JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 22, Number 6, 2015

Pp. 510–527
DOI: 10.1089/cmb.2014.0161

Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data

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

ABSTRACT

Recent cancer sequencing studies provide a wealth of somatic mutation data from a large number of patients. One of the most intriguing and challenging questions arising from this data is to determine whether the temporal order of somatic mutations in a cancer follows any common progression. Since we usually obtain only one sample from a patient, such inferences are commonly made from cross-sectional data from different patients. This analysis is complicated by the extensive variation in the somatic mutations across different patients, variation that is reduced by examining combinations of mutations in various pathways. Thus far, methods to reconstruct tumor progression at the pathway level have restricted attention to known, *a priori* defined pathways.

In this work we show how to *simultaneously* infer pathways and the temporal order of their mutations from cross-sectional data, leveraging on the exclusivity property of driver mutations within a pathway. We define the pathway linear progression model, and derive a combinatorial formulation for the problem of finding the optimal model from mutation data. We show that with enough samples the optimal solution to this problem uniquely identifies the correct model with high probability even when errors are present in the mutation data. We then formulate the problem as an integer linear program (ILP), which allows the analysis of datasets from recent studies with large numbers of samples. We use our algorithm to analyze somatic mutation data from three cancer studies, including two studies from The Cancer Genome Atlas (TCGA) on large number of samples on colorectal cancer and glioblastoma. The models reconstructed with our method capture most of the current knowledge of the progression of somatic mutations in these cancer types, while also providing new insights on the tumor progression at the pathway level.

1. INTRODUCTION

CANCER IS A DISEASE CAUSED BY THE ACCUMULATION of somatic mutations, changes in the genome that appear during the lifetime of an individual. High-throughput DNA sequencing technologies are now measuring these mutations in thousands of cancer genomes through projects such as The Cancer Genome Atlas (TCGA) (Cancer Genome Atlas Research Network, 2008), International Cancer Genome Consortium

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

(ICGC) (Zhang et al., 2011), and many others. In the analysis of somatic mutations in cancer, two important questions arise. First, which mutations are the *driver* mutations responsible for cancer and which are merely random, *passenger* mutations? Second, is there any temporal order to the driver mutations in a single cancer patient? While the first question might be addressed in part by comparing the observed frequencies of mutations across different individuals (Dees et al., 2012; Lawrence et al., 2013), the second question is much more difficult to address from *cross-sectional* data, sequencing data taken from single time-points across different individuals. Answering the question about temporal progression, and more specifically determining what mutations occur early in the progression of cancer, is essential for both a basic understanding of cancer biology and for developing targeted treatments. The ideal dataset to determine temporal progression is a *longitudinal* dataset consisting of measurements of somatic mutations from multiple time-points in a single individual. However, such datasets are nearly impossible to obtain from human tumors: It is difficult to obtain multiple samples from the same patient without also perturbing the tumor with surgery, chemotherapy, or other treatments.

A number of methods for inferring temporal progression of mutations from cross-sectional data have been introduced (Desper et al., 2000, 1999; Beerenwinkel et al., 2005a,b; Rahnenführer et al., 2005; Tofigh et al., 2011; Hjelm et al., 2006; Gerstung et al., 2009; Beerenwinkel and Sullivant, 2009; Beerenwinkel et al., 2006, 2007; Gerstung et al., 2011; Sakoparnig and Beerenwinkel, 2012; Shahrabi Farahani and Lagergren, 2013) (see section 1.1). These methods consider models of increasing complexity for cancer progression: trees, mixtures of trees, and Bayesian network models with different constraints. However, such approaches infer progression at the level of individual mutations, or individual genes. The difficulty with this approach is that cancers exhibit extensive mutational heterogeneity: the somatic mutations, including driver mutations, vary widely across individuals with the same cancer. Thus, the main signal used to infer temporal order, co-occurrence of mutations in different samples, is very weak. A major reason for this mutational heterogeneity is that somatic mutations perturb various signaling, regulatory, and metabolic pathways (Vogelstein et al., 2013). Thus, different individuals may harbor driver mutations in different genes within the same pathway. Since driver mutations target pathways, it is possible that the order in which mutations arises is at the pathway level, not at the gene level.

There has been some initial work in inferring pathway order (Gerstung et al., 2011; Cheng et al., 2012). These approaches demonstrated some advantages over gene-based approaches, but restricted attention to known, annotated pathways. Most annotated pathways are large and overlap with other pathways [e.g., KEGG (Kanehisa and Goto, 2000) contains two pathways with more than 200 genes each and 75 genes in common], thus creating problems for the discovery of mutation progression in smaller sets of interacting genes.

An alternative to known pathways is to examine sets of genes or mutations *de novo*. However, the large number of such combinations will quickly overwhelm such an exhaustive approach. Recently, it has been observed that driver mutations in pathways tend to be mutually exclusive meaning that an individual rarely has more than one driver mutation in a pathway (Yeang et al., 2008), and this observation has been used to successfully identify pathways in cancer datasets (Miller et al., 2011; Vandin et al., 2012a; Ciriello et al., 2012; Leiserson et al., 2013). Mutual exclusivity is a powerful signal to constrain the combinations of genes and mutations to examine. In this article, we design an algorithm to infer *simultaneously* pathways and their temporal order from cross-sectional data, leveraging on the expected exclusivity of mutations within pathways (and on the co-occurrence of mutations in pathways at a different stage of the temporal order). We apply this algorithm to simulated and real sequencing data from colorectal and glioblastoma cancers. The progression models produced by our method are in agreement with the current knowledge of the progression of mutations in these cancer types, and also propose some novel hypothesis. The sets identified by our method mostly correspond to known pathways or sets of interacting genes, showing the ability of our approach to simultaneously identify cancer pathways and the tumor progression they define.

1.1. Previous work

After the seminal work of Fearon and Vogelstein (1990), which proposed a model for progression of mutations in colorectal cancer, a number of computational methods have been designed to reconstruct the progression of genetic events leading to cancer from cross-sectional data, assuming that the order is at the gene level. These methods consider models of increasing complexity. The model of Fearon and Vogelstein

(1990) describes a *linear sequence* (or *path*) on genes. Desper et al. (1999) and Desper et al. (2000) considered *trees* on the genes. A number of works (Beerenwinkel et al., 2005a,b; Rahnenführer et al., 2005; Tofigh et al., 2011) have proposed the inference of *mixture of trees* on the genes to model the progression of cancer. While providing a first advance in understanding cancer progression, these methods assume that cancer progresses through disjoint paths, with no possible convergence of different paths, which is a stringent constraint on the model.

More general methods that include convergence describe the model in terms of probabilistic directed acyclic graphs, or Bayesian networks (Hjelm et al., 2006; Gerstung et al., 2009; Beerenwinkel and Sullivant, 2009; Beerenwinkel et al., 2006, 2007; Gerstung et al., 2011; Sakoparnig and Beerenwinkel, 2012; Shahrabi Farahani and Lagergren, 2013). These methods impose different restrictions on the model to limit the search space and to represent possible features of cancer progression. In practice these methods can include at most a dozen genes in their analysis; the exception is the method presented in Shahrabi Farahani and Lagergren (2013), which uses a mixed integer linear program (MILP) to infer the best (constrained) Bayesian network. However, they consider only models (networks) in which the number of parents of a node is bounded by a small value k [in Shahrabi Farahani and Lagergren (2013) values of $k \leq 4$ are used].

We note that the model we are interested in could be defined as a Bayesian network in which the genes in a pathway are the parents of all the genes in the next pathway in the progression, and the probability model should reflect the exclusivity among mutations in the parents of a particular node. None of the methods above has considered the exclusivity among genes (or mutations) in their model, and to the best of our knowledge no such model has been proposed in the machine learning literature.

Two recent works (Gerstung et al., 2011; Cheng et al., 2012) have considered the inference of the progression model at the pathway level, where the assignment of genes into pathways was defined *a priori* and provided in input to the method. As pointed out in the introduction, the *a priori* assignment of genes into pathways is complicated by the fact that such pathways are large and moreover often a gene is assigned to multiple pathways, which limit the ability to detect smaller sets of interacting genes related to tumor progression.

1.2. Contributions

This work initiates the study of the simultaneous identification of cancer pathways and their mutation order in tumor progression from cross-sectional data, without relying on the *a priori* definition of pathways, and contains the following contributions. First, we formalize the pathway linear progression model for tumor progression, in which mutations within each pathway are mutually exclusive, while they satisfy a linear progression across pathways. We show that the computational problem, which we call pathway linear progression reconstruction problem, of identifying the model that provides the best (in terms of number of errors) explanation of the observed data is NP-hard when a (minor) constraint on the solution is required. Moreover, we prove that under reasonable assumptions and with enough samples, the correct progression model is uniquely identified by the optimal solution of the pathway linear progression reconstruction problem, even when the data contains errors.

Second, we formulate the pathway linear progression reconstruction problem as an integer linear program (ILP), providing an exact solution for datasets of realistic size. Using simulated data we show that the correct progression model is identified by our algorithm under realistic assumptions for the error model when enough samples are considered. We also show that when genes that are not correlated with tumor progression are included in the analysis, our algorithm identifies the correct order among genes in pathways driving the progression.

Third, we run our algorithm on somatic mutation data from The Cancer Genome Atlas (TCGA) studies on colorectal and glioblastoma cancers, and on a different colorectal cancer study. We show that the progression models produced by our method recapitulate most of the current knowledge of the progression of mutations in colorectal cancer, and also propose novel hypothesis for the progression driving colorectal and glioblastoma cancer. In particular, on somatic mutation data from 224 TCGA samples of colorectal cancer, our algorithm identifies models that are in agreement with current knowledge of the progression of mutations in this cancer type. Moreover, our method groups members of the Raf-Ras pathways and SMADs and interacting genes in different sets. On somatic mutation data from 251 glioblastoma multiforme samples, our method defines a model with gene sets corresponding to part of the Rb1, PI3K, and p53 pathway.

