PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(5): 536~540
www.pibb.ac.cn

真核生物中的“双重编码”现象 *

申志勇　李稚锋　杭兴宜　张成岗 **
(军事医学科学院放射与辐射医学研究所，蛋白质组学国家重点实验室，北京 100850)

摘要　双重编码(dual coding)是指一段 mRNA 序列同方向上存在两个开放阅读框架，从而可编码不同氨基酸序列的现象．双重编码现象在原核生物、噬菌体和病毒中作为一种可有效利用基因组的方式而普遍存在．新近报道，在真核生物中也存在双重编码现象，对于认识真核生物基因表达调控的机制提供了新的信息．

关键词　双重编码基因，开放阅读框架，真核生物
学科分类号　Q7
DOI: 10.3724/SP.J.1206.2008.00620

双重编码(dual coding)是低等生物基因表达调控的一种重要方式，可实现对规模较小原核基因组中遗传信息的最大程度利用，对于以原核生物为主的早期生命体系的生存和进化具有重要意义[1, 2]．近年来，陆续发现真核生物中也存在利用一个转录本进行多重编码的现象，为本已复杂的真核基因表达调控领域提供了新的数据，本文主要对其中的双重编码现象进行综述．

1　双重编码基因的发现和一般类型

早期研究发现，噬菌体中部分基因可与邻近的基因重叠或共享核苷酸[3]，即存在从同一段核酸序列中编码产生不同蛋白质的情况，称之为重叠编码(overprinting)现象，大量存在于病毒基因组、线粒体基因组甚至细菌基因组中[4, 5]，从而导致很多基因被完全包含于其他基因中[6]．该现象最初在病毒、线粒体和染色体外的核元件(extrachromosomal nuclear elements)中得到鉴定，之后有研究表明细菌基因组中也同样存在非常多的重叠编码现象．随着这些原核基因中多重编码现象的发现，人们也开始关注真核生物基因组中是否也存在这种现象．从三联体密码子的基本性质考虑，对于任何一条成熟的 mRNA 序列，其三个相位上通常会包含多个潜在的开放阅读框(open reading frame，ORF)[7]．理论上，当一条序列上能够同时翻译表达多个阅读框时，就会形成多重编码现象．原核基因的多重编码现象一般仅限于长度为几个核苷酸重叠，与质粒或染色体大小有关[8]，而真核生物的基因组一般比较大，受进化选择压力而采用多重编码来承载更多遗传信息的可能性较小，因此，真核生物中双重编码现象可能是进化残留的体现，但部分具有重要功能的双重编码基因却有可能被保留而发挥作用．

一般地，双重编码基因(dual coding gene)可被定义为同一个转录本在正相位上同时存在两个 ORF 的现象，即同一个成熟的 mRNA 可通过使用不同 ORF 而编码不同蛋白质，其中一个 ORF 通常对应于文献或数据库中蛋白质序列注释，称为组成型读码框(canonical reading frame，CRF)，另一个相对于 CRF 而言则可称为选择型读码框(alternative reading frame，ARF)[9]．目前对于真核生物基因组中这种双重编码基因的意义还存在很大争议，这是因为双重编码对于形成理想的核苷酸排列顺序来说具有相当大的难度(显然将限制氨基酸组成的灵活性)[10]，尤其是其中一个 ORF 中碱基的无义突变有可能会“连带”导致另一 ORF 中氨基酸组成的改变，但对已知真核生物的研究显示，从低等到高等均存在双重编码现象，如上游可读框(upstream

* 国家重点基础研究发展规划项目(973) (2006CB504100, 2003CB715900)，国家自然科学基金资助项目(30771230, 30772293)和北京市自然科学基金重点项目(7061004).
** 通讯联系人.
Tel: 010-66931590，E-mail: zhangcg@bmi.ac.cn
收稿日期：2008-09-06，接受日期：2008-11-07

