基于可变剪接紊乱的乳腺癌亚型预测分析

许 鹏[①③] 王 兵[②] 方 刚[①] 石晓龙[①] 刘文斌*[①②]

[①](广州大学计算科技研究院 广州 510006)

[②](温州大学计算机与人工智能学院 温州 325035)

[③](黔南民族师范学院计算机与信息学院 都匀 558000)

摘 要：可变剪接与多种复杂疾病的发生、发展存在密切的联系，包括肿瘤在内的多种疾病的产生往往伴随着可变剪接的紊乱发生。现有的乳腺癌亚型分析主要是基于单个剪接异构体出发，缺少考虑亚型之间由于可变剪接紊乱造成剪接异构体在整体分布上的差异。因此该文提出了基于可变剪接紊乱的乳腺癌亚型预测方法，主要使用Jensen-Shannon(JS)散度来找寻亚型之间的可变剪接紊乱差异较大的基因，并构建反向传播(BP)神经网络模型对乳腺癌亚型进行分类。结果表明，该方法不仅能有效发现肿瘤异质性分子，在乳腺癌亚型分类方面也有较好的识别结果，其平均F1值达到0.89，且能为患者提供个性化乳腺癌亚型药物推荐。该文的研究将有效促进基于可变剪接紊乱的乳腺癌亚型研究的发展。

关键词：乳腺癌亚型预测；可变剪接；Jensen-Shannon散度；反向传播神经网络；药物推荐

中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2020)06-1348-08

DOI: 10.11999/JEIT190871

Analysis of Breast Cancer Subtypes Prediction Based on Alternative Splicing Disorders

XU Peng[①③] WANG Bing[②] FANG Gang[①] SHI Xiaolong[①] LIU Wenbin[①②]

[①](*Institute of Computing Science and Technology*, *Guangzhou University*, *Guangzhou* 510006, *China*)

[②](*College of Computer Science and Artificial Intelligence*, *Wenzhou University*, *Wenzhou* 325035, *China*)

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

Abstract: Alternative splicing is closely related to the occurrence and development of a variety of complex diseases, the emergence of various diseases including tumors is often accompanied by the occurrence of alternative splicing disorders. The existing analysis of breast cancer subtypes is mainly based on single splicing isoform, and the difference in the overall distribution of splicing isoforms caused by alternative splicing disorders among subtypes is not considered. Therefore, a prediction method of breast cancer subtypes based on alternative splicing disorders is proposed, which mainly uses Jensen-Shannon(JS) divergence to find genes with large differences in alternative splicing disorders between subtypes, then constructes Back Propagation(BP) neural network model to classify breast cancer subtypes. The results show that this method could not only effectively detect tumor heterogeneous molecules, but also had good identification results in the classification of breast cancer subtypes, with an average F1-score of 0.89, and could provide personalized drug recommendations for patients with breast cancer subtypes. This study will effectively promote the development of breast cancer subtypes based on alternative splicing disorders.

Key words: Breast cancer subtypes predicion; Alternative splicing; Jensen-Shannon(JS) divergence; Back Propagation(BP) neural network; Drug recommendation

收稿日期：2019-11-01；改回日期：2020-05-10；网络出版：2020-05-23

*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

可变剪接(Alternative Splicing, AS)是指有些基因的一个前体mRNA通过不同的剪接方式产生不同mRNA剪接异构体(isoform)的过程[1]。在基因翻译成相应蛋白质的过程中，首先由DNA转录成前体mRNA，再由前体mRNA经历内含子(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，最后用于指导蛋白质的翻译。而可变剪接使得从前体mRNA到成熟mRNA的过程变得异常复杂，同一基因往往会因此形成多种剪接异构体，最终翻译成结构、功能相异的亚型蛋白。最新研究表明，人类约95%的多外显子基因都会发生可变剪接[2]，已知的人类遗传疾病里，约35%来自于剪接遗传。总之，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，而且对细胞的增殖、分化、发育、凋亡等一系列重要的生物过程的精细调控具有非常重要的作用[3]。

