基于网络嵌入的癌症性状hub基因发现

初 妍[1],戚书豪[1],张 薇[1,2],王瀚麟[1],李 松[3]
(1. 哈尔滨工程大学计算机科学与技术学院,黑龙江省哈尔滨市 150001; 2. 黑龙江大学计算机科学与技术学院,黑龙江省哈尔滨市 150080; 3. 哈尔滨工程大学水声工程学院,黑龙江省哈尔滨市 150001)

摘 要: 研究影响癌症性状的hub基因时存在如下问题:仅关注强相关性基因进行基因信息处理,缺少对弱相关性基因和不同基因模块间共表达性的研究;仅采用度中心性判断hub基因进行分析基因网络,对蕴含数据挖掘不够全面。本文提出基因模块标签信息游走的图嵌入算法Gene2vec:选取合适软阈值,保留更多弱相关性的基因信息。联合不同种类但与性状高度正相关性的基因模块,构成基因模块共表达网络。针对传统加权基因共表达网络分析方法与图嵌入方法挖掘基因模块网络信息存在问题,利用标签参数与其他参数调节基因模块网络中的随机游走过程,分析游走生成的节点序列以挖掘基因网络的信息。实验表明,Gene2vec在hub基因的检出率上优于其他算法,得到的hub基因在癌症性状中的基因表达量高于常用生物学方法得到的hub基因。

关键词: hub基因; 数据挖掘; 信息游走; 网络嵌入; 加权基因共表达网络分析
中图分类号: TP3-0 **文献标识码:** A **文章编号:** 0372-2112(XXXX)XX-0001-08
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.xxxxxxxx *

Hub Gene Discovery of Cancer Traits based on Network Embedding

Chu Yan[1], Qi Shuhao[1], Zhang Wei[1,2], Wang Hanlin[1], Li Song[3]
(1. *College of Computer Science and Technology, Harbin Engineering University, Harbin, Heilongjiang Province* 150001, *China*;
2. *School of Computer Science and Technology, Heilongjiang University, Harbin, Heilongjiang Province* 150080, *China*;
3. *College of Underwater Acoustic Engineering, Harbin Engineering University, Harbin, Heilongjiang Province* 150001, *China*)

Abstract: The research on hub genes affecting cancer traits has such problems: only focusing on strong correlation genes for gene information processing, lack of the co-expression of weak correlation genes and different gene modules; only using degree centrality to judge hub genes to analyze gene network, not comprehensive enough for implicit data mining. We propose the graph embedding algorithm Gene2vec based on information walk with gene module label. We select the appropriate soft threshold to retain more weakly correlated gene information. The gene module co-expression network is formed by combining different kinds of gene modules with high positive correlation traits. Aiming to solve the problems of mining gene module network information by traditional weighted gene co-expression network analysis method and graph embedding method, we adjust the random walk process in the gene module network by label parameters and other parameters and analyze the node sequence generated random walk to mine the gene network information. Experiments show that Gene2vec is better than other algorithms in the hub gene´s detection rate, and the hub gene expression in cancer traits is higher than that of the hub gene obtained by common biological methods.

Key words: hub gene; data mining; information walk; network embedding; weighted gene co-expression network analysis

1 引言

深入挖掘并理解癌症发生的分子机制并找到影响癌症发生的hub基因,可为癌症治疗提供助力。生物学研究表明,细胞内分子间的相互作用会影响生物的生理机制和生理活动。现代图论知识为进一步探索生物复杂的生命活动提供了新工具和思想,不仅可以在系统水平上解释生物分子间的相互作用、弥补传统生物学研究方法的不足,还可以拓展研究的广度和深度,成为研究生命生物学特性的得力工具。加权基因共表达

收稿日期:2021-09-01;修回日期:2022-08-24;责任编辑:王天慧
基金项目:中央高校基本科研业务费研究项目(No. 3072022TS0601);黑龙江省属高校基本科研业务费基础研究项目(No. 2021-KYYWF-0043)

网络分析(WGCNA, Weighted Gene Co-expression Network Analysis)是现代图论在生物信息学领域的重要应用[1]。WGCNA以基因表达谱为基础构建基因共表达模块,通过发掘肿瘤各亚型样本中表达程度接近的基因,将这些基因分为不同的基因模块以构建基因共表达模块[2],同时还获得不同的基因模块与病体样本亚型性状间的关联信息。

但是WGCNA方法对癌症表型相关hub基因的识别存在局限性:在基因信息数据处理方面,原生物学方法仅研究强相关性基因,缺少对弱相关性基因和不同基因模块间共表达性的研究;分析处理生成的基因网络时,仅采用网络节点的度中心性定义基因网络中的hub基因,不能全面挖掘其中蕴含的数据,只利用了节点一阶相似性,而忽略了节点高阶相似性和节点间关联性。

