浙江理工大学学报，2020，43(5)：693-700
Journal of Zhejiang Sci-Tech University
DOI:10.3969/j.issn.1673－3851(n).2020.05.015

基于网络间随机游走算法的 lncRNA 与疾病关系预测

尚　敏，贺平安
（浙江理工大学理学院，杭州 310018）

摘　要：通过构造疾病之间、lncRNA 之间、基因之间的相似性网络模型，根据已有三者之间的相互关联数据，引入网络间随机游走算法预测潜在与疾病相关的 lncRNA。将该算法应用到已有数据库中的 lncRNA 与疾病关联信息数据上，通过 10 倍交叉验证和 AUC 评估该算法。结果表明：与其他算法相比，该算法在 10 倍交叉验证下有更优的 AUC 值，显示该算法具有良好的预测性能。进一步应用该算法预测了与胃癌和结直肠癌相关的 lncRNA，结果与已知的医学数据一致，表明该算法对于新疾病相关 lnRNA 的预测结果是有效的。

关键词：疾病；基因；长链非编码 RNA；网络模型；网络间随机游走；相似性；预测
中图分类号：O29　　**文献标志码**：A　　**文章编号**：1673-3851 (2020) 05-0693-08

Prediction of the relationship between lncRNA and diseases based on internetwork random walk algorithm

SHANG Min, *HE Pingan*
(School of Science, Zhejiang Sci-Tech University, Hangzhou 310018, China)

Abstract: The similarity network model of diseases, lncRNA and genes was constructed. Based on the existing correlation data among the three, an internetwork random walk algorithm was introduced to predict the potential disease-related lncRNA. The algorithm was applied to the data of lncRNA and disease association information in the existing databases. The 10-fold cross validation and AUC were applied to evaluate the effect of the algorithm. The results show that compared with other algorithms, this algorithm has better AUC value under 10-fold cross validation, demonstrating that the algorithm has a better prediction performance. Furthermore, the algorithm was used to predict the lncRNA related to gastric cancer and colorectal cancer. The predicted results are consistent with the known medical data, which shows that the algorithm is effective in predicting new disease-related lnRNA.

Key words: disease; gene; long non-coding RNA; network model; internetwork random walk; similarity; prediction

0　引　言

目前已知的蛋白质编码基因约有 2 万个，占人类基因组的比例不到 2%[1]。人类基因组中约有 70%的 DNA 被转录成非编码 RNA（Non-coding RNA，ncRNA），在很长一段时间内被认为是转录噪音[2]。随着研究的不断深入，科学家们发现 ncRNA 虽不编码蛋白，但可参与调控细胞分化、增殖、凋亡、代谢以及肿瘤的发生等多个重要生物学过程[3]。长链非编码 RNA（Long noncoding RNA，

收稿日期：2020－01－30　　网络出版日期：2020－05－08
基金项目：国家自然科学基金项目(61772027)
作者简介：尚　敏(1994－　)，女，湖北襄阳人，硕士研究生，主要从事生物信息学方面的研究。
通信作者：贺平安，E-mail：pinganhe@zstu.edu.cn

lncRNA)是一类长度大于200个核苷酸的ncRNA。lncRNA的失调伴随着多种疾病的发生,包括糖尿病[4]、心血管疾病[5]、HIV[6]、神经障碍[7]和一些癌症如肺癌[8]、乳腺癌[9]和前列腺癌[10]等。因此,识别与疾病的相关lncRNA有助于在分子水平上理解疾病相关病理过程的发生,同时有助于疾病的预防、诊断和治疗。

许多与疾病相关的重要lncRNA已被发现。例如,阿尔茨海默症是一种起病隐匿的神经性疾病,β-分泌酶-1(BACE1)的非编码反义转录本(BACE1-AS)是阿尔茨海默症病理生理学中的关键酶。BACE1-AS在体内外调节BACE1的mRNA和BACE1蛋白的表达可能驱动阿尔茨海默症的发生[11]。H19是最早被发现与癌症有关的lncRNA之一,它有抑癌和促癌的双重作用[12-13]。在肝癌中,H19上调会促进肝癌细胞的增殖,下调会抑制增殖;在结肠癌中,H19上调会促进肿瘤细胞的增殖[14]。

随着生物信息技术的不断发展,科学家们开发了多个数据库用于存储lncRNA与疾病关联数据,如lncRNA与疾病关系数据库LncRNADisease2.0[15]、lncRNA与癌症关系数据库Lnc2Cancer[16]、ncRNA与疾病网络数据库MNDR[17]等,这些数据库为建立数学模型预测潜在的lncRNA与疾病关联性提供了强有力的数据基础。预测lncRNA与疾病关联性的计算模型一般可分为基于机器学习的模型和基于网络的模型[18]。

