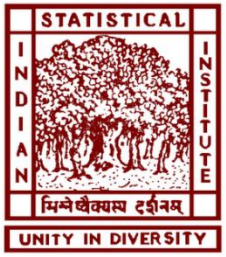
Quantitative Description of Genomic Evolution of Olfactory Receptors



Sk. Sarif Hassan
Home of Mathematical genomics,
Applied Statistics Unit,,
Indian Statistical Institute, Calcutta, India
Web: isical.ac.in/~hmg

The work is co-authored by

Pabitra Pal Choudhury, B. S. Daya Sagar, *S. Chakraborty, R. Guha and A. Goswami*

Motivation and Information

Humans recognize a gigantic variety of chemicals as having distinctive odors. Odor perception initiates in the nose, where odorants are detected by a large family of olfactory receptors (ORs).

ORs are the basis for the sense of smell, and they constitute the largest gene super family in the human genome. There are about 30,000 to 40,000 protein-coding genes in the human genome.

To gain insight into the mechanisms underlying odor perception, the amount of complexities and the quantitative differences in different genes of various species, we provide a quantitative description of three OR sequences taken from Human, Mouse and Chimpanzee.

There are many works done experimentally in different research labs across the globe but to the best of our knowledge, there is not so much of work done to decipher the quantitative content of genome.

We believe the geometry and morphology of the DNA structure are important aspects in studying their functions. We follow popular techniques to decipher quantitative aspects of DNA through fractals and mathematical morphology [3, 4, 5, 6, 7 and 8] that are employed to study many problems encountered in various branches of science and technology including the domain of biology.

In this work, We captured the evolutionary connections among ORs with the help of their quantitative descriptions.

Data Source and Data Representation

Date used and Data representation: Date are acquired from the Yale University database (ORDB: http://senselab.med.yale.edu/ordb/). This data are represented in text form which we further represented in spatial form and is coded in 2-bit (4 color). Each nucleotide is assigned with unique color such that spatially represented data show from different colors. OR for Chimpamzee is shown in Fig. 1a. Similar representations are made for other two species Human and Mouse. We consider the olfactory receptors (Ors) OR1D2, CONTIG3463.6-1888, GA_x5J8B7W3YLM-7051808 of Human, Chimpanzee and Mouse respectively for our case study [9]. It is noted that fort we have selected the OR OR1D2 from HORDE database and it was blasted in the NCBI database (http://blast.ncbi.nlm.nih.gov/Blast.cgi) to get highly similar OR sequences in Chimpanzee and Mouse and we found that the CONTIG3463.6-1888, GA_x5J8B7W3YLM-7051808 of Chimpanzee and Mouse respectively.

Some Basics

Quantitative description for ORs is done through some basics of Mathematical Morphology and Fractal Geometry. For better understanding, the basic morphological transformations are stated here.

Mathematical Morphology

- *Mathematical Morphology* based on axiomatic set theory and more relevantly lattice theory, has gained popularity out of its obvious uses in the field of image analysis which provides a quantitative description of geometrical structures.
- In this work, we apply certain morphological transformations essentially to gain the spatial distribution of different nucleotides in DNA sequence templates.
- To perform these analyses, two fundamental morphological transformations employed include morphological erosion (to shrink) and morphological dilation (to expand) as explained in equations (1) and (2).

Dilation and Erosion

$$\left(X \ominus \hat{B}\right) = \{x: B_x \subseteq X\} = \bigcap_{b \in B} X_{-b} \ldots\ldots (1)$$

$$\left(X \oplus \hat{B}\right) = \{x: B_x \cap X\} = \bigcup_{b \in B} X_{-b} \ldots\ldots (2)$$

where $B \neq \hat{B}$.

Fractal

The precise definition of "Fractal" according to Benoit Mandelbrot is, a set for which the Hausdroff Besicovitch dimension strictly exceeds the topological dimension [22].

One of the fundamental fractal parameters is 'Fractal Dimension'. There are several methods like box counting method, perimeter area dimension method and so on to compute fractal dimension of an object.

