Network-based method for inferring cancer progression at the pathway level from cross-sectional mutation data

Hao Wu, Lin Gao, Nikola Kasabov

Abstract—Large-scale cancer genomics projects are providing a wealth of somatic mutation data from a large number of cancer patients. However, it is difficult to obtain several samples with a temporal order from one patient in evaluating the cancer progression. Therefore, one of the most challenging problems arising from the data is to infer the temporal order of mutations across many patients. To solve the problem efficiently, we present a **Net**work-based method (NetInf) to **Inf**er cancer progression at the pathway level from cross-sectional data across many patients, leveraging on the exclusive property of driver mutations within a pathway and the property of linear progression between pathways. To assess the robustness of NetInf, we apply it on simulated data with the addition of different levels of noise. To verify the performance of NetInf, we apply it to analyze somatic mutation data from three real cancer studies with large number of samples. Experimental results reveal that the pathways detected by NetInf show significant enrichment. Our method reduces computational complexity by constructing gene networks without assigning the number of pathways, which also provides new insights on the temporal order of somatic mutations at the pathway level rather than at the gene level.

Index Terms—Cancer genome, cancer progression, driver mutation, driver pathways, complex network

1 INTRODUCTION

Cancer has become one of the most serious threats to human health. Cancer is driven mainly by somatic mutations, including small indels, large copy number aberrations, single nucleotide substitution, and structural aberrations that accumulate during the lifetime of an individual [1], [2], [3]. A large number of somatic mutations have been already identified in the genomes. In recent years, high-throughput DNA sequencing technologies are measuring somatic mutations in many cancer genomes as part of large projects, such as International Cancer Genome Consortium (ICGC) [4], The Cancer Genome Atlas (TCGA) [5] and so on. According to the analysis of somatic mutations in cancer genomes, three important problems appear. First, how to distinguish driver mutations, which contribute to tumorigenesis, from passenger mutations, which are merely random, functionally neutral and have no consequence for cancer [6], [7], [8], [9], [10], [11]. Second, how to detect driver pathways, which are frequently perturbed with a large number of tumor cells, and give rise to the product of tumorigenic properties, such as cell angiogenesis, proliferation or metastasis [1], [2], [3], [12], [13], [14], [15], [16], [17]. Third, how to determine temporal orders of the driver mutations in cancer patients [18], [19], [20], [21], [22]. The first question can usually be solved by comparing mutation frequencies across different individuals [6], [7], [8], [9], [10], [11]. Several methods have been developed to address the second question, based on two properties - high coverage and high exclusivity of the driver pathways [1], [2], [3], [13], [16]. However, it is almost impossible to obtain samples at multiple time-points from a single individual, therefore, it is difficult to answer the question about temporal progression and identify what mutations occur early in cancer progression [18], [19], [20], [21], [22]. One systematic approach to address the task is to identify mutually exclusive gene sets in cancer genomic data [1], [2], [3], [12], [13], [14], [15], [16], [23]. In mutually exclusive patterns, the mutations tend to occur in different patients. Such mutually exclusive gene sets have been identified in cancer data and found to be associated with synthetic lethality or functional pathways [2], [3], [12], [13]. Therefore, it is important to identify the mutually

- *H. Wu and L. Gao are with School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China. E-mail: haowu@nwsuaf.edu.cn, lgao@mail.xidian.edu.cn.*
- *N. Kasabov is with Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland, New Zealand E-mail：nkasabov@aut.ac.nz.*

Manuscript received; revised; accepted; published. For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB. Digital Object Identifier no.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

exclusive patterns for a basic understanding of cancer progression and targeted treatment [23]. Taking into account cancer phylogeny, we aim at identifying the mutually exclusive gene sets in the process of cancer progression that can help us develop new diagnostics or therapeutics targeted to specific subtypes of progression [24].

Several methods have been introduced to infer temporal progression of gene mutations from cross-sectional data [18], [19], [20], [25], [26], [27], [28]. Desper et al. [25], [26] proposed a tree model inference algorithm based on the thought of maximum-weight which relates cancer progression to measurement on gains and losses of chromosomal regions in tumor cells. Moritz et al. [27] presented a Bayesian network model to quantify cancer progression by an unobservable accumulation process which is separate from the observable mutations. However, these methods infer temporal ordering at the level of individual mutations or genes. The problem with these approaches is that cancers usually exhibit mutational heterogeneity, since clinically and histologically identical cancers often have few mutated genes in common. Therefore, Moritz et al. [18] presented a probabilistic graphical model to estimate temporal pathways during cancer progression from cross-sectional mutation data, and provided a quantitative and intuitive tumorigenesis model showing that genetic events may be related to the phenotypic progression at the pathway level, since somatic mutations, especially those oncogenic driver mutations, perturb all kinds of metabolic, signaling and regulatory pathways. Therefore, different individuals might hold driver mutations in different genes within the same pathway. Recently, many researches [1], [2], [3], [12], [13], [14], [15], [16], [23] have indicated that driver mutations in the same pathway tend to be mutually exclusive, that is, most patients have no more than one mutation within the same pathway. Therefore, Vandin et al. [20] introduced the exclusivity among mutations (genes) within the same pathway to infer cancer pathways and tumor progression from cross-sectional mutation data. They formulated the Pathway Linear Progression problem as an integer linear program. In the formulation, any partition has to satisfy two requirements: the exclusivity of mutations within each gene set, and the progression across the sets. Therefore, the Pathway Linear Progression Reconstruction problem is NP-hard to identify the best partition by simultaneously considering both exclusivity and progression.

To reduce the computational complexity and solve the NP-hard problem of the Pathway Linear Progression Model in an efficient approach, we now introduce a new network-based method to infer cancer progression at the pathway level from cross-sectional mutation data. During construction of a gene network, we introduce the definition of exclusive degree to describe how much exclusive between each pair of genes, and to take into account the coverage overlap, coverage and weight, we define weight degree to describe the ratio between weight and coverage. In the constructed gene networks, mutations of all genes in each complete subnetwork are approximately exclusive. Therefore, we just need to find a set of non-overlapping complete subnetworks which meet the linear progression between them. The specific steps of the approach are as follows. In the first step we filter the mutation matrix and obtain the critical genes which have been reported in the previous research or have a high frequency of recurrence. In the second step, a gene network is constructed by calculating the exclusive score between each pair of genes. In the network, each node is a gene and the edge between a pair of nodes will be created if the exclusive score between the pair of genes is greater than or equal to a threshold λ. In the third step, we identify all complete subnetworks and sort them from large to small according to their coverage degree. Then we use an orderly iterative method to find the driver pathways which meet the requirement for a linear progression between them.

