Chaos Game Representation (CGR) or Fractal Visualisation of DNA or RNA

Mathematically, the CGR is described by an Iterated Function System (IFS). In general, IFS represents a system that transforms a sequence of symbols into a unique set of points in the 2-dimensional space. In particular, IFS transforms DNA sequences into the 2-dimensional fractal. An important characteristic of the representation space is that there are so-called attractor points in the space, e.g. in the corners of the unit square, representing subsequences AAAA…, CCCC…, and so on. The set of equations (1) shows the four IFS transformations in the rectangular coordinate space to be aplied to successive bases of the DNA (Tino 1999):

$$\begin{aligned} \omega_T(x, y) &= (0.5x + 0.5,\ 0.5y) \\ \omega_A(x, y) &= (0.5x,\ 0.5y + 0.5) \\ \omega_G(x, y) &= (0.5x,\ 0.5y) \\ \omega_C(x, y) &= (0.5x + 0.5,\ 0.5y + 0.5) \end{aligned} \tag{1}$$

Where x and y are coordinates in a rectangular unit square. Iteration can start in arbitrary point and a corresponding transformation is applied for each successive base in the DNA sequence to the previous point x, y. A limit set of points emerging from an infinite application of the IFS is called an IFS attractor. Fig. 1 illustrates the IFS principle.

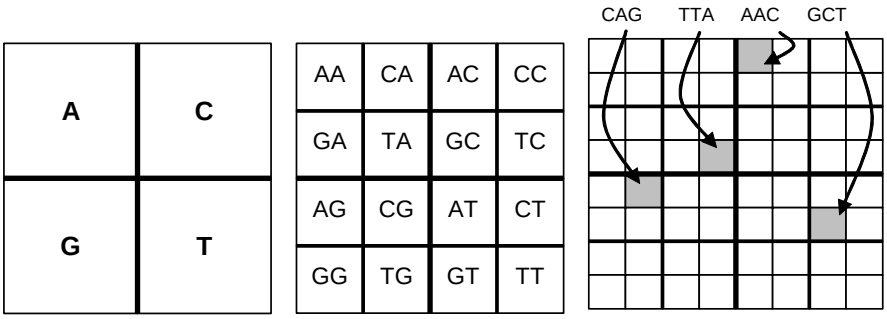

Figure 1. Correspondence between the final position of points in the IFS unit space and the suffix of the processed DNA sequence.

It can be easily shown that each transformation contracts coordinates to quarter of the original unit square. This means there is a unique relationship between the actual position in the unit space and suffix of sequence processed by the IFS. Thus, in Figure 1a, 'C' is plotted in the C-quadrant. Then 'A' is plotted in the upper right of the A-quadrant, or what might be called the 'CA' sub-quadrant (Fig. 1b). Thus, 'A' produces a copy of the C-quadrant that is one-half the size (side length) of the C-quadrant, or one-fourth of the size of the entire picture. The next letter 'G' then produces a one-half size copy of the 'CA' sub-quadrant, the 'CAG' sub-sub-quadrant, within the 'AG' sub-quadrant (Fig. 1c), etc.

In addition, due to the fact that a base is plotted in its quadrant, the converse holds as well: if two points are within the same quadrant, they correspond to sequences with the same last base; if they are in the same sub-quadrant, the sequences have the same last two bases; if they are in the same sub-sub-quadrant they have the same last three bases, etc. For example in Fig. 2 random and biological sequences are visualised. In the genome visualisation sparse and dense areas correspond to rare and frequent subsequences, respectively.

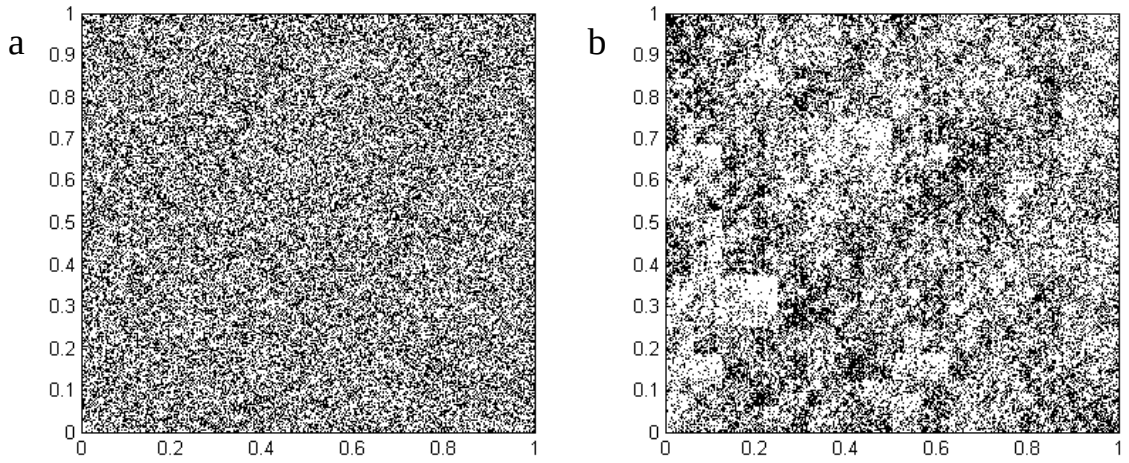

Figure 2. Chaos game visualisation using Eqs. (1): (a) the first 20 000 letters of random four-symbol sequence; (b) the first 20 000 bases of E. coli (the most common intestine bacterium) strain K12 genome. Dense and sparse areas in the right visualisation indicate a non-uniform base distribution in E. coli genome (the total count of bases is 4.639 Mbp (megabasepairs)).

