# A novel mathematical tool for generating highly conserved protein domain via different organismal genomic landscapes

*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

**Darwinian evolution hypothesizes that a short stretch of DNA was first constructed and then it expanded to give rise to a long strand. This long strand then produced a mix of exons, introns and repetitive DNA sequence. The order of production of above three kinds of DNA sequence is unknown. Reshuffling of stretches of DNA like above within organisms has given rise to different chromosomes. Till date it is not known how this process is governed. In this paper we show that starting with a sixteen base-pair human olfactory DNA sequence one can form a highly conserved protein domain. Once this domain is formed repetitive DNA sequences of a particular kind starts generating which signifies that this particular conserved protein domain will be unique in nature. The entire mathematical exercise presented in this paper is based on simplest possible context free L-System which we think has been adopted by biological system in general.**

Context free L-System [1, 2, 3, 4], considered to be one of simplest, originally proposed by Hungarian Biologist A. Lindenmayer [5] to study symmetry of plants can be used to generate a large length of DNA sequence with a definite size limit. We took a human olfactory receptor DNA sequence (OR1F1) and derived an L-System with following production rule [(Axiom: A) A ATGA, C GCGG, T GACA, G AACC] which basically covers extreme 5′-end of OR1F1. 1024 bp length DNA sequence [Fig. 1] generated from this production rule showed 3 Open Reading Frames (ORF) in 3′ 5′ direction [Fig.2].

Fig.1

ATGAGACAAACCATGAAACCATGAGCGGATGAATGAATGAGCGGGCGGATGAGACAAACCATGAATGAATGAGCGGGCGGATGAGACAAACCATGAAACCGCGGAACCAAC
CATGAGACAAACCATGAATGAGACAAACCATGAATGAGACAAACCATGAAACCGCGGAACCAACCAACCGCGGAACCAACCATGAGACAAACCATGAAACCATGAGCGGAT
GAATGAATGAGCGGGCGGATGAGACAAACCATGAATGAGACAAACCATGAATGAGACAAACCATGAAACCGCGGAACCAACCAACCGCGGAACCAACCATGAGACAAACCA
TGAAACCATGAGCGGATGAATGAATGAGCGGGCGGATGAGACAAACCATGAATGAATGAGCGGGCGGAACCGCGGAACCAACCATGAATGAGCGGGCGGATGAATGAGCGG
GCGGATGAGACAAACCATGAAACCATGAGCGGATGAATGAATGAGCGGGCGGATGAGACAAACCATGAATGAGACAAACCATGAAACCATGAGCGGATGAATGAATGAGCG
GGCGGATGAGACAAACCATGAATGAGACAAACCATGAAACCATGAGCGGATGAATGAATGAGCGGGCGGATGAGACAAACCATGAATGAATGAGCGGGCGGAACCGCGGAA
CCAACCATGAATGAGCGGGCGGATGAATGAGCGGGCGGATGAATGAGCGGGCGGAACCGCGGAACCAACCATGAATGAGCGGGCGGATGAATGAGCGGGCGGATGAGACAA
ACCATGAAACCATGAGCGGATGAATGAATGAGCGGGCGGATGAGACAAACCATGAATGAATGAGCGGGCGGATGAGACAAACCATGAAACCGCGGAACCAACCATGAGACA
AACCATGAATGAGACAAACCATGAATGAGACAAACCATGAAACCGCGGAACCAACCAACCGCGGAACCAACCATGAGACAAACCATGAAACCATGAGCGGATGAATGAATG
AGCGGGCGGATGAGACAAACCATGA

**Fig. 1** Legend. Context-free L-System [(Axiom: A) A ATGA, C GCGG, T GACA, G AACC] derived 1024 bp sequence from olfactory receptor OR1F1 sequence (ATGAGACAAACCATGA) [http://genome.weizmann.ac.il/cgi-bin/horde/showGene.pl?key=symbol&value=OR1F1].