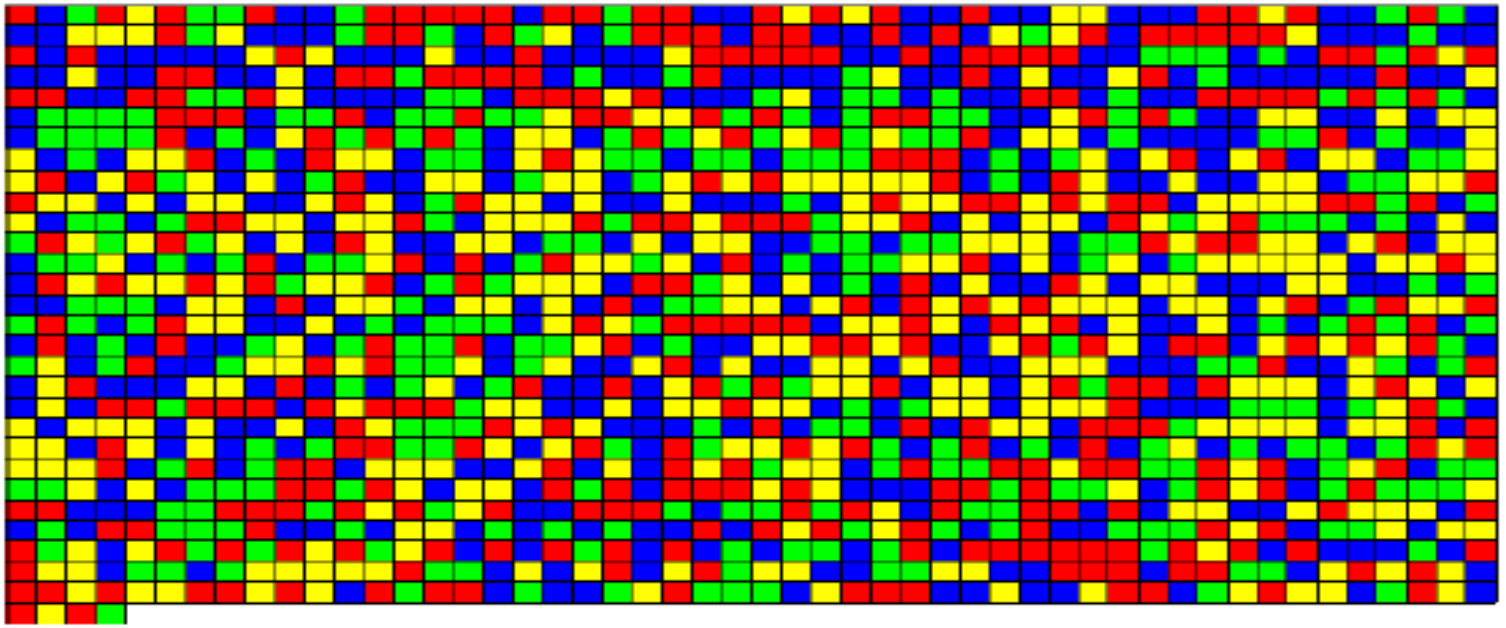
In this work, we follow Box Counting Method and is computed through well known software called **BENOIT™.**

Model Decomposition and Representation

A DNA sequence be in the form of four-letter (ATGC) nucleotides sequence (Fig. 1a). Such sequence shown in next slide as Fig. (1a) is converted as a function (Fig. 1b) depicting colors RED, BLUE, GREEN, and YELLOW respectively for A, T, G, and C.

ATGACAGGATTGAAAAATAAGAATTACACATTATTCCTTTA
ACATTGAGTTTCCCAGCTTTGAAGTAGCTGAAATAATTATA
TCGCATAAAAACTTTGTTATATTTTTCACTTTCTTATTTTC
AAAAATTATAAAATTGGGTGTAAGACATTCTTAATTCTAAG
AAAATGTTGATTTTGCTTATCTTCATGTTTTTATTCAATTA
AGGACTTTTGGTAAACATTTGCTGGTGTTAATGTTAAAAGA
GAGTTGGGGAAATGGATGGCATGGGGCTCTGGGAAGACTCC
TAGATAAACACTTTAAGAGGCT

(a)



(b)

Fig. 1. (a) a DNA sequence, and (b) function generated by proper color coding for ATGC of CONTIG3463.6-1888

Threshold Decomposition

Threshold decomposition: We decomposed the four colored image $f(x, y)$ into four binary images (Fig. 2a-d) for Chimpanzee (Fig. 1b), through the threshold decomposition function defined as:

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

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

Those decomposed binary images for chimpanzee are denoted as ${f_{CH}}^A, {f_{CH}}^T, {f_{CH}}^G \; and \; {f_{CH}}^C$. Such binary images for species Human and mouse are denoted respectively as

${f_H}^A, {f_H}^T, {f_H}^G \; and \; {f_H}^C$ (for human) and ${f_M}^A, {f_M}^T, {f_M}^G \; and \; {f_M}^C$ (for mouse)

