用生物信息学方法发现跨染色体剪接的嵌合 5′/3′ UTRs

张治华① 张 勇① 石宝晨① 邓 巍① 赵 屹② 陈润生①②*

(①中国科学院生物物理研究所, 北京 100101; ②中国科学院计算技术研究所, 北京 100080. * 联系人, E-mail: crs@sun5.ibp.ac.cn)

摘要 mRNA 的 5′/3′ UTRs 是在 mRNA 翻译过程中起重要调控作用的序列, 在 5′/3′ UTRs 上不正常的剪接模式可能导致多种严重疾病. 提出了一种基于大规模序列比对分析搜寻 5′/3′ UTRs 跨染色体剪接现象的方法, 并用此方法得到 8 例可信度高的跨染色体的 5′/3′UTRs 剪接事件, 从信息的角度证实了来自不同染色体序列剪接成为 5′/3′ UTRs 的现象是真实存在的. 对这 8 例跨染色体剪接产生的 5′/3′ UTRs 用多序列比对在剪接区域没有发现一致保守的 motif. 同时, 目前的预测算法也不能在剪接区域检测到特异性的 RNA 二级结构, 因此很难确定引导 5′/3′ UTRs 跨染色体剪接的信号.

关键词 5′/3′ UTRs mRNA 反式剪接 染色体易位 嵌合

真核生物基因的转录、表达和翻译是一个在多种因素调节控制下进行的非常复杂的过程. 目前, 已知 5′/3′ UTRs(untranslated regions)是mRNA 5′和 3′端具有复杂生物学功能的非编码序列. 它们通过其中的调控序列和调控因子(RNA附着蛋白、microRNA等)之间的相互作用, 完成调控基因的翻译、控制mRNA在胞质中的稳定性及mRNA的亚细胞定位等广泛的生物学功能. 研究显示, 在胚胎发生和发育过程中, 包括精子发生等过程都与 5′/3′ UTRs上的调控序列 或者信号因子有重要关系[1~4]. 发生在 5′/3′ UTRs (mRNA)上的突变或不正常的剪接模式, 可能造成多种严重的疾病[5,6]. 目前认为, 生命体可以通过所谓的反式剪接 (*trans*-splicing) 和染色体易位 (chromo- somal *trans*-locations)这两种机制来完成 5′/3′ UTRs的跨染色体剪接. 反式剪接最先是在体外被发现[7], 细胞内的反式剪切发现的则略晚[8]. 反式剪切现象在低等和高等的真核生物中[8,9]都有发现, 甚至在哺乳动物中也有报道[10~13]. 反式剪接的作用可能是提高(或者降低)成熟mRNA在胞质内的稳定性或者提供mRNA的亚细胞定位信号[14]. 显然, 反式剪切也提供了来自不同基因的蛋白片段发生融合的候选机 制[1,11,12]. 染色体易位的作用有可能是分离一个衔接重复(tandem duplication)[15,16], 也有可能通过染色体易位, 使得某个基因失去或者得到某个调控序列从而影响它的表达. 染色体易位、反式剪接等在很多情况下是与疾病相关的, 尤其是与肿瘤相关的疾病[15,17,18]. 发现和研究这样的表达事件是非常有意义. 然而, 由于由染色体易位, 反式剪接产生的5′/3′UTRs (mRNA), 尤其是由跨染色体剪接而成为 5′/3′ UTRs (mRNA)是在细胞内和细胞外都非常罕见的现象[8,16]. 用基于实验的方法则很难得到在基因组层次上此类事件的面貌[19,20], 因此用生物信息学的方法从数据库中挖掘这样的事件是一个非常有意思而且重要的课题.

在本文中, 我们运用基于大规模序列比对的方法在 5′/3′ UTRs数据库UTRdb[21]中选取人类的数据与人类基因组草图[22]进行相似性比对分析, 从得到的潜在的跨染色体剪接事件中去除了可能由于基因组拼接、测序, 克隆造成的错误. 最后通过查找那些在数据库注释中明确标明了出处的 5′/3′ UTRs, 作为具有可信的跨染色体剪接现象的序列.

