DOI:10.3969/j.issn.1005-202X.2022.09.019

医学生物信息

基于改进TF-IDF算法的基因通路富集方法

徐淑坦[1,2],冷银辉[1,2],陈明[1,2]
1.上海海洋大学信息学院,上海 201306;2.农业农村部渔业信息重点实验室,上海 201306

【摘要】提出一种综合考虑通路局部和全局信息的基因通路富集分析(GIGSEA)方法。首先利用基因相互作用数据,通过基因在通路的局部重要性和在通路数据库的全局特异性计算基因的影响力;然后将基因影响力和表型相关性值融合在一起,计算通路的富集分数;最后通过置换基因富集出统计学显著的通路。将GIGSEA方法运用于肝细胞癌和结肠直肠癌数据集进行风险通路富集,与基因集富集分析方法相比,GIGSEA方法能富集出一些新的相关通路,并排除无关的通路,提高疾病相关通路的富集效果。

【关键词】通路富集;基因影响力;基因集富集分析
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2022)09-1173-09

A gene pathway enrichment method based on improved TF-IDF algorithm

XU Shutan[1,2], LENG Yinhui[1,2], CHEN Ming[1,2]
1. College of Information Technology, Shanghai Ocean University, Shanghai 201306, China; 2. Key Laboratory of Fisheries Information, Ministry of Agriculture and Rural Affairs of the People's Republic of China, Shanghai 201306, China

Abstract: A gene pathway enrichment method (GIGSEA method) that comprehensively considers the local and global information of the pathways is proposed. The gene interaction data are used to calculate the gene impact based on the local importance of the gene in the pathway and its global specificity in the pathway database. Then the obtained gene impact is fused with the phenotypic correlation value to calculate enrichment score, and statistically significant pathways are identified by permutating gene. GIGSEA is applied to the data sets of hepatocellular carcinoma and colorectal cancer for the enrichment of risk pathways. Compared with the gene set enrichment analysis method, GIGSEA method can enrich some new related pathways and exclude irrelevant pathways, which improves the enrichment effect of disease-associated pathways.

Keywords: pathway enrichment; genet impact; gene set enrichment analysis

前 言

在生物医学相关研究领域,随着高通量测序技术的发展,组学数据的规模也呈指数级增长。从庞大的组学数据中,可以利用生物信息学技术挖掘与疾病发生机制相关的通路,对疾病的诊断和治疗具有重要意义。

在过去10多年,已经开发了很多基因功能富集分析方法来识别各种疾病相关的通路[1-3]。基于数据来源和算法大致可以将基因功能富集分析方法分为4大类:过代表分析(Over-Representation Analysis, ORA)、功能集打分(Functional Class Scoring, FCS)、基于通路拓扑结构(Pathway Topology, PT)和基于网络拓扑结构(Network Topology, NT)的方法[3]。ORA方法首先选定一组感兴趣的基因作为基因列表,然后对该基因列表与通路中的基因集做交集,找出它们共同的基因并进行计数,最后利用统计检验的方式来评估计数值是否显著高于随机,即待测通路在基因列表中是否显著富集。基因集富集分析(Gene Set Enrichment Analysis, GSEA)属于FCS类方法,是最具代表性的基因功能富集分析方法,该方法利用基因表达和表型数据对所有基因进行排序,然后计算基因集Kolmogorov-Smirnov(KS)统计量,即在基因排序列表中靠近两端程度的得分,最后通过置换方法评估基因集的显著性[4]。FCS类方法将通路中的基因视作独立个体,实际上通路中的基因通过复杂的相互作用来影响细胞的发育、分化或疾病等生

【收稿日期】2022-04-05
【基金项目】广东省重点领域研发计划(2021B0202070001)
【作者简介】徐淑坦,博士后,副教授,研究方向:生物信息,E-mail: stxu@shou.edu.cn
【通信作者】陈明,博士,教授,研究方向:生物信息,E-mail: mchen@shou.edu.cn

物学过程[5]。之后一些基于PT的方法开始考虑通路的信息，Liu等[6]提出一种基于定向随机游走的方法来推断通路活性，即利用基因在定向通路中的结构信息来评估每个基因在通路中的重要性，然后使用重要性对基因加权进行分析。Deng等[7]利用蛋白质互作数据和通路的基因集构建了蛋白质-通路交互网络，从全局层面优化富集分析。Yang等[8]提出一种基于PT的通路富集方法，该方法根据基因节点的全局上游或下游位置和通路中的连接度数来评估节点的重要性，富集出那些拥有更多在上游或中枢节点中的基因的通路。后来的基于NT的方法中，很多方法也借鉴了GSEA的思想。Winterhalter等[9]提出基于蛋白质互作网络的GSEA方法JEPETTO，利用网络的拓扑结构信息分析基因和通路之间的功能关联。Rahmati等[10]整合来自20个通路数据库的核心通路，构建了庞大的蛋白质互作网络，以求富集出与生物学功能相关的通路，提出pathDIP方法。Han等[11]提出基于人类基因功能网络的GSEA方法NGSEA，该方法衡量基因集的富集分数，不仅考虑单个基因的表达差异，还考虑功能网络中它们的相邻基因的表达差异。Yoon等[12]提出一种新的网络加权的GSEA，对基因集进行富集分析时，将重叠基因和蛋白质互作网络结合起来，有效地识别出与基因功能相关的基因集。Zito等[13]将图论中衡量结点的中介中心性权重引入通路富集分析，利用蛋白质互作网络，提出网络节点中心性加权的GSEA。从更高层面看，通路是生物互作网络的一部分，因此通路富集分析有必要综合考虑基因在通路的局部信息和在通路数据库的全局信息，这一点和数据挖掘领域的常用加权算法TF-IDF（Term Frequency-Inverse Document Frequency）的思想类似，该算法综合考虑字词在文件的局部重要性和文件库的全局特异性两方面来评估字词的权重。

