中国科学: 信息科学 2023 年 第 53 卷 第 2 期: 250–265
SCIENTIA SINICA Informationis
论文

基于单细胞数据的癌症协同驱动模块识别方法

陈希[1], 王峻[1,2*], 余国先[1,2], 崔立真[1,2], 郭茂祖[3]

1. 山东大学软件学院, 济南 250101
2. 山东大学 - 南洋理工大学人工智能国际联合研究院, 济南 250101
3. 北京建筑大学电气与信息工程学院, 北京 100044
* 通信作者. E-mail: kingjun@sdu.edu.cn

收稿日期: 2022–02–04; 修回日期: 2022–04–01; 接受日期: 2022–04–11; 网络出版日期: 2023–02–03

国家自然科学基金 (批准号: 62072380, 62031003) 和山东大学中央高校基本业务费项目 (批准号: 2020GN061) 资助

摘要 从大规模生物组学数据中准确识别导致癌症发生的协同驱动模块是生物信息学研究领域重大课题之一. 现有研究方法通常只基于批量组学数据进行识别, 忽视了细胞水平上的癌症异质性, 易受噪声影响. 针对上述问题, 本文提出了一种基于单细胞数据和先验知识指导的协同驱动模块识别方法 CDMFinder. 该方法首先利用基因在不同亚型及正常细胞表达数据间存在的特异性共表达信息, 融合基因交互网络, 优化形成分子功能关联网络, 在深入挖掘基因间功能关联的同时有效降低网络复杂度; 再基于重叠马尔可夫 (Markov) 聚类从该网络中挖掘功能簇, 并提出基于融合权重和贪心策略的驱动模块识别方法, 从功能簇中获得驱动模块集合; 最后, 融合功能交互网络与突变共现定义模块距离函数, 识别获取协同驱动模块. CDMFinder 充分融合评估了表达、突变、差异分析等多种因素, 展现了优良的识别性能. 在乳腺癌和胶质母细胞瘤多组学数据上的实验结果表明, 本文方法能够识别出超过对比方法 1.35 倍的驱动基因, 识别到的协同驱动模块在功能/通路水平富集度上超过现有算法 1.5 倍.

关键词 单细胞数据, 协同驱动模块, 分子功能关联网络, 马尔可夫聚类, 多组学数据融合

1 引言

癌症, 作为致死率最高的疾病之一, 其形成和发展机制十分复杂, 涉及基因组、转录组及表观组等生命体活动的各个层级 [1]. 现有研究表明癌症是由基因突变 (单核苷酸多态性, 拷贝数变异, 核苷酸序列重复、插入以及缺失等) 累积导致的, 这种对癌症的发生起促进作用的突变被称为驱动突变 (driver mutation), 驱动突变位点所处的基因被称为驱动基因 (driver gene) [2].

近年来, 随着高通量生物技术的发展, 大型癌症基因组项目, 如癌症基因组图谱 (TCGA [3])、基因组联盟 (ICGC [4]) 等, 产生和积累了丰富的高通量多组学癌症数据, 为研究者从系统层面深入解析癌

引用格式: 陈希, 王峻, 余国先, 等. 基于单细胞数据的癌症协同驱动模块识别方法. 中国科学: 信息科学, 2023, 53: 250–265, doi: 10.1360/SSI-2022-0057
Chen X, Wang J, Yu G X, et al. Cooperative driver module identification based on single cell data (in Chinese). Sci Sin Inform, 2023, 53: 250–265, doi: 10.1360/SSI-2022-0057

症机理提供了支撑. 然而, 仅依靠生物实验或者简单的统计分析方法, 很难从大规模生物组学数据中准确地识别与特定癌症类别相关的驱动突变和驱动基因. 因此, 针对大规模疾病遗传数据, 开发有效的计算方法实现精确高效的癌症驱动突变/基因集合识别, 是当前癌症信息学研究的一项重大挑战. 准确识别致癌遗传因子对癌症诊断、靶向药物开发以及癌症患者的精确、个性化治疗等诸多方面均有重要的理论和应用价值 [5].

2 相关工作

早期驱动突变/基因识别研究中, 研究人员主要关注单个驱动基因的识别, 主要采用的筛选方法是将患者基因突变的发生频率与正常样本基因的突变频率对比, 识别显著高频突变的基因. 如 Ding 等 [6] 使用在 250 个基因中鉴定的同义体细胞突变来估计背景突变率 (background mutation ratio, BMR), 以此识别了 26 个在肺腺癌中高频突变的驱动基因. 这种基于频率的方法局限性较大. 首先, BMR 受到序列上下文、突变位置、基因特异性等多种因素影响. 其次, 癌症驱动基因存在广泛的突变异质性, 同种癌症可能由不同的基因突变导致. 因此, 仅对单驱动基因进行研究存在不确定性和局限性. 近年来对癌症的深入研究表明, 癌症的产生是多基因共同作用的结果, 个体基因水平的异质性与特定基因集合 (即驱动模块) 密切相关, 癌症通常只由数量有限的驱动模块触发. 因此, 在模块水平上进行癌症发生相关遗传机制研究, 识别与癌症关联的驱动模块, 是当前探究癌症病理的关键所在.

研究发现, 癌症驱动模块具有高覆盖性和高互斥性 [7]. 高覆盖性是指驱动模块内的基因在大量样本中观测到突变, 高互斥性是指模块内的基因在同一样本中一般不同时发生突变, 通常一个基因突变就能够影响整个驱动模块. 现有驱动模块识别的研究大多基于上述先验知识进行展开 [8]. 如 Ciriello 等 [9] 提出了 MEMo (mutual exclusivity modules in cancer) 算法在基因交互网络中识别满足互斥规则的驱动模块. HotNet [10] 使用热扩散算法重新构建基因交互网络, 再从中检测具有最佳覆盖和互斥的子网络. Vandin 等 [7] 开发了 Dendrix, 通过在突变数据上建立权重函数, 再结合贪心和 MCMC 方法, 优化识别具有最高权重的基因集作为驱动模块. Zhang 等 [11] 结合表达相似的基因通常共同执行某种生物功能的特性, 提出了 MDPFinder, 该方法在 Dendrix 基础上加入基因表达数据, 并引入二元线性规划 (binary liner programming, BLP) 和遗传算法, 更充分地引入遗传信息指导, 并解决了贪心算法导致的局部最优解问题.

上述方法获取的驱动模块通常为单个或离散的多个基因集合, 未考虑模块间存在的协作关系, 因此这些方法被称为单驱动模块识别方法. 研究证明, 癌症的形成和发展过程更多地受到多个彼此间存在遗传调控或功能关联的基因簇 (驱动模块) 的协同作用影响 [12]. 因此, 识别存在协同作用的驱动模块集合能够更全面地阐释癌症相关机制. Leiserson 等 [13] 改进了 Dendrix, 提出了 Multi-Dendrix, 使用整数线性规划同时检测多个具有高权重的驱动模块, 但 Multi-Dendrix 没有考虑模块间共现性, 导致识别出的驱动模块缺乏协作性. Zhang 等 [14] 开发了 CoMDP, 定义了新权重函数, 将模块间突变高共现与模块内突变高覆盖和高互斥结合, 使用 BLP 识别存在协同作用的驱动模块. Ma 等 [15~17] 引入基因时序表达数据, 开发了多种提取方法获得多个驱动癌症发生并对应不同癌症阶段的动态模块. Yang 等 [18] 开发了 CDPath, 首先基于突变数据和基因交互网络, 利用整数线性规划获得多个驱动模块, 再使用马尔可夫 (Markov) 聚类对模块相关通路聚类, 将模块划分为协同驱动模块. 他们还提出了 CoPath [19], 使用贪婪搜索来探索具有共同下游的互斥模块, 再设计双正则双聚类方法将互斥模块判别为协同驱动模块. Liu 等 [20] 将 Vandin 等提出的最大权重基因集问题重新定义为具有连续和非凸松弛的成本函数的组合优化问题, 构建了一种从头发现驱动模块的算法 MCSS, 能够同时获得数个驱

动模块. Li 等 [21] 开发了 CDPLP, 首先对基因进行层次聚类获得协同突变模块, 通过链路预测补充通路间潜在联系, 最后基于模块和更新的通路交互网络识别协同驱动模块集.

