(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107992720 B
(45)授权公告日 2021.08.03

(21)申请号 201711336559.1

(22)申请日 2017.12.14

(65)同一申请的已公布的文献号
申请公布号 CN 107992720 A

(43)申请公布日 2018.05.04

(73)专利权人 浙江工业大学
地址 310014 浙江省杭州市下城区朝晖六区潮王路18号

(72)发明人 陈晋音 郑海斌 王桢 宣琦 应时彦 李南

(74)专利代理机构 杭州斯可睿专利事务所有限公司 33241
代理人 王利强

(51)Int.Cl.
G16B 15/30(2019.01)
G16B 40/00(2019.01)

(56)对比文件
CN 103268431 A,2013.08.28
CN 106202984 A,2016.12.07
CN 103782301 A,2014.05.07
CN 105930688 A,2016.09.07
方木云.Hadoop下基于边聚类的重叠社区发现算法研究.《计算机技术与发展》.2015,58-62.

审查员 任洪潮

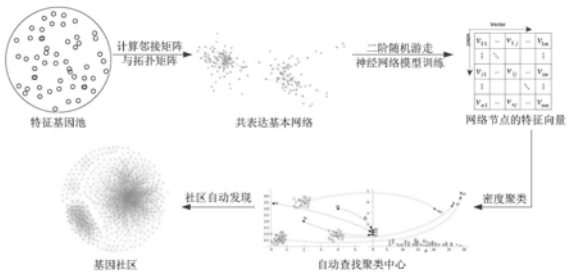
权利要求书4页 说明书8页 附图5页

(54)发明名称

基于共表达网络的癌症靶向标志物测绘方法

(57)摘要

一种基于共表达网络的癌症靶向标志物测绘方法,包括以下步骤:1)构建共表达基础网络,根据特征基因的基因表达数据计算邻接矩阵与拓扑矩阵;2)提取共表达基础网络的特征,即将拓扑网络的各个基因节点转换为特征向量作为网络的特征值;3)训练神经网络模型,根据游走序列,进行神经网络模型参数的训练;4)进行癌症靶向标志物测绘,根据基于密度峰的聚类中心自适应算法进行靶向基因社区的自动发现。本发明提供一种具有良好的普适性和精度,采用共表达基础网络构建和节点特征向量提取以及基因社区自动发现实现目标基因测绘的方法。



CN 107992720 B

1.一种基于共表达网络的癌症靶向标志物测绘方法,其特征在于:所述方法包括以下步骤:

1)构建共表达基础网络,根据特征基因的基因表达数据计算邻接矩阵与拓扑矩阵,过程如下:

1.1)将已经经过预处理与筛选的特征基因的基因表达数据作为构建共表达基础网络的源数据;

1.2)计算邻接矩阵,使用基因间表达水平的相关系数的幂指数加权值作为共表达的邻接矩阵,表示为Amatrix=[a_{ij}],计算公式如下:

$$a_{ij}\big|_{i,j}^{M_{pool3}}=|cor(gene_i,gene_j)|^{\beta} \quad (1)$$

式(1)中,M_{pool3}表示候选基因个数,即特征基因的数量;cor(·,·)表示基因i与基因j之间的相关系数;β表示加权幂指数,β的值根据无标度网络原则确定:即出现连接度为k的节点个数与该节点出现的概率p(k)反比于k的τ次方,且此时的相关系数需大于某一阈值thre;

1.3)计算拓扑矩阵,考虑基因与其它所有基因间的邻接关系,将邻接矩阵Amatrix转换为拓扑矩阵Ωmatrix=[ω_{ij}],计算公式如下:

$$\omega_{ij}=\frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}} \quad (2)$$

式(2)中,$l_{ij}=\sum_u a_{iu}a_{uj}$表示与基因i、j都存在连边的基因u的邻接系数乘积和,$k_i=\sum_u a_{iu}$表示与基因i单独连接的基因u的邻接系数和,$k_j=\sum_u a_{ju}$表示与基因j单独连接的基因u的邻接系数和;在与基因i和j之间无直接连接,且无任何其它的基因将这两个基因间接连接的情况下,取$\omega_{ij}=0$;

2)提取共表达基础网络的特征,即将拓扑网络的各个基因节点转换为特征向量作为网络的特征值,过程如下:

2.1)根据步骤1.3)中得到的拓扑矩阵确定共表达网络的基本结构;

2.2)针对网络中的每个节点进行二阶随机游走,节点总数表示为N,对于一个初始的头结点n_u,定义游走长度为$l_{randomWalk}$,C_i表示游走中的第i个节点,并以$C_0=n_u$开始,C_i的生成满足以下分布:

$$P(c_i=x\,|\,c_{i-1}=v)=\begin{cases}\frac{\pi_{vx}}{Z} & v与x间存在连边\\ 0 & 其他\end{cases} \quad (3)$$

式(3)中,x为下一步可能游走的节点,v为当前停留的节点,π_{vx}表示节点v与x间未标准化的转移概率,Z表示标准化常数;对于$C_{i-2}=t$,t表示已游走的上一个节点,定义$\pi_{vx}=\alpha_{pq(t,x)}$,其计算公式为:

$$\alpha_{pq(t,x)}=\begin{cases}1/p & if\ d_{tx}=0\\ 1 & if\ d_{tx}=1\\ 1/q & if\ d_{tx}=2\end{cases} \quad (4)$$

式(4)中,α表示带p、q参数的偏置量,d_{tx}表示节点t与x间的最短路径,且$d_{tx}\in\{0,1,2\}$;

为了避免相邻节点间的重复游走并确保游走的范围尽可能大，可将参数p设置为一个较大值，取p＞1将q设置为一个较小值，取q＜1；若π_{vx}相等，则随机选择一个节点进行游走；

2.3)根据步骤2.2)将网络中的每个节点作为头结点进行游走，得到N条长度为$l_{randomWalk}$的游走序列；

3)训练神经网络模型，根据步骤2.3)中得到的游走序列，进行神经网络模型参数的训练，过程如下：

3.1)将网络中的每一个基因节点表示成实数形式的分布式特征向量，同时使用游走序列中的节点的分布式特征向量来表示网络节点间的连接概率函数；

3.2)学习分布式特征向量与概率函数的参数，其中的训练集为步骤2.3)得到的游走序列；以一条游走序列为例，对序列中重复游走的节点仅保留第一个，处理后得到新的节点序列表示为$\{W_1,W_2,\ldots,W_T\}$，$W_T\in V$，其中V是节点集合，即大小为N的有限集合；训练目标是找到一个好的模型，使得该模型满足$f(W_t,\cdots,W_{t-n+1})=\hat{P}(W_t|W_1^{n-1})$；唯一的约束条件为：

$$\sum_{i=1}^{|V|} f(i,W_{t-1},\cdots,W_{t-n+1})=1, f>0 \qquad (5)$$

式(5)中，函数$f(W_{t-1},\cdots,W_{t-n+1})$可以分解为两个部分：第一部分为映射H(•)，其中H(i)表示节点集合中的每个节点的分布式特征向量，H实际上是一个由自由参数构成的|V|*m矩阵，其中m为自定义的向量维度；第二部分为函数g(•)，该函数将输入的节点特征向量$(H(W_{t-n+1}),\cdots,H(W_{t-1}))$映射为节点$W_t$前面n-1个节点的条件概率分布，即：

$$f(i,W_{t-1},\cdots,W_{t-n+1})=g(i,H(W_{t-1}),\cdots,H(W_{t-n+1})) \qquad (6)$$

当寻找得到满足带惩罚项的训练序列的对数似然率最大的θ，则训练结束，即：

$$L=\frac{1}{T}\sum_t \log f(W_t,W_{t-1},\cdots,W_{t-n+1})+R(\theta) \qquad (7)$$