2. METHODS AND ALGORITHMS

2.1. Model and problem definition

We are given mutation data from m samples $s_1, s_2, \ldots, s_m$, consisting of the mutation status of each of n genes $g_1, g_2, \ldots, g_n$. This data is represented by an $m \times n$ binary mutation matrix M, with samples on the rows and genes on the columns, where $M_{i,j} = 1$ if g_j is mutated in sample s_i, and $M_{i,j} = 0$ otherwise. We use r_i to denote the ith row of M, and c_j to denote the jth column of M. We now define a model in which the mutation data comes from a linear progression on sets of genes (see Fig. 1a).

Pathway linear progression model (PLPM). An $m \times n$ mutation matrix M satisfies the pathway linear progression model $PLPM(K)$ with parameter $K > 1$ if there exists a partition $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$ of the columns $\{c_1, \ldots, c_n\}$ of M into K sets such that:

1. For each row r_i of M, 1's within each set P_k are *mutually exclusive*, that is: for all $1 \le k \le K$ we have $|\{c_j \in P_k : M_{i,j} = 1\}| \le 1$;
2. Each row r_i of M satisfies the *progression* on the sets $P_1, \ldots, P_K$, that is: for all $1 < k \le K$, if $|\{c_j \in P_k : M_{i,j} = 1\}| > 0$ then $|\{c_j \in P_{k-1} : M_{i,j} = 1\}| > 0$.

Given m and $\mathcal{P}$, let $PLPM(\mathcal{P})$ be the set of $m \times n$ mutation matrices M for which the constraints 1 and 2 are satisfied. We denote a mutation matrix $M \in PLPM(\mathcal{P})$ by saying that M *satisfies* the PLPM defined by $\mathcal{P}$. Therefore a mutation matrix M satisfies the pathway linear progression model $PLPM(K)$ if there exists a partition $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$ such that $M \in PLPM(\mathcal{P})$. Each set P_k in a partition $\mathcal{P}$ defines a set of genes, or *pathway*, that by their interaction perform a certain function or process in the cell. When at least one of the columns of P_k has value 1 in r_i, we say that P_k is *mutated* in r_i.

Due to a number of factors (passenger mutations in driver genes, false positives and false negatives in mutations detection, etc.), a mutation matrix M is a noisy observation from the PLPM, and M may therefore not satisfy $PLPM(K)$, or equivalently there is no partition $\mathcal{P}$ for which $M \in PLPM(\mathcal{P})$. For a given K, we are thus interested in finding a partition $\mathcal{P} = \{P_1, P_2, \ldots, P_K\}$ such that M is *close* to satisfy the pathway linear progression model having $\mathcal{P}$ as a partition. In particular, we look for a partition $\mathcal{P}^*$ that minimizes the number of entries of M that must be changed (*flips* of $0 \to 1$ or $1 \to 0$) so that the resulting mutation matrix $M' \in PLPM(\mathcal{P}^*)$. More formally, given two binary $m \times n$ matrices M, M', let $d(M, M') = \sum_{i=1}^{m} \sum_{j=1}^{n} |M_{i,j} - M'_{i,j}|$. For a mutation matrix M and a partition $\mathcal{P}$, we define $f(M, \mathcal{P}) = \min_{M' \in PLPM(\mathcal{P})} d(M, M')$. Let $\mathcal{P}(K)$ be the set of all possible partitions (of the columns of M) into K sets.

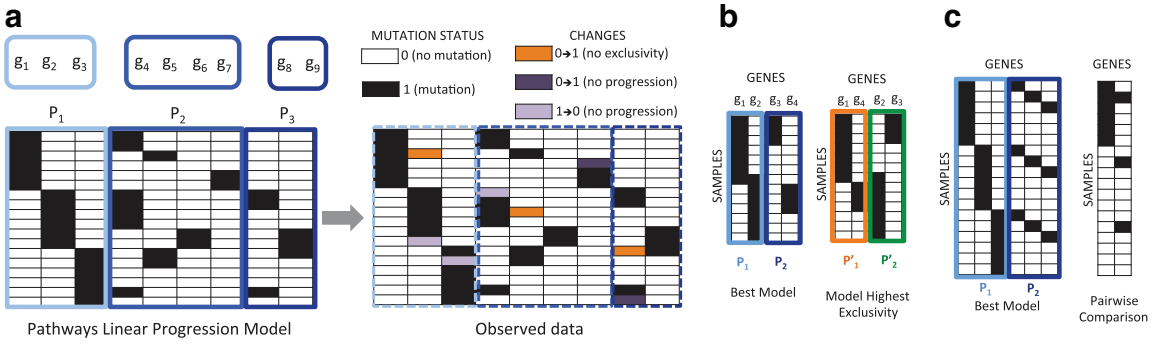

FIG. 1. The pathway linear progression model (PLPM). **(a)** A linear progression on gene sets (pathways) generates a mutation matrix with exclusive mutation within each set, and a progression of mutations across the sets. In real data errors that disrupt the exclusivity or the progression are present. **(b** and **c)** Problems of considering only exclusivity or only progression in reconstructing a PLPM. **(b)** *Left*: the correct progression model. *Right*: the (incorrect) partition that is inferred by maximizing the (total) exclusivity of sets. Since the correct model does not show perfect exclusivity (due to errors, etc., present in real data), maximizing the exclusivity does not lead to recover the correct model. **(c)** *Left*: the correct progression model. *Right*: genes pairwise comparison reveals no information about the progression. Progression at the genes level needs to appear as significant co-occurrence between 1's in two columns, while in the example for each pair of columns in $P_1 \times P_2$ the number of samples in which they are both 1 is exactly the expected number under the independence of the two columns. An arbitrary partition, most probably not correct, would then be reported by considering only the progression signal among genes.

Pathway linear progression reconstruction problem. Given an $m \times n$ mutation matrix M and an integer value $K > 1$, find $\mathcal{P}^* = \arg\min_{P \in \mathcal{P}(K)} f(M, \mathcal{P})$.

In our formulation, there are two requirements that a partition has to satisfy: the exclusivity of mutations within each set of the partition, and the progression across the sets (revealed by the co-occurrence of mutations in different sets). One may think that considering only one of the two requirements for the optimization is enough. The examples of Figure 1b and c show that this is not true.

We therefore need to identify the best partition $\mathcal{P}^*$ by simultaneously considering both exclusivity and progression. We have the following result. (All proofs are in the Appendix.)

Theorem 2.1. *The Pathway Linear Progression Reconstruction problem with the requirement that a given column $\bar{c}$ is in a given set of the solution is NP-hard for any value of K.*

2.2. Conditions for reconstruction with errors

Real mutation data has various sources of error that result in both false positive and false negative mutations. Thus, rather than observing a mutation matrix M satisfying $PLPM(K)$, we observe a perturbed mutation matrix $\tilde{M}$. A natural question is to determine conditions under which the partition $\mathcal{P}$ that defines M can be recovered as a solution of the pathway linear progression reconstruction problem when either M or $\tilde{M}$ is given as input. In this section, we prove that if the number m of samples is large enough, then with bounded error probability $\mathcal{P}^*$ is the unique solution to pathway linear progression reconstruction under two different models for generating M and $\tilde{M}$, respectively.

First, given a partition $\mathcal{P}$ of K sets $P_1, P_2, \ldots, P_K$, we define a *uniform* generative model $UPLPM(\mathcal{P})$ for an $m \times n$ mutation matrix M as follows. For each row r_i, $i = 1, \ldots, m$, we select the positions for the 1's by: i) choosing the *stage in the progression* of r_i, which is a value $t \in \{1, 2, \ldots, K\}$, uniformly at random; and ii) for each j, $1 \leq j \leq t$, choosing one of the columns in P_j uniformly at random to be 1 in r_i. The following relates the number m of rows of an $m \times n$ mutation matrix M from $UPLPM(\mathcal{P})$ to K and n (parameters of $\mathcal{P}$), and to the probability δ of not identifying $\mathcal{P}$ by finding the optimal solution to the pathway linear progression reconstruction problem.

Theorem 2.2. *Let M be an $m \times n$ mutation matrix generated from $UPLPM(\mathcal{P})$. If $m \geq Kn^2 \ln \frac{2n^2}{\delta}$, then $\mathcal{P}$ is the* unique *optimal solution to the pathway linear progression reconstruction problem with probability $\geq 1 - \delta$.*

Next, we consider the case of a mutation matrix $\tilde{M}$ that is a perturbation of an $m \times n$ mutation matrix M generated from $UPLPM(\mathcal{P})$. We generate such an $\tilde{M}$ as follows. We assume that for each $P \in \mathcal{P} : |P| = \frac{n}{K}$, and in each row one entry chosen uniformly at random has been flipped with probability q, independently for each row. We call the set of such matrices $UPLPM(\mathcal{P}, q)$). We prove the following.