2 METHODS

2.1 Exclusivity and progression

Vandin et al. [20] introduced Pathways Linear Progression Model to infer cancer pathways and tumor progression with two criteria from cross-sectional somatic mutation data. The first one is "exclusivity" which means most patients have no more than one mutation in a pathway. The second one is "progression" which means the patients with gene mutations in a pathway have certainly gene mutations in the previous pathway. Given a binary mutation matrix M with m rows (samples $s_1, s_2, \dots, s_m$) and n columns (genes $g_1, g_2, \dots, g_n$), where $M_{i,j} = 1$ if $g_j$is mutated in sample s_i, and $M_{i,j} = 0$ otherwise. For a gene g, the coverage $\Gamma(g) = \{i: M_{i,g} = 1\}$ represents the set of patients in which gene g is mutated (Fig. 1). Similarly, for a sub-matrix G of size $m \times k$ in the mutation matrix M, the coverage is denoted as $\Gamma(G) = \cup_{g \in G} \Gamma(g)$. For any pair of $g_j, g_k \in G, g_j \neq g_k$, if $\Gamma(g_j) \cap \Gamma(g_k) = \emptyset$, G is

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

mutually exclusive.

Pathways Linear Progression Model (PLPM) [20]. A mutation matrix M of size $m \times n$ satisfies the Pathways Linear Progression Model $PLPM(K)$ with parameter $K > 1$, if there is a partition $P = \{P_1, P_2, \dots, P_K\}$ of all the columns of M into K sets such that:

1. For each row s_i of M, if $\left|\{g_j \in P_k : M_{i,j} = 1\}\right| \leq 1$, then among all the rows within one set P_k are mutually exclusive, that is, for each pair of genes $g_{j_1}, g_{j_2} \in P_k, 1 \leq j_1, j_2 \leq n$ and $j_1 \neq j_2$, if $\Gamma(g_{j_1}) \cap \Gamma(g_{j_2}) = \emptyset$, among all the rows within one set P_k are mutually exclusive.
2. For all $1 < k \leq K$, if $\Gamma(P_k) \subseteq \Gamma(P_{k-1})$, then each row s_i of M satisfies the progression on the sets $P_1, \dots, P_K$, that is, for all $1 < k \leq K$, if $\left|\{g_j \in P_k : M_{i,j} = 1\}\right| > 0$, then $\left|\{g_j \in P_{k-1} : M_{i,j} = 1\}\right| > 0$.

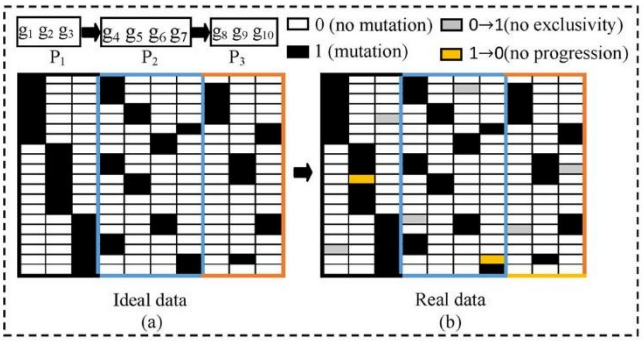


Fig. 1. Pathways linear progression model [20]. (a) A linear progression model on gene sets creates a mutation matrix with mutually exclusive mutations within each gene set, and a progression of mutations across the gene sets; (b) In real dataset, errors that disrupt the exclusivity and the progression are present.

For a sub-matrix G of size $m \times k$ in the mutation matrix M, the exclusive degree function is denoted as:

$$ED(G) = \frac{|\Gamma(G)|}{\sum_{g \in G} |\Gamma(g)|}. \quad (1)$$

For a pair of genes g_j, g_k, their exclusive degree function is denoted as:

$$ED(g_j, g_k) = \frac{|\Gamma(g_j) \cup \Gamma(g_k)|}{|\Gamma(g_j)| + |\Gamma(g_k)|}. \quad (2)$$

According to the above formula, $ED(G) = 1$ when G is mutually exclusive. That is, each row of G contains at most one mutation.

For a sub-matrix G of size $m \times k$ in the mutation matrix M, the coverage overlap [23] is denoted as:

$$\omega(G) = \sum_{g \in G} |\Gamma(g)| - |\Gamma(G)|. \quad (3)$$

For a pair of genes g_j, g_k, their coverage overlap is denoted as:

$$\omega(g_j, g_k) = |\Gamma(g_j)| + |\Gamma(g_k)| - |\Gamma(g_j) \cup \Gamma(g_k)|. \quad (4)$$

Considering both the coverage overlap $\omega(g_j, g_k)$ and the two coverages $\Gamma(g_j)$ and $\Gamma(g_k)$, the weight degree function is denoted as:

$$WD(g_j, g_k) = 1 - \frac{|\Gamma(g_j) \cap \Gamma(g_k)|}{min\{|\Gamma(g_j)|, |\Gamma(g_k)|\}}. \quad (5)$$

For a sub-matrix G of size $m \times k$ in the mutation matrix M, the coverage degree function is denoted as:

$$CD(G) = \frac{|\Gamma(G)|}{m}. \quad (6)$$

For a pair of genes g_j, g_k, their coverage degree function is denoted as:

$$CD(g_j, g_k) = \frac{|\Gamma(g_j) \cup \Gamma(g_k)|}{m}. \quad (7)$$

Note that $CD(G) = 1$, when G is the complete coverage. That is, each row of G contains at least one mutation.

For two sub-matrices M_j, M_k with $CD(M_j) > CD(M_k)$, the progression ratio of them is denoted as:

$$PR(M_j, M_k) = \frac{|\Gamma(M_j) \cap \Gamma(M_k)|}{|\Gamma(M_k)|}. \quad (8)$$

Note that $PR(M_j, M_k) = 1$ when mutations of all genes in M_k are a subset of mutations of all genes in M_j.

2.2 The proposed NetInf method

The proposed NetInf method consists of the following procedures and computational steps.

2.2.1 Constructing a gene network based on approximate exclusivity

Vandin et al. [20] introduced Pathway Linear Progression Model (PLPM) which was defined for an integer $K > 1$ as an integer linear program problem of looking for $\mathcal{P}^* = argmin_{P \in \mathcal{P}(K)} f(M, \mathcal{P})$, and showed that the problem is an NP-hard problem. To solve it more efficiently, we construct a weighted gene network based on exclusive degree between each pair of genes to simplify the relationships between the genes and to significantly reduce the computational complexity. First, we calculate the exclusive degree between each pair of genes in a mutation matrix by using formula (2). Second, we construct a weighted gene network in which each node is a gene and the weight of an edge is the exclusive degree of the two connected genes. In the process of constructing a gene network, for each pair of genes g_j, g_k, if $ED(g_j, g_k) \geq \lambda$ and $WD(g_j, g_k) \geq \gamma$, an edge will be created to link this pair of genes, otherwise, there is no an

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

edge between the pair of genes. The process is shown in Fig. 2.

