生物物理学报 第十七卷 第四期 二〇〇一年十二月
ACTA BIOPHYSICA SINICA Vol.17 No.4 Dec.2001

大肠杆菌与酵母菌基因特定序列信息参量的研究

陈颖丽， 李前忠， 马克健
（内蒙古大学理论生物物理研究室，内蒙古 呼和浩特 010021）

摘要：提出核酸序列的矩阵表示形式，按位点定义了有生物学意义的信息参数 $M_1(l)$、$M_2(l)$ 和 $M_4(l)$。着重研究了不同表达水平的大肠杆菌（*Escherichia coli*，*E.coli*）的 SD 序列（Shine－Dalgarno region，SD）以及大肠杆菌（*E.coli*）和酵母菌（Yeast）基因起始、终止密码子邻近区域核酸序列的碱基关联性与保守性，并求出相应矩阵的本征值，给出了信息参量与基因表达水平的关系。发现信息参量体现了原核生物和真核生物翻译起始区域的显著差异，而且真核生物碱基起始区域的单碱基保守性程度及碱基关联性程度要比原核生物强。

关键词：SD 区域；矩阵表示；信息参数；基因表达水平；保守位点
中图分类号：Q617 **文献标识码**：A **文章编号**：1000 6737(2001)04 0676－09

随着基因数据库的不断扩增，对基因表达的调控机制等基因信息学研究显得越来越重要。而通过多序列的比较寻找保守位点是计算生物学的主要方法之一[1]。DNA 序列在转录起始位点、终止位点，翻译起始位点和终止位点邻近两侧都有一些稳定的核苷酸保守区[2,3]。这些保守区和保守位点对基因的表达调控具有较大的作用，如原核 mRNA 在起始密码子 ATG 前的 SD 区域影响翻译过程。虽然影响基因表达的因素很多，但是从基源过程来看，表达的调节最终必然反映在碱基顺序上。本文用一种新的核酸序列的矩阵表示方法、按位点定义信息参数，着重研究模式生物 *E.coli* 的 SD 序列的信息特征与基因表达水平的关系，并进一步比较了原核生物 *E.coli* 和真核生物 Yeast 基因在起始密码子和终止密码子前后的特定序列的碱基间的关联性和保守性特点及其与基因表达水平的关系，对结果的生物学意义作了讨论。

1 矩阵表示与信息参量

1.1 单碱基矩阵表示与 M_1 值

假定将某一种生物（如 *E.coli*）中 100 个基因的起始密码子 ATG 之后 n 个碱基取出，依据位点将它们按纵向排列。生物学上将起始密码子 ATG 的 A 碱基定义为 +1 位点，T 为 +2 位点，G 为 +3 位点，下游依次为 +4、+5 …位点，ATG 上游依次为 −1、−2 …位点。分别求出各个位点四种碱基的概率，记作 $P(A)$、$P(T)$、$P(C)$、$P(G)$。将这些位点和它们的碱基概率表示成一个矩阵，用 S_1 表示。

收稿日期：2001－04－29
基金项目 国家自然科学基金资助项目（39660023）
作者简介：陈颖丽，1974 年生，硕士，电话：(0471)4992958.
通讯作者：李前忠，E mail：qzli@mail.imu.edu.cn.

$$S_1=\begin{array}{c} \\ P(A) \\ P(T) \\ P(C) \\ P(G) \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & \cdots \end{array} \\ \begin{vmatrix} P_1(A) & P_2(A) & P_3(A) & \cdots \\ P_1(T) & P_2(T) & P_3(T) & \cdots \\ P_1(C) & P_2(C) & P_3(C) & \cdots \\ P_1(G) & P_2(G) & P_3(G) & \cdots \end{vmatrix} \end{array} \tag{1}$$

其中 1、2、3 …表示位点，矩阵每一列对应该位点四种碱基概率。可知 S_1 是一 4 行 n 列的矩阵，从矩阵 S_1 中即可了解各位点单碱基分布情况。

定义一个量：

$$M_1(l)=\sum_i \frac{(P_l(i)-1/4)^2}{1/4} \qquad i=A,T,C,G \qquad l=1,2,\cdots,n \tag{2}$$

