

# 基于蚁群算法的肿瘤驱动通路搜索方法研究

潘 蕾, 秦 攀, 顾 宏\*

(大连理工大学 电子信息与电气工程学部, 辽宁 大连 116024)

**摘要:** 肿瘤的发生和发展主要是由基因突变的累积导致细胞信号通路调控紊乱而引起的。将通路中基因集的两个特性——高覆盖性和高排他性,与基因协变量相结合,提出了一种基于蚁群优化算法的驱动通路搜索方法。旨在基因突变数据及基因协变量数据的基础上,搜索满足高覆盖性和高排他性的通路基因集,从而识别致癌的驱动通路。首先,通过对基因表达水平、复制时间和染色体状态3个基因协变量相互的关联性,以及它们与基因突变频率的相关性进行分析,选择复制时间作为影响基因突变频率的权重协变量。然后,将权重协变量与现有方法结合,构造了一个新的最大权重子矩阵函数作为组合优化问题的目标函数。为克服该优化问题的NP难题,采用蚁群优化算法求解。应用该方法在肺腺癌的突变数据上进行了分析与验证。结果证明,该方法不仅比现有方法找到了更多在已证实通路中的癌基因,而且其中包含多个互斥性显著的基因对,证明了方法的有效性。

**关键词:** 细胞信号通路;覆盖性;排他性;突变频率;蚁群算法

**中图分类号:** R73-31; TP301.6 **文献标识码:** A **doi:** 10.7511/dllgxb201802011

## 0 引言

随着人们对肿瘤机制的深入研究,发现肿瘤的发生和发展与细胞信号通路有着密切的关系。细胞内癌基因突变的累积导致各种信号通路的紊乱,从而影响细胞的增殖、分化和凋亡,最终引起肿瘤的发生<sup>[1-2]</sup>。因此,信号通路的搜索不仅可以对肿瘤的形成机制有进一步的了解,还为肿瘤的治疗提供了新的分子靶点,具有重要的研究意义。

目前对于信号通路的研究方法主要有两类。一类是针对信号通路中基因集的覆盖性和排他性这两个特性。例如,Vandin等<sup>[3]</sup>首先提出了最大权重子矩阵模型,在提高覆盖性的同时抑制重叠,保证了排他性。该方法使用MCMC优化算法虽然提高了数据适应度,但随机搜索的迭代方式也带来了容易产生局部最优解的问题。Leiserson等<sup>[4]</sup>对Vandin的方法进行了改进,提出了同时识别多个满足覆盖性和排他性通路的模型,并使用线性规划的优化思想来提高算法精度和收敛速

度。但该方法并没有解决模型的NP难题。而且,上述方法都只在本地突变数据的基础上研究信号通路的两个共性特征,没有考虑基因自身的突变异质性。另一类是基于信号通路的先验知识。Babur等<sup>[5]</sup>使用通路数据库中已知的基因集数据,提出了一个用来量化基因间互斥性的度量,以此来构建信号网络用于筛选驱动通路基因集。由于目前信号通路先验知识数据库还不够完善,信号网络尚不能明确地表明基因间的相互作用关系。针对以上方法的缺陷,本文在不需要信号通路先验知识的基础上,将基因复制时间作为影响突变频率的重要协变量,重新定义最大权重子矩阵函数。在保证高覆盖性和高排他性的同时,充分考虑基因本身协变量对突变频率的影响,从而在更大程度上搜寻到癌症中的信号通路。

## 1 模型建立及优化

### 1.1 基因协变量与基因突变频率的相关性

基因突变异质性是肿瘤的特征之一,主要体

收稿日期: 2017-08-08; 修回日期: 2018-01-24.

基金项目: 国家自然科学基金资助项目(61633006,61502074);中国博士后科学基金资助项目(2016M591430);大连理工大学基本科研业务费专项资金资助项目(DUT17RC(4)09).

