Combination of L-systems: For Designing Human Olfactory Receptor Pseudo-gene, OR1D3P

Sk. Sarif Hassan[a, b], Pabitra Pal Choudhury[a], R. L. Brahmachary[b] and Arunava Goswami[b]

[a]Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. pabitrapalchoudhury@gmail.com [P. P. C.]; sarimif@gmail.com [S. S. H.];

[b]Biological Sciences Division, Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. agoswami@isical.ac.in [A.G.].

Keywords
L-system
Pseudo-gene
Olfactory receptor
Olfaction

Footnotes
To whom correspondence should be addressed E-mail: sarimif@gmail.com / agoswami@isical.ac.in

***Author contributions*:** A. G. conceptualized the experiments and performed entire research with S. S. H.; A. G. and S. S. H. acknowledges A. P., P. P. C. and R. L. B. and Ms. Rebecca French (Grad student of Prof. Michael Hochella) of Virginia Tech, USA for their enormous help while doing these series of experiments.

Conflict of interest statement: The authors declare no conflict of interest.

Abstract: Using a set of L-system mathematical rules, we have made a close relative of olfactory receptor sequence, OR1D3P and show that possibly pseudogene play an important role in either genome stability. These results might give clues to the reasons for presence of long stretches of non-coding regions of DNA in genome.

Introduction: As per conservative estimate, approximately 51-105 Olfactory receptors (ORs) loci are present in human genome occur in clusters. These clusters are like unevenly spread mosaics and spread over 21 pair of human chromosomes. Recent studies have shown that 339 full length OR genes and 297 OR pseudogenes are present in these clusters (1). OR1D3P is one of such pseudogene located near a loci where OR1D2, OR1D4 and OR1D5 OR genes are present. Pseudogenes are known to have stop codons (TAA/TAG and TGA) in all three open reading frames. Still today, there are no valid hypothesis which can account for the presence of almost equal number of OR pseudogenes in the human. This is one of the very striking facts of human genome which warrant immediate investigation. In this paper, we have made a close relative of OR1D3P using L-system starting with the template of published

sequence of OD1D3P from HORDE database (2). These results shade light on the large amount of non-coding region of the DNA present in the human genome.

2. Data and Results:

We downloaded sequence of human OR1D3P from HORDE database. (http://genome.weizmann.ac.il/cgi-bin/horde/). The sequence is as follows-

>OR1D3P (HORDE#42: 613)

