On the sample complexity of cancer pathways identification

Vandin, Fabio; Raphael, Benjamin J.; Upfal, Eli

Published in:
Journal of Computational Biology

DOI:
10.1089/cmb.2015.0100

Publication date:
2016

Document version
Peer reviewed version

Citation for pulished version (APA):
Vandin, F., Raphael, B. J., & Upfal, E. (2016). On the sample complexity of cancer pathways identification. Journal of Computational Biology, 23(1), 30-41. DOI: 10.1089/cmb.2015.0100

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 23, Number 1, 2016

Pp. 30–41
DOI: 10.1089/cmb.2015.0100

On the Sample Complexity of Cancer Pathways Identification

FABIO VANDIN,[1,2,4] BENJAMIN J. RAPHAEL,[2,3] and ELI UPFAL[2]

ABSTRACT

Advances in DNA sequencing technologies have enabled large cancer sequencing studies, collecting somatic mutation data from a large number of cancer patients. One of the main goals of these studies is the identification of *all cancer genes*—genes associated with cancer. Its achievement is complicated by the extensive mutational heterogeneity of cancer, due to the fact that important mutations in cancer target combinations of genes (i.e., *pathways*). Recently, the pattern of *mutual exclusivity* among mutations in a cancer pathway has been observed, and methods that find significant combinations of cancer genes by detecting mutual exclusivity have been proposed. A key question in the analysis of mutual exclusivity is the computation of the *minimum number of samples* required to reliably find a meaningful set of mutually exclusive mutations in the data, or conclude that there is no such set. In general, the problem of determining the *sample complexity*, or the number of samples required to identify significant combinations of features, of genomic problems is largely unexplored.

In this work we propose a framework to analyze the sample complexity of problems that arise in the study of genomic datasets. Our framework is based on tools from combinatorial analysis and statistical learning theory that have been used for the analysis of machine learning and probably approximately correct (PAC) learning. We use our framework to analyze the problem of the identification of cancer pathways through mutual exclusivity analysis. We analytically derive matching upper and lower bounds on the sample complexity of the problem, showing that sample sizes much larger than currently available may be required to identify all the cancer genes in a pathway. We also provide two algorithms to find a cancer pathway from a large genomic dataset. On simulated and cancer data, we show that our algorithms can be used to identify cancer pathways from large genomic datasets.

Key words: cancer pathways, exclusivity, PAC learning, VC dimension.

1. INTRODUCTION

High-throughput sequencing technologies now allow the measurement of somatic mutations in cancer genomes from many individuals with different cancer types (Cancer Genome Atlas Network,

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark.
[2]Department of Computer Science, Brown University, Providence, Rhode Island.
[3]Center for Computational Molecular Biology, Brown University, Providence, Rhode Island.
[4]Department of Information Engineering, University of Padova, Italy.

2012; Cancer Genome Atlas Research Network et al., 2013). One of the main objectives of large-scale cancer studies such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, 2008) is to identify all the *cancer genes* responsible for the development of cancer, and to distinguish these from genes containing only *random*, *passenger mutations* not associated with the disease.

Several approaches have been developed to predict cancer genes from the mutations measured in a collection of cancer samples. Popular methods (Dees et al., 2012; Lawrence et al., 2013) identify significant recurrently mutated genes, but while these methods have identified a number of novel cancer genes in different cancer types (Cancer Genome Atlas Research Network et al., 2013), accurate detection of cancer genes through recurrent gene analysis has two limitations. First, it requires a reasonable estimate of the background mutation rate. Unfortunately, this rate varies across cancer types (Kandoth et al., 2013), across patients, and across regions of the genome (Lawrence et al., 2013). Second, cancers exhibit extensive mutational *heterogeneity*, with mutations in different cancer genes occurring in different patients (Vogelstein et al., 2013; Garraway and Lander, 2013). The mutational heterogeneity of cancer is due, in part, to the fact that cancer mutations target signaling, regulatory, and metabolic *pathways* (Vogelstein and Kinzler, 2004). Therefore different individuals may have mutations in different genes within the same cancer pathway. Thus, an alternative to single-gene analyses is to identify recurrent groups of mutations in known pathways or protein interaction networks (Vandin et al., 2011; Ciriello et al., 2012; Shrestha et al., 2014; Vandin et al., 2012a). However, such approaches require accurate knowledge of the interactions between genes/proteins, and this information is only partially available (Raphael et al., 2014).

Ideally, one would like to identify sets of mutated genes *de novo*, without any prior knowledge of pathways, interactions, or background mutation rates. Unfortunately, the number of possible sets of genes, even of moderate size, is enormous, making exhaustive evaluation of these sets impossible due to multiple hypothesis testing considerations. Recently, it has been observed that mutations in a cancer pathway tend to be *mutually exclusive* (that is, a cancer pathway rarely has more than one mutated gene in a sample) (Yeang et al., 2008). Algorithms that identify sets of genes with mutually exclusive mutations have been introduced and used successfully to identify parts of cancer pathways *de novo* from mutation data from a large number of samples (Vandin et al., 2012b; Leiserson et al., 2013; Miller et al., 2011; Szczurek and Beerenwinkel, 2014).

A key question in mutual exclusivity analysis is to determine the *number of samples* that are required to identify (with high probability) a set of mutually exclusive mutations in the data. More generally, the problem of computing the *sample complexity*, or the number of samples required to reliably identify meaningful combinations of features in genomic data, is largely unexplored. This problem is analogous to power calculations that are performed for simple and commonly used statistical tests (Whitley and Ball, 2002). One result on this problem is the work of Ein-Dor et al. (2006) that addressed a similar question for the identification of gene expression signatures in cancer. The work of Perkins and Hallett (2010) provides a bound for the problem of inferring regulatory relationships from gene expression time-series data. While we focus here on the sample complexity of mutually exclusive sets of mutations, our work outlines a general framework for rigorously addressing a key question in computational biology—is the sample size sufficient for accepting or rejecting a postulate hypothesis on the association between genomic variation and a phenotype.

1.1. Contributions

In this article, we propose a framework to analyze the sample complexity of problems that arise in the study of genomic datasets. Our framework is based on tools from statistical learning theory (Mohri et al., 2012; Abu-Mostafa et al., 2012) and combinatorial analysis, which has been used for the mathematical analysis of machine learning and probably approximately correct (PAC) learning (Valiant, 1984). We instantiate our framework to study the problem of finding a *cancer pathway* from genomic data, where we define a *cancer pathway* to be a set of genes with mutually exclusive mutations in a collection of samples. Thus, in each sample, the binary variables defining the mutation status of genes in a cancer pathway satisfy the *exclusive or* (XOR) function. Measurements errors as well as passenger, random mutations are easily captured by our framework.

We analytically derive matching upper and lower bounds on the number of samples required to reliably identify all genes in a cancer pathway from genomic data, showing that sample sizes much larger than those currently available in TCGA and other studies (International Cancer Genome Consortium et al., 2010) may be required. Our upper bound is based on an analysis of the Vapnik-Chervonenkis (VC) dimension of the set of *exclusive or* (XOR) functions, and our lower bound is based on a second moment argument quantifying the effect of random sequencing errors; both may be of independent interest.

