Quadruple context-free L-System mathematical tools as origin of biological evolution

Arunava Goswami[$], Pabitra Pal Choudhury[#], Rajneesh Singh[$,#], Sk. Sarif Hassan[$,#]

[$]Biological Sciences Division and [#]Applied Statistics Unit, Indian Statistical Institute, 203, B. T. Road, Kolkata 700108, India

It is well known that A, T, G, C annealed together early in evolution and the long stretch of DNA was found which ultimately resulted into chromosomes of different organisms. But it is unclear till date how exons, introns, conserved protein domains was formed. Using the DNA sequences of the largest known gene-family present in human genome, i.e., olfactory receptors and simplest possible quadruple context-free L-Systems, we show that conserved protein domains and intergenic regions which lies at the heart of the biological evolution started with a sixteen base-pairs stretch of DNA.

1. Use of quadruple context-free L-Systems on olfactory receptors (OR)

ORs constitute the largest known multigene family involved in sense of smell of different organisms. There are ~388 functional and ~414 pseudo OR genes are present in human genome [1]. Researchers have found that except two chromosomes, ORs are unevenly distributed in 21 chromosomes. OR1 family contains 28 full lengths and 5 pseudo ORs as per HORDE database (https://senselab.med.yale.edu/ORDB/files/humanorseqanal.html). Fig. 1 shows the ClustalW (http://align.genome.jp/) data of extreme 5'-end stretch of 29 full length ORs.

Fig.1

```
OR1A1          ---------------------------------------------ATGAGGGAAAATAA
OR1A2          ---------------------------------------------ATGAAGAAAGAAAA
OR1N1          -----------------------------------------------------ATGGAAAA
OR1N3          -----------------------------------------------------ATGGAAAA
OR1S1          ---------------------------------------------ATGCATCAAGGAAA
OR1S2          ---------------------------------------------ATGCATCAAGAAAA
OR1L1          ---------------------------------------------ATGGGAAGAAATAA
OR1L3          ---------------------------------------------ATGGGAATGTCCAA
OR1L8          ---------------------------------------------ATGGAAAGAATCAA
OR1L4          ---------------------------------------------ATGGAGACAAAGAA
OR1L6          ---------------------------------------------ATGGAGATAAAGAA
OR1Q1          ---------------------------------------------ATGGACAACAGCAA
OR1C1          ---------------------------------------------ATGGAAAAAAGAAA
OR1F12         ---------------------------------------------ATGGAAGGGAAAAA
OR1E1          ---------------------------------------------ATGATGGGACAAAA
OR1E6          ---------------------------------------------ATGATGGGACAAAA
OR1E2          ---------------------------------------------ATGATGGGACAAAA
OR1J2          ---------------------------------------------ATGAGCCCTGAGAA
OR1J4          ---------------------------------------------ATGAAGAGGGAGAA
OR1F1          ---------------------------------------------ATGAGCGGGACAAA
OR1F2          GTATGTTTCTGAATTCACTGTCTTCTATGCAGCTGGGTCCAGACATATGAGAGGGACAAA
OR1N2          -----------------------------------ATGGGAAAACCAGGCAGAGTGAA
```

```
OR1I1           ----------------------------------------------ATGGAACCAGAAAA
OR1M1           ----------------------------------------------ATGGAACCAAGAAA
OR1D4           ----------------------------------------------ATGGATGGAGATAA
OR1D5           ----------------------------------------------ATGGATGGAGATAA
OR1D2           ----------------------------------------------ATGGATGGAGGCAA
OR1K1           ----------------------------------------------ATGGAGGCTGCCAA
OR1B1           ----------------------------------------------ATGATGAGCTTTGC

OR1A1           CC--AGTCCTCTA---CACTGGAATTCATC-CTCCTGGGAGTTACTGGTCAGCAGGAACA
OR1A2           TC--AATCCTTTA---ACCTGGATTTTATT-CTCCTGGGAGTTACTAGTCAGCAAGAACA
OR1N1           CC--AATCCAGCA---TTTCTGAATTTTTC-CTCCGAGGAATATCAGCGCCTCCAGAGCA
OR1N3           CC--AATCCAGCA---TTTCTGAATTTTTC-CTCCGAGGAATATCAGCGCCTCCAGAGCA
OR1S1           CC--AAACCACCA---TCACTGAATTCATT-CTCCTGGGATTTTTCAAGCAGGATGAGCA
OR1S2           CC--AAACCACCA---TCACTGAATTCATT-CTCCTGGGACTCTCCAACCAGGCTGAACA
OR1L1           CC--TAACAAGAC---CCTCTGAATTCATC-CTCCTTGGACTCTCCTCTCGACCTGAGGA
OR1L3           CC--TGACAAGAC---TCTCTGAATTTATT-CTCTTGGGACTCTCCTCTCGGTCTGAAGA
OR1L8           CC--ACACCAGCAGTGTCTCCGAGTTTATC-CTCCTGGGACTCTCCTCCCGGCCTGAGGA
OR1L4           TT--ATAGCAGCAGCACCTCAGGCTTCATC-CTCCTGGGCCTCTCTTCCAACCCTAAGCT
OR1L6           CT--ACAGCAGCAGCACCTCAGGCTTCATC-CTCCTGGGCCTCTCTTCCAACCCTCAGCT
OR1Q1           CT--GGACCAGTG---TGTCCCATTTTGTT-CTCTTGGGCATTTCCACCCACCCAGAAGA
OR1C1           TC--TAACAGTTG---TCAGGGAATTCGTC-CTTCTGGGACTTCCTAGCTCAGCAGAGCA
OR1F12          TC--AAACCAATA---TCTCTGAATTTCTC-CTCCTGGGCTTCTCAAGTTGGCAACAACA
OR1E1           TC--AAACCAGCA---TCTCAGACTTCCTG-CTCCTGGGCCTGCCCATCCAACCAGAGCA
OR1E6           TC--AAACCAGCA---TCTCAGACTTCCTG-CTCCTGGGCCTGCCCATCCAACCAGAGCA
OR1E2           TC--AAACCAGCA---TCTCAGACTTCCTG-CTCCTGGGCCTGCCCATCCAACCAGAGCA
OR1J2           CC--AGAGCAGCG---TGTCCGAGTTCCTC-CTTCTGGGCCTCCCCATCCGGCCAGAGCA
OR1J4           TC--AGAGCAGTG---TGTCTGAGTTCCTC-CTCCTGGACCTCCCCATCTGGCCAGAGCA
OR1F1           CC--AGTCGAGTG---TCTCCGAGTTCCTC-CTCCTGGGACTCTCCAGGCAGCCCCAGCA
OR1F2           CC--AGT-GAGTG---TCTCCGAGTTCCTC-CTCTTGGGACTCTCCAGGCAGCCCCAGCA
OR1N2           CC--AAACCACTG---TTTCAGACTTCCTC-CTTCTAGGACTCTCTGAGTGGCCAGAGGA
OR1I1           GC--AAACCGAAA---TCTCAGAATTCTTC-CTCCAGGGACTCTCAGAAAAGCCAGAGCA
OR1M1           CC--AAACCAGTG---CATCTCAATTCATC-CTCCTGGGACTCTCAGAAAAGCCAGAGCA
OR1D4           CC--AGAGTGAGA---ACTCACAGTTCCTT-CTCCTGGGGATCTCAGAGAGTCCTGAGCA
OR1D5           CC--AGAGTGAGA---ACTCACAGTTCCTT-CTCCTGGGGATCTCAGAGAGTCCTGAGCA
OR1D2           CC--AGAGTGAAG---GTTCAGAGTTCCTT-CTCCTGGGGATGTCAGAGAGTCCTGAGCA
OR1K1           TG--AGTCTTCAGAGGGAATCTCATTCGTT-TTATTGGGACTGACAACAAGTCCTGGACA
OR1B1           CCCTAATGCTTCACACTCTCCGGTTTTTTTGCTCCTTGGGTTCTCGAGAGCTAACATCTC
                                  **  *   *    *   *
```

