

文章编号: 1004-0374(2010)07-0594-04

· 评述与综述 ·

关于非编码 RNA 研究的一些思考

陈润生

(中国科学院生物物理学研究所, 北京 100101)

摘要: 近年来大量的新的转录组实验结果表明基因组中的非编码序列绝大部分是可以表达的, 同时越来越多的事实证明转录出来的非编码 RNA, 从长度为 21~24 个核苷酸的 microRNA 到长度为数千核苷酸的 Xist, 都具有重要的生物学功能。本文主要表达了作者对未来非编码 RNA 研究方向与重点的一些想法。

关键词: 生物信息学; 非编码 RNA; 双色网络

中图分类号: Q522 **文献标识码:** A

Some reflection on the non-coding RNA research

CHEN Run-sheng

(Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: The results of large scale transcriptional analyses in recent years suggest that sequences in the noncoding parts of the genome may be expressed in the form of noncoding RNAs. And accumulating evidence suggest that noncoding RNA, from 21-24 nt microRNA in length to the length of knt magnitude Xist, has important biological function. This paper mainly expresses some reflection on the future research direction and focus of the non-coding RNA study.

Key words: bioinformatics; noncoding RNA; bicolour network

快速发展的基因组研究已经度过了二十个年头。早期的研究把测序获得的基因组 DNA 序列作为对象, 综合运用数学、计算机科学和生物学的工具, 对这些信息进行分析和释读, 希望找到基因组序列中编码蛋白质的基因区及其上游的调控区等功能元件以破译隐藏在 DNA 序列中的遗传语言规律, 为认识代谢、发育、分化、进化等重要生物学问题提供依据。

进入 21 世纪后, 基因组学与生物信息学又有了很大发展, 在新一代测序设备, 如 454、Solexa、SOLiD 等的帮助下, 每天都可以获得数以亿计的碱基数据。海量的基因组序列数据进一步证实了: DNA 上编码蛋白质的区域(也就是通常说的基因)只占人类基因组的一小部分, 不会超过整个基因组的 3%, 其余 97% 左右的 DNA 序列绝大部分仍不清楚功能, 最初科学家们习惯地把这部分 DNA 统称为“Junk”DNA, 也就是基因组中的非编码序列。

近年来大量转录组的新实验结果表明, 基因组中的非编码序列是可以表达的, 其表达产物就是非编码 RNA。而且, 越来越多的事实证明, 非编码 RNA 具有重要的生物功能, microRNA 的研究就是最突出的例子。20 世纪 90 年代, 美国 Dartmouth 医学院 Victor Ambros 小组以线虫为对象用基因打靶技术研究某些基因对其发育的影响。他们找到了一个对发育有明显干扰的基因。通常线虫要通过四个幼虫阶段才能成熟, 这个基因的突变使其只停留在第一阶段。令人们惊奇的是这个基因并不编码任何蛋白质, 而是编码一个 microRNA。以后的研究证明, 这样的小 RNA 基因在果蝇、软体动物、鱼类以及人体中都存在。除了长度为 21~24 个核苷酸的 microRNA 以外, 其他的小 RNA 也不断被发现。2006 年 7 月,

收稿日期: 2010-06-24

通讯作者 E-mail: chenrs@suns.ibp.ac.cn

独立进行研究工作的小组发现了一类新的小RNA，它们与Argonaute家族的一个分支，即Piwi类蛋白发生相互作用。这些睾丸特有的小RNA，被称为piRNA，比以前所描述的小RNA稍长一些，长度为26~31个核苷酸^[1]。很多的非编码RNA具有生物学功能，与这些功能性的非编码RNA对应的基因组序列被称为非编码基因。当今，非编码序列、非编码基因和非编码RNA的研究已成为生物学领域的研究热点，重新唤起了科学家们对“RNA世界”的重视及对“生命起源于RNA分子”这一命题的兴趣。

