DOI:10.7544/issn1000-1239.2019.20180834
56(9):1889-1896,2019

基于 HeteSim 的疾病关联长非编码 RNA 预测

马 毅 郭杏莉 孙宇彤 苑倩倩 任 阳 段 然 高 琳
(西安电子科技大学计算机科学与技术学院 西安 710071)
(1006294412@qq.com)

Prediction of Disease Associated Long Non-Coding RNA Based on HeteSim

Ma Yi, Guo Xingli, Sun Yutong, Yuan Qianqian, Ren Yang, Duan Ran, and Gao Lin
(*School of Computer Science and Technology, Xidian University, Xi'an* 710071)

Abstract A growing number of evidences indicate that long non-coding RNAs (lncRNAs) play important roles in many biological processes, and mutations or dysfunction in these long non-coding RNAs can cause serious diseases in human bodies, such as various cancers. Biological methods have been exploited to predict potential associations between diseases and long non-coding RNAs, which are of great significance for the exploration of pathogenesis, diagnosis, treatment, prognosis and prevention of complex diseases. Heterogeneous information network is constructed based on the known disease-gene associations. The association strength between lncRNAs and diseases can be measured by an association score in the heterogeneous network. A simple method called HeteSim is applied to calculate the association scores between lncRNAs and diseases. The method used in this paper is based on all paths existing between a given disease and a given lncRNA. The experiments show that our method can achieve superior performance than state-of-art methods. Our predictions for ovarian cancer and gastric cancer have been verified by biological experiments, indicating the effectiveness of this method. The case studies indicate that our method can give informative clues for further investigation. In conclusion, the only paths based on known disease-gene associations are exploited, and it is can be expected that other disease associated information can also be integrated into our method, and better performance can be available.

Key words disease-gene prediction; correlation calculation; heterogeneous information networks; HeteSim; meta-path

摘 要 越来越多的研究表明,长非编码 RNA(long non-coding RNA, lncRNA)在许多生物过程中具有重要的功能,而这些长非编码 RNA 的变异或功能失调会导致一些复杂疾病的发生.通过生物信息学方法预测潜在的长非编码 RNA-疾病关联关系,对于致病机理的探索以及疾病诊断、治疗、预后和预防都具有重要的意义.基于疾病基因关联关系的异质信息网络,研究者使用了一种相关性计算法方法——HeteSim 来计算疾病基因之间的相关性,进而预测致病基因.使用的方法基于路径约束,具有可扩展性,算法效率高,留一交叉验证实验表明该方法的预测结果优于其他方法.将其应用在卵巢癌和胃癌的预测分析中,相关文献表明,所提方法的预测结果已被生物实验等验证,再次表明该方法的有效性.

收稿日期:2018-12-18;**修回日期**:2019-02-18
基金项目:国家自然科学基金面上项目(61672407,61672406);国家自然科学基金重点项目(61432010,61532014)
This work was supported by the General Program of the National Natural Science Foundation of China (61672407, 61672406) and the Key Program of the National Natural Science Foundation of China (61432010, 61532014).
通信作者:郭杏莉(xlguo@mail.xidian.edu.cn)

关键词 致病基因预测;相关性计算;异质信息网络;HeteSim 方法;元路径

中图法分类号 TP399

全基因组研究表明,2/3 的基因组能够被转录为 RNA,但其中只有一小部分可以翻译为蛋白质[1-4],非编码 RNA 大量存在于生物体内.通常,为了区别于其他短非编码 RNA,长非编码 RNA(long non-coding RNA, lncRNA)简单地定义为长度大于 200nt 且不编码蛋白质的一类 RNA 分子.lncRNA 在许多重要的生物过程中扮演关键角色,例如染色质修饰、转录和转录后调节[5].由于 lncRNA 在生命过程中发挥了重要作用,因此很大一部分人类疾病与 lncRNA 的变异以及功能失调息息相关.

随着已确定的 lncRNA 的数量持续增长,许多相关的数据库、计算方法被提出来,其中包括通用的数据库 GENCODE[6],针对 lncRNA 的专用数据库 lncRNAdb[7], LncRbase[8], LncRNA2Function[9], LncRNA2Target[10],同时包括基于网络的大规模 lncRNA 功能预测方法 lncGFP[11],以及通用的计算模型和框架[12].关于 lncRNA 在普通疾病和癌症中的作用,分别有 LncRNADisease[13] 和 Lnc2Cancer[14] 数据库.即使有一定数量的 lncRNA-疾病关联关系已经得到实验验证,不可忽略的是,绝大多数 lncRNA-疾病关联关系仍然是未知的.因此,分析 lncRNA 与疾病关联关系并预测潜在的关联关系具有重要的研究价值和社会意义.这些研究不仅可以帮助我们加深对复杂疾病在分子层面的致病机理的理解,而且可以利用 lncRNA 作为疾病诊断、预测的生物靶标以及治疗和预防的药物靶标.

