A Quantitative Model for Human Olfactory Receptors

Sk. Sarif Hassan[1], Pabitra Pal Choudhury[2], Aritra Bose[3]

Applied Statistics Unit

Indian Statistical Institute, Calcutta, India

[1]sarimif@gmail.com, [2]pabitrapalchoudhury@gmail.com and [3]aritra90@gmail.com

Corresponding author: Sk. Sarif Hassan (sarimif@gmail.com)

Abstract:

A wide variety of chemicals having distinct odours are smelled by Humans. Odour perception initiates in the nose, where they are detected by a large family of Olfactory Receptors (ORs). Based on divergence of evolutionary model a sequence of Human ORs database has been proposed by *D. Lancet et al* (2000, 2006). It is quite impossible to infer whether a given sequence of nucleotides is a Human OR or not, without any biological experimental validation. In our perspective, a proper quantitative understanding of these ORs is required to justify or nullify whether a given sequence is a Human OR or not. In this paper, the entire Human OR sequences have been quantified, and a set of clusters have been made on using the quantitative results based on two different metrics. Using this proposed quantitative model, one can easily make probable justification or deterministic nullification whether a given sequence of nucleotides is a probable Human OR homologue or not, without seeking any biological experiment. Of course a further biological experiment is essential to validate the probable Human OR homologue.

Key words: Human Olfactory Receptors, Fractal Dimension, Hurst Exponent, Gene Therapy and Chaos Game representations.

1. Introduction

1.1 Current State of Art and Authors' contribution

The *Human Genome Project (HGP)* was an international research effort, coordinated by the *National Institutes of Health and the U.S. Department of Energy* to determine the sequence of the human genome and identify the genes that it contains [1, 2, 3 and 4]. The outcomes of HGP have allowed researchers to begin to understand the blueprint of genes and genomes. Discovering the sequence of the human genome was only the first step in understanding how the instructions coded in DNA lead to a functioning human being. The next stage of genomic research will begin to derive meaningful knowledge from the DNA sequence. The information-theoretic genomic understanding will have a major impact in the fields of medicine, biotechnology, and the life sciences.

In the present days, one of the most frontier challenges is to make a revolution in medical science by introducing *Genetic Therapy* [3]. *Gene therapy* is an experimental technique that uses genes to treat or prevent disease. This method of therapy would allow us to treat a disorder by inserting a gene into a patient's cells instead of using drugs or surgery. The most usual approaches of gene therapy include

- Replacing a mutated gene that causes disease with a healthy copy of the gene.
- Inactivating, or "knocking out," a mutated gene that is functioning improperly.
- Introducing a new gene into the body to help fight a disease.

Although gene therapy is a promising treatment option for a number of diseases (including inherited disorders, some types of cancer, and certain viral infections), the technique remains risky and is still under study to make sure that it will be safe and effective. Gene therapy is currently being tested for the treatment of diseases that have no other cures [3, 4]. Prior to gene therapy as a practical approach for treating diseases, we must overcome many technical challenges. In order to do that, first we must have quantitative insight of genes and genomes. This would help us in precise characterization of a particular DNA. The quantitative study of genes will be an add-on as genetic signature of a DNA sequence.

In the present study, a mathematical quantification of human Olfactory Receptors (ORs) [7, 8, 9, and 10, 12 and 13] has been deciphered by using *Fractal Geometry* [14, 15 and 16]. Also, a set of clusters have been made on using the quantitative results based on two different metrics. So on using this proposed quantitative model, one can easily make probable justification (deterministic nullification) whether a given sequence of nucleotides is a probable Human OR homologue or not, without seeking any biological experiment.

1. 2. Model Decomposition and Representation

DNA 4-Colored Representation: Let a DNA sequence be in the form of four-letter (ATGC) nucleotides sequence (*Fig. 1A*). Such sequence shown in *Fig. 1A* is converted as a function (Fig. 1B) depicting colors Red, Blue, Green, and Yellow respectively for A, T, G, and C [17, 18]. This allows $f(x, y)$ having maximum of 4 colors, i.e. $0 \leq f(x, y) \leq 3$.

```
ATGACAGGATTGAAAAATAAGAATTACACATTATTCCTTTAACATTGAGTTTCCCAGCTTTGAAGTAGCTGAAAT
AATTATATCGCATAAAAACTTTGTTATATTTTTCACTTTCTTATTTTCAAAAATTATAAAATTGGGTGTAAGACA
TTCTTAATTCTAAGAAAATGTTGATTTTGCTTATCTTCATGTTTTTATTCAATTAAGGACTTTTGGTAAACATTT
GCTGGTGTTAATGTTAAAAGAGAGTTGGGGAAATGGATGGCATGGGGCTCTGGGAAGACTCCTAGATAAACACTT
TAAGAGGCT…
```

Fig. 1 (A): A DNA string of four variables A T C and G

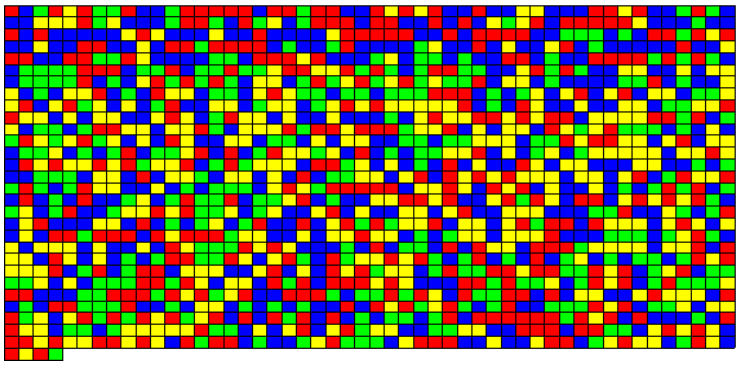


Fig.1. (B) Function generated by proper colour coding for ATGC of OR10AB1P

4-Adic Representation: Also we consider a DNA as a string of four variables 0, 1, 2 and 3 (as shown below) corresponding to A T C and G respectively. We name this string as *4-adic string* of DNA [17, 18].

023010330223000002003002201010220221122200102230322211103122230032031230002
002202021310200000122232202022222101222122022221000002202000022333232003010
221220022120030000232230222231220212210232222202210022003301222233200010222
312332322002322000030303223333000233023310233331212333003012112030200010122
200303312…

Binary Representation: We have considered a DNA as a one dimensional nucleotide sequence, and is represented as a map such that $T(A) = 00$; $T(T) = 11$, $T(C) = 01$ *and* $T(G) = 10$. This mapping yields a DNA sequence in a a binary string format. A portion of such a binary string is shown below of some fixed size (twice of the DNA sequence length). We call this representation of DNA as *2-adic string* of DNA [18].

001110000100101000111110000000000011000010000011110001000100111100111101011111110000010011
111000101111110101010010011111111000001011001001111000000011000011110011001101100100110000
000000011111111011110011001111111111010001111111011111001111111101000000000011110011000000
001111101010111011000010000100111101111…..(some more 0, 1 are there in the string).

Threshold decomposition: We have decomposed the four-colored image $f(x, y)$ into four binary images $f^i(x, y) = z$ (*Fig. 2A-D*) for a DNA through the threshold decomposition function defined as:

$$f^i(x, y) = 1\ ;\ z = i : i = 0, 1, 2 \text{ and } 3.$$

$$= 0\ ; z \neq i$$

Those decomposed binary images for one human OR are denoted as ${f_{OR}}^A, {f_{OR}}^T, {f_{OR}}^G$ *and* ${f_{OR}}^C$ are shown in the following:

Fig. 2: Threshold decomposed binary images of OR10AB1P (Black and white denote complimentary space and one of the ATGC). (A) A (B) T (C) G, and (D) C.

In the next section let us elaborate the methods applied to DNA string to extract the quantitative details.

2. Methods

The quantitative details of human ORs have been studied in the light of fractal dimension. The method of computation of each features for OR sequences are sketched in the following.

2.1 Generating Indicator Matrix and Its Quantification

A DNA is composed of four basic nucleotides namely A=Adenine, C=Cytosine, T=Thymine and G=Guanine.

Let $\mathcal{V} \stackrel{\text{def}}{=} \{A, T, G, C\}$ be the set of nucleotides and $x \in \mathcal{V}$ be any member of the alphabet.

A DNA can be thought of as a finite symbolic string $\mathcal{S} = \mathbb{N} \times \mathcal{V}$ so that $\mathcal{S} \stackrel{\text{def}}{=} x_i$, $i = 1, 2 \ldots N$ being $x_i \stackrel{\text{def}}{=} (i, x) = x(i)$, $(i = 1, 2 \ldots N;\ x \in \mathcal{V}$ the value of x at position I *and* N denote length of the string.

