DOI:10.3969/j.issn.1672-5565.201702003

一种基于基因拓扑重要性的通路识别方法

方宏源[1], 咎乡镇[2],沈良忠[2],刘文斌[1]*
(1.温州大学 物理与电子信息工程学院,浙江 温州 325035;
2.温州商学院 信息工程学院,浙江 温州 325035)

摘 要:癌症相关通路的识别是认识癌症发生发展过程机制的生物学基础。已有的通路识别方法很少考虑基因在通路中的拓扑重要性。重叠基因降权(PADOG)方法在基因集分析(GSA)方法的基础上融入了基因特异性的影响,提高了癌症相关通路的识别性能。为进一步提高癌症相关通路的识别性能,首先统计了KEGG通路数据集中基因出度的分布情况,根据基因出度的大小定义了基因的重要性。最后将基因的特异性和重要性融合在一起,提出了一种基于基因重要性和特异性的通路分析方法PAGIS。在结肠癌、肺癌和胰腺癌3个数据集上的实验结果表明,PAGIS方法比PADOG能够提高很多癌症相关的排名,从而提高癌症相关通路的识别效果。

关键词:癌症; 基因表达谱;通路分析;基因特异性;基因重要性
中图分类号:TP311.13; R730.5 **文献标志码**:A **文章编号**:1672-5565(2017)04-214-07

A pathway analysis method based on the topological importance of genes

FANG Hongyuan[1], ZAN Xiangzhen[2], SHEN Liangzhong[2], LIU Wenbin[1]*
(1. *College of Physics and Electronic Information Engineering*, *Wenzhou University*, *Wenzhou* 325035, *Zhejiang*, *China*;
2. *College of Information Engineering*, *Wenzhou Business College*, *Wenzhou* 325035, *Zhejiang*, *China*)

Abstract: Identifying cancer-related pathways is important for understanding the underling mechanisms of the development of cancers. However, current pathway analysis methods are lack of the consideration of topology characteristics of genes in the pathways. Pathway Analysis with Down-weighting of Overlapping Genes (PADOG) considered the specificity of genes based on the GSA method to improve its performance. In order to improve the performance of identifying cancer-related pathways, we first studied the out-degree distribution of genes in the KEGG pathway database. Then we defined the importance of genes based on their out-degree. Finally we proposed a pathway analysis based on the important and specificity of genes (PAGIS). Analysis results from the colorectal cancer datasets showed that our improved method could identify more cancer related pathways than PADOG.

Keywords: Cancer; Gene expression profile; Pathway analysis; Gene specificity; Gene importance

基于微阵列的高通量技术产生了大量的基因表达数据,如何从这些海量基因表达数据中获得洞察性的认识,进而理解生命现象的机制仍然是摆在世界各国科学家面前的一个严峻的挑战。生物通路是一组完成特定功能的基因之间的相互作用关系,主要有信号传导通路和代谢通路。在信号传导通路中,节点代表基因(或基因产物),边代表从一个基因转导到另一个基因的信号。在代谢通路中,节点代表生化化合物,边代表通过酶编码的化合物之间的生物化学反应,酶是为基因编码的。常用的通路数据库有KEGG[1]和Reactome[2]数据库,它们提供了基因之间相互作用的可视化形式。在过去十多年中,研究者开发了很多基于通路的基因表达差异分析方法,来识别各种癌症或疾病相关的通路。

2005年,PNAS上发表了两篇重要的通路分析方法的论文,一个是Tian等[3]提出的基于功能的显

收稿日期:2017-02-28;**修回日期**:2017-05-16.
基金项目:国家自然科学基金(61572367,61272018,61573017);浙江省自然科学基金(lq17c060001);温州大学研究生创新基金(3162014037).
作者简介:方宏源,男,硕士研究生,研究方向:生物信息学、模式识别等; E-mail: 121393069@qq.com.
* **通信作者**:刘文斌,男,教授,研究方向:生物信息学、数据挖掘、DNA计算等;E-mail: wbliu6910@126.com.

