AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO201: Intelligence of Biological Systems - 1

Chaos Game Representation

Dr. Manjusha Nair M
Amrita School of Computing, Amritapuri

Email : manjushanair@am.amrita.edu
Contact No: 9447745519

Contents

- Chaos Theory
- Butterfly Effect
- Fractals
- Chaos Game Representation (CGR)
- Sierpinski triangle
- Square Shaped CGR
- CGR for Biological sequences
- GAATT in square shaped CGR
- CGR of HUMHBB
- How to Understand the CGR of HUMHBB?
- CGR of chloroplast of quinoa plant

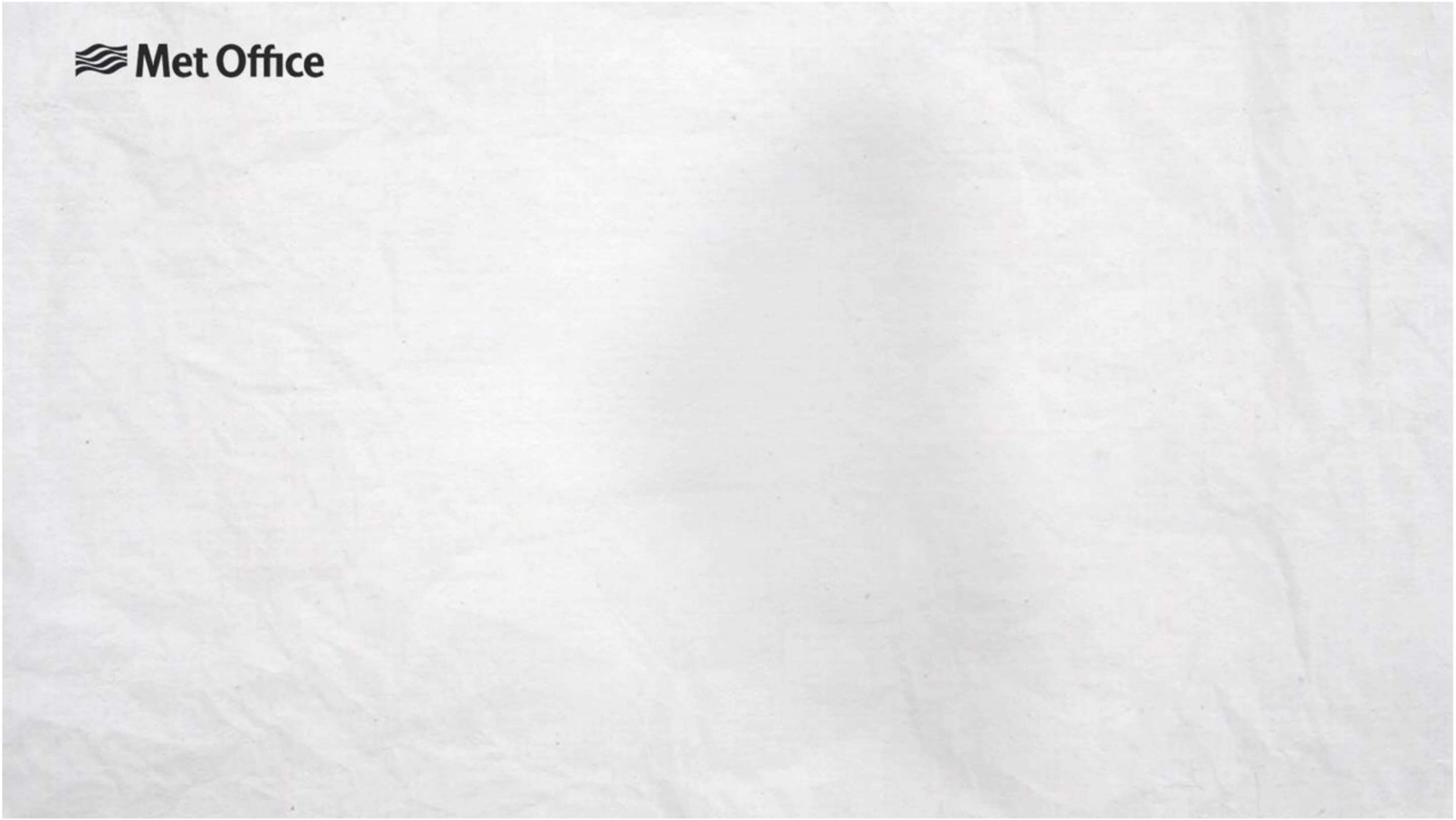
Chaos Theory

- What is Chaos?
  - It means "a state of disorder"
- Chaos is not simply disorder.
  - Chaos explores the transitions between order and disorder, which often occur in surprising ways.
  - Chaotic systems never repeat, but they always have some order.
- Everything in the universe is under control of Chaos or product of chaos