神经网络的组成包括一个隐藏层，一个映射层，以及一个可选的直连层；最底层是单一的节点，表示成one-hot编码形式，即将节点表示成一个很长的向量，向量的分量只有一个1，其他全为0，1所对应的位置就是该节点在新的节点序列中的索引，向量长度为向量集的长度|V|；然后，每个one-hot编码的向量分别与投影矩阵H相乘，则原来长度为|V|的one-hot向量，经过线性变换以后，缩短为一个长度为m的向量，其中m是预先设置的特征个数，即向量的维度，向量维度一般为2个数量级；投影完成以后，将所有的特征向量按照顺序首尾相连，形成一个长度为m*(n-1)的向量，以节点向量作为隐藏层的输入，隐藏层的激活函数取为双曲正切函数tanh(•)；输出层接受隐藏层的输出作为输入，经过softmax(•)函数进行转换，得到最终的输出P为：

$$\hat{P}(W_t|W_{t-1},\cdots,W_{t-n+1})=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \qquad (8)$$

式(8)中，y＝b+Wx+Utanh(d+Kx)；双曲正切函数逐个应用于隐藏层的各个单元；当神经网络节点间没有直连的时候，W＝0，x是首尾相连的特征向量，即：

$$x=(H(W_{t-1}),H(W_{t-2}),\cdots,H(W_{t-n+1})) \qquad (9)$$

3.3)训练结束以后，矩阵H就是需要的节点特征向量，每一行代表该位置的节点的向量；

4)进行癌症靶向标志物测绘,根据基于密度峰的聚类中心自适应算法进行靶向基因社区的自动发现,过程如下:

4.1)将步骤3)得到的特征向量作为输入;

4.2)定义待聚类的向量矩阵 $H=\{x_i\}_{i=1}^{N}$,其中x_i表示矩阵的每一行,即步骤3.3)中该位置的节点的向量,相应指标集定义为I_H={1,2,...,N},任意两行向量x_i和x_j之间的欧式距离定义为:

$$d_{ij}=dist\left(x_i,x_j\right)=\sqrt{\sum_{i=1}^{m}\left(x_i-x_j\right)^2} \qquad (10)$$

式(10)中,m表示向量的维度;对于H中的任一向量x_i,定义其对应节点的局部密度ρ_i表示H中与x_i之间的距离小于d_c的向量个数,即:

$$\rho_i=\sum_{j\in I_S\setminus\{i\}}\chi\left(d_{ij}-d_c\right),\ \text{其中}\ \chi(x)=\begin{cases}1, & x<0\\ 0, & x\geq 0\end{cases} \qquad (11)$$

式(11)中,d_c>0表示截断距离,此处指定d_c为模长最大与模长最小的两个向量x_{max}与x_{min}之间欧式距离的2%,即:

$$d_c=0.02*dist(x_{max},x_{min}) \qquad (12)$$

设$\{q_i\}_{i=1}^{N}$表示$\{\rho_i\}_{i=1}^{N}$的一个降序排列下标序,即满足$\rho_{q1}\geq\rho_{q2}\geq\cdots\geq\rho_{qN}$,则可定义对应向量的距离$\delta_i$为:

$$\delta_{q_i}=\begin{cases}\min\limits_{j\leq i}\{d_{q_iq_j}\} & i\geq 2\\ \max\limits_{j\geq 2}\left\{\delta_{q_j}\right\} & i=1\end{cases} \qquad (13)$$

4.3)对于H中的每一行向量,计算其对应的密度值和距离值(ρ_i,δ_i),$i\in I_S$;根据得到的$\{\rho_i\}_{i=1}^{N}$和$\{\delta_i\}_{i=1}^{N}$绘制决策图(以ρ为横轴,δ为纵轴),自动确定密度值和距离值都较大的基因节点作为聚类中心,对剩余的基因节点按照距离最近原则进行归类得到不同的基因模块。

2.如权利要求1所述的基于共表达网络的癌症靶向标志物测绘方法,其特征在于:所述步骤3)中,将步骤2)中得到的游走序列作为神经网络训练的训练集,其处理过程为:对每条游走序列进行节点剔除,即对于一条序列中重复出现的节点仅保留第一个,完成后得到新的节点序列。

3.如权利要求1或2所述的基于共表达网络的癌症靶向标志物测绘方法,其特征在于:在所述步骤3)中,由于步骤2)中得到的游走序列不能保证将整个网络完全遍历,因此新的节点序列不包含所有节点,即得到的节点特征向量不完整;为了保证每条游走序列都是以网络中的不同节点作为初始头节点,对整个网络进行N次重复游走,重复游走策略为:对于每次训练得到的矩阵H,只选取第一条,即该游走序列起始节点的特征向量,以N条游走序列作为N个训练集,可得到N个矩阵,选取每个矩阵的第一条,即可得到N条特征向量,分别对应于N个初始头节点。

4.如权利要求1或2所述的基于共表达网络的癌症靶向标志物测绘方法,其特征在于:在所述步骤4)中,在完成社区发现后,计算总网络和各个子网络的特征值,如平均聚类系数、平均介数等,并结合临床数据验证具有较高网络特征的基因模块与研究人员关注的表

现型之间的相关性,完成对癌症靶向标志物的测绘。

基于共表达网络的癌症靶向标志物测绘方法

技术领域

[0001] 本发明属于生物信息技术领域,具体涉及一种癌症靶向目标基因测绘方法。

背景技术

[0002] 随着近年来科技以及医疗水平的不断进步,人们对抗疾病的能力与信心不断增强,但其中仍然存在许多缺陷与技术障碍。根据世界卫生组织的癌症报告估计,过去五年内中国癌症发病人数约占全球发病总人数的五分之一,而因罹患癌症死亡的人数则已超过全球癌症死亡总人数的四分之一。癌症死亡率居高不下,一个重要原因在于我国癌症发现较多处于中晚期。因此,人们在不断研究新的癌症治疗方法的同时,对于癌症靶向基因的检测,关键基因的提取以及相关癌症标志物的鉴定需要投入更多的科研精力。

