生物物理学报　第十八卷　第三期　　二○○二年九月
ACTA BIOPHYSICA SINICA　Vol.18　No.3　　Sept. 2002

酵母、大肠杆菌和枯草杆菌基因组中短 ORF 的分布与形成原因

李　宏
(内蒙古大学理工学院物理系, 内蒙古　呼和浩特　010021)

摘要：用终止密码方法计算了酵母、大肠杆菌和枯草杆菌基因组中所有的第一类开阅读框架（记为理论 ORF），给出了理论 ORF 和已知 ORF 随长度的分布，发现长度大于 150 个氨基酸后，理论 ORF 与已知 ORF 分布基本趋于一致，小于 150 个氨基酸的理论 ORF 数目的对数随长度线性变化，并提出这些短 ORF 是随机产生的猜想；研究了组分约束下的随机 DNA 序列中 ORF 数目、ORF 的长度与随机序列总长度和 GC 含量之间的关系，证明了本文猜想的正确性；给出了短的理论 ORF 中可能的编码序列所占比例的分布曲线，这对识别短的编码序列有参考价值。

关键词：酵母；大肠杆菌；枯草杆菌；组分约束下的随机序列；开阅读框架分布
中图分类号：Q615　　**文献标识码**：A　　**文章编号**：1000-6737(2002)03-0307-06

随着人类基因组计划(Human Genome Project, HGP)的进行[1]，基因组的研究也迅猛推进。现已完成了数十种原核生物细菌和数个真核生物基因组 DNA全序列的测定工作，生命科学也随之进入到了一个新纪元—后基因组时代。基因组信息学是弄清基因结构和功能等方面信息的重要手段[2]，在实验还不能迅速穷尽一个基因组的全部基因编码区时，理论上预测可能的基因编码区显得尤为重要，人们在研究编码序列的特征、编码序列与非编码序列的联系等方面已经做了大量工作[3-5]，在此基础上，运用了许多算法并开发了相应的软件来预测全基因组中的编码序列[6-10]，比如在酵母全基因组已发现的基因编码区中，大约 60%是通过理论分析得到的。

系统地研究全基因组中全部开阅读框架（ORF）的分布规律，对全面认识基因分布特征、基因的形成和进化具有重要的理论意义。为了研究基因的分布特征和基因的形成规律，本文将计算出酵母、大肠杆菌和枯草杆菌三种生物全基因组中所有可能的 ORF 结构，研究这些 ORF 的分布情况，并着重研究短 ORF 的分布规律，阐述较短 ORF 的起源问题。

1　计算第一类 ORF 结构的理论方法

在 DNA 序列中，编码序列分两种：一是不带有内含子的编码序列，二是含有内含子的编码序列。我们把不含内含子的编码序列所对应的 ORF 称作第一类开阅读框架，把含有内含子的编码序列所对应的 ORF 称作第二类开阅读框架。已知所有的原核生物基因编码区不含内含子，酵母基因组中绝大多数的基因编码区没有内含子。因此，所选的酵母、大肠杆菌和枯草杆菌是研究第一类 ORF 很好的模式生物。

我们在研究基因组 ORF 结构的过程中，提出了寻找第一类开阅读框架 (以下将第一类开阅读框架均简记为 ORF) 的一个简便方法—终止密码法[11]。我们知道在 DNA 序列中，已知有大量的终止密码和起始密码模式存在。而在正常的 ORF 序列读码框架中间不出现终止密码模式。终止密码方法就是根据 3 个终止密码子和 ATG 在 DNA 序列

收稿日期：2002-01-08
基金项目：国家自然科学基金（10147204）和内蒙自然科学基金
作者简介：李宏，1959 年生，博士，教授，电话:(0471)4992967, E-mail:stjt@imu.edu.cn.

