维普资讯 http://www.cqvip.com

基于分形的DNA序列可视化表示研究*)

王树林[1,2] 王 戟[1] 陈火旺[1] 张波云[1]

(国防科技大学计算机学院 长沙 410073)[1] (湖南大学计算机与通信学院 长沙 410082)[2]

摘 要 DNA序列可视化表示对于研究其结构与功能具有至关重要的意义,它有助于重复子序列的识别、内含子与外显子的区分以及DNA序列进化研究等等。本文首先介绍了生成DNA序列分形图像的Hao方法和经典的混沌游戏方法,然后深入分析和比较了这两种方法的异同点,并讨论了禁止子序列中回文子序列情况;紧接着,阐述了迭代函数系统产生分形吸引子的数学机理,并根据Moore自动机与迭代函数系统定义了混沌自动机,然后详细研究了以DNA序列驱动混沌自动机产生分形图像的方法;最后提出DNA序列三联密码子的分形图像表示方法,并对其进行了初步研究。

关键词 DNA序列,迭代函数系统,混沌自动机,分形

The Research of Visualization of DNA Sequences Based on Fractal

WANG Shu-Lin[1,2] WANG Ji[1] CHEN Huo-Wang[1] ZHANG Bo-Yun[1]

(School of Computer Science, National University of Defense Technology, Changsha 410073)[1]

(College of Computer and Communication, Hunan University, Changsha 410082)[2]

Abstract The visual representation of DNA sequences plays a profound role in the research of DNA structure and its function, and it is especially helpful to the recognition of repeated subsequences, the partition into intron and exon, and the research of DNA sequences evolution, etc. At first, the two methods concerning Hao's method and the classic chaos game representation method that can generate the fractal image of DNA sequences are introduced in this paper, and further the comparison and contrast between the two methods are provided in details. On the basis of these works, the palindromes are discussed in forbidden subsequences. Second, after the mathematical principle of iterated function system generating fractal attractor is introduced, the chaos automata is defined according to the Moore finite state machine together with iterated function system, and the method of chaos automata generating fractal image driven by DNA sequences is investigated deeply and further. At last, this paper proposes and investigates the method of generating fractal image driven by codon sequences in DNA sequences.

Keywords DNA sequences, Iterated function system, Chaos automata, Fractal

1 引言

生物DNA序列所携带的大量遗传信息的传递与表达与其核苷酸序列结构之间存在密切关系,因此深入研究DNA序列结构,对于生命起源与奥秘的揭示、遗传疾病的治疗以及新一代DNA计算机的研制等都具有重要的现实意义。以人类基因组计划顺利完成为标志的后基因组时代的到来,各学科领域科学家分别从不同学科角度对DNA序列结构进行研究成为新的热点之一。抽象来看,对DNA序列结构的研究主要在微观和宏观两个层次展开。具体讲,则涉及基因识别、基因表达机理、序列特征提取、基因组演化、基因组整体功能和调节网络等方面。但由于DNA序列结构具有非线性、突变性、有序性、某种程度的随机性和异常复杂性等特点,使得研究人员在短期内完全理解由A、G、C和T4种碱基字母组成的DNA序列结构是不现实的。读懂这部仅有4个碱基字母写成的没有标点符号的大书,还需要科学家对DNA序列结构进行更深入、更全面的研究,并在此基础上建立起对所有生物体具有普适性的理论模型。

自1995年首个完整基因组测序完成以来,核苷酸序列数据呈指数级增长,大约每14个月核苷酸序列数据就会翻一番,这为DNA序列结构的研究既带来了机遇又带来巨大的挑战。在这样一种情形下,针对DNA序列内部结构分析的统计研究方法愈发显得重要,但对生物学研究人员来说,DNA序列可视化能够提供更加直观的信息。其可视化方法主要分为两种:图形表示和图像表示。图形表示是指把DNA序列映射成空间的线性结构,如DNA序列的一维、二维、三维以及多维行走等,多为曲线表示形式;而图像表示则是把DNA序列映射到空间的一个区域或体域,多为2D或3D图像表示形式。本文主要研究DNA序列的平面分形图像表示方法。

在DNA序列可视化表示方面,Jeffrey(1990)首先提出了一种称之为混沌游戏表示[10](Chaos Game Representation,简称CGR)的子序列结构可视化方法,它是一种以DNA序列驱动简单迭代函数系统(Iterated Function System,简称IFS)产生分形图像的经典方法。文[11,12]中,郝柏林院士等提出了另一种基于DNA子序列出现频率的可视化方法(简称Hao方法),同样能产生类似的分形图像。这些分形图像充分表明DNA序列具有整体和局部的结构性和长程相关性。2000

*)国家自然科学基金资助项目(60233020)。王树林 博士生,主要研究方向为:生物信息学、软件工程和复杂系统;王 戟 博士,教授,博士生导师,CCF高级会员,研究方向:高可信软件、Agent软件方法学;陈火旺 教授,博士生导师,中国工程院院士,主要研究领域为:软件理论、人工智能和软件工程;张波云 博士生,研究方向:信息安全。