预测潜在的疾病与 lncRNA 关联关系的计算方法可分为 2 大类:基于机器学习和基于网络的方法.基于机器学习的方法通常使用疾病与 lncRNA 关联关系来训练学习模型,然后用学习得到的模型来预测新的关联关系.这类方法整合了各种生物信息来注释 lncRNA.例如,Zhao 等人[15] 使用朴素贝叶斯模型来整合基因组、调节子和转录组特征,进而识别与癌症相关的潜在 lncRNA.这个方法需要阴性的训练样本(即与疾病无关的 lncRNA)来训练模型,考虑到并没有这种实验验证的阴性样本,在这项研究中,所有未知的 lncRNA-疾病关联关系被认为是阴性样本用于训练.最近,一个半监督模型——正则化最小二乘(RLS)[16] 克服了这一限制,该模型不需要阴性的训练样本.

相对于比较少的基于机器学习方法的研究,许多基于网络的方法被提出来预测与疾病相关的潜在 lncRNA.基于网络的方法通常根据 lncRNA 与疾病的关联得分大小对候选的 lncRNA 进行排序,进而预测致病基因.最常用的算法是标签传播算法,比如随机漫步(RWR)[17-21] 和 KATZ[22].这些研究的主要区别在于传播算法所应用的底层网络不同.例如:Sun 等人[17] 将 RWR 应用于 lncRNA 功能相似网络(lncRNA FSN);Liu 等人[18] 基于 lncRNA 和蛋白质编码基因表达谱构建了蛋白质编码基因-lncRNA 二部网络,然后利用 RWR 算法来预测癌症相关的 lncRNA;与此同时,Zhou 等人[19] 和 Ganegoda 等人[20] 结合 lncRNA 相似网络建立了 lncRNA-疾病异质信息网络,然后在该网络上应用 RWR 算法预测潜在疾病 lncRNA 关联关系.这些基于网络的方法是基于一种观察结果提出的,即在功能上类似的 lncRNA 通常与相同或相似的疾病联系在一起,即疾病模块原理.以上方法都是通过构建网络提出基于网络的计算模型,有的方法结合基因表达谱数据等构建网络,所构建网络结合了多种信息的逻辑关联网络,构建方法相对复杂.

本文使用了一种异质信息网络中节点相关性计算方法——HeteSim,该方法用来预测基因和疾病的关联关系,得到了很好的实验验证[23].因此,我们将这种方法应用到 lncRNA-疾病异质信息网络中,通过挖掘网络中疾病与 lncRNA 之间的关联关系,计算疾病与 lncRNA 关联得分,预测潜在疾病关联 lncRNA,预测结果优于其他方法.

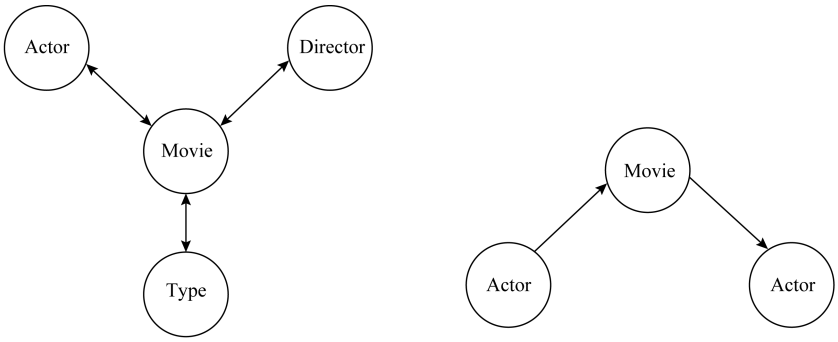
1 算 法

1.1 异质信息网络构建

预测 lncRNA 与疾病之间的关联关系可以理解为 lncRNA-疾病异质信息网络上的一个相关性搜索任务.异质信息网络是一种特殊的信息网络,下面是信息网络的定义,在此基础上可以定义得到同质信息网络和异质信息网络.

定义 1. 信息网络.给定一个模式 $S=(A, R)$,它由对象类型集合 A 和关系集合 R 构成.信息网络被抽象定义为一个有向图 $G=(V,E)$,其中,V 是所有实体节点的集合,E 是所有关系边的集合.并且

存在一个节点类型的映射函数 $\varphi: V \to A$ 和一个边类型的映射函数 $\theta: E \to R$,对于每个对象 $v \in V$ 属于一种特殊的对象类型 $\varphi(v) \in A$,每个链接 $e \in E$ 属于一种特殊的关系类型 $\theta(e) \in R$,那么这种网络类型就是信息网络.当对象类型的种类 $|A|>1$ 或者关系类型的种类 $|R|>1$ 时,这种信息网络是异质信息网络.例如图 1(a)就是由电影数据构建成电影异质信息网络.