基于机器学习的模型,通常通过训练样本提取特征训练预测器,采用交叉验证或独立数据测试其性能。例如,Chen等[19]开发了一种基于半监督学习框架的拉普拉斯正则化最小二乘法(Laplacian regularized least squares for lncRNA-disease association, LRLSLDA)计算模型,该模型假设相似的疾病倾向于与功能相似的lncRNA相关,通过整合已知的lncRNA与疾病关联性和lncRNA表达谱,对疾病相关的lncRNA候选者进行排序,从而推断潜在的与疾病相关的lncRNA。LRLSLDA在优化模型参数方面存在困难。Zhao等[20]提出了一种朴素贝叶斯分类器方法,它利用与癌症相关的lncRNA的各种信息,包括调节物组、基因组、转录组和多组数据,发现了707个潜在的与癌症相关的lncRNA。然而,这种方法需要负样本,而负样本通常是未知的。

基于网络的模型,利用lncRNA与疾病关联网络、疾病相似性网络和lncRNA相似性网络研究lncRNA与疾病之间的关联性。例如,Sun等[21]开发了一种基于全局网络的RWRlncD计算模型,该模型通过在lncRNA功能相似网络上进行重启随机游走(Random walk with restart, RWR),以推断潜在的lncRNA与疾病之间的关联性。然而,该方法不能应用于孤立疾病,即该疾病不与任何lncRNA相关。Gu等[22]提出了一种预测lncRNA与疾病关联性的全局网络随机游走模型(Global network random walk model for predicting lncRNA-disease associations, GrwLDA),该模型在lncRNA功能相似网络和疾病相似网络上执行RWR。GrwLDA在优化模型参数方面也存在困难。Xiao等[23]开发了BPLLDA模型,该模型根据在异构网络中以有限长度连接它们的路径预测lncRNA与疾病关联性。Peng等[24]提出了一种名为ThrRW的网络间随机游走算法,在三种不同的生物网络,即蛋白质相互作用网络(PIN)、域共现网络(DCN)和功能相互关系网络(FIN),分别进行不同步数的随机游走,以推断相应网络中蛋白质的功能信息。在迭代过程中,功能信息会根据不同网络中节点之间的关联从一个网络转移到另一个网络。

本文根据网络间随机游走算法,提出了一种基于不同网络之间的随机游走算法来预测潜在的lncRNA与疾病之间的关联性。首先,计算疾病与疾病、lncRNA与lncRNA以及基因与基因之间的相似性用于构造疾病相似性网络、lncRNA相似性网络和基因相似性网络。利用lncRNA、疾病以及基因之间的关联信息构造关联矩阵。然后,在3个网络中进行不同步数的随机游走,以推断疾病与相应网络中lncRNA之间的关系。将该算法应用在已知数据库中的lncRNA与疾病关联信息数据上,通过10倍交叉验证对已知的lncRNA与疾病关联性进行预测,采用AUC值作为评价指标来调试参数,获得算法中参数的最佳取值;进一步将预测结果与已有模型的预测结果比较以验证本文算法的有效性。最后,将该算法应用到与胃癌和结直肠癌相关的lncRNA预测来说明该算法的可行性。

1 数据与方法

1.1 数据

1.1.1 lncRNA与疾病关联数据和lncRNA与基因关联数据

大量的研究[4-10]已经证明,许多lncRNA参与了疾病病理过程的发生。在生物学实验中已经检测

到大量与疾病相关的 lncRNA。Bao 等[15]开发了 LncRNADisease2.0 数据库来管理这些数据，该数据库收录了超过 20 万个 lncRNA 与疾病关联，包含实验验证和计算预测的 lncRNA 与疾病关联数据；lncRNA、基因和微 RNA(MicroRNA，miRNA)之间的转录调节关联数据以及环状 RNA 与疾病关联数据。本文从 LncRNADisease2.0 数据库中下载实验验证的 lncRNA 与疾病关联和 lncRNA 与基因关联数据。

1.1.2　基因与疾病关联数据

CREEDS 数据库[25]提供了基因与疾病关联、基因表达谱、疾病与药物关联等数据。本文下载的基因与疾病关联数据均是 CREEDS 数据库中实验验证的基因与疾病关联数据。

1.1.3　数据整合

将 LncRNADisease2.0 数据库中的 lncRNA 与疾病、lncRNA 与基因以及 CREEDS 数据库中的基因与疾病关联信息数据三者合并取交集，具体方法如下：首先将 lncRNA 与疾病和基因与疾病关联数据匹配，其中有 99 种疾病相同，删除其他没有与 99 种疾病关联的 lncRNA 和基因部分；然后将 lncRNA 与疾病和 lncRNA 与基因关联数据匹配，其中有 1395 种 lncRNA 相同，保留与 1395 种 lncRNA 关联的疾病和基因部分；最后将 lncRNA 与基因和基因与疾病关联数据匹配，其中有 2902 种基因相同，删除其他没有与 2902 种基因关联的 lncRNA 和疾病部分。