基于随机游走的图嵌入方法受到Word2vec模型的启发,在网络中采用不同的随机游走模式生成网络节点序列,对节点序列进行嵌入操作表示为低维的向量[2]。Deepwalk算法采用完全随机游走的方法选取下一个访问节点,导致算法在进行游走时,对读取网络节点方向的不可控,使算法对网络的同质性与结构性的理解产生较大偏差;同时,由于游走的高度随机性,在对某个具体节点生成游走序列时更容易陷入极端情况(如在两个节点间反复游走),因此算法需要更多的迭代次数以弥补极端游走情况所产生的节点特征采样缺陷、导致算法在实际应用中效率下降;同时,算法并不能利用网络节点间的相关性权重参数指导游走,降低了算法在图嵌入中的实际效果[3]。Line算法只保留了网络节点间一阶与二阶的相似度,将注意力集中在网络的同质性上,导致节点的高阶信息利用不足[4],使得算法在节点数庞大、网络结构复杂的基因网络挖掘效能低;当节点邻居过少时,算法对该节点的特征提取不完整,致使嵌入效果不理想。Node2vec算法解决了原始算法无法兼顾挖掘网络同质性和结构对等性的问题,利用不同的in-out参数和return参数控制随机游走过程,加入网络节点间的相关性作为权重参数指导随机游走过程[5]。但Node2vec算法无法利用网络中节点的标签信息,在对节点数量庞大、节点标签信息复杂且结构复杂的网络进行多标签分类时效果不佳。

因此,本文尽量保留更多弱相关性的基因信息并深入分析关联的共表达性较强的基因模块,是准确理解触发癌症性状hub基因的关键。同时,为了使构建的加权基因共表达网络更加具有生物学意义,对作为输入的生物学数据进行预处理;根据WGCNA构建加权基因共表达网络,从基因表达相似性构建的基因共表达模块中筛选出构建图嵌入网络所需数据;采用图嵌入方法Gene2vec充分挖掘基因网络,以更准确高效地发现影响癌症性状的hub基因。论文技术路线如图1所示。

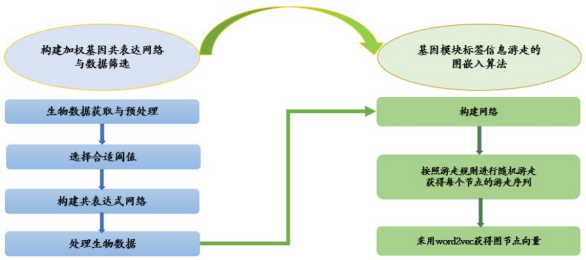


图1 技术路线图

2 构建加权基因共表达网络与数据筛选

2.1 生物数据来源与预处理

从美国国立生物信息中心(NCBI)的GEO数据库下载乳腺癌样本GSE48213的相关数据,数据中包括56个乳腺癌RNA_Seq样本、4个亚型性状和36953个基因在56个样本中的表达量。对基因数据进行预处理,将信息残缺的数据进行剔除。经过处理后的基因数据联合乳腺癌亚型性状,制作乳腺癌样本基因的分类信息与对应的亚型性状关联表,制作乳腺癌样本中所有基因的表达矩阵。

2.2 选择合适软阈值

在基因网络中,节点表示基因。若两个基因间的相关系数大于指定阈值,则以边连接表示两者间可能存在相互作用[6]。由于基因网络服从无尺度分布,在基因网络通常不是随意连接的,其中部分hub基因会连接数量极多的基因。因此,引用WGCNA中加权概念,hub基因在基因网络中拥有较大权重,这样可根据基因间相关性的强弱选择阈值。

在构建加权基因共表达网络时会选择较大的

阈值,在基因网络保持无尺度网络的同时,尽可能降低模块内的平均关联度,以排除弱相关性对计算节点连通度的影响、同时节省计算基因网络的时间。对于仅使用度中心性计算hub基因的生物学方法而言是十分有利的。但是,这样选取的软阈值虽然会强化强关系,代价却是让其他关系大幅度降低、基因间大量蕴含的中等强度关系被弱化与消除,对整体网络结构中研究基因关系不利。因此本文降低软阈值,让网络在保持无尺度网络的情况下,只消除极弱关系对基因网络构建的影响,其他关系保留。由图2和图3可知,当软阈值取到4时,网络80%的结构为无尺度分布,并且网络只将节点

