Analysis of fractal representation of genetic sequences

Amrik Sen

Department of Applied Mathematics
University of Colorado, Boulder
Boulder, CO 80309
Amrik.Sen@Colorado.edu

1 Introduction

One of the problems of interest to biologists is to determine patterns in genetic sequences that have no current explanation [Ber]. In fact experimental observation pertaining to the presence of excess oligonucleotides of some type over the expected number for random sequences can be taken as evidence for this functional significance [Sol] and hence reveal dependencies between bases that form the skeleton of any nucleic acid. Knowledge of these dependencies may be extremely useful while constructing probabilistic bounds for occurrence of a family of motifs in a gene sequence [SL]. In this paper, we employ a variant of the *chaos game representation* algorithm first coined by Barnsley [Bar] to graphically represent a typical random gene sequence. Analysis of this fractal representation enables us to answer questions like: (1) What is the probability of occurrence of a motif with a particular type of trailing subsequence, in a random sequence of fixed length ? (2) What is the conditional probability of finding a particular base given the occurrence of a certain subsequence ? In this paper, we also propose a new fractal characterization of another object of interest to probabilists , labelled Ψ_n, which denotes the number of times a rare motif may occur in a random sequence of given length, n. In spirit, this is equivalent to what is known as the *occupancy problem* in Markov process where mathematicians are concerned about the type of distribution of Ψ_n [Erh]. For biologists, a mathematical characterization of Ψ_n provides information about chances of random occurrence of mutants.

The rest of the paper is organized as follows. Section(2) introduces the chaos game algorithm and the variant employed for representing genetic sequence. In section(2.1) we provide some experimental results (plots) of this algorithm when applied to real gene sequences. In section (2.2), we propose how information may be extracted from these pictorial representations and quantified in terms of probability of occurrences of a certain class of events (subsequences). In section (3.1), we state some relevant definitions and theorems which we use to characterize Ψ_n in section (3.2). Finally, in section (4) we conclude with a brief summary and scope for future work.

2 The *chaos game* (CG) algorithm

In figure(1) we see the structure of a typical double stranded DNA. In most traditional approaches, like the one in [Ken], the occurrence of the nucleotide bases is assumed to be independent, i.e. $Pr(X_i = g|X_{i-1} = c) = Pr(X_i = g) = p_g$ and so on. Here, X_i denotes a random variable at the i^{th} instant of the random sequence of length, n; X_i may be either one of a, c, g or t. However, this assumption about independence is far from the truth, as we will see shortly.

The chaos game is an algorithm which enables one to produce fractal structures in an iterative manner. Formally, it belongs to the more general class of linear iterative function system. The basic steps of the algorithm are as follows [Jef]:

- Locate three initial points in a plane such that they are not collinear.
- Label one of the vertices with the numerals 1 and 2, the second vertex with the numerals 3 and 4, and the third vertex with the numerals 5 and 6.
- Pick a random initial starting point in the plane.
- Roll a six-sided die, the number rolled on the die picks out the corresponding vertex of the triangle.
- Place a mark halfway between the current point and the indicated vertex.
- Continue the above procedure.

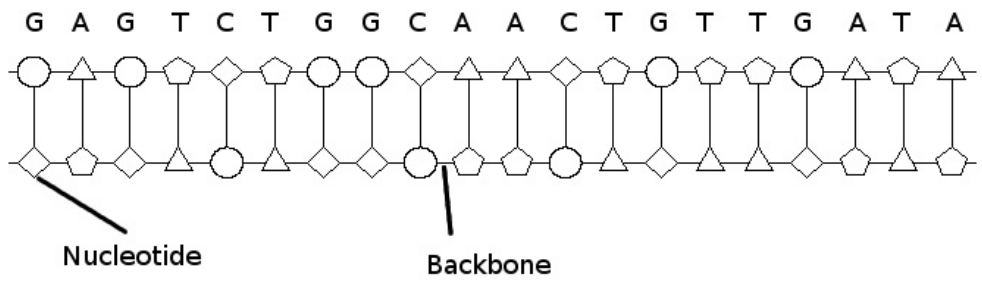

Figure 1: A typical double stranded DNA sequence.

