Designing exons for human olfactory receptor gene subfamilies using a mathematical paradigm

Sk. Sarif Hassan[a,c], Pabitra Pal Choudhury[a], Amita Pal[b], R. L. Brahmachary[c] and Arunava Goswami[c,1]

[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] Bayesian Interdisciplinary Research Unit (BIRU), Indian Statistical Institute, 203 B. T. Road, Calcutta, 700108 India. pamita@isical.ac.in [A. P.] and

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

Keywords

- Human olfactory receptor
- L-system
- ClustalW
- Star Model
- Olfaction

Footnotes

- [1]To whom correspondence should be addressed. E-mail: agoswami@isical.ac.in / srabanisopanarunava@gmail.com
- Author contributions: A. G. and S. S. H. designed research; A. G. and S. S. H. performed research; P. P. C. and A. P. analyzed data; and A. G., S. S. H., P. P. C., A. P., and R. L. B. wrote the paper.
- Conflict of interest statement: The authors declare no conflict of interest.

Abstract

Ligands for only two human olfactory receptors are known. One of them, OR1D2, binds to Bourgeonal, a volatile chemical constituent of the fragrance of mythical flower, Lily of the valley or *Our Lady's tears, Convallaria majalis (also the national flower of Finland)* [Malnic B, Godfrey P-A, Buck L-B (2004) The human olfactory receptor gene family. *Proc. Natl. Acad. Sci U. S. A.* 101: 2584-2589 and Erratum in: *Proc Natl Acad Sci U. S. A.* (2004) 101: 7205]. OR1D2, OR1D4 and OR1D5 are three full length olfactory receptors present in an olfactory locus in human genome. These receptors are more than 80% identical in DNA sequences and have 108 base pair mismatches among them. Apparently, these mismatch positions show no striking pattern using computer pattern recognition tools. In an attempt to find a mathematical rule in those mismatches, we find that L-system generated sequence can be inserted into the OR1D2 subfamily specific star model and novel full length olfactory receptors can be generated. This remarkable mathematical principle could be utilized for making new subfamily OR members from any OR subfamily. Aroma and electronic nose industry might utilize this rule in future.

Introduction

Olfactory receptors (ORs) loci in human genome occur in clusters ranging from ~51-105 and they are unevenly spread over 21 chromosomes (1-2). A conservative estimate suggests that 339 full length OR genes and 297 OR pseudogenes are present in these clusters (1). Theoretically, there are two possible ways of OR-odorant molecular binding, viz., (i) each OR binds to a large number of different odorants and (ii) each OR binds to a small number of odorants. In either case, odorant detection at the OR level follows a combinatorial rule, though the stringency of the rule would differ in the two alternatives. Experimentally, it has been demonstrated that each OR recognizes a large number of odorants and perhaps a large class of various concentrations of the odorants tested (3). OR gene (conceptually translated to protein sequences) family (>40% amino acid identity) can be divided into subfamily (>60% identity) and sub-family members might have more than 90% identity (4). Subfamily members are highly similar in DNA and protein sequences, but they are capable of recognizing different odorant molecules.

We hypothesized that there might be a nature inspired mathematical rule which determines sequences of the subfamily members which could be extended to subfamily and family. If such a rule exists, it would be of great interest for basic research; furthermore, one could construct ORs useful applied research (viz. for studies in connection with electronic nose. Three full length model subfamily OR members from HORDE database (http://genome.weizmann.ac.il/horde/), OR1D2 (Gene length: 936 bp), OR1D4 (Gene length: 936 bp) and OR1D5 (Gene length: 936 bp) were chosen from HORDE database. OR1D2 [Chromosomal position: 17p13.3; synonym: hOR17-4] recognizes odorant Bourgeonal and is perceived as Lily of the valley fragrance (1). OR1D2, OR1D4 and OR1D5 incidentally show very little or no polymorphism in the published sequence databases from different research groups (Data not shown). It is possible that these groups might have used same samples or same source, while cloning and sequencing. OR1D2, OR1D4 and OR1D5 were aligned using ClustalW and was found to contain 108 base pair mismatches out of 936 base pairs available (data not shown).

If we consider OR1D2, OR1D4 and OR1D5 each as a string of A/T/G/C, then out of 936 positions, 828 excluding 108 mismatches were found to be chosen by nature as fixed or evolutionarily conserved positions. As OR1D2, OR1D4 and OR1D5 are highly related sequences, therefore, a canonical sequence for this subfamily, termed as `star model' of OR sequence was made by using a computer C program, where 108 gaps were introduced in respective positions (Fig 1a and Fig 1b).

A context free L-system (5), was used to generate a 243 bp long DNA sequence.

L System:

Set of Variables: A, T, C, and G.

