Quantitative Understanding of Breast Cancer Genes

Antara Sengupta
MCKV Institute of Engineering
243,G.T. Road(North), Liluah
Howrah-711204,West Bengal,
India

Sk. Sarif Hassan
International Centre for
Theoretical Sciences, Tata
Institute of Fundamental
Research, Bangalore 560012,
Karnataka, India

Pabitra Pal Choudhury
Applied Statistics Unit, Indian
Statistical Institute, Calcutta
700108, West Bengal, India

ABSTRACT

Characterization of oncogenes which are responsible for Breast Cancer can help us to understand the molecular signature of those genes. As it is known that the DNA molecules are nothing but the combinations of four nucleotides A, T, C and G which encodes the genetic instructions, here in this proposed paper firstly it is tried to make a quantitative understanding of DNAs which are responsible for breast cancer by comprehending their structures, characteristics, functions, evolutions etc. which can help to identify those DNAs among several nucleotide strings. Secondly, it is tried to recognize a given nucleotide string as a probable breast cancer DNA, without any conventional biological experiment. In this case, the nucleotide strings of DNAs of thirty seven have been taken and are classified and quantified by adapting some mathematical features.

General Terms

Breast Cancer Genomics, Quantitative Understanding, Mathematical Parameters et.al

Keywords

Breast Cancer, DNA, Quantitative Analysis, Chaos Game Representation, Fractal Dimension.

1. INTRODUCTION

From last few decades cancer has taken its devastating and whereas breast cancer seeks our intensive attention as patients undergo through physical and mental trauma parallel to their life risks. Although lots of solutions are emerging, but there is huge gap between those solutions and their applications [16].When Doctors and Biologists are constantly trying booms and bursts to find out ways of diagnosis, therapy and prevention to overcome this curse, there is a talk about "Quantitative Understanding". Quantitative understanding of genes in molecular level refers to apply some Mathematical parameters to evaluate some data, to derive some values and to verify some facts. Although gene therapy [15], antisense therapy are the better ways of treatment, where in antisense therapy the genetic sequence of a particular gene is known to be causative of a particular disease, it is possible to synthesize a strand of nucleic acid that will bind to the mRNA produced by that gene and inactivate it, but still now it is in experimental level to make sure that the application will be safe and effective. So it is worthy to make a quantitative insight of the genes which are responsible for breast cancer diseases. This approach will not only work without any intervention of medicines but also able to detect the disease at its early stages which will automatically minimize the life risk. In this paper to build quantitative and deterministic model firstly it is tried to find out the underlying geometries of DNA structure and the hidden geometrical rules, so that a mapping can be made between those geometrical rules with biological activities of breast cancer genes.

1.1 Model Representation

Mathematically DNA sequences (nucleotides) are symbolic sequences and must be interpreted as digital sequences to make statistical analysis. With this aim the following model representations is taken place.

1.1.1 Binary Representation

As DNAs are one dimensional sequences, so they can be represented as binary strings [17]. It can be done so by mapping T(A)=00, T(T)=11, T(C)=01, and T(G)=10. Some portions of the DNA sequence CCND1 is shown below:

TGATCTCCTTCTGGCAACATCGCGTCACTGAGCCGGG
GAGCTCACAGAGAAAGAGGCTCCTGCAGG…
Corresponding binary representation of the sequence CCND1 is,
111000110111010111110111101001000001001101100110011
011010001111000100101101010100001001110100010010
0010000000100010100111010111100100101010…….

1.1.2 4-adic Representation

It is possible to make 4-adic representation of the DNA sequences. To do so, A, T, G, C may be considered as a string of four variables 1,2,3,4 respectively. The 4-adic representation of CCND1 is as follows:

23124244224233411412434324142313443333134241413131
11313342442341333…

2. METHODS

2.1 Generating Indicator Matrix

The notion of indicator matrix and its characterization through fractal dimension was proposed by Carlo Cattani [6][13]. DNA sequences have four basic components (*A* = adenine, *C* = cytosine, *G* = guanine, *T* = thymine) which is defined as four alphabets A, C, G, T respectively.