Chaos Theory

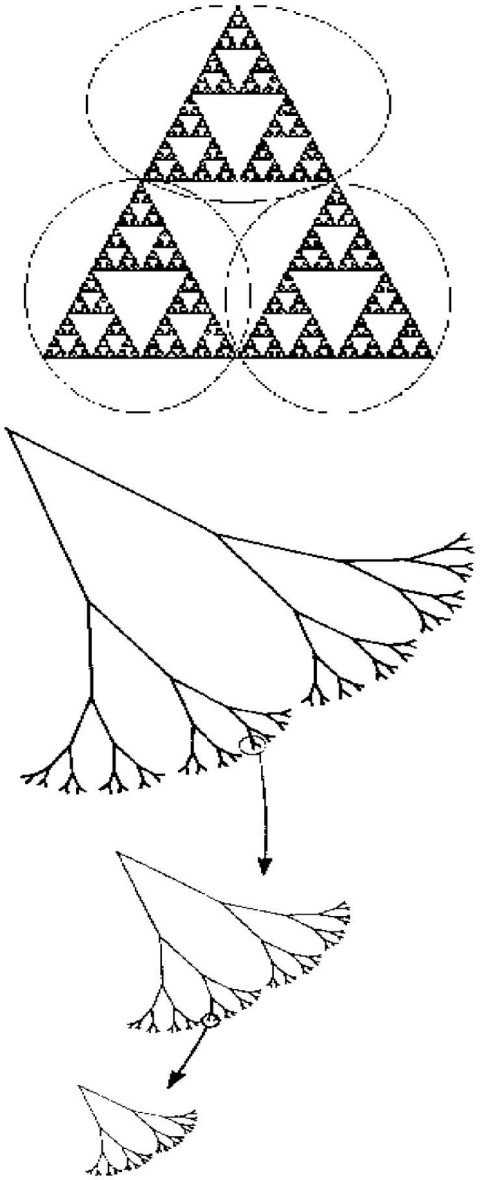
- Chaos theory is the study of non linear dynamical systems that are highly sensitive to initial conditions
- Connected with unpredictable courses of events
- Butterfly effect by Edward Lorenz
  - The basic principle is that even in an entirely deterministic system, the slightest change in the initial data can cause abrupt and random changes in the outcome .
  - Large systems remain impossible to predict with total accuracy because there are too many unknown variables to track.

Butterfly Effect



Fractals

- Driven by recursion, fractals are images of dynamic systems – the pictures of Chaos.
- Fractals are infinitely complex patterns that are self-similar across different scales.
- They are created by repeating a simple process over and over in an ongoing feedback loop.
- The random iteration algorithm for creating pictures of fractals uses iterated function system (IFS)
- Two Key concepts
  - Recursion
  - Self Similarity

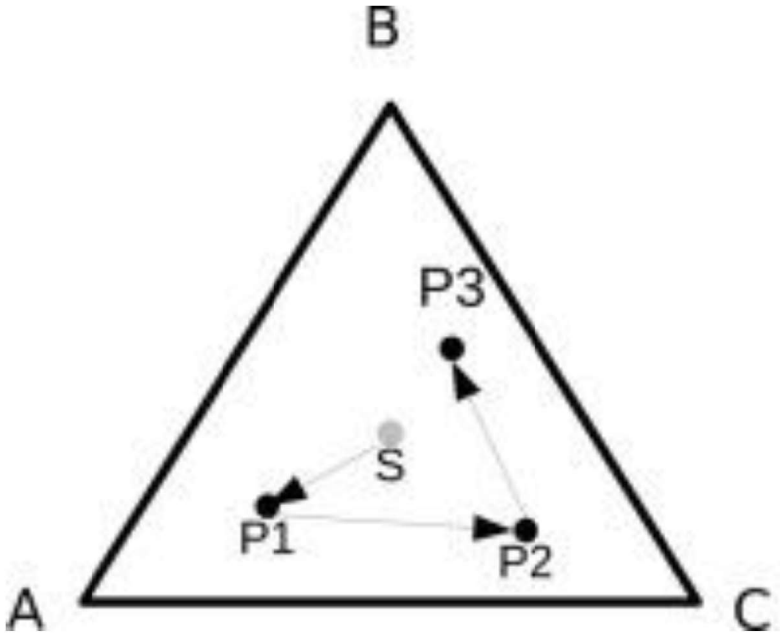


Chaos Game Representation (CGR)

- Chaos game representation (CGR) is an iterative mapping technique derived from chaos theory
- The Chaos Game is an algorithm which allows one to produce pictures of fractal structures
- It visualizes a one-dimensional sequence in a two-dimensional space.
- It was originally developed to construct the Sierpinski triangle