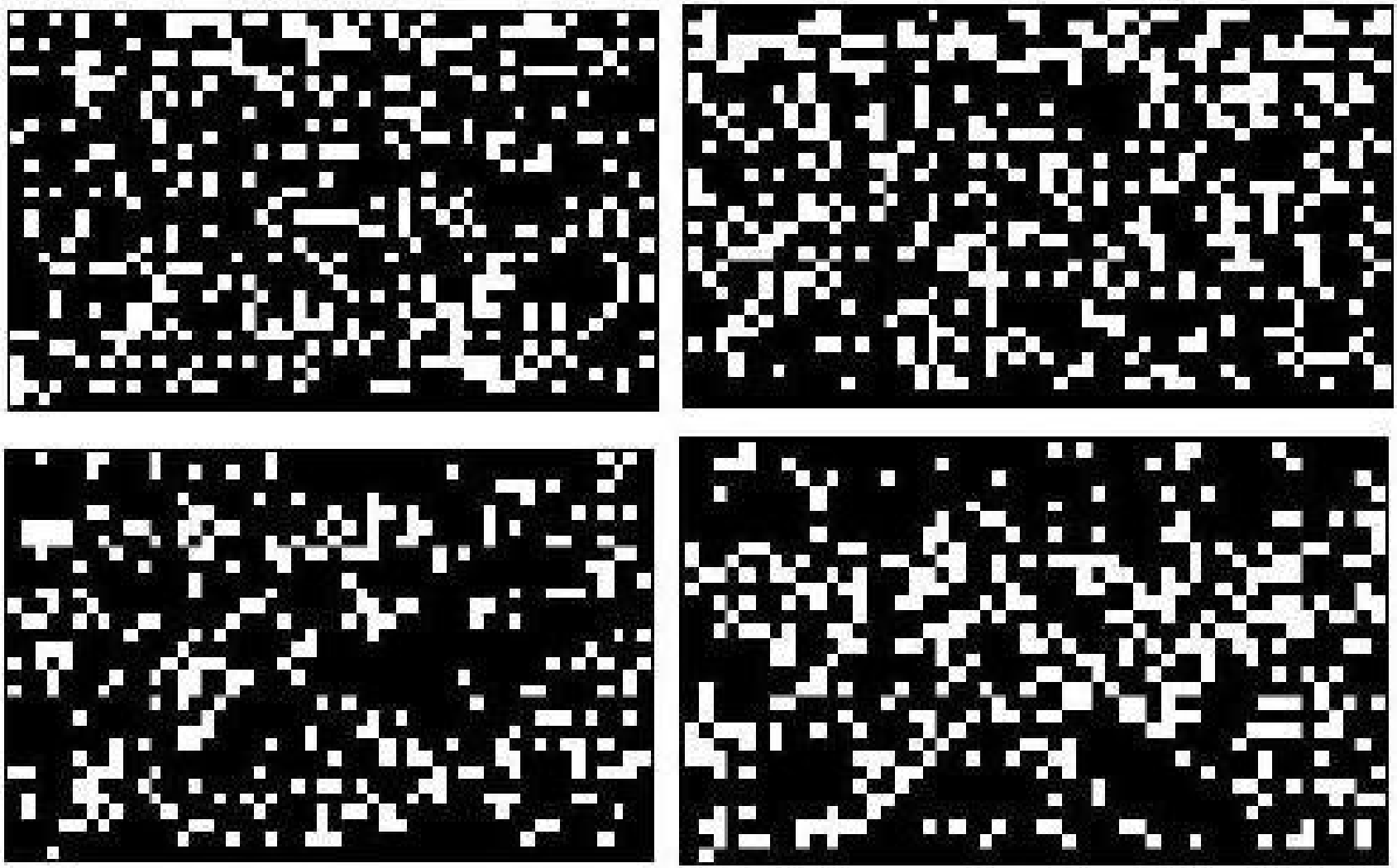


Fig. 2: Threshold decomposed binary images of ORs of Chimpanzee (Black and white denote complimentary space and one of the ATGC). (a) A (b) T (c) G, and (C).

Skeleton Decomposition

Skeleton Decomposition: Morphological skeletons (Fig. 3a-d) for threshold decomposed binary images of ORs of Chimpanzee (CONTIG3463.6-1888) shown in Fig. 2a-d are obtained according to (3).

$$SK(X) = [(X \ominus nB)(X \ominus nB) \circ B]$$

where B is a structuring element that is symmetric about the origin, and $nB = B \oplus B \oplus B \ldots \oplus B (n\ times)$.