著通路分析方法,这种方法综合考虑了一个基因集合中基因表达与集合外基因表达差异的显著性(行置换),以及该基因集基因表达与表型相关性的的显著性(列置换)。另一个是 Subramanian 等[4]提出著名的基因集富集分析方法 GSEA 方法,其主要思想是根据通路中基因表达情况与给定表型之间的相关性对所有基因进行排序,然后确定给定通路 P 的 Kolmogorov-Smirnov 统计量在排序列表中靠近极端处程度的得分。该方法中,Kolmogorov-Smirnov 统计量的显著性根据样本的列置换确定。2006 年,Zahn 等[5]使用 Van der Waerden 统计量代替 Kolmogorov-Smirnov 统计量并用自举抽样代替置换检验方法该方法考虑了通路中两个基因表达水平的相关性以及与其他因素的相关性。同年,EFRON 等[6]用最大-均值统计量替代 Kolmogorov-Smirnov 统计量来计算通路分数,然后通过行置换方法对该分数进行标准化,最后利用列置换来检验通路分值的显著性,这就是著名的 GSA 方法。

从系统生物学的角度,基因之间的相互作用及其动力学的变化是导致各种疾病及癌症发生的主要原因[7-12]。因此,癌症相关通路的识别应尽可能考虑到通路中包含基因的各种信息,如基因的上下游位置、调控基因的数量、基因之间的作用关系等等因素。2009 年,Tarca 等[13]考虑了通路中基因的上下游位置关系提出了著名的信号通路影响分析(SPIA)方法。同年,Thomas 等[14]提出了一种考虑通路中基因拓扑结构的方法,主要思想是位于上游和下游的基因比上下游中间位置具有更高权重,并且在打分上使得紧密连接的基因比不紧密连接的基因具有更高的分数。在通路中,有些基因频繁出现在很多通路中,这些基因可以看作是非特异性基因,其变化对特定通路的影响相对较小;反过来,另外一些基因仅在特定通路出现,即其特异性很高,这些基因的变化对该通路的影响往往很大。2012 年,Tarca 等[15]在 GSA 方法的基础上加入了基因特异性的影响,提出重叠基因降权的通路分析方法(PADOG)。最近,Liu 等[14]提出了称为基因相互作用富集和网络分析(GIENA)的方法,以表示协同、竞争、冗余,表达水平的依赖性的失调的基因相互作用。

由 KEGG 中的 Ras 信号传导通路,可看出其中的 Ras 基因调节该通路中的许多下游基因。由于 Ras 基因参与控制细胞分裂和细胞死亡的许多信号传导通路,已有研究表明该基因的过表达和突变与许多癌症相关,如胰腺、结肠、肺(30%)、甲状腺、膀胱、卵巢、乳腺、皮肤、肝脏、肾脏和一些白血病等。显然在通路中,调控大量基因的基因应该比仅调控少量基因的基因更为重要,它们的差异对通路的功能应该具有更大的影响。考虑这一现象,本文将基因平均出度的大小定义为基因的重要性,并和 PADOG 方法中的基因的特异性结合起来,提出了一种基于重要性和特异性的通路识别方法 PAGIS。在结肠癌、肺癌和胰腺癌 3 个数据集上的结果表明,改进后的方法能够提高癌症相关通路的识别精度。

1　材料与方法

1.1　数据集

本文主要分析了 3 个癌症数据集。

1)结肠癌数据集 GSE4107,该数据集包括 12 个结肠癌样本与 10 个正常样本(Affymetrix HG-U133 Plus 2.0 微阵列平台)。

2)肺癌数据集 GSE27262,该数据集包括 25 个肺癌样本和 25 个正常样本(Affymetrix Human Genome U133 Plus 2.0 微阵列平台)。

