文章编号:1000-582X(2008)08-0961-04

最小汉明距离译码的原核生物 DNA 表达效率分析

冯文江,初　春,龙红梅
(重庆大学 通信工程学院,重庆 400030)

摘　要:鉴于通信系统结构,将分子生物的信息传递过程用通信模型描述,采用最小汉明距离译码算法,分析核糖体 16S rRNA 的突变对原核生物 DNA 翻译效率表达的影响,仿真结果表明原核生物以 16S rRNA 作为一个标准的差错校验码对 DNA 全序列进行纠错,证明了运用通信编码理论分析原核生物的遗传信息传递的可行性。

关键词:核糖体 RNA;最小汉明距离;基因表达;基因突变
中图分类号:Q753;TN911.2　　**文献标志码**:A

Expression efficiency of prokaryotic DNA based on the minimum Hamming distance decoding

FENG Wen-jiang, *CHU Chun*, *LONG Hong-mei*
(College of Communication Engineering, Chongqing University, Chongqing 400030, P. R. China)

Abstract: We presented herein a communication theoretical model which used the translation of gene expression. Based on the minimum Hamming distance decoding, the processing of the expression efficiency of prokaryotic DNA was analyzed and simulated when the ribosome 16S RNA mutate. The results showed prokaryotes could correct the DNA based on the ribosome 16S RNA. It was demonstrated that the communications coding theory in the analysis of the genetic information transfer of prokaryotes was efficient.

Key words: ribosome RNA; the minimum Hamming distance; gene expression; gene mutation

根据现代分子生物学理论,在原核生物 DNA 的转录翻译过程中,mRNA 与核糖体的结合是 mRNA 翻译启动的基本因素。同时,mRNA 与核糖体结合效率的高低,直接影响着目标基因是否翻译以及翻译效率的高低。结合基因组信息模型与人类对复杂的网状通信系统进行控制的信令系统和编码具有一定的相似性[1]的思路,美国学者 May 对基因翻译表达的初始阶段建立了差错控制的通信模型,并设计了一种基于线形分组码的编译码算法。在此基础上,绘制了 DNA 序列的码字和基于原核生物 16S rRNA 的保守区域码字所设计的标准码字之间的最小汉明距离图,以此来达到确定基因表达翻译的准确性,诸如识别包含 mRNA 引导头的序列区域、指出和识别开放阅读框的初始位置以及区分确定翻译序列和非翻译序列等的目的。

笔者以 May 的工作为基础,进一步分析了当 16S rRNA 发生突变时,其最小汉明距离值的变化,以此验证 May 的方法可否正确反映原核生物

收稿日期:2008-04-08
基金项目:高等学校博士学科点专项科研基金资助项目(20050611022)
作者简介:冯文江(1963-),男,重庆大学教授,主要从事宽带无线接入技术、通信信号处理的研究,(Tel)023-65106498;(E-mail)fwj@ccee.cqu.edu.cn。

16S rRNA发生突变后的生物学特性。仿真分析表明,原核生物可能具有以16S rRNA作为一个标准的差错校验码对DNA全序列进行纠错检查的修复机制。

1 16S rRNA的特性

16S rRNA是所有原核生物蛋白质合成必需的一种核糖体RNA,它能够用来鉴定所有细菌的亲缘关系。结构分析、碱基修饰和突变证明了16S rRNA分子中某些碱基是核糖体的功能所必需的。如用大肠菌素(colicin)E3切除16S rRNA的3`端50个核苷酸,可丧失核糖体结合IF3、识别mRNA和结合tRNA的能力,完全阻止翻译的起始。rRNA的突变影响翻译的特异性。在16S rRNA的3`端区域存在与mRNA互补的保守序列UCCUCC,该序列的点突变可抑制翻译的进行,说明rRNA和mRNA间的互补对翻译的起始有重要作用。rRNA与mRNA间的识别作用能引起移码,越过mRNA上的一个碱基,表明在核糖体的移动过程中,rRNA与mRNA可能发生直接的碱基识别。对原核生物蛋白质翻译的启动效率有较大影响的因素包括mRNA的SD区、ATG起始密码与核糖体16S rRNA的结合性能、SD区与起始密码间的精确距离以及SD区上游和ATG下游区与核糖体RNA间配对[2]。