细胞是生物体结构与功能的基本单位, 在细胞水平上进行癌症驱动因子的研究能够更好地揭示癌症内在遗传发育机制 [22], 而现有识别方法使用的多细胞批量测序数据, 其实际测序结果是多个细胞整体水平上的平均值. 由于细胞分化的异质性, 相同表型的细胞的遗传信息可能存在显著性差异, 很多低丰度的信息会在整体表征中丢失, 难以准确描述单个肿瘤细胞的特殊性. 单细胞测序技术能够在单个细胞水平上揭示细胞基因结构和基因表达状态, 反映细胞间的异质性, 精确分离肿瘤细胞和正常细胞, 准确划分癌症分子亚型 [23]. 因此, 在驱动模块识别过程中引入单细胞测序数据将能弥补传统批量高通量测序的遗传信息缺失. 此外, 已有研究者使用多种癌症类型之间驱动基因集的共性和特异性共同推断泛癌水平的驱动模块 [24]. 这类研究证明从遗传差异性中寻找共性是可行的. 与泛癌和癌症间的关系类似, 某种特定癌症的不同亚型也存在特异性和共性 [25], 有效利用这种共性可以帮助我们寻找驱动模块.

针对现有驱动模块集合识别中低丰度遗传信息易丢失, 对癌症在亚型水平的共性和特异性利用和建模不充分等问题, 本文引入单细胞测序数据和癌症亚型数据, 提出了一种基于单细胞测序数据和亚型特异性的协同驱动模块识别方法 CDMFinder. 该算法首先基于单细胞基因表达数据对每种亚型和正常细胞分别构建细胞水平的特异性基因共表达网络, 然后将这些网络组合为能够体现亚型共性的表达关联网络, 该网络表示了在不同亚型中普遍存在且与正常细胞差异明显的基因共表达关系. 然后, 本文引入了基因功能交互网络, 将其与获取的表达关联网络融合优化为功能关联网络, 以加强基因间的功能联系, 并降低网络复杂程度. 随后, 本文将重叠马尔可夫聚类应用在该网络上, 获得多个功能簇. 为有效挖掘基因表达数据和利用驱动模块的高覆盖性和高互斥性, 本文分别引入了差异表达分析和基因突变数据, 综合构建了模块权重评估函数, 然后在功能簇上使用贪心搜索识别驱动模块. 最后, 本文基于模块间的功能联系和突变共现性, 定义了模块间的协作距离函数, 将具有较小协作距离的驱动模块判定为协同驱动模块. 在乳腺癌及胶质母细胞瘤数据集上的实验结果表明, 相比现有算法, CDMFinder 能够更准确地识别驱动基因及协同驱动模块.

3 癌症协同驱动模块识别方法

CDMFinder 的总体算法流程如图 1 所示, 主要流程可以分为 4 部分: 步骤 1, 基于单细胞数据和基因互作信息的分子功能关联网络构建; 步骤 2, 基于重叠马尔可夫聚类的功能簇获取; 步骤 3, 基于融合权重和贪心策略的驱动模块识别; 步骤 4, 基于功能关联和突变共现的协同驱动模块判定. 下文将详细介绍算法细节.

3.1 基于单细胞数据与基因互作信息的分子功能关联网络构建

本文研究中将综合利用癌症的多种亚型和正常细胞对应的单细胞基因表达数据, 增强在多个癌症亚型间普遍存在的基因共表达关系, 同时降低与癌症发育相关性较低或无影响的噪声基因关联的不利影响, 获取能体现亚型共性的基因表达关联网络. 在此基础上, 引入基因交互网络, 构建对探究癌症发病机制有指导作用的基因分子功能关联网络, 降低原始基因表达关联网络的复杂度, 增强基因间的生物学功能联系, 为后续驱动基因和驱动模块的识别提供指导, 为实验结果的可解释性提供支撑.

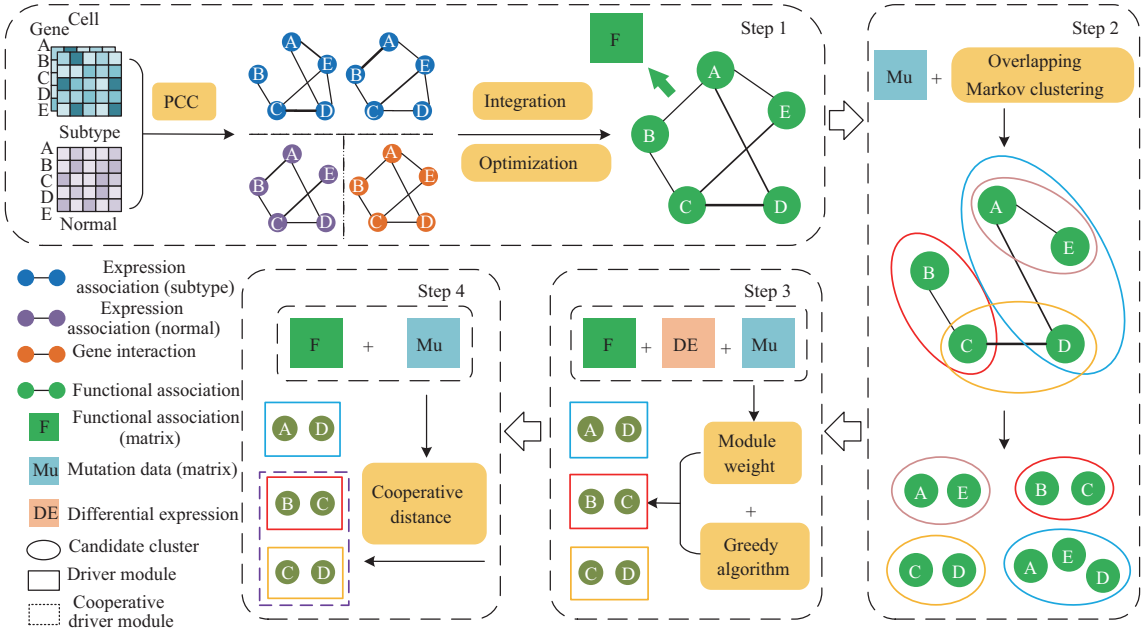

图 1 (网络版彩图) CDMFinder 流程示意图

Figure 1 (Color online) The workflow of CDMFinder

本文首先利用单细胞基因表达数据, 构建目标癌症的不同亚型对应的特异性表达关联网络. 对于基因对 i, j, 使用皮尔森 (Pearson) 相关系数的绝对值衡量基因间的共表达关系:

$$\boldsymbol{E}^s_{ij} = |\mathrm{Pcc}(\boldsymbol{X}^s_{i.}, \boldsymbol{X}^s_{j.})|, \tag{1}$$

其中, $\boldsymbol{E}^s \in \mathbb{R}^{m\times m}$ 和 $\boldsymbol{X}^s \in \mathbb{R}^{m\times d_s}$ 分别表示第 s 个亚型的表达关联网络矩阵和 TPM 格式的原始单细胞基因表达数据矩阵, m 表示考察基因总数, d_s 表示第 s 种癌症亚型的细胞数, i, j 对应基因 i 和基因 j. $\boldsymbol{X}^s_{i.}$ $(\boldsymbol{X}^s_{j.}) \in \mathbb{R}^{d_s}$ 分别代表基因 i (j) 在第 s 种亚型上的对应表达向量. Pcc() 表示皮尔森相关系数, 它描述两个基因间表达情况线性相关强弱的程度, 其绝对值越大表明相关性越强. 研究表明, 在生命活动中功能存在关联的基因往往呈现出共表达的趋势, 因此 $\boldsymbol{E}^s_{ij}$ 的值越接近 1, 表明基因 i 和基因 j 越有可能在癌症形成的过程中执行相同的功能.

同理, 本文使用正常细胞的单细胞表达数据生成其对应的共表达关联网络 $\boldsymbol{E}^n$:

$$\boldsymbol{E}^n_{ij} = |\mathrm{Pcc}(\boldsymbol{X}^n_{i.}, \boldsymbol{X}^n_{j.})|, \tag{2}$$

其中, $\boldsymbol{E}^n \in \mathbb{R}^{m\times m}$ 和 $\boldsymbol{X}^n \in \mathbb{R}^{m\times c}$ 分别表示正常细胞的表达关联网络矩阵和 TPM 格式的基因表达数据矩阵, c 表示正常细胞数. $\boldsymbol{E}^n$ 用于筛除 $\boldsymbol{E}^s$ 中的与目标癌症不相关的共表达关联.

