLF4 PML MAPK9 CHEK1 BRCA2 ING4 MDM2 MDM4 TP53 TOP1 PTEN KPNA2 STK36 GLI1 |
| 19 | ANXA1 TNK2 ERBB3 SERPINA3 SOCS1 TNC PIK3C2B PDGFRB ERBB2 NRAS VAV2 EGFR EPHA2 MET ADAM12 PIK3R1 PIK3CA CENTG1 AXL |

Table 5. Results of the Enhanced Influence Model on Lung Adenocarcinoma Samples

| | | $H_0^{\text{sample}}$ | | $H_0^{\text{gene}}$ | | |
|---|---|---|---|---|---|---|
| *s* | *No. of c.c.* $\geq s$ | $\mu$ | p-*value* | $\mu$ | p-*value* | *Enrichment* p-*value* |
| 2 | 24 | 23.4 | 0.7 | 17.67 | 0.4 | / |
| 3 | 11 | 6.51 | 0.13 | 7.27 | 0.2 | / |
| 4 | 7 | 3.21 | 0.07 | 4.98 | 0.13 | / |
| 5 | 5 | 2.09 | 0.01 | 2.18 | 0.01 | / |
| 7 | 4 | 0.54 | 0.01 | 0.56 | 0.01 | — |
| 10 | 3 | $<10^{-3}$ | $<10^{-3}$ | 0.4 | 0.02 | 0.34; $10^{-5}$; $9\times10^{-8}$ |

Columns are as described in Table 3. Last column shows, for each c.c. with $s \geq 7$, the result of the hypergeometric test for enrichment for all genes reported in significant pathways in Ding et al. (2008) (the three values correspond to 3 c.c. of size 10).

The 88 genes in the union of the connected components derived by our method overlap significantly ($P < 7\times10^{-9}$ by the hypergeometric test) with the mutated pathways reported in the network in the TSP publication (Fig. 6 in Ding et al., 2008). We identify four connected components of size $\geq 7$ (FDR $\leq 0.56$). The first connected component of size 10 contains genes in the p53 pathway, and the second one is enriched ($P < 10^{-2}$) for the MAPK pathway (Fig. 4b). The third component is the ephrin receptor gene family, a large family of membrane-bound receptor tyrosine kinases that were reported as mutated in breast and colorectal cancers (Sjoblom et al., 2006). Notably, only one of the genes in this component, EPHA3, is mentioned as significantly mutated in Ding et al. (2008). Finally, the connected component of size 7 consists exclusively of members of the Notch signaling pathway (Fig. 4c). The mutated genes include the Notch receptor (NOTCH2/3/4); Jagged (JAG1/2), the ligand of Notch; and Mastermind (MAML1/2), a transcriptional co-activator of Notch target genes. The Notch signaling pathway is a major developmental pathway that has been implicated in a variety of cancers (Axelson, 2004), including lung cancer (Collins et al., 2004). Mutations in this pathway were not noted in the original TSP publication (Ding et al., 2008), probably because no single gene in this pathway is mutated in more than three samples. Because our method exploits both mutation frequency and network topology, we are able to identify these more subtle mutated pathways, and in this case identify an entire signaling pathway.

### 3.3. Naive approach

To demonstrate the impact of the influence graph on the results, we implemented a naive approach that examines all paths in the original HPRD network that connect two tested genes and contain at most three nodes. We extracted all paths that were altered in a significant number of samples with FDR $\leq 0.01$ using the standard Benjamini-Yekutieli method (Benjamini and Yekutieli, 2001). More than 1700 paths in GBM and >2200 in lung adenocarcinoma are marked as significant with this method. A major reason for this large number of paths is the presence of highly mutated genes that are also high-degree nodes in the HPRD network (e.g., TP53). *Each* path through these high degree nodes is marked as significant. One possible solution is to remove any path that contains a subpath that is significant. Table 7 shows the results of this filtering on GBM data for $H_0^{\text{sample}}$. Table 8 shows the analogous table for lung adenocarcinoma.

