Underlying Mathematics in Diversification of Human Olfactory Receptors in Different Loci

SK. SARIF HASSAN, PABITRA PAL CHOUDHURY,

AND ARUNAVA GOSWAMI

Indian Statistical Institute, Kolkata, India

Abstract. As per conservative estimate, approximately (51-105) Olfactory Receptors (ORs) loci are present in human genome occurring in clusters. These clusters are apparently unevenly spread as mosaics over 21 pair of human chromosomes. Olfactory Receptor (OR) gene families which are thought to have expanded for the need to provide recognition capability for huge number of pure and complex odorants. ORs form the largest known multi-gene family in the human genome. Recent studies have shown that 388 full length and 414 OR pseudo-genes are present in these OR genomic clusters. In this paper, the authors report a classification method for all human ORs based on their sequential quantitative information like presence of poly strings of nucleotides bases, long range correlation and so on. An L-System generated sequence has been taken as an input into a star-model of specific subfamily members and resultant sequence has been mapped to a specific OR based on the classification scheme using fractal parameters like Hurst exponent and fractal dimensions.

Categories and Subject Descriptors: Computational Genomics, Olfaction and Molecular Biology

General Terms: Theory and Results

Additional Key Words and Phrases: L-System, Olfactory Receptors (ORs), Star Model, Hurst exponent, Poly-String Mean and Standard Deviation

Authors' addresses: Sk. Sarif Hassan and Pabitra Pal Choudhury, Applied Statistics Unit, Indian Statistical Institute, Kolkata 700108, India. e-mail: sarimif@gmail.com, pabitrapalchoudhury@gmail.com. Arunava Goswami, Biological Sciences Division, Indian Statistical institute, Kolkata 700108, India. e-mail: agoswami@isical.ac.in

1. *Introduction*

Humans can detect a very large number of odor molecules (water or air borne; pure or mixed) and give output as distinct sets. Nasal epitheliums house a large repertoire of seven transmembrane domain G protein coupled olfactory receptor (ORs). These receptors have been described to have diverse protein sequences. Researchers have classified these receptors based on simple computer based alignment tools them into several families and subfamilies. HORDE database used in this study have used a nomenclature which is not sufficient enough to cover all the kinds of receptors that have been found in nature inside human nose. More importantly, SNPs and genomic variant studies have not been performed to assign the classification very accurately. One of many advantages of the OR sequences is that they do not possess introns. Therefore, one can use them as model systems for exons as well. By a conservative estimate, it is known that there are 388 intact and 414 OR pseudogenes (Menashe I et al. [2006] and Yoshihito Niimura et al. [2003]). Researchers have reported that they are unevenly distributed among 51-105

different loci on 21 human chromosomes where each locus contains 1-3 genes (Gustavo Glusman et al [2000]). The present authors do not believe for many reasons that these are uneven distributions. It is a strong conviction of the authors that there is beautiful organization followed by some mathematical governess in diversification of human ORs in different loci. Using a number of mathematical methodologies, in this paper, the authors report that it is possible to find relationship between apparently unrelated ORs located in different chromosome. This hitherto has not been studied and this is the first report to the best of knowledge of authors.

2. *Methods and Results:*

(2A) Methods and Materials:

Any DNA sequence is clearly consisting of four textures of A, T, C, and G (nucleotide cluster) as shown in the color template (Appendix-I). There are several poly-strings having lengths 1, 2, 3, etc in all the textures of any base A, T, C and G. The frequencies of different poly-strings of different lengths for each nucleotide cluster for each OR have been calculated. Poly-string mean and standard deviation for each nucleotide cluster of A, T, C and G have been enumerated. A decreasing ordering of poly-string mean and standard deviation of different nucleotide textures were arranged. Results show that all ORs could be classified into 21 different classes using these two deterministic statistical parameters (*Data available as supplementary material-I (A) & (B)).*

Each of A, T, C and G of all OR sequences have been encoded by the following two bit information.

$A \to 00, T \to 11, C \to 01, G \to 10$. The reason for such encoding is that A pairs with T and G with C.

As for an example: AGTCG have been encoded into 0010110110.

