基于中值的JS散度可变剪接差异分析研究

刘文斌*①② 王 兵① 方 刚② 石晓龙② 许 鹏②③

①(温州大学计算机与人工智能学院 温州 325035)

②(广州大学计算科技研究院 广州 510006)

③(黔南民族师范学院计算机与信息学院 都匀 558000)

摘 要：可变剪接是一种广泛存在于生物体中造成蛋白质多样性的重要机制，它对细胞的增殖、分化、发育、凋亡等一系列重要的生物过程具有重要精细调控的作用。近年来，人们发现多种复杂疾病的产生往往伴随着剪接异构体的紊乱表达。为了研究剪接异构体在整体分布上的差异，该文提出一种基于中值的JS散度可变剪接(AS)差异分析方法。结果表明，该文的方法能够发现大量在剪接异构体整体分布上具有显著差异的基因。这些基因不仅富集在一些癌症密切相关的通路，而且也富集在一些基于可变剪接调控的信号通路、细胞分裂过程和蛋白质功能等通路。此外，与基因层次的差异分析相比，可变剪接显著差异的基因在生存分析方面也具有更好的性能。总之，该文提出基于中值的JS散度可变剪接差异分析方法，将为进一步揭示可变剪接在癌症中的机制奠定基础。

关键词：可变剪接；癌症；JS散度；KEGG通路；驱动基因

中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2020)06-1392-09

DOI: 10.11999/JEIT190941

Study on the Differential Analysis of Alternative Splicing Based on the Median Value Jensen-Shannon Divergence

LIU Wenbin①② WANG Bing① FANG Gang② SHI Xiaolong② XU Peng②③

①(*College of Computer Science and Artificial Intelligence*, *Wenzhou University*, *Wenzhou* 325035, *China*)

②(*Institute of Computing Science and Technology*, *Guangzhou University*, *Guangzhou* 510006, *China*)

③(*School of Computer Science and Information Technology*, *Qiannan Normal University for Nationalities*, *Duyun* 558000, *China*)

Abstract: Alternative splicing is an important mechanism of protein diversity in a wide range of organisms, which plays an important role in the fine regulation of cell proliferation, differentiation, development, apoptosis and a series of important biological processes. In recent years, it is found that the occurrence of multiple complex diseases is often accompanied by the disordered expression of splicing isoforms. In order to study the difference of splicing isoforms on the whole distribution, a differential analysis method of Alternative Splicing (AS) based on the median value by Jensen-Shannon (JS) divergence is proposed in this paper. The results show the method can finds plenty of genes with significant differences in the overall distribution of splicing isoforms. These genes are not only concentrated in some cancer related pathways, but also in some signaling pathways based on alternative splicing regulation, cell division process and protein function. In addition, compared with the gene-level differential analysis, the genes with significant difference in alternative splicing also have better performance in survival analysis. In conclusion, the proposed method will lay a foundation for further revealing the mechanism of alternative splicing in cancer.

Key words: Alternative splicing; Cancer; Jensen-Shannon (JS) divergence; Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway; Driver gene

收稿日期：2019-11-22；改回日期：2020-04-24；网络出版：2020-05-13

*通信作者： 刘文斌 wbliu6910@126.com

基金项目：国家重点研发计划(2019YFA0706402)，国家自然科学基金(61572367, 61573017, 61972107, 61972109)

Foundation Items: The National Key R&D Program of China (2019YFA0706402), The National Natural Science Foundation of China (61572367, 61573017, 61972107, 61972109)

1　引言

在真核生物中，一个基因翻译成相应蛋白质，需要通过3个步骤：(1)DNA转录成前体mRNA(pre-mRNA)；(2)前体mRNA再经过内含子(intron)去除外显子(exon)保留的剪接反应将保留的外显子拼接形成成熟mRNA；(3)成熟mRNA指导核糖体最终翻译成蛋白质。可变剪接(Alternative Splicing, AS)是指有些基因的一个前体mRNA通过选择不同的剪接位点组合，形成不同mRNA剪接异构体(isoform)的过程。最终，一个基因将通过可变剪接将翻译成多种蛋白质。常见的可变剪接模式主要分为7类，分别为外显子跳跃(Exon Skipping, ES)、可变供体剪接位点(Alternative Donor sites, AD)、可变受体剪接位点(Alternative Acceptor sites, AA)、内含子保留(Retained Intron, RI)、互斥外显子(Mutually Exclusive exons, ME)、可变启动子(Alternate Promoter, AP)和可变poly(A)位点(Alternate Terminator, AT)。人类的主要剪接模式为外显子跳跃(约占35%)，其次为可变受体剪接位点(约占16%)和可变供体剪接位点(约占15%)[1]。现有研究表明，大多数真核生物的蛋白质编码基因含有多个外显子，其中大约95%的基因均存在可变剪接事件[2]。根据对TCGA数据库中人类基因异构体数目的统计，仅有28.7%的基因只有一种剪接模式，剩余约71.3%的基因存在至少2个异构体，9.9%的基因存在6个以上的异构体。生物学研究表明：可变剪接是调节基因表达和产生蛋白质组多样性的重要机制，它们对细胞分化、发育、生理功能及其精细调节具有重要的影响[3]。