1 材料与方法

我们从公共的 5′/3′ UTRs数据库UTRdb (databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs, http://bighost.area. ba.cnr.it/BIG/UTRHome/)[21]下载了该数据库收集的人类mRNA 5′/3′ UTRs序列. 其中 5′端UTR 26643 条, 3′端UTR 29533 条. 同时, 我们从NCBI(National Center for Biotechnology Information)的匿名FTP服务器下载了人类基因组的草图序列[22](Aug 1, 2002, updated).

如图 1 所示, 首先把人类mRNA 5′/3′ UTRs序列用Blastn[23]与人类基因组草图序列进行相似性比对, 使用的工作站是一台双Inter P4 CPU的Linux系统(2× 2.4 GHz, 2 G RAM). 为了得到与基因组草图充分相似的序列片段, 我们把Blastn参数E值设定为小于 1×10^{-10}. 我们使用一个perl编写的工具来分析Blastn的输出结果以提取具有这样结果的mRNA 5′/3′

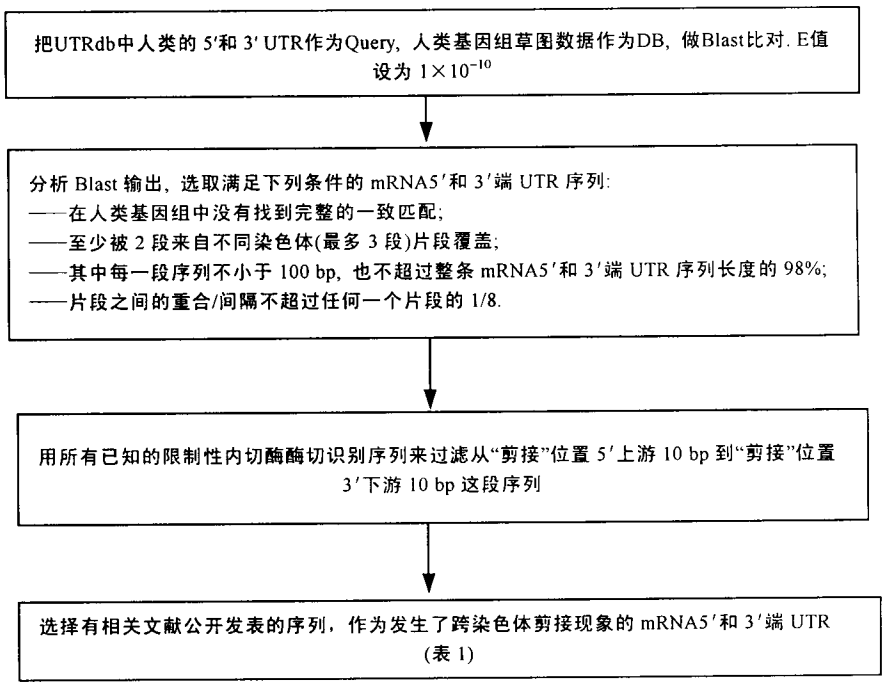

图 1 搜索 UTRdb 中跨染色体剪接的 5′和 3′ UTR 流程

UTRs 序列: (i) 该 mRNA 5′/3′ UTRs 序列没有在人类基因组中找到完整的一致匹配; (ii) 该 mRNA 5′/3′ UTRs 序列至少被 2 段(最多 3 段)来自不同染色体人类基因组相似性片段所覆盖; (iii) 其中每一段序列都不太短(大于 100 bp)，也不太长(超过整条 5′/3′ UTRs 序列的 98%); (iv) 片段之间的重合/间隔不太长，不超过 2 片段中任何一个长度的 1/8. 采取这样的策略是考虑到人类基因组草图可能的拼接错误而设定的严格限制. 这样得到可能的 mRNA 5′/3′ UTRs 序列跨染色体剪接事件的候选集，其中包括 245 条 3′ UTRs 和 268 条 5′UTR. 我们把这些候选 5′/3′ UTRs 中与基因组序列高度相似性匹配片段的两端看作是"剪接位置". 为了过滤在实验测序克隆过程中由于融合连接(fusion join)造成的虚假或者不可靠 mRNA，我们首先截取所有从"剪接位置" 5′上游 10 bp 到"剪接位置"3′下游 10 bp 的这段序列(这是因为限制性内切酶酶切识别序列长度一般在 4~6 bp)，得到一个我们称之为"剪接区域"的序列数据集. 我们用 C 语言编写的工具对这个数据集进行扫描，过滤掉包含所有已知限制性内切酶酶切识别序列的候选序列，进而得到 6 条 3′ UTRs 和 12 条 5′ UTRs. 最后，我们在这 18 条通过了筛选的序列里，选择那些注释中有确定公开发表文献的 8 条序列，认为就是可信的具有跨染色体剪接现象的 mRNA 5′/3′ UTRs(表 1).

