基于多组学数据识别癌症驱动通路的模型和算法

蔡齐荣[1] 吴璟莉[1,2]
(广西师范大学计算机科学与信息工程学院 广西 桂林 541004)[1]
(广西师范大学广西多源信息挖掘与安全重点实验室 广西 桂林 541004)[2]

摘 要 通过整合体细胞突变、拷贝数变异和基因表达等3种组学数据,提出识别癌症驱动通路的改进最大权重子矩阵模型。该模型用通路中基因平均权重调控覆盖度和互斥度,对权重大的基因集覆盖度进行加强,同时放松其高互斥度约束。引入基于贪心算法的重组算子,提出求解该模型的单亲遗传算法PGA-MWS。采用胶质母细胞瘤和卵巢癌数据集对算法PGA-MWS和GA进行实验对比分析。实验结果显示,较GA方法,基于改进模型的PGA-MWS算法能识别出覆盖度高但互斥度不太高的基因集,且其识别的基因集中,许多均参与已知信号通路,并被证实与癌细胞密切相关,同时还能识别几种潜在的候选驱动通路,因此PGA-MWS方法可作为检测癌症驱动通路的一种有效补充。

关键词 驱动通路,多组学数据,癌症,算法,模型
中图法分类号 TP301 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.09.047

Model and Algorithm for Identifying Driver Pathways in Cancer by Integrating Multi-omics Data

CAI Qi-rong[1] WU Jing-li[1,2]
(College of Computer Science and Information Technology, Guangxi Normal University, Guilin, Guangxi 541004, China)[1]
(Guangxi Key Lab of Multi-source Information Mining & Security, Guangxi Normal University, Guilin, Guangxi 541004, China)[2]

Abstract This paper proposed improved maximum weight submatrix problem model for identifying driver pathways in cancer by integrating somatic mutations, copy number variations, and gene expressions. The model tries to adjust coverage and mutual exclusion with the average weight of genes in a pathway, enhances the coverage of the gene set with large weight and relaxes its mutual exclusion constraint. By introducing a greedy based recombination operator, a parthenogenetic algorithm PGA-MWS was presented to solve the model. Experimental comparisons between PGA-MWS and GA were performed on glioblastoma and ovarian cancer datasets. Experimental results show that, compared with GA algorithm, PGA-MWS algorithm based on the improved model can identify gene sets with high coverage and less mutual exclusion. Many of the identified gene sets are involved in known signaling pathways, and have been confirmed to be closely related to cancer cells. Simultaneously, several potential drive pathways can also be discovered. Therefore, the proposed approach may become a useful complementary one for identifying driver pathways.

Keywords Driver pathway, Multi-omics data, Cancer, Algorithm, Model

1 引言

科学研究表明,癌症的形成和发展由体细胞基因组突变所导致[1]。体细胞基因组突变主要分为促使癌细胞无限增殖的“驱动突变”和对癌细胞增殖没有影响的“乘客突变”,正确识别促发癌症的关键功能突变即“驱动突变”,将为进一步了解癌症发病机理和研制抗癌药物等提供重要信息,为精准医疗或个性化医疗提供依据[2]。

近年来,癌症基因组图谱计划(The Cancer Genome Atlas,TCGA)[3]、国际肿瘤基因组协作组(the International Cancer Genome Consortium,ICGC)[4]等大规模癌症工程获得了海量癌症多组学数据,这使得利用计算方法来识别致癌“驱动突变”成为可能并得到生物信息学研究领域的广泛关注[5-6]。然而,由于基因突变存在互斥性,即便来自同一患者的两个样本数据,得到的突变基因也可能不同[5]。研究者们发现这些不同的基因突变通常靶向同一生物通路,通路中任何一个基因发生突变均会导致癌症的发生,癌症的发展实质上是由生物通路所控制[7]。因此,在通路层面上研究突变较在基因层面上研究突变更为准确,这对获取癌症的互斥模式至关重要[8]。癌症驱动通路识别问题应运而生,旨在找到驱