[0003] 基因共表达网络分析作为一种挖掘和呈现基因在不同患病样本中表达形式的有效方法,可以搜索高度共表达的基因模块,而模块中包含的关键基因则可用于该模块的信息提炼。研究人员能够以此深入探讨基因模块或其关键基因与实际样本表型之间的关联关系。而在实际应用层面,基因共表达网络构建的基础——加权基因共表达网络构建(WGCNA)算法,已被用于复杂疾病的候选标记或药物靶点的鉴定和多项疾病的研究,如家族性混合型高脂血症、自闭症、阿尔兹海默症的关联基因、生物学通路和肿瘤治疗靶点的鉴定与测绘。在胶质母细胞瘤的研究过程中,研究者利用加权基因共表达网络成功挖掘得到与已知癌症相关模块高度重叠的基因共表达模块,而其中的一个关键基因被证实为该治疗的靶点基因。在骨密度的研究中,通过对不同骨密度妇女的单细胞核mRNA基因表达数据构建共表达网络,发现了与骨密度存在显著关联关系的模块,该结论也同样得到了相关遗传学研究结果的支持。

[0004] 综上所述,深入理解基因共表达网络与WGCNA算法的基本原理,熟练掌握该方法,在其基础上进行创新与改进,并将其运用到实际的临床科学研究中,具有极其重要的理论与实践意义。

发明内容

[0005] 针对共表达网络的复杂性问题,本发明通过计算基因间表达水平的相关系数构建基础网络,利用二阶随机游走与神经网络模型训练得到网络节点的特征向量,并设计聚类中心自适应算法进行靶向基因社区的自动发现。

[0006] 为了解决上述技术问题本发明提供如下的技术方案:

[0007] 一种高效的基于共表达网络的癌症靶向标志物测绘方法,所述方法包括以下步骤:

[0008] 1)构建共表达基础网络,根据特征基因的基因表达数据计算邻接矩阵与拓扑矩阵,过程如下:

[0009] 1.1)将已经经过预处理与筛选的特征基因的基因表达数据作为构建共表达基础网络的源数据;

[0010] 1.2)计算邻接矩阵,使用基因间表达水平的相关系数的幂指数加权值作为共表达的邻接矩阵,表示为Amatrix=[a_{ij}],计算公式如下:

[0011] $$a_{ij}\big|_{i,j}^{M_{pool3}} = |cor(gene_i, gene_j)|^{\beta} \quad (1)$$

[0012] 式(1)中,M_{pool3}表示候选基因个数,即特征基因的数量;cor(·,·) 表示基因i与基因j之间的相关系数;β表示加权幂指数,β的值根据无标度网络原则确定:即出现连接度为k的节点个数与该节点出现的概率p(k)反比于k的τ次方,且此时的相关系数需大于某一阈值thre (一般取thre=0.8);

[0013] 1.3)计算拓扑矩阵,考虑基因与其它所有基因间的邻接关系,将邻接矩阵Amatrix转换为拓扑矩阵Ωmatrix=[ω_{ij}],计算公式如下:

[0014] $$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}} \quad (2)$$

[0015] 式(2)中,$l_{ij}=\sum_u a_{iu}a_{uj}$表示与基因i、j都存在连边的基因u的邻接系数乘积和,$k_i=\sum_u a_{iu}$表示与基因i单独连接的基因u的邻接系数和,$k_j=\sum_u a_{ju}$表示与基因j单独连接的基因u的邻接系数和;在与基因i和j之间无直接连接,且无任何其它的基因将这两个基因间接连接的情况下,取$\omega_{ij}=0$;

[0016] 2)提取共表达基础网络的特征,即将拓扑网络的各个基因节点转换为特征向量作为网络的特征值,过程如下:

[0017] 2.1)根据步骤1.3)中得到的拓扑矩阵确定共表达网络的基本结构;

[0018] 2.2)针对网络中的每个节点进行二阶随机游走,节点总数表示为 N,对于一个初始的头结点n_u,定义游走长度为$l_{randomWalk}$,C_i表示游走中的第i个节点,并以$C_0=n_u$开始,C_i的生成满足以下分布:

[0019] $$P(c_i = x \mid c_{i-1} = v) = \begin{cases} \frac{\pi_{vx}}{Z} & v与x间存在连边 \\ 0 & 其他 \end{cases} \quad (3)$$

[0020] 式(3)中,x为下一步可能游走的节点,v为当前停留的节点,π_{vx}表示节点v与x间未标准化的转移概率,Z表示标准化常数;对于 $C_{i-2}=t$,t表示已游走的上一个节点,定义$\pi_{vx}=\alpha_{pq(t,x)}$,其计算公式为:

[0021] $$\alpha_{pq(t,x)} = \begin{cases} 1/p & if\ d_{tx}=0 \\ 1 & if\ d_{tx}=1 \\ 1/q & if\ d_{tx}=2 \end{cases} \quad (4)$$

[0022] 式(4)中,α表示带p、q参数的偏置量,d_{tx}表示节点t与x间的最短路径,且$d_{tx}\in\{0,1,2\}$;为了避免相邻节点间的重复游走并确保游走的范围尽可能大,可将参数p设置为一个较大值(一般取p>1),将q设置为一个较小值(一般取q<1);若π_{vx}相等,则随机选择一个节点进行游走;

[0023] 2.3)根据步骤2.2)将网络中的每个节点作为头结点进行游走,得到N条长度为$l_{randomWalk}$的游走序列;

[0024] 3)训练神经网络模型,根据步骤2.3)中得到的游走序列,进行神经网络模型参数的训练,过程如下:

[0025] 3.1)将网络中的每一个基因节点表示成实数形式的分布式特征向量,同时使用游走序列中的节点的分布式特征向量来表示网络节点间的连接概率函数;

[0026] 3.2)学习分布式特征向量与概率函数的参数,其中的训练数据集为步骤2.3)中得到的N条游走序列;以一条游走序列为例,对序列中重复游走的节点仅保留第一个,处理后得到新的节点序列表示为 $\{W_1,W_2,\ldots,W_T\}$,$W_T\in V$,其中V是节点集合,即大小为N的有限集合;训练目标是找到一个良好的模型,使得该模型满足 $f\left(W_t,\cdots,W_{t-n+1}\right)=\hat{P}\left(W_t\mid W_1^{n-1}\right)$;唯一的约束条件为:

[0027] $$\sum_{i=1}^{|V|} f\left(i,W_{t-1},\cdots,W_{t-n+1}\right)=1, f>0 \qquad (5)$$

[0028] 式(5)中,函数$f(W_{t-1},\cdots,W_{t-n+1})$可以分解为两个部分:第一部分为映射H(•),其中H(i)表示节点集合中的每个节点的分布式特征向量, H实际上是一个由自由参数构成的|V|*m矩阵,其中m为自定义的向量维度;第二部分为函数g(•),该函数将输入的节点特征向量 $(H(W_{t-n+1}),\cdots,H(W_{t-1}))$映射为节点$W_t$前面n-1个节点的条件概率分布,即:

[0029] $$f(i,W_{t-1},\cdots,W_{t-n+1})=g(i,H(W_{t-1}),\cdots,H(W_{t-n+1})) \qquad (6)$$

[0030] 当寻找得到满足带惩罚项的训练序列的对数似然率最大的θ,则训练结束,即:

[0031] $$L=\frac{1}{T}\sum_t \log f\left(W_t,W_{t-1},\cdots,W_{t-n+1}\right)+R(\theta) \qquad (7)$$

