(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 108427865 B
(45)授权公告日 2022.04.22

(21)申请号 201810209311.7
(22)申请日 2018.03.14
(65)同一申请的已公布的文献号
申请公布号 CN 108427865 A
(43)申请公布日 2018.08.21
(73)专利权人 华南理工大学
地址 510640 广东省广州市天河区五山路381号
(72)发明人 周杰 徐展良
(74)专利代理机构 广州市华学知识产权代理有限公司 44245
代理人 李斌
(51)Int.Cl.
G16B 40/20(2019.01)
G16B 50/30(2019.01)
G16B 20/00(2019.01)

(56)对比文件
CN 106934252 A,2017.07.07
US 2017091382 A1,2017.03.30
US 2018039729 A1,2018.02.08
CN 102693369 A,2012.09.26
Xing Chen等.Prediction of Disease-Related Interactions between MicroRNAs and Environmental Factors Based on a Semi-Supervised Classifier.《PLoS One》.2012,第8卷(第7期),第1-10页.
Twan van Laarhoven.Gaussian interaction profile kernels for predicting drug-target interaction.《BIOINFORMATICS》.2011,第3036-3043页. (续)

审查员 王菲

权利要求书2页 说明书7页 附图2页

(54)发明名称

一种预测LncRNA和环境因素关联关系的方法

(57)摘要

本发明公开了一种预测LncRNA和环境因素关联关系的方法,包括:S1、根据已知的LncRNA-环境因素关联关系网络的邻接矩阵,分别计算出LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE;S2、把KL和KE通过拉普拉斯算子进行归一化;S3、在LncRNA空间和环境因素空间中分别定义成本函数,通过最小化该成本函数,分别得到LncRNA空间和环境因素空间的最优分类器;S4、将得到的两种最优分类器转化为统一空间的分类器,用于计算LncRNA-环境因素关联关系的概率,根据概率对LncRNA-环境因素关联关系进行排名,概率值越高说明该LncRNA-环境因素的关联关系越强。所述方法有效解决了生物实验方法的盲目性、成本高问题,对生物学家的实验研究起到了指导的作用。

CN 108427865 B

[转续页]

[接上页]

(56)对比文件

Gamage Upeksha Ganegoda.Heterogeneous Network Model to Infer Human Disease-Long Intergenic Non-Coding RNA Associations.《IEEE TRANSACTIONS ON NANOBIOSCIENCE》.2015,第175-183页.

Meng Zhou.A computational frame and resource for understanding the lncRNA-environmental factorassociations and prediction of environmental factors implicated in diseases.《Molecular BioSystems》.2014,第3264-3271页.

1.一种预测LncRNA和环境因素关联关系的方法,其特征在于,所述方法包括以下步骤:

S1、根据已知的LncRNA-环境因素关联关系网络的邻接矩阵,分别计算出LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE;

为了提高LncRNA相关的预测精度,将得到的LncRNA相似信息进行logistic函数转换,经过转换得到的LncRNA相似信息记为SL:

$$SL(l_i,l_j)=\frac{1}{1+e^{c\cdot KL(l_i,l_j)+d}}$$

其中,c=-15,d=log(9999);KL(l_i,l_j)为一对LncRNA,即LncRNAl_i和LncRNAl_j之间的高斯相互作用属性核相似性;

根据环境因素的化学性质构造环境因素之间的化学结构相似性矩阵E,E的第i行、第j列元素E(i,j)表示环境因素i和j之间的化学结构相似性分数,通过环境因素之间的化学结构相似性矩阵E和高斯相互作用属性核相似矩阵KE,构建环境因素相似矩阵SE;

使用拉普拉斯算子将SL和SE进行归一化,公式如下:

$$\begin{cases} LL=DL^{-1/2}(DL-SL)DL^{-1/2} \\ LE=DE^{-1/2}(DE-SE)DE^{-1/2} \end{cases}$$

其中,DL和DE是对角矩阵;

S2、把LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE通过拉普拉斯算子进行归一化;

S3、在LncRNA空间和环境因素空间中分别定义成本函数,通过最小化该成本函数,分别得到LncRNA空间和环境因素空间的最优分类器;

S4、将步骤S3得到的两种最优分类器转化为统一空间的分类器,用于计算LncRNA-环境因素关联关系的概率,根据概率对LncRNA-环境因素关联关系进行排名,概率值越高说明该LncRNA-环境因素的关联关系越强。

