基于突变基因网络的致癌驱动通路检测算法

吴 昊
(西北农林科技大学信息工程学院 陕西 杨凌 712100)

摘 要 通过检测突变驱动通路研究癌症的发病机理是当前癌症基础性研究的关键问题之一.该研究以人类基因组工程提供的体细胞突变数据为研究对象,结合基因组图谱中广泛存在的互斥性原理,提出一种新型的基于基因互斥网络的致癌突变驱动通路检测算法(Megnet).该算法首先利用大量癌症病人的体细胞突变数据,结合基因间互斥性原理构建突变基因网络,然后检测该网络中具有高覆盖的最大完全子图.为验证算法的效率和鲁棒性,我们将该算法应用于模拟数据中,结果显示所有模拟过程均在 15 秒内完成驱动通路检测,Megnet 算法比 Dendrix 和 Multi-Dendrix 算法运行时间更短且结果准确率更高.同时为验证算法的有效性,我们将该算法应用于肺癌数据和神经胶质瘤体细胞突变数据中,结果显示 Megnet 算法不仅比 Dendrix 和 Multi-Dendrix 算法检测的基因集合具有更高的生物相关性和统计显著性,而且还检测出一些可供生物验证的新候选基因集合,并且这些检测的基因集合与已知的 P53、RB、RAS 和 PI3K 等信号通路及细胞循环和细胞凋亡通路具有较高的重叠.Megnet 算法不需要指定通路中的基因个数和任何先验知识,为癌症发病机理研究提供新视野.该算法通过构建突变基因网络,简化了基因间相互关联关系,降低了算法复杂度,提高了致癌突变驱动通路检测的效率和准确性,对于癌症发病机理研究具有较强的理论意义和实践价值.

关键词 癌症基因组;发病机理;突变驱动通路;体细胞突变;互斥性;基因网络;生物信息学
中图法分类号 TP311 **DOI 号** 10.11897/SP.J.1016.2018.01400

Algorithm for Detecting Driver Pathways in Cancer Based on Mutated Gene Networks

WU Hao
(*College of Information Engineering, Northwest A&F University, Yangling, Shaanxi* 712100)

Abstract Recent genome sequencing studies have shown that somatic mutations drive cancer development across a large number of genes. One of the key issues of current basic research on cancer is to study the pathogenesis of cancer by detecting mutated driver pathways. This study utilizes the widely existing property of mutual exclusivity in cancer genomic spectrum, examines somatic mutation data provided by the Human Genome Projects, and proposes a novel algorithm (Megnet) for detecting mutated driver pathways on the basis of mutually exclusive gene networks. This study first constructs mutated gene networks based on mutual exclusivity between each pair of genes utilizing somatic mutation data from many cancer patients and then detects the largest complete subgraphs with high coverage. The genes in the largest complete subgraph are recurrently altered across a majority of tumor samples; they are known to or are likely to take part in the same biological process; and mutations of genes within one largest complete subgraph are mutually exclusive. To evaluate the efficiency and robustness of Megnet method, we apply it to simulated data with 300 samples and 1000 genes in which 15 mutually exclusive gene sets with different

收稿日期:2016-05-28;在线出版日期:2016-12-31.本课题得到中央高校基本科研业务费(2452017342)、博士科研启动费(2452017019)、陕西省自然科学基金面上项目(2017JM6063)、国家自然科学基金重点项目(61532014,61432010)、陕西省杨凌区科技计划项目(2017GY-03)资助.吴 昊,男,1979 年生,博士,副教授,中国计算机学会(CCF)会员,主要研究方向为系统生物信息学、复杂网络、复杂疾病和生物数据挖掘.E-mail: haowu@nwsuaf.edu.cn.

coverage degree are embedded into, and the results indicate that Megnet algorithm finishes the process of detecting driver pathways in 15 seconds in all simulations and has shorter runtime and higher accuracy than those of Dendrix and Multi-Dendrix. To further verify the efficacy of our algorithm, we apply it to somatic mutation data from the mutation profiles of lung carcinoma and glioblastoma tumor samples from the Cancer Genome Atlas (TCGA), and results show that Megnet can not only detect more biologically relevant and higher statistically significant gene sets than those of Dendrix and Multi-Dendrix, but also identify some new candidate gene sets for biological verification, which have a high degree of overlap with the known signaling pathways (like P53, RB, RAS and PI3K), cell cycle and cell apoptosis pathway. Since somatic mutations are hypothesized to target a small number of cellular signaling and regulatory pathways, a common method is to appraise whether the known pathways are enriched for the mutated gene sets. In glioblastoma multiforme cancer, we make the novel observation that TP53 alteration, copy-number amplification of MDM2 and MDM4 are mutually exclusive, and RB1 deletion, loss of CDK4 and CDKN2B are also mutually exclusive, suggesting distinct alternative causes of genomic instability in the cancer type. Overall, we develop a simple, fast and sensitive method for automatically detecting driver pathways in tumors based on mutually exclusive mutational patterns. Megnet algorithm does not need to assign the number of genes in a driver pathway neither requires any prior knowledge, thus providing insights into the research on the pathogenesis of cancer. Based on constructing mutated gene networks, the algorithm simplifies the mutual relationship between genes, reduces the algorithm complexity, and improves the efficiency and accuracy of detecting mutated driver pathways; therefore, it has high theoretical and practical value to research on the pathogenesis of cancer.

Keywords　cancer genome; pathogenesis; mutated driver pathways; somatic mutation; mutual exclusivity; gene networks; bioinformatics

1　引　言

随着基因组测序技术的快速发展和测序费用的降低,一些基因组工程(如TCGA[1-2]、ICGC[3]、CCLE[4]、TARGET[5])已对几十种癌症进行测序,并为每种癌症提供几百个样本的测序数据,这些数据的快速涌现为我们更好地理解癌症的发生、发展和恶化过程提供了前所未有的机遇和数据支撑.癌症主要由生命体存活过程中体细胞突变加速积累引起,体细胞突变主要包括单核苷酸变异、拷贝数变异、核苷酸序列重复、插入以及缺失等.基因突变可能会引起编码蛋白的氨基酸序列改变,导致蛋白结构变化而影响活性,也有一些位于调控区的基因突变可能引起功能蛋白表达量的差异,进而引发癌症[6-10].目前,癌症基因组已发现了大量的体细胞突变数据,然而解释这些数据时面临的一个关键问题是如何从这些体细胞突变数据中区分致癌的驱动突变基因和对癌症发生没有影响的乘客突变基因[11-15];另一个关键挑战是如何检测出癌症细胞中频繁突变并导致癌症发生的功能性突变基因集合,即突变驱动通路[6-8,16-18].

第一个关键问题的常用解决方法为从大量的癌症病人样本中挑选突变率较高的基因作为驱动突变基因,这种统计方式已检测出许多重要的致癌基因,却不能检测导致癌症发生的突变驱动通路.然而癌症的发生并非由单一基因突变引起,而是由一组基因突变所致[19-21].在一个重要的通路中患同种癌症的不同病人可能持有不同的突变基因.在生物体中,驱动突变通常不仅靶向单个基因组位点(如单核苷酸或基因),而且靶向多个基因组成的细胞信号传导和调控通路,这种异质现象加大了通过多组样本检测驱动突变的难度[19,22-23].判断一个基因突变是驱动突变还是乘客突变需要通过生物实验手段检测基因的生物功能,然而该检测方法,尤其是从大量病人样本中每个频发突变的基因集合位点进行生物实验验证的成本高、工作量大,在当前技术水平和研究能力中很难实现.因此,通过计算的方式预测致癌突变

驱动通路具有重要的理论意义和实践价值.

本文第 2 节对近年来的相关研究工作进行分析和总结;第 3 节提出基于构造突变基因网络的驱动通路检测算法;第 4 节通过大量的模拟仿真实验验证算法的效率和鲁棒性,并通过真实的肺癌和神经胶质瘤体细胞突变数据验证算法的有效性;第 5 节总结全文并讨论本研究的应用价值.

2 相关工作

目前,基于计算方式研究癌症的发病机理,检测互斥的致癌突变驱动通路主要包括两种方式:一种是只使用体细胞突变数据,利用基因组图谱中广泛存在的高互斥和高覆盖两种属性,不使用任何先验知识直接检测导致癌症发生的驱动通路[6-8,16-17];另一种是利用体细胞突变数据中基因间互斥性、基因表达数据中基因共表达原理和蛋白质网络分析基因间的相互作用关系,集成体细胞突变、基因表达和蛋白质网络等数据信息检测导致癌症发生的失调模块[19,24-25].