Hurst exponent for each of the encoded human OR sequences were then calculated (*Data available as supplementary material-II).*

Without loss of generality, it was assumed that all subfamily members (exons) namely OR1D2, OR1D4, and OR1D5 subfamily members of D family (as HORDE nomenclature). Then the star model of those sequences was been extracted (Sk. S. Hassan et al. [2010]). An L-System generated sequence can be taken as an input into the star model to get a full length sequence (methodology is shown in (Sk. S. Hassan et al. [2010]). The resultant sequence could be classified based on the parameters poly-string mean and standard deviation. Now the Hurst exponent was employed for exact quantification of any one or more ORs of the mapped classes.

(2B) Results and Analysis*:*

(I) Mapping Between Resultant Sequence and a Full Length OR

The star model for the ORs namely OR1D2, OR1D4, OR1D5 have been extracted which is shown below. The process of extracting a star model is shown in (Sk. S. Hassan et al. [2010]). In the following star model there are 108 mismatches which are shown as hyphened in the following star model.

```
ATGGATGGAG--AACCAGAGTGA----TCA-AGTTCCTTCTCCTGGGGAT....50
-TCAGAGAGTCCTGAGCAGCAGC-GATCCTGTTTTGGATGTTCCTGTCCA...100
TGTACCTGGTCACGGTG-TGGGAAATGTGCTCATCATCCTGGCCATCAGC...150
TCTGATTCCC-CCTGCACACCCCC-TGTACTTCTTCCTGGCCAACCTCTC...200
CTTCACTGACCTCTTCTTTGTCACCAACACAATCCCCAAGATGCTGGTGA...250
AC-TCCAGTCCCA-AACAAAGCCATCTCCTATGCAGGGTGTCTGACACAG...300
CTCTACTTCCTGGTCTCCTTGGTG-CCCTGGACAACCTCATCCTGGC-GT...350
GATGGC-TATGA-CGCTATGTGGCCA-CTGCTGCCCCCTCCACTA---CA...400
CAGCCATGAGCCCT--GCTCTGT-TCTT-CTCCT-TCCTTGTGTTGGG--...450
CT-TC-GT-CTCTATGGCCTC-T-C-CACC-TCCTC-TGACCAG-GTGAC...500
CTTCTGTGGG-C--GA-A-ATCCACTAC-TCTTCTGTGA-ATGTA--T--...550
TGCTG-GG-TGGCATGTTCCAACA--CA-AT-A-TCACACAG-G-TGATT...600
```

```
GCCAC-GGCTGCTTCATCTTCCTCA--CCCTT-GG-TTC-TGA-CA--TC...650
CTATGT-CG-ATT-TCAGA-CCATCCT---AAT-CCCTC-G-CTCTAAGA...700
AATACAAA-CCTTCTC-ACCTGTGCCTCCCATTTGGGTG--GTCTCCCTC...750
TT-TATGGGA--CTT--TATGGT-TACCT--AGCCCCTCCATACCTACTC...800
--TGAAGGACTCAGTAGCCACAGTGATGTATGCTGTG-TGACACC-ATGA...850
TGAA-CC-TTCATCTACAG-CTGAGGAACAA-GACATGCATGGGGCTC-G...900
GGAAGA-TCCTA---A-AC-CTTT-AGAGGC--A-A...936
```

Fig.1A: Star model of OR1D subfamily of OR gene sequences.

An L-System (shown below) is used to generate a sequence and that is being inputted into the above star model as proposed in (Sk. S. Hassan et al. [2010]). Consequently we got the following resultant sequence (RS-1) shown in fig 1B.