[0032] 神经网络的组成包括一个隐藏层,一个映射层,以及一个可选的直连层;最底层是单一的节点,表示成one-hot编码形式,即将节点表示成一个很长的向量,向量的分量只有一个1,其他全为0,1所对应的位置就是该节点在新的节点序列中的索引,向量长度为向量集的长度|V|;然后每个one-hot编码的向量分别与投影矩阵H相乘,则原来长度为|V|的one-hot向量,经过线性变换以后,缩短为一个长度为m 的向量,其中m是预先设置的特征个数,即向量的维度,向量维度一般为2个数量级;投影完成以后,将所有的特征向量按照顺序首尾相连,形成一个长度为m*(n-1)的向量,以节点向量作为隐藏层的输入,隐藏层的激活函数取为双曲正切函数tanh(•);输出层接受隐藏层的输出作为输入,经过softmax(•)函数进行转换,得到最终的输出P为:

[0033] $$\hat{P}\left(W_t\mid W_{t-1},\cdots,W_{t-n+1}\right)=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \qquad (8)$$

[0034] 式(8)中,y=b+Wx+Utanh(d+Kx);双曲正切函数逐个应用于隐藏层的各个单元;当神经网络节点间没有直连的时候,W=0;x是首尾相连的特征向量,即:

[0035] $$x=(H(W_{t-1}),H(W_{t-2}),\cdots,H(W_{t-n+1})) \qquad (9)$$

[0036] 3.3)训练结束以后,矩阵H就是需要的节点特征向量,每一行代表该位置的节点的向量;

[0037] 4)进行癌症靶向标志物测绘,根据基于密度峰的聚类中心自适应算法进行靶向基因社区的自动发现,过程如下:

[0038] 4.1)将步骤3)得到的特征向量作为输入;

[0039] 4.2)定义待聚类的向量矩阵 $H=\{x_i\}_{i=1}^N$,其中x_i表示矩阵的每一行,即步骤3.3)中该位置的节点的向量,相应指标集定义为$I_H=\{1,2,\ldots,N\}$,任意两行向量x_i和x_j之间的欧式

距离定义为：

[0040] $$d_{ij}=dist\left(x_i,x_j\right)=\sqrt{\sum_{i=1}^{m}\left(x_i-x_j\right)^2} \quad (10)$$

[0041] 式(10)中，m表示向量的维度；对于H中的任一向量x_i，定义其对应节点的局部密度ρ_i表示H中与x_i之间的距离小于d_c的向量个数，即：

[0042] $$\rho_i=\sum_{j\in I_S\setminus\{i\}}\chi\left(d_{ij}-d_c\right),\ \text{其中}\ \chi(x)=\begin{cases}1, & x<0\\ 0, & x\geq 0\end{cases} \quad (11)$$

[0043] 式(11)中，$d_c>0$表示截断距离，此处指定d_c为模长最大与模长最小的两个向量x_{max}与x_{min}之间欧式距离的2%，即：

[0044] $$d_c=0.02*dist(x_{max},x_{min}) \quad (12)$$

[0045] 设$\{q_i\}_{i=1}^{N}$表示$\{\rho_i\}_{i=1}^{N}$的一个降序排列下标序，即满足$\rho_{q1}\geq\rho_{q2}\geq\cdots\geq\rho_{qN}$，则可定义对应向量的距离$\delta_i$为：

[0046] $$\delta_{q_i}=\begin{cases}\min\limits_{j\leq i}\{d_{q_iq_j}\} & i\geq 2\\ \max\limits_{j\geq 2}\{\delta_{q_j}\} & i=1\end{cases} \quad (13)$$

[0047] 4.3)对于H中的每一行向量，计算其对应的密度值和距离值 (ρ_i,δ_i)，$i\in I_S$；根据得到的$\{\rho_i\}_{i=1}^{N}$和$\{\delta_i\}_{i=1}^{N}$绘制决策图(以ρ为横轴，δ为纵轴)，自动确定密度值和距离值都较大的基因节点作为聚类中心，对剩余的基因节点按照距离最近原则进行归类得到不同的基因模块。

[0048] 进一步，所述步骤3)中，将步骤2)中得到的游走序列作为神经网络训练的训练集，其处理过程为：对每条游走序列进行节点剔除，即对于一条序列中重复出现的节点仅保留第一个，完成后得到新的节点序列。

[0049] 更进一步，在所述步骤3)中，由于步骤2)中得到的游走序列不能保证将整个网络完全遍历，因此新的节点序列不包含所有节点，即得到的节点特征向量不完整；为了保证每条游走序列都是以网络中的不同节点作为初始头节点，对整个网络进行N次重复游走，重复游走策略为：对于每次训练得到的矩阵H，只选取第一条，即该游走序列起始节点的特征向量，以N条游走序列作为N个训练集，可得到N个矩阵，选取每个矩阵的第一条，即可得到N条特征向量，分别对应于 N个初始头节点。

[0050] 再进一步，在所述步骤4)中，在完成社区发现后，计算总网络和各个子网络的特征值，如平均聚类系数、平均介数等，并结合临床数据验证具有较高网络特征的基因模块与研究人员关注的表现型之间的相关性，完成对癌症靶向标志物的测绘。

[0051] 本发明的技术构思为：基于共表达网络的癌症靶向标志物测绘方法，通过对网络连续特征的学习，自动查找癌症靶向基因模块。首先构建共表达基础网络，根据特征基因的基因表达数据计算邻接矩阵与拓扑矩阵，并确定共表达网络的基本结构，再利用二阶随机游走与神经网络模型学习得到共表达基础网络中各个基因节点的特征向量。将基因节点的特征向量作为输入值，根据基于密度峰的聚类中心自适应算法进行靶向基因社区的自动发现。计算网络的相关特征值，同时结合临床数据验证具有较高网络特征的基因模块与研究人员关注的表现型之间的相关性，完成癌症靶向标志物的测绘。

[0052] 本发明的有益效果主要表现在:发现的共表达模块与动态剪切算法得到的共表达网络的吻合度十分高,说明本发明具有较好的生物信息可解释性。在真实数据上的实验结果表明,该算法具有良好的适用性和精度,能够大大缩小潜在癌症标志物的检测范围,为生物学领域的实验提供指导。

附图说明

[0053] 图1是本文算法整体框架示意图。

[0054] 图2是二阶随机游走的示意图。

[0055] 图3是神经网络模型训练示意图。

[0056] 图4是基于共表达网络的癌症靶向标志物测绘方法算法流程。

[0057] 图5(a)~5(c)是样本数据分布与目标基因测绘过程示意图,该数据集是对台湾地区患有肺癌的非吸烟女性的全基因组表达信息测量,包括配对的60个肿瘤样本和60个对照样本,每个样本具有54623维基因的表达。图5(a)为对已经经过预处理与筛选的特征基因表达数据的基因选择结果;图5(b)为用本发明方法对特征基因表达数据进行计算处理后得到的基因社区;表1是图5(b)中各个模块的网络特征的平均值,

[0058]

	Module1	Module2	Module3	Module4	Whole net
CC	0.8387	0.8156	0.5306	0.8403	0.8003
BN	9.2351	3.1875	1.6875	14.3333	8.7832
ACG	0.1752	0.2276	0.4866	0.1325	0.2108