3)胰腺癌数据集 GSE16515,包括 36 个胰腺癌样本和 16 个正常样本。

1.2　频度和平均出度的分布

如图 1 所示是 KEGG 数据库中 204 个信号通路的基因的频度和出度分布图,其中图 1(a)是基因的平均频度分布,可以看出大多数基因仅出现在一两条通路中,只有少数基因出现在多条通路中。图 1(b)是基因的平均出度分布,可以看出仅有少数基因调控大量下游基因,而大多数基因的平均出度在 0~5 之间。图 1(c)是基因的频度和平均出度的散点图,可以看出仅有部分平均出度大且频度低的基因。本文把平均出度在前 100 名的基因在 DAVID 数据库中进行 GO 功能注释,结果发现显著富集在一些癌症相关通路中,如 pathways in cancer, adipocytokine signaling pathway, neurotrophin signaling pathway, thyroid cancer, ErbB signaling pathway, PPAR signaling pathway, 和 renal cell carcinoma。这说明这些平均出度大的基因与癌症的发生发展具有密切的关系,提高它们在癌症相关通路中的权重具有生物学意义。

基因在通路中出现的频度实际上反映了一个基因的特异性,频繁出现在很多通路中的基因属于一些“公共基因”,它们对通路的影响相对较小;仅在一条或几条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基因就大。在 PADOG 方法中,文献[15]定义基因的特异性权重为

$$w_f(g) = 1 + \sqrt{\frac{\max(f) - f(g)}{\max(f) - \min(f)}}$$

式中：max(f) 、min(f) 分别为 204 条 KEGG 通路中最大频度和最小频度；$w_f(g)$ 反映基因在通路中特异程度，该值越大则基因在通路中特异程度越高，反之则特异程度越低，$w_f(g)$ 取值在 1 ~ 2 之间。

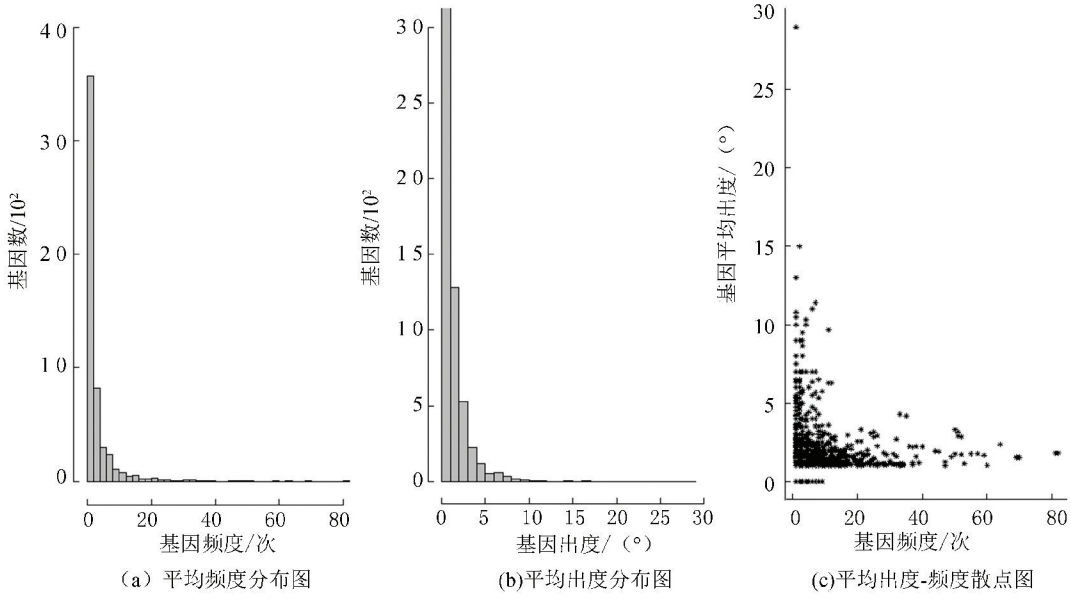

图 1 204 条 KEGG 通路基因平均出度-频度分布图

Fig.1 Distribution of the average gene out-degrees and frequencies across the 204 KEGG signaling pathways

