DOI:10.12113/202204003

基于多层基因网络的关键基因识别算法

魏丕静[1],刘晶晶[2],赵永敏[2],苏延森[3*],郑春厚[3]
(1.安徽大学 物质科学与信息技术研究院,合肥 230601;2.安徽大学 计算机科学与技术学院, 合肥 230601;
3.安徽大学 人工智能学院,合肥 230601)

摘 要:疾病关键基因可用于疾病诊断、预测和新药或新疗法有效性的评价,故识别与疾病紧密相关的关键基因十分重要。然而现在有些疾病样本数据较少,传统基于大样本的关键基因挖掘方法不适用于该类数据。本文针对含少量样本数据的疾病,首先利用单样本网络构建方法构建每个疾病样本的个体化基因网络,并通过建立基因间的层间联系构建多层基因网络。然后利用基于张量的多层网络中心性方法评估每层网络中基因间的相互作用以及层间影响,对基因进行重要性打分,识别疾病关键基因。最后将该方法应用到哮喘数据集上,并与经典算法进行比较,结果表明,利用该方法所识别的已获批准的药物靶标基因的排名较优;对所得到的新的潜在关键基因 *TP*53、*PUS*10、*MAP*3*K*1 等进行功能和通路富集分析,结果表明其与哮喘有紧密关联。

关键词:多层基因网络;随机游走;节点中心性;关键基因
中图分类号:Q343.1 **文献标志码**:A **文章编号**:1672-5565(2023)04-277-09

Key gene identification algorithm based on multi-layer network

WEI Pijing[1], LIU Jingjing[2], ZHAO Yongmin[2], SU Yansen[3*], ZHENG Chunhou[3]
(1. *Institutes of Physical Science and Information Technology*, *Anhui University*, *Hefei* 230601, *China*;
2. *School of Computer Science and Technology*, *Anhui University*, *Hefei* 230601, *China*;
3. *School of Artificial Intelligence*, *Anhui University*, *Hefei* 230601, *China*)

Abstract: Critical genes of diseases can be used to diagnose diseases, predict and evaluate the effectiveness of new drugs or new therapies, so it is very important to identify critical genes closely related to diseases. However, the samples of some diseases are limited. It is difficult to apply the traditional methods based on large sample data to mine critical genes for these diseases. In this paper, for diseases with small amount of samples, we first construct sample-specific network for each sample with the single-sample network constructing methods, and construct a multi-layer gene network by estabishing inter-layer connections between genes. A tensor-based multi-layer network centrality approach is then used to assess the interactions between genes in each layer of the network and the inter-layer effects to score the genes for importance and identify disease key genes. Finally, the method is used to two asthma datasets and compared with the classical algorithm. The results show that compared with other methods, the approved drug target genes rank higher in the gene rankings obtained by this method. And function and pathway enrichment analysis of the new potential critical genes *TP*53, *PUS*10, *MAP*3*K*1, etc. indicat that they were closely related to asthma.

Keywords: Multilayer gene interaction network; Random walk; Node centrality; Critical genes

疾病的关键基因是指在人体系统中与某种疾病密切相关的一组基因,其在人类生理过程和疾病发生过程中具有不可忽视的调控作用,了解关键基因在疾病中的功能和作用,对研究疾病的调控方式、复杂通路、治疗和预后等具有重要意义。关键基因可用于诊断疾病、判断疾病分期、预测和评价新药或新疗法的有效性等。药物靶向治疗的关键是药物能特异性作用于疾病相关基因位点,故识别与疾病紧密相关的关键基因十分重要。但由于基因数量庞大,仅通过生物实验的方法测定基因功能将会耗费巨大

收稿日期:2022-04-03;**修回日期**:2022-11-09;**网络首发日期**:2022-12-26.
网络首发地址:https://kns.cnki.net/kcms/detail23.1513.Q.2022.1223.1552.003.html
基金项目:国家重点研发计划项目(No.2021YFE0102100);安徽省自然科学基金青年项目(No.2108085QF267,No. 2008085QF294).
* **通信作者**:苏延森,女,教授,研究方向: 生物信息学,E-mail: suyansen@ ahu.edu.cn.
引用格式:魏丕静,刘晶晶,赵永敏,等.基于多层基因网络的关键基因识别算法[J].生物信息学,2023,21(4):277-285.
WEI Pijing, LIU Jingjing, ZHAO Yongmin, et al.Prediction of critical genes based on multi-layer network[J].Chinese Journal of Bioinformatics,2023,21(4):277-285.

的时间成本和经济成本。因此,基于计算模型识别疾病关键基因的预测算法亟待开发。目前有很多研究致力于发现疾病关键基因,此方面研究有助于探索人类复杂疾病的内部发病机制、研究疾病细胞存活所需的最小基因集和后续对疾病的治疗方式及治疗药物的研究[1]。

研究表明,基因并非独立的发挥生物作用,基因之间的相互作用普遍存在,并通过相互作用共同维持着生物内部整体环境的稳定性[2],故基于基因相互作用网络的基因排序技术得到了广泛的应用。Wang 等提出了在蛋白质-蛋白质相互作用网络上基于边缘聚类的关键蛋白识别方法,该方法认为节点的重要性由节点与相邻节点之间的相互作用的边缘系数与聚类系数之和决定[3]。Fan 等人提出了关键蛋白质预测方法,该方法将亚细胞室信息与基因表达信息相结合,并运用修改后的 PageRank 算法获得加权蛋白质-蛋白质相互作用网络,实验结果表明其有更好的关键蛋白质预测性能[4]。由此可以看出,将网络拓扑信息和生物学信息结合为研究关键基因提供了很好的思路。然而,虽然目前有多种技术可以用来识别疾病基因,但是大部分方法往往都是通过整合多个样本构建基因共表达网络,弱化了疾病样本与正常样本之间的差异信息,忽略了疾病样本的个体特异性。此外,个体特异性网络构建思想在揭示疾病的个体特征方面已经得到有效的验证[5-6]。