维普资讯 http://www.cqvip.com

年，Daniel Ashlock 提出并研究了新的基于迭代函数系统的 DNA 序列分形表示方法[1,8]，并引入演化计算思想，对混沌自动机进行演化以对序列进行可视化分类。本文首先介绍并比较了他们的研究工作，着重阐述了产生分形的 IFS 数学机理，并以 DNA 序列驱动混沌自动机的形式生成了许多分形图像。实验所用 DNA 序列均取自 NCBI(http://www.ncbi.nlm.nih.gov)的核苷酸序列数据库，文中所涉及的分形图像都是采用编号为 AY664416 的 Zea Mays DNA 序列来绘制的，其长度为 203581bp。

2 两种传统 DNA 序列分形表示方法

2.1 Hao 序列分形表示方法[11]

所有 DNA 序列都可看作由 4-Base={C,G,A,T}中的 4 种碱基构成的线性序列，其中不同的 k-长子序列共有 4^k 个。当 $k=1$ 时，显然只有 4 个不同的 1-长子序列，以 2×2 矩阵的形式表示如下：

$$M=\begin{pmatrix} C & G \\ T & A \end{pmatrix}$$

它的每一个元素可以表示为

$M_{0,0}=C, M_{0,1}=G,$

$M_{1,0}=A, M_{1,1}=T。$

其下标采用二进制表示。当 $k=2$ 时，共有 16 种不同的 2-长子序列，可由 M 通过⊗运算构造出来：

$$M^2=M\otimes M=\begin{pmatrix} CC & GG & GC & GG \\ CT & CA & GT & GA \\ TC & TG & AC & AG \\ TT & TA & AT & AA \end{pmatrix}$$

类似地，可通过对 M 的⊗运算构造出 M^k，即所有的 k-长子序列的矩阵可表示为：

$$M^k=M\otimes M\otimes\cdots\otimes M$$

它的任一个元素可表示为：

$$M^k_{i_1 i_2\cdots i_k, j_1 j_2\cdots j_k}=M_{i_1,j_1}\oplus M_{i_2,j_2}\oplus\cdots\oplus M_{i_k,j_k}$$

其中⊕表示字符串连接运算。

为实现上的方便，定义函数

$f: 4-Base \mapsto Binary$，其中 $Binary$ 为二进制表示的整数，即

$$f(x)=\begin{cases} 00 & x=C \\ 01 & x=G \\ 10 & x=A \\ 11 & x=T \end{cases}$$

对每一个 k-长子序列 $s_1 s_2\cdots s_i\cdots s_k$ ($s_i\in 4-Base$)，经下列子序列索引函数变换后，可得到唯一整数值与之对应：

$$h(s_1 s_2\cdots s_i\cdots s_k)=\sum_{i=0}^{k-1}4^{k-i-1}f(s_i), s_i\in 4-Base$$

为了能在笛卡尔坐标平面的 $2^k\times 2^k$ 区域上显示图像，可以通过下列两个函数建立 k-长子序列与点坐标 (x,y) 的关系：

$$x=\sum_{i=0}^{k-1}2^{k-i-1}\mathrm{low}(f(s_i)), s_i\in 4-Base$$

$$y=\sum_{i=1}^{k-1}2^{k-i-1}\mathrm{high}(f(s_i)), s_i\in 4-Base$$

其中 low 函数表示取二进制串的最低一位，high 函数表示取二进制串的最高一位。在这样一种映射下，一个 DNA k-长子序列对应坐标平面上的一个点 (x,y)，显然相同 k-长子序列对应的点也相同；很容易设计一个算法来构造 DNA 序列的分形图像。图 1 显示的是 Zea mays DNA 序列的 9-长子序列分形图像，其中的白点可能就是缺失的子序列或禁止子序列。

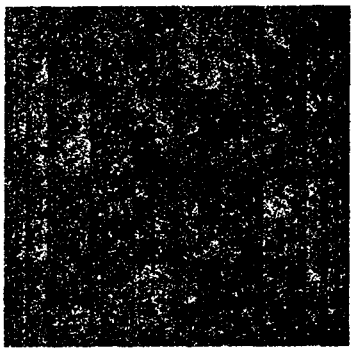


图 1 Zea mays DNA 序列 Hao 方法分形图像

2.2 混沌游戏表示(CGR)[10]

最简单的混沌游戏是由不在一条直线上的 3 个点所产生的 Serpinski 三角形，称之为 3-CGR。设 3-CGR 的 3 个固定点组成的集合为 $FP=\{(fx_1,fy_1),(fx_2,fy_2),(fx_3,fy_3)\}$。任取初始点 $(x_1,y_1)\in FP$，N 为自然数集合，则动点集合

$$MP=\{(x_{n+1},y_{n+1})\mid x_{n+1}=(x_n+fx_i)/2$$

$$y_{n+1}=(y_n+fy_i)/2, n\in N, i=1,2,3\}$$

构成该 3-CGR 的吸引子。在每次迭代计算中，$i\in\{1,2,3\}$ 是随机选取的，设 i_k 表示第 k 次迭代在集合{1,2,3}中的随机选择，则得到一序列：

$$S=i_1 i_2\cdots i_k\cdots i_n \quad (i_k\in\{1,2,3\}, 1\leqslant k\leqslant n)$$

