https://crad.ict.ac.cn
62(12): 2974−2982, 2025

面向信息系统设计的癌症阶段生物标志物识别

陈伯林[1,4] 张劲雷[1] 张建君[1] 胡悠鹏[1] 王 腾[1] 杨曼婷[2] 卞 军[3] 尚学群[1,4]

[1](西北工业大学计算机学院 西安 710072)

[2](西北工业大学软件学院 西安 710072)

[3](西安交通大学附属儿童医院普外一科 西安 710003)

[4](大数据存储与管理工信部重点实验室(西北工业大学) 西安 710072)

(blchen@nwpu.edu.cn)

Identification of the Cancer Stage Biomarkers for Information System Design

Chen Bolin[1,4], Zhang Jinlei[1], Zhang Jianjun[1], Hu Youpeng[1], Wang Teng[1], Yang Manting[2], Bian Jun[3], and Shang Xuequn[1,4]

[1](*School of Computer Science*, *Northwestern Polytechnical University*, *Xi'an* 710072)

[2](*School of Software*, *Northwestern Polytechnical University*, *Xi'an* 710072)

[3](*Department of General Surgery*, *Xi'an Jiaotong University Affiliated Children's Hospital*, *Xi'an* 710003)

[4](*Key Laboratory of Big Data Storage and Management* (*Northwestern Polytechnical University*), *Ministry of Industry and Information Technology*, *Xi'an* 710072)

Abstract Cancer is an exceptionally complex and highly heterogeneous disease with dynamic changes. Its occurrence and development are accompanied by a large number of gene mutations and functional disorders. Identifying biomarkers related to cancer stages is crucial for understanding the pathogenic and developmental mechanisms of cancer. However, the existing research on cancer biomarker recognition often treat individual genes as isolated nodes and usually only focus on the binary classification of cancer, ignoring the significant differences among different stages of cancer. To overcome the above issues, we first construct a RRN (regression residual network) for each cancer stage, and then analyze the nodes and edges of RRN in each stage. After that, the multi-source data mining is conducted in biological pathways, and the entire process of cancer evolution is characterized along with stages. By doing this, both biomarkers for cancer binary classification and multi-stage classification are obtained, and they are validated on GSE10072 and GSE42171 datasets, respectively. The experimental results show that the obtained biomarkers ALDOA and NME1 achieve competitive accuracy compared with existing methods by using only two genes for lung adenocarcinoma, and the biomarkers consisting of 17 edges achieve the improved accuracy by 14.86% by comparing with existing methods in terms of multi-stage classification.

Key words cancer evolution; network biomarker; stage biomarker; expression pattern; information system

摘 要 癌症是一种复杂且动态变化的高异质性疾病.它的发生发展伴随着大量的基因突变与功能失调.识别癌症阶段相关的生物标志物,对于了解癌症的致病机理与发展机制至关重要.然而,现有的癌症生物标志物识别方法通常将各个基因看作是孤立的节点,且仅关注癌症的二分类效果,忽略了不同基因之间的交互关系变化以及不同病理阶段的显著差别.为解决上述问题,首先为癌症各阶段构建回归残差网络

收稿日期: 2024-08-01; **修回日期**: 2025-04-15

基金项目: 国家重点研发计划项目(2021YFA1000402); 国家自然科学基金项目(62433016, 61972320); 西安市科技局资助项目(22YXYJ0057); 中央高校基本科研业务费专项资金资助(D5000250279, G25KY0612610)

This work was supported by the National Key Research and Development Program of China (2021YFA1000402), the National Natural Science Foundation of China (62433016, 61972320), the Project of Xi'an Municipal Bureau of Science and Technology (22YXYJ0057), and the Fundamental Research Funds for the Central Universities (D5000250279, G25KY0612610).

（regression residual network, RRN），分析每个阶段 RRN 的节点和边，并结合生物通路进行多源数据挖掘，刻画了癌症随阶段演化的整个过程. 通过对癌症的演化分析，分别获得癌症二分类和阶段多分类的网络生物标志物，并在 GSE10072 和 GSE42171 数据集上进行了验证. 实验结果表明，仅使用 2 个基因 ALDOA 和 NME1 组成的生物标志物，可以在肺腺癌二分类问题上获得跟现有研究结果具有同样竞争力的预测准确度，而使用由 17 条边构成的阶段生物标志物，则可以在肺腺癌阶段多分类问题上获得比现有方法高出 14.86 个百分点的预测准确度.

关键词 癌症演化；网络生物标志物；阶段生物标志物；表达模式；信息系统

中图法分类号 TP391

DOI: 10.7544/issn1000-1239.202440650 **CSTR:** 32373.14.issn1000-1239.202440650

癌症是一种具有很强异质性的复杂疾病，是危害人类健康的致命疾病之一. 资料显示，恶性肿瘤是全球发病率最高的疾病之一，死亡率也很高[1-3]. 这主要是因为大多数癌症被发现时通常已处于中晚期，此时癌细胞已经扩散和转移，导致治疗效果不佳，预后效果也较差. 智慧健康相关的医疗信息系统设计对于癌症的早诊早治非常重要. 不过现有的癌症诊断生物标志物识别往往仅关注疾病组与控制组之间的二分类效果，忽略了不同癌症病理阶段之间的显著差异，使得现有的研究仍然缺乏实际应用的条件.