ORF，uORF)就是一种双重编码现象，它会影响真菌类、植物和动物的机体发育与生长．在对水稻和拟南芥的研究中发现，表达植物 S- 腺苷甲硫氨酸脱羧酶(AdoMetDC) mRNA 的 5′端引导序列都包含一对非常保守的、重叠一个碱基的 uORF[11, 12]．对果蝇等双翅目昆虫的全基因组对比分析也发现的确存在 uORF[13]．另外，果蝇中还发现两个含有双顺反子的转录本(stoned 基因和 snapin 基因)也受双重编码调节，其中第二个顺反子的有效翻译取决于第一个顺反子 ORF 内密码子 AUG 的缺失与否[14]．现已发现，在高等哺乳动物中也存在双重编码现象，例如 GNAS1、XBP1 和 INK4a 等 3 个基因(图 1)[15, 16]，其中 GNAS1 和 XBP1 都是在一个转录本内包含有两个 ORF 的双重编码基因，GNAS1 在一定条件下甚至可同时采用这两个 ORF 编码不同的蛋白质(Xlas 和 ALEX)，而 XBP1 则一次编码产生一个蛋白质，通常是 $XBP1^U$，而在一些特殊条件下则通过选择性使用核酸内切酶 IRE1 切除转录本上一段长为 26 bp 的间隔，再连接 ORF A 和 ORF B 产生 $XBP1^S$．INK4a 则是通过产生两个不同的剪接异构体在外显子 E2 中使用不同的 ORF 以产生不同蛋白质($p16^{INK4a}$ 和 $p19^{ARF}$)[17]．

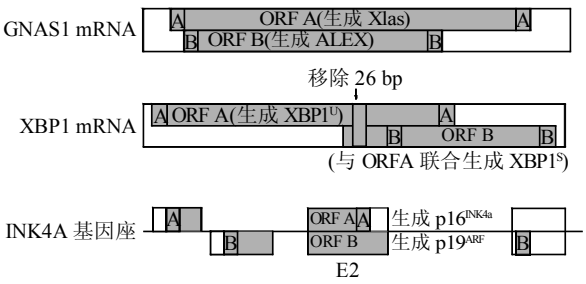

Fig. 1　Three known examples of mammalian dual-coding genes: GNAS1, XBP1 and INK4A[9]

图 1　哺乳动物中已知的 3 个双重编码基因: GNAS1, XBP1 和 INK4A[9]

研究发现，双重编码区域的重叠方式与 ORF 的选择性使用密切相关，通常可包括双 ORF相互独立、双 ORF 部分重叠和双 ORF 嵌套 3 种类型(图 2)，例如前面提到的 XBP1 基因属于部分重叠类型，分别编码产生 $XBP1^U$ 和 $XBP1^S$ 两种蛋白质[18]，而 GNAS1 基因则属于双 ORF 嵌套类型，其转录本中含有两个嵌套重叠、序列和长度不同的 ORF 分别编码产生结构上互不相关的两个蛋白质 Xlas 和 ALEX[16]．

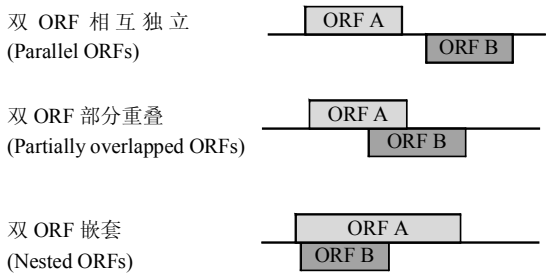

Fig. 2　Typical distribution pattern of two ORFs in dual-coding gene

图 2　双重编码基因中两个 ORF 的典型分布方式

2　真核生物双重编码现象的研究策略

利用生物信息学技术挖掘基因组中的遗传信息一直是基因组数据分析中的重要课题，目前，针对原核生物中双重编码区域内可变 ORF 的数据库有 AlterORF(a database of alternate open reading frames，http://www.alterorf.cl)，针对真核生物有显示某一基因所有外显子连接的 altGraphX 数据库[19]和包含所有已知基因的蛋白质数据库(Ensembl database，http://www.epdb.pitt.edu)[20]．研究人员可通过访问 altGraphX 数据库获取特定基因的剪接图，通过该图得到所有可能的 ORF，然后与 Ensembl 数据库中所有已知蛋白质序列进行比对分析，最终鉴定出含有 CRF 和 ARF 的双重编码外显子[21]．