Fig.2

5'3' Frame 1
Met R Q T Met K P Stop A D E Stop Met S G R Met R Q T Met N E Stop A G G Stop D K P Stop N R G T N H E T N H E
Stop D K P Stop Met R Q T Met K P R N Q P T A E P T Met R Q T Met K P Stop A D E Stop Met S G R Met R Q T Met
N E T N H E Stop D K P Stop N R G T N Q P R N Q P Stop D K P Stop N H E R Met N E Stop A G G Stop D K P Stop
Met N E R A E P R N Q P Stop Met S G R Met N E R A D E T N H E T Met S G Stop Met N E R A D E T N H E Stop D K
P Stop N H E R Met N E Stop A G G Stop D K P Stop Met R Q T Met K P Stop A D E Stop Met S G R Met R Q T Met N
E Stop A G G T A E P T Met N E R A D E Stop A G G Stop Met S G R N R G T N H E Stop A G G Stop Met S G R Met R
Q T Met K P Stop A D E Stop Met S G R Met R Q T Met N E Stop A G G Stop D K P Stop N R G T N H E T N H E Stop
D K P Stop Met R Q T Met K P R N Q P T A E P T Met R Q T Met K P Stop A D E Stop Met S G R Met R Q T Met
5'3' Frame 2
Stop D K P Stop N H E R Met N E Stop A G G Stop D K P Stop Met N E R A D E T N H E T A E P T Met R Q T Met N E
T N H E Stop D K P Stop N R G T N Q P R N Q P Stop D K P Stop N H E R Met N E Stop A G G Stop D K P Stop Met R
Q T Met N E T N H E T A E P T N R G T N H E T N H E T Met S G Stop Met N E R A D E T N H E Stop Met S G R N R
G T N H E Stop A G G Stop Met S G R Met R Q T Met K P Stop A D E Stop Met S G R Met R Q T Met N E T N H E T
Met S G Stop Met N E R A D E T N H E Stop D K P Stop N H E R Met N E Stop A G G Stop D K P Stop Met N E R A E
P R N Q P Stop Met S G R Met N E R A D E Stop A G G T A E P T Met N E R A D E Stop A G G Stop D K P Stop N H E
R Met N E Stop A G G Stop D K P Stop Met N E R A D E T N H E T A E P T Met R Q T Met N E T N H E Stop D K P
Stop N R G T N Q P R N Q P Stop D K P Stop N H E R Met N E Stop A G G Stop D K P Stop
5'3' Frame 3
E T N H E T Met S G Stop Met N E R A D E T N H E Stop Met S G R Met R Q T Met K P R N Q P Stop D K P Stop Met
R Q T Met N E T N H E T A E P T N R G T N H E T N H E T Met S G Stop Met N E R A D E T N H E Stop D K P Stop
Met R Q T Met K P R N Q P T A E P T Met R Q T Met K P Stop A D E Stop Met S G R Met R Q T Met N E Stop A G G T
A E P T Met N E R A D E Stop A G G Stop D K P Stop N H E R Met N E Stop A G G Stop D K P Stop Met R Q T Met K
P Stop A D E Stop Met S G R Met R Q T Met N E T N H E T Met S G Stop Met N E R A D E T N H E Stop Met S G R N
R G T N H E Stop A G G Stop Met S G R Met N E R A E P R N Q P Stop Met S G R Met N E R A D E T N H E T Met S G
Stop Met N E R A D E T N H E Stop Met S G R Met R Q T Met K P R N Q P Stop D K P Stop Met R Q T Met N E T N H
E T A E P T N R G T N H E T N H E T Met S G Stop Met N E R A D E T N H
3'5' Frame 1
S W F V S S A R S F I H P L Met V S W F V S W L V P R L V G S A V S W F V S F Met V C L I H G L S H G W F R G
F Met V C L I R P L I H S W F V S S A R S F I H P L Met V S W F V S S A R S F I R P L I H G W F R G S A R S F
I R P L I H P P A H S W L V P R F R P L I H S W F V S S A R S F I H P L Met V S W F V S F Met V C L I R P L I
H S S A H G F Met V C L I H G L S H P P A H S F I R S W F H G L S H P P A H S S A R S F Met V G S A V P P A H
S F Met V C L I R P L I H S S A H G F Met V C L Met V G S A V G W F R G F Met V C L I H G L S H S W F V S S A
R S F I H P L Met V S W F V S W L V P R L V G S A V S W F V S F Met V C L I H G L S H G W F R G F Met V C L I
R P L I H S W F V S S A R S F I H P L Met V S W F V S
3'5' Frame 2
H G L S H P P A H S F I R S W F H G L S H G W F R G W L V P R F H G L S H S W F V S F Met V C L Met V G S A V
S W F V S S A R S F I H G L S H P P A H S F I R S W F H G L S H P P A H S S A R S F Met V G S A V P P A H S S
A R S F I R P L I H G W F R G S A R S F I H G L S H P P A H S F I R S W F H G L S H S W F V S S A R S F I H P
L Met V S W F V S F Met V C L I R P L I H S S A H G F Met V C L I R P L I H P P A H S W L V P R F R P L I H S
W F V S S A R S F I H P L Met V S W F V S W L V P R L V G S A V S W F V S F Met V C L I H G L S H P P A H S F
I R S W F H G L S H G W F R G W L V P R F H G L S H S W F V S F Met V C L Met V G S A V S W F V S S A R S F I
H G L S H P P A H S F I R S W F H G L S H
3'5' Frame 3
Met V C L I R P L I H S S A H G F Met V C L Met V G S A V G W F R G F Met V C L I H G L S H S W F V S W L V P
R F H G L S H P P A H S F Met V C L I R P L I H S S A H G F Met V C L I R P L I H P P A H S W L V P R F R P L
I H P P A H S S A R S F Met V G S A V P P A H S F Met V C L I R P L I H S S A H G F Met V C L I H G L S H P P
A H S F I R S W F H G L S H S W F V S S A R S F I H P L Met V S W F V S S A R S F I R P L I H G W F R G S A R
S F I H G L S H P P A H S F I R S W F H G L S H G W F R G W L V P R F H G L S H S W F V S F Met V C L I R P L
I H S S A H G F Met V C L Met V G S A V G W F R G F Met V C L I H G L S H S W F V S W L V P R F H G L S H P P
A H S F Met V C L I R P L I H S S A H G F Met V C L

