IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 28, 2018, accepted April 3, 2018, date of publication April 16, 2018, date of current version May 16, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2827024

Inference of Cancer Progression With Probabilistic Graphical Model From Cross-Sectional Mutation Data

WEI ZHANG AND SHU-LIN WANG

College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082,China

Corresponding author: Shu-Lin Wang (smartforesting@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61672011, Grant 61472467, and Grant 61471169, and in part by the Collaboration and Innovation Center for Digital Chinese Medicine of the 2011 Project of Colleges and Universities in Hunan Province.

ABSTRACT With the advance of high-throughput sequencing technologies, a great amount of somatic mutation data in cancer have been produced, allowing deep analyzing tumor pathogenesis. However, the majority of these data are cross-sectional rather than temporal, and it is difficulty to infer the temporal order of gene mutations from them. In this paper, we first show a probabilistic graphical model (PGM) to infer the temporal order constrains and selectivity relation among the mutation of cancer driver genes which are presented by a directed acyclic graph. We then apply an exponential function based on the mutation probability of these driver genes to obtain their mutation waiting time which can be used to induce mutually exclusive driver pathways. Finally, we evaluate the performance of the PGM both on simulated data and real-cancer somatic mutation data. The experimental results and comparative analysis reveal that the PGM can capture most of the selectivity relation of mutated driver genes which have been validated by previous works. Furthermore, the PGM can provide new insights on simultaneously inferring driver pathways and the temporal order of their mutations from cross-sectional data.

INDEX TERMS Probabilistic graphical model, somatic mutation, cancer progression, driver pathway, waiting time.

I. INTRODUCTION

The systematic analysis of human cancer genomes in the last decade has revealed that cancer is a complex disease caused by the accumulation of somatic mutations. Common methods of exploring carcinogenesis are to integrate a large amount of data to mine the law about the accumulation of somatic mutations. Due to the rapid development of high-throughput sequencing technologies, unprecedented amount of somatic mutation data such as the data in the Cancer Genome Atlas (TCGA) are accumulated, which brings two critical challenges to the analysis and interpretation of them. The first challenge is how to distinguish driver mutations from massive passenger mutations in cancer progression [1]–[4]. The second challenge is how to identify the temporal order of these driver mutations occurred [5], [6].

Addressing of the two challenging problems is of benefit to both therapeutic decisions and the basic understanding of carcinogenesis. The first problem can be solved by comparing the driver mutation observed frequencies across different individuals, and furthermore by identifying mutually exclusive driver gene mutations, commonly referred to as pathways which consist of multiple alterations performing the same functional role in cancer progression [3], [4], [8]. Once one of the members in a pathway is altered, cancer cells gain a significant selective advantage. However, the second problem is more difficult to address, because the temporal cancer-related data from single individual at multiple time-points are nearly impossible to be obtained [9]–[11].

There have been some computational approaches to infer temporal progression of somatic mutation. Some early works reconstruct the temporal order of cancer samples by examining mostly clinical and genetic data [12]–[14]. For example, cancer progression is described as a linear model by assuming existence of a unique and most likely temporal driver gene mutation order. On the basis of the linear model, many statistical approaches considering branch like trees and graphs have been presented. These methods can be grouped into four classes: (1) Oncogenetic trees, which represent the

2169-3536 © 2018 IEEE. Translations and content mining are permitted for academic research only.
Personal use is also permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

probabilities of accumulating further mutation along divergent temporal sequences, under the assumption that each event depends on a single parent [15]–[17]. (2) Bayesian Networks, which avoid the limitation of tree-based model, do not allow differently confluent progression paths, with the cost of increased computations [18]–[21]. For instance, Conjunctive Bayesian Networks (CBNs) are generative models of cancer progression, in which allow for multiple parental nodes, thereby modeling the synergistic effects of multiple events in promoting subsequent mutations and describing the accumulation of events that are constrained in the order of their occurrence [22]. (3) Clustering and evolutionary fitting algorithms. These approaches generate graphs, in which the node denotes single gene and the edges represent the relationship between nodes [23]–[25]. (4) Other approaches, such as Progression Networks [26], RESIC [24], CAPRESE [9] and CAPRI [27]. Progression Networks is similar to Conjunctive Bayesian Networks. It employs mixed integer linear programming to reduce the difficulty of learning the Bayesian. RESIC is the evolutionary mathematical approach to explicitly consider the evolutionary dynamics of driver mutation accumulation. CAPRESE and CAPRI use a framework of probability causation to infer cancer progression at the gene level. However, these methods infer progression at the individual gene level which is hard to reflect the heterogeneity of inter-patients. In recent years, several works begin to focus on modeling cancer progression at the pathway level instead of the individual gene level. For example, Vandin et al. formulate the problem as an integer linear program (ILP) [10] and Hao Wu et al. present a Network-based method to infer cancer progression (NetInf) [28]. pathTiMEx [29] employs a stochastic optimization procedure to jointly optimize the assignment of genes to pathways and the evolutionary order constraints among pathways. But these methods infer cancer progression by considering cancer progression as single linear path at the pathway level and ignore the selectivity relationship among driver genes, which restricts the representation of carcinogenesis.

We develop probabilistic graphical model (PGM) based on causal dependency theory to infer cancer progression, which not only considers the cancer progression at the driver pathway level but also at the individual gene level that provides a better representation of carcinogenesis. First, the model utilizes directed acyclic graph (DAG) to represent the selectivity relations of driver genes, and the presence or absence of edges in DAG is determined by conditional probability which can be estimated from the cancer mutation data. When we construct the DAG, an intersection degree (ID) that describes more exact relationship between each driver gene pairs is used. Then, the waiting time between a driver gene mutation and the subsequent is estimated by a stochastic function of these genes mutation probability to reflect that the waiting time is random and independent of each other, and the impossibility to infer the exact time of the occurrence of the gene mutation from the observed data. And we use percentile to divide the waiting time into several stages. Different individuals may harbor driver mutation in different genes within a pathway, and the genes in the same pathway are likely to in the same progression stage (also called as in the same waiting time stages) [10], [30], so it is reasonable that we assume that the driver genes in the same pathway mutate at the same waiting time stage. Finally, we mark the driver genes up which are in the same pathway with the same color. From our results at DAG, we list the detailed selectivity relations between the driver genes and mark several driver pathways. So PGM constructs the DAG of cancer progression not only at the individual gene level but also at the driver pathway level, it is of benefit to understand the high inter-patient heterogeneity and carcinogenesis following different progression paths of different individuals. PGM is evaluated both on the simulate data and real cross-sectional data from colorectal and glioblastoma cancer, respectively. The experimental results indicate that our progression model can reveal the intrinsic properties of the progression of driver gene mutations in these cancer types. The workflow of our model is shown in Fig. 1a, and a schematic diagram is shown in Fig. 1b.

