基于混沌游戏表示的 DNA 序列的信息维数[1]

薛峰，田逢春，初春，龙红梅
重庆大学通信工程学院，重庆 (400030)
E-mail：xuefeng207@163.com

摘　要：本文在 DNA 序列混沌游戏表示(chaos-game representation，简称 CGR)图形化方法的基础上，提出一种计算 DNA 序列的分形信息维数的方法。实验结果表明，对同一物种的编码区序列的信息维数比非编码区序列的高。

关键词：DNA 序列，信息维数，混沌游戏表示，编码区序列
中图分类号：Q811.4　TN911.2

1．引言

伴随着人类基因组计划的完成，产生了海量的 DNA 序列数据。破译这些 DNA 序列密码，弄清 DNA 序列与生物进化细胞功能，遗传机理和疾病发生的关系，具有重要的生物学意义。

Jeffrey 提出的基于迭代函数的 DNA 序列混沌游戏表示法(chaos-game representation，简称 CGR)[1]，将 DNA 序列中一定长度字的分布规律表现为图形的分形特征，进而通过分形分析就可获取序列的分布规律，从而也成为 DNA 序列分析的一种统计方法。采用 CGR 图形方法表示基因组序列具有直观、不受序列长程相关性的影响、不依赖于序列尺度以及计算速度快等优点，因此可以克服目前 DNA 序列分析中的一些缺陷，如基因预测程序过于依赖模型、多序列比对受基因重排和计算复杂度的限制、以及不同规模的基因组之间难以比较等[2]-[3]。

本文将在 DNA 序列 CGR 图形化方法的基础上，提出一种计算 DNA 序列的分形信息维数的方法，并对同一物种的编码区序列和非编码区序列的信息维数进行了讨论。这对 DNA 序列的分形特性及生物遗传编码进化等的进一步研究提供了一种思路。

2．混沌游戏

给定平面上三点 A，B，C，再任意给定起点 Z，然后掷一粒骰子。规则：如果骰子停稳后向上的面为 1 或 2 点，则取起点与点 A 之间距离的中点为新的起点；掷出 3 或 4 点则与 B 点之间距离的中点为新的起点；5 或 6 点则与 C 点之间距离的中点为新的起点。这一过程持续下去，最后所有的随机序列对应的点就可形成一个图案，称该游戏为混沌游戏。图 1 为计算机模拟的随机投掷 2000 次后得到的图案，图 2 为随机投掷 20000 次后得到的图案。

[1] 本课题得到高等学校博士学科点专项科研基金资助（教育部）资助（项目编号：20050611022）。

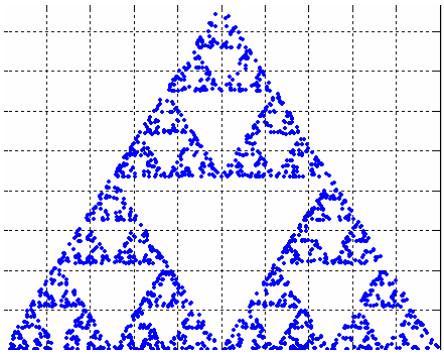


图 1 随机投掷 2000 次

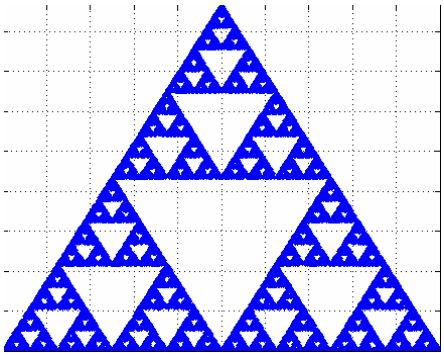


图 2 随机投掷 20000 次

游戏规则是灵活的。上述玩法用了三个定点、一个随机点和一粒骰子。从图可以看出，得到的结果是惊人的。一个混沌过程竟然产生出有序的结果！

3．DNA序列的混沌游戏表示

混沌游戏表示法是由 Jeffrey 提出的基于迭代函数系统来表示符号序列的一种方法。该方法如下：平面的四个顶点分别代表四种核苷酸{C，A，G，T}，DNA 序列的迭代函数也可产生一个平面图，每个核苷酸在平面中的位置 P_i 表示为：

$$P_i = P_{i-1} + 0.5(P_{i-1} - S_i)，\ 1 \le i \le N \qquad P_0 = (0.5, 0.5) \qquad （1）$$

其中 P_0 为任意给定的起点， N 表示序列的长度； S_i 表示序列中第 i 个核苷酸，其对应固定顶点坐标，分别为 C=（0，0），A=（1，0），G=（1，1），T=（1，0）。