许多研究表明，癌症是由多个基因、miRNA 和其他生物大分子经过突变、转录后修饰、复杂的调控关系以及环境因素共同作用的结果. 复杂的作用结果使得同种癌症在不同阶段或不同亚型之间表现出很大的差异. 随着下一代测序技术的发展，RNA 测序（RNA sequencing, RNA-Seq）技术的应用越来越广泛. 高通量测序数据的发展带动了全基因组基因表达数据的爆炸式增长，为癌症相关研究提供了强有力的数据条件[4-6].

1 相关工作

目前有多种基于机器学习的方法可用于识别生物标志物[7-10]. 然而，由于基因数量众多而样本数量较少，现有的标志物识别方法通常需要对数据进行降维，才能找到若干个由差异表达基因组成的集合作为生物标志物[11-13].

然而，基因之间通常存在较强的相互作用关系，它们共同参与某些生物功能，进而影响癌症的发生. 因此，识别这些差异表达的基因和相互作用关系对于理解癌症的演化过程至关重要[14-15]. 传统差异表达分析方法通常不考虑基因之间相互作用的变化，存在较多的假阳性和随机性等问题. 现有的微分方程方法基于不同分组间相关性的变化建立生物分子网络以描述边的差异，但却忽略了不同组间一些强相关且表达模式发生变化的边，从而导致假阴性.

针对上述问题，本文提出了一种基于表达模式差异的癌症阶段回归残差网络（regression residual network, RRN）来分析癌症的演化过程，并在此基础上实现癌症阶段相关生物标志物的识别与验证. 具体来说，本文首先使用来自 TCGA（the cancer genome atlas）数据库的基因表达数据来构建阶段相关的 RRN. 其次，使用网页排名（PageRank）中心性来评估节点在 RRN 上的重要性. 然后，考虑到网络在不同阶段的变化，本文从节点和边 2 个角度分别分析癌症的演化过程，并根据阶段间的共性和特异性分别识别癌症的二分类网络生物标志物及阶段相关的多分类网络生物标志物.

本文所提出的方法是对现有方法的改进. 传统的方法通常仅关注差异表达的基因，而忽略了基因调控关系的变化，导致预测结果的精度不高. 而 RRN 则考虑了上述 2 种类型的特征，将基因及其相互作用关系用拓扑结构的形式关联起来，并分阶段构建 RRN，挖掘基因表达水平与边的变化过程（基因间作用关系），以获得最优的标志物集合. 本文所提出的方法对于医疗信息系统设计提供了基础.

2 材料与方法

2.1 基因表达数据与预处理

高通量测序技术的快速发展积累了大量关于癌症的组学数据. 一些常用的癌症数据库包括 TCGA 数据库①和基因表达综合数据库（gene expression omnibus,

① https://portal.gdc.cancer.gov/

GEO)[①]. 这些数据库提供关于癌症的多种不同的数据类型, 例如基因组、转录组、蛋白质组、表观遗传和其他组学数据. 此外, 这些数据库还收录了肿瘤样本的病理阶段信息, 这对于研究癌症发生发展的演化机制至关重要.

本文的数据来自 TCGA 数据库中肺腺癌(lung adenocarcinoma, LUAD)的 RNA-Seq 基因表达数据[16-17], 共 525 个样本, 包括 58 个癌旁样本(正常样本)和 467 个癌症样本. 通过从临床数据中提取样本的阶段信息, 本文将癌症样本划分为 4 个阶段, 包括 58 个正常样本、256 个第 1 阶段样本、113 个第 2 阶段样本、73 个第 3 阶段样本、25 个第 4 阶段样本.

数据预处理包含以下 3 个步骤: 首先是在原始表达矩阵中对于每个基因的表达值使用对数变换进行标准化, 以减少数据的偏斜和方差. 其次是对于标准化的数据进行 Min-Max 归一化, 将数据尺度映射到 [0, 1] 区间[18] 上. 上述数据预处理步骤可以用公式化的语言描述, 具体如式(1)所示.

$$x'' = \frac{\ln(x+1) - \min(\ln(x+1))}{\max(\ln(x+1)) - \min(\ln(x+1))}, \tag{1}$$

其中, x表示某个基因在某个样本中的原始表达值, x''表示经过标准化和归一化处理后的数据. 最终进行数据过滤, 即对基因进行过滤, 删除在一半以上样本中表达为 0 的基因, 以减少噪声和提高分析的准确性, 最终获得 17 769 个待分析的基因.

2.2 阶段相关 RRN 的构建

RRN 使用线性回归模型来描述基因之间表达模式的变化. 具体网络构建流程如图 1 所示.

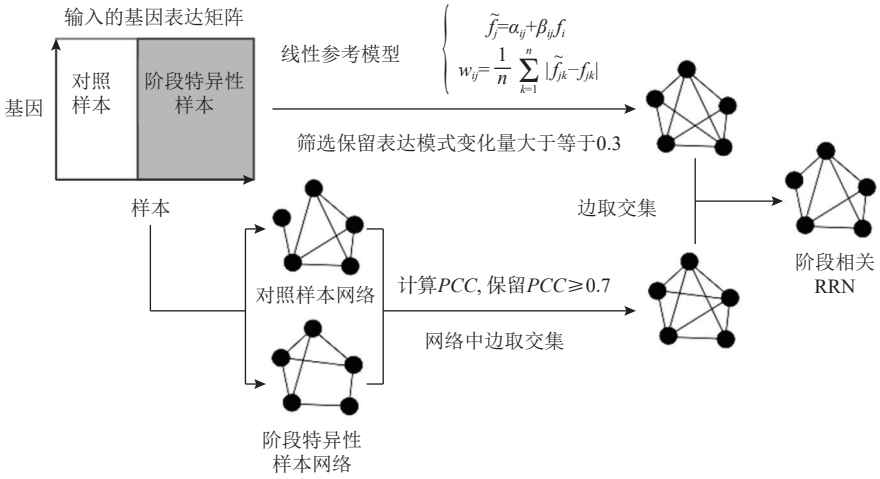


Fig. 1 Workflow of the construction of stage-specific RRN

图 1 阶段相关 RRN 的搭建流程

首先在正常样本组中构建成对基因的线性参考模型, 该线性参考模型如式(2)所示.