II. METHOD

The PGM is an inferential method that deduces the causal dependencies and the waiting time among the driver genes from cross-sectional somatic mutation data. By means of the waiting time, it can mark the driver genes that belong to the same pathway. Given the binary mutation matrix M with m rows (samples $s_1, s_2, \ldots, s_m$) and n columns (driver genes $g_1, g_2, \ldots, g_m$), where the samples on the rows and driver genes on the columns, where $M_{i,j} = 1$ if g_j is mutated in sample s_i, and $M_{i,j} = 0$ otherwise. PGM generates a probabilistic graph DAG using the putative causal dependencies theory that is described by definition 1 [9], [27], in which nodes represent driver genes and edges represent the selectivity relation of them. An edge is determined by two points: (1) driver gene marginal probability and conditional probability, and (2) the intersection degree (ID) which is defined to measure relationship between arbitrary two driver genes.

Definition 1 (Causal Dependency Theory): Given two observable genes i and j, there is selectivity relation between them if meet the two conditions: (1) if i is a prima facie cause of j, (2) if i is the probability raising of j, that is i occurs more frequently:

$$p_{j/i} > p_{j/\bar{i}} \quad and\ p_i > p_j \tag{1}$$

p_i and p_j are the marginal probability of the driver genes i and j. Besides, $p_{j/i}$ is the conditional probability of the driver gene j, which is one of the conditions to verify whether each driver gene pair i and j connect or not. This definition is part of Suppes causality theory [31], and several works have proved its role in the definition of the reconstruction problem and some of its limitations [32], [33]. It is only a necessary but not sufficient condition and additional constraints need to be imposed to filter spurious relations, e.g., it may be that for some prima facie cause A of a gene B, there is a third event C prior to both, C causes A and ultimately A causes B. C may

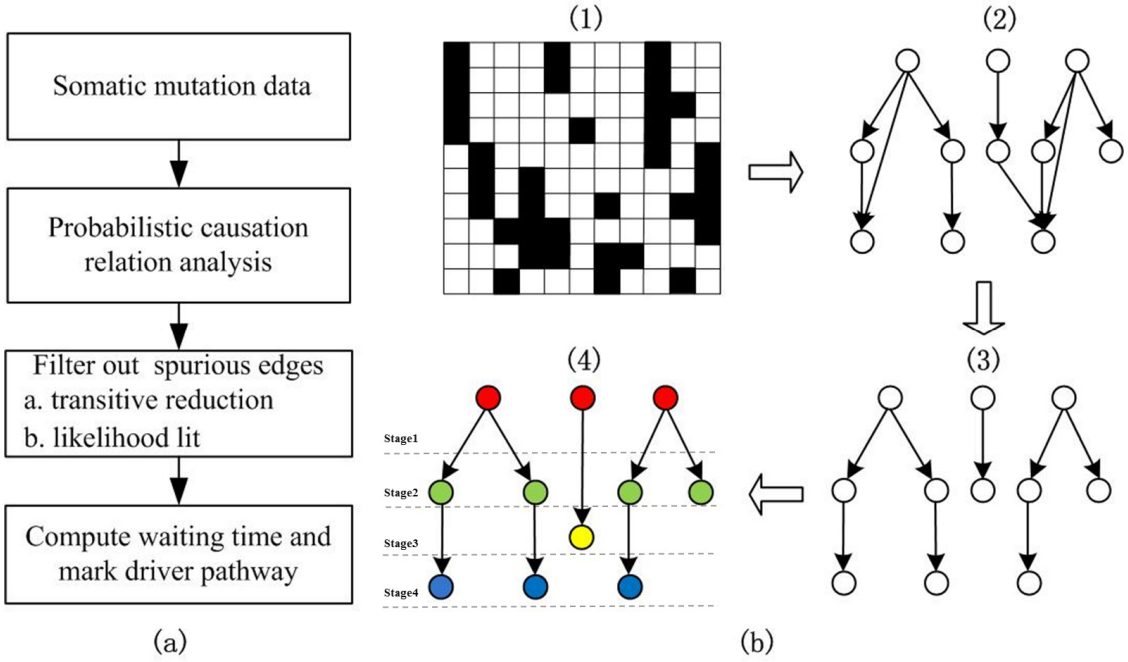

FIGURE 1. Overview of PGM. (a) The workflow of the stepwise study, (b) in the first step, a cancer driver mutation dataset is inputted in the form of a binary alteration matrix, with rows representing patients and columns representing driver mutations. A black square encodes the presence of a driver mutation, and a white square encodes its absence. In the second step, construct the initial directed acyclic graph, and edge is created if the pair of driver gene meet the definition 1 and ID > λ. In the third step, process the initial DAG, filter the spurious edges and transitive reduction. In the fourth step, estimate the waiting time, divide the driver genes with waiting time, and mark different stages up with different color.

cause both A and B independently, and the causation relationship observed from A to B is merely spurious. In this paper, standard maximum likelihood fit and Bayesian Information Criterion (BIC) are used to filter spurious relations that are intruduced at the following detailed steps.

To automatically extract the selectivity relation and reduce the computational complexity, it is necessary to filter the insignificant relations. We define ID between driver gene pair g_i and g_j to measure their relation.

Definition 2 (Intersection Degree, ID):

$$ID\left(g_i, g_j\right) = \frac{|\Gamma\left(g_i\right) \cap \Gamma\left(g_j\right)|}{|\Gamma\left(g_i\right) \cup \Gamma\left(g_j\right)|}. \tag{2}$$

$\Gamma(g)$ denotes the coverage of driver gene g, and it represents the set of patients in which the driver gene g mutated, $\Gamma(g) = \{i : M_{i,g} = 1\}$.

The two driver genes in (a), (b), (c) from Fig.2 all meet the definition 1, but (a) has the biggest *ID* value, so we consider that the driver gene pair in (a) has stronger selectivity relation than driver gene pair in (b) and (c).

In the following sections we describe PGM that adopted the notation from Szabo and Boucher [17].

- Let n be the number of driver genes.
- Let p_i and p_j respectively denote the marginal probability that the i-th and j-th driver gene mutation occur, and $i = 1, 2, \ldots, n, j = 1, 2, \ldots, n$ and $i \neq j$. Let $p_{j\cap i}$ denotes the probability that both the i-th and j-th driver gene mutation occur simultaneously.

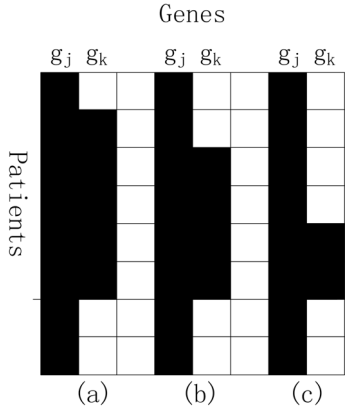

FIGURE 2. Analysis of the intersection degree. (a),(b),(c) in the figure stand for the frequency of g_i and g_j in different case, they are satisfied with definition 1, but the ID of the two genes in the three cases is (a) $ID(g_j, g_k) = 0.63$ (b) $ID(g_j, g_k) = 0.50$ (c) $ID(g_j, g_k) = 0.25$.