Theorem 2.3. *Let $q = \frac{K^2 n(K - \varepsilon n^2)}{n^3(K-1)^2 + n^3 K - 2Kn^2 + 2K^3} \geq 0$ for some $\varepsilon > 0$, and let $\tilde{M}$ be an $m \times n$ mutation matrix from $UPLPM(\mathcal{P}, q)$. If $m \geq \frac{8}{\varepsilon^2} \ln \frac{2n^2}{\delta}$, then $\mathcal{P}$ is the unique optimal solution to the pathway linear progression reconstruction problem with probability $\geq 1 - \delta$.*

2.3. Algorithm

We now formulate the pathway linear progression reconstruction problem as an integer linear program (ILP). For a partition $\mathcal{P}$, let $p_{j,k}$ be a 0-1 variable with $p_{j,k} = 1$ if column c_j is assigned to set P_k, and $p_{j,k} = 0$ otherwise. Let $a_{i,k}$ be a 0-1 variable with $a_{i,k} = 1$ if the set P_k is considered mutated (after required flips are made) in row r_i, and $a_{i,k} = 0$ otherwise. We also define auxiliary 0-1 variables $f_{i,k}$ for $1 \leq i \leq m$ and $1 \leq k \leq K$, where intuitively $f_{i,k} = 1$ if we need to flip one of the entries of columns in P_k for P_k to be mutated in row r_i, and 0 otherwise. A valid solution to our problem then satisfies the following constraints:

- each column is assigned to exactly one set: for $1 \leq j \leq n$, $\sum_{k=1}^{K} p_{j,k} = 1$;
- for each set P_k, at least one column is assigned to it: for $1 \leq k \leq K$, $\sum_{j=1}^{n} p_{j,k} = 1$;
- for each sample the progression model is satisfied: for $1 \leq i \leq m$ and $1 \leq k \leq K-1$, $a_{i,k} \geq a_{i,k+1}$;
- for each row r_i, the set P_k is considered mutated if it has a 1 in r_i or if one of its entries in row r_i is flipped to make it mutated (i.e., $f_{i,k} = 1$): for $1 \leq k \leq K$ and $1 \leq i \leq m$, $\sum_{j=1}^{n} M_{i,j} p_{j,k} + f_{i,k} \geq a_{i,k}$.

For a particular partition $\mathcal{P}$, the value of the objective function is the minimum number of entries of M that we need to flip to satisfy the constraints defined by $\mathcal{P}$. If we consider a given sample r_i and a given set P_k, after variables $p_{j,k}$, $a_{i,k}$, and $f_{i,k}$ have been fixed, the number of entries of P_k that are flipped in r_i is given by $\sum_{j=1}^{n} M_{i,j}p_{j,k} - a_{i,k} + 2f_{i,k}$. Since we want to minimize the total number of entries that are flipped, the objective function is:

$$\min \sum_{i=1}^{m}\sum_{k=1}^{K}\left(\sum_{j=1}^{n} M_{i,j}p_{j,k} - a_{i,k} + 2f_{i,k}\right).$$

The contribution of one row (i.e., one patient) to the objective function is interpreted as follows. The term $\sum_{j=1}^{n} M_{i,j}p_{j,k}$ counts the number of observed 1's in r_i for set P_k. Assume that $\sum_{j=1}^{n} M_{i,j}p_{j,k} > 0$: if we consider set P_k mutated in r_i (i.e., if $a_i = 1$), the number of entries of M that we need to flip to satisfy the progression model is $\sum_{j=1}^{n} M_{i,j}p_{j,k} - 1$, and it is $\sum_{j=1}^{n} M_{i,j}p_{j,k}$ otherwise (i.e., if $a_{i,k} = 0$). If instead $\sum_{j=1}^{n} M_{i,j}p_{j,k} = 0$ (i.e., set P_k has no mutations in row r_i), if we do not consider P_k mutated ($a_{i,k} = 0$) then the number of entries to be flipped is 0, while if we consider P_k mutated ($a_{i,k} = 1$) then the number of required flips is 1, obtained by having $f_{i,k} = 1$ as enforced by the last constraint above. Note that this reasoning assumes that $f_{i,k} = 0$ whenever $\sum_{j=1}^{n} M_{i,j}p_{j,k} > 0$, or $a_{i,k} = 0$, which is not forced by the constraints above but is obtained when the objective function is minimized.

The formulation above can be easily extended to consider the case in which not all columns (i.e., genes) are to be part of the progression model. This can be obtained by defining, for every column c_j, a 0 – 1 variable e_j that is 1 if c_j is excluded from the progression model, and 0 otherwise. In this case the first constraints above is modified so that each column c_j is either assigned to exectly one set, or the corresponding variable e_j is set to 1, that is: for $1 \leq j \leq n$, $\sum_{k=1}^{K} p_j, k + e_j = 1$. It is also necessary to introduce in the objective function a term that accounts for the mutations in genes excluded from the progression model. We obtain this by adding to the objective function the term $\sum_{j=1}^{n} W_j e_j$, where W_j is a weight assigned to columnn c_j. W_j can be used to incorporate additional information regarding the importance of column c_j and its likelihood to be in the progression model [for example, W_j can be set proportional to the expected number of driver mutations in the gene corresponding to c_j, obtained by subtracting the expected number of passenger mutations (Lawrence et al., 2013) in c_j to the observed number of mutations in c_j].

3. EXPERIMENTAL RESULTS

In this section we present the results of our experimental analysis on simulated data, and on data from cancer studies. In all cases, we solved the ILP using CPLEX v12.3 with default parameters (see Table 1 for running times).

3.1. Simulated data

We performed a number of experiments using simulated data to assess the robustness of our method to different levels of noise. We considered data coming according to a progression model $\mathcal{P}$ to which noise was added. In particular, we considered a progression model with $K = 5$ stages, each containing 5 genes, and generated 100 datasets with m samples from this model, adding noise by flipping each entry of the corresponding mutation matrix with probability p. Note that this error model is more complex and more realistic than the one we analyzed in Theorem 2.3. The progression stage for each sample was chosen uniformly at random (between 1 and 5), and for a sample the mutated gene in a stage is chosen uniformly at random. We considered values of m = 50, 100, 500, 1000, and p = 0.001, 0.01, 0.05, which are values in the expected range for passenger mutation probability given the background mutation rate and the length of the genes (Vandin et al., 2012a,b). (Note that when p = 0.05, the expected number of errors per sample is >1.)

For each combination m,p we recorded the fraction of times in which the optimal solution identified by solving the ILP (fed with the correct value for K) corresponded to $\mathcal{P}$ (i.e., all genes were reported in the correct stage). Results are shown in Figure 2a. When m = 50 samples are included in the analysis, most of the time the correct progression model is not identified, even when the error probability is not very high (p = 0.001). However, for $p \leq 0.01$, when 100 samples are analyzed the correct progression model is reported most of the times, and when 500 samples are analyzed the correct model is reported every time. In

Table 1. Running Time to Find the Optimal Solution of the ILP Using CPLEX v12.3 with Default Parameters

Dataset	K	n	p	m	*Runtime (s)*
Uniform	5	25	0.001	50	1.27
				100	1.86
				500	18.32
				1000	70.86
Uniform	5	25	0.01	50	1.64
				100	3.14
				500	347.95
				1000	42547.64
Uniform	4	16	0.001	50	0.52
				100	0.69
				500	3.36
				1000	7.34
Uniform	4	16	0.01	50	0.44
				100	0.72
				500	18.39
				1000	50.46
Increasing	4	20	0.01	50	0.58
				100	1.85
				500	14.66
				1000	60.83
Decreasing	4	20	0.01	50	0.41
				100	2.75
				500	37.92
				1000	100. 96
Extended	4	200	0.001	50	7.05
				100	16.54
				500	59.67
				1000	484.76
Extended	4	200	0.01	50	65.12
				100	2541.94
				500	870.90
				1000	494.63
Colorectal	4	8	—	95	13258.41
TCGA colorectal	5	14	—	224	86.16
TCGA colorectal (extended)	5	43	—	224	13763.49
TCGA GBM	6	27	—	251	2488.02

For each dataset, we report the number K of stages in the progression model, the number n of genes, and the number m of samples. For simulated data, we also report the noise level p. For the dataset, ''uniform'' refers to the model with equal number of genes in each stage, ''increasing'' to the model with increasing number of genes in different stages, ''decreasing'' to the model with decreasing number of genes in different stages, ''extended'' to the simulations in which we used the extended version of the ILP that does not assign all the genes to the progression model (see section 3.1 for details). ''Colorectal,'' ''TCGA colorectal,'' ''TCGA colorectal (extended),'' and ''TCGA GBM'' refer to cancer data (see section 3.2 for details).

contrast, when $p = 0.05$, with 500 samples the correct model is reported only 65% of the time, and it is reported 95% of the time when 1000 samples are considered. These results show that while data from reasonably sized cancer studies can sometimes be used to infer the correct progression model, studies of size larger than currently available may be required to identify the correct progression model if the noise level (i.e., p) is high.

To understand how the complexity (i.e., number of sets, number of genes in each set) of $\mathcal{P}$ impact the number of samples required to reliably identify the correct model, we considered a ''simpler'' model, consisting of $K = 4$ stages of progression, each including four genes (Fig. 2b). Mutations from this model were generated as for the model above, and we considered the same values for m and p. In this case, for a

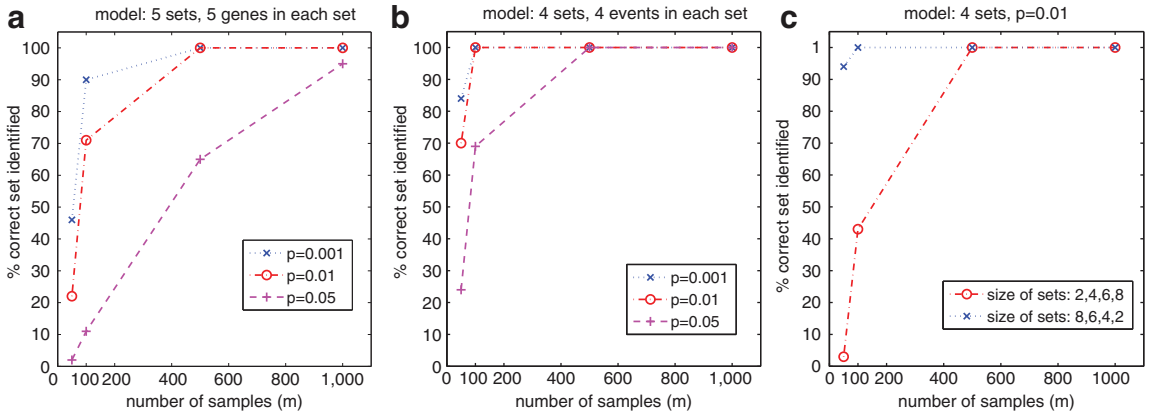

FIG. 2. Fraction of times (over 100 trials) the entire correct order is identified by the integer linear program (ILP) on m samples where the mutation matrix M comes from a progression model with K sets. Each entry of M is flipped with probability p; the results for different values of p are shown. **(a)** Results for $K=5$ and each stage containing 5 genes. **(b)** Results for $K=4$ and each stage containing 4 genes. **(c)** Results for $K=4$, $p=0.01$, and set sizes increasing (sizes: 2, 4, 6, and 8) or decreasing (sizes: 8, 6, 4, and 2) with progression stages or decreasing.

given pair (m, p), the fraction of times the correct model is reported is always greater or equal to the fraction of times the correct model with five sets and five genes in each set was identified for the same pair (m, p). For example, when $p \leq 0.01$ and $m=50$ samples are considered, the correct model is reported at least 70% of the time, while with 500 samples the correct model is reported every time even when $p=0.05$.