到稿日期:2018-07-20 返修日期:2018-10-28 本文受国家自然科学基金项目(61762015,61502111,61662007,61763003),广西自然科学基金项目(2015GXNSFAA139288,2016GXNSFAA380192),广西研究生教育创新计划项目(XYCSZ2018078),“八桂学者”工程专项,广西多源信息挖掘与安全重点实验室系统性研究基金项目(14-A-03-02,15-A-03-02),广西科技基地和人才专项(AD16380008)资助。
蔡齐荣(1993—),男,硕士生,主要研究方向为生物信息学、算法优化;吴璟莉(1978—),女,博士,教授,硕士生导师,CCF会员,主要研究方向为生物信息学、算法设计与分析,E-mail:wjlhappy@mailbox.gxnu.edu.cn(通信作者)。

动突变的组合，主要分为单驱动通路识别和协作驱动通路识别两大类[9]，本文主要研究单驱动通路识别问题。

目前大多数研究都基于先验生物通路知识来探测癌症驱动通路，然而，由于可利用的先验知识还很不完备且包含噪声数据[10]，因此，本文对不依赖任何先验知识的从头识别方法进行研究。2012 年，Vandin 等[10]利用驱动通路中基因的高覆盖性和高互斥性两个特性（高覆盖性指驱动通路应覆盖大量患者样本，即驱动通路中的基因应在大量患者中突变；高互斥性表示驱动通路中不同基因在同一样本上发生突变的几率很小[11]），提出基于突变数据求解通路识别问题的最大权重子矩阵问题，并给出基于马尔可夫链蒙特卡洛算法的求解方法 Dendrix。同年，Zhao 等[12]提出求解最大权重子矩阵问题的非线性规划算法 BLP 和遗传算法 GA，获得了比 Dendrix 方法更好的求解性能，并提出加入基因表达数据求解问题的思路。2013 年，Zhang 等[13]整合体细胞突变（Somatic Mutation）、拷贝数变异（Copy Number Variation）和基因表达（Gene Expression）3 种数据，利用高覆盖性、高互斥性、突变对其他基因的影响及基因表达高度相关等 4 个特征，提出基于网络的癌症突变核心模块识别方法，但当网络规模增大时，其求解性能下降。

综上，通过融合多组学数据信息，可在一定程度上减弱生物实验导致的测量噪声、错误突变及其误译等对识别的影响。基于该研究思路，本文重新建模最大权重子矩阵问题，使其能有效整合体细胞突变、拷贝数变异和基因表达 3 种数据，充分考虑高覆盖性、高互斥性、突变对其他基因的影响及基因表达高度相关等 4 个特征，平衡高覆盖性和高互斥性之间的关系；其次，针对该模型，引入基于贪心算法的重组算子，设计一种识别通路的单亲遗传算法 PGA-MWS。实验结果表明，基于改进模型，算法 PGA-MWS 可识别出许多具有生物意义的癌症驱动通路。

2　改进的带权子矩阵问题模型

假设有体细胞突变矩阵、拷贝数变异矩阵和基因表达矩阵，分别记为 $\boldsymbol{S}_{|p|\times|G_S|}$，$\boldsymbol{C}_{|p|\times|G_C|}$ 和 $\boldsymbol{E}_{|p|\times|G_E|}$，矩阵的行表示相同的癌症样本集 p，列分别表示候选基因集 G_S、G_C 和 G_E。矩阵 $\boldsymbol{S}$ 中每个元素 $s_{ij}\in\{0,1\}(i=1,2,\cdots,|p|,j=1,2,\cdots,|G_S|)$，若基因 j 在样本 i 中变异，则 s_{ij} 取值为 1，反之取值为 0；矩阵 $\boldsymbol{C}(\boldsymbol{E})$ 中每个元素为实数，即 $c_{ij}(e_{ij})\in R(i=1,2,\cdots,m,j=1,2,\cdots,|G_C|(|G_E|))$ 表示基因 j 在样本 i 中的拷贝数变异值（或表达量）。

令基因集 $G_A=G_S\cup G_C$，其在癌症样本集 p 上的取值记为矩阵 $\boldsymbol{A}_{|p|\times|G_A|}$，$a_{ij}\in\{0,1\}(i=1,2,\cdots,|p|,j=1,2,\cdots,|G_A|)$，即当 s_{ij} 取值为 1 或基因 j 处于样本 i 的统计显著变异区域时[12]，a_{ij} 取值为 1，反之取值为 0。为进一步整合突变矩阵 $\boldsymbol{A}$ 和表达矩阵 $\boldsymbol{E}$，采用预处理使其覆盖相同的基因集。为方便描述，预处理后的突变矩阵和表达矩阵记为 $\boldsymbol{A}_{|p|\times|G|}$ 和 $\boldsymbol{E}_{|p|\times|G|}$，其中基因集 $G=G_A\cap G_E$。