1 大量长的功能非编码RNA尚未发现

尽管长度小于50个核苷酸的非编码RNA(如microRNA和piRNA)等的研究已取得突破性进展，但具有功能且长度大于50个核苷酸的非编码RNA，当前只是作为个别例子被发现。我们实验室使用Affymetrix线虫全基因组tiling array芯片，所做的线虫转录组研究的结果表明^[2]：长度在50~500个核苷酸的全长非编码RNA有1000多个，而长度更长的非编码RNA数量就更多。这表明即使在很简单的多细胞生物中非编码RNA的数量也远大于microRNA和piRNA的数量。日本理化研究所(RIKEN)对小鼠转录组所作的实验发现了几万个全长的非编码RNA克隆，它们的长度大部分为几百到几千核苷酸，而功能未知^[3]。“DNA元件百科全书”计划[Encyclopedia of DNA Elements, ENCODE(<http://www.genome.gov/ENCODE>)]的研究结果也表明，人类基因组中有93%的DNA都会转录成RNA，众多转录本为非编码RNA^[4,5]。ENCODE计划挑战了关于人类基因组的传统理论，即我们的基因组不是由孤立的基因和大量“无用DNA片段”组成的，而是一个复杂的网络系统。单个基因、各种调控元件以及非编码的其他类型的功能DNA序列之间有着复杂的相互作用，共同控制着人类的生理活动。ENCODE计划促使我们重新考虑长期以来关于“基因”的概念和对于基因组功能元件及组织机制的认识。

在哺乳动物中，对于生殖与发育极为重要的剂量补偿现象就与名为*Xist*的长非编码RNA有关。*Xist*基因表达产生长度为千核苷酸数量级的非编码RNA转录本，该转录本可以改变哺乳动物一条X染色体的构象，从而将整条染色体沉默掉^[6]。这样一种机制可以保证与X染色体相关的基因表达在雄性与雌性之间维持一种平衡。*PCGEM1*是2000年被

发现的一个非编码基因，2004年这个基因被报道与前列腺癌有关^[7]。*His-1*是小鼠的一个非编码基因，1999年科学家报道它与小鼠白血病有关，而且证明这个基因参与了癌变代谢通路，并且参与了细胞周期调控^[8]。*MALAT-1*的转录本是一条8000多个核苷酸长度的非编码RNA，2003年证明该基因与非小细胞肺癌有关^[9]。最为极端的一个例子是“疯牛病”的致病蛋白Prion，在2003年Nature报道其正常构象向病态构象的转换过程中，有可能要借助非编码RNA的参与^[10]。Prion与疯牛病的传播在此前一直被认为是蛋白质到蛋白质的信息传递，不经过任何中介信使。Deleault等^[10]这篇报道打破了这一传统观念。这些例子说明长非编码RNA不仅数量众多且很多有明确的生物功能。

另一值得注意之点是：当RNA的长度增长时它们也会折叠，并形成固定的空间结构。比如：长度为70~80个核苷酸的tRNA就有三叶草样的结构。可以设想有空间结构的非编码RNA并不是像microRNA那样通过碱基匹配来实现功能的。

所以，进一步开展非编码序列、非编码基因和非编码RNA的研究，系统发现新的长非编码RNA，研究它们的空间结构与功能，可能为我们带来更多的创新机会。

2 非编码RNA是生物网络的元件

在传统的分子生物学研究中，认为一个基因表达一个蛋白，一个蛋白有一个结构，一个结构完成一个功能，也就是遵从由序列到结构再到功能这样一个思维方式。现在越来越多的事实说明，一个基因的单表达往往不能主宰一个生物学事件的发生，一个事件的发生是一组基因的同时表达，是一组蛋白质的协同作用。因此，实际上更真实表现生物学功能的，应当是一组相关基因、一组相互作用的蛋白质，这些相互作用着的蛋白质就必然构成一个网络。所以真正能更真实说明生物学功能的应该是一组组相互作用的生物大分子形成的网络，这就是系统生物学所带来的思维上的变化和研究思路的变化。