$P_l(i)$ 是第 l 位点碱基 i 出现的概率，当某一位点碱基随机选取时，则每一种碱基概率为 1/4，$P_l(i)-1/4$ 是此位点某一碱基概率与随机选取这一碱基概率的差值。(2)式是 $M_1(l)$ 随位点 l 的变化关系，表示 DNA 序列在该位点与完全随机序列在该位点的偏离。

如果该位点是完全保守的，设此位点上碱基都为 A，则 $P(A)=1$，$P(T)=P(C)=P(G)=0$，则该位点 $M_1(l)$ 值为 3；如果该位点碱基分布是完全随机的，即 $P(A)=P(T)=P(C)=P(G)=1/4$，则 $M_1(l)$ 值为 0。可见 $M_1(l)$ 取值在 0～3 之间，$M_1(l)$ 值越高表明该位点的保守性越强，因此，可用 $M_1(l)$ 值表征位点的保守性。

1.2 相邻双碱基矩阵表示与 M_2 值

将取出的 n 个位点的碱基，按 1－2 位点，2－3 位点，3－4 位点等，顺次统计相邻双碱基概率，每两位点间有 16 种碱基排列，即 16 个概率值，用矩阵 S_2 表示。

$$S_2=\begin{array}{c} \\ P(AA) \\ P(AT) \\ P(AC) \\ P(AG) \\ \vdots \end{array}\begin{array}{c} \begin{array}{ccccc} 1 & 2 & 3\cdots & 4\cdots \end{array} \\ \begin{vmatrix} P_{12}(AA) & P_{23}(AA) & \cdots \\ P_{12}(AT) & P_{23}(AT) & \cdots \\ P_{12}(AC) & P_{23}(AC) & \cdots \\ P_{12}(AG) & P_{23}(AG) & \cdots \\ \vdots & \vdots & \vdots \end{vmatrix} \end{array} \tag{3}$$

S_2 为一 16 行 $n-1$ 列的矩阵。

定义量：

$$M_2(l)=\sum_{ij}\frac{(P_{l,l+1}(ij)-P_l(i)P_{l+1}(j))^2}{P_l(i)P_{l+1}(j)} \tag{4}$$

$$i,j=A,T,C,G \qquad l=1,2,\cdots,n-1$$

$P_{l,l+1}(ij)$ 为在相邻位点出现一对碱基 ij 的联合概率[1]，

$$P_{l,l+1}(ij)=P_l(i)P_{l+1}(j/i) \tag{5}$$

$P_{l,l+1}(ij)$ 是在 l 位点取碱基 i 其后 $l+1$ 位点出现碱基 j 的条件概率。对于独立序列，相邻碱基的出现概率是独立事件，既 $P_{l,l+1}(ij)=P_l(i)P_{l+1}(j)$[1]。容易得出，$M_2(l)$ 值是描述序列中相邻位点双碱基关联相对于独立序列的偏离，表征了序列中双碱基的关联性。

1.3 相邻三碱基矩阵表示与 M_3 值

同样，依次计算出基因中 123 位点间，234 位点间，345 位点间等，相邻三碱基出现概率，每

三位点间有 64 种碱基排列，用矩阵 S_3 表示。

$$S_3=\begin{array}{c} \\ P(AAA) \\ P(AAT) \\ P(AAC) \\ P(AAG) \\ \vdots \end{array}\begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \left|\begin{array}{cccc} P_{123}(AAA) & P_{234}(AAA) & \cdots \\ P_{123}(AAT) & P_{234}(AAT) & \cdots \\ P_{123}(AAC) & P_{234}(AAC) & \cdots \\ P_{123}(AAG) & P_{234}(AAG) & \cdots \\ \vdots & \vdots & \vdots \end{array}\right. \end{array} \tag{6}$$