针对基因 $a_{-j}(j=1,2,\cdots,|G|)$ 取值，将样本划分为两个子集，即 $s_0^j=\{a_{i-}\mid a_{ij}=0,i=1,2,\cdots,|p|\}$，$s_1^j=\{a_{i-}\mid a_{ij}=1,i=1,2,\cdots,|p|\}$。利用 $\{e_{il}\mid e_{i-}\in s_0^j\}$ 和 $\{e_{il}\mid e_{i-}\in s_1^j\}(l=1,2,\cdots,|G|,l\neq j)$ 的差异显著性的 P 值表示基因 a_{-j} 对 a_{-l} 的影响，记为 $de(a_{-j},a_{-l})$。$de(a_{-j},a_{-l})$ 值越小，表示基因 a_{-j} 对 a_{-l} 的影响越大。因此基因 a_{-j} 的权重值 $w(a_{-j})$ 由其对其他基因影响的均值决定，如式(1)所示：

$$w(a_{-j})=|G_j^\tau|^{-1}\cdot\sum_{l\in|G_j^\tau|}(1-de(a_{-j},a_{-l})) \tag{1}$$

其中，G_j^τ 表示受基因 a_{-j} 影响最大的前 τ 个基因构成的基因集。$w(a_{-j})$ 值越大，表示基因 a_{-j} 的权重越大。

假设 $\boldsymbol{M}_{|p|\times k}$ 为矩阵 $\boldsymbol{A}$ 的任一子矩阵，令 $\Gamma(g)=\{a_{i-}\mid a_{ig}=1\}$ 记录基因 g 发生突变的样本，则 $\Gamma(\boldsymbol{M})=\bigcup_{g\in\boldsymbol{M}}\Gamma(g)$ 表示矩阵 $\boldsymbol{M}$ 覆盖的患者总数，用于衡量矩阵 $\boldsymbol{M}$ 的覆盖度。令 $\omega(\boldsymbol{M})=\sum_{g\in\boldsymbol{M}}|\Gamma(g)|-|\Gamma(\boldsymbol{M})|$ 衡量 $\boldsymbol{M}$ 中同一样本同时有多个基因突变的程度，即衡量矩阵 $\boldsymbol{M}$ 的互斥度。令 $\boldsymbol{R}(\boldsymbol{M})$ 表示 $\boldsymbol{M}$ 中基因之间的相关度，如式(2)所示：

$$\boldsymbol{R}(\boldsymbol{M})=\frac{1}{\binom{k}{2}}\sum_{g_1\neq g_2}|pcc(\boldsymbol{e}_{-g_1},\boldsymbol{e}_{-g_2})| \tag{2}$$

其中，$g_1,g_2\in\boldsymbol{M}$，$\boldsymbol{e}_{-g_1}$ 和 $\boldsymbol{e}_{-g_2}$ 为矩阵 $\boldsymbol{E}$ 的列向量，$pcc(\cdot)$ 是皮尔逊相关系数。

根据以上定义，本文对文献[10]的带权子矩阵问题模型进行改进，得到如下改进的问题模型：给定突变矩阵 $\boldsymbol{A}_{|p|\times|G|}$、表达矩阵 $\boldsymbol{E}_{|p|\times|G|}$ 和正整数 $k(k<|G|)$，在矩阵 $\boldsymbol{A}$ 中确定子矩阵 $\boldsymbol{M}_{|p|\times k}$，以使函数值 $W(\boldsymbol{M})$ 最大，如式(3)所示：

$$W(\boldsymbol{M})=\bar{w}\cdot|\Gamma(\boldsymbol{M})|-\omega(\boldsymbol{M})^{\alpha\cdot\bar{w}^{-1}}+\beta\cdot\boldsymbol{R}(\boldsymbol{M}) \tag{3}$$