针对以上问题，本研究融合基因在通路的局部重要性和在通路数据库的全局特异性定义基因影响力，提出一种基于改进TF-IDF算法的基因通路富集方法（GIGSEA方法）。首先利用基因相互作用数据来计算通路基因的影响力，然后通过基因表达数据和表型数据来计算基因与表型的相关性值，并利用基因影响力和相关性值计算富集分数，最后通过统计学方法计算通路的显著性P值。在肝细胞癌（Hepatocellular Carcinoma, HCC）和结肠直肠癌（Colorectal Cancer, CRC）数据集的实验结果表明该方法能有效识别出与疾病相关的通路，对于今后研究疾病的发生发展机制有重要的指导意义。

1 材料与方法

1.1 数据集及预处理

本研究的通路数据来自KEGG数据库[14]，KEGG数据库是代谢组学和蛋白质组学研究中常用的生化通路数据库，从GSEA网站下载通路数据集，共包括186个通路，共有5 245个基因。

本研究的基因相互作用数据从STRING数据库[15]中获得，下载地址为https://www.string-db.org/cgi/download，下载的数据是最新的11版本的人类蛋白互作数据，共11 938 498对，从中提取包含KEGG通路基因的基因相互作用，去除重复的相互作用对，共计977 800对。

本研究的基因表达数据包括HCC和CRC数据集。HCC数据集来自TCGA数据库，包括来自肝癌患者的肿瘤邻近组织的50对配对样本和324个肿瘤样本，本研究只使用50对配对样本进行测试，即包括50个肿瘤样本和50个正常样本，该数据集下载自UCSC Xena网站（https://xena.ucsc.edu/）[16]。CRC数据集来自GEO（Gene Expression Omnibus）数据库的GSE8671表达谱数据，包括32个结肠癌腺瘤组织和32个配对的癌旁组织。

1.2 方法

1.2.1 基于改进TF-IDF算法计算基因影响力 在TF-IDF算法中，TF是词频（Term Frequency），即词条t在文件中出现的频率，TF越大，意味着t可能是文件的关键词，说明t越重要；IDF是逆文件频率指数（Inverse Document Frequency），如果包含词条t的文件越少，IDF越大，说明t具有很好的类别区分能力，t的权重相应地越大。TF是对词条t在本文件的评估，IDF是对t在文件库的评估。

类似地，本研究的基因影响力由基因在本通路的局部重要性和在通路数据库的全局特异性来衡量。通路基因的相互作用的关系可以抽象为一个图，本研究定义基因的重要性为在通路中与其他基因产生相互作用的数量，即在通路图中的连接度。采用基因频率GF（Gene Frequency）来表示基因重要性，GF表示为：

$$\mathrm{GF}_{i,j}=\frac{n_{i,j}}{\sum_{k\in P}n_{i,k}} \tag{1}$$

其中，$n_{i,j}$是基因j在通路p_i中的度数，分母则是通路p_i中所有基因的度总和。

一个基因的特异性表现为基因在所有通路中出现的频度，频繁出现在很多通路中的基因，它们对通路的影响相对较小；仅在少数通路中出现的基因其特异性高，它们的差异表达对通路的影响就大。本

研究定义逆通路频率(Inverse Pathway Frequency, IPF)来表示基因特异性,IPF表示为:

$$\mathrm{IPF}_i = \log\left(\frac{|P|}{1+|i:g_j \in p_i|}\right) \quad (2)$$

其中,$|P|$是数据库中的基因通路的总数;$|i:g_j \in p_i|$表示包含基因g_j的通路数目,即$n_{i,j} \neq 0$的通路数目。

某基因在特定通路内的重要性越高,并且在通路数据库中的特异性越高,那么该基因影响力越大。合并GF和IPF计算基因影响力GI($g_{i,j}$),表示为:

$$\mathrm{GI}(g_{i,j}) = \mathrm{GF}_{i,j} * \mathrm{IPF}_j \quad (3)$$

其中,GI($g_{i,j}$)表示基因j对通路i的影响力大小。

1.2.2 GIGSEA方法　假定有N个基因、T个样本的基因表达数据,T个样本包括两种表型pos和neg,样本数量分别为$t1$和$t2$;给定一个通路p_i,包括Q个基因。GIGSEA方法主要过程如下:

(1)利用基因相互作用数据统计每个基因在通路中的连接度数量,计算出GF(图1a)。然后统计每个基因在全部通路中出现的频数,计算出基因IPF。最后合并GF和IPF计算GI($g_{i,j}$)。

