基因表达调控与选择性剪接机制研究

闻 芳,李衍达
(清华大学生物信息研究所智能技术与系统国家重点实验室,北京 100084)

摘 要: 随着人类基因组计划(HGP)的完成,生物信息学的研究进入了后基因组时代,用计算方法对基因表达调控和基因功能进行研究成为生物信息学研究的核心内容. 由于在真核基因表达调控中的特殊地位,选择性剪接成为研究真核基因表达调控的重要内容之一. 本文从收集选择性剪接基因的数据出发,尽可能的收集已知的选择性剪接的基因和它们的各种转录产物,并进行了适当的筛选以保证数据的质量和统计分析的可靠性. 对挑选出的 371 个人类基因,提取各种转录产物的编码区(coding regions,或简称 cds),应用一种新的针对选择性剪接的多序列比对程序 ASALIGN 进行多序列比对来揭示不同 cds 间的剪接关系,提出其中的可变区域与不可变区域,并对可变区域与不可变区域的长度分布,可变区域在 cds 中出现的位置,由于选择性剪接引起的同一段序列读码框相位的变化以及可变区域与不可变区域及二者边界上的密码子使用频率进行了统计分析,得到了一些很有意思的结果. 这些统计结果对于选择性剪接机制的进一步研究以及选择性剪接基因的预测提供了良好的线索.

关键词: 选择性剪接; 多序列比对; 基因结构; 剪接模式
中图分类号: Q11 **文献标识码**: A **文章编号**: 0372-2112 (2001) 12A-1735-05

A Bioinformatic Analysis of Alternatively Spliced Genes of Human

WEN Fang, LI Yan-da
(Institute of Bioinformatics, Tsinghua University State key laboratory of intelligent technology and system, Beijing 100084, China)

Abstract: With the completion of Human Genome draft, bioinformatic research has entered the post-genomic era. Studying of the control of gene expression and function using computational methods has gradually become the mainstream of research. For its special role in the control of eukaryotic gene expression, alternative splicing has become an important topic in genomic research. In this study, we analyzed alternative splicing of human genes based on complete coding sequences (cds) of alternatively spliced genes. As reported previously we have set up a database AsMamDB. Transcripts probably belonging to one gene were put into a single cluster. We also developed a new multi-alignment tool, ASALIGN, aiming to reveal alternative splicing patterns of transcripts belonging to a same gene and identify their alternative and non-alternative regions. Here, through filtering we chose 371 genes that have two or more different cds. Using Asalign we have identified all alternatively spliced regions in the cds of selected genes. We also studied lengths, positions of alternatively spliced regions, shifting of reading frames due to alternative splicing, as well as codon frequencies of non-alternative, alternatively spliced regions and their boundaries, and ended with some interesting results. Our study may provide important insights into the alternative splicing phenomenon and clues for prediction of alternative splicing.

Key words: alternative splicing; multiple alignment; gene structure; splice pattern

1 引言

2 月 12 日,参与人类基因组计划的 6 国科学家宣布了有关人类基因组的初步研究结果,其中有一项内容引起了公众和媒体的广泛关注,人类基因为什么不是原来预计的 8 万至 10 万个,而是只有 3.5 万个左右. 人类基因组计划首席科学家柯林斯认为,人类基因数比预计的少得多说明,人类在使用基因上很节约,与其他物种相比更高效. 与过去一个基因编码一种蛋白的假设相比,看来每个基因平均负责制造三种蛋白. 人类不是靠自我开发新基因来获取新功能,而是通过重新编码或扩充已有的可靠资源来达到创新的目的. 选择性剪接现象的存在正是人类节约使用基因的最好的例证.