其中，$\bar{w}=k^{-1}\cdot\sum_{g\in\boldsymbol{M}}w(g)$，$\bar{w}$ 越大表示矩阵 $\boldsymbol{M}$ 中基因的平均权重越大，其构成驱动通路的可能性也越大；同时其倒数 $\bar{w}^{-1}$ 可适当放松驱动通路的高互斥性约束，缓解覆盖度和互斥度的潜在矛盾；参数 α 和 β 用于平衡函数 $W(\boldsymbol{M})$ 中的各项取值。

3　PGA-MWS 算法

本节基于改进的带权子矩阵问题模型，提出求解驱动通路识别问题的单亲遗传算法 PGA-MWS。该算法的输入为矩阵 $\boldsymbol{A}_{|p|\times|G|}$，$\boldsymbol{E}_{|p|\times|G|}$ 及参数 k，输出为子矩阵 $\boldsymbol{M}_{|p|\times k}$。下面首先介绍 PGA-MWS 算法的关键要素，然后给出算法流程。

3.1　染色体编码及初始种群

染色体采用十进制编码方式，以 k 个基因构成的集合表示一个问题解，即 $X=\{x_1,x_2,\cdots,x_k\}(x_i=1,2,\cdots,|G|)$。初始染色体生成方法如下：1）随机生成 1 到 $|G|$ 的全排列，按排列顺序将基因划分为 $k^{-1}\cdot|G|$ 个集合，得到对应 $k^{-1}\cdot|G|$ 个子矩阵 $\boldsymbol{M}_1,\boldsymbol{M}_2,\cdots,\boldsymbol{M}_{|G|/k}$；2）令 $\max=\arg\max_{1\leqslant i\leqslant|G|/k}W(\boldsymbol{M}_i)$，则选择子矩阵 $\boldsymbol{M}_{\max}$ 的基因构成初始染色体。利用此方法生成 N 个初始染色体以构成初始种群。

3.2　适应度函数

由于每个染色体表示一个选定的驱动通路，因此应对其进行评估。给定染色体 X，适应度函数 $Fitness(X)$ 的定义如式(4)所示：

$$Fitness(X)=W(\boldsymbol{M}_X) \tag{4}$$

其中，$\boldsymbol{M}_X$ 表示染色体 X 对应的子矩阵。

3.3　选择算子

本文采用轮盘赌选择和精英策略来产生新一代种群。适

应度最高的个体直接从父代遗传到子代，然后运用轮盘赌选择算子来生成其余 $N-1$ 个个体。

3.4 重组算子

本节提出一种基于贪心策略的重组算子。具体步骤如下：首先，给定一个父代染色体 $X=\{x_1,x_2,\cdots,x_k\}$（$x_i=1,2,\cdots,|G|$），由此确定候选基因集合 $C_X=\{g\mid g\in G,g\notin X\}$；其次，从基因集 X 中随机删除一个基因，得到基因集 X'；最后，基于贪心策略，从候选集合 C_X 中选出基因 i，以使函数值 $Fitness(\boldsymbol{M}_{X'\cup\{i\}})$ 最大，对应于子矩阵 $\boldsymbol{M}_{X'\cup\{i\}}$ 的基因集 $X'\cup\{i\}$ 为后代染色体，即 $X=X'\cup i$。

3.5 算法流程

根据上述算法要素，给出 PGA-MWS 算法的流程，如算法 1 所示。

算法 1 PGA-MWS

输入：矩阵 $\mathbf{A}_{|p|\times|G|}$ 和 $\mathbf{E}_{|p|\times|G|}$，参数 k

输出：子矩阵 $\mathbf{M}_{|p|\times k}$

1. 设置最大演化代数 maxgen、最优值保持恒定的阈值 maxt，种群规模 N。生成初始种群 pop_0，计算 pop_0 中每个个体的适应值，将 pop_0 的最优个体保存到变量 best 中。初始化迭代次数 gen=0，最优值保持恒定的代数 t=0。
2. 若 gen>maxgen 或 t>maxt，转入步骤 4，否则转入步骤 3。
3. 将 best 个体放入 pop_{gen+1}，并运用轮盘赌选择算子选出 N−1 个个体放入 pop_{gen+1} 中。对 pop_{gen+1} 中每个个体 X_i(i=1,2,…,N)执行重组算子，若得到的新个体 X'_i 的适应值大于 X_i，则 $X_i=X_i'$。若 pop_{gen+1} 最优个体适应值大于 best 个体适应值，则更新 best 个体，t=0；否则 t=t+1。gen=gen+1，返回步骤 2。
4. 将 best 个体转换为基因集，由此得到子矩阵 **M**，并将其输出。