中的分布来标识 ORF 结构的，对任意给定的 DNA 序列，分别以第 x、第 x+1 和第 x+2 位碱基开始读码（x 任意），步长为一个密码子，寻找三个终止密码子和 ATG 的位置，得到 3 个位置序列；在每一个位置序列中，假设第一类 ORF 结构的起始密码子为此序列段上游（同相位）紧邻最后一个终止密码子（记为 L-Ter）的第 1 个 ATG，终止密码子就是此序列段下游遇到的第 1 个终止密码子（记为 Ter），这一假设也称作第一 ATG 规则。因此我们用此方法系统地找出了酵母、大肠杆菌和枯草杆菌全基因组中所有的 ORF 结构，记做理论 ORF，把已知 ORF 和已知编码序列记做已知 ORF。

为了研究理论 ORF 与已知 ORF 的关系，须对第一 ATG 规则的正确性有一个了解。这方面的工作将在另外的文章中阐述，本文仅给出主要结论：第一 ATG 规则的正确率对酵母为 99.9%；对大肠杆菌以 ATG 起始的 ORF 为 92.9%；对枯草杆菌以 ATG 起始的 ORF 为 81%。

下面会看到，非第一 ATG 起始和非 ATG 起始的 ORF 不会影响我们的主要研究结果，因为我们研究的是理论 ORF 的数目和已知 ORF 的数目之间的关系，非第一 ATG 起始的已知 ORF 包含在相应的理论 ORF 之中，非 ATG 起始的已知 ORF 包含了绝大多数的理论 ORF，结构数目不发生变化，只是长度有差别。终止密码方法不能含盖的 ORF 结构数目很小，对大肠杆菌为 0.5%，枯草杆菌为 0.8%，在这些已知 ORF 中，L-Ter 下游只有 GTG 或 TTG 作为起始密码子，整个序列中没有 ATG；酵母为 5%，均为有内含子的 ORF。

上述分析表明，比较这三种生物理论 ORF 和已知 ORF 分布的理论基础是可靠的。

2 三种生物基因组中 ORF 的分布

从 GenBank 上调取了酵母、大肠杆菌和枯草杆菌的全基因组序列（ftp://ftp.cbi.pku.edu.cn/pub/ncbigenomes/），将双链 DNA 转化成 6 个由 ATG 和终止密码子组成的位置序列，用终止密码方法得到每个位置序列中包含各种长度的第一类 ORF。

经计算，酵母 16 条染色体上，共有各种长度的理论 ORF 281 405 个，其中长度大于 90 个氨基酸的 ORF 是 8 228 个；大肠杆菌基因组中，共有 106 615 个理论 ORF，枯草杆菌的理论 ORF 有 112 078 个。数据库中给出的已知 ORF 数目为：酵母 6 620 个，大肠杆菌 4 289 个，枯草杆菌 4 100 个。现分别将这三类生物基因组中理论 ORF 和已知 ORF 按其长度作分布图(图 1)。图 1 中，横坐标是 ORF 的长度，以氨基酸为单位，以 10 个氨基酸为一个区组，如长度在 0–9 的记做 10；10–19 的记做 20 等，纵坐标 y 是相应区组内 ORF 个数 N 的对数，y=lg$N$。

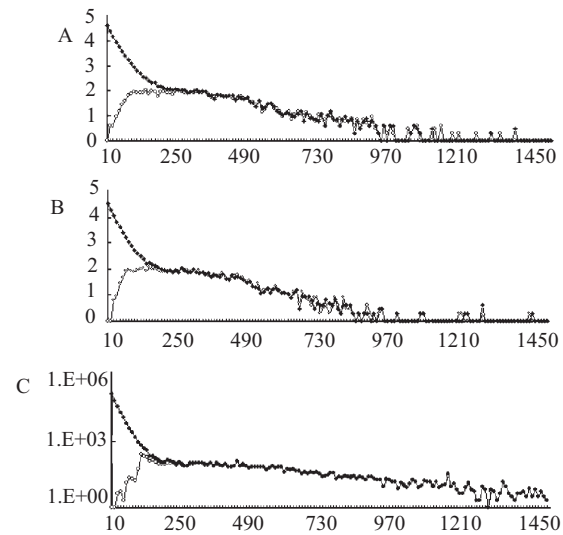

Fig.1 The known ORF and the theoretical ORF number verses its length. The vertical coordinate denotes the logarithm of ORF number and the abscissa—the ORF length. (A) *Escherichia coli*; (B) Bacillus subtilis; (C) yeast. —●— known ORF; —○— theoretical ORF