[0059] 表1

[0060] 由此可以进一步看出,模块1和模块4具有较好的网络特征。图 5(c)为对拓扑网络进行特征向量提取,利用PCA进行向量的主成分提取后的可视化效果展示。图中横坐标和纵坐标分别是主成分的前两个维度,在共表达网络中的度值越大,图中对应点的半径越大。图中各个模块的颜色与图5(b)相对应。可以看出,不同模块的区分度较高,说明模块内部的功能连接更加紧密。同时,最重要的一点,在已有研究中发现的,可以作为癌症检测生物标志物的基因SEMA5A,同样在本实验中作为潜在分析目标被发现。在图5(c)中可以直观看出,表示为黄色点的基因SEMA5A处在模块1的中心处,十分接近聚类中心,这也从侧面反应了本发明方法的有效性。

具体实施方式

[0061] 下面结合附图对本发明作进一步描述。

[0062] 参照图1～图5(c),一种基于共表达网络的癌症靶向标志物测绘方法,包括以下步骤:

[0063] 1)构建共表达基础网络,根据特征基因的基因表达数据计算邻接矩阵与拓扑矩阵,过程如下:

[0064] 1.1)将已经经过预处理与筛选的特征基因的基因表达数据作为构建共表达基础网络的源数据;

[0065] 1.2)计算邻接矩阵,使用基因间表达水平的相关系数的幂指数加权值作为共表达的邻接矩阵,表示为Amatrix=$[a_{ij}]$,计算公式如下:

[0066] $$a_{ij}\left|_{i,j}^{M_{pool3}}=\left|cor(gene_i,gene_j)\right|^{\beta}\right. \quad (1)$$

[0067] 式(1)中,M_{pool3}表示候选基因个数,即特征基因的数量;cor(·,·) 表示基因i与基因j之间的相关系数;β表示加权幂指数,β的值根据无标度网络原则确定:即出现连接度为k的节点个数与该节点出现的概率p(k)反比于k的τ次方,且此时的相关系数需大于某一阈值thre (一般取thre=0.8);

[0068] 1.3)计算拓扑矩阵,考虑基因与其它所有基因间的邻接关系,将邻接矩阵Amatrix转换为拓扑矩阵Ωmatrix=[ω_{ij}],计算公式如下:

[0069] $$\omega_{ij}=\frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}} \quad (2)$$

[0070] 式(2)中,$l_{ij}=\sum_u a_{iu}a_{uj}$表示与基因i、j都存在连边的基因u的邻接系数乘积和,$k_i=\sum_u a_{iu}$表示与基因i单独连接的基因u的邻接系数和,$k_j=\sum_u a_{ju}$表示与基因j单独连接的基因u的邻接系数和;在与基因i和j之间无直接连接,且无任何其它的基因将这两个基因间接连接的情况下,取$\omega_{ij}=0$;

[0071] 2)提取共表达基础网络的特征,即将拓扑网络的各个基因节点转换为特征向量作为网络的特征值,过程如下:

[0072] 2.1)根据步骤1.3)中得到的拓扑矩阵确定共表达网络的基本结构;

[0073] 2.2)针对网络中的每个节点进行二阶随机游走,节点总数表示为 N,对于一个初始的头结点n_u,定义游走长度为$l_{randomWalk}$,C_i表示游走中的第i个节点,并以$C_0=n_u$开始,C_i的生成满足以下分布:

[0074] $$P(c_i=x\,|\,c_{i-1}=v)=\begin{cases}\frac{\pi_{vx}}{Z} & v与x间存在连边\\ 0 & 其他\end{cases} \quad (3)$$

[0075] 式(3)中,x为下一步可能游走的节点,v为当前停留的节点,π_{vx}表示节点v与x间未标准化的转移概率,Z表示标准化常数;对于 $C_{i-2}=t$,t表示已游走的上一个节点,定义$\pi_{vx}=\alpha_{pq(t,x)}$,其计算公式为:

[0076] $$\alpha_{pq(t,x)}=\begin{cases}1/p & if\ d_{tx}=0\\ 1 & if\ d_{tx}=1\\ 1/q & if\ d_{tx}=2\end{cases} \quad (4)$$

[0077] 式(4)中,α表示带p、q参数的偏置量,d_{tx}表示节点t与x间的最短路径,且$d_{tx}\in\{0,1,2\}$;为了避免相邻节点间的重复游走并确保游走的范围尽可能大,可将参数p设置为一个较大值(一般取p>1),将q设置为一个较小值(一般取q<1);若π_{vx}相等,则随机选择一个节点进行游走;

[0078] 2.3)根据步骤2.2)将网络中的每个节点作为头结点进行游走,得到N条长度为$l_{randomWalk}$的游走序列。

[0079] 3)训练神经网络模型,根据步骤2.3)中得到的游走序列,进行神经网络模型参数的训练,过程如下:

[0080] 3.1)将网络中的每一个基因节点表示成实数形式的分布式特征向量,同时使用游

走序列中的节点的分布式特征向量来表示网络节点间的连接概率函数；

[0081] 3.2)学习分布式特征向量与概率函数的参数，其中的训练集为步骤2.3)得到的游走序列；以一条游走序列为例，对序列中重复游走的节点仅保留第一个，处理后得到新的节点序列表示为$\{W_1,W_2,...,W_T\}$，$W_T \in V$，其中V是节点集合，即大小为N的有限集合；训练目标是找到一个好的模型，使得该模型满足$f(W_t,\cdots,W_{t-n+1})=\hat{P}(W_t|W_1^{n-1})$；唯一的约束条件为：

[0082] $$\sum_{i=1}^{|V|} f(i,W_{t-1},\cdots,W_{t-n+1})=1, f>0 \qquad (5)$$

[0083] 式(5)中，函数$f(W_{t-1},\cdots,W_{t-n+1})$可以分解为两个部分：第一部分为映射H(•)，其中H(i)表示节点集合中的每个节点的分布式特征向量，H实际上是一个由自由参数构成的|V|*m矩阵，其中m为自定义的向量维度；第二部分为函数g(•)，该函数将输入的节点特征向量$(H(W_{t-n+1}),\cdots,H(W_{t-1}))$映射为节点$W_t$前面n-1个节点的条件概率分布，即：

[0084] $f(i,W_{t-1},\cdots,W_{t-n+1})=g(i,H(W_{t-1}),\cdots,H(W_{t-n+1}))$ (6)

[0085] 当寻找得到满足带惩罚项的训练序列的对数似然率最大的θ，则训练结束，即：

[0086] $$L=\frac{1}{T}\sum_t \log f(W_t,W_{t-1},\cdots,W_{t-n+1})+R(\theta) \qquad (7)$$

[0087] 神经网络的组成包括一个隐藏层，一个映射层，以及一个可选的直连层；最底层是单一的节点，表示成one-hot编码形式，即将节点表示成一个很长的向量，向量的分量只有一个1，其他全为0，1所对应的位置就是该节点在新的节点序列中的索引，向量长度为向量集的长度|V|。然后，每个one-hot编码的向量分别与投影矩阵H相乘，则原来长度为|V|的one-hot向量，经过线性变换以后，缩短为一个长度为m的向量，其中m是预先设置的特征个数，即向量的维度，向量维度一般为2个数量级；投影完成以后，将所有的特征向量按照顺序首尾相连，形成一个长度为m*(n-1)的向量，以节点向量作为隐藏层的输入，隐藏层的激活函数取为双曲正切函数tanh(•)；输出层接受隐藏层的输出作为输入，经过softmax(•)函数进行转换，得到最终的输出P 为：