由于人类基因和蛋白质相互作用知识尚不够完善,因此存在的基因通路数据库和相互作用网络尚不能精确显示细胞内的通路和相互作用关系[8].因此,把多种生物数据集成研究致癌驱动通路可能会限制发现某些具有显著生物功能的基因集合.近年来,研究者对基因组图谱进行分析,研究驱动基因是如何出现在一个通路中,提出了将高互斥性和高覆盖性相结合的模式[6,8,10,16-17,26-28].互斥现象存在于多种癌症中,即在一个基因集合中,每位病人最多只有一个基因发生突变.比如,在 p53 信号通路中,TP53 突变、MDM2 和 MDM4 的拷贝数增殖几乎不会同时出现在同一位神经胶质瘤病人中;在细胞循环和细胞凋亡通路中,RB1 删除、CDK4 和 CDKN2B 的缺失也几乎不会同时出现在同一位神经胶质瘤病人中.相对于高互斥性,高覆盖性是突变驱动通路另一个关键特征,即一个驱动通路中的基因趋向于在多数病人中发生突变.目前多数生物学者认为互斥的突变基因集合为功能性突变基因和生物通路之间存在的关联关系提供了强有力的证据[27].因此,近几年计算生物学研究领域基于突变基因间互斥性提出了多种计算方式用以检测癌症的突变驱动通路或者失调模块.

利用这种组合模式,Vandin 等人[16]提出了 Dendrix 算法以体细胞突变数据为研究对象检测突变驱动通路.为了得到高互斥的基因集合,该算法引入了惩罚重叠同时悬赏覆盖的权重函数 W,即 $W(M)=|\Gamma(M)|-\omega(M)=2|\Gamma(M)|-\sum_{g\in M}|\Gamma(g)|$,这里 $\Gamma(g)$ 表示基因 g 发生突变的病人样本集合,$|\Gamma(M)|$ 表示基因集合 M 的覆盖度,$\omega(M)$ 表示基因集合 M 覆盖重叠样本个数.找到最大 $W(M)$ 的问题被称为最大权重子阵列问题.作者引入了贪心算法和马尔科夫链蒙特卡洛(MCMC)思想查找满足最大覆盖互斥的子阵列.该算法计算每个基因集合的权重分数,选取权重分数最高的基因集合作为突变驱动通路,然后移除这些节点,再次重复上述步骤进行迭代,然而这种随机搜索的迭代方式只能产生局部最优解,而且计算时需要事先指定驱动通路中基因的数量.

Zhao 等人[6]集成体细胞突变数据和基因表达数据研究致癌驱动通路,提出了 MDPFinder 算法.该算法利用遗传算法确定最优边界范围问题,结合相同通路中的基因集合通常共同执行某一生物功能的特性表示基因表达数据中基因间的关系,结合体细胞突变数据中基因间互斥关系共同衡量基因间的关系,利用线性规划算法解决最大权重子阵列问题,解决了 Dendrix 算法存在的局部最优解问题.

Szczurek 等人[26]开发了一个随机模型来分析癌症突变数据,并检测互斥基因集合.作者引入了两个模型:一是允许观测误差的突变基因互斥性随机生成模型;二是假定基因突变相互独立的空模型.随机生成模型假定一个集合中的基因具有相同的突变概率,使用互斥的统计测试方法对两种模型进行比较以检测突变驱动通路.因此,该模块检测方法趋向于把突变率相近的基因划分于一个基因集合.

Leiserson 等人[8]改善了 Dendrix 算法,提出了一种同时检测多个驱动通路的 Multi-Dendrix 算法.该算法优化了权重函数使用线性规划思想同时检测多个满足高互斥和高覆盖的基因集合.该算法把最大权重子阵列问题作为线性规划问题,定义目标函数和设定限制条件,以找到满足条件的最大权重子阵列,属于 NP 难问题.

为了降低算法的复杂度和解决最大权重子阵列的 NP 难问题,本文提出一个基于构造突变基因网络以检测驱动通路算法,该算法不需要任何先验知识,直接从体细胞突变数据中检测突变驱动通路.具体方法为:首先,过滤掉突变率低的基因,保留满足一定突变率标准的基因,因为突变率很低的基因常为对癌症发生没有影响的乘客突变[3].本研究借鉴前期的研究成果将 MAF 设为 2.5%,即删除样本中

MAF 低于 2.5%的基因[8,16-17]. 其次，计算每对基因间的互斥度和权重函数值，如果一对基因间的互斥度大于等于阈值 λ，并且权重函数值大于等于阈值 γ，则认为这对基因满足互斥关系，并建立连边关系，构成基因相互作用网络. 最后，检测该网络中满足高覆盖的最大完全子图，由于网络中任何两个节点都是互斥的，因此每个完全子图都符合高互斥性，每个高覆盖的最大完全子图很可能就是一个突变驱动通路[16]，这是因为每个癌症病人包含为数不多的干扰着多个细胞通路或者调控通路的驱动突变基因. 由于每一个病人的每个突变驱动通路大约只包含一个驱动突变基因，因此在一个驱动通路中，突变驱动基因间存在着互斥关系. 另外，一个重要的突变驱动通路应需覆盖大多数病人，这就是突变基因的高覆盖模式. 也就是说，一个突变驱动通路对应着一个基因集合，这个基因集合中的基因在大多数病人中发生突变，并且这些突变基因间互斥或近似互斥. 这些高互斥高覆盖的基因集合通常比大多数信号通路和调控通路包含的基因个数更少.

3　突变驱动通路检测算法

3.1　互斥度和覆盖度

对基因组图谱中可能存在的大量基因集合进行检测和分类，发现体细胞突变模式具有以下两个方面的特征：(1) 在一个驱动通路内部，一个驱动基因突变通常干扰整个通路，结合驱动突变相对较少的事实，大多数病人在一个通路中仅展示出一个驱动突变，即突变基因互斥性；(2) 一个重要的致癌通路应该在多数癌症病人中受干扰，也就是说，大部分病人在这个通路中至少有一个基因发生突变，即整个通路具有高覆盖性[10,26,28]. 互斥的基因集合通常共同执行某一生物功能，这些基因集合在一定数量的样本中发生突变才具有生物意义[17,23,27]. 假定存在一个具有 $m\times n$ 的二元突变矩阵 $\boldsymbol{A}$，m 表示病人样本个数，n 表示基因个数，$A_{i,j}=1$ 表示样本 s_i 中的基因 g_j 发生突变，否则 $A_{i,j}=0$，如图 1 所示. 在实际生物测序数据处理中，由于误差的存在干扰着互斥性和覆盖性，因此本算法将通过近似互斥思想有效的解决这一问题.

对于基因 g，覆盖函数 $\Gamma(g)=\{i:A_{ig}=1\}$ 表示基因 g 发生突变的病人集合. 对于突变矩阵 $\boldsymbol{A}$ 中的 $m\times k$ 子阵列 M，覆盖函数 $\Gamma(M)=\bigcup\limits_{g\in M}\Gamma(g)$ 表示 k 个基因中发生突变的病人集合. 对于任意一对基因 g_i，

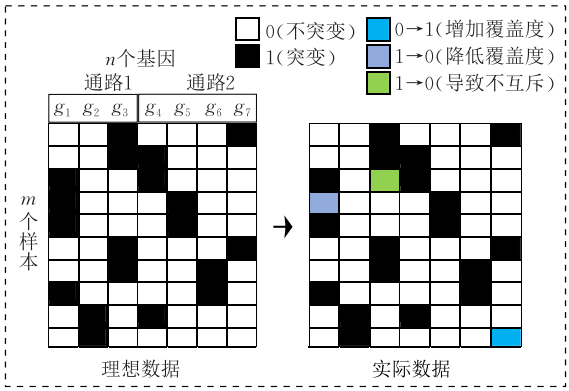

图 1　突变矩阵

$g_j\in M$，$g_i\neq g_j$，$1\leq i,j\leq n$，如果 $\Gamma(g_i)\cap\Gamma(g_j)=\varnothing$，则 M 中基因是互斥的. 突变矩阵中具有高互斥和高覆盖的基因集合表示一个突变驱动通路. 互斥度和覆盖度函数定义如下：

对于突变矩阵 $\boldsymbol{A}$ 中的 $m\times k$ 子阵列 M，M 中基因间互斥度函数 $ED(M)$ 表示为

$$ED(M)=\frac{|\Gamma(M)|}{\sum\limits_{g\in M}|\Gamma(g)|}\tag{1}$$

对于突变矩阵 $\boldsymbol{A}$ 中的任一基因对 g_i，g_j，其互斥度函数 $ED(g_i,g_j)$ 表示为

$$ED(g_i,g_j)=\frac{|\Gamma(g_i)\cup\Gamma(g_j)|}{|\Gamma(g_i)|+|\Gamma(g_j)|}\tag{2}$$

上述分析可知，如果 $ED(M)=1$，子阵列 M 中的基因是互斥的，也就是说，M 中每位病人最多只有一个基因发生突变. 同理，如果 $ED(g_i,g_j)=1$，则基因对 g_i，g_j 是互斥的.

对于突变矩阵 $\boldsymbol{A}$ 中的 $m\times k$ 子阵列 M，M 中基因间覆盖度函数 $CD(M)$ 表示为

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

对于突变矩阵 $\boldsymbol{A}$ 中的任一基因对 g_i，g_j，其覆盖度函数 $CD(g_i,g_j)$ 表示为

$$CD(g_i,g_j)=\frac{|\Gamma(g_i)\cup\Gamma(g_j)|}{m}\tag{4}$$

如果 $CD(M)=1$，则子阵列 M 是完全覆盖的，即对于子阵列 M 中的基因，所有样本中至少有一个基因发生突变.