(a) Movie data　　(b) Meta-path: *AMA*

Fig. 1　Heterogeneous information network instance and meta-path[24]

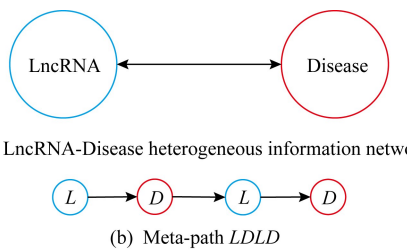
图 1　异质信息网络实例和元路径[24]

在信息网络中,我们将对象的类型和关系的类型明确区分开,不同类型对象之间存在的关系可以用网络模式清晰地描述.我们把类型 A 和类型 B 之间的关系 R 表示为 $A \xrightarrow{R} B$,其中 A 和 B 分别是关系 R 的源类型和目标类型,逆关系 R^{-1} 可以表示为 $B \xrightarrow{R^{-1}} A$.一般情况下关系 R 不等于关系 R^{-1},除非 R 是对称的并且关系两端的对象类型是相同的.此外,元路径是基于网络模式定义的,表示对象类型之间的关系,如图 1(b)就表示电影异质信息网络里的一种元路径 *AMA*,表示演员之间的合作关系.

基于已知的 lncRNA 与疾病关联关系,构建 lncRNA-疾病异质信息网络,如图 2(a)所示.网络中包含 2 种类型节点,分别为 lncRNA 和疾病,包含 1 种类型的边,即 lncRNA-疾病关联关系.为了集成更多的疾病相关的基因信息,类似地,我们集成了 OMIM(online mendelian inheritance in man)数据库中已知的编码基因与疾病的关联关系,将上面所构建的异质信息网络进行了扩展.扩展后的网络中包含 2 种类型节点,分别为基因和疾病,其中基因包括 lncRNA 和从 OMIM 中集成的编码基因.相应的边扩展为基因-疾病关联关系.lncRNA 与疾病的关联预测在基因-疾病关联异质信息网络上进行.

1.2　元路径选择

由于 HeteSim 是一种路径约束的相关性计算方法,所以选择相关路径是非常重要的.构建了异质信息网络之后,我们的目的是要研究 lncRNA 和疾病的相关关系,即通过现有的异质信息网络预测出 lncRNA 是否和其他疾病相关联,因此我们选择 lncRNA-疾病-lncRNA-疾病(*LDLD*)作为元路径,如图 2 所示.在此路径下使用 HeteSim 算法计算 lncRNA 和疾病之间的相关性,就能根据已有的关系预测出潜在的 lncRNA-疾病关联关系.



(a) LncRNA-Disease heterogeneous information network

(b) Meta-path *LDLD*

Fig. 2　LncRNA-Disease heterogeneous information network and meta-path *LDLD*

图 2　LncRNA-疾病异质信息网络和元路径 *LDLD*

1.3　模型描述

石川等人[24]提出了 HeteSim 算法来计算异质信息网络中任意节点对的相关性,该方法具有对称特性而且可以计算相同或不同类型对象之间的相关性,从而适用于很多的应用.HeteSim 是一种基于双向随机游走(pair-wise random walk)的相关性计算方法,它将元路径 P 分割成 2 条相等长度的元路径 P_L 和 P_R,之后将对象 s 和 t 分别沿着元路径 P_L 和 P_R 进行随机游走,最后将 2 个对象走到相同中间节点的概率作为 s 和 t 的相关性.

给定一个相关路径 $P=A_1A_2\cdots A_{l+1}$,该路径可以被分解为 2 条相等长度的路径 P_L 和 P_R.$P=$

$P_L P_R$, $P_L = A_1 A_2 \cdots A_{mid-1} M$, $P_R = M A_{mid+1} \cdots A_{l+1}$. M 为路径中的中间类型对象,当路径长度为偶数时 $mid = \frac{l}{2}+1$,当路径长度为奇数时 $mid = \frac{l+1}{2}+1$.

对于我们选择的元路径 lncRNA-疾病-lncRNA-疾病($LDLD$),由于路径长度是奇数,元路径两端的 2 个节点始终都不会在 1 个点相遇,因此我们需要插入中间类型 M 从而使路径可以等分成路径 $P_L = LDM$ 和 $P_R = MLD$,如图 3 所示:

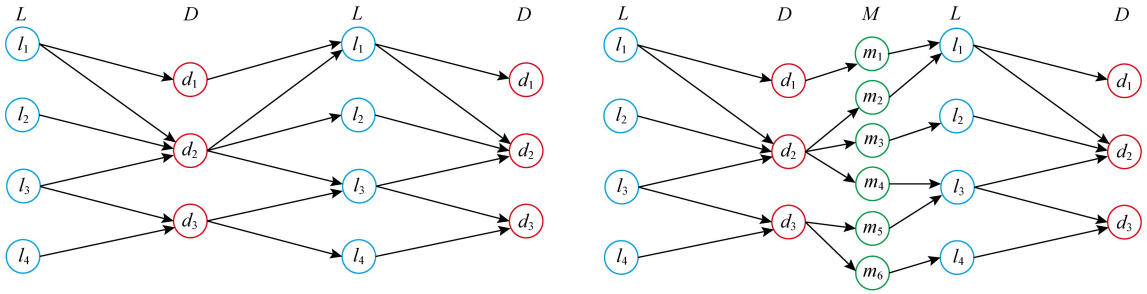


Fig. 3 Before and after insertion of the intermediate type M

图 3 插入中间类型 M 前后

下面介绍如何利用矩阵乘法计算 lncRNA 和疾病之间的关联得分.首先,我们定义 2 类矩阵:转移概率矩阵和可达概率矩阵.