The algorithm produces the well known *Sierpinski triangle* with 3 initial points; however with 5,6 or 7 initial points, the chaos game code produces a pentagon within a pentagon, a hexagon within a hexagon, a heptagon within a heptagon respectively. The case with 4 initial points is quite different though. In fact with 4 initial points, the space (square) gets filled up uniformly and randomly with dots. Thus we see that the patterns produced depend heavily on the initial number of vertices, hence the name *chaos* game.

It may be important to note that the picture produced by the chaos game is known as the *attractor*. A more formal treatment of the chaos game in terms of iterated function systems may be found in [Edg] but has been omitted from discussion in this paper.

In this paper, we employ the chaos game algorithm with some modification to reveal certain patterns in genetic sequences. We refer to figure(2). We start by assigning the tags corresponding to each of the bases in a DNA sequence (i.e. a, c, g and t) to the four vertex of a square. We take the center of the square as the starting point of our algorithm and read the genetic sequence character by character, each time placing a dot half way between the current point and the vertex corresponding to the character being read out from the sequence and continue the process until we have read the entire sequence. A few initial instances of this algorithm is shown in figure(2).

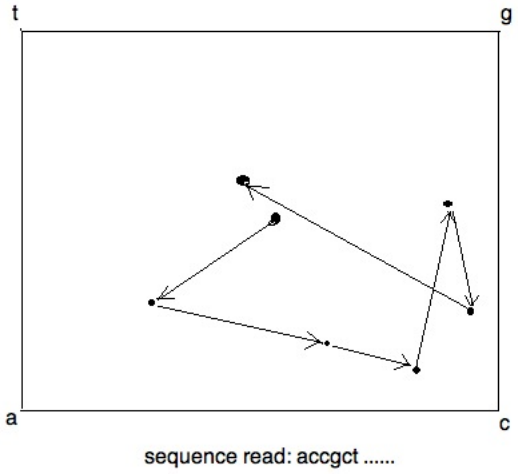

Figure 2: modified chaos game algorithm

2.1 CGR on experimental sequences

If the sequence being read were to be truly random with independent occurrence of the nucleotide base, we would expect absence of any interesting pattern(s). To ascertain our claim, we generated a random sequence comprised of characters from the alphabet, $\mathcal{A} = \{a, c, g, t\}$ by using *Matlab's* pseudo random generator, `randi()` and obtained the picture shown in figure(3). Clearly, we see that the space within the square was uniformly and randomly filled with dots, thereby implying the lack of any inherent dependencies in the occurrence of the bases.

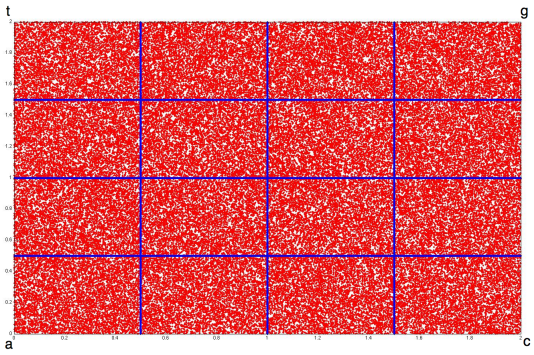

Figure 3: Chaos game representation of a truly random sequence.

