生物物理学报 第十七卷 第三期 二〇〇一年九月
ACTA BIOPHYSICA SINICA Vol.17 No.3 Sept.2001

Yeast 基因组编码区特征参数的研究

张 颖， 李 宏， 吕 军， 罗辽复
(内蒙古大学理论物理和理论生物物理研究室，内蒙古 呼和浩特 010021)

摘要：以碱基成分偏移量 D 值[1]为基本参数定义参数 d，以 d 为 *Yeast* 编码区的特征参数，对 *Yeast* 的第 1、2、3 类 ORF(open reading frame)进行了统计，得到 d 的特征参数区间。并且，以此区间为标准对 *Yeast* 的 6 类 ORF，以及 5′帽、3′尾、内含子、组分随机序列等非编码序列进行了检验。结果表明，用 d 作编码区的特征参数是可行的，它可以很好地区分编码序列和非编码序列。另外，又讨论了参数 d 与基因表达水平(用 *CAI* 值来衡量)的关系。发现，参数 d 与基因表达水平成很好的正相关关系；发现密码子的第 1 位点和第 2 位点的某些碱基分布与基因表达水平有关。

关键词：*Yeast* 基因组；编码区；碱基成分偏移；正相关
中图分类号：Q617 **文献标识码**：A **文章编号**：1000－6737(2001)03－0535－07

随着基因组研究的不断进展，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，如大肠杆菌(*Escherichia coli*)、酵母(*Saccharomyces cerevisiae*)和线虫(*Caenorhabditis elegans*)。到 2002 年对果蝇、2003 年对人类、2008 年对小鼠的全基因组序列的测序也将完成。但是，目前已经完成和正在实施的基因组计划给人们留下的只是一部“天书”。如何读懂这部即无段落又无标点的“天书”，已成为研究的重点。理论上对 DNA 全序列的分析主要是探讨寻找新基因的理论方法，而在这个方法确立之前，对已知基因即编码序列的特征进行分析就显得很有必要。

Yeast 已知的 ORF 中，分成 6 类。第 1 类 ORF 是指有确定蛋白质对应的核酸序列；第 2 类 ORF 是指与已知蛋白质强相似的核酸序列；第 3 类 ORF 是指与已知蛋白质弱相似的核酸序列；第 4 类 ORF 是指与未知蛋白质强相似的核酸序列；第 5 类 ORF 是指与未知蛋白质弱相似的核酸序列；第 6 类是指有疑问的 ORF。我们选取 *Yeast* 全基因组数据库中编号以“Y”打头的(如：YAL001c，YBR299w 等)第 1、2、3 类 3587 个 ORF 作为统计对象，得到 *Yeast* 编码区的特征参数，从而试图去预测 *Yeast* 中尚未发现的 ORF 是编码区的可能性，或预测 *Yeast* 中 6 类已知的 ORF 中非真实基因编码区的可能性，并将此方法进一步推广去预测其它生物基因组的编码区。

1 特征参数的选取

文献[2, 3]用偏好模分析法，以对独立序列的偏离 U_{Ki} 为参数，判定外显子中存在三重性读码框架，而内含子、5′帽、3′尾等非编码区不存在读码框架；文献[4, 5]又引入关联谱分析法，对关联函数 $C(\tau)$进行谱分析，发现编码区的谱 $P(k)$在 $k=N/3$ 处存在峰，表明碱基在密码子

收稿日期：2000－12－18
基金项目：国家自然科学基金资助项目(39660035)
作者简介：张颖，1973 年生，硕士研究生，电话：(0471)6519888，E－mail：lujun8210@263.net.

的三个位点上的不均匀分布；文献 [6, 7] 定义了非均匀指数 $HI(l)$，进而定义了 $F(l)=\frac{HI(l)}{3(l-1)}$，发现编码区 $F(l)$ 在 $l=3$ 时取极大值，表明编码区的三重读码框架。

由于不同的氨基酸及其不同的同义密码子使用频率各不相同，导致了在编码序列中密码子的三个位置上四种碱基出现的概率不均匀，而且三个位置上各有其特征碱基概率分布，这一现象称作碱基成分偏移[8~11]。可以利用编码序列的这一特征将其与非编码序列相区分。