Note that these filtered paths include *none* through important, highly mutated, and high degree genes like TP53. Since our influence graph uses both mutation frequency and local topology of the network, we are able to recover subnetworks containing these genes without also reporting an extremely large number of

Table 6. Connected Components of Size $\geq 7$ for Lung Adenocarcinoma Using the Diffusion Model with $\delta = 0.01$

| *Size* | *Genes* |
|---|---|
| 10 | WT1 CDKN2A TP53 CCNG1 KLF6 ATR CDKN2C TP73L TFDP1 CHEK1 |
| 10 | RAP2B PIK3CA HRAS RASSF2 NRAS MRAS PIK3CG BRAF NF1 RHOB |
| 10 | EPHB1 EPHB6 EPHA7 EPHA6 EPHA5 EPHA4 EPHA3 EPHA2 EPHA1 FGFR4 |
| 7 | MAML2 MAML1 NOTCH4 NOTCH2 NOTCH3 JAG2 JAG1 |

Table 7. Statistically Significant Mutated Paths (FDR = 0.01) Using the HPRD Network (Keshava Prasad et al., 2009) and the Glioblastoma Mutations Dataset (TCGA, 2008)

| *Genes* | *No. of mutated samples* | p-*value* |
|---|---|---|
| PDGFRB, PIK3CA | 8 | $10^{-4}$ |
| PIK3CA, PRKCD, EP300 | 10 | $6\times10^{-5}$ |
| PIK3CA, IRS4, PRKCZ | 8 | $10^{-4}$ |

For each significant path, the genes in the path, the number of samples with at least one mutation in the path, and the (non-corrected) *p*-value are shown.

other extraneous subnetworks. Finally, we note that finding larger, statistically significant subnetworks (e.g., those with 10 or 20 nodes) with the naive approach is impossible in the GBM and lung datasets because of the severe multiple hypotheses correction for the large number of subnetworks tested; e.g., the number of connected components with 10 tested genes in the HPRD network is $>10^{10}$. For the same reason, the enumeration of all the paths or connected components of reasonable size is impossible.

## 4. CONCLUSION

We present an approach to identify significantly mutated pathways in a large, unannotated interaction network. The subnetworks derived by our method share significant overlap with known cancer pathways. Remarkably, we automatically extracted a large fraction of these pathways with a modest number (100–200) of samples (Fig. 4). Our approach has two key advantages over the common strategy of testing the overlap between mutated genes and genes from known pathways, using a hypergeometric or similar test. First, we incorporate biological information that is not presently represented in existing well-characterized pathways, while accounting for the uncertainty in large gene interaction networks. Second, we are able to assign significance to genes that are altered at low frequency but are part of a larger subnetwork that is altered at significant frequency. The latter advantage was demonstrated in the lung adenocarcinoma dataset where we identify the Notch signaling pathway as significant, even though the individual genes were not mutated at significant frequency.

There are numerous ways to extend the model presented here. First, additional data types may be considered, such as copy number aberrations, genome rearrangements, gene expression changes, or epigenetic alterations. Second, a different null model could be used. Nearly any null model used for single-gene tests of significance could be adapted to the network context. For example, one could employ a null model where mutations in gene occur at a fixed ''background'' rate, meaning that longer genes would be more likely to harbor mutations. Using such a model on the GBM and lung adenocarcinoma data produced results extremely close to those presented here (data not shown). Finally, the network model could be expanded to include