(2)考虑到通路中有一些基因不在基因表达数据的基因集中,对两个集合取交集,计算交集内的基因与表型的相关性值,计算公式为:

$$r_j = \frac{\mathrm{mean}(\mathrm{pos}) - \mathrm{mean}(\mathrm{neg})}{\mathrm{std}(\mathrm{pos}) - \mathrm{std}(\mathrm{neg})} \quad (4)$$

其中,函数mean(x)、std(x)分别指表型x的基因表达值的平均值和标准差。最后按相关性值从大到小排序得到列表L=[[g_1, r_1], …, [g_j, r_j], …, [g_m, r_m]],包括M个基因(图1b)。

(3)计算通路的富集分数。从列表L的第一个基因开始,当遇到一个在通路p_i里面的基因(hits),则增加分数;遇到一个不在p_i里面的基因(misses),则减少分数,具体公式为:

$$\mathrm{ES}_0(p_i) = \max_{1 \leqslant k \leqslant M}\left|\sum_{\substack{g_j \in p_i \\ 1 \leqslant j \leqslant k}}\left[\frac{|r_j|\left(1+\mathrm{GI}(g_{i,j})\right)}{\sum_{g_j \in p_i}|r_j|\left(1+\mathrm{GI}(g_{i,j})\right)}\right] - \sum_{\substack{g_j \notin p_i \\ 1 \leqslant j \leqslant k}}\frac{1}{M-Q}\right| \quad (5)$$

最终得到一个分数曲线(图1c),曲线上的点到横坐标距离的最大值即为$\mathrm{ES}_0(p_i)$。

(4)随机置换基因N_{perm}次(实验中,N_{perm}取基因富集分析常用的值,1 000次)。随机置换基因指随机在列表L挑选Q个基因作为通路基因。重复步骤(3),计算置换后的通路富集分数$\mathrm{ES}_{\mathrm{perm}}(p_i)$。如图1d所示,最后统计$|\mathrm{ES}_{\mathrm{perm}}(p_i)| > |\mathrm{ES}_0(p_i)|$的数量$N_{\mathrm{sign}}$,$P$值等于$N_{\mathrm{sign}}$与$N_{\mathrm{perm}}$的比值。

为了方便比较,本研究按GSEA方法[4]计算校正后的富集分数(Normalized Enrichment Score, NES)和错误发现率(False Discovery Rate, FDR)。