2.根据权利要求1所述的一种预测LncRNA和环境因素关联关系的方法,其特征在于,步骤S1的具体过程为:基于功能相似的LncRNA与相似的环境因素之间具有关联关系的假设,利用已知的LncRNA-环境因素关联关系网络,构建LncRNA的高斯相互作用属性核相似矩阵KL,首先,每一个LncRNA的IP表示在已知的LncRNA-环境因素关联关系网络中的一个二进制向量编码,'1'代表存在关联关系,'0'代表不存在关联关系,对于一个给定的LncRNAl_i,它的IP(l_i)被定义为LncRNA-环境因素关联关系的邻接矩阵A的第i列,如果已知LncRNAl_i和环境因素e_j之间存在关联,则A(l_i,e_j)为1,否则为0;然后,计算每个LncRNA对,即LncRNAl_i和LncRNAl_j之间的高斯相互作用属性核相似性:

$$KL(l_i,l_j)=\exp\left(-\gamma_l\left\|IP(l_i)-IP(l_j)\right\|^2\right)$$

$$\gamma_l=\gamma_l'\Big/\left(\frac{1}{nl}\sum_{i=1}^{nl}\|IP(l_i)\|^2\right)$$

其中,γ_l用于控制高斯相互作用属性核相似性的频宽,它表示基于新的频宽参数γ'_l的正规化的高斯相互作用属性核相似性频宽,γ'_l取值为1;nl表示LncRNA的数量;KL表示

LncRNA的高斯相互作用属性核相似矩阵，元素KL(l_i,l_j)表示LncRNAl_i和LncRNAl_j的高斯相互作用属性核相似性；对于一个给定的LncRNAl_j，IP(l_j)为LncRNA-环境因素关联关系的邻接矩阵A的第j列；

同样地，基于功能相似的LncRNA与相似的环境因素之间具有关联关系的假设，利用已知的LncRNA-环境因素关联关系网络，构建环境因素的高斯相互作用属性核相似矩阵KE：

$KE(e_i,e_j)=\exp(-\gamma_e\|IP(e_i)-IP(e_j)\|^2)$

$$\gamma_e=\gamma'_e\Big/\left(\frac{1}{ne}\sum_{i=1}^{ne}\|IP(e_i)\|^2\right)$$

其中，γ_e表示基于新频宽参数γ'_e的正规化的高斯相互作用核相似性频宽；ne表示环境因素的数量；KE表示环境因素的高斯相互作用属性核相似矩阵，元素KE(e_i,e_j)表示环境因素e_i和环境因素e_j的高斯相互作用属性核相似性。

3.根据权利要求2所述的一种预测LncRNA和环境因素关联关系的方法，其特征在于，步骤S3中，LncRNA空间的最小化成本函数为：

$$\min_{FL}[\|A^T-FL\|_F^2+\eta L\|FL\cdot LL\cdot FL^T\|_F^2]$$

环境因素空间的最小化成本函数为：

$$\min_{FE}[\|A-FE\|_F^2+\eta E\|FE\cdot LE\cdot FE^T\|_F^2]$$

其中$\|\cdot\|_F$代表弗罗贝尼乌斯范数；ηL和ηE是权重参数，取值为1；通过计算目标函数的导数来解这两个最优化问题，分别得到LncRNA空间和环境因素空间的最优分类器如下；

$FL^*=SL(SL+\eta L\cdot LL\cdot SL)^{-1}A^T$

$FE^*=SE(SE+\eta E\cdot LE\cdot SE)^{-1}A$。

4.根据权利要求3所述的一种预测LncRNA和环境因素关联关系的方法，其特征在于，将步骤S3得到的两种最优分类器FL^*和FE^*通过一个加权操作转化为统一空间的分类器：

$F^*=lw\cdot FL^{*T}+(1-lw)\cdot FE^*$

其中，参数lw代表LncRNA空间和环境因素空间在整合分类函数中的权重系数，F^*是一个概率矩阵，大小为nl*ne，代表预测的LncRNA-环境因素的关联关系网络，最后，利用这个矩阵计算LncRNA-环境因素关联关系概率，根据概率对LncRNA-环境因素关联关系排名，概率值越高说明该LncRNA-环境因素的关联关系越强。

一种预测LncRNA和环境因素关联关系的方法

技术领域

[0001] 本发明涉及生物信息学领域，具体涉及一种预测LncRNA和环境因素关联关系的方法。

背景技术

[0002] 生物个体的先天本性和后天发展出来的行为习惯的不同主要是由遗传和环境的差异造成的。生物学家普遍认为，表型变异不是单纯由遗传或环境的差异产生，而是由两者的相互作用共同影响的；表型和疾病是由遗传因素(genetic factors，GFs)和环境因素(environmental factors，EFs)的复杂相互作用决定的。如今人们普遍认为，几乎所有的疾病都是由个体的遗传因子与其环境暴露之间复杂的相互作用引起的。例如，癌症、心脏病、阿尔茨海默病和糖尿病等人类疾病均是由GFs和EFs之间复杂的相互作用引起的。