近年来随着RNA-Seq技术的广泛应用，为深入研究可变剪接在生命过程的调控机制及其对疾病的影响提供了丰富的转录组数据。通过对RNA-Seq数据的分析，发现剪接异构体之间存在一种平衡性。可变剪接紊乱会导致各异构体之间原有的比例失衡，某些异构体表达量的改变，甚至产生新的异构体，从而破坏原有的生物学进程。当某些基因发生不规则剪接时，即通过可变剪接产生多种不同的转录异构体，各个不同的转录异构体编码出结构不同的蛋白质亚型，甚至具有不同的功能，如Bcl-X基因存在多种异构体，其中有2种具有拮抗功能的异构体，抑制细胞凋亡的异构体Bcl-Xl和促进细胞凋亡的异构体Bcl-Xs，现已在多种癌症类型中观察到Bcl-Xl的上调和Bcl-Xs的下调[4]。可变剪接的研究进一步促进了疾病的诊断和治疗，如Nek2基因的剪接异构体Nek2C与乳腺癌的发生密切相关，抑制Nek2C的表达是乳腺癌的潜在治疗手段，或者通过空间阻滞寡核苷酸治疗策略[5]，阻止剪接复合体对pre-mRNA的错误剪接，恢复正确mRNA亚型的表达。

可变剪接紊乱可能对关键转录因子、信号分子、膜蛋白、分泌蛋白等的生成产生巨大的影响，从而导致多种人类疾病甚至是癌症的产生[6]，如帕金森综合征、乳腺癌相关脾脏络氨酸激活酶isoform-S[7]、人类上皮增长因子受体等[8]。因此，找寻异常的可变剪接有助于进一步揭示疾病发生、发展的机制。现阶段对于可变剪接差异分析主要是基于表达水平和剪接模式上的差异。Shen等人[9]提出了rMATS(robust Multivariate Analysis of Transcript Splicing)模型针对重复的RNA-Seq数据来检测差异可变剪接基因。欧书华等人[10]提出基于KL散度(Kullback-Leibler divergence)的RNA-Seq数据差异异构体比例检测方法来测差异可变剪接基因。Liu等人[11]通过对基因各剪接模式的PSI(Percent-Spliced-In)值进行统计学分析检测差异可变剪接基因。Zong等人[12]使用多因素比例风险回归法来检测差异可变剪接基因。Zhang等人[13]将深度学习与贝叶斯假设检验相结合来检测差异可变剪接基因。

本质上，可变剪接各异构体概率分布的差异是导致可变剪接紊乱的原因。简单的从一个剪接异构体或一种剪接类型的差异出发，无法从整体上认识可变剪接紊乱对疾病的影响。现有的研究也主要从单个基因或单个异构体层次来考虑基因的差异，缺少在整体上分析基因各异构体的差异。在统计应用中，经常用一个简单的、近似的概率分布来描述另一个复杂的、真实的概率分布，而KL散度就被应用于比较这两个概率分布的差异程度。但由于KL散度具有非对称性等不足，通过使用KL散度的一种变体JS散度(Jensen-Shannon divergence)能更确切地衡量两个概率分布的差异性。借助该思想，本文提出了一种基于中值的JS散度可变剪接(AS)差异分析方法。通过使用癌症和正常样本的表达中值，来构建两种状态下基因异构体的代表表达向量，再根据代表向量中各异构体的百分比计算基因的JS散度，来研究剪接异构体在整体分布上的差异。且从单个基因和单个异构体角度构建差异分析模型与本方法进行对比，本文检测到了其它方法不能检测出的差异基因。通过KEGG通路分析，发现了一些与代谢、蛋白质等相关的一些通路，如代谢通路、蛋白质消化吸收等通路，这些通路与可变剪接紊乱和癌症发生均存在密切关联。根据驱动基因分析，发现驱动基因不仅仅通过变异促进癌症的发展，而且有一部分癌症驱动基因的可变剪接也与癌症的发生密切相关。在癌症分类上，本方法同样具

有较好的分类效果。在生存分析上，本方法在显著性和稳定性方面都要好于基因层次的差异分析方法。

2 研究方法

2.1 实验数据集

本文从TCGA数据库网站(https://tcga-data.nci.nih.gov/tcga)分别下载了乳腺癌(BReast invasive CArcinoma, BRCA)、肝癌(LIver Hepatocellular Carcinoma, LIHC)和子宫内膜癌(Uterine Corpus Endometrial Carcinoma, UCEC)的基因表达数据和异构体表达数据，使用TPM(Transcripts Per Million)来衡量样本的表达丰度。且从TCGA SpliceSeq网站(https://bioinformatics.mdanderson.org/TCGASpliceSeq/)获取相应癌症样本各基因的剪接模式PSI表达数据。由于过低表达量的基因或异构体在癌症调控过程中往往起不到调节作用，且发生可变剪接的基因至少含2个异构体。本文筛选在50%以上样本中表达量至少为0.1 TPM的异构体，再剔除掉只有单个异构体的基因，最终各癌症数据集的样本个数、基因个数和对应的异构体个数如表1所示。

2.2 JS散度

KL散度常被用于描述随机变量的理论分布与真实分布的差异。假设随机离散变量x有k种取值情况，$x\in\{x_1,x_2,\cdots,x_k\}$, $p(x)$与$q(x)$是关于随机离散变量x取值的两个概率分布，则KL散度的计算公式为

$$\mathrm{KL}(p||q)=\sum_{i=1}^{k}p(x_i)\log_2\frac{p(x_i)}{q(x_i)} \tag{1}$$

在信息理论中，KL散度等价于两个概率分布信息熵的差值。KL散度具有非对称性，即$\mathrm{KL}(p||q)\neq\mathrm{KL}(q||p)$。仅当两个概率分布相同时，KL散度为零。反之，当两个概率分布的差别越大，则其KL散度值也越大。