Sierpinski triangle

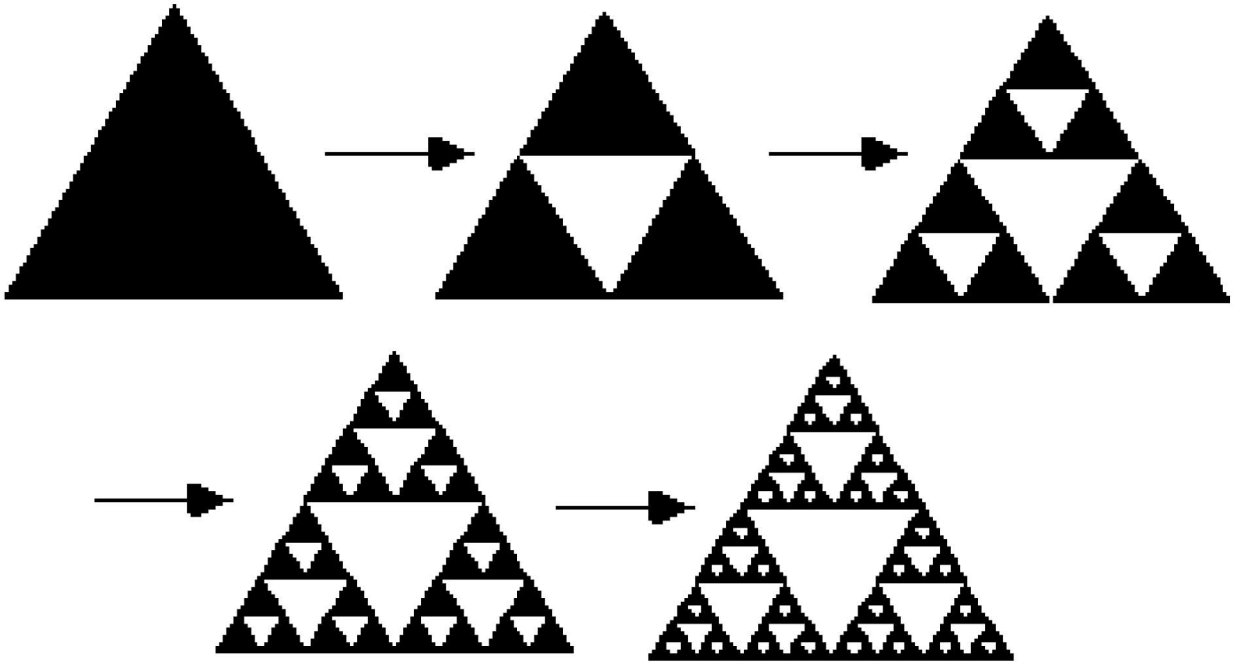
- Numbers from one to three are assigned on the vertices of a triangle.
- Based on a randomly selected start point (S-Seed), a vertex is randomly chosen (V1), and a point P1 is drawn in half the distance to the vertex V1.
- This process is repeated, with P1 as the new starting point.
- The second point (P2) is drawn at half the way to the second randomly selected vertex (V2).

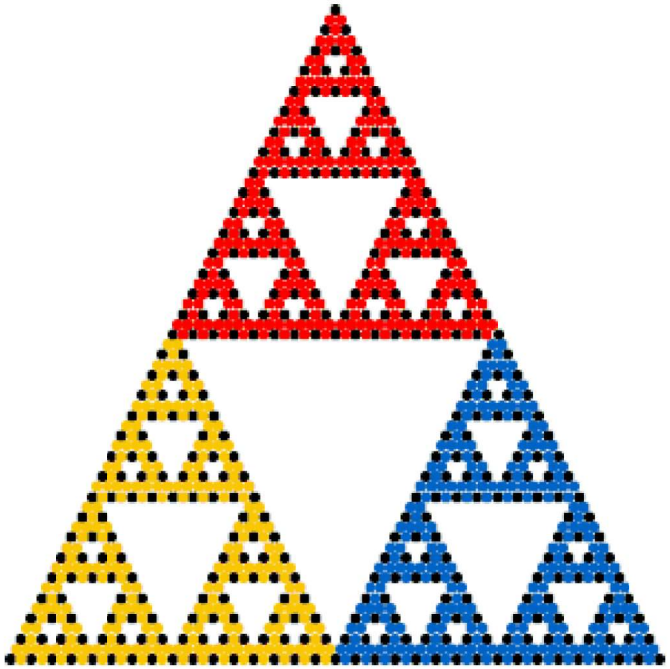


Sierpinski triangle

- Begin to record the track of these traveling points after each roll of the die.
- One might expect that this procedure, if repeated many times, would yield
  - a paper covered with random dots or,
  - a triangle filled with random dots.
- The points form what mathematicians call the Sierpinski triangle
- This figure results no matter what seed is used to begin the game
- Chaos game produces a figure with visible patterns.
- The formation of a complex pattern often has a very simple process responsible for it.
  - These complex patterns are called fractals.