基于现有生物学研究成果 [26, 27], 本文提出以下假设:

(1) 在不同亚型中都保持紧密关联的基因对癌症发展具有更大贡献.

(2) 在肿瘤和正常细胞中均表达相似的基因可能与生命体的一般活动相关.

(3) 在肿瘤和正常细胞中共表达关系差异大的基因更有可能驱动癌症的发生.

基于上述假设, 本文将构建的癌症亚型和正常细胞对应的多个共表达网络进行融合, 综合利用单

细胞测序数据中包含的丰富遗传信息, 定义癌症基因共表达关联网络:

$$\boldsymbol{E}=\left|\sum_{s=1}^{u}(\boldsymbol{E}^{s}-\boldsymbol{E}^{n})\right|, \tag{3}$$

其中 $\boldsymbol{E}\in\mathbb{R}^{m\times m}$ 表示融合的共表达关联网络矩阵, u 表示亚型总数, $\boldsymbol{E}^{s}-\boldsymbol{E}^{n}$ 剔除了亚型关联网络潜在的与癌症无关的共表达关联.

仅利用皮尔森相关系数难以准确描述基因间的遗传关联, 且生成的共表达网络规模过于庞大, 可能保留大量低质量的基因共表达关系, 进行后续计算分析时也将面临较高的时空复杂度. 因此, 本文引入基因交互网络 $\boldsymbol{G}\in\mathbb{R}^{m\times m}$ 作为先验知识, $\boldsymbol{G}$ 描述了基因间共同参与信号传递, 能量和物质代谢及细胞周期调控等生命过程时的相互作用强度. 将 $\boldsymbol{G}$ 作为先验知识指导 $\boldsymbol{E}$ 进行优化, 能够加强功能关联网络中基因间的功能协作关系, 使最终的识别结果更具可解释性, 也能有效降低网络复杂程度, 提高算法的运行效率. 具体优化方程如下:

$$\boldsymbol{F}=\boldsymbol{E}\odot\boldsymbol{G}, \tag{4}$$

其中, $\odot$ 为矩阵间的哈达玛积 (Hadamard product), $\boldsymbol{F}\in\mathbb{R}^{m\times m}$ 为优化的癌症关联基因功能关联网络矩阵.

通过式 (3), 本文整合了来自不同亚型和正常细胞的表达关联网络, 获取了能够描述特定癌症各亚型中广泛存在且和正常样本存在明显差异的基因共表达关系的表达关联网络, 再利用式 (4) 引入基因交互网络, 对表达关联网络进行优化, 支撑后续癌症驱动模块识别.

3.2 基于重叠马尔可夫聚类的功能簇获取

在生物学数据上应用聚类技术能够有效识别具有相似或协同功能的生物分子簇, 帮助研究者探索生物分子间可能存在的相互作用 [28]. 其中, 马尔可夫聚类 (Markov clustering, MCL) 作为常用的聚类算法, 被广泛应用于蛋白质相互作用网络、基因交互网络等重要生物网络的分析 [29]. MCL 是一种基于不同节点之间的转移概率 (随机游走) 的图聚类算法, 它对转移矩阵进行一系列的膨胀、扩展等操作, 使其达到收敛. 在最终状态下所在行不全为零的节点被称为吸引节点 (attractor), 它聚集着所在行中所有值为正的点, 可以被视为聚类中心. 相比 K-means 等传统聚类算法, 马尔可夫聚类更能容忍数据中存在的噪声, 并能有效发现高质量的功能簇 [30].

传统马尔可夫聚类簇间不共享元素, 同一生物分子如基因或蛋白质只能被划分到一个功能簇. 而大量研究表明, 生物分子通常共同行使多种不同的生命功能, 联合驱动生命活动 [31]. 针对这一情况, Shih 等 [32] 提出了重叠马尔可夫聚类, 迭代地运行马尔可夫聚类来产生大量簇, 并通过惩罚参数对迭代得到的吸引节点进行惩罚, 以保证每次迭代生成的簇与之前不一致, 以此保证了在不同的迭代中, 相同因子能够有机会被划分到不同的功能簇中. 本文在 3.1 小节生成的功能关联网络 $\boldsymbol{E}$ 上应用重叠马尔可夫聚类, 以获得多个保持紧密功能关联, 又具有互斥性的功能簇, 具体流程如算法 1 所示.

算法中引入标识基因对之间是否存在互斥关系的互斥网络指示矩阵 $\boldsymbol{M}\in\mathbb{R}^{m\times m}$, 以保证生成的功能簇保持基因间的互斥性, 同时也能够降低网络的复杂程度, 减小功能簇的规模, 方便后续的模块识别. 对于任意两个基因 i,j, 本文基于目标癌症突变数据 $\mathbf{Mu}$ ($\mathbf{Mu}\in\mathbb{R}^{m\times b}$ 为突变矩阵, b 表示批量样本总数, 具体定义见补充材料 A), 将其对应的互斥值定义为下式:

$$\boldsymbol{M}_{ij}=\begin{cases}0, & \frac{|\Gamma(i)\cup\Gamma(j)|}{|\Gamma(i)|+|\Gamma(j)|}<\delta,\\ 1, & \frac{|\Gamma(i)\cup\Gamma(j)|}{|\Gamma(i)|+|\Gamma(j)|}\geqslant\delta,\end{cases} \tag{5}$$

Algorithm 1 Functional cluster acquisition based on overlapping Markov clustering

Require: Functional association network $\boldsymbol{F}$, mutual exclusion network $\boldsymbol{M}$, the number of iterations t (usually set to 10), the inflation parameter r, the penalized ratio β.

Ensure: The set of functional clusters $\mathcal{C}$.

$\mathcal{C} = \{\}$;
$\boldsymbol{a} = [0, 0, \ldots, 0]$; //The indicator vector of attractor nodes.
$\boldsymbol{F} = \boldsymbol{F} \odot \boldsymbol{M}$;
$\boldsymbol{P} = \text{Transition}(\boldsymbol{F})$; //Convert $\boldsymbol{F}$ to transition matrix.
For iter $= 1 \rightarrow t$ do
  Repeat
    $\boldsymbol{P} = \boldsymbol{P} \times \boldsymbol{P}$;
    $\boldsymbol{P} = \text{Inflate}(\boldsymbol{P}, r, \boldsymbol{a}, \beta)$; //Penalize the attractor nodes.
    $\boldsymbol{P} = \text{Prune}(\boldsymbol{P})$; //Renormalize $\boldsymbol{P}$ to transition matrix.
  Until $\boldsymbol{P}$ converges;
  For $v \in \text{attractors}(\boldsymbol{P})$ //Label the attractor nodes.
    $\boldsymbol{a}(v) = \boldsymbol{a}(v) + 1$;
  $\mathcal{C}_{\text{iter}} = \text{clusters}(\boldsymbol{P})$; //Collect functional clusters for each iteration.
$\mathcal{C} = \mathcal{C} \cup \mathcal{C}_{\text{iter}}$.

其中, $\Gamma(i)$ 和 $\Gamma(j)$ 分别代表基因 i, j 在 **Mu** 上发生突变的样本集合, $||$ 表示对集合取模, δ 表示阈值, 通常设置为 0.95. $\frac{|\Gamma(i) \cup \Gamma(j)|}{|\Gamma(i)| + |\Gamma(j)|}$ 描述了基因之间的互斥程度, 取最高值 1 时表明 i, j 在 **Mu** 上没有共享的样本.