平均连通度较高的约50个关联节点保留。这样选

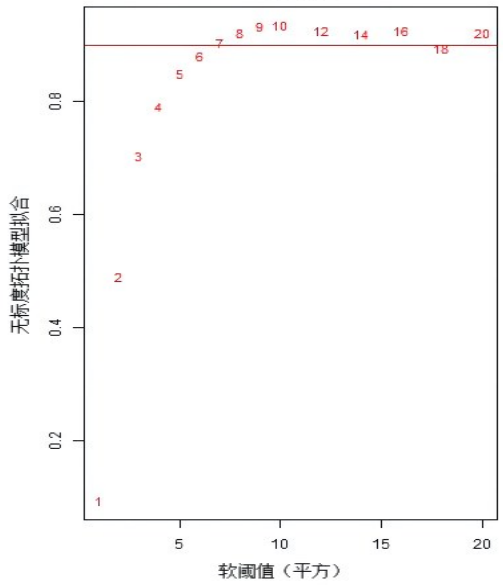


图2　无标度性分布图

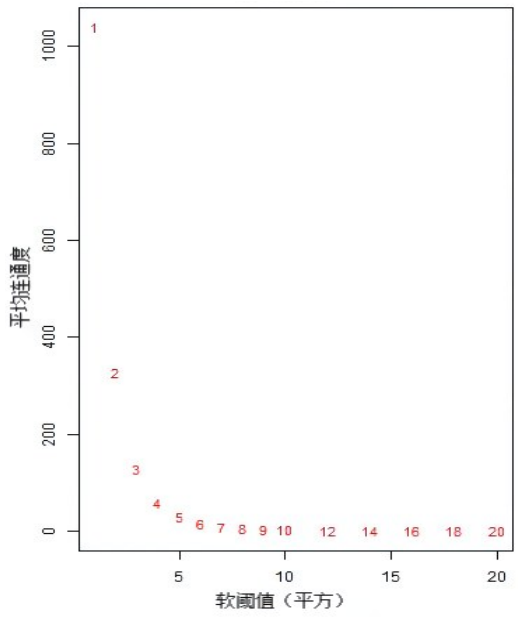


图3　平均连通性分布图

取的软阈值权衡了挖掘基因节点间较高关联性与保持无尺度分布两方面的需求。

2.3　构建共表达矩阵

构建共表达矩阵是实现对庞大复杂的基因数据进行剪枝与分类。计算基因间的关联相似度,得到基因间的相似性和相异性系数,建立基因间的系统聚类树[7]。为保证基因模块的实验效果,本文设置系统聚类树每个分支里至少含有30个基因。根据动态剪支法剔除关联性较弱的基因模块,依次计算每个基因模块的特征向量值,将特征向量距离较近的基因模块合并为具有高度相关性的基因模块。

2.4　生物数据处理

将基因模块与乳腺癌亚型性状间的相关系数用可视化的方法展示。“Basal”“Claudin-low”“Luminal”“Non-malignant”“unknown”这5类亚型性状所对应的不同模块的基因列表。同时,得到与每一种乳腺癌亚型性状强烈相关的基因模块,可以作为表达标志,模块里的基因可做下游分析,作为基因网络的数据来源。

由于Luminal亚型跟MEpurple基因模块相关性高达0.88(如图4所示),呈现极其显著的正相关性,所以认为Luminal亚型有较高的研究价值,设置为后续研究挖掘的对象。

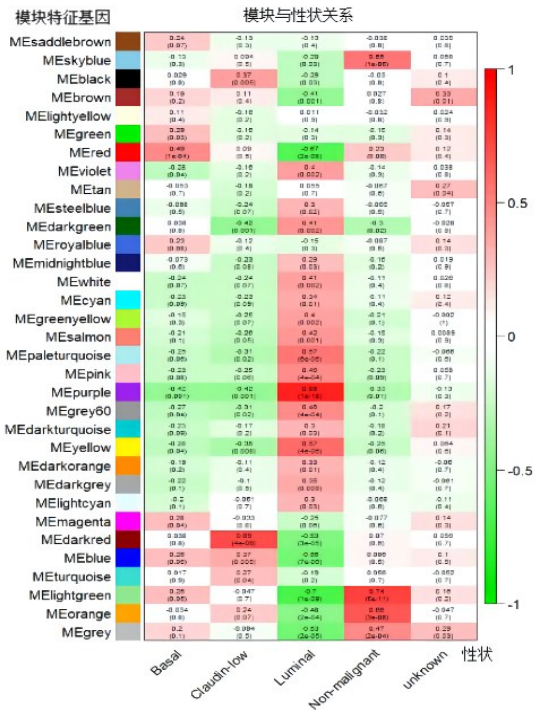


图4　基因模块与性状关系图

同时,为了弥补传统加权基因共表达方法对基因网络数据分析不够全面,研究基因模块间的共表达情况和基因间相互关系。将对Luminal亚型相关性较高的MEyellow基因模块(0.57)、MEpaleturquoise基因模块(0.57)和MEpink基因模块(0.48)联合组成共表达模块网络,作为构建图嵌入网络以及图嵌入算法分析实验所需要的基因网络数据。