从图中看出，在长度大于 150 个氨基酸以后，酵母基因组中，预测的理论 ORF 与基因组中已知 ORF数目吻合的很好，大肠杆菌和枯草杆菌基因组中，长度在 150–650 之间吻合很好，长度大于 650 个氨基酸时，吻合度差些，主要原因是原核生物中有非 ATG 起始和非第一 ATG 起始的 ORF 存在造成的，本文在此不做具体说明。

与已知 ORF 对比可见，在长度小于 150 个氨基酸的理论 ORF 中，绝大多数不是编码区的侯选者。我们发现，理论 ORF 小于 150 个氨基酸特别是小于 110 个氨基酸的 ORF 数目的对数随长度的分布是一条直线。用长度小于等于 110 个氨基酸的 11 个区组做线性回归，结果见表 1。表中 a、b 和 R^2 分别是回归方程 $y=ax+b$ 的回归系数、回归截矩和决定系数。决定系数在 99%的置信水平下的临界值是 R_0^2=0.54，而三个拟合的 R^2 值远大于 R_0^2，说明二者之间线性关系极为显著。我们认为：基因

Table 1 The linear regression coefficient of the shorter ORF distribution (less than 110 amino acids) for the 3 organisms and the ORF distribution for component-constrained random sequences correspondingly

GC content	Length (million bp)	a	b	R^2	Maximum length of ORFs	Name
0.40	12.2	-0.0231	5.2589	0.9962	228	Yeast
		-0.0297	5.4827	0.9921	185	Random
0.52	4.6	-0.0181	4.7364	0.9918	262	*E.coli*
		-0.0206	4.8516	0.9935	236	Random
0.44	4.2	-0.0192	5.6235	0.9918	293	*B.subtilis*
		-0.0245	4.9148	0.9877	200	Random

组中众多短的 ORF 是随机构成的，这些 ORF 数目的对数与长度的线性关系是序列的随机性的体现，下面通过与随机序列的 ORF 分布对比，将证明我们的猜测。

我们用组分约束下的随机 DNA 序列进行模拟，研究随机形成的 ORF 的个数随其长度分布，随机形成的 ORF 的总数与 GC 含量以及 DNA 的总长度的关系，对于给定的 GC 含量和 DNA 总长度，能够随机产生的 ORF 的最大长度等问题。

3　组分约束下随机 DNA 序列的长度对 ORF 组成的影响

本节探讨组分约束下的随机 DNA 序列中形成的第一类 ORF 的数目和最大长度随序列总长度的变化。将碱基 A、C、G 和 T 含量都固定在 25%，用计算机产生 4 条不同种子的随机（伪随机）DNA 序列，其长度分别为 10 万、50 万、100 万和 400 万个碱基作为 DNA 的正链（W 链），再由这 4 条正链构造出 4 条 DNA 负链（C 链），形成随机的双链 DNA。选取的 4 个长度段作为采样点是考虑到酵母基因 16 条染色体单链的长度在 20 万到 230 万个碱基范围内变化，大肠杆菌和枯草杆菌基因组的单链长度约为 400 万个碱基，通过对这些随机 DNA 序列的研究，可对上述三种生物作参考对照。

对每一种长度的 DNA 序列，分别用终止密码方法[11]，找出 W 链和 C 链上全部的 ORF 后，与图 1 的作图方法一样，以 10 个氨基酸为一个区组，分别统计在各个区组内 ORF 的个数，然后以每个区组长度为横坐标，如（0–9）记为 10,（10–19）记为 20，……，以每个区组内 ORF 个数的对数为纵坐标作图，再将图上的各个点作线性拟合。计算结果见表 2。

表 2 中 R_0^2 是在相应的自由度下，99%置信水平的决定系数的临界值。需要说明的是，随着

Table 2 The relation between the number of theoretical ORFs for component-constrained random DNA sequences with different DNA length

GC content	Length (million bp)	a	b	R^2	R_0^2	Maximum length of ORFs	Total number of ORFs
0.5	0.1	-0.0175	2.9374	0.9767	0.33	168	2100
0.5	0.5	-0.0183	3.6438	0.9810	0.29	199	10774
0.5	1.0	-0.0190	3.9837	0.9886	0.28	210	21568
0.5	4.0	-0.0205	4.6939	0.9924	0.28	229	86206

ORF长度的增加，其数目趋于 0 时，N 的涨落很大，在做对数线性拟合时，没有计入这几个点。显然，拟合直线与横坐标的交点所对应的长度就是在此条件下，随机产生的 ORF 长度的理论最大值。

