Underlying Mathematics in Diversification of Human Olfactory Receptors in Different Loci

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, *and A. Goswami*

About Olfactory Receptors

- 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.

About Olfactory Receptors

- 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.

Problem

We propose 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.

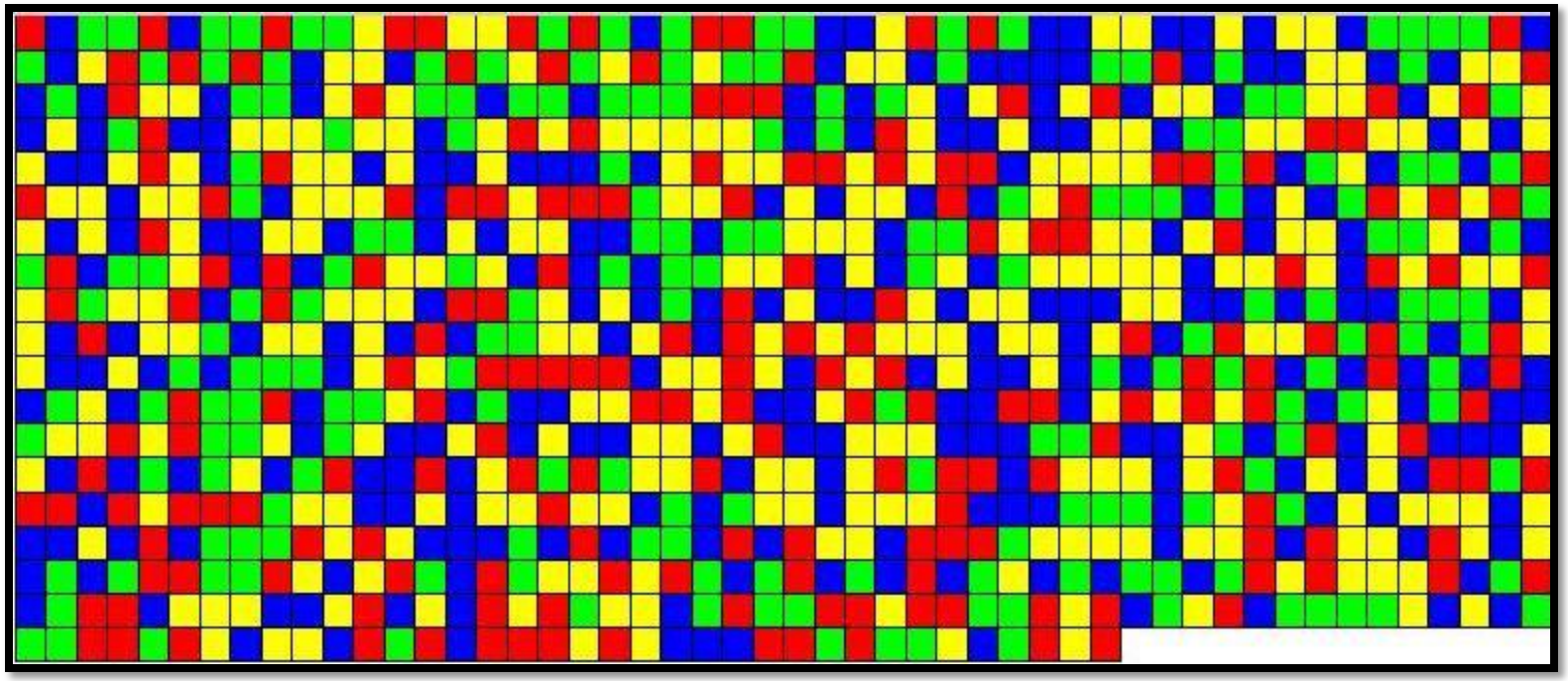
An L-System generated sequence has been taken as an input into a star-model of specific subfamily members and resultant sequence has been mapped to a specific OR based on the classification scheme using fractal parameters like Hurst exponent and fractal dimensions.

Methods and Results

- Any DNA sequence is clearly consisting of four textures of A, T, C, and G (nucleotide cluster). 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.

Color template

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



- 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 and supplementary material-II*).

- Each of A, T, C and G of all OR sequences have been encoded by the following two bit information.
- 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.
- An L-System generated sequence can be taken as an input into the star model to get a full length sequence.
- 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.

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. In the following star model there are 108 mismatches which are shown as hyphened in the following star model.

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
```

- An L-System (shown below) is used to generate a sequence and that is being inputted into the above star model.
- Consequently we got the following resultant sequence (RS-1) shown below.
- **L-System**: *Axiom*: A: *Production Rules*: A→CTG, C→CCA, T→TGC, G→GAC.

- ATGGATGGAGCCAACCAGAGTGAGTCCTCACAGTTCCTTCTCCTGGGGATGTCAGAGAGTCCTGAGCAGCAG
  CAGATCCTGTTTTGGATGTTCCTGTCCATGTACCTGGTTCACGGTGCTGGGAAATGTGCTCATCATCCTGGCCAT
  CAGCTCTGATTCCCCCCTGCACACCCCCGTGTACTTCTTCCTGGCCAACCTCTCCTTCACTGACCTCTTCTTTGTC
  ACCAACACAATCCCCAAGATGCTGGTGAACCTCCAGTCCCAGAACAAAGCCATCTCCTATGCAGGGTGTCTGA
  CACAGCTCTACTTCCTGGTCTCCTTGGTGACCCTGGACAACCTCATCCTGGCCGTGATGGCCTATGATCGCTATG
  TGGCCAGCTGCTGCCCCCTCCACTACGCCACAGCCATGAGCCCTGCGCTCTGTCTCTTCCTCCTGTCCTTGTGTT
  GGGCGCTGTCAGTCCTCTATGGCCTCCTGCCCACCGTCCTCATGACCAGCGTGACCTTCTGTGGGCCTCGAGA
  CATCCACTACGTCTTCTGTGACATGTACCTGGTGCTGCGGTTGGCATGTTCCAACAGCCACATGAATCACACAG
  CGCTGATTGCCACGGGCTGCTTCATCTTCCTCACTCCCTTGGGATTCCTGACCAGGTCCTATGTCCCCATTGTCA
  GACCCATCCTGGGAATACCCTCCGCCTCTAAGAAATACAAAGCCTTCTCCACCTGTGCCTCCCATTTGGGTGGA
  GTCTCCCTCTTATATGGGACCCTTCCTATGGTTTACCTGGAGCCCCTCCATACCTACTCCCTGAAGGACTCAGTA
  GCCACAGTGATGTATGCTGTGGTGACACCCATGATGAACCCGTTCATCTACAGCCTGAGGAACAAGGACATGC
  ATGGGGCTCAGGGAAGACTCCTACGCAGACCCTTTGAGAGGCAAACA

The resultant sequence (RS-1)

Result

- We have found the following data corresponding to (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

Result

- Therefore the decreasing ordering based on mean and SD is (CGAT) and (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 in link….

Result

In the link 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.

Conclusion

- In this work, 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 work,
- 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.

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.