$$\tilde{f}_j = \alpha_{ij} + \beta_{ij} f_i,\ w_{ij} = \frac{1}{n}\sum_{k=1}^{n}|\tilde{f}_{jk} - f_{jk}|, \tag{2}$$

其中: i 和 j 分别代表 2 个不相同的基因; α 和 β 为线性模型中的系数, 即使用线性模型通过基因 i 的表达值预测基因 j 的表达值, 并且将使用除基因 j 外的其余基因的预测值与基因 j 表达的差值之和 w 作为表达模式的变化量. 为了提高边的显著性水平, 每个阶段中表达模式大于等于 0.3 的边被保留为 RRN 的差异边.

为了克服线性模型无法描述基因对之间可能存在的非线性关系以及由此导致的假阳性问题, RRN 在构建时还计算了正常组与各个阶段的基因对的皮尔逊相关系数(Pearson correlation coefficient, PCC). 根据 PCC 值筛选得到各自对应 PCC 网络, 本文设置的筛选条件为 $PCC\geqslant 0.7$.

将线性参考模型得到的有效边和 PCC 相关系数筛选得到的 PCC 网络中的边取交集, 得到 4 个阶段相关的 RRN, 其中, 第 1 阶段有 7 679 个节点和 72 022 条边; 第 2 阶段有 8 836 个节点和 107 565 条边; 第 3 阶段有 9 036 个节点和 121 612 条边; 第 4 阶段有 11 484 个节点和 145 647 条边.

2.3 阶段相关 RRN 的演化分析

在生物系统的发生发展过程中, 不同癌症阶段的生物网络动态变化着. 一些节点或边会在不同的

① www.ncbi.nlm.nih.gov/geo/

时刻出现或消失. 本文通过 RRN 中的节点和边的变化这 2 种不同的角度分析癌症的演化过程.

首先是基于 RRN 节点的演化分析. 癌症的发展总是伴随着基因表达及其相关的生物功能的变化. 在肿瘤发展过程中, 一些基因发生突变, 导致表型差异. 这些突变基因的代谢通路和生物功能在癌症发展过程中不断变化. 分析不同阶段差异表达的基因、生物功能和代谢路径变化, 有助于探索癌症的演化机制. 功能富集分析可以根据功能相似性将基因分为不同的类别. 通过对基因进行功能富集分析, 可以在癌症的各个阶段识别出与疾病最相关的生物分子, 也可以在癌症的各个阶段判定哪些生物功能是重要的. 同时, 通过分析疾病各个阶段之间生物功能的变化过程, 可以了解哪些生物功能参与了疾病的癌变过程, 这对研究癌症的各个阶段具有重要意义. 本文使用 DAVID(the database for annotation, visualization, and integrated discovery)数据库进行 GO(gene ontology)和 KEGG(Kyoto encyclopedia of genes and genomes)功能富集分析[19-20].

其次是基于 RRN 边的演化分析. 特定基因对之间的调控关系也随着癌症的发展而不断变化, 因此 RRN 发现的差异边在不同阶段具有不同的权重. 边的权重越大, 这对基因在正常组和癌症特异性阶段之间的表达模式差异就越大. 这些边可以作为癌症发展的重要研究对象, 对于研究癌症的动态演化具有重要意义. 本文将回归残差较大的边作为研究对象, 选择 RRN 中权重大于 0.6 的边进行进一步分析. 根据回归残差, 对边的讨论可以分为 2 种情况: 一种是由基因表达水平差异导致的边的差异; 另一种则是真正由于基因表达模式变化导致的边的差异. 每个阶段的差异基因可以通过差异表达分析获得, 而网络中的节点可以分为 3 种类型基因, 分别是非差异基因、差异上调基因、差异下调基因. 这样一来, 根据节点的类型, 就可以将边分为 6 类, 如图 2 所示, 其中: “Up”代表上调基因; “Down”代表下调基因; “notDEGs”代表非差异表达基因.

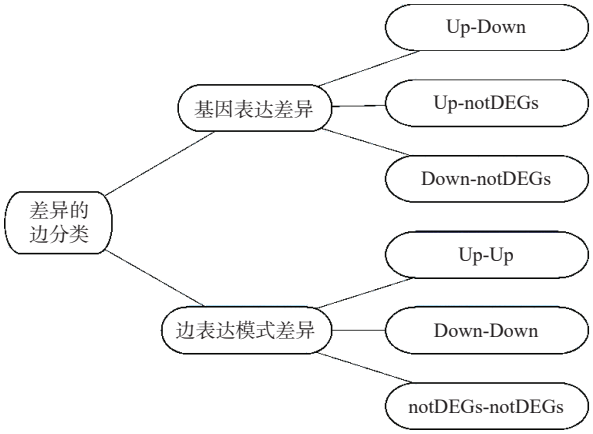


Fig. 2 Classification of difference edges

图 2 差异边的分类

2.4 网络标志物识别

通过网络获得的一组具有分类能力的节点或边称为网络生物标志物, 可用于区分癌症正常样本和肿瘤样本的二分类, 也可以对癌症的各个阶段进行多分类. 生物标志物的识别方法如图 3 所示.