分析表 2 可以得到如下结论:（1）ORF 长度与 ORF 的数量的对数呈极显著的线性关系;（2）随着 DNA 序列长度的增加，ORF 的总数基本上按比例增加，在 GC 含量为 0.5 的条件下，序列长度每增加 1 万个碱基，ORF 的数目平均增加 214 个;（3）ORF 长度的最大值，随着序列长度的增加而增大，如序列长度是 10 万个碱基时，最大长度的 ORF 是 168 个氨基酸，当长度增加到 400 万个碱

基时，最大长度的 ORF 是 229 个氨基酸；（4）回归系数的绝对值也随着序列长度的增加略有增加，从 0.0175 增加到 0.0205，拟合曲线逐渐变陡，这说明随着序列长度的增加，较短的 ORF 相对增多。

4 组分约束随机序列的 GC 含量对 ORF 组成的影响

研究在组分约束下的随机 DNA 序列中，ORF 的组成随 GC 含量的变化。将序列的长度固定在 100 万个碱基。计算方法与上节类似，用计算机产生 4 条长度为 100 万个碱基的 DNA 正链，它们的 GC 含量分别为 0.3、0.4、0.6 和 0.7，再由这 4 条正链构造出相应的负链，形成组分随机的双链 DNA。需要说明的是：在我们统计酵母和大肠杆菌基因组中碱基的含量时，发现碱基 G 的含量与碱基 C 的含量非常接近，碱基 A 和碱基 T 的含量非常接近，因此本文在生成组分随机序列时，让 G 和 C 的含量相等，A 和 T 的含量相等。比如 GC 含量等于 0.3 是指碱基 G 和 C 的含量分别为 0.15，那么碱基 A 和 T 的含量分别就是 0.35 了。

与上节的做法一样，对每一种长度的 DNA 序列，分别用终止密码方法，找出 W 链和 C 链上全部的 ORF 后，以 10 个氨基酸为步长，分别统计在每一个步长范围内 ORF 的个数，然后以步长为横坐标，以每个步长内 ORF 个数的对数为纵坐标作图，再将图上的各点作线性拟合，结论见表 3。表 3 中，GC 含量为 0.5 的数据来自表 2。

Table 3 The relation between the numbers of theoretical ORFs for component-constrained random DNA sequences with different GC content

GC content	Length (million bp)	a	b	R^2	R_0^2	Maximum length of ORFs	Total number of ORFs
0.3	1.0	-0.0357	4.5624	0.9922	0.47	128	28317
0.4	1.0	-0.0263	4.2475	0.9848	0.35	162	26283
0.5	1.0	-0.0190	3.9837	0.9886	0.28	210	21568
0.6	1.0	-0.0140	3.7655	0.9944	0.23	269	16055
0.7	1.0	-0.0083	3.3097	0.9850	0.17	399	10191

仔细分析表 3 并与表 2 相比，可以得到以下结论：（1）随着 GC 含量的不同，ORF 数目的对数与其长度之间仍保持极显著的线性关系；（2）ORF 的总数随 GC 含量的上升而减少，GC 含量越大，ORF 减少的数目越多；（3）ORF 长度的最大值随着 GC 含量的增加显著变长，如从 GC 含量为 0.3 的 128 个氨基酸增加到 GC 含量为 0.7 的 399 个氨基酸；（4）回归系数的绝对值随着 GC 含量的增加显著变小；就是说随着 GC 含量的增加，相对较短的 ORF 数目明显减少，相对较长的 ORF 数目增加。

5 与三种生物基因组对比

将表 3、表 2 和表 1 进行对比，可见三种生物短 ORF 的分布与组分约束下的随机序列具有相同的规律，回归系数和回归截距的取值基本一致，都有显著的线性关系。这就证明了生物基因组中，短 ORF 的组成与随机序列中 ORF 的组成是相同的。

为进一步解释我们的猜测，严格按照大肠杆菌、枯草杆菌和酵母各个染色体的 DNA 长度、GC 含量，构造出与三种生物对照的组分约束下的随机序列，用终止密码方法求出所含的 ORF，与上面的过程一样，找出区组内 ORF 数目的对数与长度的线性拟合关系，与实际序列的情况对比，见表 1。实际序列短的 ORF 分布与组分约束下的随机序列的 ORF 分布规律相同，都有极显著的线性关系，再次证明了我们的猜测。