The notion of indicator matrix and its characterization through fractal dimension was proposed by C. Cattani [19] as follows

$$f: \mathcal{S} \times \mathcal{S} \rightarrow \{0, 1\} \text{ such that } f(x_{\mathrm{h}}, x_{\mathrm{k}}) \stackrel{\text{def}}{=} \begin{cases} 1 \; if \; x_{\mathrm{h}} = x_{\mathrm{k}} \\ 0 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}} \end{cases} \quad x_{\mathrm{h}}, x_{\mathrm{k}} \in \mathcal{S}$$

Therefore, the indicator matrix of an N-length string can be easily described as N×N sparse symmetric, binary matrix which results from

$$M_{\mathrm{hk}} = f_{x_h}(x_k) x_h, x_k \in \mathcal{S}, h, k = 1, 2, 3 \ldots, N$$

This definition of indicator matrix does not help us differentiate between zeros formed by distinct base pairs. A slightly modified definition of f is proposed as follows [17, 18]:

$$f: \mathcal{S} \times \mathcal{S} \rightarrow \{0,1,2,3\} \text{ such that } f(x_{\mathrm{h}}, x_{\mathrm{k}}) \stackrel{\text{def}}{=} \begin{cases} 0 \; if \; x_{\mathrm{h}} = x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \mathcal{S} \\ 1 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \{G, T\} or \{A, C\} \\ 2 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \{T, C\} or \{A, G\} \\ 3 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \{C, G\} or \{A, T\} \end{cases}$$

Consequently, the matrix M_{hk} corresponding to a given DNA is a four-threshold matrix, namely 0, 1, 2 and 3. Let us decompose the matrix M_{hk} into four binary matrices A1, A2, A3 and A4 as follows:

$$A1_{\mathrm{hk}} = \begin{cases} 1 \; where \; x_{\mathrm{h}} = x_{\mathrm{k}}, \; x_{\mathrm{h}}; x_{\mathrm{k}} \in \mathcal{S} \\ 0 \; otherwise \end{cases}$$

$$A2_{\mathrm{hk}} = \begin{cases} 1 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \{G, T\} or \{A, C\} \\ 0 \; otherwise \end{cases}$$

$$A3_{\mathrm{hk}} = \begin{cases} 1 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \{T, C\} or \{A, G\} \\ 0 \; otherwise \end{cases}$$

$$\text{and } A2_{\mathrm{hk}} = \begin{cases} 1 \; if \; x_{\mathrm{h}} \neq x_{\mathrm{k}}, x_{\mathrm{h}}; x_{\mathrm{k}} \in \{C, G\} or \{A, T\} \\ 0 \; otherwise \end{cases}$$

From the indicator matrix we have an idea of fractal-like distribution of nucleotides in DNAs. The corresponding fractal dimensions for the graphical representation of indicator matrices can be computed through *Box Counting Method* which is briefly stated as follows.

Box-Counting Method: This method computes the number of cells required to entirely cover an object, with grids of cells of varying size. Practically, this is performed by superimposing regular grids over an object and by counting the number of occupied cells. The logarithm of N(r), the number of occupied cells, versus the logarithm of 1/r, where r is the size of one cell, gives a line whose gradient corresponds to the box dimension [15, 16]. To calculate the dimension for a fractal S, the Box-Counting dimension is defined as,

$$\text{Dim}_{\text{box}}(S) = \lim_{n\to 0} \frac{\log N(r)}{\log(\frac{1}{r})}$$

Let us understand through an example considering the sequence one OR sequence viz. OR10AB1P having the sequence ATGGGCAATCACACTG... (Continuing)

The indicator matrices A1, A2, A3, and A4 for the OR10AB1P are as follows.

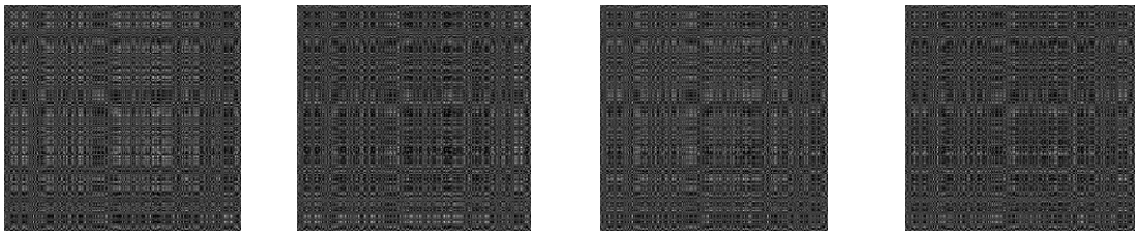


Fig. 3A *Fig. 3B* *Fig. 3C* *Fig. 3D*

Fig. 3: Indicator Matrices of OR10AB1P

We have calculated the fractal dimensions of the images using one of the well-known methods called *'Box-Counting method'*.

The fractal dimensions of the indicator matrices A1 A2 A3 and A4 for OR10AB1P are 1.83359, 1.82657, 1.83072 and 1.82465 respectively [18].

In the similar fashion the fractal dimensions of the indicator matrices for all human OR DNA strings have been computed.

2.2 DNA Walk of the DNA sequences

The DNA walk is defined as a sum of the progression $\sum Y_n, n = 1,2, \ldots\ldots, N \;\&\; Y_n \epsilon \{1,2,3,4\}$ which is the cumulative sum on the DNA string representation$\{Y_1, Y_1 + Y_2, \ldots\ldots, \sum_{m=1}^{n-1} Y_m, \ldots\ldots, \sum_{m=1}^{N} Y_m\}$ [19].
Also we define $a_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(A, x_i), g_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(G, x_i), c_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(C, x_i)$ & $t_n \overset{\text{def}}{=} \sum_{i=1}^{n} f(U, x_i)$. It has been resulted by plotting (W_n, V_n) as we have defined two functions:

$$W_n \overset{\text{def}}{=} \sin a_n^2 - \sin g_n^2 \text{ and } V_n \overset{\text{def}}{=} \sin t_n^2 - \sin c_n^2.$$

Here we compute the Fractal dimension of all DNA walk for the *4-adic string* of all ORs. The plot of the DNA walk for the OR1AB1P string is shown in *Fig. 4*.

Fig. 4: DNA Walk (W_n, V_n) for OR10AB1P

The box-counting dimension for the DNA walk of OR1AB1P is 1.94601. In the similar manner we have computed all the Fractal dimension of all the human OR DNA strings.

2.3 Hurst Exponent of the DNA sequences

Hurst exponent is referred to as the "index of dependence," and is the relative tendency of a time series either to regress strongly to the mean or to cluster in a direction. It is a measure of long range correlation of one-dimensional time series [19, 20].

Let us consider a string $X = \{x_i\},\ i = 1,2, \ldots, n$

$$m_{X,n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$Y(i,x) = \sum_{j=1}^{i}\{x_{j-}m_{x,n}\}$$

$$R(n) = \max Y(i,n) - \min Y(i,n)\ \ 1 \le i \le n$$

$$S(n) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - m_{x,n}\right)^2}$$

The Hurst exponent H is defined as :$(\frac{n}{2})^H = \frac{R(n)}{S(n)}$, where n is the length of the string. The range for which the Hurst exponent, H indicates negative, positive auto-correlation are $0 < H < 0.5$ and $0.5 < H < 1$ respectively. A value of $H=0.5$ indicates a true random walk, where it is equally likely that a decrease or an increase will follow from any particular value [20].

Here we consider 2-adic and 4 adic string of DNA for computation of Hurst exponent.

The Hurst exponents of the *4-adic and 2-adic string* of OR10AB1P are 0.5635 and 0.5765 respectively. This is how we have computed Hurst exponent for all the human OR DNA strings [18].

2.4 Succolarity

The degree of percolation of an image (how much a given fluid can flow through this image) can be measured through Succolarity, a fractal parameter [21].

The succolarity of a binary image is defined as

$$\sigma(BS(k), dir) = \frac{\sum_{k=1}^{n} OP(BS(k)) \times PR(BS(k), pc)}{\sum_{k=1}^{n} PR(BS(k), pc)}$$

where 'dir' denotes direction; $BS(n)$ where n is the number of possible divisions of a binary image in boxes. The occupation percentage (OP) is defined as, for each box size, k, then the sum of the multiplications of the $OP\big(BS(k)\big)$, where k is a number from 1 to n, by the pressure $PR(BS(k), pc)$, where *pc* is the position on x or y of the centroid of the box on the scale of pressure) applied to the box are calculated. Therefore for any binary decomposed images of $f(x, y)$, the succolarity can be obtained.

Here we compute succolarity of the decomposed images for DNA as shown in the previous section. The succolarity of the four decomposed images ${f_{OR}}^{A}, {f_{OR}}^{T}, {f_{OR}}^{G}$ and ${f_{OR}}^{C}$ of OR10AB1P are 1.227665, 1.07841875, 0.1156 and 0.749545938 respectively.

Similarly we have computed the succolarity of the decomposed images for all the human ORs.