[0003] 根据分子生物学的中心法则，遗传信息主要存储于DNA序列中。遗传信息从DNA转录成RNA，再从RNA翻译成蛋白质。RNA是DNA序列与其编码蛋白质之间的中间体。基因组序列分析表明，人类基因组中，编码蛋白质的序列占DNA序列的比例不到2％，其余约98％的DNA序列都不编码蛋白质。因此，由DNA转录的RNA中，绝大多数为不编码蛋白质的RNA。生物学中将不编码蛋白质的RNA称为非编码RNA(non-coding RNAs，ncRNAs)。在ncRNA中，长度在200-100000nt之间的ncRNA分子被称为长非编码RNA(Long non-coding RNA，LncRNA)。LncRNA占总RNA的比例可达4％-9％。分子生物学研究表明，LncRNA占总RNA的比例随着生物体复杂性的增加而增加。作为ncRNAs的一个重要子集，LncRNAs最近被确定为最大的具有显著多样化的RNA家族之一，并且已经成为不同物种基因组信息的重要组成部分。近年研究表明，LncRNA参与了X染色体沉默、染色体修饰和基因组修饰、转录激活、转录干扰、核内运输等过程，同时在细胞增殖分化、染色质重塑、表观遗传调控、基因组剪接、转录、翻译等许多重要生物过程中发挥着至关重要的作用。LncRNA是一类重要的调控生命过程的ncRNA，它在多层面上(表观遗传调控、转录调控以及转录后调控等)调控基因的表达。LncRNA被认为主要参与mRNA调控，并参与调节发育和疾病。在某些疾病中LncRNAs也被确定为药物靶点或预后因素。然而，由于LncRNA的调控网络复杂，其调控的潜在机制仍然不清楚。大多数LncRNAs的功能仍然未知，需要进一步的探索研究。

[0004] 然而，与基因和miRNA相比，利用生物信息学方法以及计算方法研究与疾病有关的LncRNA和EFs之间的关联关系却相对较少。Zhou等人设计了RWREFD(基于重启随机游走模型的LncRNA-EF关联关系预测模型)预测与疾病相关的LncRNA-EFs关联关系并开发了一个LncRNA-EFs关联关系数据库：LncEnvironmentDB，这是一个基于Web的数据库，旨在为LncRNA和EF提供全面的资源平台。Zhou和Shi设计了一个基于二分网络和资源转移的方法来预测LncRNA-EFs的关联关系，预测的结果覆盖了更多被实验证实的LncRNA-EFs的关联关系。存储LncRNA-EFs关联关系数据库已经被建立起来，越来越多LncRNA和环境因素的联系被实验所证实，因此，基于这些可用的生物数据发明有效的计算方法来预测潜在的LncRNA和环境因素之间的联系就显得非常重要。

发明内容

[0005] 本发明的目的是针对现有技术的不足，提供了一种预测LncRNA和环境因素关联关系的方法，所述方法基于半监督学习方法设计了预测LncRNA-环境因素关联关系的拉普拉斯正则化最小二乘法分类器，能更准确地预测出LncRNA和环境因素的关联关系，并且可以大规模地一次预测出多对LncRNA-环境因素之间关联关系的概率。

[0006] 本发明的目的可以通过如下技术方案实现：

[0007] 一种预测LncRNA和环境因素关联关系的方法，所述方法包括以下步骤：

[0008] S1、根据已知的LncRNA-环境因素关联关系网络的邻接矩阵，分别计算出LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE；

[0009] S2、把LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE通过拉普拉斯算子进行归一化；

[0010] S3、在LncRNA空间和环境因素空间中分别定义成本函数，通过最小化该成本函数，分别得到LncRNA空间和环境因素空间的最优分类器；

[0011] S4、将步骤S3得到的两种最优分类器转化为统一空间的分类器，用于计算LncRNA-环境因素关联关系的概率，根据概率对LncRNA-环境因素关联关系进行排名，概率值越高说明该LncRNA-环境因素的关联关系越强。

[0012] 进一步地，步骤S1的具体过程为：基于功能相似的LncRNA与相似的环境因素之间具有关联关系的假设，利用已知的LncRNA-环境因素关联关系网络，构建LncRNA的高斯相互作用属性核相似矩阵KL，首先，每一个LncRNA的IP(Interaction Profile)表示在已知的LncRNA-环境因素关联关系网络中的一个二进制向量编码，'1'代表存在关联关系，'0'代表不存在关联关系，对于一个给定的$lncRNAl_i$，它的$IP(l_i)$被定义为LncRNA-环境因素关联关系的邻接矩阵A的第i列，如果已知$lncRNAl_i$和环境因素e_j之间存在关联，则A(i,j)为1，否则为0；然后，计算每个LncRNA对，即$lncRNAl_i$和$lncRNAl_j$之间的高斯相互作用属性核相似性：