构成一个随机序列，因此动点集合 MP 可以看成是由 S 随机序列驱动 3-CGR 产生的吸引子，如图 2 所示。

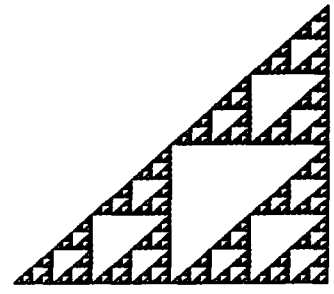


图 2 三角混沌游戏吸引子

类似地，四点混沌游戏(简称 4-CGR)以正方形的 4 个顶点构成 4 个固定点：

$$FP=\{(fx_1,fy_1),(fx_2,fy_2),(fx_3,fy_3),(fx_4,fy_4)\}$$

按照类似的方法可得到 4-CGR 吸引子，所得到的动点集合为：

$$MP=\{(x_{n+1},y_{n+1})\mid x_{n+1}=(x_n+fx_i)/2$$

$$y_{n+1}=(y_n+fy_i)/2, n\in N, i=1,2,3,4\}$$

它将随机地充满整个矩形。假设正方形的边长为 1，则动点集合 MP 中的 x 和 y 坐标都具有 $\sum_{k=1}^{\infty}\frac{b_k}{2^k}$，$b_k\in\{0,1\}$ 的形式。而这样的数被称之为二进有理数，二进有理数在区间[0,1]内是稠密的。同理，动点集合 MP 也可看成是由序列

$$S=i_1 i_2\cdots i_k\cdots i_n \quad (i_k\in\{1,2,3,4\}, 1\leqslant k\leqslant n)$$

驱动的。由于 i_k 在每次迭代中是随机地取值，因此序列 S 是一个随机序列。但当 S 不是一个随机序列时，则正方形将不能被动点集合 MP 充满。定义 DNA 序列映射：

维普资讯 http://www.cqvip.com

$$S(x)=\begin{cases}1 & x=C\\ 2 & x=G\\ 3 & x=A\\ 4 & x=T\end{cases}$$

则 S(DNA 序列)为以集合{1,2,3,4}中字符数字组成的一个序列。例如:S(CGAT)=1234。以 S(Zea Mays DNA 序列)驱动 4-CGR,则可得到具有明显分形结构的图像。图 3 显示的是 DNA 序列中碱基在图像中的位置分布情况,图 4 是 Zea Mays DNA 序列的混沌游戏吸引子。同样,图像中的白点表明 DNA 序列中碱基子序列的缺失情况,或是该基因组中的禁止子序列。需注意的是,这里的 DNA 子序列中的碱基不是绝对相邻的。

<table>
<tr><td>CC</td><td colspan="2">CG</td><td rowspan="2">G</td></tr>
<tr><td>CT</td><td colspan="2">CA</td></tr>
<tr><td rowspan="2">TC</td><td>TGC</td><td>TGG</td><td rowspan="3">A</td></tr>
<tr><td>TGT</td><td>TGA</td></tr>
<tr><td>TT</td><td colspan="2">TA</td></tr>
</table>

图 3　DNA 子序列在分形图中的分布情况

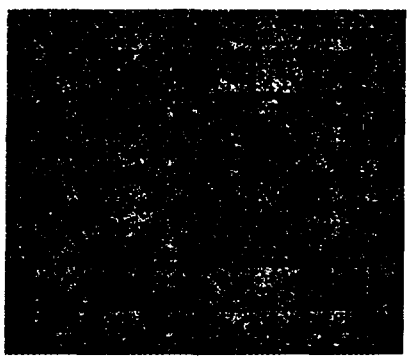


图 4　Zea Mays DNA 序列的 CGR 分形图像

2.3　两种方法的比较

虽然 Hao 方法和 CGR 方法生成 DNA 序列分形图像的过程有很大的区别,但生成的分形图像却是非常相似的,都表明 DNA 序列具有一定的结构而不是纯粹随机的序列。下面对这两种方法进行比较。

(1)在生成图像的分辨率方面,Hao 方法采用固定图像分辨率,其分辨率取决于子序列长度 k,一个像素点对应一个确定的 k-长 DNA 子序列。因此,当 k 值较大时,将无法显示完整的分形图像。而 CGR 方法所生成的图像分辨率不是固定的,可以有很大的变化,但要受屏幕分辨率的限制;虽然在绘制时,一个碱基绘制一个像素点,但实际上一个像素点也代表一个以本碱基字母为后缀的 k-长子序列;为了揭示更多细节信息,可以采用鹰眼技术把部分图像放大以查看更精细的结构。

(2)在图像的上色方面,Hao 方法为图像上色是比较容易的,比如为不同出现频率的点设置不同的颜色,频率越高颜色越深,因此从分形图像上很直观地看出哪些是高频出现的子序列,哪些是极少出现的子序列;而采用 CGR 方法为图像上色需要计算点密度,相对来说复杂一些。

(3)在 Hao 方法生成的分形图像中,相邻的 k-长子序列所对应的像素点不一定相邻;而在 CGR 生成的图像中,相邻碱基对应的像素点也不一定相邻,但在一个小的邻域内所有像素点所表示的 k-长子序列在一定的尺度下是相同的。