作者简介: 潘蕾(1992-),女,硕士生,E-mail:398105253@qq.com;秦攀(1978-),男,副教授,硕士生导师;顾宏\*(1961-),男,教授,博士生导师,E-mail:guhong@dut.edu.cn.

现在两个方面. 第一,不同肿瘤类型之间基因突变类型的异质性. 例如,肺癌患者体内基因突变多为 C→A 类型的突变,而胃肠道肿瘤的突变则多为 C→G 类型的突变<sup>[6]</sup>. 第二,肿瘤内部基因组的区域异质性,即基因不同时,突变频率也有很大的区别<sup>[7]</sup>. 而造成区域异质性的基因协变量主要有 3 个:基因表达水平、复制时间和染色体状态<sup>[8-9]</sup>. 这 3 个协变量的影响,会导致每个基因发生突变的

频率有所不同.

本文中,针对这 3 个基因协变量相互之间的关系,和它们对基因突变频率的影响程度进行了数值实验分析. 结果如图 1 所示,其中基因协变量数据来自于肿瘤基因组图谱数据库(The Cancer Genome Atlas,TCGA),数据信息详见本文 2.1 节. 该协变量数据已经证明适用于其他癌症实验分析<sup>[10]</sup>.

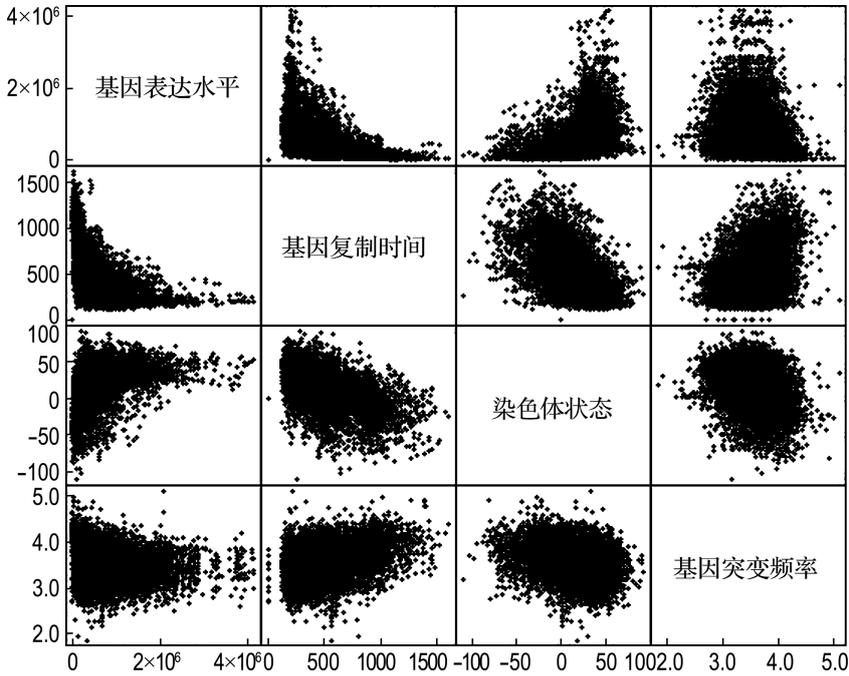


图 1 基因协变量与基因突变频率的相关性

Fig. 1 The relationship between gene covariates and gene mutation frequency

根据图 1,发现基因表达水平和染色体状态与基因突变频率之间均呈负相关关系,相关系数分别为 $-0.216\ 226\ 1$ 和 $-0.275\ 197\ 7$ . 基因复制时间和基因突变频率呈正相关关系,相关系数为 $0.352\ 776\ 1$ . 而且,3 个协变量之间的互相关关系图显示,各个基因协变量之间也存在着很强的关联性. 相关研究表明,基因组不同区域的复制时间与基因表达水平、染色体状态有着密切的关系. 复制时间较早的基因,染色体高度螺旋,基因表达水平较高. 而复制时间较晚的基因,染色体状态疏松,基因表达水平较低,甚至不表达<sup>[11]</sup>. 因此,为了减少算法的复杂度,选择基因复制时间作为对基因突变频率影响最为重要的协变量而结合到本文方法中.