表 1 跨染色体剪接的 5′和 3′ UTR列表[a)]

EMBL AC	UTRdb AC	5′或 3′	染色体	序列(5′→3′)
M12996	CC005643	3′	X; 17	CCCCTCCCAACCTCAATGCCCTGTAAGGATTTTGCCCA
AF111848	CC079141	3′	1; 20; 7	GGGAGGGCAGAGCAAGGACAACCCCACCACCACC
Y08201	CC049069	3′	1; 7	TTGTAAAATAAAGACCCTTTACAACAACAGC
AF057352	BB063054	5′	13; 3	AGTCAAGCAACTCAACGGAGGAGGCGAGGAG
U29943	BB002182	5′	6; 9	GAAAAGGCAGTTGAAGGAGGCAGAGAAGGGGTTGG
D49372	BB057162	5′	15;17	TGGAACAAAAATAAACCAGAAACCACCACCTCTCA
AF264784	BB107675	5′	19; 8	AGGGGGATGGACGGAAGGAAAGACCTTTTTCTC
U76368	BB010036	5′	4; 8	CAAGATAGAACCTTTAGATGTCTCACCACGAAAC

a) 所列为跨染色体剪接的 5′/3′ UTRs 的相关信息. EMBL AC 为 5′/3′UTRs 的 mRNA 在 EMBL 的登录号; UTRdb AC 为 5′/3′UTRs 在 UTRdb 的登录号; 序列为"剪接位置" 5′上游 10 bp 到"剪接位置" 3′下游 10 bp 的序列

2 结果与讨论

如上所述，目前认为有2种可能的机制可以使得作为mRNA非翻译部分的5′/3′ UTRs由来自不同染色体的序列剪接而成：(i) mRNA反式剪接[7,8,24]；(ii) 染色体易位[15]. 这些嵌合的mRNA有些在细胞的发育过程中起到非常重要的作用，有些则是致命的疾病根源[25]. 目前在实验中发现过少数几例跨染色剪接剪接[19,20]. 我们期望用生物信息学的方法得到基因组水平上跨染色体剪接的面貌. 纯粹的生物信息学的数据库搜索方法必须考虑这样几个来源的噪音. 首先，我们知道人类基因组的完成图至今仍然还在制作中[22]，以前的分析已经显示了人类基因组数据的2个版本是存在差异的[25]. 随着拼接的完善，更多的错误会被排除. 为了过滤所有潜在的由于序列拼接造成的伪mRNA反式剪接，我们把嵌合的模式限制在不同的染色体之间. 就目前已知的嵌合机制来看，mRNA反式剪接和染色体易位都是低概率事件，因此产生复杂的契合模式的可能性是极低的，因此，我们把复杂的(由3段以上不同片段嵌合而成)剪接模式作为噪声过滤. 当有多条来自不同染色体的序列和5′/3′ UTRs的某个相近区域相似并且相互覆盖时，这样的序列有很大的几率是在基因组中大量出现的重复序列，而重复序列是产生拼接错误的主要部分，为此，我们限定相互覆盖的长度不能超过任何一个片段长度的1/8.

除此之外，在分子克隆过程中，不可避免的会因为随机重组或者端对端的随机连接而出现融合连接的mRNA. 这样的mRNA具有的特征往往是在剪切位置具有限制性内切酶酶切识别序列，而这些酶切识别序列是已知的，我们利用这些酶切识别序列作为识别信号就可以过滤可能出现的融合连接mRNA. 我们扫描的范围是从"剪接"位置5′上游10 bp到"剪接"位置3′下游10 bp的这段序列，这样做是因为限制性内切酶酶切识别序列长度一般在4~6 bp，对于这样短的序列，在完整的mRNA或者5′/3′ UTRs中是完全可能随机大量出现的.