对双重编码现象进行研究时，通常首先利用一些分析软件如 EMBOSS、ORFfinder、ClustalW 等对物种的基因组进行 ORF、进化树及结构域分析，寻找全部可能的 ORF，然后通过一些指标(如偏移比率 *K*a/*K*s、长度范围、打分算法)或构建相应的密码子模型等，筛查得到真正具有双重编码的序列．在数据的挖掘上可以采用如贝叶斯等算法来评估编码的可能性[22]，也可通过计算外显子中基于无义到同义的偏移比率(*K*a/*K*s)来评估基因组区域中蛋白质编码的可能性．研究发现，在某些情况下，双重编码基因可能在功能约束上比较宽松，即不局限于某种特定功能，或者蛋白质水平上产物多样，即因无义突变形成不完全的蛋白质产物多，因此，其 *K*a/*K*s 比率不可能小于 1[23]．另外还有通过计算 CRF 和 ARF 中每个密码子位置上的核苷酸组成和氨基酸的使用情况，然后与整个基因组的使用对比来挖掘序列 GC 含量数据，发现 CRF 在第一和第三密码子位置上 GC 含量丰富，与之对应的 ARF

则在第二密码子位置上 GC%增加显著[24]. 最近有人专门针对哺乳动物中双重编码现象进行分析，以 NCBI 上人与小鼠的 14 159 个同源基因为筛查对象，首次明确定义了双重编码基因概念并构建了基于重叠 ORF 的密码子模型，通过创建一个 CCRT (codon column replacement test)指标(用来评估筛查的基因序列含有 ARF 的可能性)的统计筛查方法，最终鉴定出 40 个进化上含高度保守重叠编码区域的候选基因，从而得出双重编码现象在真核生物中并非偶然存在的结论[9]，认为这很可能是一种进化残留现象.

3 真核生物双重编码现象与可变剪接及进化的关系

在真核生物中，可变剪接作为基因表达调控的重要机制，同双重编码一样都能够增加单一基因编码蛋白质产物的多样性. 已知人类基因中 40%～60%都能够发生可变剪接[25]，通过使用不同的外显子组合，可变剪接将会产生出一些双重编码区域，使用不同 ORF 来编码氨基酸，这些区域中同一外显子序列将被不同的转录本之间所共享[24]. 研究还发现，不仅是人的一些可变剪接基因中包含有多重编码区域，在秀丽线虫和果蝇的可变剪接基因中也分别至少有 6.8%和 2.3%的基因包含有多重编码区域[24]，这说明多重编码与可变剪接密切相关. 有研究对所有的导致读码框转换的剪接事件进行归类，发现外显子跳过(exon skipping)是其中最普遍的剪接事件，占整个剪接事件的 57%，而可变供体 / 受体剪接位点和内含子保留的剪接事件则大致相等，各占 14%[21]. 可见对于一个典型双重编码基因而言，很可能在进化中采用不同的剪接模式，从而导致 ARF 的出现. 因此，理论上而言，通过检测同源基因中该 ORF 是否在其他物种中也表达，就有可能直接确定该双重编码的进化起源. 目前，已有研究通过鉴定其他动物物种与人的相应 ORF 中同源序列是否被翻译表达，来追踪确定人基因组中双重编码区域的进化过程[22]. 研究者通过对比分析黑猩猩、大鼠、小鼠、鸡和狗的基因组，发现除了极少数情况外，在这些正向同源区域中，终止密码子都会在 CRF 和 ARF 当中的某一个出现，经统计发现，在物种的双重编码区域中两个 ORF 之一含有一个或多个终止密码子的现象为鸡 80%、狗 42%、小鼠 45%、大鼠 54%、黑猩猩 4%，说明双重编码区域内的许多 ARF 是在哺乳动物中出现的，这当中又有很大一部分是在灵长类和啮齿类动物分化后才出现的，其中一些可能只存在于灵长类，而且每个谱系看来都会存在一些特有的、在人类基因组中未出现的双重编码区域[24].