4 实验结果

本文利用两种真实的癌症数据进行实验测试，对改进模型和 PGA-MWS 算法的有效性进行验证，并与原模型及 GA 算法进行对比分析。实验在一台工作站（Intel(R) Core(TM) i5-6500 3.20 GHz CPU，内存为 8 GB）上进行，操作系统为 Windows 7，编译运行工具为 R3.4.1。

实验数据来自于文献[12]提供的胶质母细胞瘤和卵巢癌样本集。胶质母细胞瘤样本集包含 91 名患者样本的 SM 数据及 206 名患者样本的 CNV 和 GE 数据。经过预处理后（见改进的带权重子矩阵问题模型），得到覆盖 90 个样本和 1126 个基因的突变矩阵和表达矩阵，进一步过滤掉在样本中突变率低于 3%的基因，最终得到覆盖 90 个样本和 99 个基因的突变矩阵及表达矩阵。卵巢癌样本集包含 313 个患者样本的 SM 数据，以及 489 个患者样本的 CNV 和 GE 数据。经上述同样处理后，最终得到覆盖 313 个样本和 274 个基因的突变矩阵及表达矩阵。此外，在卵巢癌数据中，基因 TP53 的突变率超过 80%，远高于其他基因低于 25%的突变率；文献[15]报道基因 TTN 的突变可能是假象。因此，考虑到 TP53 的过高突变率和 TTN 突变的不准确性，将其从候选基因中移除，得到剩下的 272 个基因。

PGA-MWS 算法的参数设置如下：$N=100$，$maxgen=50$，$maxt=10$，$\tau=n/5$，$\alpha=0.7$，$\beta=10$。GA 算法的参数设置与识别过程同文献[12]一致，识别时首先根据突变矩阵得到初步驱动通路，再利用表达矩阵数据对初步结果进行分析以确定最终结果[12]。

4.1 胶质母细胞瘤样本

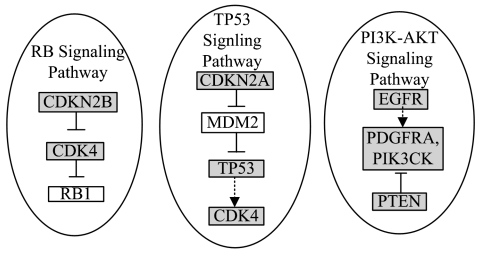
在表 1 中，针对参数 k 设置了 3 组实验：$k=2$，$k=3$ 和 $k=4$，其中加黑字体表示基因处于相同的生物通路。

表 1 胶质母细胞瘤样本实验结果比较

Table 1 Comparison results on glioblastoma samples

k	GA	PGA-MWS
2	**(CDK4,CDKN2B)** **(TP53,CDKN2A)**	**(CDK4,CDKN2B)**
3	**(CDK4,CDKN2B,RB1)**	**(CDK4,CDKN2A,TP53)**
4	(PIK3R1,NF1, TSPAN31,MTAP)	(PIK3R1,NF1, TSPAN31,MTAP)

当 $k=2$ 时，GA 算法首先根据突变数据得到基因集（CDK4，CDKN2B），（TP53，CDKN2A）和（TSPAN31，CDKN2B），再通过表达数据分析将（TSPAN31，CDKN2B）排除。虽然 CDK4 和 TSPAN31 属于一对元基因，但是 PGA-MWS 算法依据改进模型能够直接识别出更重要的基因对（CDK4，CDKN2B）（权值 $W(\cdot)=60.93$），而排除基因对（TSPAN31，CDKN2B）（权值 $W(\cdot)=60.36$）。根据 KEGG 数据库显示，（CDK4，CDKN2B）为 RB 信号通路（Signaling Pathway）的一部分[16]（见图 1），而无明显证据表明 TSPAN31 和 CDKN2B 之间的关系[12]。



