下一代测序技术数据中的选择性剪切计算识别方法研究

邹 权,李旭斌,林子雨,江 弋,林 琛
(厦门大学信息科学与技术学院,福建厦门 361005)

摘 要: 随着测序技术的发展,下一代测序技术(Next-Generation Sequencing)给生物信息学领域研究带来了新的机遇和挑战.由于选择性剪切(alternative splicing,AS)在真核生物基因表达和蛋白质多样性方面的重要性,识别选择性剪切位点一直都是研究的重点.下一代测序技术的出现,使得选择性剪切研究的计算方法不断地变化.介绍了选择性剪切过去和目前研究的状况,然后总结了基于 RNA-seq 数据的选择性剪切研究方法、软件以及数据库,并利用了 RNA-seq 数据比较了相关软件,最后讨论了选择性剪切中计算方法的发展方向和前景.

关键词: 下一代测序技术; RNA_seq; 选择性剪切; 剪切位点; 读段定位; 生物信息学
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2012)02-0350-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2012.02.023

Computational Identification of Alternative Splicing in Next-Generation Sequencing Data

ZOU Quan, LI Xu-bin, LIN Zi-yu, JIANG Yi, LIN Chen
(*School of Information Science and Technology, Xiamen University, Xiamen, Fujian* 361005, *China*)

Abstract: With the development of sequencing technology, Next-Generation Sequencing brought opportunities and challenges to the bioinformatics. As the importance of alternative splicing in gene expression and protein diversity in eukaryotes, alternative splicing identification has been the focus of bioinformatics research. The computation methods for alternative splicing need to be improved since the appearance of Next-Generation Sequencing. We introduce the past and current research of AS firstly; then conclude and compare the research methods, software and databases of AS based on RNA-seq data; finally try to discuss the development and prospect of the computational methods on AS.

Key words: next-generation sequencing; RNA-seq; alternative splicing; splice sites; read mapping; bioinformatics

1 引言

选择性剪切即基因转录产生的 pre-mRNA 根据不同的剪切方式产生不同的成熟 mRNA,从而导致蛋白质的多样性.自从发现选择性剪切以来,研究结果表明,人类基因转录过程存在着大量选择性剪切事件[1],根据高通量深度测序发现,约有 95% 的人类基因存在选择性剪切[2].选择性剪切是真核生物调节基因表达和产生蛋白质组多样性的重要调控机制[3,4].在医学研究方面,选择性剪切与许多疾病有着密切的联系,例如:癌症,神经系统疾病等等[5,6].由此,医学、遗传学、生物信息学等领域的学者对其投注大量精力,希望能找到更多的剪切事件,深入了解其调控机制.

剪切位点识别是选择性剪切研究的关键步骤,通过对外显子/内含子结构的定位以及剪切位点特征来预测剪切位点,是传统预测选择性剪切位点的研究方法.随着第一代测序方法运用,出现了许多用于序列比对的算法,软件以及数据库.专门用于选择性剪切研究的资源逐渐丰富,例如常见的 ASD[7]选择性剪切数据库.但是,第一代测序方法的代价较高,因此,后基因组时代的测序技术努力迈向千美元基因组和百美元基因组的目标.下一代高通测序技术的高通量低成本特点为科学研究提供了新的舞台.近年来,RNA-seq(高通量 RNA 测序)逐渐成为了基因表达和转录组分析的新手段,这一时

收稿日期:2011-04-25; 修回日期: 2011-09-22
基金项目:国家自然科学基金(No.61001016,No61102136);福建省自然科学基金(No.2011J05158);厦门大学基础创新科研基金(中央高校基本科研业务费专项资金)(No.2011121049)

期,也出现了许多用于短序列比对以及基于 RNA-seq 选择性剪切位点预测的软件和数据库.

本文概述两代测序技术下选择性剪切研究的方法、软件以及数据库,分析它们对选择性剪切研究的影响.以 Illumina/Solexa 测序平台产生的 RNA-seq 数据为例,在不同的测序深度以及序列读长条件下,对常见的三种剪切位点预测软件(HMMsplicer[8], SOAPsplice, TopHat[9])进行对比.通过比较每种软件的准确预测位点个数,准确率和错误率来反映它们在不同条件下的性能.最后,讨论在 RNA-seq 数据下研究选择性剪切面临的问题和挑战.

2　下一代测序技术

第一代测序技术基于荧光标记的 Sanger 法[3].诺贝尔化学奖得主 Sanger 于 20 世纪 70 年代发明了末端终止法测序技术,使得人们得以对生命遗传领域进行实质性研究.之后,出现了自动测序仪,例如 ABI 370, ABI 3730 等.虽然,这些测序仪的读长可以达到 1000bp,准确率也很高,但是成本和速度成为了限制其发展的根本因素.