[0013] $$KL(l_i,l_j)=\exp(-\gamma_l||IP(l_i)-IP(l_j)||^2)$$

[0014] $$\gamma_l=\gamma'_l/(\frac{1}{nl}\sum_{i=1}^{nl}\|IP(l_i)\|^2)$$

[0015] 其中，γ_l用于控制高斯相互作用属性核相似性的频宽，它表示基于新的频宽参数γ'_l的正规化的高斯相互作用属性核相似性频宽，γ'_l取值为1；nl表示LncRNA的数量；KL表示LncRNA的高斯相互作用属性核相似矩阵，元素$KL(l_i,l_j)$表示$lncRNAl_i$和$lncRNAl_j$的高斯相互作用属性核相似性；

[0016] 同样地，基于功能相似的LncRNA与相似的环境因素之间具有关联关系的假设，利用已知的LncRNA-环境因素关联关系网络，构建环境因素的高斯相互作用属性核相似矩阵KE：

[0017] $$KE(e_i,e_j)=\exp(-\gamma_e||IP(e_i)-IP(e_j)||^2)$$

[0018] $$\gamma_e=\gamma'_e/(\frac{1}{ne}\sum_{i=1}^{ne}\|IP(e_i)\|^2)$$

[0019] 其中，γ_e表示基于新频宽参数γ'_e的正规化的高斯相互作用核相似性频宽；ne表示环境因素的数量；KE表示环境因素的高斯相互作用属性核相似矩阵，元素$KE(e_i,e_j)$表示环境因素e_i和环境因素e_j的高斯相互作用属性核相似性。

[0020] 进一步地，为了提高LncRNA相关的预测精度，将得到的LncRNA相似信息进行logistic函数转换，经过转换得到的LncRNA相似信息记为SL：

[0021] $$SL(l_i,l_j)=\frac{1}{1+e^{c\cdot KL(l_i,l_j)+d}}$$

[0022] 其中，c＝-15，d＝log(9999)；

[0023] 根据环境因素的化学性质构造环境因素之间的化学结构相似性矩阵E，E的第i行、第j列元素E(i，j)表示环境因素i和j之间的化学结构相似性分数，通过环境因素之间的化学结构相似性矩阵E和高斯相互作用属性核相似矩阵KE，构建环境因素相似矩阵SE：

[0024] $$SE(i,j)=\begin{cases} ew\cdot E(e_i,e_j)+(1+ew)\cdot KE(e_i,e_j) \\ KE(e_i,e_j) \end{cases}$$

[0025] 其中，ew代表两种环境因素关联关系信息在SE中的权重参数；

[0026] 使用拉普拉斯算子将SL和SE进行归一化，公式如下：

[0027] $$\begin{cases} LL=(DL)^{-1/2}(DL-SL)(DL)^{-1/2} \\ LE=(DE)^{-1/2}(DE-SE)(DE)^{-1/2} \end{cases}$$

[0028] 其中，DL和DE是对角矩阵，DL(i，i)和DE(i，i)分别表示SL和SE第i行的总和。

[0029] 进一步地，步骤S3中，LncRNA空间的最小化成本函数为：

[0030] $$\min_{FL}\left[\left\|A^T-FL\right\|_F^2+\eta L\left\|FL\cdot LL\cdot FL^T\right\|_F^2\right]$$

[0031] 环境因素空间的最小化成本函数为：

[0032] $$\min_{FE}\left[\left\|A-FE\right\|_F^2+\eta E\left\|FE\cdot LE\cdot FE^T\right\|_F^2\right]$$

[0033] 其中$||\cdot||_F$代表弗罗贝尼乌斯范数；ηL和ηE是权重参数，取值为1；通过计算目标函数的导数来解这两个最优化问题，分别得到LncRNA空间和环境因素空间的最优分类器如下：

[0034] $FL^*=SL(SL+\eta L\cdot LL\cdot SL)^{-1}A^T$

[0035] $FE^*=SE(SE+\eta E\cdot LE\cdot SE)^{-1}A$。

[0036] 进一步地，将步骤S3得到的两种最优分类器FL^*和FE^*通过一个加权操作转化为统一空间的分类器：

[0037] $F^*=lw\cdot FL^{*T}+(1-lw)\cdot FE^*$

[0038] 其中，参数lw代表LncRNA空间和环境因素空间在整合分类函数中的权重系数，F*是一个概率矩阵，大小为nl*ne，代表预测的LncRNA-环境因素的关联关系网络，最后，利用这个矩阵计算LncRNA-环境因素关联关系概率，根据概率对LncRNA-环境因素关联关系排名，概率值越高说明该LncRNA-环境因素的关联关系越强。

