真核生物 DNA 非编码区的组分分析

刘 蓉[1] 齐 震[2] 朱小蓬[2] 凌伦奖[2]* 韩汝珊[1]*

([1]北京大学物理系，北京 100871；[2]中国科学院生物物理研究所，北京 100101）

摘要 在全基因组水平上，用直方图、混沌表示灰度图、距离差异度和信息熵差异度四种方法，研究了拟南芥、线虫、果蝇的 DNA 内含子、基因间隔区 DNA、外显子三种区域的核苷酸短序列组分及组分复杂度. 结果表明：a. 不同基因组之间，不管基因数目多少，用 4 种方法得到的外显子部分其组分复杂度都比较接近，而非编码区部分的组分复杂度却很大. 这一点定量地说明了物种之间的复杂程度，主要不体现在编码区部分，而体现在非编码区部分. b. 同一基因组中，内含子的核苷酸短序列组分复杂度都是相似的，外显子和 intergenic DNA 部分的组分复杂度也是相似的. c. 内含子和 intergenic DNA 在转录、剪切、二级结构等方面有很大的不同，但它们在核苷酸短序列组分上的差异却很小，说明内含子和 intergenic DNA 在转录、剪切、二级结构上的不同并不通过核苷酸短序列组分来进行限制.

关键词 非编码区，生物复杂性，碱基组分，距离差异度，信息熵差异度，混沌表示

学科分类号 Q7

高等真核生物，特别是人的基因组中，非编码区占了 95%～97%. 对非编码区的研究是后基因组时代的巨大挑战[1～3]. 而非编码区研究，一个最基本的工作就是它的碱基组分分析. 以前对碱基组分的分析主要集中在碱基的长程关联和短程关联等方面[4～7]. 以前工作的一个共同特点就是对非编码区的研究不够，大多是将编码区和非编码区不加区分，混在一起进行分析；而将非编码区分成内含子和基因间隔区 DNA（intergenic DNA）两个区域的研究更少. 同时它们的研究对象多集中在古细菌和真细菌里，而复杂真核生物的相关报道较少. 基于非编码区研究的重要意义，及几个复杂真核生物全基因组 DNA 序列的测定，已经有可能从一个更高的水平上来重新审视这个问题. 本文在全基因组水平，研究了几种复杂真核生物的 DNA 非编码区在物种内与物种间的共性与差异.

1 材料与方法

选取进行本研究时已完全测序的 3 个真核生物：拟南芥（*Arabidopsis thaliana*）、线虫（*Caenorhabditis elegans*）和果蝇（*Drosophila melanogaster*），以它们的基因组 DNA 序列为素材，包括拟南芥的Ⅱ，Ⅳ号染色体，AE005172，AE005173 两个 contig，线虫的 6 条染色体，果蝇的 19 条大的 contig，全部数据来自 genbank release 123.0. 将其中的内含子、外显子、intergenic DNA（本文指基因组中除内含子和外显子以外的区域）分别抽取出来，以下简称 3 种区域. 对于一个物种，全基因组中属于同一区域的所有数据序列的总和称为该种区域的全集，而各染色体或 contig 的相应序列称为该种区域的子集. 考察核苷酸短序列，即由 A、T、C、G，4 个字母组成的字串. 比如：长度为 4 的字串有 AAAA，AAAT，AAAC，……GGGC，GGGG，共 $4^4=256$ 种排列模式. 字长为 n 的字串共有 4^n 种排列模式. 如下例所示计算该序列中各种字串出现的总次数：

GTAGGAAACTGGCAAATT…AG

设第 i 条内含子的序列长度为 L_i，我们研究其中字长为 n 的字串，应有 L_i-n+1 个. 若某字串 word 在该序列中出现的次数为 $N_{i,\mathrm{word}}$，则它在该内含子序列中出现的频率为 $F_{i,\mathrm{word}}=N_{i,\mathrm{word}}/(L_i-n+1)$. 设共有 M 条内含子序列，则字串 word 在所有内含子中出现的概率为 $P_{\mathrm{word}}=(F_{1,\mathrm{word}}+F_{2,\mathrm{word}}+\cdots\cdots+F_{M,\mathrm{word}})/M$.

在截取内含子和外显子时，取的是编码链的序列，intergenic DNA 同时取了正链和反向互补链，结果再除以 2 进行平均.