定义 2. 转移概率矩阵.定义有向元路径 $A \xrightarrow{R} B$,对象 A 和对象 B 之间的连接关系为 R(A 和 B 表示同一类型对象构成的集合),A 和 B 之间的关系可以用 0/1 邻接矩阵 $\boldsymbol{W}_{AB}$ 表示,元素 1 表示 2 节点连通,元素 0 表示 2 节点不连通.将 0/1 邻接矩阵 $\boldsymbol{W}_{AB}$ 分别按照行向量和列向量进行标准化操作,得到矩阵 $\boldsymbol{X}_{AB}$ 和 $\boldsymbol{Y}_{AB}$. $\boldsymbol{X}_{AB}$ 和 $\boldsymbol{Y}_{AB}$ 就是转移概率矩阵,分别表示 $A \xrightarrow{R} B$ 和 $B \xrightarrow{R^{-1}} A$ 这 2 种有向关系.根据矩阵的性质,可以得到:

$$\boldsymbol{X}_{AB} = \boldsymbol{Y}_{BA}^{\mathrm{T}}, \boldsymbol{Y}_{AB} = \boldsymbol{X}_{BA}^{\mathrm{T}}.$$

定义 3. 可达概率矩阵.转移概率矩阵是可达概率矩阵的特例.转移概率矩阵用来描述长度为 1 的元路径节点间的关系,而可达概率矩阵则用来衡量在元路径长度大于 1(复合关系 $R = R_1 \circ R_2 \circ \cdots \circ R_l$)的情况下节点间的关系.基于复合关系 R 给定元路径 $P = A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} A_{l+1}$,元路径 P 上的对象 A_1 与 A_{l+1} 之间的可达概率矩阵定义为 $\boldsymbol{Q}_P = \boldsymbol{X}_{A_1A_2}\boldsymbol{X}_{A_2A_3}\cdots\boldsymbol{X}_{A_lA_{l+1}}$,它表示 A_1 沿着路径 P 随机游走到 A_{l+1} 的概率.

根据 HeteSim 的定义,类型 L 中的节点基于元路径 $P = LDLD$ 到类型 D 中的节点之间的相似度为类型 L 的节点和类型 D 中的节点随机游走恰好在元路径中间类型 M 相遇的概率,计算公式为

$$HeteSim(L, D|P) = HeteSim(L, D|P_L P_R) = \boldsymbol{X}_{LD}\boldsymbol{X}_{DM}\boldsymbol{Y}_{ML}\boldsymbol{Y}_{LD} = \boldsymbol{X}_{LD}\boldsymbol{X}_{DM}\boldsymbol{X}_{LM}^{\mathrm{T}}\boldsymbol{X}_{DL}^{\mathrm{T}} = \boldsymbol{X}_{LD}\boldsymbol{X}_{DM}(\boldsymbol{X}_{DL}\boldsymbol{X}_{LM})^{\mathrm{T}} = \boldsymbol{Q}_{P_L}\boldsymbol{Q}_{P_R^{-1}}^{\mathrm{T}}, \quad (1)$$

式(1)表明 L 和 D 之间基于路径 P 的相关性是 2 个概率分布的内积.

对于 lncRNA 和疾病类型中具体的对象 l, d,基于路径 P 的关联得分计算为

$$HeteSim(l, d|P) = \boldsymbol{Q}_{P_L}(l,:)\boldsymbol{Q}_{P_R^{-1}}^{\mathrm{T}}(d,:), \quad (2)$$

其中 $\boldsymbol{Q}_P(l,:)$ 为矩阵 $\boldsymbol{Q}_P$ 中对象 l 所对应的行向量.

为了使得 HeteSim 得分取值位于区间[0,1],还需要对计算出的关联得分进行标准化处理:

$$HeteSim(l, d|P) = \frac{\boldsymbol{Q}_{P_L}(l,:)\boldsymbol{Q}_{P_R^{-1}}^{\mathrm{T}}(d,:)}{\sqrt{\|\boldsymbol{Q}_{P_L}(l,:)\|\|\boldsymbol{Q}_{P_R^{-1}}^{\mathrm{T}}(d,:)\|}}. \quad (3)$$

由式(1)~(3)我们就可以计算出 lncRNA 和疾病之间的关联得分.可以看到,计算 HeteSim 得分的过程主要包括 3 个部分:邻接矩阵标准化运算、矩阵连乘运算、相似度标准化运算.