Let us consider $D \overset{\text{def}}{=} \{A, C, G, T\}$ be the set of nucleotides and x$\in D$ where, x is any alphabet of D. A DNA sequence is the finite symbolic string $S = \mathbb{N} \times D$ *so that* S $\overset{\text{def}}{=} \{x_h\}_{h=1,\ldots,N}$,$N<\infty$ being $x_h \overset{\text{def}}{=}$ *(h, x)=x(h) where h=1,2,…,N and* x$\in D$the value *x* at the position *h*. According to C.Cattani the Indicator Matrix can be characterized through fractal dimension as follows,

$$f: S \times S \to \{0,1\} \text{such that } f(x_h, x_k) \overset{\text{def}}{=} \begin{cases} 1 & if x_h = x_k \\ 0 & if x_h \neq x_k \end{cases}$$

where x_h, $x_k \in S$

So now it is possible to easily describe a N×N sparse binary matrix from an indicator matrix of N length, which may be written as,

$$M_{hk} = f(x_k)x_{h},x_k \in S, h, k = 1,2,3, \ldots N$$

But it is not possible to differentiate between zeros formed by distinct base pairs. So slight modification is needed into it, which is [7],

$$f: S \times S \rightarrow \{1,2,3,4\}$$

such that

$$f(x_{h,}x_k) \overset{\text{def}}{=} \begin{cases} 1 \ \ if x_h = x_k & ; \ x_{h,}x_k \in S \\ 2 \ \ if x_h \neq x_k & ; x_{h,}x_{k \in \{G,T\} or \{A,C\}} \\ 3 \ \ if x_h \neq x_k & ; x_{h,}x_{k \in}\{T,C\} or \{A,G\} \\ 4 \ \ if x_h \neq x_k & ; \ x_{h,}x_k \in \{C,G\} or \{A,T\} \end{cases}$$

So the matrix M_{hk} will be decomposed into four binary matrices namely, A1, A2, A3, A4 and corresponding four binary images will be made for each DNA sequence. And then the fractal dimension[11][12][14] will be calculated using “Box-Counting” method where Box-Counting is a method of collection of data for analyzing complex pattern like dataset, images etc. and breaking them into smaller part or pieces and analyze them individually.

2.2 Fractal Dimension of DNA Walk

DNA Walk is a vector representation of DNA sequences transformed into a planer trajectory [2]. DNA Walk is defined as a series $\Sigma Y_{n,} n=1,2….N \& Y_n \in \{1,2,3,4\}$which is the cumulative sum on the DNA string representation[3]

$$\{Y_{1,}Y_{2,} \ldots, \sum_{m=1}^{n-1} Y_{m,\ldots,} \sum_{m=1}^{n} Y_{m,\ldots}\}$$

We can also define $a_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(A, x_{i)}, t_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(T, x_{i)}, g_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(G, x_{i)}, c_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(C, x_{i)}.$

It has been resulted by plotting (W_n, V_n) as two functions $W_n \overset{\text{def}}{=} \ \sin a_{n}^{2} {}_{-} \sin g_n^2$

$and \ \ V_n sint_n^2{}_{-} sinc_n^2$ are defined.

Accordingly the fractal dimensions of all 37 human breast cancer genes are derived. The DNA walk of gene CCND1 is shown below (see figure 1).

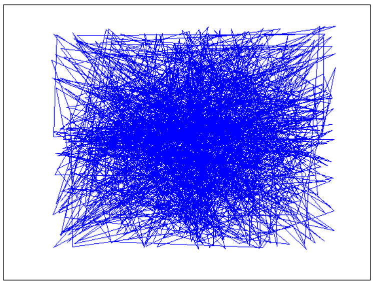

Fig 1: DNA Walk of gene CCND1

The box-counting dimension for the DNA walk of CCND1 is 1.87522. Similarly the box counting dimensions are computed for all genes.