但由于KL散度具有非对称性，这使得用KL散度值来衡量两个概率分布的差异存在一定不足。而KL散度的一种变体JS散度能很好解决非对称性问题，其思想是构造$p(x)$与$q(x)$的平均概率分布来解决该问题，则JS散度的计算公式为

表 1 癌症数据集统计信息

癌症	癌症样本	正常样本	基因个数	异构体个数
BRCA	1100	112	10178	33481
LIHC	373	50	8871	26234
UCEC	117	24	9765	30953

$$\mathrm{JS}(p||q)=\frac{1}{2}\mathrm{KL}\left(p||\frac{p+q}{2}\right)+\frac{1}{2}\mathrm{KL}\left(q||\frac{p+q}{2}\right) \tag{2}$$

JS散度具有对称性，即$\mathrm{JS}(p||q)=\mathrm{JS}(q||p)$。且JS散度的值域范围是[0, 1]，当两个概率分布相同则是0，当两个概率分布相反则是1。当两个概率分布的差别越大，则其JS散度值也越大。相较于KL散度，JS散度无论在对称性还是值域范围，能更确切地判别两个概率分布的差异性。

2.3 基于中值的JS散度可变剪接差异分析方法

在癌症和正常两种状态下，一个基因剪接异构体分布的差异可以反映由于可变剪接种类或模式的差异。因此，本文将利用异构体表达谱数据来研究基因的可变剪接的JS散度。给定一个有k个异构体的基因x, $\boldsymbol{x}^{\mathrm{T}}$和$\boldsymbol{x}^{\mathrm{N}}$分别为在癌症和正常状态的异构体表达量，则很容易根据每个剪接异构体的组成百分比$p({x_i}^{\mathrm{T}})$与$q({x_i}^{\mathrm{N}})$，计算基因x在这对样本中的JS散度。然后，根据基因x的JS散度大小，研究其在癌症和正常两种状态剪接异构体分布差异的显著性。

但是，TCGA数据库的大部分表达谱数据集均为非配对数据，即癌症和正常样本均来自不同的个体，因此，无法直接计算一个基因在两种状态下的可变剪接的JS散度。在统计分析中，由于中值不受极端值的影响，具有稳定性和可靠性的优点，常常用于刻画一个变量的代表值。因此，本文首先根据基因x的k个异构体分别在癌症组和正常组表达量中值，构建基因x在癌症和正常两种状态的代表表达向量$\boldsymbol{x}^{\mathrm{T}}$和$\boldsymbol{x}^{\mathrm{N}}$；然后，计算代表表达向量$\boldsymbol{x}^{\mathrm{T}}$和$\boldsymbol{x}^{\mathrm{N}}$中每个异构体的百分比；最后，计算基因$x$的可变剪接的JS散度。

本文提出的基于中值的JS散度可变剪接差异分析方法步骤如下：

步 骤 1　确定基因x在癌症和正常两种状态下的代表异构体表达向量$\boldsymbol{x}^{\mathrm{T}}$和$\boldsymbol{x}^{\mathrm{N}}$；

步 骤 2　计算每个异构体的百分比$p({x_i}^{\mathrm{T}})$与$q({x_i}^{\mathrm{N}})$

$$p({x^{\mathrm{T}}}_i)=\frac{x_i^{\mathrm{T}}}{\sum_{i=1}^{k}x_i^{\mathrm{T}}},\ q({x^{\mathrm{N}}}_i)=\frac{x_i^{\mathrm{N}}}{\sum_{i=1}^{k}x_i^{\mathrm{N}}} \tag{3}$$

步 骤 3　通过式(1)和式(2)计算基因x的JS散度值；

步 骤 4　重复步骤1—步骤3，对每个基因的JS散度按照从大到小排序。

总体上，本方法的通过中值构造基因各异构体在癌症和正常两种状态下的代表表达向量，在之后的运算中仅需对这对向量进行分析，无需对众多样

本整体分析，使得本文方法的计算复杂度较低。本方法利用JS散度能找寻剪接异构体在整体分布差异较大的基因，避免了从单个基因或单个异构体分析导致忽略了基因内部异构体的整体分布差异。且与从剪接模式差异分析相比，由于一个剪接模式会产生两个剪接异构体，当一个基因仅发生一个剪接模式时，则能考虑到异构体整体分布的差异，当存在多个剪接模式时，则无法考虑到整体分布的差异，所以从剪接模式上考虑差异基因仍存在一定不足。

3　实验结果及分析

3.1　差异基因分析

本文使用基于中值的JS散度可变剪接差异分析方法构建AS Model对癌症进行差异基因分析。同时分别对单个基因、单个异构体和Liu等人[11]方法根据各剪接模式的PSI值均使用t检验构建Gene Model, Isoform Model和PSI Model与AS Model进行对比。其中Gene Model使用基因表达数据，通过校正后的Pvalue(FDR)显著性进行重要差异基因排序。Isoform Model使用了异构体表达数据求得每个异构体的FDR，由于同一基因含多个异构体，则选取FDR值最小的异构体作为该基因的显著性值。同理，PSI Model中同一基因可能存在多个剪接模式，选取FDR值最小的剪接模式作为该基因的显著性值。因为FDR和JS散度值没有可比性，所以本文按排名来评价基因的差异性。