因此，本文整合后的数据为 99 种疾病、2902 种基因和 1395 种 lncRNA 之间的关联数据。

1.2　疾病语义相似性、lncRNA 功能相似性及基因功能相似性

1.2.1　疾病语义相似性

本文采用层次有向无环图来计算两种疾病之间的相似性[26]。对于一种疾病 d，$G_d=(d,\ T_d,\ E_d)$ 是它的有向无环图，其中：T_d 表示疾病 d 的所有祖先节点集合(包括疾病 d 本身)；E_d 表示 G_d 中连接各个节点边的集合。对于任意的疾病 $t\in T_d$，假设 t 是 d 的祖先，或者 $t=d$，则 t 对 d 的贡献 $D_d(t)$ 的计算公式为：

$$D_d(t)=\begin{cases}1, & t=d\\ \max\{D_e\times D_d(t')\mid t'\in t\text{ 的孩子}\}, & t\neq d\end{cases}\tag{1}$$

其中：$D_e\in[0,1]$ 是边 $e(e\in E_d)$ 的语义贡献因子。疾病 d_1 和 d_2 的相似性 r_{d_1,d_2} 被定义为：

$$r_{d_1,d_2}=\frac{\sum_{t\in T_{d_1}\cap T_{d_2}}(D_{d_1}(t)+D_{d_2}(t))}{\sum_{t\in T_{d_1}}D_{d_1}(t)+\sum_{t\in T_{d_2}}D_{d_2}(t)}\tag{2}$$

$$\boldsymbol{R}_{dd}=\begin{bmatrix}r_{d_1,d_1} & \cdots & r_{d_1,d_n}\\ \vdots & \ddots & \vdots\\ r_{d_n,d_1} & \cdots & r_{d_n,d_n}\end{bmatrix}\tag{3}$$

其中：$\boldsymbol{R}_{dd}$ 表示疾病语义相似性矩阵；n 表示疾病个数。

1.2.2　lncRNA 功能相似性

本文通过与 lncRNA 相关的疾病的相似性来表示 lncRNA 相似性[21,26]。假设与 lncRNA l_1 关联的疾病集合为：$D_1=\{d_{11},\ d_{12},\cdots,\ d_{1m}\}$，与 lncRNA l_2 关联的疾病集合为：$D_2=\{d_{21},\ d_{22},\cdots,\ d_{2n}\}$。首先，计算一个疾病 d_{11} 和一个疾病集合 D_2 的相似性 $S(d_{11},\ D_2)$，计算公式为：

$$S(d_{11},D_2)=\max_{d\in D_2}(\boldsymbol{R}_{dd}(d_{11},d))\tag{4}$$

然后，lncRNA l_1 和 l_2 的功能相似性 r_{l_1,l_2} 的计算公式为：

$$r_{l_1,l_2}=\frac{\sum_{1\leqslant k\leqslant m}S(d_{1k},D_2)+\sum_{1\leqslant l\leqslant n}S(d_{2l},D_1)}{m+n}\tag{5}$$

$$\boldsymbol{R}_{ll}=\begin{bmatrix}r_{l_1,l_1} & \cdots & r_{l_1,l_s}\\ \vdots & \ddots & \vdots\\ r_{l_s,l_1} & \cdots & r_{l_s,l_s}\end{bmatrix}\tag{6}$$

其中：$\boldsymbol{R}_{ll}$ 表示 lncRNA 功能相似矩阵；s 表示 lncRNA 的个数。

1.2.3　基因功能相似性

基因本体(Gene ontology，GO)[27]是在生物医学中广泛应用的一个本体，它分三个层面：生物进程、细胞组分和分子功能，对基因产物进行注释，这些注释的 GO 节点提供了一种计算基因相似性的方法。本文采用 Wang 等[28]提出的一种基于 GO 术语语义相似性的基因功能相似性评价算法计算基因功能相似性。

对于一个 GO 节点 A，$G_A=(d,\ T_A,\ E_A)$ 是它的有向无环图，其中：T_A 表示 G_A 中 GO 节点的集合，包括 GO 节点 A 和它的所有祖先 GO 节点；E_A 表示 G_A 中连接各个 GO 节点边的集合。对于任意的 GO 节点 $t\in T_A$，假设 t 是 A 的祖先，或者，$t=A$，t 对 A 的贡献 $S_A(t)$ 的计算公式为：

$$S_A(t)=\begin{cases}1, & t=A\\ \max\{W_e\times S_A(t')\mid t'\in t\text{ 的孩子}\}, & t\neq A\end{cases}\tag{7}$$

其中：W_e 为节点关联权重，取值范围为[0，1]。定义节点 A 的语义贡献 $V(A)$ 的计算公式为：

$$V(A)=\sum_{t\in T_A}S_A(t) \tag{8}$$

分别给定节点 A 和节点 B 的有向无环图 $G_A=(d, T_A, E_A)$ 和 $G_B=(d, T_B, E_B)$，节点 A 和节点 B 的语义相似性 $S_{GO}(A, B)$ 的计算公式为：