Sierpinski triangle





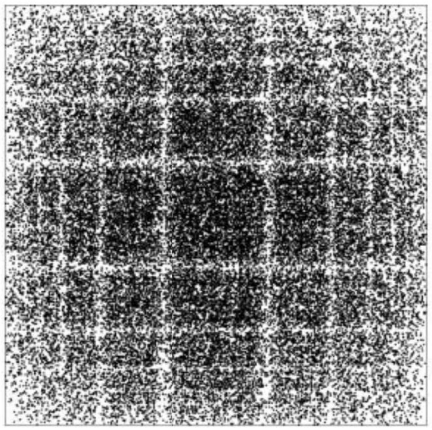
- Play It: http://thewessens.net/ClassroomApps/Main/chaosgame.html

Sierpinski triangle - Dividing rate , r

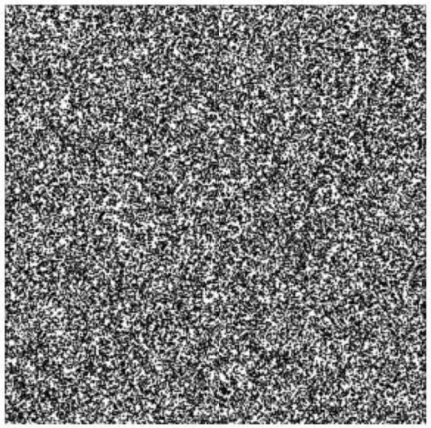
- The new point can be placed anywhere within the line segment created by the two points of reference (or even outside)
  - When the new point is placed halfway between the current point and the vertex, the **dividing rate** is r = 0.5
  - If the new point is placed closer to the location of the current point, then the dividing rate r < 0.5,
  - If the point is closer to the vertex, then the dividing rate r > 0.5

Square Shaped CGR

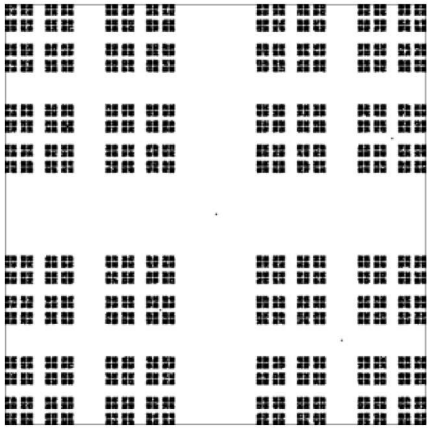
- Square shaped CGR did not exhibit any fractal patterns when r = 0.5



r = 0.4



r = 0.5



r = 0.6

CGR for Biological sequences

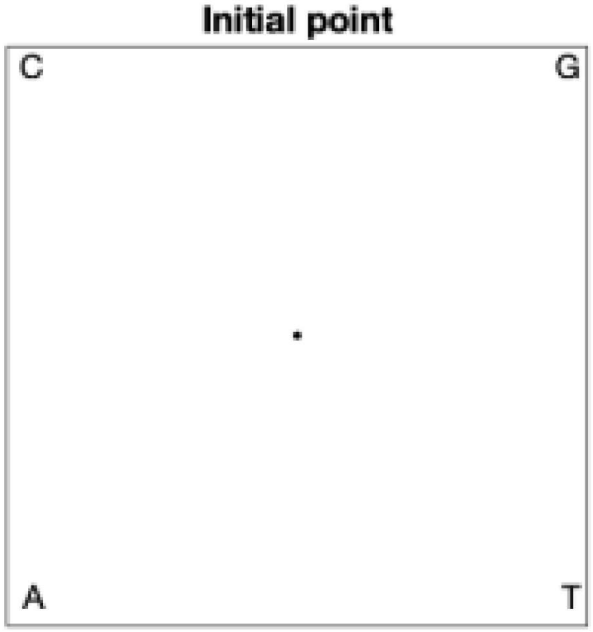
- It is possible to use the chaos game algorithm to represent any well-defined sequence
- Concept of the CGR algorithm is to map
  - a sequence, i.e., a 1D representation to a higher dimensional space, typically to the 2D space
- for a sequence with a number N of distinct elements, it is possible to play the chaos game on an N-sided polygon
  - assigning each element to a vertex
  - and playing the game choosing the vertices following the progression of the sequence (instead of choosing a random vertex)

CGR for Biological sequences

- Let N = sides of the polygon and r = the ratio between two distances (ratio between the length of the side of the first subscale polygon and the side of the original polygon),
- For Sierpinski triangle, N = 3, r = 1/2
- To represent a genomic sequence (DNA /Gene), N = 4, r = 1/2
- To represent amino acid sequences of Proteins, N=20, r = 0.8

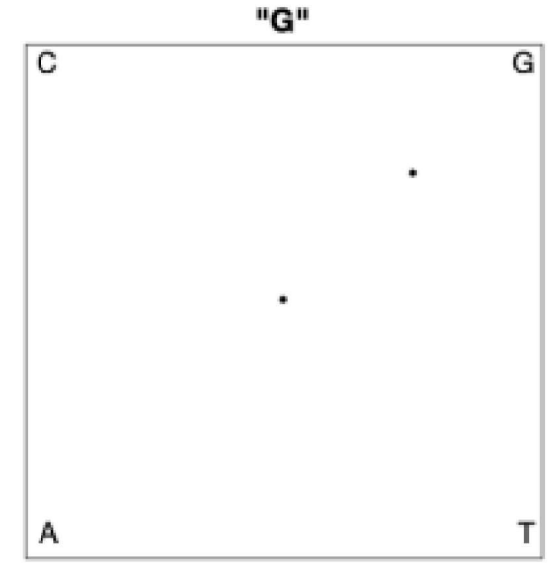
CGR of genetic sequences (DNA/RNA)