L-System:

Axiom: A

Production Rules: A→CTG, C→CCA, T→TGC, G→GAC.

```
ATGGATGGAGCCAACCAGAGTGAGTCCTCACAGTTCCTTCTCCTGGGGATGTCAGAGAGTCCTGAGCAGCAGCAGATCCTGTTTTGGATGTT
CCTGTCCATGTACCTGGTTCACGGTGCTGGGAAATGTGCTCATCATCCTGGCCATCAGCTCTGATTCCCCCCTGCACACCCCCGTGTACTTC
TTCCTGGCCAACCTCTCCTTCACTGACCTCTTCTTTGTCACCAACACAATCCCCAAGATGCTGGTGAACCTCCAGTCCCAGAACAAAGCCAT
CTCCTATGCAGGGTGTCTGACACAGCTCTACTTCCTGGTCTCCTTGGTGACCCTGGACAACCTCATCCTGGCCGTGATGGCCTATGATCGCT
ATGTGGCCAGCTGCTGCCCCCTCCACTACGCCACAGCCATGAGCCCTGCGCTCTGTCTCTTCCTCCTGTCCTTGTGTTGGGCGCTGTCAGTC
CTCTATGGCCTCCTGCCCACCGTCCTCATGACCAGCGTGACCTTCTGTGGGCCTCGAGACATCCACTACGTCTTCTGTGACATGTACCTGGT
GCTGCGGTTGGCATGTTCCAACAGCCACATGAATCACACAGCGCTGATTGCCACGGGCTGCTTCATCTTCCTCACTCCCTTGGGATTCCTGA
CCAGGTCCTATGTCCCCATTGTCAGACCCATCCTGGGAATACCCTCCGCCTCTAAGAAATACAAAGCCTTCTCCACCTGTGCCTCCCATTTG
GGTGGAGTCTCCCTCTTATATGGGACCCTTCCTATGGTTTACCTGGAGCCCCTCCATACCTACTCCCTGAAGGACTCAGTAGCCACAGTGAT
GTATGCTGTGGTGACACCCATGATGAACCCGTTCATCTACAGCCTGAGGAACAAGGACATGCATGGGGCTCAGGGAAGACTCCTACGCAGAC
CCTTTGAGAGGCAAACA
```

[**Fig 1B**: The resultant sequence (RS-1)]

We have found the following data corresponding to the Resultant Sequence (RS-1):

Nucleotide Texture (NT)	*Poly-String Mean (PSM)*	*Poly String Standard Deviation (PS-SD)*
A	1.176471	0.459008
T	1.195000	1.195000
C	1.594872	0.838175
G	1.344156	0.658411

[**Table-I**: Poly String mean and SD for the resultant sequence]

Therefore the decreasing ordering based on mean and SD is $C > G > A > T$ (CGAT) and $C > G > T > A$ (CGTA) respectively. Also the Hurst exponent of the encoded two bit sequence corresponding to the RS-1 is **0.613191**. Consequently the RS-1 is mapped into the class CGAT and CGTA based on mean and SD respectively (See the supplementary material-I). Now, the sequences OR1D2, OR1D4, OR1D5, and OR1N2 belong to the union of two classes CGAT and CGTA. According to Hurst exponent the resultant sequence is mapped into OR1N2 (See the supplementary material-II).

In the same manner as above some other results are given below:

Subfamily Members	*L-System*	*Data*			*Hurst Exponent of RS*	*Mapped OR*
OR1D2, OR1D4, OR1D5, OR1E1	Axiom: A A->CTGCTG C->CCACCA T->TGCTGC G->GACGAC	***NT***	***PSM***	***PS-SD***	0.608632	OR1R1P
		A	1.182390	0.460453		
		T	1.208556	0.531702		
		C	1.656388	1.080867		
		G	1.229508	0.524590		
OR1D2, OR1D4, OR1D5, OR1E1, OR1E2	Axiom: A A->ATGC C->CGCG T->GTCG G->CCCA	***NT***	***PSM***	***PS-SD***	0.578986	OR1ABAP
		A	1.256000	0.618437		
		T	1.200000	0.445769		
		C	1.672646	1.094428		
		G	1.349693	0.678628		
OR1D2, OR1D4, OR1D5, OR1E1, OR1E2, OR1E6	Axiom: A A->ATGC C->CGCG T->GTCG G->CCCA	***NT***	***PSM***	***PS-SD***	0.58051	OR1J2
		A	1.221311	0.535805		
		T	1.351485	0.605440		
		C	1.538462	0.939753		
		G	1.356643	0.662751		
OR1F1, OR1F2, OR1F12	Axiom: A A->GCGC C->CCTC T->AAAG G->GAAC	***NT***	***PSM***	***PS-SD***	0.546249	OR1Q1
		A	1.253247	0.553169		
		T	1.237180	0.717118		
		C	1.663934	1.083487		
		G	1.265306	0.563354		
OR1F1, OR1F2, OR1F12, OR1I1	Axiom: A A->GCAC C->CTCC T->GGGC G->CTTC	***NT***	***PSM***	***PS-SD***	0.511355	OR1F2
		A	1.114286	0.397954		
		T	1.237374	0.681070		
		C	1.807087	1.215949		
		G	1.279221	0.540523		