Now, we use the gene database from the National Center for Biotechnology Information to test the CGR code on some real gene sequences. The results are shown in the figures in this section. A few interesting features of

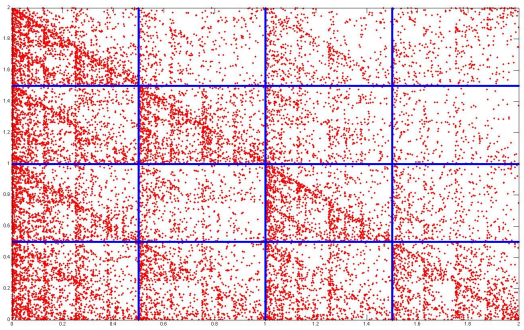

Figure 4: CGR of mitochondrial RNA sequence from Atlantic Hagfish.

the above representation produced by the CG algorithm are noted as below.

- neighboring points on the CGR are not close by in the actual sequence.
- subsequences ending with a common *trailing sequence* are mapped to their respective sub-quadrant as shown in figure(6).
- evidence of fractal nature of the plots imply presence of dependence in occurrence of the bases.
- figure(6) reveals regions of sparsity, for eg. the cg sub-quadrant; this implies that the likelihood of a g occurring after an occurrence of c is less likely in comparative terms. Similar arguments may be made for other subsequences.
- The features/patterns observed in the human DNA have also been found in the DNA sequences of vertebrates and those of certain viruses like HIV.

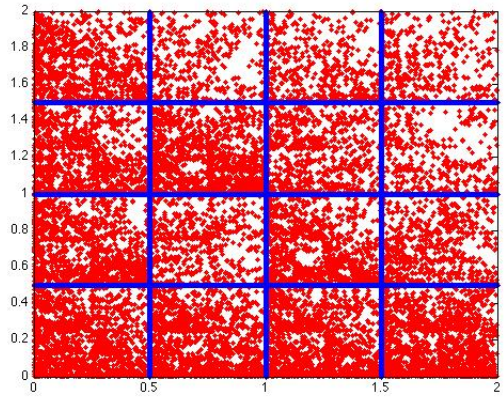

Figure 5: CGR of mitochondrial RNA sequence from Homo Sapiens Neanderthalensis.

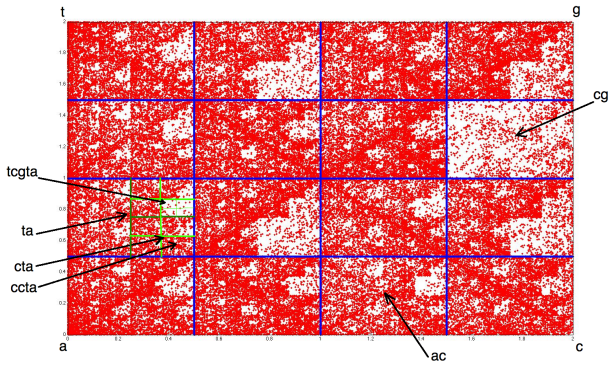

Figure 6: CGR of human DNA extracted from clone cell.

2.2 Probabilistic assertions about the occurrence of motifs

While [Jef] did report similar results, he relied on visual characterization of the patterns found in CGRs. In fact it was noted by [Jef] that a more mathematical and measure theoretic approach would be an extremely useful direction of future research. In the light of the above statement by [Jef], we make a novel attempt in that direction. It may be useful to note that the following information may be quantified using the CGR,

- $Pr(occurrence \quad of \quad a \quad subsequence) = \frac{no. \quad of \quad dots \quad in \quad the \quad corresponding \quad sub-quadrant}{total \quad no. \quad of \quad dots \quad in \quad the \quad plane}$
- In a Markov chain model of the genetic sequence, i.e. if $\{X_i\}_{i\geq 0}$ is a Markov process with finite state space $\{a, c, g, t\}$, the *joint distribution* of the chain (as an example) is computed as follows [Fu]:

  $$Pr(X_n = g, X_{n-1} = c, ..., X_1 = g, X_0 = t) = Pr(X_n = g|X_{n-1} = c)...Pr(X_1 = g|X_0 = t)Pr(X_0 = t)$$

  Clearly, the terms in the right hand side can be easily derived from the CGR of the sequence, for eg. $Pr(X_n = g|X_{n-1} = c) = \frac{no. \quad of \quad dots \quad in \quad the \quad cg \quad sub-quadrant}{no. \quad of \quad dots \quad in \quad the \quad c \quad sub-quadrant}$. The above is with respect to a first order Markov model. For higher order Markov models, it should be clear that the above arguments can be extended by looking at subsequence lengths of 3 or more and their respective sub-quadrants.