真核基因在结构上的不连续性是近 10 年来生物学上的重大发现之一. 而进一步的大量实验表明,对于同一个基因,其剪接位点和拼接方式可以有所改变,从而导致同一个基因可以表达出多个不同的相关蛋白产物,行使不同的生理功能. 这就 RNA 的选择性剪接(alternative splicing),也称可变剪接. RNA 剪接特别是选择性剪接是真核基因表达调控研究的重要内容之一. 由于 RNA 的选择性剪接不牵涉到遗传信息的永

收稿日期:2001-06-27;修回日期:2001-09-10
基金项目:国家自然科学基金(No. 69872018,69935020);北京市科技合同项目(No. 9550213000);清华大学基础和创新研究基金

久性改变,所以是真核基因表达调控中一种比较灵活的方式.对选择性剪接机制的研究并在此基础上预测同一基因的不同剪接形式对于进一步分析基因的功能及其表达调控具有重要的理论意义,同时,这一研究对于针对不同剪接形式进行组织特异性的药物设计也具有重大的实际意义.另一方面,考虑到选择性剪接在真核生物中,尤其在人类这样的高等生物中存在的普遍性(据研究至少 30 %的人类基因存在不同的剪接产物[1]),对选择性剪接机制的研究也有助于进一步提高当前基因预测程序的准确性.

长期以来,生物学上对选择性剪接的研究仍多停留在对单个基因的选择性剪接现象和机制的研究,而缺少更为系统更为综合的分析.究竟什么是影响基因产生选择性剪接的因素,这些基因在发生选择性剪接的过程中是否存在共同的机制或共同的规律,选择性剪接导致的产物间究竟会产生多大的不同,对其功能有什么影响,大量的问题靠一个一个基因地实验是难以解决的.近两年来,应用计算方法对选择性剪接现象进行研究开始引起人们的关注.一方面,一些研究小组通过将 EST 数据与已知的人类基因的 DNA 数据进行比对估计人类基因可变剪接现象的出现频率,统计可变剪接现象发生的区域(编码区或非翻译区),验证剪接位点的可变性,对选择性剪接的类型进行分类,或力图发现一些可能的新的剪接形式[2~5].另一方面,一些小组开始收集选择性剪接的数据,并建立相应的二级数据库,如 AsDB[1,6], Intronerator[7], AsMamDB[8].但是,现有数据库在收集的数据的数量和质量上都存在一定的问题,而对于应用 EST 数据进行的统计分析而言,由于 EST 数据本身质量不高,而且缺少完整转录本的信息,统计结果的可靠性没有保证,并且只能获得局部信息.

本文中,将从完整 cds 的角度出发对人类的选择性剪接基因进行较为系统、综合的研究.在前期工作中,我们也进行了选择性剪接数据的收集工作,建立了 AsMamDB 数据库[8],收集了 1563 个已知的人、小鼠和大鼠的选择性剪接的基因,将属于一个基因的不同转录本归为一族.并且开发了一种针对选择性剪接的多序列比对程序 ASALIGN[9],用以揭示同一个基因若干转录产物之间的剪接关系.进一步的,对数据库中人类 899 个选择性剪接基因数据进行了质量控制和数据筛选,提取了具有明确的两种或两种以上不同转录产物(这里指 cds 不同)的 371 个基因,应用 ASALIGN 进行多序列比对找到不同转录本的可变区域,对可变区域的长短,可变区域的发生位置,由于选择性剪接引起的同一段序列的读码框相位变化的情况以及可变区域与不可变区域中密码子使用频率进行了统计分析.这些统计结果对于对选择性剪接现象与剪接机制的进一步研究与认识具有重要意义,也为选择性剪接基因的预测提供了线索.

2 数据的收集 ——AsMamDB 的建立

为了对选择性剪接这一问题进行更为系统、深入的分析,我们收集了 899 个人类的、431 个小鼠和 233 个大鼠的选择性剪接的基因,构建了选择性剪接数据库 ——AsMamDB[8] (http://166.111.30.65/ASMAMDB.html).数据收集从 GenBank 出发,选择标注为选择性剪接的序列项,然后利用 UniGene 等现有数据库,通过序列比对尽可能收集同属于一个基因的不同转录本,将其归类为一个序列族. GenBank 中关于这个基因的各个序列项的其他信息,如基因结构,表达产物,表达组织器官,调控元件等,以及可能与此基因相关的 EST 数据也都格式化的记录在 AsMamDB 中.

AsMamDB 的建立为我们对人类选择性剪接基因进行更深入系统的分析和研究提供了基础.