$$S_{GO}(A,B)=\frac{\sum_{t\in T_A\cap T_B}(S_A(t)+S_B(t))}{V(A)+V(B)} \tag{9}$$

定义一个 GO 节点 o 和一个 GO 节点集合 $G=\{o_1, o_2, \cdots, o_k\}$ 的语义相似性 $U(o, G)$ 的计算公式为：

$$U(o,G)=\max_{1\leqslant i\leqslant k}(S_{GO}(o,o_i)) \tag{10}$$

假设已知基因 g_1 和 g_2 的 GO 术语集注释分别为：$G_1=\{o_{11}, o_{12}, \cdots, o_{1m}\}$ 和 $G_2=\{o_{21}, o_{22}, \cdots, o_{2n}\}$，则基因 g_1 和 g_2 的相似性 r_{g_1,g_2} 的计算公式为：

$$r_{g_1,g_2}=\frac{\sum_{1\leqslant i\leqslant m}U(o_{1i},G_2)+\sum_{1\leqslant j\leqslant n}U(o_{2j},G_1)}{m+n} \tag{11}$$

$$\boldsymbol{R}_{gg}=\begin{bmatrix} r_{g_1,g_1} & \cdots & r_{g_1,g_q} \\ \vdots & \ddots & \vdots \\ r_{g_q,g_1} & \cdots & r_{g_q,g_q} \end{bmatrix} \tag{12}$$

其中：$\boldsymbol{R}_{gg}$ 表示基因功能相似性矩阵；q 表示基因个数。

显然，$\boldsymbol{R}_{dd}$、$\boldsymbol{R}_{ll}$ 和 $\boldsymbol{R}_{gg}$ 都是对称矩阵。

1.3 网络间随机游走算法

令矩阵 $\boldsymbol{R}_{ld}$、$\boldsymbol{R}_{gd}$ 和 $\boldsymbol{R}_{lg}$ 分别表示已知的 lncRNA 与疾病、基因与疾病和 lncRNA 与基因关联矩阵。其中 $\boldsymbol{R}_{ld}$ 矩阵行名为 lncRNA，列名为疾病；$\boldsymbol{R}_{gd}$ 矩阵行名为基因，列名为疾病；$\boldsymbol{R}_{lg}$ 矩阵行名为 lncRNA，列名为基因。如果相应节点之间存在关联，则这些矩阵中的元素值为 1，否则为 0。例如，如果基因 i 和疾病 j 之间存在关联，则矩阵 $\boldsymbol{R}_{gd}$ 中的元素 $\boldsymbol{R}_{gd}(i, j)$ 值为 1。$\boldsymbol{P}_{ld}$ 和 $\boldsymbol{P}_{gd}$ 分别表示预测的 lncRNA 与疾病和基因与疾病关联矩阵，其中：$\boldsymbol{P}_{ld}$ 矩阵行名为 lncRNA，列名为疾病；$\boldsymbol{P}_{gd}$ 矩阵行名为基因，列名为疾病。

网络间随机游走算法通过建立疾病、lncRNA 和基因三者之间的网络，并在三个网络中进行不同步数的随机游走来实现目标，本文结构框架见图 1。图 1 中，正方形、圆形和三角形节点分别代表疾病、lncRNA 和基因；实线连接 2 个相同形状的节点，如连接 2 个正方形节点的边用疾病语义相似性来构造；虚线连接 2 个不同形状的节点，包括已知的 lncRNA 与疾病、lncRNA 与基因、基因与疾病关联和预测的 lncRNA 与疾病关联。

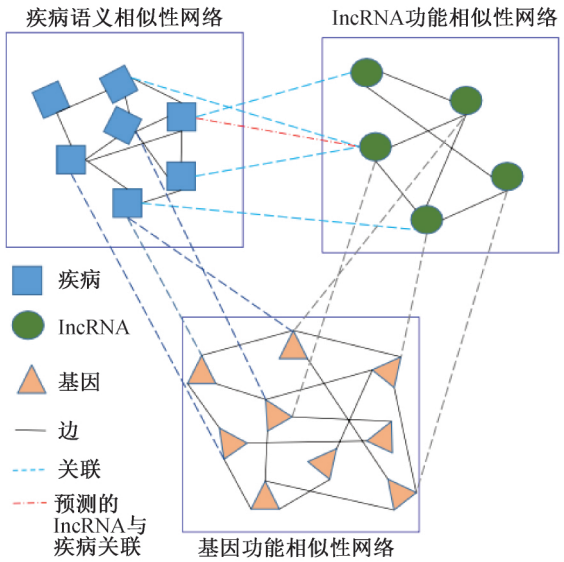

图 1 网络间随机游走算法结构框架

网络间随机游走算法根据矩阵 $\boldsymbol{R}_{dd}$、$\boldsymbol{R}_{ld}$、$\boldsymbol{R}_{ll}$、$\boldsymbol{R}_{lg}$ 和 $\boldsymbol{P}_{gd}$ 迭代更新矩阵 $\boldsymbol{P}_{ld}$ 的值来实现目标。因此，有三种方法可以更新矩阵 P_{ld} 的值。首先，在疾病与疾病语义相似性网络中进行几个随机游走步骤(步长用 l_1 表示)，如式(13)；其次，在 lncRNA 与 lncRNA 功能相似性网络中进行几个随机游走步骤(步长用 r_1 表示)，如式(14)；第三，在 lncRNA 与基因关联网络中进行几个随机游走步骤(步长用 l_1 表示)，如式(15)。由于疾病与疾病语义相似性网络和 lncRNA 与 lncRNA 功能相似性网络之间的差异，在两个网络中的步长是不同的(疾病与疾病语义相似性网络中为 l_1 步，lncRNA 与 lncRNA 功能相似性网络中为 r_1 步)。在数学上，该过程可以表示如下：