Skeletons

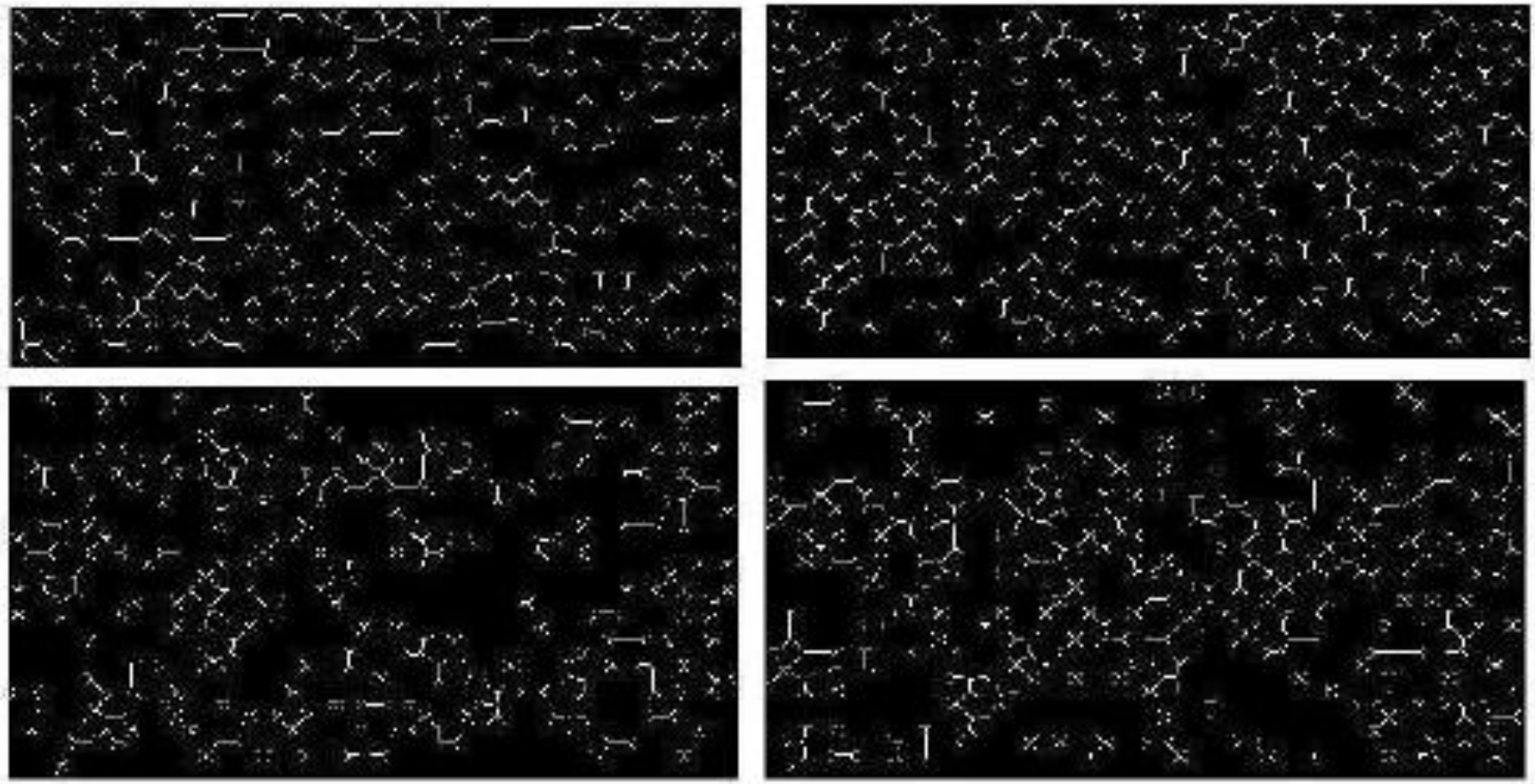


Fig. 3. Skeletons of OR Chimpanzee

Binary Decomposition

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).

11110101010010011111111000001011001001111000000011
00001111001100110110010011000000000011111111101111
00110011111111110100011111110111110011111111010000
00000011110011000000001111101010111011000010000100
1111011111…..(some more 0, 1 are there in the string).

Methods

- The techniques employed to derive indexes are from the fields of mathematical morphology and fractal geometry further to retrieve geometric characteristics in quantitative manner.
- In particular, we employed skeletonization and granulometries from mathematical morphology and fractal dimension, Hurst exponent, succolority measures from fractal geometry to derive six quantitative indexes that also include poly-string mean and standard deviation.
- Data represented in spatial form enables rich geometric characteristics.

Poly String Mean and SD

Poly-String Mean and Standard Deviation (A classification method): Let total number of poly-strings of different size $k_1, k_2, k_3, k_4 \ldots . k_n$ of nucleotide 'N' are $m_1, m_2, m_3, m_4 \ldots , m_n$. Then poly-string mean $\left(P_m{}^N\right)$ and poly-string standard deviation $(P_{SD}{}^N)$ of N are defined as

$$P_m{}^N = \frac{\sum_{i=1:n} m_i k_i}{\sum_{1:n} m_i} \text{ and } P_{SD}{}^N = \sqrt{\frac{1}{n}\sum_{i=1}^{n} m_i (k_i - P_m)^2}$$

After calculating four nucleotides poly string mean and standard deviation will get a strict ordering relation among the $P_M{}^N$ and $P_{SD}{}^N$. We can classify any DNA sequence according to the order of $P_M{}^N$ and $P_{SD}{}^N$.

Fractal Dimension

Fractal Dimension of Indicator Matrix: We have plotted a DNA sequences in two axes and defined a indicator map $f: \{X, Y\} \rightarrow \{0, 1\}$ *as*

$$f(X, Y) = 0 \; if \; Y \neq X$$

$$= 1 \; otherwise$$

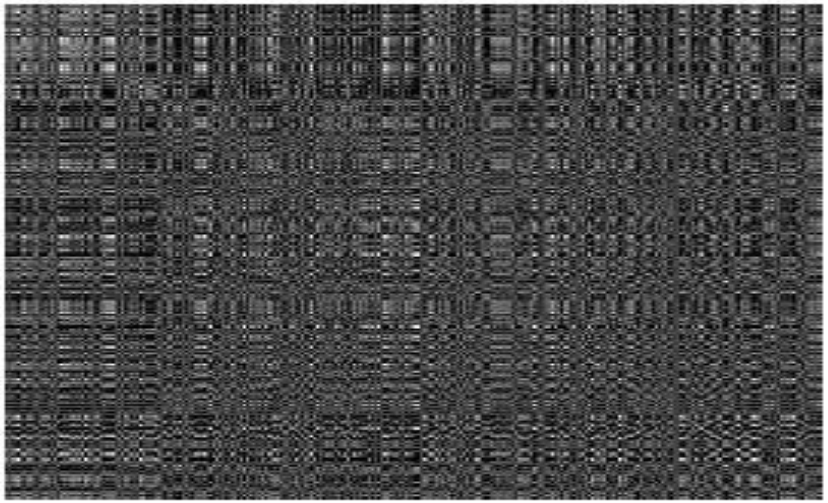


Fig-4: Indicator matrix

Then we have calculated the box-counting dimension of the indicator matrix on using the *Benoit* software.

Hurst Exponent

Let a string $x = \{x_i\}_{i=1\ to\ n}$, defines the following entities regarding the sequence as:

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

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

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

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

Then Hurst Exponent (H) is defined as $\left(\frac{n}{2}\right)^H = \frac{R(n)}{S(n)}$

In our case 'n' denotes the length of the binary strings corresponding to each DNA sequence and x_i are the binary digits of the strings. The Hurst exponent varies between 0 and 1. In our present study we can calculate the Hurst exponent for any binary string (as shown in III (A)).

Succolarity

Succolarity index: 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. The succolarity of a binary image is defined below [22]

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

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 (BS(k)), 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.

Morphometric Dimension

It is conspicuous that the skeletal networks (Fig. 3a-d) decomposed from threshold decomposed binary images (Fig. 2a-d) of ORs of Chimpanzee possess branching pattern. The open-ended segments in the skeleton network are designated as first order segments, and second order segments begin from the point where two first order segments join, and where two second order segments join a third order segment begins, and so on. If any lower order segment joins to a higher order segment, higher order designation is maintained.