We also considered the case in which the different stages contained a different number of genes. In particular, for $p=0.01$, we considered a model with $K=4$ stages and $2i$ genes in stage i, and a model with $K=4$ stages and $8-2(i-1)$ genes in stage i (Fig. 2c). These results we obtained show that when later stages of the progression include more genes, a larger number of samples may be required to identify the correct progression model.

These results show that the number of samples required to identify the correct model is sensitive to the parameters of the progression model, confirming and extending the analytical results of section 2.2, and moreover show that currently available cancer studies have sufficient samples sizes to identify progression models where the number of sets and the number of genes in each set is not too high, while more samples may be required to correctly identify models with a large number of sets (stages) and a large number of genes, or when the probability of false positives and false negatives is very high.

We also used simulations to assess the impact of the inclusion of genes not related to the progression on the accuracy of our method. We considered the progression model with $K=$ five stages, each consisting of five genes related to the progression, described above, and also included mutations for 25 additional genes, not related to the progression, each mutated in 5% of the samples independently of all other events. We generated 100 datasets from this model for each of the values $m=50$, 100, and 500, fixing $p=0.001$. For $m=50$, the inferred model never corresponded to the correct model on the 25 genes related to progression; for $m=100$, the inferred model on the 25 genes related to the progression was reported 41% of the time, while for $m=500$ this happened 100% of the time. This shows that even when genes not associated with the progression model are included in the analysis, our method is able to correctly reconstruct the relationship between the genes associated with the progression when the number of samples is sufficiently high. Our analysis also shows that spurious associations are more likely to be observed in late stages of the inferred progression (data not shown).

We also used simulations to test the extension of our model to not include all genes in the progression model. To this end, we considered a progression model with $K=4$ stages, each containing 10 genes (40 genes in total in the model), and we also considered 160 *random* genes containing passenger mutations. For each *random* gene g, we assumed it was mutated in a patient, independently of every other event, with a passenger probability r_g chosen between 0.1 and 0.01. As before, we generated 100 datasets with m samples from this model, adding noise by flipping each entry of the corresponding mutation matrix with probability p. For the ILP, we set W_j by considering for each gene j the probability p_j that j contains passenger

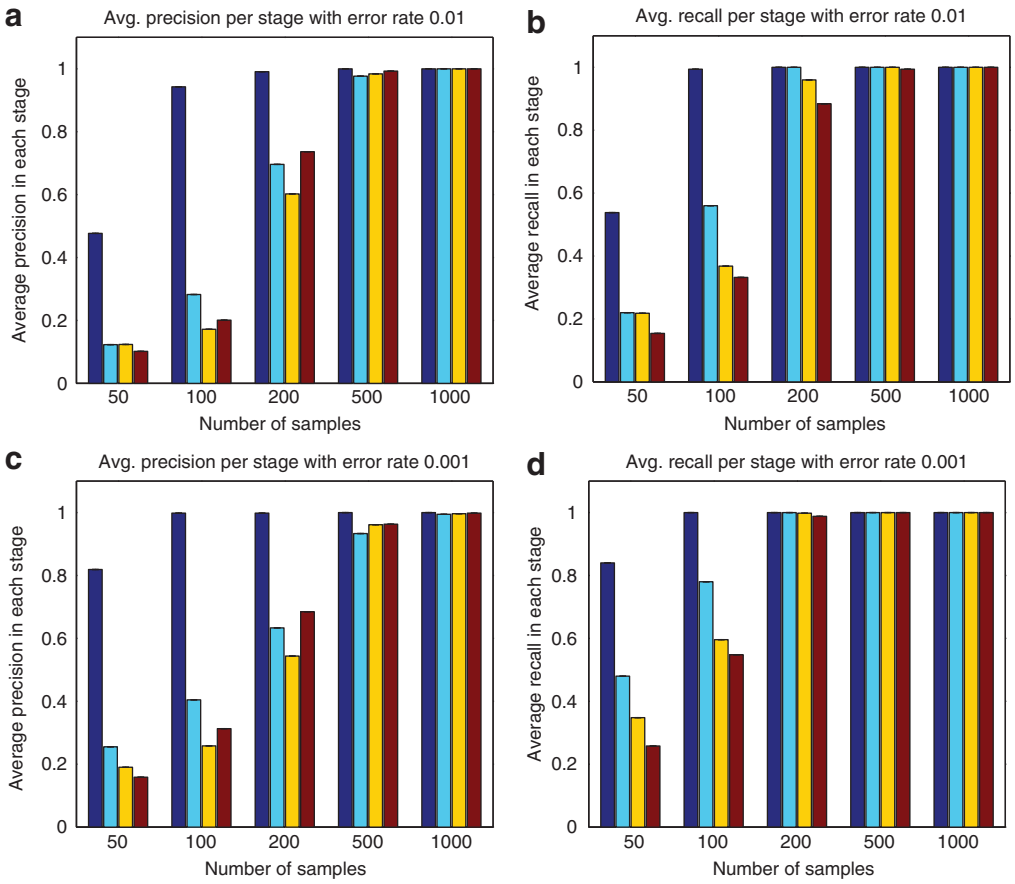

FIG. 3. Results for ILP where not all genes are assigned to the model, with $K=4$ stages of 10 genes each and 160 random genes. Results are averaged over 100 random datasets. Results for different stages are represented by different colors (*dark blue, blue, yellow,* and *red*, for stage 1, 2, 3, and 4, respectively). **(a)** Precision with noise $p=0.01$, for different number m of samples. **(b)** Recall with noise $p=0.01$, for different number m of samples. **(c)** Precision with noise $p=0.001$, for different number m of samples. **(d)** Recall with noise $p=0.001$, for different number m of samples.

mutations; that is, $p_j=p$ if the gene is in the progression model that generated the data, and $p_j=r_j$ otherwise; we then set $W_j=\max\{0,obs(j)-mp_j\}$, where $obs(j)$ is the observed number of samples with a mutation in gene j (and mp_j is the expected number of patients with a mutation in j). We considered values of m = 50, 100, 200, 500, 1000, and p = 0.001, 0.01. For each combination m,p and each stage i of the progression model ($1 \le i \le 4$) we recorded the average fraction of genes in the i-th stage of the progression model that are correctly identified, as well as the average fraction of random genes that are reported in stage i (Fig. 3). The results show that while with m = 200 samples almost all genes associated with the progression are correctly reported in the corresponding stage of the model, more samples are required to avoid the erroneous assignment of random genes to some stage of the progression.

3.2. *Cancer data*

We used our ILP to analyze somatic mutation data from published cancer studies. We first analyzed the dataset from a colorectal cancer study (Wood et al., 2007) considered in Gerstung et al. (2011). We then analyzed two large datasets from The Cancer Genome Atlas (TCGA) studies on colorectal cancer (Cancer Genome Atlas Network, 2012) and glioblastoma multiforme (Brennan et al., 2013).

For all these datasets we used the ILP to identify the set $\mathcal{P}_K^*$ of cardinality K of minimum weight for $K=2, \ldots, 8$ and then considered the best progression model to be the set $\mathcal{P}^*$ of minimum weight among the different solutions obtained: $\mathcal{P}^* = \arg\min_{K\in\{2,\ldots,8\}} f(\mathcal{P}_K^*)$; while in theory there could be multiple optimal solutions and multiple best progression models, in the instances below the optimal solution was

unique. To assess the statistical significance of our observation we computed a p-value using a permutation test, estimating the probability of obtaining a set of size K of weight less or equal to $\mathcal{P}^*$ when the mutations are placed independently in the samples preserving the mutation frequency of the genes. For each gene we also computed the fraction of times it is reported in a particular stage of the progression using *bootstrap* datasets (Efron and Tibshirani, 1994); this measures the stability to random fluctuations in the samples population of the assignment of a particular gene to a stage in the progression.

3.2.1. Colorectal cancer. We analyzed the mutations reported from the 95 samples considered in Wood et al. (2007), for the eight genes mutated with frequency above 5%: APC, EPHA3, EVC2, FBXW7, KRAS, PIK3CA, TCF7L2, and TP53. The PLPM of minimum weight is shown in Figure 4a. The progression model inferred with our method shares some similarities with the one inferred in Gerstung et al. (2011) (Fig. 4b), and is consistent with the proposed linear order of mutations in colorectal cancer: mutations in APC occur early in the progression, while KRAS mutations appear later.