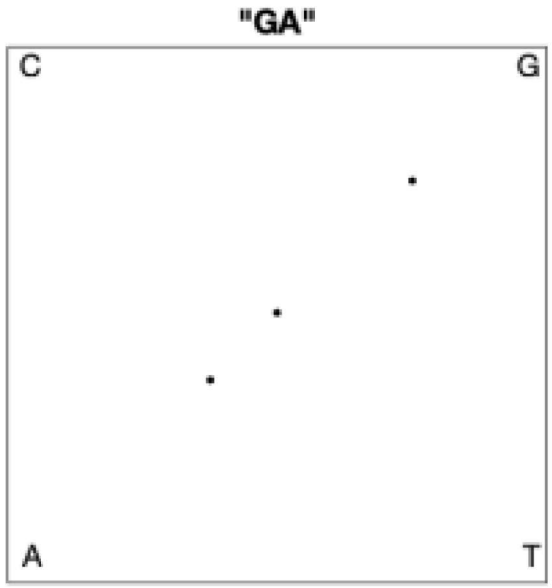
- Use a square-shaped CGR
- Mark the center as initial point.
- Label the four corners with the name of each base.
- A is in the bottom-left corner
- C in the top left
- G in the top right,
- and U/T in the bottom right.



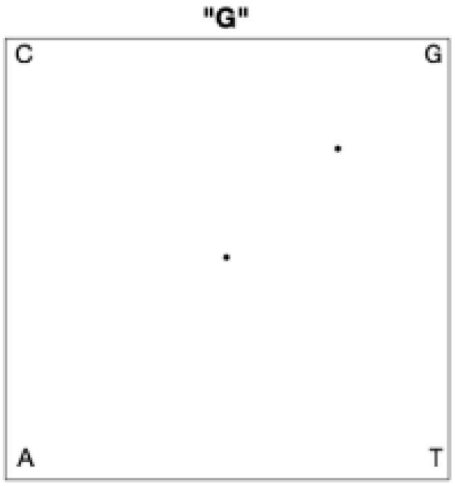
GAATT in square shaped CGR

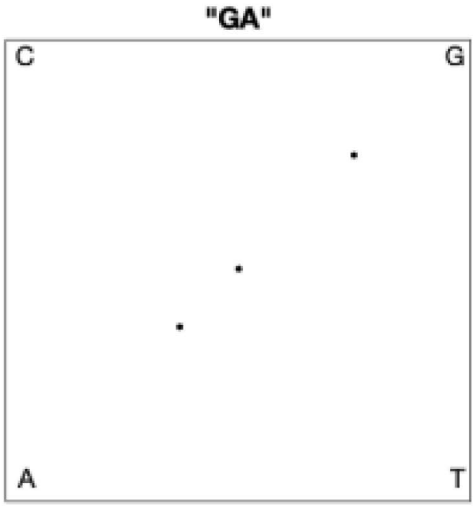
- The first base in the sequence is ‘G’, so plot a point halfway between our initial point and the G corner.
- The next base in the sequence is ‘A’, so plot a point halfway between the point that we just plotted and the ‘A’ corner.

