

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

Sk. Sarif Hassan, Pabitra Pal Choudhury, Amita Pal, R. L. Brahmachary and Arunava Goswami

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. Copyright © 2011 Yang's Scientific Research Institute, LLC. All rights reserved.

Index Terms—L-system, Pseudo-gene, Olfactory receptor, Olfaction.

I. 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 [3]. 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.

II. DATA AND RESULTS:

We downloaded sequence of human OR1D3P from HORDE database. (<http://genome.weizmann.ac.il/cgi-bin/horde/>). The sequence is shown in Fig. 1 (see the next page).

Following table show that the degree of matches and mismatches of OR1D3P gene sequence with that of other

Manuscript received September 24, 2009; revised March 01, 2010.

Sk. Sarif Hassan, Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India & cBiological Sciences Division, Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. Pabitra Pal Choudhury, Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. Amita Pal, Bayesian Interdisciplinary Research Unit (BIRU), Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. R. L. Brahmachary and Arunava Goswami, Biological Sciences Division, Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. Corresponding emails address: sarimif@gmail.com (S.S. Hassan), agoswami@isical.ac.in (A. Goswami).

Publisher Item Identifier S 1542-5908(11)10310-8/\$20.00

Copyright ©2011 Yang's Scientific Research Institute, LLC. All rights reserved. The online version posted on March 08, 2011 at <http://www.YangSky.com/ijcc/ijcc91.htm>

full length ORS in the same OR loci as well as to that of Chimpanzee (Table I).

TABLE I: MOST SIMILAR GENES IN ORID AND CONTIG RECEPTORS

The OR genes	Most Similar Gene	Matching length	%PID
OR1D2	OR1D3P	313	84.98
	OR1D4	310	82.26
	OR1D5	310	82.58
	CONTIG3463.6-1888	312	98.4
OR1D3P	OR1D2	313	84.98
	OR1D4	303	83.83
	OR1D5	303	83.17
	CONTIG1419.13-6395	318	93.17
OR1D4	OR1D5	312	98.4
	CONTIG1419.10-10399	312	97.44
	CONTIG3601.2-11308	312	94.55
OR1D5	OR1D4	312	98.4
	CONTIG1419.10-10399	312	97.12
	CONTIG3601.2-11308	312	94.23

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 shown in Table II (see the next two pages).

The Crucial changes in the production rule in the first L-system for each member are shown in the following table.

TABLE III: FOR L-SYSTEM FOR ITERATION #1

OR receptor	The Production Rule for T in the iteration #1
OR1D2 & OR1D3p	A → ATGG, C → ATGG, T → AGGC, G → AACC
OR1D4 & OR1D5	A → ATGG, C → ATGG, T → AGAT, G → AACC
CONTIG1419.13-6395	A → TCAT, C → AGAT, T → ATTA, G → TTCC
CONTIG3463.6-1888	A → ATGA, C → CAGG, T → ATTG, G → AAAA
CONTIG1419.10-10399	A → GAAG, C → ATGG, T → CAGG, G → ATTA
CONTIG3601.2-11308	A → TGCA, C → GGAT, T → TAAA, G → AATA

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 III) amongst full length ORs in the OR1D2, OR1D4 & OR1D5 loci and the pseudo-gene, OR1D3P. One more important point can be described as

>OR1D3P (HORDE#42: 613)