注：实线箭头表示直接促进作用，虚线箭头表示间接促进作用，没有箭头的实线表示抑制作用，灰色矩形表示 PGA-MWS 识别的基因。线条与符号表示下同

图 1 PGA-MWS 识别的通路（胶质母细胞瘤）

Fig. 1 Pathways identified by PGA-MWS method (glioblastoma)

当 $k=3$ 时，GA 算法从突变矩阵中得到基因集（CDKN2B，RB1，TSPAN31）和（CDK4，CDKN2B，RB1），再经过表达矩阵相关性分析后得到后者基因集；PGA-MWS 算法直接得到基因集（CDK4，CDKN2A，TP53）。文献[12]的中的方法对覆盖度及互斥度的约束较强，旨在找到覆盖度大且互斥度高的基因集，如（CDK4，CDKN2B，RB1）的覆盖度 $|\Gamma(\cdot)|$ 和互斥度 $\omega(\cdot)$ 分别为 66 和 3。然而，覆盖度和互斥度的潜在矛盾会导致漏选某些潜在的驱动通路，如（CDK4，CDKN2A，TP53）具有很高的覆盖度（$|\Gamma(\cdot)|=72$），但其互斥度并不高（$\omega(\cdot)=12$）。本文模型通过指数 $\alpha\cdot\bar{w}^{-1}$ 来放松对高互斥度的约束，使互斥度不太高的通路被选中。根据 KEGG 数据库显示，（CDK4，CDKN2A，TP53）是 p53 信号通路的一部分[17]（见图 1），TP53 在细胞周期调控中起重要作用，并且大多数 TP53 突变会降低肿瘤抑制活性并促进肿瘤生长[18]。

当 $k=4$ 时，与文献[12]的处理方式相同，首先剔除已识别的基因，然后在余下基因中进行识别。GA 算法首先识别

出(CYP27B1，MTAP，PIK3R1，COL6A2)和(PIK3R1，NF1，TSPAN31，MTAP)，然后通过相关性分析保留了后者基因集；PGA-MWS 算法直接得到基因集(PIK3R1，NF1，TSPAN31，MTAP)。由于这两种算法一次探测得到的结果基因集均不在同一通路中，因此分别将这些基因剔除后，在剩余基因中进一步识别基因数为 4 的基因集。GA 算法识别出基因集(PIK3CA，EGFR，PTEN，MDM4)，其前 3 个基因存在于 PI3K-AKT 信号通路中[19]；PGA-MWS 算法识别出基因集(PIK3CA，PDGFRA，EGFR，PTEN)，其全部包含在 PI3K-AKT 信号通路中[19](见图 1)。PTEN 是重要的肿瘤抑制基因。其突变或缺失可能会加快细胞增殖和减缓细胞死亡，TP53 和 PTEN 基因的失活会促进胶质母细胞瘤的发展[20]。图 2 为 k 取不同值时识别的胶质母细胞瘤基因集的覆盖度和互斥度。其中，斜线条表示互斥突变，灰色条表示共同发生的突变，白色条表示无突变。

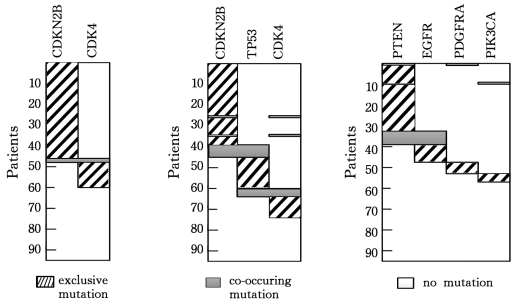


图 2　胶质母细胞瘤样本的基因集

Fig. 2　Gene sets identified from glioblastoma samples

4.2　卵巢癌样本

表 2 给出参数 k 在 3 种设置下的实验结果比较。当 $k=2$ 时，GA 和 PGA-MWS 得到的结果相同，均识别出基因集(MYC，CCNE1)。MYC 和 CCNE1 是参与细胞周期进程的两个重要基因[21](见图 3)。MYC 是一种强大的原癌基因，编码一种转录因子，经常在许多类型的癌症中持续表达[22]；基因 CCNE1 的扩增与生存率有关，其可能是卵巢癌的潜在治疗靶点[23]。