- Let $p_{j/i}$ denotes the conditional probability that the j-th driver gene mutation occurs given that the i-th driver gene mutation has occurred. The definition of conditional probability $p_{j/i}$ is

$$p_{j/i} = \frac{p_{j\cap i}}{p_i}, \quad i, j = 1, 2, \ldots, n;\ i \neq j. \tag{3}$$

- Let $p_{j/\bar{i}}$ denotes the conditional probability that the j-th driver gene mutation occurs given that the i-th driver

gene mutation does not have occurred. The definition of conditional probability $p_{j/\bar{i}}$ is

$$p_{j/\bar{i}} = \frac{p_j - p_{j\cap i}}{1 - p_i}, \quad i, j = 1, 2, \ldots, n;\ i \neq j. \tag{4}$$

The detailed steps of PGM are described as follows.

Step 1: Calculate the marginal probability p_i, p_j, conditional probability of mutation in each gene pair $p_{j/i}, p_{j/\bar{i}}$ and their *ID* of driver gene mutation from the data using the above definition 2 ($i, j = 1, 2, \ldots, n$; $i \neq j$). The marginal probability of a driver gene mutation is calculated as the frequency of the driver mutation in a dataset.

Step 2: Construct the initial DAG on the vertices $\{v_1, v_2, \ldots, v_n\}$ which represent driver genes, and an edge is created between the two vertices if them meets definition 1 and ID $>$ λ, otherwise, there is no edge between the driver genes. λ is an experienced threshold, which can be used to remove some insignificant edges from the initial DAG.

Step 3: Update the obtained DAG. We filter the redundant edges through transitive reduction and spurious edges through likelihood fit. The edges a $\rightarrow$ b $\rightarrow$ c can be described with a $\rightarrow$ b and b $\rightarrow$ c if the mutation of driver gene a cause b, b cause c, but a does not cause c.

The transitive reduction algorithm [34] is adopted to eliminate the redundant edge in the initial DAG by traversing vertices in inversely topological sorting order. All of the vertices reachable from the vertex v must be processed before process v. For the vertex v, we preferentially process the vertex with the bigger number in all of the vertices that are reachable from v. If an edge is firstly computed, it is marked as reachable one. If a vertex v is reachable from an edge e and the vertex v has been marked as reachable one, the edge e is redundant and should be deleted.

For any selectivity structure of DAG, spurious edges can obviously contribute to a reduction in the likelihood-fit with respect to true edges. Thus, we adopt a standard maximum likelihood fit and Bayesian Information Criterion (BIC) to prune the DAG. The BIC score is measured by (5).

$$BIC = ln\widehat{L(M)} - \ln(m) * dim(M)/2. \tag{5}$$

Here M is an input observed data, m denotes the number of patients, and dim(M) is the number of parameters in the model, which depends on the number of parents of each node has. $\widehat{L(M)}$ is the maximized value of the likelihood function. We first process the obtained DAG through adding, deleting and reversing single edge and obtain a set of new DAGs, then, we compute the BIC score of the obtained DAG and new DAGs using equation (5), that is compute the score of a node and its parents given completely observed data. Finally, we select the best score DAG.

Step 4: Compute the mutation probability of each driver gene. Let $p(i)$ denotes the probability of the i-th driver gene mutation ($i = 1, 2, \ldots, n$), which can be computed with four cases. First, if the driver gene does not have in-degree, the probability of this driver gene is equal to marginal probability $p(i) = p_i$. Second, if the driver gene has one in-degree and the number of its precursor is more than 1, the probability of this driver gene can be computed by $p(i) = \sum_{k=1}^{u} p_{i/j_k} + \prod_{k=1}^{u} p(j_k)$, where u is the number of $i's$ precursor and $u > 1$, p_{i/j_k} is the conditional probability of its precursor, the so-called precursors are the driver genes that mutated before the target driver gene. In this case, the mutated probability of driver gene i is affected by any of a precursor (it computed by $\sum_{k=1}^{u} p_{i/j_k}$) or simultaneously affected by all precursors (it computed by $\prod_{k=1}^{u} p(j_k)$). Third, if the driver gene has only one in-degree and the number of its precursor is equal to 1, the mutation probability of this driver gene only can be affected by a precursor driver gene, so it can be computed by $p(i) = p_{i/j_1}$. Last, if the driver gene has several in-degrees, the mutation probability of this driver gene can be computed by $p(i) = \sum_{t=1}^{w} p(t)$, where w denotes the number of in-degrees, and $p(t)$ is described in the same as the second or third situation. The mutation probability of driver genes represents the strength of the selectivity relation between a mutated driver gene and its parents excluding the driver genes without parents.

Step 5: Estimate the waiting time of driver gene mutation as realization of an exponential process which is inversely proportional to the driver gene mutation probability. For each mutated driver gene pair that connection in the DAG draw a realization Δt from an exponential distribution.

$$\Phi \sim \mu \exp(-\mu \Delta t), \quad \mu \equiv p(i). \tag{6}$$

Here Δt is the waiting time of the driver gene j mutated. The rate parameter μ is the mutation probability computed by Step 4 of the mutated driver gene j. We produce a set of random values (the number of random values equal to driver gene numbers) which meets exponential distribution with the rate parameter μ to represent the waiting time of the mutated gene.

Step 6: Identify and mark driver pathways. We use percentile to divide the waiting time into k stages (we use the value of k set by Raphael and Vandin [10]). Step 5 and Step 6 run n ($n = 1000$ used in this paper) times repeatedly to obtain n pathway distribution cases, the pathway distribution cases are then ranked according to the frequency of each pathway distribution case. The statistical significance test is adopted to evaluate the ranked pathway distribution case, and the optimal pathway distribution case with minimum *p-value* is selected. The final DAG is derived not only at the pathway level but also at the gene level after marking the optimal pathway distribution.

III. RESULTS

To assess the robustness of the proposed model, our experiment analysis is conducted on the simulated datasets with addition of different levels of noise and three real cancer datasets.

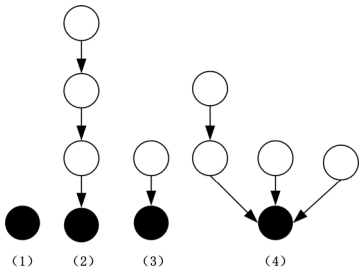

FIGURE 3. The schematic diagram for four cases computation of driver gene mutation probability. The filled circles represent the target driver genes for computation mutation probability and the hollow circles represent the precursor mutated driver genes of the target driver genes. (1) In the first case, there is no in-degree for the target gene, (2) in the second case, there is one in-degree for the target driver and its precursor is more than 1, means that several driver genes mutated before the target driver gene and they depend on one path, (3) in the third case, there is one in-degree for the target driver gene and its precursor is 1, (4) in the fourth case, there are several in-degrees for the target driver gene.