a)疾病与疾病语义相似性网络，即：

$$\boldsymbol{P}_{ld}^{t}=\alpha\boldsymbol{P}_{ld}^{t-1}\cdot\boldsymbol{R}_{dd}+(1-\alpha)\boldsymbol{R}_{ld} \tag{13}$$

b)lncRNA 与 lncRNA 功能相似性网络，即：

$$\boldsymbol{P}_{ld}^{t}=\alpha\boldsymbol{R}_{ll}\cdot\boldsymbol{P}_{ld}^{t-1}+(1-\alpha)\boldsymbol{R}_{ld} \tag{14}$$

c)lncRNA 与基因关联网络，即：

$$\boldsymbol{P}_{ld}{}^{t}=\boldsymbol{R}_{lg}\cdot\boldsymbol{P}_{gd}{}^{t-1} \tag{15}$$

其中：参数 α 调节已知 lncRNA 与疾病关联或已知基因与疾病关联迭代过程的程度，取值范围为[0, 1]；$\boldsymbol{P}_{ld}$ 初始矩阵为 $\boldsymbol{R}_{ld}$。

另一方面，类似于矩阵 $\boldsymbol{P}_{ld}$ 的方法计算 $\boldsymbol{P}_{gd}$。该过程可以表示如下：

a)疾病与疾病语义相似性网络，即：

$$\boldsymbol{P}_{gd}^{t}=\alpha\boldsymbol{P}_{gd}^{t-1}\cdot\boldsymbol{R}_{dd}+(1-\alpha)\boldsymbol{R}_{gd} \tag{16}$$

b)基因与基因功能相似性网络，即：

$$\boldsymbol{P}_{gd}^{t}=\alpha\boldsymbol{R}_{gg}\cdot\boldsymbol{P}_{gd}^{t-1}+(1-\alpha)\boldsymbol{R}_{gd} \tag{17}$$

c)lncRNA 与基因关联网络，即：

$$\boldsymbol{P}_{gd}^{t}=\boldsymbol{R}'_{lg}\cdot\boldsymbol{P}_{ld}^{t-1} \tag{18}$$

其中：疾病与疾病语义相似性网络中步长为 l_2，如式(16)；基因与基因功能相似性网络中步长为 r_2，如式(17)；lncRNA 与基因关联网络中步长为 l_2，如式(18)；$\boldsymbol{P}_{gd}$ 初始矩阵为 $\boldsymbol{R}_{gd}$。

1.4　模型评估

预测的准确性取决于预测函数与实际函数的匹配程度，本文采用受试者工作特征曲线(ROC 曲线)下面积(AUC 值)的大小来衡量。ROC 曲线以伪阳性率(假正类率，False positive rate)为横坐标，以真阳性率(真正类率，True positive rate)为纵坐标。伪阳性率是指判定为负例却不是真负例的概率，真阳性率是指判定为正例也是真正例的概率。AUC 值一般在 0.500～1.000 之间。AUC 值越大，说明该模型的性能越好。

1.5　交叉验证

为了评估预测方法的性能，本文采用 10 倍交叉验证方法。10 倍交叉验证是把样本随机均分为 10 份，其中 9 份作为训练集，其余 1 份作为测试集，重复 10 次，直到 10 份中每一份都曾作为测试集，将 10 次结果取均值作为对算法精度评估的依据。

本文将 $\boldsymbol{R}_{ld}$ 矩阵中为 1 的元素随机均分为 10 份，选取其中的 9 份作为训练集，将另外的一份值设为 0。基于 $\boldsymbol{R}_{dd}$ 和 $\boldsymbol{R}_{ld}$ 矩阵计算 $\boldsymbol{R}_{ll}$ 矩阵，将 10 次 ROC 曲线下的面积取均值作为本文的 AUC 值。

2　结果与讨论

将本文算法应用于 LncRNADisease2.0 数据库中的 lncRNA 与疾病关联、lncRNA 与基因关联和 CREEDS 数据库中的基因与疾病关联数据上，计算得到疾病语义相似性、lncRNA 功能相似性和基因功能相似性，部分疾病语义相似性数据见表 1，部分 lncRNA 功能相似性数据见表 2，部分基因功能相似性数据见表 3。

表 1　部分疾病语义相似性数据

疾病	乳腺癌	结直肠癌	胃癌	黑素瘤	肺癌
乳腺癌	1.0000	0.3041	0.4220	0.3192	0.4220
结直肠癌	0.3041	1.0000	0.4756	0.2437	0.3041
胃癌	0.4220	0.4756	1.0000	0.3192	0.4220
黑素瘤	0.3192	0.2437	0.3192	1.0000	0.3192
肺癌	0.4220	0.3041	0.4220	0.3192	1.0000

