生物物理学报 第十八卷 第一期 二〇〇二年三月
ACTA BIOPHYSICA SINICA Vol.18 No.1 Mar.2002

大肠杆菌、酵母和果蝇基因保守位点的信息熵分析

吕 军[1,2], 李 宏[1], 马克健[1]
(1. 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;
2. 内蒙古工业大学物理教研室, 内蒙古 呼和浩特 010062)

摘要 对大量的大肠杆菌(*Escherichia coli*)、酵母(Yeast)和果蝇(*Drosophila melanogaster*)已知基因起始密码子和终止密码子上、下游各 30 个碱基序列,用重新定义单碱基信息冗余(记为 $D_1(l)$,l 是位点)和紧邻碱基的信息冗余(记为 $D_2(l)$),统计计算每个位点的 $D_1(l)$和 $D_2(l)$值。从结果看,双碱基比单碱基携带更多的信息。酵母和果蝇基因起始密码子上游 -3 位点 $D_1(-3)$ 和 $D_2(-3)$ 有一明显峰值;大肠杆菌基因起始密码子上游 SD 区域 $D_1(l)$和 $D_2(l)$有明显峰值,与他人结论相同。发现酵母基因起始密码子下游的 $+4$ 位点与 $+5$ 位点的紧邻碱基的 $D_2(l)$有一峰值,其关联模式为 TC(联合概率为 0.211)。这说明用重新定义的信息冗余去确认 DNA 序列中存在的保守位点是完全可行的。

关键词:信息熵;关联;保守位点
中图分类号:Q617 **文献标识码**:A **文章编号**:1000-6737(2002)01-0071-05

DNA 序列中保守位点的确认及分析,对于认识基因表达调控机制有极其重要的作用。这些位点称为顺式元件;识别或结合的蛋白质(酶)称为反式作用因子,他们的相互作用是调控网络必不可少的组成单元。DNA 序列在转录的起始位点、终止位点、翻译的起始位点、终止位点邻近及内含子两侧都有一些特定的核苷保守区,已经被大家所知[1]。文献[2]中给出了不同物种起始密码子和终止密码子邻近及内含子两侧的核苷保守区。本文用重新定义的各位点的信息熵去分析大肠杆菌、酵母和果蝇已知基因起始密码子和终止密码子上、下游各 30 个碱基的序列,这些核苷保守区可以明显地显示出来,从而验证定义的有效性。

文中数据选自 Genbank 数据库。所用的大肠杆菌、酵母和果蝇基因均取自它们的全基因组数据库。我们从大肠杆菌基因组数据库给出的 4289 个 ORF 中,选取了 2933 个已知基因的编码序列、从酵母全基因组数据库给出的 6620 个 ORF 中,选出了第一类 2513 个已知基因序列、选取了果蝇 2L 染色体的 2310 个 ORF(包括已知基因编码区和理论预测的 ORF)序列为统计样本。

1 定义某一位点上的信息熵和关联熵以及信息冗余

首先以文献[3]为基本出发点,定义位点 l 的单碱基信息熵和位点 l 与 $l+1$ 的紧邻碱基关联的信息熵。

位点 l 的单碱基信息熵为

$$H_0(l) = -\sum_{\alpha} P_l(\alpha)\mathrm{Log}_2 P_l(\alpha) \quad (\alpha = \mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}) \tag{1}$$

位点 l 与 $l+1$ 的紧邻碱基关联的信息熵为

$$H_1(l) = -\sum_{\alpha}\sum_{\beta} P_l(\alpha\beta)\mathrm{Log}_2 P_l(\alpha\beta) \quad (\alpha,\beta = \mathrm{A},\mathrm{C},\mathrm{G},\mathrm{T}) \tag{2}$$

所谓某一位点的信息熵和关联熵,指的是,以 N 个基因为统计样本,纵向统计出基因编码区(或非编码区)l 位点上出现碱基 α 的概率,l 位点上出现碱基 α 和 $l+1$ 位点上出现碱基 β 的联合概率,然后依据公式(1)和(2)计算得到,公式中的对数均以 2 为底。在以上公式中,$P_l(\alpha)$为位点 l 上出现碱基 α 的概率,$P_l(\alpha\beta)$为位点 l 上出现碱基 α 后位点 $l+1$ 上出现碱基 β 的联合概率。显然,$P_l(\alpha)=1/4$ 或 $P_l(\alpha\beta)=1/16$ 时,$H_0(l)$ 或 $H_1(l)$ 将达到最大值 2 或 4,它对应于随机序列。当 l 位点碱基保守使用时,熵 $H_0(l)$ 和 $H_1(l)$ 将对应一个与 2 或 4 相差相对较大

收稿日期:2001-05-25
基金项目:国家自然科学基金(10147204)资助项目,内蒙古自治区自然科学基金(20001301)资助项目
作者简介:吕军,1973 年生,硕士研究生,讲师,电话:(0471)6552521,E-mail:lujun8210@263.net.