Since our analysis shows that conclusive results require the processing of a large number of samples, we also provide two algorithms to identify the cancer pathway *de novo* from large sequencing data. The first algorithm is based on an integer linear problem formulation (ILP) for the problem of finding the XOR function (of k variables among n) that is satisfied by the largest number of samples. The second algorithm is a polynomial time algorithm that identifies the cancer pathway through careful covariance analysis, provided that the number of samples in the dataset satisfies the general upper bound that we derive. We note that while we focus here on exact XOR functions, previous work, including ours (Vandin et al., 2012b; Leiserson et al., 2013; Vandin et al., 2012c; Zhao et al., 2012), considered heuristic approaches that use scoring functions for gene sets to approximate the XOR function considered in this work.

We run our algorithms on simulated and real cancer data, showing that for certain combinations of the problem's parameters the ILP algorithm identifies the cancer pathway using a number of samples that is near the number that may soon be available. On thyroid cancer data, our ILP algorithm identifies a set of genes that overlaps with a key pathway in the pathogenesis of thyroid cancer.

2. METHODS AND ALGORITHMS

2.1. Model

Let $\mathcal{G}$ be the set of genes, with $|\mathcal{G}| = n$. Let $\mathcal{P} \subset \mathcal{G}$ be a *cancer pathway*, that is, a set of genes whose mutations cause cancer. For every cancer sample, we assume that its mutations are generated as follows, independently of all other events:

1. with probability f, exactly one gene in $\mathcal{P}$ is mutated, and the probability that $g \in \mathcal{P}$ is the (only) mutated gene is f_g, with $f = \sum_{g \in \mathcal{P}} f_g \leq 1$ (with probability $1 - f$ the number of mutated genes in $\mathcal{P}$ is $\neq 1$);
2. for each gene $g \in \mathcal{G} \backslash \mathcal{P}$, g is mutated with probability p_g independent of other events.

The model above captures errors in the mutation calling process, due to sequencing errors as well as to false positives/negatives in mutation calls (that may lead a sample to have no mutations in $\mathcal{P}$). Moreover, the model allows for random passenger mutations (not associated with the disease) for genes in $\mathcal{G} \backslash \mathcal{P}$ (i.e., p_g's capture the passenger mutation rate) as well as for genes in $\mathcal{P}$ (i.e., when $f < 1$ there may be multiple mutations, including passenger ones, in $\mathcal{P}$).

2.2. Upper bound on the sample size

Our goal is to use mutation data from the model above to identify $\mathcal{P}$. We study this problem in the *probably approximately correct* (PAC) learning framework. For a gene $g \in \mathcal{G}$, we define a 0-1 variable x_g. Mutations in a sample S define an assignment $x^{(S)}$ of the variables $\{x_g : g \in \mathcal{G}\}$, with $x_g^{(S)} = 1$ if g is mutated in S, and $x_g^{(S)} = 0$ otherwise. Given a set $\mathcal{C} = \{g_1, g_2, \ldots, g_k\}$ of k genes in $\mathcal{G}$ we define the k-XOR function $h_{\mathcal{C}}$ of the corresponding k 0-1 variables. Let $h_{\mathcal{C}}(S)$ be the XOR function defined on $\mathcal{C}$ evaluated on the assignment $x^{(S)}$: $h_{\mathcal{C}}(S) = XOR(x_{g_1}^{(S)}, x_{g_2}^{(S)}, \ldots, x_{gk}^{(S)})$. We say that sample S *satisfies* $h_{\mathcal{C}}$ if $h_{\mathcal{C}}(S) = 1$.

Let $\mathcal{D}$ be the probability distribution on the assignment $x^{(S)}$ defined by the mutation model of section 2.1. Note that by the definition of the model $\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{P}}(S) = 1] = f$.

Theorem 1. *Assume that* $p_g < 0.5$ *for all* $g \in \mathcal{G}$, *and let* $\mathcal{C} \subset \mathcal{G} \backslash \mathcal{P}$. *Then* $\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S) = 1] < 0.5$.

Proof. The proof is by induction on $k = |\mathcal{C}|$. Let's assume that $k = 2$, and let $\mathcal{C} = \{g_0, g_1\}$. Without loss of generality, let $p_{g_1} \leq p_{g_0}$. Then

$$\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S) = 1] = p_{g_0}(1 - p_{g_1}) + p_{g_1}(1 - p_{g_0}) = p_{g_0} + p_{g_1}(1 - 2p_{g_0}).$$

Now let $p_{g_0} = 0.5 - \varepsilon$ for $\varepsilon > 0$. Then

$$\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S) = 1] = p_{g_0} + p_{g_1}(1 - 2p_{g_0}) = 0.5 - \varepsilon + 2\varepsilon p_{g_1} \leq 0.5 - 2\varepsilon^2 < 0.5.$$

Now assume that $\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S)=1] < 0.5$ for all $\mathcal{C} \subset \mathcal{G}\backslash\mathcal{P}$ with $|\mathcal{C}| < k$. Consider $\mathcal{C} \subset \mathcal{G}\backslash\mathcal{P}$ with $|\mathcal{C}|=k$. Consider $g=\arg\min_{g\in C} p_g$, and let $\mathcal{C}_0=\mathcal{C}\backslash\{g\}$. Let $p_{\mathcal{C}_0}=\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}_0}(S)=1]$; by inductive hypothesis, $p_{\mathcal{C}_0} < 0.5$. Then we have:

$$\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S)=1]=p_{\mathcal{C}_0}(1-p_g)+p_g(1-p_{\mathcal{C}_0})=p_{\mathcal{C}_0}+p_g(1-2p_{\mathcal{C}_0}). \quad (1)$$

Analogously to the case $k=2$, we have $\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S)=1] < 0.5$. ■

Let $p^*=\min\{p_g : g \in \mathcal{G}\backslash\mathcal{P}\}$. Let f^* be the minimum contribution to f of a gene in $\mathcal{P}$, that is, $f^*=\min_{g\in\mathcal{P}}\{f_g\}$. We have the following.

Theorem 2. *Assume that $p_g < 0.5$ for all $g \in \mathcal{G}$, and that $f-f^*=\frac{1}{2}+c$ for $c > 0$. Then* $\max_{\mathcal{C}\neq\mathcal{P},|\mathcal{C}|=k} \mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S)=1] \leq \max\{f-f^*-2cp^*, 0.5\}$.

Proof. Consider a set $\mathcal{C} \neq \mathcal{P}$, and let $\mathcal{C}_{\mathcal{P}}=\mathcal{C}\cap\mathcal{P}$ and $\mathcal{C}_G=\mathcal{C}\cap(\mathcal{G}\backslash\mathcal{P})$. Note that the probability that $\mathcal{C}_{\mathcal{P}}$ has exactly 1 mutation in a sample S from $\mathcal{D}$ is at most $f-f^*$, and that the probability that $\mathcal{C}_G$ has exactly 1 mutation in a sample is at most 0.5 (by Theorem 1). Also note that mutations in $\mathcal{C}_G$ and in $\mathcal{C}_{\mathcal{P}}$ are independent. Let $p_{\mathcal{C}_G}$ (resp., $p_{\mathcal{C}_{\mathcal{P}}}$) be the probability that $\mathcal{C}_G$ (resp., $\mathcal{C}_{\mathcal{P}}$) is mutated in a sample. The $\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S)=1]=p_{\mathcal{C}_{\mathcal{P}}}+p_{\mathcal{C}_G}(1-2p_{\mathcal{C}_{\mathcal{P}}}) \leq f-f^*-2cp^*$. Moreover, if $\mathcal{C}_{\mathcal{P}}=\emptyset$, then $\mathbf{Pr}_{\mathcal{D}}[h_{\mathcal{C}}(S)=1]<0.5$ by Theorem 1, and the result follows. ■

From the above, we can conclude that if[1] $f>0.5$, with enough samples from $\mathcal{D}$ we will be able to identify the set $\mathcal{P}$ by identifying the set of k genes whose XOR function is satisfied by the largest number of samples (due to the concentration of binomial random variables). In the following we estimate the number of samples required to identify $\mathcal{P}$. In order to estimate the number of samples required to identify $\mathcal{P}$, we bound the VC dimension of the set of k-XOR functions on n variables.