由图 1 可以看到，显然绝大多数短的理论 ORF 不是可能的编码序列。下面探讨短的理论 ORF 序列中，能成为编码序列的比例。我们知道，酵母和大肠杆菌是重要模式生物，基因组的研究相对最完善，已经识别的编码序列比例最大，理论预测的编码序列的可靠程度很高。用酵母和大肠杆菌已知的 ORF（包括编码序列）为基础，研究短 ORF 序列中可能的编码序列所占的比例是合理的。在每个长度段内，将相应区组内已知的 OEF 与短的理论 ORF 相比，其比值分布见图 2。

可见，三种生物的分布曲线随着 ORF 长度增加，在理论 ORF 中，已知 ORF 所占比例逐步增

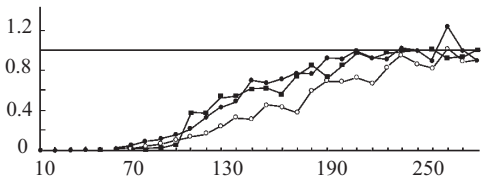

Fig.2 The ratio of the number of known and theoretical ORFs for the 3 organisms
—■— yeast —○— *E.coli* —●— *B.subtilis*

加，在长度大于 250 个氨基酸后，比值接近 1，即预测的 ORF 数目与实际 ORF 数目趋于一致。结合对图 1 的分析，在大于 250 氨基酸的区间内，预测的 ORF 数目基本都是编码序列。但初步计算后发现，情况并不全是这样，预测 ORF 数目要大于实际 ORF 的数目，多出的 ORF 中，情况比较复杂，是否是编码序列还需认真研究。

在长度小于 250 个氨基酸内，分布大体呈 S 型，大肠杆菌偏离 S 型大一些。比如长度小于 60 个氨基酸内的理论 ORF 中，几乎没有编码序列；长度在 100–110 的区组内，已知 ORF 与理论 ORF 的比值分别是酵母 5%、大肠杆菌 10%、枯草杆菌 15%；长度在 190–200 区组内，这一比值分别上升到：酵母 72%、大肠杆菌 69%、枯草杆菌 92%，当长度达到 220 个氨基酸后，酵母和枯草杆菌 95%以上的理论 ORF 是已知 ORF 了；长度达到 250 个氨基酸后，大肠杆菌 95%以上的理论 ORF 是已知 ORF 了。

在基因识别的各种方法中，确定生物基因组中短 ORF 是否是编码序列是非常困难的。期望我们给出的比值分布可能对短基因编码序列的识别有所帮助。这个比值分布只是一个初步估计，并不严格准确。因为已知 ORF 的数据不完善，酵母中带有内含子序列的影响，大肠杆菌和枯草杆菌中非 ATG 起始和非第一 ATG 起始的编码序列的影响等，都会影响估计的准确性。

这组分布曲线，可为我们理论识别短的编码序列提供一个数量上的限制，为理论上估计低等生物（无内含子的编码序列）基因组中编码序列的总数提供一个依据，也可以估计已知短的 ORF 中假阳性基因的比例。例如在研究中我们发现，酵母基因组中，长度在 110–130 个氨基酸范围内的已知 ORF 比例偏大，在图 2 和图 1 中也能看出这一点，如果将图 1 中已知 ORF 数目的对数分布换成数目随长度的分布，这一点就非常明显。就是说，在这一区间内已知的 ORF 中，假阳性的比例最大。由此可以估计酵母基因组中，基因总数的实际值应比已公布的数目要少。

6 讨　　论

组分约束下的随机序列的 ORF 都有一个最大长度限制。与三种生物对照，可以肯定，凡是大于这一长度的 ORF 都不是随机形成的，它们是基因进化的必然结果。至于它们是不是编码区，还要综合其它条件才能确定。

随机形成的 ORF 的最大长度随序列长度的增加而变长。从理论上讲是可以理解的，因为序列越长，随机产生较长的 ORF 的概率就越大；随机形成的 ORF 的数量与序列长度成正比。因为序列越长，ATG 和终止密码子的个数就越多，能构造出来的 ORF 就越多。