## 1.2 方法介绍

细胞信号通路中的基因集有两个特性,即高覆盖性和高排他性<sup>[12]</sup>. 其中高覆盖性是指一个通路中的基因应该尽可能多地覆盖样本,高排他性则是指每个通路中基因的突变对于每个病人来说,要尽可能呈现唯一性. 根据通路中基因集的两个特性,Vandin 等<sup>[3]</sup>提出了 Dendrix(De novo driver exclusivity)方法,该方法的缺陷是忽略了基因自身协变量对突变频率的影响. 为了解决上述问题,本文基于基因协变量的影响,提出了一种改进方法(ACO covariant driver pathway, ACDP). 方法模型建立步骤如下:

(1)构造突变矩阵  $A_{m \times n}$ ,  $m$  是独立病人样本编号,  $n$  是基因名,如图 2 所示.  $a_{ij} = 1$  表明第  $i$  个

病人的第  $j$  个基因发生了突变。

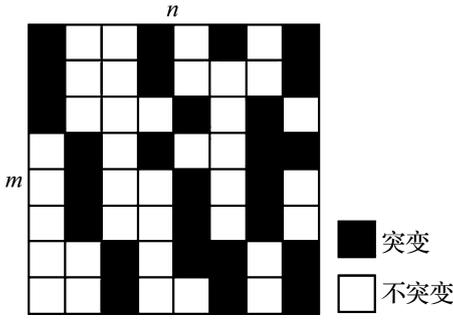


图 2 突变矩阵

Fig. 2 Mutation matrix

(2) 定义基于基因协变量影响的最大权重子矩阵函数:

$$W(\mathbf{M}) = |\Gamma(\mathbf{M})| - \omega(\mathbf{M}) = 2|\Gamma(\mathbf{M})| - \sum_{j \in M} \lambda_j |\Gamma(j)| \quad (1)$$

式中:  $\mathbf{M}_{m \times N}$  是从突变矩阵  $\mathbf{A}$  中得到的  $N$  列最大权重子矩阵;  $\Gamma(j) = \{i; a_{ij} = 1\}$  表示基因  $j$  发生突变时的所有病人样本集;  $|\Gamma(\mathbf{M})| = \bigcup_{j \in M} \Gamma(j)$  是覆盖性的一个度量, 表明  $\mathbf{M}$  中所有基因突变对应的病人样本集;  $\omega(\mathbf{M}) = \sum_{j \in M} \lambda_j |\Gamma(j)| - |\Gamma(\mathbf{M})|$  是对排他性的一个度量, 表明  $\mathbf{M}$  中所有样本重复覆盖的个数;  $\lambda_j$  是基因  $j$  的协变量权重值。

上述模型也可转化为一个二元线性规划问题进行求解:

$$\begin{aligned} \max f(\mathbf{x}, \mathbf{y}) &= 2 \sum_{i=1}^m x_i - \sum_{j=1}^n \left( y_j \sum_{i=1}^m \lambda_j a_{ij} \right) \\ \text{s. t. } x_i &\leq \sum_{j=1}^n a_{ij} y_j; \quad i = 1, \dots, m \\ \sum_{j=1}^n y_j &= N \end{aligned} \quad (2)$$

式中:  $N$  表示  $\mathbf{M}$  矩阵中基因集的元素个数;  $x_i \in \{0, 1\}$  代表第  $i$  个病人样本落入  $\mathbf{M}$  矩阵中的基因是否发生了突变, 突变为 1, 否则记为 0;  $y_j \in \{0, 1\}$  代表第  $j$  个基因是否落入  $\mathbf{M}$  矩阵, 落入则为 1, 否则记为 0;  $\mathbf{x}$  与  $\mathbf{y}$  分别是由  $x_i$  与  $y_j$  构成的向量。