We define a *range space* as a pair (X, R) where X is a set and R is a family of subsets of X. Given $I \subset X$, the *projection* $P_R(I)$ of R on I is defined as $P_R(I)=\{r\cap I : r \in R\}$. if $P_R(I)=2^I$ (or, equivalently, $|P_{R(I)}|=2^{|I|}$), then I is said to be shattered by R.

Definition 1 (Vapnik and Chervonenkis, 1971). *Let $H=(X, R)$ be a range space. The VC dimension $VC(H)$ of H is the maximum cardinality of a shattered subset of X. If there are arbitrary large shattered subsets, then $VC(H)=\infty$.*

Let $XOR_{n,k}$ be the set of $\binom{n}{k}$ k-XOR functions on n variables. We define the *range space* $H=(X, R)$ where X is the set of all Boolean vectors on n variables, and each $r \in R$ is a set of vectors in X that satisfy a given k-XOR function. Note that $|X|=2^n$ and $|R|=|XOR_{n,k}|=\binom{n}{k}$.

Theorem 3. *The VC dimension $VC(H)$ of $H=(X, R)$ is $\Theta(k\log n - k\log k)$.*

Proof. *Upper bound.* Assume that ℓ vectors are shattered, then there is a distinct k-XOR function for each of the 2^ℓ subsets of the ℓ vectors. Thus, $2^\ell \leq \binom{n}{k}$, or $\ell \leq \log_2\binom{n}{k}=O(k\log n - k\log k)$.

Lower Bound. We construct a set of $\ell=k\log n - k\log k$ vectors, and a set of k-XOR functions that shatters this set of vectors. Let $C=\{v_1, \ldots, v_\ell\}$ be the set of vectors. We partition the set C into k disjoint sets $C_1, \ldots, C_k$, each with $c=\log n - \log k$ vectors. Let $S_i^1=\emptyset, S_i^2, \ldots, S_i^{2^c}$ be all the $2^c=\frac{n}{k}$ subsets of the set C_i. The vectors v are constructed as follows:

1. v has 0 in position $1+(i-1)\frac{n}{k}$ for $i=1, \ldots, k$;
2. for each set S_i^j such that $v \in S_i^j$, v has a 1 in position $(i-1)(\frac{n}{k}-1)+j$.

Consider a dichotomy on the set C, and let D be the set of vectors for which the value of the dichotomy is 1. We construct an XOR function that expresses that dichotomy as follows: for each $i=1,\ldots,k$,

[1]While $f>0.5$ may not be satisfied by all cancer pathways, it is a reasonable assumption for the most important cancer pathways (Ciriello et al., 2013).

1. if $D \cap C_i = \emptyset$ then add $X_{(i-1)\frac{n}{k}+1}$ to the XOR; otherwise
2. if $D \cap C_i = S_i^j$ then add $X_{(i-1)(\frac{n}{k}-1)+j}$ to the XOR.

The XOR has exactly k variables. Consider a vector $v \in C_i$. The 1's in v are in the range $(i-1)(\frac{n}{k}+2)$ to $i(\frac{n}{k}-1)-1$. If $v \in S_i^j \subset D$ then v has a 1 in the location $(i-1)(\frac{n}{k}-1)+j$ and the value of the XOR is 1, otherwise it's 0. ■

We also generalize the result above to the case of functions that are AND or OR of multiple XOR functions. Let $AXOR_{n,k,h}$ (respectively, $OXOR_{n,k,h}$) be the set of $\left(\binom{n}{k}\right)^h$ functions on n variable, where each function is an AND (resp., OR) of h k-XOR functions on n variables. Let (X, R_1) (resp., (X, R_2)) be a range space, where X is the set of all Boolean vectors on n variables, and each $r \in R_1$ (resp., $r \in R_2$) is a set of vectors in X that satisfy a function in $AXOR_{n,k,h}$ (resp., $OXOR_{n,k,h}$). We have the following.

Theorem 4. *The VC dimension of the range space* (X, R_1) *and of the range space* (X, R_2) *is* $\Theta(hk \log n - hk \log k)$.

Proof (sketch). We provide the proof for the range space (X, R_1). The proof for the range space (X, R_2) is analogous.

Upper bound. Assume that ℓ vectors are shattered, then there is a distinct (h, k)-AXOR function for each of the 2^ℓ subsets of the ℓ vectors. Thus, $2^\ell \leq \left(\binom{n}{k}\right)^h$, or $\ell \leq h \log_2 \binom{n}{k} = O(hk \log n - hk \log k)$.

Lower bound. We construct a set of $\ell = hk \log n - hk \log k - hk \log h$ vectors, and a set of (h, k)-AXOR functions that shatters this set of vectors.

The construction is similar to the one in the proof of the previous theorem, only here we partition the ℓ vectors to hk disjoint sets of $\frac{n}{hk}$ vectors each. Let $C = \{v_1, \ldots, v_\ell\}$ be the set of vectors. For $i = 1, \ldots, k$ and $j = 1, \ldots, h$ let $C_{i,j}$ denote the sets, each with $c = \log n - \log hk$ vectors. Let $S_{i,j}^1 = \emptyset, S_{i,j}^2, \ldots, S_{i,ji}^{2^c}$ be all the $2^c = \frac{n}{hk}$ subsets of the set $C_{i,j}$.

The construction of the vectors is the same as before with the exception that if a vector is not in $\cup_{i=1}^k C_{i,j}$, for some j, then the position corresponding to $S_{1,j}^1$ is set to 1 (so the corresponding AND gives 1). ■

Let $\mathcal{T}$ be a collection of m samples from the model of section 2.1. Let $\hat{\mathcal{T}}$ be the probability distribution on the assignment $x^{(S)}$ defined by taking a sample S uniformly at random from $\mathcal{T}$. The following result bounds the difference between the fraction of samples in $\mathcal{T}$ that satisfy $h_\mathcal{C}$ and the probability that a random sample from $\mathcal{D}$ satisfies $h_\mathcal{C}$.

Theorem 5. *With probability* $\geq 1 - \delta$ *the following are satisfied for all* $\mathcal{C} \subset \mathcal{G}$, $|\mathcal{C}| = k$ *simultaneously:*

- $|\mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) = 1] - \mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1]| \leq \sqrt{\frac{\log \binom{n}{k} + \log \frac{2}{\delta}}{2m}}$;
- $|\mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) = 1] - \mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1]| \leq \sqrt{\frac{VC(H)(1 + \log m) + \log \frac{4}{\delta}}{m}}$;
- $|\mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) = 1] - \mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1]| \leq \sqrt{\frac{VC(H)(1 + \log \frac{2m}{VC(H)}) + \log \frac{1}{4\delta}}{m}}$.