3 数据的筛选

AsMamDB 中针对每个选择性剪接的基因,尽可能地收集了它的已知的转录本,并将其归为一个序列族.但在数据质量上仍然存在一些问题会妨碍我们进一步的统计分析.如有些序列族中只含有一种 cds(编码序列,即 mRNA 中真正翻译为蛋白质的序列段),序列族的序列间存在冗余,有些序列族中存在不属于同一个基因的 mRNA 等等,因此我们对这 899 个基因又进行了筛选.筛选规则如下:

(*a*)代表此基因的序列族中至少含有两种不同的 cds,即此基因至少有两种不同的产物.

(*b*)去除同一序列族中的冗余序列.由于对于只有 5′UTR 区起始或 3′UTR 区终止不同的转录本,因为无法判断是由于测序不完全还是由于它们具有不同的转录起始位点或终止位点,如果两条 mRNA 只有转录起始或转录终止不同,且 cds 区完全一致,将只在此序列族中保留较长的一条.

(*c*)过滤掉不属于同一序列族的噪声序列.如果一条 cds 只有很小的一部分可以和此序列族中的其它序列比对上,或者它和另外一个序列族中序列的相似程度大于它与此序列族中的任何一条序列,则认为它并不属于此序列族,将被去除或合并到另一个序列族中.

(*d*)要求不同 cds 的差别不是由于个别碱基的插入或缺失造成的,因为这样会影响到对读码框相位的分析.

由此,我们筛选出 371 个基因,后文中的统计分析将主要针对这 371 个基因进行.

4 多序列比对算法 —ASALIGN

由于人类全基因组序列还没有完全公开,我们研究的 371 个基因中大部分的 mRNA 序列还无法在基因组上进行精确定位.因此,研究的第一步是通过多序列比对揭示同一个基因产生的 mRNA 间的剪接关系,从比对结果中就可以看出哪些区域是不同 mRNA 间的可变区域,而哪些又是它们共同的区域.为了比对结果更加准确可靠,采用了一种新的针对选择性剪接的多序列比对算法 —ASALIGN[9](可从下面站点下载:ftp://166.111.30.65/pub/asalign).与传统的多序列比对算法相比,ASALGIN 具有下面两个特点.首先,考虑到在多个选择性剪接转录本的比对中 gap 的大小可以有很大的变化(从几个碱基到几千个碱基),通过引入 Morgenstern 等[10]提出的段与段(segment-to-segment)比对的思想,避免了给不同长度的 gap 处以不同罚分的困难,从而得到更为准确的比对结果.此外,在计算序列间相似性时不进行两两比对,而根据两序列匹

配词的个数之间进行估计,极大地缩短了计算时间. ASALGIN 具有的这两个特点使它无论在比对效果和运行时间上都要优于现在常用的多序列比对算法.

考虑到发生在 cds 区的选择性剪接才会真正影响到基因翻译的产物,而另一方面,由于缺乏足够的基因组序列,现有 mRNA 的非翻译区(UTR)的序列完整性和数据质量尚难以保证,这里主要研究一个基因由于选择性剪接产生的不同 cds 间的关系. 因此,首先对这 371 个基因提取不同的 cds,然后应用 ASALIGN 进行多序列比对. 图 1 给出其中一个基因(pre-T cell receptor)的两种 cds 间的比对结果.

```
seq1 >- * * 1 * (1) * 426 * *---------- * * 427 * (1) * 846 * *--
seq2 >- * * 1 * (2) * 426 * *-- * * 427 * ( 2) * 471 * *-- * * 472 * (2) * 891 * *--
```

图 1　某基因的两种 cds 的比对结果图

可以看到,在比对结果图中,每条序列被间隔成了不同的片段. 有些片段是各条序列所共有的,它对应于各转录产物中的不可变区域,而另一些片段是某些序列所特有的,它对应于转录产物中的可变区域. 如在上图中,从比对结果图中得到了三个序列片段,第 1 个片段是两条 cds 的不可变区域,此片段在两条序列中的位置都是从第一个碱基到第 426 个碱基,第 2 个片段是第 2 条 cds 所特有的,即为可变区域,它在序列中的位置是从第 427 个碱基到第 471 个碱基,同样,第 3 个片段也是不可变区域,它在第 1 条序列中的位置是从第 427 个碱基到第 846 个碱基,而在第 2 条序列中的位置为第 472 个碱基到第 891 个碱基.