本文以文献 [1] 中给出表示碱基成分偏移的 D 值为参数，进一步定义参数 $d=D-D_0$（D_0 的具体定义见第 2 节），以 d 作为编码区特征参数。其中：

$$D=\sum_j\sum_i|M_j-f_{ij}| \quad (j=A、C、G、T;\ i=1、2、3) \tag{1}$$

M_j 表示的是同一碱基 j 出现在三联体密码子的不同位置的平均概率，f_{ij} 表示的是碱基 j 出现在三联体密码子的第 i 位的概率。

显然，对于均匀序列 D 值接近于 0，对于非均匀序列 D 值一般在 0 到 1 之间取值。

2 编码区特征参数取值区间的统计

我们对 *Yeast* 的第 1、2、3 类共计 3587 个 ORF 进行了统计。其四种碱基在三联体密码子的不同位置上的平均概率见表 1。表中的 1－3 行数值表示的是碱基 j 出现在三联体密码子的第 i 位的概率即 f_{ij}，第 4 行表示的是同一碱基 j 出现在三联体密码子的不同位置的平均概率即 M_j。

Table 1 The average probability of codon at 1, 2 and 3 site in Yeast coding region

	P(A)	P(C)	P(G)	P(T)
1	0.331	0.157	0.291	0.221
2	0.350	0.226	0.143	0.281
3	0.290	0.193	0.186	0.331
Mean	0.324	0.192	0.207	0.278

根据表 1 中的 f_{ij} 和 M_j 计算得到的 D 值记为 D_0，D_0 就可以作为编码区的一个标准参数。对于 *Yeast* 的第 1、2、3 类的每一个 ORF 计算其相应的 D 值，计 D 与 D_0 的距离为 d，定义 $d=D-D_0$。统计相同 d 值下 ORF 出现的频率 $f(d)$，结果见图 1。统计中我们做了这样的约定：1) 将 d 值准确到 1/100；2) 将 $d>1$ 的 ORF 排除在统计之外（$d>1$ 的 ORF 非常少，只有 YDL130w、YDR534c、YKL096w－a、YMR173w、YOR053w 和 YOR383c 共 6 个序列，为了作图方便，我们作了这样的约定）。从图 1 看，统计数据不是标准的正态分布。

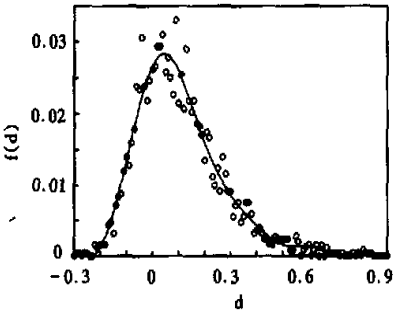

Fig.1 The curve of relation between parameter d and its relevant frequency $f(d)$

由平均值 $\bar{d}$ 和方差 σ 的计算公式：

$$\bar{d}=\sum_i f(d_i)\times d_i \tag{2}$$

$$\sigma=\sqrt{\sum_i f(d_i)\times(d_i-\bar{d})^2} \tag{3}$$

可得：$\bar{d}=0.105$，$\sigma=0.164$，取$(\bar{d}\pm\sigma)$作为特征参数 $\bar{d}$ 的取值区间，即特征参数的取值区间为$(-0.059,0.269)$。

3　对特征参数的检验

确定了特征参数的取值区间后，我们对 Class＝1、2、3、4、5、6 的所有 ORF 序列及第 1 类 ORF 前后的 5′帽、3′尾序列以及内含子、组分随机序列进行了检验，结果见表 2。3′尾、5′帽序列的长度为 600bp。组分随机序列是指所产生的随机序列与原序列所含的 A、C、G、T 组分百分比相同，即单碱基出现概率 P(n)(n＝A、C、G、T)相同。我们选取第 1 类以"Y"开头的 2513 条序列作为原序列，产生一系列与之一一对应的组分随机序列。表中的 c 为各类序列的总条数，cd 为 d 值落在特征参数的取值区间$(-0.059,0.269)$内的序列条数，$X=(100*\mathrm{cd}/\mathrm{c})\%$

Table 2　The verification of different type sequence parameter d

Sequence	c	cd	X(%)
Class＝1	2513	1823	72.5
Class＝2	340	248	72.9
Class＝3	734	551	75.1
Class＝4	652	424	65.0
Class＝5	1544	1021	66.1
Class＝6	399	208	52.1
intron	231	97	42.0
3′tail	2513	772	30.7
5′cap	2513	416	16.6
Component random sequence	2513	67	2.7