观察表 1 可以看到疾病语义相似性数据的范围均在 0 到 1 之间，对角线元素取值均为 1.0000。除对角线数据外，表 1 中数据最大为 0.4756，最小为 0.2437。最大数据是胃癌和结直肠癌的语义相似性数据，最小数据是结直肠癌和黑素瘤的语义相似性数据。胃癌和结直肠癌的语义相似性较高，是因为它们都属于消化系统癌症且有较多的相同祖先节点；结直肠癌和黑素瘤的语义相似性较低，是因为它们有较少的相同祖先节点。

表 2 为部分 lncRNA 功能相似性数据。观察表 2 可以看到 lncRNA 功能相似性数据的范围均在 0 到 1 之间，对角线元素取值均为 1.0000。除对角线数据外，表 2 中数据最大为 1.0000，最小为 0.2554。最大数据为 1.0000 的有两组：A2M-AS1 和 ABHD11-AS1 的相似性数据、AC000067.1 和 AC002070.1 的相似性数据。A2M-AS1 和 ABHD11-AS1 都只与一种疾病胃癌相关联，因此两者之间相似性较高。

表 2　部分 lncRNA 功能相似性数据

lncRNA	A2M-AS1	ABALON	ABHD11-AS1	AC000067.1	AC002070.1
A2M-AS1	1.0000	0.3192	1.0000	0.2554	0.2554
ABALON	0.3192	1.0000	0.3192	0.5275	0.5275
ABHD11-AS1	1.0000	0.3192	1.0000	0.2554	0.2554
AC000067.1	0.2554	0.5275	0.2554	1.0000	1.0000
AC002070.1	0.2554	0.5275	0.2554	1.0000	1.0000

表 3 为部分基因功能相似性数据。观察表 3 可以看到 lncRNA 功能相似性数据的范围均在 0 到 1 之间，对角线元素取值均为 1.0000。除对角线数据外，表 3 中数据最大为 0.6708，最小为 0.3455。

网络间随机游走算法引入了 5 个参数(α，l_1，r_1，l_2，r_2)。本文参数取值范围如下：参数 α 在[0，1]区间取值，间距设为 0.1；参数 l_1 在[1，15]区间取值，间距设为 1；参数 r_1 在[1，15]区间取值，间距设为 1；参数 l_2 在[1，15]区间取值，间距设为 1；参数 r_2 在[1，15]区间取值，间距设为 1，部分参数选择及结果见表 4。

表3 部分基因功能相似性数据

基因	A2M	AACS	AADAT	AAGAB	AAK1	AAMP
A2M	1.0000	0.5675	0.3653	0.6708	0.6134	0.5260
AACS	0.5675	1.0000	0.3455	0.4928	0.6434	0.4167
AADAT	0.3653	0.3455	1.0000	0.3636	0.3962	0.3503
AAGAB	0.6708	0.4928	0.3636	1.0000	0.5978	0.6477
AAK1	0.6134	0.6434	0.3962	0.5978	1.0000	0.4876
AAMP	0.5260	0.4167	0.3503	0.6477	0.4876	1.0000

表4 参数 l_1,r_1,l_2,r_2 取不同值得到的AUC值

l_1	r_1	l_2	r_2	AUC值
2	2	2	2	0.946
2	3	2	2	0.932
2	4	2	2	0.860
2	5	2	2	0.843
3	2	2	2	0.917
3	3	2	2	0.943
3	4	2	2	0.831

注：$\alpha=0.1$。

通过实验发现：当 α 从0.1取到0.9时，AUC值并无明显的变化；当固定 l_1,r_1 时，变动 l_2,r_2 对AUC值的影响并不大；当固定 r_1,l_2,r_2 时，变动 l_1（从4开始）AUC值随着参数 l_1 的增加而下降；当固定 l_1,l_2,r_2 时，变动 r_1（从4开始）AUC值随着参数 r_1 的增加而下降。因此，本文将参数 α 的值设置为0.1，当 $l_1=2$,$r_1=2$,$l_2=2$,$r_2=2$ 时，AUC值达到最大，此时的AUC值为0.946。

为了与其他计算模型做比较，将本文数据应用于Gu等[22]的GrwLDA模型（参数选择 $\gamma=0.9$,$\alpha=0.1$,$\beta=0.1$,$\eta=0.3$）、Chen等[19]的LRLSLDA模型（参数选择 $lw=0.1$）和Xiao等[23]的BPLLDA模型（参数选择 $L=3$,$T=0.2$）。实验结果如下：网络间随机游走算法的AUC值为0.946，比LRLSLDA的AUC值0.820提高了12.6%；比GrwLDA的AUC值0.782提高了16.4%；比BPLLDA的AUC值0.872提高了7.4%，AUC值结果如图2所示。因此，本文算法对于lnRNA与疾病关联的预测性能优于LRLSLDA、GrwLDA和BPLLDA模型。