5'- ATGGATGGAGGCAACCAGAGTGAAGGTTCAGAGTTCCTTCTCCTGGGGAT

CTCAGAGAGTCCTGAGCAGCAGCAGATGCTGTTTTGGATGTTCCTGGTCA

GGTACCTGGTCACGGTGCTGGGAAATGTGCTCATCATCCTGGCCATCAGC

TCTGATTCCCGCCTGCACACCCCCATGTACTTCTTCCTGGCCAACCTCTC

CTTCACTGACCTCTTCTTTGTCACCAACACAATCCCCAAGATGCTGGTGA

ACCTCCAGTCCCAGAACAAAGCCATCTCCTACACAGGGTGTCTGACACAG

CTCTACTTCCTGGTCTCCTTGGTGGCCCTGGACAACCTCAACCTGGCCGT

GATGGCGTATGATCGCTATGTGGCCATCTGCCGTCCCCTCCACTATGTCA

CAGCCATGATCCCTGGGCTCTGTATCTTGCTCCTCTCCTTGTGTTGGGTG

TTCTCTGCCCTCTATGGCCTCATCCATATCCTCCTCATGACCAGGTGACC

TTCTGTGGGTCTCAAAAGATCCACTACCTCTTCTGTGAGATGTACTTCCT

GCTAAGGCTGGCATGTTCCAACATCCACGTCAACCACACAGTACTGGTTG

CCACGGGCTGCTTCATCTTCCTCATCCCCTTAGGTTTCATGATCACATCC

TACGCCCGCATTGTCAGAGCCATCCTCCAAATACCCTCAGCCACTGGGAA

GTACAAAGCCTTCTCCACCTGTGCTTCCCATTTGGCTGTGGTCTCCCTCT

TCTATGGGACTCTGGGTATGGTGTACCTGCAGCCCCTCCAAACCTACTCC

ATGAAGGACTCAGTAGCCACAGTGATGTATGCGGTGGTGACGCCATGATT

AACCCTTTCATCTACAGCCTGAGGAACAAGGACATGCATGGGGCTCTGGG

AAGACTTCGCCAAGGAAAAGCCTTCCAGAAGTTGACATGAGGGGTAATTT

TG-3'

Following table show that the degree of matches and mismatches of OR1D3P gene sequence with that of other full length ORS in the same OR loci as well as to that of Chimpanzee (Table 1).

The OR genes	Most Similar Gene	Matching length	%PID

OR1D2	OR1D3P OR1D4 OR1D5 CONTIG3463.6-1888	313 310 310 312	84.98 82.26 82.58 98.4
OR1D3P	OR1D2 OR1D4 OR1D5 CONTIG1419.13-6395	313 303 303 318	84.98 83.83 83.17 93.17
OR1D4	OR1D5 CONTIG1419.10-10399 CONTIG3601.2-11308	312 312 312	98.4 97.44 94.55
OR1D5	OR1D4 CONTIG1419.10-10399 CONTIG3601.2-11308	312 312 312	98.4 97.12 94.23

[Table 1: Most similar genes in OR1D and contig receptors]

Using following construction methodology of L-system, we could make a sequence which is 100% similar to that of OR1D3P.

The governing set of L-systems is as follows:

L System for iteration # 1:

$A \rightarrow ****$
$C \rightarrow ****$
$T \rightarrow ****$
$G \rightarrow ****$

The * would followed by Table 2

L System for iteration # 2 :

$A \rightarrow TGGG$
$C \rightarrow GATG$
$T \rightarrow ACCC$
$G \rightarrow TCAC$

L System for iteration # 3:

$A \rightarrow ACCC$
$C \rightarrow TCAC$
$T \rightarrow GTCT$
$G \rightarrow CTCC$

L System for iteration # 4 :

$A \rightarrow GTCT$
$C \rightarrow CTCT$
$T \rightarrow CAGA$
$G \rightarrow TCTG$

L System for iteration # 5 :

$A \rightarrow GTCT$
$C \rightarrow AGAT$
$T \rightarrow CTGA$
$G \rightarrow GCAC$

L System for iteration # 6:

$A \rightarrow TCTC$
$C \rightarrow GAGC$
$T \rightarrow CTGT$
$G \rightarrow ATTA$

L System for iteration # 7:

$A \rightarrow TCCG$
$C \rightarrow AGCG$
$T \rightarrow TATT$
$G \rightarrow ACTG$

L System for iteration # 8:

$A \rightarrow CCGA$
$C \rightarrow GACG$
$T \rightarrow GGGC$
$G \rightarrow TTCC$

L System for iteration # 9:

$A \rightarrow CCAA$
$C \rightarrow CGGG$
$T \rightarrow CTTC$
$G \rightarrow ATTA$

L System for iteration # 10:

$A \rightarrow CAAC$
$C \rightarrow GGGC$
$T \rightarrow CGCA$
$G \rightarrow AGAA$

L System for iteration # 11:

$A \rightarrow CACG$

$C \rightarrow GCCC$
$T \rightarrow AAAT$
$G \rightarrow GTTA$

L System for iteration # 12:

$A \rightarrow CACG$
$C \rightarrow AAAT$
$T \rightarrow TTAT$
$G \rightarrow TTTA$

L System for iteration # 13:

$A \rightarrow CTG$
$C \rightarrow AATTT$
$T \rightarrow ATTAC$
$G \rightarrow TTAGA$