为落在特征参数的取值区间内的序列条数的百分比。

从表 2 中我们可以清楚地看出，对于内含子、3′尾、5′帽、组分随机序列等非编码序列，其 d 值落入特征参数区间的可能很小，尤其组分随机序列是一个均匀序列，其参数 d 值几乎全部不落入特征参数区间内。而相对于此，编码序列的参数 d 值则大部分处于我们的特征参数区间内，并且随着 ORF 编码蛋白质的可能性的降低，X 值也逐渐减小。所以参数 d 很好地区分了编码序列和非编码序列，可以作为编码区的特征参数。

4　特征参数与表达水平的关系

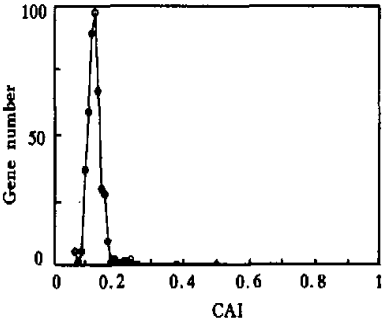

Fig.2　The curve of relation between CAI value in gene d ≤－0.059 and its relative number of gene.relevant CAI average value after sectionalizing by d value

Table 3 The average probability of codon at 1,2 and 3 site in Yeast coding region, d average value and its CAI average value

		P(A)	P(C)	P(G)	P(T)	d average value	CAIaverage value
The	1	0.337	0.172	0.247	0.244		
first	2	0.346	0.206	0.146	0.302	−0.149	0.13
section	3	0.312	0.183	0.189	0.316		
	Mean	0.332	0.187	0.194	0.287		
The	1	0.331	0.159	0.293	0.217		
second	2	0.354	0.222	0.142	0.281	0.002	0.18
section	3	0.291	0.189	0.187	0.332		
	Mean	0.325	0.190	0.208	0.277		
The	1	0.318	0.128	0.344	0.211		
third	2	0.329	0.283	0.142	0.245	0.282	0.39
section	3	0.247	0.238	0.161	0.354		
	Mean	0.298	0.217	0.216	0.270		

4.1 我们以特征参数的取值区间(−0.059,0.269)为依据,将酵母的 1、2、3 类的 3587 个 ORF 分为三段: $d \leqslant -0.059$ 的为第一段共 441 个 ORF, $-0.059 < d < 0.269$ 的为第二段共 2638 个 ORF, $d \geqslant 0.269$ 的为第三段共 508 个 ORF。然后,再分别求出这三段四种碱基在三联体密码子的不同位置上的平均概率、d 平均值以及相应的 CAI 平均值(见表 3)。另外,我们给出了按特征参数 d 分段后,基因的 CAI 值与相应的基因数关系曲线,见图 2、图 3 和图 4。

Table 4 The average probability of codon at 1,2 and 3 site in Yeast coding region, d average value and its relevant CAI average value after sectionalizing by CAI value

		P(A)	P(C)	P(G)	P(T)	d average value	CAI average value
The	1	0.340	0.168	0.270	0.221		
first	2	0.349	0.222	0.146	0.283	−0.064	0.12
section	3	0.302	0.183	0.196	0.319		
	Mean	0.330	0.191	0.204	0.274		
The	1	0.325	0.150	0.305	0.221		
second	2	0.354	0.226	0.139	0.280	0.047	0.20
section	3	0.284	0.195	0.180	0.340		
	Mean	0.321	0.191	0.208	0.280		
The	1	0.310	0.116	0.355	0.220		
third	2	0.327	0.253	0.148	0.272	0.292	0.64
section	3	0.230	0.263	0.159	0.348		
	Mean	0.289	0.210	0.221	0.280		

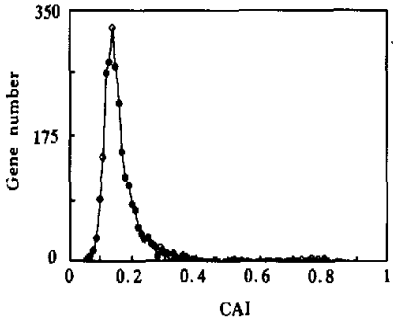

Fig.3 The curve of relation between CAI value in gene $-0.059<d<0.269$ and its relative number of gene.relevant CAI average value after sectionalizing by the d value