由以上公式可知，最理想的情况是：在构建突变基因网络时，基因对 g_i，g_j 的互斥度 $ED(g_i,g_j)=1$；在检测突变驱动通路时，能够检测到互斥度 $ED(M)=1$ 且覆盖度 $CD(M)=1$ 的基因集合 M. 但在实际生物数据中，由于测序和计算时存在误差，部分基因对的互斥度接近 1，却通常共同行使某一生

物功能,这些基因对中两个基因的覆盖度相对比较接近时,才能达到驱动通路的标准[26]. 图 2 给出了具有相同互斥度和覆盖度的基因对,可能存在一个基因覆盖包含于另一个基因覆盖的可能. 因此,我们定义了覆盖重叠、权重函数和非重叠比重函数如下:

对于突变矩阵 $\boldsymbol{A}$ 中的 $m\times k$ 子阵列 M,覆盖重叠函数 $\omega(M)$ 表示为

$$\omega(M)=\sum_{g\in M}|\Gamma(g)|-|\Gamma(M)| \tag{5}$$

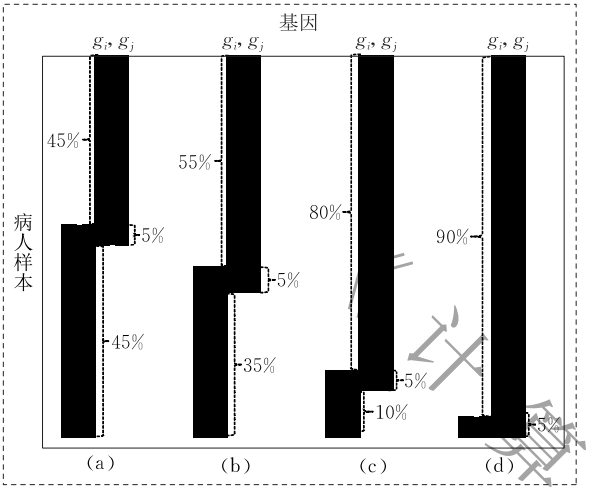

图 2 近似互斥类型分析

对于突变矩阵 $\boldsymbol{A}$ 中的任一基因对 g_i,g_j,这两个基因间覆盖重叠函数 $\omega(g_i,g_j)$ 表示为

$$\omega(g_i,g_j)=|\Gamma(g_i)\cap\Gamma(g_j)| \tag{6}$$

对于突变矩阵 $\boldsymbol{A}$ 中的 $m\times k$ 子阵列 M,考虑到覆盖度 $CD(M)$ 和覆盖重叠 $\omega(M)$ 两个方面的因素,权重函数 $WD(M)$ 表示为

$$WD(M)=1-\frac{\omega(M)}{CD(M)} \tag{7}$$

对于突变矩阵 $\boldsymbol{A}$ 中的任一基因对 g_i,g_j,考虑到基因覆盖 $\Gamma(g_i),\Gamma(g_j)$ 和覆盖重叠 $\omega(g_i,g_j)$ 两个方面的因素,非重叠比重函数 $RD(g_i,g_j)$ 表示为

$$RD(g_i,g_j)=1-\frac{|\Gamma(g_i)\cap\Gamma(g_j)|}{\min\{|\Gamma(g_i)|,|\Gamma(g_j)|\}} \tag{8}$$

构建突变基因网络时,使用非重叠比重函数可以避免图 2(c)和(d)的情况发生,只取图 2(a)和(b)作为近似互斥的基因对和 $ED(g_i,g_j)=1$ 的互斥基因对. 图 2 中 4 种情况具有相同的覆盖度 95%、覆盖重叠度 5%和互斥度 95%,具有不同的非重叠比重函数分别是(a)$RD(g_i,g_j)=90\%$,(b)$RD(g_i,g_j)=87.5\%$,(c)$RD(g_i,g_j)=66.7\%$,(d)$RD(g_i,g_j)=0$.

3.2 突变基因网络构建

Leiserson 和 Vandin 等人[8,16]引入了最大权重子阵列问题,即检测满足具有最大权重 $W(M)$ 的子阵列 M,该问题是 NP 难问题. 为了更有效地解决此问题,对于体细胞突变矩阵,本文利用式(2)计算任意一对基因间的互斥度,利用式(8)计算该对基因间的非重叠比重值,如果 $ED(g_i,g_j)\geq 0.95$ 且 $RD(g_i,g_j)\geq 0.85$,则节点 g_i,g_j 间连边,否则不连边,以此建立基因相互作用网络. 此基因网络中,每个节点代表一个基因,每条连边代表这对基因满足互斥关系,如果一个基因和其它基因都没有连接关系,则该基因不出现在网络中. 该过程的算法描述如下.

算法 1. 突变基因网络构建.

输入:m 位病人样本 n 个基因的突变矩阵 $\boldsymbol{A}$

输出:n 行 n 列的基因网络 G

```
1.  Initialization: x←m; y←n; i←1;
2.  DO WHILE
3.    FOR j←i+1 to y DO
4.      exclusive_degree=ED(i,j);
5.      rate_degree=RD(i,j);
6.      IF(exclusive_degree >= 0.95 and
              rate_degree >= 0.85)
7.        THEN genenetwork[i][j]=1;
8.      END IF
9.    END FOR
10. i=i+1;
11. END WHILE (i>=y);
```

该算法通过构建突变基因网络,简化基因间相互关联关系,从而降低了算法复杂度. 突变基因网络构建和驱动通路检测过程如图 3 所示.

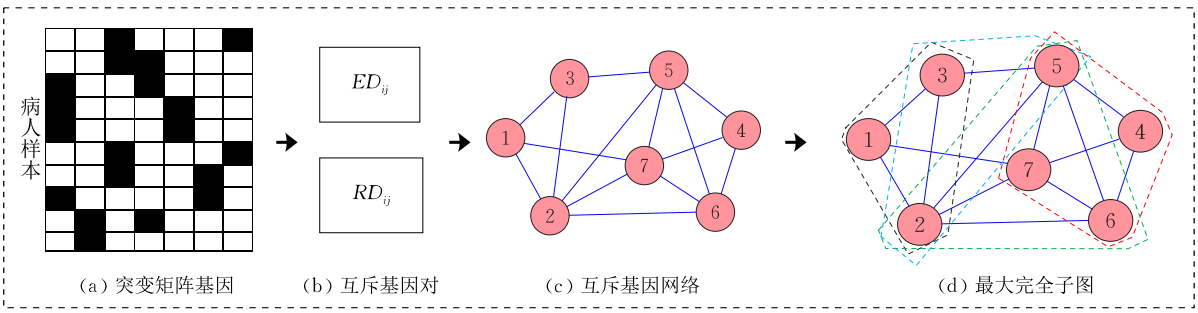

图 3 突变驱动通路检测过程

3.3　突变驱动通路子图挖掘算法

上述构建的基因网络中，任意两个相连的基因是互斥的，因此，该基因网络中任意一个完全子图中的基因集合都是互斥的. 我们首先找到具有最大覆盖且相互连接的 3 个基因作为种子基因集合，对于基因集合外任意一个节点，如果该节点和基因集合中每个节点相连，且具有最大的覆盖度，则将该节点加入到基因集合中，直到不存在与基因集合中所有节点相连的基因，则该过程结束，以检测出最大完全子图. 对于该基因集合中的每个节点，如果删除某个节点，基因集合的权重函数值增加，则从基因集合中删除该节点. 查找一个突变驱动通路过程的算法描述如下.

算法 2.　突变驱动通路检测.

输入：突变矩阵 $\boldsymbol{A}$ 和基因网络 G

输出：突变驱动通路

1. Initialization: $x=n$, $t=0$;
2. $V_t=\{i,j,k\}$;
3. DO WHILE
4. FOR $i\leftarrow 1$ to x DO
5. IF (V_t and $\{i\}$ all connection and having max coverage)
6. THEN $V_{t+1}=V_t\cup\{i\}$;
7. END IF
8. $t=t+1$;
9. END WHILE (no new node added)
10. FOR $i\leftarrow 1$ to $|V_t|$ DO
11. IF ($WD(V_t)<=WD(V_t-\{i\})$) THEN
12. $V_t=V_t-\{i\}$; //delete node i
13. END IF
14. END FOR
15. IF ($CD(V_t)\geq 0.3$ and $|V_t|\geq 3$) THEN
16. V_t is a MDP;
17. END IF

该过程可以找出最优的一组突变驱动通路. 如果可以找到一组相互连接的 3 个基因，而且它们具有最高的覆盖度且不包含于任何一组已检测出的驱动通路中，则重复上述过程进行迭代. 反之，若不能找出满足条件的种子基因集合，则终止驱动通路检测. 若该基因集合中所有基因的覆盖度大于等于 0.3 且基因个数大于等于 3，这个基因集合就是一个突变驱动通路. 如果一个基因集合的覆盖度太低或者基因数量少于 3，通常认为不是一个突变驱动通路[8,17].