[0039] 本发明与现有技术相比，具有如下优点和有益效果：

[0040] 本发明采用半监督学习方法，通过引入高斯相互作用属性核相似性和拉普拉斯正则化最小二乘法分类器，利用已知的LncRNA与环境因素关联关系网络的拓扑结构，有效利用顶点和边蕴含的信息，训练最优分类器；作为一种全局测量方法，本发明对所有可能相关的LncRNA和环境因素的关联关系进行了优先级排序，这对生物学家的实验研究能够起到指

导的作用,生物学家可以针对关联关系概率较大的LncRNA和环境因素对进行试验测试,避免了盲目的测试,大大减少了工作量。

附图说明

[0041] 图1为本发明实施例预测LncRNA和环境因素关联关系方法的流程图。

[0042] 图2为使用本发明方法预测得到的LncRNA和环境因素关联关系与使用其他方法预测得到的LncRNA和环境因素关联关系的ROC曲线和AUC值对比示意图。

[0043] 图3为使用本发明方法预测得到的LncRNA和环境因素关联关系网络的度分布示意图。

具体实施方式

[0044] 下面结合实施例及附图对本发明作进一步详细的描述,但本发明的实施方式不限于此。

[0045] 实施例:

[0046] 本实施例提供了一种预测LncRNA和环境因素关联关系的方法,所述方法的流程图如图1所示,包括以下步骤:

[0047] S1、根据已知的LncRNA-环境因素关联关系网络的邻接矩阵,分别计算出LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE;

[0048] S2、把LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE通过拉普拉斯算子进行归一化;

[0049] S3、在LncRNA空间和环境因素空间中分别定义成本函数,通过最小化该成本函数,分别得到LncRNA空间和环境因素空间的最优分类器;

[0050] S4、将步骤S3得到的两种最优分类器转化为统一空间的分类器,用于计算LncRNA-环境因素关联关系的概率,根据概率对LncRNA-环境因素关联关系进行排名,概率值越高说明该LncRNA-环境因素的关联关系越强。

[0051] 其原理是通过引入高斯相互作用属性核相似性和拉普拉斯正则化最小二乘法分类器,利用已知的LncRNA-环境因素关联关系网络的拓扑信息,以及蕴藏在网络中的顶点和边的信息,训练最优分类器,从而计算LncRNA与环境因素之间的相关性。本实施例是一种全局测量方法,能对所有可能相关的LncRNA和环境因素的关联关系进行优先级排序,对生物学家的实验研究起到指导的作用,生物学家可以针对关联关系概率较大的LncRNA和环境因素对进行试验测试,避免了盲目的测试,大大减少了工作量。

[0052] 本实施例需要的数据从LncEnvironmentDB数据库中下载得到,该数据库包含5649个LncRNA-环境因素之间的关联关系,其中包含820个LncRNA和209种环境因素。

[0053] 根据上面的数据,具体实施包括以下步骤:

[0054] 步骤1、构建LncRNA-环境因素关联网络的邻接矩阵A。

[0055] 步骤2、基于功能相似的LncRNA与相似的环境因素之间具有关联关系的假设,利用已知的LncRNA-环境因素关联关系网络,构建LncRNA的高斯相互作用属性核相似矩阵KL,首先,每一个LncRNA的IP(Interaction Profile)表示在已知的LncRNA-环境因素关联关系网络中的一个二进制向量编码,‘1’代表存在关联关系,‘0’代表不存在关联关系,对于一个给

定的LncRNAl$_i$，它的IP(l$_i$)被定义为LncRNA-环境因素关联关系的邻接矩阵A的第i列，如果已知LncRNAl$_i$和环境因素e$_j$之间存在关联，则A(i,j)为1，否则为0；然后，计算每个LncRNA对，即LncRNAl$_i$和LncRNAl$_j$之间的高斯相互作用属性核相似性：

[0056] $KL(l_i,l_j)=\exp(-\gamma_l||IP(l_i)-IP(l_j)||^2)$

[0057] $$\gamma_l=\gamma'_l/(\frac{1}{nl}\sum_{i=1}^{nl}\|IP(l_i)\|^2)$$

[0058] 其中，γ_l用于控制高斯相互作用属性核相似性的频宽，它表示基于新的频宽参数γ'_l的正规化的高斯相互作用属性核相似性频宽，γ'_l取值为1；nl表示LncRNA的数量；KL表示LncRNA的高斯相互作用属性核相似矩阵，元素KL(l$_i$,l$_j$)表示LncRNAl$_i$和LncRNAl$_j$的高斯相互作用属性核相似性；

[0059] 同样地，基于功能相似的LncRNA与相似的环境因素之间具有关联关系的假设，利用已知的LncRNA-环境因素关联关系网络，构建环境因素的高斯相互作用属性核相似矩阵KE：