[**Table-II**: Detail results]

In the above table, it is observed that a star model of a set of subfamily members together with the contribution of an L-System would lead to a resultant sequence which is mapped to either a pseudo-gene or a full length gene. It is worth noting that all star models of full-length genes have been considered. Interestingly the contribution of L-System input into the star model would map to either a full length gene or a pseudo-gene. Also an interesting significant observation is that a full gene could be mapped to either full length gene or pseudo gene and same is true for pseudo gene too.

3. Conclusion

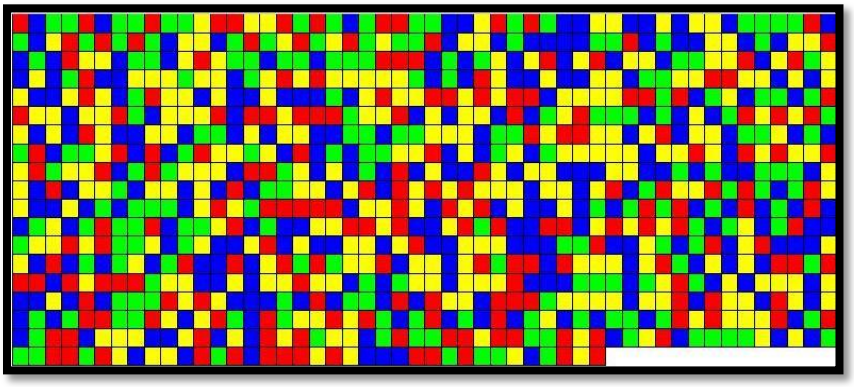
In this paper a classification scheme is explored based on fractal and deterministic statistical parameters. Starting from a particular set of sequences which are taken from a subfamily of ORs, a unique resultant sequence corresponding to an L-system could be generated and mapped to another OR which may or may not be in the same loci where from the subfamily members were chosen. In this regard a conclusion could be drawn that this is how diversification of human ORs were done, which is governed by the mathematical principle as described in the paper, of course there might be different principle which really followed by nature to make the diversification of ORs in the

different loci but this is our humble try to understand nature, we believe nature is beyond of all our artificial engineering.

Acknowledgement: Authors are grateful to the visiting students ***Ranita Guha***, ***Shantanav Chakrovorty*** for their research and technical help and other windows supports in this research work.

Appendix-I

The color template of DNA as shown below: Coding scheme is as follows: A=Red, T=Blue, G=Green and C=Yellow.



Appendix-II (Supplementary materials)

The classification tables for Poly-String Mean and Standard deviation are given as supplementary material-I. Also list of Hurst exponent for all encoded 2 bit information corresponding to all OR sequences is also attached as supplementary material-II.

REFERENCES

Menashe I, Aloni R, Lancet D. 2006. A probabilistic classifier for olfactory receptor pseudogenes. *BMC Bioinformatics.* 7- 393.

Yoshihito Niimura and Masatoshi Nei. Evolution of olfactory receptor genes in the human genome, 2004. *PNAS, U. S. A. vol. 100 no. 21 12235-12240*

Bettina Malnic, Paul A. Godfrey, and Linda B. Buck, 2004The human olfactory receptor gene family, *PNAS U. S. A.* 101(8): 2584–2589.

Gustavo Glusman et al. The olfactory receptor gene super family: data mining, classification, and nomenclature 2000. *Mammalian Genome* 11, 1016–1023.

Hassan, S. Sk. Choudhury, P.P., Pal, A., Brahmachary, R.L. and Goswami. A. (2010) Designing exons for human olfactory receptor gene subfamilies using a mathematical paradigm. *Journal of Biosciences*, Vol 35 (3), 389-393.