3.4　Megnet 算法时间复杂度分析

由算法 1 可知，计算两个基因间互斥度和非重叠比重值的时间复杂度为 $O(m)$，选择两个基因的时间复杂度为 $O(n^2)$，因此构建突变基因网络的时间复杂度为 $O(m\cdot n^2)$. 由算法 2 可知，检测种子基因集合的时间复杂度为 $O(n^3)$，聚类过程的时间复杂度是 $O(n^3)$. 因此 Megnet 算法总的时间复杂度为 $O((m+n)\cdot n^2)$，而 Dendrix 算法的时间复杂度为 $O(m\cdot n^2\cdot\exp(n^{(k-1)/k})$，$k$ 表示一组驱动通路中包含的基因个数，Multi-Dendrix 算法的时间复杂度为 $O(m^2\cdot n^4)$.

4　实验验证和结果分析

为验证算法的效率和鲁棒性，我们将该算法应用于模拟体细胞突变数据中，检测预设于模拟数据中的通路，并将 Megnet 算法的运行时间和结果准确率与 Dendrix 和 Multi-Dendrix 算法进行比较. 为验证算法的有效性，将 Megnet 算法分别应用于真实的肺癌和神经胶质瘤体细胞突变数据中，检测出导致癌症发生的突变通路. 同时，将 Megnet 算法检测的突变驱动通路和 Dendrix 和 Multi-Dendrix 算法的结果进行比较，分析每个驱动通路内基因间相互作用关系，并验证每个驱动通路的生物相关性、统计显著性及其与已知重要信号通路的关系.

4.1　模拟数据

为了验证算法的效率和鲁棒性，我们生成 $m=300$ 位病人样本，$n=1000$ 个基因的突变数据，并嵌入 15 组通路 $P=\{P_1,P_2,\cdots,P_{14},P_{15}\}$，其中 4 组通路各包含 4 个基因，3 组通路各包含 6 个基因，4 组通路各包含 8 个基因，4 组通路各包含 10 个基因，初始生成每组通路的覆盖度 $CD(P_i)$ 依次为 0.95，0.92，0.89，0.86，0.83，0.80，0.77，0.74，0.71，0.68，0.65，0.62，0.59，0.56，0.53. 这些数据集的大小和覆盖度的取值依据常见的真实生物突变数据. 对于每一组通路 P_i，随机选取 $|\Gamma(P_i)|$ 位病人样本，只把这组通路 P_i 内的某一个基因设置为驱动突变. 因此，在每组通路 P_i 中，驱动基因间都是互斥的. 在真实数据中，由于测序和计算误差，总是存在一些噪声，因此，我们分别以 $q=0.0001$，0.0005，0.001，0.005，0.01，0.015，0.02 的概率设置乘客突变. 对于每个乘客突变概率 q，分别生成 20 组突变数据. 随着噪声加入到这些数据中，上述给出的覆盖度会略有变化. 这个模拟过程在 64 位操作系统、2.50 GHz 处理器和 8 GB 内存的计算机上运行，在不同的乘客突变概率下，Megnet 算法在 15 s 内完成

驱动通路检测，Multi-Dendrix 算法的运行时间约为 Megnet 算法的 2 倍，而 Dendrix 算法具有最长的运行时间. 表 1 给出 3 种算法分别在不同乘客突变概率 q 下检测驱动通路需要的时间，时间取值来自于每种算法在一种乘客突变率下对 20 组突变数据操作的平均值.

表 1 3 种算法分别在模拟数据中的运行时间

乘客突变概率 q	运行时间/s		
	Dendrix	Multi-Dendrix	Megnet
0.0001	963.42	15.28	9.65
0.0005	1045.23	19.74	9.58
0.001	1235.81	21.34	10.25
0.005	1345.79	25.08	11.02
0.01	1538.51	29.63	12.57
0.015	1749.58	35.56	13.45
0.02	2135.59	40.95	14.38

在模拟的体细胞突变数据中，本算法与 Dendrix 和 Multi-Dendrix 算法在不同的乘客突变率下对算法的运行时间和结果的正确率方面进行比较. 对于 Dendrix 算法，在不同的突变率下，使用迭代方式运行，每次找出一组权重分数最高的基因集合，然后从数据集中移除这些基因，再次运行 Dendrix 算法，重复上述步骤 15 次以得到 15 组最优的基因集合.

图 4 给出 3 种算法分别在不同乘客突变率 q 下检测驱动通路的正确率，正确率取值来自于每种算法在一种乘客突变率下对 20 组突变数据操作的平均值. 当乘客突变率小于等于 0.01 时，Megnet 算法能检测到完全正确的基因集合. 随着噪声增加，检测驱动通路的正确率有所下降. 如果噪声低于临界值，我们就能得到理想的驱动通路. Multi-Dendrix 和 Megnet 算法得到结果的正确率相近，即当突变率不高于 0.015 时，这两种算法都能得到较理想的结果. 而 Dendrix 算法只有在乘客突变率低于 0.001 时能得到较理想的结果. 由此可见在不同乘客突变率下 Megnet 算法在正确率和运行时间方面优于其它两种算法.

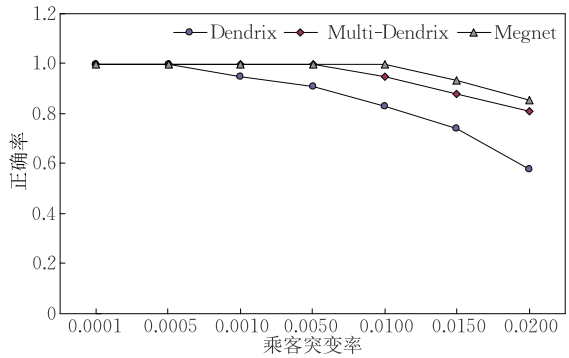

图 4 不同噪声概率下检测驱动通路的正确率

当乘客概率 q 为 0.001 时，生成 $m=300$ 位病人样本，n 分别为 1000，2000，3000 个基因的突变数据，所嵌入的通路个数、每条通路包含的基因个数及每条通路的覆盖率和上述情况一致，并生成 20 组突变数据. 表 2 给出了 Megnet、Dendrix 和 Multi-Dendrix 算法在 20 组突变数据上运行的平均时间. 由这些结果可知，当基因数目增大到 2000 或 3000 时，Megnet 算法呈现出较高的效率.

表 2 3 种算法在不同基因个数下的运行时间

基因数目	运行时间/s		
	Dendrix	Multi-Dendrix	Megnet
1000	1235.81	21.34	10.25
2000	3638.56	158.95	29.19
3000	8468.45	865.76	71.75

4.2 肺癌数据(LC)

近些年，肺癌发病率与死亡率增长速度很快，对人类健康和生命威胁产生巨大影响的恶性肿瘤之一[6]. Vandin 等人[16]分析了来自 188 位肺癌患者的 1013 组体细胞突变数据，发现 356 个基因至少在一位患者中发生突变. 该突变矩阵包括 163 个样本，356 个基因，所有样本的最大突变基因数为 64，每个样本的平均突变基因数为 6.0，所有基因的平均突变样本数为 2.75. 其中只有 4 个基因(TP53，KRAS，STK11 和 EGFR)的突变样本数多于 20，它们的突变样本数分别是 64，60，34 和 30，因此该矩阵为稀疏矩阵，所以本文选取 $MAF>1.5\%$ 的基因进行操作，即对 163 位病人 95 个基因分别运行 Megnet、Dendrix 和 Multi-Dendrix 这 3 种算法.

对于 Dendrix 算法[16]，参数 k 表示一组驱动通路中包含基因的个数. 当 $k=2$ 时，得到两组最优突变基因集合(EGFR，KRAS)和(TP53，ATM)，第 1 组基因集合的突变样本数为 90，因此其覆盖率为 55.2%；第 2 组基因集合的突变样本数为 76，因此其覆盖率为 46.6%. 当 $k=3$ 时，得到一组最优基因集合(EGFR，KRAS，STK11)，该基因集合的突变样本数为 109，因此其覆盖度为 66.9%，但该基因集合的重叠样本数为 15，由此可以看出，该集合中增加一个基因 STK11，产生了较多的重叠. 当 $k\geq 4$ 时，没有检测到最优基因集合.

对于 Multi-Dendrix 算法[8]，需要设置两个参数 t 和 $k_{\max}$，t 表示生成基因集合的个数，$k_{\max}$ 表示一个基因集合中包含基因的最大个数. 当 $t=2$，$k_{\max}=3$ 时，得到基因集合(EGFR，KRAS，STK11)和(TP53，ATM，PAK4)；当 $t=3$，$k_{\max}=4$ 时，得到基因集合(KRAS，EGFR，NTRK3，EPHB1)、(TP53，

ATM,PAK4,ACVR1B)和(STK11,LRP1B,NF1,NTRK1);当 $t=4,k_{max}=5$ 时,得到基因集合(KRAS,EGFR,ERBB4,PRKCG,NRAS)、(TP53,ATM,PAK4,ACVR1B,PAK6)、(STK11,NF1,MYO3B,NTRK1,PTEN)和(LRP1B,CDKN2A,GNAS,NTRK3,MSH6).对于模拟数据,已知嵌入到数据集中驱动通路个数以及每组驱动通路中的基因个数,Multi-Dendrix 算法能够得到较理想的结果.但对于真实的突变数据,不能事先知道驱动通路个数和每组驱动通路中基因个数,所以很难精确设置参数 t 和 k_{max} 的值以得到合理的突变驱动通路.