We designated the skeletal networks according to this segment ordering scheme, and we computed order-wise number of skeletal segments , and order-wise mean lengths of skeletal segments . By using these two quantities we estimated bifurcation ( ) and mean length ( ) ratios of order-designated segments as follows:

( ) and

Further, we estimated fractal dimension using the ratio of logarithms of these topological quantities (i.e.———). For the four skeletal networks decomposed from four threshold decomposed binary images (Fig. 3a-d), the estimated fractal dimensions are given in Fig. 7.

Morphological Entropy

Here let us see how A, T, G, C are spatially distributed in a DNA sequence. We need to calculate the area (# of blocks) of the images by successive morphological opening on

We then calculate the probability through the probability density function:

$$(\mid) \quad \frac{(\qquad)}{\qquad\qquad}$$

where X is the decomposed images of ( ). The *Morphological Entropy* (ME) is defined as – where N is defined as

[( )] [(( ))] .

Results and Discussions

We estimated six quantities viz. poly-string mean, poly-string standard deviation, Hurst exponent, fractal dimension, succolarity index and morphological entropy for ORs of three species namely Human, Mouse and Chimpanzee.

Here we will discover quantitative connections through the measures, described in the section-III.

Evolutionary Connection of ORs of Mouse and Chimpanzee with Human ORs

It is to be noted that we have classified all the human ORs based on classification methodology on the poly-string mean and standard deviation as proposed in [15].

Using the same we have classified OR1D2 (Human), GA_x5J8B7W3YLM-7052533-7051808 (Mouse) and CONTIG3463.6-1888(Chimpanzee) and the results are shown in table-I.

Olfactory Receptors	***Class According to Poly-String Mean/SD***	***Hurst Exponent (H)***	***Maps to OR W.r.to H***
OR1D2	CGTA/CGAT	0.598911	OR1D2
GA_x5J8B7W3YLM-7052533-7051808	GCTA/GCAT	0.645594	OR4D2
CONTIG3463.6-1888	CGAT/ACTG	0.539152	OR3A3

Table-1: Evolutionary Connection of ORs with Human

Inference

- The Mouse OR (GA_x5J8B7W3YLM-7052533-7051808) maps to a human OR OR4D2 based on classification and closest Hurst exponent.
- But it is to be noted that GA_x5J8B7W3YLM-7052533-7051808 is more similar to OR1D2. But as far as Hurst exponent is concerned (amount of long range correlation in the sequence) the mouse OR maps to OR4D2.
- In this connection, it is our strong conviction that, OR4D2 and OR1D2 are structurally similar in sequence despite the fact that they belong to different families as per HORDE qualitative classification. Also we discovered that mouse and human ORs are significantly similar in structure and in function.
- The Chimpanzee OR (CONTIG3463.6-1888) maps to a human OR OR3A3 according to the classification (shown in table –I). Although OR3A3 and OR1D2 belong to different families but with respect to evolution in connection with Chimpanzee OR CONTIG3463.6-1888, they are structurally almost same as per quantification shown above.

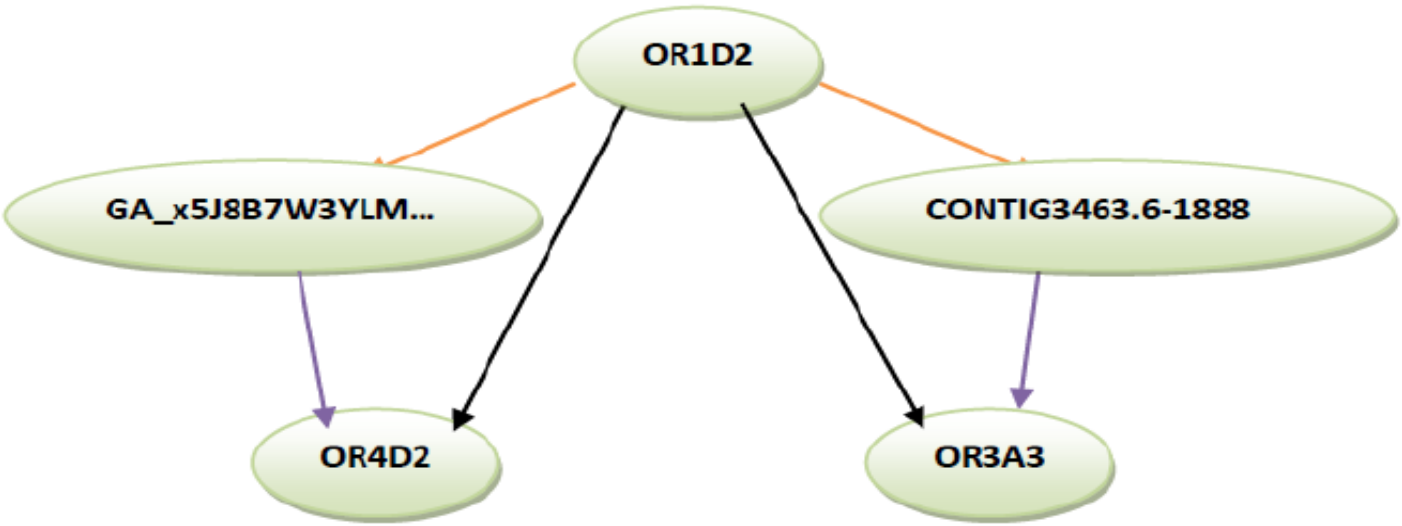


Fig-5: Evolutionary connect ion among Human, Mouse and Chimpanzee ORs

OR1D2, GA_x5J8B7W3YLM-7052533-7051808 and CONTIG3463.6-1888 are also most similar to OR4D2 and OR3A3 as shown above. They are evolutionarily connected and hence through biological evolution CONTIG3463.6-1888 and GA_x5J8B7W3YLM-7052533-7051808 are updated as OR3A3 and OR4D2 respectively.