5' - ATGGATGGAGGCAACCAGAGTGAAGGTTGAGAGTTCCTTCTCCTGGGGAT
 CTCAGAGAGTCTGAGCAGCAGCAGATGCTGTTTTGGATGTTCTGGTCA
 GGTACCTGGTCACGGTGTGGGAAATGTGCTCATCATCCTGGCCATCAGC
 TCTGATTCGCGCTGCACACCCCATGACTTCTTCTGGCCAACCTCTC
 CTTCACTGACCTCTTCTTTGTCACCAACACAATCCCCAAGATGCTGGTGA
 ACCTCCAGTCCCAGAACAAAGCCATCTCCTACACAGGGTGTCTGACACAG
 CTCTACTTCTGGTCTCCTTGGTGGCCCTGGACAACCTCAACCTGGCCGT
 GATGGCGTATGATCGTATGTGGCCATCTGCCGTCCTCCACTATGTCA
 CAGCCATGATCCCTGGGCTCTGTATCTTGCTCCTCTCCTTGTGTTGGGTG
 TTCTCTGCCCTCTATGGCCTCATCCATATCCTCCTCATGACCAGGTGACC
 TTCTGTGGGTCTCAAAGATCCACTACCTCTTCTGTGAGATGTACTTCTC
 GCTAAGGCTGGCATGTTCCAACATCCACGTCAACCACACAGTACTGGTTG
 CCACGGGCTGCTTCATCTTCTCATCCCCTTAGGTTTCATGATCACATCC
 TACGCCCGCATTGTGAGCCATCCTCAAATACCCTCAGCCACTGGGAA
 GTACAAAGCCTTCTCCACCTGTGCTTCCCATTTGGCTGTGGTCTCCCTCT
 TCTATGGGACTCTGGGTATGGTGTACCTGCAGCCCCTCAAACCTACTCC
 ATGAAGGACTCAGTAGCCACAGTGTATGCGGTGGTGACGCCATGATT
 AACCTTTTCATCTACAGCCTGAGGAACAAGGACATGCATGGGGCTCTGGG
 AAGACTTCGCCAAGGAAAAGCCTTCCAGAAGTTGACATGAGGGGTAATTT
 TG-3'

Fig. 1: Sequence of human OR1D3P.

TABLE II: GOVERNING SET OF L-SYSTEMS

L-system for iteration #1	L-system for iteration #2
$A \rightarrow ***$	$A \rightarrow TGGG$
$C \rightarrow ***$	$C \rightarrow GATG$
$T \rightarrow ***$	$T \rightarrow ACCC$
$G \rightarrow ***$	$G \rightarrow TCAC$
The * would followed by Table II.	
L-system for iteration #3	L-system for iteration #5
$A \rightarrow ACCC$	$A \rightarrow GTCT$
$C \rightarrow TCAC$	$C \rightarrow AGAT$
$T \rightarrow GTCT$	$T \rightarrow CTGA$
$G \rightarrow CTCC$	$G \rightarrow GCAC$
L-system for iteration #4	L-system for iteration #6
$A \rightarrow GTCT$	$A \rightarrow TCTC$
$C \rightarrow CTCT$	$C \rightarrow GAGC$
$T \rightarrow CAGA$	$T \rightarrow CTGT$
$G \rightarrow TCTG$	$G \rightarrow ATTA$
L-system for iteration #7	L-system for iteration #9
$A \rightarrow TCCG$	$A \rightarrow CCAA$

$C \rightarrow AGCG$	$C \rightarrow CGGG$
$T \rightarrow TATT$	$T \rightarrow CTTC$
$G \rightarrow ACTG$	$G \rightarrow ATTA$
L-system for iteration #8	L-system for iteration #10
$A \rightarrow CCGA$	$A \rightarrow CAAC$
$C \rightarrow GACG$	$C \rightarrow GGGC$
$T \rightarrow GGGC$	$T \rightarrow CGCA$
$G \rightarrow TTCC$	$C \rightarrow AGAA$
L-system for iteration #11	L-system for iteration #13
$A \rightarrow CACG$	$A \rightarrow CTG$
$C \rightarrow GCCC$	$C \rightarrow AATTT$
$T \rightarrow AAAT$	$T \rightarrow ATTAC$
$G \rightarrow GTTA$	$G \rightarrow TTAGA$
L-system for iteration #12	
$A \rightarrow CACG$	
$C \rightarrow AAAT$	
$T \rightarrow TTAT$	
$G \rightarrow TTTA$	

follows- the differences in all the four contig sequences can be brought to focus by only the L system in the iteration #1.

III. 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 [1]. 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] Sk. S. Hassan, P. Pal Choudhury, A. Pal, R. L. Brahmachary, and A. Goswami. Designing exons for human olfactory receptor gene subfamilies using a mathematical paradigm. *Journal of Bioscience*, 35(5), 2009.
- [2] B. Malnic, J. Hirono, T. Sato, and L-B. Buck. Combinatorial receptor codes for odors. *Cell*, 96:713–723, 1999.
- [3] P. Prusinkiewicz and A. Lindenmayer. *The algorithmic beauty of plants*. Springer-Verlag, 1990.