Interestingly, in Gerstung et al. (2011) the order of TP53 mutations were reported as independent of other mutations, and mutations in PIK3CA were reported as independent of KRAS mutations, while in our model mutations in TP53 and PIK3CA are reported to appear after APC mutations, but before KRAS mutations. TP53 mutations have been reported to appear after APC in Attolini et al. (2010), while TP53 mutations are considered to appear after KRAS mutations (Fearon and Vogelstein, 1990; Fearon, 2011). TP53 mutations and PIK3CA mutations are significantly exclusive in this dataset ($p < 0.008$ by Fisher exact test), and are therefore potentially related. Of the 71 samples that contain a TP53 mutation or a PIK3CA mutation (or both - 1 sample), 51 also present a KRAS mutation, while only 8 samples with a KRAS mutation do not have a TP53 mutation or a PIK3CA mutation. Therefore, the most reasonable explanation, assuming a linear order among pathways, is that KRAS mutations come after TP53/PIK3CA mutations.

Since the model inferred in Gerstung et al. (2011) considers TP53 mutations as independent of the other mutations, we assessed how well the data is described by the two models when TP53 is ignored. In particular, we found that 12 samples have mutations that (ignoring TP53 and assuming no errors) do not conform to the PLPM model in Figure 4a, while 22 samples have mutations that (ignoring TP53 and assuming no errors) do not conform to the model of Gerstung et al. (2011). For example, none of the five samples with EPHA3 mutations come from the model of Gerstung et al. (2011), while only one such sample does not come from the PLPM model in Figure 4a. Therefore, our model provides a better explanation of the colorectal cancer data from Wood et al. (2007).

3.2.2. TCGA colorectal cancer. We analyzed 224 colorectal samples from the TCGA study on this cancer type. We download mutation data from the Broad GDAC Firehose, including single nucleotide variants and indels. We restricted our analysis to the 14 genes identified as recurrently mutated by MutSigCV (Lawrence et al., 2013).

The progression model inferred by our method is shown in Figure 5a. Interestingly, the progression model restricted to the genes APC, TP53, PIK3CA, and KRAS is the same we identify from the smaller dataset of Wood et al. (2007). Moreover, the bootstrap analysis reveals that these genes and NRAS have the most stable assignments to the different stages of the progression.

As noted before, TP53 mutations are usually reported as appearing after KRAS mutations. However, even considering only (TP53, PIK3CA) in the second stage of the progression model, and considering

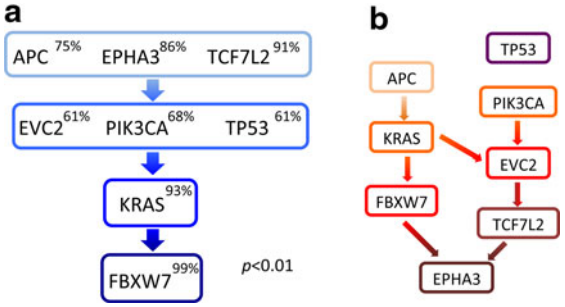

FIG. 4. Progression models for colorectal data (Wood *et al.*, 2007). **(a)** PLPM inferred using our method. The p-value from the permutation test is reported. For each gene, the fraction of times it appears in the same stage out of 100 bootstrap dataset is shown. **(b)** Model from Gerstung *et al.* (2011). In Gerstung *et al.* (2011) all parents of a node must be mutated for a gene to be mutated. TP53 is independent of other genes.

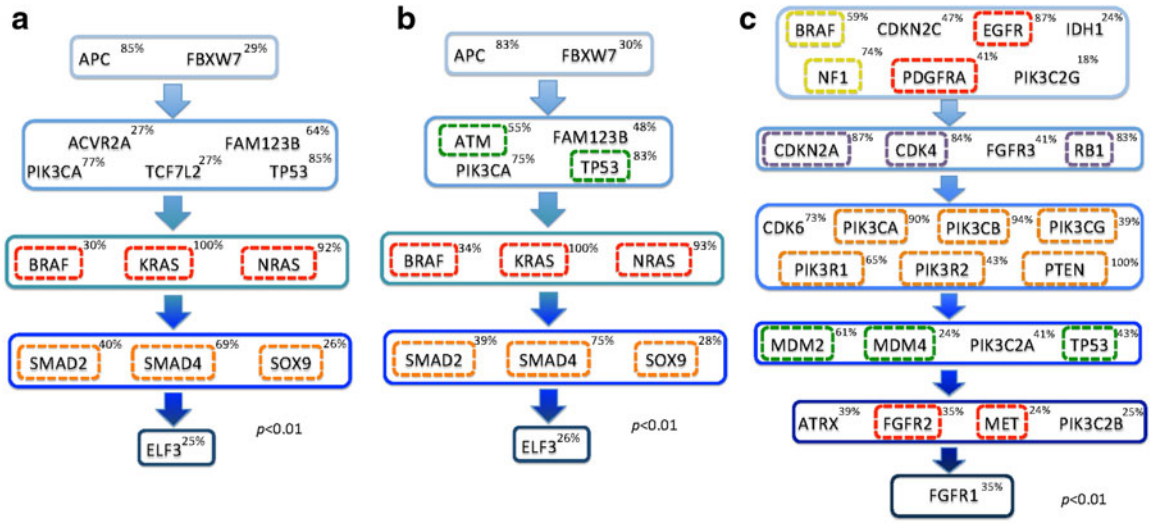

FIG. 5. PLPM models from TCGA cancer datasets. **(a)** PLPM for TCGA colorectal data (Cancer Genome Atlas Network, 2012) using the 14 genes recurrently mutated by MutSisCV (Lawrence *et al.*, 2013). Dashed boxes identify genes in the same pathway, with different colors for different pathways. **(b)** PLPM for TCGA colorectal data using 43 genes and the extension of our method that does not include all the genes in the model. Dashed boxes identify genes in the same pathway, with different colors for different pathways. **(c)** PLPM for TCGA glioblastoma multiforme data (Brennan *et al.*, 2013). Dashed boxes identify genes that are in a set with at least another gene in the same pathway (as annotated in Brennan *et al.,* 2013), with different colors for different pathways. For each PLPM, the *p*-value from the permutation test is reported, and the fraction of times genes appear in the same stage out of 100 bootstrap dataset is shown.

(BRAF, NRAS, KRAS) in the third stage, we have that 58 samples contain a mutation in the set (TP53, PIK3CA) and not in (BRAF, NRAS, KRAS), while 48 samples contain a mutation in (BRAF, NRAS, KRAS) and not in (TP53, PIK3CA); therefore, it is more reasonable to assume that mutations in (TP53, PIK3CA) (that show again significant exclusivity of mutations - $p < 0.0032$ by Fisher exact test) appear before mutations in (BRAF, NRAS, KRAS). Moreover, a recent analysis (Kandoth et al., 2013) suggested that in three other cancer types mutations in TP53 appear early during tumorigenesis, while KRAS mutations appear later in the tumor development.

Two sets in our model contain genes all in the same pathway or interacting. In particular, (BRAF, KRAS, NRAS) is part of the Ras-Raf pathway, and SOX9 interacts with SMAD2, which interacts with SMAD4. For both these sets, the probability that these genes are assigned to the same set in the partition under a random assignment is < 0.05. This shows that our method identifies sets that correspond to pathways or sets of interacting genes without any *a priori* information about the interactions among genes and their assignment to pathways.

For this dataset we also have robust passenger probability estimates for each gene provided by MutSigCV, which can be used to derive a weight for the version of our method that does not force every gene to be part of the model. Therefore, even if according to the results on simulated data the number of samples in this study may be too small, we extended the set of genes in the analysis to consider the 14 recurrently mutated genes as well as all genes mutated in > 21 patients, for a total of 43 genes considered in the analysis, and for gene j we set $W_j = \max\{0, obs(j) - mp_j\}$, where $obs(j)$ is the observed number of samples with a mutation in gene j (and mp_j is the expected number of patients with a mutation in j). The progression model inferred by our method is shown in Figure 5b. In this case the number of stages in the model is the same as in the model inferred by using only the 14 recurrently mutated genes, and for the most part the two models are equal. The only difference is in the second stage, where ACVR2A and TCF7L2 (that had the smallest bootstrap frequency among genes in the second stage) are now not reported in the model, while ATM (not among the 14 recurrently mutated genes by MutSigCV) is assigned to the second stage. Interestingly ATM is known to interact with TP53. These results show that even when genes (probably) not associated with the disease are considered in the analysis, our method identifies a meaningful progression model on sets of interacting genes without using information on the interactions among genes.

3.2.3. TCGA glioblastoma multiforme. We analyzed 251 samples from a recent TCGA study on this cancer type (Brennan et al., 2013). We restricted our analysis to the 27 genes reported in Brennan et al. (2013) as part of the landscape of pathway alterations in GBM, mostly obtained from manual curation.

For each gene, we considered single nucleotide variants, indels, and copy number aberrations (CNAs) consistent with the report in Brennan et al. (2013) for these genes. For CNAs we only considered the type (amplification or deletion) that appears in the majority of samples (discarding CNAs in case the two types appear in the same number of samples). The progression model inferred by our method is shown in Figure 5b. In four of the six sets in the progression model inferred by our method, at least 50% of the genes are part of the same pathway [as annotated in Brennan et al. (2013)], and each set has such genes coming from a different pathway; for three of these sets, all but one gene are part of the same pathway. In particular, the second set in the progression contains mostly genes from the Rb1 pathway, the third set contains mostly genes in the PI3K pathway, and the fourth set in the progression contains mostly genes in the p53 pathway. Moreover, five of the six sets in the model (i.e., all sets with at least two genes) contain at a least a pair of genes that are in the same pathway. The bootstrap analysis reveals that on average the assignment (in terms of progression stage) of genes that are in a set with at least another gene in the same pathway is more stable than the assignment of the other genes. For the first four sets in the progression model, this is true also considering only the genes in the specific set. This shows that the model reported by our method identifies pathway relations among genes in the different stages.