2.5 Chaos Game Representation

Chaos Game Representation (CGR) can recognize patterns in the nucleotide strings using the techniques of fractal structures and by considering DNA sequences as strings composed of four units A, T, G and C. Such recognition of patterns relies on visual identification [22]. It is an application of non-random input to an iterated function system. The original CGR method is an algorithm which produces pictures revealing patterns in DNA sequences. Basically, the whole set of frequencies of the words found in a given genomic sequence can be displayed in the form of a single image in which each pixel is associated with a specific word. Frequencies of words found in a sequence are displayed in a square image, with the location of a given word being chosen according to a recursive procedure. Thus, the image is divided into four quadrants in which sequences ending with the appropriate base are collected. This gives the base composition of the sequence. Each quadrant is subsequently divided into four sub quadrants, each containing sequences ending with a given dinucleotide, such that sequences differing only in the first letter are in adjacent sub quadrants. The sequence is read base by base so that all available words are considered.

For example in the CGR of the sequence "ATTGCAGGCT" the sixth point represents the sequence "ATTGCA". Thus there is a one to one correspondence between the subsequences and the points in the CGR. 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.

First we design the CGR of all OR strings then the fractal dimension of CGR have been calculated by Box-counting method. The CGR of the OR strings OR10AB1P is given in *Fig. 5* and corresponding fractal dimension is 1.9407.

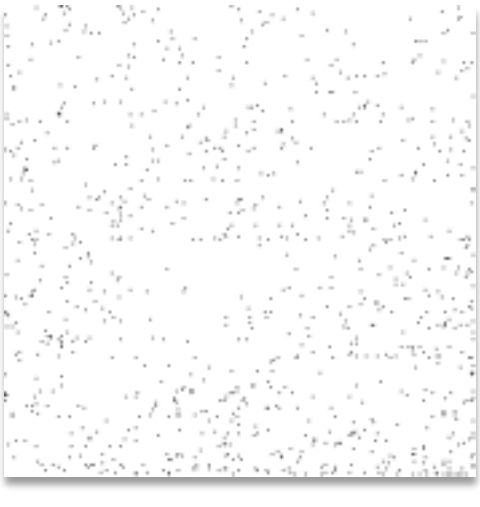


Fig. 5: Chaos Game Representation of OR10AB1P

3. Results and Discussions

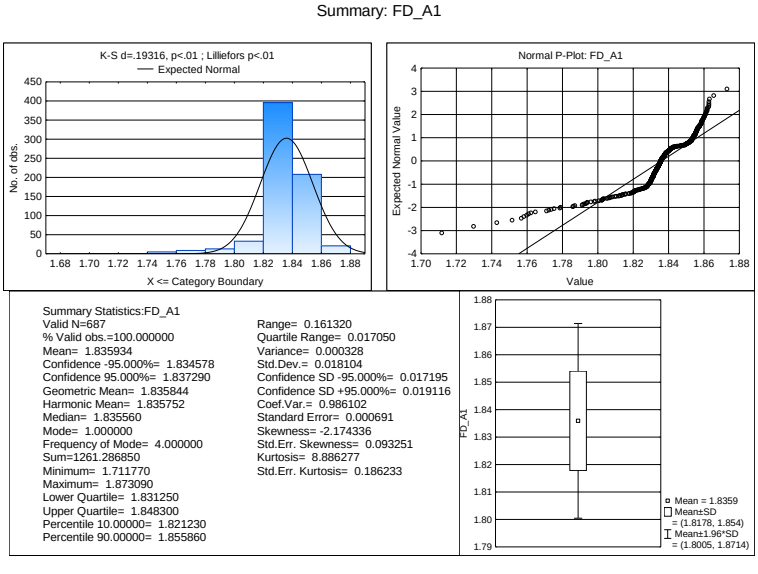
Let us now elaborate in detail, the result obtained for all human ORs using the above stated methods.

3.1 Fractal Dimension of Indicator Matrices

Here we compute the indicator matrices A1, A2, A3 and A4 for all the human OR sequences. Then we found the fractal dimension for each of those indicator matrices. The results are elucidated in the following. It is noted that the descriptive statistics for all the features are obtained using the software *Statistica*.

3.1.1. FD of Indicator Matrix (A1)

The fractal dimensions (FD) for the entire human OR sequences (687) are ranging from 1.71 to 1.87. The detailed result is given in the table 1. It is noted that the harmonic and geometric mean are almost same 1.83.



Tab. 1: Descriptive Statistics of FD of Indicator Matrix A1

It shows that the FD of Indicator matrix A1 is not normally distributed over the ORs. It follows *non-parametric* distribution as shown in *Fig. 6.*

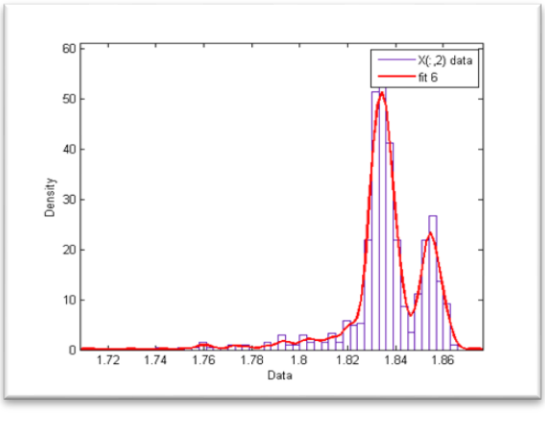
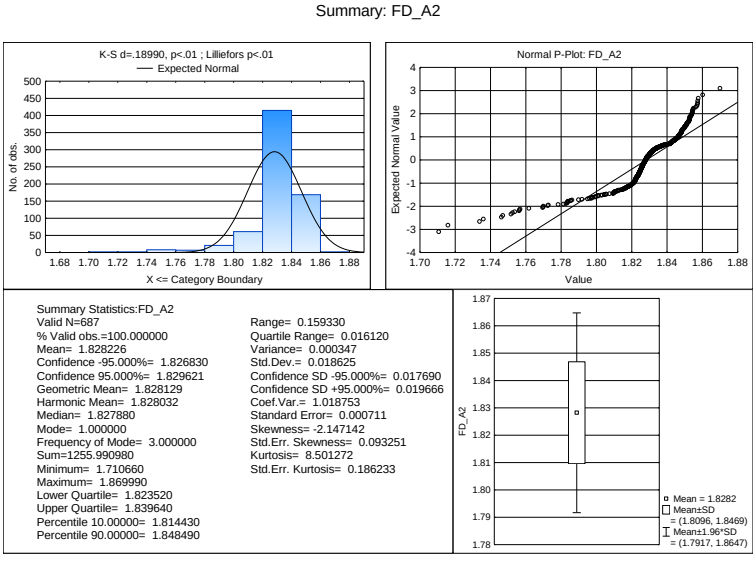


Fig. 6: Non-Parametric Distribution of FD of Indicator matrices (A1)

3.1.2. FD of Indicator Matrix (A2)

The fractal dimensions (FD) of the indicator matrices A2 for all ORs is illustrated in the below in the *Tab.2*. The estimated interval for FD of A1 is (1.71, 1.87).



Tab. 2: Descriptive Statistics of FD of Indicator Matrix A2

As we saw in the previous case, the distribution is following non-parametric distribution as resulted in the *Fig. 7*. It is seen that the harmonic and geometric mean of the distribution is nearly same.

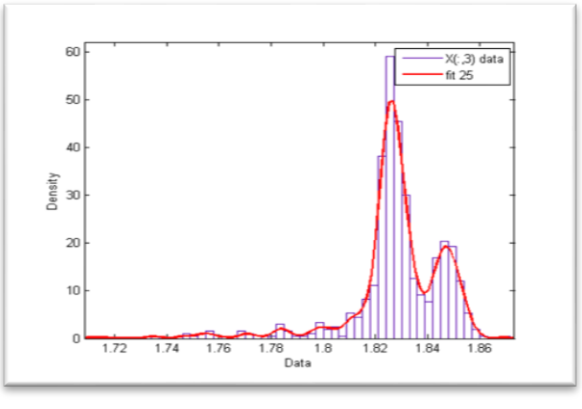
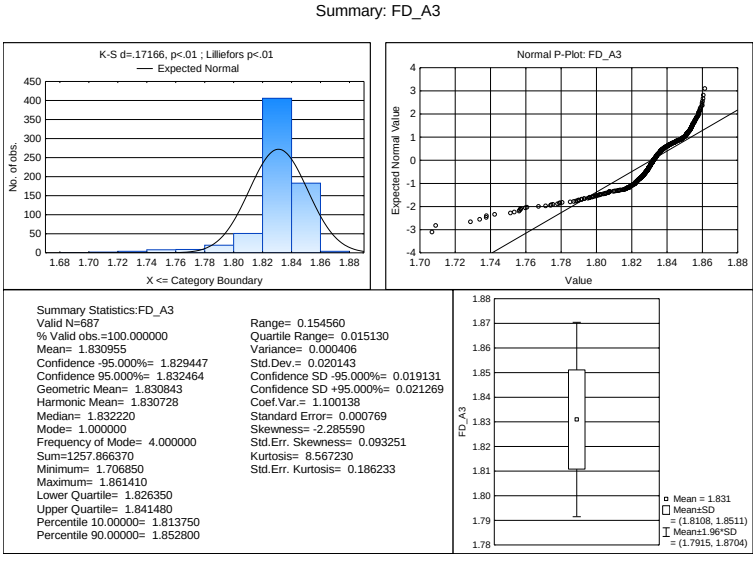


Fig. 7: Non-parametric distribution of FD of Indicator Matrices (A2)

3.1.2. FD of Indicator Matrix (A3)

The fractal dimensions (FD) of the indicator matrices A3 for all ORs is illustrated in the below in the *Tab.3*. The estimated interval for FD of A1 is (1.70, 1.86).