2 实验结果与分析

2.1 实验数据

实验中所使用疾病与基因关联数据均来自文献[25],包括 lncRNA 与疾病关联数据以及已知的编码基因与疾病关联数据.lncRNA 与疾病关联数据包括 2 个部分:1)来自 LncRNADisease 数据库[13]的数据,其中包含 480 条实验验证的 lncRNA 与疾

病关联关系,涉及到 166 种疾病和 118 种 lncRNA;2)在 PubMed 上进行文本挖掘得到的 lncRNA 与疾病关联数据,其中包含 380 条 lncRNA-疾病关联的数据,包括 226 种 lncRNA 和 145 种疾病.

整合上述 2 种数据集,最终得到了 578 条 lncRNA-疾病关联关系,其中包括 295 种 lncRNA 和 214 种疾病,构成了 lncRNA-疾病异质信息网络.

编码基因与疾病关联数据来自 OMIM 数据库[26].针对上述 lncRNA-疾病关联数据中涉及到的 214 种疾病,其中 160 种疾病可通过 MIM 编号在 OMIM 数据库中找到该疾病的致病基因,Yang 等人[25]提取了 OMIM 数据库中这 160 种疾病与编码基因的关联关系,得到 980 条编码基因与疾病关联的数据条目,包括 801 个编码基因和 160 种疾病.

通过整合上述 lncRNA 与疾病关联数据、编码基因与疾病关联数据,得到 1 558 条编码-长非编码基因与疾病的关联关系,其中包括 214 种疾病和 1 096 种基因(编码基因或 lncRNA),根据以上数据构建基因-疾病异质信息网络.

上述 2 个网络中的具体信息如表 1 所示:

Table 1　Specific Information in the LncRNA/Gene-Disease Heterogeneous Information Network

表 1　lncRNA/基因-疾病异质信息网络中的具体信息

Heterogeneous Information Networks	Number of LncRNA/Coding Gene	Number of Disease	Number of Edges
LncRNA-Disease	295/0	214	578
Gene-Disease	295/801	214	1 558

2.2　性能分析

对基因-疾病异质信息网络中不存在连边的基因与疾病对,采用 HeteSim 算法计算疾病与基因之间的关联得分,预测潜在的 lncRNA 和疾病关联关系.对每一个疾病,选取关联得分在 top10 的基因认为是其潜在的致病基因.

HeteSim 在 lncRNA-疾病异质信息网络中的性能通过留一交叉验证(leave-one-out cross validation, LOOCV)实验来评估.由于二部网络中度为 1 的节点所关联边被移除后会成为孤立节点,不能通过网络方法和计算模型得到任何信息,因此本文的预测方法无法计算这些边的得分值.所以,在进行留一交叉验证之前应过滤这类边.最后,我们保留了 532 条边,其中包括 103 个疾病和 163 个基因(包括 44 个 lncRNA 和 119 个编码基因).对于保留的每一条关联关系中的疾病,我们在没有边相连的 lncRNA 中随机选取 1 个 lncRNA 与该疾病相连,构造本文实验的负样本.

在每次留一交叉验证运行过程中,我们删除 1 个已知的 lncRNA-疾病关联边,然后在剩下的网络中应用 HeteSim 算法计算出删除边的 HeteSim 关联得分.这个被删除的边被认为是测试样本,剩下的网络结构被认为是训练样本.通过设定不同的阈值(top k%,$1 \leqslant k \leqslant 100$),我们使用 ROC 曲线和 ROC 曲线下的区域(AUC)来评估 HeteSim 在网络上的表现.ROC 曲线的横轴是“假阳性率”(FPR),它是实际负样本中错误地识别为正样本的比例;纵轴是“真阳性率”(TPR),它是所有实际正样本中正确识别的正样本的比例.二者的计算公式为

$$FPR=\frac{FP}{FP+TN}, \tag{4}$$

$$TPR=\frac{TP}{TP+FN}. \tag{5}$$

TPR 表示的是移除的关联边排名在 k% 以内的比率;FPR 表示的是不存在的关联边排名在 k% 以内的比率.当阈值 k 在 1～100 之间变化时可以得到相应的 TPR 和 FPR.通过这种方式,可以绘制 ROC 曲线,从而计算 AUC.按照以上步骤,我们在 lncRNA-疾病异质信息网络上进行了留一交叉验证,并取得了 0.682 8 的 AUC.相应的 ROC 曲线如图 4 所示:

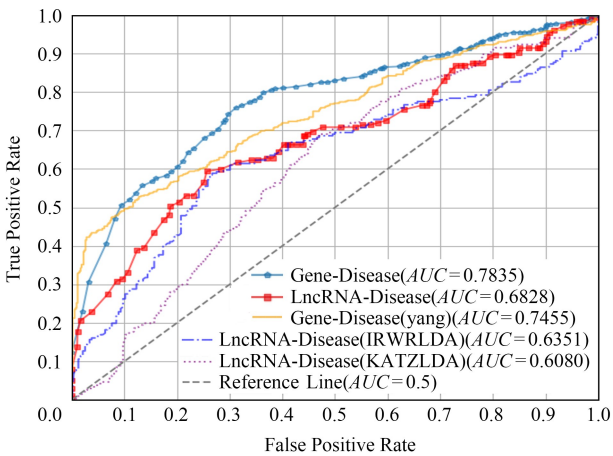


Fig. 4　Leave-one-out cross validation ROC curve

图 4　留一交叉验证 ROC 曲线图

为了提高方法的性能,我们将编码基因和疾病关联关系整合到 lncRNA-疾病网络中得到基因-疾病异质信息网络.我们在基因-疾病异质信息网络上进行了留一交叉验证,负样本的构造方法与之前类似,得到的 AUC 值为 0.783 5,如图 4 所示.很明显,