3　基于基因模块标签信息游走的图嵌入算法

基因网络与普通网络的结构不同,基因节点数目庞大且为复杂网络结构,同时存在大量蕴含生物学意义、种类复杂的标签信息。若忽视节点的标签信息,随机游走过程在生物学意义上将失去控制。来自不同模

块的基因被认为具有高度的同质性，因此利用节点的标签信息，对随机游走过程具有指导作用会使图嵌入结果更准确、更具有生物学意义。本文提出基于基因模块标签信息游走的图嵌入算法Gene2vec。

3.1 算法原理

从宏观层面挖掘网络结构，设置控制读取网络同质性与结构对等性的in-out参数和return参数。同时，设置节点间相关性参数与标签参数来利用节点信息，在微观层面挖掘网络信息。

定义基因网络为G(V, E)，其中V代表网络的节点，E代表网络的边，跳转节点邻域是与起始节点和当前跳转节点同时成边的节点集合。设置in-out参数q和return参数p，权衡当前跳转节点跳转时深入游走离开起始节点邻域与返回起始节点的概率。节点间的相互关系设置游走权重参数weight，节点中的基因标签信息为label。同时，设置标签参数$k(0<k<1)$指导当前跳转节点对处于邻域内其他节点的随机游走概率。

当k=0时，不考虑当前跳转节点与邻域内节点的标签信息，在邻域内进行只考虑游走权重参数的随机游走。当$0<k<1$时，节点标签信息在邻域内与节点间相关性同样具有指导随机游走过程的作用，依据标签参数k的权重赋予游走和当前跳转节点具有相同标签信息邻居节点的概率。随着k的增加，跳转节点更倾向于游走到邻域内有相同标签的邻居节点。当k=1时，跳转节点只选择邻域内具有相同标签的邻居节点进行随机游走，如图5所示。

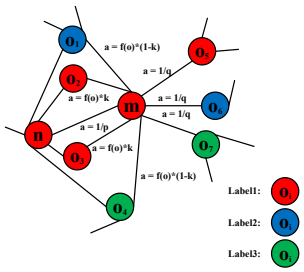


图5 基因网络局部图

游走起始节点为n，当前跳转节点为m，根据m邻居节点集合o内下一跳转节点j的距离与标签信息，定义m与j游走倾向系数a，如公式(1)所示。

$$a(o)=\begin{cases}\dfrac{1}{p} & d_{nj}=0\\ f(o)*k & d_{nj}=1\cap m_{label}=j_{label}\\ f(o)*(1-k) & d_{nj}=1\cap m_{label}\neq j_{label}\\ \dfrac{1}{q} & d_{nj}=2\end{cases}\tag{1}$$

其中d_{nj}为n与j点间的距离。定义符合$d_{nj}=1\cap m_{label}=j_{label}$的节点构成一个集合$U_1(o)$以及符合$d_{nj}=1\cap m_{label}\neq j_{label}$的节点构成另一个集合$U_2(o)$。$f(o)$为补偿函数，如公式(2)所示。

$$f(o)=\frac{\sum\limits_{i\in U_1(o)\cup U_2(o)} weight_i}{\sum\limits_{i\in U_1(o)} k*weight_i+\sum\limits_{i\in U_2(o)}(1-k)*weight_i}\tag{2}$$

对o内节点的游走倾向系数进行归一化后得出，当前节点m的跳转概率分布函数$g(o)$，如公式(3)所示。

$$g(o)=\frac{a_j\cdot weight_j}{\sum\limits_{i\in o} a_i\cdot weight_i}\ j\in o\tag{3}$$

根据跳转概率分布函数$g(o)$，指导每一步游走过程，进而获取游走序列。

3.2 算法描述

首先载入经过WGCNA预处理、筛选后的基因网络数据，读入每一个网络节点的基因模块标签信息及对应的邻居节点，节点间的相关性权重数据。同时，对所有节点进行预处理，从而节省实际游走过程中计算下一个游走节点概率的时间。然后依据图的边表得出节点对，边表的弧头节点作为当前游走节点，弧尾节点为当前节点的前驱，根据当前节点和其邻居节点的信息以及概率分布函数g计算该节点对下一步游走情况的概率序列P。使用Alias采样方法对该概率序列P进行二次处理。令每个节点作为初始节点按预处理结果进行随机游走，得到若干条游走路径的节点序列，通过反复迭代确保该算法对图中每个节点提取特征的完整性。最后调用Word2vec方法对随机游走节点序列进行训练，以节点向量的形式输出每个图节点。计算节点对每个邻居节点被游走的概率，并生成节点对Alias边字典，如算法1所示。