5　人类选择性剪接基因的统计分析

从比对结果图中,可以提出对于每个基因的不同产物中的可变区域和不可变区域. 针对这些序列片段对如下几个特征进行了统计分析,得到了一些有趣的结果. 希望通过这些分析,能够对选择性剪接这一现象有更深入的理解,并为选择性剪接基因的预测提供一定的线索.

5.1　不同基因的 mRNA 与 cds 数目分布

前面已经谈到,从 AsMamDB 中的 899 个人类基因中筛选出了符合要求的 371 个基因. 它们转录产生的 mRNA 和 cds 的数目分布如下:

表 1　cluster 中 mRNA 数目的分布

Num. Of mRNAs	2	3	4	5	6	7	8	10	11	13	19	21	23	32
Num. Of Clusters	157	96	57	28	17	7	2	1	1	1	1	1	1	1

371 个 cluster 中总共包含了 1266 条 mRNA.

表 2　cluster 中 cds 数目的分布

Num. Of cds	2	3	4	5	6	7	15	17	20
Num. of Clusters	252	69	27	9	9	1	2	1	1

371 个 cluster 中总共包含了 992 条 cds.

从上面的数字可以看到,研究的 371 个基因的转录产物中包含了 1266 条不同的 mRNA 序列,却只有 992 种不同的 cds,即至少(1266 - 992) / 1266 = 21.6 %的序列的选择性剪接只发生在序列的 UTR 区,这里还没有包括只有转录起始和终止不同的引起的不同转录本. Mironov 等[2]通过 EST 数据与 genomic 数据的比对发现 80 %的选择性剪接的基因在 5′UTR 区,19 %的选择性剪接的基因在 3′UTR 区存在可变剪接的现象. 这一结果与我们得到的并不矛盾,因为在我们的分析中去除了同时发生在 UTR 区和 cds 区的可变剪接现象.

我们知道,只有 cds 区发生的选择性剪接才会真正影响基因的翻译产物,那么为什么会有这么多的选择性剪接只发生在 mRNA 的 UTR 区,这是否与基因的表达水平与表达调控有关,这些问题还有待我们进一步的分析和实验的验证.

5.2　发生在翻译区的可变区域与不可变区域的长度分布

可变区域对应于 premRNA 序列上的选择性剪接现象的发生,或者采用了不同的剪接位点,或者干脆跳过完整的外显子,可变区域的长度则对应了选择性剪接位点间的距离或可选择性外显子的大小. 另一方面,可变区域和不可变区域又分别代表了基因的各种产物间的不同与相同的片段. 这里,将这 371 个基因的 cds 中提出的可变区域与不可变区域的长度分别进行了统计并将二者进行了对比.

在 371 个基因的不同产物中,总共存在 914 个可变区域和 584 个不可变区域,它们的长度分布如图 2 所示. 图中,横轴代表片段长度(取对数坐标),纵轴代表某种长度的片段出现的频率.

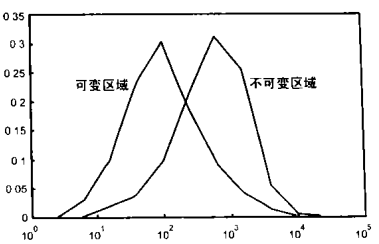


图 2　可变区域与不可变区域的长度分布

从图中可以看到,可变区域的长度多集中在 10^2 量级,而不可变区域的程度则多集中在 10^3 量级,即与不可变区域的长度相比,大部分可变区域要小得多. 值得注意的,有 258 个可变区域的片段长度小于或等于 50 个碱基,占总数的 28 %. 这意味着,有相当大比例的选择性剪接现象的发生是由于存在两个距离很近的可选择的剪接位点或在某种剪接形式中跳过了一个很小的外显子. 我们知道,就目前的基因预测的算法而言,预测较长的外显子比较容易,预测较短的外显子较难,预测外显子的大致位置相对容易一些,精确定位剪接位点则较难[11]. 现在得到的这一统计结果一方面可能是造成这一状况的原因之一(学习集数据的噪音),另一方面也向我们预示了利用序列的统计特征进行选择性剪接区域预测的困难性.