Fractal and Morphological Quantification of ORs

- The fractal dimensions of DNA nucleotide sequences---- OR1D2 CONTIG3463.6-1888, GA_x5J8B7W3YLM-7052533-7051808 of Human, Chimpanzee, and Mouse respectively-----generated by plotting the sequences in two axes (fig.-II) and that yields respectively 1.77687, 1.81916 and 1.82963.
- Here we observe that fractal dimensions of ORs of Chimpanzee and Mouse are significantly similar. Through genomic evolution they got updated into OR1D2 in human the fractal dimension of is reduced by a small amount 0.04 i.e. through genomic evolution amount of complexity or disorderliness got decremented.

Results on Succolarity

For three olfactory receptor DNA sequences of Human, Mouse and Chimpanzee succolarity index are calculated. A DNA sequence can be thought of as a texture of four disjoint templates of A, T, C and G. For their four different templates of each DNA sequence the succolarity for each of those three sequences are shown in fig.-6.

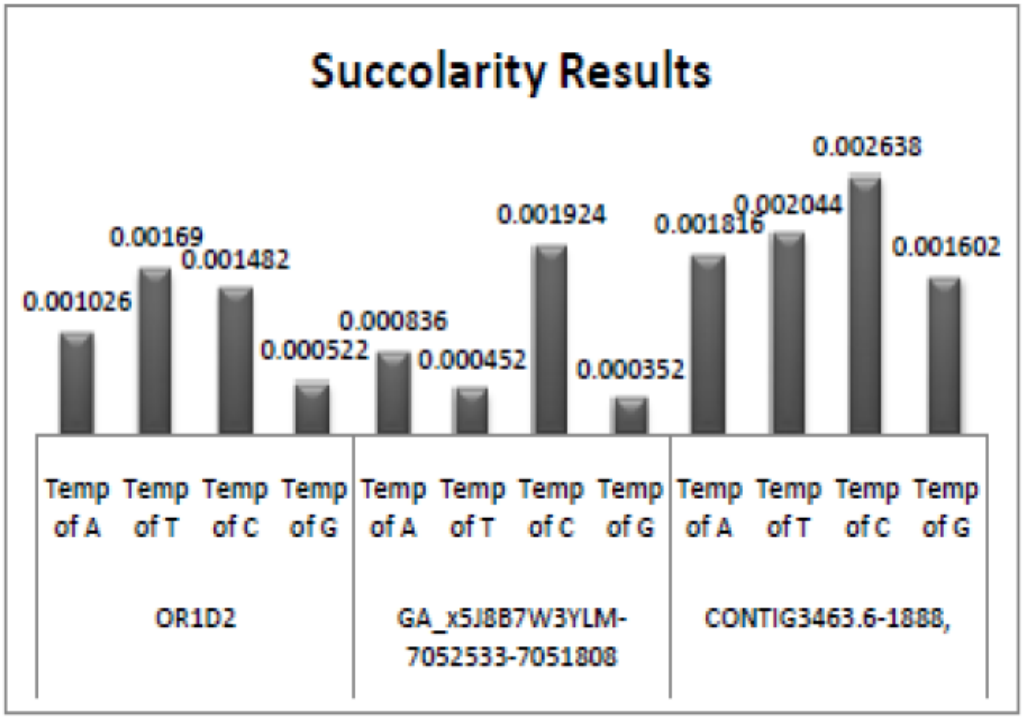


Figure-6: Succolarity of ORs

The succolarity of all the textures of A, T, C, and G are almost same for Mouse and Chimpanzee ORs but in case of Human OR OR1D2 are less than that of other two ORs. It is seen that over genomic evolution the succolarity (amount of continuous density) in sequence structure in Human OR is smaller than the other similar sequences of Mouse and Chimpanzee.

Results on Morphometric Dimension of Skeleton

- Morphometry-based fractal d*imensions* for four skeletons are computed and found similarity among the species as shown in Fig.-7.