膨胀函数 $\text{Inflate}(\boldsymbol{P}, r, \boldsymbol{a}, \beta)$ 表示对 $\boldsymbol{P}$ 执行膨胀操作, 对任意基因对 i, j, 该操作具体如下所示:

$$\boldsymbol{P}_{ij} = \boldsymbol{P}_{ij}^{r \times \beta^{\boldsymbol{a}(i)}}, \tag{6}$$

其中, 膨胀系数 r 和惩罚比率 β 分别设置为 2 和 1.25, $\boldsymbol{a}(i)$ 指示在此前的迭代过程中基因 i 为吸引节点的次数. 通过膨胀函数, 重叠马尔可夫聚类实现了每次迭代获得不同的功能簇, 单个基因能够被多个功能簇共有, 符合生物学中基因在不同条件下执行不同功能的特性. 算法 1 最终获得一个包含多个功能簇的集合 $\mathcal{C}$ 用于后续驱动模块识别.

3.3 基于融合权重和贪心策略的驱动模块识别

癌症驱动模块具有功能联系紧密, 模块内基因在癌症和正常细胞中表达情况存在差异, 在突变数据上具备高覆盖性和高互斥性等特性. 为有效识别候选驱动模块, 本文在前文构建的基因功能关联网络的基础上, 引入差异表达分析和基因突变数据, 定义了一种新的模块权重函数, 在 3.2 小节中获取的每个功能簇上基于贪心策略识别具有最大权重的基因子集作为候选模块, 再将所有候选模块按权重排序获取驱动模块集合, 具体步骤如下.

首先, 依据 3.1 小节构建的基因功能关联网络, 定义网络中基因子集 (候选模块) 的功能关联权重:

$$R(\mathcal{M}) = \frac{1}{2} \sum_{i,j \in \mathcal{M}} \boldsymbol{F}_{ij} \times \boldsymbol{M}_{ij}, \tag{7}$$

其中 $\mathcal{M}$ 表示考察的候选模块, 即候选基因子集, $\boldsymbol{F}_{ij} \times \boldsymbol{M}_{ij}$ 代表基因 i, j 之间的关联强度, 通过再次引入互斥网络 $\boldsymbol{M}$, 降低功能簇中非互斥的基因权重. $R(\mathcal{M})$ 是考察模块中所有互斥基因对间边权重

的和. $R(\mathcal{M})$ 值越大表明 $\mathcal{M}$ 内基因更可能执行相同功能, $\mathcal{M}$ 是驱动模块的可能性也越高. $R(\mathcal{M})$ 可帮助寻找网络中功能联系紧密的基因子集.

如果一个基因相比正常细胞在肿瘤细胞中显著高表达或低表达, 那么它就有可能对癌症的发生起促进作用 (原癌基因) 或失去了原本的抑制癌症功能 (抑癌基因). 基因差异表达分析可通过比较相同基因在不同条件下的表达水平, 以能体现生物学意义的方式识别癌症驱动基因. 为利用这种性质, 本文引入倍数变化法 (fold change) 对每个基因生成一个差异表达权重 DE:

$$\mathrm{DE}(i) = \left| \log_2 \left(\frac{\overline{\boldsymbol{Y}}_i^t}{\overline{\boldsymbol{Y}}_i^n} \right) \right|. \tag{8}$$

$\boldsymbol{Y}^t \in \mathbb{R}^{m\times d}$ 和 $\boldsymbol{Y}^n \in \mathbb{R}^{m\times c}$ 分别表示肿瘤和正常细胞对应的 Counts 格式的基因表达数据矩阵, d 表示所有肿瘤细胞的个数. $\overline{\boldsymbol{Y}}_i^t$ 和 $\overline{\boldsymbol{Y}}_i^n$ 分别表示基因 i 在肿瘤和正常样本中的平均表达量.

候选模块的差异表达权重由对应基因子集的差异表达权重的和定义:

$$D(\mathcal{M}) = \sum_{i\in\mathcal{M}} \mathrm{DE}(i). \tag{9}$$

$R(\mathcal{M})$ 和 $D(\mathcal{M})$ 与模块中基因数量均呈正相关, 为了限制模块内基因数, 以及更进一步利用驱动模块的高覆盖性和高互斥性, 本文再次利用目标癌症对应的突变数据 **Mu**, 计算获取候选模块对应的突变权重 [7]:

$$W(\mathcal{M}) = |\Gamma(\mathcal{M})| - \omega(\mathcal{M}) = 2|\Gamma(\mathcal{M})| - \sum_{i\in\mathcal{M}} |\Gamma(i)|, \tag{10}$$

其中 $\Gamma(\mathcal{M}) = \bigcup_{i\in\mathcal{M}} \Gamma(i)$ 表示考察基因集 $\mathcal{M}$ 在 **Mu** 上发生突变的全体样本集合. $|\Gamma(\mathcal{M})|$ 描述了 $\mathcal{M}$ 的覆盖性, $\omega(\mathcal{M}) = \sum_{i\in\mathcal{M}} |\Gamma(i)| - |\Gamma(\mathcal{M})|$ 描述了 $\mathcal{M}$ 中覆盖重叠 (即不互斥) 的部分. 显然, 随着模块内基因数的增加, $W(\mathcal{M})$ 并不总是保持增长.

结合式 (7), (9) 和 (10), 可定义候选模块的权重评估函数为

$$L(\mathcal{M}) = R(\mathcal{M}) + \lambda D(\mathcal{M}) + \gamma W(\mathcal{M}), \tag{11}$$

其中, λ 和 γ 为权重调节参数, 分别直接影响差异表达权重及突变权重在 $L(\mathcal{M})$ 中所占的比例, 同时共同作用间接调整功能关联权重. $R(\mathcal{M})$ 和 $D(\mathcal{M})$ 分别在细胞水平强调了考察集合 $\mathcal{M}$ 在多种亚型之间存在的功能共性和不同状况下的表达差异性, $W(\mathcal{M})$ 则体现了 $\mathcal{M}$ 的覆盖性和互斥性.

本文通过结合贪心策略和式 (11) 中定义的模块权重评估函数, 在功能簇集合 $\mathcal{C}$ 上进行驱动模块集合获取. 该算法首先在每个功能簇上寻找具有最大模块权重的基因对作为种子集合; 再基于贪心搜索策略, 对该簇内所有集合外节点进行判断, 若节点的加入使新模块具有最大权重且大于原有模块权重, 则将该节点作为新识别的候选基因加入基因集合中; 重复搜索步骤至该功能簇内没有可加入的基因, 将最终获取的基因集合作为该簇的候选模块; 在所有簇都搜寻完毕后, 依据每个候选模块的整体权重, 将排名靠前的候选模块判定为驱动模块. 该算法的具体流程见算法 2. 相比应用在全部基因上的 Dendrix 等方法, 本文提出的 CDMFinder 将贪心策略应用在由重叠马尔可夫聚类获得的功能簇上, 在保留贪心算法易实现、可解释性强及效率较高的优点的同时, 还在一定程度上避免了局部最优问题, 能够高效准确地获取目标癌症密切关联的多个驱动模块.

Algorithm 2 Driver module identification based on greedy algorithm

Require: The set of functional clusters $\mathcal{C}$, module comprehensive weight $L(\mathcal{M})$.

Ensure: The set of driver modules $\mathcal{D}$.

$\mathcal{D} = \{\}$;
For iter $= 1 \to |\mathcal{C}|$ do //Get a candidate module for each functional cluster.
  $\mathcal{D}_{\text{iter}} = \{i, j\}$; //$L(i \cup j)$ is the maximum weight of the seed module.
  Repeat
    For $g \in \mathcal{C}_{\text{iter}}$ do
      If ($L(\mathcal{D}_{\text{iter}} \cup g)$ is max weight)
        If ($L(\mathcal{D}_{\text{iter}} \cup g) > L(\mathcal{D}_{\text{iter}})$)
          $\mathcal{D}_{\text{iter}} = \mathcal{D}_{\text{iter}} \cup g$;
  Until no new gene added;
  $\mathcal{D} = \mathcal{D} \cup \mathcal{D}_{\text{iter}}$; //Collect candidate modules.
$\mathcal{D} = \text{process}(\mathcal{D})$; //Filter redundant modules.
$\mathcal{D} = \text{driver}(\mathcal{D})$; //Select driver modules by weight.

3.4 基于功能关联和突变共现的协同驱动模块识别

第 3.3 小节获取的驱动模块对应了考察基因集合中与癌症发生发育存在潜在关联的单个基因子集, 为进一步探索与癌症存在关联的不同模块间的协同作用机制, 本文将基于模块间的功能关联及不同模块在突变数据 **Mu** 上的共现性构建模块间距离函数, 进一步识别存在协同作用的驱动模块集合.