Table 8. Statistically Significant Mutated Paths (FDR = 0.001) Using the HPRD Network (Keshava Prasad et al., 2009) and the Lung Adenocarcinoma Mutation Dataset (Ding et al., 2008)

| *Genes* | *No. of mutated samples* | p-*value* |
|---|---|---|
| CDKN2A, E4F1, RB1 | 15 | $10^{-6}$ |
| CDKN2A, WRN, PRKDC | 15 | $10^{-6}$ |
| EPHA7, EFNA1, EPHA3 | 15 | $10^{-6}$ |
| PRKDC, HSP90AA1, KDR | 15 | $10^{-6}$ |
| EPHA3, EFNA2, EPHA5 | 15 | $10^{-6}$ |
| NTRK3, DYNLL1, NTRK1 | 14 | $6\times10^{-6}$ |
| NTRK1, CAV1, KDR | 14 | $6\times10^{-6}$ |
| KDR, ITGB3, PDGFRA | 14 | $6\times10^{-6}$ |

For each significant path, the genes in the path, the number of samples with at least one mutation in the path, and the (non-corrected) *p*-value are shown.

additional interaction types (e.g., regulatory or miRNA), directed interactions or weighted interactions. The later can be included naturally in our diffusion model by adding weights, or reliabilities, on the edges.

We anticipate that our method will become even more useful as larger datasets become available. Several recent studies (Wood et al., 2007; Jones et al., 2008; Parsons et al., 2008) have surveyed a much larger number of genes than considered here (approximately 20,000), but in a relatively small number of samples (one to two dozen per cancer type). Continuing decline in DNA sequencing costs and the development of targeted exon-capture techniques (Hodges et al., 2007) will soon enable global surveys of all protein-coding genes in hundreds to thousands of cancer samples.

# 5. APPENDIX

## A. *Proofs*

In this Appendix, we report the proofs of the theorems stated in the article.

**Theorem 1.** *The connected maximum coverage problem on star graphs is NP-hard.*

**Proof.** The proof is by reduction from the maximum coverage problem. Given an instance of the maximum coverage problem, consisting of $I$, $F$, and $k$, we build an instance of the connected maximum coverage problem. We define $I' = I \cup \{v_0\}$, with $v_0 \notin I$; and $F' = F \cup \{v_0\}$. Moreover, we build the graph $G = (V, E)$ where $V = F'$ and $E = \{(v_0, s) | s \in F\}$. It is easy to verify that $G$ is a star graph, and then each non-trivial (i.e., with more than 1 vertex) subgraph of $G$ will contain the vertex $v_0$. The solution $X$ to the connected maximum coverage problem on the graph $G$ is then of the form $X = Y \cup \{v_0\}$, where $Y \subseteq F$. It is easy to verify that $X$ is a connected maximum coverage of size $k+1 > 1$ if and only if $Y$ is maximum coverage of size $k > 0$. ■

**Theorem 2.** *The combinatorial algorithm gives a* $\frac{1}{cr}$*-approximation for the connected maximum coverage problem on G, where* $c = \frac{2e-1}{e-1}$ *and r is the radius of optimal solution in G.*

**Proof.** We first analyze the solution obtained assuming the nodes in the solution are inserted one at the time (i.e., $|\ell_v(u) \backslash \mathcal{C}_v| = 1$ for each node $u$ inserted in the solution). We will then show that, when the nodes are not inserted one at the time, the solution obtained cannot have a worse value.

Let $z^*(v)$ be the value of the best solution $OPT(v)$ that can be found starting at node $v$. For $1 \leq i \leq r_v$, define $OPT_i(v) = \{v_j \in OPT(v) : d(v_j, v) = r_v - i + 1\}$, and

$$z_i^*(v) = \left| \left\{ \bigcup_{g_j \in OPT_i(v)} P_j - \bigcup_{\ell < i} \left( \bigcup_{g_j \in OPT_\ell(v)} P_j \right) \right\} \right|,$$