A. SIMULATED DATA

We perform a large number of experiments using simulated data with different levels of noise. Mutation datasets are produced according to the progression model. In particular, we consider two kinds of progression models with $k = 5$ stages and $n = 25$ genes, as well as $k = 4$ stages and $n = 16$ genes, respectively. We generate 100 simulation datasets with the form of the binary alteration matrix of m samples for each of the two progression models, in which 1 represents the presence of a mutation, 0 represents its absence. Without loss of generality, we randomly generate these datasets that allow the mutation frequency in one gene between 0.1 % and 60 %. Noises are added by flipping some entry of the corresponding mutation data with different probability q. We consider values of $m = 100, 200, 400, 600, 800, 1000$ and $q = 0.001, 0.01, 0.05$.

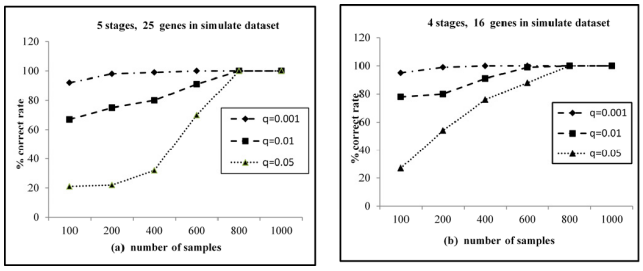

FIGURE 4. Correct ratio for different number of samples and different probabilities of noise addition. Axis *x* represents the number of samples, and axis *y* represents correct ratio, and the mutation matrix *M* comes from a progression model with *k* stages, and include *n* genes. Noise is added to the each matrix *M* with a probability *q*. We don't fix the number of genes in each stages. (a) Result for $k = 5$, *M* containing 25 genes and different values of *q* and different number of samples. (b) Result for $k = 4$, *M* containing 16 genes and different values of *q* and different number of samples.

For each combination of m and q, we record correct times that the genes belong to the corresponding stages. In these experiments we only consider the final results on the relationship between pathways. Experimental results are shown in Fig. 4. It is clear that the correct rate increases when the value of q decreases, and the correct rate increases when the number of samples increases. For $q \leq 0.01$, when 100 samples are analyzed the correct progression model is reported in most cases. When 800 samples are analyzed the correct model is reported in every case. When $q = 0.05$, the correct rate is no more than 80% with 600 samples are analyzed, and the 100% correct rate can be obtained when 800 samples are considered. These results indicate that while data from reasonably sized can be used to infer the correct progression model. If the noise level is high, a larger number of data may be required to identify the correct progression model. To better understand how the number of stages and genes impact the complexity of PGM, we compare the model with $k = 5$ (including 25 genes) in dataset and the one with $k = 4$ (including 16 genes) in dataset (shown in Fig. 4). For the same m and q combination, the correct rate of the latter is always greater or equal to the former. For example, when $q = 0.01$ and $m = 400$ samples are considered, the latter is greater than 80% and the former is lower than 80%. The simulated experiments demonstrate that correct progression model can be obtained with low noise levels. Note that the assignment of genes number to pathway is not fixed, with the sole restriction that each pathway contains at least one gene.

In the experiments on simulated data, when noise probability q is set to be 0.001 and λ is set to the 65th percentile of all *ID* values, we obtain exact results from 99 runs out of 100, when noise probability q is set to 0.01 and λ is set to 60th percentile of all *ID* values, we obtain exact results from 90 runs out of 100. When q is set to 0.05 and λ is set to the 55th percentile of all *ID* values, we obtain exact results from 85 runs out of 100. When noise probability q is set to 0.001 and λ is set to the 55th percentile of all *ID* values, we obtain exact results from 99 runs out of 100. The results show when λ is set to the 65th we obtain ideal results only if the datasets have low noise. As λ is set to the 55th, we can obtain ideal results even if noise probability q is relatively high. We compute *ID* in the gene pair, and set an adjustable percentile value of 55th $\leq \lambda \leq$ 65th, which should yield good results in conducting the experiments.

Model performance is compared with other models using Hamming distance (HD). HD measures the structural similarity among the progression stage distribution inferred from the dataset with noise and the progression stage distribution inferred from dataset without noise, in terms of the minimum-cost sequence of gene edit operations that transforms the former progression stage distribution into the latter. This measure is bounded above by n when all genes have incorrect distribution. In Fig 5, we show the performance comparison of PGM and other models (including ILP [10], NetInf [28], and pathTiMEx [29]) that are all recent models to infer cancer progression at the pathway level, in terms of Hamming distance, on simulation datasets with two progression models. Particularly, we show the performance at different values of the sample size m with a fixed noise probability. As is shown in Fig. 5, PGM outperforms all the competing models with respect to all the possible combinations of noise probability

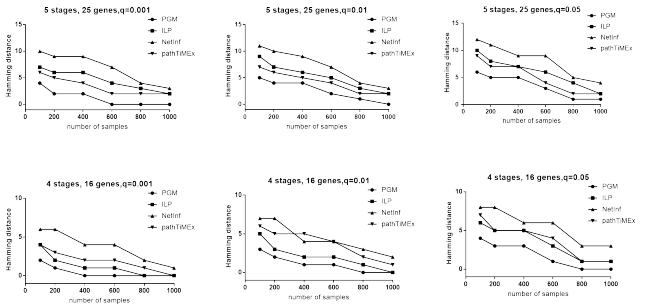

FIGURE 5. The performance of PGM and other models is compared using Hamming distance. The first row is the comparison results of the progression model with k = 5 stages and n = 25 genes with different noise probability, and the second row is the comparison results of the progression model with k = 4 stages and n = 16 genes with different noise probability. We obtain the average Hamming distance (HD) with 1000 runs between the progression stage distribution of noise dataset and no noise dataset with m = 100, 200, 400, 600, 800, 1000. The lower the HD, the smaller is the total rate of false inference progression stages among genes.

and sample size. In other words, we prove on the basis of extensive simulation tests that PGM needs a much lower number of samples than other models to converge to the correct progression stage distribution and also that it is much more robust even in the presence of significant amount of noise in the dataset, irrespective of the underlying topology.

TABLE 1. Overview of the datasets used in this study and basic information about these datasets.