算法1需要接收图数据G、标签信息L以及用于指导游走的in-out参数q、return参数p和标签参数k等输入参数。对图中每条边进行如下操作：取一条边的弧头节点作为当前游走节点，弧尾节点为当前节点的前驱，根据当前节点和邻居节点的信息以及游走倾向系数$a(o)$生成未归一化的权重序列。对邻域内节点进行补偿，归一化生成游走概率序列P。为提高游走采样效率，使用Alias采样方法对该概率序列P二次处理，建立一个长度和概率序列P等长，宽度为2的二维数组Alias查找表。具体步骤如下：设整个概率序列长度为N，将整个概率分布按均值归一化，得到一个总面积为N的图形，保证每一列中只存在两个概率事件，将图形建立成一个1*N的长方形。该长方形储存为Alias查找表，Alias_Table[0][i]里存着第i列对应的游走节点矩形占总游走的概率，Alias_Table[1][i]里储存第i列不游走节点时节点的序号。以节点对为键，该节点对的Alias查找表即为建立字典Alias边。

算法1 generate_dictionary

输入:Graph: G(V, E), labels info: L (V, label_list), the specified value: p, q, k

输出:Alias_Edges

FOR each edge(V1, V2) in E

neighbors=G.neighbors(V2)　#读取邻居节点

Initial unnormalized_probs,normalized_probs and book with zeros, its length equals to neighbors　#初始化辅助数组

Initial $d1$_sum_weight and $d1$_sum_kweight with zeros　#初始化辅助变量

FOR each neighbor in neighbors　#历遍邻居节点生成未归一化权重序列

IF neighbor==V1　#游走节点为起始节点

append weight/p to unnormalized_probs

append 0 to book

ELSE IF　E(V1,neighbor) existed　#游走节点为邻域节点

d1_sum_weight+=weight

IF　L[V2]and L [neighbor] has common label　#邻域节点标签相同

append weight *k to unnormalized_probs

$d1$_sum_kweight+=weight*k

ELSE

append weight*k to unnormalized_probs　#邻域节点标签不同

$d1$_sum_kweight+=weight*(1-k)

append 1 to book

ELSE

append weight/q to unnormalized_probs　#游走节点为邻域其他节点

append 0 to book

FOR each mark in book　#对处在V2节点邻域内节点进行权重补偿

IF book[mark]==1

change the unnormalized_probs[mark] to unnormalized_probs[mark]* = ($d1$_sum_weight/$d1$_sum_kweight)

FOR each prob in unnormalized_probs　#生成归一化的游走概率序列

append prob/sum(unnormalized_probs) to normalized_probs

Alias_Edges [edge(V1,V2)]=alias_setup(normalized_probs)　#建立Alias边字典

随机游走采样过程如算法2所示。

依据Alias_Edges变量得到起始节点begin_node和跳转节点next_node建立节点对,以游走权重参数进行首次随机游走。通过Alias Sampling Method产生随机数:第一个为1~N之间的整数i,决定选择Alias查找表中哪一列进行采样;第2个为0~1之间的任意数,判断其与Alias查找表中Alias_Table[0][i]的大小。如果小于AliasTable[0][i],则采样i;如果大于AliasTable[0][i],则采样AliasTable[1][i]。在确定采样结果后,返回采样节点的下标索引,选取对应下标索引的邻居节点作为下一游走节点。一次游走完成,调整起始节点和跳转节点并重复上述步骤,直到游走的节点数量达到预定游走序列长度即停止,游走路径即为游走序列。

算法2 probability_wandering

输入:Graph: G(V, E) , the wandering starts node: begin_node, the length of wandering: walk_length, the dictionary of alias sampling method: Alias_Edges

输出:walk_list

walk_list append begin_node　#选择游走节点

wandering_node = begin_node

while the length of walk_list＜walk_length　#进行随机游走

neighbors_list= G.neighbors(wandering_node)

IF length of walk_list==1　#起始节点

index = Alias_Sample (Alias_ Edges [E(null, wandering_node)][0], Alias_ Edges [wandering_node][1])　#使用Alias Sampling Method采样

ELSE

prev_node = walk_list[2]　#其他节点

index = Alias_Sample (Alias_Edges[E (prev_node, wandering_node)][0], Alias_Edges[E(prev_node, wandering_node)][1])　#使用Alias Sampling Method采样

next_node = neighbors_list[index]

walk_list append next_node

wandering_node = next_node

表现最好的Node2vec算法无法处理多标签分类问题,本文提出改进的基因标签信息游走算法Gene2vec。在图节点中游走时,选取下一个游走节点采取策略和对应概率分布函数,在随机游走开始前建立Alias边字典,由随机游走生成节点序列,利用Word2vec对游走序列进行训练最终输出图节点向量。