S_3 为一 64 行 $n-2$ 列的矩阵。

定义量：

$$M_3(l)=\sum_{ijk}\frac{(P_{l,l+1,l+2}(ijk)-P_l(i)P_{l+1}(j)P_{l+2}(k))^2}{P_l(i)P_{l+1}(j)P_{l+2}(k)} \tag{7}$$

$$i,j,k=A,T,C,G \qquad l=1,2,\cdots,n-1$$

同理 $P_{l,l+1,l+2}(ijk)$ 为在相邻位点出现三个碱基 ijk 的联合概率，

$$P_{l,l+1,l+2}(ijk)=P_l(i)P_{l+1}(j/i)P_{l+2}(k/ij) \tag{8}$$

$P_{l+2}(k/ij)$ 是在 l 位点取碱基 i，$l+1$ 位点取碱基 j 其后 $l+2$ 位点出现碱基的条件概率。对于独立序列，相邻碱基的出现概率是独立事件，既 $P_{l,l+1,l+2}(ijk)=P_l(i)P_{l+1}(j)P_{l+2}(k)$。$M_3(l)$ 值用来描述序列中相邻位点三碱基关联相对于独立序列的偏离，亦表征了三碱基的关联性。

2 应 用

2.1 应用于大肠杆菌 SD 序列的研究

SD 序列（Shine－Dalgarno region）是特定的核苷保守区，由原核生物起始密码子上游的一些碱基构成，富含嘌呤[4]，距起始密码子的距离可在 5 到 13 个碱基范围内变化[5]，它与 16SrRNA 的 3′－OH 尾的九个碱基互补，是非常重要的核糖体结合位点。我们将核酸序列的矩阵表示与信息参量 M 具体应用于大肠杆菌（*E.coli*）的 SD 序列，观察定义的 $M(l)$ 值与不同基因表达水平的 SD 序列的关系。

由大肠杆菌基因库 Mapshearch 2.0 给出的 *E.coli* 基因序列，求出 1373 个编码序列的修正的密码子适应性指数（Codon Adaption Index，CAI）[3]值。CAI 值标志基因的表达水平，范围在 0～1 之间，其值越接近 1，表明基因表达水平越高。对 E.coli 基因，CAI 值与基因表达水平的实验值保持有很好的正比关系[3]。选出以 ATG 或 GTG 为起始密码子的 CAI≥0.65 的高表达基因 233 个，CAI≤0.40 的低表达基因 189 个，0.40＜CAI＜0.65 的中表达基因 350 个。分别将高、中、低表达基因起始密码子 ATG 前 25 个位点的碱基取出，将它们按纵向排列，求出各位点四种碱基的概率。将概率值代入(1)式中，可以给出 25 个位点的矩阵 S_1（具体形式略），根据(2)式计算出 $M_1(l)$ 值。

图 1(a)是这三个 $M_1(l)$ 值的比较，横轴表示位点，纵轴为 $M_1(l)$ 值。（图 1 中用实线表示高表达基因，用长虚线表示中表达基因，用短虚线表示低表达基因）。对应高、中、低表达基因，在－10 位点附近约 5 个位点范围 $M_1(l)$ 值都很大，高表达基因在－10 位点达到峰值，中表达基因在－9 位点达到峰值，低表达基因峰值在－11 位点，高、中、低表达基因的峰值依次降低。低表达基因此 5 个位点范围内 $M_1(l)$ 值较其它位点的 $M_1(l)$ 值大的程度不如高、中表达基因

明显。从图 1(a)中可以清楚地看到 - 10 位点附近约 5 个位点范围既为 SD 序列。SD 序列单碱基保守性强弱与基因表达水平的高低成正相关关系，其回归方程为：$\hat{y} = 0.8224x - 0.0797$（其中 x 表示 CAI 值，$\hat{y}$ 表示 $M_1(l)$值，相关系数 $r = 0.990$）。

对于高、中、低表达基因，由(3)式给出 25 个位点的矩阵 S_2(具体形式略)，(4)式计算出 $M_2(l)$值。图 1(b)为高、中、低表达基因位点间 $M_2(l)$值的比较。对于高表达基因 $M_2(l)$值在 - 8、- 9 位点间达到峰值，中表达基因在 - 9、- 10 位点间达到峰值，而低表达基因在此区间已没有峰值，但在 - 1 位点 $M_2(l)$值却非常高。从图 1(b)中看出 SD 区域相邻双碱基关联性很强，且双碱基关联性强弱与基因表达水平成正相关关系，其回归方程为：$\hat{y} = 0.0032x - 0.0008$（其中 x 表示 CAI 值，$\hat{y}$ 表示 $M_2(l)$值，相关系数 $r = 0.979$）。从图 1(a)、(b)的比较可知：对于单碱基保守性强的区域双碱基关联性也强。