编码基因-疾病关联关系的整合可以提高我们方法的性能,分析原因主要是通过集成编码基因-疾病关联数据增加了网络中边的数量,使网络结构变得更紧密,潜在的基因可以从其他基因和疾病中获得更多信息传播,从而可以更好地进行预测.因此,在我们做链路预测相关方面研究时,通过整合多种数据,结合更有意义的语义信息,可以有效地提升预测的准确性.

在这里我们与 Yang 等人[25]提出的方法在相同的数据集上进行比较,这 2 种方法都是基于已知的基因与疾病之间的关联,不借助其他的信息进行疾病与基因的关联预测,图 4 给出了本文方法与 Yang 等人的方法预测结果比较,本文方法优于 Yang 等人的方法.此外.我们又与 IRWRLDA[21]和 KATZLDA[22]这 2 种方法进行比较,这 2 种方法除了已知的 lncRNA-疾病关联数据,还加入了 lncRNA 相似性和疾病相似性的数据来进行预测,本文的方法优于这 2 种方法,比较结果如图 4 所示.

2.3 案例分析

为进一步验证本文方法的可靠性和实用性,分别对卵巢癌和胃癌 2 种疾病做案例分析.对每一种疾病,所有未与该疾病有关联连边的基因按照其与该疾病的关联得分从大到小进行排序,排名 top10 的基因被认为是与该疾病潜在关联的基因.

卵巢恶性肿瘤是女性常见的恶性肿瘤之一,发病率仅次于子宫颈癌和子宫体癌.而卵巢上皮癌死亡率占各类妇科肿瘤的首位,对妇女的生命造成非常严重的威胁.表 2 显示了卵巢癌中排名 top10 的基因,包括 4 个 lncRNA,目前这 4 个已有文献通过生物实验等证实确实与该疾病有关,对应的 PubMed 唯一标识码(PubMed unique identifier, PMID)也在表 2 中给出,通过 PMID 可以在 PubMed 搜索引擎中查阅对应的文献.例如:Zhou 等人[27]通过研究发现 MALAT-1 在卵巢肿瘤中高表达,会促进卵巢癌细胞的生长和迁移,表明 MALAT-1 可能是卵巢癌发展的重要因素;Yang 等人[28]通过实验发现 UCA1 在上皮性卵巢癌组织和细胞中异常上调,研究表明 UCA1 是上皮性卵巢癌的新预后生物标志物;Xiu 等人[29]发现 MEG3 的表达在上皮性卵巢癌中较低,通过调节 ATG3 活性和诱导自噬在上皮性卵巢癌中充当肿瘤抑制剂,并可能被认为是卵巢癌的生物标志物;Zhang 等人[30]研究发现在患有卵巢癌的患者中,HOTAIR 显著上调.此外,HOTAIR 的上调增加了卵巢癌细胞的增殖、迁移和侵袭,从而促成了卵巢癌细胞的恶性进展.

Table 2 Top10 Genes Linked to Ovarian Cancer

表 2 Top10 与卵巢癌有关的基因

Gene	Rank	PMID	Gene	Rank	PMID
MALAT-1	1	27 227 769	HOTAIR	6	27 484 896
UCA1	2	26 867 765	CASP8	7	
KRAS	3		KCNQ1OT1	8	
MEG3	4	28 423 647	KLF6	9	
TP53	5		CHEK2	10	

胃癌是起源于胃黏膜上皮的恶性肿瘤,在我国各种恶性肿瘤中发病率居首位,对人类的健康造成巨大威胁.表 3 显示了胃癌中排名 top10 的基因,包括 5 个 lncRNA,其中有 3 个目前已有文献证实确实与该疾病有关.例如:Okugawa 等人[31]通过实验发现在腹膜播散的胃癌细胞中,HOTAIR 的 SiRNA 抑制细胞增殖、迁移和侵袭,为 HOTAIR 表达作为鉴定腹膜转移患者的潜在生物标志物的生物学和临床意义提供了新的证据,并且作为胃肿瘤患者的新治疗靶点;Chen 等人[32]通过实验发现 MALAT-1 在胃癌细胞系和组织中上调;此外,MALAT-1 在高转移潜能胃癌细胞系 SGC7901M 中的表达高于在低转移潜能胃癌细胞系 SGC7901NM 中的表达,结果表明 MALAT-1 可能部分通过调节上皮间质转化(EMT)促进胃癌细胞的迁移和侵袭;Xu 等人[33]通过实验证明 MEG3/miR21 通过调节 EMT 参与胃癌的肿瘤进展和转移.