生物研究表明，在特定的组织器官中或生理状态下，某些基因的可变剪接存在规律性的变化。所以某些基因存在特异性的可变剪接模式，与细胞所处的环境或状态相适应，以完成特定的生物学功能[3]。越来越多的证据表明，包括癌症在内的多种疾病的发生、发展都与可变剪接的异常调节密切相关[4]。如脊髓性肌肉萎缩症(Spinal Muscular Atrophy, SMA)[5]、肺癌相关的超大型B细胞淋巴瘤(B-cell lymphoma extra-Large, Bcl-xL)[6]、人类上皮增长因子受体(Human Epidermal growth factor Receptor, HER-2)[7]等。可变剪接紊乱是指原有的剪接异构体比例失衡，会使某些异构体表达量发生改变，甚至产生新的异构体。本质上，某些紊乱的可变剪接会导致基因调控作用的改变和异常蛋白质的翻译，从而改变原有的生命进程，促使肿瘤的发生[8,9]。因此，找出异常剪接与特定疾病间的关系，有利于疾病的诊断和治疗，进一步揭示疾病发生、发展的机制。

1999年国家癌症研究所(National Cancer Institute, NCI)最早提出肿瘤分子分型。肿瘤的分子分层是发展个性化治疗的关键，发现肿瘤异质性分子是个性化和有效的癌症治疗手段[10]。而现有大部分癌症研究都是基于基因表达谱的分析，忽略了人类转录组的复杂性，缺少从更深层次的可变剪接的角度来深入癌症亚型层次的研究。目前有研究学者通过统计学、机器学习建模方法来实现癌症亚型分型预测。Pal等人[11]基于异构体表达谱数据，使用非负矩阵分解(Non-negative Matrix Factorization, NMF)来实现胶质瘤亚型分类[12]。Zhao等人[13]通过构造线性回归模型来检测发生异构体交换(isoform switching)的基因，进而用于乳腺癌亚型聚类分析[14]。Stricker等人[15]通过假设检验筛选与乳腺癌亚型相关的异构体，再使用逻辑回归(Logistics Regression, LR)对乳腺癌亚型分类[16]。

乳腺癌是全世界女性恶性肿瘤致死的主要原因，是当前社会重大公共健康问题之一[17]。由于乳腺癌的异质性和复杂性，在临床、病理和分子方面不同的乳腺癌亚型均表现出很大的区别。根据Perou等人[12]和Sorlie等人[14]提出的乳腺癌分子分类，目前主流是将乳腺癌分为5个亚型，三阴性乳腺癌(Basal)、人表皮生长因子受体2(Her2)、导管A性(LumA)、导管B性(LumB)和正常乳腺亚型(Normal)。其中，三阴性乳腺癌的恶化和转移速度最快，属于侵袭性乳腺癌。Her2型的患者容易发生腋窝淋巴结处转移。LumA型乳腺癌对内分泌的治疗方式十分敏感[14]。LumB型患者的激素表达要低于LumA型，主要是高龄患者。因此，利用乳腺癌的高度异质性，实现乳腺癌亚型分型有利于诊断、预后和治疗方法的选择，且是各种相关研究的基础。

本质上，可变剪接紊乱会直接导致基因各异构体所占比例的发生变化。现有的研究仅是从单个剪接异构体或一种剪接模式的出发，缺少从整体上认识可变剪接紊乱对乳腺癌异质性的影响。考虑现有方法的不足，本文通过在整体上分析基因各异构体的差异，提出了基于可变剪接紊乱的乳腺癌亚型预测方法。该方法主要分为两个部分，可变剪接紊乱基因的筛选和乳腺癌亚型分类模型的构建。由于(JS)Jensen-Shannon散度，能很好地衡量两个概率分布的差异性，本文使用JS散度来筛选两两亚型之间重要的可变剪接紊乱基因。之后将筛选基因对应的异构体表达数据作为输入特征，构建基于反向传播(Back Propagation, BP)神经网络[16]的乳腺癌亚型分类模型进行训练。本文通过JS散度值分析了亚型之间的差异性，实现了乳腺癌亚型的聚类、分类，并结合差异情况分析了聚类和分类结果。最后，根据不同乳腺癌亚型患者基因的可变剪接紊乱程度，为患者找寻重要的癌症亚型标志物和提供个性化乳腺癌亚型药物推荐。

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)下载了乳腺癌(BReast invasive CArcinoma, BRCA)患者异构体表达数据，使用每百万转录本(Transcripts Per Million, TPM)来衡量样本的表达丰度。其中共有20531个基因，73599个