Proof. For a given set $\mathcal{C} \subset \mathcal{G}$ and the corresponding function $h_\mathcal{C}$, define the in-sample error $E_{in} = \frac{1}{m} \sum_{S \in \mathcal{T}} \mathbb{1}[h_\mathcal{C}(S) \neq 1]$ where $\mathbb{1}[\cdot]$ is the indicator function. Define the out-of-sample error $E_{out} = \mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) \neq 1]$. Note that $\mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1] = \frac{1}{m} \sum_{S \in \mathcal{T}} \mathbb{1}[h_\mathcal{C}(S) = 1] = 1 - \frac{1}{m} \sum_{S \in \mathcal{T}} \mathbb{1}[h_\mathcal{C}(S) \neq 1] = 1 - E_{in}$ and that $\mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) = 1] = 1 - \mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) \neq 1] = 1 - E_{out}$. Then $|\mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) = 1] - \mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1]| = |E_{in} - E_{out}|$, and the results follow directly from known (generalization) bounds (Bousquet et al., 2003; Mohri et al., 2012; Abu-Mostafa et al., 2012) on $|E_{in} - E_{out}|$. ■

Combining Theorem 2 and Theorem 5, we show that if $\mathcal{T}$ consists of $m = O(k \log n - k \log k)$ samples, with high probability the XOR function that is satisfied by the largest number of samples in $\mathcal{T}$ is given by the pathway $\mathcal{P}$.

Corollary 1. *If* $m = O(k \log n - k \log k)$, *with high probability:* $\mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{P}(S) = 1] > \mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1]$ *for all* $\mathcal{C} \neq \mathcal{P}$, $|C| = k$.

Proof (sketch). Consider bound (1) in Theorem 5. Let $\varepsilon = f^* + 2p^*(f - f^* - 0.5)$, and $m \geq 4 \frac{\log \binom{n}{k}}{\varepsilon^2}$. Then by Theorem 2: $\mathbf{Pr}_\mathcal{D}[h_\mathcal{P}(S) = 1] - \mathbf{Pr}_\mathcal{D}[h_\mathcal{C}(S) = 1] > \varepsilon$ for all $\mathcal{C} \neq \mathcal{P}$, $|C| = k$. By Theorem 5, with probability at least $1 - O(\frac{1}{n})$ we have $\mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{P}(S) = 1] > \mathbf{Pr}_\mathcal{D}[h_\mathcal{P}(S) = 1] - \varepsilon/2$ and for all $\mathcal{C} \neq \mathcal{P}$, $|C| = k$: $\mathbf{Pr}_{\hat{\mathcal{T}}}[h_\mathcal{C}(S) = 1] < \mathbf{Pr}_\mathcal{D}[h_C(S) = 1] + \varepsilon/2$, and the result follows. Similar derivations apply for bounds (2) and (3) of Theorem 5. ■

2.3. Lower Bound on the minimum sample size

Using the VC dimension, which is a combinatorial property of the set of functions, one can obtain a lower bound on the sample complexity (matching the upper bound of Corollary 1) that applies to the worst case input distribution (Mohri et al., 2012). In this section we show that $\Omega(k \log n - k \log k)$ samples are required to identify the cancer pathway $\mathcal{P}$ even for the special case of input distribution defined by our model. In particular, we show that with a dataset of smaller size with high probability there is a set $\mathcal{C}$ of k genes from $\mathcal{G} \backslash \mathcal{P}$ whose function $h_{\mathcal{C}}$ is satisfied by at least fm samples, and thus cannot be distinguished from $\mathcal{P}$.

Theorem 6. *Let $\mathcal{T}$ be a dataset of $m = o(k \log n - k \log k)$ samples. Then with high probability there exists $\mathcal{C} \subset \mathcal{G} \backslash \mathcal{P}$, $|\mathcal{C}| = k$ such that $h_{\mathcal{C}}$ is satisfied by at least fm samples, that is: $\mathbf{Pr}_{\hat{\mathcal{T}}}[h_{\mathcal{C}}(S) = 1] \geq fm$.*

Proof. To simplify the presentation we assume that there are n nonsignificant genes, all mutated randomly with the same probability $0 < p < 1/2$. We also assume $k = O(1)$. In our model, for a set of nonsignificant genes $\mathcal{C}$, $\alpha = \mathbf{Pr}_{\mathcal{D}}(h_{\mathcal{C}}(S) = 1) = kp(1-p)^{k-1} < 1/2$.

We now define a collection of $\binom{n}{k}$ Bernoulli random variables $Z_{\mathcal{C}}$, for each $\mathcal{C} \subset \mathcal{G}$, such that $Z_{\mathcal{C}} = 1$ if the function $h_{\mathcal{C}}$ is satisfied by at least fm samples, and $Z_{\mathcal{C}} = 0$ otherwise. Then $\mathbf{Pr}(Z_{\mathcal{C}}) = \sum_{j=fm}^{m} \binom{m}{j} \alpha^j (1-\alpha)^{m-j} \geq \alpha^m$. Denote the expected number of sets whose functions are satisfied by at least fm samples by $\mu(m) = \mathbf{E}\left[\sum_{\mathcal{C} \subset \mathcal{G}} Z_{\mathcal{C}}\right] = \binom{n}{k} Pr(Z_{\mathcal{C}} = 1)$. Note that since $\alpha < 1/2$ and $f > 1/2$, the expectation $\mu(m)$ is monotonically decreasing in m. Furthermore, since $Pr(Z_{\mathcal{C}} = 1) \geq \alpha^m$, there is a constant $c_1 > 0$ such that for $m = c_1(k \log n - k \log k)$, $\mu(m) > 2$. For our proof we use $m = c(k \log n - k \log k)$ where $c = \min[c_1, c_3]$, for a constant $c_3 > 0$ defined below.

Next, we will apply the second moment method (Mitzenmacher and Upfal, 2005) [Theorem 6.7] to bound the probability that $\sum_{\mathcal{C} \subset \mathcal{G}} Z_{\mathcal{C}} = 0$. To apply this method we define for each set $\mathcal{C}$ the *neighborhood set* of $\mathcal{C}$: $I(\mathcal{C}) = \{\mathcal{C}' | \mathcal{C} \cap \mathcal{C}' \neq \emptyset, |\mathcal{C}'| = |\mathcal{C}| = k\}$. If $\mathcal{C}' \notin I(\mathcal{C})$ then $Z_{\mathcal{C}}$ and $Z_{\mathcal{C}'}$ are independent.

Applying Mitzenmacher and Upfal (2005) [Lemma 6.9]:

$$\mathbf{Var}\left[\sum_{\mathcal{C} \in \mathcal{G}} Z_{\mathcal{C}}\right] \leq \mu(m) + \sum_{\mathcal{C}} \sum_{\mathcal{C}' \neq \mathcal{C} \in I(\mathcal{C})} \mathbf{E}[Z_{\mathcal{C}} Z_{\mathcal{C}'}].$$

For two sets $\mathcal{C}$ and $\mathcal{C}'$ such that $|\mathcal{C} \cap \mathcal{C}'| = k - \ell$: $Pr(h_{\mathcal{C}'}(S) = 1 | h_{\mathcal{C}}(S) = 1) = \frac{k-\ell}{k}(1-p)^{\ell} + \frac{\ell}{k}\ell p(1-p)^{\ell-1} \leq \frac{\ell}{k}(1-p)^{\ell}$.