4　实验验证与分析

4.1　实验数据集

数据集是WGCNA预处理后筛选组成的共表达模块图数据,该图数据是由852个节点,41085条边所组成的带权无向图。

4.2　分类效果对比分析

将Gene2vec的分类结果与Deepwalk、Line、Node2vec进行比较。其中,Deepwalk的参数设置为:游走长度为10,每个节点的游走重复次数为50,嵌入向量维度为128。Line的工作参数为:训练样本大小为1024,训练次数为50,每个节点使用2阶相似度。Node2vec的参数设置:游走长度为10,游走次数50,嵌入向量长度为128,P=11,Q=8。五个评价指标结果如表1所示。

表1 四个算法实验对比结果

算法名称	Micro	Macro	Samples	Weighed	Acc
Gene2vec	**0.97660**	**0.97948**	**0.97660**	**0.97636**	**0.97660**
Node2vec	0.93491	0.93923	0.93491	0.93455	0.93491
Line	0.82456	0.79573	0.82456	0.82526	0.82456
Deepwalk	0.80117	0.76712	0.80116	0.76722	0.80116

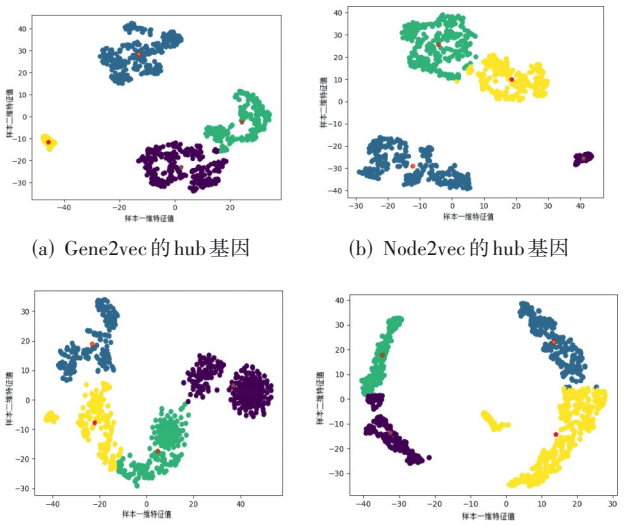
从对比结果可见，Gene2vec的实验效果优于Node2vec，且明显优于Deepwalk和Line。Deepwalk和Line算法嵌入效果相似，在0.8左右；Node2vec算法嵌入效果最好的在为0.93到0.94之间；而Gene2vec嵌入效果最好的在0.97到0.98之间。

4.3 hub基因的验证与评价

4.3.1 提取hub基因

确定Gene2vec和其他对比算法的参数后，进行后续实验以确定hub基因。选取与Luminal亚型显著相关的MEpurple基因模块(0.88)、MEyellow基因模块(0.57)、MEpaleturquoise基因模块(0.57)和MEpink基因模块(0.46)。因此，在共表达基因网络中找到这个四基因模块的聚类中心基因，即找到对Luminal亚型性状影响最大的四个hub基因。

为了验证提取的hub基因更具可信度，将Gene2vec与其他三种算法在聚类效果上进行对比。在共表达情况下的四种算法根据k-means算法得出四个基因模块聚类效果，如图6所示。



(a) Gene2vec的hub基因 (b) Node2vec的hub基因

(c) Line的hub基因 (d) Deepwalk的hub基因

图6 四种算法获得的hub基因

由于随机游走算法在读取网络节点及网络嵌入阶段存在不确定性，因此每次实验可能会得到不同的聚类中心。为了减少误差，增强其可信度和生物学意义，本文分别对Gene2vec和其他三种算法进行了1000次实验。检出率为基因在1000次实验中出现的概率，使用检出率说明所提取hub基因的可信度，其结果如表2所示。

表2 图嵌入算法四个聚类中心的实验结果

算法名称	Hub基因	MEpurple聚类中心	MEyellow聚类中心	Mepaleturquoise聚类中心	MEpink聚类中心
Gene2Vec	基因名称	CAV1	ACOT2	NRP1	ALAD
	检出率	**0.905**	**0.848**	**0.812**	**0.831**
Node2vec	基因名称	CAV1	ACOT2	NRP1	ALAD
	检出率	0.814	0.764	0.698	0.732
Line	基因名称	CAV1	ACOT2	NRP1	ALAD
	检出率	0.627	0.582	0.635	0.658
Deepwalk	基因名称	CAV1	ACOT2	KLF10	PTBP1
	检出率	0.578	0.442	0.537	0.512