2 基因表达过程中的通信模型

遗传学将遗传信息的流动方向称为信息流。信息流的方向可以用科学家Francis Crick于1954年提出的"中心法则"来描述。有关遗传信息的传递过程如图1所示[2]。

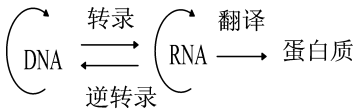

图1 遗传信息的传递过程

在翻译的起始,核糖体和mRNA的结合位被称为引导头,引导头区域由起始密码子前面的上游碱基组成。常用的初始密码子是AUG,标明了mRNA的编码区域开始,即蛋白质从这点开始编码,直至遇到终止密码子为止。翻译起始阶段如图2所示[3]。

核糖体识别引导头区域信号和初始密码子这一复杂的生物机制可以用通信系统模型来表征[4-6],如图3所示。

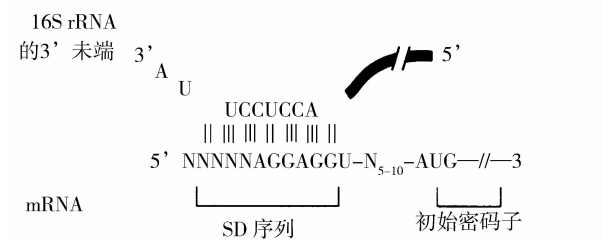

图2 原核生物翻译的起始

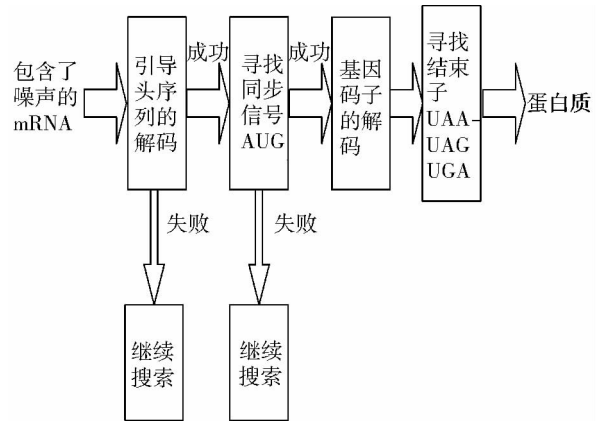

图3 原核翻译初始阶段的通信模型

3 确定的编译码算法

为了应用May提出的DNA序列编译码算法,先将碱基序列数字化。考虑到RNA序列中存在次黄嘌呤I,RNA中存在5种碱基I,A,C,G,U,量化过程在有限域5中进行。定义I=0,A=1,C=2,G=3,U=4。数字化DNA序列时,取$T=4$,其他几位碱基的数字表达同RNA[7-10]。

May提出,利用分组码模型,把基因译码模型化为一个(5,2)线性分组码[11],标准码字表的设计满足以下条件[12-13]:

1)以16S rRNA 3`端的最后13位碱基(其中包含了与SD序列互补的部分)的互补序列5'UAAGGAGGUGAUC…3'来做校验碱基。校验位碱基对应的数值如表1。

表1 16S rRNA互补序列中所取出的3位碱基校验位

校验位碱基	校验位碱基和	校验位碱基	校验位碱基和
UAA	1	GGU	3
AAG	4	GUG	3
AGG	0	UGA	2
GGA	0	GAU	2
GAG	0	AUC	3