第二代测序技术基于循环阵列合成测序法,又称下一代测序技术.20 世纪 90 年代末到 21 世纪初,出现了 Illumina 的 Genome Analyzer,罗氏 454 基因组测序仪,AB Life Technologies 的 SOLiD 系统等新一代的测序仪.它们产生的数据都有通量大、读长短、成本低的特点.针对于读长短,出现了一些单分子测序技术,例如 HeliScope 等.第三代测序技术即直接测序目前仍处于论证阶段.

测序技术应用不久,1977 年发现了选择性剪切[10].随后,研究人员意识到选择性剪切有着重要意义,不仅对基因表达起到调节作用,还是产生蛋白质多样性的原因[11,12].在过去的十几年间,研究人员通过实验以及数据分析手段对选择性剪切研究投入了大量研究.下一代测序技术的低成本、高通量优势给测序带来了新的舞台.与此同时,对于 RNA-seq 技术产生的海量数据进行研究成为了同行们的主要研究方向.选择性剪切的研究由此进入了新的历程.

RNA-seq 数据以 FASTQ 格式[13]保存,与 FASTA 格式不同的是,FASTQ 用 4 行信息描述每一条读段。第一行以'@'开头,加上读段的描述信息,一般由文件名,读段编号以及读段长度等等组成。第二行由具体的碱基序列组成。第三行以'+'开头,同样可以保存读段描述信息。第四行则为测序的质量分数,质量分数个数和碱基个数一致,分数通过相应算法计算得来。根据测序方法的不同,FASTQ 格式还有细微差异。

RNA 测序在基因表达水平研究,选择性剪切研究以及新基因发现等方面应用广泛.RNA 测序基础上评估基因表达水平主要是统计读段定位到有注释的基因外显子上的数量.根据 RNA 测序的原理,统计读段数量还需要考虑到测序深度以及本身基因的长度.因此,通常人们用 RPM 和 RPKM 来衡量某个基因表达水平[14]:

$$\mathrm{RPKM} = \frac{\text{Total exon reads}}{\text{Mapped reads(millions)} \times \text{Exon length(KB)}}$$

Total exon reads 表示定位到外显子上的读段个数, Mapped reads 表示定位到基因上的所有读段个数, Exon length 表示外显子长度.

选择性剪切事件和新基因发现都和 RNA-seq 数据读段定位相关密切.根据结合区读段的定位能识别出剪切位点信息,根据定位策略的不同可以验证剪切位点或者发现新的剪切位点.在读段定位的过程中,发现有些读段并不能定位到已有注释的基因上,这时候人们考虑是不是存在新的基因没有被研究者发现.

3　基于传统 Sanger 测序数据下的选择性剪切研究

除了通过实验手段确认选择性剪切事件之外,研究者主要通过 EST 表达序列标签以及基因序列的比对来预测潜在的选择性剪切事件.通过大量分析研究,发现了剪切位点 3'端受体位点和 5'端供体位点在剪切事件中的重要意义,同时总结出了五种选择性剪切形式,如图 1 所示.

除此之外,Fairbrother 和 Wang Z 等人对人类基因组中的外显子进行研究,发现剪切增强子 ESE 和 ESS 对选择性剪切起着重要调控作用[15].Black D L 等人的研究发现内含子中的剪切增强子 ISE 和沉默子 ISS 对于剪切位点的选择以及外显子、内含子的识别也十分重要[16].由此我们相信真核基因的选择性剪切过程不只是剪切因子决定的,而是一个复杂的调控过程.

选择性剪切的研究手段主要有以下几种:

(1)基于 ESTs, mRNA 和基因片段之间的比对分析.EST 比对分析是最早的选择性剪切研究方法之一,这种方法虽然能识别一定的选择性剪切事件,但是存在着自身的局限性,比如 EST 数据的不完整性,来自基因污染的影响,3'端敏感,以及代价高昂等[17~19].

(2)利用基因芯片等高通量技术.基因芯片技术带来了全基因转录组研究的热潮,目前,通过这个技术已经识别大量的选择性剪切事件.Johnson 等人通过分析基因芯片数据发现了许多的盒式外显子(exon-skipping)事件[1,20].不过,其缺点是,探针密度限制,并且需要根据已知序列设计探针,数据分析困难等.

(3)利用机器学习方法进行理论预测.常见的利用机器学习的算法模型有:支持向量机 SVM[21],权重矩

阵,隐马氏模型,二次判别法[22,23],神经网络模型[24]等等.