另外，从进化角度来看，对比分析不同物种间双重编码区域内同一编码区中各 ORF 的选择强度，即物种更多地倾向于选择哪一个 ORF 进行翻译也很有意义. Han 等通过统计人与黑猩猩的同源序列编码区域中各氨基酸的替换率发现，处于 CRF 内的氨基酸序列通常要比处于 ARF 内的在进化上更加保守，仅有 9%的处于 ARF 中区域内的氨基酸替换率比相应的处于 CRF 中的低，这就意味着 CRF 在进化上所受到的选择压力较 ARF 更大[24]，并不容易发生改变.

4 真核生物双重编码基因的翻译调控

从同一个成熟 mRNA 上可以翻译出两个不同蛋白质现象的存在，表明在生物体内必然存在一个调控机制来决定究竟哪个 ORF 被翻译或者两个 ORF 均被翻译. 目前，在病毒与酵母中已发现一些调控机制，如程序性翻译移位(programmed translational frameshifting)[1]和核糖体支路(ribosome shunting)机制[26]，对于更高等的真核生物的研究也发现一种扫描遗漏(leaky scanning)[27]机制，即对于一个 mRNA 来说，如果从其第一个起始密码子 ATG 位置所获得的序列对翻译起始来说不是最佳的话，40 S 的核糖体亚基可能会继续扫描该 mRNA 并在下一个 ATG 位置开始翻译，例如研究发现，大鼠中 Xlas/ ALEX mRNA 可能就受该机制调节，40 S 的核糖体亚基通过跳过第一个起始密码子 AUG 而扫描到 ALEX 蛋白的起始密码子 AUG 处开始翻译该蛋白质. 目前比较认同的观点，是在真核生物中蛋白质的合成起始是一个动态、灵活的过程，包含多种机制以满足真核细胞各种生命活动的需求，而进化上保留存在的双重编码机制正是其灵活的体现. 双重编码基因所产生的两个蛋白质通常(或理论上)具有功能相关性，如 INK4a 所产生的两个蛋白质 $p16^{INK4a}$ 和 $p19^{ARF}$ 各自调控相对独立的肿瘤抑制通路，对维持癌症和生理性的衰亡之间的平衡发挥重要作用[21]. 对人体内具有双重编码的 XBP1 基因所产生的两个蛋白质 $XBP1^{U}$ 和 $XBP1^{S}$ 进行功能研究，发现 $XBP1^{U}$ 作为 $XBP1^{S}$ 的负反馈调控因子，能够抑制 $BP1^{S}$ 并促其降解，从而导致受 $XBP1^{S}$ 调控的基因不能转录，最终达到阻止未

折叠蛋白反应(unfolded protein response，UPR)激活的目的.

5 从信息熵的角度看待双重编码现象

基因组是遗传信息的载体. 在“生老病死”的整个生命过程中，表型是基因型展示的最终结果，而这一点基本上可以从信息流的角度进行相对度量和借鉴. 换言之，基因组 DNA 序列中存储的应该是一种信息的表征方式，而包括转录、翻译等过程在内的基因表达调控则是这种“内嵌式(遗传)信息”展示的具体过程. 一个物种的复杂程度(或进化等级)与基因组的信息量是正相关的，但双重编码现象显然打破了这种认识. 从信息量角度而言，Shannon 提出将“信息熵”作为信息的量化标准，认为一个系统越是有序，信息熵就越低，反之，系统越混乱，信息熵就越高. 而双重编码所造成的序列信息的重叠可能引起信息的简并化，尤其是密码子的简并在理论上会降低信息熵. 因此，长期以来基因组内各种重叠编码现象就一直困扰着进化学家、遗传学家和序列分析人员. 目前对该现象的解释是：重叠编码作为一种有效机制可增加基因序列单位长度内的信息量[27]，而重叠中的双重编码现象之所以能够在进化的选择过程中保留下来，其对于生物体最大意义就在于此，尤其对于那些基因组较小的原核生物和病毒来说双重编码的生物学意义就在于能够更加有效地利用空间，编码更多生命活动所必须的信息. 有研究发现，当双重编码区域中包含有共同的重要残基时，该方式还能够减少 DNA 序列中的一些不稳定点，同时它还能够增强一个基因的部分功能而不需要产生全长产物[24, 28].