Ct	Spl	Ge	Am	Ag	Mf
CRC1	95	8	2.52	29.9	78
CRC2	223	14	2.77	44.1	165
GBM	232	25	1.99	18.4	90

In the table, Ct, Cancer type; Spl, number of samples; Ge, number of genes; Am, average number of mutations per sample; Ag, average of mutation frequency for all genes; Mf, maximum of mutation frequency of all genes; CRC1, colorectal cancer data reported in [1]; CRC2: colorectal cancer data from TCGA[2]; GBM: glioblastoma multiforme data from UCSC[7] cancer browser for TCGA glioblastoma multiforme dataset.

B. REAL CANCER DATA

To assess the performance of PGM on real somatic mutation data, we test PGM on three public available cancer datasets including colorectal cancer dataset (CRC1) [1], colorectal mutation dataset (CRC2) [2] and giloblastoma multiforme dataset (GBM) [7]. Table 1 summarizes the information of the three datasets, including number of samples, number of driver genes. For the convenience of describing the results, we show the results in two forms: linear diagram and DAG. In the DAG, the mutated driver genes may appear two or more times to facilitate the expression of the relationship among them. We adopt the permutation test to assess the statistical significance of our results, estimating the probability of obtaining a set of DAGs in which the sum of mutation probability less or equal to the sum of mutation probability of the DAG in our result when the mutations are placed independently in the samples preserving the mutation frequency of the driver genes. Furthermore, to measure the stability of the assignment of a particular driver gene to a stage in the progression, we compute the correct rate of driver genes that appear in a particular stage of the progression using bootstrap re-sampling [35].

The experimental datasets we tested are also used in ILP [10], NetInf [28], and pathTiMEx [29], which inferred cancer progression at the pathway level. The results of linear diagram in PGM are highly consistent with the three previous models on the three datasets. Most of all, our model reflect cancer progression not only at the pathway level but also at the individual gene level. The selectivity relation between each driver gene pair that is shown in our DAGs is highly consistent with the existing biomedical literature. Here we take ILP as an example to compare with PGM.

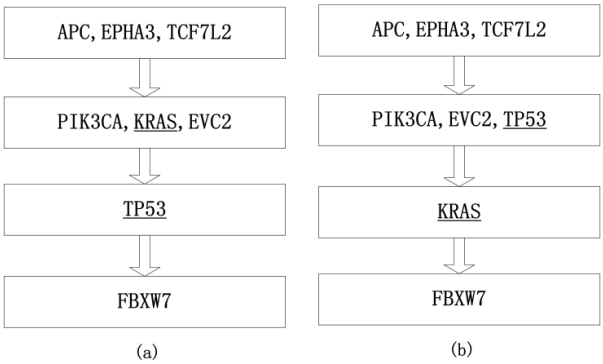

FIGURE 6. Progression model built with the use of colorectal cancer data that showed by linear form. (a) The result obtained by applying the proposed PGM; (b) the result obtained by ILP.

1) COLORECTAL CANCER DATA

PGM is firstly evaluated on the CRC1 dataset, which consists of mutations from 95 samples, including eight driver genes with mutation frequency above 5%. The eight driver genes are TP53, KRAS, EVC2, APC, TCF7L2, PIK3CA, FBXW7 and EPHA3. The pathway-based evolutionary progression of colorectal cancer obtained by PGM is shown in Fig. 6a. It is very similar to the progression model inferred by the integer linear program (ILP) model shown in Fig. 6b. The threshold λ of the intersection degree is set to the 55th percentile of all ID values. For all genes, the average correct rate they appear in the same stage out of 100 bootstrap datasets is 82%.

The reconstruction of progression model reaffirms the current knowledge on colorectal cancer. For example, APC mutation is an early event, followed by those in KRAS and TP53 [29]. However, the only difference between the two progression models is that KRAS mutation occurs earlier than TP53, inferred by PGM, while TP53 mutation occurs earlier than KRAS, inferred by ILP. Actually, some literatures [18], [36], [37] show that the temporal order of the TP53 and KRAS mutations are not clear yet.

Our model can derive not only the linear temporal order diagram but also the DAG of cancer progression. The DAG of CRC1 ($p - value < 0.01$) obtained by PGM is shown in

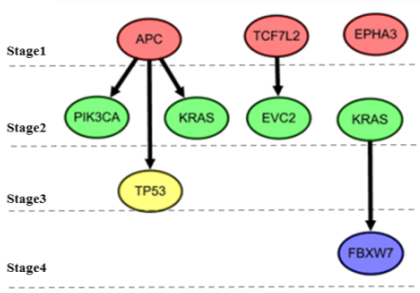

FIGURE 7. Progression model is showed by DAG. Red circles represent the first stage, green circles represent the second stage, yellow circles represent the third stage, and blue circle represents the fourth stage.

Fig. 7, in which we mark the different pathway with different color and there is no edge between two driver genes belonging to the same pathway. The dashed lines indicate different progression stages. The KRAS and PIK3CA mutations occur after APC, and they are independent with each other [18]. TCF7L2 is earlier than EVC2, which is consistent with the current knowledge on colorectal tumorigenesis [36], [37]. We derive the selectivity relation (temporal order) between mutations in KRAS and FBXW7 shown in Fig. 7, and it has been proved by Gerstung *et al.* [18]. Among these genes EPHA3 is an independent node, which indicates that it has no relation with other nodes. Therefore, our model derives not only the relationship among driver genes but also between driver pathways.

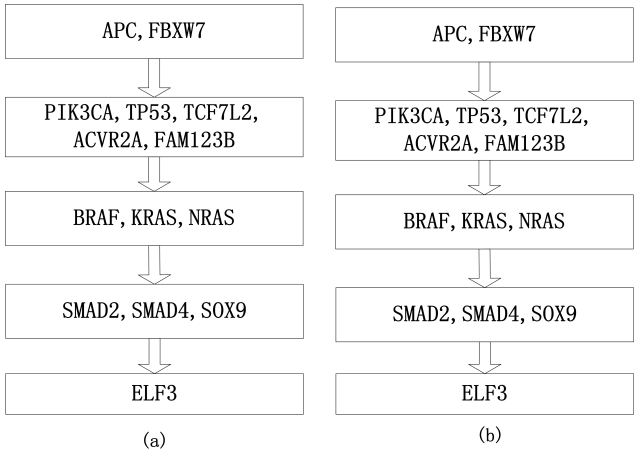

FIGURE 8. Progression model built with the use of colorectal cancer data that showed by linear form. (a) the result obtained by applying the proposed PGM method; (b) the result obtained by ILP method.