5.3　可变区域在 cds 中的发生位置

基因的选择性剪接使得同一个基因可以产生不同的转录产物,这些转录产物间的不同可以表现在 cds 的起始片段、终止片段,也可以表现在 cds 的中部. 对应于 371 个基因中的 914 个可变区域,表 3 列出了它们在 cds 的不同位置出现的数目和频率.

表 3　可变区域在 cds 中发生位置

	起始片段	中部	终止片段
片段数目	189	368	357
出现频率	21 %	40 %	39 %

可以看到,选择性剪接造成产物 cds 的可变区域在起始、

中部和结尾发生的概率远不是均等的,发生在 cds 的中部或结尾部分的概率差不多,而 cds 起始部分不同的则要少得多.

对这种现象我们认为可能是从 mRNA 到蛋白质翻译过程中的保护机制导致在同一个基因的不同产物间在蛋白质的起始部分发生变化的概率要远小于在中部或结尾部分发生变化的概率.

5.4 由于选择性剪接引起的同一段序列读码框相位的变化

选择性剪接使得同一条 premRNA 在剪接时会采用不同的剪接位点或干脆跳过一个或多个外显子而得到不同的成熟 mRNA. 这样,由于这些可变区域的存在,在翻译时,很有可能发生某一段核酸序列在不同的转录产物中由于读码框相位不同而翻译成不同的氨基酸序列. 这里对这一现象进行了统计分析.

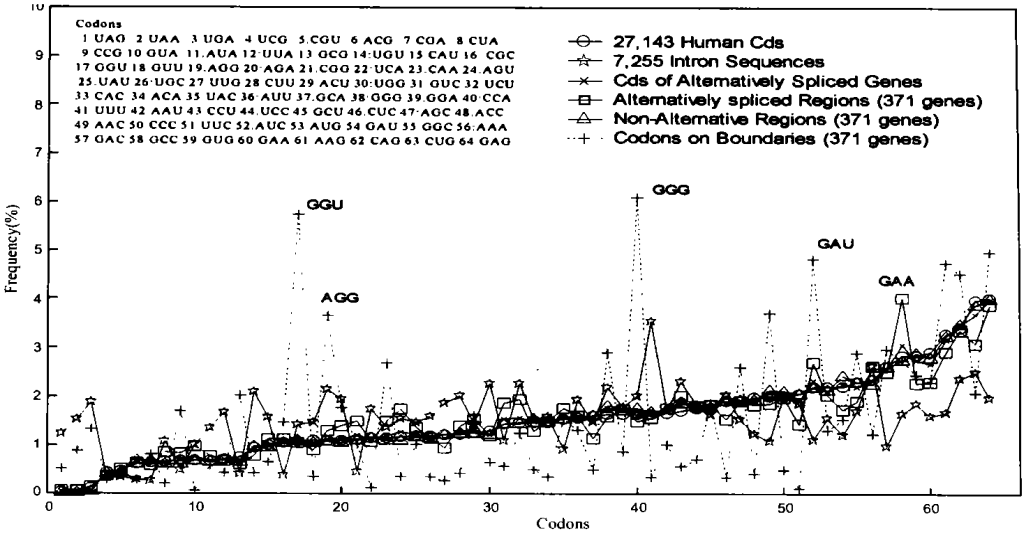
在 371 个基因中,只有 51 个基因的选择性剪接引起同一段序列读码框相位的变化,占总数的 14 %. 而在这 51 个基因中,又有 42 个基因由于序列读码框相位的变化遇到终止密码子而导致翻译的终止,另有 2 个基因以两种相位读码的序列片段发生在 cds 的起始部分. 也就是说,选择性剪接尽管采用不同的剪接位点或干脆跳过外显子,但不同 mRNA 的序列片段在翻译时仍倾向于采用同一种读码框相位,而采用不同相位时往往会导致读码框的终止. 这一结果可能对进行选择性剪接预测算法的设计有所帮助.

早期研究表明,读码框的三相位子序列的统计特性,例如依赖 GC 含量的 6 碱基熵,存在差异,这种差异正是判别阅读相位与测序时检查碱基插入缺失的算法基础(http://genio.informatik.uni-stuttgart.de/GENIO/frame/). 对选择性剪接基因的这一统计结果从一定程度上肯定了这一算法的可行性,即使对于选择性剪接的基因,不同 mRNA 的同一序列片段在翻译时的确倾向于采用同一种读码相位. 但同时也反映了这种算法在某种程度上的不可靠性,毕竟有一定比例的基因会由于选择性剪接而发生同一段序列具有两种读码相位的情况.