thus, $OPT(v) = \cup_{i=1}^{r_v} OPT_i(v)$, where $r_v$ is the maximum distance between $v$ and a node in $OPT(v)$, and $z^*(v) = \sum_{i=1}^{r_v} z_i^*(v)$. We divide the execution of our algorithm in $r_v$ phases: in phase $i$ our algorithm inserts $|OPT_i(v)|$ new nodes in the solution. Note that in phase $i$, our algorithm always has the possibility to reach each node in $OPT_i(v)$. Thus, in phase $i$, the algorithm above is equivalent to the greedy algorithm for the maximum coverage problem where the sets that can be chosen are all the sets at distance at most $r_v - i + 1$, and then all the sets in $OPT_i(v)$ can be chosen by the greedy algorithm. Let $A_i(v)$ be the increment in the value of the solution found by our algorithm between the end of phase $i$ and the end of phase $i - 1$. The value of the solution of our algorithm starting from node $v$ is then $A(v) = \sum_{i=1}^{r_v} A_i(v)$. Since the approximation factor for the maximum coverage is $1 - 1/e$ and each element in $OPT_i(v)$ is seen with weight reduced of a factor $1/(r_v - i + 1)$ (since it is at distance $r_v - i + 1$), in phase $i$ our algorithm improves the current solution of a factor

$$A_i(v) \geq \frac{1}{r_v} \left(1 - \frac{1}{e}\right) (z_i^*(v) - \sum_{j=1}^{i-1} A_j(v)).$$

Summing the terms above for all $i$, $1 \leq i \leq r_v$ we obtain:

$$\begin{aligned}
A(v) &\geq \frac{1}{r_v}\left(1-\frac{1}{e}\right)\left(\sum_{i=1}^{r_v} z_i^*(v) - \sum_{j=1}^{r_v-1}(r_v-j)A_j(v)\right)\\
&\geq \frac{1}{r_v}\left(1-\frac{1}{e}\right)\sum_{i=1}^{r_v} z_i^*(v) - \frac{1}{r_v}\left(1-\frac{1}{e}\right)\sum_{j=1}^{r_v-1}(r_v-j)A_j(v)\\
&\geq \frac{1}{r_v}\left(1-\frac{1}{e}\right)z^*(v) - \frac{1}{r_v}\left(1-\frac{1}{e}\right)r_v A(v)\\
&\geq \frac{1}{r_v}\left(1-\frac{1}{e}\right)z^*(v) - \left(1-\frac{1}{e}\right)A(v).
\end{aligned}$$

We then obtain

$$\frac{2e-1}{e}A(v) \geq \frac{1}{r_v}\left(\frac{e-1}{e}\right)z^*(v)$$

that is

$$A(v) \geq \frac{1}{r_v}\left(\frac{e-1}{2e-1}\right)z^*(v).$$

Now, let $v^*$ be such that (i) $v^* = \operatorname{argmax}_v\{z^*(v)\}$ (i.e., $z^*(v^*) = z^*$), and (ii) $r_{v^*} = r$. Then

$$A = \max_v A(v) \geq A(v^*) \geq \frac{1}{r_{v^*}}\left(\frac{e-1}{2e-1}\right)z^*(v^*) = \frac{1}{r}\left(\frac{e-1}{2e-1}\right)z^*.$$

Now consider the case $|\ell_v(u)\backslash\mathcal{C}_v| > 1$: this means that we insert a path whose weight, divided by $|\ell_v(u)\backslash\mathcal{C}_v|$, is higher than the weight of any other node (currently) reachable from $v$. Then we have that the value of the solution found by our algorithm can only improve, since we are inserting $|\ell_v(u)\backslash\mathcal{C}_v|$ nodes such that the average value of the inserted nodes is greater than the maximum value of any (currently) reachable node in $OPT(v)$ divided by its distance (that is, at most $r_v$). ■

**Theorem 3.** *$\beta_1, \beta_2, \ldots, \beta_{\mathcal{K}}$ such that $\sum_{i=1}^{\mathcal{K}} \beta_i = \beta$. Let $s^*$ be the first $s$ such that $\tilde{r}_s \geq \frac{\mathbf{E}[r_s]}{\beta_s}$. If we return as significant all connected components of size $\geq s^*$, then the FDR of the test is bounded by $\beta$.*

