Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-28 (2015) 234–238

Inference of cancer progression from somatic mutation data

Hao Wu[1,2,3], Lin Gao[1*], Nikola Kasabov[3]

[1]*School of Computer Science and Technology, Xidian University, Xi'an, Shaanxi, China*
[2]*College of Information Engineering, Northwest A&F University, Yangling, Shaanxi, China*
[3]*Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, Auckland, New Zealand*

*Corresponding author: lgao@mail.xidian.edu.cn

Abstract: Large-scale cancer genomics projects are providing a wealth of somatic mutation data. Therefore, one of the most challenging problems arising from the data is to infer the temporal order of somatic mutations. In the paper, we present a network-based method (NetInf) to infer cancer progression at the pathway level. We apply it to analyze somatic mutation data from real cancer studies. Experimental results show that these detected pathways overlap with known pathways, including RB, P53 signaling pathways. Our method reduces computational complexity and also provides new insights on the temporal order of somatic mutations at the pathway level.

Keywords: Cancer genome; Cancer progression; Network models; Dynamic problem; Driver mutation

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]. 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) [2], The Cancer Genome Atlas (TCGA) [3], Kyoto Encyclopedia of Genes and Genomes (KEGG) [4] and so on. According to the analysis of somatic mutations in cancer genomes, one important problem appears. How to determine temporal orders of the driver mutations in cancer patients? It is almost impossible to obtain samples at multiple time-points from a single individual, therefore, it is very difficult to answer the question about temporal progression and identify what mutations occur early in cancer progression [5]. Taking into account cancer phylogeny, we aim at identifying common mutation events in the process of cancer progression that can help us develop new diagnostics or therapeutics targeted to specific subtypes of progression [6].

Several methods have been introduced to infer temporal progression of gene mutations from cross-sectional data [5, 7-12]. Desper *et al*. [7, 8] 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*. [9] 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*. [10] 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, 13-16] 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.* [5] 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 introduce a new network-based method to infer driver progression at the pathway level from cross-sectional mutation data. In the constructed gene networks, mutations of all genes in each complete subnetwork are approximately exclusive. Therefore, we just need to find a

10.1016/j.ifacol.2015.12.131

set of non-overlapping complete subnetworks which meet the linear progression between them.

2. METHODS

2.1 Exclusivity and progression

Vandin *et al.* [5] 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* $\Gamma(g) = \{i: A_{ig} = 1\}$ 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 *mutually exclusive.*

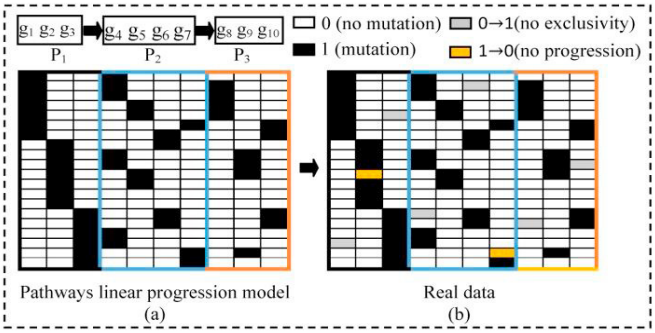

Fig. 1. Pathways linear progression model [5]**.** A linear progression model on gene sets creates a mutation matrix with mutually exclusive mutations within each gene set, and a progression of mutations across the gene sets. In real dataset, errors that disrupt the exclusivity and the progression are present.

Pathways Linear Progression Model (PLPM) [5]. 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 $\Gamma(P_{k-1}) \supseteq \Gamma(P_k)$, 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 $\left|\{g_j \in P_{k-1}: M_{i,j} = 1\}\right| > 0$.

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, the *exclusive degree* between the pair of genes 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 degree* function is denoted as:

$$CD(M) = \frac{|\Gamma(M)|}{m}. \quad (3)$$

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:

$$PR(M_j, M_k) = \frac{|\Gamma(M_j) \cap \Gamma(M_k)|}{|\Gamma(M_k)|}. \quad (4)$$

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.* [5] 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 reduce significantly 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, if the exclusive degree between each pair of genes is greater than or equal a *threshold* λ, an edge will be created to link this pair of genes, otherwise, there is no an edge between the pair of genes. The process is shown in Fig. 2.