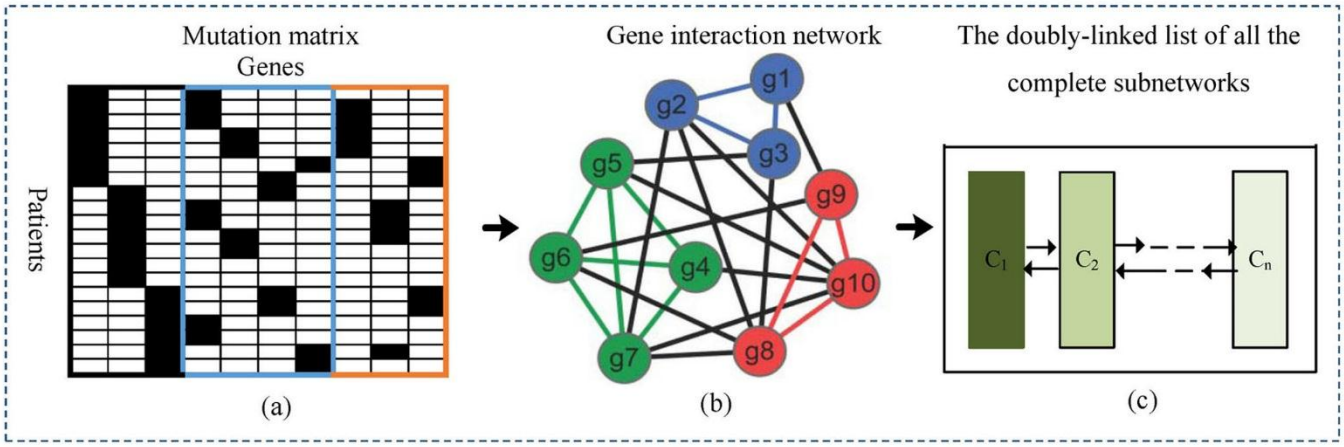


Fig. 2. An illustration of the steps in the process of inferring cancer progression at the pathway level. (a) A mutation matrix is created using somatic mutation data from multiple patients; (b) The exclusive degree between each pair of genes is calculated and a gene interaction network is constructed according to the exclusive degrees. If the exclusive degree between a pair of genes is greater than or equal to λ, an edge is created between the two genes and the exclusive degree is represented as its weight; (c) All complete subnetworks are detected and sorted from large to small according to their coverage degree.

2.2.2 Detecting pathways which meet the requirement for a linear progression

In a gene network, according to the process of construction in the previous step, mutations of all genes in each complete subnetwork are approximately exclusive. Therefore, we need to find the non-overlapping complete subnetworks which meet the definition of PLPM. Firstly, we find all gene sets in which each gene set can constitute a complete subnetwork in the gene network. Secondly, we sort the gene sets found in the previous step from large to small according to their coverage degree and create a doubly-linked list. Finally, we identify gene sets which meet the linear progression from the doubly-linked list.

The step-by-step description of the algorithm for identifying gene sets which meet the linear progression starting from the first gene set M_0 in the doubly-linked list is as follows.

Step 1: Create a null doubly-linked list N, and set the step number $s = 0$ for doubly-linked list M and $t = 0$ for doubly-linked list N.

Step 2: Let $N_t = M_s$.

Step 3: If there is no intersection between genes in M_{s+1} and N, calculate progression ratio between M_{s+1} and N_t using formula (8). Otherwise, $s = s + 1$ and continue Step 3.

Step 4: If $PR(N_t, M_{s+1}) \geq \delta$, $N_{t+1} = M_{s+1}, t = t + 1$ and $s = s + 1$. Otherwise, $s = s + 1$ and return to Step 3.

Step 5: If the number of genes in doubly-linked list N is less than the number of all genes and there is no end of the doubly-linked list M, return to Step 3. Otherwise, go to Step 6.

Step 6: If the number of genes in doubly-linked list N is less than the total number of genes, delete node $N_t, t = t - 1, s = s + 1$ and return Step 2. Otherwise, output the gene sets in doubly-linked list N and end the process.

We can identify the gene sets in which mutations within each gene set are approximately exclusive and mutations across them meet the linear progression.

2.3 Parameter settings

In the NetInf method, a threshold λ is applied to decide whether there exists an edge between each pair of genes according to their exclusive degree. A threshold γ is applied to describe the ratio between non-overlap (weight) and coverage. Another threshold δ is used to determine whether there exists a linear progression between two pathways. If $\lambda = 1, \gamma = 1$ and $\delta = 1$, this is an ideal case for the gene sets to satisfy the Pathways Linear Progression Model. For a real mutation data, there always exist errors which disrupt the exclusivity or progression. Therefore, λ, γ and δ are usually less than 1. In the process of constructing a gene network, we set $\lambda = 0.95$ as reported in [1]. In order to avoid the case where two connected genes have a high exclusive degree

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

but a low weight degree, we create formula (5) and analyze different weight degrees of two genes with the same coverage, coverage overlap and exclusive degree (Fig. 3).

Although the coverage, coverage overlap and exclusive degree of the two genes g_j, g_k in the four cases in Fig. 3 are the same, the two genes in Fig. 3a&b are usually regarded as exclusive, while the two genes in Fig. 3c&d are not regarded as exclusive [1]. We obtain ideal results on simulated data and biological data when we attempt to set $\gamma = 0.8$ and δ as adjustable value, that is, Fig. 3b is regarded as a boundary instance.

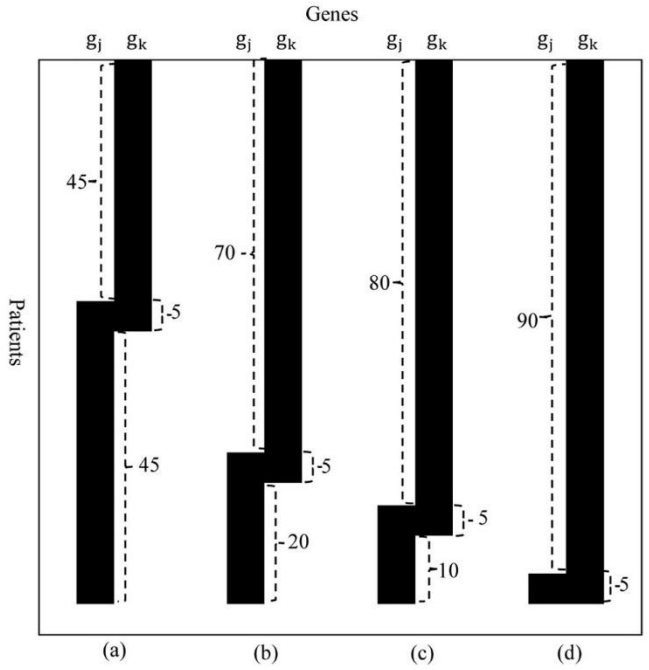


Fig. 3. Analysis of the weight degrees in two genes with the same coverage, coverage overlap and exclusive degree. The numbers in the figure stand for coverage in different cases. The coverage of the two genes in the four cases is $|\Gamma(g_j, g_k)| = 95$; the coverage overlap of the two genes in the four cases is $\omega(g_j, g_k) = 5$; the exclusive degree of the two genes in the four cases is $ED(g_j, g_k) = 0.95$. But the weight degrees are different in the four cases, (a) $WD(g_j, g_k) = 0.9$; (b) $WD(g_j, g_k) = 0.8$; (c) $WD(g_j, g_k) = 0.67$; (d) $WD(g_j, g_k) = 0$.