表 2　卵巢癌样本的实验结果比较

Table 2　Experimental results comparison on ovarian carcinoma samples

k	GA	PGA-MWS
2	(**MYC**，**CCNE1**)	(**MYC**，**CCNE1**)
3	(**RYR2**，**PPP2R2A**，KRAS)	(**RYR2**，**PPP2R2A**，BRD4)
4	(**MAPK8IP2**，**NF1**，STMN3，CASC1)	(**KRAS**，**MAPK8IP2**，**NOTCH3**，PRPF6)

当 $k=3$ 时，首先剔除基因 MYC 和 CCNE1，然后在余下基因中进行识别。GA 算法识别出基因集(RYR2，PPP2R2A，KRAS)，其覆盖度、互斥度和基因相关度分别为 97，8，2.25。PGA-MWS 算法识别出基因集(RYR2，PPP2R2A，BRD4)，其覆盖度、互斥度及基因相关度分别为 96，8，4.41。由于后者比前者的基因相关度大且改进模型利用基因集的平均权重对覆盖度进行调节，同时放松其高互斥度约束，这对(RYR2，PPP2R2A，BRD4)基因集获得较高的目标函数值 $W(\cdot)$ 起到积极作用，从而使其被识别。基因 RYR2 和 PPP2R2A 是肾上腺素信号通路的一部分[24](见图 3)，研究表明 PPP2R2A 的失调对儿童畸胎瘤中至少一部分生殖细胞肿瘤具有致病作用[25]；基因 BRD4 被认为是卵巢癌的潜在治疗靶点[26]。

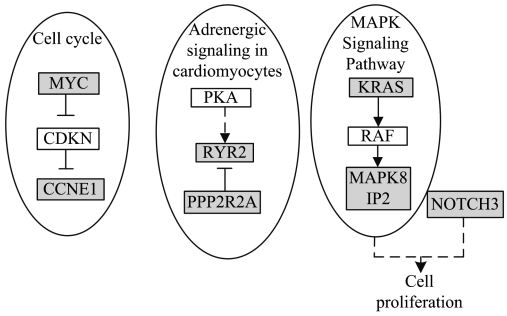


图 3　PGA-MWS 识别的通路(卵巢癌)

Fig. 3　Pathways identified by PGA-MWS method (ovarian carcinoma)

当 $k=4$ 时，分别将上述已识别基因剔除，然后在余下基因中进行识别。GA 识别出了 $\Gamma(\cdot)=106$ 及 $\omega(\cdot)=7$ 的基因集(MAPK8IP2，NF1，STMN3，CASC1)。PGA-MWS 识别出基因集(KRAS，MAPK8IP2，NOTCH3，KCNQ2)，其覆盖度和互斥度分别为 110，19。实验结果进一步表明改进模型能找出互斥度较低的基因集。基因 KRAS、MAPK8IP2 和 NOTCH3 均为 Apelin 信号通路的一部分[27](见图 3)，它们能调节细胞的增殖和分化；NOTCH3 与 KRAS 有共同的作用，它们都参与调节细胞增殖、分化和凋亡[28]。图 4 为不同 k 取值下识别的卵巢癌基因集的覆盖度和互斥度。

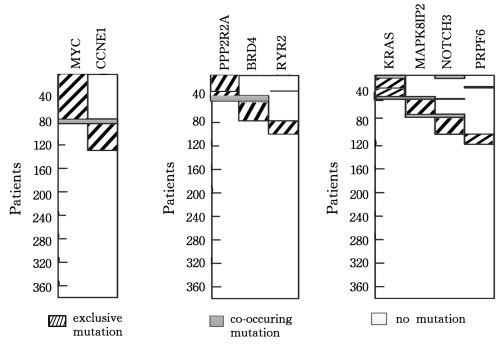


图 4　卵巢癌样本的基因集

Fig. 4　Gene sets identified from ovarian carcinoma samples

结束语　癌症驱动通路识别问题是生物信息学中的重要研究问题。本文整合体细胞突变、拷贝数变异和基因表达等 3 种组学数据，提出改进的最大权重子矩阵问题模型。该模型用通路中基因的平均权重对覆盖度和互斥度进行调控，加强平均权重大的基因集覆盖度，同时放松其高互斥度约束，使得覆盖度高但互斥度不太高的通路可以被识别。基于该改进模型，提出一种识别驱动通路的单亲遗传算法 PGA-MWS，该算法引入一种基于贪心策略的重组算子。