异构体。为了避免低表达量异构体的干扰，且考虑到发生可变剪接的基因至少含两个异构体，本文筛选在50%以上样本中表达量至少为0.1 TPM的异构体，再剔除掉只含单个异构体的基因，最终保留33481个异构体，对应的基因个数为10178个。从UCSC xena网站(https://xenabrowser.net/datapages/)下载相应样本的临床数据，对于各种乳腺癌亚型样本数如表1所示。

表 1 不同乳腺癌亚型的样本数

乳腺癌亚型	样本数
Basal	140
Her2	67
LumA	432
LumB	194
Normal	117

2.2 JS散度

JS散度是Kullback-Leibler(KL)散度的一种变体，它可以用来表示两个概率分布之间的差异。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散度和JS散度的计算公式分别为

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

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

JS散度利用了两个KL散度的叠加，且使用到了$p(x)$与$q(x)$的平均概率分布。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散度解决了KL散的非对称性问题，且其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JS散度的可变剪接紊乱基因分析方法

对于已有5组亚型样本，同一基因各异构体在不同组样本中的分布差异可以反映该基因可变剪接的紊乱程度。而JS散度能很好的用于刻画两个概率分布的差异，所以本文采用JS散度来计算同一基因各异构体在任意两种亚型中的分布差异。假设存在一基因x有k个剪接异构体，$\boldsymbol{x}^{\mathrm{A}}$和$\boldsymbol{x}^{\mathrm{B}}$表示该基因在某两种亚型样本中的异构体表达量，可以分别求出每个异构体在该基因中所占比例$p(x_i^{\mathrm{A}})$与$q(x_i^{\mathrm{B}})$，则可以计算基因x在对应两种亚型状态下的JS散度。最后，可以根据基因x的JS散度大小，研究其在两种亚型下剪接异构体分布的差异性。

但TCGA中大部分样本数据均为非配对样本，即任意两种乳腺癌亚型的样本很有可能来自不同的个体。我们无法直接计算一个基因在两种亚型状态下的可变剪接的JS散度。在统计分析中，由于中值有稳定性和可靠性的优点，且不受极端值的影响，常常被用作为一组数据的代表值。因此，本文首先根据基因x的k个异构体分别在某两种亚型样本中的表达量中值，构建基因x在两种状态的代表表达向量$\boldsymbol{x}^{\mathrm{A}}$和$\boldsymbol{x}^{\mathrm{B}}$；然后，计算代表表达向量$\boldsymbol{x}^{\mathrm{A}}$和$\boldsymbol{x}^{\mathrm{B}}$中每个异构体的比例；最后，计算基因$x$的可变剪接的JS散度。

本文所提基于JS散度的可变剪接紊乱基因分析方法步骤如下：

(1) 确定基因x在某两种亚型状态下的代表异构体表达向量$\boldsymbol{x}^{\mathrm{A}}$和$\boldsymbol{x}^{\mathrm{B}}$；

(2) 计算每个异构体所占比例$p(x_i^{\mathrm{A}})$与$q(x_i^{\mathrm{B}})$

$$p(x_i^{\mathrm{A}})=\frac{x_i^{\mathrm{A}}}{\sum_{i=1}^{k}x_i^{\mathrm{A}}},\ q(x_i^{\mathrm{B}})=\frac{x_i^{\mathrm{B}}}{\sum_{i=1}^{k}x_i^{\mathrm{B}}} \tag{3}$$

(3) 通过式(1)、式(2)计算两种亚型状态下基因x的JS散度值

2.4 基于可变剪接紊乱的乳腺癌亚型分类模型

由于共有5组类型样本，根据2.3节的计算方法，计算两两类型之间的JS散度则可得到10组基因在两种状态下的JS散度，本文选择这10组基因中JS散度大于0.3的基因集合对应的异构体作为特征。在模型设计上，首先将样本顺序打乱，使用5折交叉验证，其中先求得训练集各特征的均值与标准差，采用求得的均值和标准差将训练集和测试集各特征进行z-score标准化。训练模型采用含3层隐藏层的BP神经网络，3个隐藏层的节点数分别为2000, 5000和100，使用的激活函数为线性整流函数(Rectified Linear Unit, ReLU)，并重复100次5折交叉验证，则950个样本每个样本均有100个预测结果，最后对每个样本的100个预测结果选择众数作为该样本的最终预测结果。