本文以基因间表达相似性为基础构建基因网络,并用来筛选有价值的生物标志物或关键基因,探索基因和疾病之间复杂关系。具体来说,首先利用正常样本的基因表达数据构建参考基因共表达网络,然后依次将每个疾病样本的基因表达数据与正常样本组合,构建疾病样本扰动网络,根据此扰动网络和参考网络,得到每一个疾病样本的个体特异性网络。然后将个体特异性网络作为单层网络,并将单层网络之间的基因联系起来,从而得到多层基因网络,这样既保留了疾病样本的特异性又将多个疾病样本联系在一起。最后,利用 Wu 等[7]提出的基于张量的多层网络中心性的计算方法,对多层网络中的基因节点中心性进行打分,从而得到关键基因集。与其他经典算法的对比分析表明该方法在预测药物靶标基因上具有一定的优势,功能和通路富集分析证明关键基因集与疾病联系紧密。

1 数据与方法

1.1 数据集

基因表达数据集来源于基因表达综合数据库 GEO(https://www.ncbi.nlm.nih.gov/geo/)。本文主要考虑样本量偏少的数据集,因此从 GEO 数据库中获取哮喘疾病的基因表达数据集 GSE31773 和 GSE43696。在哮喘疾病样本选取的过程中,由于 mRNA 在 CD8+T 细胞中的表达差异性大于在 CD4+T 细胞中,因此选择的疾病样本为 CD8+类型的。此外,根据控制变量的原则,尽量使得正常样本和异常样本的其他生物信息如年龄,性别等保持一致。因此,在 GSE31773 中选取了 8 个正常样本和 6 个疾病样本,每个样本包含 8 789 个基因。同理,在 GSE43696 中选取 20 个正常样本和 6 个疾病样本,每个样本包含9 194 个基因。

疾病相关的基因来源于 DisGeNet(https://www.disgenet.org/)和 Phenopedia(https://phgkb.cdc.gov/PHGKB/startPagePhenoPedia.action)数据库。从两个数据库中获取与哮喘相关的2 712 个基因,并与 GSE31773 和 GSE43696 数据集中的数据进行整合,分别得到2 522 个基因和2 478 个基因的表达数据。

此外,从 TTD(http://db.idrblab.net/ttd/)数据库获取 11 个针对哮喘已获批准的药物靶标。

1.2 方法

1.2.1 多层基因网络构建

多层基因网络构建主要分为四步,具体构建过程如图 1 所示。

第一步是获取疾病相关基因的表达数据。首先从 GEO 数据库获取正常样本和疾病样本的基因表达数据,从疾病基因相关数据库获取所有与所要研究的疾病潜在相关的基因,从正常样本和疾病样本的表达数据中筛选出疾病相关基因的表达数据。

第二步是利用所有正常样本构建参考基因网络[5]。设参考网络为 $G_{ref}(V,E,W)$,其中点集 V 是由与疾病相关的基因所构成,边集 E 表示基因对之间的边集,W 表示边权,即基因对间的皮尔逊相关系数,其计算方式如式(1)。

$$\omega(ij)=\frac{\left| n\sum C_{ik}C_{jk}-\sum C_{ik}\sum C_{jk}\right|}{\sqrt{n\sum C_{ik}^{2}-\left(\sum C_{ik}\right)^{2}}\sqrt{n\sum C_{jk}^{2}-\left(\sum C_{jk}\right)^{2}}} \tag{1}$$

其中,C_{ik} 表示基因 i 在第 k 个正常样本中的表达值,n 为基因节点的总数。

第三步是针对每个疾病样本构建个体特异性网络[5]。个体特异性网络的构建参考 Liu 等[5]提出的方法。具体而言,首先在所有正常样本的表达数据中加入一个疾病样本的表达数据,根据第二步的公式(1),求新的表达数据中基因之间的皮尔逊相关系数,

构建一个新的基因网络,并将其看作是加入该疾病样本后的扰动网络[5]。然后根据参考网络和扰动网络构建个体特异性网络,其中边权值为扰动网络和参考网络的边权值的差值绝对值[5]。接着利用拐点分析法设置阈值对网络中的边进行选择,删除一些不显著的边。这种方法考虑到了每个样本的个体特异性,体现了参考网络受到疾病样本的干扰程度,有效衡量了基因间相互作用关系与疾病的相关程度。

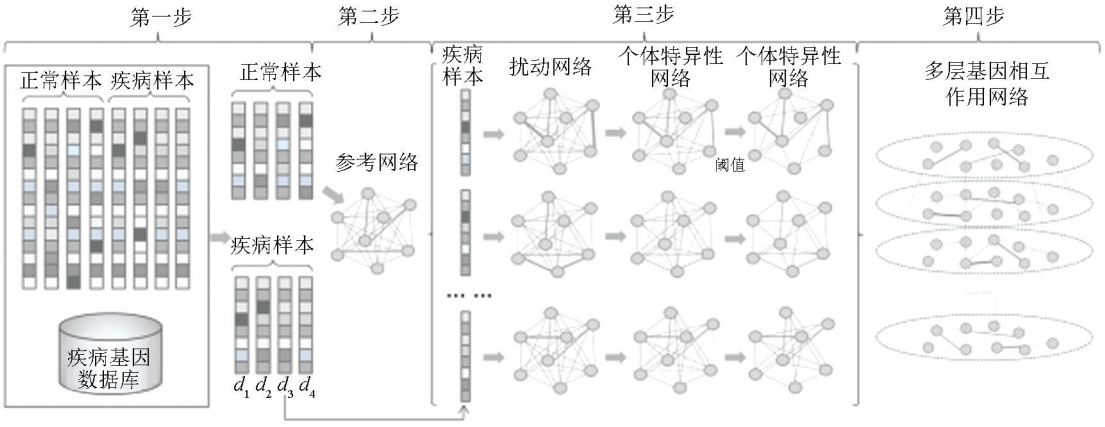


图 1　多层基因网络构建示意图

Fig.1　Schematic diagram of multilayer network construction

注:多层基因网络构建分成四步:第一步是数据收集,即在 GEO 数据库中选择正常样本和疾病样本,在疾病基因数据库中选择出与哮喘相关的基因;第二步是利用正常样本构建参考网络;第三步是构建疾病样本的个体特异性网络,首先利用疾病样本构建扰动网络,再用扰动网络减去参考网构建个体特异性网络,然后根据阈值去除部分异常边权值后得到最终的个体特异性网络。第四步是整合所有个体特异性网络得到多层基因相互作用网络,层间边连接每层网络中的相同基因,边的权值为 1.