[0060] $KE(e_i,e_j)=\exp(-\gamma_e||IP(e_i)-IP(e_j)||^2)$

[0061] $$\gamma_e=\gamma'_e/(\frac{1}{ne}\sum_{i=1}^{ne}\|IP(e_i)\|^2)$$

[0062] 其中，γ_e表示基于新频宽参数γ'_e的正规化的高斯相互作用核相似性频宽；ne表示环境因素的数量；KE表示环境因素的高斯相互作用属性核相似矩阵，元素KE(e$_i$,e$_j$)表示环境因素e$_i$和环境因素e$_j$的高斯相互作用属性核相似性。

[0063] 步骤3、为了提高LncRNA相关的预测精度，将得到的LncRNA相似信息进行logistic函数转换，经过转换得到的LncRNA相似信息记为SL：

[0064] $$SL(l_i,l_j)=\frac{1}{1+e^{c\cdot KL(l_i,l_j)+d}}$$

[0065] 其中，c=-15，d=log(9999)；

[0066] 根据环境因素的化学性质构造环境因素之间的化学结构相似性矩阵E，E的第i行、第j列元素E(i,j)表示环境因素i和j之间的化学结构相似性分数，通过环境因素之间的化学结构相似性矩阵E和高斯相互作用属性核相似矩阵KE，构建环境因素相似矩阵SE：

[0067] $$SE(i,j)=\begin{cases}ew\cdot E(e_i,e_j)+(1+ew)\cdot KE(e_i,e_j)\\KE(e_i,e_j)\end{cases}$$

[0068] 其中，ew代表两种环境因素关联关系信息在SE中的权重参数；

[0069] 使用拉普拉斯算子将SL和SE进行归一化，公式如下：

[0070] $$\begin{cases}LL=(DL)^{-1/2}(DL-SL)(DL)^{-1/2}\\LE=(DE)^{-1/2}(DE-SE)(DE)^{-1/2}\end{cases}$$

[0071] 其中，DL和DE是对角矩阵，DL(i,i)和DE(i,i)分别表示SL和SE第i行的总和。

[0072] 步骤4、在LncRNA空间和环境因素空间中分别定义成本函数，通过最小化该成本函数，分别得到LncRNA空间和环境因素空间的最优分类器，其中LncRNA空间的最小化成本函数为：

[0073] $$\min_{FL}\left[\left\|A^{T}-FL\right\|_{F}^{2}+\eta L\left\|FL\cdot LL\cdot FL^{T}\right\|_{F}^{2}\right]$$

[0074] 环境因素空间的最小化成本函数为：

[0075] $$\min_{FE}\left[\left\|A-FE\right\|_{F}^{2}+\eta E\left\|FE\cdot LE\cdot FE^{T}\right\|_{F}^{2}\right]$$

[0076] 其中$||\cdot||_{F}$代表弗罗贝尼乌斯范数;ηL和ηE是权重参数,取值为1;通过计算目标函数的导数来解这两个最优化问题,分别得到LncRNA空间和环境因素空间的最优分类器如下：

[0077] $FL^{*}=SL(SL+\eta L\cdot LL\cdot SL)^{-1}A^{T}$

[0078] $FE^{*}=SE(SE+\eta E\cdot LE\cdot SE)^{-1}A$。

[0079] 步骤5、,将得到的两种最优分类器FL^{*}和FE^{*}通过一个加权操作转化为统一空间的分类器：

[0080] $F^{*}=lw\cdot FL^{*T}+(1-lw)\cdot FE^{*}$

[0081] 其中,参数lw代表LncRNA空间和环境因素空间在整合分类函数中的权重系数,F^{*}是一个概率矩阵,大小为nl*ne,代表预测的LncRNA-环境因素的关联关系网络,最后,利用这个矩阵计算LncRNA-环境因素关联关系概率,根据概率对LncRNA-环境因素关联关系排名,概率值越高说明该LncRNA-环境因素的关联关系越强。

[0082] 通过留一验证对本实施例预测LncRNA和环境因素关联关系的方法进行性能评估,在留一验证中,5949个LncRNA-环境因素之间的关联关系中,依次去掉其中一个并将它当作测试样例,剩余的作为训练集。如果测试样例的排名高于特定阈值,则可以认为对该测试LncRNA-环境因素进行正确的预测。用ROC曲线下面的面积(AUC)定量评估本实施例所述方法的性能,从图2可以看出,利用留一验证法,通过本方法预测得到的LncRNA和环境因素关联关系的ROC曲线优于用其他方法得到的LncRNA-环境因素关联关系的ROC曲线,本方法预测得到的LncRNA-环境因素关联关系的AUC为0.9096,其他方法预测得到的LncRNA和环境因素关联关系的AUC为0.7732。