由表2可见，在检出率方面Gene2vec效果最好，在0.81-0.91之间。而Node2vec效果略差且相对不稳定，检出率在0.7-0.83之间，而MEpaleturquoise聚类中心检出率已低于0.7，Line和Deepwalk则效果更差：Line检出率已经降低到0.58-0.71之间，而Deepwalk不仅检出率降至0.44-0.61之间，其MEpaleturquoise和MEpink聚类中心和其他三种算法的差异较大。正是由于Gene2vec在随机游走时考虑了基因标签与基因间相关性，才使其在基因共表达网络聚类时效果远远优于其他图嵌入算法。

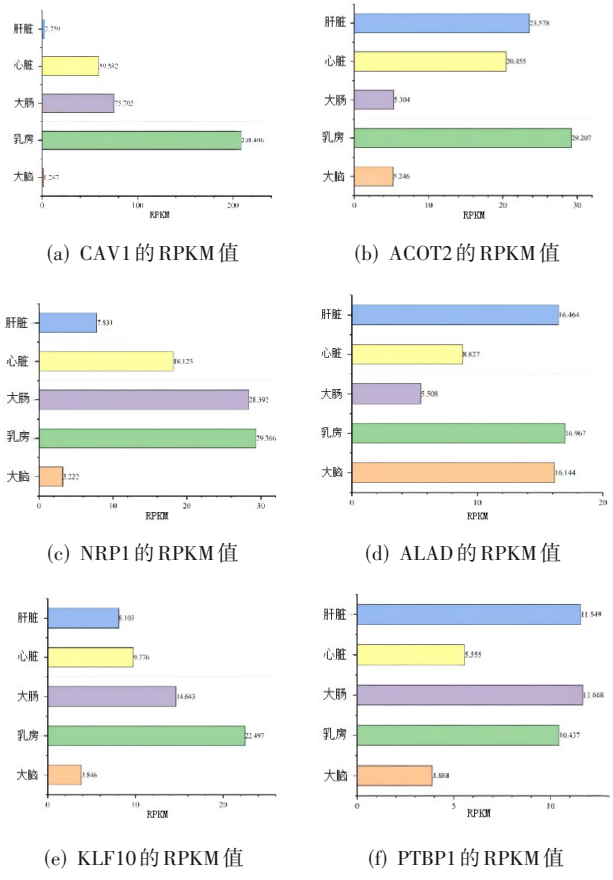
4.3.2 生物学方法与生物学数据验证hub基因

为保证结果的生物学意义，本文将上述结果与生物学方法所得到的hub基因进行对比。该方法寻找hub基因利用Cytoscape软件对基因网络进行分析并计算网络中节点的度中心性，度中心性最高的节点为hub基因。

生物学方法计算出的hub基因的基因编号结果为：MEpurple的hub基因为CAV1，MEyellow的hub基因为ACOT2，MEpaleturquoise的hub基因为KLF10，MEpink的hub基因为ENSG00000011304。

从实验结果来看,在乳腺癌样本GSE48213中生物学方法的hub基因与Gene2vec、Node2vec和Line有两项相同,与Deepwalk有四项相同。这是因为生物学方法与Deepwalk在计算时无法计算节点间相关性。在ME-paleturquoise和MEpink中,KLF10与PTBP1是度中心性最高的基因,却不是结合基因标签与基因节点间相关性综合的hub基因。

除了聚类中心检出率实验,还使用生物学数据证实了实验效果。在Gene2vec的实验中,本文认为与乳腺癌Luminal亚型性状影响相关性最高的四个基因模块hub基因的基因编号为CAV1、ACOT2、NRP1和ALAD。该结果与Node2vec和Line所得结果相同。生物学方法和Deepwalk所得到hub基因为CAV1、ACOT2、KLF10和PTBP1。五种方法共计获得6个基因,6个基因在不同器官中的RPKM值如图7所示。



(a) CAV1的RPKM值　(b) ACOT2的RPKM值

(c) NRP1的RPKM值　(d) ALAD的RPKM值

(e) KLF10的RPKM值　(f) PTBP1的RPKM值

图7　基因在不同器官中的RPKM值

可见6个基因都在乳腺癌样本中进行表达,其中CAV1、ACOT2、NRP1基因以及KLF10在乳腺中表达情况更为明显。Gene2vec的hub基因CAV1、ACOT2、NRP1和ALAD在乳腺癌样本中RPKM值均大于16,其中CAV1更是超过200,可以被认为在所有的乳腺癌样本中均有极高的表达量。同时,Gene2vec的hub基因NRP1和ALAD在乳腺癌样本中RPKM值比生物学方法和Deepwalk识别的hub基因KLF10、PTBP1的RPKM值多约7。Gene2vec发现的hub基因在病体样本中的基因表达量高于生物学方法,其生物学意义更高。