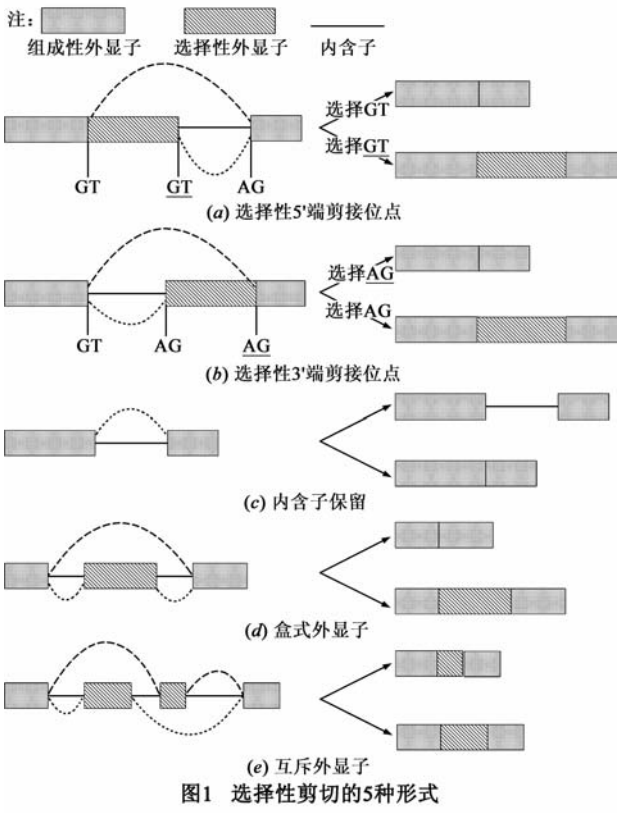

图1 选择性剪切的5种形式

几个比较常用的选择性剪切数据库,如表 1 所示.

表 1 常见的选择性剪切数据库

数据库名称	网址	文献	下载服务
ASTD	http://www.ebi.ac.uk/astd/	[7]	Yes(fasta)
ASPicDB	http://www.caspur.it/ASPicDB	[25]	No
H-DBAS	http://jbirc.jbic.or.jp/h-dbas/	[26]	Yes(fasta)
ASAP II	http://bioinfo.mbi.ucla.edu/ASAP2/	[27]	Yes(sql)
ProSplicer	http://prosplicer.mbc.nctu.edu.tw/	[28]	No
ECgene	http://genome.ewha.ac.kr/ECgene/	[29]	Yes(txt)
AS-ALPS	http://as-alps.nagahama-i-bio.ac.jp/	[30]	Yes(txt)
ASIP	http://www.plantgdb.org/ASIP/	[31]	Yes(sql)
SpliceNest	http://splicenest.molgen.mpg.de/	[32]	No
MAASE	http://maase.genomics.purdue.edu/	[33]	No

目前,数据库通常具有三种服务:查询 query,图形化分析 visual,以及提供序列下载 download.ASTD 是其中较丰富的选择性剪切数据库,但是,随着各类数据库的涌起,数据库开发者逐渐将零散的数据集用描述文件代替完整的序列文件.由此,ASTD 也在今年停止维护,将选择性剪切数据合并到 Ensembl 数据库中,用户可以在 Ensembl 数据库中筛选出满足的选择性剪切序列.绝大部分的选择性剪切数据库都将注意力投向人类,鼠等动物,而 ASIP 数据库则主要研究植物的选择性剪切情况,对于植物类的选择性剪切研究具有重要参考意义.在可视化方面,AS-ALPS 数据库具体标识出选择性剪切的类型和位点,图形简洁易懂,但是,内容还不够丰富,包括 SNP 位点信息等等,并且不能对用户序列进行分析.综上所述,以上选择性剪切数据库各有所长,各有所短.我们仍期待一个,更新及时,数据丰富,种类齐全,可轻松获取的选择性剪切数据库.

4 基于新一代高通量 RNA 测序数据下的选择性剪切研究

Sanger 测序法由于通量低、价格贵等原因使得局限性越加显现,因而新一代高通量测序技术出现后便得到快速地发展,使得测序技术又向千美元基因组计划迈近了一步,同时,RNA 测序成为了基因表达和转录组研究分析的新途径.在这个阶段,传统的选择性剪切研究方法与新一代的研究方法并存发展,越来越多的学者投入到新时期新算法的研究上[34].

下一代测序技术的高通量,高效以及廉价的优点,给选择性剪切的研究提供了广阔的平台.但是.RNA-seq 也有其自身的缺点,最主要的挑战来自读长短和错误率偏高的问题.第一代测序即 Sanger 测序的读长可达 1000bp 左右,然而,RNA-seq 刚开始的读长只有 25bp 左右,目前虽然可达 100bp,利用 Illumina/Solexa 双端测序可达双倍,但是还是相对很短.

获得 RNA-seq 原始数据之后,需要进行组装,读段定位到基因组上才可以进行其他分析。读段定位过程的准确性直接影响到后续分析的可靠性。因而,读段短的问题给读段定位造成了相当的难度.同时,测序中发生错误是不可避免的,并且受到单核苷酸的多态性的影响,因此,在适当的范围内允许读段定位存在误差.

不仅仅是读段长度问题,测序的深度也影响着定位效果。测序深度是测序得到的碱基总量与基因组大小的比值。因而,测序深度越深,获得的数据量越大,读段定位能覆盖到的转录组范围越广,识别的选择性剪切事件越完整.