2) TCGA COLORECTAL CANCER

We further analyze the colorectal mutation data including 223 samples (CRC2) from TCGA. The 14 driver genes identified as recurrent mutation in previous work are analyzed. The linear result of our model is shown by Fig. 8a and the progression model inferred by ILP is shown in Fig. 8b. The *ID* threshold λ is set to the 65th percentile of all *ID* values. For all driver genes, the average correct rate they appear in the same stage out of 100 bootstrap datasets is 89%. It is obvious that our result of linear form is the same with the result obtained

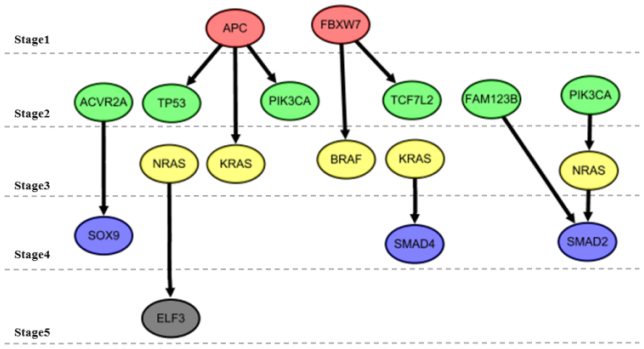

FIGURE 9. Progression model is showed by DAG. Red circles represent the first stage, green circles represent the second stage, yellow circles represent the third stage, blue circles represent the fourth stage, and black circle represents the fifth stage.

by ILP in CRC2, the difference is that we list the relationship of among driver genes, ILP do not list.

Interestingly, the progression model restricted to the driver genes APC, PIK3CA is the same as the one that we identify from the smaller dataset CRC1. KRAS and TP53 have the different stage from the smaller dataset. We consider that this dataset add some driver genes like BRAF and NRAS which influence the result, because BRAF, NRAS, KRAS are considered in the same pathway [28] and they are part of the Ras-Raf pathway. Moreover, the bootstrap analysis reveals that TP53 and NRAS have the most stable assignments to the different stages of the progression.

The most likely DAG ($p-value < 0.01$) obtained by PGM is shown in Fig. 8. Actually, it is highly accordance with our findings on the CRC1 dataset, and also highly consistent with the current knowledge on colorectal tumorigenesis [36], [37]. Specifically, APC mutation is the earliest event, followed by mutations in KRAS, TP53 and PIK3CA [38]. SMAD4 and SMAD2 mutations are followed by those in KRAS and NRAS, respectively [39]. FBXW7 mutations correlated positively with BRAF mutations and PIK3CA correlate positively with NRAS. FAM123B involves in the Wnt signaling pathway, and it is positively influence SMAD2 [40]. For the pair of NRAS and ELF3 ($p - value < 0.01$), $p_{NRAS} > p_{ELF3}$ and $p_{ELF3/NRAS} > p_{ELF3/\overline{NRAS}}$, moreover, their *ID* is 0.21, it is greater than 65th percentile of all *ID* values. For the pairs of ACVR2A and SOX9 ($p - value < 0.01$), $p_{ACVR2A} > p_{SOX9}$ and $p_{SOX9/ACVR2A} > p_{SOX9/\overline{ACVR2A}}$, moreover, their *ID* is 0.19. It is greater than the 65th percentile of all *ID* values. So, the pair of NRAS and ELF3 and the pair of ACVR2A and SOX9 very likely have selectivity relation. In conclusion, PGM provides a reasonable linear model of colorectal cancer and offers a better explanation of colorectal tumorigenesis than the existing models [10], [18].

3) TCGA GLIOBLASTOMA MULTIFORME

We download GBM dataset from UCSC cancer browser, and restrict our analysis to the driver genes reported in previous work [41]. We filter the samples and genes with all zeros in the dataset, and derive the data consisted by 25 driver genes and 232 samples. The pathway-based evolutionary

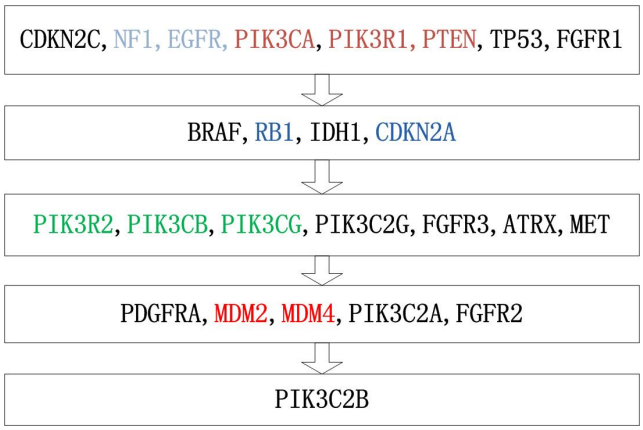

FIGURE 10. Progression model built with the use of TCGA Glioblastoma Multiforme that showed by linear form.

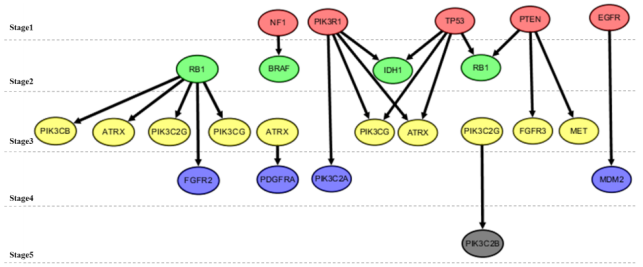

FIGURE 11. Progression model is showed by DAG. Red circles represent the first stage, green circles represent the second stage, yellow circles represent the third stage, blue circles represent the fourth stage, and black circle represents the fifth stage.

progression of giloblastoma multiforme obtained by PGM is shown in Fig. 10 and the DAG ($p - value < 0.01$) is shown in Fig. 11. The *ID* threshold λ is set to the 65th percentile of all *ID* values. For all driver genes, the average correct rate is 80% in 100 bootstrap datasets. EGFR and NF1 are the core members of the MAPK signaling pathway, EGFR expression is associated with the development of the Schwann cell-derived tumors characteristic of NF1 [42]. PIK3CA, PTEN and PIK3R1 are the core members of the RTK/RAS/PI(3)K signaling pathway which is prominently altered in glioblastoma. RB1 and CDKN2A are from the Rb1 pathway, PIK3R2, PIK3CB and PIK3CG are in the PI3K pathway; MDM4 and MDM2 are the core members of the p53 signaling pathway. The results shown in Fig. 11 are consistent with the model proposed by Misra *et al.* [43], which indicates that TP53, PTEN, EGFR, NF1 and PIK3R1 mutations are the initiating event in secondary glioblastomas. RB1 follows most commonly TP53 and PTEN mutations, IDH1 follows TP53 [43], ATRX mutation occurs behind TP53 mutation [44]. Somatic mutation of PIK3R1 provides tumors with an additional mechanism to deregulate PI3K signaling and promote tumor progression [45], so PIK3R1 has positive influence on PIK3CG and PIK3C2A. GBM with NF1 mutations might make benefit for a RAF inhibitor as part of a combination, as shown for BRAF mutant cancers [46]. And EGFR is positive to MDM2 [47]. In conclusion, the above analyses suggest that the PGM can identify driver pathways and can provide the most major selectivity relation among driver genes from GBM.