的值，极限情形，比如某一位点只使用某一碱基而不使用其他碱基，则熵 $H_0(l)$ 为零。

由式(1)和(2)我们进一步定义两个信息冗余量 $D_1(l)$ 和 $D_2(l)$

$$D_1(l) = H_{0\max} - H_0(l) = 2 + \sum_{\alpha} P_l(\alpha)\mathrm{Log}_2 P_l(\alpha) \quad (3)$$

$$D_2(l) = H_{1\max} - H_1(l) = 4 + \sum_{\alpha}\sum_{\beta} P_l(\alpha\beta)\mathrm{Log}_2 P_l(\alpha\beta) \quad (4)$$

其中 $D_1(l)$ 为 l 位点的一阶信息冗余，$D_2(l)$ 为 l 位点的二阶信息冗余(我们这里仍沿用文献[1]中的符号及表述，这里形式上仍把 $D_2(l)$ 称为二阶信息冗余，但与文献[3]中的 D_2 内容上有所不同，要注意区别)。下面，我们以 $D_1(l)$ 和 $D_2(l)$ 为参数去分析大肠杆菌、酵母和果蝇基因序列中起始密码子以及终止密码子邻近序列的保守位点。

统计之前，对起始密码子上下游和终止密码子上下游的碱基位点标记作以下约定：起始密码子 ATG 的碱基 A 约定为 +1 位点，T 为 +2 位点，G 为 +3 位点，ATG 下游依次为 +4，+5，…位点，ATG 上游依次为 −1，−2，…位点；终止密码子 TAA(TAG、TGA)的碱基 T 约定为 +1 位点，上游依次为 −1，−2，…位点，下游依次为 +2，+3，+4，+5，…位点，具体如图 1 所示。另外，不失一般性，我们令：

$D_1(+1) = D_1(+2) = D_1(+3) = 0$；

$D_2(-1) = D_2(+1) = D_2(+2) = D_2(+3) = 0$。

图 2 − 图 4 中 0 位点为无意义位点。

2 大肠杆菌起始密码子以及终止密码子邻近序列的保守位点

对于大肠杆菌基因，已知起始密码子上游 −10 位点前后有富嘌呤区[4]称为 SD 序列(Shine − Dalgarno region)，它与 16S rRNA 的 3′−OH 尾的 9 个碱基有很强的互补性，是非常重要的核糖体结合位点。SD 区域到起始密码子的距离可在 5 到 13 个碱基范围内变化[5]，文献[6]研究了它与表达水平的关系。我们这里用信息熵再次讨论这一保守区。

大肠杆菌基因起始密码子和终止密码子上下游序列 $D_1(l)$ 和 $D_2(l)$ 的统计结果见图 2a 和 b。

2.1 大肠杆菌起始密码子邻近序列的保守位点

由图 2(a)可知，在大肠杆菌起始密码子上游 −9 位点 $D_1(-9)$ 和 $D_2(-9)$ 都有一峰值，这个峰不那么尖锐，看上去有点钝，因为它是由 5 个点组成的，这些位点的强偏置表示一个保守区域，这个保守区域就是 SD 区域，−9 位点是这个保守区域的中心。这一结果与[7]中结果一致。大肠杆菌的 SD 区域不仅是单碱基的强保守区，同时也是多碱基的强关联区。

在起始密码子下游 +4 位点 $D_1(+4)$ 和 $D_2(+4)$ 同时出现极大值，$D_1(+4)$ 的极值与该位点为 A 优势的说法相一致[6]；$D_2(+4)$ 的极值说明 ATG 之后第 1 个密码子的 1、2 位出现较强关联。这个关联可能与 SD 序列一起构成转录的起始信号。