In the experiment on simulated data, when noise probability p is set to be 0.001 and δ is set to be 0.95, we obtain exact results from 12 runs out of 20; when noise probability p is set to be 0.001 and δ is set to be 0.92, we obtain exact results from 20 runs; when noise probability p is set to be 0.05 and δ is set to be 0.9, we obtain the results from 6 runs out of 20; when noise probability p is set to be 0.05 and δ is set to be 0.85, we obtain the results from 18 runs out of 20. The results show when δ is set to be 0.95, we can obtain ideal results only if the progression model has very low noise; when δ is set to be 0.85, we can obtain ideal results even If noise probability p is relatively high. Given the close correlation between the δ value and the noise level, we set an adjustable value of $0.85 \leq \delta \leq 0.95$. Therefore, we set $\lambda = 0.95, \gamma = 0.8$ and an adjustable value of $0.85 \leq \delta \leq 0.95$ which yield ideal results in conducting the experiments.

3 Results

To assess the robustness of the proposed NetInf method, we apply it on simulated data with the addition of different levels of noise [20]. When executing the method on a conventional computer, NetInf can obtain ideal results. To verify the performance of NetInf, we apply it on three biological datasets (Table 1) and compare the results with ILP. The detailed comparison is elaborated below.

3.1. Simulated data

We perform a large number of experiments on simulated data with different levels of noise. We generate mutation data according to a progression model $\mathcal{P}$ to which noise is added. First, we consider a progression model with $k = 4$, $k = 5$ stages, each containing $n = 4, n = 5$ genes, respectively, and generate 20 mutation data with m samples, adding noise with different probabilities p to the corresponding mutation data. We consider values of $m = 50, 100, 200, 400, 600, 800, 1000$ and $p = 0.001, 0.01, 0.05$. For each combination m, p, we record the correct ratio which is the ratio between the number of genes belonging to corresponding sets and the total number of genes (Fig. 4).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

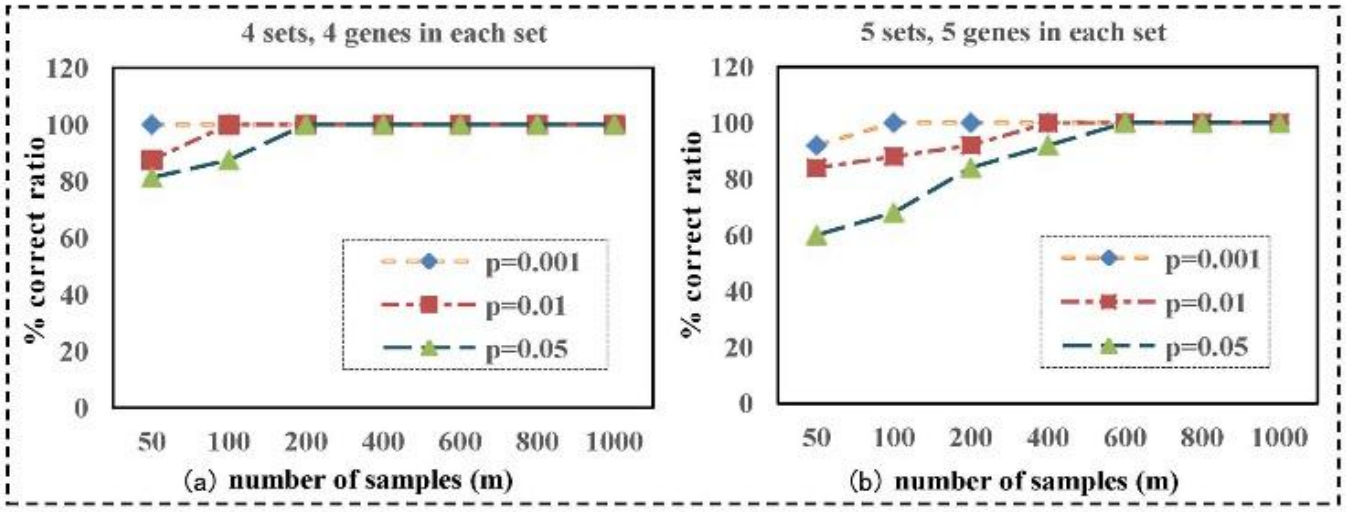


Fig. 4. Correct ratio for different number of samples and different probabilities of noise addition. Correct ratio is shown on the graph on m samples, where the mutation matrix M is based on a linear progression model with k sets, each containing n genes. Noise is added to the matrix M with a probability p. (a) Results for $k = 4, n = 4$ and different values of p and m; (b) Results for $k = 5, n = 5$ and different values of p and m.

As we can see, the correct ratio is very high when the number of samples increases and the probability of error decreases. In Fig. 4a, we can obtain absolutely right results when the number of samples is not less than 200 or the probability of error $p = 0.001$. In fact, if the probability of error is not very high or the number of samples is not very small, we can get ideal results for $k = 4$ and $n = 4$. In Fig. 4b, we can obtain absolutely right results when the number of samples is not less than 600 or the probability of error $p = 0.001$. Actually, if the probability of error is less than 0.01 or the number of samples is more than 200, we can get ideal results for $k = 5$ and $n = 5$. These results show that data from the reasonable number of cancer samples can be used to infer the correct progression model. If the noise level is relatively high, the number of cancer patients is required to be large to infer the correct progression model.

3.2 Real data

To assess the performance of our NetInf on real biological data, we analyze three somatic mutation data from published cancer studies. In table 1, we present the information about number of genes, number of samples, maximum mutation frequency for all genes, average of mutation frequency for all genes and average mutation number of each sample.

TABLE 1

Biological Datasets Used in This Study

Cancer type	#Patient	#Gene	MMF	AMF	AMN
CRC 1	94	8	78	29.9	2.52
CRC 2	223	14	165	44.1	2.77
GBM	290	27	90	17.7	1.65

In the table, #Patient: number of patients; #Gene: number of genes; MMF: maximum mutation frequency for all genes; AMF: average of mutation frequency for all genes; AMN: average mutation number of each sample; CRC 1: Colorectal cancer data reported in [29]; CRC 2: Colorectal cancer data from TCGA [5]; GBM: glioblastoma multiforme data from TCGA [5].

3.2.1 SCIENCEMAG: Colorectal cancer data

We first apply NetInf to colorectal cancer data reported in [29]. The data contains 94 samples and eight genes for which mutation frequency is over 5%. They are TP53, KRAS, EVC2, APC, EPHA3, FBXW7, PIK3CA and TCF7L2. The progression model inferred with the use of the ILP method is shown in Fig. 5a, and it shares close similarities with the model inferred with the use of the proposed NetInf method (Fig. 5b).