(4)这两种方法所生成图像中的像素点与 DNA 序列有很好的对应关系,因此分形图像的数学特征就是 DNA 序列的潜在特征。如果 DNA 序列分形图像相似,也就表明 DNA 序列相似。因此,这两种方法都能为 DNA 序列的进化关系提供实验依据。

(5)实际上,这两种方法所生成的分形图像的分形维是等价的。比如,运用 Hao 方法所生成的分形图像,分形维数可以直接转换为 CGR 方法所生成的分形图像的维数。

2.4　禁止子序列探讨

通过实验研究表明,在一个自然 DNA 序列中存在许多一次也没有使用的子序列,称之为禁止子序列。而且在这些子序列中,大多包含回文子序列。回文子序列是指沿 5′→3′方向读一条链与沿 5′→3′方向读其互补链,得到相同结果的子序列。实际上回文子序列长度只能是偶数,这可以简单证明。假设回文子序列

$$S=5'-s_1s_2\cdots s_i\cdots s_k-3'$$

其中 $1\leqslant i\leqslant k, s_i\in 4-Base$,且 $|S|=k$ 为奇数。当 $k=1$ 时,显然结论成立。设 $k\geqslant 3$,则存在唯一整数 $m=\lfloor k/2\rfloor+1$。设 S 的互补序列为 $S'=3'-s'_1s'_2\cdots s'_i\cdots s'_k-5'$,其中 $1\leqslant i\leqslant k, s'_i\in 4-Base$,且 s'_i 与 s_i 符合 Watson-Crick 互补性原则,所以 s'_m 与 s_m 互补,即 s'_m 与 s_m 不可能相同。因此当 k 为奇数时,沿 5′→3′读 S 序列与沿 5′→3′读 S' 其互补序列所得结果不可能相同。所以 k 只可能为偶数。

2 个字母的碱基序列构成的回文子序列太普遍,而且只可能是 gc、cg、at 和 ta 4 种情况,而不可能是其它。因此,定义像这样的子序列为平凡回文子序列,实际情况下可不予考虑。根据这一原理,假定回文子序列长度都不小于 4。

例如,针对 Zea Mays DNA 序列采用 Hao 方法来研究 8-长子序列的情况。如 5'-cccactag-3'在这个序列中就没有出现,其中包含了回文 5'-ctag-3'子序列,它的互补子序列为 3'-gatc-5';又如 5'-cgcagatc-3'在此 DNA 序列中也没有出现,其中包含回文 5'-gatc-3',它的互补子序列为 3'-ctag-5'。在序列中出现 22 次的子序列 5'-aaaatttt-3'和 5'-aattaatt-3',其回文子序列却长达 8 个碱基。实验表明,禁止子序列中大多包含 4 碱基回文子序列,因此可以认为 DNA 序列是在尽力避免较短回文子序列的出现,这是因为许多酶能够识别回文子序列,并把它剪断。

3　迭代函数系统分形(IFS)

实际上,混沌游戏就是一种简单迭代函数系统产生分形的方法。下面先介绍一下 IFS 产生分形吸引子的数学机理。

3.1　压缩映射

变换 $W:R^2\to R^2$ 具有形式为 $W(x)=Ax+t$,则称 W 为一个二维仿射变换。其中

$$A=\begin{pmatrix}a & b\\ c & d\end{pmatrix}, t=\begin{pmatrix}e\\ f\end{pmatrix}$$

而,$a,b,c,d,e,f\in R$

若其谱半径 $r_\sigma(A)<1$ 时,变换 W 是压缩的。例如,常用的二维仿射变换有:缩放变换 A_s 和旋转变换 A_θ,分别是

$$A_s=\begin{pmatrix}s & 0\\ 0 & s\end{pmatrix}, A_\theta=\begin{pmatrix}\cos\theta & -\sin\theta\\ \sin\theta & \cos\theta\end{pmatrix}$$

维普资讯 http://www.cqvip.com

一个同时具有缩放、旋转和平移的仿射变换可以表示为：

$$W(x)=A_s(A_\theta x+t'),t'=\begin{pmatrix}\Delta x\\ \Delta y\end{pmatrix}$$

其中 $\Delta x=e/s,\Delta y=f/y$。

显然，一个这样的复合压缩映射可以由一个四元组$(s,\theta,\Delta x,\Delta y)$决定。规定 $0\leqslant s\leqslant 1,0\leqslant\theta\leqslant 2\pi,-1\leqslant\Delta x,\Delta y\leqslant 1$。

3.2 迭代函数系统

一个迭代函数系统由一个完备的度量空间(X,d)和一个有限的压缩映射集 $W_n:X\to X$ 及其相应的压缩因子 $s_n,n=1,2,\cdots,N$ 所组成，简称为 IFS，表示为$\{X;W_n,n=1,2,\cdots N\}$，其压缩因子 $s=\max\{s_n:n=1,2,\cdots N\}$。