4.3.3　实验结果分析

相比于生物学方法和Deepwalk,Gene2vec考虑了节点间的相关性,将其作为权重参数指导游走,让具有强相关性的基因节点在嵌入向量方向即生物学意义上更接近,减少数量多、相关性极弱的基因节点对分析hub基因的影响。这导致Gene2vec的结果NRP1和ALAD虽然不是度中心性最高的基因,但RPKM值却比生物学方法和Deepwalk结果KLF10和PTBP1高。

相比于Node2vec和Line,Gene2vec在多标签且网络结构复杂的基因网络中发现hub基因效果更好。Line仅限于1阶与2阶相似度,让其无法挖掘有相关性但图上距离较远的节点,导致了Line检出率较低;Node2vec不能利用基因节点的标签信息,这让节点有较大概率游走到图上距离较近且相关性权重较高的异标签节点。当联合的基因模块数目增加,基因网络中标签种类复杂、基因节点数据足够庞大时,Node2vec的效果将进一步下降。因此Gene2vec对hub基因CAV1、ACOT2、NRP1和ALAD检出率均高于Node2vec,远高于Line。

5　结论

通过改进处理基因信息的方法,本文选定合适软阈值加强对基因模块中弱相关性基因数据的利用,让基因网络既维持无尺度网络结构同时也保留弱相关性基因。对不同基因模块间的共表达性进行研究,使用WGCNA处理基因表达谱得到基因共表达网络,从中选取与乳腺癌Luminal亚型性状高度正相关的四个基因模块,联合成基因模块网络作为后续数据集。

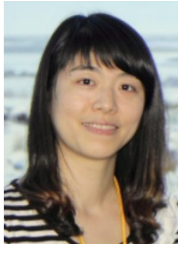
为更全面地挖掘基因模块网络中节点的高阶相似性和节点间关联性,提出基于基因模块标签信息游走的图嵌入算法Gene2vec。通过评价指标的设定和参数调节,获得Gene2vec算法在数据集上的最优效果,对比实验表明在乳腺癌数据集上Gene2vec算法的性能优于其他算法。为验证其生物学意义,通过NCBI查询实验结果在器官中的RPKM值大小,得到Gene2vec算法优于生物学在加权基因共表达网络分析中寻找hub基因的Degree方法和其他图嵌入算法。实验结果表明Gene2vec算法可更好地帮助发现影响癌症性状的hub基因。

参考文献

[1] 高攀. 关于乳腺癌的基因共表达网络分析及药物预测

[D]. 大连海事大学, 2018.
GAO P. Gene Co-expression Network Analysis and Drug Prediction for Breast Cancer. [D]. Dalian Marine University, 2018. (in Chinese)
[2] YIN X, WANG P, YANG T S, LI G, TENG X, HUANG W, YU H F. Identification of key modules and genes associated with breast cancer prognosis using WGCNA and ceRNA network analysis. [J]. Aging, 2020 Dec 9, 13(2): 2519-2538.
[3] 祁志卫, 王箭辉, 岳昆, 乔少杰, 李劲. 图嵌入方法与应用: 研究综述[J]. 电子学报, 2020, 48(04): 808-818.
QI Z W, WANG J H, YUE K, QIAO S J, LI J. Methods and applications of graph embedding: A survey. [J]. ACTA ELECTRONICA SINICA, 2020, 48(04): 808-818. (in Chinese)
[4] PALASH G, EMILIO F. Graph embedding techniques, applications, and performance: A survey. [J]. Knowledge-Based Systems, 2018 July 1, Volume 151: 78-94.
[5] MENG L Q, MASUDA N. Analysis of node2vec random walks on networks. [J]. Proceedings of The Royal Society A Mathematical Physical and Engineering Sciences, 476 (2243): 20200447.
[6] 谢勇军. 水稻非生物逆境差异表达基因分析及共表达网络构建[D]. 华中农业大学, 2018.
XIE Y J. Analysis of Differentially Expressed Genes and Construction of Gene Co-Expression Network in Rice under Abiotic Stresses. [D]. Huazhong Agricultural University, 2018. (in Chinese)
[7] YUAN Q H, ZHOU Q, REN J, WANG G, YIN C L, SHANG D, XIA S L. WGCNA identification of TLR7 as a novel diagnostic biomarker, progression and prognostic indicator, and immunotherapeutic target for stomach adenocarcinoma. [J]. Cancer med., 2021 Jun; 10(12): 4004-4016.

作者简介



初 妍 女,1979年出生,哈尔滨人. 哈尔滨工程大学计算机科学与技术学院副教授. 研究方向:机器学习、推荐系统. E-mail: chuyan@hrbeu.edu.cn

张 薇 (通讯作者) 女,1992年出生,黑龙江省牡丹江市人. 黑龙江大学计算机科学技术学院副教授. 研究方向:网络表征学习、基因预测. E-mail: zhangwei_jsj@hrbeu.edu.cn