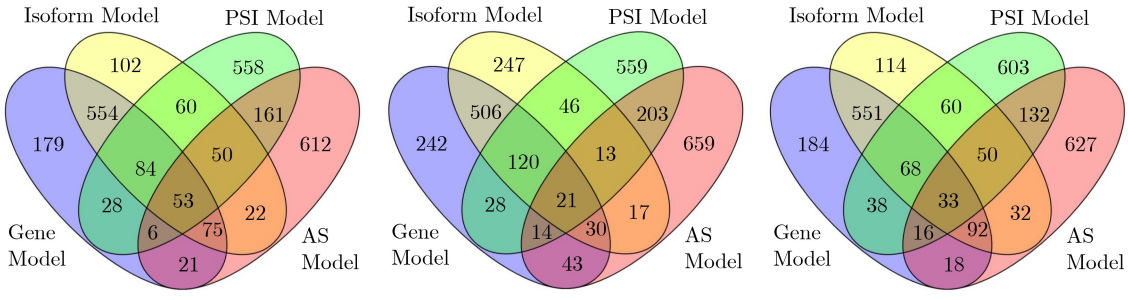
图1为Gene Model, Isoform Model, PSI Model, AS Model, 4种方法在3种癌症数据集上排名靠前1000的差异基因的韦恩图。对应的这些基因在Gene Model, Isoform Model, PSI Model的 FDR分别小于$5.41\times10^{-10}, 3.76\times10^{-12}, 1.47\times10^{-7}$，在AS Model的JS散度均大于0.033。

首先，Gene Model和Isoform Model在检测到的差异基因3种癌症中的重合度均高达约70%，主要由于Isoform Model是根据基因中最显著的异构体来寻差异基因，而通常在基因表达差异较大的情况下，其内部异构体也很大概率上存在较大差异。Isoform Model仍能检测出部分Gene Model无法检测到的差异基因，是由于它们基因层次上不显著差异，但其内部存在显著差异的异构体。一个剪接模式对应产生两个剪接异构体，PSI Model是根据剪接模式找寻差异基因，这使得该模型检测到的差异基因与Isoform Model, AS Model重合度相对较高。AS Model与其它3种方法找寻差异基因的方式不同，这使得找寻的差异基因存在较大的差异。存在大部分使用AS Model检测到的差异基因却无法被Gene Model, Isoform Model和PSI Model检测到，是由于这些基因在基因表达水平上差异不大，且其内部异构体存在一定差异但无明显差异，而AS Model恰能考虑到同一基因各异构体的整体分布差异从而找到该差异基因。因此，从单个异构体的差异出发，无法检测出在剪接异构体整体分布上具有显著差异的基因。

3.2　KEGG通路分析

因为Gene Model, Isoform Model, PSI Model的差异分析原理大致相同，且得到的差异基因重合度较高，两者间的差异仅是使用了不同表达水平的癌症数据集。所以在之后的分析中主要针对基于基因表达谱的Gene Model与本文所提出基于异构体表达谱的AS Model进行比较，分析从基因表达水平和异构体表达水平检测到的差异基因在多方面的异同点。

分别选取Gene Model和AS Model在3种癌症中排名靠前500的差异基因，使用在线网站KOBAS 3.0(http://kobas.cbi.pku.edu.cn)对这些差异基因分别进行KEGG通路分析，来比较Gene Model和AS Model的差异基因在生物学上的通路富集情况。分别按显著性选取排名靠前10的通路进行分析，结果如表2所示。由表2可知，Gene Model的差异基因在3种癌症中主要富集了在与癌症相关的通路，如癌症通路、Rap1信号通路、cAMP信号通路、AMPK信号通路等通路。而AS Model差异基



(a) BRCA中各方法差异基因对比　(b) LIHC中各方法差异基因对比　(c) UCEC中各方法差异基因对比

图 1 4种方法差异基因的韦恩图

表 2 KEGG通路分析

癌症	通路(Gen Model)	通路(AS Model)
BRCA	Focal adhesion	Cell cycle
	PI3K-Akt signaling pathway	p53 signaling pathway
	Tight junction	Pathways in cancer
	Regulation of lipolysis in adipocytes	Oocyte meiosis
	Pathways in cancer	Viral carcinogenesis
	Rap1 signaling pathway	Adherens junction
	cAMP signaling pathway	Purine metabolism
	ABC transporters	PI3K-Akt signaling pathway
	Cell adhesion molecules (CAMs)	Hippo signaling pathway
	Leukocyte transendothelial migration	Metabolic pathways
LIHC	Metabolic pathways	Metabolic pathways
	Fatty acid degradation	Phagosome
	Protein processing in endoplasmic reticulum	Fc gamma R-mediated phagocytosis
	Proteasome	Leishmaniasis
	mTOR signaling pathway	Homologous recombination
	AMPK signaling pathway	Sphingolipid metabolism
	Valine, leucine and isoleucine degradation	ECM-receptor interaction
	Spliceosome	Cell cycle
	Ubiquitin mediated proteolysis	Fanconi anemia pathway
	Insulin signaling pathway	Ribosome biogenesis in eukaryotes
UCEC	Vascular smooth muscle contraction	Osteoclast differentiation
	cGMP-PKG signaling pathway	Cell cycle
	Focal adhesion	Adherens junction
	MAPK signaling pathway	Axon guidance
	Proteoglycans in cancer	Phagosome
	Calcium signaling pathway	Rheumatoid arthritis
	Platelet activation	AMPK signaling pathway
	Adherens junction	PPAR signaling pathway
	Oxytocin signaling pathway	ECM-receptor interaction
	Ras signaling pathway	Platelet activation