定义变换 $W:H(X)\to H(X)$为 $W(B)=\bigcup_{n=1}^{N}W_n(B)$，$\forall B\in H(X)$是完备空间$(H(X),(d))$上具有压缩因子 s 的压缩映射，其中 $H(X)$为由 X 的全体非空紧子集组成的空间。它的唯一不动点 $C\in H(X)$满足 $C=W(C)=\bigcup_{n=1}^{N}W_n(C)$，而且不动点可以通过迭代得到，即 $C=\lim_{n\to\infty}W^n(B),B\in H(X),C$ 就是IFS的吸引子。

单一的压缩映射所产生的分形吸引子是简单的，而一组压缩映射所组成的IFS能够产生较为复杂的分形吸引子。表1中给出了8个压缩映射变换，每一个变换都是由4个参数组成的四元组$(s,\theta,\Delta x,\Delta y)$表示。每次迭代中随机选取这8个压缩映射中的任何一个，得到如图5所示的吸引子。如果把每次迭代所选用的变换索引 $W_i,i\in\{1,2,\cdots,8\}$中的下标记录下来，则可以得到一个随机序列 $S=i_1i_2\cdots i_k\cdots i_n,(i_k\in\{1,2,\cdots,8\},1\leqslant k\leqslant n)$其中 n 表示总迭代次数。换句话说，图5中的吸引子图像可以看成是由随机序列 S 驱动得到的。

表1 8个压缩变换的参数表

映射索引	旋转角 θ	比例因子 s	平移量$(\Delta x,\Delta y)$
W_1	3.6	0.25	(0.58,0.5)
W_2	1.0	0.5	(−0.828,0.1)
W_3	3.0	0.75	(0.3,0.2)
W_4	0.8	0.2	(−0.5,0.3)
W_5	4.0	0.4	(0.8,0.6)
W_6	2.5	0.7	(−0.35,0.45)
W_7	0.5	0.9	(0.8,0.3)
W_8	3.5	0.7	(−0.7,0.4)

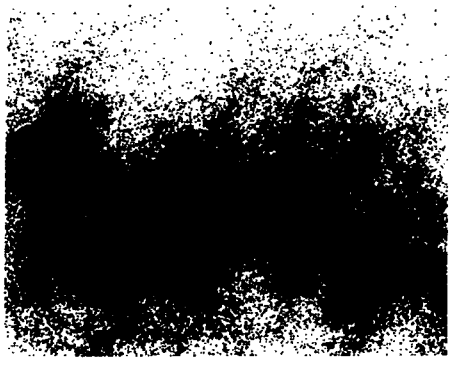


图5 根据表1中给出的参数绘制的随机序列驱动迭代函数系统产生的吸引子

可以看出，图中的点集在平面上的分布比较散。

3.3 混沌自动机(Chaos Automata，简称 CA)[1,8]

CA是根据 Moore 有限自动机来定义的。设 CA=$(S,\Sigma,\delta,s_0,W,M)$是一个六元组，其中 S 是有穷状态集合；Σ 是输入字母表；δ 是 $S\times E\mapsto S$ 的状态转移函数，表示如果 CA 处于状态 $q\in S$ 且输入字符为 $a\in\Sigma$，则 $\delta(q,a)\in S$ 是 CA 要转移到的唯一确定的状态；$s_0\in S$ 是初始状态；W 是输出符号集，每一个符号代表一个压缩映射，且 $|S|=|W|$；$f:S\mapsto W$ 是输出函数，表示当 CA 转移到新状态 $\delta(q,a)\in S$ 时，施加 $f(\delta(q,a))\in W$ 于动点，并在平面上绘制一个点。

例如：假设 $S=\{0,1,\cdots,7\}$，输入字母表 $\Sigma=4\text{-}Base$，状态转移函数 δ 如表2所示，如 $\delta(0,C)=2,\delta(0,G)=5$ 等等，初始状态 $s_0=6$，压缩变换集 $W=\{W_1,W_2,\cdots,W_8\}$就是表2中的映射索引集；输出函数是：$f(0)=W_1,f(1)=W_2$ 等等。当用 Zea Mays DNA 序列驱动这台混沌自动机工作时，产生如图6所示的吸引子。显然，由 DNA 序列驱动 CA 所得吸引子与随机序列驱动迭代函数系统所得吸引子有显著区别。在压缩映射变换参数满足 3.1 节中所给约束的情况下，适当调整参数可以生成不同的分形吸引子。但对同一 CA，不同的自然 DNA 序列(甚至是随机生成的 DNA 序列)所生成的吸引子差异比较小，从视觉上还不能区分。那么，如何从 IFS 吸引子来区分不同的 DNA 序列呢？这需要进一步研究。至于说，由 DNA 序列驱动 CA 所生成的分形吸引子的生物学含义，目前还不能确切地对其进行解释。

表2 8个状态的混沌自动机状态转移表

状态	C	G	A	T	映射索引
0	2	5	6	0	W_1
1	5	7	1	2	W_2
2	2	1	2	2	W_3
3	4	4	4	7	W_4
4	3	3	3	3	W_5
5	2	5	7	1	W_6
6	1	0	2	1	W_7
7	1	4	5	2	W_8