基于此，本文提出了基于可变剪接紊乱的乳腺癌亚型分类模型，主要步骤如下：

(1) 根据2.3节中方法求得乳腺癌两两亚型间的JS散度；

(2) 选择JS散度大于0.3的基因作集合，提取基因集合对应的异构体表达数据作为输入特征；

(3) 构建BP神经网络分类模型，设置隐藏层和激活函数；

(4) 将样本顺序打乱，使用5折交叉验证划分训练集和测试集并标准化，使用(3)中模型进行训练预测出每个样本的类型；

(5) 重复(4)100次，对于每个样本的100个预测结果选择众数作为该样本的最终预测结果。

3　实验结果及分析

3.1　JS散度分布

由于JS散度的值域在0～1之间，这使得即使不同基因的异构体数目不唯一，但仍可以很好地比较不同基因的可变剪接紊乱程度。为了比较5组乳腺癌样本两两间各基因可变剪接的差异，其中共有10组结果，本文选择这10组中排名靠前100的差异可变剪接基因，根据各自的JS散度值作出箱线图，如图1所示。由图1可知，排名靠前100基因的JS散度主要集中在0.1～0.5之间。当某组JS散度整体偏小时，则表示对应的两种乳腺癌亚型的可变剪接差异较小，反之则较大。可以看出Normal与LumA, LumA与LumB, LumA与Her2, LumB与Her2的可变剪接的差异程度较小，而其它乳腺亚型之间差异程度较大。

为进一步观察不同乳腺癌亚型与正常乳腺亚型之间差异基因的交并情况，本文分别筛选了不同乳腺癌亚型与正常型之间JS散度大于0.3的基因，并作出韦恩图如图2所示。由图2可知，LumA与Her2型乳腺癌的特异性差异可变剪接基因数目较少，仅分别为3个、9个。而Basal与LumB型乳腺癌异质性基因相对较多。4种亚型公共的差异基因仍有12个，说明乳腺癌亚型之间仍存在一定的相似性。

3.2　乳腺癌亚型聚类

本文选择了10组结果中JS散度大于0.3的基因进一步分析，其筛选结果如图3所示。并将所有JS散度大于0.3的基因作集合，最终共有160个基因，对应共483个剪接异构体。由于异构体表达数据并不呈现正态分布，直接使用进行层次聚类作图会导致展示效果很差。因此，本文首先将这483个异构体表达数据首先进行$\log_2(x+0.1)$转化，使其大致呈现正太分布，再进行z-score标准化，使得每个异构体转化为标准正太分布[13]。最终，将处理后的异构体表达数据进行层次聚类作图，其结果如图4所示，其中列为样本，行为异构体。

由图4可以看出，总体上5种类型的乳腺癌样本根据层次聚类具有一定的可分辨性。其中，Basal的聚类效果最好，结合图3可知，是由于Basal与其它类型的差异可变剪接程度较大。Normal的聚类效果也相对较好，仅有部分LumA型样本与其聚在一起，是由于Normal与LumA间JS散度大于0.3的基因仅有25个，二者间的可变剪接差异相对较小。LumB和Her2的聚类效果较差，是由于LumA与LumB, LumA与Her2, LumB与Her2间筛选后的差异基因仅分别为12个、23个、10个，这主要使得LumB与LumA型样本易聚在一起，Her2与LumB型样本易聚在一起。

3.3　乳腺癌亚型分类

本文根据所提方法进行训练，最终各乳腺癌亚型分类的准确率如图5所示和分类的精确率、召回率、F1值如表2所示。结合图表可知，Basal, LumA和Normal的准确率较高，均在0.91之上。而Her2和LumB识别率较低，其中67个Her2型样本中有11个被预测成LumB，这使得Her2的召回率仅为0.75。194个LumB型样本中有39个样本被预测成LumA，且错误地将11个Her2样本和21个LumA样本预测成LumB，所以LumB的精确率和召回率分别仅有0.81和0.79。总体上，乳腺癌各亚型分类的F1值都在0.79之上，其中Basal的F1值更是高达0.97，且通过计算总的平均F1值达到0.89。所以基于可变剪接紊乱的乳腺癌亚型预测很好的可行性。