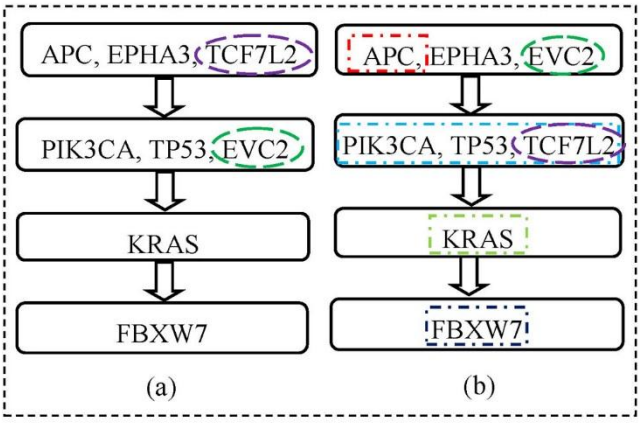


Fig. 5. Progression models built with the use of colorectal cancer data. (a) Results obtained by applying the ILP method [20]; (b) Results obtained by applying the proposed here NetInf method. Dashed oval boxes show the differences between the results of ILP and NetInf. Dashed rectangular boxes in Fig. 5b&6b show the same order of the six genes appearing in the two colorectal cancer datasets.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

The only difference between the two progression models is that mutations in EVC2 occur early in the progression model inferred by NetInf, while TCF7L2 mutations appear later. These results seem to be reasonable because EVC2 mutations have been reported to be precursor node of TCF7L2 in colorectal cancer [18].

Interestingly, gene set (APC, EPHA3, EVC2) in our results has the same coverage degree but larger exclusive degree (90.9%) than gene set (APC, EPHA3, TCF7L2) in [20], showing that the gene set (APC, EPHA3, EVC2) is more likely in the same functional pathway [30]. APC and EPHA3 have stable co-expression together with a cytoplasmic form of BirA for efficient biotinylation of AP-tagged EPHA3 C-terminus [31]. However, gene set (APC, EPHA3, EVC2) shows no significant enrichment using the DAVID functional annotation tool [32]. Gene set (PIK3CA, TP53, TCF7L2) with p-value=1.1E-4 shows significant enrichment, and they are the core members of the pathways in cancer and Wnt/Notch signaling pathways. The functions of PIK3CA and TP53 are related to age at disease onset [33]. The details, including coverage degree, exclusive degree and p-value of each pathway, are presented in Table 2. We want to find the gene sets whose coverage degree and exclusive degree are simultaneously large, but it is necessary to point out that exclusive degree increases at the expense of declining coverage degree, and vice versa [1]. The analyses show that the NetInf method proposed here obtains a more accurate tumor progression model of colorectal cancer than the ILP method.

TABLE 2

Results of ILP and NetInf Methods for Colorectal Cancer Data

Gene sets	Results of ILP			Results of NetInf		
	CD	ED	P	CD	ED	P
Set 1	85.1%	88.8%	2.1E-2	85.1%	90.9%	N/A
Set 2	78.7%	90.2%	9.7E-3	78.7%	88.1%	1.1E-4
Set 3	62.8%	100%	N/A	62.8%	100%	N/A
Set 4	8.5%	100%	N/A	8.5%	100%	N/A

In the table, CD: Coverage degree; ED: Exclusive degree; P: p-value, which is obtained using the DAVID functional annotation tool ("http://david.abcc.ncifcrf.gov/summary.jsp"). The contents of Set (i) are corresponding to the pathways displayed in Fig. 5 respectively. N/A represents no significant enrichment.

3.2.2 TCGA: Colorectal cancer data

We download colorectal mutation data from TCGA study and analyze 223 samples on this type of cancer. We choose 14 genes identified as recurrent mutation by MutSigCV [34]. The progression model inferred with the use of the ILP is shown in Fig. 6a, and it shares some similarities with the one inferred with the use of the NetInf (Fig. 6b). The details, including coverage degree, exclusive degree and p-value of each pathway, are presented in Table 3.

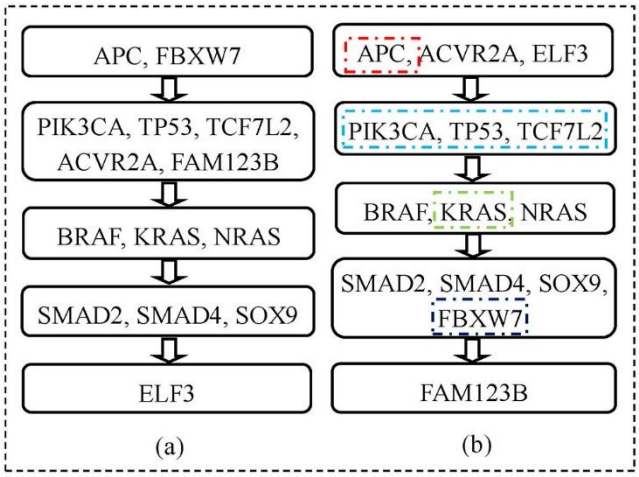


Fig. 6. Progression models built with the use of colorectal cancer data from TCGA. (a) The results from the ILP method [20]; (b) The results from the proposed here NetInf method.

In the first stage of the progression, mutations in APC always occur early in tumor progression [20]. ELF3 is a member of the E-twenty-six family of transcription factors and it drives β-catenin transactivation [35]. In the hypermutated tumors, APC and ACVR2A are frequent targets of mutation, along with most BRAF mutations, and they commonly target specific genes MIR192, MIR215 and MAPK8 [36]. The gene set is altered in 76.2% with large exclusive degree at 92.4%. In the second stage of the progression, TP53 binds to the PIK3CA promoter and inhibits its activity. Up-regulation of PIK3CA and inactivation of TP53 contribute to the pathophysiology of many human cancers [1]. SiRNA-mediated reduction in TCF7L2 activity results in increased expression of TP53, and results in increased p53 protein activity and an elevated expression of the p53 target gene Tp53inp1 [37]. PIK3CA mutations are associated with over-expression of TCF7L2 involved in the Wnt signaling pathway [38]. The same gene set has been identified in the first experiment, and PIK3CA, TP53 and TCF7L2 are core members of the pathways in cancer and Wnt/Notch signaling pathways. In the non-hypermutated tumors, the three genes are usually regarded as most frequently mutated genes [36]. The gene set is altered in 78.7% with p-value= 1.1E-4. In the third stage of the progression, the same results have been identified in [20] and BRAF, KRAS and NRAS are the core members of the RAS/RAF/MAPK signaling pathways. The three mutated genes usually have

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

oncogenic codon 12 and 13 or codon 61 mutations [36]. The gene set is altered in 59.6% with large exclusive degree at 95.7% and p-value=2.7E-5. In the fourth stage of the progression, SMAD2 and SMAD4 are the core members of the WNT signaling pathway. SMAD4 interacts with SMAD2, and SMAD2 interacts with SOX9 [20]. Mutations in FBXW7 have been reported to appear after KRAS mutations in [18], [20].