第四步是整合单层基因网络得到多层基因网络。将得到的个体特异性网络作为多层基因网络的每一层,依次连接每两个单层网络中的相同节点构建层与层之间的边,边的权值为 1,得到多层复用基因网络。

1.2.2　基因节点中心性计算

参考 Wu 等[7]的方法,利用四阶邻接张量 E 表示多层网络的点和边,其中 $E_{j\beta}^{i\alpha}$ 表示第 α 层中的节点 i 到第 β 层中的节点 j 的加权边,用以存储网络的结构信息;利用二阶张量 W 表示层之间的影响,其中 $W_{\beta}^{iter,\alpha}$ 表示第 $iter$ 次迭代中第 α 层网络对第 β 层网络的影响,其计算方式为两层网络重要性之比,单层网络的重要性定义为第 $iter$ 次迭代中该层中的所有节点中心性均值。当 $W_{\beta}^{iter,\alpha}>1$ 时,表示第 α 层对第 β 层有增强效果;当 $W_{\beta}^{iter,\alpha}<1$ 时,表示第 α 层对第 β 层有削弱效果;当 $W_{\beta}^{iter,\alpha}=1$ 时,表示第 α 层对第 β 层无影响。最后整个网络之间的交互张量表示为 H,其中 $H_{j\beta}^{iter,i\alpha}$ 表示第 α 层中的节点 i 到第 β 层中的节点 j 的整体相互作用。根据随机游走的思想,H 的计算方式如式 2。

$$H_{j\beta}^{iter,i\alpha}=W_{\beta}^{iter,\alpha}\times\left(d\times E_{j\beta}^{'i\alpha}+(1-d)\times\frac{ones(size(E_{j\beta}^{'i\alpha}))}{M\times N}\right)\tag{2}$$

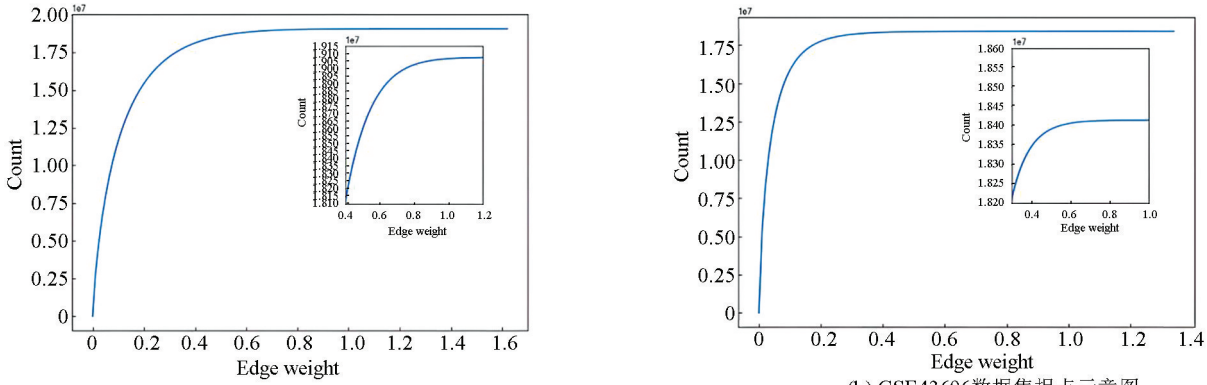
其中,E' 表示归一化后的邻接张量,$ones(size(E_{j\beta}^{'i\alpha}))$ 表示与 E' 维度相同的全“1”四阶张量,M 表示多层网络的层数,N 表示每层节点数,d 表示阻尼系数,一般取 0.85。

根据单层网络中 PageRank 算法的幂法求解过程,交互张量 H 相当于转移概率矩阵,求解张量方程 $H\Phi=\lambda\Phi$ 得到中心性二阶张量 Φ,其中 $\Phi_{i\alpha}$ 表示当前迭代中第 α 层的第 i 个基因节点的中心性值,λ 表示特征系数,λ 这里取值为 1,保证二阶张量 Φ 的存在性和唯一性。迭代结束后,将每个基因节点在所有层中的中心性均值作为该基因的最终中心性值,降序排序后选取排名靠前的基因作为关键基因,分值越高说明基因在疾病中发挥的作用越重要。

2　结果与讨论

2.1　网络边阈值的选取

通过设置皮尔逊相关系数的阈值得到多层网络。具体而言,针对 GSE31773 和 GSE43696 两个数据集,分别利用拐点分析法选择拐点,并将其作为筛选边的阈值。根据图 2 可以发现,GSE31773 数据集拐点示意图中,当边权值大于 1 时,趋势不再有明显上升,因此构建网络的阈值选择为 1。同理对于数据集 GSE43696 阈值选择为 0.6。确定数据集 GSE31773 和 GSE43696 构建 6 层网络的阈值分别为 1 和 0.6。



(a) GSE31773数据集拐点示意图　　(b) GSE43696数据集拐点示意图

图 2　拐点分析图

Fig.2　Analysis diagram of inflection point

注:图中小图表示拐点附近放大图.

2.2　多层基因网络有效性分析

利用本文的方法,针对两个独立数据集 GSE31773 和 GSE43696 分别构建多层网络,其信息如表 1 所示,其中层间边连接每层的相同基因,例如数据集 GSE43696,其中层间的边数是每层节点连接其他五层中相同节点,即总计 37 170 条边。以数据集 GSE43696 构建的多层网络为例,将其可视化后如图 3 所示。

表 1　多层网络信息

Table 1　Information of multilayer network

数据集	层数	单层节点数	层内边数	层间边数
GSE31773	6	2 522	[3 763,276,100,1 316,1 280,4 711]	37 830
GSE43696	6	2 478	[901,87,24,6 514,169,25]	37 170