4. CONCLUSIONS

In this article, we study the problem of the simultaneous identification of cancer pathways and the tumor progression from cross-sectional mutation data. We formally define a model in which mutations within each pathway are exclusive, while they satisfy a linear progression at the pathway level. We prove that the problem of reconstructing the best model is NP-hard when a (minor) constraint on the solution is required, and provide an ILP formulation to solve the problem for reasonably sized datasets. Moreover, we show analytically and with synthetic data that under reasonable assumptions on the progression model and on the errors occurring in real data the optimal solution provided by our method captures the correct progression model when enough samples are considered.

We analyze somatic mutations data from three cancer studies and show that most of the current knowledge of the mutation progression in these studies is captured by the models produced by our method. Most of the sets in the models obtained from these datasets correspond to interacting genes or part of known pathways, showing the ability of our method to correctly infer cancer pathways while inferring the progression of genetic events leading to cancer.

There are many directions in which our work can be extended. In certain cases more information about the probability of false positives and false negatives is known, even for each single gene. Our ILP formulation can easily incorporate such information whenever available.

Moreover, while our current formulation assumes that the mutated pathway are shared by all patients with the same tumor type, this may only be partially true, and models that assume different pathway progression models for different groups of patients could be considered. Finally, more complex models at the pathway level could be studied. However, the challenges of simultaneous reconstruction of cancer pathways and complex models among them from a finite number of samples imply that these generalizations, and the comparison of the different models one can obtain, will not be straightforward.

ACKNOWLEDGMENTS

This work is supported by NIH grant R01HG007069-01 and by NSF grant IIS-1247581.

AUTHOR DISCLOSURE STATEMENT

The authors declare they have no competing financial interests.

5. APPENDIX

5.1. Exclusivity or progression alone are not enough: examples

Let's assume first that we try to identify the progression by identifying the sets of maximum exclusivity first (the progression among the sets is defined afterward). For example, assume that we have a mutations matrix M describing the mutations in 4 genes g_1, g_2, g_3, g_4, and m samples $s_1, \ldots, s_m$, and want to find the best partition of the columns of M into $K=2$ sets. Let's assume that c_1 is 1 in rows $r_1, r_2, \ldots, r_{\lceil m/2 \rceil}$; c_2 is 1 in rows $r_{\lceil m/2 \rceil}, r_{\lceil m/2 \rceil+1}, \ldots, r_m$; c_3 is 1 in rows $r_1, r_{\lfloor m/4 \rfloor}$; and c_4 is 1 in rows $c_{\lceil m/2 \rceil+1}, \ldots r_{\lceil m/2 \rceil+\lfloor m/4 \rfloor}$. It is easy to verify (see Fig. 1b) that the best (and correct) progression model is given by the partition $\mathcal{P}=\{\{c_1, c_2\}, \{c_3, c_4\}\}$. However, since sets $\{c_1, c_4\}$ and $\{c_2, c_3\}$ both present perfect mutual exclusivity of 1's, while $\{c_1, c_2\}$ does not, the partition $\{\{c_1, c_4\},\{c_2, c_3\}\}$ would be reported when only mutual exclusivity is taken into account, therefore leading to an incorrect inferred progression model.

On the other end, if one infers the progression model by looking for evidence of progression among single genes, without considering exclusivity among mutations, there are instances in which the best (and correct) partition would not be inferred (Fig. 1b). The reason is that progression at the genes level needs to appear as significant co-occurrence between 1's in two columns, which is not necessarily present in our model, as the following example shows. Consider the case of data coming from a progression model defined as partition $\mathcal{P}=\{P_1, P_2\}$, with $P_1=\{c_1, c_2, c_3\}$ and $P_2=\{c_4, c_5, c_6\}$; 1's in P_1 are perfectly exclusive, and 1's in P_2 are as well. Assume that each of c_1, c_2, c_3 is 1 in one-third of the rows (all rows have a 1 in P_1). Assume that each of c_4, c_5, c_6 appear in one-sixth of all rows (half of the rows have a 1 in P_2), and for each pair (c, c') $P_1 \times P_2$, 1/18th of the rows have 1's in both c and c'. It is to verify that for each pair (c, c'), the 1's in (c, c') appear as independent [since the number of rows with 1's in both (c, c') is equal to the expectation under the assumption of independence between c and c'], therefore providing no evidence of progression. Pairs of columns in the same set P_k $(k=1, 2)$ are perfectly exclusive, with no significant co-occurrence of 1's. Therefore, without including the exclusivity of mutations in the problem, an arbitrary partition would be inferred, corresponding with very high probability to an incorrect partition.

5.2. Proofs

In the following we denote the minimum number of changes of mutations in sample s to have it satisfy the constraint defined by a partition $\mathcal{P}$ with $f(s, \mathcal{P})$.

Theorem 2.1. *The pathway linear progression reconstruction problem with the requirement that a given column $\bar{c}$ is in a given set of the solution is NP-hard for any value of K.*

Proof. The proof is by reduction from the *edge bipartization problem*, a well-known NP-complete problem (Guo et al., 2006). For clarity, we first present the proof for $K=2$, and then we generalize it to any value of K. In the edge bipartization problem, one is given in input an undirected graph $G=(V, E)$ with $|V|=n$ and $|E|=m$, and looks for the minimum cardinality set of edges to be removed from E to make G bipartite. It's easy to see that this is equivalent to find a partition of the vertices V of G into sets V_1, V_2, which minimizes the sum of cardinality of the set E_{V_1} of edges among vertices in V_1 and the cardinality of the set E_{V_2} of edges among vertices in V_2 (that is, $E_{V_1}=\{\{u, v\} \in E : u \in V_1, v \in V_1\}$, and E_{V_2} is defined analogously).

Given the input graph G for the edges bipartization problem, we build an instance of the pathway linear progression reconstruction problem over the set of genes $\{g_v : v \in V\} \cup \{\bar{g}\}$. We define two sets of samples $\mathcal{S}_1$ and $\mathcal{S}_2$, where to each edge $\{u, v\}$ in E we associate one sample $s_1^{\{u,v\}} \in \mathcal{S}_1$ and one sample $s_2^{\{u,v\}} \in \mathcal{S}_2$. The mutations in $s_1^{\{u,v\}}$ are defined to be g_u, g_v, $\bar{g}$, while mutations in $s_2^{\{u,v\}}$ are g_u, g_v. Let M be the mutation matrix defined by these genes and samples.

Now, let's consider a solution of the pathway linear progression reconstruction problem with $K=2$, which is a partition $\mathcal{P}=\{P_1, P_2\}$ for the mutation data above. We require $\bar{g}$ to be in P_1. Note that $\mathcal{P}$ induces a partition (V_1, V_2) of the vertices V of G, with $V_1=\{v \in G : g_v \in P_1\}$ and $V_2=\{v \in G : g_v \in P_2\}$. Note that for pathway linear progression reconstruction, the order of V_1, V_2 is relevant, but we now show that the cost of the solution where vertices in V_1 and V_2 are swapped is the same. Consider an edge $\{u, v\} \in E$, and the samples $s_1^{\{u,v\}}$ and $s_2^{\{u,v\}}$. We now consider $f(s_1^{\{u,v\}}, \mathcal{P})$: if u, v are both in V_1, then $f(s_1^{\{u,v\}}, \mathcal{P})=2$ (two flips are needed to make mutations of V_1 exclusive); if $u \in V_1$ and $v \in V_2$, then $f(s_1^{\{u,v\}}, \mathcal{P})=1$ (one flip is

needed to make mutations in P_1 exclusive); if u, v are both in V_2, then $f(s_1^{\{u,v\}}, \mathcal{P}) = 1$ (one flip is needed to make mutations in V_2 exclusive). We now consider $f(s_2^{\{u,v\}}, \mathcal{P})$: if u, v are both in V_1, then $f(s_2^{\{u,v\}}, \mathcal{P}) = 1$ (one flip is needed to make mutations of V_1 exclusive); if $u \in V_1$ and $v \in V_2$, then $f(s_2^{\{u,v\}}, \mathcal{P}) = 0$ (no flips are needed); if u, v are both in V_2, then $f(s_2^{\{u,v\}}, \mathcal{P}) = 2$ (we need one flip to make mutations in V_2 exclusive and one flip to make V_1 mutated, or two flips to remove all mutations). Summarizing, we have that if u and v are both in V_1 or both in V_2, then $f(s_1^{\{u,v\}}, \mathcal{P}) + f(s_2^{\{u,v\}}, \mathcal{P}) = 3$, while if $u \in V_1$ and $v \in V_2$, then $f(s_1^{\{u,v\}}, \mathcal{P}) + f(s_2^{\{u,v\}}, \mathcal{P}) = 1$. Since summing over all samples corresponds to summing the term $f(s_1^{\{u,v\}}, \mathcal{P}) + f(s_2^{\{u,v\}}, \mathcal{P})$ over all edges $\{u, v\} \in E$, we have:

$$f(M, \mathcal{P}) = \sum_{\{u,v\} \in E} f(s_1^{\{u,v\}}, \mathcal{P}) + f(s_2^{\{u,v\}}, \mathcal{P}) = 3|E_{V_1}| + 3|E_{V_2}| + |C(V_1, V_2)| = m + 2|E_{V_1}| + 2|E_{V_2}|, \quad (1)$$

where $C(V_1, V_2)$ denotes the set of edges $\{u, v\}$ with $u \in V_1$ and $v \in V_2$, and the last equality follows from the fact that E_{V1}, E_{V2}, $C(V_1, V_2)$ is a partition of the m edges of G. Note that the cost of the solution depends only on the partition of vertices into sets V_1, V_2, and not on the order of V_1 and V_2 into $\mathcal{P}$ (i.e., if vertices in V_1 and V_2 are switched, the cost of the solution is the same). Therefore a partition $\mathcal{P}$ minimizes $f(M, \mathcal{P})$ if and only if the corresponding partition $\{V_1, V_2\}$ of V minimizes $|E_{V1}| + |E_{V2}|$.