图6 Zea Mays DNA 序列驱动 CA 得到的吸引子图像

3.4 CA 分形实现算法

下面给出 DNA 序列驱动 CA 实现分形的算法。

CurrentState:=s0; //指定当前状态为初始状态

$x:=0;y:=0$; //指定初始当前动点(x,y)为(0,0)

Ch:=Read(File); //串行读取 DNA 序列文件中第一个碱基

While not eof(File)do //还有未读 DNA 序列

Begin

CurrentState:=δ(*CurrentState*,*ch*);

//根据当前状态和读取的碱基 ch 按照状态转移函数指定新的当前状态；

$x_{new}:=s*(x*\cos(\theta)-y*\sin(\theta)+\Delta x$

$y_{new}:=s*(x*\sin(\theta)+y*\cos(\theta)+\Delta y)$

维普资讯 http://www.cqvip.com

//根据与当前状态相关的压缩映射以及当前动点计算新的当前动点；

Plot(X_{new},Y_{new})；//绘制当前动点；

x：=X_{new}；y：=Y_{new}；//为下一步迭代做准备；

Ch：=read(file)；//在文件中读取下一个碱基；

End;

4 三联密码子(codon)序列分形

按照生物学中心法则，DNA 序列转录为核糖核苷酸序列 mRNA，它进一步被翻译成各种蛋白质而行使生物体的生命调控功能。在基因表达过程中，序列中每三个碱基翻译为一个氨基酸。不同的氨基酸共有 20 种。为了便于研究 DNA 序列分形结构，暂不考虑太多生物学意义而直接按 DNA 序列顺序。每三个碱基依次映射为三联密码子序列，即氨基酸序列(忽略停止密码子等其它结构)。

建立 DNA 序列三联密码子映射：

translation：*DNA* 序列↦氨基酸序列

也就是

{*ccc*,*ccg*,*cca*,*cct*,*cgc*,*cgg*,*cga*,*cgt*,*ctc*,*ctg*,*cta*,*ctt*,*cac*,*cag*,*caa*,*cat*,*gcc*,*gcg*,*gca*,*gct*,*ggc*,*ggg*,*gga*,*ggt*,*gtc*,*gtg*,*gta*,*gtt*,*gac*,*gag*,*gaa*,*gat*,*acc*,*acg*,*aca*,*act*,*agc*,*agg*,*aga*,*agt*,*atc*,*atg*,*ata*,*att*,*aac*,*aag*,*aaa*,*aat*,*tcc*,*tcg*,*tca*,*tct*,*tgc*,*tgg*,*tga*,*tgt*,*ttc*,*ttg*,*tta*,*ttt*,*tac*,*tag*,*taa*,*tat*}↦{2,6,6,6,15,15,15,15,2,2,2,9,10,10,9,17,17,17,17,20,20,20,20,16,16,16,16,18,19,19,18,7,7,7,7,5,15,15,5,3,4,3,3,11,12,12,11,5,5,5,5,13,14,0,13,1,2,2,1,8,0,0,8}。其中每一个编号(1～20)代表一个氨基酸，0 表示停止密码子。

如何建立氨基酸序列的分形图像，的确是一件困难的事情。例如，一种方法是采用 20 角混沌游戏(20-*CGR*)。实验结果表明，这种方法不能得到分形图像，动点将以一定概率充满 20 边形。另一种方法是把 20 种氨基酸分为 4 大类，每类对应 4-*CGR* 的一个角点，运用 4-*CGR* 方法可以得到氨基酸序列的分形图像，但图像中的像素点的意义不明确，图像特征不能很好地反映子序列结构。

如果从 20 种氨基酸中任选 4 种，采用 4-*CGR* 方法仍可以得到相应的分形图像，但供选择的方法多达 $C_{20}^{4}=4845$ 种。这种方法的优点是其简单性，但损失了许多其它未考虑的氨基酸子序列的结构信息。具体方法是：任选 4 种氨基酸{A_i}，$i\in\{1,2,3,4\}$，把其它氨基酸从序列中删除，得到一条只由这 4 种氨基酸组成的序列。设 $C(A_i)$表示 A_i，$i\in\{1,2,3,4\}$在氨基酸序列中出现的次数，则这 4 种氨基酸在序列中出现的总数 total=$\sum_{i=1}^{4}C(A_i)$，A_i 在序列中出现的概率 $P(A_i)=C(A_i)$/total，$i\in\{1,2,3,4\}$。显然，$\sum_{i=1}^{4}P(A_i)=1$，且 $0\leqslant P(A_i)\leqslant 1$，$i\in\{1,2,3,4\}$然后以下面两种方法分别获得分形图像作为对比：

(1)以任意选定的 4 种氨基酸所组成的氨基酸序列驱动 4-CGR 得到一种分形图像；

(2)以概率 $P(A_i)$，$i\in(1,2,3,4\}$、迭代总数为 Total 的随机序列驱动 4-CGR，得到另一种随机分形图像。实际上，这是带概率的迭代函数系统