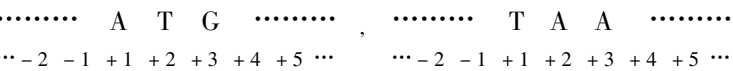

Fig.1 The site marks of the upstream and down-stream bases of the initiation codon ATG and termination codon

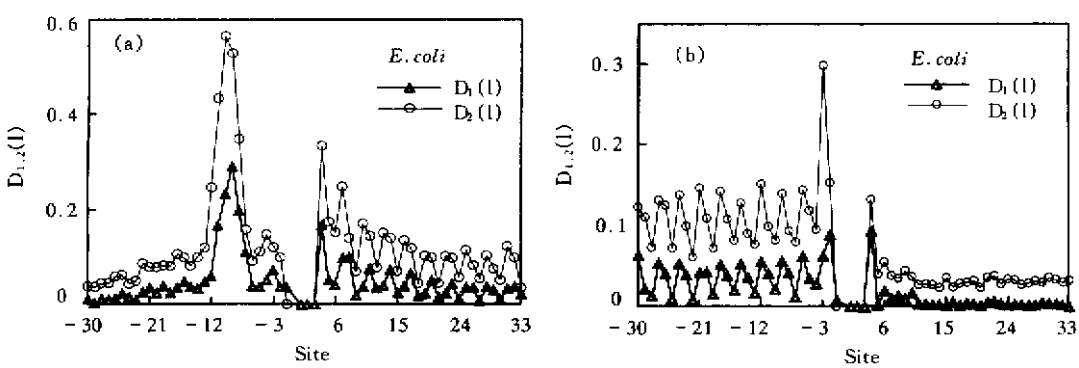

Fig.2 The curves of $D_1(l)$ and $D_2(l)$ as a function of site l nearby the coding start region (a) and the coding terminal region (b) of *E.coli* genes

整体上看密码子第 1 位点处的 $D_2(+4, +7, +10, \cdots)$均有一个极值。这说明了编码区密码子内 1、2 位点的关联强于 2、3 位点和 3、1 位点的关联，这也正是编码区的 3 周期[8,9]特性的体现。

2.2　大肠杆菌终止密码子邻近序列的保守位点

图 2(b)是大肠杆菌终止密码子邻近区域 $D_1(l)$和 $D_2(l)$随位点的变化曲线。与图 2(a)相比，大肠杆菌终止密码子邻近区没有明显的单碱基保守位点出现，但在 -3 位点 $D_2(-3)$的峰很高，这表明大肠杆菌终止密码子上游 -3 位与 -2 位碱基存在一较强关联，这在以前未见报道。-3 位与 -2 位碱基的强关联可能与基因终止有关。

终止密码子后 $D_1(+4)$和 $D_2(+4)$同时出现了极值，但不明显。这与文献[2]中指出的大肠杆菌终止密码子后第 1 位存在 T 优势的说法一致。

另外，终止密码子上游 $D_2(l)$ 的分布呈明显的周期振荡，$D_2(-3, -6, -9, \cdots)$处均有极值，说明密码子 1、2 位的关联 > 2、3 位的关联 > 3、1 位的关联，这也正是编码区的 3 周期特性的体现[8,9]。

3　酵母基因起始密码子和终止密码子邻近序列的保守位点

对酵母这种真核模式生物 DNA 序列的研究具有重要的意义。尤其对起始密码子和终止密码子邻近序列特征的研究，已经有不少有意义的结果[9,10]。文献[7]用矩阵法对这一区域的保守位点进行了研究，我们这里再次用第 1 节定义的信息熵和关联熵对这一区域的保守位点进行分析。酵母基因起始密码子和终止密码子上下游序列 $D_1(l)$ 和 $D_2(l)$的统计结果见图 3 的(a)和(b)。

3.1　酵母起始密码子邻近序列的保守位点

从图 3(a)来看，ATG 上游 -3 位点处 $D_1(-3)$有一峰，非常明显。说明这一位点是一个保守位点，这与文献[4]所得结论一致，而由我们以前的工作[8]可知，这一位点应是碱基 A 的保守位点，其概率超过 65%，这也与[2]中结论一致。-4，-3 位点的 $D_1(-3)$，$D_2(-3)$ 的峰也很高，表明 -3 与 -4 和 -2 这 3 个位点之间的碱基存在强关联，其原因是 -3 位点的高度保守。