从这些具有两种阅读相位的序列片段的长度看,从最短的 12 个碱基到最长的 804 个碱基,它的分布与一般可变区域片段的长度分布基本一致. 值得注意的是,这种可以采用不同相位的序列片段也是可以很长的,这里有 6 条片段的长度大于 200,最长的可达 800 个碱基之多.

5.5 可变区域与不可变区域密码子使用频率

外显子与内含子序列的区分是基因预测的重要任务之一,早期的研究表明,外显子与内含子序列在密码子使用的倾向性上具有差别. 为了检验 cds 中的可变区域与不可变区域是否具有不同的密码子使用频率,我们计算了这 371 个基因的完整 cds 区,可变区域与不可变区域中,以及出现在二者交界处的密码子的使用频率,并将它们与从 CUTG 数据库(http://www.kazusa.or.jp/codon/)[12,13]提取的 27,143 条人类基因的 cds 以及从 EID 数据库(http://mcb.harvard.edu/gilbert/EID)[14]中提取的 7,255 条人类内含子序列中的密码子使用频率进行了对比. 在计算内含子序列的密码子使用频率时,假定内含子没有被剪接掉而得到其读码相位. 比较结果如图 3 所示. 为清楚起见,我们根据密码子在 27,143 条人类基因的 cds 中的出现频率进行了排序.



图中横坐标代表了 64 中密码子,纵坐标则代表了密码子的使用频率

图 3 可变区域与不可变区域密码子使用频率

从图 3 中可以看出,这 371 个基因的完整 cds 以及它们的不可变区的密码子使用频率与 27,143 个人类基因的 cds 的密码子使用频率非常接近,而它们的 cds 的可变区域中的密码子使用频率与之相比则略有变化,特别是对某几个特定密码子如 GAA. 不论是可变区域还是不可变区域,它们的密码子使用频率都与内含子序列有很大区别,而二者本身却很相近. 事实上,我们对二种区域中翻译成同一种氨基酸的不同密码子使用的倾向性也进行了比较,差别也很小(篇幅所限,具体数据这里不再列出).

值得注意的是,出现在可变区域与不可变区域边界上的密码子的使用有较强的倾向性,有几个密码子的使用频率要远大于其它,如 GGG,GGU,GAU 和 AGG. 其中,GGU 与 AGG 的

频繁出现与外显子连接处的碱基使用保守性[15]是一致的,但 GGG 与 GCU 在使用上的倾向性则需要进一步的研究. 可变区域与不可变区域边界上的密码子使用的倾向性有可能为识别 cds 中可变区域与不可变区域提供一定的线索.

6　小结

后基因组时代,基因表达调控和基因功能的研究成为生物信息学研究的核心内容. 由于选择性剪接在真核基因表达调控中的特殊地位以及其在真核生物特别是人类等高等生物中发生的普遍性,使其成为研究真核基因表达调控的重要内容之一.

目前,对于选择性剪接问题的研究还存在大量的问题没有解决. 仅仅依靠实验的方法对单个基因的选择性剪接现象和机制进行研究很难解决选择性剪接机制与选择性剪接预测的问题,必须进行更为系统更为综合的分析.

从收集选择性剪接基因的数据出发,尽可能的收集已知的选择性剪接的基因和它们的各种转录产物,并进行了适当的筛选以保证数据的质量和统计分析的可靠性. 对挑选出的 371 个人类基因,提取其各种转录产物的 cds 区,应用 ASALIGN 进行多序列比对来揭示不同 cds 间的剪接关系,提出其中的可变区域与不可变区域,并对可变区域与不可变区域的长度分布,可变区域在 cds 中出现的位置,由于选择性剪接引起的同一段序列读码框相位的变化以及可变区域与不可变区域以及二者边界上的密码子使用频率进行了统计分析,得到了一些很有意思的结果:

