一种基于突变基因网络的癌症驱动通路识别算法

郭 炳[1] 郑文萍[2] 韩素青[1]
(太原师范学院计算机科学与技术系 山西 晋中 030619)[1]
(山西大学计算机与信息技术学院 太原 030006)[2]

摘 要 大型癌症基因组项目(TCGA,ICGC等)产生了大量的癌症组学数据,使人们深入研究癌症变为可能,其中寻找引发癌症的相关突变基因是一个重要挑战。在癌细胞中,基因变异可分为两类:一类是可导致癌症发生的驱动突变(driver mutation),另一类是对癌症发生扩散没有影响的乘客突变(passenger mutation)。识别癌症驱动基因有利于理解癌症发病原理和发展进程以及研发癌症药物或进行靶向治疗,是生物信息学中的重要问题。文中提出一种基于突变基因网络的癌症驱动通路识别算法GNDP,对癌症病人的体细胞突变数据进行分析。该算法定义了非重叠平衡度来度量基因对的位于同一驱动通路的可能性;根据基因对的非重叠平衡度、互斥和覆盖度,构建基因互斥网络,很大程度上减少了网络边数,提高了计算效率;在所构造的基因互斥网络中将查找到的极大团作为潜在驱动通路基因集合;用覆盖度和互斥度对潜在驱动通路基因集合进行筛选,得到其极大权重子团,并将其作为识别出的驱动通路。分别在模拟数据、肺腺癌以及多形性成胶质细胞瘤突变数据上对GNDP算法进行有效性验证,并将其与经典驱动通路识别算法Dendrix和Multi-Dendrix进行实验对比。结果表明,GNDP不需要指定驱动通路的基因个数,能在模拟数据上准确检测出所有人工设置的驱动通路;针对肺腺癌和多形性成胶质细胞瘤突变数据,GNDP在不需要任何先验知识的情况下达到较高的识别准确率,能高效地识别出主要驱动通路,其结果优于对比算法。

关键词 癌症基因组,体细胞突变,基因互斥网络,极大团,驱动通路
中图法分类号 TP311 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2018.07.040

Driver Pathway Identification Algorithm Based on Mutated Gene Networks for Cancer

GUO Bing[1] ZHENG Wen-ping[2] HAN Su-qing[1]
(Department of Computer Science and Technology, Taiyuan Normal University, Jinzhong, Shanxi 030619, China)[1]
(School of Computer & Information Technology, Shanxi University, Taiyuan 030006, China)[2]

Abstract Large cancer genome projects such as The Cancer Genome Atlas(TCGA) and International Cancer Genome Consortium(ICGC) have produced big amount of data collected from patients with different cancer types. The identification of mutated genes causing cancer is a significant challenge. Genovariation in cancer cells can be divided into two types: functional driver mutation and random passenger mutation. Identifcation of driver genes is benefit to understand the pathogenesis and progression of cancer, as well as research cancer drug and targeted therapy, and it is an essential problem in the field of bioinformatics. This paper proposed a driver pathway identification algorithm based on mutated gene networks for cancer(GNDP). In GNDP, a nonoverlap balance metric is defined to measure the possibility of two genes lying in the same driver pathway. To reduce the complexity of the constructed mutually exclusive gene networks, the nonoverlap balance metric, the exclusivity and the coverage of a gene pair are computed first, and then the edges with low nonoverlap balance metric, low exclusivity and low coverage are deleted. Then, all maximal cliques which might be potential driver pathways are found out. After that, the weight of each clique is assigned as the product of its exclusive degree and coverage degree and then every node of a clique will be checked to judge whether is's deletion might obtain a larger weight. At last, the maximal weight cliques are obtained in mutually exclusive gene networks as the final driver pathways. This paper compared GNDP algorithm with classical algorithm Dendrix and Multi-Dendrix on both simulated data sets and somatic mutation data sets. The results show that GNDP can detect all artificial pathways in simulated data. For Lung adenocarcinoma and Glioblastoma data, GNDP shows higher efficiency and accuracy than the comparison algorithms. In addition, GNDP does not need any prior knowledge and does not need to set the number of genes in driver pathways in advance.

Keywords Cancer genome, Somatic mutation, Gene networks, Maximal cliques, Driver pathways

到稿日期:2018-01-28 返修日期:2018-05-03 本文受山西省回国留学人员科研基金(2017-014),国家自然科学基金(61572005),山西省软科学研究项目(2016041036-4)资助。

郭 炳(1985—),男,硕士,助教,CCF会员,主要研究方向为生物信息学与机器学习,E-mail:guobing@tynu.edu.cn(通信作者);**郑文萍**(1979—),女,博士,副教授,硕士生导师,主要研究方向为图论算法与生物信息学;**韩素青**(1964—),女,博士,副教授,硕士生导师,主要研究方向为数据挖掘与机器学习。

外，由于 Astra3D 传感器支持多个操作平台，因此该算法还可以在 Windows，Linux 等平台上实现。

为了验证该坐姿检测系统的可行性与有效性，让一名志愿者坐在传感器前正常学习，使用本系统对其进行实时检测，系统测试结果如图 11 所示。

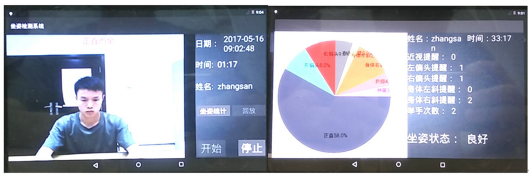

图 11 系统测试结果

Fig. 11 Results of system test

通过测试结果可知，本系统能够对人体的坐姿进行有效检测和对不良坐姿进行及时提醒以及对坐姿进行准确统计。除此之外，本系统还具有便携性、小型化等优点。

结束语 本文以检测不良坐姿和分析人们的坐姿习惯为引导，设计了一种基于深度传感器的坐姿检测系统。首先，基于坐姿深度图像设计了一种前景提取方法，对背景和干扰进行了有效去除；随后，对提取到的坐姿前景图进行投影特征与金字塔 HOG 特征的提取，并运用随机森林进行分类识别；最后，运用坐姿深度图像数据库进行统一测试与交叉测试，测试结果表明该方法具有很好的识别率与识别速度。出于实用化考虑，本文基于 Android 平台设计了坐姿检测系统应用软件，实现了坐姿的有效检测与不良坐姿的及时提醒等功能。未来的工作主要是建立更大的坐姿数据库，提高坐姿识别精度，以及分析提取更多的坐姿特征对一些复合姿态进行识别。

参 考 文 献

[1] KAMIYA K，KUDO M，NONAKA H，et al. Sitting posture analysis by pressure sensors[C]// International Conference on Pattern Recognition. IEEE，2008：1-4.