豪斯多夫距离 (Hausdorff distance) 是一种针对集合的距离度量, 它描述一个集合到另一个集合中最近点的最大距离 [33]. 与该概念相似, 本文将选择两个驱动模块中所有基因对在功能关联网络 $\boldsymbol{F}$ 上所有最大值间的最小值作为两个模块间的功能关联, 因此基于豪斯多夫距离定义驱动模块之间的功能联系如下:

$$\begin{aligned} h(\mathcal{I}, \mathcal{J}) &= \min_{i \in \mathcal{I}} \{\max_{j \in \mathcal{J}} \boldsymbol{F}_{ij}\}, \\ h(\mathcal{J}, \mathcal{I}) &= \min_{j \in \mathcal{J}} \{\max_{i \in \mathcal{I}} \boldsymbol{F}_{ji}\}, \\ H(\mathcal{I}, \mathcal{J}) &= \min\{h(\mathcal{I}, \mathcal{J}), h(\mathcal{J}, \mathcal{I})\}, \end{aligned} \tag{12}$$

其中, $\mathcal{I}$ 和 $\mathcal{J}$ 表示两个驱动模块 (基因集合), $h(\mathcal{I}, \mathcal{J})$ 和 $h(\mathcal{J}, \mathcal{I})$ 表示分别获取模块 $\mathcal{I}$ 到模块 $\mathcal{J}$, 模块 $\mathcal{J}$ 到模块 $\mathcal{I}$ 的最小功能关联. $H(\mathcal{I}, \mathcal{J})$ 代表消除不对称性后的最小功能关联. 显然, $H(\mathcal{I}, \mathcal{J})$ 的值越大, 表明 $\mathcal{I}$ 和 $\mathcal{J}$ 之间存在更多的功能交互, 有可能具有相同/类似的功能或在生命活动中共同执行某种功能, 也更可能是协同驱动模块.

研究发现, 协同驱动模块在突变数据上的基因突变存在大量共享样本, 即协同驱动模块之间存在突变共现性 [14]. 本文将模块间共现性定义为在 **Mu** 上模块共享样本数与考察模块对中较小模块样本数的比值:

$$\text{CO}(\mathcal{I}, \mathcal{J}) = \frac{|\Gamma(\mathcal{I}) \cap \Gamma(\mathcal{J})|}{\min\{|\Gamma(\mathcal{I})|, |\Gamma(\mathcal{J})|\}}. \tag{13}$$

$\text{CO}(\mathcal{I}, \mathcal{J})$ 揭示了模块 $\mathcal{I}$ 和模块 $\mathcal{J}$ 之间的共现程度, 其取值大小直接影响 $\mathcal{I}$ 和 $\mathcal{J}$ 之间协同性判定.

$H(\mathcal{I}, \mathcal{J})$ 和 $\text{CO}(\mathcal{I}, \mathcal{J})$ 均属于权重函数, 不能直接作为距离度量衡量模块间的协作性, 因而本文结

表 1　CDMFinder 在乳腺癌和胶质母细胞瘤数据上实验结果

Table 1　Experimental results of CDMFinder on BRCA and GBM datasets

Module set	Module	Genes	Module set	Module	Genes
1	1	ASPM, INCENP, TACC1, **RARA**	4	12	CBR1, GCLM, **TP53**, ALDH3A2, SORD
	2	ASPM, INCENP, TACC1, SHCBP1, KIF4A		13	**NF1**, **SMAD4**, SMAD9, PTPRK, **MAPK1**
	3	CASKIN2, IGHMBP2, **MAP3K1**		14	**EEF1A1**, RPL7, RPS15, UPF3A, EIF2B2
2	4	**ERBB3**, **NOTCH2**, **PIK3CA**		15	**EEF1A1**, RPL7, RPS15, UPF3A, EIF3B
	5	GRB2, MET, **PIK3CA**		16	**EEF1A1**, RPL7, RPS15, UPF3A, WDR13
	6	LPAR6, **PIK3CA**, TSHZ2		17	CDCA7L, PSIP1, ZNF197, FNBP1L, EAPP
	7	MET, **PIK3CA**, PPM1D	5	18	**MTOR**, STAT5B, **TP53**, GRB2
3	8	**BRD7**, DPF2, **TP53**		19	**MTOR**, **TP53**, XIAP, GCLM
	9	**FOXA1**, **GATA3**, **TBX3**, **PIK3CA**		20	STAT5B **TP53** ULBP1 GRB2 KIAA1549
	10	**FOXA1**, POU2F3, **TBX3**, **MTOR**		21	F3, STAT5B, **TP53**, TIMP2, THBS1
	11	**MAP3K1**, SHANK2, ZNF425, IBTK		22	F3, SHC1, **TP53**, **ERBB3**

合功能关联网络和突变数据, 定义描绘模块间协作性的距离函数如下:

$$S(\mathcal{I},\mathcal{J}) = \mathrm{e}^{-(H(\mathcal{I},\mathcal{J})+\theta \mathrm{CO}(\mathcal{I},\mathcal{J}))}, \tag{14}$$

其中 θ 为参数变量, 负责调控模块间功能关联和突变共现性的贡献.

结合融合功能关联和突变共现的距离函数 $S(\mathcal{I},\mathcal{J})$, 本文利用 K-means 对所有驱动模块进行聚类划分, 获取的模块簇聚集了具有相同/相似功能或协同完成某一生物功能的驱动模块. 被分配到同一聚簇中的驱动模块在完成对应生物功能过程中具有协同性, 这些驱动模块协同地影响癌症的发生发育. 因此, 本文最终将分配到相同簇中的驱动模块识别为协同驱动基因模块.

4　实验验证和结果分析

为验证本文提出的 CDMFinder 的有效性, 我们选择乳腺癌 (breast invasive carcinoma, BRCA) 和胶质母细胞瘤 (glioblastoma, GBM) 这两种危害性大、异质程度高的癌症作为研究对象, 具体数据来源及预处理细节见补充材料 A. 最终, 本文选择了 515 个细胞的单细胞表达数据、965 个批量样本的突变及表达数据和 2000 个基因作为乳腺癌实验数据; 选择了 859 个细胞的单细胞表达数据、146 个批量样本的突变及表达数据和 1000 个基因作为胶质母细胞瘤实验数据.

4.1　实验结果

本文在乳腺癌和胶质母细胞瘤多组学数据上应用 CDMFinder 以评估其效果. 为了分析式 (11) 中参数 λ 和 γ 对实验结果的影响, 以及确定贪心算法应选择的模块数, 本文对相关参数进行了深入的分析, 具体实验结果与分析见附录 B. 基于参数分析结果, CDMFinder 在乳腺癌数据集上设定 λ 和 γ 分别为 0.1 和 0.05, 在胶质母细胞瘤数据集上设定 λ 和 γ 分别为 0.1 和 0.2, 在两个数据集上最终选择的模块数均设定为 11. 最终 CDMFinder 将两组数据集上获取的驱动模块 (每组 11 个模块, 编号 1∼22) 分别划分为 3 组和 2 组协同驱动模块集. 具体实验结果如表 1 所示, 其中协同模块集 1, 2, 3 和协同模块集 4, 5 分别为乳腺癌和胶质母细胞瘤数据集的识别结果, 粗体表示已验证的驱动基因.

4.1.1　驱动基因识别分析

如表 1 所示, CDMFinder 识别的驱动模块包含 PIK3CA, TP53, MTOR, TBX3, MAP3K1 等

11 种乳腺癌关键驱动基因和 TP53, PTEN, MTOR, MAPK1, NF1 等 8 种胶质母细胞瘤关键驱动基因. 在两种数据集上, 绝大多数 (20/22=90.9%) 的驱动模块都包含至少 1 个或以上已验证的驱动基因. 尽管模块 2 不包含任何已知驱动基因, 但仍有实验证明 [34], KIF4A 参与多个重要的细胞过程, 能够调控癌细胞的增殖及细胞凋亡. 类似的, 模块 17 中, EAPP 能够通过抑制糖皮质激素激活促进胶质瘤细胞的增殖, 该激素分泌可能导致神经元细胞死亡 [35].