### 1.3 目标函数优化

由于上述组合优化是一个 NP 难题<sup>[13]</sup>, 本文使用启发式的蚁群算法来求解组合最优化问题。

蚁群算法是一种随机优化算法, 受到自然界真实蚂蚁的行为启发而提出的模拟进化算法, 其根据蚂蚁在路径上释放的信息素指导蚁群进行路径寻优<sup>[14]</sup>。蚁群算法不仅可以解决局部最优问题, 而且其后期快速收敛的特性适合解决这种大规模数据的优化。本文提出的优化目标函数实际上是一个 0-1 背包问题, 即当限制背包的承重时, 寻找使得背包中总价值最大物品问题。针对细胞信号通路搜索问题, 将每个基因的“质量”  $w$  设置为 1, 用承重值控制基因集中基因个数  $N$ ; 每个基因的“价值”  $c$  是在协变量影响下的基因突变次数; 把基因是否落入限制大小的基因集中描述为某个物品是否装入限定承重的背包。

本文使用蚁群算法对目标函数进行优化, 当某个基因上累积的信息素越来越多时, 这个基因最终落入结果基因集的概率就越大。在一次迭代中, 蚁群中的每只蚂蚁都是按基因选择概率的大小来决定要选择的基因。下式表示第  $k$  只蚂蚁对基因  $g$  的选择概率:

$$p_g^k(t) = \begin{cases} \frac{(\tau_g(t))^\alpha (\eta_g(t))^\beta}{\sum_{s \notin T(k)} (\tau_s(t))^\alpha (\eta_s(t))^\beta}; & g \notin T(k) \\ 0; & g \in T(k) \end{cases} \quad (3)$$

式中:  $T(k)$  是禁忌表, 是第  $k$  只蚂蚁在一次迭代中选择基因的历史记录表, 作用是避免重复选择已落入基因集的基因;  $\tau_g(t)$  是基因  $g$  在  $t$  次迭代过程中的信息素强度;  $\eta_g(t)$  是启发函数, 在解决背包问题时, 常令  $\eta_g(t) = c_g / w_g$ ,  $c_g$  是基因  $g$  的“价值”,  $w_g$  是基因  $g$  的“质量”, 则  $\eta_g$  代表基因  $g$  的“单位价值”;  $\alpha$  是信息启发因子, 决定信息素的重要性;  $\beta$  是期望启发因子, 决定启发函数的重要性。

每只蚂蚁在选择一个基因后, 需要判断此时背包的质量是否超出承重值, 也就是判断已选择基因的个数是否超出设置的基因集大小  $N$ 。当蚁群中每只蚂蚁都完成一次迭代中的所有选择后, 每个基因上累积的信息素要进行一次调整, 调整公式为

$$\begin{aligned} \tau_g(t+1) &= (1-\rho)\tau_g(t) + \Delta\tau_g(t) \\ \Delta\tau_g(t) &= \sum_{k=1}^{n_a} \Delta\tau_g^k(t) \end{aligned} \quad (4)$$

式中:  $\rho$  是信息素挥发系数,  $\rho \in (0, 1)$ ;  $\tau_g(t+1)$  表示在下一迭代过程中基因  $g$  的信息素强度;  $n_a$

表示蚁群中蚂蚁的个数； $\Delta\tau_g^k(t)$ 表示第  $k$  只蚂蚁在本次迭代过程中在基因  $g$  上释放的信息素，计算公式为

$$\Delta\tau_g^k = \begin{cases} c_g Q / c^k; & g \in g^k \\ 0; & g \notin g^k \end{cases} \quad (5)$$

式中： $Q$ 表示信息强度，是一个常数； $g^k$ 表示第  $k$  只蚂蚁在本次迭代中选择的基因列表； $c^k$ 是第  $k$  只蚂蚁所选基因的“总价值”。本文中实验参数设置分别为  $\alpha=2, \beta=5, \rho=0.5, Q=100$ ，蚁群规模为 30。