We now consider the case of a general $K > 2$. The reduction is again from the edge bipartization problem. Given the edge bipartization problem input in the form of an undirected graph $G = (V, E)$ with $|V| = n$ and $|E| = m$, we build an instance of the pathway linear progression reconstruction problem over the set of genes $\{g_v : v \in V\} \cup \{\bar{g}_1, \bar{g}_2, \ldots, \bar{g}_{K-1}\}$. We define two sets of samples $\mathcal{S}_1$ and $\mathcal{S}_2$, where to each edge $\{u, v\}$ in E we associate one sample $s_1^{\{u,v\}} \in \mathcal{S}_1$ and one sample $s_2^{\{u,v\}} \in \mathcal{S}_2$. The mutated genes in $s_1^{\{u,v\}}$ are defined to be $g_u, g_v, \bar{g}_1, \bar{g}_2, \ldots, \bar{g}_{K_1}$, while mutated genes in $s_2^{\{u,v\}}$ are $g_u, g_v, \bar{g}_1, \bar{g}_2, \ldots, \bar{g}_{K_2}$. Let M be the mutation matrix defined by these genes and samples. We require $\bar{g}_{K_1}$ to be in P_{K-1}. It's simple to see that the optimal solution $\mathcal{P}^*$ to the pathway linear progression reconstruction problem (i) has the genes $\{g_v : v \in V\}$ (corresponding to vertices of V of G) assigned only to sets P^*_{K-1} and P^*_K; (ii) one gene $\bar{g}_i$ assigned to each P^*_k; and (iii) the cost of the optimal solution is $m + |E_{V_1}| + |E_{V_2}|$, where E_{V_1} are vertices of V with both genes assigned to P^*_{K-1}; and E_{V_2} are vertices of V with both genes assigned to P^*_K. Therefore, with the same reasoning as for the case $K = 2$, the optimal solution of the pathway linear progression reconstruction defines the optimal solution of the edges bipartization problem. ■

Theorem 2.2. *Let M be an $m \times n$ mutation matrix generated from UPLPM($\mathcal{P}$). If $m \geq Kn^2 \ln \frac{2n^2}{\delta}$, then $\mathcal{P}$ is the unique optimal solution to the pathway linear progression reconstruction problem with probability $\geq 1 - \delta$.*

Proof. Note that for any number m of samples, $f(M, \mathcal{P}^*) = 0$. We prove that whenever m satisfied the lower bound defined above, $f(M, \mathcal{P}) > 0$ for all $\mathcal{P} \neq \mathcal{P}^*$. Note that for every $\mathcal{P} = \{P_1, \ldots, P_K\} \neq \mathcal{P}^*$, one of the following is true:

(1) there exist $g \in P^*_i$ and $g' \in P^*_j$ with $i \neq j$ for which $g \in P_k$ and $g' \in P_k$ for a certain k;
(2) there exist P' and P'' such that $P^*_{i_1} = P'$ and $P^*_{i_2} = P''$ with $i_1 < i_2$ in $\mathcal{P}^*$, but $P_{i_1} = P'$ and $P^*_{i_2} = P''$ with $i_1 > i_2$ in $\mathcal{P}$.

Note that for (1) when sample s_i has mutations in both g and g', then $f(M, \mathcal{P}) > 0$; for (2) when there is a sample s_i with P' mutated and P'' not mutated, then $f(M, \mathcal{P}) > 0$.

Let's define the events: $\mathcal{E}_1 =$ ''there exists a pair of genes g, g' as in (1) that does not cause $f(M, \mathcal{P}) > 0$ (for any $P \neq \mathcal{P}$),'' and $\mathcal{E}_2 =$ ''there exists a pair of sets P' and P'' as in (2) that does not cause $f(M, \mathcal{P}) > 0$.'' We now prove that when $m \geq Kn^2 \ln \frac{2n^2}{\delta}$, $\Pr(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \delta$.

Consider two genes g, g' as in (1). The probability that they are not both mutated [not resulting in $f(M, \mathcal{P}) > 0$] in a sample is $\leq 1 - \frac{1}{K}\frac{1}{n^2}$, and if $m \geq Kn^2 \ln \frac{2n^2}{\delta}$ the probability that there is no sample in which they are both mutated is $\leq (1 - \frac{1}{K}\frac{1}{n^2})^m \leq e^{-\frac{m}{Kn^2}} \leq \frac{\delta}{2n^2}$. Since there are $\leq n^2$ possible pairs g, g', by union bound $\Pr(\mathcal{E}_1) \leq n^2 \frac{\delta}{2n^2} = \frac{\delta}{2}$.

Now consider two sets P' and P'' as in (2) above. The probability that in a sample P' is mutated and P'' is not [resulting in $f(M, \mathcal{P}) > 0$] is $\geq \frac{1}{K}$. Therefore the probably that in $m \geq Kn^2 \ln \frac{2n^2}{\delta}$ samples there is no sample for which P' is mutated and P'' is not is $\leq (1 - \frac{1}{K})^m \leq e^{-\frac{m}{K}} \leq \frac{\delta}{2K^2}$. Since there are $\leq K^2$ possible pairs P', P'', by union bound $\Pr(\mathcal{E}_2) \leq K^2 \frac{\delta}{2K^2} = \frac{\delta}{2}$.

The result follows by union bound: $\Pr(\mathcal{E}_1 \cup \mathcal{E}_2) \leq \delta$. ■

Theorem 2.3. *Let* $q = \frac{K^2 n(K-\varepsilon n^2)}{n^3(K-1)^2 + n^3 K - K^2 n^2 + 2K^3} \geq 0$ *0 for some* $\varepsilon > 0$, *and let* $\tilde{M}$ *be an* $m \times n$ *mutation matrix from UPLPM*$(\mathcal{P}, q)$. *If* $m \geq \frac{8}{\varepsilon^2} \ln \frac{2n^2}{\delta}$, *then* $\mathcal{P}$ *is the unique optimal solution to the pathway linear progression reconstruction problem with probability* $\geq 1-\delta$.

Proof. The outline of the proof is the following: we begin by bounding the expectation of the contribution of a sample s_i to the cost of the solution $\mathcal{P}^*$; we then bound the expectation of the contribution of a sample s_i to the cost of the solution $\mathcal{P}$ or any $\mathcal{P} \neq \mathcal{P}^*$; we then show that when m is large enough (i.e., as requested in the statement of the theorem), the probability that the cost of $\mathcal{P}^*$ is larger than the cost of any $\mathcal{P}$ is at most δ. In what follows we denote the fact that driver mutations in sample s_i have been selected (by the stochastic process generating the mutations) to have mutations only in $P_1^*, \ldots, P_k^*$ by saying that s_i is in *stage k*.

Consider $\mathbf{E}[f(s_i, \mathcal{P}^*]$. We have that:

$$\mathbf{E}[f(s_i, \mathcal{P}^*] = \sum_k \mathbf{E}[f(s_i, \mathcal{P}^*) | s_i \text{ in stage } k] \Pr(s_i \text{ in stage } k) = \frac{1}{K} \sum_k \mathbf{E}[f(s_i, \mathcal{P}^*) | s_i \text{ in stage } k].$$

Now assume that s_i is in stage k with $1 \leq k \leq K$. Then $f(s_i, \mathcal{P}^*) = 1$ only if s_i contains an error, and the error is in one of the genes in $P_1^*, \ldots, P_{K-1}^*$, or one of the genes in P_K^* other than the one containing the correct mutation, or one of the sets P_{k+1}, for $2 \leq i \leq K-k$. (Note that the latter event does not occur for $k = K-1$ and $k = K$). Therefore, we have:

$$\begin{aligned}\mathbf{E}[f(s_i, \mathcal{P}^*] &= \frac{1}{K}\left[(K-1)\frac{q}{n}\left(\frac{n(K-1)}{K} - 1\right)\right] + \frac{1}{K}\left[\frac{q}{n}(n-1)\right] \\ &= \frac{q}{K}\left[\frac{(K-1)^2}{K} + 1 - \frac{K}{n}\right].\end{aligned}$$

As in the proof of Theorem 2.2, we consider the partitions $\mathcal{P} \neq \mathcal{P}^*$ according to the fact that:

(1) there exist $g \in P_i^*$ and $g' \in P_j^*$ with $i \neq j$ for which $g \in P_k$ and $g' \in P_k$;
(2) there exist P' and P'' such that $P_{i_1}^* = P'$ and $P_{i_2}^* = P''$ with $i_1 < i_2$ in $\mathcal{P}^*$, but $P_{i_1}^* = P'$ and $P_{i_2}^* = P''$ with $i_1 > i_2$ in $\mathcal{P}$.