首先是二分类网络标志物的识别. 为了分析癌症演化过程, 根据不同阶段网络之间的共性特征, 挖掘在肿瘤发生和发展中起重要作用的基因和差异边作为二分类的网络生物标志物. 图 4 阐述了 RRN_Node 方法通过生物功能筛选节点来识别网络标记的过程. 在基于 RRN 的节点演化分析中, 根据节点的 PageRank

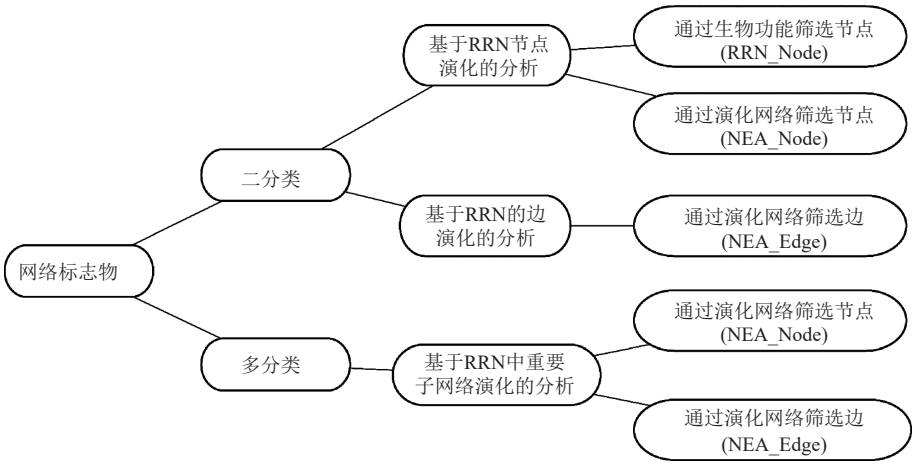


Fig. 3 Several methods for identifying network biomarkers proposed in this study

图 3 本文提出的用于识别网络生物标志物的几种方法

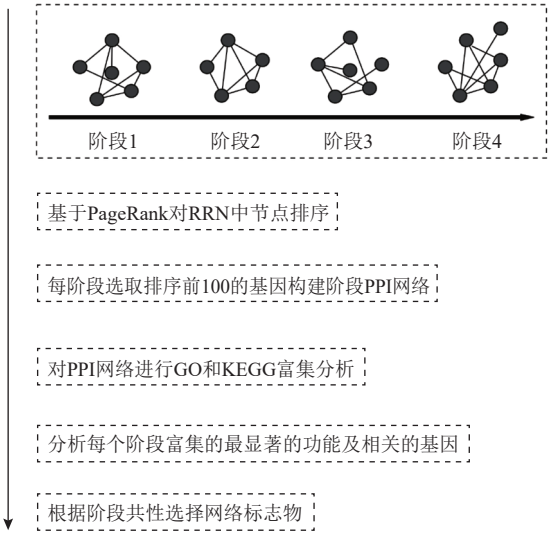


Fig. 4 Method flow of RRN_Node

图 4 RRN_Node 方法流程

中心性筛选出每个阶段的前 100 个基因, 并对每个阶段进行 GO 和 KEGG 功能富集分析. 再通过演化分析首先获得关键的癌症阶段演化子网络, 然后再提取每个子网络中阶段之间的公共节点和公共边作为网络生物标志物.

其次是多分类网络标志物的识别. 将癌症按照病理阶段划分对癌症的诊断和治疗都有重要的影响. 癌症在发生发展过程中, 其分子网络也在不断变化, 某些节点或边在特定阶段出现或消失, 而某些节点或边则存在于癌症发展的全过程中, 这就导致存在阶段共性和阶段特异性的点与边. 基于对癌症演化的分析, 本文提出了 2 种方法来识别阶段多分类的网络标记物, 分别是通过演化网络筛选节点(NEA_Node)和通过演化网络筛选边(NEA_Edge), 其中, 方法 NEA_Node 选择关键演化子网络中的所有节点作为阶段网络生物标志物, 方法 NEA_Edge 选择关键演化子网络中的所有边作为阶段网络生物标志物.

3 结 果

3.1 基于节点演化的结果分析

每个阶段构建好 RRN 后, 通过 PageRank 中心性选择每个阶段前 100 个基因进行后续分析, 并使用 DAVID 进行功能富集分析.

本文对 4 个阶段的关键基因进行了 GO 功能富集分析, 分别富集了 26, 18, 24, 17 个重要的 GO 通路. 发现 4 个阶段的基因共同富集了 4 种生物功能, 它们是 GO: 0044822 poly(A) RNA binding, GO: 0051056 regulation of small GTPase mediated signal transduction, GO: 0005829 cytosol 和 GO: 0005515 protein binding. 每个阶段特有的生物功能数量分别为 8, 2, 8, 3. 在这些 GO 通路中, GO: 0044822 poly(A) RNA binding 是最显著富集的一项. 研究发现, poly(A) RNA binding 参与调节多种转录机制, 其异常调节与癌症有关.