{X;W_1,W_2,W_3,W_4,$P(A_1)$,$P(A_2)$,$P(A_3)$,$P(A_3)$}所产生的分形吸引子。其中

$W_1(x,y)=(0.5x,0.5y)$

$W_2(x,y)=(0.5x+0.5,0.5y)$

$W_3(x,y)=(0.5x,0.5,0.5y+0.5)$

$W_4(x,y)=(0.5x,0.5y+0.5)$

通过实验发现，氨基酸序列的随机程度相对较大，但仍存在明显的分形结构特征。

例 1 在 translation(Zea Mays DNA 序列)序列中选择 4 种出现频率差别较大的氨基酸{5,10,15,20}，其出现频数依次分别是{5771,2160,5619,3266}。图 7 中左图是采用第一种方法绘制的分形图像，右图是采用第二种方法绘制的分形图像。对比这两幅图像可明显看出，氨基酸使用偏性也产生了明显的序列分形结构特征。

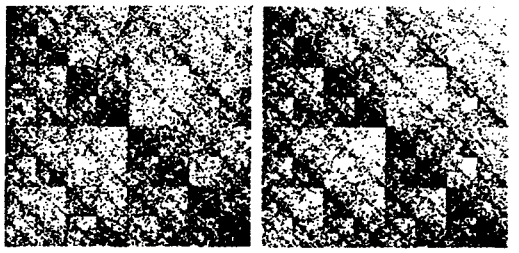


图 7 左图是 4 种氨基酸{5,10,15,20}组成的序列驱动 4-CGR 绘制的分形图像；右图是相应的随机序列驱动 4-CGR 绘制的图像

例 2 在 translation(Zea Mays DNA 序列)序列中选择 4 种出现频率差别较小的氨基酸{3,6,7,20}，其出现频数依次分别是{3447,3154,3770,3266}。采用类似方法绘制的分形图像如图 8 所示。对比这两幅图像可明显看出，左图中依然存在着某种结构(特别是对角线)，而右图中没有明显的结构。左图中除对角线外的其余图像与右图对应部分图像的区别不是很明显。左图中的分形结构特征表明，虽然从总体上说序列中没有明显的氨基酸使用偏性，但是存在许多特异的氨基酸子序列。

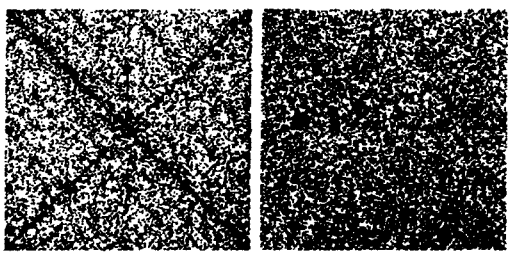


图 8 左图是 4 种氨基酸{3,6,7,20}组成的序列驱动 4-CGR 绘制的分形图像；右图是相应的随机序列驱动 4-CGR 绘制的图像

5 进一步的研究工作

本文主要综述研究了 DNA 序列分形图像的表示方法，还未对 DNA 序列分形图像做定量的深入研究，而且实验也还不够充分。因此，进一步要做的研究工作还很多。比如，采用分形方法定量研究 DNA 序列分形图像，可获得像分形维数等一些数字特征，并进一步研究这些数字特征与序列结构的关系；CGR 所产生的分形图像中，高频出现的子序列与低频出现的子序列代表什么含义？禁止子序列有什么规律？混沌自动机 CA 中 IFS 参数如何确定？是否能通过确定一组 IFS 参数使得不同的 DNA 序列产生不同的分形图像？是否能通过 IFS 分形图像来区分不同的 DNA 序列？在序列驱动

维普资讯 http://www.cqvip.com

CA 产生分形图像的过程中如何利用已知的 DNA 序列结构信息？更好的 DNA 三联密码子序列分形图像表示方法是如何实现的？如果把 DNA 子序列所出现的频率也看成一维，这样二维分形图像就可以转化成三维图像，从而能研究不同子序列出现频率的分布情况，这些都是下一步要做的研究工作。

结束语 基因组是由四种不同的脱氧核糖核苷酸按照特定的编码规则串联成的一维线性序列，其中蕴藏着生物体中绝大部分的结构信息和控制信息。因此，可以说基因组是生物体内的信息和控制中心。对 DNA 序列结构规律的探究是至关重要的研究工作。但面对海量的 DNA 序列数据，传统的统计研究方法还不能很好地满足 DNA 序列研究的需要；序列可视化研究无疑是一种能够揭示更多隐藏信息的数据表示方法。因为人类获得外在世界信息的 80%是通过视觉通道获得的，可视化依然继续着借助形象化表达人类意图的传统。