实验结果中包含的其他未验证基因也与乳腺癌/胶质母细胞瘤的发生发展存在潜在关联. 例如, 模块 5, 7 中的 MET 基因, 其高表达对激活包括 PI3K-AKT, RAP1 和 MAPK 等信号通路在内的多条乳腺癌相关通路均有积极影响, 抑制 MET 的表达被证实能有效降低癌症死亡率 [36]; 模块 6 中的 LPAR6 的表达在乳腺癌细胞中显著降低, 其表达较高的乳腺癌患者预后效果较好 [37]; 模块 11 中的 IBTK 能够赋予细胞凋亡抗性, 引发细胞不受控的恶性增殖, 最终导致癌症 [38]; 模块 12 中的 CBR1 基因能通过调整上皮间质转化来抑制癌细胞的侵袭 [39]; 模块 18, 20 中的 STAT5B 基因被证实能通过调节基因表达参与调控胶质母细胞瘤细胞的生长、侵袭和迁移 [40]. 上述分析结果均证明 CDMFinder 识别的基因与癌症的发生发展密切相关, 其识别到的未被验证的基因也有可能是潜在的驱动基因.

4.1.2 驱动模块识别分析

CDMFinder 识别的驱动模块内基因间存在明显的功能交互.

在乳腺癌数据集上: 模块 1, 2 中, ASPM 和 INCENP 共同参与细胞增殖、分裂和调节过程, 而乳腺癌正是乳腺上皮细胞在多种因素作用下, 发生增殖失控所导致的; 模块 4 中, PIK3CA 和 ERBB3 同属 ErbB 信号通路, 该通路的激活可以调节乳腺上皮细胞中 EMT 相关的侵袭和迁移, 从而促进转移过程中肿瘤细胞的迁移、内渗和外渗 [41]; 模块 5 中, PIK3CA 和 GRB2 同属 mTOR 信号通路, 它能够调节细胞蛋白质、酯类的合成, 控制细胞周期进程、细胞的扩增及迁移, 其异常激活会导致细胞无限扩增, 进而导致肿瘤 [42]; 模块 6 中的 LPAR6 与 PIK3CA 在 PI3K-Akt 信号通路上富集, 该通路不仅可以调控癌细胞的增殖, 还与肿瘤的侵袭密切相关, 在大量乳腺癌患者中显示出广泛的激活 [42]; 模块 5, 7 中 PIK3CA 和 MET 均与 Rap1 信号通路相关, 该通路能够抑制著名抑癌基因 Ras 的活性, 该基因能有效地抑制癌细胞的转移并启动癌细胞的程序性凋亡和自噬作用, 进而杀死癌细胞 [43].

在胶质母细胞瘤数据集上: 模块 13 中, SMAD4, SMAD9 和 MAPK1 均属于 TGF-β 信号通路, 它可以诱导胶质瘤起始细胞群进行自我更新, 同时不会促进正常细胞的增殖, 从而在胶质母细胞瘤发生过程中起关键作用 [44, 45]; 模块 18, 19 中 TP53 和 MTOR 同时与 PI3K/AKT 信号通路相关, 该通路激活有助于胶质瘤细胞不受限制地生长, 避免细胞凋亡, 并增强癌细胞的侵袭能力 [46]; 模块 22 中, TP53 和 ERBB3 不仅同属于 PI3K/AKT 信号通路, 还均与 MAPK 信号通路存在关联, 后者能够阻止癌症抑制基因的表达, 减弱癌细胞的凋亡, 从而诱发癌症 [47].

上述实验结果均证明了本文所识别的模块并非彼此独立的癌症驱动基因的集合, 而是一组存在功能合作, 关联紧密的协同驱动癌症发生的功能基因集合. 综上所述, 本文提出的 CDMFinder 算法能有效地识别与癌症发生发展相关的驱动模块.

4.1.3 协同驱动模块识别分析

为进一步探索驱动基因模块间的协作关系, 验证 CDMFinder 最终识别的协同驱动模块的有效性, 本文使用 STRING 数据库对获取的每个协同驱动模块集进行了 GO 和 KEGG 富集分析, 具体实验结果和详细分析在附录 C 给出. 实验分析结果表明, CDMFinder 识别的所有驱动模块集都在多个与癌症发生相关的生理功能/通路中富集. 更重要的是, 不同模块中的基因富集到了相同的功能/通路中,

表 2 驱动基因识别效果对比

Table 2 Comparison of driver gene identification

	BRCA				GBM			
Method	Number	Increase (%)	P-value	$F1$-score	Number	Increase (%)	P-value	$F1$-score
CDMFinder	**11**	–	**9.9319E−12**	**0.2366**	**8**	–	**9.3245E−08**	**0.2581**
Dendrix	4	175	3.5439E−06	0.1081	2	300	2.2981E−03	0.1026
MDPfinder	5	120	1.7741E−07	0.1333	3	167	1.0662E−04	0.1538
Multi-Dendrix	4	175	3.5439E−06	0.1081	3	167	2.4134E−04	0.1463
CoMDP	7	57	3.4456E−11	0.1867	3	167	2.4134E−04	0.1463
CDPath	6	83	7.9367E−08	0.1519	4	100	9.4676E−05	0.1702
MCSS	8	38	8.9031E−10	0.1928	4	100	1.2151E−06	0.1818

这证实了本文识别的协同驱动模块的确共同参与关键生命活动, 模块之间存在可证的协作关系, 对癌症的发生发展起到了驱动促进作用. CDMFinder 有着良好的驱动模块识别能力, 可以准确地识别多个驱动模块, 并揭示这些模块间的协作关系, 有效地探索离散模块间的潜在联系, 从而获取驱动癌症产生的协同驱动模块集.

4.2 对比实验分析

为了进一步验证 CDMFinder 的效果, 本文还将其与前文介绍的 6 种相关方法进行了对比, 其中包括两种单驱动模块识别方法: Dendrix [7] 和 MDPfinder [11], 4 种多模块识别方法: CDPath [18], MCSS [20], Multi-Dendrix [13] 和 CoMDP [14]. 这些方法的参数按照原文给出的范围设置, 单模块识别方法 Dendrix 和 MDPfinder 运行两次以获取模块对, Multi-Dendrix, CoMDP, CDPath 和 MCSS 均运行一次. 我们分别在基因水平和功能/通路水平两方面进行了对比分析.

4.2.1 基因水平对比分析

在基因水平上, 本文对比了各方法识别的驱动基因数量, 对比结果见表 2, 其中 P 值是在已验证驱动基因集上经超几何检验计算得出的统计值, 其值越小表明结果越显著, $F1$-score 为比对模块包含的基因与已验证的驱动基因计算得出的 $F1$ 分数, 其值越大表明方法精度越高. 增幅 (Increase) 表示 CDMFinder 相对其他方法识别驱动基因数量提升的幅度, 以百分数形式给出.

在乳腺癌数据集上, 如表 2 所示, 6 种现有对比方法中 Dendrix 和 Multi-Dendrix 识别的驱动基因最少, 且对应的 $F1$ 分数和 P 值最差, 这是因为它们只是单纯地找到多个具有最大权重的驱动模块, 没有考虑模块之间是否存在功能交互; 在单模块识别方法中, MDPfinder 引入了基因表达数据, 考虑了基因间存在的共表达关系, 因而效果优于前两种方法. 在多模块识别方法中, CDPath 表现稍差, 这可能是因为它对模块内基因的要求较为苛刻, 从而导致性能的损失. CoMDP 表现较好, 这是因为它除了考虑模块内的高互斥性和高覆盖性外, 还在优化目标中加入了模块间的突变共现性, 从而能找到一对功能联系紧密的驱动模块. 在所有对比方法中识别效果最好的是 MCSS, 它结合基因表达数据重新阐述了 Dendrix 提出的最大权重模块问题, 使用非凸优化技术计算新构建的成本函数, 能根据不同初始值同时获得多个具有最小损失的驱动模块.