Let's first consider $\mathcal{P}$ for which (1) holds. Then if sample s_i has both g and g' mutated after an error is introduced, its contribution to the cost of the solution $\mathcal{P}$ is at least 1. That is, let's define $\mathcal{E}_1 =$ ''g, g′ are mutated in s_i after errors are introduced.'' Then

$$\begin{aligned}\mathbf{E}[f(s_i, \mathcal{P})] &= \mathbf{E}[f(s_i, \mathcal{P}) | \mathcal{E}_1] \Pr(\mathcal{E}_1) + \mathbf{E}[f(s_i, \mathcal{P}) | \bar{\mathcal{E}}_1] \Pr(\bar{\mathcal{E}}_1) \\ &\geq \Pr(\mathcal{E}_1) \mathbf{E}[f(s_i, \mathcal{P}) | \mathcal{E}_1] \\ &\geq \frac{1}{K}\frac{K^2}{n^2}(1 - q + q(1 - 2/n))) \\ &= \frac{K}{n^2}\left(1 - \frac{2q}{n}\right).\end{aligned}$$

Let's now consider $\mathcal{P}$ for which (2) holds. Then if in s_i P' is mutated and P'' is not mutated after an error is introduced, the contribution of s_i to the cost of the solution $\mathcal{P}$ is at least 1. That is, let's define $\mathcal{E}_2 =$ ''P' is mutated in s_i; P'' is not mutated in s_i.'' Then

$$\begin{aligned}\mathbf{E}[f(s_i, \mathcal{P})] &\geq \Pr(\mathcal{E}_2) \mathbf{E}[f(s_i, \mathcal{P}) | \mathcal{E}_2] \\ &\geq \frac{1}{K}\left[1 - q + q\left(1 - \frac{\frac{n}{K} + 1}{n}\right)\right] \\ &= \frac{1}{K}\left(1 - \frac{(n+K)q}{nK}\right).\end{aligned}$$

Now for $\mathcal{P} \neq \mathcal{P}^*$ for which (1) holds, let's define $X_i = \mathbf{E}[f(M, \mathcal{P}) - f(M, \mathcal{P}^*) | s_0, s_1, \ldots, s_i]$. The sequence $X_0, \ldots, X_m$ defines a *martingale*. By the definition of q, $X_0 = \mathbf{E}[f(M, \mathcal{P}) - f(M, \mathcal{P}^*)] \geq m\varepsilon$. Note that X_m is the difference between the cost of the solution $\mathcal{P}$ and the cost of solution $\mathcal{P}^*$. Since the cost of the solution

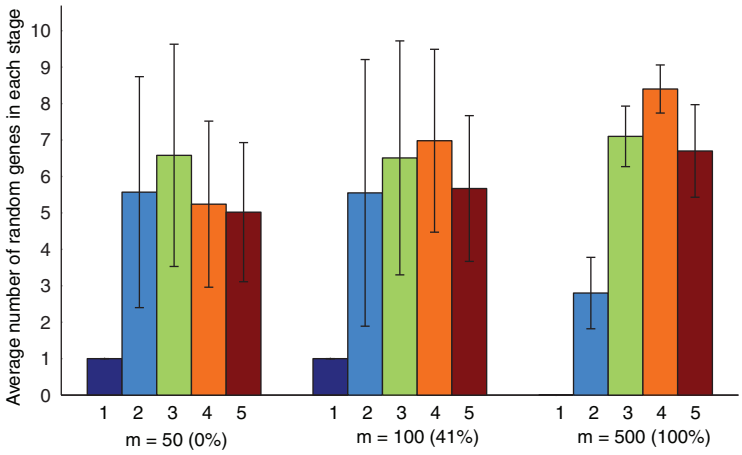

FIG. 6. Average and standard deviation (over 100 trials) of the number of random genes that are reported in each stage when $m = 50$, 100, or 500 samples with mutations generated from progression model with 5 stages and 5 genes in each stage, and mutations from 25 random (i.e., not related to progression) mutation genes with frequency 5% are included in the analysis. An error probability of $p = 0.001$ is considered.

$\mathcal{P}^*$ can increase by at most one fixing the mutations in one sample, and since for the events we are considering the cost of a solution $\mathcal{P} \neq \mathcal{P}^*$ increases by at least one, assuming $X_{i+1} - X_i \leq 1$ for all i gives us a conservative bound (since the value of $f(M, \mathcal{P})$ could increase by more than one). Therefore, by the Azuma inequality we have that with m samples

$$\Pr\left(f(M,\mathcal{P}) - f(M,\mathcal{P}^*) < m\frac{\varepsilon}{2}\right) \leq e^{-\frac{\varepsilon^2 m^2}{8m}} \leq \frac{\delta}{2n^2}.$$

By union bound on the pairs g, g', the probability that there exists a solution $\mathcal{P}$ for which (1) holds of cost less than $f(M, \mathcal{P}^*)$ is $\leq \frac{\delta}{2}$. A similar argument shows that the probability that there exists a solution $\mathcal{P}$ for which (2) holds of cost less than $f(M, \mathcal{P}^*)$ is $\leq \frac{\delta}{2}$. The result follows by union bound. ■

5.3. *Synthetic data with random genes*

We also used simulations to assess the impact on the accuracy of our method of the inclusion genes not related to the progression. We considered the progression model with $K = 5$ stages, each consisting of five genes related to the progression, described above, and also included mutations for 25 genes, each mutated in 5% of the samples independently of all other events. We generated 100 datasets from this model for each of the values $m = 50$, 100, 500, fixing and $p = 0.001$. For $m = 50$, the inferred model never corresponded to the correct model on the 25 genes related to progression; for $m = 100$, the inferred model on the 25 genes related to the progression was reported 41% of the time, while for $m = 500$ this happened 100% of the time. This shows that even when genes not associated with the progression model are included in the analysis, our method is able to correctly reconstruct the relationship between the genes associated with the progression when the number of samples is sufficiently high.

Figure 6 shows the average number of random genes reported in each stage. We observe that in general fewer random genes are reported in the first stage (no random gene is ever reported in the first stage for $m = 500$). This shows that spurious associations are more likely to be observed in late stages of the inferred progression.

5.4. *ILP running time*

Table 1 shows the running time to find the optimal solution of the ILP using CPLEX v12.3 with default parameters, using 1 CPU.

REFERENCES

Attolini, C.S.-O., Cheng, Y.-K., Beroukhim, R., et al. 2010. A mathematical framework to determine the temporal sequence of somatic genetic events in cancer. *Proc. Natl. Acad. Sci. USA* 107, 17604–17609.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. 2006. Evolution on distributive lattices. *J. Theor. Biol.* 242, 409–420.

Beerenwinkel, N., Eriksson, N., and Sturmfels, B. 2007. Conjunctive bayesian networks. *Bernoulli* 13, 893–909.

Beerenwinkel, N., Rahnenführer, J., Däumer, M., et al. 2005a. Learning multiple evolutionary pathways from cross-sectional data. *J. Comput. Biol.* 12, 584–598.

Beerenwinkel, N., Rahnenführer, J., Kaiser, R., et al. 2005b. Mtreemix: a software package for learning and using mixture models of mutagenetic trees. *Bioinformatics* 21, 2106–2107.

Beerenwinkel, N., and Sullivant, S. 2009. Markov models for accumulating mutations. *Biometrika* 96, 645–661.

Brennan, C.W., Verhaak, R.G.W., McKenna, A., et al. 2013. The somatic genomic landscape of glioblastoma. *Cell* 155, 462–477.

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.

Cheng, Y.-K., Beroukhim, R., Levine, R.L., et al. 2012. A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis. *PLoS Comput. Biol.* 8, e1002337.

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

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

Desper, R., Jiang, F., Kallioniemi, O.P., et al. 1999. Inferring tree models for oncogenesis from comparative genome hybridization data. *J. Comput. Biol.* 6, 37–51.

Desper, R., Jiang, F., Kallioniemi, O.P., et al. 2000. Distance-based reconstruction of tree models for oncogenesis. *J. Comput. Biol.* 7, 789–803.

Efron, B., and Tibshirani, R. 1994. *An Introduction to the Bootstrap*, 1st edition. Chapman and Hall, New York.

Fearon, E.R. 2011. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* 6, 479–507.

Fearon, E.R., and Vogelstein, B. 1990. A genetic model for colorectal tumorigenesis. *Cell* 61, 759–767.

Gerstung, M., Baudis, M., Moch, H., and Beerenwinkel, N. 2009. Quantifying cancer progression with conjunctive bayesian networks. *Bioinformatics* 25, 2809–2815.

Gerstung, M., Eriksson, N., Lin, J., et al. 2011. The temporal order of genetic and pathway alterations in tumorigenesis. *PLoS One* 6, e27136.

Guo, J., Gramm, J., HÃffner, F., et al. 2006. Compression-based fixed-parameter algorithms for feedback vertex set and edge bipartization. *J. Comput. Syst. Sci.* 72, 1386–1396.

Hjelm, M., Höglund, M., and Lagergren, J. 2006. New probabilistic network models and algorithms for oncogenesis. *J. Comput. Biol.* 13, 853–865.

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

Kanehisa, M., and Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30.

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., and Raphael, B.J. 2013. Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.* 9, e1003054.

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.

Rahnenführer, J., Beerenwinkel, N., Schulz, W.A., et al. 2005. Estimating cancer survival and clinical outcome based on genetic tumor progression scores. *Bioinformatics* 21, 2438–2446.

Sakoparnig, T., and Beerenwinkel, N. 2012. Efficient sampling for bayesian inference of conjunctive bayesian networks. *Bioinformatics* 28, 2318–2324.

Shahrabi Farahani, H., and Lagergren, J. 2013. Learning oncogenetic networks by reducing to mixed integer linear programming. *PLoS One* 8, e65773.

Tofigh, A., Sjölund, E., Höglund, M., and Lagergren, J. 2011. A global structural em algorithm for a model of cancer progression. *Adv. Neural Inform. Process. Syst.* 24, 163–171.

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

Vandin, F., Upfal, E., and Raphael, B.J. 2012b. Finding driver pathways in cancer: models and algorithms. *Algo. Mol. Biol.* 7, 23.

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

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

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

Zhang, J., Baran, J., Cros, A., et al. 2011. International cancer genome consortium data portal—a one-stop shop for cancer genomics data. *Database (Oxford)* 2011, bar026.

Address correspondence to:
Dr. Fabio Vandin
Department of Mathematics and Computer Science
University of Southern Denmark
Campusvej 55
DK-5230, Odense M, Denmark

E-mail: vandinfa@imada.sdu.dk