Assume that $Z_{\mathcal{C}} = 1$ and $Z_{\mathcal{C}'} = 1$, then among the fm samples that satisfy $h_{\mathcal{C}}$ there are t samples that satisfied both functions, and there is an additional set of $fm - t$ samples, such that each of these samples satisfies $h_{\mathcal{C}'}$. Therefore,

$$\begin{aligned}
\mathbf{E}[Z_{\mathcal{C}} Z_{\mathcal{C}'}] &\leq Pr(Z_{\mathcal{C}} = 1) \sum_{i=0}^{fm} \binom{fm}{t} \binom{m - fm}{fm - t} (kp(1-p)^{k-1})^{fm-t} \left(\frac{\ell}{k}(1-p)^{\ell}\right)^t \\
&= (Pr(Z_{\mathcal{C}} = 1))^2 \sum_{i=0}^{fm} \binom{fm}{t} \frac{\binom{m-fm}{fm-t}}{\binom{m}{fm}} \left(\frac{\frac{\ell}{k}(1-p)^{\ell}}{kp(1-p)^{k-1}}\right)^t \\
&\leq (Pr(Z_{\mathcal{C}} = 1))^2 \sum_{i=0}^{fm} \frac{(fm)^t}{t!} \left(\frac{f}{1-f}\right)^t \left(\frac{1}{kp(1-p)^{k-1}}\right)^t \\
&\leq (Pr(Z_{\mathcal{C}} = 1))^2 e^{\mathcal{C}_2 \frac{m}{k}}
\end{aligned}$$

for $\mathcal{C}_2 = f^2/((1-f)p(1-p)^{k-1})$.

Let $\mathcal{C}_3 = 1/2\mathcal{C}_2$. With our choice of m,

$$\begin{aligned}
\sum_{\mathcal{C}} \sum_{\mathcal{C}' \neq \mathcal{C} \in I(\mathcal{C})} \mathbf{E}[Z_{\mathcal{C}} Z_{\mathcal{C}'}] &\leq \binom{n}{k} \sum_{\ell=1}^{k-1} \binom{k}{k-\ell} \binom{n-k}{\ell} (Pr(Z_{\mathcal{C}} = 1))^2 e^{\mathcal{C}_2 \frac{m}{k}} \\
&\leq \mu(m)^2 \frac{2^{2k}}{n} e^{\mathcal{C}\frac{m}{k}} = \mu(m)^2 O\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

$\mathbf{Var}\left[\sum_{\mathcal{C}\in\mathcal{G}} Z_{\mathcal{C}}\right] \leq \mu(m) + \sum_{\mathcal{C}} \sum_{\mathcal{C}'\neq\mathcal{C}\in I(\mathcal{C})} \mathbf{E}[Z_{\mathcal{C}} Z_{\mathcal{C}'}] \leq \mu(m) + \mu(m)^2 O\left(\frac{1}{\sqrt{n}}\right)$, and applying the second moment method: $Pr(\sum_{\mathcal{C}\in\mathcal{G}} Z_{\mathcal{C}} = 0) \leq Pr(|\sum_{\mathcal{C}\in\mathcal{G}} Z_{\mathcal{C}} - \mu(m)| \geq \mu(m)) \leq \frac{Var[\sum_{\mathcal{C}\in\mathcal{G}} Z_{\mathcal{C}}]}{\mu(m)^2} \leq O\left(\frac{1}{\sqrt{n}}\right)$, and with probability $1 - O\left(\frac{1}{\sqrt{n}}\right)$ there exists $\mathcal{C} \subset \mathcal{G}\backslash\mathcal{P}$ of k genes whose XOR function is satisfied by as many samples as the actual cancer pathway. ■

2.4. An ILP to find the best XOR

The results of section 2.2 show that when enough samples are provided the *best* XOR (i.e., the one that is satisfied by the largest number of samples) on k genes identifies the cancer pathway $\mathcal{P}$. However, these results do not provide an algorithm to identify the *best* XOR. We provide the integer linear program (ILP) formulation for the problem of identifying the *best* XOR of k genes. Let $M_{i,j} = 1$ if gene j is mutated in sample i, and $M_{i,j} = 0$ otherwise. Let x_j be a 0-1 variable with $x_j = 1$ if gene j is in the solution, and $x_j = 0$ otherwise. Let y_i be an auxiliary 0-1 variable, with $y_i = 1$ if the solution contains at least one 1 in sample i, and $y_i = 0$ otherwise. Let $z_i = 1$ be an auxiliary 0-1 variable, with $z_i = 1$ if the solution contains more than one 1 in sample i, and $z_i = 0$ otherwise. A solution to our problem then satisfies the following constraints:

1. k genes are included in the solution: $\sum_{j=1,\ldots,n} x_j = k$;
2. for every sample, the solution is mutated in the sample if at least one of the genes in the solution is mutated in the sample: $\forall i, 1 \leq i \leq m : \sum_{j=1,\ldots,n} M_{i,j} x_j \geq y_i$;
3. for every sample i, if there is more than one mutation in the solution, then $z_i = 1 : \forall_i, 1 \leq i \leq m : kz_i \geq (\sum_{j=1,\ldots,n} M_{i,j} x_j) - y_i$.

The objective function is $\max \sum_{i=1,\ldots,m} (y_i - z_i)$, counting the number of samples for which the solution has exactly 1 mutation (i.e., $y_i = 1$ and $z_i = 0$). (This assumes that $z_i = 0$ when the solution contains exactly 1 mutation in a sample that is not enforced by the constraints but is achieved when the objective function is maximized.)

2.5. Polynomial time algorithm for identifying the significant pathway

We also provide a polynomial time algorithm (Algorithm 1) that identifies the cancer pathway $\mathcal{P}$ when the number of samples is as derived in section 2.2. We consider a slightly more detailed model than the one in section 2.1, that is, we assume that mutations in a sample are generated as follows, independently of all other events:

1. with probability f, exactly one gene in $\mathcal{P}$ is mutated, and the probability that $g \in \mathcal{P}$ is the (only) mutated gene is f_g, with $f = \sum_{g\in\mathcal{P}} f_g \leq 1$ (with probability $1 - f$ the number of mutated genes in $\mathcal{P}$ is $\neq 1$);
2. for each gene $g \in \mathcal{G}$ (not mutated in 1), g is mutated with probability p_g independent of other events.

Note that a gene $g \in \mathcal{P}$ has two chances to be mutated, once as a unique mutation in $\mathcal{P}$, and once as a random mutation. We require $p_g \leq f_g$; that is, the error rate is no larger than the actual signal.