[2] WANG C Y. Research on the Monitoring of Setting Posture Based on Image Technology[D]. Xi'an：Xidian University，2013.（in Chinese）
王春阳. 基于图像技术的人体坐姿监测研究[D]. 西安：西安电子科技大学，2013.

[3] WU S L，CUI R Y. Human Behavior Recognition Based on Sitting Postures[C]// International Symposium on Computer，Communication，Control and Automation Proceedings. 2010：138-141.

[4] YUAN D B，DAI Y，CHEN T Q. Multi-feature fusion recognition of incorrect sit posture[J]. Computer Engineering and Design，2017，38(2)：528-523.（in Chinese）
袁迪波，戴永，陈统乾. 不规范书写坐姿的多类特征融合与识别[J]. 计算机工程与设计，2017，38(2)：528-523.

[5] ZHANG H Y，LIU W，XU W，et al. Depth hllage Based Gesture Recognition for Multiple Lesrners[J]. Computer Science，2015，42(9)：299-302.（in Chinese）
张鸿宇，刘威，许炜，等. 基于深度图像的多学习者姿态识别[J]. 计算机科学，2015，42(9)：299-302.

[6] HUANG J Y，HSU S C，HUANG C L. Human upper body posture recognition and upper limbs motion parameters estimation[C]// Signal and Information Processing Association Summit and Conference. 2013：1-9.

[7] ABDI H，WILLIAMS L J. Principal component analysis[J]. Wiley Interdisciplinary Reviews Computational Statistics，2010，2(4)：433-459.

[8] DALAL N，TRIGGS B. Histograms of Oriented Gradients for Human Detection[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition，2005. IEEE，2005：886-893.

[9] BREIMAN L. Random Forest[J]. Machine Learning，2001，45：5-32.

（上接第 236 页）

[25] ZENG X X，LIAO Y L，LIU Y S，et al. Prediction and validation of disease genes using HeteSim Scores[J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics，2016，PP(99)：1.

[26] CHEN X，YAN G Y，REN W，et al. Modularized random walk with restart for candidate disease genes prioritization[C]// The Third International Symposium on Optimization and Systems Biology(OSB'09). 2009：353-360.

[27] CHEN X，HUANG L，XIE D，et al. EGBMMDA：Extreme gradient boosting machine for miRNA-disease association prediction[J]. Cell Death & Disease，2018，9(1)：3.

[28] MILLER C A，SETTLE S H，SULMAN E P，et al. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors[J]. BMC Medical Genomics，2011，4(1)：34.

[29] ZHAO J F，ZHANG S H，WU L Y，et al. Efficient methods for identifying mutated driver pathways in cancer[J]. Bioinformatics，2012，28(22)：2940-2947.

[30] 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.

[31] WU H，GAO L，LI F，et al. Identifying overlapping mutated driver pathways by constructing gene networks in cancer[J]. Bmc Bioinformatics，2015，16(S5)：S3.

[32] WU H. Algorithm for detecting driver pathways in cancer based on mutated gene networks[J/OL]. Chinese Journal of Computer，(2017-02-24). http://kns.cnki.net/kcms/detail/11.1826.TP.20170224.1408.101.html.（in Chinese）
吴昊. 基于突变基因网络的致癌驱动通路检测算法[J/OL]. 计算机学报，(2017-02-24). http://kns.cnki.net/kcms/detail/11.1826.TP.20170224.1408.101.html.

[33] SZCZUREK E，BEERENWINKEL N. Modeling mutual exclusivity of cancer mutations[J]. Plos Computational Biology，2014，10(3)：e1003503.

1　引言

癌症是一种严重威胁人类健康的复杂疾病。研究证实，癌症与基因突变有关[1]，通过这些突变，癌细胞会无限增殖，且在体内通过血液和淋巴进行传播扩散[2]。随着第二代DNA测序技术的发展，特别是大型癌症基因组项目TCGA[3-5]和ICGC[6]等的建立，科学家已经得到了大量癌症细胞的组学数据[7-9]。从这些数据中找到促进癌症增殖过程的驱动基因和突变通路，是当前的研究热点之一[10-12]。

癌细胞中的突变通常分为两种：一种对肿瘤的增殖扩散有质的影响，可以使肿瘤获得选择性的生长优势，称为驱动突变(driver mutation)，突变的基因称为驱动基因(driver gene)；另外一种对肿瘤增殖扩散的影响较小或者没有质的影响，称为乘客突变(passenger mutation)，突变的基因称为乘客基因(driver gene)[13]。

对于驱动突变，可以在大量患者数据中通过统计突变基因的发生频率(也称为覆盖率)进行识别，一般将覆盖率较高的突变基因识别为驱动突变(或者驱动基因)。根据突变在病人中的覆盖度进行驱动基因识别，目前已经获得了一些重要的驱动基因，如黑色素瘤中的PIK3CA基因、肺癌中的KRAS基因、恶性成胶质细胞瘤中的ERBB2基因等[14-16]。

然而，并不是所有的驱动基因都在病人中有较高的覆盖率，例如肺腺癌的驱动基因之一ATM，在163个病人样本中，只有13个样本发生突变，覆盖率仅为8.0%[17]。通常，一个信号通路上的某一个基因突变，就可能导致癌症的发生，而一条通路上的基因突变在不同的个体中呈现出很强的异质性，即使是在同一位癌症患者体内的两个不同的癌细胞，它们发生突变的基因可能也是不同的[3,18]。因此，统计单个突变基因的发生频率或覆盖率并不能识别出全部的驱动突变。这意味着，检测肿癌的驱动突变、驱动基因以及驱动通路(Driver pathway)具有重要意义，对理解癌症发病原理和发展进程并设计癌症药物或进行靶向治疗非常关键[11,19-21]。