由于基因出度表示的是一个基因调控的下游基因的数量，因此，出度越大的基因，对通路的影响就越大。为此，本文定义基因重要性的权重为

$$w_d(g) = 1 + \sqrt{\frac{d(g) - \min(d)}{\max(d) - \min(d)}}$$

式中：max(d) 、min(d) 分别为 204 条 KEGG 通路中基因最大平均出度和最小平均出度；$w_d(g)$ 反映基因在通路中的重要性，该值越大则基因在通路中重要程度越高，值越小则基因在通路中重要程度越低，取值也在 1~2 之间。

1.3 癌症相关通路分析方法

本文简要介绍 GSEA 方法、GSA 方法、PADOG 方法，进而引出本文的改进方法。假定所有基因总数为 N，给定一个通路 S，通路中基因数为 M，GSEA 的主要过程如下。

Step1 按照每个基因 g 与表型间相关性 r（或 t 统计量）对 N 个基因排序 $w_d(g)$ $L = [g_1, \ldots, g_j, \ldots g_N]$。

Step2 用带权值的 Kolmogorov-Smirnov 统计量计算通路的富集分数 $\mathrm{ES}_0(S)$ 为

$$\mathrm{ES}_0(S) = \max_{1 \leqslant i \leqslant N} \left| \sum_{\substack{g_j \in S \\ 1 \leqslant j \leqslant i}} \left[|r_j|^p / \sum_{g_j \in S} |r_j|^p \right] - \sum_{\substack{g_j \notin S \\ 1 \leqslant j \leqslant i}} \frac{1}{N - M} \right|$$

式中 p 为用来校正 ES 的权值，p 一般取 1。

Step3 随机置换样本标签 N_{ite} 次，并重新计算通路 S 的分数 $\mathrm{ES}_{\mathrm{ite}}(S)$。

Step4 计算该通路富集分数 $\mathrm{ES}_0(S)$ 的显著性 p-value。

在 GSA 方法中，文献[6]使用“最大均值”统计量代替 Kolmogorov-Smirnov 统计量来计算通路分数 ES。公式如下：

$$\mathrm{ES}_0(S) = \max\left\{ \frac{1}{M} \sum_{g_j \in S} z_j^{(+)}, \left| \frac{1}{M} \sum_{g_j \in S} z_j^{(-)} \right| \right\}$$

式中：$\frac{1}{M} \sum_{g_j \in S} z_j^{(+)}$ 为通路中基因的正得分平均值；$\frac{1}{M} \sum_{g_j \in S} z_j^{(-)}$ 为通路中基因的负得分平均值。GSA 与 GSEA 的另一不同之处在于，通过行随机化方法标准化通路得分并且使用样本列置换方法来确定通路的显著性。

1.4 基于重叠基因降权通路分析方法(PADOG)

使用通路中所有基因的加权绝对矫正 t 分数和的均值来计算通路 S 分数 $\mathrm{ES}_0(S)$，公式如下：

$$\mathrm{ES}_0(S) = \frac{1}{M} \sum_{g_j \in S} |T(g_j)| \cdot w_f(g_j)$$

式中：$T(g_j)$ 为基因 g_j 在两类样本中矫正 t 分数；$w_f(g_j)$ 为基因 g_j 的权重。

利用行随机化和置换排列方法计算通路显著性 p-value。公式如下：

$$P_{\mathrm{PADOG}}(ES)=\frac{\sum_{\mathrm{ite}} I(ES^{*}_{\mathrm{ite}}(S)\geqslant ES^{*}_{0}(S))}{N_{\mathrm{ite}}}$$

式中:函数 I 中表达式为真则返回结果 1,否则返回 0; $ES^{*}_{ite}(S)$ 为通路 S 在第 ite 次置换排列中的标准化得分。

1.5　基于基因重要性和特异性的通路分析方法(PAGIS)

为将基因的平均出度引入到 PADOG 方法框架中,本文合并权重 $w_f(g)$ 和 $w_d(g)$ 成 $w(g)$,公式如下:

$$w(g)=\sqrt{w_f(g)\cdot w_d(g)}$$

式中: $w_f(g)$ 为基因频度的权重; $w_d(g)$ 为基因平均出度的权重; $w(g)$ 为合并权重且值取 1 ~ 2; $w(g)$ 反映基因在通路中的重要性和特异性的程度,基因在通路中重要程度和特异程度越高则该值越大,相反基因的重要程度或特异程度越低则该值越小。本文将 $w(g)$ 作为 PADOG 计算通路分数的新权重并提出 PAGIS 方法。