**Proof.** Let $V_s$ be the number of erroneous rejections of connected components of size $s$, i.e., the number of connected components of size $s$ that were flagged erroneously as significant. Note that $E[V_s] \leq E[r_s]$, since if these hypothesis were erroneously rejected they were generated by the null distribution. Let $E_s \equiv$ "$\tilde{r}_s$ conforms with the distribution of $r_s$", and $\bar{E}_s$ be the complementary event.

$$\begin{aligned}
FDR &= \sum_{s=1}^{|\mathcal{K}|} E\left[\frac{V_s}{\tilde{r}_s}\right]\Pr(\bar{E}_s, E_{s-1}, \ldots, E_1)\\
&\leq \sum_{s=1}^{|\mathcal{K}|} \frac{\beta_s E[r_s \mid \bar{E}_s E_{s-1}, \ldots, E_1]}{E[r_s]}\Pr(\bar{E}_s, E_{s-1}, \ldots, E_1)\\
&= \sum_{s=1}^{|\mathcal{K}|} \frac{\beta_s \sum_j j \Pr(r_s = j, \bar{E}_s, E_{s-1}, \ldots, E_1)}{E[r_s]}\\
&\leq \sum_{s=1}^{|\mathcal{K}|} \frac{\beta_s E[r_s]}{E[r_s]} \leq \beta.
\end{aligned}$$

■

## ACKNOWLEDGMENTS

This work is supported by NSF grant 11S-1016648, the Department of Defense Breast Cancer Research Program, the Alfred P. Sloan Foundation, and the Susan G. Komen Foundation. F.V. was supported in part by the MIUR of Italy (project AlgoDEEP prot. 2008TFBWL4) and the Ing. Aldo Gini Foundation, Padova, Italy; this work was done in part while he was affiliated with the Dipartimento di Ingegneria dell'Informazione, Università di Padova, Padova, Italy. B.J.R. was supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

## DISCLOSURE STATEMENT

No competing financial interests exist.

## REFERENCES

Axelson, H. 2004. Notch signaling and cancer: emerging complexity. *Semin. Cancer Biol.* 14, 317–319.

Bader, G.D., Donaldson, I., Wolting, C., et al. 2001. BIND—The Biomolecular Interaction Network Database. *Nucleic Acids Res.* 29, 242–245.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate. *J. R. Stat. Soc.* Ser. B 57, 289–300.

Benjamini, Y., and Yekutieli, D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29, 1165–1188.

Chuang, H.Y., Lee, E., Liu, Y.T., et al. 2007. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.* 3, 140.

Chung, F. 2007. The heat kernel as the pagerank of a graph. *Proc. Nat. Acad. Sci. USA* 104, 19735.

Collins, B.J., Kleeberger, W., and Ball, D.W. 2004. Notch in lung development and lung cancer. *Semin. Cancer Biol.* 14, 357–364.

Cui, Q., Ma, Y., Jaramillo, M., et al. 2007. A map of human cancer signaling. *Mol. Syst. Biol.* 3, 152.

Ding, L., Getz, G., Wheeler, D.A., et al. 2008. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069–1075.

Doyle, P., and Snell, J. 1984. *Random Walks and Electric Networks*. The Mathematical Association of America, Washington, DC.

Feige, U., Kortsarz, G., and Peleg, D. 1999. The dense k-subgraph problem. *Algorithmica* 29, 2001.

Greenman, C., Stephens, P., Smith, R., et al. 2007. Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153–158.

Hahn, W.C., and Weinberg, R.A. 2002. Modelling the molecular circuitry of cancer. *Nat. Rev. Cancer* 2, 331–341.