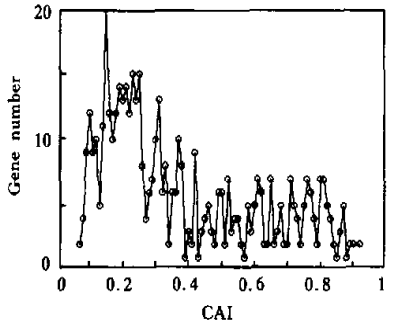

Fig.4 The curve of relation between CAI value in gene $d \geqslant 0.269$ and its relative number of gene.relevant CAI average value after sectionalizing by the d value

从表 3 我们可以看出，d 值与表达水平成正相关关系，即随着 d 值的增大表达水平也增大。而且我们还注意到，按特征参数的取值区间分段以后，个别位点的碱基概率对这个分段很敏感，呈显著变化。比如，第 3 位点的 P(A)、第 1 位点的 P(C)、第 2 位点的 P(T)都随着 d 值与表达水平的增大而减小，相对于平均值，减小幅度都在 10%左右；第 2 位点的 P(C)、第 1 位点的 P(G)、第 3 位点的 P(T) 都随着 d 值与表达水平的增大而增大，相对于平均值，增大幅度也都在 10%左右。普遍的观点认为，密码子的第 3 位点与表达有关，第 2 位点与进化有关。在我们这里发现，密码子的第 1 位点和第 2 位点也与表达有关，就如表 3 中指出的碱基 C 的第 1 位点概率和碱基 T 的第 2 位点概率随着表达水平的增大而减小，碱基 C 的第 2 位点概率和碱基 G 的第 1 位点概率随着表达水平的增大而增大。

4.2 为了进一步证实 d 值与表达水平成正相关关系，我们把上面的过程倒过来。即，首先按 CAI 值将酵母的 1、2、3 类的 3587 个 ORF 分为三段：$CAI<0.15$ 的为第一段共 1584 个 ORF，$0.15 \leqslant CAI<0.4$ 的为第二段共 1688 个 ORF，$CAI \geqslant 0.4$ 的为第三段共 315 个 ORF。然后，再分别统计这三段四种碱基在密码子的三个位点上的平均概率并且求出其相应的 d

Table 5 The average probability of codon at 1,2 and 3 site in Yeast coding region, d average value and its relevant CAI average value after sectionalizing by CAI value in gene $d \geqslant 0.269$

		P(A)	P(C)	P(G)	P(T)	d average value	CAIaverage value
The	1	0.325	0.135	0.328	0.212		
first	2	0.332	0.294	0.142	0.232	0.237	0.22
section	3	0.259	0.224	0.168	0.348		
	Mean	0.306	0.218	0.212	0.264		
The	1	0.306	0.108	0.370	0.215		
second	2	0.322	0.266	0.142	0.270	0.367	0.65
section	3	0.219	0.270	0.154	0.357		
	Mean	0.283	0.215	0.222	0.281		

值，结果见表 4。由表 4 可以清楚地看出，d 值与表达水平成正相关关系，而且，表 3 中那些碱基概率对按特征参数的取值区间分段敏感的位点，在表 4 中这些位点的碱基概率依然对按 CAI 值分段敏感。比如：第 3 位点的 P(A)、第 1 位点的 P(C)、第 2 位点的 P(T)都随着 d 值与表达水平的增大而减小，相对于平均值，减小幅度都在 10% 左右；第 2 位点的 P(C)、第 1 位点的 P(G)、第 3 位点的 P(T) 都随着 d 值与表达水平的增大而增大，相对于平均值，增大幅度也都在 10% 左右。

4.3 统计中我们还注意到，$d \geqslant 0.269$ 这个区段的 508 条基因所对应的 CAI 值比较分散 ($CAI_{min}=0.07$; $CAI_{max}=0.92$)，CAI 值与相应的基因数的对应关系见图 4。我们把这个区段的基因按 $CAI<0.4$(有 307 条)和 $CAI \geqslant 0.4$(有 201 条)又分为两个部分，分别统计了第一部分($CAI<0.4$)和第二部分($CAI \geqslant 0.4$)四种碱基在密码子的三个位点上的平均概率，并且求出其相应的 d 值，结果见表 5。

由表 5 也可以得到 4.1 和 4.2 同样的结论，比如：第 3 位点的 P(A)、第 1 位点的 P(C)、第 2 位点的 P(T)都随着 d 值与表达水平的增大而减小，只不过与 4.1 和 4.2 相比，这里，相对于平均值减小幅度更大，都在 10% 以上；第 2 位点的 P(C)、第 1 位点的 P(G)、第 3 位点的 P(T) 都随着 d 值与表达水平的增大而增大，相对于平均值，增大幅度也都在 10% 以上。表 5 中第一部分 CAI 值不比表 3 的第二段高多少，但碱基偏置比第二段要强很多，其原因还不清楚。