当完成所有迭代过程后，计算每只蚂蚁的背包价值，价值最大的背包中对应的基因则为选择的信号通路基因集。

#### 1.4 ACDP 方法的伪代码实现

采用 Matlab 实现本文 ACDP 方法，表 1 是对程序中部分变量的解释。ACDP 方法的具体伪代码如下：

表 1 程序中部分变量解释

Tab.1 The explanation of part of the variables in the program

变量	解释
$S_m$	最大迭代次数
$C_{mr}$	第 $t$ 次迭代过程中基因集的最大价值
$C_m$	一次迭代中，每只蚂蚁选中基因集的最大价值
$M_v$	所有迭代过程中最大价值基因集中基因编号
$G_s$	基因集

输入： $w, c, m, n, N, S_m, n_a$

输出： $M_v, C_m$

1. 初始化  $t=1, T=\emptyset, C_m=0$
2. while  $t \leq S_m$  do
3. for  $k=1:n_a$  do
4.  $T(k) = \{g_i\}$
5. for  $j=2:N$  do
6. 
$$p_g^k(t) = \begin{cases} \frac{(\tau_g(t))^\alpha (\eta_g(t))^\beta}{\sum_{s \in T(k)} (\tau_s(t))^\alpha (\eta_s(t))^\beta}; & g \notin T(k) \\ 0; & g \in T(k) \end{cases}$$
7.  $T(k) = T(k) + \{g | \max(p_g^k(t))\}$
8. end
9. end

10. 
$$C_{mr} = \max f(x, y) = 2 \sum_{i=1}^m x_i - \sum_{j=1}^n \left( y_j \sum_{i=1}^m \lambda_j a_{ij} \right)$$
11.  $C_m = \max(C_m, C_{mr})$
12.  $M_v = \{G_s | C_m\}$
13.  $\tau_g(t+1) = (1-\rho)\tau_g(t) + \Delta\tau_g(t)$
14.  $T = \emptyset$
15.  $t = t + 1$
16. end

## 2 实验结果与分析

### 2.1 实验数据集

使用肺腺癌数据对所提方法进行了验证。肺腺癌基因突变数据来自于肿瘤测序计划数据库(The Tumour Sequencing Project, TSP)，由此构

造突变矩阵  $A_{m \times n} = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}$ ， $m=163, n=$

319,  $a_{ij} \in \{0, 1\}$ ；基因协变量数据来自于肿瘤基因组图谱数据库(TCGA)，由此构造协变量矩阵  $C_{p \times q} = (G \ E \ R \ H \ F)$ ， $p=17\ 668, q=5$ ，其中  $G$  为基因数据， $E$  为基因表达数据， $R$  为基因复制时间数据， $H$  为染色体状态数据， $F$  为基因突变频率数据；本文实验结果验证过程中的通路信息来自于全基因组及代谢途径数据库(Kyoto Encyclopedia of Genes and Genomes, KEGG)。其中肺腺癌通路信息选用 KEGG 数据库中与肺腺癌显著相关的一些通路，分别为 MAPK 信号通路、p53 信号通路、Wnt 信号通路、细胞循环通路和 mTOR 通路<sup>[15]</sup>。

### 2.2 实验结果对比

本文实验是在肺腺癌突变数据上分别运行 ACDP、Dendrix<sup>[3]</sup>、Multi-Dendrix<sup>[4]</sup> 和 Mutex<sup>[5]</sup>，并对 4 种方法的通路搜索结果在样本中的覆盖性、在肺腺癌相关通路中存在性，以及互斥性进行了对比分析。其中，基因对的互斥性由费希尔精确检验得到的  $P$  值来度量， $P$  值越小，互斥性越显著。对于不同的基因集大小  $N(4 \sim 10)$ ，实验结果精度均在 78% 以上，以  $N$  取 9 为例，实验结果如表 2 所示。其中括号中表示显著性很高的基因对，加粗部分表示该基因在肺腺癌相关信号通路中。