对 4 个阶段富集该功能的基因进行比较分析, 发现了 11 个常见基因, 其名称分别是: CCT3, HSPE1, ILF2, NME1, MRPL3, UBE2O, MRPL9, TOP1, ALDOA, TNS1, DAP3.

本文对 4 个阶段的关键基因进行了 KEGG 通路富集分析, 分别得到了 1, 1, 4, 4 个通路. 这些通路中只有 2 个通路相同. 在第 1, 2, 4 阶段的基因同时富集到 1 个共同的通路, 即: hsa01130: Biosynthesis of antibiotics, 且基因 ALDOA 和 NME1 在这 3 个阶段都参与了该通路.

对于 ALDOA 和 NME1 这 2 个基因, 不同阶段的表达水平如图 5 所示, 这 2 个基因在癌症各个阶段都显著上调, 对于判定患者是否患癌有着重要的作用.

3.2 基于边演化的结果分析

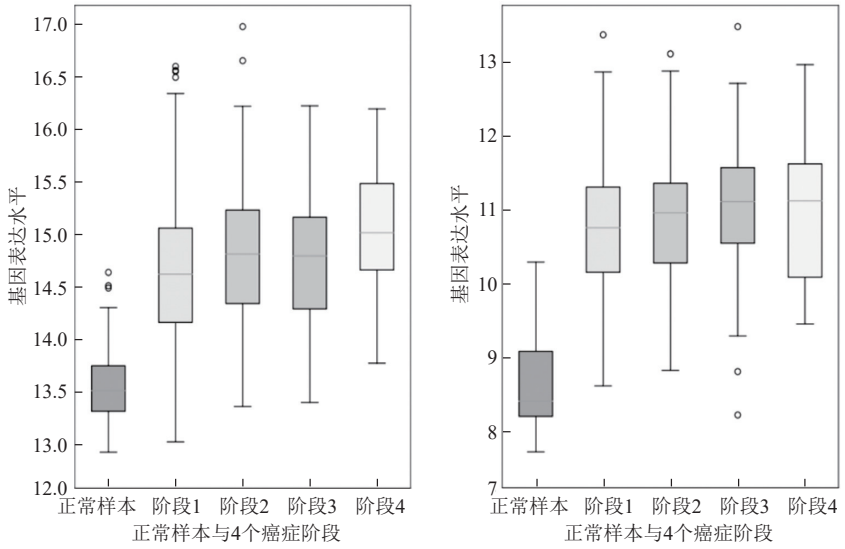
本文分别采用回归残差和差异共表达关系这 2 种方法对边进行比较和分析. 首先利用 *PCC* 的变化量构建各阶段的差异共表达网络, 过滤掉疾病组和对照组 *PCC* 变化差值小于等于 0.5 的边, 仅保留差异显著的边. 表 1 中列出 3 种边的表达模式的变化信息.

将上述 3 种类型的边进行合并, 得到 2 个较大的连通分支(分别称为子网络 1 和子网络 2), 如图 6 所示.

针对子网络 1, 其共包含 5 个基因节点, 分别为: GOLM1, DNAH2, HMGB3, C10orf81, WDR66. 这 5 个基因都是差异上调基因. 该子网络的阶段演化关系可以由图 7 表示, 此时坐标轴从左到右依次表示癌症发展的各个阶段. 随着癌症的发展, 可见子网络 1 在每个阶段都发生了变化.

针对子网络 2, 其共包含 10 个基因, 其名称分别为: DDX3Y, ZFY, KDM5D, RPS4Y1, TTTY15, UTY, USP9Y, CYorf15A, ARSD, STS. 差异表达分析表明, 尽管这些基因都不是差异基因, 但是它们表达模式的改变对于癌症的发生发展有着重要的影响, 还跟患者的性别有着重要的关系.

人类有 2 条性染色体, 分别是 X 染色体和 Y 染色体. 女性有 2 条 X 染色体, 而男性有 1 条 X 和 1 条 Y 染色体. 这 2 条染色体上有许多与肿瘤相关的基因位点. 研究表明, 癌症风险随着各种细胞中这 6 个关键



(a) ALDOA表达水平箱线图 (b) NME1表达水平箱线图

Fig. 5 Box plot of ALDOA and NME1 expression levels

图 5 ALDOA 和 NME1 表达水平箱线图

Table 1 Number of Different Sides of the Three Expression Modes at Different Stages