Tab. 3: Descriptive Statistics of FD of Indicator Matrix A3

The FD of A3 follows non-parametric distribution across the ORs *Fig. 8*. The harmonic and geometric mean of the distribution is nearly same as in the previous occasions.

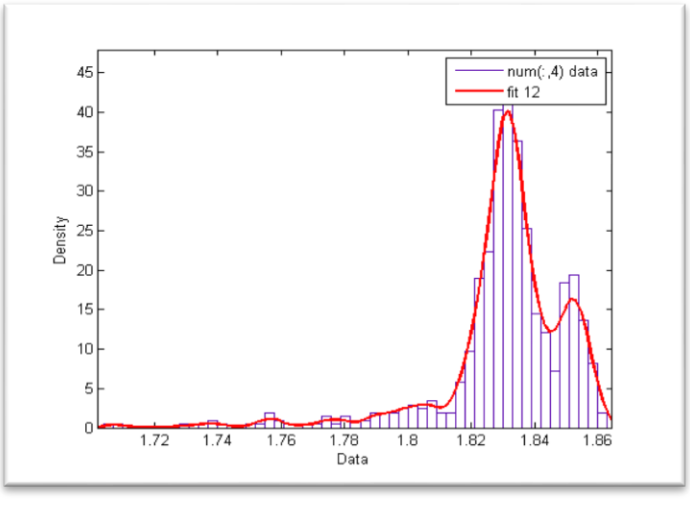
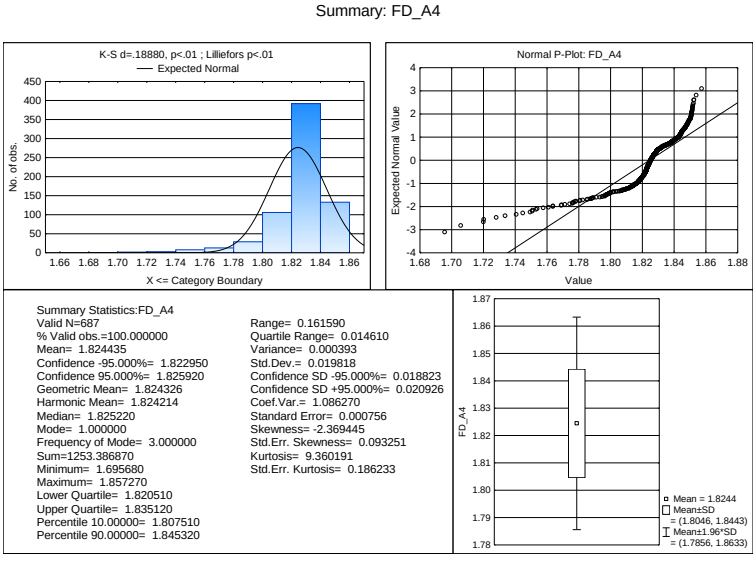


Fig. 8: Non-parametric distribution of FD of Indicator Matrices (A3)

3.1.2. FD of Indicator Matrix (A4)

The fractal dimension (FD) of the indicator matrices A4 for all ORs is illustrated in the below in the *Tab.4*. The estimated interval for FD of A1 is (1.69, 1.85).



Tab. 4: Descriptive Statistics of FD of Indicator Matrices (A4)

The FD of A4 follows non-parametric distribution as well as resulted in the *Fig. 9*. The harmonic and geometric mean of the distribution follows the same as before.

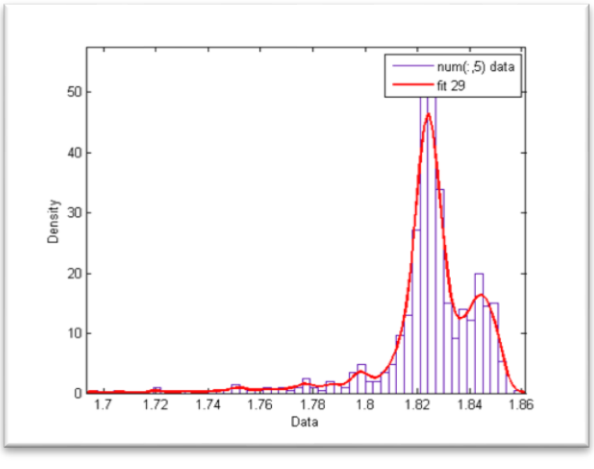


Fig. 9: Non-parametric distribution of FD of Indicator Matrices (A4)

The fractal dimensions (FD) of A1, A2 A3 and A4 are highly positively correlated as we have seen in *Fig. 10*. The correlation coefficients are tabulated in *Tab. 5* with graphical representation in *Fig. 10*.

	FD of A1	**FD of A2**	**FD of A3**	**FD of A4**
FD of A1	1	0.9705	0.9554	0.9552
FD of A2	0.9705	1	0.9259	0.9480
FD of A3	0.9554	0.9259	1	0.9865
FD of A4	0.9552	0.9480	0.9865	1

Tab. 5: Correlation coefficients for Indicator matrices A1, A2, A3 and A4

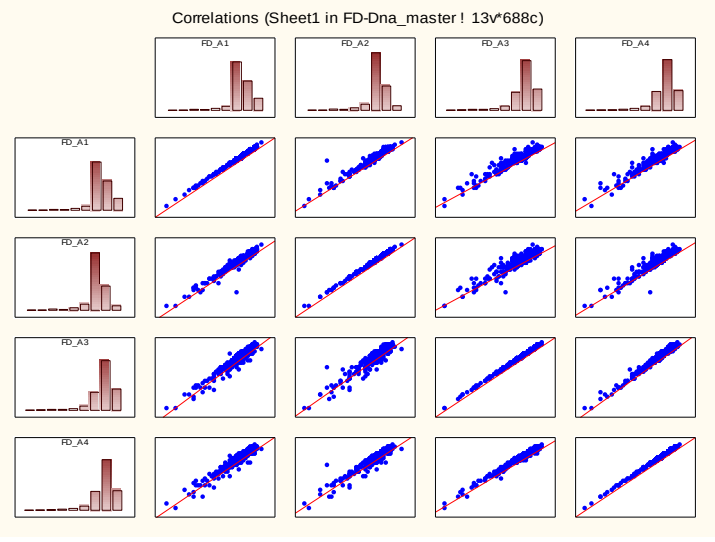
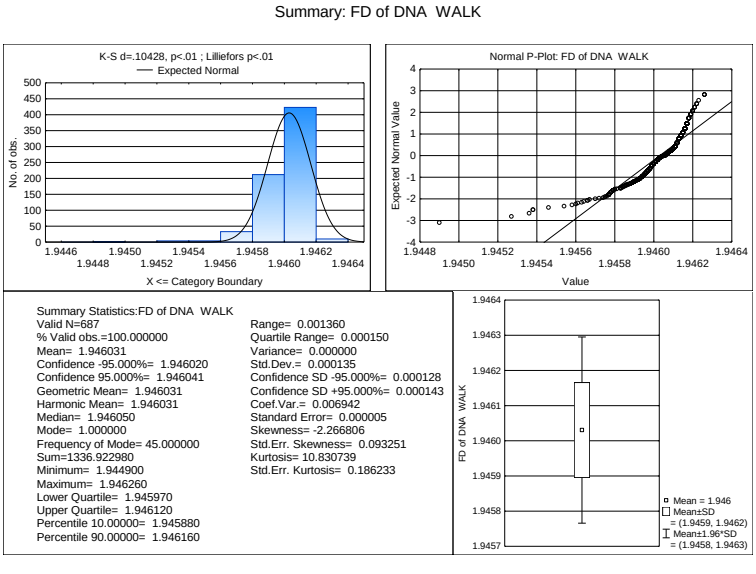


Fig. 10: Correlation graph for FD of A1, A2, A3 and A4

3.2. Fractal Dimension of DNA Walk

For the entire human ORs the fractal dimension of the DNA walks are almost same and that is 1.946 as shown in *Tab. 6.*