随着大量非编码RNA的发现，通过RNA与蛋白质相互作用或RNA与RNA相互作用来实现生物功能的事例也越来越多，如snRNA U1、U2、U4、U5、U6同多达75种蛋白质组成剪接复合物，负责pre-mRNA的剪接^[11]；小鼠的NRON RNA同11种蛋白质结合，控制NFAT蛋白的转运^[12]。非编码

RNA 还可以通过与 target RNA 序列匹配来定位靶标, 并进一步招募功能蛋白质来行使功能。如 C/D box snoRNA 通过其上的互补序列定位 rRNA 上的作用位点, 并招募蛋白质来对 rRNA 进行甲基化修饰^[13]; miRNA 通过其种子序列定位于特定 mRNA 的 3'UTR (untranslated region) 区域, 并通过其招募的 RISC 蛋白复合物来控制 mRNA 的稳定性或抑制 mRNA 的翻译^[14]。大量非编码 RNA 的加入将使网络结点迅速增加, 网络规模成倍的扩大。更重要的是非编码 RNA 的加入增加了各种新的相互作用, 使网络的内容更加丰富。使用非编码 RNA 构建网络还有助于研究非编码 RNA 本身的功能。过去, 人们一直以为生物功能主要是由蛋白质实现的, 因此, 以蛋白质为中心开展了大量的功能与调控研究, 发现了大量关于蛋白质产生、调控和代谢的 pathway 及相关网络。这些发现为人类认识生命活动的本质做出了巨大贡献。未来若干年对非编码 RNA 生物功能的发现将会越来越多, 对非编码 RNA 与蛋白质的相互作用的了解也, 越来越多, 并得到丰富的有非编码 RNA 参与的 pathway 及相关网络。几年前, 我们就提出了双色网络的概念^[15]: 它是指生物网络应由蛋白质与功能 RNA 两类物质构成。因为这两类物质的区别非常明显, 在网络中可以简单地用两种颜色的节点表示, 故称为双色网络。当前有关生物网络的研究都是只把蛋白质作为网络节点, 但近年来大量新研究成果表明非编码 RNA 是许多生命过程中富有活力的参与者。一些科学家认为成千上万非编码 RNA 分子组成了巨大的分子网络调节着细胞中的生命活动, 它们与蛋白质网络相对应, 同时这两类网络必然有紧密的相互作用从而构成更复杂的网络。所以未来的网络至少是双色的才更符合生命活动的实际。

为了探讨使用当前的实验数据构建双色网络的可能性, 我们进行了一些理论研究, 提出了一个新的假设: miRNA 可以调控一类特殊的非编码 RNA, 即 mRNA-like RNA 的转录水平, 从而构造了一个同时含有非编码 RNA 和蛋白质的调控网络^[16]。实验发现, 基因组上存在着大量的长的非编码转录本。其中一些和编码蛋白的 mRNA 有相似之处: 长度都很长, 都由 RNA 聚合酶 II 转录, 转录后都存在剪接、加帽加尾现象, 但是又没有蛋白编码框, 因此被称为 mRNA-like 非编码基因。少数 mRNA-like 非编码基因的功能已经得到证实, 然而绝大部分 mRNA-like 非编码基因的功能和作用机制仍然是未知的。