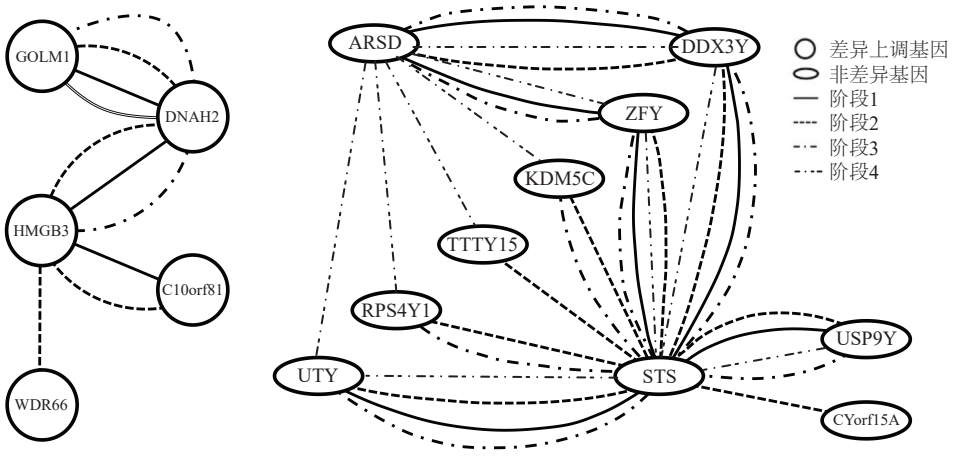
表 1 3 种表达模式在不同阶段的差异边数量

边的类型	阶段 1	阶段 2	阶段 3	阶段 4	总和
Up-Up	3	5	3	2	13
Down-Down	6	22	26	55	109
notDEGs-notDEGs	12	25	19	39	95

Y 染色体基因功能的丧失而增加，这些关键基因分别为：DDX3Y, KDM5D, RPS4Y1, UTY, ZFY, EIF1AY. 前 5 个基因都存在于子网络 2 中. 这 6 个基因都位于 Y 染色体的非同源重组区，并且在功能上与细胞周期相关. 例如，DDX3Y 负责 DNA 解旋，而 KDM5D 和 UTY 则与组蛋白去甲基化有关. 其中 4 个在 X 染色体上也有同源基因. 这 6 个 Y 染色体基因参与了细胞周期的转录调控. 它们与肿瘤发生密切相关，它们的缺失增加了患癌症的风险并促进了肿瘤的发展.

为了确认子网络 2 中的这 10 个基因是否与性别有关，本文从 TCGA 数据库下载的肺腺癌临床数据中的样本性别信息查看它们在男性和女性中的表达变化，并绘制了相应的箱线图，如图 8 所示.

从图 8 中可以看出，只有 2 个基因 ARSD 和 STS，在性别方面差异不大，其余 8 个基因的表达水平在男性和女性中存在显著差异，且它们中的大多数在女性中没有表达. 进一步分析可以发现，在该子网络中，ARSD 和 STS 是度数最大的 2 个基因，其余 8 个基因均与这 2 个基因相关. 因此可以推测 Y 染色体



(a) 子网络1 (b) 子网络2

Fig. 6 Two connected components

图 6 2 个连通分支

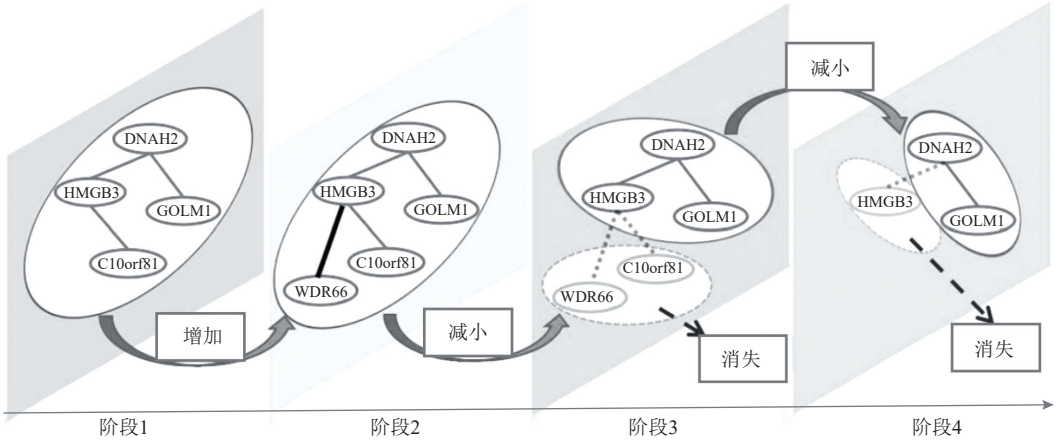


Fig. 7 Evolution process of subnet 1

图 7 子网络 1 的演化过程

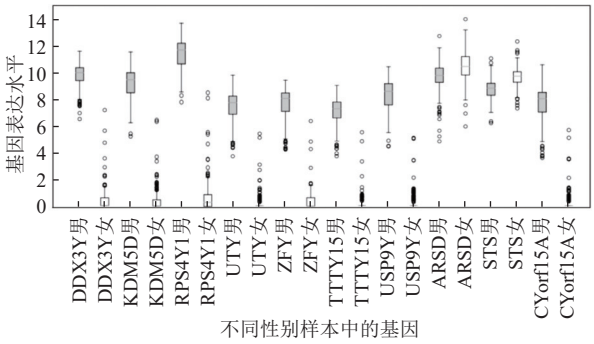


Fig. 8 Box plot of ten genes in different sex samples in subnet 2

图 8 子网 2 中 10 个基因在不同性别样本中的箱线图

上的这组基因可能与癌症的发生、发展和演化密切相关.

3.3 二分类网络标志物结果分析

节点演化分析表明，基因 ALDOA 和 NME1 是一组可行的二分类网络标志物(记作 RRN_Node 方法). 而边演化模型则分别选择子网络 1 和子网络 2 中的公共顶点与边，分别得到 8 个基因和 6 条边作为二分类网络标志物(分别记作 NEA_Node 和 NEA_Edge 方法).

使用肺腺癌数据集(GSE10072)对上述方法进行 5 折交叉验证，得到的结果与传统的基于差异表达分析的 DEA 方法进行对比，结果如表 2 所示.