2　结果与分析

本文比较 PADOG 和 PAGIS 方法在 3 个癌症数据集上的结果,PADOG 的 R 语言包由文献[15]开发。由于不同方法 p 值计算有所不同,仅仅比较 p 值不够合理。本文基于通路的 p 值升序排列并比较排名,通路排名越靠前则该通路倾向被认为与癌症显著相关。表 1~3 列出 PADOG 和 PAGIS 方法在前 30 名中与癌症相关的通路排名。在 3 个癌症数据集中,PADOG 和 PAGIS 共识别出 21、23、15 条癌症相关通路。

表 1　PAGIS 和 PADOG 方法在结肠癌数据集中前 30 名癌症相关通路和排名

Table 1　The rank of top 30 cancer-related pathway in colorectal cancer

Pathway No(通路排名)	Pathway Name(通路)	PAGIS	PADOG
1	Metabolic pathways	1	82
2	Cell cycle	2	10
3	Ribosome biogenesis in eukaryotes	3	1
4	Bile secretion	4	2
5	Purine metabolism	5	11
6	Fatty acid degradation	6	7
7	Fatty acid elongation	8	3
8	Pathways in cancer	11	79
9	DNA replication	13	16
10	Pyrimidine metabolism	14	42
11	p53 signaling pathway	16	31
12	Apoptosis	18	13
13	Mismatch repair	20	17
14	Ubiquitin mediated proteolysis	22	114
15	RNA polymerase	25	27
16	Colorectal cancer	26	14
17	Sulfur metabolism	27	24
18	Base excision repair	29	49
19	Drug metabolism-other enzymes	30	46
20	One carbon pool by folate	55	19
21	B cell receptor signaling pathway	35	26
	Average Rank(平均排名)	17.62	30.14

图 2(a)~(c)分别是 PADOG 和 PAGIS 方法在结肠癌、肺癌和胰腺癌数据集中癌症相关通路的排名折线图。该图中横轴对应表 1~3 中 Pathway No 字段,纵轴对应表 1~3 中 PADOG 和 PAGIS 方法中的通路排名。由图 3 可看出,相比 PADOG 方法 PAGIS 能够显著提高某些癌症相关通路的排名。如图2(a)所示,通路 Metabolic pathways, Pathways in cancer 和 Ubiquitin mediated proteolysis 在 PADOG 方法中排名是 82、79 和 114,而 PAGIS 是 1、11 和 22;在肺癌数据集(图 2(b))中 ECM-receptor interaction

和 Metabolic pathways 在 PADOG 方法中排名分别是 53、195，而 PAGIS 是 20、29；在胰腺癌数据集中（图 2(c)）中通路 ECM-receptor interaction，Cell cycle 和 Regulation of actin cytoskeleton，在 PADOG 方法中排名分别是 25、35 和 31，而 PAGIS 是 5、15 和 16。

表 1~3 列出 PADOG 和 PAGIS 方法在 3 个癌症数据集中识别出癌症相关通路的平均排名，PADOG 方法识别出癌症相关通路的平均排名分别为30.14、29.43和15.87，而 PAGIS 分别为17.62、16.91和14.13，排名值越小越靠近排名列表的顶端位置，意味着总体与癌症相关程度越高；排名值越大越靠近排名列表的底端位置，意味着总体相关程度越低。显然在 3 个癌症数据集中 PAGIS 方法识别出的癌症相关通路平均排名位置比 PADOG 方法更靠近顶端位置。