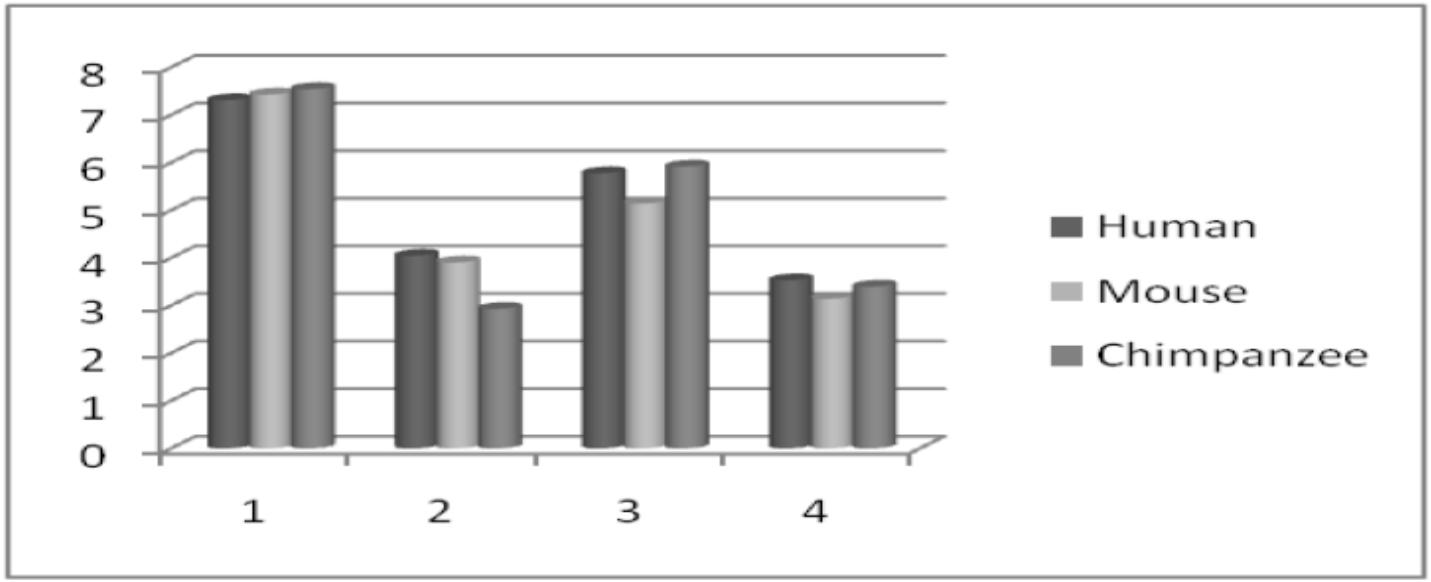


Fig.-7: Histogram of Morphometric Dimensions.

In Fig.-7, it is apparent that they do not follow a strict order. We believe this parameter provide a distinction between the functions of ORs.

Results on Morphological Entropy (Spatial Distribution)

Itern./ Area	*0*	*1*	*2*	*3*	*4*	*5*	*6*
f_H^A	170	26	6	1	1	0	
f_H^C	189	67	15	4	2	0	
f_H^G	147	39	8	2	0		
f_H^T	214	36	8	1	0		
f_M^A	120	19	5	2	0		
f_M^C	165	52	7	2	1	0	
f_M^G	101	30	6	2	0		
f_M^T	172	34	6	2	0		
f_{CH}^A	278	59	21	8	5	1	0
f_{CH}^C	256	81	17	5	3	0	
f_{CH}^G	221	59	14	2	0		
f_{CH}^T	317	79	20	6	2	0	

Table-2: Area over successive opening

From the above Table-2, we then have calculated the ME for all the decomposed images of $f(x, y)$ as shown below in the Fig. 8.

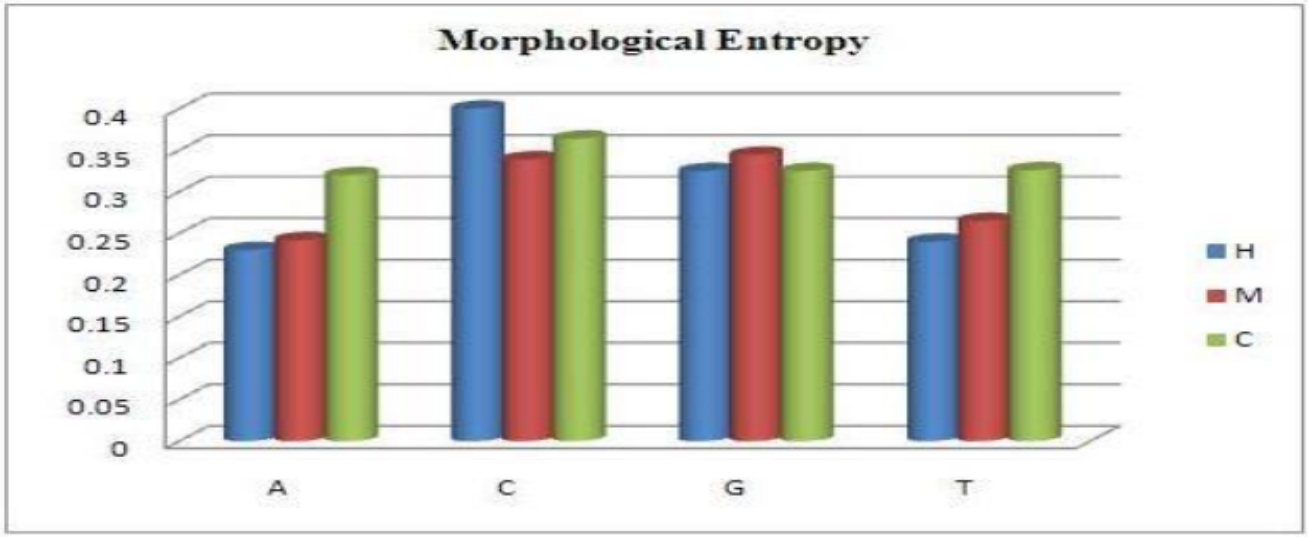


Fig.-8: ME of decomposed images of $f(x, y)$

From the Fig-8, it is noticed that the morphological entropy of the decomposed images of A and T for all the species are almost same. In other words we can say that the spatial distribution of A and T in three sequences are almost same due to the fact of homologue relationship among them. But the spatial distribution of C and G are different from others individually, and it is our strong conviction that this differentiation makes them distinct species-wise.