6 展 望

真核生物中存在的双重编码现象对于理解本已十分复杂的真核基因表达调控提供了新的信息，提示真核基因的编码能力仍然存在认识深化的空间. 真核生物的基因表达调控是复杂的，基因组中除了双重编码外，还存在假基因[29]、嵌套基因以及密码子通读等现象，这些都还有待于进一步研究. 总之，双重编码的研究将有助于进一步揭示真核基因的表达机制，对研究基因组的进化、演示生命进化的历史起到积极作用，从而有助于人类更好地理解生命的起源和进化.

参 考 文 献

1 Choudhuri S. Gene regulation and molecular toxicology. Toxicol Mech Method, 2004, **15**(1): 1～23

2 Huynen M A, Konings D A, Hogeweg P. Multiple coding and the evolutionary properties of RNA secondary structure. J Theor Biol, 1993, **165**(2): 251～267

3 Barrell B G, Air G M, Hutchison CA 3rd. Overlapping genes in bacteriophage phiX174. Nature, 1976, **264**(5581): 34～41

4 Krakauer D C. Stability and evolution of overlapping genes. Evolution, 2000, **54**(3): 731～739

5 Rogozin I B, Spiridonov A N, Sorokin A V, *et al*. Purifying and directional selection in overlapping prokaryotic genes. Trends Genet, 2002, **18**(5): 228～232

6 Pavesi A, de Laco B, Granero M I, *et al*. On the informational content of overlapping genes in prokaryotic and eukaryotic viruses. J Mol Evol, 1997, **44**(6): 625～631

7 Pedroso I, Rivera G, Lazo F, *et al*. AlterORF: a database of alternate open reading frames. Nucleic Acids Res, 2008, **36**(Database issue): D517～518

8 Johnson Z I, Chisholm S W. Properties of overlapping genes are conserved across microbial genomes. Genome Res, 2004, **14**(11): 2268～2272

9 Chung W Y, Wadhawan S, Szklarczyk R, *et al*. A first look at ARFome: dual-coding genes in mammalian genomes. PLoS Comput Biol, 2007, **3**(5): e91

10 Keese P K, Gibbs A. Origins of genes: "big bang" or continuous creation?. Proc Natl Acad Sci USA, 1992, **89**(20): 9489～9493

11 Franceschetti M, Hanfrey C, Scaramagli S, *et al*. Characterization of monocot and dicot plant S-adenosyl-l-methionine decarboxylase gene families including identification in the mRNA of a highly conserved pair of upstream overlapping open reading frames. Biochem J, 2001, **353**(Pt 2): 403～409

12 Hanfrey C, Elliott K A, Franceschetti M, *et al*. A dual upstream open reading frame-based autoregulatory circuit controlling polyamine-responsive translation. J Biol Chem, 2005, **280**(47): 39229～39237

13 Hayden C A, Bosco G. Comparative genomic analysis of novel conserved peptide upstream open reading frames in *Drosophila melanogaster* and other dipteran species. BMC Genomics, 2008, **9**: 61

14 Wall A A, Phillips A M, Kelly L E. Effective translation of the second cistron in two *Drosophila* dicistronic. J Biol Chem, 2005, **280**(30): 27670～27678