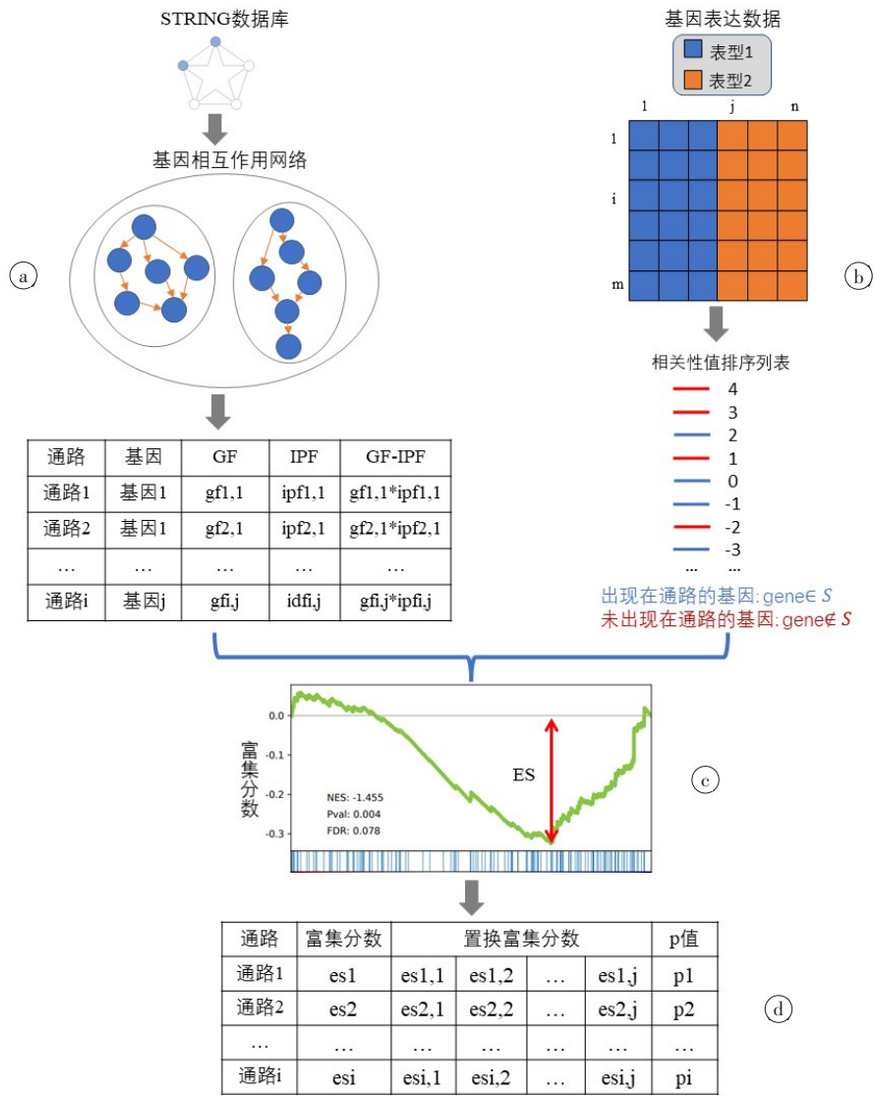

图1 GIGSEA方法流程图

Figure 1 GIGSEA flowchart

1.3 结果评价

本研究将显著性*P*值、FDR和|NES|的阈值分别设为0.05、0.25和1.00，筛选出具有显著性意义的通路，为了检验方法的有效性，与GSEA方法进行比较。两种方法的显著通路大部分是重叠的，重叠通路只是排名的略微差异，因此关注显著通路的差集更有意义，将显著通路差集内的通路定义为差异通路。本研究主要从差异通路的3个方面来说明方法的有效性，包括(1)生物学解释验证。通过查阅关于HCC或CRC相关的生物学研究文献来验证通路与HCC或CRC存在的某种联系。大部分富集方法都通过这种方式来解释，如果通路中的某些基因或者产物对疾病产生影响，那么该通路与疾病是相关的。(2)相关文献数量。通过PubMed生物医学论文数据库检索通路和HCC或CRC存在联系的文献，利用文献数量的多少来表示差异通路与疾病相关性的强弱。如果检索词全部出现在文献中，那么该条文献会出现在PubMed检索结果中，因此文献数量从一定程度上可以反映通路与疾病的相关性。比如对于Jak Stat Signaling Pathway与HCC的相关性，通过关键词Jak Stat、Hepatocellular Carcinoma来进行搜索，排除掉Signaling Pathway这些冗余的词；而对于Asthma和Peroxisome通路，则采取补全关键词Pathway来检索。(3)通路基因集对应的表达数据对表型的分类性能。每一个通路基因集对应的基因表达数据和样本表型数据输入到构建的支持向量机(Support Vector Machine, SVM)分类模型中，求出AUC，结合计算的*P*值、FDR和|NES|，利用分类性能对两种方法进行有效性比较[17]。AUC是分类模型中常用的评价指标，其取值范围为0.5~1.0，AUC越大，分类效果越好，意味着通路基因的表达值对疾病分类越准确。比如HCC数据集包括100个样本，有Normal和Cancer两种表型，各50个；而通路Jak Stat Signaling Pathway中有155个基因，则输入SVM模型的数据为：100个样本，每个样本有155个特征，对应155个基因，每一个样本的标签对应于表型，比如Normal为0、Cancer为1。因此这是一个针对样本表型的二分类问题，如果两种表型的基因表达值存在的差异更高，则更能够把表型区别开来，产生更高的AUC，意味着通路内的基因和HCC疾病的相关性更高。

2 结果与讨论

2.1 HCC数据集的结果与讨论

HCC数据集的富集结果如表1所示，表中共有33个通路，为两种方法显著通路的并集。GSEA富集出30条通路，GIGSEA富集出29条通路，且富集出3个新的通路：Jak Stat信号通路(Jak Stat Signaling Pathway)、过氧化物酶体通路(Peroxisome)、半胱氨酸和蛋氨酸代谢通路(Cysteine and Methionine Metabolism)，分别排在第5、8、11位，与GSEA相比，它们的排名也都提高了。

经过查阅文献发现，GIGSEA的差异通路都与HCC疾病的发生机制有一定的联系。Jak Stat信号通路的失调可能导致包括HCC在内的各种癌症[18]。Tang等[19]发现Jak Stat信号通路在HCC中维持具有肿瘤增殖能力的癌症干细胞以及创建免疫抑制微环境，该通路中的STAT3基因对靶基因和蛋白质的调节促成肿瘤发生的概率。目前针对HCC，已经开发出多种JAK或STAT小分子抑制剂和RNA疗法。过氧化物酶体通路中，过氧化物酶体包含至少50种不同的酶[20]。Xu等[21]发现过氧化物酶体增殖物激活受体δ(PPARδ)是一种核转录因子，与肿瘤发生有关，通过PPARδ和前列腺素信号通路之间的串扰，两个通路共同调节人体HCC细胞的生长。Wirtz等[22]及Zhuang等[23]发现HCC细胞转移与半胱氨酸和甲硫氨酸代谢通路中代谢物水平和代谢酶表达的改变有关，该通路的ENOPH1的过表达促进细胞迁移和侵袭，而ENOPH1下调抑制细胞迁移和侵袭，研究表明ENOPH1可以促进HCC进展，可以作为HCC的生物标志物和治疗靶点。

相关文献可以反映已有研究对于通路和疾病具有相关性的支持，因此本研究对两种方法的差异通路进行相关文献统计，验证其与HCC的相关性。统计结果如表2中文献数量一列所示，GIGSEA和GSEA的差异通路的文献数量均值分别为129和6，GIGSEA明显高于GSEA，表明GIGSEA的差异通路与HCC的相关性是更高的。

接着利用SVM模型来测试差异通路对HCC数据集的两种表型的分类性能。GIGSEA的平均*P*值和平均FDR远小于GSEA，且|NES|大于GSEA，从统计学上表明GIGSEA差异通路与HCC相关性更强。两种方法的差异通路在SVM模型都表现出良好的分类效果，GIGSEA和GSEA的AUC分别达到99.22%和96.99%，表明GIGSEA差异通路的分类性能同样优于GSEA差异通路(表2)。