Conclusion

In the present work, we have shown an evolutionary connection among Human, Mouse and Chimpanzee ORs. These sequences have very close sequential similarity but they do different in different species due to their intricate details of the structures in the DNA sequence. Those intricate details are illustrated here. In our further studies, we provide a quantitative classification based on Fractals and Mathematical Morphology with some more details about all the ORs of Human, Chimpanzee and Mouse.

On Going ….
Cancer Quantitative Genomics

In recent past it is reported that a genomic sequence (PTPN11/Shp2 of Human) is acting as Oncogene (responsible for cancer) as well as it works as tumour-suppressor function in liver.

This is a first kind of report published in the journal *Cell*.

We have made an effort to characterize the genomic landscape in the gene sequence, quantitatively through Fractal and Mathematical morphology.

Acknowledgment

We are grateful to **Prof. Jean Serra,** Emeritus Professor, ESIEE-Engineering, University Paris-Est, Former Director of Centre for Mathematical Morphology, France for his kind advice and suggestions and also express their gratitude to **Mr. Pratap Vardhan** for his technical help in writing programs.

References

[1] B. Malnic and PA Godfrey, "The olfactory receptor gene family," *Proc. Natl. Acad. Sc*, Vol. 101 pp. 2584-2589 2004.

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

[3] G Matheron and J Serra, "History of Mathematical Morphology", *http://cmm.ensmp.fr/~serra/pdf/birth_of_mm.pdf*

[4] B. S. Daya Sagar "Fractal relations of a morphological skeleton", *Chaos, Solitons & Fractals*, Vol. 7, no. 11, pp. 1871-1879.

[5] B. S. Daya Sagar, et al, "Morphometric relations of fractal-skeletal based channel network model", "*Discrete Dynamics in Nature and Society*", Vol 2, no. 2, pp. 77-92. 1998

[6] P. Radhakrishnan, B. S. Daya Sagar, and Teo Lay Lian, "Estimation of fractal dimension through morphological decomposition", *Chaos, Solitons & Fractals*, Vol 21, no. 3, pp. 563-572. 2004

[7] B. S. Daya Sagar and Tay Lea Tien, " Allometric power-law relationships in a Hortinian fractal digital elevation model," *Geophysical Research Letters*, Vol. 31, no. 6, L06501.

[8] Shih and Mitchell "Threshold Decomposition of Gray -Scale Morphology into Binary Morphology", *IEEE transactions on Pattern Analysis and Machine Intelligence* Vol. 11, No. 1. Pp. 31-42, 1989.

[9] Yoav Gilad, Orna Man and Gustavo Glusman, "A comparison of the human and chimpanzee olfactory receptor gene repertoires," *Genome Res*. Vol. 15 no.2 pp. 224-30, 2005.

[10] J. G. Venkateswaran, B. Song, and T. Kahveci, "A Tool for Finding Distant Structural Similarities," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*," Vol: 8 no.3 pp: 819-831, 2011.

[11] G. Cardona, M. Llabres, and F. Rossello, "Comparison of Galled Trees," *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"* Vol. 8 no: 2 pp: 410-427, 2011

References

[12] J. Stoye and R. Wittler, "A Unified Approach for Reconstructing Ancient Gene Clusters," *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"* Vol. 6 no: 3 pp. 387-400, 2009

[13] Q. Zhu, Z. Adam and V. Choi, "Generalized Gene Adjacencies, Graph Bandwidth, and Clusters in Yeast Evolution," *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"*Vol. 6 no: 2 pp. 213-220, 2009

[14] R. Boscolo, C. L James and P. V. Roychowdhury, "An information theoretic exploratory method for learning patterns of conditional gene coexpression from microarray data." *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"* Vol. 5 no: 1 pp. 15-24, 2008

[15] A. Phaedra; B. N. Kreiswirth, and N. Steve, "Typing Staphylococcus aureus using the spa gene and novel distance measures," *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"* Vol. 4 no. 4 pp: 693-704, 2007

[16] M. A. Alekseyev and P. A. Pevzner, "Colored de Bruijn graphs and the genome halving problem," *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"* Vol. 4 no. 1 pp. 98-107, 2007

[17] O. Abul, R. Alhajj and F. Polat, "A powerful approach for effective finding of significantly differentially expressed genes," *"IEEE/ACM Transactions on Computational Biology and Bioinformatics"* Vol. 3 no. 3 pp: 220-231, 2006.

[18] J. Serra, "Image Analysis and Mathematical Morphology", *Academic Press, Inc.* Orlando, FL, USA ©1983 ISBN: 0126372403. 1982.

[19] C. Carlo, "Fractals and Hidden Symmetries in DNA", *Mathematical Problems in Engineering* Vol. 2010, Article ID 507056.

[20] P. P. Choudhury, S.S. Hassan, S. Sahoo and B. K. Nayak, "Carry *Value Transformation (CVT): It's Application in Fractal formation"*, *Advance Computing Conference*, pp. 971 – 976, 978-ISBN: 1-4244-2927-1, IEEE. 2009.

References

[21] P. Choudhury, S.S. Hassan, S. Sahoo and B. K. Nayak, "Act of CVT and EVT In the Formation of Number-Theoretic Fractals", *International Journal of Computer Cognition*, Vol 9, No. 1, pp. 1-8 2011.

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

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

[24] Yu Zu-Guo, "Fractals in DNA sequence analysis", *Chinese Physics,* Vol 11, no. 12, pp.1313-1318, 2002.

Thanks