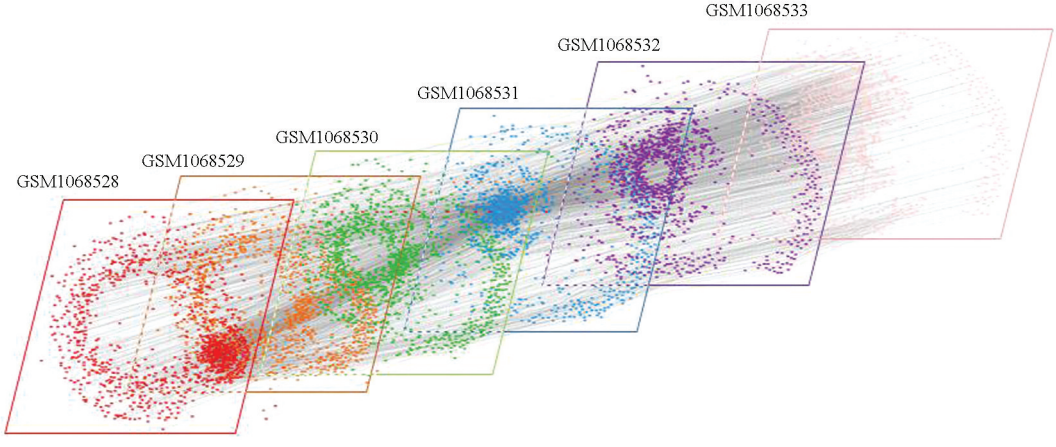


图 3　GSE43696: 6 层基因网络示意图

Fig.3　GSE43696: Diagram of 6-layers gene network

在多层网络构建过程中,多层基因网络的层数有多种选择。为了验证本文构建 6 层网络的有效性,本文在数据集 GSE31773 中随机选择不同数量的疾病样本构建了不同层数的多层基因网络,并对比已知哮喘药物靶标基因在不同层数的网络中的排名结果,如表 2 所示。其中在选择哮喘药物靶标基因时,首先选择有效治疗哮喘的药物,并在数据库中寻找药物关键基因靶标,最终选择包含在本文数据集中的 11 个靶标基因。表 2 中“排名 1”和“排名 2”指随机选择了两次相同数量样本的结果。从表中看出,在六层基因网络中,有 5-*LOX*、*IL*17、*CCR*4、*IL5RA*、*ROS* 等 5 个哮喘的药物靶标基因排名更靠前;在五层基因网络中,有 *H*1*R*、*IL*5、*JAK*-1 等 3 个基因排名更优,在四层基因网络中,只有基因 *CAMP* 有更好的排名;在三层基因网络中,有 2 个基因 *JAK*-2、*IL*4*R* 排名更优。综上所述,在识别关键基因集时,构建六层网络的效果更好。

表 2　哮喘靶标在不同层网络排名情况

Table 2　Ranking of asthma targets in different layers of networks

基因	6 层排名	5 层排名 1	5 层排名 2	4 层排名 1	4 层排名 2	3 层排名 1	3 层排名 2
5-*LOX*	**1 123**	1 391	1 168	1 225	1 226	1 633	1 633
JAK-2	1 155	154	204	191	183	135	135
H1R	1 296	**1 098**	1 379	1 426	1 463	1 222	1 222
IL4R	1 604	2 037	1 649	1 288	1 288	**1 256**	1 256
*IL*17	**818**	1 060	981	921	948	1 269	1 269
IL5	1 753	**1 071**	1 854	1 963	1 968	1 257	1 257
JAK-1	2 364	2 445	**2 349**	2 444	2 444	2 437	2 437
*CCR*4	**1 738**	2 131	1 850	1 967	1 963	2 437	2 437
CAMP	1 791	2 160	1 487	**1 420**	1 425	2 055	2 055
IL5RA	**804**	1 072	975	913	956	1 258	1 258
ROS	**1 995**	2 356	2 108	2 127	2 121	2 411	2 411

2.3　算法对比分析

根据哮喘基因数据集,利用本文提出的方法,可以得到哮喘相关的基因。为了进一步评估已知的疾病特异性通路或基因是否在预测的关键基因上具有显著的优先级,本研究利用 GSEA 软件的 GSEAPreranked 工具对结果进行分析。GSEA 富集分析主要是用来评估一个预先定义的基因集在与表型相关的基因排序列表中的分布趋势,它不需要进行基因过滤,输入数据主要包括两部分,一种是预先定义的基因集,一种是给定的基因排序列表。本文中,预先定义的基因集是 KEGG 通路数据库中的哮喘特异性相关的基因集,基因排序列表是本文预测的所有基因排序结果。通过 GSEA 富集分析揭示我们的模型结果和 KEGG 通路数据库中哮喘特异性相关的基因集之间的关联,以 GSE31773 数据集为例,根据其所有基因排名和 KEGG 通路数据库中哮喘特异性相关的基因集进行加权 K-S 检验得到 p 值,结果如图 4 所示。结果表明,与其他预测关键基因的排序方法 MI[8], t-Test[9], PCC[10], SCC, FC[11], NetRank[12], MarkRank[13] 相比,本研究中的算法在对疾病关键特异性基因进行优先排序时具有显著的p 值。

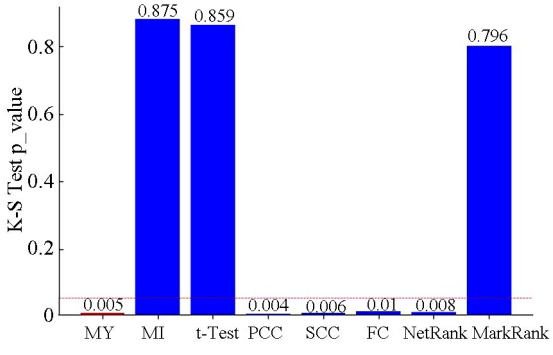


图 4　不同对比算法的哮喘通路富集分析

Fig.4　Enrichment analysis of asthma pathways with different comparison algorithms