表 2 4 种方法实验结果

Tab. 2 The experimental results of four methods

方法	基因集	准确度	覆盖性	互斥性 $P$ 值
ACDP	(EGFR KRAS) STK11 (TP53 ATM) APC NF1 LRP1B MASTL	78% (7/9)	84% (137/163)	(EGFR KRAS) $<0.001$ (TP53 ATM)0.012
Dendrix	(EGFR KRAS) NF1 NRAS TSC2 PTEN BAX CHEK1 BRCA2	67% (6/9)	66% (108/163)	(EGFR KRAS) $<0.001$
Multi-Dendrix	(EGFR KRAS) STK11 (TP53 ATM) VAV1 FES ERBB4 PTPRD	56% (5/9)	83% (135/163)	(EGFR KRAS) $<0.001$ (TP53 ATM)0.012
Mutex	(EGFR KRAS) NF1 BRAF TP53 MAPK1 RIT1 HGF WRN	67% (6/9)	76% (124/163)	(EGFR KRAS) $<0.001$

表 2 结果显示,本文方法比另外 3 种方法找到了更多在肺腺癌通路中的癌基因,其中包括 NF1、STK11、APC 等基因.很多相关的医学研究表明,这些基因在肺腺癌中有着很高的突变频率,会影响机体对细胞增殖、分化和凋亡过程的正常调控,从而导致肿瘤的发生<sup>[16-20]</sup>.更重要的是,本文方法可以直接搜索到多个互斥基因对,而 Dendrix 方法则需要在数据中删除找到的基因对,才能搜索其他的基因对,这样做忽略了已搜索的互斥基因对和其他基因之间的互斥性.如表 2 所示,在搜索到的基因集结果中,有两个互斥性较为显著的基因对(EGFR KRAS)和(TP53 ATM).医学研究人员通过实验在两个基因对中发现了很明显的负

相关性,充分说明了基因对中任意一个基因发生突变,足以使得相关通路所控制的生物功能失控.另一方面,也指出了这两个基因对分别在 MAPK 信号通路和细胞循环通路中有着直接影响的作用关系<sup>[21]</sup>.如 ATM 基因是一个重要的抑癌基因,负责细胞循环检测点酶蛋白的编码,参与细胞周期调控.有实验证据显示 DNA 损失程度信号是通过 ATM 来传递给下游的 TP53 基因,再由 TP53 来修复受损的 DNA,促进癌细胞的凋亡,发挥其抑癌基因的作用<sup>[22]</sup>.本文还对两个互斥性显著基因对的覆盖性和在细胞信号通路中作用关系做了可视化表达,更形象地体现本文方法结果的可靠性及生物意义,结果分别如图 3、4 所示.