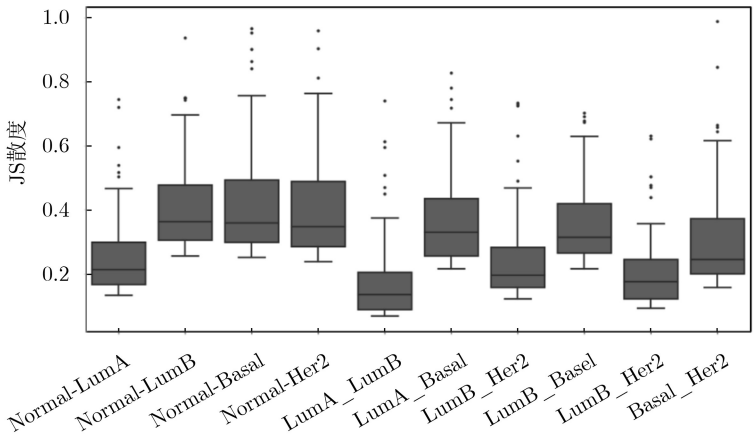


图 1 排名靠前100基因的JS散度分布情况

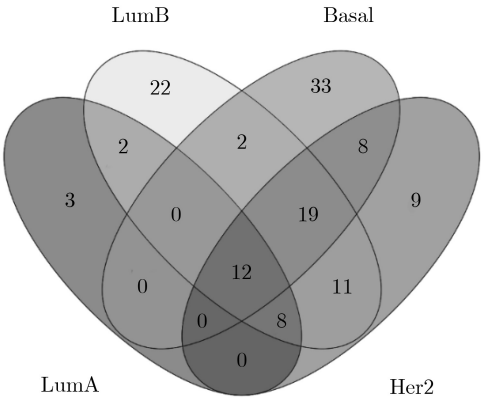


图 2 不同乳腺癌亚型与正常型之间差异基因的韦恩图

3.4 基于乳腺癌亚型的药物推荐

对于乳腺癌亚型患者的精准用药是有效治疗乳腺癌的一大难题，实现合理的乳腺癌亚型药物推荐具有重要临床意义。基于本文所提方法，首先从DrugBank数据库网站(https://www.drugbank.ca/)[18]下载了与乳腺癌相关的药物和药物对应的靶基因数据。然后，分别选择正常型与其它乳腺癌亚型间JS散度大于0.3的基因并作集合得到129个差异可变剪接基因。本文将靶基因和这129个基因作交集，发现8个公共基因。如表3所示，展示了基于这8个基因相应的药物推荐和这些基因在不同乳腺癌亚型中与正常型之间JS散度。由表3可知，这8个基因均在Basal型乳腺癌患中发生的可变剪接紊乱程度较大，所以可推荐的药物种类较多。主要是因为Basal型乳腺癌侵袭性较强，死亡率相对较高，使得Basal型乳腺癌成为主要研究对象。对于LumA型乳腺癌，发生可变剪接紊乱的基因最少，适合推荐的药物有Enzastaurin[19], AT9283[20]。同样在现实中，LumA型乳腺癌的预后效果在乳腺癌中是最好的，一般仅需针对内分泌的治疗即可，所以可变剪接紊