同样计算出高、中、低表达基因中相邻位点三碱基出现概率，由(6)式给出矩阵 S_3 的形式(具体形式略)，由(7)式计算出 $M_3(l)$值。图 1(c)为高、中、低表达基因位点间 $M_3(l)$值的比较。对于高、中表达基因在 - 9 位点左右的四个 $M_3(l)$值非常大，其余位点的值都接近于 0，而低表达基因所有位点的 $M_3(l)$值都接近 0，只有 - 1 位点的值相对最大，它是否蕴涵特定的生物学含义，有待于进一步研究。从图 1(c)中看出 SD 区域相邻三碱基关联性仍很强，三碱基关联性强弱与基因表达水平亦成正相关关系，其回归方程为：$\hat{y} = 0.0033x - 0.0009$（其中 x 表示 CAI 值，$\hat{y}$ 表示 $M_3(l)$值，相关系数 $r = 0.998$）。

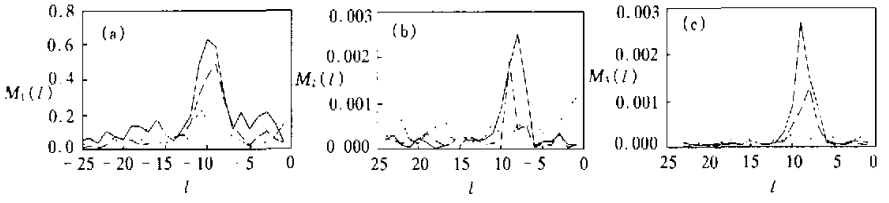

Fig.1 The curves of $M(l)$ as function of sites l in *E.coli* genes (solid line, long dash line and dot line are respectively high, middle and low expressed genes)

2.2 应用于 *E.coli* 和 Yeast 基因特定序列碱基保守性、关联性的对比研究

我们以原核生物 *E.coli* 和真核生物 Yeast 这两种模式生物为例，研究两种不同生物基因特定序列的一些特性。生物学上认为，多肽链生物合成的起始、延伸和终止三个阶段中，控制翻译的主要环节在肽链合成的起始阶段，起始密码子 ATG 附近的结构特征与起始作用的调控有密切的关系[6]。因而本文着重研究两种生物起始密码子上、下游各 30 个碱基和终止密码子上、下游各 30 个碱基的分布情况，且与基因的表达水平联系起来。选出 1306 个大肠杆菌基因，这些基因的起始密码子上、下游和终止密码子上、下游至少各有 30 个碱基，其中 CAI≥0.65 的高表达基因 227 个，0.40<CAI<0.65 的中表达基因 900 个，CAI≤0.40 的低表达基因 179 个。在酵母基因组 DNA 全序列数据中[7]，依据 CAI 值，选出高表达基因 93 个，中表达

基因 131 个，低表达基因 5832 个，这些基因起始与终止密码子上、下游也至少各有 30 个碱基，CAI 值选取范围与 *E. coli* 基因相同。

2.2.1 起始区域

选 *E. coli* 基因的起始密码子 ATG(也使用 GTG、TTG)左右和 Yeast 基因的起始密码子 ATG 左右各 30 个碱基，此段序列都各含 63 个碱基，计算 $M_1(l)$、$M_2(l)$、$M_3(l)$值，为了便于比较原核生物和真核生物起始密码子邻近区域碱基的组成情况及碱基的保守性、关联性，将 *E. coli* 和 Yeast 的高、中、低表达基因的各个 $M(l)$值放在一起研究。

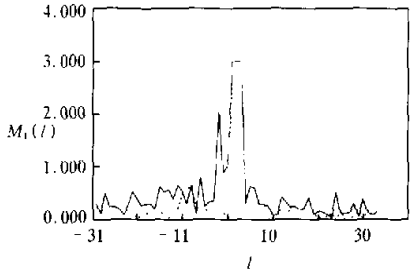

Fig.2 The curves of $M_1(l)$ ~ l of the sequences near the initiation codon of high expressed genes of Yeast(solid line) and *E. coli*(dot line)