在 $+4$ 位点没有保守单碱基出现，这与文献[11]以及我们以前的工作[8]所得结论一致，而与文献[2]对高等生物 PRI、ROD、MAM、及 VRT 类以及 Kozak 对脊椎生物的统计[10]结论不同，即 ATG 后第 1 位为 G 优势的所谓 Kozak 规则对酵母这样的低等真核生物不明显。我们认为这一位点的 G 优势可能与进化有关。值得注意的是 $D_2(+4)$值很大，即在 $+4$ 处出现一个较高的峰，说明 $+4$ 和 $+5$ 位点的碱基间出现强关联。这在以前的报道中未见过，仔细研究表明，$+4$ 和 $+5$ 位点出现的是嘧啶保守关联，其关联模式为碱基 TC(联合概率为 0.211)。

另外，凡是密码子第 1 位 $D_2(+4, +7, +10, \cdots)$处均有一个极值，表明密码子第 1 位与第 2 位的关联要强于 2、3 位和 3、1 位的关联，仍体现了编码区的周期特性[8,9]，而起始密码子前即非编码区则无这一特征。

3.2　酵母终止密码子邻近序列的保守位点

酵母终止密码子邻近区域单碱基信息冗余和紧邻碱基的关联信息冗余随位点的变化曲线见图 3(b)。由图可见，酵母终止密码子邻近区没有明显的保守位点出现。文献[2]中指出真核终止密码子后第 1 位为嘌呤优势，与我们以前工作[8]结论相同，不过，从图 2(b)可知这种优势并不明显。与起始密码子后面的编码区相同，终止密码子前在密码子的第 2 位与第 1 位的关联处 $D_2(-3, -6, -9, \cdots)$ 均出现极值，也体现了编码区的周期特性，而终止密码子后即非编码区则无这一特征。

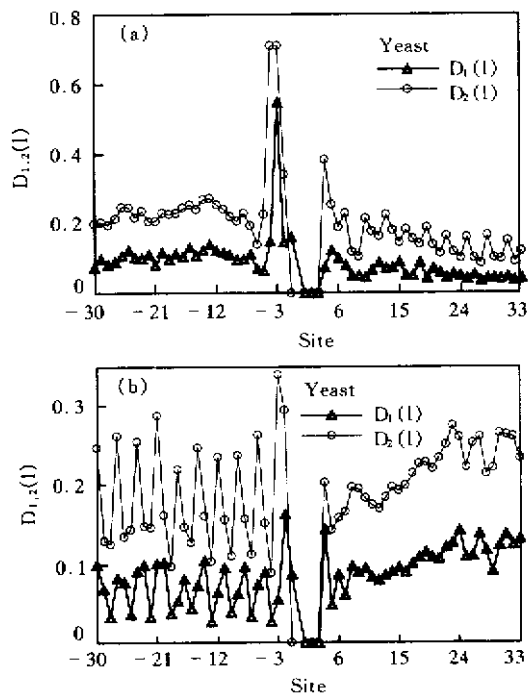

Fig.3　The curves of $D_1(l)$ and $D_2(l)$ as a functions of site l nearby the coding start region(a) and the coding terminal regian (b) of yeast genes

4 果蝇基因起始密码子和终止密码子邻近序列的保守位点

图 4(a),(b) 为果蝇基因的起始密码子和终止密码子前后邻近序列的 $D_1(l)$ 和 $D_2(l)$ 随着 l 的变化曲线,可以看出,果蝇与酵母相比其保守位点的位置是一致的。只是果蝇编码区的 3 周期特性与大肠杆菌和酵母相比不明显,其原因在结果和讨论中给出分析。

另外,值得注意的是,酵母和果蝇基因终止密码子后的 $D_1(l)$ 和 $D_2(l)$ 随着 l 的增大而呈上升趋势,这种上升趋势一直延续到 40 个碱基左右后保持稳定。这一点与大肠杆菌终止密码子后面的序列有明显差别,其原因尚不明,似乎是原核生物与真核生物的一种差别,当然,这一结论还有待进一步证实。