4.1 读段定位

利用 RNA-seq 数据进行选择性剪切位点预测的第一步就是要将读段(reads)定位到参考转录组(reference transcriptome)上,但是由于转录组本身不够完整,因而一般的分析工具常常将它们定位到参考基因组(reference genome)上[9].RNA-seq 读长短,加上转录组本身并不完整,使得这一步骤的准确性直接影响到预测的效果.

其中,读段定位中存在某些读段,它们会跨越两个外显子结合区(exon-exon juctions)[35],这些“结合区读段”无法被直接定位到基因组序列上,如图 2 所示.由于这些读段中包含着潜在的剪切位点,因而,这部分是研

究选择性剪切位点的重点;也是发掘尚未被发现的剪切事件的关键区域,由此剪接结合区读段的处理策略是预测剪切位点的关键[36].在处理结合区读段这个问题上,一种做法是:根据目前已知的外显子注释,将这些读段定位到参考基因组;ERANGE[14]软件就是采用这种做法,但是,由于外显子注释本身不够完善,因而这种做法很难发现新的剪切事件.另一种做法是:不根据已有的注释进行定位.例如,TopHat 做法,首先将那些能完整定位到参考基因组上的读段分成不同的几个聚类(cluster),然后,把具有重叠的区域的读段归类到同一个聚类中,在每个聚类中划定一个外显子区域[35],最后将那些结合区读段定位到可能的结合区域上.由于这种做法不是基于已知的外显子注释,因而能发现一些新的剪切事件.剪切位点预测软件 TopHat[9]同时具备这两种策略.

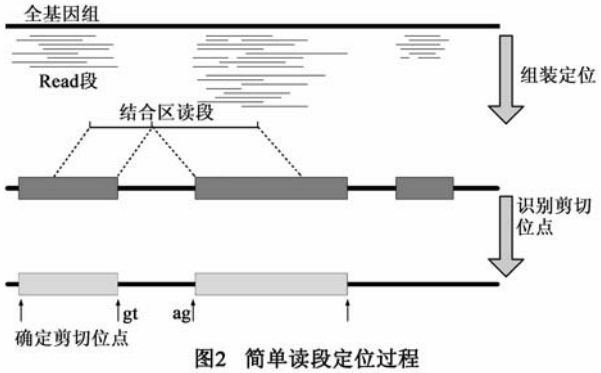

图2　简单读段定位过程

4.2　数据分析软件

目前,有许多专门用于对 RNA-seq 数据进行读段定位的软件,这些软件基本上采用以下几种算法:(1) Smith-Waterman 算法,例如 BFAST[37], SHRiMP[38];(2) 2way-BWT 算法,例如 SOAPAligner[39];(3) BWT(Burrows-Wheeler Transform)算法,例如 Bowtie[40] 和 BWA[41];(4) spaced-seed 空位种子算法,例如 MAQ[42].对于软件来说,还要考虑的问题是数据兼容性.由于不同测序平台产生的 RNA-seq 数据格式的不同[13],导致软件能否支持多种格式成为了影响软件通用性的因素.Bowtie 和 BWA 相对比较高效,而 SOAPAligner, BFAST, 和 MAQ 具有良好的容错能力(mismatches).

除了采用读段定位的手段外,还有一种软件专门用于读段组装,即 *de novo* assembly.在读段组装的基础上研究选择性剪切的方法并不多,但是,读段组装在其他生物信息学方面有其特殊作用.典型的读段组装软件有:SHARCGS[43], SSAKE[44], ALLPATHS[45] 等等,前两者只针对单序列数据进行组装,而后者可以对双端测序的一对序列进行组装.MAQ 也有读段组装的用途.

选择性剪切数据库方面,SRA(Sequence Read Archive)是 NCBI 专门用来存放 RNA-seq 相关数据的数据库.SRA 的数据以.sra 格式存放,可以通过 NCBI 提供的 SRA Toolkit 软件来转化成相应的格式,如:SOLiD 本地数据格式,Illumina 本地数据格式,FASTQ 格式,SFF 格式,以及 txt 文本格式.当然这些数据格式也可以通过该软件转化成.sra 格式,上传至数据库.

4.3　选择性剪切位点预测软件

几种常用的剪切位点预测的软件有:ERANGE, QPALMA[46], TopHat, MapSplice[47], SpliceMap, SOAPsplice, SplitSeek[48], HMMSplicer 等等.当前 RNA-seq 选择性剪切研究集中于寻找剪切位点,尽可能发掘新的剪切位点,然后进行下一步的选择性剪切研究.对于预测软件来说,预测的准确性和效率是非常关键的因素,在预测更多的剪切位点的同时要提高准确率,降低错误概率,这和所选算法的优劣有很大关系.