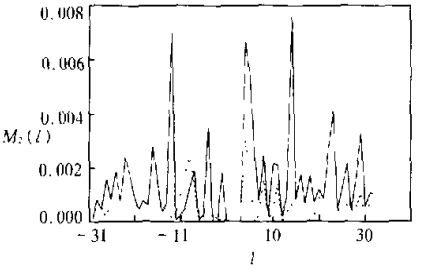

Fig.3 The curves of $M_2(l)$ ~ l of the sequences near the initiation codon of high expressed genes of Yeast(solid line) and *E. coli*(dot line)

图 2 是 *E. coli* 和 Yeast 的高表达基因的起始密码子邻近区域 $M_1(l)$值比较。横坐标是碱基位点，+1，+2，+3 位点为起始密码子的位置，纵坐标是 $M_1(l)$的值。(图 2 - 图 5 中实线表示 Yeast 基因，虚线表示 *E. coli* 基因)。整体来看，Yeast 基因的 $M_1(l)$值普遍比 *E. coli* 基因的值高，由此可知真核生物基因起始区域的单碱基保守性程度比原核生物强。对于 *E. coli* 基因 ATG 上游有一 $M_1(l)$ 的高值区域，它与 SD 区域相对应，是原核生物起动机制所特有的，它对翻译起始具有重要影响。Yeast 基因中 ATG 上游 -3 位点有一 $M_1(l)$的极值，说明此位点保守性极强。生物学上已证实，对于真核生物，ATG 的旁侧序列与翻译起始效率有密切关系，-3 位点为 A 是普遍规律，对于 ATG 被起始识别有最显著的促进作用。我们定义的 $M_1(l)$值，很明显的体现了原核生物与真核生物翻译起始区域的显著差异。根据 Kozak[8]等系统分析 mRNA 有效翻译的序列特点，一般认为，起始密码子附近的序列（包括起始密码子）为 GCAGCCAUGG 时，对翻译起始最为有效，尤其是 -3 位和 +4 位的核苷酸影响最大，即 -3 位为嘌呤核苷酸，+4 位为 G 时，翻译效率最佳。

图 3 是 *E. coli* 和 Yeast 高表达基因的起始密码子邻近区域 $M_2(l)$值的比较。Yeast 基因中 ATG 上、下游都有几个 $M_2(l)$值较高的峰，而 *E. coli* 基因中 ATG 上游只在体现 SD 区域处有一 $M_2(l)$高值区，下游 +4 位点有一峰值，其余位点 $M_2(l)$值明显比 Yeast 基因的值低很多。从图中可知，Yeast 基因中关联片段多，关联性强，碱基构成较 *E. coli* 基因复杂，这也体现了真核生物较原核生物高等的一个方面。

图 4 为 *E. coli* 和 Yeast 的高表达基因 $M_3(l)$值的比较。*E. coli* 基因中 ATG 上游 SD 区域 $M_3(l)$值明显，体现了这一特定区域碱基之间的关联性，ATG 下游 +3，+4，+5 三位点间

的 $M_3(l)$值相当高，图 3 中此位点处 $M_2(l)$值也有类似情况。一些研究指出，*E.coli* 基因转译起始是由起始密码子上游 SD 区域与 ATG 下游紧邻的特殊构成的十几个碱基一起来完成[5]，文献[9]也指出，*E.coli* 基因 ATG 下游第 2、第 3 个密码子这段区域除了与 SD 区域一起承担着转译起始功能外，还是基因表达调控区域的一部分，本文得出的 $M_2(l)$、$M_3(l)$值体现了这一结论。对于 Yeast 基因，$M_3(l)$值普遍比 *E.coli* 基因的值高，ATG 上游有两段区域 $M_3(l)$值很高，ATG 下游 +3、+4、+5 三位点间的 $M_2(l)$值非常高，此处碱基构成值得关注。生物学实验已证实，对于真核生物，+4 位点的 G 对于 ATG 被起始识别亦有显著的促进作用，若 −3 位点不是 A，则 +4 位点的 G 对有效的翻译起始作用是必需的[6]。图 4 中 +3、+4、+5 三位点得出的 $M_3(l)$值极高，就蕴含着 +4 位点较特殊这一现象，图 3 中该处的 $M_2(l)$值也体现了此位点的特殊性，由此说明我们定义的体现碱基关联性的 $M_2(l)$、$M_3(l)$ 的值对辨别生物基因中特殊位点有一定的作用。