[0083] 图3表明通过本方法预测得到的LncRNA和环境因素关联关系网络的度分布符合幂律分布,显示了生物网络的一般特征,说明通过本方法预测得到的LncRNA和环境因素关联关系网络不是随机网络,具有生物学意义。通过本方法预测得到的LncRNA和环境因素关联关系的网络拓扑特征参数如表1所示：

[0084]

特征	特征参数
节点数目	412
边数目	5649
聚集系数	0
连通数量	1
直径	4
半径	3
中心度	0.393
最短路径百分比	100%
特征路径长度	2.341
平均邻结节点数量	33.102
密度	0.081
网络异质性	1.232
整体效率	0.457
调和平均数	2.186

[0085]

传递性	0
中心点优势	0.275

[0086] 表1

[0087] 从表1可以看出，通过本方法预测得到的LncRNA和环境因素关联关系网络显示出短的特征路径长度，一个连通分支，低的直径和密度，表明与其他生物网络一样具有小世界和无标度性以及模块化结构，另外，通过本方法预测得到的LncRNA和环境因素关联关系网络具有较高的集中度，平均邻居数，全局效率和中心点优势；这表明通过本方法预测得到的LncRNA和环境因素关联关系网络中存在重要性更高的结点或边，即使部分网络损坏，网络的其他部分仍可以连通。

[0088] 以上所述，仅为本发明专利较佳的实施例，但本发明专利的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明专利所公开的范围内，根据本发明专利的技术方案及其发明专利构思加以等同替换或改变，都属于本发明专利的保护范围。

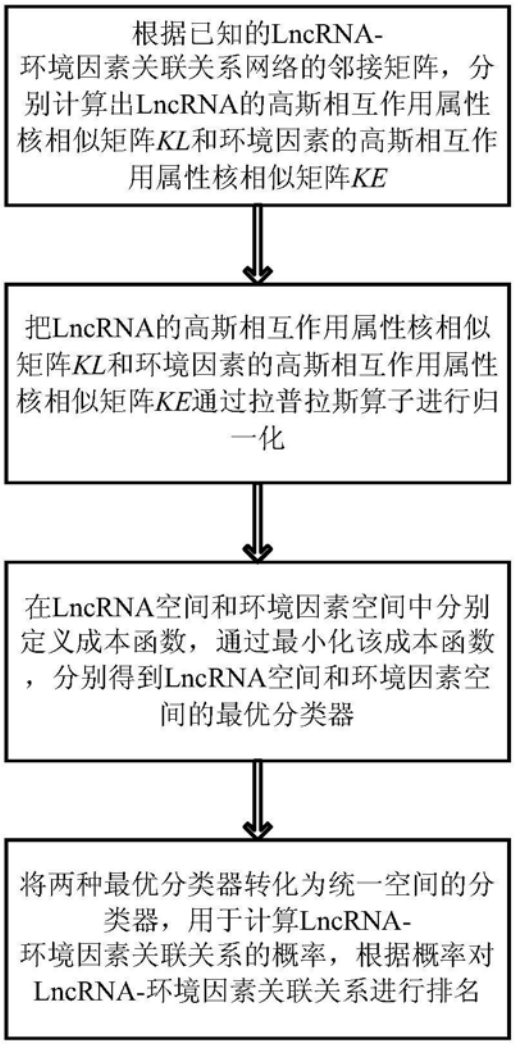
根据已知的LncRNA-环境因素关联关系网络的邻接矩阵，分别计算出LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE
把LncRNA的高斯相互作用属性核相似矩阵KL和环境因素的高斯相互作用属性核相似矩阵KE通过拉普拉斯算子进行归一化
在LncRNA空间和环境因素空间中分别定义成本函数，通过最小化该成本函数，分别得到LncRNA空间和环境因素空间的最优分类器
将两种最优分类器转化为统一空间的分类器，用于计算LncRNA-环境因素关联关系的概率，根据概率对LncRNA-环境因素关联关系进行排名