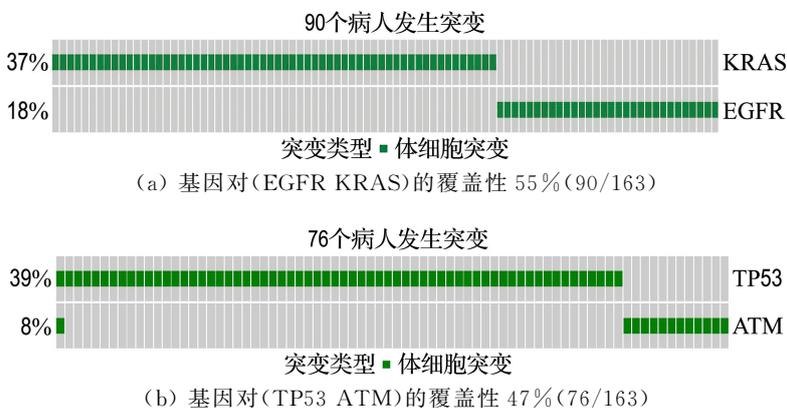


图 3 基因对的覆盖性统计

Fig. 3 The coverage statistic of gene pairs

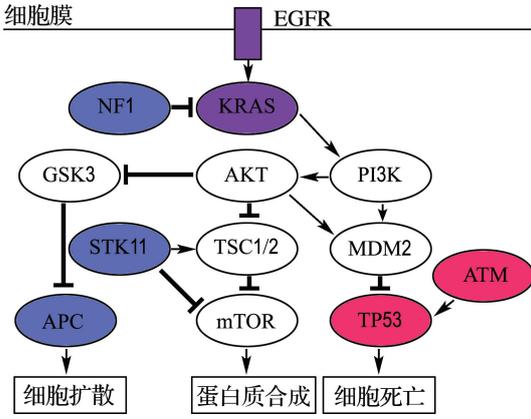


图4 肺腺癌信号通路

Fig. 4 Signal pathways in lung adenocarcinoma

### 3 结 语

本文提出了一种新的用于细胞信号通路中基因集搜索的方法。在考虑基因集覆盖性和排他性的基础上,用基因复制时间作为影响基因突变频率的权重协变量,结合到方法当中。由于本文提出的目标函数属于 NP 难题,使用蚁群算法进行优化。实验结果表明,方法不仅在癌症相关通路中搜索到更多的癌基因,更重要的是比现有方法能找到更多的已经证实在细胞信号通路中存在直接作用的互斥基因对。

### 参考文献:

[1] WEINSTEIN I B. Addiction to oncogenes; The Achilles heal of cancer [J]. *Science*, 2002, **297**(5578):63-64.

[2] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways [J]. *Nature*, 2008, **455**(7216):1061-1068.

[3] VANDIN F, UPFAL E, RAPHAEL B J. De novo discovery of mutated driver pathways in cancer [J]. *Genome Research*, 2012, **22**(2):375-385.

[4] LEISERSON M D M, BLOKH D, SHARAN R, *et al*. Simultaneous identification of multiple driver pathways in cancer [J]. *PLoS Computational Biology*, 2013, **9**(5):e1003054.

[5] BABUR Ö, GÖNEN M, AKSOY B A, *et al*. Systematic identification of cancer driving signaling pathways based on mutual exclusivity of genomic alterations [J]. *Genome Biology*, 2015, **16**(1):45.

[6] PLEASANCE E D, STEPHENS P J, O' MEARA S, *et al*. A small-cell lung cancer genome with complex signatures of tobacco exposure [J]. *Nature*, 2010, **463**(7278):184-190.

[7] Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer [J]. *Nature*, 2015, **487**(7407):330-337.

[8] PLEASANCE E D, CHEETHAM R K, STEPHENS P J, *et al*. A comprehensive catalogue of somatic mutations from a human cancer genome [J]. *Nature*, 2010, **463**(7278):191-196.

[9] STAMATOYANNOPOULOS J A, ADZHUBEI I, THURMAN R E, *et al*. Human mutation rate associated with DNA replication timing [J]. *Nature Genetics*, 2009, **41**(4):393-395.

[10] LAWRENCE M S, STOJANOV P, POLAK P, *et al*. Mutational heterogeneity in cancer and the search for new cancer-associated genes [J]. *Nature*, 2013, **499**(7457):214-218.

[11] NEPH S, VIERSTRA J, STERGACHIS A B, *et al*. An expansive human regulatory lexicon encoded in transcription factor footprints [J]. *Nature*, 2012, **489**(7414):83-90.

[12] VOGELSTEIN B, KINZLER K W. Cancer genes and the pathways they control [J]. *Nature Medicine*, 2004, **10**(8):789-799.