Tab. 6: Descriptive Statistics of FD of DNA Walk

The distribution of the FD of DNA walk is non-parametric as shown in *Fig. 10.*

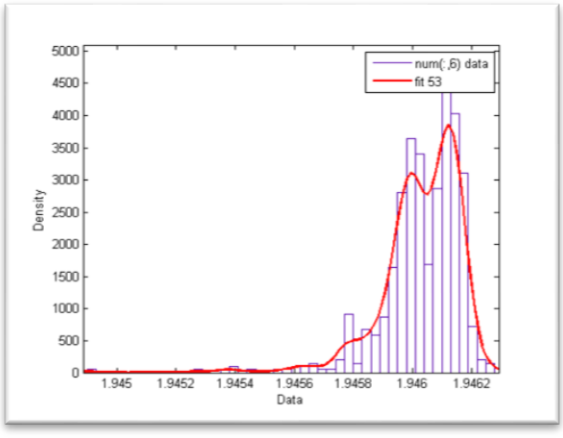


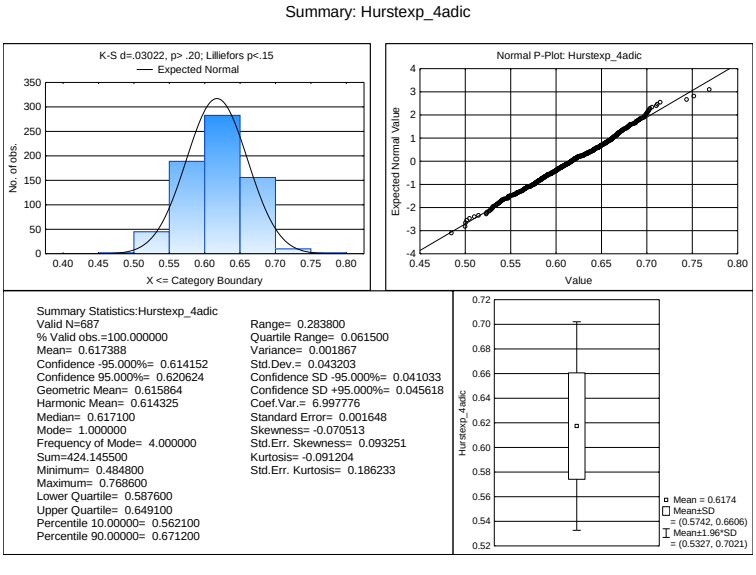
Fig. 10: Non-parametric distribution of FD of DNA Walk

3.3 Hurst Exponent

We have calculated the *Hurst exponent* of 4-adic as well as 2-adic strings of DNA for all the ORs. The details result is given as follows.

3.3.1 Hurst Exponent of 4-adic Strings

The Hurst exponent of *4-adic string* of ORs follows normal distribution as illustrated in the *Tab. 6*. The Hurst exponents for 2-adic strings range from 0.484 to 0.768.

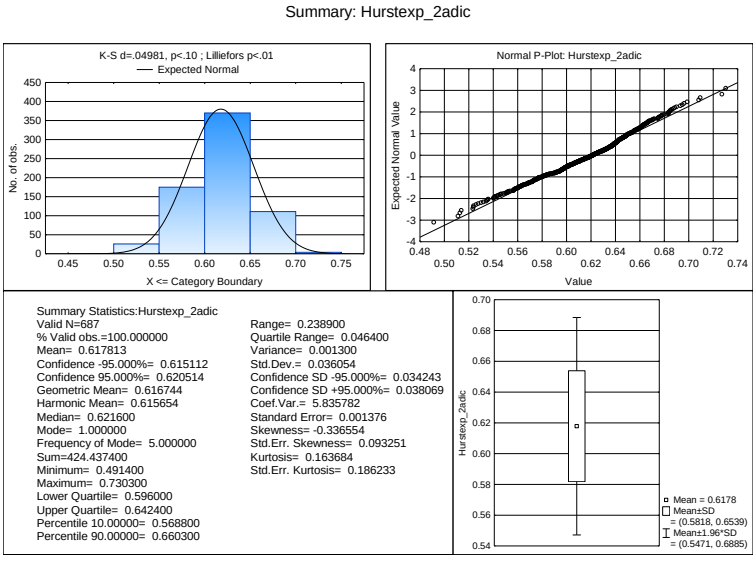


Tab. 6: Descriptive Statistics of Hurst exponent (4-adic)

Here also geometric and harmonic mean are same for the above distribution of 4-adic Hurst exponent.

3.3.2 Hurst Exponent of 2-adic Strings

The Hurst exponents (HEs) of *2-adic strings* normally distributed over the ORs as illustrated in the *Tab. 7*. The HEs for 2-adic strings range from 0.491 to 0.730. It is noted that the geometric and harmonic mean are same.



Tab. 7: Descriptive Statistics of Hurst exponent (2-adic)

The correlation coefficient of HEs for 4-adic and 2-adic string is 0.8052 as shown in *Fig. 11*.

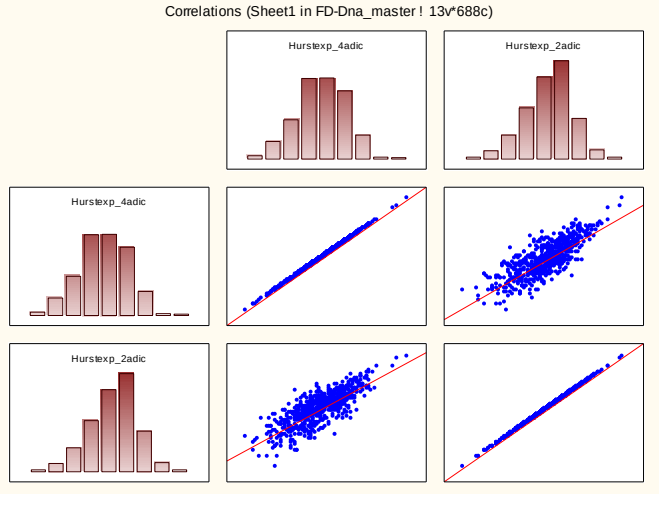


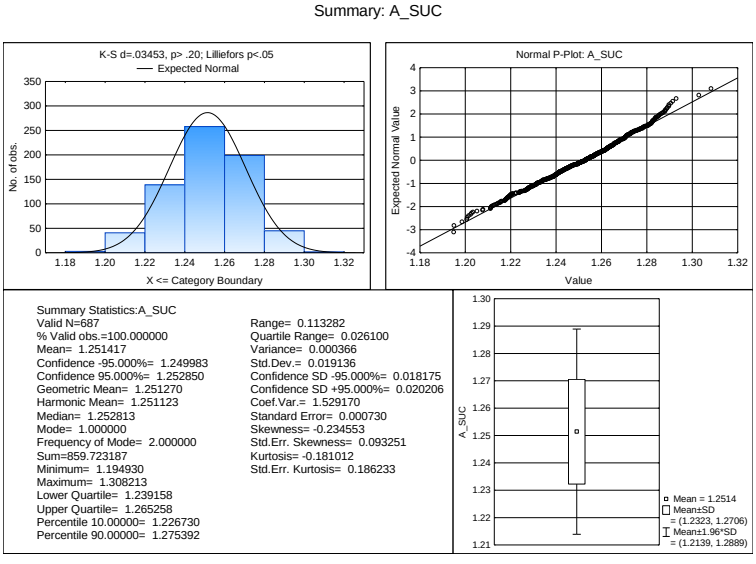
Fig. 11: Graph of correlation coefficient of 4-adic and 2-adic HEs.

3.4 Succolarity Indices

The succolarity indices for all ORs have been enumerated as illustrated below.

3.4.1 Succolarity of A

The succolarity indices follow normal distribution across the ORs as figured in *Tab. 8*. The succolarity of A for all ORs lies in the interval (1.19, 1.30).

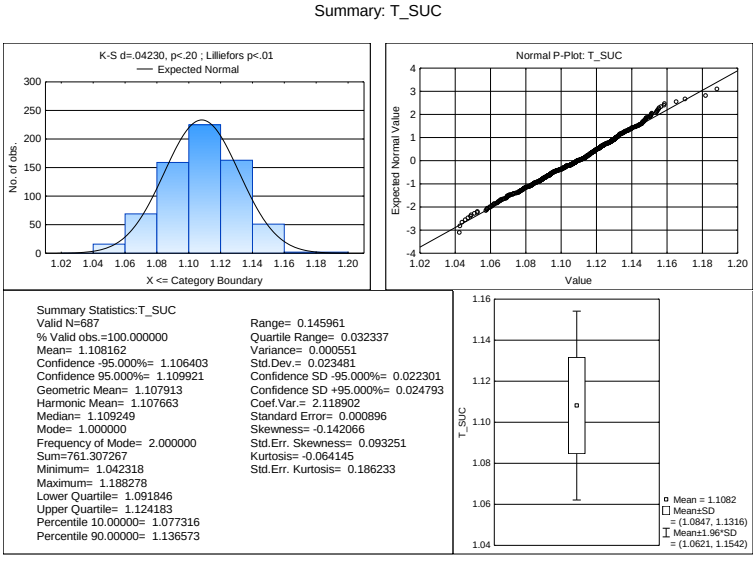


Tab. 8: Descriptive Statistics of Succolarity of A

It is shown that geometric and harmonic mean of the distribution again follows the same as we have obtained.

3.4.2 Succolarity of T

The succolarity indices adhere to a normal distribution for all the ORs as shown in Tab 9. Succolarity of T across all ORs lies in the interval (1.04, 1.18).