在胶质母细胞瘤数据集上, 如表 2 所示, Dendrix 所识别驱动基因数量及 $F1$ 分数和 P 值仍是对比方法中最差的; 其次是 Multi-Dendrix 和 CoMDP, CoMDP 在此数据集上表现不佳的原因是使用的突变数据的样本数较少, 难以提供较高的基因区分度. 对比表 2 中不同数据集的识别效果也能看出, 突变

表 3 方法变种识别驱动基因对比

Table 3 Comparison of driver gene identification of variant methods

	BRCA				GBM			
Method	Number	Decrease (%)	P-value	$F1$-score	Number	Decrease (%)	P-value	$F1$-score
CDMFinder	**11**	–	**9.9319E−12**	**0.2366**	**8**	–	**9.3245E−08**	**0.2581**
CDMFinder (noSC)	9	−18.1	1.9672E−07	0.1731	4	−50	5.1778E−03	0.1143
CDMFinder (noGI)	4	−63	1.3900E−02	0.0727	3	−62.5	3.7261E−02	0.0811
CDMFinder (noST)	7	−36.4	2.6752E−06	0.1474	5	−37.5	2.6558E−04	0.1587

数据可靠性的下降导致所有方法的性能都出现了降低, 尤其是对于 Dendrix, Multi-Dendrix 和 CoMDP 这些使用单一突变数据的方法, 该影响更为显著. 例如 MDPfinder 和 Multi-Dendrix 及 CoMDP 识别数量相当, 但在两种显著性指标上优于后两者. 在所有对比方法中效果最好的依然是 MCSS, CDPath 在显著性方面略逊一筹.

与上述对比方法相比, 本文提出的 CDMFinder 能够显著识别出更多的驱动基因. 此外, 尽管已验证驱动基因相比全部基因数量较少, 导致所有 7 种方法的 P 值和 $F1$ 分数都相对较低, 但 CDMFinder 仍然在统计显著性上显示出了一定的优势. 这是因为 CDMFinder 不仅利用单细胞亚型特异性表达数据分析了基因间的功能联系, 还引入多种先验知识和方法对基因模块进行了高互斥、高覆盖、差异表达等不同方面的综合度量, 这不仅有助于从多维度识别模块, 也能降低单一数据来源不可靠所带来的损失. 同时, 基于重叠马尔可夫聚类进行功能簇获取也使得 CDMFinder 能够允许同一基因出现在不同的模块中, 更符合生物学事实. 以上基因水平上的实验分析可以证实, CDMFinder 能够有效获取更多更准确的驱动基因及未验证的潜在驱动基因. 利用这些与癌症发生密切关联的驱动基因, CDMFinder 可以获取比其他方法更为多样化的驱动模块.

4.2.2 功能/通路水平对比分析

在功能/通路水平上, 本文仍使用 STRING 数据库对各方法进行 GO 和 KEGG 功能和通路富集分析, 比较各方法识别与癌症相关联的功能/通路的能力, 对比结果及分析见附录 D. 这些结果证实在乳腺癌和胶质母细胞瘤数据集上, CDMFinder 所识别的协同驱动模块均富集了更多的癌症相关功能/通路. 这是因为 CDMFinder 能从突变共现及功能关联等多个层面定义综合模块间的协作关系, 而其他方法往往仅基于某个特定维度. CDMFinder 不仅能识别出更多的驱动基因模块, 还能更准确地将存在协同功能关联的驱动模块判定为协同驱动模块. 另外, 实验分析表明, 相比其他方法, CDMFinder 所识别的协同驱动模块与目标癌症间存在更为密切的关联.

4.3 消融实验分析

为证明引入单细胞数据, 癌症亚型特异性和基因交互网络的重要性和有效性, 本文在两个数据集上对 CDMFinder 进行了仅使用批量数据 (noSC), 不对癌症进行分型 (noST) 和不引入基因交互网络 (noGI) 3 种情况下的变种实验对比, 并与 CDMFinder 的结果进行了基因水平和功能/通路水平的对比分析. 其中基因水平对比结果见表 3, 由于消融实验分析主要研究缺失信息对原方法造成的不利影响, 因此对比实验分析中的增幅在这里修改为更直观的降幅 (Decrease), 表示变种方法所识别驱动基因数相比原方法识别驱动基因数降低的幅度, 以负百分数表示. 在功能/通路水平上的具体对比结果及分析见附录 E.

从表 3 中可以看出, 在移除单细胞数据、基因交互网络和癌症亚型特异性后, CDMFinder 识别到的已验证乳腺癌驱动基因数量分别下降了 18.1%, 63% 和 36.4%, 识别到的已验证胶质母细胞瘤驱动基因数量分别下降了 50%, 62.5% 和 37.5%. 两种统计指标 P 值和 $F1$-Score 也显现出了明显的下降.

补充材料中的分析也表明, 在移除不同数据后, 尽管识别的驱动模块仍能富集到功能/通路, 但在功能/通路水平上 CDMFinder 的性能均出现了不同程度的下降. 这说明剔除这 3 种数据中的任何一种都会导致 CDMFinder 的性能损失, 有效利用这 3 类数据有助于准确识别协同驱动模块.

5 结束语

基于高通量生物数据挖掘协同驱动模块是当前癌症信息学研究的重点, 能够帮助研究者理解癌症发病机制, 辅助癌症的精确诊断和患者的个性化医疗. 本文提出了一种基于单细胞测序数据和多组学数据融合的协同驱动模块识别方法 CDMFinder. 该方法有效地利用单细胞数据刻画不同亚型的共表达模式, 通过融合基因交互网络, 差异表达分析和突变数据, 基于重叠马尔可夫聚类和贪心算法识别驱动模块, 最后通过融合的距离函数实现协同驱动模块的判定. 在乳腺癌和胶质母细胞瘤多组学数据集上基因和功能/通路水平的实验显示, 相比现有方法, CDMFinder 不仅能找到更多的驱动基因和驱动模块, 还能找到更具生物学意义的协同驱动模块. 在后续研究中, 我们还将对癌症亚型的共性和异质性融合及单细胞数据的特性展开更深入的探索, 识别不同亚型/细胞类型的特异性协同驱动模块.

补充材料 本文的补充材料包括以下内容: (A) 数据收集与预处理; (B) 参数分析; (C) 协同驱动模块识别分析; (D) 功能/通路水平对比分析; (E) 消融实验分析. 本文的补充材料见网络版 infocn.scichina.com. 补充材料为作者提供的原始数据, 作者对其学术质量和内容负责.

参考文献

1 Vogelstein B, Papadopoulos N, Velculescu V E, et al. Cancer genome landscapes. Science, 2013, 339: 1546–1558

2 Greenman C, Stephens P, Smith R, et al. Patterns of somatic mutation in human cancer genomes. Nature, 2007, 446: 153–158

3 Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature, 2011, 474: 609–615

4 International Cancer Genome Consortium. International network of cancer genome projects. Nature, 2010, 464: 993–998

5 Zhang J H, Zhang S H. The discovery of mutated driver pathways in cancer: models and algorithms. IEEE ACM Trans Comput Biol Bioinf, 2016, 15: 988–998

6 Ding L, Getz G, Wheeler D A, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature, 2008, 455: 1069–1075

7 Vandin F, Upfal E, Raphael B J. De novo discovery of mutated driver pathways in cancer. Genome Res, 2012, 22: 375–385

8 Wu H. Algorithm for detecting driver pathways in cancer based on mutated gene networks. Chin J Comput, 2018, 41: 1400–1414 [吴昊. 基于突变基因网络的致癌驱动通路检测算法. 计算机学报, 2018, 41: 1400–1414]

9 Ciriello G, Cerami E, Sander C, et al. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res, 2012, 22: 398–406

10 Vandin F, Upfal E, Raphael B J. Algorithms for detecting significantly mutated pathways in cancer. J Comput Biol, 2011, 18: 507–522