表 2 PAGIS 和 PADOG 方法在肺癌数据集中前 30 名癌症相关通路和排名

Table 2 The rank of top 30 cancer-related pathway in lung cancer

Pathway No(通路排名)	Pathway Name(通路名称)	PAGIS	PADOG
	Focal adhesion	1	16
2	Cell cycle	2	3
3	Malaria	3	1
4	Adherens junction	4	6
5	Vascular smooth muscle contraction	5	5
6	Endocytosis	6	22
7	DNA replication	8	11
8	Pathways in cancer	10	33
9	Axon guidance	12	4
10	p53 signaling pathway	13	24
11	Tight junction	14	28
12	mRNA surveillance pathway	16	23
13	Arrhythmogenic right ventricular cardiomyopathy (ARVC)	17	12
14	Dilated cardiomyopathy	18	26
15	ECM-receptor interaction	20	53
16	Chemokine signaling pathway	21	54
17	Ubiquitin mediated proteolysis	23	38
18	Renin-angiotensin system	26	57
19	Homologous recombination	28	7
20	Metabolic pathways	29	195
21	ErbB signaling pathway	30	20
22	Vitamin B6 metabolism	47	9
23	Proteasome	36	30
	Average Rank(平均排名)	16.91	29.43

表 3 PAGIS 和 PADOG 方法在胰腺癌数据集中前 30 名癌症相关通路和排名

Table 3 The rank of top 30 cancer-related pathway in pancreatic cancer

Pathway No(通路排名)	Pathway Name(通路名称)	PAGIS	PADOG
1	Pathways in cancer	1	1
2	Focal adhesion	2	2
3	ECM-receptor interaction	5	25
4	p53 signaling pathway	6	3
5	Axon guidance	7	10
6	Notch signaling pathway	13	21
7	Mucin type O-Glycan biosynthesis	14	6
8	Cell cycle	15	35
9	Regulation of actin cytoskeleton	16	31
10	Pancreatic cancer	17	13
11	Wnt signaling pathway	18	17
12	Apoptosis	20	12
13	Tight junction	23	11
14	Hepatitis C	24	27
15	ErbB signaling pathway	31	24
	Average Rank(平均排名)	14.13	15.87