因在3种癌症中的共同的通路有细胞周期通路、代谢通路、p53信号通路。已有的研究表明，癌症的产生及恶化与细胞周期的失调密切相关，细胞周期失调导致细胞增生[14]。基于可变剪接差异基因在细胞周期通路的富集表明：可变剪接对细胞周期的精细调控具有重要影响。研究这些剪接异构体相应的蛋白质对细胞周期的影响将有助于癌蛋白质亚型水平破译癌症等复杂疾病的产生机制。对于代谢通路，是由于可变剪接通过控制某些剪接异构体相应的蛋白质表达量，进而调控改变原有的新陈代谢以驱动肿瘤发生。且Inoue等人[15]发现p53信号通路中抑癌基因的异常剪接会促进肿瘤发生。

AS Model差异基因的通路中还有一些与细胞分裂和蛋白质相关的通路，如卵母细胞减数分裂通路、破骨细胞分化通路、蛋白质消化吸收通路、泛素介导的蛋白水解等通路。Munding等人[16]发现在减数分裂的不同阶段，有些基因的可变剪接效率会发生着改变。Chu等人[17]发现Fbx4基因的剪接异构体会干扰人类癌症中的细胞周期蛋白D1蛋白的水解。总的来说，AS Model差异基因富集的通路不仅与癌症相关，同样与可变剪接存在密切联系。

3.3 驱动基因分析

“驱动基因”是指那些对癌症的发生、发展起到关键“驱动”作用的基因，现有的驱动基因数据库[18]共收集了299个驱动基因。由于驱动基因和可变剪接都与癌症存在着千丝万缕的联系，本文尝试

将发生差异可变剪接的驱动基因关联起来。即将AS Model中排名靠前1000的差异可变剪接基因结合驱动基因数据库，分别找寻到属于BRCA, LIHC, UCEC的26, 18, 21个驱动基因。

为了进一步研究这些基因之间的互作关系，本文将BRCA, LIHC和UCEC3种癌症的所关联的驱动基因结合STRING数据库(http://string-db.org/)进行蛋白互作用网络(Protein Protein Interaction network，PPI network)分析。以乳腺癌结果为例，结果如图2所示，AS Model识别的差异驱动基因形成的PPI网络的关系更为密切，其结果要明显好于Gene Model，在另两种癌症中有此结果。由AS Model构建的PPI网络可知，这些基因并不是孤立存在，说明这些属于差异可变剪接的驱动基因并不是单一对癌症的发生、发展起到决定性作用，更多的是通过基因间的相互作用，形成调控网络共同对癌症进行调节。总的来说，有一部分癌症驱动基因的可变剪接紊乱是诱发癌症产生的重要因素，且这些基因之间存在密切的联系，可变剪接的紊乱会导致翻译后的蛋白质相互作用发生改变。

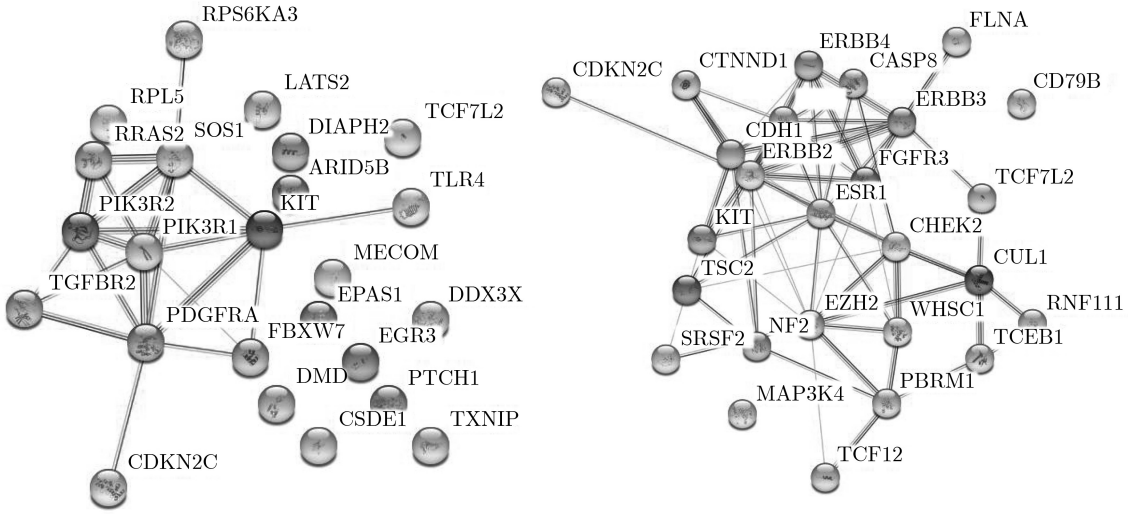
3.4 癌症分类

差异基因对癌症分类有着重要作用，不同于常规使用基因表达谱数据对癌症进行分类，本文尝试使用异构体表达谱数据进行癌症分类。随着特征基因数目的增加，分类精度会趋近于100%，会导致计算量增大和可比性较差，所以选取数目较少且排名靠前的基因进行癌症分类。由于数据样本存在不平衡现象，使用准确度不足以作为分类评分标准，则本文通过交叉验证采用AUC值为癌症分类评分指标。以BRCA和LIHC为例，首先将样本顺序随机打乱，分别选取Gene Model中排名靠前1、前3、前5的差异基因结合基因表达数据，以及AS Model中前1、前3、前5的差异基因对应异构体表达数据作为输入特征。再使分别使用BP神经网络(Back Propagation Neural Network, BPNN)[19]、随机森林(Random Forest, RF)、支持向量机(Support Vector Machine, SVM)[20]、K最近邻(K-Nearest Neighbor, KNN)算法进行五折交叉验证，根据两种方法在各算法中分类的AUC值进行比较，分类结果如图3所示。