Fig. 1 Legend. ClustalW analysis of 29 (OR1E1 and OR1E6 are same) OR1 family members of OR present in HORDE. Asterisks indicate conserved nucleotides.

Fig. 1 shows that out of aforesaid 28, 23 ORs has distinct (each with a set of 4 bp) 16 bp sequence at the extreme 5′-end which could be used to design production rules of L-Systems [2]. Fig. 2 shows three different classes of L-Systems, which have been derived from 28 ORs of family 1.

First class, designated as 'Class-I' consists of 19 following OR1 family sequences- OR1A1, OR1B1, OR1E1 (same as OR1E6), OR1E2, OR1F1, OR1F12, OR1I1, OR1J2, OR1J4, OR1K1, OR1L1, OR1L3, OR1L4, OR1L6, OR1L8, OR1N2, OR1Q1, OR1S1, OR1S2. For these sequences, when context free L-System production rules were made from the extreme 5′-end of the DNA sequences, with axiom: A, the rules were found to non-overlapping. For example, for OR1A1 the production rules are A ATGA, C GGGA, T AAAT, G AACC. One of the observations, in this class is OR1S1, OR1S2. Although their first sixteen base-pair sequence same (i.e., their production rules are same), OR1S1 and OR1S2 has different DNA sequences in their exons. Using context-free L-Systems over a number of iterations one can introduce mutations in terms

of insertions to form OR1S1 to OR1S2 and vice-versa [3, 4, 5, 6]. As context-free L-System is an irreversible system so when OR1S1/ OR1S2 will be made from OR1S2/OR1S1 after the exonic sequence a large intergenic sequence would be generated. OR1S1 and OR1S2 are present in chromosome 11. Although no subfamily members of OR1S are present in the chromosome 11, it may be that from a large number of other ORs present in the same chromosome (~390 ORs in HORDE) have given rise to OR1S1/OR1S2. From this study it is clear that after OR1S1 and OR1S2, it is difficult to generate same subfamily members and a large intergenic region would be produced after the generation of OR1S1/OR1S2.