**Fig. 2.** Legend. Expasy [http://expasy.org/tools/dna.html] generated six reading frames from Fig.1 show that there are three ORFs in the 3' 5' direction.

As DNA synthesis occurs in 5' 3' direction therefore we made complementary sequence of Fig. 1 as shown in Fig.3. We then generated 1024 bp sequence from another L-System [(Axiom: A) A TACT, C CTGT, T TTGG, G TACT] which is the extreme 5'-end of Fig.3.

TACTCTGTTTGGTACTTTGGTACTCGCCTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTTGGCGCCTTGGTTG
GTACTCTGTTTGGTACTTACTCTGTTTGGTACTTACTCTGTTTGGTACTTTGGCGCCTTGGTTGGTTGGCGCCTTGGTTGGTACTCTGTTTGGTACTTTGGTACTCGCCTA
CTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTACTCTGTTTGGTACTTACTCTGTTTGGTACTTTGGCGCCTTGGTTGGTTGGCGCCTTGGTTGGTACTCTGTTTGGT
ACTTTGGTACTCGCCTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTACTTACTCGCCCGCCTTGGCGCCTTGGTTGGTACTTACTCGCCCGCCTACTTACTCGCC
CGCCTACTCTGTTTGGTACTTTGGTACTCGCCTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTACTCTGTTTGGTACTTTGGTACTCGCCTACTTACTTACTCGC
CCGCCTACTCTGTTTGGTACTTACTCTGTTTGGTACTTTGGTACTCGCCTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTACTTACTCGCCCGCCTTGGCGCCTT
GGTTGGTACTTACTCGCCCGCCTACTTACTCGCCCGCCTACTTACTCGCCCGCCTTGGCGCCTTGGTTGGTACTTACTCGCCCGCCTACTTACTCGCCCGCCTACTCTGTT
TGGTACTTTGGTACTCGCCTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTACTTACTCGCCCGCCTACTCTGTTTGGTACTTTGGCGCCTTGGTTGGTACTCTGT
TTGGTACTTACTCTGTTTGGTACTTACTCTGTTTGGTACTTTGGCGCCTTGGTTGGTTGGCGCCTTGGTTGGTACTCTGTTTGGTACTTTGGTACTCGCCTACTTACTTAC
TCGCCCGCCTACTCTGTTTGGTACT

**Fig.3** Legend. The complementary strand of Fig.1 in the 5′ 3′ direction.

This sequence when conceptually translated and it gave rise to three ORFs in 5′ 3′ direction (Fig.4).