由于导致癌症发生的靶标信号与通路的调控由多个基因共同作用[15,22]，因此识别癌症驱动基因的问题可以转化为识别突变驱动通路的问题。2010年以来，Vaske等提出了基于已有基因组学数据和基因表达数据，利用因子图模型知识来识别癌症驱动通路的方法；Bashashati提出了基于整合基因突变数据、基因互作用网络、基因表达数据等已有知识来进行驱动通路的识别，并将驱动基因求解问题转化为贪心算法来求解二分图的最大匹配问题[23-24]。然而，通路数据库目前还不完备，无法为识别癌症驱动通路提供足够的支持。近年来，曾湘祥等[25]和陈兴等[26-27]将多种网络(如基因疾病网络、miRNA疾病网络、蛋白质互作用网络、疾病网络)与链路预测方法相结合进行癌症相关基因或miRNA的推断，取得了较好的效果。

Miller等[28]于2011年提出了基于周期性(recurrent)和互斥性(mutually exclusive)的RME算法，该算法利用癌症基因突变数据识别驱动通路。首先，利用Winnow方法将突变矩阵转化为基因互斥网络，然后在该网络中搜索RME模块，其中，每个RME模块表示一个驱动通路。但是，该算法在处理较大规模的数据时效率很低。2012年，Vandin等[17]提出了Dendrix算法，该算法将驱动通路识别问题转化为最大权重子矩阵求解问题(The maximum weight submatrix problem)，并采用马尔科夫蒙特卡洛方法(MCMC)对Dendrix算法的目标函数进行优化，将在癌症病人体细胞突变数据中的识别具有高互斥度(high exclusivigy)和高覆盖度(high coverage)的k个基因组成的集合作为驱动通路。Dendrix算法需要事先指定驱动通路所包含的基因个数k，且要求识别的驱动通路之间没有公共基因。然而，一个驱动通路通常包含的基因个数从2到6不等，且驱动通路之间存在部分公共基因，因此Dendrix算法的应用受到限制。2012年，赵俊飞等[29]使用遗传算法(GA)来解决Vandin等提出的最大权重子矩阵的问题。鉴于Dendrix算法和GA算法在识别驱动通路时需要去掉上一次迭代的最优解，降低了识别的准确率，Mark等[30]于2013年提出了Multi-dendrix算法，该算法通过线性规划求解最大权重子矩阵问题，旨在同时识别多条驱动通路。由于最大权重子矩阵求解问题属于NP困难问题，为了在允许的时间范围内求解该问题，Multi-dendrix算法需要预先指定每条驱动通路中的最大基因个数$k_{\max}$和识别的驱动通路数m，而这需要有合适的方法来实现。2015年，吴昊[31]给出了基于癌症病人体细胞突变数据的NBM算法，该算法利用互斥度大于给定阈值的基因对构建基因互斥网络，然后在该网络上搜索基因覆盖度高于给定阈值的子网作为驱动通路的候选集，从而得到具有重复基因的突变驱动通路。但是，NBM算法利用基因互斥度构造的基因互斥网络的边数较多，因此，计算效率相对较低。2017年，吴昊等[32]提出了Megnet算法，该算法首先对两个互斥度接近1的基因对的覆盖情况进行深入分析，定义了覆盖重叠函数和非重叠比重函数，然后基于互斥度$ED>0.95$且非重叠比重函数值$RD>0.85$的基因对构造基因互斥网络，有效缩减了基因互斥网络的规模。然而，在基因突变矩阵中，互斥度为1的基因对(即两个基因完全互斥)大量存在，如在包括163个病人共356个基因的肺腺癌基因突变矩阵中[17]，互斥度为1的基因对有57553对，占所有可能的基因对的91.08%，其中含有大量突变次数相差悬殊的基因对，而两个突变次数相差悬殊的基因往往位于不同的驱动通路中[33]，因此，所构造的基因互斥网络应该排除此类边。例如，对于基因对(STK11，MDM2)，其互斥度为1，STK11在35个病人样本中发生突变，位于mTOR通路；MDM2在2个病人样本中发生突变，位于P53通路。

此外，由于病人的个体差异，在基因突变矩阵中存在大量覆盖度较小的基因对，如在包括163个病人共356个基因的肺腺癌基因突变矩阵中[17]，有61012对覆盖度小于0.09的基因对，占所有可能的基因对的96.55%。例如，对于基因对CDKN2B和ERBB2，其覆盖度CD为0.025，但基因CDKN2B属于PB信号通路，而基因ERBB2属于RTK/RAS信号通路，它们不属于同一驱动通路，因此，此类边也应被排除。

基于上述分析，本文以癌症病人的体细胞突变数据为基础，提出一种基于突变基因网络的癌症驱动通路识别算法，以下称为驱动通路识别算法(Driver Pathways Identify based on Gene Networks，GNDP)。该算法首先定义了非重叠平衡度

B(nonoverlapbalance metric),用来衡量突变矩阵中基因对的互斥度和覆盖接近程度,它不仅考虑了有部分重叠的基因对之间的覆盖度,也考虑了完全互斥的基因对之间的覆盖接近程度。其含义为:若两个基因的互斥度越高,则该基因对的非重叠平衡度值越高;若两个基因的覆盖度越接近,则该基因对的非重叠平衡度值越高。GNDP 算法利用非重叠平衡度 $B>0.13$ 且覆盖度 $CD>0.09$ 的基因对构造基因互斥网络,并在该网络上搜索极大团作为候选驱动通路;然后,根据互斥度和覆盖度对所得到的极大团进行排序,选择具有较高互斥度和覆盖度的极大团作为最终发现的驱动通路。在人工模拟数据集上对算法的性能进行测试,结果表明,本文给出的 GNDP 算法在人工模拟数据上能够在很短的时间内准确发现所有人工设定的驱动通路;在肺腺癌和多形性成胶质细胞瘤真实数据集上,GNDP 算法能够在较短的时间内识别更完整的癌症驱动通路,且不需要预先指定驱动通路的基因个数和过滤突变次数极少的基因,也不需要任何已知生物学知识并指定驱动基因的个数。

2 相关知识

癌症病人的体细胞突变数据用一个 $m\times n$ 的矩阵 $A_{m\times n}$ 来表示,其中包含 m 个病人样本$\{p_1,\cdots,p_m\}$和 n 个突变基因$\{g_1,\cdots,g_n\}$。若病人 p_i 的基因 g_j 发生突变,则 $A_{ij}=1$,否则 $A_{ij}=0$。如图 1 所示,其中灰色节点对应 1,白色节点对应 0。