与 Dendrix 和 Multi-Dendrix 算法不同,Megnet 算法不需要指定驱动通路的数目以及每组驱动通路中包含的基因数目,只需要根据数据本身的特征进行聚类,产生 4 组最优基因集合如表 3 所示.对于每组检测的驱动通路,使用置换检验(Permutation Test)检测驱动通路的互斥程度.对于基因 g,$Cover(g)$ 表示基因 g 突变的样本数目,对于一组驱动通路 A,$Cover(A)=\bigcup_{g\ in\ A} Cover(g)$.Megnet 算法根据 $|Cover(A)|$ 是否比置换实例覆盖样本数目大,来判断这组基因集合的互斥性.对每组基因集合,执行 10000 次置换,得到置换检验的 P 值如表 3 所示.

表 3　肺癌数据实验结果

最优基因集合	基因数量	互斥度	覆盖度	置换检验 P 值
TP53,STK11,LRP1B	3	0.952	0.601	0.238
STK11,NF1,EPHA7	3	0.964	0.485	0.195
KRAS,LRP1B,EPHA3,MYO3B	4	0.952	0.368	0.113
EGFR,STK11,LRP1B,PRKDC,EPHA7	5	0.958	0.417	0.063

Dendrix 和 Multi-Dendrix 算法选取不同的参数组合 $2\leq t\leq 4$ 和 $3\leq k_{max}\leq 5$,对得到的 27 组基因集合进行置换检验,得到基因集合(EGFR,KRAS,STK11)的 P 值(0.098)最小,得到基因集合(TP53,ATM,PAK4,ACVR1B,BAP1)的 P 值(0.312)最大,这 27 组基因集合的平均 P 值为 0.206.而 Megnet 算法检测的 4 组最优基因集合中,置换检验最小 P 值为 0.063,最大 P 值为 0.238,平均 P 值为 0.152.可以看出,Megnet 算法检测的基因集合具有较高的互斥度和统计显著性,且不需要事先指定驱动通路个数及每组驱动通路中包含的基因个数,因此 Megnet 算法检测驱动通路的性能优于 Dendrix 和 Multi-Dendrix 算法.

4.3　神经胶质瘤(GBM)

神经胶质瘤是一种常见的浸润性很强的原发性中枢神经系统肿瘤,是所有神经上皮来源的肿瘤,约占所有颅内原发肿瘤的一半.原发性胶质母细胞瘤的分子改变以表皮生长因子受体(EGFR)的扩增和过量表达为主,而继发性胶质母细胞瘤则以 p53 突变为主[29-33].

为了进一步验证算法的有效性,我们选择真实的神经胶质瘤体细胞突变数据.实验数据集来源于 TCGA[34] 中的单核苷酸变异和 DNA 拷贝数变异数据.该突变矩阵包括 362 个样本和 18009 个基因,选取 $MAF>2.5\%$ 的基因后,保留 344 个基因,所有基因中突变样本数最多的是 113,突变样本数大于 70 的基因有 TTN,PETN,MUC16,TP53,DUX4L19 和 EGFR,突变样本数分别是 113,101,92,85,73 和 71 次.数据详细信息如表 4 所示.

表 4　神经胶质瘤数据信息

	病人数量	基因数量	样本平均突变基因数	基因平均突变样本数
所有基因	362	18009	189.00	3.79
选取 $MAF>1.5\%$ 的基因	362	1076	54.14	18.21
选取 $MAF>2.0\%$ 的基因	362	484	29.55	22.10
选取 $MAF>2.5\%$ 的基因	362	344	26.64	28.04
选取 $MAF>3.0\%$ 的基因	362	171	14.94	31.62

基于上述神经胶质瘤数据集,对 362 位病人样本包含 344 个基因分别运行 Megnet、Dendrix 和 Multi-Dendrix 这 3 种算法,Megnet 算法得到 9 组最优的基因集合,而 Dendrix 和 Multi-Dendrix 算法分别得到 4 组最优的基因集合,表 5 给出了 3 种算法分别得到的最优基因集合,以及每组最优基因集合对应的基因个数、基因间互斥度、权重度、覆盖度和 P 值.Megnet 和 Dendrix 算法都检测出最优基因集合(CDK4,CDKN2B,RB1),其互斥度、权重度和覆盖度分别为 0.956、0.954 和 0.722,而 Multi-Dendrix

算法检测出的最优基因集合(CDK4,CDKN2B,RB1,CDKN2A)包含基因 CDKN2A,使得他们的互斥度和权重度分别降低到 0.579 和 0.318,而覆盖度仅增加了 0.011,说明基因 CDKN2A 和 CDK4,CDKN2B,RB1 这 3 个基因的互斥度很低,并且基因集合(CDK4,CDKN2B,RB1)和(CDK4,CDKN2B,RB1,CDKN2A)的 P 值均为 2.50e-04,因此,(CDK4,CDKN2B,RB1)作为最优基因集合是合理的.

表 5 神经胶质瘤实验结果

检测算法	最优基因集合	基因数量	互斥度	权重度	覆盖度	P 值
Megnet	MDM2,CPT1B,PIK3R1,DST,MET,PIK3CA	6	0.971	0.970	0.367	9.00e-06
	TP53,QKI,MDM2,MDM4	4	0.932	0.912	0.456	2.74e-05
	EGFR,PDGFRA,PIK3R1,COL6A2	4	1.000	1.000	0.367	4.90e-05
	CYP27B1,NF1,EGFR,PIK3R1	4	0.902	0.913	0.511	5.32e-05
	TP53,CDKN2A,CDK4,RB1	4	0.824	0.787	0.834	6.20e-05
	MDM2,TP53,PIK3R1,CPT1B	4	0.889	0.900	0.445	4.30e-04
	PTEN,PIK3CA,EGFR	3	0.867	0.846	0.577	9.20e-05
	PIK3CA,PTEN,PIK3R1	3	0.980	0.979	0.533	9.20e-05
	CDK4,CDKN2B,RB1	3	0.956	0.954	0.722	2.50e-04
Dendrix	CDKN2B,RB1,CYP27B1	3	0.943	0.939	0.733	3.30e-02
	CDK4,CDKN2B,RB1	3	0.956	0.954	0.722	2.50e-04
	TP53,CDKN2A	2	0.901	0.891	0.711	7.50e-03
	NF1,EGFR	2	0.968	0.968	0.333	5.40e-02
Multi-Dendrix	CDK4,CDKN2B,RB1,CDKN2A	4	0.579	0.318	0.733	2.50e-04
	PTEN,PIK3CA,PIK3R1,IDH1,PDPN,PRDM2	6	0.979	0.979	0.533	1.10e-03
	TP53,MDM2,MDM4,NLRP3,AKAP6,NPAS3	6	0.974	0.973	0.411	4.50e-04
	EGFR,PDGFRA,RB1	3	0.968	0.967	0.333	1.60e-04

表皮生长因子受体 EGFR 的增殖和过量表达是导致原发性胶质母细胞瘤发生的主要原因[35].Megnet 算法检测的基因集合(EGFR,PDGFRA,PIK3R1,COL6A2)的互斥度和权重度都为 1,具有完全互斥性,这组基因集合的 P 值为 4.90e-05,且全部是 PI3K 信号通路的核心成员;Dendrix 算法检测的基因集合(EGFR,NF1)的 P 值为 5.40e-02;Multi-Dendrix 算法检测的基因集合(EGFR,PDGFRA,RB1)的 P 值为 1.60e-04,因此 Megnet 算法检测的基因集合(EGFR,PDGFRA,PIK3R1,COL6A2)具有更高的生物相关性和统计显著性.

肿瘤抑制基因 TP53 突变是导致继发性胶质母细胞瘤发生的主要原因[36],Megnet 算法检测的基因集合(TP53,QKI,MDM2,MDM4)的 P 值为 2.74e-05,(TP53,CDKN2A,CDK4,RB1)的 P 值为 6.20e-05;Dendrix 算法检测的基因集合(TP53,CDKN2A)的 P 值为 7.50e-03,是基因集合(TP53,CDKN2A,CDK4,RB1)的子集;Multi-Dendrix 算法检测的基因集合(TP53,MDM2,MDM4,NLRP3,AKAP6,NPAS3)的 P 值为 4.50e-04,和基因集合(TP53,QKI,MDM2,MDM4)重叠 3 个基因(TP53,MDM2,MDM4).在 p53 信号通路中,这 3 个基因存在紧密的相互关联关系,TP53 激活 MDM2,MDM2 和 MDM4 存在相互作用关系,MDM2 和 MDM4 共同抑制 TP53,并且 TP53 直接调控 QKI 的基因表达.因此 Megnet 算法检测的基因集合(TP53,QKI,MDM2,MDM4)具有更强的生物相关性和统计显著性.

Megnet 算法检测的基因多位于癌症的重要通路中,如 p53、RB、RAS/PI3K 信号通路以及细胞循环和细胞凋亡通路.该算法检测的多组基因集合属于同一个通路,比如第 4 组基因集合(CYP27B1,NF1,EGFR,PIK3R1)是 RAS/PI3K 信号通路的核心成员;第 5 组基因集合(TP53,CDKN2A,CDK4,RB1)是细胞循环和细胞凋亡通路的核心成员;第 9 组基因集合(CDK4,CDKN2B,RB1)是 RB 信号通路的核心成员.这些信号通路对癌症的发生、发展和恶化起着至关重要的作用.