IV. CONCLUSION AND DISCUSSION

Inference of temporal order on driver gene mutations is a valuable research in exploring cancer pathogenesis, clinical diagnosis and therapy, as well as pharmaceutical research and development. In this paper, we propose PGM to infer the cancer progression not only at the individual gene level but also at the pathways level, and can infer the selectivity relation in each driver gene pairs from different pathways. The experimental analysis suggests that the temporal order of driver gene mutation obtained by the proposed method may reflect the essential properties of cancer progression including branches, confluences and independent progressions.

In summary, PGM can deduce the causal dependencies among driver gene mutation, and reflect the dynamics of the driver mutation accumulation process during cancer progression more closely. In PGM, the intersection degree between arbitrary two genes is defined to filter some insignificant relation. Marginal probability and conditional probability can be estimated directly from the data, and the mutation waiting time of the driver genes can be estimated by stochastic function of their probability.

Our model is beneficial in the following four aspects. First, PGM use causal dependency theory to generate a DAG to better representation of tumorigenesis. Second, our model constructs DAG with the intersection degree which is used to prune the redundant edges to discover more significance selectivity relation. Third, our model use the assumption to derive driver pathway that driver genes at the same pathway is at same waiting time stage. Fourth, our model not only can captures the selectivity relation among genes but also derive driver pathway, it can also be said that our model consider the cancer progression not only at the individual gene level but also at the pathway level. In general, PGM can recapitulates current knowledge, also offering new insights on the order constrains among pathways in cancer progression.

Compared with the existing methods and models, our model considers the selectivity relation among driver genes, so it can discover more temporal orders among driver genes. For example, in Fig. 11, the mutation of RB1 is the earlier event than PIK3CB, ATRX, PIK3C2G, FGFR2 and PIK3CG, and RB1 is divided into Stage 2, the mutation of PIK3CB, ATRX, PIK3C2G, and PIK3CG are divided into Stage 3, the mutation of FGFR2 is divided into Stage 4 with the correct rate greater than 85% by using bootstrap re-sampling. However, to our best knowledge there is not relevant medical report to verify some of our novel hypotheses which allow researchers to be interrogated and tested with clinical and biochemical trials. Such as we discover that exist likely selectivity relation of mutation on the pair of NRAS and ELF3 and the pair of ACVR2A and SOX9 in colorectal cancer.

The presented work has demonstrated that it is feasible using cross-sectional mutation data to study cancer progression, but there are some potential unresolved problems.

Our future work will focus on incorporating currently available molecular data (e.g., mRNA, microRNA, copy number, methylation, clinical data and gene expression data) to construct more complex and precise cancer progression DAG. Furthermore, the number of progression stages in our model need to be predefined, however, usually it is difficult to determine the value, automatically identifying them will also be a logical expansion.

REFERENCES

[1] L. D. Wood *et al.*, "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, no. 5853, pp. 1108–1113, Nov. 2007.

[2] M. S. Lawrence *et al.*, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214–218, Jul. 2013.

[3] X. Hua, H. Xu, Y. Yang, J. Zhu, P. Liu, and Y. Lu, "DrGaP: A powerful tool for identifying driver genes and pathways in cancer sequencing studies," *Amer. J. Hum. Genet.*, vol. 93, no. 3, pp. 439–451, Sep. 2013.

[4] S. Nambara *et al.*, "Omics approach to identify driver genes for peritoneal dissemination of gastric cancer cells," *Cancer Res.*, vol. 75, p. 5169, Aug. 2015.

[5] X. D. Dai, Y. Gong, M. D. Galsky, E. E. Schadt, W. K. Oh, and J. Zhu, "Inferred regulatory interaction network from prostate cancer reveals potential regulators coordinating progression and metastasis," *Cancer Res.*, vol. 74, no. 19, p. 368, Oct. 2014.

[6] C. B. Diep *et al.*, "The order of genetic events associated with colorectal cancer progression inferred from meta-analysis of copy number changes," *Genes Chromosomes Cancer*, vol. 45, no. 1, pp. 31–41, Jan. 2006.

[7] W. J. Kent *et al.*, "The human genome browser at UCSC," *Genome Res.*, vol. 12, no. 6, pp. 996–1006, Jun. 2002.

[8] W. Zhang and S. Wang, "An integrated framework for identifying mutated driver pathway and cancer progression," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, p. 1, Dec. 2017.

[9] L. O. Loohuis *et al.*, "Inferring tree causal models of cancer progression with probability raising," *PLoS ONE*, vol. 9, no. 10, p. e108358, Oct. 2014.

[10] B. J. Raphael and F. Vandin, "Simultaneous inference of cancer pathways and tumor progression from cross-sectional mutation data," *J. Comput. Biol.*, vol. 22, no. 6, pp. 510–527, Jun. 2015.

[11] J. Y. Shi *et al.*, "Inferring the progression of multifocal liver cancer from spatial and temporal genomic heterogeneity," *Oncotarget*, vol. 7, no. 3, pp. 2867–2877, Jan. 2016.

[12] B. Vogelstein *et al.*, "Genetic alterations during colorectal-tumor development," *New England J. Med.*, vol. 319, no. 9, pp. 525–532, Sep. 1988.

[13] P. Ding, J. Luo, Q. Xiao, and X. Chen, "A path-based measurement for human miRNA functional similarities using miRNA-disease associations," *Sci. Rep.*, vol. 6, Sep. 2016, Art. no. 32533.

[14] J. Luo and J. Wu, "A new algorithm for essential proteins identification based on the integration of protein complex co-expression information and edge clustering coefficient," *Int. J. Data Mining Bioinf.*, vol. 12, no. 3, pp. 257–274, 2015.

[15] R. Desper, F. Jiang, O. P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer, "Distance-based reconstruction of tree models for oncogenesis," *J. Comput. Biol.*, vol. 7, no. 6, pp. 789–803, 2000.

[16] N. Beerenwinkel, J. Rahnenführer, R. Kaiser, D. Hoffmann, J. Selbig, and T. Lengauer, "Mtreemix: A software package for learning and using mixture models of mutagenetic trees," *Bioinformatics*, vol. 21, no. 9, pp. 2106–2107, May 2005.

[17] A. Szabo and K. Boucher, "Estimating an oncogenetic tree when false negatives and positives are present," *Math. Biosci.*, vol. 176, no. 2, pp. 219–236, Apr. 2002.

[18] M. Gerstung, N. Eriksson, J. Lin, B. Vogelstein, and N. Beerenwinkel, "The temporal order of genetic and pathway alterations in tumorigenesis," *PLoS ONE*, vol. 6, no. 11, p. e27136, Nov. 2011.

[19] M. D. Radmacher *et al.*, "Graph models of oncogenesis with an application to melanoma," *J. Theor. Biol.*, vol. 212, no. 4, pp. 535–548, Oct. 2001.