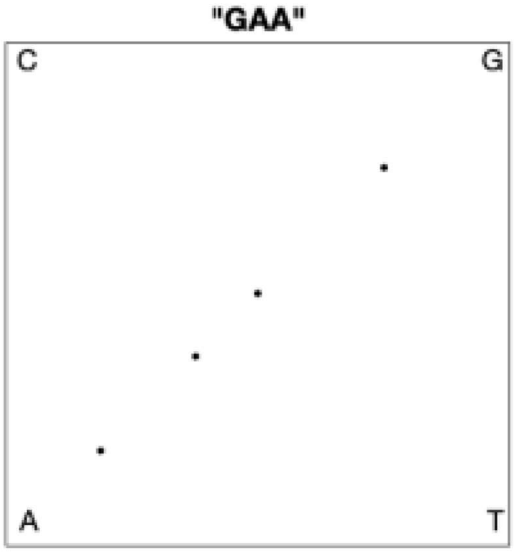


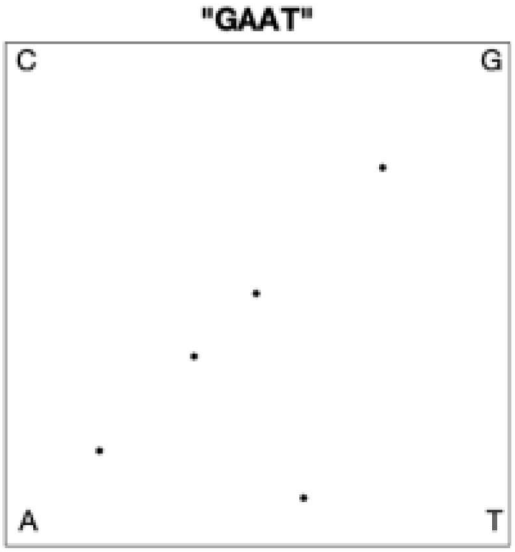


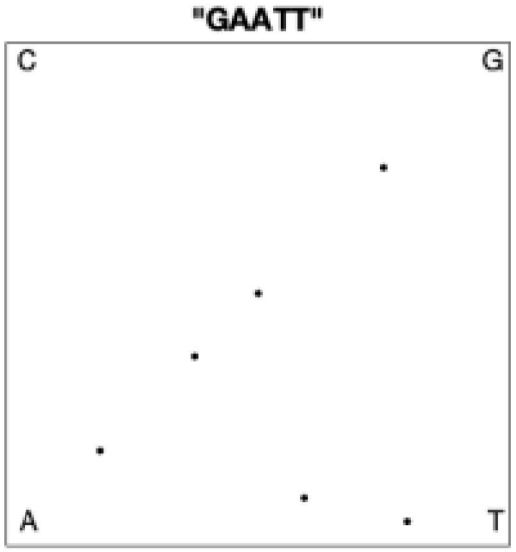
GAATT in square shaped CGR





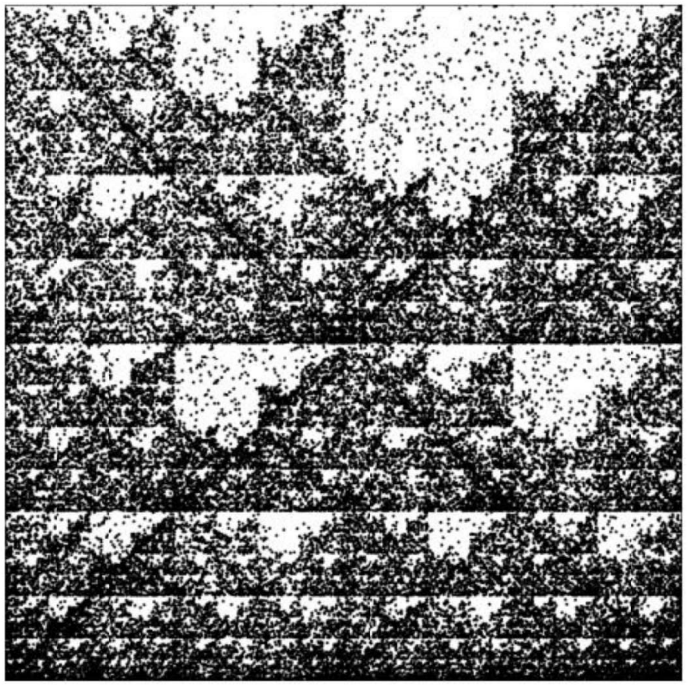






CGR of HUMHBB

- GAATT is the first five bases of the DNA sequence HUMHBB (human beta globin region on chromosome 11)



GSR of DNA of (HUMHBB)---73308 bp

CGR - Another Detailed Example

- Let us consider a given DNA sequence composed of N nucleotides S={ S1,S2,....,S_N }.
- Usually, the starting point x_0 is placed at the center of the square while the choice of the corners is arbitrary and can be assigned in any other way.

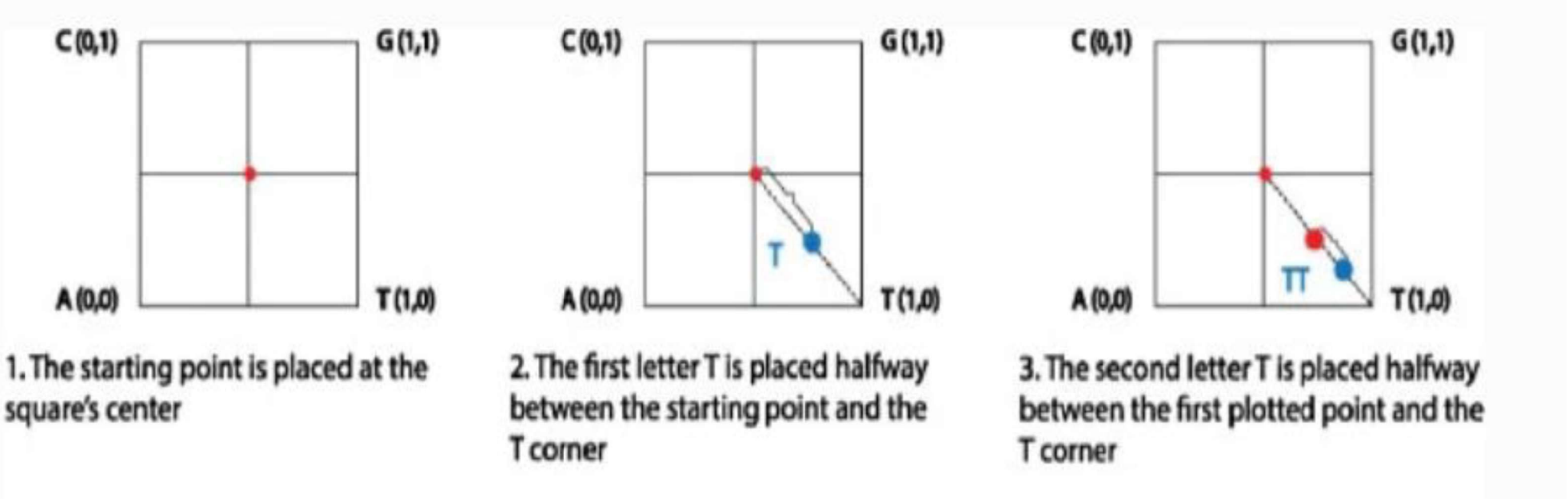
$$\begin{cases} x_0 = (0.5, 0.5) \\ x_i = x_{i-1} + \frac{1}{2}(y_i - x_{i-1}), i = 1, \ldots, N \end{cases}$$