我们考察了上述 3 个物种，三种区域的全集与全集，子集与全集，子集与子集之间字长为 4～6 的各种字串组分和组分复杂度之间的关系. 为了测度这种关系，分别用直方图、混沌表示灰度图

* 通讯联系人.

凌伦奖 Tel：010-64888544，E-mail：ling@sun5.ibp.ac.cn

韩汝珊 Tel：010-62752333，E-mail：rshan@pku.edu.cn

收稿日期：2001-11-26，接受日期：2002-01-28

（FCGR）定性地进行了研究，同时定义了距离差异度和信息熵差异度进行定量研究．第一种方法见文献［5］，第二种方法见文献［8］，后两种方法分别描述如下．

距离差异度：设 $L=4^n$，表示字长为 n 的所有字串的花样数目，第 j 和 k 个集合间字串组分的距离差异度定义为：$D_{jk}=(|P_{j,1}-P_{k,1}|+|P_{j,2}-P_{k,2}|+\cdots\cdots+|P_{j,L}-P_{k,L}|)$，其中 $P_{j,2}$ 表示第二种花样的字串在第 j 个集合中出现的概率．可见 D_{jk} 表示字长为 n 的各种字串在 j，k 两个集合中的概率分布差的绝对值之和．D_{jk} 在表 1～3 中简记为 D．

信息熵差异度：设 $L=4^n$，表示字长为 n 的所有字串的花样数目，第 j 个集合中字串概率分布的信息熵为：$H_j=-(P_{j,1}\ln P_{j,1}+P_{j,2}\ln P_{j,2}+\cdots\cdots+P_{j,L}\ln P_{j,L})$，则第 j 和 k 个集合的信息熵差异度为：$E_{jk}=|H_j-H_k|$．这样定义的差异度可以衡量两个集合的组分复杂度差异．E_{jk} 在表 1～3 中简记为 E．

以上 4 种方法分别从不同侧面对组分进行分析：直方图侧重于直观地表示各字串的概率分布，FCGR 能清楚地给出小于或等于字串长度的各字串概率分布灰度图，距离差异度是一种不加权重的距离量度，信息熵差异度是加了以各字串概率分布为权重的一种距离量度．它们均分析了 DNA 序列集合之间的组分共性与差异．

2 结果与讨论

以下给出了字长为 4 的分析结果，字长为 5，6 的分析结果与之类似．所有相关数据，感兴趣的读者可以通过匿名 FTP 从网址 1（ftp://159.226.118.105）获取．

2.1 同一物种内三个区域间的比较

图 1、图 2 和表 1 分别从不同侧面揭示了：同

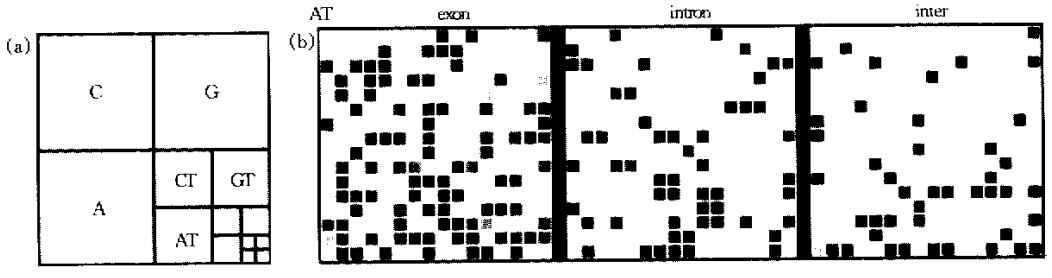

Fig.1 FCGR of *Arabidopsis thaliana*

(a) the arrangement of string of 1～4 nucleotide in blocks；(b) Intron，inter and exon represent separately the three regions.