2.4　哮喘关键基因集分析

为了验证本方法所识别的疾病相关基因的重要性,针对 GSE31773 和 GSE43696 两个数据集,分别选择排名前 10 的关键基因(见表 3),分析是否已有研究证实其为哮喘关键基因。研究发现 *TP*53、*MAP*3*K*1、*COL*18*A*1、*DACT*1、*CD*40*LG*、*ANKRD*55、*CD*4 以及 *TNFSF*18、*AFM*、*NKX*2-1、*SCGB*1*A*1、*RAG*1、*FRAS*1、*HSD*11*B*2、*GSTO*2、*SOAT*1、*IL*19 等基因在哮喘发生发展过程中起重要作用。例如,Yuan 等[14]的研究表明,与迟发性哮喘临床表型相关的 *TP*53 差异甲基化位点是早期筛选的有效生物标志物。Zhang 等的研究证明 *DACT*1 可能是治疗哮喘的潜在靶点[15]。对于 *CD*40*LG*,有研究表明 *CD*86 和 *CD*40*LG* 之间的相互作用会促进过敏性哮喘的发展[16]。*CD*4 *T* 细胞淋巴细胞活化在严重哮喘发病机制中起重要作用[17]。*SCGB*1*A*1 是肺重要的防御分子,防止 *SCGB*1*A*1 被抑制可有效的改善哮喘[18]。有研究表明 *GSTO*2 是哮喘易感基因,*GSTO*2 基因的多态性和哮喘有关[19]。此外,有研究证实,IL-19 基因在哮喘中高度表达,在变应性疾病中起着重要作用[20]。研究还发现,嗜酸性粒细胞的凋亡在支气管哮喘病理生理中发挥至关重要的作用[21],并且 *PUS*10 基因能够调节 *Trail* 诱导的细胞凋亡过程[22]。轴突或突触结构调控哮喘的激发机制[23],且 F5 蛋白在膜-细胞骨架相互作用和突触结构或功能的动态方面发挥重要作用[24]。此外,“下丘脑-垂体-肾上腺”轴功能与肺功能改善程度相关[25],并且 *CRHBP* 调节促肾上腺皮质激素控制“下丘脑-垂体-肾上腺”轴功能[26]。由此推断,*PUS*10、*F*5 及 *CRHBP* 等基因也与哮喘发生发展紧密相关。

2.5　GO 功能富集分析

为了分析本算法预测的疾病关键基因的功能相关性,利用本算法分别在两个独立数据集上选择排名前 100 的基因,使用基因功能分析工具 DAVID 对其作 GO 功能富集分析。基于 DAVID 分析工具,得

到与前 100 个基因显著相关的基因本体,图 5 展示了排名前 10 的基因本体。图的纵坐标展示了 GO 的功能注释,横坐标上的值表示 GO 在关键基因集中的富集显著性值-log(p)。由图可以发现,在排名前 10 的基因本体中,免疫反应、调控 T 细胞增殖、T 细胞刺激以及细胞因子活性均被证实与哮喘有密切联系[38]。具体来说,哮喘是由免疫系统对环境因子和不同的基因表达的联合反应引起的呼吸系统疾病。T 细胞是哮喘中过敏性气道炎症的关键介质[39],T 细胞的增殖会引起免疫球蛋白水平增加和支气管高反应性即哮喘发作,细胞因子也会辅助 T 细胞增殖的反馈控制。此外,炎症反应也与哮喘相关,在哮喘恶化过程中伴随着循环嗜酸性粒细胞、嗜碱性粒细胞及其前体细胞的变化等各种炎症反应[40-41]。除上述机制外还有几种潜在的新机制,例如药物反应,内皮细胞分化,蛋白质磷酸化调控,信号调控,应对缺氧,转录调控等在哮喘发展过程中都起着重要的作用。

表 3 排名前 10 的关键基因集

Table 3 Top 10 critical gene sets

GSE31773			GSE43696		
基因 ID	基因 Symbol	文献支持	基因 ID	基因 Symbol	文献支持
7157	*TP53*	[14]	8995	*TNFSF18*	[27]
150962	*PUS10*	---	173	*AFM*	[28]
4214	*MAP3K1*	[29]	7080	*NKX2-1*	[30]
2153	*F5*	---	7356	*SCGB1A1*	[18]
80781	*COL18A1*	[31]	5896	*RAG1*	[32]
51339	*DACT1*	[15]	80144	*FRAS1*	[33]
959	*CD40LG*	[16]	3291	*HSD11B2*	[34]
1393	*CRHBP*	---	119391	*GSTO2*	[19]
79722	*ANKRD55*	[35]	6646	*SOAT1*	[36]
920	*CD4*	[17]	29949	*IL19*	[37]

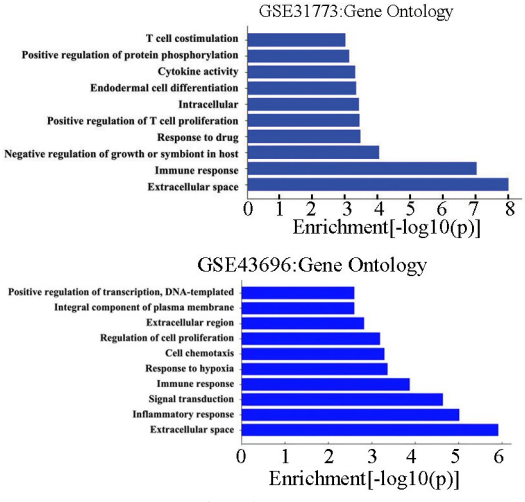


图 5 关键基因富集 Gene Ontology

Fig.5 Key gene enrichment Gene Qntology

2.6 通路富集分析

为了定位关键通路的关键基因,本文基于 DAVID 平台对两个独立数据集排名前 100 的基因进行通路富集分析,得到与 100 个基因显著相关($p\_val \leqslant 0.05$)的通路,表 4 和表 5 展示了显著相关的通路。由上述通路富集分析结果可知,细胞因子受体相互作用、趋化因子信号通路、T 细胞受体信号通路、原发性免疫不全四条通路都与哮喘紧密相关。肿瘤坏死因子(TNF)信号通路、TGF-beta 信号通路、Th1/Th2 分化等通路也被证明与哮喘有关。TNF 信号通路的坏死因子 α 是免疫和炎症反应的有效调节剂,可以引起包括哮喘在内的多种自身免疫性疾病[42]。哮喘会通过 TGF-beta 信号通路促进小鼠脉络膜血管新生[43]。T 淋巴细胞介导的对过敏原的免疫应答是哮喘发病机制的早期关键因素,而 Th1/Th2 平衡是哮喘发病机制的核心[44]。此外,还有若干个与哮喘潜在相关的通路,包括黏着连接、焦点粘连、鞘脂类信号通路等。