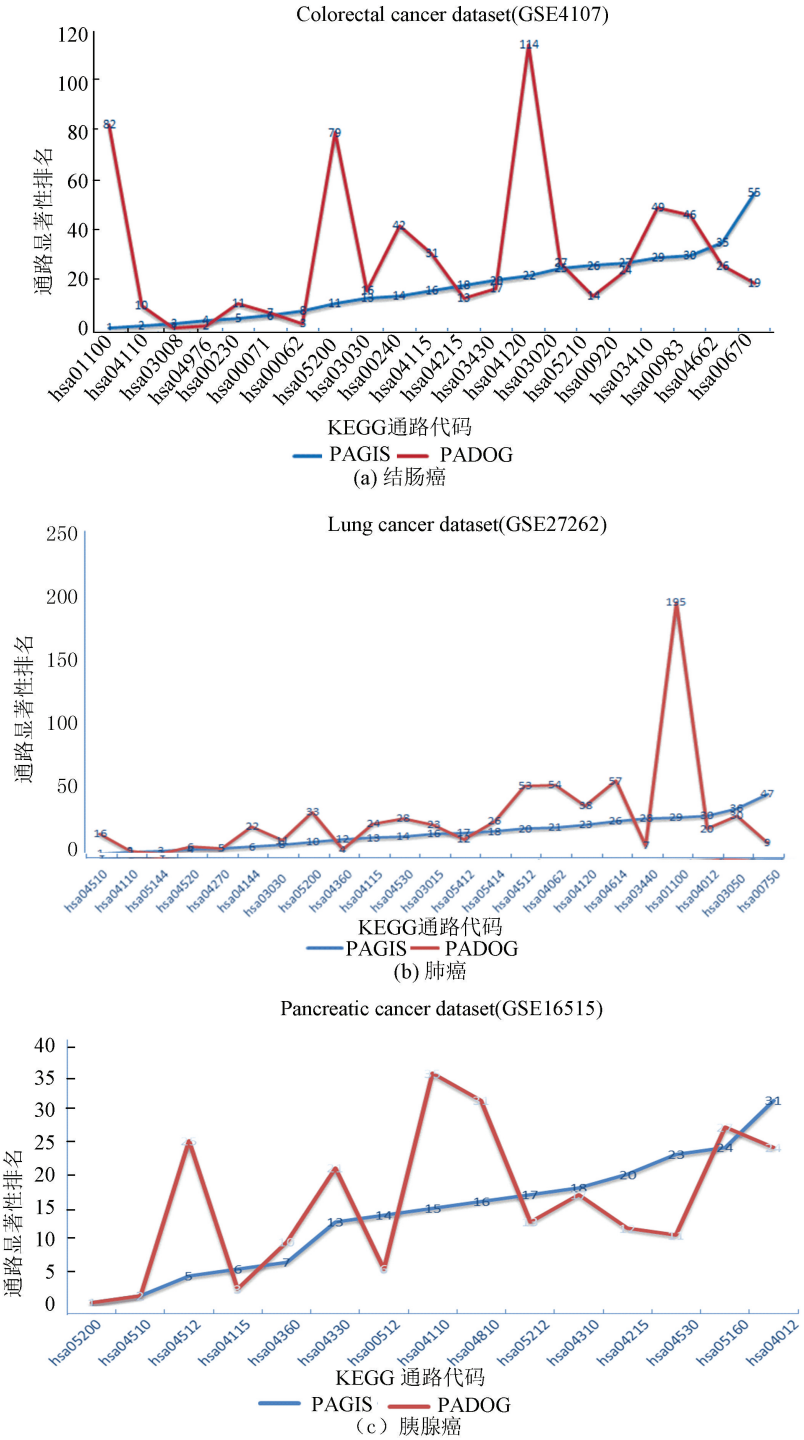

图 2　PADOG 和 PAGIS 方法在 3 个数据集中癌症相关通路排名折线图

Fig.2　Ranks of the cancer related pathways by PAGIS and PADOG in the three cancer datasets

另一方面如图 2 中虚线所示,在结肠癌数据集中,排名在 30 名后的通路 PADOG 方法有 7 条,而 PAGIS 仅有 2 条;在肺癌数据集中 PADOG 方法有 6 条,PAGIS 仅有 2 条;在胰腺癌数据集中 PADOG 方法有 2 条而 PAGIS 有 1 条。显然 PAGIS 方法能识别出更多的癌症相关通路。为进一步比较 PADOG 和 PAGIS 方法在 3 个癌症数据集中的性能,本文分别列出 PADOG 和 PAGIS 方法在前 10、20 和 30 名中识别出的癌症相关通路的数目,见表 4。表中在结肠癌数据集中前 10 名与癌症相关的通路 PAGIS 方法识别出 7 条,PADOG 识别出 5 条,前 20 名中 PAGIS 方法识别出 13 条而 PADOG 识别出 11 条,前 30 名中 PAGIS 方法识别出 19 条而 PADOG 识别出 14 条。其他两个数据集的结果和结肠癌数据集类似,这说明在各段排名中 PAGIS 方法能稳定的识别出比 PADOG 更多的癌症相关通路,PAGIS 具有比 PADOG 更好的性能优势。

表 4 PADOG 和 PAGIS 方法在前 10、20、30 名中识别癌症相关通路数目

Table 4 Numbers of cancer-related pathway in top 10, 20, 30 identified by PADOG and PAGIS

Rank(排名)	Colorectal cancer(结肠癌)		Lung cancer(肺癌)		Pancreatic cancer(胰腺癌)	
	PADOG	PAGIS	PADOG	PAGIS	PADOG	PAGIS
10	5	7	7	8	5	5
20	11	13	11	15	9	12
30	14	19	17	21	13	14

3 结 论

1)本文统计了 KEGG 数据库中 204 条信号通路中基因的频度和出度,并计算出每个基因的平均出度。

2)在基因特异性加权的通路分析方法(PADOG)基础上引入基因的平均出度,并用平均出度表示基因在通路中的重要程度。

3)合并基因特异性和重要性的权值,提出一种基于基因拓扑重要性的通路识别方法(PAGIS),并将该方法应用在结肠癌、肺癌和胰腺癌数据集中。

4)总体上 PAGIS 方法比 PADOG 方法识别出更多的癌症相关通路,能稳定提高癌症相关通路的识别率。

参考文献 (References)

[1]KANEHISA M, FURUMICHI M, TANABE M, et al. KEGG: new perspectives on genomes, pathways, diseases and drugs[J]. Nucleic Acids Research, 2017, 45(D1): D353-D361. DOI: 10.1093/nar/gkw1092.