表1是我们最后在UTRdb中搜索得到的8条相对可信的跨染色体剪接的5′/3′ UTRs序列. 它们是从分别为245条和268条可能的5′/3′ UTRs"剪接"中过滤得到的，对比5′端和3′端UTRs的总数可以看出，跨染色体的剪接现象在目前的数据库中所占的比例是非常低的(小于 1.42×10^{-4}). 这个结论和以前关于mRNA的分析是相符合的[26]. 用多序列比对软件CLUSTALW[27]比较这些跨染色体剪接的5′/3′ UTRs序列的剪接区域，除了具有一个"AA"相对保守之外，没有其他明显的保守性区域(图2). 另外，我们用RNA二级结构预测软件RNAStructure[28]预测这8个个剪接区域，都能形成类似图3这样的单股茎环结构. 作为对照，我们随机生成了10条长度为35 bp的序列，RNAStructure的预测结果是其中9条形成单股茎环结构，一条形成双股茎环结构. 而对相对较长的全mRNA或者5′/3′ UTRs的二级结构预测目前并没有一致公认的有效算法，所以，就目前的数据而言，这些剪接区域的二级结构没有显著的特异性，因此我们认为目前的证据都不足以支持RNA二级结构是引导跨染色体剪接充分的信号.

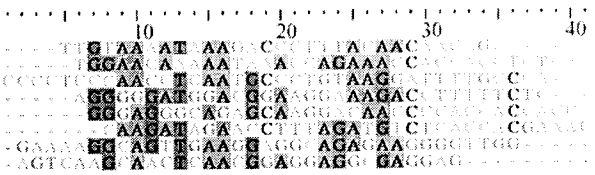

图2　跨染色体剪接的mRNA5′和3′端UTRs剪接区域序列的多序列比对结果

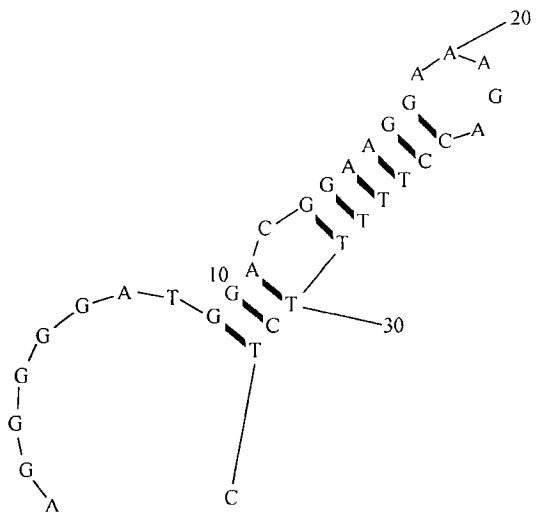

图3　RNAStructure[28]对序列EMBL AC:AF264784 (UTRdb:BB107675)剪接区域二级结构预测结果
其他剪接区域的预测结果也是形成类似的的单股茎环结构

致谢　本工作为国家重点基础研究发展规划(批准号：2003CB715907)和中国科学院知识创新工程(批准号：KSCX2-2-07)资助项目.

参 考 文 献

1 Kuersten S, Goodwin E B. The power of the 3′ UTR: Translational control and development. Nat Rev Genet, 2003, 4(8): 626~637[DOI]

2 Gray N K. Translational control by repressor proteins binding to the 5′UTR of mRNAs. Methods Mol Biol, 1998, 77: 379~397

3 Mazumder B, Seshadri V, Fox P L. Translational control by the 3′-UTR: The ends specify the means. Trends Biochem Sci, 2003, 28(2): 91~98[DOI]

4 Wilkie G S, Dickson K S, Gray N K. Regulation of mRNA translation by 5′- and 3′-UTR-binding factors. Trends Biochem Sci, 2003, 28(4): 182~188[DOI]

5 Latsi P, Pantelidis P, Vassilakis D, et al. Analysis of IL-12 p40 subunit gene and IFN-gamma G5644A polymorphisms in Idiopathic Pulmonary Fibrosis. Respir Res, 2003, 4(1): 6[DOI]

6 Lahiri D K, Chen D, Vivien D, et al. Role of cytokines in the gene expression of amyloid beta-protein precursor: Identification of a 5′-UTR-Binding nuclear factor and its implications in Alzheimer's disease. J Alzheimers Dis, 2003, 5(2): 81~90