图1

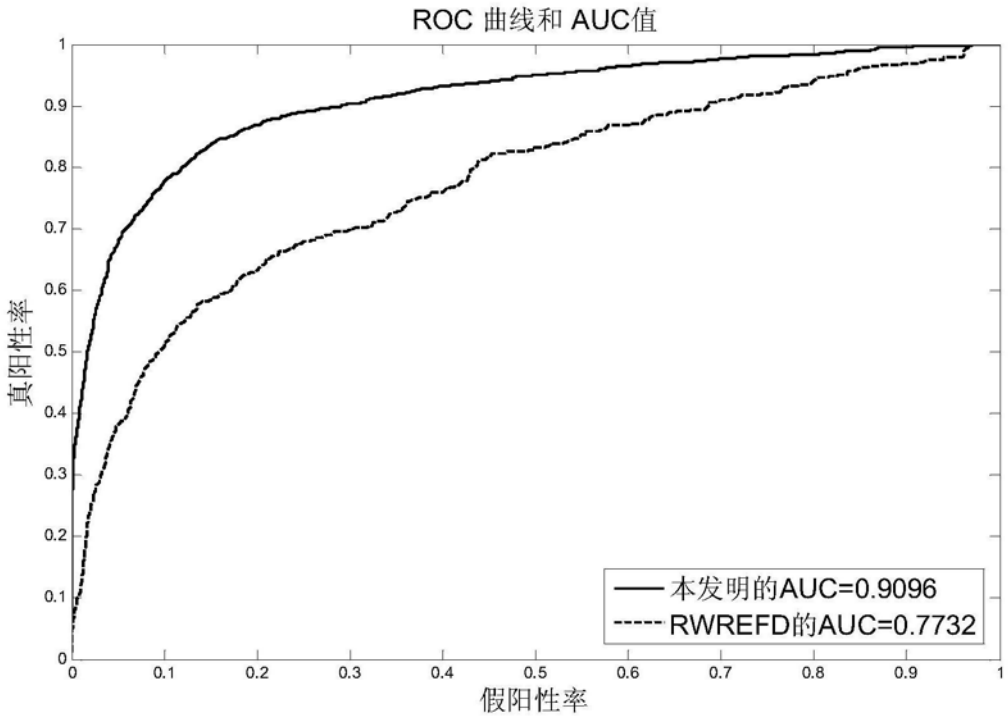
ROC 曲线和 AUC值
真阳性率
假阳性率
本发明的AUC=0.9096
RWREFD的AUC=0.7732

图2

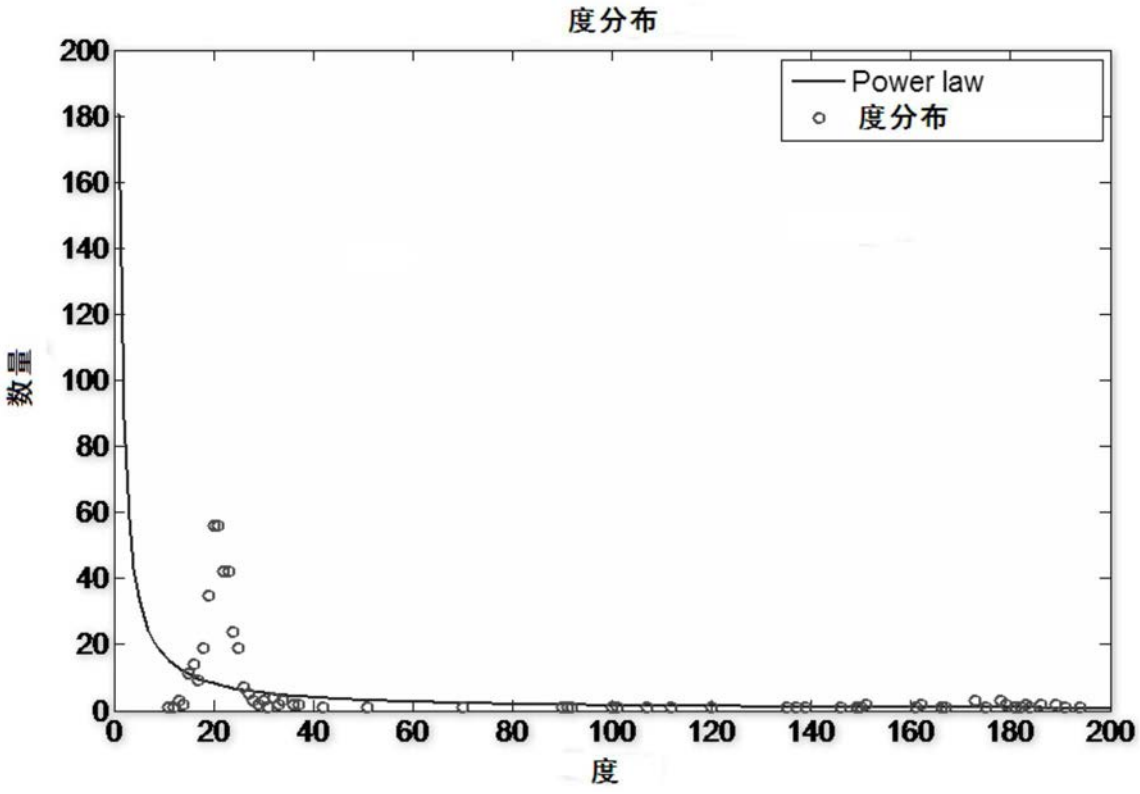
度分布
Power law
度分布
数量
度

图3