ERANGE 是很早出现的一种方法,它使用第一种读段定位方法,根据已知的外显子注释将读段定位到参考基因组上,因而这种方法不能发现新的剪切位点.QPALMA 采用机器学习的策略,利用已知的剪切位点来训练支持向量机进行位点识别,在定位方面采用 Vmatch 的方法.由于 Vmatch 不是专门用来比对读段短的序列,因而相对于 Bowtie 来说效率不够高.TopHat 首先用 Bowtie 将序列定位到参考基因组,然后采用 MAQ 将成功定位到参考基因组的序列组装,接着根据相邻外显子识别可能的剪切位点;同时搜集那些没有定位到参考基因组的序列建立空位种子索引,最后采用空位—扩展比对得到可能的剪切位点.根据 TopHat 作者测试,TopHat 每小时处理 2.2 百万个读段,而 QPALMA 大概在 18 万左右.但是,TopHat 由于算法中运用到外显子簇(islands),这在测序深度低或者内含子很短的情况下表现不佳.

SpliceMap 主要分为四个步骤:半读段定位(half-read mapping),种子选择(seeding selection),位点搜索,以及双端过滤(paired-end filtering).SpliceMap 将读段分割成两半,将每个部分和基因序列进行比对定位,在最长内含子的长度范围内尝试将余下部分定位到下游区域.这种做法,要求读长至少为 50bp 以上,因而 50bp 以下的读长序列 SpliceMap 无法处理.SpliceMap 论文将其同 ERANGE 作对比,ERANGE 发现 160899 个位点的同时,SpliceMap 能准确预测 127043 位点.在发掘新的剪切位点方面,SpliceMap 发现的 151317 个位点中有 24274 个没有被 ERANGE 发现,其中有 23020 个是新剪切位点,但是这只是预测出的新位点,并没有经过确认.在 TopHat 和 SpliceMap 之后,出现了 MapSplice 软件,它不是基于剪切位点特性或者内含子长度,同时也有发现新位点的潜力,并且对于长短读段都能适应.

SOAPsplice的出现将剪切位点预测软件的评判标准提高,不仅仅是依赖于识别剪切位点的个数,更加强调高准确率低错误率.在接下一节的实验中我们可以看出,SOAPsplice表现相对突出.SplitSeek对输入数据比较严格,目前只支持ABI SOLiD产生的数据,而且输入数据要通过ABI的whole transcriptime analysis tool进行处理,因而应用面相对不是很广.HMMSplicer做法类似SpliceMap,但是有其创新的地方.它首先还是将读段分成两半,然后将一半和基因组序列对比,利用HMM(隐马氏模型)获得外显子的边界即5'端.第二步,将余下的部分定位到第一部分的下游,从而确定内含子的另一个边界3'端,这个过程中常见(GT-AG,GC-AG,AT-AC)和非常见的剪切位点都被记录,最后利用打分算法给候选位点打分.

4.4 利用模拟数据对比三种软件

我们用Illumina/Solexa输出数据对HMMSplicer,SOAPsplice,TopHat这三种软件做了如下测试.参考基因组数据取自人类第10号染色体,首先利用MAQ软件将基因序列处理成不同读段长度以及不同测序深度的一对模拟RNA-seq序列,作为SOAPsplice和TopHat的测试数据.由于HMMsplicer目前还不支持双端测序数据,因而将每对FASTQ数据简单合并成一个FASTQ文件当作HMMsplicer的测试数据.结果如表2和图3所示.

表2 HMMSplicer,SOAPsplice,TopHat三种软件的比对结果

		HMMSplicer			SOAPsplice			TopHat		
Read len	Depth	TP	# TP	% FP	TP	# TP	% FP	TP	# TP	% FP
40bp	1X	3278	56.41	1.18	2961	50.96	0.70	1816	31.25	4.22
	5X	6766	82.41	1.66	6790	82.70	0.90	5633	68.61	4.77
	10X	7405	89.66	3.04	7751	93.85	1.22	6032	73.04	10.05
	25X	7541	91.23	5.61	8021	97.04	2.10	5467	66.14	24.84
	50X	7567	91.54	8.84	8044	97.31	3.07	4908	59.38	43.95
50bp	1X	3583	61.66	0.75	3378	58.13	0.91	1951	33.57	5.66
	5X	7062	86.02	1.05	7240	88.19	1.04	6485	78.99	5.47
	10X	7468	90.42	1.69	7911	95.79	1.30	7693	93.15	7.28
	25X	7555	91.40	3.13	8073	97.67	2.45	8120	98.23	10.64
	50X	7574	91.63	5.00	8093	97.91	3.36	8169	98.83	14.15
60bp	1X	3721	64.03	0.64	3545	61.00	0.70	1875	32.27	5.30
	5X	7130	86.85	0.93	7365	89.71	1.11	6469	78.79	6.04
	10X	7470	90.45	1.65	7947	96.22	1.41	7655	92.69	7.32
	25X	7559	91.45	2.44	8070	97.63	2.18	8081	97.76	11.37
	50X	7576	91.65	4.00	8094	97.92	3.12	8163	98.75	14.23
70bp	1X	3832	65.94	0.49	3562	61.30	0.56	3574	61.90	2.94
	5X	7103	86.52	0.98	7342	89.43	1.06	7455	90.80	5.43
	10X	7460	90.33	1.31	7928	95.99	1.29	8041	97.36	7.20
	25X	7533	91.13	2.30	8051	97.4	2.16	8180	98.96	12.19
	50X	7563	91.50	3.43	8085	97.81	2.74	8194	99.13	16.17

注:Read len:读长;Depth:测序深度;TP:预测正确个数;# TP=正确个数/总的个数;% FP=错误个数/预测出的个数.