7 Konarska M M, Padgett R A, Sharp P A. *Trans*-splicing of mRNA precursors *in vitro*. Cell, 1985, 42: 165~171

8 Bonen L. *Trans*-splicing of pre-mRNA in plants, animals, and protists. Faseb J, 1993, 7(1): 40~46

9 Pirrotta V. *Trans*-splicing in *Drosophila*. Bioessays, 2002, 24(11): 988~991[DOI]

10 Bruzik J P, Maniatis T. Spliced leader RNAs from lower eukaryotes are *trans*-spliced in mammalian cells. Nature, 1992, 360(6405): 692~695[DOI]

11 Bruzik J P, Maniatis T. Enhancer-dependent interaction between 5′ and 3′ splice sites in *trans*. Proc Natl Acad Sci USA, 1995, 92(15): 7056~7059

12 Puttaraju M, Jamison S F, Mansfield S G, et al. Spliceosome-mediated RNA trans-splicing as a tool for gene therapy. Nat Biotech, 1999, 17(3): 246~252[DOI]

13 Caudevilla C, Serra D, Miliar A, et al. Natural trans-splicing in carnitine octanoyltransferase pre-mRNAs in rat liver. Proc Natl Acad Sci USA, 1998, 95(21): 12185~12190[DOI]

14 Hyde M, Block-Alper L, Felix J, et al. Induction of secretory pathway components in yeast is associated with increased stability of their mRNA. J Cell Biol, 2002, 156(6): 993~1001[DOI]

15 Rabbitts T H, Stocks M R. Chromosomal translocation products engender new intracellular therapeutic technologies. Nat Med, 2003, 9(4): 383~386[DOI]

16 Lewin B. Genes Ⅶ, 8th ed. New York: Wiley, 2000

17 Lin R J, Sternsdorf T, Tini M, et al. Transcriptional regulation in acute promyelocytic leukemia. Oncogene, 2001, 20(49): 7204~7215[DOI]

18 Nakamura S, Matsumoto T, Jo Y, et al. Chromosomal translocation *t*(11;18)(q21;q21) in gastrointestinal mucosa associated lymphoid tissue lymphoma. J Clin Pathol, 2003, 56(1): 36~42[DOI]

19 Sit K H, Wong H B. Translocation dicentric chromosomes in prostaglandin E2 induced abortuses and possible aneusomy through asynchronous centromeric divisions. Cytogenet Cell Genet, 1981, 29(1): 60~64

20 Pergolizzi R G, Alexander E R, Rachel D, et al. *In vivo* trans-splicing of 5′ and 3′ segments of pre-mRNA directed by corresponding DNA sequences delivered by gene transfer. Mol Ther, 2003, 8: 999~1008[DOI]

21 Pesole G, Liuni S, Grillo G, et al. UTRdb and UTRsite: Specialized databases of sequences and functional elements of 5′ and 3′ untranslated regions of eukaryotic mRNAs. Update 2002. Nucleic Acids Res, 2002, 30(1): 335~340[DOI]

22 Lander E S, Linton L M, Birren B, et al. Initial sequencing and analysis of the human genome. Nature, 2001, 409(6822): 860~921[DOI]

23 Altschul S F, Madden T L, Schäffer A A, et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res, 1997, 25(17): 3389~3402[DOI]

24 Solnick D. Trans-splicing of mRNA precursors. Cell, 1985, 42(1): 157~164

25 Li S, Liao J, Cutler G, et al. Comparative analysis of human genome assemblies reveals genome-level differences. Genomics, 2002, 80(2): 138~139[DOI]

26 Romani A, Guerra E, Trerotola M, et al. Detection and analysis of spliced chimeric mRNAs in sequence databanks. Nucleic Acids Res, 2003, 31(4): e17~17[DOI]

27 Thompsom J D, Higgins D G, Gibson T J, et al. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weitghting, position specific gap penalties and weight matrichoic. Nucleic Acids Res, 1994, 22 (22): 4673~4680

28 Mathews D H, Sabina J, Zuker M, et al. Turner, expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol, 1999, 288: 911~940[DOI]

(2003-12-15 收稿, 2004-03-15 收修改稿)