Algorithm 1: FindDriverPathway

Data: m vectors $x^1, \ldots, x^m$, where $x_i^\ell = 1$ if gene i is mutated in sample ℓ, otherwise $x_i^\ell = 0$
Result: set $\mathcal{O}$ of genes
$\mathcal{O} \leftarrow \emptyset$; **for** $i = 1, \ldots, n$ **do** $r_i \leftarrow \frac{1}{m}\sum_{\ell=1}^{m} x_i^\ell$;
for $i = 1, \ldots, n$ **do**
  for $j = 1, \ldots, n$ **do**
    if $i \neq j$ **then** $\mathcal{C}_{i,j} \leftarrow \frac{1}{m}(\sum_{\ell=1}^{m} x_i^\ell x_j^\ell) - r_i r_j$;
  end
  $H \leftarrow k - 1$ genes corresponding to the $k - 1$ smallest elements in $\{\mathcal{C}_{i,j} \mid i \neq j\}$ (ties broken arbitrarily) ;
  if $\sum_{j\in H} \mathcal{C}_{i,j} < -\sqrt{\frac{3r_i(\sum_{j\in H} r_j)\log\left(\frac{2n}{\delta}\binom{n-1}{k-1}\right)}{m}}$ **then** $\mathcal{O} \leftarrow \mathcal{O} \cup \{i\}$;
end
return $\mathcal{O}$;

The following theorem shows that when m is large enough, Algorsreithm 1 identifies $\mathcal{P}$ with high probability.

Theorem 7. *If $m \in O(k \log n - k \log k)$, then $\mathcal{O} = \mathcal{P}$ with probability $\geq 1 - \delta$.*

Proof. Let $X_i = 1$ if gene i is mutated in a random sample, else $X_i = 0$. For any pair of genes $i \neq j$, if either i and/or j are not in $\mathcal{P}$ then $Cov[X_i, X_j] = 0$. If both i and j are in $\mathcal{P}$ then $E[X_i] = f_i + (1 - f_i)_{pi}$, and

$$Cov[X_i, X_j] = f_i p_j + f_j p_i + (1 - f_i - f_j) p_i p_j - (f_i + (1 - f_i) p_i)(f_j + (1 - f_j) p_j) = -f_i f_j (1 - p_i)(1 - p_j).$$

We need to show that with $m = O(k \log n - k \log k)$ independent samples the empirical (observed) estimates of the covariances are sufficiently concentrated to distinguish between the two cases. We cannot prove that for individual pairs of genes, instead we use the linearity of the covariance. For each gene i, we compare the sum of the $k - 1$ smallest empirical covariances (among the $n - 1$ covariances of gene i) to a predefined threshold. We show that these values are sufficiently concentrated to identify the pathway.

The empirical estimate of $E[X_i]$ is $r_i = \frac{1}{m} \sum_{\ell=1}^{m} x_i^\ell$, and for the empirical estimate of $Cov\ [X_i, X_j]$ we use $\mathcal{C}_{i,j} = \frac{1}{m} \left(\sum_{\ell=1}^{m} x_i^\ell x_j^\ell \right) - r_i r_j$ (dividing by $m - 1$ just complicates the calculation).

Let $r = (r_1, \ldots, r_n)$. Fix a gene i and a set H of $k - 1$ other genes. Conditioned on r, we bound the probability that the empirical estimate $\sum_{j \in H} \mathcal{C}(i, j)$ is far from the correct value of $\sum_{j \in H} Cov[X_i, X_j]$ by more than $t_{i,H} = \sqrt{\frac{3 r_i \left(\sum_{j \in H} r_j \right) \log \left(\frac{2n}{\delta} \binom{n-1}{k-1} \right)}{m}}$. Equivalently we bound the probability that the observed value of $\sum_{j \in H} \sum_{\ell=1}^{m} x_i^\ell x_\ell^\ell$ is far from its expectation by at least $t_{i,H}$.

$\sum_{j \in H} \sum_{\ell=1}^{m} x_i^\ell x_j^\ell$ is a sum of $0 - 1$ random variables subject to the condition on r. When i is not in $\mathcal{P}$, and when i is in $\mathcal{P}$ and none of the genes in H is in $\mathcal{P}$, the expectation of the sum is $m r_i \left(\sum_{j \in H} r_j \right)$. Otherwise, the expectation of the sum is less than that value.

For a given i and H, and $\epsilon_{i,H} = \sqrt{\frac{3 \log (2n \binom{n-1}{k-1} / \delta)}{m r_i (\sum_{j \in H} r_j)}}$,

$$Pr\left(\left| \sum_{j \in H} \mathcal{C}(i, j) - \sum_{j \in H} Cov[X_i, X_j] \right| \geq t_{i,H} \right) \leq 4 e^{-m r_i (\sum_{j \in H r_j}) \epsilon_{i,H}^2 / 3} \leq \frac{\delta}{2n \binom{n-1}{k-1}}.$$

(To satisfy the conditioning on r we use a Chernoff bound on a Poisson approximation.)

Thus, with probability $1 - \delta/2$ the above holds for all genes and all subsets of $k - 1$ covariances. In particular, with that probability the algorithm does not include in the path any gene that is not in $\mathcal{P}$.

To show that the algorithm chooses the correct genes to include in the pathway it remains to show that for i and H in $\mathcal{P}$: $-f_i (1 - p_i)(\sum_{j \in H} f_j (1 - p_j)) + t_{i,H} < -t_{i,H}$, or

$$f_i (1 - p_i)(\sum_{j \in H} f_j (1 - p_j)) \ > \ 2 \sqrt{\frac{3 r_i (\sum_{j \in H} r_j) \log (n \binom{n-1}{k-1} / \delta)}{m}}. \tag{2}$$

Since we required for all i, $p_i \leq f_i$: $E[r_i] = f_i + p_i - f_i p_i \leq 2 f_i$. For $m \geq \frac{12}{f^*} \log \frac{2n}{\delta}$, we have for all i: $n Pr(r_i \geq 4 f_i) \leq e^{-m f_i / 12} \leq \frac{\delta}{2}$.

Let $\hat{p} = \max_{g \in \mathcal{G}} \{p_g\}$. To satisfy Equation (2) we need

$$m = 4 \frac{3 r_i (\sum_{j \in H} r_j) \log (n \frac{n-1}{k-1} / \delta)}{(f_i (1 - p_i)(\sum_{j \in H} f_j (1 - p_j)))^2} \leq 4 \frac{48 f_i (\sum_{j \in H} f_j) \log (n \binom{n-1}{k-1} / \delta)}{(f_i (1 - p_i)(\sum_{j \in H} f_j (1 - p_j)))^2}$$

$$\leq \frac{192 \ \log (n \frac{n-1}{k-1} / \delta)}{f^* (f - f^*)(1 - \hat{p})^4} = O(k \log n - k \log k).$$

■

3. RESULTS

In this section we present the results of our experimental analysis on simulated data and on data from thyroid cancer. The ILP formulation was solved using CPLEX v12.3 with default parameters.