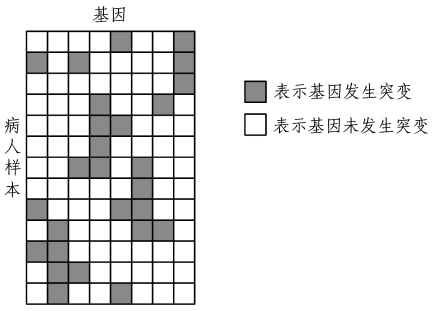

图 1 突变矩阵

Fig. 1 Mutation matrix

一个癌症驱动通路中的基因之间往往存在较高的互斥度和覆盖度[17]。在图 G 中,基因 g_i 的覆盖集 $\Gamma(g_i)=\{p_j\mid A_{ij}=1,1\leqslant j\leqslant m\}$,表示基因 g_i 发生突变的病人集合;对于矩阵 A 的 $m\times k$ 子矩阵 M,基因子集 M 的覆盖集 $\Gamma(M)=\bigcup\limits_{g\in M}\Gamma(g)$,表示 k 个基因中每个基因至少有一位病人发生突变的病人集合。

基因子集 M 的覆盖度表示在子集 M 中发生基因突变的病人占所有病人数的比例,记作:

$$CD(M)=\frac{|\Gamma(M)|}{m} \tag{1}$$

特别地,任意基因对(g_i,g_j)的覆盖度为:

$$CD(g_i,g_j)=\frac{|\Gamma(g_i,g_j)|}{m} \tag{2}$$

如果 $CD(M)=1$,则基因子集 M 对病人而言是完全覆盖的,即对于 M 中的基因,所有样本中至少有一个基因发生突变。

对于 M 中任意一对基因 $g_i,g_j\in M(g_i\neq g_j)$,如果$\Gamma(g_i)\cap\Gamma(g_j)=\varnothing$,则称 M 中的基因是互斥的。基因子集 M 的互斥度记作:

$$ED(M)=\frac{|\Gamma(M)|}{\sum\limits_{g\in M}|\Gamma(g)|} \tag{3}$$

特别地,对于任意基因对(g_i,g_j),其互斥度函数为:

$$ED(g_i,g_j)=\frac{|\Gamma(g_i)\bigcup\Gamma(g_j)|}{|\Gamma(g_i)|+|\Gamma(g_j)|} \tag{4}$$

如果 $ED(M)=1$,则子矩阵 M 中的基因是互斥的,即 M 中的每位病人最多有一个基因发生突变。

对于任意基因对(g_i,g_j),其非重叠比重函数 $RD(g_i,g_j)$[32]被定义为:

$$RD(g_i,g_j)=1-\frac{|\Gamma(g_i)\bigcap\Gamma(g_j)|}{\min\{|\Gamma(g_i)|,|\Gamma(g_j)|\}} \tag{5}$$

根据基因突变矩阵构造基因互斥网络 $G=(V,E)$,其中节点集 $V=\{v_1,v_2,\cdots,v_n\}$表示突变的基因,边集 $E=\{e=(v_i,v_j)\mid 1\leqslant i,j\leqslant n\}$代表基因之间的联系集合,寻找驱动通路的问题可被转化为在基因互斥网络图 G 中寻找极大团的问题。为了方便叙述,首先给出一些必要的图论术语。

设 $G=(V,E)$和 $G'=(V',E')$是两个图,如果 $V'\subseteq V$ 且 $E'\subseteq E$,则称 G'是 G 的子图,记作 $G'\subseteq G$。在无向图 G 中,存在一个子图 G_i'且子图 G_i'中任意两个不同节点之间都恰有一条连边,则称子图 G_i'为一个团。在无向图 G 中,子图 G_i'为 G 的一个团,不存在另一个团 $G_j'(i\neq j)$,使得 $G_i'\subseteq G_j'$,则称子图 G_i'为极大团,用 G^M 表示,包含顶点数目最多的团即为 G 的最大团。显然,一个图的最大团必然是该图的极大团,但极大团不一定是最大团。

在此基础上,可以给团赋予权重来衡量其优劣。在无向图 G 中,子图 G_i'为 G 的一个团,该团的权重为 $W(G_i')=ED(G_i')\times CD(G_i')$,如若不存在另一个团 $G_j'(i\neq j)$,使得 $W(G_i')\leqslant W(G_j')$,则称子图 G_i'为极大权重团。

本文算法所寻找的驱动通路就是基因互斥网络 $G=(V,E)$的极大权重团。

3 驱动通路识别算法

在癌症基因突变数据中,识别驱动通路的本质是在突变矩阵中搜索满足高覆盖性和高互斥性的基因组合。随着突变数据的增长,直接在突变矩阵中搜索这种组合的计算复杂度增加且效率降低。鉴于此,本文提出一种基于突变基因网络的癌症驱动通路识别算法,该算法包括基因互斥网络构建、种子子图初始化、极大权重团挖掘 3 个主要过程。

3.1 突变基因互斥网络的构建

2011 年以来,一些学者根据基因对之间满足的高覆盖度、高互斥度等条件构建了突变基因网络,以降低算法的复杂度[28,32]。其中,吴昊提出的 Megnet 算法将满足 $ED>0.95$ 且 $RD>0.85$ 的基因对连接,以构造网络,但是在该条件下构造的网络边数比较稠密。比如在肺腺癌基因数据集中[17],按照 $ED>0.95$ 且 $RD>0.85$ 的条件构建的网络的点数为 356,边数为 57583。图 2 中情况(a)-(c)给出了覆盖度 $CD=0.95$、互斥度 $ED=1$ 的 3 对基因。显然,情况(c)所示的基因对之间的突变次数相差悬殊,在基因互斥网络中不应存在连

边。然而,根据文献[32]给出的式(5),该基因对的 $RD=1$,因此其对应的边包含在所构造的基因互斥网络中。为了将这类多余的连边从基因互斥网络中去掉,本文提出式(6)所示的非重叠平衡度来对此类情况进行准确定义(不混淆的情况下,简称为 B)。

$$B(g_i,g_j)=\frac{\min(|\Gamma(g_i)|,|\Gamma(g_j)|)-|\Gamma(g_i)\bigcap\Gamma(g_j)|}{\frac{1}{2}(|\Gamma(g_i)|+|\Gamma(g_j)|)} \tag{6}$$

可以看出,两个基因的覆盖度相差越悬殊,B 值就越小。在图 2 所示的(a)—(c),3 种情形中,非重叠平衡度分别为 1,0.74,0.11。当 $B>0.13$ 时,可以排除情形(c)。