The question of when two points close in the CGR represent similar sequences is a bit complicated. In general, two close points may correspond to different sequences. However, this situation can only occur if the two points, although close, are in different quadrants of the picture. Since a base is always plotted in its quadrant, any sequence will always be plotted somewhere in the quadrant of its last base, and conversely any two points in the same quadrant must have the same last base. Further, the notion of quadrant is recursive; each quadrant can be divided into sub-quadrants, etc. Thus, from the biological point of view, similar oligomers, i.e. small fraction of DNA sequence of fixed size K, are situated spatially close to each other in the visualised quadrant as a result of suffix similarity.

Limited precision while rendering image on a computer inevitably causes loss of information. Especially regularities in the form of self-similar subsequences in genomic data result in a lot of overlapping points in visualisation. In addition, since in CGR the number of plotted points always equals to the number of bases, very long sequences can potentially fill out most of the plane and no fine details would be resolvable. Straightforward solution to this problem is to divide visualisation space into smaller quadrants, count occurrences of points in each quadrant and visualise their density either in greyscale or colour, as it was suggested in (Hao et al. 2000, Shen et al. 2004, Makula & Benuskova 2009).

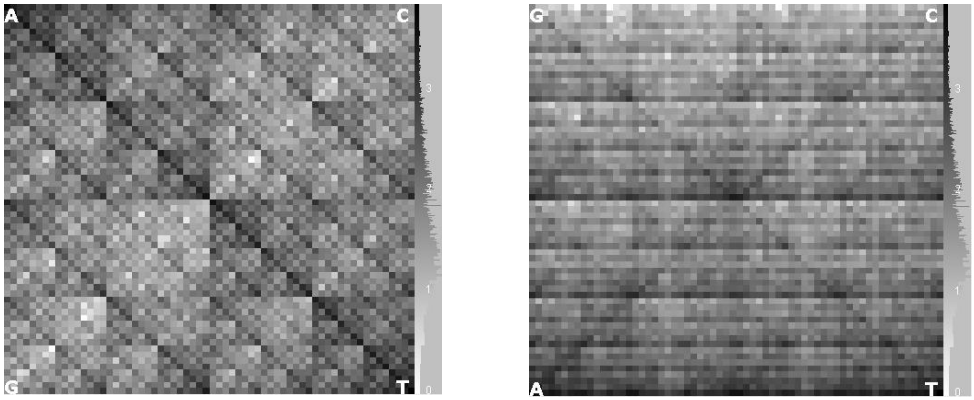

Figure 3. The images represent oligomer frequency fractals created by the CGR visualisation of Aster Yellows Witches' Broom phytoplasma genome for different corner base attractors, from left to right: (left) A, C, G, T, (right) G, C, A, T. Black colour means more frequent oligomers, and white colour stands for less frequent oligomers ("oligo" means few, in this case $K = 6$).

Probability or frequencies of occurrence of oligomers in arbitrary DNA (RNA) sequence are not equal and they also reflect specific biological characteristics. When the division of CGR space is based on the layout from Fig. 1, i.e. the space is divided into $2^K \times 2^K$ regions, frequencies of all oligomers of length K in analysed sequence will be displayed in the resulting plot, which also forms a fractal (see Fig. 3). Another important part of the CGR visualisation, which was not mentioned earlier, is the layout of image related to the location of the so-called attractor points. These points represent subsequences of identical bases e.g. AAAA…, CCCC…, etc. and are located in the corners of visualisation space. Positions of attractors can be easily rearranged by changing bases assignments in transformations (Eqs. 1). Different assignments can significantly change the resulting plot. In Fig. 3 two different layouts for visualisation of Aster Yellows Witches' Broom phytoplasma genome were chosen. The abundance of A-rich and T-rich sequences at opposite corners and diagonal is immediately evident in the left image (Fig. 3, left). On contrary similar feature is represented as horizontal dark patterns between A and T attractor points in the right image (Fig. 3, right). Unfortunately there is no general rule which CGR layout is the best, and it has to be chosen according to the analysed sequence.

References

[1] Hao, B., Lee, H. C., & Zhang, S. (2000). Fractals related to long DNA sequences and complete genomes. *Chaos, Solitons and Fractals, 11*(6), 825-836.
[2] Makula M and Benuskova L (2009) Interactive visualisation of oligomer frequency in DNA. *Computing and Informatics*, 28: 695-710. (http://mato.elet.sk/ifs/index.php)
[3] Shen, J., Zhang, S., Lee, H. C., & Hao, B. (2004). SeeDNA: Visualisation of k-string content of long DNA sequences and their randomized counterparts. Genomics, Proteomics & Bioinformatics, 2(3), 192-196.
[4] Tino, P. (1999). Spatial representation of symbolic sequences through Iterative Function Systems. *IEEE Trans. Systems, Man, and Cybernetics Part A*, 29(4), 386-392.