Megnet 算法检测的 9 组最优基因集合比 Dendrix 和 Multi-Dendrix 算法分别检测的 4 组最优基因集合的 P 值要小很多,说明 Megnet 算法检测的最优基因集合具有更高的统计显著性和生物相关性.

这 9 组互斥的突变驱动通路中,每组通路都具有较高的互斥度和较强的统计显著性和生物相关性(P 值$<$0.01),每组通路的 P 值均通过 DAVID 功能标注工具(https://david.ncifcrf.gov/tools.jsp)得到.所有通路共涉及 21 个驱动基因,所有这些基因突变都发生在细胞存活状态下,其中原癌基因 6 个,分别为 MDM2,MET,PIK3CA,MDM4,EGFR,

PDGFRA，占所有基因的 28.6%. 原癌基因是细胞内与细胞增殖相关的基因，当原癌基因的结构或调控区发生突变，基因产物增多或活性增强时，使细胞过度增殖，从而形成肿瘤[37-39]. 其余 15 个基因是肿瘤抑制基因或抗体，占所有基因的 71.4%. 抑癌基因是一类抑制细胞过度生长、增殖从而遏制肿瘤形成的隐性基因，这些基因突变体现在基因的缺失或失活，不能调控原癌基因增殖，从而不能协调配合原癌基因共同维持细胞的正常活动[40-42]. 图 5 给出了 Megnet 算法检测的 9 组突变驱动通路基因间的相互作用关系、基因的性质、基因和已知信号通路的关系、以及每组突变驱动通路的 P 值、互斥度、权重度和覆盖度等. 每组突变驱动通路的详细说明如下：

第 1 组驱动通路包括 MDM2 和 MET 的增殖、CPE1B 和 DST 的缺失、PIK3R1 的插入和 PIK3CA 的删除，这个模块中 36.7%(133/362)的样本发生突变. 其中 MDM2，MET，PIK3CA 是原癌基因，是 PI3K 信号通路的核心成员，PI3K 是神经胶质瘤中最显著突变的信号通路，其中 PIK3R1 的肿瘤抑制基因分数是 37%，而 MET 和 PIK3CA 的致癌基因分数分别是 61%和 95%[43]. MDM2 是细胞循环和细胞凋亡通路的核心成员，通过细胞增殖导致癌变的原癌基因. CPT1B，PIK3R1，DST 是抑癌基因，PIK3R1 也是 PI3K 信号通路的核心成员，CPT1B 是 PPARA 通路中的成员，DST 是血小板溶素家族成员，主要在中枢神经系统中转录[44]，基因 CPT1B 和 DST 还没有被报道在神经胶质瘤中起重要的作用. 在 PI3K 信号通路中，PIK3R1 和 PIK3CA 存在相互作用关系，并且共同激活 MDM2，MDM2 抑制 p53 信号通路，MET 能够激活 PI3K 信号通路. 这组基因集合具有较高的互斥度(0.971)和权重度(0.970)，P 值是 9.00e-06，具有很强的生物相关性. 驱动通路中基因间相互作用关系如图 5(a)所示.

第 2 组驱动通路包括 MDM2 和 MDM4 的增殖、QKI 的删除和 TP53 的缺失，这个模块中 45.6%(165/362)的样本发生突变. 其中 MDM2 和 MDM4 是细胞循环和细胞凋亡通路的核心成员，是通过细胞增殖导致癌变的原癌基因. TP53 是细胞循环和细胞凋亡以及 p53 等多个信号通路的核心成员，是人类癌症最常见的突变肿瘤抑制基因，它调控多种重要的生物活动，并通过转录后修饰进行自身调控[45-46]. TP53 的肿瘤抑制基因分数是 20%，而致癌基因分数是 73%[43]，这充分说明 TP53 的表达和突变在癌症发生、发展和恶化过程中起着非常重要的作用. QKI 是神经胶质瘤中新发现的一个抑制基因，是信号传导和激活 RNA 绑定蛋白质的 RNA 家族核心成员之一. 在 p53 信号通路中，TP53 直接调控 QKI 的基因表达，TP53 激活 MDM2，MDM2 和 MDM4 共同抑制 TP53，并且 MDM2 和 MDM4 存在相互作用关系. 这组基因集合具有较高的互斥度(0.932)和权重度(0.912)，P 值是 2.74e-05，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(b)所示.

第 3 组驱动通路包括 PIK3R1 和 COL6A2 的插入、EGFR 的删除和 PDGFRA 的增殖，这个模块中 36.7%(133/362)的样本发生突变. 这 4 个基因 EGFR、PDGFRA、PIK3R1 和 COL6A2 是 PI3K 信号通路的核心成员，其中 EGFR 和 PDGFRA 是原癌基因，它们的致癌基因分数分别是 97%和 84%[43]，而 PIK3R1 和 COL6A2 是抑制基因，PIK3R1 的肿瘤抑制基因分数是 37%[43]. EGFR 基因表达和源于细胞的肿瘤特性密切相关；PDGFRA 参与胚胎发育、血管生成、细胞增殖和变异等多种生物过程；PIK3R1 是磷酸化脂质激酶家族成员之一，负责协调细胞增殖和细胞生存等功能转换；COL6A2 在人类大脑皮层和肌肉活组织中表达，对肌阵挛性癫痫疾病起重要作用[47]. 在 PI3K 信号通路中，EGFR 激活 PIK3R1. 这组基因集合是完全互斥的，互斥度和权重度值都为 1，P 值是 4.90e-05，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(c)所示.

第 4 组驱动通路包括 PIK3R1 的插入、EGFR 的删除、NF1 的缺失和 CYP27B1 的增强，这个模块中 51.1%(185/362)的样本发生突变. 其中 EGFR 和 PIK3R1 是 PI3K 信号通路的核心成员，其中 EGFR 是原癌基因，致癌基因分数是 97%；PIK3R1 是抑制基因，肿瘤抑制基因分数是 37%[43]. EGFR 和 NF1 是 RAS 信号通路的核心成员，其中 NF1 是抑制基因，肿瘤抑制基因分数是 73%[43]. CYP27B1 和新陈代谢通路相关，在正常骨骼生长、钙质新陈代谢和组织变异中扮演重要的角色，还没有被报道在神经胶质瘤中起重要的作用[47]. 在 PI3K 信号通路中，EGFR 激活 PIK3R1. 在 RAS 信号通路中，NF1 抑制 RAS，RAS 和 EGFR 存在相互作用关系. 这组基因集合的互斥度和权重值分别为 0.902 和 0.913，P 值是 5.32e-05，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(d)所示.