参 考 文 献

1 Ashlock D, Golden J B. Iterated function system fractals for the detection and display of DNA reading frame. In: Proceedings of the 2000 Congress on Evolutionary Computation, 2000
2 Back T, Hammel U, Schwefel H-P. Evolutionary computation: Comments on the history and currents state. IEEE Transactions on Evolutionary Computation, 1997, 1(1): 3～17
3 Chellapilla K, Czarnecki D. A preliminary investigation into evolving modular finite state machines. In: Proceedings of the 1999 Congress on Evolutionary Computation, 1999. 1349～1356
4 Nandy A. Graphical analysis of DNA sequence structure: III. Indications of evolutionary distinctions and characteristics of introns and exons. Curr. Sci, 1996(70): 611～668
5 Leong P M, et al. Morgenthaler. Random walk and gap plots of DNA sequences. Computer appl Biosciences, 1995, 11(5): 503～507
6 National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health. http://www.ncbi.nlm.gov/; NCBI Gen Bank. http://www.ncbi.nlm.nih.gov/Genbank/; NCBI Genomes, http://www.ncbi.nlm.nih.gov/Genomes/
7 Berger J A, Mitra S K, et al. Visualization and analysis of DNA sequences using DNA walks. Journal of the Franklin Institute, 2004, 341: 37～53
8 Ashlock D, Golden J. Chaos automata: iterated function systems with memory. Physica D, 2003, 181: 274～285
9 Koza J R. Genetic Programming. Cambridge, MA: The MIT Press, 1992
10 Jeffrey H J. Chaos game representation of gene structure. Nucleic Acid Research, 1990, 18(8): 2163～2170
11 Hao Bai-Lin. Fractals from genomes- exact solutions of a biology-inspired problem. Physica A, 2000, 282: 225～246
12 Hao Bai-lin, Lee H C, Zhang Shu-yu. Fractals related to long DNA sequences and Complete genomes. Chaos, Solitons and Fractals, 2000, 11: 825～836

(上接第 154 页)

图为提取的彩色水印图像，效果清晰，易于识别。

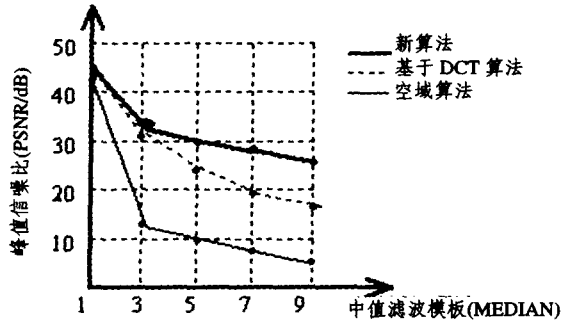


图 5 不同滤波情况下提取水印图像的 PSNE 值比较

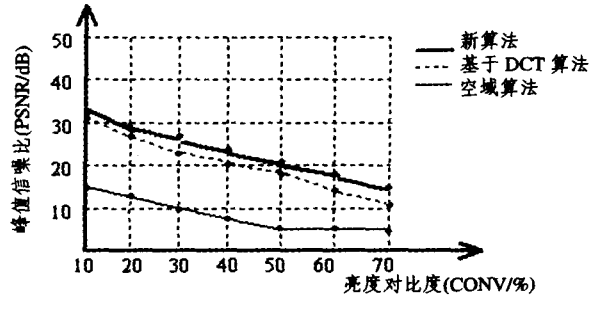


图 6 亮度、对比度调整后提取水印图像的 PSNR 值

图 5 给出了图像在进行不同程度的滤波情况下，各个算法提取出的水印图像 PSNR 值结果。图 6 给出了图像在进行不同程度的亮度对比度变换情况下，各个算法提取出的水印图像 PSNR 值结果。从图 5 可以看出，当中值滤波的模板稍微增大的时候，空域算法的效果就明显下降，可见其在这方面的健壮性不好，而新算法提取出水印图像质量虽有所下降，但总的效果还是比较好的。从图 6 可以看出，亮度、对比度的调整对空域算法的影响最大，而新算法在一定程度上能够抵抗这种变化对提取水印图像质量的影响。

结论 互联网技术及电子商务、电子政务技术的高速发展，多媒体数字产品的版权保护日益引起人们的重视。对于该问题的解决，数字水印技术是一种很好的思路。小波分析所具有的多分辨分析能力使其广泛地用于图像处理领域。

大多数字水印算法均是以灰度图像或者随即序列作为嵌入对象，而面向彩色图像的水印算法很少。由于彩色水印容易辨别，因此将其作为嵌入对象具有极强的实际意义。根据人视觉的多分辨率特性，本文提出了一种新的数字水印算法，新算法吸收了零树小波编码得重要思想，将水印信息嵌入到重要系数中，实验表明采用新算法既能保证水印的鲁棒性，而且在看压缩、抗裁减、抗滤波方面均表现出良好的特性。

参 考 文 献

1 Schyndel R G Van, Tirkel A Z. A digital watermark. In: Proc. IEEE Int. Conf, on Image Processing, Piscataway: IEEE Press, 1994, 2: 86～89
2 Wolfgang R B, Delp E J. A watermark for digital image. In: Proc. IEEE Int. Conf. on Image Processing, Piscataway: IEEE Press, 1996, 3: 219～223
3 Nikoladis N, Pitas I. Robust image watermarking in the spatial domain. Signal Processing, 1998, 66(3): 386～405
4 Hsu C T, Wu J L. Hidden signayure in images. IEEE Trans. On image Processing, 2001, 9(2): 58～65
5 Ruanaidh J K O, Pun T. Rotation, scale and translation invariant spread spectrum digital image watermark. Signal Processing, 2002, 86(4): 303～315
6 Shaprio J M. Embedded image coding using zerotree of wavelet coefficients [J]. IEEE Trans on Signal Processing, 1993, 41(12): 3445～3461