[0088] $$\hat{P}(W_t|W_{t-1},\cdots,W_{t-n+1})=\frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \qquad (8)$$

[0089] 式(8)中，y＝b+Wx+Utanh(d+Kx)；双曲正切函数逐个应用于隐藏层的各个单元；当神经网络节点间没有直连的时候，W＝0，x是首尾相连的特征向量，即：

[0090] $x=(H(W_{t-1}),H(W_{t-2}),\cdots,H(W_{t-n+1}))$ (9)

[0091] 3.3)训练结束以后，矩阵H就是需要的节点特征向量，每一行代表该位置的节点的向量；

[0092] 其中，由于步骤2)得到的游走序列不能保证将整个网络完全遍历，因此新的节点序列不包含所有节点，即得到的节点特征向量不完整；为了保证每条游走序列都是以网络中的不同节点作为初始头节点，对整个网络进行N次重复游走，重复游走策略为：对于每次训练得到的矩阵H，只选取第一条，即该游走序列起始节点的特征向量，以N条游走序列作为N个训练集，可得到N个矩阵，选取每个矩阵的第一条，即可得到N条特征向量，分别对应于N个初始头节点。

[0093] 4)进行癌症靶向标志物测绘，根据基于密度峰的聚类中心自适应算法进行靶向基

因社区的自动发现，过程如下：

[0094] 4.1)将步骤3)得到的特征向量作为输入；

[0095] 4.2)定义待聚类的向量矩阵$H=\{x_i\}_{i=1}^{N}$，其中x_i表示矩阵的每一行，即步骤3.3)中该位置的节点的向量，相应指标集定义为$I_H=\{1,2,...,N\}$，任意两行向量x_i和x_j之间的欧式距离定义为：

[0096] $$d_{ij}=dist(x_i,x_j)=\sqrt{\sum_{i=1}^{m}(x_i-x_j)^2} \quad (10)$$

[0097] 式(10)中，m表示向量的维度；对于H中的任一向量x_i，定义其对应节点的局部密度ρ_i表示H中与x_i之间的距离小于d_c的向量个数，即：

[0098] $$\rho_i=\sum_{j\in I_S\setminus\{i\}}\chi(d_{ij}-d_c),\ \text{其中}\ \chi(x)=\begin{cases}1, & x<0\\ 0, & x\geq 0\end{cases} \quad (11)$$

[0099] 式(11)中，$d_c>0$表示截断距离，此处指定d_c为模长最大与模长最小的两个向量x_{max}与x_{min}之间欧式距离的2%，即：

[0100] $d_c=0.02*dist(x_{max},x_{min})$ (12)

[0101] 设$\{q_i\}_{i=1}^{N}$表示$\{\rho_i\}_{i=1}^{N}$的一个降序排列下标序，即满足$\rho_{q1}\geq\rho_{q2}\geq\cdots\geq\rho_{qN}$，则可定义对应向量的距离$\delta_i$为：

[0102] $$\delta_{q_i}=\begin{cases}\min\limits_{j\leq i}\{d_{q_iq_j}\} & i\geq 2\\ \max\limits_{j\geq 2}\{\delta_{q_j}\} & i=1\end{cases} \quad (13)$$

[0103] 4.3)对于H中的每一行向量，计算其对应的密度值和距离值 (ρ_i,δ_i)，$i\in I_S$。根据得到的$\{\rho_i\}_{i=1}^{N}$和$\{\delta_i\}_{i=1}^{N}$绘制决策图(以ρ为横轴，δ为纵轴)，自动确定密度值和距离值都较大的基因节点作为聚类中心，对剩余的基因节点按照距离最近原则进行归类得到不同的基因模块。

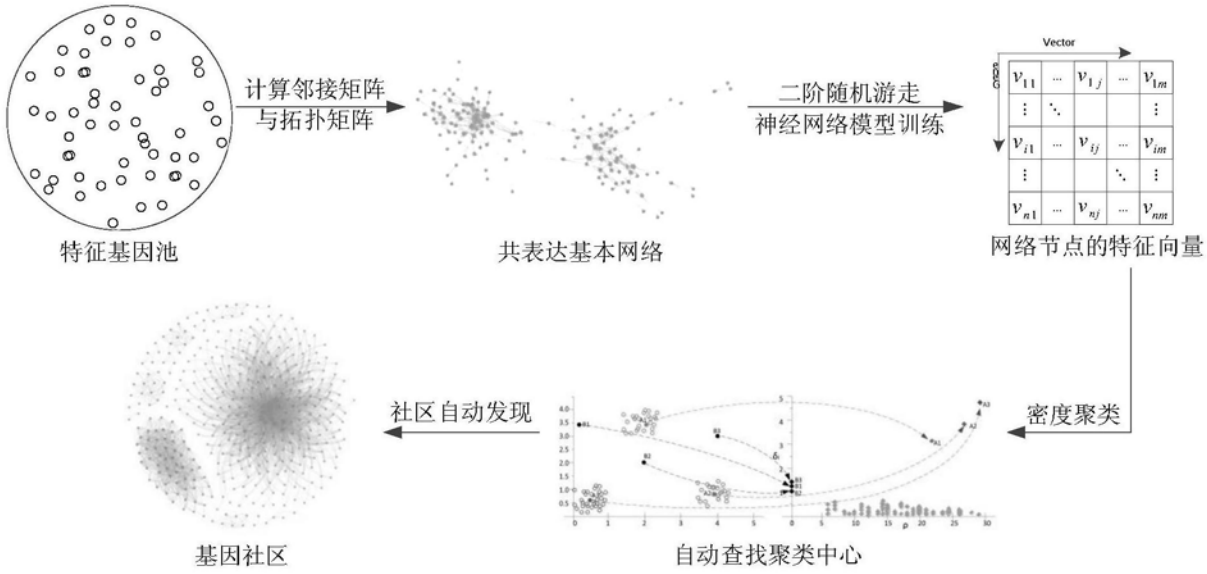

特征基因池
计算邻接矩阵
与拓扑矩阵
共表达基本网络
二阶随机游走
神经网络模型训练
Vector
网络节点的特征向量
密度聚类
自动查找聚类中心
社区自动发现
基因社区