5 结果讨论

5.1 从以上的讨论我们认为参数 d 作为编码区的特征参数是比较成功的。用它可以很好地区分编码序列和非编码序列，而且用参数 d 作统计计算量不大，也可以很方便地推广。

5.2 文章又讨论了参数 d 与基因表达水平(用 CAI 值来衡量)的关系。发现，参数 d 与基因表达水平成很好的正相关关系，即随着 d 值的增大 CAI 值也增大。

5.3 文章还发现密码子的第 1 位点和第 2 位点也与表达水平有关。从 DNA 全序列中准确找出编码区是十分重要的工作，也是摆在研究人员面前的艰巨任务。2000 年 7 月，人类基因组图谱的草图已完成。最近，有报道[12]说人类基因数超过 14 万，超过近几年流行的估计数 10 万个左右近 40%，这也提示我们要从理论上寻找更好的方法来确定编码区。在基因组计划中，寻找新基因是非常有意义且有价值的工作，本文的结论给由理论所预测的新基因提供了一种理论检验依据和方法。

参考文献：

[1] Staden R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes[J]. *Nucleic Acids Res*, 1984, 12: 551 - 567.

[2] Luo LF, Ji FM. The preferential mode analysis of DNA sequence[J]. *J Theo Biol*, 1997, 188: 343 - 353.

[3] Ji FM, Luo LF. The ordered fragments in nucleotide sequence and molecular molecular evolution[J]. 内蒙古大学学报(自然科学版), 1997, 28: 493 - 504.

[4] Luo LF, Ji FM. The correlation spectrum of nuclotide sequences[J]. 内蒙古大学学报(自然科学版), 1995, 26: 419 - 426.

[5] Lee WJ, Luo LF. The periodicity of base correlation in nucleotide sequennces[J]. *Phys Rev E*, 1997, 56: 848 - 851.

[6] 李炜墨. Heterogeneity analysis of nucleotide sequences[J]. 内蒙古大学学报(自然科学版),1996,28:729-731.
[7] Lee WJ, Luo LF. Inhoomogeneity analysis on DNA sequence[A]. In:Luo L F. Collected Works on Theoretical Biophyysics[C]. Inner Mongolia Univercty Press, 1999.
[8] 李宏, 罗辽复. 大肠杆菌编码区 5'端碱基片段的统计分析[J]. 内蒙古大学学报, 1998,29(6):777-781.
[9] 李宏, 罗辽复. 大肠杆菌终止密码子前后序列碱基的统计分析[J]. 内蒙古大学学报, 1999,30(2):179-184.
[10] 李宏, 罗辽复. 大肠杆菌基因编码区碱基分布非均匀性的研究[J]. 内蒙古大学学报, 1999,30(5):579-583.
[11] 李宏, 罗辽复. 大肠杆菌基因编码序列碱基片段的统计分析[J]. 内蒙古大学学报, 2000,31(2):176-184.
[12] 贾弘提. 人类基因数超过 14 万[J]. 生命的化学, 2000,20(2):49.

RESEARCH IN THE CHARACTERISTIC PARAMETER OF GENE CODING REGION OF YEAST GENOME

ZHANG Ying,　LI Hong,　LU Jun,　LUO Liao-fu

(Laboratory of Theoretical Physics and Biology, NeiMongol University, Hohhot 010021, China)

Abstract: The parameter d is defined as basic parameter D, the bias count of base component. The characteristic parameter interval of d is obtained statistically by of the ORFs(open reading frame) of the 1st, 2nd, 3rd type and regarded as the characteristic parameter of Yeast gene coding regions. Using this interval as standard, 6 type ORFs, 5' cap, 3' tail, intron, component random sequence and other non-coding region sequence are verified. It is found d can be regarding as characteristic parameter of gene coding region, and d can clearly specify the coding region and non-coding region. In addition, the relationship between parameter d and gene expression level is discussed. It is found that parameter d is corrlated with gene expression level. And some base bias of the 1st and the 2nd site of a codon also are related to gene expression level as well.

Key Words: Yeast genome; Gene coding region; Bias count of base component; Direct proportion