TABLE 3

Results Obtained with the Use of the ILP and NetInf Methods for Colorectal Cancer Data from TCGA

Gene sets	Results of ILP			Results of NetInf		
	CD	ED	P	CD	ED	P
Set 1	80.7%	87.8%	N/A	76.2%	92.4%	N/A
Set 2	75.3%	78.5%	6.4E-2	69.1%	87.0%	1.1E-4
Set 3	59.6%	95.7%	2.7E-5	59.6%	95.7%	2.7E-5
Set 4	21.5%	94.1%	1.4E-2	34.5%	84.6%	2.9E-2
Set 5	3.6%	100%	N/A	11.7%	100%	N/A

The contents of Set (i) are corresponding to the pathways displayed in Fig. 6 respectively.

Interestingly, six genes APC, TP53, PIK3CA, TCF7L2, KRAS and FBXW7 in the dataset also appear in the first experiment. Moreover, we find that these genes have the same assignments in different stages of the two progression models, that is, mutations in APC occur in the first stage, mutations in TP53, PIK3CA and TCF7L2 occur in the second stage, mutations in KRAS occur in the third stage and mutations in FBXW7 occur in the fourth stage, and we obtain the same gene set (TP53, PIK3CA and TCF7L2) located in the second stage in the two progression models (Fig. 5b&6b). From the results, we can find the exclusive degree of the gene sets in NetInf is slightly higher than that of gene sets in ILP, and p-value of the gene sets in NetInf is slightly smaller than that of gene sets in ILP, so the gene sets in NetInf indicate more significant enrichment than that in ILP. The analyses show that the NetInf method proposed here obtains a more accurate tumor progression model of colorectal cancer based on the used data than the ILP method.

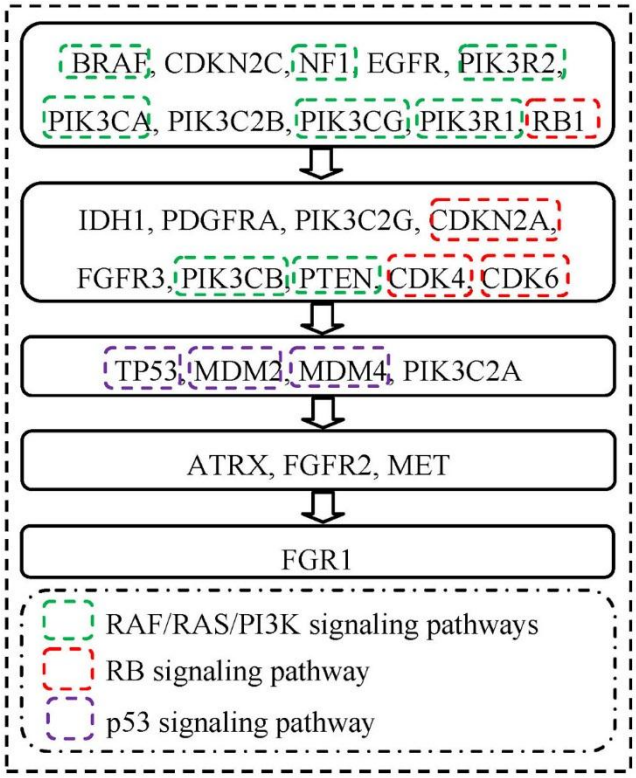


Fig. 7. Progression models built with the use of glioblastoma multiforme data from TCGA [39]. Dashed boxes identify genes in the same signaling pathway, with different colors used to denote different signaling pathways.

3.2.3 TCGA: Glioblastoma Multiforme data

We download glioblastoma multiforme data from the TCGA study and analyze 290 samples of this cancer type. We restrict the analysis to the 27 genes reported in [39] as a part of pathway alterations in GBM. The progression model inferred with NetInf is shown in Fig. 7, and the details, including coverage degree, exclusive degree and p-value of each pathway, are shown in Table 4.

TABLE 4

Results Obtained with the Use of the ILP and NetInf Methods for Glioblastoma Multiforme Cancer Data from TCGA

Gene sets	CD	ED	P
Set 1	61.0%	76.3%	9.8E-11
Set 2	43.8%	93.4%	1.2E-8
Set 3	29.7%	97.7%	4.5E-4
Set 4	7.50%	100%	3.6E-2
Set 5	0.30%	100%	N/A

The contents of Set (i) are corresponding to the pathways displayed in Fig. 7 respectively.

In the first stage of the progression, PIK3CA, PIK3CG, PIK3R1, PIK3R2, NF1 and BRAF are the core members of the RAF/RAS/PI3K signaling pathways. The gene set is altered in 61.0% with p-value=9.8E-11. In the second stage

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics

of the progression, PIK3CB and PTEN are the members of the RAF/RAS/PI3K signaling pathways, and CDKN2A, CDK4 and CDK6 are the core members of the RB signaling pathway. CDKN2A inhibits CDK4, CDK4 inhibits p27, and p27 inhibits CDK6 in the RB signaling pathway. The gene set is altered in 43.8% with large exclusive degree at 93.4% and p-value=1.2E-8. In the third stage of the progression, TP53, MDM2 and MDM4 are the core members of the p53 signaling pathway. MDM4 interacts with MDM2, MDM4 and MDM2 inhibit TP53, and TP53 activates MDM2 in the p53 signaling pathway. The gene set is altered in 29.7% with large exclusive degree at 97.7% and p-value=4.5E-4. For the first three sets in the progression model, most genes in one set belong to the same known signaling pathway. The analyses show that the NetInf method proposed here identifies pathway relations among genes in the different progression stages and obtains an ideal cancer progression model based on the used data.

4 Discussion and Conclusions

Inference of cancer progression at the pathway level is an essential problem in computational biology. In this study, we use a progression model in which mutations within each gene set (pathway) are approximately exclusive, and they follow a linear progression at the pathway level. The problem of reconstructing the best progression model has been proved to be NP-hard [20], so we introduce a novel method called NetInf for automatic inference of cancer progression at the pathway level from cross-sectional mutation data without any prior biological knowledge. In this algorithm, the critical genes are firstly determined from mutation matrix by meeting a certain frequency of recurrence requirement or being reported in the previous research. Secondly, a gene network is constructed according to high exclusivity of mutations between each pair of genes to solve the problem of high complexity that the previous methods encounter. Thirdly, all the complete subnetworks in the gene network are identified and sorted from large to small according to their coverage degree, and then an orderly iterative method is used to find the pathways which meet the linear progression between them. The results show that integrative analysis of cross-sectional mutation data has the potential to identify gene sets which are closely related to cancer phenotypes in the process of cancer progression. Moreover, our algorithm makes it possible to find the function-related oncogene sets at different stages of cancer progression.