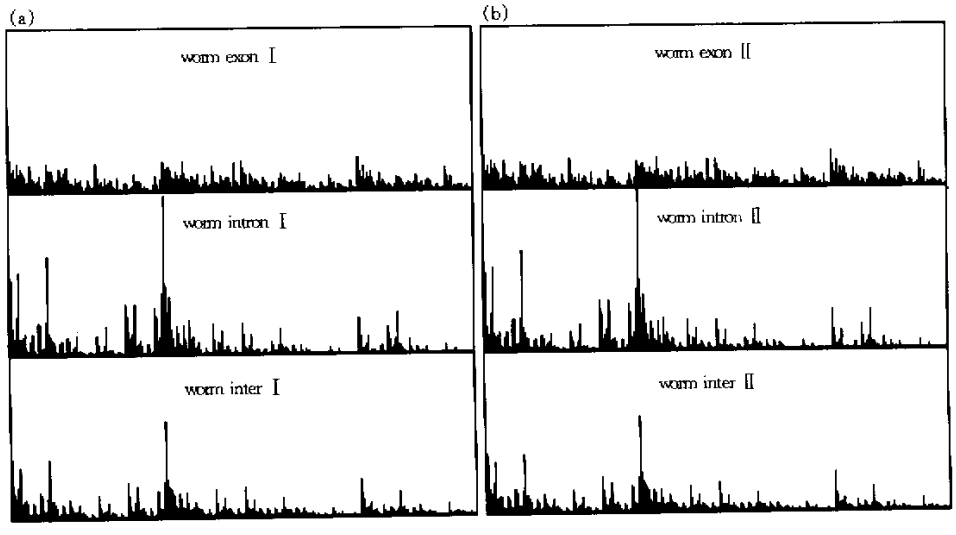

Fig.2 Composition histogram of tetranucleotides over three regions of worm chromsome Ⅰ and Ⅱ

The order of tetranucleotides is AAAA，AAAT，AAAC，AAAG，AATA，AATT…

一物种内，内含子和 intergenic DNA 的组分差别很小. 图 1 中，拟南芥的内含子与 intergenic DNA 的混沌表示灰度图的花样差不多. 它们与外显子的差别就较大. 图 2a，图 2b 中每一个的后两个子图间的差别都比较小，即线虫的内含子与 intergenic DNA 的直方图形状差不多. 表 1 中，拟南芥的内含子与 intergenic DNA 的距离差异度最小（0.2968），内含子和外显子的距离差异度最大（0.5353），intergenic DNA 和外显子之间的距离差异度界于二者之间（0.4290）. 线虫和果蝇的结果类似（其余数据和图表见网址 1）.

Table 1 Discrepancy of distance（*D*）and entropy（*E*）between regions within organisms

	Intron-inter		Intron-exon		Inter-exon	
	D	*E*	*D*	*E*	*D*	*E*
AT	0.2968	0.0235	0.5353	0.2514	0.4290	0.2279
worm	0.2615	0.1609	0.6157	0.3695	0.3876	0.5305
fly	0.2131	0.4451	0.6157	0.1407	0.4971	0.5858

2.2 同一物种内同一种区域间的比较

同一种区域中，同一物种内的组分及其复杂度都非常相似，这种相似性以往是未见报道的.

图 2（a）的每一个子图和图 2（b）的相应子图的差别都很小，表明线虫Ⅰ、Ⅱ号染色体的外显子组分差别很小，内含子、intergenic DNA 的组分差别也很小. 其余染色体之间以及染色体与全集之间的结果类似（网址 1）.

表 2 中，线虫内部，6 条染色体与全集之间内含子的距离差异度均小于 0.0890，intergenic DNA 的距离差异度均小于 0.0517，外显子的距离差异度均小于 0.0744. 6 条染色体与全集之间内含子的信息熵差异度均小于 0.0325，intergenic DNA 的信息熵差异度均小于 0.0811，外显子的信息熵差异度均小于 0.0159，从这些数值看出，它们之间非常接近. 我们还研究了各区域中，各染色体或 contig 之间组分及其复杂度的关系. 结果表明：它们之间的差异也相当小. 拟南芥和果蝇也有类似的结果，相应的图和具体数据见网址 1.

Table 2 Discrepancy of distance（*D*）and entropy（*E*）between each chromsome and whole sets of worm

	Chromosome Ⅰ		Chromosome Ⅱ		Chromosome Ⅲ		Chromosome Ⅳ		Chromosome Ⅴ		Chromosome Ⅹ	
	D	*E*	*D*	*E*	*D*	*E*	*D*	*E*	*D*	*E*	*D*	*E*
intron	0.0354	0.0049	0.0279	0.0005	0.0475	0.0238	0.0223	0.0148	0.0120	0.0017	0.0890	0.0325
inter	0.0317	0.0700	0.0182	0.0256	0.0331	0.0369	0.0181	0.0643	0.0305	0.0811	0.0517	0.0595
exon	0.0364	0.0064	0.0240	0.0061	0.0447	0.0011	0.0245	0.0115	0.0744	0.0158	0.0614	0.0159

2.3 同一种区域不同物种间的差异

表 3 中，外显子区域，三个物种之间的信息熵差异度均小于 0.0617；而内含子和 intergenic DNA 区域，三个物种之间信息熵差异度均大于 0.1271.