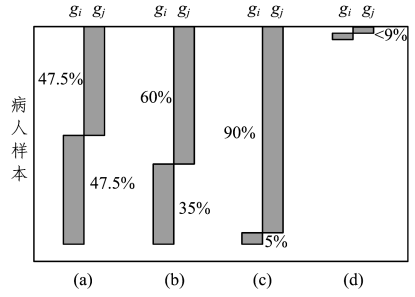

图 2　互斥类型分析

Fig. 2　Analysis of mutual exclusion type

本文给出的非重叠平衡度也可以在构造基因互斥网络时排除那些不完全互斥且突变次数相差悬殊的基因对,比如,图 3 给出了覆盖度 $CD=0.95$、互斥度 $ED=0.95$ 时的基因对,其非重叠平衡度值分别为 0.90,0.70,0.10,0。当 $B>0.13$ 时,可以排除(c)和(d)两种情形。

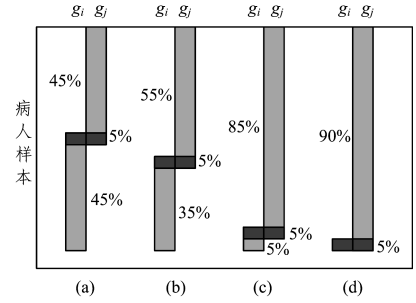

图 3　近似互斥类型分析

Fig. 3　Analysis of approximate mutual exclusion type

此外,经过大量分析发现,实际数据中存在大量覆盖度很小的基因对。由于个体差异,这样的基因对通常位于不同的驱动通路。如对于肺腺癌数据中的基因对 CDKN2B 和 ERBB2,二者的覆盖度 CD 均为 0.025,但基因 CDKN2B 属于 PB 信号通路而基因 ERBB2 属于 RTK/RAS 信号通路,所构造的基因互斥网络应该排除此类边。在肺腺癌数据中,覆盖度 $CD<0.09$ 的基因对在所有可能的基因对中占比96.55%。因此,本文在构造基因互斥网络连边时,考虑了两个基因的覆盖度,只有覆盖度 $CD>0.09$ 时才有可能在这两个基因之间构造连边,如图 2 中情况(d)所示。

综合以上分析,首先根据基因对的互斥度 ED、覆盖度 CD 和非重叠平衡度 B 构建基因互斥网络 $G_{n\times n}$,其中,顶点代表突变基因。对于每一对突变基因,计算其互斥度 ED、覆盖度 CD 和非重叠平衡度 B。若基因对 $ED>0.85$,$B>0.13$ 且 $CD>0.09$,则在该基因对之间添加一条无向边。

3.2　种子子图的初始化

在上一步构造的突变基因互斥网络 $G_{n\times n}$ 中,一个团可能是一条潜在的驱动通路。因此,GNDP 算法选择该网络上的所有极大团作为寻找驱动通路的种子子图。由于去掉了大量的多余边,$G_{n\times n}$ 网络比较稀疏,因此可以在较短时间内寻找到其中的所有极大团。假设找到的所有极大团的集合为 $SG=\{SG_1,\cdots,SG_t\}$,每个极大团是一个种子子图。

3.3　极大权重团的挖掘

根据驱动通路满足高覆盖度和高互斥度的思想,可以定义极大团权重函数来对其进行评价。

对于 G 中的一个团 C,令其顶点集为 V_c,则团 C 的权重定义为:

$$W(C)=ED(V_c)\times CD(V_c) \tag{7}$$

显然,一个团的顶点的导出子图仍然为团。在团 C 的所有顶点导出子图中,权重最大的子团称为团 C 的极大权重子团。一个驱动通路上的基因集合往往对应于基因互斥网络中的一个极大权重子团。

对于 G 的每个极大团 SG_i,如果从 SG_i 中删除一个顶点 v 所得的子团的权重比 SG_i 的权重更高,则认为子团 $SG_i-\{v\}$ 比极大团 SG_i 更优;从 SG_i 中迭代删除顶点,直到权重不再增加,即可以得到 SG_i 的极大权重子团 S_i。

由于驱动通路中包含的基因数通常不小于 3,因此若 SG_i 的极大权重子团 S_i 满足条件 $|V(S_i)|\geqslant 3$ 且覆盖度 $CD(S_i)>0.3$[32],则输出 S_i 是一个得到的驱动通路。

算法 1　驱动通路识别算法 GNDP

输入:基因突变矩阵 $A_{m\times n}$(m 个病人样本,n 个基因)

输出:突变驱动通路集合 Path

Step1　初始化基因互斥网络 $G_{n\times n}=\{0\}$,驱动通路集合 Path=∅。

Step2　计算所有基因对的互斥度 ED、非重叠平衡度 B、覆盖度 CD。

Step3　对于任意两个基因 g_i 和 g_j,若 ED>0.85,B>0.13 且 CD>0.09,则 G[i][j]=G[j][i]=1。

Step4　在网络 G 上搜索所有的极大团,设 SG 为所找到的极大团集合。

Step5　对于于 SG 中的每个极大团 SG_i

Step5.1　对于 SG_i 中的所有点 v,令 f(v)=W($SG_i\backslash\{v\}$)为其子团 $SG_i\backslash\{v\}$的权重。

Step5.2　令 t=arg max{f(v)|v∈SG_i}。若 W($SG_i\backslash\{t\}$)>W(SG_i),则令 $SG_i=SG_i\backslash\{v\}$,返回 Step 5.1;否则,SG_i 为所发现的一个极大权重团。

Step5.3　计算 SG_i 的覆盖度,若 CD(SG_i)>0.3 且 $|SG_i|\geqslant 3$,则 Path=Path∪{SG_i}。

Step 6　输出驱动通路集合 Path,结束。

4　实验过程与结果分析

4.1　实验过程及实验环境

图 4 是驱动通路检测的一个实验过程示意图。其中,图 4(a)为癌症病人的基因突变矩阵,行表示病人样本,列表示基因,深色单元格表示某样本在该基因处发生突变;图 4(b)为算法根据疾病基因突变矩阵所构造的基因互斥网络;图