2.2 CRC数据集的结果与讨论

CRC数据集的富集结果如表3所示，表中共有26个通路，包括GIGSEA和GSEA的显著通路。GSEA富集出20条通路，GIGSEA富集出23条通路，且富集出6个新的通路：肌动蛋白细胞骨架的调节通路(Regulation of Actin Cytoskeleton)、白细胞跨内皮迁移通路(Leukocyte Transendothelial Migration)、补

表1 HCC数据集中两种方法的富集结果

Table 1 Enrichment results of two methods in HCC dataset

通路	*P*值1	排名1	*P*值2	排名2
Fatty Acid Metabolism	0.000 0	1	0.001 0	6
Glycolysis Gluconeogenesis	0.000 0	2	0.001 0	7
Arginine and Proline Metabolism	0.000 0	3	0.002 0	8
Graft Versus Host Disease	0.000 0	4	0.002 1	9
Jak Stat Signaling Pathway*	0.000 0	5	0.062 4	40
Histidine Metabolism	0.000 0	6	0.002 1	10
Starch and Sucrose Metabolism	0.000 0	7	0.004 0	11
Peroxisome*	0.000 0	8	0.123 6	88
Drug Metabolism Other Enzymes	0.001 0	9	0.004 0	12
Pyruvate Metabolism	0.001 0	10	0.005 0	13
Cysteine and Methionine Metabolism*	0.001 0	11	0.090 9	56
Spliceosome	0.001 1	12	0.000 0	1
Proximal Tubule Bicarbonate Reclamation	0.002 0	13	0.000 0	2
Nicotinate and Nicotinamide Metabolism	0.003 0	14	0.022 7	22
Cell Cycle	0.003 0	15	0.000 0	4
Pyrimidine Metabolism	0.003 0	16	0.001 0	5
Base Excision Repair	0.003 0	17	0.005 1	14
Nitrogen Metabolism	0.004 0	18	0.009 3	15
Intestinal Immune Network for Iga Production	0.011 0	19	0.012 0	16
Alanine Aspartate and Glutamate Metabolism	0.011 0	20	0.013 0	17
Dna Replication	0.011 1	21	0.013 1	18
Autoimmune Thyroid Disease	0.012 0	22	0.015 2	19
Retinol Metabolism	0.012 0	23	0.022 7	21
Drug Metabolism Cytochrome P450	0.013 4	24	0.025 8	23
Complement and Coagulation Cascades	0.025 9	25	0.033 9	26
Allograft Rejection	0.035 3	26	0.044 5	27
Valine Leucine and Isoleucine Degradation	0.037 0	27	0.045 1	28
Glycine Serine and Threonine Metabolism	0.038 1	28	0.046 7	29
Tryptophan Metabolism	0.047 8	29	0.049 8	30
Hematopoietic Cell Lineage#	0.600 0	99	0.000 0	3
Progesterone Mediated Oocyte Maturation#	0.062 2	50	0.018 0	20
Glyoxylate and Dicarboxylate Metabolism#	0.092 1	78	0.032 8	24
Vasopressin Regulated Water Reabsorption#	0.078 6	74	0.033 3	25

*P*值1及排名1为GIGSEA富集出的通路的*P*值和*P*值排名；*P*值2及排名2为GSEA富集出的通路的*P*值和*P*值排名。*为GIGSEA的差异通路；#为GSEA的差异通路

体和凝血级联通路（Complement and Coagulation Cascades）、朊病毒病通路（Prion Diseases）、哮喘通路（Asthma）、肾素血管紧张素系统通路（Renin Angiotensin System），它们的排名在GIGSEA中都有一定的提高。

文献检索显示除了哮喘通路，GIGSEA的差异通路都与CRC有一定的联系。Kanaan等[24]发现肌动蛋白细胞骨架调节通路通过细胞骨架蛋白，如Fascin-1，参与转移性散发性CRC的发展；与散发性CRC相比，该通路的调控基因之间的相似遗传多态性和突变也可能与CRC的发育不良、癌变以及侵袭和转移的易感性增加有关。Tremblay等[25]发现E-选

表2 HCC数据集差异通路的对比

Table 2 Comparison of the differential pathways in HCC dataset

方法	差异通路	文献数量	*P*值	FDR	|NES|	AUC
GIGSEA	Jak Stat Signaling Pathway	205	0.000 0	0.096 8	1.112 2	0.995 4
	Peroxisome	144	0.000 0	0.006 3	2.174 8	0.986 1
	Cysteine and Methionine Metabolism	37	0.001 0	0.057 2	1.429 1	0.995 1
	平均值	129	0.000 3	0.053 5	1.572 0	0.992 2
GSEA	Hematopoietic Cell Lineage	18	0.000 0	0.155 2	1.229 9	0.985 9
	Progesterone Mediated Oocyte Maturation	5	0.018 0	0.208 0	1.362 6	0.944 4
	Glyoxylate and Dicarboxylate Metabolism	2	0.032 8	0.024 2	1.535 6	0.976 9
	Vasopressin Regulated Water Reabsorption	0	0.033 3	0.153 2	1.416 3	0.972 2
	平均值	6	0.021 0	0.135 2	1.386 1	0.969 9