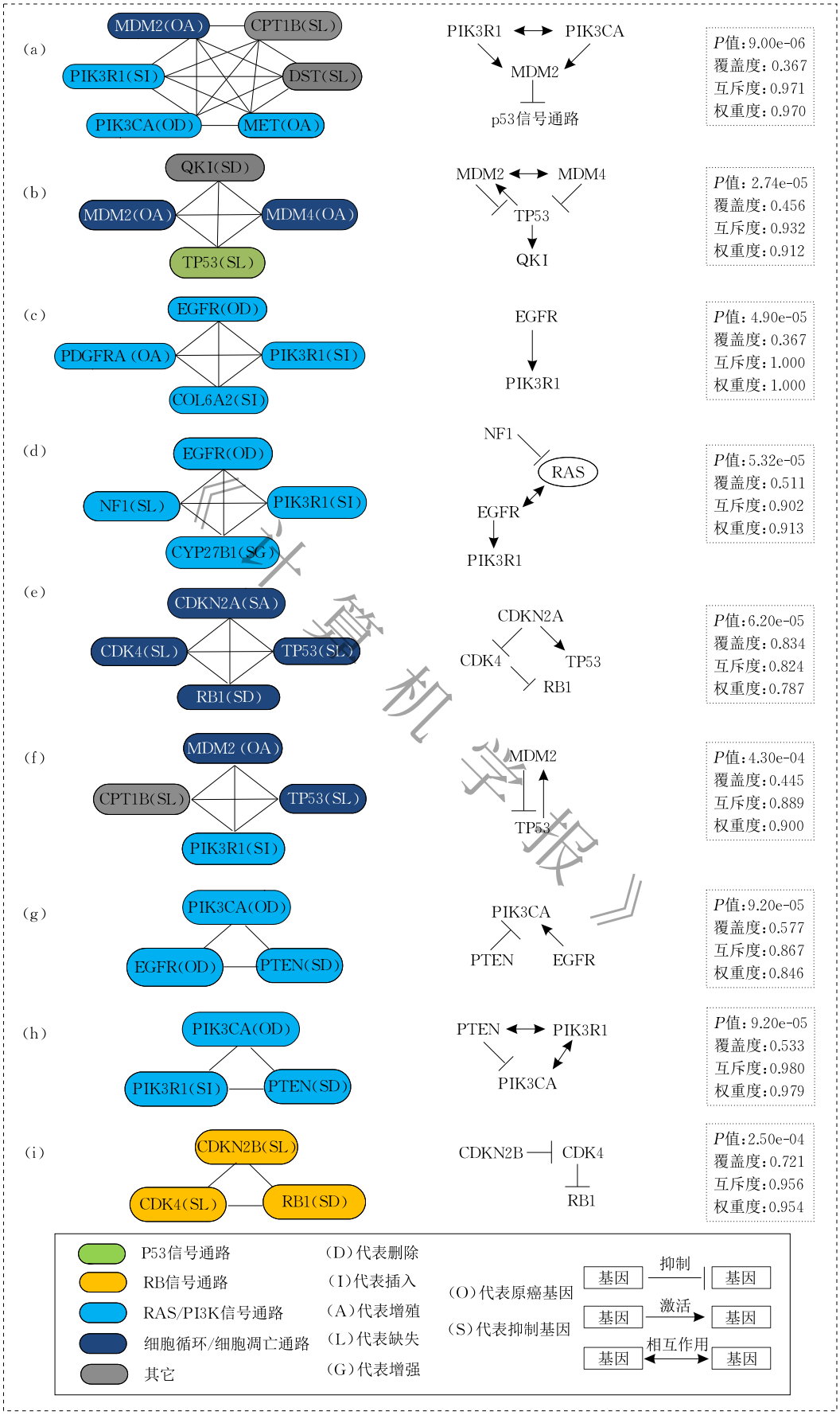

图 5 突变驱动通路内基因间相互关系

第 5 组驱动通路包括 TP53 和 CDK4 的缺失、RB1 的删除和 CDKN2A 的增殖，这个模块中 83.4%(302/362)的样本发生突变. 这 4 个基因 TP53、CDKN2A、CDK4 和 RB1 是信号通路 p53 和 RB 等的核心成员，参与细胞循环和细胞凋亡的整个过程，且全部是重要的抑制基因，其肿瘤抑制基因分数分别为 20%、49%、53% 和 80%[43]. CDKN2A 和 CDK4、CDK6 具有紧密的相互作用关系，共同作为正常细胞增殖的负调控基因，在多种癌症中频繁执行突变和删除. RB1 被钙调磷酸酶脱去磷酸，因此它通过 CDK4 和 CDK6 磷酸化来调节细胞循环和细胞凋亡[47]. 在 RB 信号通路中，CDKN2A 抑制 CDK4 且激活 TP53，CDK4 抑制 RB1. 这组基因集合的互斥度和权重值分别为 0.824 和 0.787，具有最高的覆盖度 0.834，P 值是 6.20e-05，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(e)所示.

第 6 组驱动通路包括 TP53 和 CPT1B 的缺失、PIK3R1 的插入和 MDM2 的增殖，这个模块中 44.5%(161/362)的样本发生突变. 其中 TP53 和 MDM2 是细胞循环和细胞凋亡通路的核心成员，PIK3R1 是 PI3K 信号通路的核心成员. TP53 和 PIK3R1 是抑制基因，肿瘤抑制基因分数分别是 20% 和 37%[43]，而 MDM2 是通过增殖突变的原癌基因. 在 p53 信号通路中，MDM2 抑制 TP53，TP53 激活 MDM2. 这组基因集合的互斥度和权重值分别为 0.889 和 0.900，P 值是 4.30e-04，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(f)所示.

第 7 组驱动通路包括 PIK3CA、EGFR 和 PTEN 的删除，这个模块中 57.7%(209/362)的样本发生突变. 这 3 个基因 PTEN、PIK3CA 和 EGFR 是 PI3K 信号通路的核心成员，其中 PIK3CA 和 EGFR 是原癌基因，致病基因分数分别是 95% 和 97%[43]；PTEN 在多种癌症中发生高突变的抑制基因，肿瘤抑制基因分数是 55%[43]. 在 PI3K 信号通路中，PTEN 抑制 PIK3CA，EGFR 激活 PIK3CA. 这组基因集合的互斥度和权重值分别为 0.867 和 0.846，P 值是 9.20e-05，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(g)所示.

第 8 组驱动通路包括 PIK3CA、PIK3R1 和 PTEN 的删除，这个模块中 53.3%(193/362)的样本发生突变. 这 3 个基因 PIK3CA、PTEN 和 PIK3R1 是 PI3K 信号通路的核心成员，其中 PIK3CA 是原癌基因，致病基因分数是 95%[43]；PTEN 和 PIK3R1 是抑制原因，肿瘤抑制基因分数分别是 55% 和 37%[43]. 在 PI3K 信号通路中，PTEN 和 PIK3R1 存在相互作用关系，PIK3R1 和 PIK3CA 存在相互作用关系，PTEN 抑制 PIK3CA. 这组基因集合的互斥度和权重值分别为 0.980 和 0.979，P 值是 9.20e-05，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(h)所示.

第 9 组驱动通路包括 CDKN2B 和 CDK4 的缺失、RB1 的删除，这个模块中 72.1%(261/362)的样本发生突变. 这 3 个基因 CDK4，CDKN2B 和 RB1 是 RB 信号通路的核心成员，参与细胞循环和细胞凋亡的整个过程，且全部是重要的抑制基因. 肿瘤抑制基因分数分别是 53%、41% 和 80% [43]. CDKN2B 和 CDKN2A 互为同源基因，在多种癌症中频繁的突变和删除，与多种癌症的发生、发展和恶化密切相关[48]. 在 RB 信号通路中，CDKN2B 抑制 CDK4，CDK4 抑制 RB1. 这组基因集合的互斥度和权重值分别为 0.956 和 0.954，且具有很高的覆盖度为 0.722，P 值是 2.50e-04，具有较高的生物相关性和统计显著性，驱动通路中基因间相互作用关系如图 5(i)所示.

5　总结与展望

互斥的基因组突变模式为理解癌症发生、发展和恶化过程提供了重要的线索. 近年来这种性质被广泛应用于检测致癌驱动通路和失调模块中. 本文提出了一种利用计算方式从大量样本的体细胞突变数据中检测突变驱动通路算法，该算法利用构建突变基因网络的方式检测具有高互斥和高覆盖的基因集合. 本研究不借助任何先验生物知识，只关注癌症的体细胞突变数据，从数据本身固有属性研究癌症的发病机理. 本算法利用基因间互斥性特征描述基因间关系，构建互斥突变基因网络，简化了数据间复杂关系，降低了时间复杂度. 在检测最大完全子图过程中，本算法在解决重叠节点归属问题时优先考虑互斥度和权重函数值高的节点进入完全子图中的情况，从而提高了检测突变驱动通路的效率和准确性.

相对于其它检测突变驱动通路算法，该算法具有以下两方面的优点：(1) 通过构建互斥突变基因网络，简化了数据关系，降低了复杂度；(2) 不需要事先指定突变驱动通路中的基因个数，而是根据数

据本身特征来决定.该算法为癌症数据分析提供了有益补充,该突变驱动通路检测研究有利于加深对癌症发病机理的理解和认识,为开展癌症演变过程等相关研究奠定了基础.对下一步工作的建议是结合临床数据中癌症不同发展阶段信息,建立合理的数学模块,推理癌症发展和恶化过程中不同阶段的失调模块.

参 考 文 献

[1] McLendon R, Friedman A, Bigner D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature, 2008, 455(7216): 1061-1068

[2] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature, 2011, 474(7353): 609-615

[3] Hudson T J, Anderson W, Aretz A, et al. International network of cancer genome projects. Nature, 2010, 464(7291): 993-998

[4] Barretina J, Caponigro G, Stransky N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature, 2012, 483(7391): 603-607

[5] Mullighan C G, Su X, Zhang J, et al. Deletion of IKZF1 and prognosis in acute lymphoblastic leukemia. New England Journal of Medicine, 2009, 360(5): 470-480

[6] Zhao J, Zhang S, Wu L Y, et al. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics, 2012, 28(22): 2940-2947

[7] Zhang J, Wu L Y, Zhang X S, et al. Discovery of co-occurring driver pathways in cancer. BMC bioinformatics, 2014, 15(1): 271

[8] Leiserson M D M, Blokh D, Sharan R, Raphael B J. Simultaneous identification of multiple driver pathways in cancer. PLoS Computational Biology, 2013, 9(5): e1003054

[9] Foo J, Liu L L, Leder K, et al. An evolutionary approach for identifying driver mutations in colorectal cancer. PLoS Computational Biology, 2015, 11(9): e1004350

[10] Constantinescu S, Szczurek E, Mohammadi P, et al. TiMEx: A waiting time model for mutually exclusive cancer alterations. Bioinformatics, 2016, 32(7): 968-975

[11] Chen Y, Hao J, Jiang W, et al. Identifying potential cancer driver genes by genomic data integration. Scientific Reports, 2013, 3(12): 3538

[12] Hou J P, Ma J. DawnRank: Discovering personalized driver genes in cancer. Genome Medicine, 2014, 6(7): 56

[13] Sakoparnig T, Fried P, Beerenwinkel N. Identification of constrained cancer driver genes based on mutation timing. PLoS Computational Biology, 2015, 11(1): e1004027

[14] Tamborero D, Gonzalez P A, Perez L C, et al. Comprehensive identification of mutational cancer driver genes across 12 tumor types. Scientific Reports, 2013, 3(10): 134