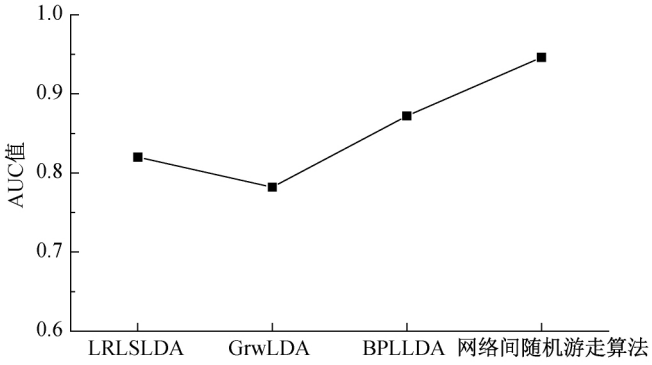

图2 LRLSLDA、GrwLDA、BPLLDA和网络间随机游走算法的AUC值

为了评估网络间随机游走算法预测新疾病相关lncRNA的能力，本文选择2种疾病：胃癌和结直肠癌，删除了与这2种疾病相关联的lncRNA，参数 α 的值设置为0.1，当 $l_1=2$,$r_1=2$,$l_2=2$,$r_2=2$ 时分别预测了排名前5的与疾病相关的lncRNA，实验结果见表5，这些与疾病关联的lncRNA均在LncRNADisease2.0数据库和相关文献中得到证实，进一步体现了本文算法的有效性。

表5 网络间随机游走算法预测预测出排名前5的与疾病相关的lncRNA

疾病	预测出排名前5的与疾病相关的lncRNA				
胃癌	HOTAIR[29]	H19[30]	MEG3[31]	MALAT1[32]	TUG1[33]
结直肠癌	HOTAIR[34]	H19[35]	MEG3[36]	UCA1[37]	MALAT1[38]

3 结 论

lncRNA是生物体的重要调控因子，可作为很多疾病的诊断标记物。研究lncRNA与疾病关联性有助于诊断、预后和治疗这些疾病。本文基于已知数据库中的疾病、lncRNA和基因相互之间的关联信息，构造疾病、lncRNA和基因自身之间的相似性网络，提出了一种基于网络的随机游走算法，预测与疾病相关的lncRNA。为了证明网络间随机游走算法的有效性，将该算法与其他模型应用到LncRNADisease2.0数据库中的lncRNA与疾病关联数据上，得到比其他模型更优的AUC值为0.946，结果表明本文算法的预测性能优于其他模型。

将本文算法应用于新疾病预测研究中，预测的lncRNA均在LncRNADisease2.0数据库和相关文

献中得到证实，进一步体现了该算法的有效性。另外，本文算法也可预测潜在基因与疾病关联性。

参考文献：

[1] Claverie J M. Fewer genes, more noncoding RNA [J]. Science, 2005, 309(5740):1529-1530.

[2] Clark M B, Mattick J S. Long noncoding RNAs in cell biology [J]. Seminars in Cell & Developmental Biology, 2011, 22(4):366-376.

[3] Mercer T R, Dinger M E, Mattick J S. Long non-coding RNAs: Insights into functions [J]. Nature Reviews Genetics, 2009, 10(3):155-159.

[4] Pasmant E, Sabbagh A, Vidaud M, et al. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS [J]. The FASEB Journal, 2011, 25(2):444-448.

[5] Congrains A, Kamide K, Oguro R, et al. Genetic variants at the 9p21 locus contribute to atherosclerosis through modulation of ANRIL and CDKN2A/B [J]. Atherosclerosis, 2012, 220(2):449-455.

[6] Zhang Q, Chen C Y, Yedavalli V S R K, et al. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression [J]. mBio, 2013, 4(1):e00596.

[7] Johnson R. Long non-coding RNAs in Huntington's disease neurodegeneration [J]. Neurobiology of Disease, 2012, 46(2):245-254.

[8] Ji P, Diederichs S, Wang W B, et al. MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer [J]. Oncogene, 2003, 22(39):8031-8041.

[9] Barsyte-Lovejoy D, Lau S K, Boutros P C, et al. The c-myc oncogene directly induces the H19 noncoding RNA by allele-specific binding to potentiate tumorigenesis [J]. Cancer Research, 2006, 66(10):5330-5337.

[10] de Kok J B, Verhaegh G W, Roelofs R W, et al. DD3 (PCA3), a very sensitive and specific marker to detect prostate tumors [J]. Cancer Research, 2002, 62(9): 2695-2698.

[11] Faghihi M A, Modarresi F, Khalil A M, et al. Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of β-secretase [J]. Nature Medicine, 2008, 14(7):723-730.

[12] Hao Y, Crenshaw T, Moulton T, et al. Tumour-suppressor activity of H19 RNA [J]. Nature, 1993, 365(6448):764-767.

[13] Brannan C I, Dees E C, Ingram R S, et al. The product of the H19 gene may function as an RNA [J]. Molecular Cellular Biology, 1990, 10(1):28-36.