- One interesting aspect of the CGR pointed out earlier is the fact that points close in the sense of the Euclidean norm in the CGR plane may be far apart in the sequence space. An alternative measure in the CGR space may be to use the *Hausdorff* measure. In this regard, it may be useful to point out the strong similarities between the CGR of the RNA sequence from the Atlantic Hagfish and the *Sierpinski gasket*; and therefore we can estimate the Hausdorff dimension of the CGR of the RNA from the Hagfish to be about $\frac{\log 3}{\log 2} = 1.58$ which is numerically verified by using the box-counting algorithm from problem set 10. [Jia] also provide tight bounds on the estimates of the Hausdorff measure of such fractal sets.

3 Fractal characterization of Ψ_n

Recall that Ψ_n denotes the number of times a particular type of rare motif may occur in a random sequence of length, n. The cumulative distribution of Ψ_n defined here as $Pr(\sum_{m\geq 0}\Psi_n^{(m)} \leq c)$ is of immense interest to probabilists; however the nature of the rare events does not enable one to compute this explicitly and hence many like [Erh] have proposed approximate distributions with error bounds. Here, we propose a novel fractal characterization of Ψ_n in terms of a set whose elements are described based on $\sum_{m\geq 0}\Psi_n^{(m)} \leq c$ in some limiting sense stated shortly below. Such results have recently been of much interest in the context of random walks.

3.1 Fractal geometry in a probability space, $(\Omega, \mathcal{F}, \mu)$[DT]

Definition: A subset $A \subset \Omega$ is said to be a *fractal* with respect to a measure (probability measure) μ over Ω if $\mu(A) = 0$ and $dim_\mu(A) = Dim_\mu(A) = constant$. Here, $dim(.)$ is the Hausdorff dimension and $Dim(.)$ is the packing dimension as described in [DT]

Clearly, our knowledge about the *Cantor* set being a fractal satisfies this definition because the lebesgue measure, $\lambda(A) = 0$ and $dim_\lambda(A) = Dim_\lambda(A) = \frac{\log 2}{\log 3}$. It may also be useful to recall the equivalence of the lebesgue measure over the unit real line and the probability measure over Ω [Qu].

In the same spirit, we state the following theorem.

Theorem [Bill][DT]: For $\omega \in \Omega$, let $u_n(\omega) = \{\omega_0 : X_i(\omega_0) = X_i(\omega); i = 1, 2, ..., n\}$ and let $A = \{\omega : X_k(\omega) = a_k; k = 1, ..., n\} \subset \Omega$ s.t. $\mu(A) = 0$ for some n, then $\exists \nu(A) = p(a_1, ..., a_n)$ s.t. $\nu(u_n(\omega)) \leq \mu(u_n(\omega))^c \; \forall \omega \in \Omega$ and $c > 0$ constant.
Also, for some $M_0 \subset \{\omega : \lim_{n\to\infty} \frac{\log \nu(u_n(\omega))}{\log \mu(u_n(\omega))} = c\}$, if $\nu(M_0) > 0$; then M_0 is a *fractal* and $dim_\mu(M_0) = Dim_\mu(M_0) = c$.

3.2 Characterization of Ψ_n

Let Z_i denote a finite length querry subsequence (or motif) and S_0 be the *rarely* occurring target motif of the same length. $\Psi_n := \sum_{i\geq 1}[Z_i \in S_0] \equiv \sum_{i\geq 1} I_i$; where I_i has success probability p_{s_0}. Here, [.] refers to the *indicator* function.