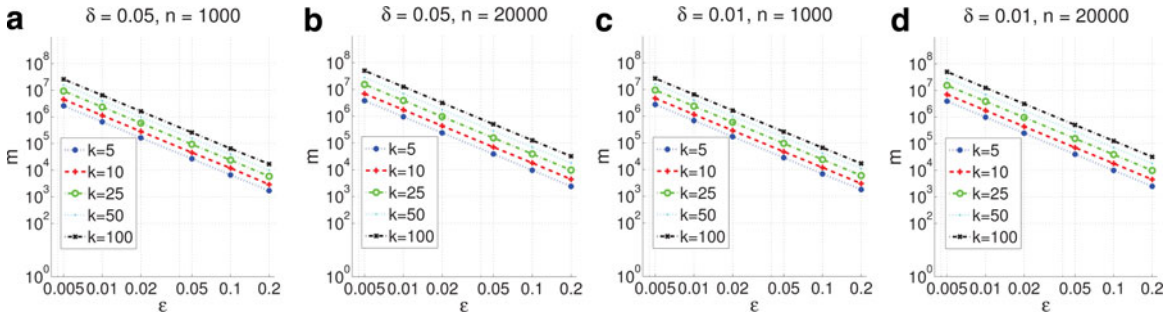

FIG. 1. Number of samples required to find all genes in the cancer pathway $\mathcal{P}$, obtained from analytical bounds. For every pair (k, ε), with $k = |\mathcal{P}|$ and $\varepsilon = f^* + 2(f - f^* - 0.5)p^*$, where f is the probability that exactly one gene in $\mathcal{P}$ is mutated, f^* is the minimum frequency of mutation of a gene in $\mathcal{P}$, and p^* is the minimum probability of mutation of any analyzed gene, we show the number m of samples required to identify $\mathcal{P}$ with probability $\geq 1 - \delta$ when n total genes are analyzed. **(a)** Results for $\delta = 0.05$, $n = 1000$. **(b)** Results for $\delta = 0.05$, $n = 20000$. **(c)** Results for $\delta = 0.01$, $n = 1000$. **(d)** Results for $\delta = 0.01$, $n = 20000$.

3.1. Simulated data

We used the bounds obtained in section 2.2 to estimate the number of samples required to identify all genes in the cancer pathway $\mathcal{P}$. In particular, we considered the case of k genes in $\mathcal{P}$ for $k = 5$, 10, 25, 50, 100 [these are reasonable values for cancer pathways (Vogelstein et al., 2013)], and a total of n genes analyzed. We considered different values for the difference ε between the probability of exclusive mutations in $\mathcal{P}$ and the probability of exclusive mutations in any other set of genes of cardinality k. As from Theorem 5, ε is a function of the minimum frequency f^* of mutation of a gene in $\mathcal{P}$, the probability f that a sample has exactly one mutation in $\mathcal{P}$, and the minimum probability p^* of mutation of any analyzed gene: $\varepsilon = f^* + 2(f - f^* - 0.5)p^*$.

For each pair (k, ε) we estimated the number m of samples required to identify $\mathcal{P}$ with probability at least $1 - \delta$ when n total genes are considered, for values of $\delta = 0.01$, 0.05. We considered the cases $n = 1000$, 20000, corresponding to the case where the most mutated genes are analyzed and to the case where all genes are analyzed, respectively.

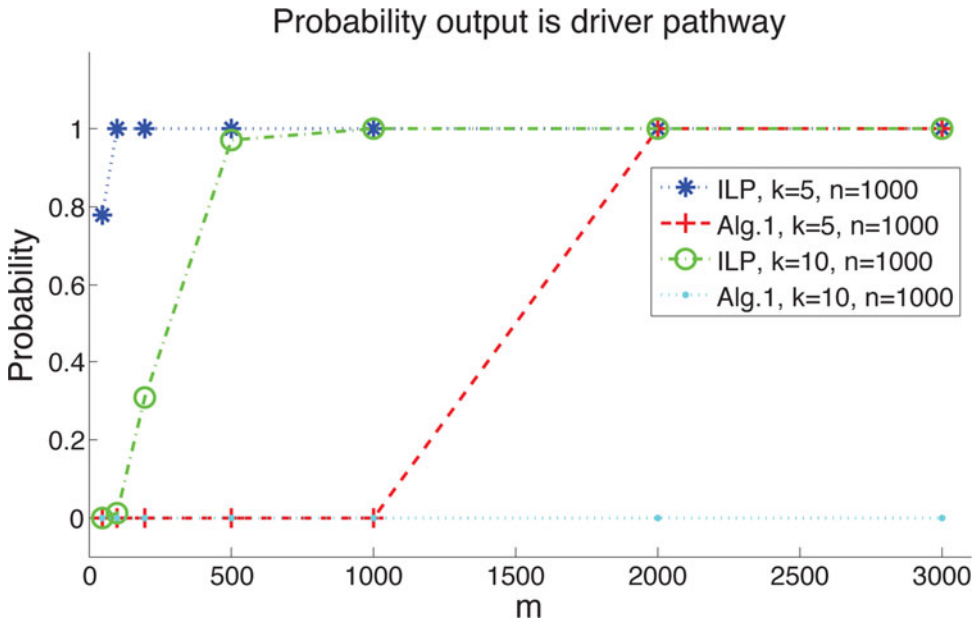

FIG. 2. Probability that the cancer pathway is identified by an algorithm, as function of m. $k = 5$, $n = 1000$ refers to a model with $k = 5$ genes in $\mathcal{P}$ and $n = 1000$ passenger genes with the following parameters: $f = 0.95$, $f^* = 0.1$, $p^* = 0.1$, $\hat{p} = 0.2$; Also, $k = 10$, $n = 1000$ refers to a model with $k = 10$ genes in $\mathcal{P}$ and $n = 1000$ passenger genes with the following parameters: $f = 0.95$, $f^* = 0.01$, $p^* = 0.01$, $\hat{p} = 0.05$. ''ILP'' denotes the results obtained solving the ILP formulation, while ''Alg. 1'' denotes the results with Algorithm 1. Probabilities are estimated using 100 permutations.

The results (Fig. 1) show that even for small values of the pathway size k and for reasonable values of ε [e.g., if the probability that $\mathcal{P}$ contains exactly one mutation in a sample is 0.95, the minimum frequency of mutation of a gene in $\mathcal{P}$ is 0.01, and only genes with mutation probability ≥ 0.01 are analyzed to focus on clinically important frequencies (Lawrence et al., 2013), then $\varepsilon = 0.019$] the number m of samples required to reliably identify all genes in $\mathcal{P}$ is much larger than the sample sizes currently available (i.e., <1000 samples for a given cancer type). We also used simulated data to compare the ability of the ILP formulation and Algorithm 1 to identify $\mathcal{P}$. For different values of k and ε, we estimated the probability that the best XOR identifies $\mathcal{P}$ when m samples are provided, for different values of m; for the same values of k,ε and m we estimated the probability that Algorithm 1 identifies $\mathcal{P}$ (Fig. 2). The results show that the ILP formulation requires less samples than Algorithm 1 to identify $\mathcal{P}$. Moreover, for some choice of the parameters, even when the number of samples is much lower than provided by the analytical bounds, the ILP formulation and Algorithm 1 are able to reliably identify $\mathcal{P}$.

3.2. *Cancer data*

We analyzed cancer data from 399 samples of thyroid carcinoma from TCGA, available through the International Cancer Genome Consortium data portal. We considered somatic mutations, discarding synonymous variants, noncoding exon variants, and variants in intergenic regions. We only considered genes mutated in at least 1% of the samples, for a total of 163 genes, and $k = 4$. Due to the relatively small sample size, we only used the ILP algorithm. We identified the set of genes {BRAF,CSDE1,EIF1AX,HRAS} that present perfectly exclusive (i.e., exactly 1) mutations in 72% of the samples ($p < 0.01$ by permutation test that preserves the frequency of mutation of the single genes). HRAS and BRAF are two well-known thyroid cancer genes (Cohen et al., 2003; Kimura et al., 2003) while EIF1AX has not been previously reported in thyroid cancer, but its recurrent mutation in other cancer types (Martin et al., 2013) suggests that EIF1AX is a novel thyroid cancer gene.