11 Zhao J H, Zhang S H, Wu L Y, et al. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics, 2012, 28: 2940–2947

12 Kanehisa M, Goto S, Sato Y, et al. KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res, 2012, 40: D109–D114
13 Leiserson M D M, Blokh D, Sharan R, et al. Simultaneous identification of multiple driver pathways in cancer. PLoS Comput Biol, 2013, 9: e1003054
14 Zhang J H, Wu L Y, Zhang X S, et al. Discovery of co-occurring driver pathways in cancer. BMC Bioinf, 2014, 15: 271
15 Li D, Zhang S, Ma X. Dynamic module detection in temporal attributed networks of cancers. IEEE ACM Trans Comput Biol Bioinf, 2022, 19: 2219–2230
16 Huang Z, Wang Y, Ma X. Clustering of cancer attributed networks by dynamically and jointly factorizing multi-layer graphs. IEEE ACM Trans Comput Biol Bioinf, 2021. doi: 10.1109/TCBB.2021.3090586
17 Ma X, Sun P G, Gong M. An integrative framework of heterogeneous genomic data for cancer dynamic modules based on matrix decomposition. IEEE ACM Trans Comput Biol Bioinf, 2022, 19: 305–316
18 Yang Z Y, Yu G X, Guo M Z, et al. CDPath: cooperative driver pathways discovery using integer linear programming and Markov clustering. IEEE ACM Trans Comput Biol Bioinf, 2019, 18: 1384–1395
19 Yang Z Y, Yu G X, Yu J, et al. CoPath: discovering cooperative driver pathways using greedy mutual exclusivity and bi-clustering. In: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019. 165–170
20 Liu B, Wu C, Shen X, et al. A novel and efficient algorithm for de novo discovery of mutated driver pathways in cancer. Ann Appl Stat, 2017, 11: 1481
21 Li S F, Wang J, Guo M Z, et al. Cooperative driver pathway discovery by hierarchical clustering and link prediction. In: Proceedings of IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2020. 115–120
22 Stuart T, Butler A, Hoffman P, et al. Comprehensive integration of single-cell data. Cell, 2019, 177: 1888–1902.e21
23 Navin N, Kendall J, Troge J, et al. Tumour evolution inferred by single-cell sequencing. Nature, 2011, 472: 90–94
24 Zhang J H, Zhang S H. Discovery of cancer common and specific driver gene sets. Nucleic Acids Res, 2017, 45: e86
25 Wang X, Wang J, Yu G X, et al. Network regularized bi-clustering for cancer subtype categorization. Chin J Comput, 2019, 42: 1274–1288 [王星, 王峻, 余国先, 等. 基于网络约束双聚类的癌症亚型分类. 计算机学报, 2019, 42: 1274–1288]
26 van Dam S, Võsa U, van der Graaf A, et al. Gene co-expression analysis for functional classification and gene-disease predictions. Brief Bioinform, 2018, 19: 575–592
27 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol, 2005, 4: 1–45
28 Jiang D X, Tang C, Zhang A D. Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng, 2004, 16: 1370–1386
29 Satuluri V, Parthasarathy S, Ucar D. Markov clustering of protein interaction networks with improved balance and scalability. In: Proceedings of the 1st ACM International Conference on Bioinformatics and Computational Biology, 2010. 247–256
30 Vlasblom J, Wodak S J. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. BMC Bioinf, 2009, 10: 1–4
31 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: tool for the unification of biology. Nat Genet, 2000, 25: 25–29
32 Shih Y K, Parthasarathy S. Identifying functional modules in interaction networks through overlapping Markov clustering. Bioinformatics, 2012, 28: i473–i479
33 Huttenlocher D P, Klanderman G A, Rucklidge W J. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Machine Intell, 1993, 15: 850–863
34 Hou P F, Jiang T, Chen F, et al. KIF4A facilitates cell proliferation via induction of p21-mediated cell cycle progression and promotes metastasis in colorectal cancer. Cell Death Dis, 2018, 9: 477
35 Chen K, Ou X M, Wu J B, et al. Transcription factor E2F-associated phosphoprotein (EAPP), RAM2/CDCA7L/JPO2 (R1), and simian virus 40 promoter factor 1 (Sp1) cooperatively regulate glucocorticoid activation of monoamine oxidase B. Mol Pharmacol, 2011, 79: 308–317
36 Gherardi E, Birchmeier W, Birchmeier C, et al. Targeting MET in cancer: rationale and progress. Nat Rev Cancer,

2012, 12: 89–103
37 Tao K, Guo S, Chen R, et al. Lysophosphatidic acid receptor 6 (LPAR6) expression and prospective signaling pathway analysis in breast cancer. Mol Diagn Ther, 2019, 23: 127–138
38 Vecchio E, Golino G, Pisano A, et al. IBTK contributes to B-cell lymphomagenesis in *Eμ-myc* transgenic mice conferring resistance to apoptosis. Cell Death Dis, 2019, 10: 320
39 Murakami A, Yakabe K, Yoshidomi K, et al. Decreased carbonyl reductase 1 expression promotes malignant behaviours by induction of epithelial mesenchymal transition and its clinical significance. Cancer Lett, 2012, 323: 69–76
40 Liang Q C, Xiong H, Zhao Z W, et al. Inhibition of transcription factor STAT5b suppresses proliferation, induces G1 cell cycle arrest and reduces tumor cell invasion in human glioblastoma multiforme cells. Cancer Lett, 2009, 273: 164–171
41 Hardy K M, Booth B W, Hendrix M J C, et al. ErbB/EGF signaling and EMT in mammary development and breast cancer. J Mammary Gland Biol Neoplasia, 2010, 15: 191–199
42 Paplomata E, O'Regan R. The PI3K/AKT/mTOR pathway in breast cancer: targets, trials and biomarkers. Ther Adv Med Oncol, 2014, 6: 154–166
43 Zhang Y L, Wang R C, Cheng K, et al. Roles of Rap1 signaling in tumor cell migration and invasion. Cancer Biol Med, 2017, 14: 90–99
44 Peñuelas S, Anido J, Prieto-Sánchez R M, et al. TGF-β increases glioma-initiating cell self-renewal through the induction of LIF in human glioblastoma. Cancer Cell, 2009, 15: 315–327
45 Eichhorn P J A, Rodón L, González-Juncá A, et al. USP15 stabilizes TGF-β receptor I and promotes oncogenesis through the activation of TGF-β signaling in glioblastoma. Nat Med, 2012, 18: 429–435
46 McDowell K A, Riggins G J, Gallia G L. Targeting the AKT pathway in glioblastoma. Curr Pharm Des, 2011, 17: 2411–2420
47 Zohrabian V M, Forzani B, Chau Z, et al. Rho/ROCK and MAPK signaling pathways are involved in glioblastoma cell migration and proliferation. Anticancer Res, 2009, 29: 119–123

Cooperative driver module identification based on single cell data

Xi CHEN[1], Jun WANG[1,2*], Guoxian YU[1,2], Lizhen CUI[1,2] & Maozu GUO[3]

1. *School of Software, Shandong University, Jinan* 250101*, China*;
2. *Joint SDU-NTU Centre for Artificial Intelligence Research* (*C-FAIR*)*, Shandong University, Jinan* 250101*, China*;
3. *College of Electrical and Information Engineering, Beijing University of Civil Engineering and Architecture, Beijing* 100044*, China*
* Corresponding author. E-mail: kingjun@sdu.edu.cn

Abstract Identifying cooperative driver modules from large-scale cancer omics data is one of the key topics in bioinformatics. Current methods usually only focus on discovering driver modules from batch omics data and ignore the cancer heterogeneity on the cell level, which are venerable to batch-level noises. To overcome these limitations, we propose a cooperative driver module identification method (CDMFinder) based on single-cell data and prior knowledge. CDMFinder first utilizes the gene co-expression specificity of different cancer subtypes and normal cell expression data to construct expression association networks and fuses these networks with a gene interaction network to obtain a gene functional association network. Thus, it effectively reduces network complexity while capturing in-depth functional associations between genes. It then adopts an overlapping Markov clustering on this functional network to mine functional clusters, along with a greedy strategy with a hybrid weight function to identify driver modules from the clusters. Finally, it introduces an interaction and mutation co-occurrence-based distance function on driver module sets to identify cooperative driver modules. CDMFinder fully integrates a variety of genetic factors (i.e., expression, mutation, and subtype specificity) and manifests a prominent performance. Experimental results on the multi-omics data of breast cancer and glioblastoma show that the number of driver genes identified by CDMFinder is 1.35 times larger than that of competitive methods. The identified cooperative driver modules are enriched at the pathway/functional level and are over 1.5 times more than compared methods.

Keywords single-cell data, cooperative driver module, molecular functional association network, Markov clustering, multi-omics data fusion