图1

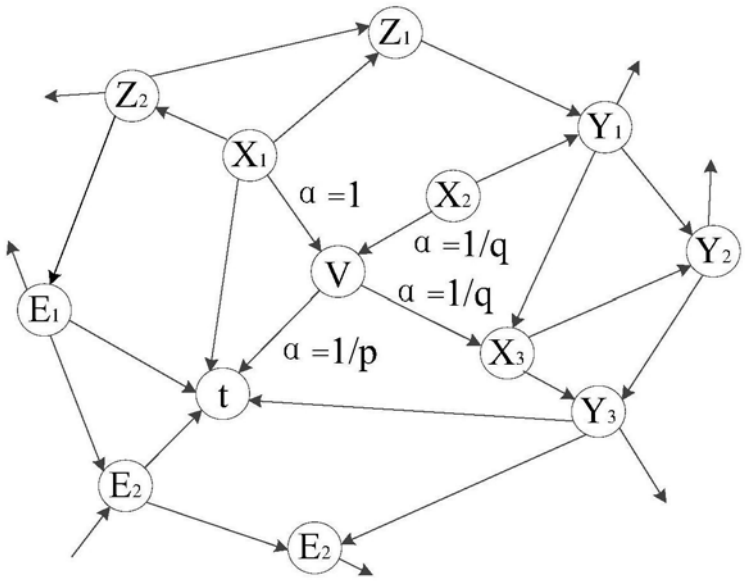

Z1
Z2
X1
Y1
α =1
X2
α =1/q
V
α =1/q
Y2
E1
α =1/p
X3
t
Y3
E2
E2

图2

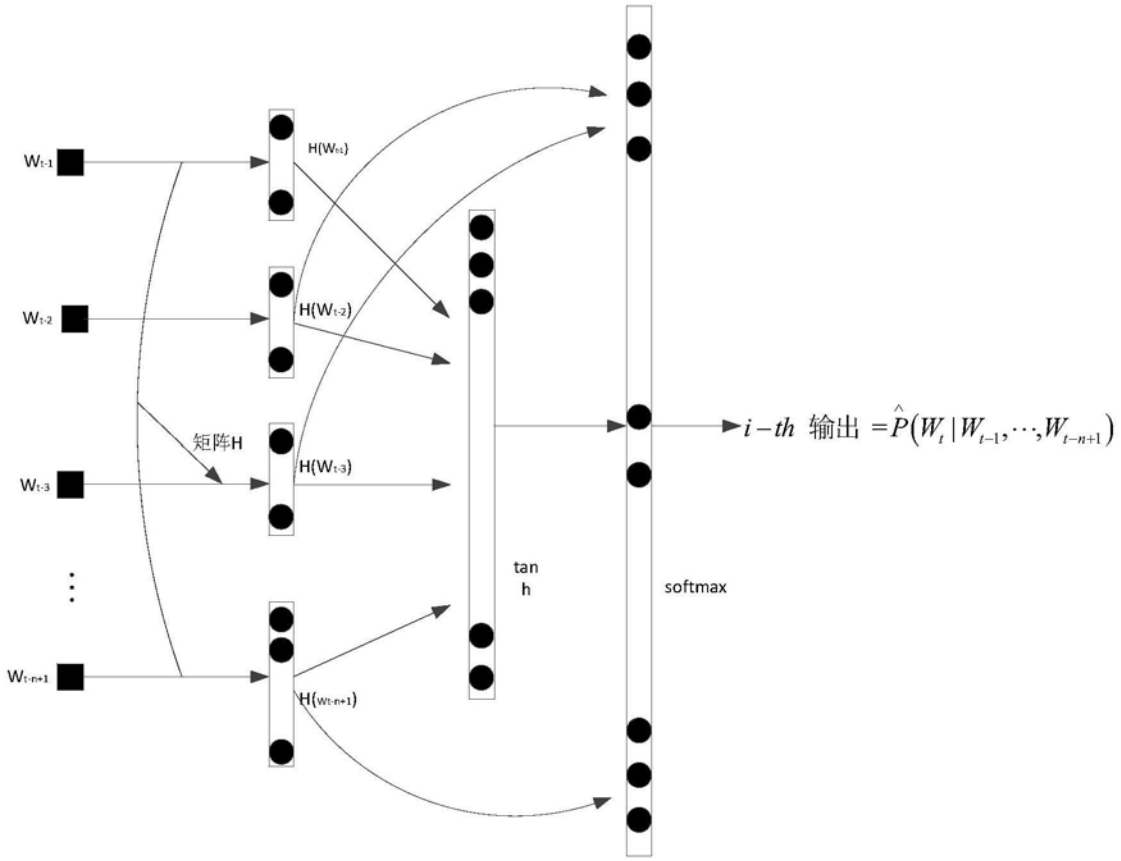

Wt-1
Wt-2
Wt-3
Wt-n+1
H(Wt-1)
H(Wt-2)
H(Wt-3)
H(Wt-n+1)
矩阵H
tan h
softmax
i-th 输出 $=\hat{P}(W_t|W_{t-1},\cdots,W_{t-n+1})$