4. CONCLUSION

In this article, we propose a framework to analyze the sample complexity of problems that arise in the study of genomic datasets. Our framework is based on tools from combinatorial analysis and statistical learning theory that have been used for the theoretical analysis of machine and PAC learning. Using our framework, we derive matching analytical upper and lower bounds on the sample complexity of the identification of cancer pathways using mutual exclusivity. To simplify the presentation we focus on the sample complexity as a function of the two major factors, the total number n of genes analyzed and the number k of genes in the cancer pathway; more elaborate calculations express the complexity also as a function of the probability f of exclusivity in the cancer pathway and the passenger mutation probabilities p_g, and will be presented in the full version of this extended abstract. Our results show that sample sizes much larger than those currently available in large cancer studies (e.g., TCGA) may be required. Our upper bound relies on an analysis of the VC dimension of XOR functions, and we derive our lower bound using a second moment argument that quantifies the impact of random sequencing errors on XOR functions; both may be of independent interest. We also provide two algorithms for finding cancer pathways from large sequencing data.

Directions for the extension of this work include the analysis of other problems that arise in the study of genomic datasets using our framework, the analysis of more complicated and realistic models of mutations in cancer pathways (e.g., including multiple pathways with mutual exclusive mutations, and copy number aberrations), and the employment of more advanced statistical learning techniques (e.g., Rademacher averages; see Koltchinskii, 2001) to study the sample complexity of finding cancer pathways.

ACKNOWLEDGMENTS

This work is supported by NSF grant IIS-1247581 and NIH grant R01-CA180776 and, in part, by the University of Padova under project CPDA121378/12.

AUTHOR DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Abu-Mostafa, Y.S., Magdon-Ismail, M., and Lin, H.-T. 2012. *Learning From Data.* AMLBook, Singapore.

Bousquet, O., Boucheron, S., and Lugosi, G. 2003. Introduction to statistical learning theory, 169–207. *In* Bousquet, O., von Luxburg, U., and Rätsch, G., eds. *Advanced Lectures on Machine Learning*, Volume 3176 of Lecture Notes in Computer Science. Springer, New York.

Cancer Genome Atlas Network. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature*. 487, 330–337.

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

Cancer Genome Atlas Research Network, Weinstein, J.N., Collisson, E.A., et al. 2013. The cancer genome atlas pan-cancer analysis project. *Nat. Genet.* 45, 1113–1120.

Ciriello, G., Cerami, E., Sander, C., et al. 2012. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res*. 22, 398–406.

Ciriello, G., Miller, M.L., Aksoy, B.A., et al. 2013. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.* 45, 1127–1133.

Cohen, Y., Xing, M., Mambo, E., et al. 2003. Braf mutation in papillary thyroid carcinoma. *J. Natl. Cancer Inst.* 95, 625–627.

Dees, N.D., Zhang, Q., Kandoth, C., et al. 2012. Music: Identifying mutational significance in cancer genomes. *Genome Res*. 22, 1589–1598.

Ein-Dor, L., Zuk, O., and Domany, E. 2006. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl. Acad. Sci. USA* 103, 5923–5928.

Garraway, L.A., and Lander, E.S. 2013. Lessons from the cancer genome. *Cell*. 153, 17–37.

International Cancer Genome Consortium, Hudson, T.J., Anderson, W., et al. 2010. International network of cancer genome projects. *Nature*. 464, 993–998.

Kandoth, C., McLellan, M.D., Vandin, F., et al. 2013. Mutational landscape and significance across 12 major cancer types. *Nature*. 502, 333–339.

Kimura, E.T., Nikiforova, M.N., Zhu, Z., et al. 2003. High prevalence of braf mutations in thyroid cancer: Genetic evidence for constitutive activation of the ret/ptc-ras-braf signaling pathway in papillary thyroid carcinoma. *Cancer Res*. 63, 1454–1457.

Koltchinskii, V. 2001. Rademacher penalties and structural risk minimization. *IEEE Trans. Inf. Theory.* 47, 1902–1914.

Lawrence, M.S., Stojanov, P., Polak, P., et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*. 499, 214–218.

Leiserson, M.D.M., Blokh, D., Sharan, R., et al. 2013. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9, e1003054.

Martin, M., Maßhöfer, L., Temming, P., et al. 2013. Exome sequencing identifies recurrent somatic mutations in eif1ax and sf3b1 in uveal melanoma with disomy 3. *Nat. Genet.* 45, 933–936.

Miller, C.A., Settle, S.H., Sulman, E.P., et al. 2011. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics*. 4, 34.

Mitzenmacher, M., and Upfal, E. 2005. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, UK.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. 2012. *Foundations of Machine Learning*. MIT Press, New York.

Perkins, T.J., and Hallett, M.T. 2010. A trade-off between sample complexity and computational complexity in learning boolean networks from time-series data. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7, 118–125.

Raphael, B.J., Dobson, J.R., Oesper, L., et al. 2014. Identifying driver mutations in sequenced cancer genomes: Computational approaches to enable precision medicine. *Genome Med*. 6, 5.

Shrestha, R., Hodzic, E., Yeung, J., et al. 2014. Hit'ndrive: Multi-driver gene prioritization based on hitting time, 293–306. *In Research in Computational Molecular Biology*. Springer, New York.

Szczurek, E., and Beerenwinkel, N. 2014. Modeling mutual exclusivity of cancer mutations. *PLoS Comput. Biol*. 10, e1003503.

Valiant, L.G. 1984. A theory of the learnable. *Commun. ACM* 27, 1134–1142.

Vandin, F., Clay, P., Upfal, E., et al. 2012a. Discovery of mutated subnetworks associated with clinical data in cancer. *Pac. Symp. Biocomput.* 2012, 55–66.

Vandin, F., Upfal, E., and Raphael, B.J. 2011. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol*. 18, 507–522.

Vandin, F., Upfal, E., and Raphael, B.J. 2012b. *De novo* discovery of mutated driver pathways in cancer. *Genome Res*. 22, 375–385.

Vandin, F., Upfal, E., and Raphael, B.J. 2012c. Finding driver pathways in cancer: Models and algorithms. *Algorithms Mol. Biol*. 7, 23.

Vapnik, V.N., and Chervonenkis, A.Y. 1971. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probability Its Appl.* 16, 264–280.

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

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., et al. 2013. Cancer genome landscapes. *Science*. 339, 1546–1558.

Whitley, E., and Ball, J. 2002. Statistics review 4: Sample size calculations. *Crit. Care*. 6, 335–341.

Yeang, C.-H., McCormick, F., and Levine, A. 2008. Combinatorial patterns of somatic gene mutations in cancer. *FASEB J*. 22, 2605–2622.

Zhao, J., Zhang, S., Wu, L.-Y., et al. 2012. Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*. 28, 2940–2947.

Address correspondence to:
Dr. Fabio Vandin
Department of Information Engineering
University of Padova
Via Gradenigo 6/B
I-35131 Padova, Italy

E-mail: vandinfa@imada.sdu.dk