Fig.4


Fig.4 Legend. Expasy generated six reading frames from Fig.3 show that there are three ORFs in the 5′ 3′ direction.

Fig.5

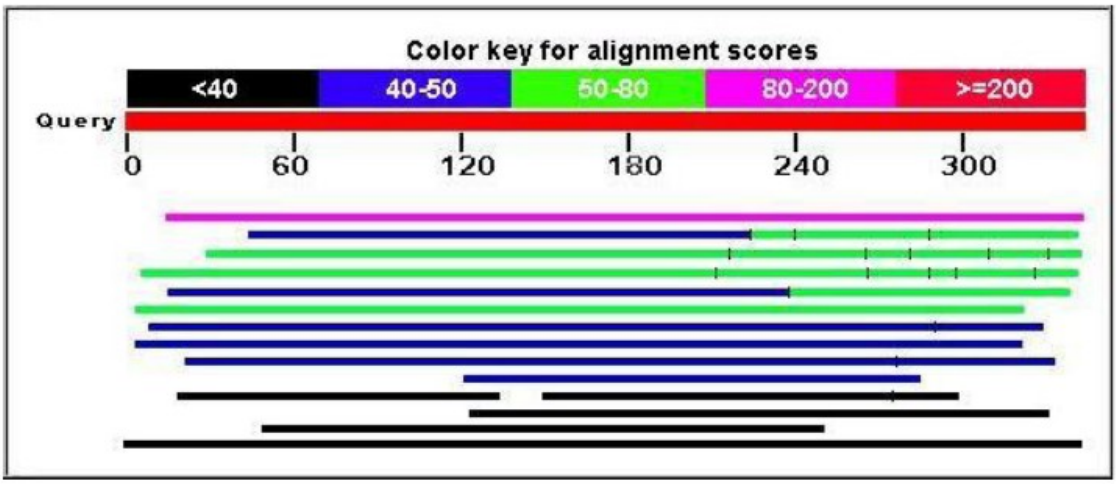


Fig. 5. Legend: First ORF of Fig. 4 gives 80%-200% homology with a protein domain of nearly 350 amino acids long. This protein has been classified by NCBI as a hypothetical protein (1081 aa long) from Gray short-tailed opossum, *Monodelphis domestica*. Similarly other two ORFs also exhibited high homology with other organisms (data not shown).

The extreme 5' end sequence in Fig. 3 was again used as a template (like Fig.1) to generate 1024 bp sequence with following production rule [(Axiom: A) A TTGG, C TTGG, T TACT, G TACT]. On careful observation it is clear the production rule of A and C are same as well as T and G. As expected the L-System produced was found to be a highly repetitive DNA sequence (Fig.6).

Fig.6

TTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTACTCTGTTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTG
GTTGGTACTCTGTTTGGTTGGTTGGTACTTACTTTGGTACTCTGTTTGGCTGTTTGGTACTTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTT
GGCTGTTTGGTACTTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTTGGTACTT
ACTTTGGTACTCTGTTTGGCTGTTTGGTACTTTGGTTGGTTGGTACTTACTCTGTTTGGTACTTTGGTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTTGGTACT
TACTTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTACTCTGTTTGGCTGTTTGGTACTTTGGTTGGTTGGTACTTACTTTGGTACTCTG
TTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTACTCTGTTTGGTTGGTTGGTACTTACTTTGGTACTCT
GTTTGGCTGTTTGGTACTTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTACTCTGTTTGGTTGGTTGGT
ACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTACTCTGTTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGTTGGTACT
CTGTTTGGTTGGTTGGTACTTACTTTGGTACTCTGTTTGGCTGTTTGGTACTTTGGTTGGTTGGTACTTACTTTGGTTGGTACTTACTTTGGTACTCTGTTTGGCTGTTTG
GTACTTTGGTTGGTTGGTACTTACT

Fig. 6 Legend. Repetitive DNA sequence generated following [(Axiom: A) A TTGG, C TTGG, T TACT, G TACT] which showed no significant homology when searched in NCBI database.