Axiom: C (C is the starting symbol)

Production Rule: A → CTG, C→CCA, T→TGC and G→GAC

Following aforesaid production rule, 1st and 2nd iteration, would give CCA (03 bp) and CCACCACTG (09 bp) respectively. Four iterations yield 81 base pair sequences. This is insufficient to answer for 108 mismatches. Five such iterations generate the following 243 bp sequence-

CCACCACTGCCACCACTGCCATGCGACCCACCACTGCCACCACTGCCATGCGACCCACCACT
GTGCGACCCAGACCTGCCACCACCACTGCCACCACTGCCATGCGACCCACCACTGCCACCAC
TGCCATGCGACCCACCACTGTGCGACCCAGACCTGCCACCACCACTGCCACCACTGCCATGC
GACTGCGACCCAGACCTGCCACCACCACTGGACCTGCCACCATGCGACCCACCACTG…(i)

Using a C computer program, nucleotides present in sequence (i) was introduced from 5'- end of the sequence into the star model gaps shown in Fig. 1 sequentially. Briefly, the

Step 1: First, in all the gaps (with 1 bp, 2 bp, 3 bp and 4 bp) in star model, only one nucleotide would be inserted.

Step 2: 1 bp gaps in star model would become 0 gap. Then in the remaining gaps (1 bp, 2 bp and 3 bp) would be filled up and the process would be repeated until all gaps are filled.

The resultant OR sequence is shown in (ii) below.

ATGGATGGAGCCAACCAGAGTGAGTCCTCACAGTTCCTTCTCCTGGGGATGTCAGAGAGTCC
TGAGCAGCAGCAGATCCTGTTTTGGATGTTCCTGTCCATGTACCTGGTCACGGTGCTGGGAA
ATGTGCTCATCATCCTGGCCATCAGCTCTGATTCCCCCCTGCACACCCCCGTGTACTTCTTCC
TGGCCAACCTCTCCTTCACTGACCTCTTCTTTGTCACCAACACAATCCCCAAGATGCTGGTGA
ACCTCCAGTCCCAGAACAAAGCCATCTCCTATGCAGGGTGTCTGACACAGCTCTACTTCCTG
GTCTCCTTGGTGACCCTGGACAACCTCATCCTGGCCGTGATGGCCTATGATCGCTATGTGGCC
AGCTGCTGCCCCCTCCACTACGCCACAGCCATGAGCCCTGCGCTCTGTCTCTTCCTCCTGTCC
TTGTGTTGGGCGCTGTCAGTCCTCTATGGCCTCCTGCCCACCGTCCTCATGACCAGCGTGACC
TTCTGTGGGCCTCGAGACATCCACTACGTCTTCTGTGACATGTACCTGGTGCTGCGGTTGGCA
TGTTCCAACAGCCACATGAATCACACAGCGCTGATTGCCACGGGCTGCTTCATCTTCCTCACT
CCCTTGGGATTCCTGACCAGGTCCTATGTCCCCATTGTCAGACCCATCCTGGGAATACCCTCC
GCCTCTAAGAAATACAAAGCCTTCTCCACCTGTGCCTCCCATTTGGGTGGAGTCTCCCTCTTA
TATGGGACCCTTCCTATGGTTTACCTGGAGCCCCTCCATACCTACTCCCTGAAGGACTCAGTA
GCCACAGTGATGTATGCTGTGGTGACACCCATGATGAACCCGTTCATCTACAGCCTGAGGAA
CAAGGACATGCATGGGGCTCAGGGAAGACTCCTACGCAGACCCTTTGAGAGGCAAACA (ii)

Sequence (ii) was blasted using DNA-DNA and translated protein-protein (Blastx) search engines in HORDE and NCBI database from where initial OR1D2, OR1D4 and OR1D5 sequences were obtained. Results of blast searches show that with the search parameters available in the HORDE website (which could not be changed by remote user), the (ii) sequence showed 92%, 92% and 91% identity with OR1D2, OR1D4 and OR1D5 respectively. Significantly, these insertions do not produce any stop codon in the exon sequence. It is interesting to note following rules that might govern this biological process-

(i) If one utilizes a production rule which starts with C→ CCC, then a viable OR could be produced. It is tempting to check whether the long poly C containing region of telomere serve as template for insertion as in case of DNA replication.

(ii) It seems that each OR subfamily utilizes a specific star model. We have tested OR10J, OR10K and OR3A loci (data not shown). Rules that govern the formation of the star model for each subfamily members are in the process of analysis.

(iii) Sense of smell and taste are primordial in nature. Our current hypothesis is based on the idea that star model or conserved region of the ORs was produced following an as yet unidentified mathematical rule quite early in evolution. Then mathematical rules like L system and their variants were used to make the variable regions which contribute to odorant ligand binding domains of the ORs. This process of insertion might have happened at DNA polymerization level.

We have already mentioned earlier in the text that there are 2-5 highly related yet diverse OR subfamily sequences clustered in human genome. The reason and significance of this special genomic architectural plan has to be searched for in an evolutionary framework at the theoretical level. The results obtained following the aforesaid production rule as spelt out tempt us to test the hypothesis- whether nature followed this procedure or not. A comparative study of usage of the L-systems in olfactory subgenomes of lower vertebrates like mouse to that of human might offer clues in this direction.