我们也计算比较了两种生物，中、低表达基因 $M_1(l)$、$M_2(l)$、$M_3(l)$值的情形，有类似高表达基因的结论，这里不再赘述。

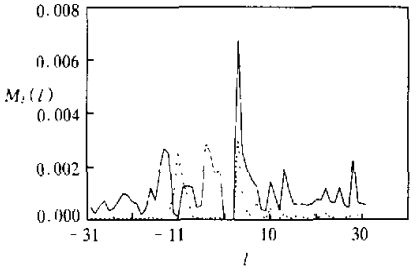

Fig.4 The curves of $M_3(l) \sim l$ of the sequences near the initiation codon of high expressed genes of Yeast(solid line) and *E.coli*(dot line)

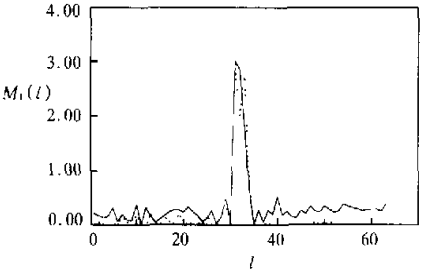

Fig.5 The curves of $M_1(l) \sim l$ of the sequences near the termination codon of high expressed genes of Yeast(solid line) and *E.coli*(dot line)

2.2.2 终止区域

选两种生物基因的终止密码子 TAA(TAG、TGA)左右各 30 个碱基，研究这一特定区域，含 63 个碱基，计算 *E.coli* 和 Yeast 高、中、低表达基因的 $M_1(l)$、$M_2(l)$、$M_3(l)$的值，作为代表，我们只给出了 *E.coli* 和 Yeast 高表达基因终止密码子区域 $M_1(l)$ 值的对比，如图 5 所示。图中横坐标为碱基位点，31、32、33 位点为终止密码子所处位置，在其左右各有 30 个碱基位点，纵坐标是 $M_1(l)$的值。结果表明，Yeast 基因的 $M_1(l)$值普遍比 *E.coli* 基因的值高，这也说明真核生物较原核生物碱基构成复杂。在终止密码子旁侧序列，对于 *E.coli* 和 Yeast 两种生物的各个 $M_1(l)$值分析来看没有发现较特殊的规律，是否蕴含着与终止作用有关的调控区域，还有待于进一步研究。

2.3 利用矩阵及其本征值之和作为衡量基因表达水平与基因碱基结构的一个可能指标

2.3.1 SD 序列矩阵 P 及其本征值

由图 2 中已经得知 *E.coli* 基因 ATG 上游 −7～−13 位点 $M_1(l)$ 值非常高，即为 SD 序

列，本文取这七个位点表示成矩阵 S_1：

$$S_1=\begin{array}{c} \\ P(A)\\ P(T)\\ P(C)\\ P(G)\end{array}\begin{array}{c}\begin{array}{cccc} -13 & -12 & \cdots & -7\end{array}\\ \begin{bmatrix} P_{-13}(A) & P_{-12}(A) & \cdots & P_{-7}(A)\\ P_{-13}(T) & P_{-12}(T) & \cdots & P_{-7}(T)\\ P_{-13}(C) & P_{-12}(C) & \cdots & P_{-7}(C)\\ P_{-13}(G) & P_{-12}(G) & \cdots & P_{-7}(G)\end{bmatrix}\end{array} \tag{9}$$

此处 S_1 为一 4 行 7 列的矩阵，对于大肠杆菌基因，选出 CAI≥0.65 的高表达基因 227 个，0.40<CAI<0.65 的中表达基因 900 个，CAI≤0.40 的低表达基因 179 个，根据基因表达水平将高、中、低表达基因表示成矩阵 $S_{1高}$、$S_{1中}$、$S_{1低}$。