Therefore L-System generation started with human olfactory receptor DNA OR1F1 has reached its limit. We then blastp-ed the hypothetical protein (1081 aa long) from Gray short-tailed opossum, *Monodelphis domestica* into NCBI database (Fig. 5 with pink color line) and found that it has a high level of domain homology with *Forkhead box* protein E4 of *Xenopus laevis* [Fig.7].

Fig.7

```
Query  735   PAAVGAVGGGRRRKRPVQRGKPPFSYIALIAMAIAHSAERRLTLGGIYRFITERFAFYRD  794
             P      GGRRRKRPVQRGKPP+SYIALIAMAIA+S ER+LTLGGIY+FI ERF FYR+
Sbjct  28    PEEHNQASGGRRRKRPVQRGKPPYSYIALIAMAIANSPERKLTLGGIYKFIMERFPFYRE  87

Query  795   NPRKWQNSIRHNLTLNDCFVKIPREPGHPGKGNYWALDPAAQDMFDSGSFLRRRKRFKRS  854
             N +KWQNSIRHNLTLNDCFVKIPREPGHPGKGNYW LDPAA+DMFD+GSFLRRRKRFKR+
Sbjct  88    NSKKWQNSIRHNLTLNDCFVKIPREPGHPGKGNYWTLDPAAEDMFDNGSFLRRRKRFKRT  147
```

This domain was re-blastp-ed to NCBI and we found that this domain is highly conserved amongst a number of organisms who are distantly related in evolution [Fig.8].

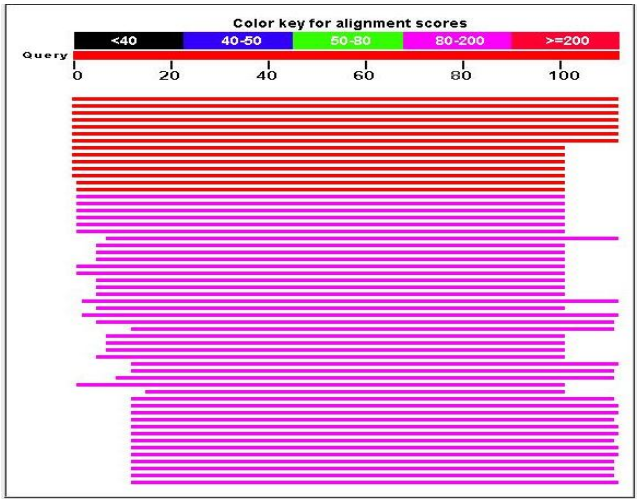


Fig. 8 Legend. Conserved protein domain obtained from Fig. 7 when blastp-ed into NCBI protein database

This result shows that starting with a sixteen base-pair unique stretch of DNA and using L-System production rule one can make a conserve protein domain which was hitherto unknown to biologists. We predict that this is how all conserved protein domains in living organisms have been produced. This mathematical exercise also clearly shows after a conserved protein domain is formed the biological system does not allow any more conserved domain to be formed from the same sequence by producing a repetitive DNA sequence.

Therefore we can conclude that L-System is an important mathematical tool which can be explored to find out the genomic domain shuffling, protein domain formation and repetitive DNA evolution of different organisms. It is tempting to speculate that natural systems might have used this kind of context free L-System derived methodology to generate genomes of different organisms. This method also could be used by synthetic biologist to find correlation between different organismal DNA, protein domain and finally change them at will.

**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. 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)

2. 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 (to be published as cover page article in September 2010 issue).

3. 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).

4. Goswami, Arunava, Singh, Rajneesh, Choudhury, Pabitra, and Hassan, Sk. Sarif Hassan. 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).

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