由于 mRNA-like ncRNA (mlRNA) 同 mRNA 在序列和结构上的相似性, 我们猜测 mlRNA 很可能也是 miRNA 的靶基因。为了验证这一猜想, 借鉴了已有对 miRNA 调控 mRNA 进行验证的方法, 通过基因芯片检测所预测 miRNA 靶基因的 RNA 水平来评估预测结果的可靠性。整个工作选取了 FANTOM 数据库中收集的 3.4 万条 mRNA-like 非编码 RNA 作为研究对象。在这 3.4 万条序列中有约 1.1 万条序列在 20 个组织中有基因表达谱数据。选取了 8 条已经确认的存在组织特异性表达的 miRNA 作为我们的 miRNA 研究集合。由于 miRNA 能够显著下调其靶基因的 RNA 水平, 所以对于组织特异表达的 miRNA 在其特异表达组织中其靶基因的表达水平应该显著低于其他组织。我们所分析的 8 条 miRNA, 根据 miRanda 预测结果和靶位点序列保守性结果预测了它们在 1.1 万条 mRNA-like 非编码 RNA 集合上的靶基因。然后, 对预测靶基因的表达谱用 Wilcoxon's rank sum test 进行检验, 结果表明有 3 条 miRNA 在 4 个特异表达的组织中其靶基因的表达水平显著下调, 显著水平和 miRNA 在 mRNA 上的调控水平相当, 这就验证了我们对 miRNA 能够调控 mRNA-like 非编码 RNA 的猜想。这一结果大大扩展了 miRNA 参与的转录后调控网络。更为有趣的是, 结果表明在 mRNA-like 非编码 RNA 中存在大量的 miRNA-encoding RNA, 而 microRNA 对这些自身的初级转录本也存在调控关系, 因此形成了一个复杂的 microRNA 间相互调控网络。

所以, 关注非编码 RNA 与传统 pathway 的相互作用, 将会使生物网络更加真实、更加完整。

3 非编码 RNA 调节的多样性

非编码 RNA 在发挥生物学功能时的作用方式是多种多样的, 以下仅以 miRNA 为例叙述其不同的调节作用。miRNA 是基因组编码的内源性小分子 RNA, 在转录时它是以前体 (pri-microRNA) 的形式被表达的, 经过 RNase III 家族的 Drosha 和 Dicer 酶的两次剪切后成为成熟的 miRNA。miRNA 主要的生物功能是在翻译水平对靶 mRNA 的表达进行负调控, 也就是抑制靶 mRNA 的表达。miRNA 的负调控作用是通过 RNA 诱导基因沉默复合物 (RNA-induced silencing complex, RISC) 完成的, 复合物的核心成员是 Ago 家族蛋白。在这一过程中 miRNA 作为特征识别因子, 将 RISC 复合物引导向互补的 mRNA 目标, 其识别位点存在于 mRNA 的 3'UTR。

miRNA 的负调控作用不仅存在于靶 mRNA 的 3' UTR 区, 也可发生在 5' UTR 区。我们实验室使用人乳腺癌细胞株进行的 miRNA-10a 对 *Hoxd4* 基因表达调节的研究结果显示, 这一抑制作用是发生在 *Hoxd4* 基因的启动子区域^[17]。

miRNA 不仅具有负调控作用, 也可以激活基因的表达。Vasudevan 实验室发现, 在细胞周期过程中, miRNA 效应在抑制作用和活化作用间摆动; 在静态细胞中(G₀期), miRNA 活化翻译和上调基因表达, 而在其他细胞循环/增殖期则继续发挥抑制作用^[18, 19]。miRNA 激活作用与富含腺嘌呤/尿嘧啶元件(adenylate/uridylate-rich elements, AREs)相关。ARE 是 miRNA 活化翻译的信号, 在 miRNA 指导下, miRISC 复合物成员, 如 Ago、FXRP 被招募到 ARE 上, 激活翻译、上调基因表达。ARE 元件是一种 mRNA 不稳定元件, 位于 mRNA 3' UTR, 对很多 mRNA 来说, 它都是保守的。

Eiring 等^[20]的研究表明, miRNA 不仅可通过调节 mRNA 的表达发挥生物作用, 还可以通过调节蛋白来发挥生物功能。俄亥俄州立大学的 Danilo Perrotti 实验室及其合作者在以淋巴细胞白血病为对象, 研究 miRNA 对 mRNA 表达的影响时发现: 当 miR-328 显著缺失时, 激酶依赖性的 MAPK1-hnRNP E2 路径被激活, 慢性粒细胞白血病原始细胞出现危象, 如果重新表达 miR-328, 它就会与核不均一核糖核蛋白(heterogeneous ribonucleoproteins, hnRNPs)结合, 将及时挽救危象。这个相互作用的过程并不依赖 microRNA 的种子序列, 并且会激发 CCAAT/增强子结合蛋白 a (CEBPA) mRNA 的释放, 促使粒细胞分化。这些结果表明, miRNA 具有控制细胞命运的双重路径, 它不仅像通常的那样, 通过 mRNA 来发挥作用, 还通过调节蛋白质来发挥作用。

miRNA 的作用方式尚且如此复杂, 其他非编码 RNA 行使功能的途径可能更为多样, 特别值得注意的是长非编码 RNA, 且不要把经典 miRNA 的作用方式作为定式。