Table 3 Top10 Genes Linked to Gastric Cancer

表 3 Top10 与胃癌有关的基因

Gene	Rank	PMID	Gene	Rank	PMID
MALAT-1	1	28 276 823	SLC22A1L	6	
FGFR3	2		HOTAIR	7	25 280 565
BRCA2	3		PRKN	8	
MEG3	4	29 749 532	BC043430	9	
CASP8	5		BC017743	10	

3 结 论

长非编码 RNA 在许多生物过程中具有重要的功能,这些长非编码 RNA 的变异或功能失调会导致一些复杂疾病的发生.因此,通过生物信息学方法预测潜在的长非编码 RNA-疾病关联关系,这对于

致病机理的探索以及疾病诊断、治疗、预后和预防都具有重要的意义.

近年来,针对这一问题,很多研究者已提出了其他基于网络的预测方法,并且在网络模型的基础上集成基因表达数据或者基因与 miRNA 之间的调控关系数据,实现 lncRNA 与疾病关联的预测.

本文使用了一种异质信息网络中的相关性计算方法——HeteSim,用来预测 lncRNA 与疾病之间的关联.该方法基于路径约束,通过元路径两端节点随机游走到中间节点相遇的概率作为疾病与 lncRNA 之间的关联得分,发掘潜在的疾病与 lncRNA 关联关系.实验结果表明该计算方法有较高的预测准确性和鲁棒性,并且该方法可以很好地集成其他类型的关联数据,例如基因间的蛋白质相互作用[34]、lncRNA 和编码基因的共表达、miRNA 对 lncRNA 和编码基因的调控、疾病之间的相似性信息等.集成这些关联数据,从而对元路径进行扩展,可以使更多与 lncRNA 疾病相关的语义信息被用来预测,有利于预测的准确性,这也是本文工作进一步深入研究的方向.

参　考　文　献

[1] Bertone P, Stolc V, Royce T E, et al. Global identification of human transcribed sequences with genome tiling arrays [J]. Science, 2004, 306(5705): 2242-2246

[2] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project [J]. Nature, 2007, 447(7146): 799-816

[3] Kapranov P, Cheng J, Dike S, et al. RNA maps reveal new RNA classes and a possible function for pervasive transcription [J]. Science, 2007, 316(5830): 1484-1488

[4] Carninci P, Kasukawa T, Katayama S, et al. The transcriptional landscape of the mammalian genome [J]. Science, 2005, 309(5740): 1559-1563

[5] Taft R J, Pang K C, Mercer T R, et al. Non-coding RNAs: Regulators of disease [J]. The Journal of Pathology, 2010, 220(2): 126-139

[6] Derrien T, Johnson R, Bussotti G, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression [J]. Genome Research, 2012, 22(9): 1775-1789

[7] Quek X C, Thomson D W, Maag J L, et al. lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs [J]. Nucleic Acids Research, 2015, 43(D1): 168-173

[8] Chakraborty S, Deb A, Maji R K, et al. LncRBase: An enriched resource for lncRNA information [J]. PloS One, 2014, 9(9): e108010

[9] Jiang Qinghua, Ma Rui, Wang Jixuan, et al. LncRNA2-Function: A comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data [J]. BMC Genomics, 2015, 16(Suppl 3): S2

[10] Jiang Qinghua, Wang Jixuan, Wu Xiaoliang, et al. LncRNA2Target: A database for differentially expressed genes after lncRNA knockdown or overexpression [J]. Nucleic Acids Research, 2014, 43(D1): D193-D196

[11] Guo Xingli, Gao Lin, Liao Qi, et al. Long non-coding RNAs function annotation: A global prediction method based on bi-colored networks [J]. Nucleic Acids Research, 2013, 41(2): e35

[12] Guo Xingli, Gao Lin, Wang Yu, et al. Advances in long noncoding RNAs: Identification, structure prediction and function annotation [J]. Briefings in Functional Genomics, 2015, 15(1): 38-46

[13] Chen Geng, Wang Ziyun, Wang Dongqing, et al. LncRNADisease: A database for long-non-coding RNA-associated diseases [J]. Nucleic Acids Research, 2012, 41(D1): D983-D986

[14] Ning Shangwei, Zhang Jizhou, Wang Peng, et al. Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers [J]. Nucleic Acids Research, 2015, 44(D1): D980-D985

[15] Zhao Tingting, Xu Jinyuan, Liu Ling, et al. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features [J]. Molecular BioSystems, 2014, 11(1): 126-136

[16] Chen Xing, Yan Guiying. Novel human lncRNA-disease association inference based on lncRNA expression profiles [J]. Bioinformatics, 2013, 29(20): 2617-2624

[17] Sun Jie, Shi Hongbo, Wang Zhenzhen, 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