The Crucial changes in the production rule in the first L-system for each member are as follows:

OR receptor	The Production Rule for T in the iteration #1
OR1D2 & OR1D3P	$A \rightarrow ATGG, C \rightarrow ATGG, T \rightarrow AG\mathbf{GC}, G \rightarrow AACC$ $A \rightarrow ATGG, C \rightarrow ATGG, T \rightarrow AG\mathbf{GC}, G \rightarrow AACC$
OR1D4 & OR1D5	$A \rightarrow ATGG, C \rightarrow ATGG, T \rightarrow AG\mathbf{AT}, G \rightarrow AACC$
CONTIG1419.13-6395	$A \rightarrow TCAT, C \rightarrow AGAT, T \rightarrow ATTA, G \rightarrow TTCC$
CONTIG3463.6-1888	$A \rightarrow ATGA, C \rightarrow CAGG, T \rightarrow ATTG, G \rightarrow AAAA$
CONTIG1419.10-10399	$A \rightarrow GAAG, C \rightarrow ATGG, T \rightarrow CAGG, G \rightarrow ATTA$
CONTIG3601.2-11308	$A \rightarrow TGCA, C \rightarrow GGAT, T \rightarrow TAAA, G \rightarrow AATA$

[Table 2: For L System for iteration #1]

We find that the first sixteen bases of OR1D2 and OR1D3P are same. This is the primary reason for the difference in iteration # 1 (L-system; Table 2) amongst full length ORs in the OR1D2, OR1D4 & OR1D5 loci and the pseudo-gene, OR1D3P. One more important point can be described as follows- the differences in all the four contig sequences can be brought to focus by only the L system in the iteration #1.

3. Conclusion:

In summary, we show in this paper that a pseudogene can be formed by using a set of L-systems. In this case, 100% similarity was observed as because we followed the template of the OR1D3P. If we use other set of L-systems, we can make a close relative of OR1D3P, which can be thought of as equivalent to evolutionary changes incorporated during the course of evolution (data not shown). Similarly, we have also developed a set of L-systems with which a full length OR can be made (data not shown). The earlier manuscript which deals with making a full length OR using star model can be viewed as making a pseudogene with star model as areas of OR, where if A,T, G, C falls in wrong positions could generate stop codons (3). Both these results immediately allow us to speculate that because of the stability of the long chain of DNA, DNA chains are made by following a set of L systems and OR genes come automatically because of the restrictions imposed by the L systems. These results however do not violate any of the current dogma of

molecular biology (data not shown). Finally, we would like to make a long chain of DNA where such rules can be applied such as that of human mitochondrial genome.

Acknowledgements

This work was supported by Department of Biotechnology (DBT), Govt. of India grants (BT/PR9050/NNT/28/21/2007 & BT/PR8931/NNT/28/07/2007 to A. Goswami) & NAIP-ICAR-World Bank grant (Comp-4/C3004/2008-09; Project leader: A. Goswami) and ISI plan projects for 2001-2011. Authors are grateful to their visiting students Rajneesh Singh and Snigdha Das for their technical help in making advanced C programs and other computer applications on Windows support used for this study.

References:

1. Prusinkiewicz P, Lindenmayer A (1990), The algorithmic beauty of plants, *Springer-Verlag ISBN 978-0387972978.*

2. Malnic B, Hirono J, Sato T, Buck L-B (1999) Combinatorial receptor codes for odors. *Cell* 96: 713-723.

3. Sk. S Hassan, P Pal Choudhury, A Pal, R. L. Brahmachary and A. Goswami (2009) Designing exons for human olfactory receptor gene subfamilies using a mathematical paradigm, Journal of Bioscience , Vol 35 (3) pp 389-393

4. Sk. S Hassan, P Pal Choudhury, A Pal, R. L. Brahmachary and A Goswami (2009) L-Systems: A Mathematical Paradigm for Designing Full Length Genes and Genomes, *GJCST (2010)* ***10****: 119-122.*