再将这七个位点表示成矩阵 S_2：

$$S_2=\begin{array}{c} \\ P(AA)\\ P(AT)\\ P(AC)\\ P(AG)\\ \vdots\end{array}\begin{array}{c}\begin{array}{ccccc} & -13 & & -12 & \cdots & -7\end{array}\\ \begin{bmatrix} & P_{-13-12}(AA) & P_{-12\ -11}(AA) & \cdots\\ & P_{-13-12}(AT) & P_{\ 12\ -11}(AT) & \cdots\\ & P_{\ 13-12}(AC) & P_{-12-11}(AC) & \cdots\\ & P_{\ 13-12}(AG) & P_{\ 12\ -11}(AG) & \cdots\\ & \vdots & \vdots & \vdots\end{bmatrix}\end{array} \tag{10}$$

此处 S_2 为一 16 行 6 列的矩阵，同理高、中、低表达基因表示成矩阵 $S_{2高}$、$S_{2中}$、$S_{2低}$。利用 $P=S\times S^*$[10]和(9)式得 $P_1=S_1\times S_1^*$，对应高、中、低表达基因 P_1 写为 $P_{1高}$、$P_{1中}$、$P_{1低}$（具体形式略）；同样利用(10)式可得 $P_2=S_2\times S_2^*$，对应高、中、低表达基因 P_2 写为 $P_{2高}$、$P_{2中}$、$P_{2低}$（具体形式略）。

求出各个 P_1 和 P_2 的本征值，P_1 有 4 个本征值，P_2 有 16 个本征值，再将每个矩阵的本征值加起来，用符号 Z_1 和 Z_2 表示。表 1 给出 SD 序列高、中、低表达基因 Z_1 和 Z_2 值的对比情况。

Table 1 The values of Z_1 and Z_2 in SD region for high, middle and low expressed genes

expression level	Z_1	Z_2
high	2.361213	0.810532
middle	2.231365	0.676033
low	1.994676	0.539340

从表中值的对比可知，Z_1 和 Z_2 值随基因表达水平的降低而依次降低，SD 序列高表达基因的 P 矩阵本征值之和最大。文献[10]中的 P_1 为一 4 行 4 列的对角矩阵，对角元素即为四种碱基出现的相对概率，称为单碱基概率矩阵；P_2 为一 16 行 16 列的对角矩阵，称二阶紧邻概率矩阵，对角元素为紧邻碱基相对概率。本文的 P_1、P_2 都为实对称矩阵，也体现了碱基构成及碱

基间的关联，求得的本征值之和越大说明基因的表达水平越高，它提供了一个比较基因表达水平量的可能指标。

2.3.2 *E.coli* 和 Yeast 基因的矩阵 P 及其本征值之和

取图 2 中 ATG 前后各 30 个碱基，表示成矩阵 S_{1f} 和 S_{1b}，下标 f 代表 ATG 之前的位置，下标 b 代表 ATG 之后的位置，*E.coli* 基因的表示为 S_{E1f} 和 S_{E1b}，Yeast 基因的表示为 S_{Y1f} 和 S_{Y1b}，下标 E 代表 *E.coli* 基因，Y 代表 Yeast 基因；取图 3 中 ATG 前后碱基表示成矩阵 S_{2f} 和 S_{2b}，*E.coli* 基因的表示为 S_{E2f} 和 S_{E2b}，Yeast 基因的表示为 S_{Y2f} 和 S_{Y2b}，可知 S_1 为 4×30 维的矩阵，S_2 为 16×29 维的矩阵。利用 $P=S\times S^*$，表示出各个 P 矩阵：

$$\begin{aligned} P_{E1f}&=S_{E1f}\times S^*_{E1f} \qquad P_{E1b}=S_{E1b}\times S^*_{E1b}\\ P_{Y1f}&=S_{Y1f}\times S^*_{Y1f} \qquad P_{Y1b}=S_{Y1b}\times S^*_{Y1b}\\ P_{E2f}&=S_{E2f}\times S^*_{E2f} \qquad P_{E2b}=S_{E2b}\times S^*_{E2b}\\ P_{Y2f}&=S_{Y2f}\times S^*_{Y2f} \qquad P_{Y2b}=S_{Y2b}\times S^*_{Y2b} \end{aligned} \tag{11}$$