[2] FABREGAT A, SIDIROPOULOS K, VITERI G, et al. Reactome pathway analysis: a high-performance in-memory approach[J]. BMC Bioinformatics, 2017, 18(1): 142. DOI: 10.1186/s12859-017-1559-2.

[3]TIAN Lu, GREENBERG S A, KONG S W, et al. Discovering statistically significant pathways in expression profiling studies[J].Proceedings of the National Academy of Sciences of the United States of America, 2005, 102(38), 13544-13549.DOI:10.1073/Pnas.0506577102.

[4]SUBRAMANIAN A, TAMAYO P, MOOTHA V K, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles[J]. Proceedings of the National Academy of Sciences, 2005, 102(43):15545-15550. DOI:10.1073/pnas.0506580102.

[5]ZAHN J M,SONU R VOGEL H,et al. transcriptional profiling of aging in human muscle reveals a common aging signature[J]. PLoS Genetics, 2016,2(7):e115.DOI:10.1371/journal.pgen.0020115.

[6]EFRON B B, TIBSHIRANI R. On testing the significance of sets of genes[J].The Annals of Applied Statistics, 2007, 1(1):107-129. DOI: 10.1214/07-AOAS101.

[7]KHATRI P, SIROTA M, BUTTE A J. Ten years of pathway analysis: current approaches and outstanding challenges-supplementary notes[J]. Plos Computational Biology, 2012, 8(2):e1002375. DOI: 10.1371/journal.pcbi.1002375.

[8]TURNBULL C, SEAL S, RENWICK A, et al. Gene-gene interactions in breast cancer susceptibility[J]. Human Molecular Genetics, 2012, 21(4):958-962. DOI: 10.1093/hmg/ddr525.

[9]JEONG H H, LEEM S, WEE K, et al. Integrative network analysis for survival-associated gene-gene interactions across multiple genomic profiles in ovarian cancer[J]. Journal of Ovarian Research, 2015, 8(1):42.DOI: 10.1186/s13048-015-0171-1.

[10]ZHANG Jigang, LI Jian, DENG Hongwen. Identifying gene interaction enrichment for gene expression data[J]. Plos One, 2009, 4(11): e8064. DOI: https://doi.org/10.1371/journal.pone.0008064.

[12]DUTTA B, WALLQVIST A, REIFMAN J. PathNet: a tool for pathway analysis using topological information[J]. Source Code for Biology and Medicine, 2012, 7(1):10. DOI: 10.1186/1751-0473-7-10.

[13]TARCA A L,DRAGHICI S,KHATRI P,et al. A novel signaling pathway impact analysis[J]. Bioinformatics, 2009, 25(1): 75-82.DOI:10.1093/bioinformatics/BTN577.

[14] THOMAS R, GOHLKE J M, STOPPER G F, et al. Choosing the right path: enhancement of biologically relevant sets of genes or proteins using pathway structure[J]. Genome Biology, 2009, 10(4):R44. DOI: 10.1186/gb-2009-10-4-r44.

[15]TARCA A L, DRAGHICI S, BHATTI G, et al. Down-weighting overlapping genes improves gene set analysis[J]. BMC Bioinformatics, 2012, 13(1):136. DOI: 10.1186/1471-2105-13-136.

[16]LIU Yu, KOYUTÜRK M, BARNHOLTZ-SLOAN J S, et al. Gene interaction enrichment and network analysis to identify dysregulated pathways and their interactions in complex diseases[J]. BMC Systems Biology, 2012, 6(1): 65. DOI: 10.1186/1752-0509-6-65.