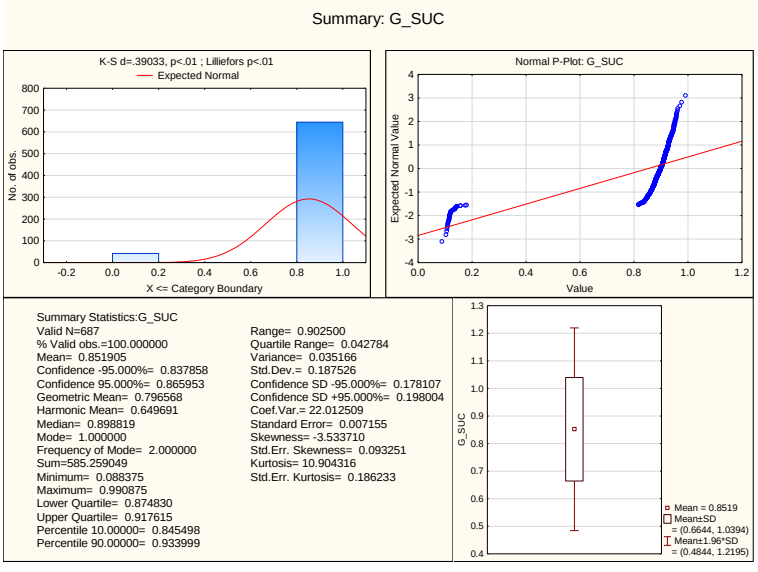


Tab. 9: Descriptive Statistics of Succolarity of T

The Geometric Mean and the Harmonic Mean of the distribution are nearly similar to each other.

3.4.3 Succolarity of G

The succolarity indices for all ORs are shown in the following, Tab 10. It is seen that the Succolarity of G for all ORs lies in the interval (0.08, 0.99).



Tab. 10: Descriptive Statistics of Succolarity of G

It is observed that the succolarity of G follows a non-parametric distribution. We show this in *Fig. 12*.

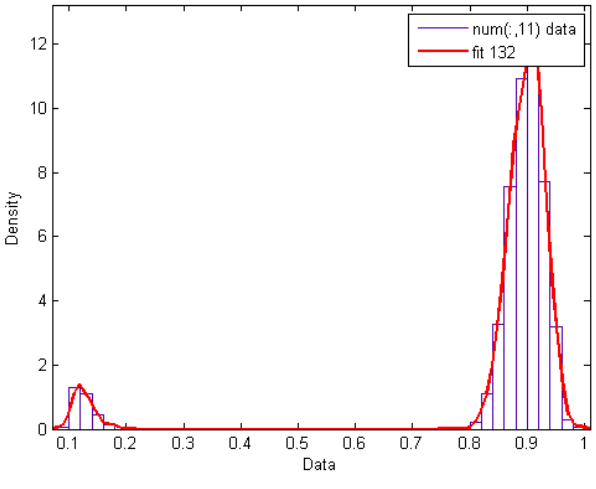
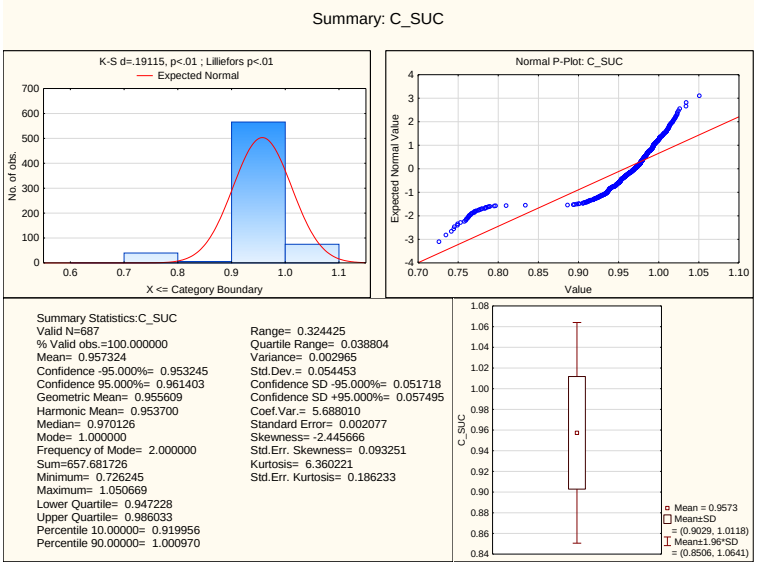


Fig. 12: Non-parametric distribution of succolarity of G

3.4.4 Succolarity of C

The succolarity indices of C are computed for all ORs and illustrated below in Tab 11.It is observed that the indices range from 0.726 to 1.05.



Tab. 11: Descriptive Statistics of Succolarity of C

Geometric Mean and the Harmonic Mean of succolarity of C are seen to be almost equal. Also, it is observed that the distribution follows a non-parametric pattern as shown in *Fig 13*.

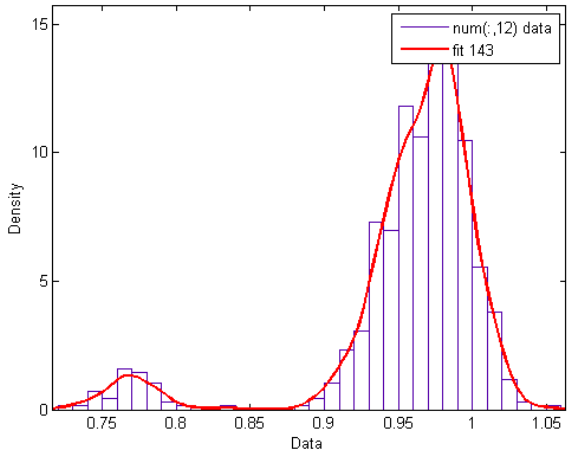


Fig. 13: Non-parametric distribution of succolarity of C

The correlation among succolarities of A, T, C and G are illustrated in the *Tab. 12* with graphical representation in *Fig. 10.*

	A_Suc	T_Suc	C_Suc	G_Suc
A_Suc	1	0.9783	0.5351	0.2213
T_Suc	0.9783	1	0.5274	0.2067
C_Suc	0.5351	0.5274	1	0.9383
G_Suc	0.2213	0.2067	0.9383	1

Tab. 12: Correlation coefficients for Sucolarities of A, T, C and G

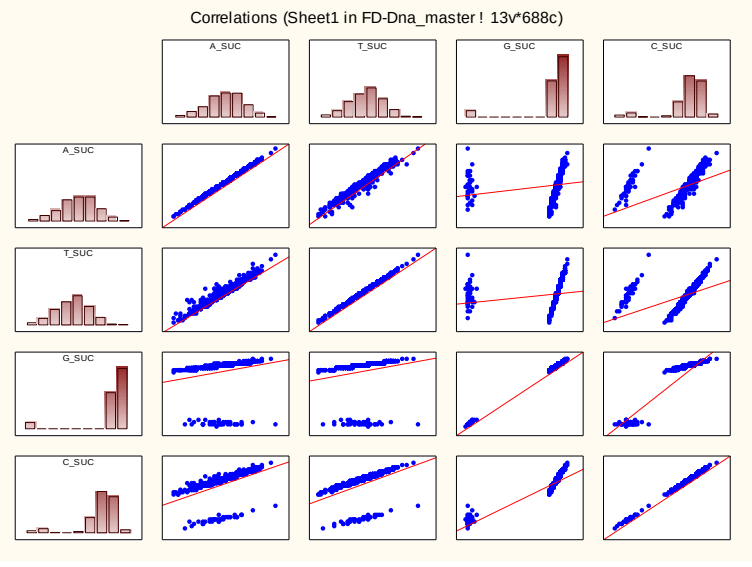
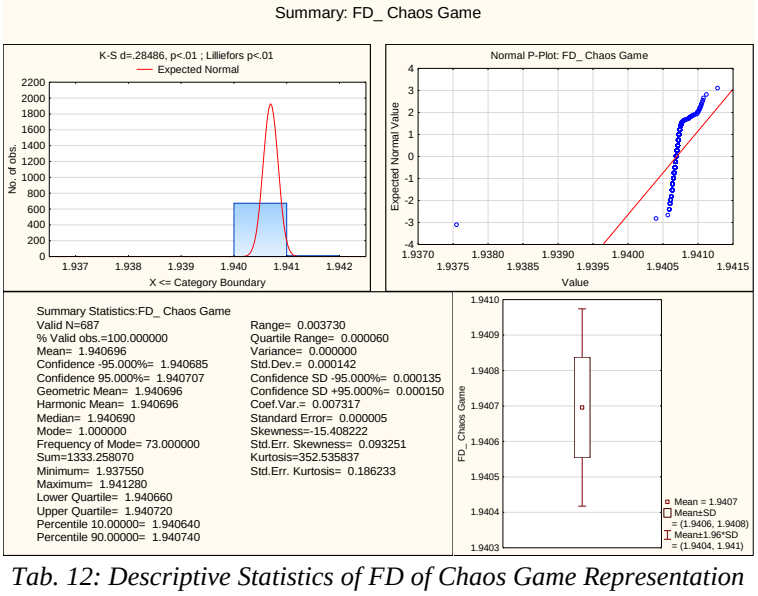


Fig. 13: Correlation coefficient among Succolarities of A, T, C, and G.