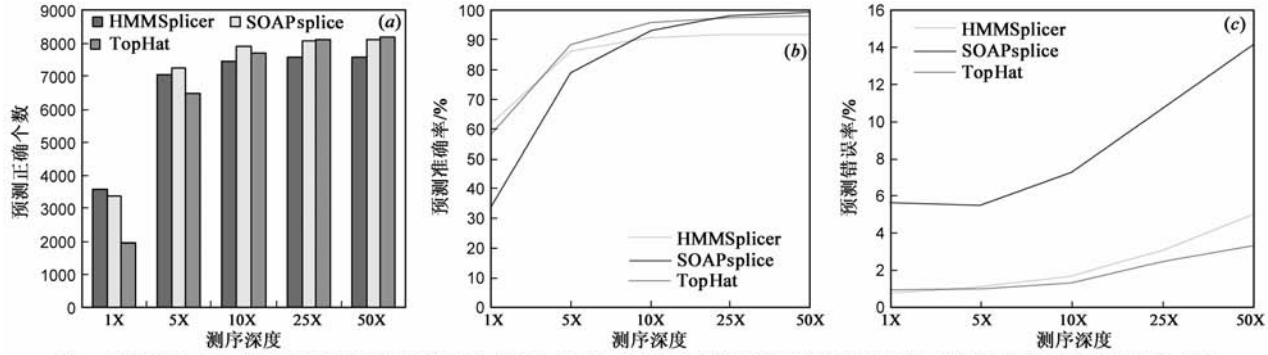

图3 在读长为50bp的前提下,进行实验数据分析;(a),(b),(c)分别为测序深度与预测正确个数,预测准确率以及错误率的关系

从图3中我们可以看出,在读长为50bp的前提下,随着测序深度的加深,每种软件能预测的位点个数逐渐增多.在测序深度为1X到10X的范围内,HMMSplicer和SOAPsplice表现得比TopHat好.由于TopHat首先用

Bowtie 将读段进行全基因组定位,接着,将能完整定位到基因组上的读段归纳为许多 cluster 簇,每个簇代表一个外显子区域,最后在簇之间搜索剪切位点.由此,当测序深度很低时,形成簇的参考读段有限,使得簇的精确度不高,导致寻找剪切位点时准确度不高.

虽然三者在测序深度加深之后识别的位点都增多,但是,在预测错误率方面也逐步攀升.其中 TopHat 错误率增高最为明显.结合图 3(a)和图 3(b)我们可以发现,测序深度达到 25X 以上时预测位点正确个数并没有明显增加,与此同时,预测错误的位点却大为增加.错误率主要来自于深度加深之后测序错误率本身影响加大.可见,HMMSplicer 和 SOAPsplice 在这种情况下表现得更加稳定.

本文实验的具体数据和结果可以从 ftp://datamining.xmu.edu.cn/pub/AS_exp/ 下载.

5　总结

本文通过对目前选择性剪切的相关算法和软件进行介绍和对比,对选择性剪切的研究现状进行了简单的概述.选择性剪切涉及到的读段定位以及位点识别算法将仍然是目前研究的重点;提高算法的质量,从而使得预测位点个数尽量多,并且满足高准确率,是学者们努力的目标.随着下一代测序技术的不断发展,RNA-seq 数据量十分庞大,这对于选择性剪切的研究以及生物信息学的其他领域研究,都是一个广阔的平台.希望本文关于选择性剪切的实验手段和研究方法的综述对其他研究学者有所帮助.

虽然高通量测序给选择性剪切的研究带来了前所未有的契机,但是,目前根据 RNA-seq 数据研究选择性剪切的学者还不是很多.许多算法和软件也没有之前根据 EST/cDNA 理论得来的丰富.新时期的算法和软件在比对这个步骤差距很大,也是基于 RNA-seq 数据研究的关键一步.同时,许多基于 RNA-seq 数据的专门数据库还不太完善.相信在接下来的时间里,随着对选择性剪切研究的不断深入,相应的新的研究方法,以及软件,数据库都会得到完善.

参考文献

[1] Johnson JM, Castle J, Garrett-Engele P, et al. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays[J]. Science, 2003, 302(5653): 2141 – 2144.

[2] Pan Q, Shai O, Lee LJ, et al. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing[J]. Nat Genet, 2008, 40(12): 1413 – 1415.

[3] Modrek B, Lee C. A genomic view of alternative splicing[J]. Nat Genet, 2002, 30(1): 13 – 19.