Table 3 Discrepancy of entropy（*E*）between organisms

E	AT worm	AT fly	Worm fly
intron	0.2896	0.1455	0.2905
inter	0.3322	0.2051	0.1271
exon	0.0269	0.0348	0.0617

说明不同物种间，外显子区域的组分复杂度最接近；内含子、intergenic DNA 区域组分复杂度的差别较大，比外显子区域的组分复杂度大一个数量级. 从直方图、混沌表示灰度图、距离差异度方面也可以看到这一点.

综上所述，我们的研究结果可概括为：a. 同一物种内，内含子和 intergenic DNA 的组分很接近，与外显子相比，差别较大. 其中，内含子和外显子的差别最大，intergenic DNA 和外显子的差别次之. b. 内含子、intergenic DNA、外显子这三个区域的同一区域中，同一物种内各染色体或 contig 之间，任一染色体或 contig 与它们所组成的全集之间组分和组分复杂度的差别很小，说明它们之间具有很强的相似性现象. c. 不同物种之间，外显子之间的组分复杂度最接近，内含子之间和 intergenic DNA 之间的组分复杂度的差别都较大. 编码区和非编码区的组分复杂度在信息熵差异度定义下相差一个数量级.

3 讨 论

a. 结论暗示了：内含子和 intergenic DNA 可

能受到相似的限制，反映了整个非编码区对整个选择压力的反应. 内含子和 intergenic DNA 在转录、剪切、二级结构等方面有很大的不同，但它们在核苷酸短序列组分上的差异却很小，说明内含子和 intergenic DNA 在转录、剪切、二级结构上的不同并不通过核苷酸短序列组分来进行限制. 这一点从不同侧面验证了文献［9，10］的结论.

b. 结论表明同一物种各染色体或 contig 之间具有某种自相似性现象，这种自相似现象是以往文献未见报道的. 它的一个潜在应用价值是，由已完全对编码区和非编码区进行定位的区域，来推测那些没有经过实验验证的计算机预测定位是否准确. 我们认为果蝇的数据少部分有例外，就是因为现在还有很多基因没有完全定位的原因.

c. 结论中外显子之间的组分复杂度最接近，这是因为编码区受到三联体密码子的限制. 内含子之间和 intergenic DNA 之间组分复杂度的差别都较大，表明它们不受一种简单的规律所限制. 不同物种之间，外显子之间的组分复杂度远远小于内含子和 intergenic DNA 之间的组分复杂度，这一点定量地说明了物种之间的复杂程度，无论是动物（果蝇、线虫）还是植物（拟南芥）主要不体现在编码区部分，而体现在非编码区部分.

到目前为止，已完全测序的真核基因组还有酵母和人，由于酵母的内含子相对较少，而人的大量基因和非编码区还没有完全定位，所以本文没选用. 但我们从人的已定位的部分数据中也得出了与本文类似的结论. 由于全基因组数据运算量较大，本文没有分析 7 个字串以上的组分. 下一步工作就是要改进计算方法，提高运算速度，改变存储方法，研究 7 到几十字长的字串规律. 同时将选择的物种范围进一步扩大到未完全测序的物种中去. 由于非编码区中存在着很多调控元件和编码 RNA 的 DNA 序列，通过研究不同物种间组分的差别，有可能揭示出这些有价值的部分，同时为基因识别算法和进化研究提供一定的依据.