首先,从这 371 个基因转录产生的 mRNA 和 cds 的数目比较可以发现,有 20 %左右的选择性剪接只发生在序列的 UTR 区,而并不影响 cds 的产生. 其次,选择性剪接造成产物 cds 的可变区域更倾向于发生在 cds 区的中间或结尾部分,而可变区域的长度分布说明有相当大比例的选择性剪接现象的发生是由于存在两个距离很近的可选择的剪接位点或在某种剪接形式中跳过了一个很小的外显子. 再有,通过读码框相位的分析,我们发现选择性剪接尽管采用不同的剪接位点或干脆跳过外显子,但不同 mRNA 的序列片段在翻译时仍倾向于采用同一种读码框相位,而采用不同相位时往往会导致读码框的终止. 最后,可变区域与不可变区域密码子使用频率的统计结果表明,二者在密码子使用频率上基本是一致的并且与所有目前已知的人类基因的 cds 的密码子使用频率十分相似,但在二者的交界处密码子的使用频率却有一定的倾向性.

虽然这些统计分析目前还仅仅是很初步的,但不可否认,这些结果对于我们对选择性剪接机制与选择性剪接基因预测的进一步深入的研究提供了良好的线索. 同时也希望通过这一工作为选择性剪接问题的研究提供一条新的思路.

参考文献:

[1] Gelfand M S,et al. ASDB:database of alternatively spliced genes [J]. Nucl. Acids. Res. 1999,27:301 - 302.

[2] Mironov A A,et al. Frequent alternative splicing of human genes [J]. Genome Research. 1999,9:1288 - 1293.

[3] Croft L,et al. ISIS,the intron information system,reveals the high frequency of alternative splicing in the human genome [J]. Nat. Genet. 2000,24:340 - 341.

[4] Hanke J,et al. Alternative splicing of human genes more the rule than the exception [J]? Genome Analysis. 1999,15:389 - 390.

[5] Thanaraj TA. A clean data set of EST-confirmed splice sites from Homo sapiens and standards for clean-up procedures [J]. Nucleic Acids Res. 1999,27,2627 - 2637.

[6] Dralyuk I,et al. ASDB:database of alternatively spliced genes [J]. Nucl. Acids. Res. 2000,28:296 - 297.

[7] James Kent W,et al. The Intronerator:exploring introns and alternative splicing in Caenorhabditis elegans [J]. Nucleic Acids Res 2000,28:91 - 93.

[8] Ji H,et al. AsMamDB:An alternative splice database of mammals [J]. Nucleic Acids Res. 2001,29:260 - 263.

[9] 计宏凯,等. 针对基因选择性剪接的多序列比对算法研究 [J]. 清华大学学报(已接收).

[10] Morgenstern B,et al. Multiple DNA and protein sequence alignment based on segment-to-segment comparision [J]. Proc. Natl. Acad. Sci. USA,1996,93:12098 - 12103.

[11] Thanaraj TA. Positional characterisation of false positives from computational prediction of human splice sites [J]. Nucleic Acids Res. 2000,28:744 - 754.

[12] Nakamura Y,et al. Codon usage tabulated from the international DNA sequence databases [J]. Nucleic Acids Res. 1996,24:214 - 215.

[13] Nakamura Y,et al. Codon usage tabulated from the international DNA sequence databases;its status 1999 [J]. Nucleic Acids Res. 1999,27:292.

[14] Saxonov S,et al. EID:the Exon-Intron Database - an exhaustive database of protein-coding intron-containing genes [J]. Nucleic Acids Res. 2000,28:185 - 190.

[15] Solovyev VV,et al. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames [J]. Nucleic Acids Res. 1994,22:5156 - 5163.

作者简介:

闻　芳　女. 1975 年 4 月生于北京. 1997 年获清华大学自动化系工学学士学位,现于本校自动化系攻读模式识别与智能控制专业博士学位. 主要研究兴趣为图象处理、模式识别、生物信息学.

李衍达　男. 1936 年 10 月出生于广东省东莞市. 中国科学院院士,清华大学自动化系教授,博士生导师. 现任国务院学位委员会委员、北京生物工程学会生物信息学专业委员会主任、IEEE 高级会员、中国自然科学基金委员会委员. 主要学术方向为:信号处理,生物信息学与智能信息处理. 已发表论文百余篇及多部著作.