举一个简单的例子，对只有 7 个核苷酸的序列 CACGTGA，图 3 显示了其按照迭代函数系统（1）得到的 CGR 图形的绘制路线。为直观起见，图中我们画出了点的走向，实际绘制时，每个字符仅用一个点来表示。

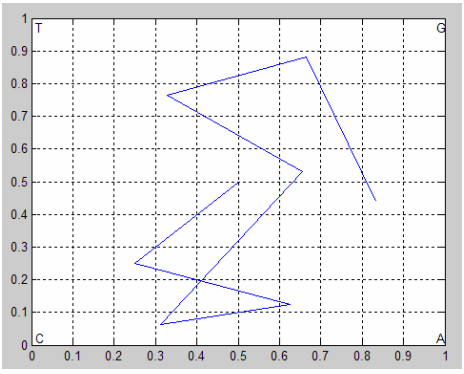


图 3 序列 CACGTGA 的图形绘制

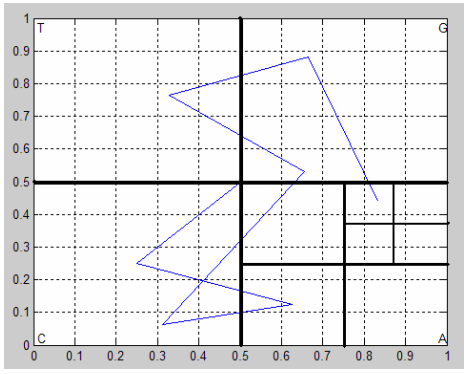


图 4 恢复原始的 DNA 序列

不同的 DNA 序列得到的 CGR 图形是不一样的。我们可以根据迭代函数系统（1）恢复原始的 DNA 序列。实际上，如果精度允许，只需要最后一个点的坐标即可恢复原始的 DNA 序列。如图 4 所示。最后一个点的坐标为 $P_7 = (0.8320, 0.4414)$。先将单位正方形一分为四， P_7 位于右下象限（A 象限），将 A 象限一分为四， P_7 位于右上象限（G 象限），再一分为四， P_7 位于左上象限（T 象限），以此类推。这样就得到原始序列从后至前为 AGT……。

从 GenBank 数据库中选取编号为 AC148573 的序列，即 Taeniopygia guttata clone TG_Ba-372P20 的序列，其长度为 129282bp，由 CGR 法得到的图形如图 5 所示。可以看到，图形具有明显的分形特征。

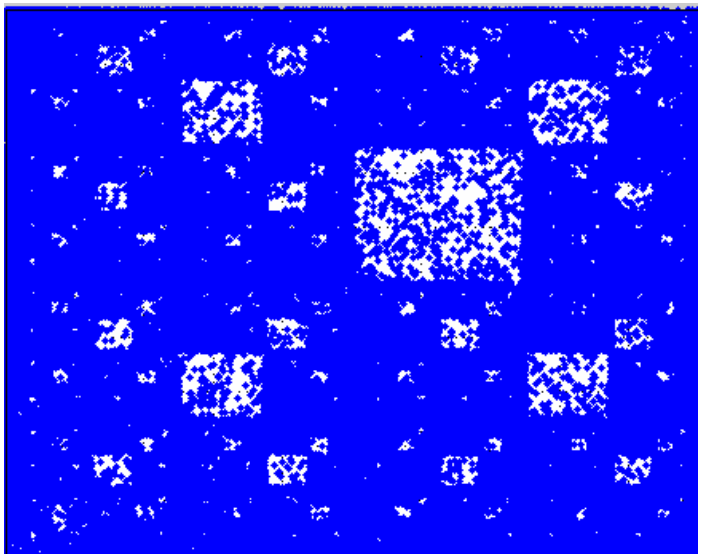
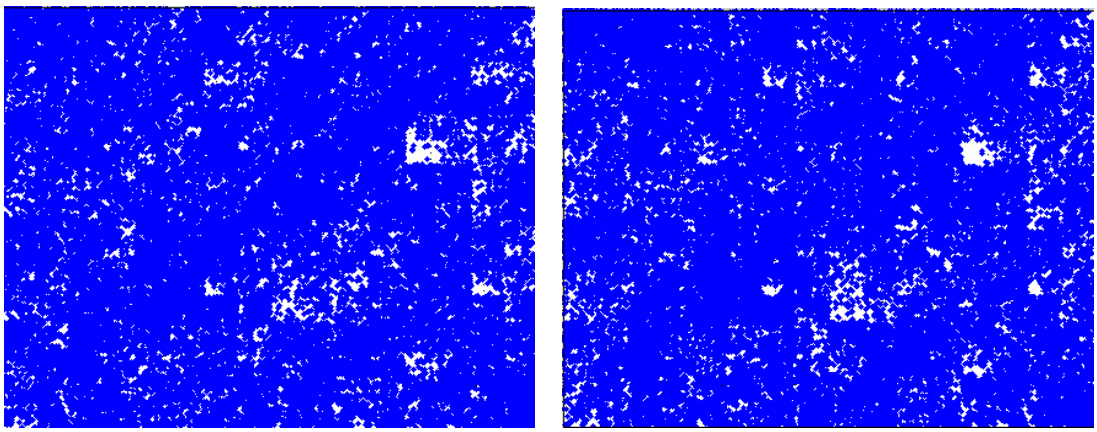