In summary, in this paper, we report a relatively simple model of context free L-system for making variable region of the OR and this could be adopted for making artificial ORs. Many more advanced context free L-system could be designed once it is experimentally established that this is the kind of rule OR utilizes for generating subfamily members, at least, if not subfamily and family members, i.e., more divergent ORs in the genome. Here we observe that the computer generated star model sequence, sequentially filled with A, T, G, C in the way described above from a sequence generated by L-system could generate a sequence which is highly similar to those of OR1D2, OR1D4 and OR1D5. Therefore, most likely, this work is purely mathematical in nature at this stage and a large number of experimental evidences are necessary.

Appendix

While writing the computer programme, L-System satisfied following rules- A. To fill the single gap of the Star Model: System will check two previous states as well as the two past states of the gap- (a) If the second previous is ‘T’ and the first previous is ‘A’ and the first past is ‘A’ and the second past is either ‘A’ or ‘G’, then the chosen L-System must produce ‘C’ at the gap. e.g., …TA_AA(/G)… (b) If the second previous is ‘T’ and the first previous is ‘A’ and the first past is ‘G’ and the second past is ‘A’, then the chosen L-System must produce ‘C’ at the gap. e.g., …TA_GA… (c) If the second previous is ‘T’ and the first previous is ‘A’ and the first past is either ‘T’ or ‘C’, then the chosen L-System must produce ‘C’ or ‘T’ at the gap. e.g., …TA_T(/C)… (d) If the second previous is ‘T’ and the first previous is ‘G’ and the first past is ‘A’ and the second past is either ‘A’ or ‘G’, then the chosen L-System must produce ‘C’ or ‘G’ at the gap. e.g., …TG_AA(/G)… (e) If the second previous is ‘T’ and the first previous is ‘G’ and the first past is ‘G’ and the second past is ‘A’, then the chosen L-System must produce ‘C’ or ‘G’ at the gap. e.g., …TG_GA… (f) If the second previous is ‘T’ and the first previous is ‘G’ and the first past is either ‘T’ or ‘C’, then the chosen L-System must produce ‘C’ or ‘G’ or ‘T’ at the gap. e.g., …TG_T(/C)… (g) If the first previous is ‘T’ and the first past is ‘A’ and the second past is either ‘C’ or ‘T’, then the chosen L-System must produce ‘T’ or ‘C’ at the gap. e.g., …T_AC(/T)… (h) If the first previous is ‘T’ and the first past is ‘A’ and the second past is either ‘A’ or ‘G’, then the chosen L-System must produce ‘C’ at the gap. e.g., …T_AA(/G)… (i) If the first previous is ‘T’ and the first past is ‘G’ and the second past is ‘A’, then the chosen L-System must produce ‘G’ or ‘C’ at the gap. e.g., …T_GA… (j) If the first previous is ‘T’ and the first past is ‘G’ and the second past is ‘C’ or ‘T’ or ‘G’, then the chosen L-System must produce ‘T’ or ‘C’ or ‘G’ at the gap. e.g., …T_GC(/T/G)… (h) If the first previous is ‘C’ and the first past is ‘A’ and the second past is either ‘A’ or ‘G’, then the chosen L-System must produce ‘C’ or ‘G’ or ‘A’ at the gap. e.g., …C_AA(/G)… (i) If the first previous is ‘C’ and the first past is ‘G’ and

the second past is ‘A’, then the chosen L-System must produce ‘C’ or ‘G’ or ‘A’ at the gap. e.g., …C_GA… (j) Else the gap can be filled by any state like ‘A’ or ‘C’ or ‘T’ or ‘G’. B. To fill the double or more than double gap of the Star Model (one gap fill at a time): System should check only two previous states of the gap- (a) If the second previous is ‘T’ and the first previous is ‘A’, then the chosen L-System must produce ‘C’ or ‘T’ at the gap. e.g., …TA_ _… (b) If the second previous is ‘T’ and the first previous is ‘G’, then the chosen L-System must produce ‘C’ or ‘T’ or ‘G’ at the gap. e.g., …TG_ _… (c) Else the gap can be filled by any state like ‘A’ or ‘C’ or ‘T’ or ‘G’. This rule would be applicable until the number of gap becomes one. When the no. of gap becomes one, then the rule (A) is applicable.

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, Snigdha Das and Somnath Mukherjee for their technical help in making advanced C programs and other computer applications on Windows support used for this study.

References

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

2. Young J-M, Endicott R-M, Parghi S-S, Walker M, Kidd J-M, Trask B-J (2008) Extensive copy-number variation of the human olfactory receptor gene family. *Am. J. Hum. Genet.* 83: 228–242.

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

4. Glusman G, Yanai I, Rubin I, Lancet D (2001) The complete human olfactory subgenome. *Genome Research* 11: 685-702.

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