Comparing with the previous algorithms of inferring cancer progression, our algorithm is beneficial in the following three aspects. First, complexity of the solution is reduced by constructing gene networks according to high exclusivity of mutations between each pair of genes. Second, our algorithm does not need to assign the number of pathways in the progression model. Third, our algorithm infers cancer progression at the level of pathways rather than individual mutations or genes. It is necessary to note that this algorithm does not use gene expression data, known pathways, gene interaction data and other biological knowledge. The method may provide a supplement to the analyses of cancer data and it will be helpful in producing hypotheses which will drive some specific biological experiments and increase understanding for cancer progression [1], [40]. Further research is anticipated for the development of new machine learning techniques specific for this task [41]. We also plan to analyze the cancer progression models as binary temporal sequences modeled and visualized with the use of spiking neural networks, where a mutation of a gene can be represented as a spike at a certain time of the progression [42].

Competing Interests

The authors declare no competing interests exist.

Author Contributions

H.W. and L.G. conceived the study; H.W. developed the theoretical results and was the lead writer of the manuscript; H.W. and N.K. developed the algorithms and performed the actual experimental data analysis; all the authors modified the manuscript and approved the final version.

Acknowledgments

This work was supported by the China Scholarship Council (CSC) under Grant No. 201306305017, the National Natural Science Foundation of China under Grant Nos. 61532014, 91130006, 61432010, 61402349, 61303118, 61303122, 61202174 and the Fundamental Research Funds for the Central Universities under Grant No. BDZ021404. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The work of HW and NK is also supported by the Knowledge Engineering and Discovery Research Institute of the Auckland University of Technology, New Zealand. We thank Jihua Dong for

her careful proofreading, Bingbo Wang, Xiaofei Yang, Songwei Jia and Feng Li for their helpful advice and discussions, and the anonymous reviewers for their valuable comments and suggestions. Parts of this paper appeared at the 17th Symposium on System Identification of the International Federation of Automatic Control (SYSID 2015).

REFERENCES

[1] H. Wu, L. Gao, F. Li, F. Song, X. Yang, and N. Kasabov, "Identifying Overlapping Mutated Driver Pathways by Constructing Gene Networks in Cancer," *BMC Bioinformatics*, vol. 16, no. Suppl 5, pp. S3, 2015.

[2] F. Vandin, E. Upfal, and B. J. Raphael, "De novo discovery of mutated driver pathways in cancer," *Genome research*, vol. 22, no. 2, pp. 375-385, 2012.

[3] J. Zhao, S. Zhang, L.-Y. Wu, and X.-S. Zhang, "Efficient methods for identifying mutated driver pathways in cancer," *Bioinformatics*, vol. 28, no. 22, pp. 2940-2947, 2012.

[4] J. Zhang, J. Baran, A. Cros, J. M. Guberman, S. Haider, J. Hsu, Y. Liang, E. Rivkin, J. Wang, and B. Whitty, "International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data," *Database*, vol. 2011, pp. bar026, 2011.

[5] R. McLendon, A. Friedman, D. Bigner, E. G. Van Meir, D. J. Brat, G. M. Mastrogianakis, J. J. Olson, T. Mikkelsen, N. Lehman, and K. Aldape, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061-1068, 2008.

[6] L. Ding, G. Getz, D. A. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, and M. B. Morgan, "Somatic mutations affect key pathways in lung adenocarcinoma," *Nature*, vol. 455, no. 7216, pp. 1069-1075, 2008.

[7] A. Youn, and R. Simon, "Identifying cancer driver genes in tumor genome sequencing studies," *Bioinformatics*, vol. 27, no. 2, pp. 175-181, 2011.

[8] Y. Chen, J. Hao, W. Jiang, T. He, X. Zhang, T. Jiang, and R. Jiang, "Identifying potential cancer driver genes by genomic data integration," *Scientific reports*, vol. 3, no. 3538, pp. srep03538, 2013.

[9] D. Tamborero, A. Gonzalez-Perez, C. Perez-Llamas, J. Deu-Pons, C. Kandoth, J. Reimand, M. S. Lawrence, G. Getz, G. D. Bader, and L. Ding, "Comprehensive identification of mutational cancer driver genes across 12 tumor types," *Scientific reports*, vol. 3, no. 2650, pp. srep02650, 2013.

[10] T. Sakoparnig, P. Fried, and N. Beerenwinkel, "Identification of Constrained Cancer Driver Genes Based on Mutation Timing," *PLoS computational biology*, vol. 11, no. 1, pp. e1004027, 2015.

[11] J. P. Hou, and J. Ma, "DawnRank: discovering personalized driver genes in cancer," *Genome Med*, vol. 6, no. 7, pp. 56, 2014.

[12] G. Ciriello, E. Cerami, C. Sander, and N. Schultz, "Mutual exclusivity analysis identifies oncogenic network modules," *Genome research*, vol. 22, no. 2, pp. 398-406, 2012.

[13] M. D. Leiserson, D. Blokh, R. Sharan, and B. J. Raphael, "Simultaneous identification of multiple driver pathways in cancer," *PLoS computational biology*, vol. 9, no. 5, pp. e1003054, 2013.

[14] J. Zhang, L.-Y. Wu, X.-S. Zhang, and S. Zhang, "Discovery of co-occurring driver pathways in cancer," *BMC bioinformatics*, vol. 15, no. 1, pp. 271, 2014.

[15] R. D. Melamed, J. Wang, A. Iavarone, and R. Rabadan, "An information theoretic method to identify combinations of genomic alterations that promote glioblastoma," *Journal of molecular cell biology*, vol. 7, no. 3, pp. 203-213, 2015.

[16] C.-H. Yeang, F. McCormick, and A. Levine, "Combinatorial patterns of somatic gene mutations in cancer," *The FASEB Journal*, vol. 22, no. 8, pp. 2605-2622, 2008.

[17] J. Zhang, S. Zhang, Y. Wang, and X.-S. Zhang, "Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data," *BMC systems biology*, vol. 7, no. Suppl 2, pp. S4, 2013.

[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, pp. e27136, 2011.

[19] S. J. Baker, and E. P. Reddy, "Understanding the temporal sequence of genetic events that lead to prostate cancer progression and metastasis," *Proceedings of the National Academy of Sciences*, vol. 110, no. 37, pp. 14819-14820, 2013.

[20] B. J. Raphael, and F. Vandin, "Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data," *Journal of Computational Biology*, vol. 22, no. 00, pp. 250-264, 2014.

[21] N. Navin, A. Krasnitz, L. Rodgers, K. Cook, J. Meth, J. Kendall, M. Riggs, Y. Eberling, J. Troge, and V. Grubor, "Inferring tumor progression from genomic heterogeneity," *Genome research*, vol. 20, no. 1, pp. 68-80, 2010.

[22] A. Ashworth, C. J. Lord, and J. S. Reis-Filho, "Genetic interactions in cancer progression and treatment," *Cell*, vol. 145, no. 1, pp. 30-38, 2011.

[23] E. Szczurek, and N. Beerenwinkel, "Modeling mutual exclusivity of cancer mutations," *PLoS computational biology*, vol. 10, no. 3, pp. e1003503, 2014.

[24] Y. Park, S. Shackney, and R. Schwartz, "Network-based inference of cancer progression from microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 200-212, 2009.