图 5 序列 AC148573 的 CGR 图形

进一步讨论同种生物基因组的不同片段的 CGR 图形。我们从 GenBank 数据库中选取编号为 U00096 的序列，即大肠杆菌（Escherichia coli）基因组序列，长度为 4639675bp，我们以每 50000bp 长度作一个 CGD 图形，得到的前四个图形如图 6 所示。从图中我们可以看出，大肠杆菌基因组的不同片段的 CGD 图形直观上具有很强的相似性。



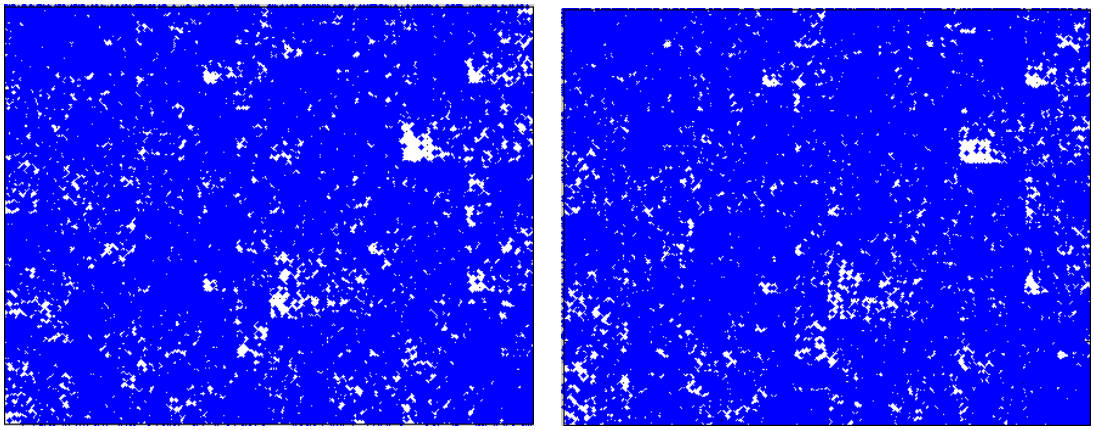


图 6 大肠杆菌基因组不同片段序列的 CGR 图形

4. 信息熵和信息维数

信息熵和信息维数可以对图形的分形特征进行定量的描述。

4.1 信息熵

考虑概率不同的一组互斥事件的集合，用符号 $B_1,.....,B_N$ 表示事件，用 $p_1,.....p_N$ 表示概率（假定它们的总和为 1）。根据信息熵的定义，有

$$I=\sum_{i=1}^{N} p_i \log_2 \frac{1}{p_i} \qquad (2)$$

我们将图 5 划分成 n^2 个盒子，DNA 序列中的字符在每个盒子中的概率为 $p_{i,j}$（$i,j=1,\cdots,n$）。在这种情况下，信息熵为：

$$I=\sum_{i,j=1}^{n} p_{i,j} \log_2 \frac{1}{p_{i,j}} \qquad (3)$$

显然，当盒子的个数增加（即盒子尺寸 s 减小）时，信息熵 $I(s)$ 必定增加。考虑用分形维数来度量图形的有序性和不规则性。

4.2 信息维数

分形维数是分形几何学中的一个十分重要的参数，利用分维对图形的复杂性进行度量的计算方法很多。在复杂图形的研究中，常常应用的是容量维(计盒维)。容量维的计算仅考虑了覆盖整个图形的盒子数，而未考虑一个盒子内所包含的点数，而且实际计算中，计盒维的数值往往会随着研究的点数的增加而变大。显然，应用容量维数描述复杂图形的特征有其局限性。下面考虑对信息维数计算的一种改进。

信息维数的计算公式为[4]：

$$D_I=\lim_{s\to 0}\frac{I(s)}{\log_2 s} \qquad (4)$$

其中，s 为盒子的尺寸，例如，当盒子个数为 $2^2=4$ 时，$s=\frac{1}{2}$。显然，当盒子的尺寸 s 减小时，信息熵 $I(s)$ 必定增加。对复杂图形的实际计算时，很难做到（4）式中的盒子的