2.3 Mean and Standard Deviation Ordering of Gene Sequences

A gene is a string constituting of different permutations of the base pairs A, C, T and G where repetition of a base pair is allowed. It is possible to classify the sequences based on the ordering of poly-string mean of A, C, T, and G in the string [4][5].

$$\text{Mean} N_u = \frac{2(N_{u1}+N_{u2}+N_{u3}+\cdots.N_{um})}{m}.(m+1)$$

where $N_{ui} \in \{A,T,C,G\}$, i=1,2,3,…,m and m is the length of longest poly-string over the string.

2.4 Chaos Game Representation

Chaos Theory is the combination of methods, results and visualization of dynamical system. As the name implies chaos theory tries to explain how chaos is produced and how it can be controlled. Chaos Game Representation is a graphical representation of one dimensional sequence into graphical form. It considers DNA sequences as strings of four units (A,T,C,G) and recognizes the pattern using the technique of fractal structure. The Chaos Game Representation of Breast cancer gene AIB1 is shown below:

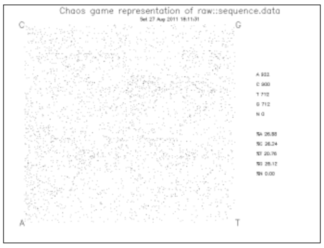

Fig 2: Chaos Game Representation of breast cancer DNA sequence AIB1

3. RESULTS AND DISCUSSIONS

3.1 Calculating Fractal Dimension of Indicator Matrix

Fractal dimensions of A1, A2, A3, A4 for each breast cancer (BC) genes are derived using BENOIT software. Descriptive Statistical analysis has been done using software named STATISTICA.

3.1.1 The FD of Indicator Matrix A1 and A2

The FDs for A1 and A2 are deducted for all the BC genes (Supplementary I), where it is observed that the FDs for A1 are between (1.79264 and 1.95478) and for A2 are between (1.78874 and 1.89672).So from here it can stated that an oncogene which is responsible for BC, its fractal dimension for A1 and A2 will be within that specified range.

Table 1. Correlation Coefficient Matrices of A1, A2, A3, A4.

	FD of A1	*FD of A2*	*FD of A3*	*FD of A4*
FD of A1	1	0.91522	0.81998	0.75850
FD of A2	0.91522	1	0.88650	0.81803
FD of A3	0.81998	0.88650	1	0.88304
FD of A4	0.75850	0.81803	0.88304	1

The descriptive statistical analysis is shown in the Supplementary I.

3.1.2 The FD of Indicator Matrix A3 and A4

The FDs for A3 and A4 are deducted for all the BC genes (Supplementary I), where it is observed that the FDs for A3 are between (1.78882 and 1.93334) and for A4 are between (1.78882 and 1.93334). So from here it can be stated that an oncogene which is responsible for BC, its fractal dimension for A3 and A4 will be within that specified range. The descriptive statistical analysis is shown in Supplementary I.

3.2 Fractal Dimension of DNA Walk

Fractal Dimension of DNA Walk has been derived for all 37 cancer genes (Supplementary I) and it is observed that the FD of DNA walks are lying between 1.78648 and 1.90967. So from here it can be concluded that any string of DNA having the FD of DNA Walk within these rang can be a BC gene. The descriptive statistical analysis is shown in the Supplementary I from which it is very cleared that in spite of the length of the DNA sequences varies but the genes are having almost same box-counting dimensions.

3.3 Statistical Autocorrelation

The statistical autocorrelation of all 37 breast cancer genes have been derived and it is found that the statistical autocorrelation (σ) values are lying between 1.1089 and 1.3740.The descriptive statistical analysis is shown in the Supplementary I from which it is very clear that the sigma value of all 37 breast cancer genes are equally distributed.As per following tabular representation the fractal dimensions of A1, A2, A3, A4 matrices are positively correlated.

3.4 Chaos Game Representation

The Chaos Game Representation (CGR) of the 37 breast cancer genes are being made (Supplementary I), where it can be found that the FDs of the CGR are lying between 1.92202 to 1.93455. The descriptive statistical analysis is shown in the Supplementary I from which it is found that although the DNA sequences are having differences in length and order but the fractal pictures which we are getting have approximately similar fractal deviations of CGR.