where $y_i =$

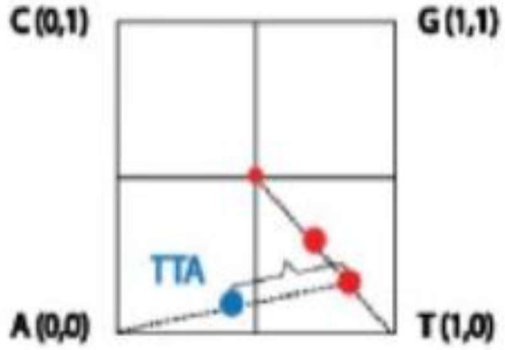
$$\begin{cases} (0,0) & \text{if} \quad S[i] = \text{A} \\ (0,1) & \text{if} \quad S[i] = \text{C} \\ (1,0) & \text{if} \quad S[i] = \text{T} \\ (1,1) & \text{if} \quad S[i] = \text{G} \end{cases}$$

CGR - Another Detailed Example

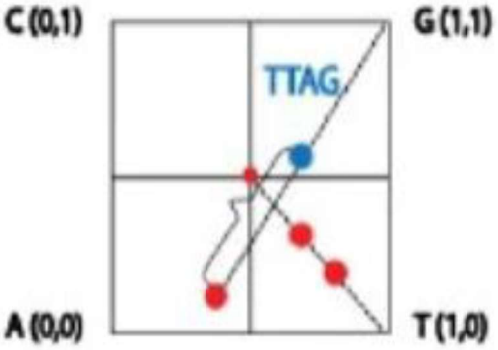
- The figure given below shows the procedure to draw the sequence 'TTAGC'.



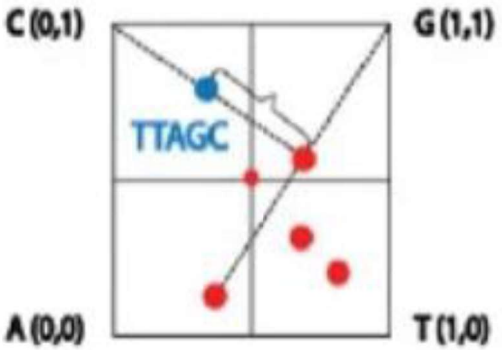
CGR - Another Detailed Example



4. The third letter A is placed halfway between the second plotted point and the point A corner



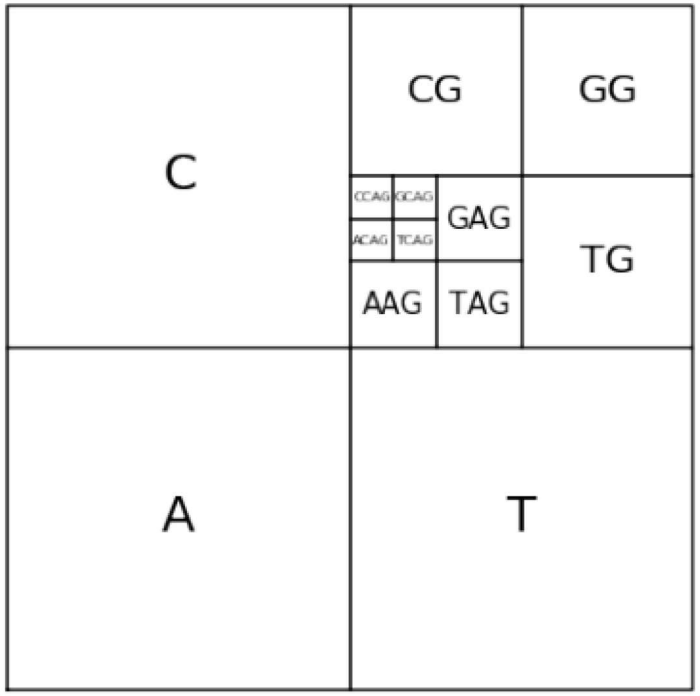
5. The fourth letter G is placed halfway between the third plotted point and the G corner



6. The final letter C is placed halfway between the previous plotted point and the C corner

How to Understand the CGR ?