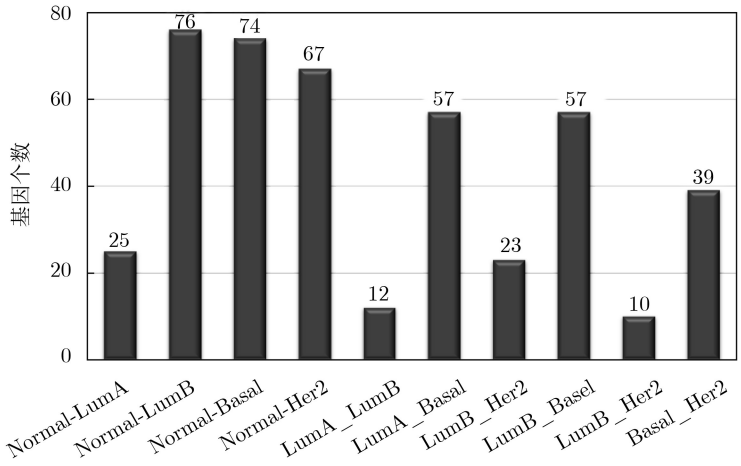


图 3 JS散度大于0.3的基因个数

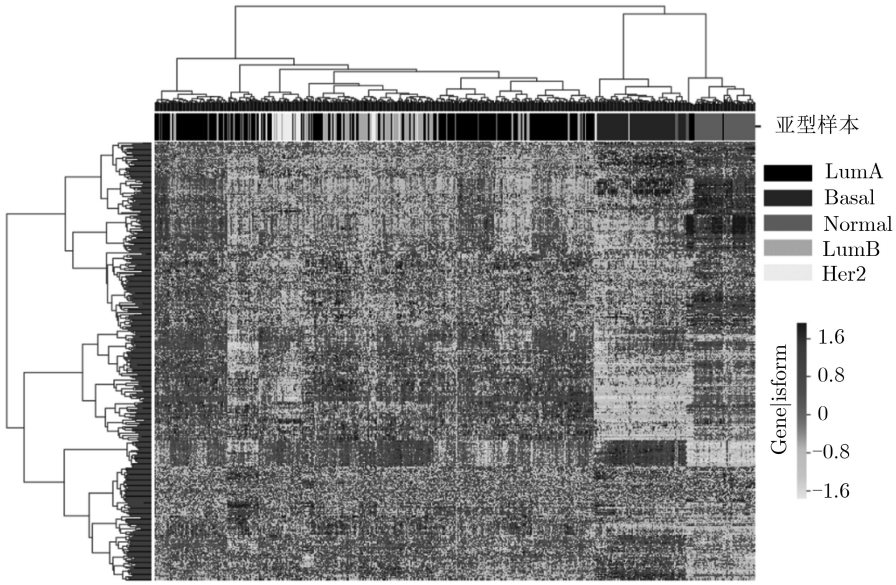
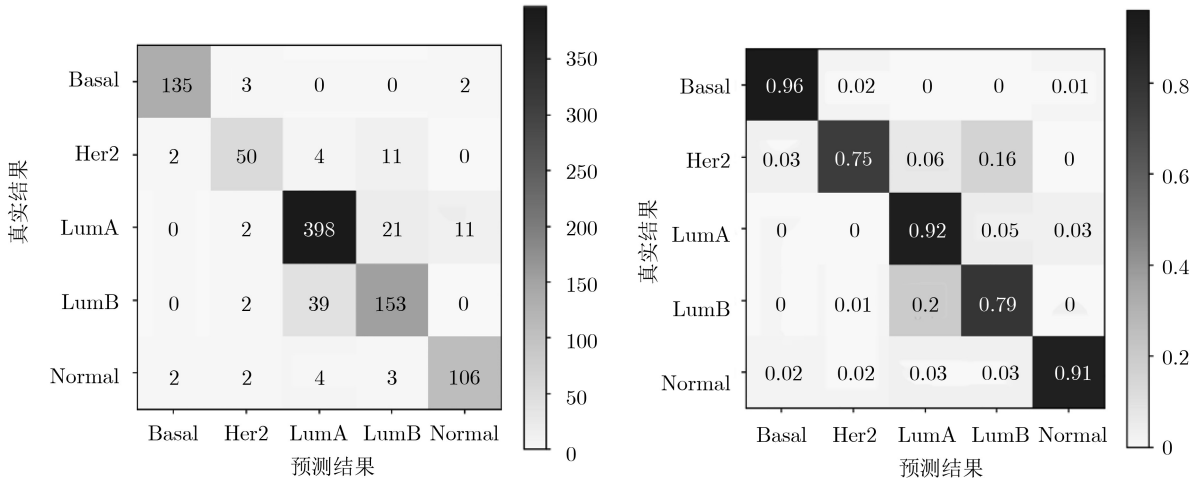


图 4 乳腺癌亚型聚类



(a) 混淆矩阵　　(b) 归一化的混淆矩阵

图 5 乳腺癌亚型分类结果

表 2　乳腺癌亚型分类

乳腺癌亚型	精确率	召回率	F1值
Basal	0.97	0.96	0.97
Her2	0.85	0.75	0.79
LumA	0.89	0.92	0.91
LumB	0.81	0.79	0.80
Normal	0.89	0.91	0.90