然后再求出所有 P 矩阵的本征值，表 2 给出 *E.coli* 和 Yeast 基因中 P 矩阵本征值之和的对比关系（用符号 Z_{1f}、Z_{1b}、Z_{2f}、Z_{2b} 分别表示矩阵 P_{1f}、P_{1b}、P_{2f}、P_{2b} 的本征值之和）

Table 2 The sum of the eigenvalues in matrix P for *E.coli* and Yeast genes

	Z_{1f}	Z_{1b}	Z_{2f}	Z_{2b}
E.coli	8.63634	8.45022	2.61202	2.53520
Yeast	11.08438	9.37163	4.20721	3.24284

从表中值的对比可知，Yeast 基因所有 P 矩阵的本征值之和都比 *E.coli* 基因矩阵本征值大。得出结论，求得的本征值之和越大说明单碱基的保守性与双碱基的关联性越强，这一矩阵的特性值，提供了一个比较原核基因与真核基因碱基结构的一个可能指标。

3 结　论

文中矩阵表示 S_1、S_2、S_3 很明显地体现了核酸序列中碱基之间的关系；按位点定义的信息参数 $M_1(l)$、$M_2(l)$ 和 $M_3(l)$ 有明显的生物学意义，是研究核酸序列的保守位点，序列碱基关联的新方法，应用于 *E.coli* 基因中的 SD 序列得出结论：$M_1(l)$ 值都与基因表达水平成正相关关系，定义的信息参量体现了原核生物和真核生物翻译起始区域的显著差异，利用矩阵的本征值，发现了衡量基因表达水平及衡量原核生物和真核生物基因间碱基结构的一个可能指标。矩阵表示是一个严谨而又直观的数学理论形式，矩阵理论本身既是一个十分庞大的逻辑体系，将其应用于对核酸序列的研究，将可能为 DNA 序列的研究开辟一个新的途径。

参考文献：

[1] 罗辽复　生命进化的物理观[M]　上海：上海科学技术出版社，2000.265－268.

[2] Heijne G. Sequence analysis in molecular biology[M]. New York: Acad press Inc, 1987.

[3] Li H, Luo L F. The relation between codon usage, base correlation and gene expression level in

Escherichia coli and yeast[J]. *J Theor Boil*, 1996,181:111-124.

[4] Rudd KE, Miller W, Werner C, et al. Mapping sequence E.coli genes by computer:Software strategies and examples[J]. *Nucleic Acids Res*, 1991,19:637-647.

[5] Gold L. Posttranscriptional regulatory mechanisms in *Escherichia coli*[J]. *Annu Rev Biochem*, 1988,57:199 233.

[6] 沈珏旺,方得福. 真核基因表达调控[M]修订版. 北京:高等教育出版社,施普林格出版社,1997.63-64.

[7] THE YEAST GENOME THE Munich Information Center for Protein Sequences (MIPS).

[8] Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs[J]. *Nucleic Acids Research*, 1987,15:8125-8132.

[9] 李宏,罗辽复. 大肠杆菌编码区 5'端碱基的统计分析[J]. 内蒙古大学学报(自然科学版), 1998,29(6) 777 781.

[10] 李前忠. 核酸序列的矩阵表示[J]. 内蒙古大学学报(自然科学版), 1999,30(1):41-44.

A STUDY OF THE INFORMATIONAL PARAMETERS FOR THE SPECIFIC SEQUENCES OF *E.coli* GENES AND YEAST GENES

CHEN Ying-li, LI Qian-zhong, MA Ke-jian

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

Abstract: The matrix representation of nucleic acid sequences and the definition by $M_1(l)$, $M_2(l)$ and $M_3(l)$) are presented. The definition is biological of significance. The E.coli in different expression level were investigated for SD region. The conservative and correlative properties of the bases of the specific sequences between *E.coli* and Yeast genes were studied comparatively, and the eigenvalues of the corresponding matrix were calculated. The relation between informational parameters and gene expression level was given. The results show that the informational parameters indicate remarkable difference in the translated starting regions, and the conservativity of a single base and the correlation between base pair in the starting regions of eukaryote are stronger than those of prokaryote.

Key Words: Shine-Dalgarno region; Matrix representation; Informational parameter; Gene expression level; Conserved site