2）标准码字满足在有限域 5 中和为 0 的校验准则，即从序列中取到的信息位碱基 U_{genetic} 和表 1 中的校验位碱基 U_{Parity} 之和为 0。$\sum_{1}^{K} U_{\text{genetic}} + \sum_{1}^{n-k} U_{\text{parity}} = 0$。满足上述 2 个条件的所有码字构成系统的标准码字表 S_c。

3）研究对象是起始位置前后各 30 位碱基，即 $[b_{-30}\quad b_{-29}\quad \cdots\quad b_{-1}\quad \text{ATG}\quad b_{+3}\quad \cdots\quad b_{+29}]$，共取 n 条，$n>100$。

对其中一条序列所取的第 p 个 5 位码字为

$\boldsymbol{r}_p = [b_p\quad b_{p+1}\quad b_{p+2}\quad b_{p+3}\quad b_{p+4}]$（$-30 \leqslant p \leqslant +29$）。

依照最小距离译码 $d_{\text{minp}} = \min[d(r_p;\ S_c)]$，比如第一个 5 位码字为

$\boldsymbol{r}_{-30} = [b_{-30}\quad b_{-29}\quad b_{-28}\quad b_{-27}\quad b_{-26}]$。

将在此位点上所取的每个码字和每个标准码字的不同位数的个数记为汉明距离，取其中最小的一个记为此位点的最小汉明值 $d_{\text{min}-30}$，下一个码字则从 b_{-29} 开始取，得到

$\boldsymbol{r}_{-29} = [b_{-29}\quad b_{-28}\quad b_{-27}\quad b_{-26}\quad b_{-25}]$，重复上述步骤得到 $d_{\text{min}-29}$，以此类推，取完 60 个碱基，得到此序列上的所有码字位点的最小汉明距离值。依照此方法，求出所取的 n 条序列的每个碱基位点的最小汉明距离值，最后把所有对应位点的汉明距离值取平均，获得最小汉明距离平均值 M，仿真结果如图 4 所示[12]。图中横轴代表碱基位置；纵轴代表平均最小汉明距离 M；0 点代表起始位点。幅度动态变化越小，表明 DNA 和 16S rRNA 之间的结合力越大，结合紧密，可以保证整个序列的翻译能正常进行。2 个波谷发生的位置正好分别是 SD 序列和翻译的起点位置。两者之间相距 12，这正好是 SD 序列和翻译起点之间距离的最佳值。

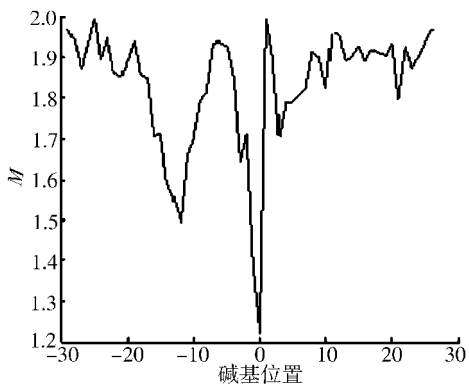

图 4　平均最小汉明距离译码的基因表达

4　仿真实验分析

May 把上述算法用来区别确定翻译序列、理论预测翻译序列和非翻译序列。标明原核翻译初始阶段的 SD 区域和 ATG 起始密码子。笔者则将这个算法拓展运用到了 16S rRNA 发生突变时对确定基因的翻译表达所产生不良影响的分析中，通过仿真印证了生物学实验所证实的相关理论，进一步说明用通信编译码理论来分析生物体遗传信息的传递过程是可行的[14]。具体仿真实验和结果如下。

仿真分析所用数据来自于 http://www.ncbi.nlm.nih.gov/下载的 DNA 序列的确认基因。基因包括理论预测基因和确定基因，所谓理论预测基因是没有在实验上完全确定的。在基因的注释中带有诸如“putative, possible, probable, probably, hypothetical, predicted, like, orf, uncharacterized, Similar to”之类说明的 ORF，又把理论预测基因分为标有基因名称的和标有 ORF 的基因；而确定基因就是在实验上得到完全确认的基因。使用这些确定基因的优势在于：它们和 16S rRNA 之间的结合关系是经实验证明其正确性的，同时它们的突变特性也是经过实验证明了的。运用确定基因，有利于体现实验特性。