乱的基因相对较少。因此，可以根据基因的可变剪接紊乱程度为患者找寻重要的癌症亚型标志物，为患者提供个性化乳腺癌亚型药物推荐。

4　结束语

近年来随着新一代测序技术的发展，为人们从可变剪接角度认识疾病分子机理提供了海量的数据资源。从基因水平深入剪接异构体水平的研究必将进一步揭示人类基因组深层的表达调控机制[21]。因此，通过对可变剪接的研究能加深人们对疾病发生、发展的认识，为癌症患者的诊断、治疗找寻重要的肿瘤异质性分子。

由于可变剪接紊乱是多种疾病产生的重要因素，且目前主流是仅考虑单个异构体的差异，本文提出了基于可变剪接紊乱的乳腺癌亚型预测方法。其思想是通过JS散度计算两种乳腺癌亚型中基因各异构体的整体比例分布差异，再根据JS散度值的大小研究基因可变剪接紊乱。之后筛选可变剪接紊乱较大基因对应的异构体表达数据，使用构建基于BP神经网络的乳腺癌亚型分类模型进行训练。结果表明，乳腺癌亚型间存在一定的相似性和异质性，如LumA和LumB型乳腺癌具有较高的相似性，Basal型乳腺癌与其它亚型存在较高的异质性，且具有较高的复杂性。从差异可变剪接角度来进行乳腺癌亚分类取得了较好的结果，其中分类结果更是达到了0.89，且与现实理论大致吻合，证明了本文方法可行性，为亚型的预测提供了新的方法。最后，本方法通过比较基因的可变剪接紊乱程度能为患者找寻重要的癌症亚型标志物和提供个性化乳腺癌亚型药物推荐。

表 3　乳腺癌亚型的药物推荐

靶基因	药物	Basal	Her2	LumA	LumB
CHEK1	Enzastaurin	**0.393**	**0.369**	0.195	**0.316**
ESR1	Melatonin,Homosalate,Estradiol,2-Amino-1-methyl-6-phenylimidazo(4,5-b)pyridine,Danazol,Fulvestrant,Raloxifene,Custirsen,Tamoxifen,Estrone sulfate,Methyltestosterone,Fluoxymesterone,Afimoxifene	**0.305**	0.133	0.028	0.025
FOLR2	Folic acid,Methotrexate	**0.842**	0.025	0.108	**0.354**
GPER1	Estradiol	**0.442**	**0.438**	0.021	0.013
GSN	Latrunculin A	**0.443**	**0.419**	0.224	**0.476**
PPARG	Curcumin,Isoflavone,Valproic acid,Mesalazine,Nabiximols,Cannabidiol	**0.668**	**0.645**	0.030	**0.637**
AURKB	Enzastaurin,AT9283	**0.640**	**0.569**	**0.352**	**0.591**
ABCC11	Methotrexate,Folic acid	**0.431**	0.036	0.013	0.040

本文所提方法设计简单，考虑了整体上发生可变剪接紊乱的基因，为乳腺癌亚型分型、找寻肿瘤异质性分子和个性化亚型药物推荐提供了新的思路。但本文方法仍在一定不足，主要体现在乳腺癌亚型聚类方面，如LumA和LumB亚型难以区分。后续我们将进一步从可变剪接角度为相似度较高的亚型找寻更有潜力的肿瘤异质性分子，揭示更多亚型层次的可变剪接分子机制。

参 考 文 献

[1] ULE J and BLENCOWE B J. Alternative splicing regulatory networks: Functions, mechanisms, and evolution[J]. *Molecular Cell*, 2019, 76(2): 329–345. doi: 10.1016/j.molcel.2019.09.017.

[2] PAN Qun, SHAI O, LEE L J, *et al*. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing[J]. *Nature Genetics*, 2008, 40(12): 1413–1415. doi: 10.1038/ng.259.

[3] HIRSCH C L, AKDEMIR Z C, WANG Li, *et al*. Myc and SAGA rewire an alternative splicing network during early somatic cell reprogramming[J]. *Genes & Development*, 2015, 29(8): 803–816. doi: 10.1101/gad.255109.114.

[4] VENABLES J P. Aberrant and alternative splicing in cancer[J]. *Cancer Research*, 2004, 64(21): 7647–7654. doi: 10.1158/0008-5472.CAN-04-1910.

[5] NARYSHKIN N A, WEETALL M, DAKKA A, *et al*. *SMN2* splicing modifiers improve motor function and longevity in mice with spinal muscular atrophy[J]. *Science*, 2014, 345(6197): 688–693. doi: 10.1126/science.1250127.