[25] R. Desper, F. Jiang, O.-P. Kallioniemi, H. Moch, C. H. Papadimitriou, and A. A. Schäffer, "Inferring tree models for oncogenesis from comparative genome hybridization data," *Journal of computational biology*, vol. 6, no. 1, pp. 37-51, 1999.

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

[27] M. Gerstung, M. Baudis, H. Moch, and N. Beerenwinkel, "Quantifying cancer progression with conjunctive Bayesian networks," *Bioinformatics*,

vol. 25, no. 21, pp. 2809-2815, 2009.

[28] R. Diaz-Uriarte, "Identifying restrictions in the order of accumulation of mutations during tumor progression: effects of passengers, evolutionary models, and sampling," *BMC Bioinformatics*, vol. 16, no. 1, pp. 41, 2015.

[29] L. D. Wood, D. W. Parsons, S. Jones, J. Lin, T. Sjöblom, R. J. Leary, D. Shen, S. M. Boca, T. Barber, and J. Ptak, "The genomic landscapes of human breast and colorectal cancers," *Science*, vol. 318, no. 5853, pp. 1108-1113, 2007.

[30] Y.-A. Kim, D.-Y. Cho, P. Dao, and T. M. Przytycka, "MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types," *Bioinformatics*, vol. 31, no. 12, pp. i284-i292, 2015.

[31] P. W. Janes, S. H. Wimmer-Kleikamp, A. S. Frangakis, K. Treble, B. Griesshaber, O. Sabet, M. Grabenbauer, A. Y. Ting, P. Saftig, and P. I. Bastiaens, "Cytoplasmic relaxation of active Eph controls ephrin shedding by ADAM10," *PLoS biology*, vol. 7, no. 10, pp. 2264, 2009.

[32] G. Dennis Jr, B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, H. C. Lane, and R. A. Lempicki, "DAVID: database for annotation, visualization, and integrated discovery," *Genome biol*, vol. 4, no. 5, pp. P3, 2003.

[33] T. Nome, A. M. Hoff, A. C. Bakken, T. O. Rognum, A. Nesbakken, and R. I. Skotheim, "High frequency of fusion transcripts involving TCF7L2 in colorectal cancer: Novel fusion partner and splice variants," *PloS one*, vol. 9, no. 3, pp. e91264, 2014.

[34] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, and S. A. Roberts, "Mutational heterogeneity in cancer and the search for new cancer-associated genes," *Nature*, vol. 499, no. 7457, pp. 214-218, 2013.

[35] J. Wang, Z. Chen, H. Chen, M. Wang, X. Kong, Y. Wang, T. Sun, J. Hong, W. Zou, and J. Xu, "Elf3 drives β-catenin transactivation and associates with poor prognosis in colorectal cancer," *Cell death & disease*, vol. 5, no. 5, pp. e1263, 2014.

[36] C. G. A. Network, "Comprehensive molecular characterization of human colon and rectal cancer," *Nature*, vol. 487, no. 7407, pp. 330-337, 2012.

[37] Y. Zhou, E. Zhang, C. Berggreen, X. Jing, P. Osmark, S. Lang, C. M. Cilio, O. Göransson, L. Groop, and E. Renström, "Survival of pancreatic beta cells is partly controlled by a TCF7L2-p53-p53INP1-dependent pathway," *Human molecular genetics*, vol. 21, no. 1, pp. 196-207, 2012.

[38] M. Cizkova, G. Cizeron-Clairac, S. Vacher, A. Susini, C. Andrieu, R. Lidereau, and I. Bièche, "Gene expression profiling reveals new aspects of PIK3CA mutation in ERalpha-positive breast cancer: major implication of the Wnt signaling pathway," *PLoS One*, vol. 5, no. 12, pp. e15647, 2010.

[39] C. W. Brennan, R. G. Verhaak, A. McKenna, B. Campos, H. Noushmehr, S. R. Salama, S. Zheng, D. Chakravarty, J. Z. Sanborn, and S. H. Berman, "The somatic genomic landscape of glioblastoma," *Cell*, vol. 155, no. 2, pp. 462-477, 2013.

[40] H. Wu, L. Gao, J. Dong, and X. Yang, "Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks," *PloS one*, vol. 9, no. 3, pp. e91856, 2014.

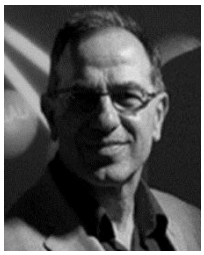
[41] N. Kasabov, *Springer Handbook of Bio-/neuro-informatics*: Springer Science & Business Media, 2013.

[42] N. K. Kasabov, "Neucube: A spiking neural network architecture for mapping, learning and understanding of spatio-temporal brain data," *Neural Networks*, vol. 52, no. 4, pp. 62-76, 2014.

Hao Wu received BSc and MSc degree in Computer Sciences in 2006 from Computer College of Inner Mongolia University and has been working at Information Engineering College of Northwest A&F University. He is a PhD student at School of Computer Science and Technology, Xidian University from 2012. He has published over 15 works in the areas of bioinformatics, complex networks, rough set, rough-fuzzy clustering, driver pathways and cancer progression. He worked as a visiting scholar at the Knowledge Engineering and Discovery Research Institute in Auckland University of Technology from July, 2014 to August, 2015. His main research interests include data mining, bioinformatics and systems biology. He is working on mining models and analysing algorithms related to cancer based on integrated biological networks.

Lin Gao received the B.Sc and M.Sc. in Computational Mathematics from Xi'an Jiaotong University and Northwest University, respectively, and the. Ph. D. degree in Circuit and System from Electronics Engineering Institute from Xidian University. She was a visiting scholar at University of Guelph, Canada from 2004 to 2005. At present, she is a professor in the Department of Computer Science and Technology, Xidian University. Her research interests include data mining, bioinformatics, graph theory and optimization. She has over 100 publications in professional journals and conferences. Her research has been funded by NSFC, 863 Program, ROSC and others. She has also served on various conference program committees and reviewer for various journals.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520934, IEEE/ACM Transactions on Computational Biology and Bioinformatics



Nikola K. Kasabov (**FIEEE, 2010**) received MSc degree in Electrical Eng., spec. Computer Science, in 1971 and PhD degree in Mathematical Sciences in 1975 from the Technical University in Sofia. He has published over 600 works in the areas of intelligent systems, neural networks, connectionist and hybrid connectionist systems, fuzzy systems, expert systems, bioinformatics, neuroinformatics. He is a Fellow of IEEE, Fellow of the Royal Society of New Zealand and Distinguished Visiting Fellow of the RAE UK. He is a Past President of the International Neural Network Society (INNS) and the Asia Pacific Neural Network Assembly (APNNA) and currently - a member of the INNS and APNNA Governing Boards. Kasabov is the Director of the Knowledge Engineering and Discovery Research Institute (www.kedri.aut.ac.nz) and Personal Chair of Knowledge Engineering in Auckland University of Technology, New Zealand. He is also a Marie Curie Fellow for the EvoSpike project at ETH/UZH Zurich and Advisory Professor at Shanghai Jiao-Tong University.