Second class, designated as 'Class-II' consists of three following OR1 family sequences- OR1C1, OR1N1, OR1N3. The production rules for this class has a unique feature that one of the production rule gives poly A sequences. For example, for OR1C1 the production rules are A ATGG, C AAAA, T AAGA, G AATC.

Third class, designated as 'Class-III' consists of five following OR1 family sequences- OR1A2, OR1D2, OR1D4, OR1D5, OR1M1. Unique feature of class-III is that two of the production rules are similar. For example, for OR1A2 the production rules are A ATGA, C AGAA, T AGAA, G AATC. As the production rule for T and C are same, so successive iterations will increase the amount of repetitive DNA sequence generated by the class-III production rules. OR1D subfamily has a single pseudogene OR1D3P which is most similar to OR1D2. As described above OR1D3P was produced from OR1D2. We took OR1D3P sequence and it also fall in the class-III category of ORs.

Therefore from the above discussion it is clear this is a new classification system based on only first 16 bp sequence of ORs and not on the sequence homology of either DNA or protein of the entire OR sequence.

2. Evolutionary implications

Let us assume that 16 bp sequence of all ORs (Class-I, Class-II and Class-III) of OR family 1 were present as individual sequences somewhere in the evolutionary scale. This is not an assumption per se as they are present in all the subfamily members of the largest multigene family known in the genome i.e., OR which is present in the human genome till date. Each of these 16 bp sequences will give rise to evolution of unique protein domains (supplementary data provided). For example we took OR1L subfamily members which have 5 members (OR1L1, OR1L3, OR1L4, OR1L6 and OR1L8). Applying the same method as described in Ref. 6 (and supplementary information provided herewith) we generated following conserved protein domains. OR1L1 gave 80%-200% alignment score with a hypothetical protein from Giant Panda. Similarly OR1L3 gave a composite conserved protein domain from California purple sea urchin with homology ranging from 40%-200%. Using the similar methodology OR1L4 and OR1L6 derived L-Systems could not produce any ORFs and therefore consequently no conserved

protein domains were formed. We conjecture that these two aforesaid and following L-Systems (Axiom: A) A ATGG, C AGAC, T AAAG, G AATT and A ATGG, C AGAT, T AAAG, G AACT have not been so far for producing conserved protein domains but may have been used for making intergenic DNA sequence. OR1L8 when was used in the same model was found to pick up suprabasin protein from three different organisms, viz. Dog, Horse and Giant Panda.

Therefore one can visualize very very early in evolution if nature could derive any of the aforesaid three classes (class-I, class-II, class-III) L-Systems which means generation of a simple stretch of 16 bp DNA sequence then extension of the DNA chain using DNA polymerase can happen. We showed in this paper that extension of this kind of L-System can lead to meaningful conserved protein domains as well as intergenic DNA sequences. To the best of our knowledge, this is the first report which showed how evolution started with a 16 bp DNA sequences.

ACKNOWLEDGMENTS: This work was supported by the Department of Biotechnology (DBT), New Delhi, grants (BT/PR9050/NNT/28/21/2007 and BT/PR8931/NNT/28/07/2007 to AG) and NAIP-ICAR-World Bank grant (Comp-4/C3004/2008-09; Project leader: AG) and ISIplan projects for 2001—2011 to A. G. Authors would like to thank visiting students, Ranita Guha and Shantanav Chakrovorty for their valuable inputs in computations.

References

1. Malnic B, Godfrey P-A and Buck L-B (2004) The human olfactory receptor gene family; *Proc. Natl. Acad. Sci. USA* **101** 2584—2589 Erratum in: *Proc. Natl. Acad. Sci. USA* 2004 **101** 7205

2. Prusinkiewicz, P. and Lindenmayer, A. (1990) in the algorithmic beauty of plants (New York: Springer-Verlag).

3. Hassan, S. Sk., Choudhury, P. P., Pal, A., Brahmachary, R. L. and Goswami, A. (2010) L-Systems: A Mathematical Paradigm for Designing Full Length Genes and Genomes. Global Journal of Computer Science and Technology, 10: 119-122, category: I.2.1, J.3, and G.1.0. (Published in June, 2010)

4. 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, volume 35, number 3: 389-393.

5. Hassan, S. Sk., Choudhury, P. P., Pal, A., Brahmachary, R. L. and Goswami, A. (2010) Combination of L-systems: For Designing Human Olfactory Receptor Pseudo-gene, OR1D3P. International Journal of Computational Cognition, (Publisher: Yang's Scientific Research Institute, USA) (Accepted for publication in 2011).

6. Goswami, Arunava, Singh, Rajneesh, Choudhury, Pabitra, and Hassan, Sk. SarifHassan. Designing L-Systems for making three and six open reading frames from the leading strand of a single DNA molecule. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2010.4844.1> (2010).