[4] 闻芳,李衍达.基因表达调控与选择性剪接机制研究[J].电子学报,2001,29(12A):1735 – 1739.
Wen Fang, Li Yan-da. A bioinformatic analysis of alternatively spliced genes of human[J]. Acta Electronica Sinica, 2001, 29(12A): 1735 – 1739. (in Chinese)

[5] Dutertre M, Vagner S, Auboeuf D. Alternative splicing and breast cancer[J]. RNA Biol, 2010, 7(4): 403 – 411.

[6] Dredge BK, Polydorides AD, Darnell RB. The splice of life: alternative splicing and neurological disease[J]. Nat Rev Neurosci, 2001, 2(1): 43 – 50.

[7] Stamm S, Riethoven JJ, Le Texier V, et al. ASD: Bioinformatics resource on alternative splicing[J]. Nucleic Acids Res, 2006, 34(Database issue): D46 – 55.

[8] Dimon MT, Sorber K, DeRisi JL. HMMSplicer: A tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data[J]. PLoS One, 2010, 5(11): e13875.

[9] Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq[J]. Bioinformatics, 2009, 25(9): 1105 – 1111.

[10] Chow LT, Gelinas RE, Broker TR, et al. An amazing sequence arrangement at the 5 ' ends of adenovirus 2 messenger RNA (Reprinted from Cell, vol 12, pg 1 – 12, 1977)[J]. Reviews in Medical Virology, 2000, 10(6): 362 – 369.

[11] Lee C, Wang Q. Bioinformatics analysis of alternative splicing[J]. Brief Bioinform, 2005, 6(1): 23 – 33.

[12] Kriventseva EV, Koch I, Apweiler R, et al. Increase of functional diversity by alternative splicing[J]. Trends in Genetics, 2003, 19(3): 124 – 128.

[13] Cock PJ, Fields CJ, Goto N, et al. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants[J]. Nucleic Acids Res, 2010, 38(6): 1767 – 1771.

[14] Mortazavi A, Williams BA, McCue K, et al. Mapping and quantifying mammalian transcriptomes by RNA – Seq[J]. Nat Methods, 2008, 5(7): 621 – 628.

[15] Fairbrother WG, Yeh RF, Sharp PA, et al. Predictive identification of exonic splicing enhancers in human genes[J]. Science, 2002, 297(5583): 1007 – 1013.

[16] Black DL. Mechanisms of alternative pre-messenger RNA splicing[J]. Annual Review of Biochemistry, 2003, 72: 291 – 336.

[17] Bonizzoni P, Rizzi R, Pesole G. Computational methods for alternative splicing prediction[J]. Brief Funct Genomic Proteomic, 2006, 5(1): 46 – 51.

[18] Modrek B, Resch A, Grasso C, et al. Genome-wide detection of alternative splicing in expressed sequences of human genes[J]. Nucleic Acids Res, 2001, 29(13): 2850 – 2859.

[19] 邹权,郭茂祖,张涛涛.RNA 二级结构预测方法综述[J].电子学报,2008,36(2):333 – 337.
Zou Quan, Guo Mao-zu, Zhang Tao-tao. A review of RNA secondary structure prediction algorithms[J]. Acta Electronica Sinica, 2008, 36(2): 333 – 337. (in Chinese)

[20] Wang LG, Xi YX, Yu J, et al. A statistical method for the detection of alternative splicing using RNA-Seq[J]. PLoS One, 2010, 5(1): e8529.

[21] Dror G, Sorek R, Shamir R. Accurate identification of alternatively spliced exons using support vector machine[J]. Bioinformatics, 2005, 21(7): 897 – 901.

[22] XING Yong-qiang ZL-r, LUO Liao-fu. Prediction of alternative splicing sites of cassette exon and intron retention in human genome[J]. Acta Biophysica Sinca, 2008, 24: 393 – 400.

[23] 李建伏,郭茂祖.系统发生树构建技术综述[J].电子学报,2006,34(11):2047 – 2052.
Li Jian-fu, Guo Mao-zu. A review of phylogenetic tree reconstruction technology [J]. Acta Electronica Sinica, 2006, 34 (11): 2047 – 2052. (in Chinese)

[24] Wang M, Marin A. Characterization and prediction of alternative splice sites[J]. Gene, 2006, 366(2): 219 – 227.

[25] Castrignano T, D'Antonio M, Anselmo A, et al. ASPicDB: A database resource for alternative splicing analysis[J]. Bioinformatics, 2008, 24(10): 1300 – 1304.

[26] Takeda J, Suzuki Y, Nakao M, et al. H – DBAS: Alternative splicing database of completely sequenced and manually annotated full-length cDNAs based on H-Invitational[J]. Nucleic Acids Res, 2007, 35(Database issue): D104 – 109.

[27] Kim N, Alekseyenko AV, Roy M, et al. The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species[J]. Nucleic Acids Res, 2007, 35 (Database issue): D93 – 98.

[28] Huang HD, Horng JT, Lee CC, et al. ProSplicer: A database of putative alternative splicing information derived from protein, mRNA and expressed sequence tag sequence data[J]. Genome Biol, 2003, 4(4): R29.