确定了研究对象后，再运用 GENEMARK 分析软件确定每条确定翻译的基因序列的翻译起始位置，对每个确定的起始位置前后各取 30 个位点作为研究对象。

当 16S rRNA 中的保守区域 GGAGG 分别突变为 GUGUG（Ⅰ型突变）和 CCUCC（Ⅱ型突变）时，对基因序列号为 u00096 的大肠杆菌 K12 的确定翻译的 DNA 序列运用上述算法仿真分析，所得结果如图 5、6 所示。

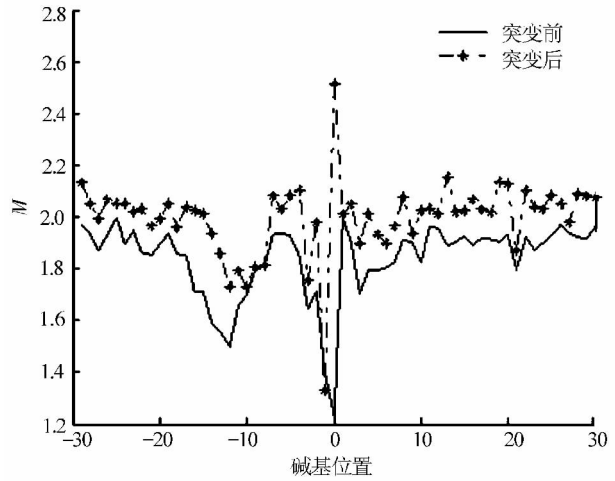

图 5　Ⅰ型突变后最小汉明距离的平均值变化

由图 5、6 可知：16S rRNA 发生突变后，平均最小汉明距离值曲线的最低点的位置发生了改变，由正常的 0 位变成了－1 位，次低点的位置也由正常的－12 位，变成了－9 位和－7 位。把图中的变化对应翻译成生物学的语言就是由于突变的产生，起