3.5 Chaos Game Representations

For the entire human ORs the fractal dimension (FD) of Chaos Game Representation are computed as shown in the following *Tab. 12*.



Tab. 12: Descriptive Statistics of FD of Chaos Game Representation

Geometric Mean and Harmonic Mean are observed to be the same. FD of Chaos Game Representation for all ORs lies in the interval (1.937, 1.9412). It is seen that the values are almost same around 1.9406. Also, this follows a non-parametric distribution; this is shown in the following figure.

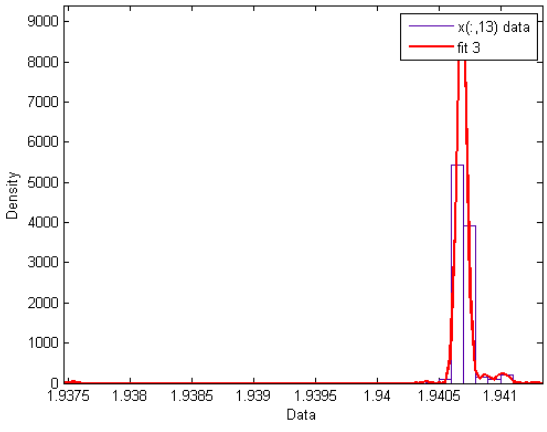


Fig. 13: Non-parametric distribution of FD of Chaos Game Representation

So far we have determined different measures as demonstrated above. Now, we make clusters of ORs based on the obtained data for the above features using K-mean clustering technique.

3.6 K-Means Clustering of Human ORs

Using K-means clustering method [21], we have clustered all 687 Human ORs into fourteen (14) different clusters. Each cluster contains more than 11 and less than 89. The mean, SD and variance are framed in the *Tab. 13.*

Variable	**Cluster 1 (88 members)**			**Cluster 2 (80 members)**			**Cluster 3 (68 members)**		
	Mean	SD	Variance	Mean	SD	Variance	Mean	SD	Variance
FD_A1	1.834781	1.834781	0.000019	1.835275	0.004594	0.000021	1.830382	0.011278	0.000127
FD_A2	1.826786	1.826786	0.000027	1.827137	0.004612	0.000021	1.821464	0.012051	0.000145
FD_A3	1.829568	1.829568	0.000038	1.832214	0.005923	0.000035	1.825728	0.012542	0.000157
FD_A4	1.822843	1.822843	0.000032	1.824691	0.005370	0.000029	1.818435	0.012652	0.000160
FD of DNA WALK	1.946009	1.946009	0.000000	1.946052	0.000109	0.000000	1.946044	0.000114	0.000000
Hurstexp_4adic	0.598911	0.598911	0.000211	0.614762	0.008670	0.000075	0.575422	0.018225	0.000332
Hurstexp_2adic	0.603058	0.603058	0.000167	0.630578	0.010301	0.000106	0.580509	0.014160	0.000201
A_SUC	1.242705	1.242705	0.000012	1.254021	0.003272	0.000011	1.230102	0.005009	0.000025
T_SUC	1.097671	1.097671	0.000017	1.111353	0.003906	0.000015	1.083019	0.005011	0.000025
G_SUC	0.885586	0.885586	0.000026	0.902679	0.004882	0.000024	0.867263	0.006263	0.000039
C_SUC	0.957255	0.957255	0.000028	0.974657	0.005153	0.000027	0.940223	0.005898	0.000035
FD_ Chaos Game	1.940688	1.940688	0.000000	1.940687	0.000035	0.000000	1.940706	0.000075	0.000000
Variable	**Cluster 4** (46 members)			**Cluster 5**(63 members)			**Cluster 6** (55 members)		
	Mean	SD	Variance	Mean	SD	Variance	Mean	SD	Variance
FD_A1	1.836022	0.016157	0.000261	1.833293	0.005561	0.000031	1.854721	0.004184	0.000018
FD_A2	1.827244	0.017272	0.000298	1.826656	0.006147	0.000038	1.847335	0.005268	0.000028
FD_A3	1.833777	0.016475	0.000271	1.827757	0.006818	0.000046	1.850491	0.005962	0.000036
FD_A4	1.825910	0.017561	0.000308	1.822145	0.006980	0.000049	1.843775	0.005530	0.000031
FD of DNA WALK	1.946092	0.000073	0.000000	1.946002	0.000191	0.000000	1.946027	0.000130	0.000000
Hurstexp_4adic	0.539689	0.021020	0.000442	0.645259	0.011145	0.000124	0.580365	0.020137	0.000405
Hurstexp_2adic	0.547715	0.020485	0.000420	0.628544	0.010468	0.000110	0.591176	0.017149	0.000294
A_SUC	1.214851	0.008163	0.000067	1.260488	0.003202	0.000010	1.240336	0.006280	0.000039
T_SUC	1.062087	0.008493	0.000072	1.120073	0.003592	0.000013	1.089476	0.007793	0.000061
G_SUC	0.841086	0.010618	0.000113	0.913591	0.004497	0.000020	0.875339	0.009747	0.000095
C_SUC	0.916435	0.010792	0.000116	0.980674	0.004643	0.000022	0.949082	0.009151	0.000084
FD_ Chaos Game	1.940710	0.000068	0.000000	1.940694	0.000047	0.000000	1.940686	0.000034	0.000000

Variable	Cluster 7 (61 members)			Cluster 8 (18 members)			Cluster 9 (42 members)		
	Mean	SD	Variance	Mean	SD	Variance	Mean	SD	Variance
FD_A1	1.854920	0.003976	0.000016	1.806145	0.011799	0.000139	1.831019	0.019870	0.000395
FD_A2	1.847557	0.004609	0.000021	1.795747	0.016003	0.000256	1.822722	0.019397	0.000376
FD_A3	1.850807	0.005329	0.000028	1.797033	0.011873	0.000141	1.825163	0.025385	0.000644
FD_A4	1.844290	0.005156	0.000027	1.792964	0.011027	0.000122	1.817218	0.025308	0.000640
FD of DNA Walk	1.946046	0.000099	0.000000	1.945938	0.000174	0.000000	1.946097	0.000124	0.000000
Hurstexp_4adic	0.623067	0.018529	0.000343	0.639222	0.014918	0.000223	0.611838	0.046162	0.002131
Hurstexp_2adic	0.628525	0.014909	0.000222	0.631217	0.021983	0.000483	0.606474	0.036003	0.001296
A_SUC	1.260482	0.005090	0.000026	1.252335	0.007545	0.000057	1.245407	0.019822	0.000393
T_SUC	1.114530	0.006187	0.000038	1.117178	0.008480	0.000072	1.102454	0.024526	0.000602
G_SUC	0.906651	0.007734	0.000060	0.909988	0.010609	0.000113	0.126203	0.018513	0.000343
C_SUC	0.977547	0.007275	0.000053	0.977679	0.011530	0.000133	0.770134	0.019892	0.000396
FD_ CGR	1.940679	0.000030	0.000000	1.940749	0.000135	0.000000	1.940713	0.000083	0.000000

Variable	**Cluster 10** (35 members)			**Cluster 11** (69 members)			**Cluster 12** (38 members)		
	Mean	SD	Variance	Mean	SD	Variance	Mean	SD	Variance
FD_A1	1.855107	0.004943	0.000024	1.832815	0.005321	0.000028	1.839379	0.011107	0.000123
FD_A2	1.848171	0.005596	0.000031	1.825642	0.005575	0.000031	1.833002	0.011760	0.000138
FD_A3	1.848677	0.006562	0.000043	1.828294	0.006155	0.000038	1.835316	0.009268	0.000086
FD_A4	1.842826	0.005423	0.000029	1.822547	0.005217	0.000027	1.829769	0.009623	0.000093
FD of DNA Walk	1.946018	0.000121	0.000000	1.946036	0.000109	0.000000	1.946027	0.000095	0.000000
Hurstexp_4adic	0.662323	0.012018	0.000144	0.658942	0.011382	0.000130	0.687568	0.011115	0.000124
Hurstexp_2adic	0.650523	0.008975	0.000081	0.650617	0.009642	0.000093	0.671032	0.010364	0.000107
A_SUC	1.275423	0.004507	0.000020	1.269536	0.003126	0.000010	1.283599	0.004323	0.000019
T_SUC	1.133572	0.005032	0.000025	1.131283	0.003682	0.000014	1.147056	0.004334	0.000019
G_SUC	0.930460	0.006294	0.000040	0.927594	0.004602	0.000021	0.947314	0.005419	0.000029
C_SUC	0.997494	0.005710	0.000033	0.994757	0.004483	0.000020	1.012250	0.005340	0.000029
FD_ CGR	1.940675	0.000034	0.000000	1.940632	0.000378	0.000000	1.940684	0.000034	0.000000