ORF 最大长度的增加明显依赖 GC 含量的增加，而 ORF 的总数目随着 GC 含量的增加明显减少。其原因是 GC 含量增加，AT 含量就减少，决定 ORF 长度的关键是起始密码子 ATG 和终止密码子 TAA、TGA 和 TAG 的数目。而密码子 ATG、TAG 和 TGA 的三个碱基中，有两个是 A 或 T，只有一个是 G，在密码子 TAA 中，三个全是 A 或 T。因此，当 GC 含量增加（AT 含量减少）时，ATG、TAA、TGA 和 TAG 这四个密码子的数目会明显减少，构成 ORF 的数目必然会减少；由于在一定的序列长度下，密码子的总数不变，决定 ORF 头部和尾部的密码子分布变稀，所组成的 ORF 长度必然要增大。

与研究的三种生物 GC 含量和序列长度相同的条件下构造随机序列，两者 ORF 的线性拟合方程并不完全相同，主要表现在回归系数上，真实基因组的回归系数绝对值比随机序列的小，而 ORF 最大长度比随机序列长，原因之一是生物基因组中 GC 含量的分布是非均匀的，基因组内有大量的规则片段（如编码序列，RNA 等），这都会造成线性回归参数与随机序列不完全一样。

参考文献:

[1] Cllins FS, Patrinos A, Jordan E, et al. New goals for the U. S. human genome project:1998-2003 [J]. *Science*, 1998,282 (5389):682-689.

[2] Hieter P, Boguski M. Functional genomics: its all how you read it[J]. *Science*, 1997,278(5338):601-602.

[3] 刘新文，童坦君. 真核生物中的基因流动现象[J]. 生理科学进展，1998,29:324-330.

[4] Kozak M. An analysis of 5'-nocoding sequences from 699 vertebrate messager RNAs[J]. *Nucleic Acids Res*, 1987,15: 8125-8147.

[5] 李宏，罗辽复. 大肠杆菌编码区碱基片段的分析研究[J]. 生物物理学报，2001,17:167-173.

[6] Zhang CT, Wang J. Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve[J]. *Nucleic Acids Res*, 2000,28:2804-2814.

[7] Salzberg SL, Delcher AL, Kasif S, et al. Microbial gene identification using interpolated Markov models[J]. *Nucleic Acids Res*, 1998,26:544-548.

[8] Dong S, Searls DB. Gene structure prediction by linguistic methods[J]. *Genomics*, 1994,23:540-551.

[9] Krogh A, Mian IS, Haussler D. A hidden Markov model that finds genes in *E.coli* DNA[J]. N*ucleic Acids Res*, 1994,22: 4768-4778.

[10] Snyder EE, Stormo GD. Identification of coding region in genomic DNA[J]. *J Mol Biol*, 1995,248:1-18.

[11] 李宏，罗辽复. 酵母全基因组中新的 ORF 结构的预测[J]. 内蒙古大学学报， 2001,32:498-503.

THE ORIGIN AND DISTRIBUTION OF SHORT ORFS IN YEAST, *E. coli* AND *B. subtilis* GENOMES

LI Hong

(Department of Physics, College of Sciences and Technology, Inner Mongolia University, Huhhot 010021, China)

Abstract: Using the terminal codon method proposed by us, the first kind of ORFs (denoted theoretical ORF) are predicted in yeast, *Escherichia coli* and *Bacillus subtilis* genomes. The theoretical ORF number and known ORF number verses its length are given. The two distributions are consistent with each other while ORF is length larger than 150 amino acids. There is a good linear relation between the logarithm of theoretical ORF numbers and its length for the theoretical ORF shorter than 150 amino acids. We suppose that the theoretical ORFs and their linear relation with the length of ORFs come from the randomness of the DNA sequences. The relation between ORF distribution and GC content, and between ORF distribution and length for component-constrained random sequences are analyzed. The results show that our supposition is correct. The ratio of the number of known ORF and short theoretical ORF are given. The relation may be useful for gene identification.

Key Words: Yeast; *Escherichia coli; Bacillus subtilis;* Component-constrained random DNA sequences; Open reading frame distribution

致谢：作者衷心感谢罗辽复先生的指点！