利用胶质母细胞瘤和卵巢癌数据集对 PGA-MWS 和 GA 算法进行实验测试，对比分析其在 $2\leqslant k\leqslant 4$ 参数设置下识别的驱动通路。实验结果表明，基于改进模型，PGA-MWS 算法能够有效识别出一些覆盖度高但互斥度不太高的重要基因集。根据 KEGG 数据库显示，已识别的许多基因参与已知信

号通路，并已被证实有致癌作用。综上所述，改进模型和PGA-MWS算法可成为识别癌症驱动通路的有效工具。

参 考 文 献

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

[2] GREENMAN C,STEPHENS P,SMITH R,et al. Patterns of somatic mutation in human cancer genomes [J]. European Journal of Cancer Supplements,2008,6(9):153-158.

[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 International Cancer Genome Consortium. International network of cancer genome projects [J]. Nature,2010,464(7291):993-998.

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

[6] DEES N D,ZHANG Q,KANDOTH C,et al. MuSiC:Identifying mutational significance in cancer genomes [J]. Genome Research,2012,22(8):1589-1598.

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

[8] 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.

[9] ZHANG J,ZHANG S. The Discovery of Mutated Driver Pathways in Cancer:Models and Algorithms [J]. IEEE/ACM Transactions on Computational Biology & Bioinformatics,2018,15(3):988-998.

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

[11] YEANG C H,MCCORMICK F,LEVINE A. Combinatorial patterns of somatic gene mutations in cancer [J]. Faseb Journal,2008,22(8):2605-2622.

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

[13] 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 [J]. Bmc Systems Biology,2013,7(S2):S4.

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

[15] THE CANCER GENOME ATLAS RESEARCH NETWORK. Integrated genomic analyses of ovarian carcinoma [J]. Nature,2011,474(7353):609-615.

[16] KEGG(Release86.1)[OL]. https://www.kegg.jp/kegg-bin/show_pathway?query=RB&map=map05200&scale=1.0&show_description=hide.

[17] KEGG(Release86.1)[OL]. http://www.kegg.jp/dbget-bin/www_bget? map04115.

[18] WARREN R S,ATREYA C E,NIEDZWIECKI D,et al. Association of TP53 mutational status and gender with survival after adjuvant treatment for stage III colon cancer:results of CALGB 89803 [J]. Clinical Cancer Research An Official Journal of the American Association for Cancer Research,2013,19(20):5777-5787.

[19] KEGG(Release86.1) [OL]. http://www.genome.jp/dbget-bin/www_bget?pathway:map04151.

[20] 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.

[21] KEGG(Release86.1)[OL]. http://www.kegg.jp/dbget-bin/www_bget?map04110.

[22] NAKAYAMA N,NAKAYAMA K,SHAMIMA Y,et al. Gene amplification CCNE1 is related to poor survival and potential therapeutic target in ovarian cancer [J]. Cancer,2010,116(11):2621.

[23] ENGLER D A,GUPTA S,GROWDON W B,et al. Genome Wide DNA Copy Number Analysis of Serous Type Ovarian Carcinomas Identifies Genetic Markers Predictive of Clinical Outcome [J]. Plos One,2012,7(2):e30996.

[24] KEGG(Release86.1)[OL]. http://www.kegg.jp/dbget-bin/www_bget?map04261.

[25] JIN Y,MERTENS F,KULLENDORFF C M,et al. Fusion of the Tumor-Suppressor Gene CHEK2 and the Gene for the Regulatory Subunit B of Protein Phosphatase 2 PPP2R2A in Childhood Teratoma [J]. Neoplasia,2006,8(5):413-418.

[26] BARATTA M G,SCHINZEL A C,ZWANG Y,et al. An in-tumor genetic screen reveals that the BET bromodomain protein,BRD4,is a potential therapeutic target in ovarian carcinoma [J]. Proceedings of the National Academy of Sciences of the United States of America,2015,112(1):232.

[27] KEGG(Release86.1) [OL]. http://www.genome.jp/dbget-bin/www_bget?pathway:map04371.

[28] RICCIARDELLI C,OEHLER M K. Diverse molecular pathways in ovarian cancer and their clinical significance [J]. Maturitas,2009,62(3):270-275.