Variable	**Cluster 13** (12 members)			**Cluster 14** (12 members)		
	Mean	SD	Variance	Mean	SD	Variance
FD_A1	1.804247	0.017256	0.000298	1.755853	0.019662	0.000387
FD_A2	1.801109	0.018344	0.000337	1.747164	0.019622	0.000385
FD_A3	1.785269	0.021408	0.000458	1.744545	0.020790	0.000432
FD_A4	1.782754	0.022247	0.000495	1.739677	0.022333	0.000499
FD of DNA Walk	1.945750	0.000191	0.000000	1.946018	0.000153	0.000000
Hurstexp_4adic	0.705092	0.022784	0.000519	0.607267	0.035632	0.001270
Hurstexp_2adic	0.694867	0.015313	0.000235	0.622275	0.040351	0.001628
A_SUC	1.282116	0.009554	0.000091	1.228809	0.014985	0.000225
T_SUC	1.156956	0.011089	0.000123	1.101092	0.018642	0.000348
G_SUC	0.959758	0.013888	0.000193	0.889861	0.023318	0.000544
C_SUC	1.023424	0.012137	0.000147	0.960509	0.023229	0.000540
FD_ CGR	1.940843	0.000192	0.000000	1.941023	0.000052	0.000000

Tab. 13: Descriptive statistics of 14 clusters (K-means)

The mean of each cluster is plotted in the *Fig. 14.*

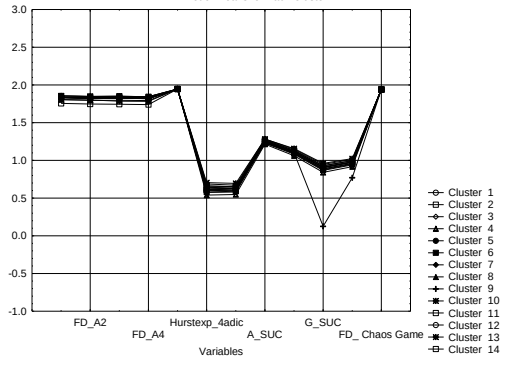


Fig. 14: Means of each clusters

Members and the distances from the centre of each cluster are available as *supplementary file* (1).

3.7. Deterministic Model Representation

We have figured out twelve features for all the human OR sequences. We have found closed-intervals for each of the above stated mathematical features. So basically, we now have a twelve dimensional rectangular model through which one can accept or reject a given sequence of reasonable length. As we proposed in the introduction, a given sequence of nucleotides can be probably justified or deterministically nullified as a human OR sequence.

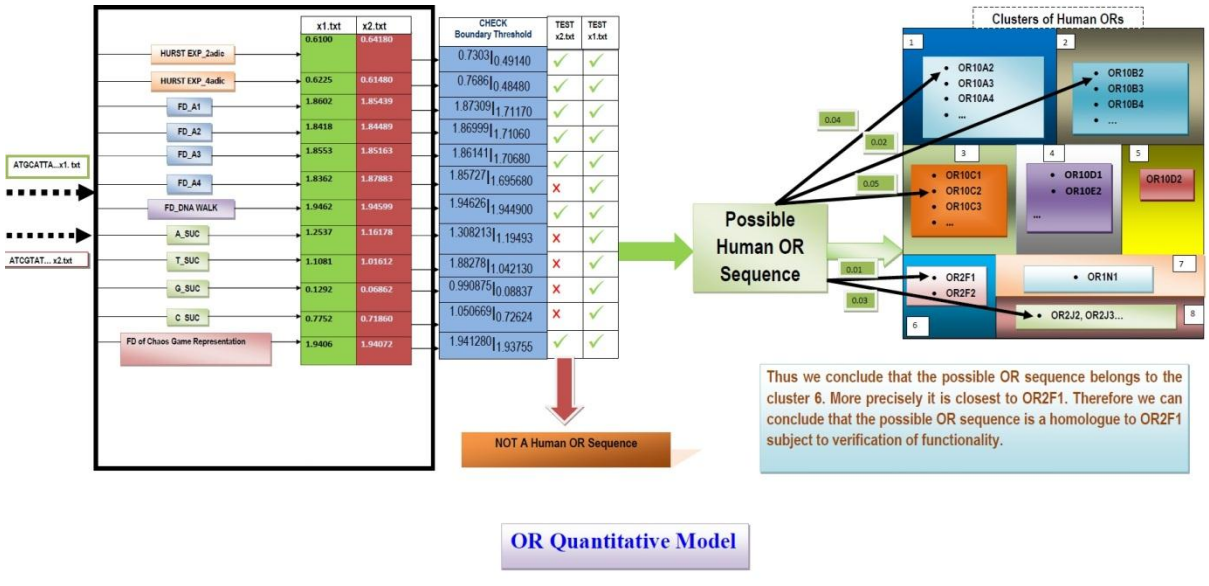


Fig. 15: A deterministic quantitative model representation

If a sequence of nucleotides does not pass through the twelve-dimensional rectangle then we readily conclude that it is not a human OR. Otherwise, the given sequence can be thought as a probable human OR homologue subject to the biological validation. Also, we have mapped the probable candidate to one or more human OR (s) by finding the minimum length from the clusters and members. The protocol of acceptance and rejection of an input sequence of nucleotides is explained in the *Fig. 15 (also available as supplementary file (2).*

4. Conclusion and Future Endeavours

In this paper, we have proposed a quantitative deterministic model through which a given string of nucleotides can be inferred as a human OR or not without seeking any biological experiment. This would help us in screening any given stretch of nucleotides of length ~1000bp as a Human OR homologue. In human ORs globe, there are almost 1:1 pseudogenes and coding genes. We are in a strong conviction that each functional OR is associated with one or many pseudogene (s) [7, 8]. This fact can be established in our future endeavours through our proposed model. It is noted that the proposed deterministic model is not only meant for human ORs but also can be treated as a standard prototype for other genes and genomes.

Authors Contributions: *Sk. S. Hassan* conceptualized the problem and experiments and performed entire research with *A. Bose and P. Pal Choudhury*. *Sk. S. Hassan and A. Bose* wrote the article. The entire work is checked by *P. Pal Choudhury*.

Acknowledgement

The authors are grateful to *Dr. Arunava Goswami and Dr. B. S Daya Sagar* of *Indian Statistical Institute, India* for their kind guidance and suggestions.

References

[1] JD Watson (1990) "The human genome project: past, present, and future" *Science* Vol. 248 (4951), 44-49.

[2] MD Mark P. Sawicki, MD Ghassan Samara[1], MD Michael Hurwitz[1] and MD Edward Passaro Jr (1993) "Human Genome Project" *Am J Surg* **165** (2), 258–264.

[3] Francis S. Collins, Michael Morgan and Aristides Patrinos (2003) "The Human Genome Project: Lessons from Large-Scale Biology", *Science* **300** (5617), 286-290

[4] Francis S. Collins and Victor A. McKusick (2001) "Implications of the Human Genome Project for Medical Science" *JAMA* **285** (5), 540.

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

[6] U. S. National Library of medicine (2012) "Genetics Home Reference" http://ghr.nlm.nih.gov/

[7] B. Malnic, PA Godfrey and L. Buck (2004) "The olfactory receptor gene family," *Proc. Natl. Acad. Sc*, **101**, 2584-2589.

[8] Xinmin Zhang and Stuart Firestein (2002) "The olfactory receptor gene superfamily of the mouse" *Nat. Neurosci* **5** (2), 124-133.

[9] P Kitts, EV Koonin, I Korf, D Kulp and D Lancet, (2001) "Initial sequencing and analysis of the human genome", *Nature*, **409**, 860-921.

[10] Yoav Gilad, Orna Man and Gustavo Glusman, (2005) "A comparison of the human and chimpanzee olfactory receptor gene repertoires," *Genome Res.* **15** (2) 224-30.

[11] G. Glusman, I. Yanai, I. Rubin and D. Lancet (2001) "The complete human olfactory subgenome" *Genome Res.* **11** (5), 685-702

[12] C. Crasto, M. S. Singer and G. Shepherd (2001) "The olfactory receptor family album", *Genome Bio.* **2** (10), 1-4.

[13] I. Gaillard, S. Rouquier and D. Giorgi (2004) "Olfactory receptors" *Cell Mol. life Sci*, **61**, 456-469.

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

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

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

[17] Sk. S. Hassan, P. Pal Choudhury and A. Goswami, (2012), "Underlying Mathematics in Diversification of Human Olfactory Receptors in Different Loci ", *Interdisc. y Sc.s: Comptnl. Life Sc.*), In Press.

[18] Sk. S. Hassan, P. Pal Choudhury, B.S. Daya Sagar, S.Chakraborty, R.Guha, and A.Goswami, (2011), "Quantitative Description of Genomic Evolution of Olfactory Receptors ", (*Under Review*).

[19] C. Carlo, (2010) "Fractals and Hidden Symmetries in DNA", *Math. Prblm. in Engng.* **2010**, 507056.

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

[21] 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.

[22] 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.