表3 CRC数据集中两种方法的富集结果

Table 3 Enrichment results of two methods in CRC dataset

通路	*P*值1	排名1	*P*值2	排名2
Proximal Tubule Bicarbonate Reclamation	0.000 0	1	0.002 1	5
Regulation of Actin Cytoskeleton*	0.001 0	2	0.047 9	23
Nitrogen Metabolism	0.001 2	3	0.007 9	9
Tight Junction	0.005 6	4	0.021 7	15
Axon Guidance	0.006 0	5	0.046 0	19
Pancreatic Secretion	0.007 2	6	0.000 0	1
Inositol Phosphate Metabolism	0.011 5	7	0.000 0	2
Chemokine Signaling Pathway	0.012 3	8	0.007 5	8
MapkSignaling Pathway	0.013 8	9	0.001 2	4
Insulin Signaling Pathway	0.014 2	10	0.003 1	6
GnrhSignaling Pathway	0.016 4	11	0.027 9	16
Leukocyte TransendothelialMigration*	0.018 1	12	0.198 1	47
Phosphatidylinositol Signaling System	0.018 9	13	0.048 0	20
Long Term Depression	0.023 1	14	0.013 1	13
Calcium Signaling Pathway	0.024 4	15	0.033 1	17
Cell Adhesion Molecules Cams	0.026 1	16	0.016 5	14
VegfSignaling Pathway	0.028 1	17	0.011 0	12
Complement and Coagulation Cascades*	0.029 2	18	0.059 6	41
Prion Diseases*	0.033 2	19	0.323 1	56
Asthma*	0.033 3	20	0.052 3	33
Maturity Onset Diabetes of The Young	0.034 6	21	0.045 3	18
Renin Angiotensin System*	0.044 1	22	0.119 0	45
B Cell Receptor Signaling Pathway	0.046 4	23	0.008 3	10
Fc Epsilon Ri Signaling Pathway#	0.073 1	34	0.006 6	7
Fc Gamma R-Mediated Phagocytosis#	0.101 3	42	0.000 0	3
Gap Junction#	0.101 9	43	0.009 1	11

*P*值1及排名1为GIGSEA富集出的通路的*P*值和*P*值排名；*P*值2及排名2为GSEA富集出的通路的*P*值和*P*值排名。*为GIGSEA的差异通路；#为GSEA的差异通路

择蛋白被结肠癌细胞激活会触发p38和ERK MAP激酶的激活，从而诱导细胞骨架重塑，导致内皮层破裂，促进粘附的结肠癌细胞的外渗，该研究表明E-选择蛋白介导的p38和ERK MAP激酶激活对结肠癌细胞跨内皮迁移的调节。Matilda等[26]发现补体和凝血级联通路、参与脂质代谢的通路、急性期反应信号通路是高C反应蛋白(C-reactive Protein, CRP)CRC患者的主要干扰通路。细胞朊病毒蛋白(Cellular Prion Protein, PrPc)是一种细胞表面蛋白，由朊病毒通路中的PRNP基因编码[27]。Ong等[28]发现PrPc的过表达可能通过诱导内皮增殖-分化开关的方式参与CRC诱导的血管生成。Chen等[29]发现肾素血管紧张素系统抑制剂的使用与CRC风险和死亡率降低有关，该研究通过实验证明肾素血管紧张素系统抑制剂使用持续时间每增加一年，CRC风险降低6%。

同样地，通过Pubmed检索和SVM模型分类来验证差异通路与CRC之间的相关性，结果如表4所示。GIGSEA差异通路的相关文献数量的均值为55，而GSEA为34；GSEA的P值要小于GIGSEA，两种方法的差异通路平均FDR和|NES|都接近；GIGSEA的平均AUC明显高于GSEA，分别为91.32%和85.67%，表明GIGSEA的差异通路的分类性能优于GSEA。综合考虑，GIGSEA的差异通路与CRC的相关性比GSEA强。

表4 CRC数据集差异通路的结果对比

Table 4 Comparison of the differential pathwaysin CRC dataset

方法	差异通路	文献数量	P值	FDR	|NES|	AUC
GIGSEA	Regulation of Actin Cytoskeleton	220	0.0010	0.2160	1.5788	0.8170
	Leukocyte Transendothelial Migration	12	0.0181	0.2149	1.6165	0.9738
	Complement and Coagulation Cascades	12	0.0292	0.2265	1.5556	0.9679
	Prion Diseases	7	0.0332	0.2009	1.5206	0.9952
	Asthma	16	0.0333	0.2473	1.6363	0.7274
	Renin Angiotensin System	61	0.0441	0.1952	1.5198	0.9976
	平均值	55	0.0265	0.2168	1.5713	0.9132
GSEA	Fc Epsilon Ri Signaling Pathway	14	0.0066	0.2232	1.6557	0.7762
	Fc Gamma R-Mediated Phagocytosis	3	0.0000	0.2427	1.6089	0.9952
	Gap Junction	86	0.0091	0.2023	1.5292	0.7988
	平均值	34	0.0052	0.2227	1.5979	0.8567

2.3 基因通路富集分析网站

为了方便用户使用GIGSEA进行通路富集分析，本研究基于SSM框架开发了一个在线的基因通路富集分析的网站，SSM框架是Java企业级开发领域Spring、Spring MVC和MyBatis框架的缩写。本研究利用bootstrap-tablejs组件来绘制富集分析的结果展示页面，可以对富集结果进行排序、搜索等功能(图2)。