表 4 GSE31773:关键基因通路富集分析

Table 4 GSE31773: Pathways enrichment analysis of critical genes

ID	通路	*P*_value
hsa04060	Cytokine-cytoking receptor interaction	2.2×10^{-6}
hsa05202	Transcriptional misregulation in cancer	4.7×10^{-3}
hsa05219	Bladder cancer	6.6×10^{-3}
hsa04062	Chemokine signaling pathway	7.9×10^{-3}
hsa04350	TGF-beta signaling pathway	7.9×10^{-3}
hsa04550	Signaling pathway regulating pluripotency of stem cells	1.0×10^{-2}
hsa05161	Hepatitis B	1.1×10^{-2}
hsa05213	Endometrial cancer	1.2×10^{-2}
hsa04660	T cell receptor signaling pathway	1.4×10^{-2}
hsa05223	Non-small cell lung cancer	1.5×10^{-2}
hsa04668	TNF signaling pathway	1.8×10^{-2}
hsa05145	Toxoplasmosis	1.9×10^{-2}
hsa05210	Colorectal cancer	2.0×10^{-2}
hsa05214	Glioma	2.3×10^{-2}
hsa05212	Pancreatic cancer	2.3×10^{-2}
hsa04722	Neurotrophin signaling pathway	2.6×10^{-2}
hsa04071	Sphingolipid signaling pathway	2.6×10^{-2}
]hsa04917	Prolactin signaling pathway	2.9×10^{-2}
hsa05218	Melanoma	2.9×10^{-2}
hsa05520	Chronic myeloid leukemia	3.0×10^{-2}
hsa05216	Thyroid cancer	2.9×10^{-2}
hsa05166	HTLV-I infection	3.2×10^{-2}
hsa04380	Osteoclast differentiation	3.5×10^{-2}
h_41BBPathway	The 4-1BB-dependent immune response	3.7×10^{-2}
hsa05205	Proteoglycans in cancer	4.0×10^{-2}
hsa04510	Focal adhesion	4.4×10^{-2}
h_th1th2Pathway	Th1/Th2 Differentiation	4.5×10^{-2}
hsa05222	Small cell lung cancer	4.6×10^{-2}
hsa05330	Allograft rejection	4.7×10^{-2}

表 5　GSE43696:关键基因通路富集分析

Table 5　GSE43696: Pathways enrichment analysis of critical genes

ID	通路	*P*_value
hsa04060	Cytokine-cytokine receptor interaction	3.0×10^{-3}
hsa04520	Adherens junction	1.9×10^{-2}
hsa04068	Fox0 signaling pathway	2.1×10^{-2}
hsa05340	Primary immunodeficiency	3.0×10^{-2}
h_cytokinePathway	Cytokine Network	4.0×10^{-2}

3　结　论

1)复杂疾病的发生发展本质上与基因和生物功能过程的改变密切相关,疾病关键基因的识别对于研究疾病机理尤其是药物靶向治疗具有重要意义。哮喘作为全球范围内发病率最高的慢性呼吸道疾病之一,其发病率在逐年上升。识别出与哮喘成因紧密相关的基因有助于提高治疗效果。然而临床研究中由于疾病样本数较少,通常导致疾病相关基因识别困难。针对上述问题,本研究提出基于少数样本构建多层网络,进而利用多层网络随机游走识别疾病相关的关键基因的方法。该方法有助于挖掘样本数量受限条件下的疾病相关基因,加深对疾病致病机理的理解。

2)构建的多层网络对识别小样本疾病的致病基因可行且有效。本文利用皮尔逊相关系数计算出每条边的权值;为增强网络结构的稳定性,采用拐点分析法寻找最佳阈值,保留扰动程度较大的边;通过比较对已知疾病关键基因的排序选取最优的网络层数。例如针对数据集 GSE31773 的实验分析表明,构建六层基因网络效果最佳。

3)与其他方法相比,本算法识别的哮喘相关基因的排名更具显著性。利用本算法分别在 GSE31773 和 GSE43696 数据集中挖掘排名前 10 的关键基因,研究发现 *TP*53、*MAP*3*K*1、*COL*18*A*1、*DACT*1、*CD*40*LG*、*ANKRD*55、*CD*4 以及 *TNFSF*18、*AFM*、*NKX*2-1、*SCGB*1*A*1、*RAG*1、*FRAS*1、*HSD*11*B*2、*GSTO*2、*SOAT*1、*IL*19 等基因在哮喘发生发展过程中起重要作用,并推断 *PUS*10、*F*5 及 *CRHBP* 等基因也与哮喘发生发展紧密相关。

4)对分别从 GSE31773 和 GSE43696 两个数据集中所得排名前 100 的关键基因进行通路富集分析和 GO 功能富集分析,分析结果表明所识别的基因能够显著富集到与哮喘相关的通路和功能中。

参考文献(References)

[1]LI Xingyi, LI Wenkai, ZENG Min, et al. Network-based methods for predicting essential genes or proteins: A survey [J]. Briefings in Bioinformatics, 2020, 21(2): 566-583. DOI: 10.1093/bib/bbz017.

[2]KOONIN E V, WOLF Y I, KAREV G P. Power laws, scale-free networks and genome biology [M]. Berlin: Springer, 2006.

[3]WANG Jianxin, LI Min, WANG Huan, et al. Identification of essential proteins based on edge clustering coefficient[J]. IEEE/ACM transactions on Computational Biology and Bioinformatics, 2011, 9(4): 1070-1080. DOI: 10.1109/TCBB.2011.147.

[4]FAN Yetian, HU Xiaohua, TANG Xiwei, et al. A novel algorithm for identifying essential proteins by integrating subcellular localization[C].//in 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016.

[5]LIU Xiaoping, WANG Yuetong, JI Hongbin, et al. Personalized characterization of diseases using sample-specific networks[J].Nucleic Acids Research, 2016, 44(22): e164-e164. DOI: 10.1093/nar/gkw772.

[6]LIU Rui, YU Xiangtian, LIU Xiaoping, et al. Identifying critical transitions of complex diseases based on a single sample[J].Bioinformatics, 2014, 30(11): 1579-1586. DOI: 10.1093/bioinformatics/btu084.