由图可知，在对乳腺癌患者和肝癌患者分类中，Gene Mode和AS Model在各算法分类的AUC(Area Under Curve)值均在0.9以上，两种方法在癌症分类中表现差距较小。图3(b)所示在BPNN和SVM算法模型中，分类的AUC值均接近1，具有较好的性能。在RF算法模型中，分类AUC值相对差些，是特征数数目较少对算法分类结果的影响。而KNN算法分类结果较低是由于该算法自身性能较弱。总体上，经过多种算法和多种数据集上的分类验证，AS Model在癌症分类上的表现仍具有较高的精确度和鲁棒性，根据差异可变剪接基因的异构体表达谱数据同样可以应用于癌症诊断。

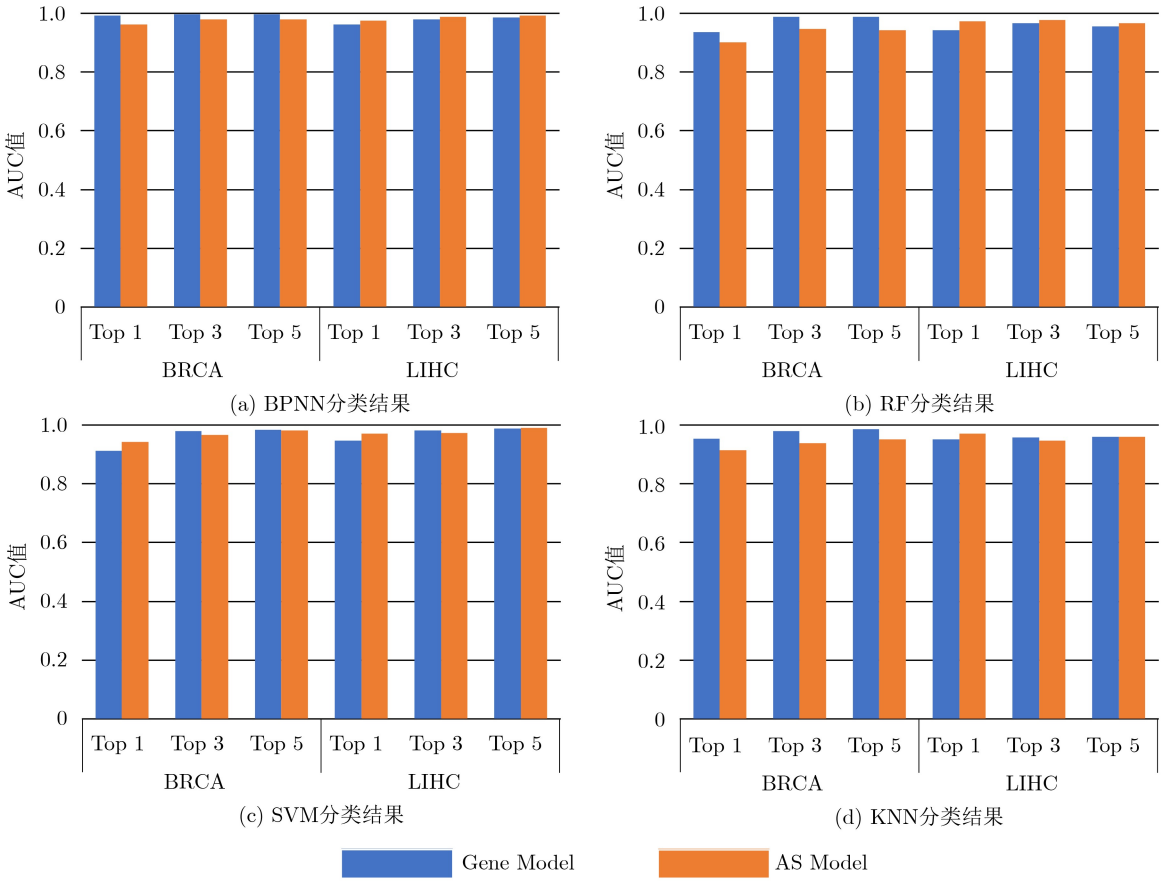
3.5 生存分析

以往的研究表明，显著差异的基因往往与癌症患者的生存周期密切相关。本文分别选取Gene Model和AS Model中排名第一的差异基因，以BRCA和UCEC为例，使用生存分析网站(http://kmplot.com/analysis/)来进行生存分析[21]研究对比。绘制的生存曲线如图4所示，由图可知，显著性方面，AS Model排名第一的差异基因在生存分析中P值均小