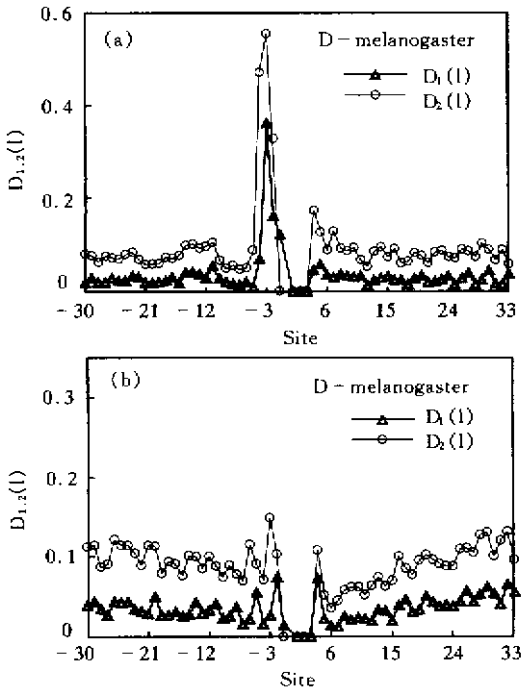

Fig.4 The curves of $D_1(l)$ and $D_2(l)$ as a function of site l nearby the coding start region (a) and coding terminal region (b) of *Drosophila* genes

5 涨落限的估算

依据文献[3] 估计 $D_1(l)$ 和 $D_2(l)$ 的涨落限为(N 为序列数)

$$D_1(f.b.)=\frac{11.3}{2N\ln 2}=\frac{8.15}{N},\text{(99\%置信度)} \quad (5)$$

$$D_2(f.b.)=\frac{21.7}{2N\ln 2}=\frac{15.65}{N}。\text{(99\%置信度)} \quad (6)$$

由上两式可得:

大肠杆菌(N = 2933) $D_1(l)$ 和 $D_2(l)$ 的涨落限为

$D_1(E.coli, f.b.) = 0.0028$,

$D_2(E.coli, f.b.) = 0.0053$;

酵母(N = 2513) 和的涨落限为

$D_1(Yeast, f.b.) = 0.0032$,

$D_2(Yeast, f.b.) = 0.0062$;

果蝇 (N = 2310) $D_1(l)$ 和 $D_2(l)$ 的涨落限为

$D_1(D-melanogaster, f.b.) = 0.0035$,

$D_2(D-melanogaster, f.b.) = 0.0068$。

所有的涨落限非常接近于 0, 由此可见,我们所讨论的峰值是有明显的统计意义的。

6 结果和讨论

6.1 综上所述,用本文定义的某位点的单碱基信息冗余和紧邻碱基的关联信息冗余为参数,对酵母、大肠杆菌和果蝇基因的起始密码子和终止密码子上、下游各 30 个位点进行统计,再次用信息论的观点验证了酵母和果蝇起始密码子上游 - 3 位点为保守位点,大肠杆菌起始密码子上游的 SD 区域为保守区域。还给出了酵母起始密码子下游的 + 4 位点与 + 5 位点的紧邻碱基的模式为 TC;验证了编码区内密码子的 1、2 位关联强于 2、3 位 3、1 位的关联。也验证了文献 [2] 中的部分结论。用本文的参数去讨论保守位点比用文献[7, 10]的参数效果要明显,保守位点的峰值更突出,噪声更小,反映出的信息更多。这说明,用某一位点的信息熵和关联熵来寻找基因序列的功能片段是有效的。

6.2 我们发现,大肠杆菌终止密码子上游 - 3 位点 $D_2(-3)$ 的峰很高,也就是说 - 3 位与 - 2 位碱基存在一较强关联;紧邻酵母基因起始密码子下游 + 4 位点没有保守单碱基出现,即 $D_1(+4)$ 值不大,与文献[2]对高等生物 PRI、ROD、MAM、及 VRT 类以及 Kozak 对脊椎生物的统计[11]结论不同,Kozak 指出 ATG 后第 1 位为 G 优势的所谓 Kozak 规则对酵母这样的低等真核生物不明显。我们认为这一位点的 G 优势可能与进化有关,生物越高级,G 优势越明显。在 + 4 和 + 5 位点的关联 $D_2(+4)$,在此处出现一个较高的峰,是强关联。仔细研究发现,+ 4 和 + 5 位点出现的是嘧啶保守关联,其关联模式为碱基 TC(关联条件概率为 0.734 联合概率为 0.211)

6.3 酵母基因的终止密码子前和大肠杆菌基因的终止密码子前的编码区,其 $D_1(l)$ 和 $D_2(l)$ 分布特性相似,都有明显的 3 周期特性,而且 3 周期震荡的幅度比编码开始区域大的多;酵母和果蝇基因终止