[20] C. Pei, S. L. Wang, J. Fang, and W. Zhang, "GSMC: Combining parallel Gibbs sampling with maximal cliques for hunting DNA motif," *J. Comput. Biol.*, vol. 24, no. 12, pp. 1243–1253, 2017.

[21] J. Luo, Q. Xiao, C. Liang, and P. Ding, "Predicting MicroRNA-disease associations using kronecker regularized least squares based on heterogeneous omics data," *IEEE Access*, vol. 5, pp. 2503–2513, 2017.

[22] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive Bayesian networks," *Bioinformatics*, vol. 25, no. 21, pp. 2809–2815, Nov. 2009.

[23] F. Michor, Y. Iwasa, and M. A. Nowak, "Dynamics of cancer progression," *Nature Rev. Cancer*, vol. 4, no. 3, pp. 197–205, Mar. 2004.

[24] C. S. Attolini *et al.*, "A mathematical framework to determine the temporal sequence of somatic genetic events in cancer," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 41, pp. 17604–17609, Oct. 2010.

[25] Y. K. Cheng, R. Beroukhim, R. L. Levine, I. K. Mellinghoff, E. C. Holland, and F. Michor, "A mathematical methodology for determining the temporal order of pathway alterations arising during gliomagenesis," *PLoS Comput. Biol.*, vol. 8, no. 1, p. e1002337, Jan. 2012.

[26] H. S. Farahani and J. Lagergren, "Learning oncogenetic networks by reducing to mixed integer linear programming," *PLoS ONE*, vol. 8, no. 6, p. e65773, Jun. 2013.

[27] D. Ramazzotti *et al.*, "CAPRI: Efficient inference of cancer progression models from cross-sectional data," *Bioinformatics*, vol. 31, no. 18, pp. 3016–3026, Sep. 2015.

[28] H. Wu, L. Gao, and N. K. Kasabov, "Network-based method for inferring cancer progression at the pathway level from cross-sectional mutation data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 6, pp. 1036–1044, Nov. 2016.

[29] S. Cristea, J. Kuipers, and N. Beerenwinkel, "pathTiMEx: Joint inference of mutually exclusive cancer pathways and their progression dynamics," *J. Comput. Biol.*, vol. 24, no. 6, pp. 603–615, Dec. 2016.

[30] J. L. Fleck, A. B. Pavel, and C. G. Cassandras, "Integrating mutation and gene expression cross-sectional data to infer cancer progression," *BMC Syst. Biol.*, vol. 10, p. 12, Jan. 2016.

[31] P. Suppes, "A probabilistic theory of causality," *J. Amer. Stat. Assoc.*, vol. 67, no. 337, 1970.

[32] L. Glynn, "A probabilistic analysis of causation," *Brit. J. Philosophy Sci.*, vol. 62, no. 2, pp. 343–392, Jun. 2011.

[33] E. Horvitz, "Probability, causality, and intelligence," *IEEE Intell. Syst.*, vol. 26, p. 14, Jul./Aug. 2011.

[34] A. V. Aho, M. R. Garey, and J. D. Ullman, "The transitive reduction of a directed graph," *SIAM J. Comput.*, vol. 1, no. 2, pp. 131–137, 1972.

[35] A. Linden, J. L. Adams, and N. Roberts, "Evaluating disease management program effectiveness—An introduction to the bootstrap technique," *Disease Manage. Health Outcomes*, vol. 13, no. 3, pp. 159–167 2005.

[36] E. R. Fearon and B. Vogelstein, "A genetic model for colorectal tumorigenesis," *Cell*, vol. 61, no. 5, pp. 759–767, Jun. 1990.

[37] E. R. Fearon, "Molecular genetics of colorectal cancer," *Annu. Rev. Pathol. Mech. Disease*, vol. 6, pp. 479–507, Nov. 2011.

[38] S. Cristea, J. Kuipers, and N. Beerenwinkel, "pathTiMEx: Joint inference of mutually exclusive cancer pathways and their dependencies in tumor progression," in *Proc. Int. Conf. Res. Comput. Mol. Biol.*, 2016, pp. 65–82.

[39] A. Dallol *et al.*, "Clinical significance of frequent somatic mutations detected by high-throughput targeted sequencing in archived colorectal cancer samples," *J. Trans. Med.*, vol. 14, no. 1, p. 118, 2016.

[40] D. Chisanga *et al.*, "Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D969–D974, Jan. 2016.

[41] C. W. Brennan *et al.*, "The somatic genomic landscape of glioblastoma," *Cell*, vol. 155, no. 2, pp. 462–477, Oct. 2013.

[42] J. E. DeClue *et al.*, "Epidermal growth factor receptor expression in neurofibromatosis type 1âĂŞrelated tumors and NF1 animal models," *J. Clin. Invest.*, vol. 105, no. 9, pp. 1233–1241, May 2000.

[43] N. Misra, E. Szczurek, and M. Vingron, "Inferring the paths of somatic evolution in cancer," *Bioinformatics*, vol. 30, no. 17, pp. 2456–2463, Sep. 2014.

[44] X.-Y. Liu *et al.*, "Frequent *ATRX* mutations and loss of expression in adult diffuse astrocytic tumors carrying *IDH1/IDH2* and *TP53* mutations," *Acta Neuropathol.*, vol. 124, no. 5, pp. 615–625, Nov. 2012.

[45] S. N. Quayle *et al.*, "Somatic mutations of *PIK3R1* promote gliomagenesis," *PLoS One*, vol. 7, no. 11, p. e49466, 2012.

[46] R. McLendon *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061–1068, 2008.

[47] R. M. Montgomery, L. de Souza Queiroz, and F. Rogerio, "EGFR, p53, IDH-1 and MDM2 immunohistochemical analysis in glioblastoma: Therapeutic and prognostic correlation," *Arquivos Neuro-Psiquiatria*, vol. 73, no. 7, pp. 561–568, Jul. 2015.

WEI ZHANG received the B.Sc. degree in computer science from the Kunming University of Science and Technology in 2010 and the M.Sc. degree in computer science from Hunan University in 2013, where she is currently pursuing the Ph.D. degree with the College of Computer Science and Electronic Engineering. Her research interests include driver pathways, cancer progression, and bioinformatics and its applications.

SHU-LIN WANG received the B.Sc. degree in computer application from the China University of Geosciences, and the M.Sc. degree in computer application and the Ph.D. degree in computer science and technology from the National University of Defense Technology, China. He was a Post-Doctoral Researcher with the Department of Application of Bioinformatics, The University of Kansas, from 2012 to 2013. He is currently a Professor with the College of Computer Science and Electronics Engineering, Hunan University, China. His researcher has been funded by NSFC. He has over 60 publications in professional journals and conferences. His research interests include big data analysis, bioinformatics, software engineering, artificial intelligence, and complex systems.