(a) Gene Model 乳腺癌驱动基因的PPI网络　　(b) AS Model 乳腺癌驱动基因的PPI网络

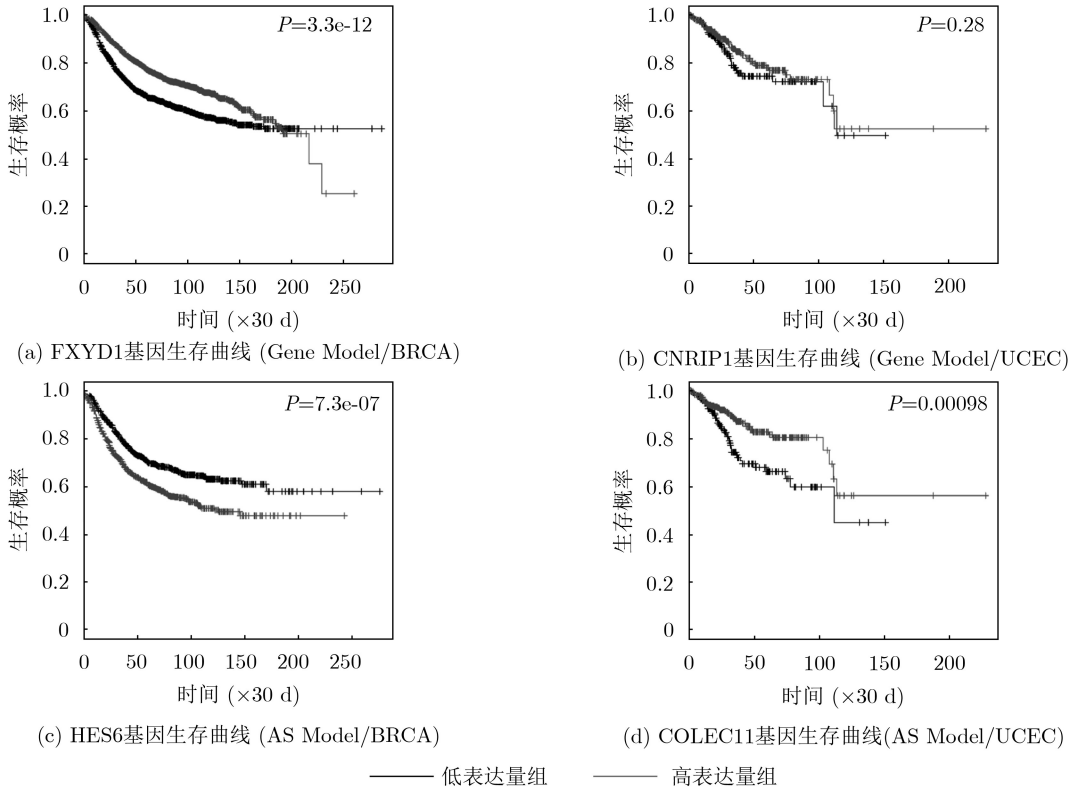
图 2 乳腺癌驱动基因的PPI网络



(a) BPNN分类结果 (b) RF分类结果

(c) SVM分类结果 (d) KNN分类结果

图 3 癌症分类结果比较



(a) FXYD1基因生存曲线 (Gene Model/BRCA)

(b) CNRIP1基因生存曲线 (Gene Model/UCEC)

(c) HES6基因生存曲线 (AS Model/BRCA)

(d) COLEC11基因生存曲线(AS Model/UCEC)

图 4 差异基因生存分析曲线

于0.05。而Gene Model中UCEC的CNRIP1基因在子宫内膜癌中P值大于0.05，未表现出显著差异。从曲线走势来看，在整个生存周期中，AS Model各基因的2组生存曲线均分离明显，表现稳定。而Gene Model中BRCA的FXYD1基因在乳腺癌中，约在第180月时，2组生存曲线交错在一起，稳定性较差。所以从异构体表达水平发现的差异可变剪接基因有着更好的预后价值研究，可以为癌症治疗提够了新的潜在靶基因。

4　结束语

本文提出的基于中值的JS散度可变剪接差异分析方法具有简单、计算复杂度低的优点，可以方便地找寻到在可变剪接整体上发生显著变化的可变剪接紊乱基因。且结果表明：(1)与基于单个异构体的差异分析相比，本文提出的差异分析方法能够发现大量在单个异构体表达差异不显著，但在整体分布上具有显著差异的基因，为进一步研究可变剪接差异与癌症的关系奠定了基础；(2)通路的富集分析揭示，这些在可变剪接整体上具有显著差异的基因不仅仅富集在很多与癌症密切相关的通路，而且富集在一些有可变剪接参与调控的信号通路，如细胞周期通路、代谢通路、p53信号通路。此外，还有一些与细胞分裂和蛋白质相关的通路；(3)癌症驱动基因不仅仅通过变异促进癌症的发展，而且有一部分癌症驱动基因的可变剪接也与癌症的发生密切相关；(4)与基因层次的显著差异分析相比，可变剪接显著差异基因在生存分析方面具有更好的结果，因此，有可能成为更具潜力的癌症标志物。

参 考 文 献

[1] WANG E T, SANDBERG R, LUO Shujun, *et al*. Alternative isoform regulation in human tissue transcriptomes[J]. *Nature*, 2008, 456(7221): 470–476. doi: 10.1038/nature07509.

[2] ZHOU Yujie, ZHU Guiqi, ZHANG Qingwei, *et al*. Survival-associated alternative messenger RNA splicing signatures in pancreatic ductal adenocarcinoma: A study based on RNA-sequencing data[J]. *DNA and Cell Biology*, 2019, 38(11): 1207–1222. doi: 10.1089/dna.2019.4862.

[3] XIE Zucheng, WU Huayu, DANG Yiwu, *et al*. Role of alternative splicing signatures in the prognosis of glioblastoma[J]. *Cancer Medicine*, 2019, 8(18): 7623–7636. doi: 10.1002/cam4.2666.

[4] LI Mingxue, WANG Dun, HE Jianhua, *et al*. Bcl-X_L: A multifunctional anti-apoptotic protein[J]. *Pharmacological Research*, 2020, 151: 104547. doi: 10.1016/j.phrs.2019.104547.