4 NcRNA 与疾病紧密相关

非编码 RNA, 特别是 miRNA 与疾病相关的研究已数不胜数, 有些已用于临床, 这里不再赘述。这里只想说, 非编码 RNA 研究是一个紧密结合实际, 又可迅速用于实际的领域, 应特别关注从基础到应用的转化。

[参 考 文 献]

- [1] Girard A, Sachidanandam R, Carmell MA, et al. A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, 2006, 442: 199
- [2] He H, Wang J, Liu T, et al. Mapping the *C. elegans* non-coding transcriptome with a whole genome tiling microarray. *Genome Res*, 2007, 17(10): 1-7
- [3] Okazaki Y, Furuno M, Kasukawa T, et al. Analysis of the mouse transcriptome based on functional annotation of 60, 770 full-length cDNAs. *Nature*, 2002, 420: 563-73
- [4] The ENCODE Project Consortium. *Science*, 2004, 306: 636
- [5] The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 2007, 447: 799-816
- [6] Chureau C, Prissette M, Bourdet A, et al. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res*, 2002, 12(6): 894-908
- [7] Petrovics G, Zhang W, Makarem M, et al. Elevated expression of *PCGEM1*, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene*, 2004, 23: 605-11
- [8] Xu F, McFarland M, Askew DS. His-1: a noncoding RNA implicated in mouse leukemogenesis. *Histol Histopathol*, 1999, 14: 235
- [9] Ji P, Diederichs S, Wang W, et al. MALAT-1, a novel noncoding RNA, and thymosin β 4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 2003, 22: 8031-41
- [10] Deleault NR, Lucassen RW, Supattapone S. RNA molecules stimulate prion protein conversion. *Nature*, 2003, 425: 717-20
- [11] Will CL, Lüthmann R. Spliceosomal UsnRNP biogenesis, structure and function. *Curr Opin Cell Biol*, 2001, 13: 290-301
- [12] Willingham AT, Orth AP, Batalov S, et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, 2005, 309: 1570-3
- [13] Chen CL, Liang D, Zhou H, et al. The high diversity of snoRNAs in plants: identification and comparative study of 120 snoRNA genes from *Oryza sativa*. *Nucleic Acids Res*, 2003, 31: 2601-13
- [14] Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 2004, 116: 281-97
- [15] 吴涛, 何顺民, 陈润生. 非编码基因的功能、调控与双色网络[M]//科学发展报告. 北京: 科学出版社, 2007: 103-8
- [16] Zhao Y, He S, Liu C, et al. MicroRNA regulation of messenger-like noncoding RNAs: a network of mutual microRNA control. *Trends Genetics*, 2008, 24: 323-7
- [17] Tan Y, Zhang B, Wu T, et al. Transcriptional inhibition of *Hoxd4* expression by miRNA-10a in human breast cancer cells. *BMC Mol Biol*, 2009, 10: 1-9
- [18] 赵爽, 刘默芳. MicroRNA作用机制研究的新进展. *中国科学C辑: 生命科学*, 2009, 39(1): 109-13
- [19] Vasudevan S, Tong Y, Steitz JA. Switching from repression to activation: microRNAs can up-regulate translation. *Science*, 2007, 318(5858): 1931-4
- [20] Eiring AM, Harb JG, Neviani P, et al. miR-328 functions as an RNA decoy to modulate hnRNP E2 regulation of mRNA translation in leukemic blasts. *Cell*, 2010, 140: 652-65