本研究利用EChartsjs组件的关系图来进行基因通路的可视化，如图3所示，通路图中的节点表示基因，节点之间的连线表示基因之间的相互作用关系，基因节点的大小与基因相关性值的绝对值成比例，基因节点的颜色与基因列表相关性值正负相关，红色代表相关性值为正，蓝色代表相关性值为负，灰色代表该基因不在基因列表中。

3 结 论

本研究提出一种基于改进TF-IDF算法的GIGSEA方法。首先利用通路基因相互作用数据，考虑基因在通路的局部重要性和在通路数据库的全局特异性，计算基因的影响力；然后利用基因表达数据和表型数据计算基因与表型的相关性值；接着融合基因影响力和表型相关性值，计算通路的富集分数；最后通过置换基因的方式，考察通路是否和疾病相关。本研究利用HCC和CRC数据集来测试GIGSEA的效果。与GSEA比较，本研究发现了与HCC相关的3个新通路，以及与CRC相关的6个新通路。除了哮喘通路，本研究都找到研究文献来证实通路与疾病之间的相关性。利用PubMed检索相关文献的结果显示在两个数据集中，GIGSEA的文献数量都远远多于GSEA。利用SVM模型分类的结果显示在两个数据集中，GIGSEA通路对应的表达数据的分类效果都优于GSEA。GIGSEA方法不仅丰富了富集分析方法，更重要的是为发现与疾病相关的通路提供了一种新思路。

sign

Pathway	es	nes	pval	fdr	geneset_size	matched_size
kegg_type_i_diabetes_mellitus	-0.575	-2.048	0.000	0.002	43	36
kegg_ribosome	-0.516	-2.077	0.000	0.003	88	73
kegg_graft_versus_host_disease	-0.579	-1.920	0.000	0.009	41	28
kegg_n_glycan_biosynthesis	0.537	1.963	0.000	0.023	46	24
kegg_neuroactive_ligand_receptor_interaction	-0.393	-1.799	0.000	0.032	272	185
kegg_asthma	-0.543	-1.734	0.003	0.039	30	24
kegg_allograft_rejection	-0.534	-1.742	0.005	0.043	37	28
kegg_cell_adhesion_molecules_cams	-0.422	-1.758	0.000	0.044	133	92
kegg_systemic_lupus_erythematosus	-0.452	-1.653	0.004	0.064	139	43
kegg_calcium_signaling_pathway	-0.369	-1.643	0.001	0.064	178	145

图2 富集结果展示页面

Figure 2 Enrichment result display interface

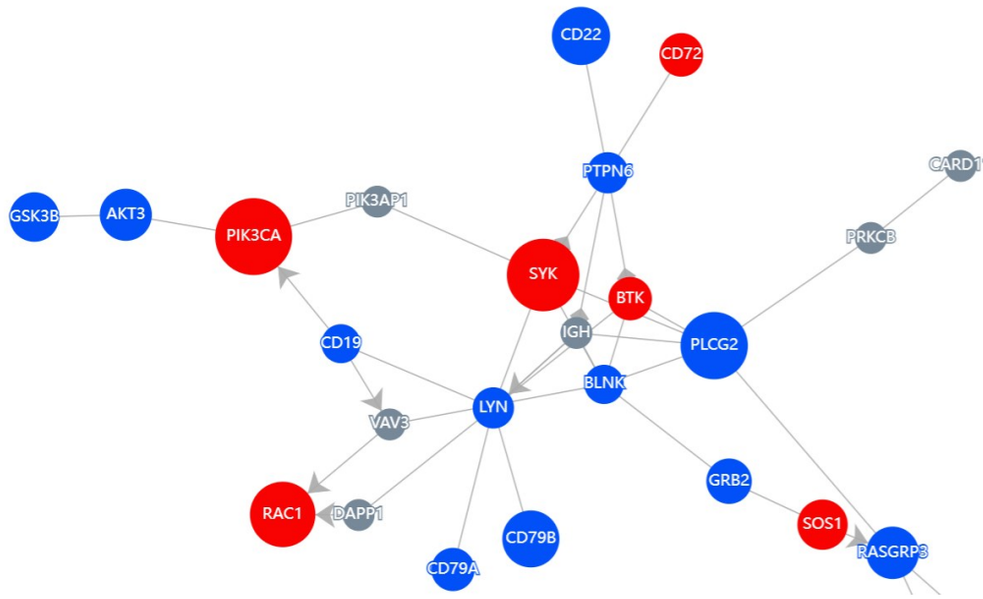

图3 Echart可视化的通路局部示意图

Figure 3 Partial schematic diagram of Echart visualized pathways

【参考文献】

[1] Haynes WA, Higdon R, Stanberry L, et al. Differential expression analysis for pathways[J]. PLoS Comput Biol, 2013, 9(3): e1002967.

[2] Khatri P, Sirota M, Butte AJ, et al. Ten years of pathway analysis: current approaches and outstanding challenges[J]. PLoS Comput Biol, 2012, 8(2): e1002375.

[3] Wang X, Yin TS, Boyi LI, et al. Progress in gene functional enrichment analysis[J]. Scientia Sinica (Vitae), 2016, 46(4): 363-373.

[4] Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proc Nat Acad Sci U S A, 2005, 102(43): 15545-15550.