密码子后的 $D_1(l)$和 $D_2(l)$分布很相象，都是随着l的增大而呈上升趋势，这种上升趋势一直延续到40bp 处，作为原核的大肠杆菌则没有这一性质。这说明原核生物与真核生物在这一区域有明显的差别。

6.4 果蝇基因编码区，$D_1(l)$和 $D_2(l)$分布的 3 周期特征与大肠杆菌和酵母相比，不明显。其原因是：果蝇基因编码区虽然 4 种碱基各自在密码子的 3 个位点的分布呈 3 周期性变化，但在密码子各个位点上 4 种碱基的概率很接近，故而在信息冗余图中果蝇基因编码区的 3 周期特性不明显；另外，我们在选取果蝇基因编码序列时，有一些序列是理论预测的 ORF，它们只是可能的编码区，统计的 $D_1(l)$和 $D_2(l)$分布 3 周期特征不明显，说明了有些理论预测的编码区是错误的，因非编码区没有 3 周期特征，在作统计时，必然会削弱 3 周期的特点。

总之，用重新定义某一位点的单碱基信息冗余和紧邻碱基的信息冗余来寻找基因序列的功能片段和保守位点是直观有效的。由此得到的一些新的结论是值得注意的。

参考文献：

[1] Heijne G. Sequence Analysis in Molecular Biology[J]. *Acad Press Inc*, 1987.

[2] 胜利，罗辽复. 核酸序列的保守位点及双螺旋结构的局域偏差[J]. 内蒙古大学学报(自然科学版)，1991,22：234－247.

[3] 罗辽复. 生命进化的物理观[M]. 上海：上海科学技术出版社，2000.168－175.

[4] Rudd KE，Miller W，Werner C，et al. Mapping sequenced E.coli genes by computer：Software Strategies and examples[J]. *Nucleic Acids Res*，1991,19：673－647.

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

[6] 李宏，杨体强，达赖，罗辽复. 大肠杆菌 SD 序列与基因表达水平的关系[J]. 内蒙古大学学报(自然科学版)，1998,29(2)：172－176.

[7] 陈颖丽，李前忠. SD 序列矩阵表示与保守性[J]. 内蒙古大学学报(自然科学版)，1999,30(3)：329－334.

[8] 吕军，李宏，马克健. 酵母编码区及非编码区的统计分析[J]. 内蒙古大学学报(自然科学版)，2001,32(2)：147－151.

[9] Li H，Luo LF. The relation between codon usage base correlation and gene expression level in Escherichia coli and yeast[J]. *J Theor Biol*，1996,181：111－124.

[10] Kozak M. An analysis of 5'－noncoding sequences from 699 vertebrate messenger RNAs[J]. *Nucleic Acids Research*，1987,15(20)：8125－8132.

[11] 陈颖丽，李前忠. E.coli 和 Yeast 基因起始与终止密码子邻近序列碱基保守性、关联性的对比研究[J]. 内蒙古大学学报(自然科学版)，2000,31(2)：164－169.

AN INFORMATION ENTROPY ANALYSIS OF CONSERVATIVE SITES OF *E.coli*、YEAST AND *Drosophila* GENES

LU Jun[1,2]，　LI Hong[1]，　MA Ke－jian[1]

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

2. Teaching and Research Section of Physics, NeiMongol Industry University, Hohhot 010021, China)

Abstract: The formulation of the single base information redundancy $D_1(l)$ and the adjacent base related information redundancy $D_2(l)$ are revised. For the sequences of upstream and downstream the start codon and the terminal codons of *E.coli*, yeast and *Drosophila* genes, the $D_1(l)$ and $D_2(l)$ for each site l $(l = -30, -29, \cdots, +32, +33)$ are calculated. The results shown that $D_2(l)$ have more information than $D_1(l)$. In site -3 of coding start sequences, $D_1(-3)$ and $D_2(-3)$ have a distinct peak value for yeast and *Drosophila*. In the SD region of *E.coli* gene sequences, $D_1(l)$ and $D_2(l)$have obvious peak value distribution, which is consistent with the others' results. $D_2(l)$ in site $+4$ of coding start sequences in yeast also have a peak value, whose related mode is TC (the combined probability is 0.211). Therefore, the revised information redundancies applied in this thesis are feasible to confirm the conservative sites in DNA sequence.

Key Words: Information entropy；Correlation；Conservative sites