[7]WU Mincheng, HE Shibo, ZHANG Yongtao, et al. A tensor-based framework for studying eigenvector multicentrality in multilayer networks[J].Proceedings of the National Academy of Sciences, 2019, 116(31): 15407-15413. DOI: 10.1073/pnas.1801378116.

[8]张焕萍, 王惠南, 卢光明, 等. 基于互信息的差异共表达致病基因挖掘方法[J].东南大学学报: 自然科学版, 2009, 39(1): 151-155. DOI: 10.3969/j.issn.1001-0505.2009.01.029.
ZHANG Huanping, WANG Huinan, LU Guangming, et al. Finding differentially co-expressed disease-related genes based on mutual information[J]. Journal of Southeast University (Natural Science Edition), 2009, 39(1): 151-155. DOI: 10.3969/j.issn.1001-0505.2009.01.029.

[9]JAFARI P,AZUAJE F. An assessment of recently published gene expression data analyses: Reporting experimental design and statistical factors[J].BMC Medical Informatics and Decision Making, 2006, 6(1): 1-8. DOI: 10.1186/1472-6947-6-27.

[10]HUANG Hungchung, ZHENG Siyuan, ZHAO Zhongming. Application of Pearson correlation coefficient (PCC) and Kolmogorov-Smirnov distance (KSD) metrics to identify disease-specific biomarker genes[J].BMC Bioinformatics, 2010, 11(Suppl 4):23. DOI: 10.1186/1471-2105-11-S4-P23.

[11]DEMBELE D, KASTNER P. Fold change rank ordering statistics: A new method for detecting differentially expressed genes[J].BMC Bioinformatics, 2014, 15(1): 1-15. DOI: 10.1186/1471-2105-15-14.

[12]WINTER C, KRISTIANSEN G, KERSTING S, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes[J]. Plos Computational Biology, 2012, 8(5): e1002511. DOI: 10.1371/journal.pcbi.1002511.

[13]SUN D, REN X, ARI E, et al. Discovering cooperative biomarkers for heterogeneous complex disease diagnoses[J]. Briefings in Bioinformatics, 2019, 20(1): 89-101. DOI: 10.1093/bib/bbx090.

[14]YUAN Lin, WANG Leyuan, DU Xizi, et al. The DNA methylation of FOXO3 and TP53 as a blood biomarker of late-onset asthma[J]. Journal of Translational Medicine, 2020, 18(1):467. DOI: 10.1186/s12967-020-02643-y.

[15]ZHANG Cunxue, YANG Peili, CHEN Yan, et al. Expression of DACT1 in children with asthma and its regulation mechanism[J]. Experimental and Therapeutic Medicine, 2018, 15(3): 2674-2680. DOI: 10.3892/etm.2018.5706.

[16]LEE S H, LEE E B, SHIN E S, et al. The interaction between allelic variants of CD86 and CD40LG: A common risk factor of allergic asthma and rheumatoid arthritis[J]. Allergy, Asthma & Immunology Research, 2014, 6(2): 137-141. DOI: 10.4168/aair.2014.6.2.137.

[17]CORRIGAN C, KAY A. CD4 T lymphocyte activation in acute severe asthma[J].International Archives of Allergy and Immunology, 1991, 94(1-4): 270-271. DOI: 10.1164/ajrccm/141.4_Pt_1.

[18]WARREN R, O'REILLY M A. An elusive fox that suppresses Scgb1a1 in asthma has been found[J]. American Journal of Respiratory cell and Molecular Biology, 2019, 60(6):615-617.DOI:10.1165/rcmb.2019-0019ED.

[19]PIACENTINI S, VERROTTI A, POLIMANTI R, et al. Functional polymorphisms of GSTA1 and GSTO2 genes associated with asthma in Italian children[J].Clinical Chemistry and Laboratory Medicine, 2012, 50(2): 311-315. DOI: 10.1515/cclm.2011.774.

[20]HUANG F, WACHI S, THAI P, et al. Potentiation of IL-19 expression in airway epithelia by IL-17A and IL-4/IL-13: Important implications in asthma[J].Journal of Allergy and Clinical Immunology, 2008, 121(6): 1415-1421. DOI: 10.1016/j.jaci.2008.04.016.

[21]包海鹏, 阎玥, 史琦, 等. 嗜酸性粒细胞凋亡在支气管哮喘作用中的研究进展[J].中华中医药学刊, 2019, 37(5):1095-1098. DOI: 10.13193 /j.issn.1673-7717.2019.05.018.

BAO Haipeng, YAN Yue, SHI Qi, et al., Research progress on role of eosinophil apoptosis in bronchial asthma [J], Chinese Archives of Traditional Chinese Medicine, 2019, 37(5): 1095-1098. DOI: 10.13193 /j.issn.1673-7717.2019.05.018.

[22]SONG Jinghui, ZHUANG Yuan, ZHU Chenxu, et al. Differential roles of human PUS10 in miRNA processing and tRNA pseudouridylation [J]. Nature Chemical Biology, 2020, 16(2): 160-169. DOI: 10.1038/s41589-019-0420-5.

[23]李慧敏, 王健, 卢银忠, 等. PM 2.5 对中枢神经系统功能及脂质代谢紊乱的影响[J].环境与职业医学, 2020, 37(10):1022-1029. DOI:10.13213/j.cnki.jeom.2020.20146.

LI Huimin, WANG Jian, LU Yinzhong, et al., Effects of PM2.5 on function of central nervous system and lipid metabolism disorder[J]. Journal of Environmental and Occupational Medicine, 2020, 37(10): 1022-1029. DOI: 10.13213/j.cnki.jeom.2020.20146.

[24]ARAI M, COHEN J. Subcellular localization of the F5 protein to the neuronal membrane-associated cytoskeleton[J]. Journal of Neuroscience Research, 1994, 38(3): 348-357. DOI: 10.1002/jnr.490380313.

[25]陈伟林, 张允健, 林小燕. 雾化吸入布地奈德对哮喘患者下丘脑-垂体-肾上腺轴功能的影响[J].海南医学, 2018, 29(10): 3. DOI: 10.3969/j.issn.1003-6350.2018.10.033.