参 考 文 献

1 International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature，2001，**409**（6822）：860～921

2 Hardison R C. Conserved noncoding sequences are reliable guides to regulatory elements. Trends in Genetics，2000，**16**（9）：369～372

3 陈润生. 生物大分子结构的理论分析与模拟. 见：郝柏林，刘寄星，编. 理论物理与生命科学. 上海：上海科学技术出版社，1997. 45～61
Chen R S. Theory analysis and simulation of the structure of biological macro moleculars. In：Hao B L，Liu J X，eds. Theory Physics and Life Science. Shanghai：Shanghai Scientific and Technological Press，1997. 45～61

4 Luo L F，Lee W J，Jia L J，*et al*. Statistical correlation of nucleotides in a DNA sequence. Phycial Review E，1998，**58**（1）：861～871

5 Gentles A J，Karlin S. Genome-scale compositional comparisons in eukaryotes. Genome Research，2001，**11**（4）：540～546

6 Hao B L. Fractals from genomes-exact solutions of a biology-inspired problem. Physica A，2000，**282**（1～2）：225～246

7 Hao B L，Lee H C，Zhang S Y. Fractals related to long DNA sequences and complete genomes. Chaos Solitions Fractals，2000，**11**（6）：825～836

8 Almeida J S，Carriço J A，Maretzek A，*et al*. Analysis of genomic sequences by chaos game representation. Bioinformatics，2001，**17**（5）：429～437

9 Bergman C M，Kreitman M. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. Genome Reseach，2001，**11**（8）：1335～1345

10 Clark A G. The search for meaning in noncoding DNA. Genome Research，2001，**11**（8）：1319～1320

Genome-scale Compositional Comparisons in Noncoding Regions of Eukaryotes

LIU Rong[1)]，QI Zhen[2)]，ZHU Xiao-Peng[2)]，LING Lun-Jiang[2)]*，HAN Ru-Shan[1)]*

([1)] *Department of Physics，Peking University，Beijing* 100871，China；

[2)] *Institute of Biophysics，The Chinese Academy of Sciences，Beijing* 100101，China)

Abstract By using four methods (histogram，chaos game representation，discrepancy of distance and discrepancy of entropy) at genomic level，the composition of short oligonucleotides and their compositional complexities in three different regions (introns，intergenic DNAs and exons) of genomic DNA from *Arabidopsis thaliana*，*Caenorhabditis elegans* and *Drosophila melanogaster* were studied. It can be concluded that：(1) although the genome sizes and gene numbers are quite different，the compositional complexities of exons are

similar，while that of noncoding regions are quite different between eukaryotic genomes. From quantitative perspective，this finding means that the degree of organismal complexity is mainly reflected by noncoding regions，but not by exons；(2) in the same regions of genomic DNAs，composition and compositional complexity are highly similar between chromosomes or contigs within eukaryotic genomes；(3) composition differ remarkably little between introns and intergenic DNAs. This suggests that the effects of transcription，splicing，second structure contribute minimally to the constraints operating on these sequences.

Key words noncoding region，organismal complexity，base composition，discrepancy of distance，discrepancy of entropy，chaos game representation

* Corresponding author.
LING Lun-Jiang，Tel：86-10-64888544，E-mail：ling@sun5.ibp.ac.cn
HAN Ru-Shan，Tel：86-10-62752333，E-mail：rshan@pku.edu.cn
Received：November 26，2001 Accepted：January 28，2002

欢迎订阅 2003 年《动物学杂志》

《动物学杂志》是由中国科学院动物研究所、中国动物学会主办的科技期刊，亦是中国自然科学核心期刊，2001 年入闱“中国期刊方阵”.《动物学杂志》是以普及与提高相结合、基础性和应用性并重为宗旨的综合性学术刊物. 力求及时报道动物科学领域具有创造性和重要意义的最新研究成果，介绍有创见的新思想、新学说、新技术、新方法，开展学术交流与争鸣。主要栏目有：研究报告、动物资源与管理、珍稀濒危动物、动物养殖、有害动物防治、技术与方法、研究简报和快讯、基础资料、自然保护区、综述与进展、学术论坛、专题知识讲座、科技动态、新书评介等. 读者对象为：动物科学领域的研究、教学、技术、管理人员及广大业余爱好者.

《动物学杂志》为 16 开，双月刊，2003 年增到 96 页，每册定价 14 元，全年 84 元. 国内外公开发行，国内邮发代号：2-422；国外发行代号（Code No.）：BM58，全国各地邮局均可订阅. 如未能在当地邮局订到，亦可与编辑部直接联系订阅.

联系地址：北京中关村路 19 号 100080，《动物学杂志》编辑部

电话：010-62581475；传真：010-62569682；E-mail：journal@panda.ioz.ac.cn

欢迎订阅《动物学杂志》；欢迎您在《动物学杂志》上刊登广告.