[15] Youn A, Simon R. Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics, 2011, 27(2): 175-181

[16] Vandin F, Upfal E, Raphael B J. De novo discovery of mutated driver pathways in cancer. Genome Research, 2012, 22(2): 375-385

[17] Wu H, Gao L, Li F, et al. Identifying overlapping mutated driver pathways by constructing gene networks in cancer. BMC Bioinformatics, 2015, 16(Supplment 5): S3

[18] Kim Y A, Cho D Y, Dao P, et al. MEMCover: Integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types. Bioinformatics, 2015, 31(12): 284-292

[19] Miller C A, Settle S H, Sulman E P, et al. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Medical Genomics, 2011, 4(1): 34

[20] Wu H, Gao L, Kasabov N. Network-based method for inferring cancer progression at the pathway level from cross-sectional mutation data. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2016, 13(6): 1036-1044

[21] Yu X, Zeng T, Li G. Integrative enrichment analysis: A new computational method to detect dysregulated pathways in heterogeneous samples. BMC Genomics, 2015, 16(1): 918

[22] Srihari S, Ragan M A. Systematic tracking of dysregulated modules identifies novel genes in cancer. Bioinformatics, 2013, 29(12): 1553-1561

[23] Wu H, Gao L, Dong J H, et al. Detecting overlapping protein complexes by rough-fuzzy clustering in protein-protein interaction networks. PloS One, 2014, 9(3): e91856

[24] Zhang J, Zhang S, Wang Y, et al. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. BMC Systems Biology, 2013, 7(Supplment 2): S4

[25] Gerstung M, Pellagatti A, Malcovati L, et al. Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. Nature Communications, 2015, 6(23): 5901

[26] Szczurek E, Beerenwinkel N. Modeling mutual exclusivity of cancer mutations. PLoS Computational Biology, 2014, 10(3): e1003503

[27] Vandin F, Upfal E, Raphael B J. Finding driver pathways in cancer: Models and algorithms. International Conference on Algorithms in Bioinformatics, 2011, 7(1): 23

[28] Ciriello G, Cerami E, Sander C, et al. Mutual exclusivity analysis identifies oncogenic network modules. Genome Research, 2012, 22(2): 398-406

[29] Parsons D W, Jones S, Zhang X, et al. An integrated genomic analysis of human glioblastoma multiforme. Science, 2008, 321(5897): 1807-1812

[30] Munoz J L, Bliss S A, Greco S J, et al. Delivery of functional anti-miR-9 by mesenchymal stem cell—derived exosomes to glioblastoma multiforme cells conferred chemosensitivity. Molecular Therapy-Nucleic Acids, 2013, 2(10): e126

[31] Yuan X, Gao M, Zhao H, et al. Glioblastoma multiforme in association with a dural arteriovenous fistula. Neurology India, 2016, 64(7): 121

[32] Gilbert M R, Dignam J J, Armstrong T S, et al. A randomized trial of bevacizumab for newly diagnosed glioblastoma. New England Journal of Medicine, 2014, 370(8): 699-708

[33] Brennan C W, Verhaak R G, McKenna A, et al. The somatic genomic landscape of glioblastoma. Cell, 2013, 155(2): 462-477

[34] Weinstein J N, Collisson E A, Mills G B, et al. The cancer genome atlas pan-cancer analysis project. Nature Genetics, 2013, 45(10): 1113-1120

[35] Fan Q W, Cheng C K, Gustafson W C, et al. EGFR phosphorylates tumor-derived EGFRvIII driving STAT3/5 and progression in glioblastoma. Cancer Cell, 2013, 24(4): 438-449

[36] England B, Huang T, Karsy M. Current understanding of the role and targeting of tumor suppressor p53 in glioblastoma multiforme. Tumor Biology, 2013, 34(4): 2063-2074

[37] Viale A, Pettazzoni P, Lyssiotis C A, et al. Oncogene ablation-resistant pancreatic cancer cells depend on mitochondrial function. Nature, 2014, 514(7524): 628-632

[38] Lee H J, Zhuang G, Cao Y, et al. Drug resistance via feedback activation of Stat3 in oncogene-addicted cancer cells. Cancer Cell, 2014, 26(2): 207-221

[39] Figlioli G, Landi S, Romei C, et al. Medullary thyroid carcinoma (MTC) and RET proto-oncogene: mutation spectrum in the familial cases and a meta-analysis of studies on the sporadic form. Mutation Research/Reviews in Mutation Research, 2013, 752(1): 36-44

[40] Mamo A, Cavallone L, Tuzmen S, et al. An integrated genomic approach identifies ARID1A as a candidate tumor-suppressor gene in breast cancer. Oncogene, 2012, 31(16): 2090-2100

[41] Bar P L, Chantranupong L, Cherniack A D, et al. A Tumor suppressor complex with GAP activity for the Rag GTPases that signal amino acid sufficiency to mTORC1. Science, 2013, 340(6136): 1100-1106

[42] Jiang W, Zhao S, Jiang X, et al. The circadian clock gene Bmal1 acts as a potential anti-oncogene in pancreatic cancer by activating the p53 tumor suppressor pathway. Cancer Letters, 2016, 371(2): 314-325

[43] Vogelstein B, Papadopoulos N, Velculescu V E, et al. Cancer genome landscapes. Science, 2013, 339(6127): 1546-1558

[44] Li R, Ochs M F, Ahn S M, et al. Expression microarray analysis reveals alternative splicing of LAMA3 and DST genes in head and neck squamous cell carcinoma. PLoS One, 2014, 9(3): e91263

[45] Rausch T, Jones D T, Zapatka M, et al. Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell, 2012, 148(1): 59-71

[46] Leroy B, Girard L, Hollestelle A, et al. Analysis of TP53 mutation status in human cancer cell lines: A reassessment. Human Mutation, 2014, 35(6): 756-765

[47] Jiao X, Sherman B T, Huang D W, et al. DAVID-WS: A stateful web service to facilitate gene/protein list analysis. Bioinformatics, 2012, 28(13): 1805-1806

[48] Nicolae-Cristea A R, Benner M F, Zoutman W H, et al. Diagnostic and prognostic significance of CDKN2A/CDKN2B deletions in patients with transformed mycosis fungoides and primary cutaneous CD30-positive lymphoproliferative disease. British Journal of Dermatology, 2015, 172(3): 784-788

WU Hao, born in 1979, Ph. D., associate professor. His main research interests include systems bioinformatics, complex networks, complex diseases and biological data mining.

Background

Mutated driver pathways (MDPs) play an important role in cancer pathogenesis, cancer subtype, cancer therapy and drag targets etc, and it has become a research field in complex diseases and computational bioinformatics. The cancer pathogenesis in human is still poorly understood. With the rapid development of high-throughput sequencing technologies, genome-wide measurements of somatic mutations in large numbers of cancer patients have been generated. Many scholars understand cancer pathogenesis by identifying MDPs, the different combinations of driver mutations observed in different patients with the same cancer type. Mutual exclusivity of gene mutations has been observed in

various cancer types and has been used as an important property of MDPs. Maximum Weight Sub-matrix Problem was introduced to find the MDPs with mutual exclusivity and high coverage of gene mutations, but it is an NP-hard problem.

In order to reduce the complexity of the problem, we present a novel method (Megnet) for detecting mutated driver pathways on the basis of mutually exclusive gene networks. The method simplifies the relationship among genes by constructing gene networks based on mutual exclusivity between each pair of genes. Therefore, gene mutations in each complete subgraph in the gene networks are mutually exclusive. We just need to find the maximum complete subgraph with high coverage. Megnet can detect more biologically relevant and higher statistically significant gene sets.

The author of this paper has conducted the research in this field since 2010, and the research is supported by the Fundamental Research Funds for the Central Universities (No. 2452017342), the Startup Fund for Doctoral Scholars (No. 2452017019), the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2017JM6063), the National Natural Science Foundation of China (Nos. 61532014, 61432010) and Science and Technology Project in Yangling District of Shaanxi Province (No. 2017GY-03). The research team has done some creative work and published more than 20 papers in international or domestic journals and conference proceedings in recent several years. From 2014 to 2016, Wu Hao published six papers in "IEEE/ACM Transactions on Computational Biology and Bioinformatics", "BMC Bioinformatics", "PLoS One" etc.

From July, 2014 to August, 2015, Wu Hao acted as a visiting scholar in KEDRI Lab of Computer Science Department, Auckland University of Technology, New Zealand. During the period, he was mainly involved in the project "Complex disease deterioration process" (NSFC, No. 91130006). He concentrated on the research of driver pathways and cancer progression, and published three papers on "IEEE TCBB", "BMC Bioinformatics" and "SYSID 2015" as the first author.

Comparing with the previous methods of finding mutated driver pathways, Megnet is superior in the following three aspects. Firstly, the algorithmic complexity is reduced by constructing mutated gene networks from somatic mutation data. Secondly, the algorithm does not need to assign the number of genes in a driver pathway. Thirdly, the algorithm does not use gene interaction data, known pathways and other biological information. The algorithm provides a supplement to cancer data analysis. We also anticipate that this method will be helpful in producing hypotheses that will drive some specific experiments and increase understanding for cancer pathogenesis.