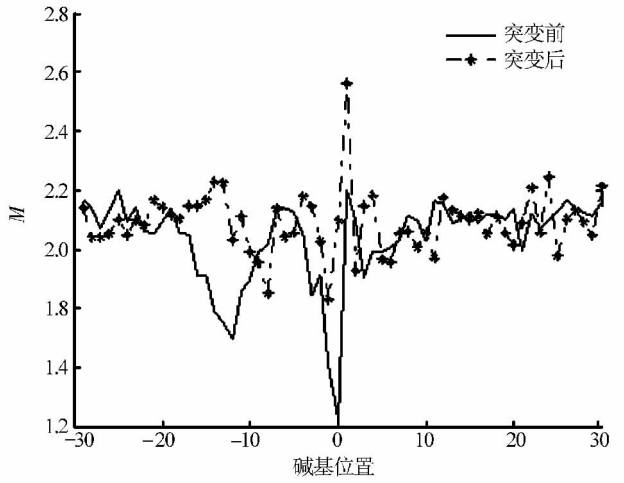

图6 Ⅱ型突变后最小汉明距离的平均值变化

始密码子和核糖体结合力最大的这一点的位置(最低点位置),由0变成了-1,意味着翻译的起始位点也发生变化,变成了从-1位开始翻译。其次受突变的影响,SD区域与核糖体结合的起始位置也发生了改变(次低点位置),所以SD区域与起始密码子间的精确距离变小,这对整个翻译过程的准确性而言,也产生了不良的影响。再次发生突变后在SD区域起始位置和翻译起始位置的最小汉明距离的平均值变大,对应的生物学意义是核糖体和DNA的结合力的减弱,使得翻译过程极易终止。综合上述因素的影响,突变最终导致的结果就是翻译阅读框错位,翻译出的蛋白质错误,翻译提前终止。笔者所得结论与美国生化专家Hui和Boer研究证明当16S rRNA的保守区域中的第4-8位由GGAGG突变为CCUCC或GUGUG时,会导致生物体的蛋白质合成翻译错误[15]相符,也与德国学者Zaher Dawy在文献[3]中基于16S rRNA的逐步移位码表,利用碱基的最小自由能所得到的结果相符,验证了16S rRNA对DNA分子表达效率的校验和识别作用。整个研究方法表明,用通信编译码理论来分析生物体遗传信息的传递过程是可行的。

5 结 语

实践证明用通信的编码理论来正确分析原核生物遗传信息的传递是可行的。随着人们对分子生物学知识的不断了解,相信通信的理论和知识在分析生物分子的信息传递过程中会发挥更大的作用。

参考文献:

[1] GUPTA M K. The quest for error correction in biology[J]. IEEE Engineering in Medicine and Biology Magazine, 2006, 25(1): 46-53.

[2] 李宁. 动物遗传学[M]. 2版. 北京:中国农业出版社,2003.

[3] 翟中和,王喜忠,丁明孝. 细胞生物学[M]. 北京:高等教育出版社, 2000.

[4] ZATHER D, FARUCK M G, JOACHIM H, et al. Modeling and analysis of gene expression mechanisms: a communication theory approach [C] // IEEE International Conference on Communications. Seoul: [s. n.],2005:815-819.

[5] OLGICA M, BANE V. Information theory and coding problems in genetics[M]. Mexico: The IEEE EMBS Cancun ,2004:60-65.

[6] ROSEN G L, MOORE J D. Investigation of coding structure in DNA [C] // Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing. Mexico: The IEEE EMBS Cancun , 2003 : 361-364.

[7] COSTA L D F, BARBOSA M S, MANOEL E T M, et al. Mathematical characterization of three-dimensional gene expression patterns [J]. Bioinformatics, 2004, 20(11):1653-1662.

[8] BOLSHAKOVA N, AZUAJE F. Cluster validation techniques for genome expression data [J]. Signal Processing,2003, 83: 825-833.

[9] MACDONAIL D A. Digital parity and the composition of the nucleotide alphabet [J]. Engineering in Medicine and Biology Magazine,2006, 25(1):54-61.

[10] 饶妮妮,邱丽君. DNA序列数值映射方法的研究[J]. 生物医学工程学杂志,2005,22(4):681-685.
RAO NI-NI, QIU LI-JUN . Study of numerical mapping methods for DNA sequences[J]. Journal of Biomedical Engineering,2005,22(4):681-685.

[11] MAY E E. Optimal generators for a systematic block code model of prokaryotic translation initiation [C]// Engineering in Medicine and Biology Society, 2003. Proceedings of the 25th Annual International Conference of the IEEE. Mexico: The IEEE EMBS Cancun,2003: 3858- 3860.

[12] MAY E E, VOUK M A, BITZER D L, et al. Coding theory based models for protein Translation initiation in prokaryotic organisms [C] // Pre-proceedings of the IPCAT 2003 (Fifth International Workshop on Information Processing in Cells and Tissues). Lausanned:EPFL,2003: 371-389.

[13] MAY E E. Towards a biological coding theory discipline[J]. New Thesis,2004,1(1) :19-38.

[14] PAUL D C. Large scale features in DNA genomic signals [J]. Signal Processing,2003, 83(4):871-888.

[15] HUI A, BOER H D. Specialized ribosome system: preferential translation of a single mRNA species by a subpopulation of mutated ribosome in Escherichia coli [J]. Proc Natl Acad Sci, 1987, 84(14):4762-4766.

(编辑 李胜春)