The *strong law of large numbers* implies $\frac{1}{m}\sum_{m\geq 1}\Psi_n^{(m)} \to np_{s_0}$ as $m \to \infty$, n fixed. Next we define, $A := \{\omega : \frac{1}{m}\sum_{m\geq 1}\Psi_n^{(m)}(\omega) \nrightarrow np_{s_0}\}$ and $B_c := \{\omega : \lim_{m\to\infty}\frac{1}{m}\sum_{m\geq 1}\Psi_n^{(m)}(\omega) \leq c \in [0, np_{s_0})\}$. Clearly, $B_c \subset A; Pr(B_c) = Pr(A) = 0$. Hence, using the theorem above we have a new measure ν s.t. $\nu(B_c) > 0, dim_\mu(B_c) = f(c)$ and B_c is a *fractal*.

It must be pointed out here that while the theorem above is only an existence statement of the new measure ν, no general construction of such a measure has been reported in the literature; however some example specific constructions have been proposed for ν by [Bill].

4 Conclusion

In this paper, we have presented a fractal representation of genetic sequence via the modified CGR algorithm and have thereby shown the inherent dependencies amongst the nucleotide bases. We have made an attempt to extend the visual features of the CGR to more probabilistic and measure theoretic assertions about the same. We have also proposed a novel fractal based characterization of Ψ_n which was defined as the number of times a rare motif may be seen in a random gene sequence. Future work may include extending this characterization and establishing a stronger mathematical relation between the fractal characterization and the recent works on approximating probability distribution for Ψ_n. A comparative study of the CGR based representation of genes and other graphical representations of genes may also be very interesting.

References

[Ber] Berthelsen, Cheryl L et. al., *Global fractal dimension of human DNA sequences treated as pseudorandom walks*, Physical Review A, (Vol. 45, Number 12), 1992.

[Sol] Solovyev, Victor V., *Fractal graphical representation and analysis of DNA and protein sequences*, BioSystems, 30 (1993) 137-160, ©Elsevier Scientific Publishers Ireland Ltd.

[SL] Sen, A. and Lladser, M., *On some Chen-Stein like bounds for probability of finding a family of motifs in a random genetic sequence*, (manuscript in preparation).

[Bar] Barnsley, M., *Fractals Everywhere*, 1988, ©Morgan Kauffmann.

[Erh] Erhardsson, T., *Compound Poisson approximation for Markov chains using Stein's method*, The Ann. of Prob., 1999, Vol 27, No 1.

[Ken] Kennedy, R., *Calculating RNA motif probabilities and recognizing patterns in sequence data*, Honor's thesis, 2009, Univ. of Colorado, Boulder.

[Jef] Jeffrey, Joel H., *Chaos game representation of gene structure*, Nucleic Acid Research, Vol. 18, No. 8, p: 2163-2170, ©1990 Oxford University Press.

[Edg] Edgar, G., *Measure, Topology and Fractal Geometry*, 2^{nd} edition, 2007, ©Springer.

[Fu] Fu, J. et al., *Distribution Theory of Runs and Patterns and Its Applications*, 2003, ©World Scientific Press.

[Jia] Jia, B., *Bounds of Hausdorff measure of the Sierpinski gasket*, J. Math. Anal. Appl., 330 (2007), 1016-1024.

[DT] Dai, C. and Taylor, S.J., *Defining fractals in a probability space*, Illinois J. Math., 1994, Vol 38, No. 3.

[Qu] Qu, C.Q. et al., *Hausdorff measure of homogeneous Cantor set*, Acta Mathematica Sinica, 2001, Vol 17, No. 1, 15-20.

[Bill] Billingsley, P. *Hausdorff dimension in probability theory II*, Illinois J. Math., 1961, Vol 5, 291-298.