3.5 Mean and SD Ordering of Gene Sequences

All the breast cancer genes are classified based on their poly-string standard deviation (P_{SD}^{N}) and poly-string mean (P_{M}^{N}). The breast cancer genes can be classified into 11on the basis of orders poly-string mean and can be classified into 16 on basis of orders poly-string standard deviation (Supplementary I).

In some cases order of poly-string mean and order of poly-string SD are same but in some cases they differ. From the observation it is also cleared that the ordering of nucleotides in a gene is very important feature to distinguish them from each other and to make them unique.

3.6 K-Means Clustering of Breast Cancer genes

K-Mean clustering method is a method of vector quantization which aims to partition 'N' observations into C numbers of clusters in which each observation belongs to the cluster with the nearest mean. Using K Mean Clustering method[8]all thirty Seven breast cancer genes have been clustered into 15 different clusters (Supplementary I). The mean, SD and variance are stated in supplementary page.

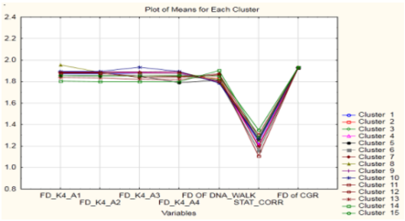

Fig 3: Mean of Clusters

4. CONCLUSION AND FUTURE ENDEAVOUR

In this paper it is tried to make quantitative understanding of breast cancer genes by constructing quantitative and deterministic model which can identify a given string of nucleotide as breast cancer gene without any intervention of biological investigations or experiments. Moreover, the model may be a standard prototype to identify all kinds of genes which can help the biologists to take an insight before starting further biological experiments if needed. In near future the study can be extended by validating the model through biological experiments and by constructing such models for rest of the cancer genes which are already there in cancer gene repository.

5. REFERENCES

[1] Sariego J (2010). "Breast cancer in the young patient".The American surgeon 76 (12): 1397–1401.

[2] Mandelbrot,"The Fractal Geometry of Nature."W.H. Freeman and Company..ISBN 0-7167-1186-9.

[3] Avnir (1998) "Is the geometry of Nature fractal".

[4] K Develi, T Babadagli (1998) "Quantification of natural fracture surfaces using fractal geometry" Math.

[5] Carlo Cattani, "Fractals and Hidden Symmetries in DNA" Mathematical Problems in Engineering Volume 2010, Article ID 507056

[6] Sk. S. Hassan, P. Pal Choudhury and A. Goswami, (2012), "Underlying Mathematics in Diversification of Human Olfactory Receptors in Different Loci ". Under disciplinary Sciences:Life Sciences December 2013, Volume 5, Issue 4, pp 270-273

[7] R. H. C. de Melo and A. Conci, (2008)"Succolarity: Defining a Method to calculate this Fractal Measure," ISBN: 978-80-227-2856-0 291- 294.

[8] PJ Deschavanne, A Giron, J Vilain, G Fagot (1999) "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences" Mol. Bio. Evo. 16 (10), 1391-1399.

[9] Sk. Sarif Hassan, Pabitra Pal Choudhury and Aritra Bose, (2011), "*A Quantitative Model for Human Olfactory Receptors* ", Nature Precedings, npre20126967-2.

[10] Yu Zu-Guo, (2002) "Fractals in DNA sequence analysis", Chinese Physics, 11 (12), 1313-1318.

[11] B. B. Mandelbrot, "The fractal geometry of nature". New York, ISBN 0-7167-1186-9, 1982.

[12] D. Avnir (1998) "Is the geometry of Nature fractal", Science 279, 39.

[13] K Develi, T Babadagli (1998) "Quantification of natural fracture surfaces using fractal geometry" Math. Geology 30 (8), 971-998.

[14] Yu Zu-Guo, (2002) "Fractals in DNA sequence analysis", Chinese Physics, 11 (12), 1313-1318.

[15] Friedmann, T.; Roblin, R. (1972). "Gene Therapy for Human Genetic Disease?". Science Vol 175 (4025), 949.

[16] Suzanne A Eccles et.al" Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer" Breast Cancer Research 2013, 15:R92.