CHEN Weilin, ZHANG Yunjian, LIN Xiaoyan, Effect of nebulized budesonide inhalation on hypothalamic-pituitary-adrenal axis function in asthmatic patients[J]. Hainan Medical Journal, 2018, 29(10): 3. DOI: 10.3969/j.issn.1003-6350.2018.10.033.

[26]ROY A, HODGKINSON C A, DELUCA V, et al. Two HPA axis genes, CRHBP and FKBP5, interact with childhood trauma to increase the risk for suicidal behavior[J]. Journal of Psychiatric Research, 2012, 46(1): 72-79. DOI: 10.1016/j.jpsychires.2011.09.009.

[27]OGASAWARA N, POPOSKI J A, KLINGLER A I, et al. Role of TNFSF11 and group 2 innate lymphoid cells in type 2 inflammation in chronic rhinosinusitis with nasal polyps [J]. Journal of Allergy and Clinical Immunology, 2018, 141(2): AB1. DOI: 10.1016/j.jaci.2017.12.002.

[28]ZEMŁA J, STACHURA T, GROSS-SONDEJ I, et al. AFM-based nanomechanical characterization of bronchoscopic samples in asthma patients[J].Journal of Molecular Recognition, 2018, 31(12): e2752. DOI: 10.1002/jmr.2752.

[29]SZCZEPANKIEWICZ A, SOBKOWIAK P, RACHEL M, et al. Multilocus analysis of candidate genes involved in neurogenic inflammation in pediatric asthma and related phenotypes: a case-control study[J]. Journal of Asthma, 2012, 49(4): 329-335. DOI: 10.3109/02770903.2012.669442.

[30]GRAS D, JONARD L, ROZE E, et al. Benign hereditary chorea: phenotype, prognosis, therapeutic outcome and long term follow-up in a large series with new mutations in the TITF1/NKX2-1 gene[J].Journal of Neurology, Neuro-

surgery & Psychiatry, 2012, 83(10): 956-962. DOI: 10.1136/jnnp-2012-302505.

[31] CASTRO-GINER F, BUSTAMANTE M, RAMON GONZÁLEZ J, et al. A pooling-based genome-wide analysis identifies new potential candidate genes for atopy in the European Community Respiratory Health Survey (ECRHS)[J].BMC Medical Genetics, 2009, 10(1): 1-9. DOI: 10.1186/1471-2350-10-128.

[32] CHRISTIANSON C A, GOPLEN N P, ZAFAR I, et al. Persistence of asthma requires multiple feedback circuits involving type 2 innate lymphoid cells and IL-33[J].Journal of Allergy and Clinical Immunology, 2015, 136(1): 59-68. DOI: 10.1016/j.jaci.2014.11.037.

[33] WINKLER C, HOCHDöRFER T, ISRAELSSON E, et al. Activation of group 2 innate lymphoid cells after allergen challenge in asthmatic patients[J].Journal of Allergy and Clinical Immunology, 2019, 144(1): 61-69. DOI: 10.1016/j.jaci.2014.11.037.

[34] JAHNKE J R, TERáN E, MURGUEITIO F, et al. Maternal stress, placental 11β-hydroxysteroid dehydrogenase type 2, and infant HPA axis development in humans: Psychosocial and physiological pathways[J].Placenta, 2021, 104: 179-187. DOI: 10.1016/j.placenta.2020.12.008.

[35] CLARK A D,NAIR N, ANDERSON A E, et al. Lymphocyte DNA methylation mediates genetic risk at shared immune-mediated disease loci[J].Journal of Allergy and Clinical Immunology, 2020, 145(5): 1438-1451. DOI: 10.1016/j.jaci.2019.12.910.

[36] POLIMANTI R, PIACENTINI S, MOSCATELLI B, et al. GSTA1, GSTO1 and GSTO2 gene polymorphisms in Italian asthma patients[J].Clinical and Experimental Pharmacology and Physiology, 2010, 37(8): 870-872. DOI: 10.1111/j.1440-1681.2010.05385.x.

[37] LIAO S C, CHENG Y C, WANG Y C, et al. IL-19 induced Th2 cytokines and was up-regulated in asthma patients[J].The Journal of Immunology, 2004, 173(11): 6712-6718. DOI: 10.4049/jimmunol.173.11.6712.

[38] QUESADA D. A mathematical model for T-cell differentiation in Asthma episodes[C].//in APS March Meeting Abstracts, 2005. DOI: 10.1016/j.immuni.2015.11.004.

[39] MEDOFF B D, THOMAS S Y,LUSTER A D. T cell trafficking in allergic asthma: The ins and outs[J].Annual Review of Immunology, 2008, 26: 205-232. DOI: 10.1146/annurev.immunol.26.021607.090312.

[40] GIBSON P, DOLOVICH J, GIRGIS-GABARDO A, et al. The inflammatory response in asthma exacerbation: Changes in circulating eosinophils, basophils and their progenitors[J].Clinical & Experimental Allergy, 1990, 20(6): 661-668. DOI: 10.1111/j.1365-2222.1990.tb02705.x.

[41] BORISH L. The inflammatory theory of asthma[J].Immunological Investigations, 1987, 16(6): 501-532. DOI: 10.3109/08820138709087099.

[42] ALBUQUERQUE R, HAYDEN C, PALMER L, et al. Association of polymorphisms within the tumour necrosis factor (TNF) genes and childhood asthma[J].Clinical and Experimental Allergy: Journal of the British Society for Allergy and Clinical Immunology, 1998, 28(5): 578-584. DOI: 10.1046/j.1365-2222.1998.00273.x

[43] YANG F,SUN Y, BAI Y, et al. Asthma promotes choroidal neovascularization via the transforming growth factor beta1/smad signalling pathway in a mouse model[J].Ophthalmic Research, 2022, 65(1): 14-29. DOI: 10.1159/000510778.

[44] LIAO X, TANG S, THRASHER J B, et al. Small-interfering RNA-induced androgen receptor silencing leads to apoptotic cell death in prostate cancer[J].Molecular Cancer Therapeutics, 2005, 4(4): 505-515. DOI: 10.1158/1535-7163.