尺寸 $s \to 0$，因而考虑下面的一种改进计算。对图 5 的情况，表 7 给出了各种 s 情况下信息维数的估值。

表 7 信息的维数的估值

k	s_k	$I(s_k)$	D_I
1	1/2	1.9768	
2	1/4	3.9295	1.9527
3	1/16	5.8546	1.9251
4	1/32	7.7274	1.8728

我们利用(2)式计算信息熵 $I(s_k)$。最后一栏是信息维数的估值。最后一栏（D_I）中的第 k 项是按 $I(s_k)-I(s_{k-1})$ 算出的。这些数据揭示了一个幂律，即当 $s \to 0$ 时，$I(s)$ 随 $1/s$ 的对数线性增加。换句话说，

$$I(s) \approx I_0 + D_I \log_2 \frac{1}{s} \qquad (5)$$

其中 I_0 是一个常数，而 D_I 是 $I(s)$ 与 $\log_2 1/s$ 的关系曲线的斜率。D_I 刻画了这一信息的增长，D_I 即为信息维数。这样计算的 D_I 不会随盒子的尺寸 s 的变化而变化，一般来说，对二维图形，$D_I < 2$。

5. 计算和讨论

根据上面的讨论，对图 5 的序列，根据（5）式进行线性回归，得到 $D_I = 1.8614$。同理，对图 6 的 4 个序列，进行计算得到的信息维数 D_I 依次为：1.9575，1.9564，1.9589，1.9619。

进一步的计算可以知道，对不同物种的 DNA 序列的信息维数并没有看出明显的规律。不过，对同一物种的编码区序列和非编码区序列的信息维数具有明显的统计规律。但序列片段长度的不同，也会对信息维数的计算造成差异。我们以 Oenothera argillicola（一种陆生植物）、Alligator mississippiensis（鳄鱼）、Canis lupus familiaris（犬属）、Drosophila melanogaster（果蝇）为例，从其已经确定编码位置的 DNA 上每个物种各选取编码序列和非编码序列若干段，每段长度约为 1000bp，然后计算各段的信息维数。表 8 显示了对这几个物种的计算结果。

表 8 对几个物种的计算结果

类别	编码区序列 信息维数平均值	非编码区序列 信息维数平均值
Oenothera argillicola	1.9391	1.9043
Alligator mississippiensis	1.9314	1.9077
Canis lupus familiaris	1.8548	1.8247
Drosophila melanogaster	1.9182	1.8984

从表 8 可以看出明显的统计规律：同一物种的编码区序列的信息维数比非编码区序列的高。

6. 结论

DNA 序列的 CGR 图形具有明显的分形特征，不同物种的 DNA 序列的信息维数尽管没有明显的规律，但对同一物种的编码区序列的信息维数比非编码区序列的高。而由信息熵和信息维数的定义可知，当所有事件完全随机均匀发生时，信息熵和信息维数的值达到最大。这就说明，同一物种中编码区序列的核苷酸排列的随机性和不规则性比非编码区序列的高。

利用混沌游戏表示可以把任意一条 DNA 序列映射到二维平面的一个图形，得到序列的分形图形。序列片段的信息维数，可以作为我们区分编码区序列和非编码区序列的一个指标。当然，更准确的验证还要通过生物学的手段。这对序列的进一步分析提供了一些新的有益的思路。

参考文献

[1] Almeida J S,Carrico J A,Maretzek A. Analysis of genomic sequences by chaos game representation[J]. Bioinformatics. 2001, 17(5):429-437.
[2] 奚襄君等.人类 DNA 序列图形化参数的提取和应用[J].复旦学报（自然科学版）. Vol. 44 No.6 Dec.2005:960-963.
[3] 李小妹等.DNA 序列的分形结构[J].计算机科学. Vol.32 No.8 2005:326-328.
[4] Heinz-Otto Peitgen,Hartmut Juergens, Dietmar Saupe. Chaos and Fractals: New Frontiers of Science[M]. Springer. 2004.

Information Dimension of DNA Sequence based on the Chaos-Game Representation

Xue Feng，Tian Fengchun，Chu Chun，Long Hongmei
Communication Engineering College of Chongqing University，Chongqing (400030)

Abstract
This paper presents a calculation method of fractal information dimension of the DNA sequence based on the chaos-game representation (CGR). The experimental results show that for the same species the information dimension of Coding sequence (CDS) is bigger than that of the non-coding sequences.
Keywords: DNA sequence, information dimension, chaos-game representation，coding sequence