从表 2 中可知，从预测的准确率、AUC 以及所使用的特征数量这 3 个方面进行比较可知，本文所提出的 RRN_Node 算法识别的 2 个标志物基因 ALDOA 和 NME1 可以作为网络标志物来区分正常样本和肺腺癌样本.

3.4 多分类网络标志物结果分析

对于多分类网络标志物分析，本文提出的 RRN_Node 算法识别出的基因作为一个二分类网络标志物，

Table 2 Four Binary Classification Methods and Corresponding Indicators

表 2 4 种二分类方法及相应的指标

方法	准确率	AUC	特征数量
DEA	0.960	0.960	**2**
RRN_Node（本文）	**0.962**	**0.990**	**2**
NEA_Node（本文）	0.887	0.950	8
NEA_Edge（本文）	0.944	**0.990**	6

注：黑体数值表示该列的最优结果.

无法完成多分类任务，因此在多分类网络标志物中选用方法 NEA_Node 和 NEA_Edge 进行多分类网络标志物结果分析. 因此，根据阶段的特殊性，选择子网络 1 和子网络 2 中的节点和边作为 NEA_Node 和 NEA_Edge 方法识别的网络生物标志物. NEA_Node 共识别了 14 个节点，NEA_Edge 共识别了 17 条边. 为了验证本文提出的 2 种方法识别多阶段网络生物标志物的有效性，还将它们与传统基于差异表达分析的 DEA 方法进行了比较.

使用另一个带有癌症分阶段信息的数据集 GSE41271 进行 5 折交叉验证，得到的结果与传统的 DEA 方法进行比较，结果如表 3 所示.

Table 3 Three Multi-Classification Methods and Corresponding Indicators

表 3 3 种多分类方法及相应的指标

方法	准确率	AUC	特征数量
DEA	0.44	0.76	38
RRN_Node（本文）	0.51	0.76	**14**
RRN_Edge（本文）	**0.59**	**0.77**	17

注：黑体数值表示该列的最优结果.

从表 3 中可知，本文提出的基于演化分析的网络标志物识别方法 NEA_Node 和 NEA_Edge，其准确率均高于 DEA，分别比 DEA 高 7 个百分点和 15 个百分点，并且使用的特征数量均比 DEA 少. 这里 NEA_Edge 方法的准确率最高，AUC 值也最高，比其他 2 个方法高 0.01. 因此，考虑到准确率和 AUC，NEA_Edge 方法识别的网络生物标志物对 GSE41271 数据集中的癌症分阶段具有最佳的多分类效果. 基于癌症演化分析提出的 NEA_Node 和 NEA_Edge 方法对比表明，使用子网络中的边作为特征的效果优于使用节点，准确率可以提高 8 个百分点. 综上所述，NEA_Edge 方法识别的差异边在 GSE41271 数据集对肺腺癌分阶段的多分类效果最佳，为 0.59.

4 结　　论

本研究主要面向医疗信息系统设计，探索了癌症的分阶段演化过程，设计了癌症阶段相关生物标志物识别算法. 差异表达分析存在假阳性，差异相关分析存在假阴性，本研究针对每个癌症阶段构建 RRN 以降低上述问题的概率，并识别出一些相关性没有差异但表达模式不同的边. 基于 RRN，本文从节点和边 2 个不同的角度分析了癌症的演化，得到了 2 个关键的演化子网络和一些与演化相关的关键基因. 基于以上对癌症演化过程的分析，从阶段网络之间的共性和特征出发，提出了二分类和多分类癌症的网络生物标志物方法. 此外，本文还做实验验证了本文提出的基于表达模式的网络分析方法是有效的.

在基于差异边的演化分析中，通过筛选得到了 2 个随阶段演化的关键子网络. 子网络 1 的演化过程图展示了癌症在不同阶段的动态演化过程. 子网络 2 的分析发现大部分节点是 Y 染色体基因，这可能与性别有关，这是本文的一个重要发现. 最后结合已有文献对本文发现的与癌症演化相关的关键基因进行分析验证. 此外，在随后对 2 个子网络的分析中获得了网络生物标志物.

本文多分类识别阶段网络标志物 NEA_Edge 方法的准确率仅为 0.59，其与癌症二分类（正常和疾病）相比并没有很优秀，因为研究发现在正常状态和疾病状态之间存在预疾病状态，在此期间会产生突变，所以更容易区分正常样本和患病样本. 现有的癌症分阶段通常基于癌变的组织学位置和医生对癌变恶化程度的判断. 也就是说，一个癌症样本可能会被不同的医生分成不同的阶段，这可能会影响癌症的阶段分类准确性. 同时，样本不平衡也是影响准确性的重要因素. 对现有数据库中具有病理分阶段信息的多种癌症进行统计分析发现，癌症分阶段与样本数量呈负相关. 癌症所处阶段越晚，对应的样本数越少，这也可能是较少影响 NEA_Edge 方法准确率的 1 个因素.

作者贡献声明：陈伯林提出了算法思路和实验方案；张劲雷、张建君、胡悠鹏、王腾、杨曼婷完成实验并撰写论文；卞军、尚学群提出指导意见并修改论文.