[29] Lee Y, Kim B, Shin Y, et al. ECgene: an alternative splicing database update[J]. Nucleic Acids Res, 2007, 35(Database issue): D99 – 103.

[30] Shionyu M, Yamaguchi A, Shinoda K, et al. AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse [J]. Nucleic Acids Res, 2009, 37(Database issue): D305 – 309.

[31] Wang BB, O'Toole M, Brendel V, et al. Cross-species EST alignments reveal novel and conserved alternative splicing events in legumes[J]. BMC Plant Biol, 2008, 8: 17.

[32] Coward E, Haas SA, Vingron M. SpliceNest: visualization of gene structure and alternative splicing based on EST clusters [J]. Trends Genet, 2002, 18(1): 53 – 55.

[33] Zheng CL, Kwon YS, Li HR, et al. MAASE: An alternative splicing database designed for supporting splicing microarray applications[J]. RNA, 2005, 11(12): 1767 – 1776.

[34] 庄永龙,马飞,周敏,等.基于多 Agent 的生物信息数据整合系统-BioAgent1[J].电子学报,2005,33(1):78 – 82.
Zhuang Yong-long, Ma Fei, Zhou Min, et al. BioAgent: A biological data integration system based on multi-agent[J]. Acta Electronica Sinica, 2005, 33(1): 78 – 82. (in Chinese)

[35] Au KF, Jiang H, Lin L, et al. Detection of splice junctions from paired-end RNA-seq data by SpliceMap [J]. Nucleic Acids Res, 2010, 38(14): 4570 – 4578.

[36] 王曦,汪小我,王立坤,等.新一代高通量 RNA 测序数据的处理与分析[J].生物化学与生物物理进展.2010,37(8):834 – 846.
Wang, X, X W Wang, L K Wang, et al. A review on the processing and analysis of next-generation RNA-seq data[J]. Progress in Biochemistry and Biophysics, 2010, 37(8): 834-846. (in Chinese)

[37] Homer N, Merriman B, Nelson SF. BFAST: An alignment tool for large scale genome resequencing[J]. PLoS One, 2009, 4 (11): e7767.

[38] Rumble SM, Lacroute P, Dalca AV, et al. SHRiMP: accurate mapping of short color-space reads[J]. PLoS Comput Biol, 2009, 5(5): e1000386.

[39] Li R, Li Y, Kristiansen K, et al. SOAP: Short oligonucleotide alignment program [J]. Bioinformatics, 2008, 24(5): 713 – 714.

[40] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome[J]. Genome Biol, 2009, 10(3): R25.

[41] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform [J]. Bioinformatics, 2009, 25 (14): 1754-1760.

[42] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores[J]. Genome Research, 2008, 18(11): 1851 – 1858.

[43] Dohm JC, Lottaz C, Borodina T, et al. SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing[J]. Genome Res, 2007, 17(11): 1697 – 1706.

[44] Warren RL, Sutton GG, Jones SJM, et al. Assembling millions of short DNA sequences using SSAKE[J]. Bioinformatics, 2007, 23(4): 500-501.

[45] Butler J, MacCallum I, Kleber M, et al. ALLPATHS: De novo assembly of whole-genome shotgun microreads[J]. Genome Research, 2008, 18(5): 810 – 820.

[46] De Bona F, Ossowski S, Schneeberger K, et al. Optimal spliced alignments of short sequence reads[J]. Bioinformat-

ics,2008,24(16):i174-180.
[47] Wang K,Singh D,Zeng Z,et al. MapSplice:accurate mapping of RNA-seq reads for splice junction discovery[J]. Nucleic Acids Res,2010,38(18):e178.
[48] Ameur A,Wetterbom A,Feuk L,et al. Global and unbiased detection of splice junctions from RNA-seq data[J]. Genome Biol,2010,11(3):R34.

作者简介

邹　权　男,1982 年生于黑龙江佳木斯.厦门大学计算机科学系助理教授、硕士生导师.研究方向为生物信息学与数据挖掘.
E-mail:zouquan@xmu.edu.cn

李旭斌　男,1990 年生于福建福安.厦门大学计算机系硕士研究生,主要研究方向为生物信息学.
E-mail:xubinli@stu.xmu.edu.cn

林子雨　男,1978 生于吉林柳河.厦门大学计算机科学系助理教授、硕士生导师,主要研究方向为数据仓库,联机分析技术,数据挖掘等.
E-mail:ziyulin@xmu.edu.cn

江　弋　男,1960 年生于福建福州.厦门大学计算机系副教授、硕士生导师,主要研究方向为数据库技术与应用、数据挖掘、电子商务、多媒体技术及应用、嵌入式系统等.
E-mail:jiangyi@xmu.edu.cn

林　琛(通讯作者)　女,1982 年生于福建厦门.厦门大学计算机科学系助理教授、硕士生导师,主要研究领域为 Web 数据检索,数据挖掘和管理等.
E-mail:chenlin@xmu.edu.cn