[5] Saket N, Carl K. The power of protein interaction networks for associating genes with diseases[J]. Bioinformatics, 2010, 26(8): 1057-1063.

[6] Liu W, Li C, Xu Y, et al. Topologically inferring risk-active pathways toward precise cancer classification by directed random walk[J]. Bioinformatics, 2013, 29(17): 2169-2177.

[7] Deng GL, Xu YJ, Zhang CL, et al. A network-based strategy from the global perspective for identification of risk pathways in complex diseases[J]. Prog Biochem Biophys, 2015, 42(3): 286-296.

[8] Yang Q, Wang S, Dai E, et al. Pathway enrichment analysis approach based on topological structure and updated annotation of pathway[J]. Brief Bioinform, 2019, 20(1): 168-177.

[9] Winterhalter C, Widera P, Krasnogor N. JEPETTO: a Cytoscape plugin for gene set enrichment and topological analysis based on interaction networks[J]. Bioinformatics, 2014, 30(7): 1029-1030.

[10] Rahmati S, Abovsky M, Pastrello C, et al. PathDIP: an annotated resource for known and predicted human gene-pathway associations and pathway enrichment analysis[J]. Nucleic Acids Res, 2017, 45 (D1): D419-D426.

[11] Han H, Lee S, Lee I. NGSEA: network-based gene set enrichment analysis for interpreting gene expression phenotypes with functional gene sets[J]. Mol Cells, 2019, 42(8): 579-588.

[12] Yoon S, Kim J, Kim SK, et al. GScluster: network-weighted gene-set clustering analysis[J]. BMC Genom, 2019, 20(1): 352.

[13] Zito A, Lualdi M, Granata P, et al. Gene set enrichment analysis of interaction networks weighted by node centrality[J]. Front Genet, 2021, 12: 577623.

[14] Ogata H, Goto S, Sato K, et al. KEGG: kyoto encyclopedia of genes and genomes[J]. Nucleic Acids Res, 1999, 27(1): 29-34.

[15] Damian S, Gable AL, Nastou KC, et al. The STRING database in 2021: customizable protein-protein networks, and functional characterization

of user-uploaded gene/measurement sets[J]. Nucleic Acids Res, 2021, 49(D1): D605-D612.
[16] Cline MS, Craft B, Swatloski T, et al. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser[J]. Sci Rep, 2013, 3(1): 1-6.
[17] Yu N. GSEMT: a gene set enrichment analysis method based on mantel test[J]. J Phys, 2021, 1828(1): 012048.
[18] Calvisi DF, Ladu S, Gorden A, et al. Ubiquitous activation of Ras and Jak/Stat pathways in human HCC[J]. Gastroenterology, 2006, 130(4): 1117-1128.
[19] Tang JJ, Thng DK, Lim JJ, et al. JAK/STAT signaling in hepatocellular carcinoma[J]. Hepat Oncol, 2020, 7(1): HEP18.
[20] Gabaldón T, Snel B, Zimmeren FV, et al. Origin and evolution of the peroxisomal proteome[J]. Biol Direct, 2006, 1(1): 1-14.
[21] Xu L, Han C, Lim K, et al. Cross-talk between peroxisome proliferator-activated receptor δ and cytosolic phospholipase A2α/Cyclooxygenase-2/Prostaglandin E2 signaling pathways in human hepatocellular carcinoma cells[J]. Cancer Res, 2006, 66(24): 11859-11868.
[22] Wirtz M, Droux M. Synthesis of the sulfur amino acids: cysteine and methionine[J]. Photosynth Res, 2005, 86(3): 345-362.
[23] Zhuang H, Qiang Z, Shao X, et al. Integration of metabolomics and expression of enolase-phosphatase 1 links to hepatocellular carcinoma progression[J]. Theranostics, 2019, 9(12): 3639-3652.
[24] Kanaan Z, Qadan M, Eichenberger MR, et al. The actin-cytoskeleton pathway and its potential role in inflammatory bowel disease-associated human colorectal Cancer[J]. Genet Test Mol Bioma, 2010, 14(3): 347-353.
[25] Tremblay PL, Auger FA, Huot J. Regulation of transendothelial migration of colon cancer cells by E-selectin-mediated activation of p38 and ERK MAP kinases[J]. Oncogene, 2006, 25(50): 6563-6573.
[26] Matilda H, Mayank S, Sakari J, et al. Colorectal cancer patients with different C-reactive protein levels and 5-year survival times can be differentiated with quantitative serum proteomics[J]. PLoS One, 2018, 13(4): e0195354.
[27] Ryskalin L, Busceti CL, Biagioni F, et al. Prion protein in glioblastoma multiforme[J]. Int J Mol Sci, 2019, 20(20): 5107.
[28] Ong SH, Goh KW, Chieng KL, et al. Cellular prion protein and γ-synuclein overexpression in LS 174T colorectal cancer cell drives endothelial proliferation-to-differentiation switch[J]. Peer J, 2018, 6: e4506.
[29] Chen X, Yi CH, Ya KG. Renin-angiotensin system inhibitor use and colorectal cancer risk and mortality: a dose-response meta analysis[J]. J Renin-Angio-Aldo Syst, 2020, 21(3): 147032031989564.

（编辑：谭斯允）