15 Nekrutenko A, He J. Functionality of unspliced XBP1 is required to explain evolution of overlapping reading frames. Trends Genet, 2006, **22**(12): 645～648

16 Nekrutenko A, Wadhawan S, Goetting-Minesky P, *et al*. Oscillating evolution of a mammalian locus with overlapping reading frames: an XLalphas/ALEX relay. PLoS Genet, 2005, **1**(2): e18

17 Kozak M. Extensively overlapping reading frames in a second mammalian gene. EMBO Rep, 2001, **2**(9): 768～769

18 Ku S C, Lee J, Lau J, *et al*. XBP-1, a novel human T-lymphotropic virus type 1 (HTLV-1) tax binding protein, activates HTLV-1 basal and tax-activated transcription. J Virol, 2008, **82**(9): 4343～4353

19 Sugnet C W, Kent W J, Ares M Jr, *et al*. Transcriptome and genome conservation of alternative splicing events in humans and mice. Pac Symp Biocomput, Hawaii, 2004. 66～77
20 Kelley L A, Sutcliffe M J. OLDERADO: on-line database of ensemble representatives and domains. On Line Database of Ensemble Representatives And DOmains. Protein Sci, 1997, **6**(12): 2628～2630
21 Szklarczyk R, Heringa J, Pond S K, *et al*. Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. Proc Natl Acad Sci USA, 2007, **104**(31): 12807～12812
22 Hanada K, Zhang X, Borevitz J O, *et al*. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res, 2007, **17**(5): 632～640
23 Nekrutenko A, Makova K D, Li W H. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. Genome Res, 2002, **12**(1): 198～202
24 Liang H, Landweber L F. A genome-wide study of dual coding regions in human alternatively spliced genes. Genome Res, 2006, **16**(2): 190～196
25 Mironov A A, Fickett J W, Gelfand M S. Frequent alternative splicing of human genes. Genome Res, 1999, **9**(12): 1288～1293
26 Kozak M. New ways of initiating translation in eukaryotes?. Mol Cell Biol, 2001, **21**(6): 1899～1907
27 Weiland J J, Dreher T W. Infectious TYMV RNA from cloned cDNA: effects *in vitro* and *in vivo* of point substitutions in the initiation codons of two extensively overlapping ORFs. Nucleic Acids Res, 1989, **17**(12): 4675～4687
28 Peleg O, Kirzhner V, Trifonov E, *et al*. Overlapping messages and survivability. J Mol Evol, 2004, **59**(4): 520～527
29 许 亮, 卢向阳, 易 克, 等. 假基因的研究进展. 生命的化学, 2003, **23**(6): 406～409
Xu L, Lu X Y,Yi K, *et al*. Chemistry of Life, 2003, **23**(6): 406～409

Dual Coding Genes in Eukaryote*

SHEN Zhi-Yong, LI Zhi-Feng, HANG Xing-Yi, ZHANG Cheng-Gang**
(*Beijing Institute of Radiation Medicine, State Key Laboratory of Proteomics, Beijing* 100850, China)

Abstract Dual coding is a phenomenon which refers to a mature mRNA that contains two open reading frames (ORFs) in the same direction to code two different proteins. This phenomenon is quite common in prokaryote, bacteriophages and viruses as an economical and effective way to present the genome information. However, recent reports suggest that the dual coding phenomenon does also occur in eukaryote. Dual coding will help us to further understand the complicated regulation of eukaryotic genes or even the law of molecular evolution.

Key words dual coding gene, open reading frame, eukaryote
DOI: 10.3724/SP.J.1206.2008.00620

*This work was supported by grants from National Basic Research Project (2006CB504100, 2003CB715900), The National Natural Science Foundation of China (30771230, 30772293), Major Program for Science and Technology Research of Beijing Municipal Bureau (7061004).
**Corresponding author.
Tel: 86-10-66931590, E-mail: zhangcg@bmi.ac.cn
Received: September 6, 2008 Accepted: November 7, 2008