4(c)－图 4(g)和图 4(i)－图 4(j)分别为由基因互斥网络上两个点数为 3 的团扩展得到极大权重团的过程。

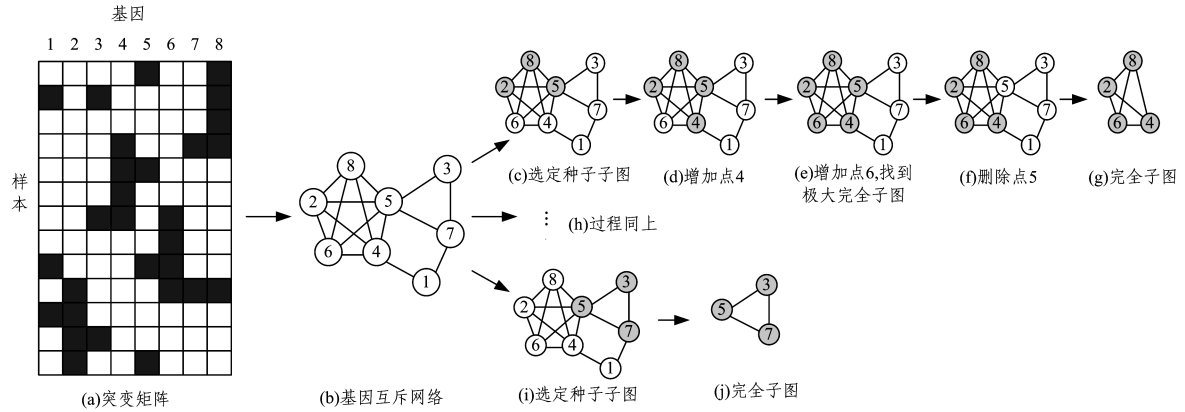

图 4 驱动通路的检测过程

Fig. 4 Detection process of driver pathways

算法的运行环境为 32 位 Win7 操作系统，4 GB 内存，3.60GHz 处理器。

4.2 结果分析

本文在模拟数据、肺腺癌以及多形性成胶质细胞瘤突变数据上对 GNDP 算法的运算效率、性能、准确率进行验证，并将其与经典驱动通路识别算法 Dendrix 和 Multi-Dendrix 进行实验对比。

(1)模拟数据

首先，生成样本数 $m=600$、基因数 $n=2000$ 的突变数据，嵌入 12 组通路 $P=\{P_1,P_1,\cdots,P_{12}\}$，其中通路 P_1,P_2,P_3 各包含 6 个驱动基因($k=6$)，通路 P_4,P_5,P_6 各包含 5 个驱动基因($k=5$)，通路 P_7,P_8,P_9 各包含 4 个驱动基因($k=4$)，通路 P_{10},P_{11},P_{12} 各包含 3 个驱动基因($k=3$)，每条通路的覆盖度 $CD(P_i)$ 分别为 0.90，0.87，0.84，0.80，0.77，0.74，0.70，0.67，0.64，0.60，0.57，0.54。为了更接近真实数据，在生成的样本数据中加入大量的乘客突变基因或者由于测量误差导致的噪声数据。实验中，设置乘客突变概率 q 分别为0.0001，0.0005，0.001，0.005，0.01，再在每个 q 值分别生成的 20 个基因突变矩阵上计算平均运行时间，并将其作为算法的运行时间来使用。通过与 Dendrix 算法、Multi-Dendrix 算法进行对比(见表 1)，可以看出本文提出的 GNDP 算法的运行效率得到了大幅提高。

表 1 3 种算法在不同 q 值下的平均运行时间

Table 1 Average running time of three algorithms with different q

乘客突变概率 q	平均运行时间/s		
	Dendrix	Multi-Dendrix	GNDP
0.0001	1662.26	44.32	21.75
0.0005	1744.77	89.79	35.06
0.001	1826.59	155.47	66.05
0.005	1886.38	326.28	181.69
0.01	1841.41	368.26	232.40

本文对乘客突变概率相同、基因个数不断增加的情况也进行了性能比较，结果如表 2 所列，其中样本数为 600，基因数分别为 2000，2500，3000，乘客突变概率为 $q=0.001$。可以看出，相较于 Dendrix 算法和 Multi-Dendrix 算法，GNDP 算法获得了更高的计算效率。

表 2 3 种算法在乘客突变概率($q=0.001$)相同且基因个数不同的情况下的平均运行时间

Table 2 Average running time of three algorithms with different number of genes and $q=0.001$

基因个数	运行时间/s		
	Dendrix	Multi-Dendrix	GNDP
2000	1826.59	155.47	66.05
2500	1714.14	207.18	105.69
3000	1916.06	221.27	157.66

本文对 3 种算法在不同 q 值下检测驱动通路的准确率进行了统计，结果如图 5 所示，其中每种算法在每个 q 值的准确率由 20 组数据求平均得出。当 $q\leqslant 0.0001$ 时，3 种算法都能检测出 12 条通路。然而，随着 q 的增大，噪声数据增多，各算法的准确率逐渐下降，可以看出本文算法 GNDP 在噪声量增加的情况下有更好的表现。

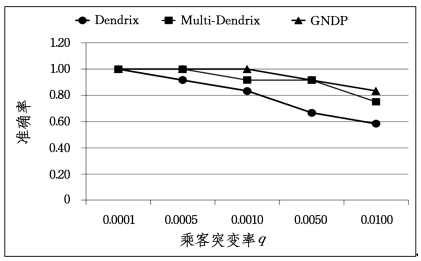

图 5 3 种算法在不同 q 值下检测驱动通路的准确率

Fig. 5 Accuracy of detecting driver pathways of three algorithms with different q

(2)肺腺癌(Lung Adenocarcinoma，LA)数据

肺癌是发病率和死亡率增长最快、对人体健康和生命威胁最大的恶性肿瘤之一。Vandin 等[17]分析了来自 188 位肺腺癌患者的 1030 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 个样本共 356 个基因。为了验证算法的有效性，分别对该数据集运行 Dendrix 算法、Multi-Dendrix 算法和 GNDP 算法，并对结果进行对比，如表 3 所列。可以看出，本文算法 GNDP 可以找到更完整的驱动通路基因集合，如参与对 mTOR 通路调节的基因中，Dendrix 和 Multi-Dendrix 均检测出了其中的 3 个(EGFR，