[5] KOLE R, KRAINER A R, and ALTMAN S. RNA therapeutics: Beyond RNA interference and antisense oligonucleotides[J]. *Nature Reviews Drug Discovery*, 2012, 11(2): 125–140. doi: 10.1038/nrd3625.

[6] SONG Jukun, LIU Yongda, SU Jiaming, *et al*. Systematic analysis of alternative splicing signature unveils prognostic predictor for kidney renal clear cell carcinoma[J]. *Journal of Cellular Physiology*, 2019, 234(12): 22753–22764. doi: 10.1002/jcp.28840.

[7] DOU Tonghai, XU Jiaxi, GAO Yuan, *et al*. Evolution of peroxisome proliferator-activated receptor gamma alternative splicing[J]. *Frontiers in Bioscience* (*Elite Edition*), 2010, 2: 1334–1343. doi: 10.2741/e193.

[8] LI Ji, CHOI P S, CHAFFER C L, *et al*. An alternative splicing switch in FLNB promotes the mesenchymal cell state in human breast cancer[J]. *eLife*, 2018, 7: e37184. doi: 10.7554/eLife.37184.

[9] SHEN Shihao, PARK J W, LU Zhixiang, *et al*. rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2014, 111(51): E5593–E5601. doi: 10.1073/pnas.1419161111.

[10] 欧书华, 刘学军, 张礼. 基于KL散度的RNA-Seq数据差异异构体比例检测[J]. 计算机工程与科学, 2017, 39(1): 158–164. doi: 10.3969/j.issn.1007-130X.2017.01.022.
OU Shuhua, LIU Xuejun, and ZHANG Li. Differential isoform ratio detection based on KL divergence for RNA-Seq data[J]. *Computer Engineering & Science*, 2017, 39(1): 158–164. doi: 10.3969/j.issn.1007-130X.2017.01.022.

[11] LIU Jingwei, LI Hao, SHEN Shixuan, *et al*. Alternative splicing events implicated in carcinogenesis and prognosis of colorectal cancer[J]. *Journal of Cancer*, 2018, 9(10): 1754–1764. doi: 10.7150/jca.24569.

[12] ZONG Zhen, LI Hui, YI Chenghao, *et al*. Genome-wide profiling of prognostic alternative splicing signature in colorectal cancer[J]. *Frontiers in Oncology*, 2018, 8: 537. doi: 10.3389/fonc.2018.00537.

[13] ZHANG Zijun, PAN Zhicheng, YING Yi, *et al*. Deep-learning augmented RNA-seq analysis of transcript splicing[J]. *Nature Methods*, 2019, 16(4): 307–310. doi: 10.1038/s41592-019-0351-9.

[14] WHITFIELD M L, GEORGE L K, GRANT G D, *et al*. Common markers of proliferation[J]. *Nature Reviews Cancer*, 2006, 6(2): 99–106. doi: 10.1038/nrc1802.

[15] INOUE K and FRY E A. Aberrant splicing of the DMP1-ARF-MDM2-p53 pathway in cancer[J]. *International Journal of Cancer*, 2016, 139(1): 33–41. doi: 10.1002/ijc.30003.

[16] MUNDING E M, SHIUE L, KATZMAN S, *et al*.

Competition between pre-mRNAs for the splicing machinery drives global regulation of splicing[J]. *Molecular Cell*, 2013, 51(3): 338–348. doi: 10.1016/j.molcel.2013.06.012.

[17] CHU Xiufeng, ZHANG Ting, WANG Jie, *et al*. Alternative splicing variants of human Fbx4 disturb cyclin D1 proteolysis in human cancer[J]. *Biochemical and Biophysical Research Communications*, 2014, 447(1): 158–164. doi: 10.1016/j.bbrc.2014.03.129.

[18] BAILEY M H, TOKHEIM C, PORTA-PARDO E, *et al*. Comprehensive characterization of cancer driver genes and mutations[J]. *Cell*, 2018, 173(2): 371–385. e18. doi: 10.1016/j.cell.2018.02.060.

[19] 曾勇, 舒欢, 胡江平, 等. 基于BP神经网络的自适应伪最近邻分类[J]. 电子与信息学报, 2016, 38(11): 2774–2779. doi: 10.11999/JEIT160133.

ZENG Yong, SHU Huan, HU Jiangping, *et al*. Adaptive pseudo nearest neighbor classification based on BP neural network[J]. *Journal of Electronics & Information Technology*, 2016, 38(11): 2774–2779. doi: 10.11999/JEIT160133.

[20] 陈素根, 吴小俊. 基于特征值分解的中心支持向量机算法[J]. 电子与信息学报, 2016, 38(3): 557–564. doi: 10.11999/JEIT150693.

CHEN Sugen and WU Xiaojun. Eigenvalue proximal support vector machine algorithm based on eigenvalue decoposition[J]. *Journal of Electronics & Information Technology*, 2016, 38(3): 557–564. doi: 10.11999/JEIT150693.

[21] ZHANG Yangjun, YAN Libin, ZENG Jin, *et al*. Pan-cancer analysis of clinical relevance of alternative splicing events in 31 human cancers[J]. *Oncogene*, 2019, 38(40): 6678–6695. doi: 10.1038/s41388-019-0910-7.

刘文斌：男，1969年生，教授，研究方向为生物信息学.
王　兵：男，1993年生，硕士生，研究方向为生物信息学.
方　刚：男，1969年生，教授，研究方向为生物信息学.
石晓龙：男，1975年生，教授，研究方向为生物信息学.
许　鹏：男，1986年生，博士后，研究方向为生物信息学.