Hescott, B.J., Leiserson, M.D.M., Cowen, L., et al. 2009. Evaluating between-pathway models with expression data. Proc. *RECOMB* 2009 372–385.

Hochbaum, D.S., ed. 1997. *Approximation Algorithms for NP-Hard Problems*. PWS Publishing Co., Boston.

Hodges, E., Xuan, Z. Balija, V., et al. 2007. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.* 39, 1522–1527.

Ideker, T., Ozier, O., Schwikowski, B., et al. 2002 Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18, Suppl 1, S233–S240.

Jensen, L.J., Kuhn, M., Stark, M., et al. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 37, D412–D416.

Jones, S., Zhang, X., Parsons, D.W., et al. 2008. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321, 1801–1806.

Jonsson, P.F., and Bates, P.A. 2006. Global topological features of cancer proteins in the human interactome. *Bioinformatics* 22, 2291–2297.

Kanehisa, M., and Goto, S. 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27–30.

Karni, S., Soreq, H., and Sharan, R. 2009. A network-based method for predicting disease-causing genes. *J. Comput. Biol.* 16, 181–189.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. 2009. Human Protein Reference Database—2009 update. *Nucleic Acids Res.* 37, D767–D772.

Kirsch, A., Mitzenmacher, M., Pietracaprina, A., et al. 2009. An efficient rigorous approach for identifying statistically significant frequent itemsets. *PODS* 117–126.

Kondor, R.I., and Lafferty, J. 2002. Diffusion kernels on graphs and other discrete structures. *Proc. ICML* 315–322.

Lin, J., Gan, C.M., Zhang, X., et al. 2007. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.* 17, 1304–1318.

Liu, M., Liberzon, A., Kong., S.W., et al. 2007. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.* 3, e96.

Lovász, L. 1993. Random walks on graphs: a survey, 1–46. Combinatorics. Paul Erdös Is Eighty (Volume 2).

Ma, X., Lee, H., Wang, L., et al. 2007. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics* 23, 215–221.

McCormick, F. 1999. Signalling networks that cause cancer. *Trends Cell Biol.* 9, M53–M56.

Nabieva, E., Jim, K., Agarwal, A., et al. 2005. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics* 21, Suppl 1, i302–i310.

Nacu, S., Critchley-Thorne, R., Lee, P., et al. 2007. Gene expression network analysis and applications to immunology. *Bioinformatics* 23, 850–858.

Parsons, D.W., Jones, S., Zhang, X., et al. 2008. An integrated genomic analysis of human glioblastoma multiforme. *Science* 321, 1807–1812.

Qi, Y., Suhail, Y., Lin, Y.Y., et al. 2008. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.* 18, 1991–2004.

Salwinski, L., Miller, C.S., Smith, A.J., et al. 2004. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* 32, D449–D451.

Shuai, T.-P. and Hu, X.-D. 2006. Connected set cover problem and its applications. *Proc. AAIM* 243–254.

Sjoblom, T., Jones, S. Wood, L.D., et al. 2006. The consensus coding sequences of human breast and colorectal cancers. *Science* 314, 268–274.

TCGA (The Cancer Genome Atlas Research Network). 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068.

Tsuda, K., and Noble, W. S. 2004. Learning kernels from biological networks by maximizing entropy. *Bioinformatics* 20, Suppl 1, i326–i333.

Ulitsky, I., Karp, R. M., and Shamir, R. 2008. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. *Proc. RECOMB* 2008 347–359.

Vogelstein, B., and Kinzler, K. W. 2004. Cancer genes and the pathways they control. *Nat. Med.* 10, 789–799.

Wood, L.D., Parsons, D.W., Jones, S., et al. 2007. The genomic landscapes of human breast and colorectal cancers. *Science* 318, 1108–1113.

Address correspondence to:
*Dr. Fabio Vandin*
*Department of Computer Science*
*Brown University*
*Providence, RI 02912*

*E-mail:* vandinfa@cs.brown.edu