KRAS,STK11);而 GNDP 检测出了(EGFR,FGFR4,KRAS,NF1,STK11),这些基因都参与了 mTOR 通路的调节[18]。对于细胞周期检控点调控基因,Dendrix 算法和 Multi-Dendrix 算法只检测出了两个(ATM,TP53),GNDP 算法检测出了 3 个(ATM,TP53,EGFR)。图 6[18] 给出了 Dendrix 算法和 Multi-Dendrix 算法检测到的主要驱动通路,浅灰色基因为参与调控 mTOR 通路的基因,深灰色基因为细胞周期检控点调控基因。图 7 为 GNDP 算法检测到的主要驱动通路,浅灰色基因为参与调节 mTOR 通路的基因,深灰色基因为细胞周期检控点调控基因,黑色基因 EGFR 为两个通路共有[3]。

表 3　3 种算法对肺腺癌数据的检测结果

Table 3　Detection results of three algorithms for lung adenocarcinoma

算法	最优基因集合	基因数量	互斥度	覆盖度
Dendrix	EGFR,KRAS,STK11	3	0.89	0.67
	ATM,TP53	2	0.99	0.47
Multi-Dendrix	EGFR,KRAS,STK11	3	0.89	0.67
	KRAS,STK11,EGFR,PRKCG	4	0.89	0.70
	TP53,ATM,PAK4,ACVR1B	4	0.99	0.50
	KRAS, STK11, EGFR, EPHB1, MAP3K3	5	0.89	0.72
	TP53,ATM,PAK4,AIFM3,BAP1	5	0.99	0.51
GNDP	EGFR,KRAS,NF1	3	0.95	0.60
	EGFR,ATM,TP53	3	0.88	0.58
	EGFR,KRAS,NF1,STK11	4	0.86	0.72
	EGFR,FGFR4,KRAS,NF1,STK11	5	0.84	0.74
	ATM,CDKN2A,EGFR,ERBB4,LRP1B,NF1,STK11	7	0.83	0.63

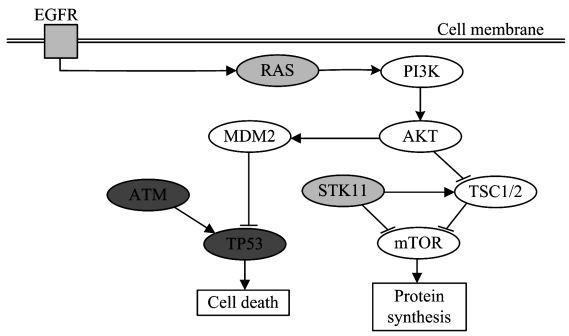

图 6　Dendrix 算法和 Multi-Dendrix 算法检测到的肺腺癌主要驱动通路

Fig. 6　Main lung adenocarcinoma driver pathways detected by Dendrix and Multi-Dendrix

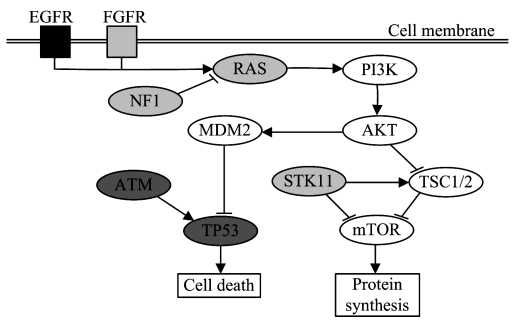

图 7　GNDP 算法检测到的肺腺癌主要驱动通路

Fig. 7　Main lung adenocarcinoma driver pathways detected by GNDP

(3) 多形性成胶质细胞瘤(Glioblastoma,GBM)数据

为进一步验证算法的有效性,选择多形性成胶质细胞瘤体细胞突变数据,实验数据集来源于 TCGA[3]。该突变矩阵中包括 84 个样本共 178 个基因,对该数据集分别运行 Dendrix,Multi-Dendrix 和 GNDP 3 种算法。表 4 给出了 3 种算法得出的最优基因集合、每组最优基因集合对应的基因数量、基因集合的互斥度和覆盖度。3 种算法都检测出最优基因集合为(CDKN2B,CYP27B1,RB1);Dendrix 算法和 GNDP 算法检测出了基因集合(CDKN2B,CDK4,RB1),这 3 个基因都是 RB 信号通路的核心成员[3]。基因 CDKN2A,MDM2,MDM4,TP53 都是 P53 信号通路的核心成员,Dendrix 算法检测出了其中两个基因 TP53 和 CDKN2A;Multi-Dendrix 算法检测出了基因集合(CDKN2A,TP53,DTX3),包含 TP53 和 CDKN2A;而 GNDP 算法检测出了基因集合(CDKN2A,TP53,MDM2)和(CDKN2A,MDM2,MDM4,TP53,ERBB2),前者全部为 P53 信号通路的核心成员[3],后者包含 P53 信号通路的 4 个核心成员。Dendrix 算法和 Multi-Dendrix 算法的检测结果基本一致,即检测到 RB 通路中的 3 个基因,以及 P53 通路、RTK 通路各两个基因。图 8 为这两种算法检测到的主要驱动通路,灰色基因表示检测结果,白色基因表示未检测到的基因。GNDP 算法检测到 RB 通路中的 3 个基因,以及 P53 通路中的 4 个基因。图 9 为 GNDP 算法检测到的主要癌症驱动通路[3]。

表 4　3 种算法对多形性成胶质细胞瘤数据的检测结果

Table 4　Detection results of three algorithms for glioblastoma

算法	最优基因集合	基因数量	互斥度	覆盖度
Dendrix	CDKN2B,CYP27B1,RB1	3	0.94	0.79
	CDKN2B,CDK4,RB1	3	0.94	0.77
	TP53,CDKN2A	2	0.91	0.70
	EGFR,NF1	2	0.98	0.61
Multi-Dendrix	CDKN2B,CYP27B1,RB1	3	0.94	0.79
	CDKN2A,TP53,DTX3	3	0.88	0.79
	CDKN2B,MARCH9,RB1,ERBB2	4	0.92	0.83
	CDKN2A,TP53,DTX3,CDC123	4	0.88	0.81
	EGFR,NF1	2	0.98	0.61
GNDP	CDKN2B,CYP27B1,RB1	3	0.94	0.79
	CDKN2B,RB1,CDK4	3	0.94	0.77
	CDKN2A,TP53,MDM2	3	0.86	0.79
	CDKN2A,MDM2,MDM4,TP53,ERBB2	5	0.80	0.86