[18] Liu Yongjing, Zhang Rui, Qiu Fujun, et al. Construction of a lncRNA-PCG bipartite network and identification of cancer-related lncRNAs: A case study in prostate cancer [J]. Molecular BioSystems, 2015, 11(2): 384-393

[19] Zhou Meng, Wang Xiaojun, Li Jiawei, et al. Prioritizing candidate disease-related long non-coding RNAs by walking on the heterogeneous lncRNA and disease network [J]. Molecular BioSystems, 2015, 11(3): 760-769

[20] Ganegoda G U, Li Min, Wang Weiping, et al. Heterogeneous network model to infer human disease-long intergenic non-coding RNA associations [J]. IEEE Transactions on Nanobioscience, 2015, 14(2): 175-183

[21] Chen Xing, You Zhuhong, Yan Guiying, et al. IRWRLDA: Improved random walk with restart for lncRNA-disease association prediction [J]. Oncotarget, 2016, 7(36): 57919

[22] Chen Xing. KATZLDA: KATZ measure for the lncRNA-disease association prediction [J]. Scientific Reports, 2015, 5(1): No.16840

[23] Zeng Xiangxiang, Liao Yuanlu, Liu Yuansheng, et al. Prediction and validation of disease genes using HeteSim Scores [J]. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2017, 14(3): 687-695

[24] Shi Chuan, Kong Xiangnan, Huang Yue, et al. HeteSim: A general framework for relevance measure in heterogeneous networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(10): 2479-2492

[25] Yang Xiaofei, Gao Lin, Guo Xingli, et al. A network based method for analysis of lncRNA-disease associations and prediction of lncRNAs implicated in diseases [J]. PloS One, 2014, 9(1): e87797

[26] Hamosh A, Scott A F, Amberger J S, et al. Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders [J]. Nucleic Acids Research, 2005, 33(Suppl 1): D514-D517

[27] Zhou Yanqing, Xu Xiaying, Lü Huabing, et al. The long noncoding RNA MALAT-1 is highly expressed in ovarian cancer and induces cell growth and migration [J]. PLoS One, 2016, 11(5): e0155250

[28] Yang Yijun, Jiang Yi, Wan Yicong, et al. UCA1 functions as a competing endogenous RNA to suppress epithelial ovarian cancer metastasis [J]. Tumor Biology, 2016, 37(8): 10633-10641

[29] Xiu Yinling, Sun Kaixuan, Chen Xi, et al. Upregulation of the lncRNA Meg3 induces autophagy to inhibit tumorigenesis and progression of epithelial ovarian carcinoma by regulating activity of ATG3 [J]. Oncotarget, 2017, 8(19): 31714-31725

[30] Zhang Zhongbao, Cheng Jiajing, Wu Yi, et al. LncRNA HOTAIR controls the expression of Rab22a by sponging miR-373 in ovarian cancer [J]. Molecular Medicine Reports, 2016, 14(3): 2465-2472

[31] Okugawa Y, Toiyama Y, Hur K, et al. Metastasis-associated long non-coding RNA drives gastric cancer development and promotes peritoneal metastasis [J]. Carcinogenesis, 2014, 35(12): 2731-2739

[32] Chen Di, Liu Lili, Wang Kai, et al. The role of MALAT-1 in the invasion and metastasis of gastric cancer [J]. Scandinavian Journal of Gastroenterology, 2017, 52(6/7): 790-796

[33] Xu Gang, Meng Lei, Yuan Dawei, et al. MEG3/miR 21 axis affects cell mobility by suppressing epithelial mesenchymal transition in gastric cancer [J]. Oncology Reports, 2018, 40(1): 39-48

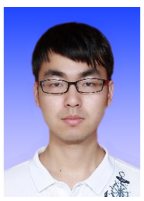
[34] Li Min, Meng Xiangmao. Progress in the construction, analysis and application of dynamic protein networks [J]. Journal of Computer Research and Development, 2017: 54(6): 1281-1299 (in Chinese)
(李敏, 孟祥茂. 动态蛋白质网络的构建, 分析及应用研究进展[J]. 计算机研究与发展, 2017, 54(6): 1281-1299)

Ma Yi, born in 1996. Master candidate. His main research interests include bioinformatics, data mining.

Guo Xingli, born in 1979. PhD, associate professor. Member of CCF. Her main research interests include data mining algorithm, complex network model and its application in bioinformatics research. (xlguo@mail.xidian.edu.cn)

Sun Yutong, born in 1995. Master candidate. Her main research interests include data mining, bioinformatics, long non-coding RNA analysis.

Yuan Qianqian, born in 1996. Master candidate. Her main research interests include data mining, bioinformatics, long non-coding RNA analysis.



Ren Yang, born in 1993. Master candidate. His main research interests include data mining, bioinformatics.

Duan Ran, born in 1990. PhD candidate. His main research interests include multi-omics data integration, cancer integrative analysis, and machine learning.

Gao Lin, born in 1964. PhD, professor and PhD supervisor. Her main research interests include bioinformatics, data mining, graph theory and combinatorial optimization algorithm and applications. (lgao@mail.xidian.edu.cn)