[6] GILLINGS A S, BALMANNO K, WIGGINS C M, *et al*. Apoptosis and autophagy: BIM as a mediator of tumour cell death in response to oncogene-targeted therapeutics[J]. *The FEBS Journal*, 2009, 276(21): 6050–6062. doi: 10.1111/j.1742-4658.2009.07329.x.

[7] MENON R and OMENN G S. Proteomic characterization of novel alternative splice variant proteins in human epidermal growth factor receptor 2/neu-induced breast cancers[J]. *Cancer Research*, 2010, 70(9): 3440–3449. doi: 10.1158/0008-5472.CAN-09-2631.

[8] DAVID C J and MANLEY J L. Alternative pre-mRNA splicing regulation in cancer: Pathways and programs unhinged[J]. *Genes & Development*, 2010, 24(21): 2343–2364. doi: 10.1101/gad.1973010.

[9] BLACK D L. Mechanisms of alternative pre-messenger RNA splicing[J]. *Annual Review of Biochemistry*, 2003, 72: 291–336. doi: 10.1146/annurev.biochem.72.121801.161720.

[10] ALMENDRO V, MARUSYK A, and POLYAK K. Cellular heterogeneity and molecular evolution in cancer[J]. *Annual Review of Pathology: Mechanisms of Disease*, 2013, 8: 277–302. doi: 10.1146/annurev-pathol-020712-163923.

[11] PAL S, BI Yingtao, MACYSZYN L, *et al*. Isoform-level gene signature improves prognostic stratification and accurately classifies glioblastoma subtypes[J]. *Nucleic Acids Research*, 2014, 42(8): e64. doi: 10.1093/nar/gku121.

[12] PEROU C M, SØRLIE T, EISEN M B, *et al*. Molecular portraits of human breast tumours[J]. *Nature*, 2000, 406(6797): 747–752. doi: 10.1038/35021093.

[13] ZHAO Wei, HOADLEY K A, PARKER J S, *et al*. Identification of mRNA isoform switching in breast cancer[J]. *BMC Genomics*, 2016, 17: 181. doi: 10.1186/s12864-016-2521-9.

[14] SØRLIE T, TIBSHIRANI R, PARKER J, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets[J]. *Proceedings of the National Academy of Sciences of the United States of America*, 2003, 100(14): 8418–8423. doi: 10.1073/pnas.0932692100.

[15] STRICKER T P, BROWN C D, BANDLAMUDI C, *et al*. Robust stratification of breast cancer subtypes using differential patterns of transcript isoform expression[J]. *PLoS Genetics*, 2017, 13(3): e1006589. doi: 10.1371/journal.pgen.1006589.

[16] 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.

[17] AKRAM M, IQBAL M, DANIYAL M, *et al*. Awareness and current knowledge of breast cancer[J]. *Biological Research*, 2017, 50: 33. doi: 10.1186/s40659-017-0140-9.

[18] 刘文斌, 陈杰, 方刚, 等. 基于药物互作网络的协同与拮抗预测研究[J]. 电子与信息学报, 2020, 42(6): 1428–1435. doi: 10.11999/JEIT190867.
LIU Wenbin, CHEN Jie, FANG Gang, *et al*. Prediction of drug synergy and antagonism based on drug-drug interaction network[J]. *Journal of Electronics & Information Technology*, 2020, 42(6): 1428–1435. doi: 10.11999/JEIT190867.

[19] CALIFANO A and ALVAREZ M J. The recurrent architecture of tumour initiation, progression and drug sensitivity[J]. *Nature Reviews Cancer*, 2017, 17(2): 116–130. doi: 10.1038/nrc.2016.124.

[20] KIMURA S. AT-9283, a small-molecule multi-targeted kinase inhibitor for the potential treatment of cancer[J]. *Current Opinion in Investigational Drugs*, 2010, 11(12): 1442–1449.

[21] ZHOU Donghu, JIANG Ying, and HE Fuchu. Alternative splicing in lifeomics era[J]. *Scientia Sinica Vitae*, 2015, 45(12): 1177–1184. doi: 10.1360/N052015-00135.

许 鹏：男，1986年生，博士后，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.
刘文斌：男，1969年生，教授，研究方向为生物信息学.