[14] Gibb E A, Brown C J, Lam W L. The functional role of long non-coding RNA in human carcinomas [J]. Molecular Cancer, 2011, 10(1):38.

[15] Bao Z Y, Yang Z, Huang Z, et al. LncRNADisease2.0: An updated database of long non-coding RNA-associated diseases [J]. Nucleic Acids Research, 2019, 47(D1):D1034-D1037.

[16] Ning S W, Zhang J Z, Wang P, et al. Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers [J]. Nucleic Acids Research, 2016, 44(D1):D980-D985.

[17] Wang Y, Chen L, Chen B, et al. Mammalian ncRNA-disease repository: A global view of ncRNA-mediated disease network [J]. Cell Death & Disease, 2013, 4 (8):e765.

[18] Chen X, Yan C C, Zhang X, et al. Long non-coding RNAs and complex diseases: From experimental results to computational models [J]. Briefings in Bioinformatics, 2017, 18(4):558-576.

[19] Chen X, Yan G Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles [J]. Bioinformatics, 2013, 29(20):2617-2624.

[20] Zhao T T, Xu J Y, Liu L, et al. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features [J]. Molecular BioSystems. 2015, 11(1): 126-136.

[21] Sun J, Shi H B, Wang Z Z, et al. Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network [J]. Molecular BioSystems, 2014, 10(8): 2074-2081.

[22] Gu C L, Liao B, Li X Y, et al. Global network random walk for predicting potential human lncRNA-disease associations [J]. Scientific Reports, 2017, 7(1): 12442.

[23] Xiao X F, Zhu W, Liao B, et al. BPLLDA: Predicting lncRNA-disease associations based on simple paths with limited lengths in a heterogeneous network [J]. Frontiers in Genetics, 2018, 9:411.

[24] Peng W, Wang J X, Chen L, et al. Predicting protein functions by using unbalanced bi-random walk algorithm on protein-protein interaction network and functional interrelationship network [J]. Current Protein & Peptide Science, 2014, 15(6):529-539.

[25] Wang Z C, Monteiro C D, Jagodnik K M, et al. Extraction and analysis of signatures from the gene expression omnibus by the crowd [J]. Nature

Communications, 2016, 7:12846.

[26] Wang D, Wang J, Lu M, et al. Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases [J]. Bioinformatics, 2010, 26(13):1644-1650.

[27] Gene Ontology Consortium. Gene ontology annotations and resources [J]. Nucleic Acids Research, 2013, 41 (D1): D530-D535.

[28] Wang J Z, Du Z, Payattakool R, et al. A new method to measure the semantic similarity of GO terms [J]. Bioinformatics, 2007, 23(10):1274-1281.

[29] Xu Z Y, Yu Q M, Du Y A, et al. Knockdown of long non-coding RNA HOTAIR suppresses tumor invasion and reverses epithelial-mesenchymal transition in gastric cancer [J]. International Journal of Biological Sciences, 2013, 9(6):587-597.

[30] Wu M S, Wang H P, Lin C C, et al. Loss of imprinting and overexpression of IGF2 gene in gastric adenocarcinoma [J]. · Cancer Lett, 1997, 120(1):9-14.

[31] Sun M, Xia R, Jin F, et al. Downregulated long noncoding RNA MEG3 is associated with poor prognosis and promotes cell proliferation in gastric cancer [J]. Tumor Biology, 2014, 35(2):1065-1073.

[32] Wang J, Su L, Chen X, et al. MALAT1 promotes cell proliferation in gastric cancer by recruiting SF2/ASF [J]. Biomedicine & Pharmacotherapy, 2014, 68(5): 557-564.

[33] Zhang E, He X, Yin D, et al. Increased expression of long noncoding RNA TUG1 predicts a poor prognosis of gastric cancer and regulates cell proliferation by epigenetically silencing of p57 [J]. Cell Death & Disease, 2016, 7(2):e2109.

[34] Kogo R, Shimamura T, Mimori K, et al. Long noncoding RNA HOTAIR regulates polycomb-dependent chromatin modification and is associated with poor prognosis in colorectal cancers [J]. Cancer Research, 2011, 71(20):6320-6326.

[35] Hibi K, Nakamura H, Hirai A, et al. Loss of H19 imprinting in esophageal cancer [J]. Cancer Research, 1996, 56(3):480-482.

[36] Yin D D, Liu Z J, Zhang E, et al. Decreased expression of long noncoding RNA MEG3 affects cell proliferation and predicts a poor prognosis in patients with colorectal cancer [J]. Tumor Biology, 2015, 36 (6):4851-4859.

[37] Han Y, Yang Y N, Yuan H H, et al. UCA1, a long non-coding RNA up-regulated in colorectal cancer influences cell proliferation, apoptosis and cell cycle distribution [J]. Pathology, 2014, 46(5):396-401.

[38] Xu C, Yang M, Tian J, et al. MALAT-1: A long non-coding RNA and its important 3' end functional motif in colorectal cancer metastasis [J]. International Journal of Oncology, 2011, 39(1):169-175.

(责任编辑:康 锋)