- Each point plotted in the CGR corresponds to a base.
- Depending on where it is placed, we can trace back and figure out parts of the sequence we are examining.
- Any point that corresponds with base G will be located in the upper right quadrant of the CGR plot.
- To see what the previous base is, divide the quadrant into sub-quadrants. depending on where the point is, determine the previous base of the sequence.



How to Understand the CGR ?

K=1

C	G
A	T

K=2

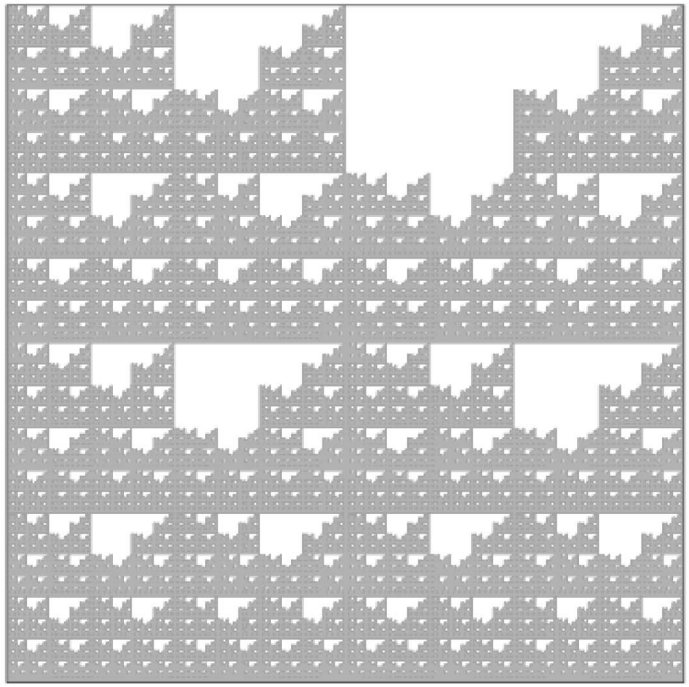
CC	CG	GC	GG
CA	CT	GA	GT
AC	AG	TC	TG
AA	AT	TA	TT

K=3

CCC	CCG	CGC	CGG	GCC	GCG	GGC	GGG
CCA	CCT	CGA	CGT	GCA	GCT	GGA	GGT
CAC	CAG	CTC	CTG	GAC	GAG	GTC	GTG
CAA	CAT	CTA	CTT	GAA	GAT	GTA	GTT
ACC	ACG	AGC	AGG	TCC	TCG	TGC	TGG
ACA	ACT	AGA	AGT	TCA	TCT	TGA	TGT
AAC	AAG	ATC	ATG	TAC	TAG	TTC	TTG
AAA	AAT	ATA	ATT	TAA	TAT	TTA	TTT

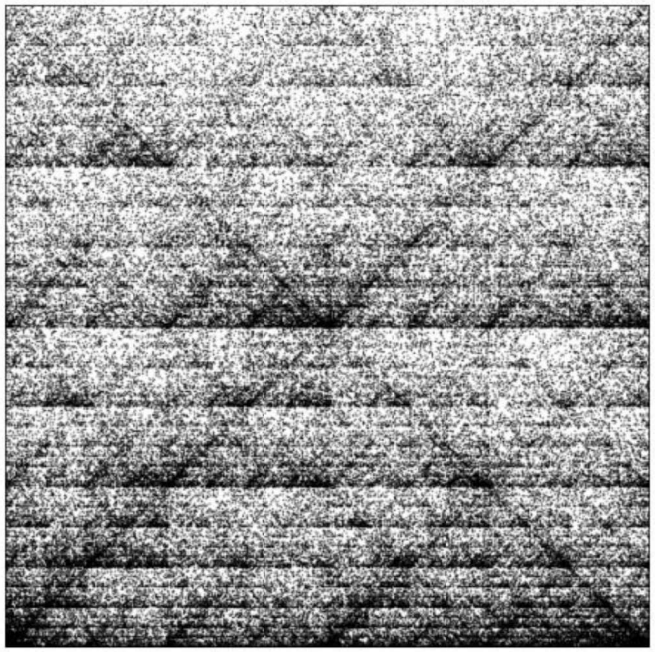
How to Understand the CGR of HUMHBB?

- We can repeat this step again and again to find the order in which the bases appear in the sequence.
- Figure shows a CGR square where all the CG quadrants are unfilled.
- **Double scoop Pattern**
  - The most prominent Fractal pattern here, which appears in almost all vertebrate DNA sequences.
  - This pattern is due to the fact that there is a comparative sparseness of guanine following cytosine in the gene sequence since CG dinucleotides are prone to methylation and subsequently mutation.



CGR of chloroplast of quinoa plant

- CGR of the DNA sequence of the chloroplast of quinoa plant and does not exhibit a double scoop pattern.
- By using the CGR of DNA sequences, we are able to distinguish between different species



DNA of Chenopodium quinoa cultivar Real Blanca chloroplast, complete sequence, whole genome shotgun sequence---152282 bp

AMRITA VISHWA VIDYAPEETHAM
DEEMED TO BE UNIVERSITY

22BIO201: Intelligence of Biological Systems - 1