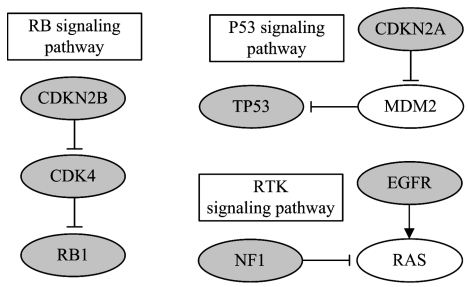

图 8　Dendrix 算法和 Multi-Dendrix 算法检测到的多形性成胶质细胞瘤的主要驱动通路

Fig. 8　Main glioblastoma driver pathways detected by Dendrix and Multi-Dendrix

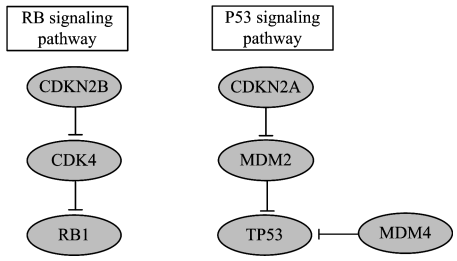

图 9 GNDP 算法检测到的多形性成胶质细胞瘤的主要驱动通路

Fig. 9 Main glioblastoma driver pathways detected by GNDP

结束语 本文提出的驱动通路识别算法 GNDP 将矩阵近似计算问题转化为子图挖掘问题。首先，由于构建基因互斥网络时，将满足高覆盖性和高互斥性条件的基因对连边，而将不满足高覆盖性和高互斥性条件的基因自动从搜索空间去除，因此有效降低了算法的复杂度；其次，通过在构建的基因互斥网络中搜索极大权重团，并将找到的极大权重团作为识别癌症的驱动通路，提高了识别的效率和准确率。

在模拟数据、肺腺癌以及多形性成胶质细胞瘤突变数据上对 GNDP 算法的运算效率、性能、准确率进行了验证。与经典的驱动通路识别算法 Dendrix 和 Multi-Dendrix 的实验对比结果表明，在模拟数据上，GNDP 不需要指定驱动通路的基因个数，能准确检测出所有人工设置的驱动通路；在肺腺癌和多形性成胶质细胞瘤突变数据上，GNDP 不需要任何先验知识就能达到较高的识别准确性。

参 考 文 献

[1] HANAHAN D, WEINBERG R A. The hallmarks of cancer[J]. Cell, 2000, 100(1): 57-70.

[2] FIDLER I J. The pathogenesis of cancer metastasis: the 'seed and soil' hypothesis revisited[J]. Nature Reviews Cancer, 2003, 3(6): 453-458.

[3] MCLENDON R, FRIEDMAN A, BIGNER D, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. [J]. Nature, 2008, 455(7216): 1061-1068.

[4] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma[J]. Nature, 2011, 474, 609-615.

[5] MEYERSON M, GABRIEL S, GETZ G. Advances in understanding cancer genomes through second-generation sequencing [J]. Nature Reviews Genetics, 2010, 11(10): 685-696.

[6] HUDSON T J, ANDERSON W, ARTEZ A, et al. . International network of cancer genome projects[J]. Nature, 2010, 464 (7291): 993-998.

[7] MARDIS E R, WILSON R K. Cancer genome sequencing: a review[J]. Human Molecular Genetics, 2009, 18(R2): R163.

[8] BARRETINA J, CAPONIGRO G, STRANSKY N, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity[J]. Nature, 2012, 483(7391): 603-607.

[9] MENG J, LI R, HAO H. Gene microarray data classification based on intersecting neighborhood rough set[J]. Computer Science, 2015, 42(6): 37-40. (in Chinese)
孟军，李锐，郝涵. 基于相交邻域粗糙集的基因微阵列数据分类[J]. 计算机科学，2015，42(6)：37-40.

[10] OVERDEVEST J B, THEODORESCU D, LEE J K. Utilizing the Molecular Gateway: The path to personalized cancer management[J]. Clinical Chemistry, 2009, 55(4): 684-697.

[11] SWANTON C, CALDAS C. Molecular classification of solid tumours: towards pathway-driven therapeutics[J]. British Journal of Cancer, 2009, 100(10): 1517-1522.

[12] ZHENG C H, YANG W, CHONG Y W, et al. Identification of mutated driver pathways in cancer using a multi-objective optimization model[J]. Computers in Biology & Medicine, 2016, 72: 22-29.

[13] GREENMAN C, STEPHENS P, SMITH R, et al. Patterns of somatic mutation in human cancer genomes[J]. Nature, 2007, 446(7132): 153-158.

[14] BATCHELOR T T, BETENSKY R A, ESPOSITO J M, et al. Age-dependent prognostic effects of genetic alterations in glioblastoma[J]. Clinical Cancer Research: An Official Journal of the American Association for Cancer Research, 2004, 10(1): 228-233.

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

[16] XIE Q Q, LI D F, ZHANG W. Two novel tree structure-based methods for gene selection[J]. Computer Science, 2015, 42(7): 250-253. (in Chinese)
谢倩倩，李订芳，章文. 两种基于树结构的基因选择算法[J]. 计算机科学，2015，42(7)：250-253.

[17] VANDIN F, UPFAL E, RAPHAEL B J. De novo discovery of mutated driver pathways in cancer[J]. Genome Research, 2012, 22(2): 375-85.

[18] DING L, GETZ G, WHEELER D A, et al. Somatic mutations affect key pathways in lung adenocarcinoma[J]. Nature, 2008, 455 (7216): 1069-1075.

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

[20] BOCA S M, KINZLER K W, VELCULESCU V E, et al. Patient-oriented gene set analysis for cancer mutation data[J]. Genome Biology, 2010, 11(11): R112.

[21] EFRONI S, BENHAMO R, EDMONSON M, et al. Detecting cancer gene networks characterized by recurrent genomic alterations in a population[J]. Plos One, 2011, 6(1): e14437.

[22] HAHN W C, WEINBERG R A. Modelling the molecular circuitry of cancer[J]. Nature Reviews Cancer, 2002, 2(5): 331-341.

[23] VASKE C J, BENZ S C, SANBORN J Z, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM [J]. Bioinformatics, 2010, 26(12): i237-i245.

[24] BASHASHATI A, HAFFARI G, DING J, et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer[J]. Genome Biology, 2012, 13(12): R124.

(下转第 242 页)