参　考　文　献

[1] Siegel R L, Miller K D, Jemal A. Cancer statistics, 2019[J]. CA: A Cancer Journal for Clinicians, 2019, 69(1): 7−34

[2] Miller K D, Nogueira L, Devasia T, et al. Cancer treatment and survivorship statistics, 2022[J]. CA: A Cancer Journal for Clinicians, 2022, 72(5): 409−436

[3] Allemani C, Weir H K, Carreira H, et al. Global surveillance of cancer survival 1995 —2009: Analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2)[J]. The Lancet, 2015, 385(9972): 977−1010

[4] Ramos J, Felty Q, Roy D. Integrated Chip-Seq and RNA-Seq data analysis coupled with bioinformatics approaches to investigate regulatory landscape of transcription modulators in breast cancer cells[G]//MIMB 2102: Molecular Toxicology Protocols. New York: Humana, 2020: 35−59

[5] Hong Mingye, Tao Shuang, Zhang Ling, et al. RNA sequencing: New technologies and applications in cancer research[J]. Journal of Hematology & Oncology, 2020, 13: 166

[6] Marioni J C, Mason C E, Mane S M, et al. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays[J]. Genome Research, 2008, 18(9): 1509−1517

[7] Guyon I, Elisseeff A. An introduction to variable and feature selection[J]. Journal of Machine Learning Research, 2003, 3: 1157−1182

[8] Huang Shujun, Cai Nanguang, Pacheco P P, et al. Applications of support vector machine (SVM) learning in cancer genomics[J]. Cancer Genomics & Proteomics, 2018, 15(1): 41−51

[9] Marcos-Zambrano L J, Karaduzovic-Hadziabdic K, Loncar Turukalo T, et al. Applications of machine learning in human microbiome studies: A review on feature selection, biomarker identification, disease prediction and treatment[J]. Frontiers in Microbiology, 2021, 12: 634511

[10] Rehman O, Zhuang Hanqi, Muhamed Ali A, et al. Validation of miRNAs as breast cancer biomarkers with a machine learning approach[J]. Cancers, 2019, 11(3): 431

[11] Chen Bolin, Yang Manting, Shang Xuequn. Identifying functional evolution processes of cancer according to regression residuals network[C]//Proc of the IEEE BIBM 2021. Piscataway, NJ: IEEE, 2021: 3713−3719

[12] Chakraborty D, Maulik U. Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning[J]. IEEE Journal of Translational Engineering in Health and Medicine, 2014, 2: 4300211
[13] Liu Yihui, Aickelin U, Feyereisl J, et al. Wavelet feature extraction and genetic algorithm for biomarker detection in colorectal cancer data[J]. Knowledge-Based Systems, 2013, 37: 502−514
[14] Merlo L M, Pepper J W, Reid B J, et al. Cancer as an evolutionary and ecological process[J]. Nature Reviews Cancer, 2006, 6(12): 924−935
[15] Arshad Z, McDonald J F. Changes in gene-gene interactions associated with cancer onset and progression are largely independent of changes in gene expression[J]. iScience, 2021, 24(12): 103522
[16] Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (TCGA): An immeasurable source of knowledge[J]. Contemporary Oncology, 2015, 19(1A): A68−A77
[17] Barrett T, Wilhite S E, Ledoux P, et al. NCBI GEO: Archive for functional genomics data sets—Update[J]. Nucleic Acids Research, 2013, 41(Database issue): D991−D995
[18] Durinck S, Spellman P T, Birney E, et al. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt[J]. Nature Protocols, 2009, 4(8): 1184−1191
[19] Dennis G Jr, Sherman B T, Hosack D A, et al. DAVID: Database for annotation, visualization, and integrated discovery[J]. Genome Biology, 2003, 4(5): P3
[20] Huang Dawei, Sherman B T, Lempicki R A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources[J]. Nature Protocols, 2009, 4(1): 44−57



Chen Bolin, born in 1984. PhD, associate professor. His main research interests include data mining, machine learning, and bioinformatics.
陈伯林, 1984 年生. 博士, 副教授. 主要研究方向为数据挖掘、机器学习、生物信息学.



Zhang Jinlei, born in 2001. Master candidate. His main research interests include data mining and bioinformatics.
张劲雷, 2001 年生. 硕士研究生. 主要研究方向为数据挖掘、生物信息学.



Zhang Jianjun, born in 2001. Master candidate. Her main research interests include data mining and bioinformatics.
张建君, 2001 年生. 硕士研究生. 主要研究方向为数据挖掘、生物信息学.



Hu Youpeng, born in 2003. Undergraduate. His main research interests include data mining and bioinformatics.
胡悠鹏, 2003 年生. 本科生. 主要研究方向为数据挖掘、生物信息学.



Wang Teng, born in 1998. Master. Her main research interests include data mining and bioinformatics.
王　腾, 1998 年生. 硕士. 主要研究方向为数据挖掘、生物信息学.



Yang Manting, born in 1997. Master. Her main research interests include data mining and bioinformatics.
杨曼婷, 1997 年生. 硕士. 主要研究方向为数据挖掘、生物信息学.



Bian Jun, born in 1977. Master, associate chief physician. His main research interests include comprehensive surgery and thoracic surgery.
卞　军, 1977 年生. 硕士, 副主任医师. 主要研究方向为综合外科、胸外科.



Shang Xuequn, born in 1973. PhD, professor. Her main research interests include data mining, machine learning, and bioinformatics.
尚学群, 1973 年生. 博士, 教授. 主要研究方向为数据挖掘、机器学习、生物信息学.