图3

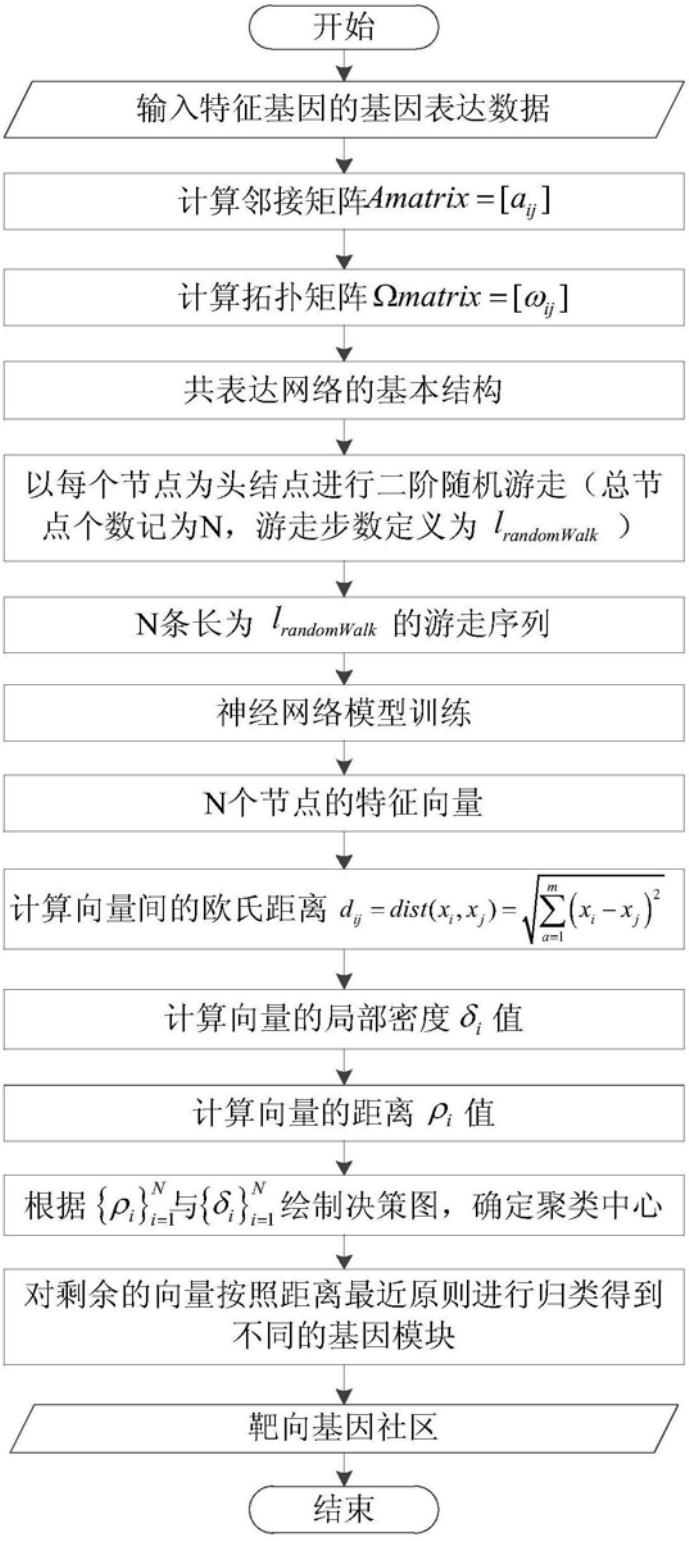

开始
输入特征基因的基因表达数据
计算邻接矩阵$Amatrix=[a_{ij}]$
计算拓扑矩阵$\Omega matrix=[\omega_{ij}]$
共表达网络的基本结构
以每个节点为头结点进行二阶随机游走（总节点个数记为N，游走步数定义为 $l_{randomWalk}$ ）
N条长为 $l_{randomWalk}$ 的游走序列
神经网络模型训练
N个节点的特征向量
计算向量间的欧氏距离 $d_{ij}=dist(x_i,x_j)=\sqrt{\sum_{a=1}^{m}(x_i-x_j)^2}$
计算向量的局部密度 δ_i 值
计算向量的距离 ρ_i 值
根据 $\{\rho_i\}_{i=1}^{N}$与$\{\delta_i\}_{i=1}^{N}$ 绘制决策图，确定聚类中心
对剩余的向量按照距离最近原则进行归类得到不同的基因模块
靶向基因社区
结束

图4

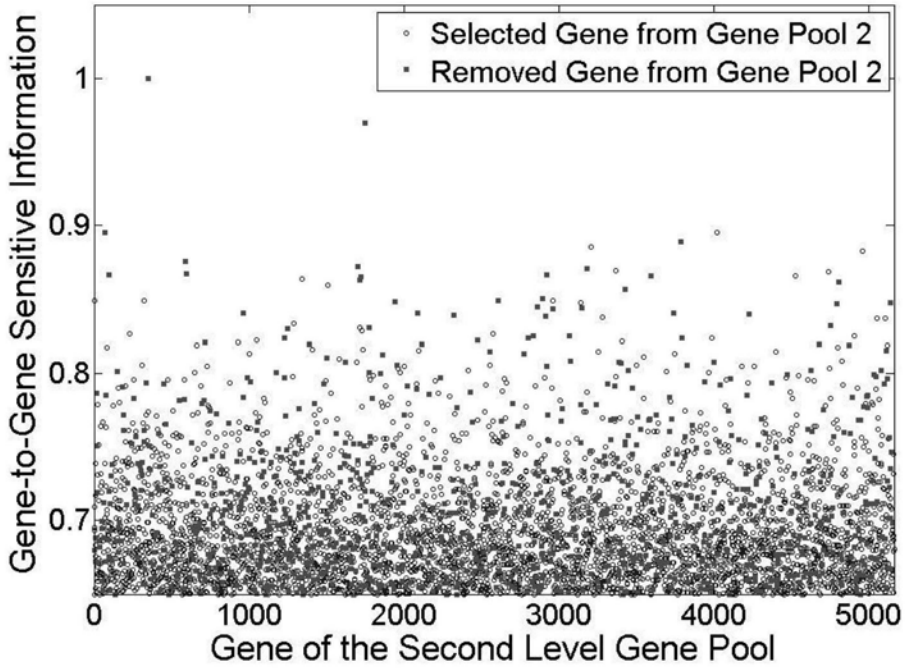

Selected Gene from Gene Pool 2
Removed Gene from Gene Pool 2
Gene-to-Gene Sensitive Information
1
0.9
0.8
0.7
0
1000
2000
3000
4000
5000
Gene of the Second Level Gene Pool

图5(a)

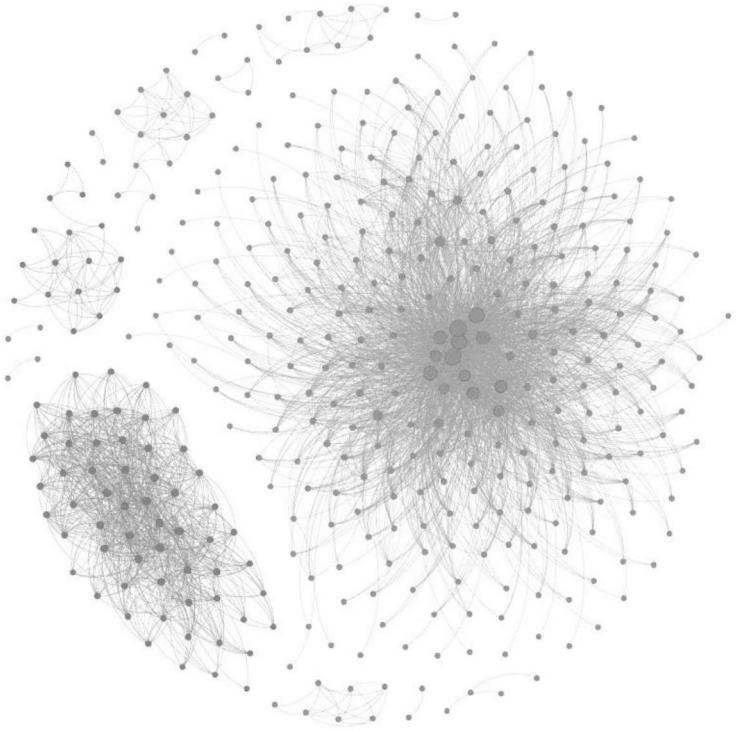


图5(b)

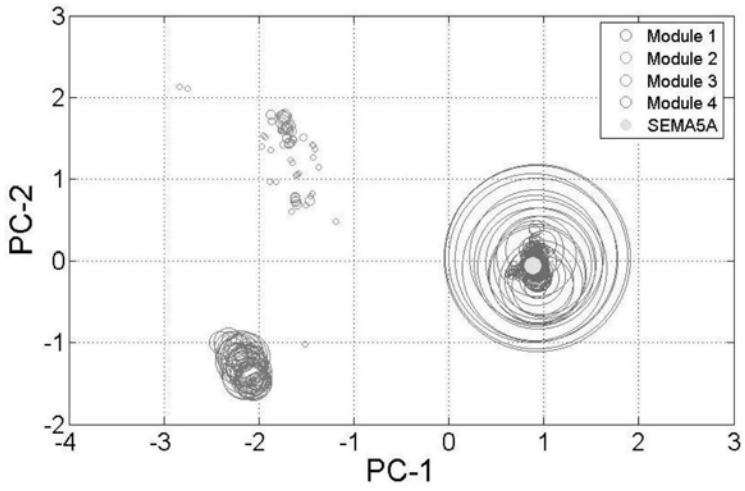

Module 1
Module 2
Module 3
Module 4
SEMA5A
PC-2
PC-1
3
2
1
0
-1
-2
-4
-3
-2
-1
0
1
2
3

图5(c)