[13] 高海昌,冯博琴,朱 利. 智能优化算法求解 TSP 问题[J]. 控制与决策, 2006, **21**(3):241-247, 252. GAO Haichang, FENG Boqin, ZHU Li. Reviews of the meta-heuristic algorithms for TSP [J]. *Control and Decision*, 2006, **21**(3):241-247, 252. (in Chinese)

[14] 丁建立,陈增强,袁著社. 基于自适应蚂蚁算法的动态最优路由选择[J]. 控制与决策, 2003, **18**(6):751-753, 757. DING Jianli, CHEN Zengqiang, YUAN Zhuzhi. Dynamic optimization routing method based on ant adaptive algorithm [J]. *Control and Decision*, 2003, **18**(6):751-753,757. (in Chinese)

[15] TAKAHASHI T, NAU M M, CHIBA I, *et al*. p53: A frequent target for genetic abnormalities in lung cancer [J]. *Science*, 1989, **246**(4929):491-494.

[16] AHRENDT S A, HU Y C, BUTA M, *et al*. p53 mutations and survival in stage I non-small-cell

- lung cancer; results of a prospective study [J]. **Journal of the National Cancer Institute**, 2003, **95**(13):961-970.
- [17] VELCHETI V, GOVINDAN R. Hedgehog signaling pathway and lung cancer [J]. **Journal of Thoracic Oncology**, 2007, **2**(1):7-10.
- [18] STEWART D J. Wnt signaling pathway in non-small cell lung cancer [J]. **Journal of the National Cancer Institute**, 2014, **106**(1):djt356.
- [19] TOMASINI P, WALIA P, LABBE C, *et al.* Targeting the KRAS pathway in non-small cell lung cancer [J]. **Oncologist**, 2016, **21**(12):1450-1460.
- [20] PACKENHAM J P, TAYLOR J A, WHITE C M, *et al.* Homozygous deletions at chromosome 9p21 and mutation analysis of p16 and p15 in microdissected primary non-small cell lung cancers [J]. **Clinical Cancer Research**, 1995, **1**(7):687-690.
- [21] DING Li, GETZ G, WHEELER D A, *et al.* Somatic mutations affect key pathways in lung adenocarcinoma [J]. **Nature**, 2008, **455**(7216):1069-1075.
- [22] 唐光明. p53 及其上游基因 ATM、下游基因 PUMA 在大肠癌中的表达及意义[D]. 南充:川北医学院, 2015.
- TANG Guangming. The expression and significance of p53 and upstream gene ATM, downstream gene PUMA in colorectal cancer [D]. Nanchong: North Sichuan Medical College, 2015. (in Chinese)

## Research on searching method of driver pathways in tumor based on ant colony algorithm

PAN Lei, QIN Pan, GU Hong\*

( Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China )

**Abstract:** The occurrence and development of tumor are mainly caused by the accumulation of gene mutations, which leads to the disorder of cell signal pathways. There are two properties of gene sets in a pathway, i. e. , high coverage and high exclusivity. A driver pathway searching method is proposed based on ant colony optimization algorithm by combining gene covariates with these two properties. In this way, cancer-causing driver pathways are identified by searching for highly covered and highly exclusive gene sets in a pathway based on the gene mutation data and covariate data. First, by analyzing the correlations between the three gene covariates, i. e. , gene expression level, replication time and hic compartment, and their correlations with the gene mutation frequency, replication time is selected as the weight covariate of the gene mutation frequency. Then, a novel maximum weight submatrix function is constructed by combining the weight covariate with existing methods as the objective function of the combinatorial optimization problem. Finally, the ant colony optimization algorithm is introduced to overcome the NP problem of this optimization problem. The proposed method is applied to the lung adenocarcinoma mutation data and the results show that compared with the existing methods the proposed method can identify more cancer genes, some of which are involved in known pathways. In addition, the detected gene pairs with significant exclusivity are contained, all of these prove the efficiency of the proposed method.

**Key words:** cell signal pathway; coverage; exclusivity; mutation frequency; ant colony algorithm