L-Systems: A Mathematical Paradigm for Designing Full Length Genes and Genomes

Sk. Sarif Hassan[a,c,1], 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
- Assembly Method
- Human mitochondrial DNA

Footnote

- [1]To whom correspondence should be addressed. E-mail: sarimif@gmail.com / srabanisopanarunava@gmail.com

Abstract: We have shown how L-Systems can generate long genomic sequences. This has been achieved in two steps. In the first, a single L-System turned out to be sufficient to construct a full length human olfactory receptor gene. This however is only an exon. The more complicated problem of generating a genome containing both exon and intron, such as that of human mitochondrial genome has been now addressed. We have succeeded in establishing a set of L-Systems and thus solved the problem. In this context we have discussed the recent attempts at Craig Venter's group to experimentally construct long DNA molecules adopting a number of working hypotheses. A mathematical rule for generating such long sequences would shed light on various fundamental problems on various advanced areas of biology, viz. evolution of long DNA chains in chromosomes, reasons for existence of long stretches of non-coding regions as well as usher in automated methods for long DNA chains preparation for chromosome engineering. However, this mathematical principle must have room for editing / correcting DNA sequences locally in those areas of genomes where mutation and / or DNA polymerase has introduced errors over millions of years. We present the whole mitochondrial genome (exon and introns) generated by a set of L-Systems.

1. Introduction: As a working model we fasten our attention on OR1D2, human olfactory receptor gene. This consists of a sequence of 936 bp of A, T, C and G. it is worth noting that 4^{936} different combinations of the four bases A, T, C, G are possible. Of these enormous numbers of possibilities, only three have been selected as a human olfactory receptor of length 936 bp namely OR1D2, OR1D4, and OR1D5 [http://genome.weizmann.ac.il/horde/]. We will show that a single L-System can generate the sequence of OR1D2.

Gibson DG et al. (2008) [1] from The Craig J. Venter Institute in USA surprised us in a paper [2] by preparing whole genome of *Mycoplasma genitalium* within yeast cells where he and his colleagues could stitch 25 overlapping DNA fragments to form a complete synthetic genome (524 kb). This was a phenomenal discovery. Gibson et al (2009) [2] in a paper published in Nature Methods showed long DNA chain could made easily using an elegant experimental method in which concerted action of a 5' exonuclease, a DNA polymerase and a DNA ligase lead to a thermodynamically favored isothermal single reaction. In the beginning, researchers recessed DNA fragments and this process yielded single-stranded DNA overhangs that specifically annealed, and then covalently joined them. In this process, they could assemble multiple overlapping DNA molecules and surely enough mechanism of action behind making a full chromosome is now ready. But initially to organize *assembly method* proposed by Gibson DG et al. (2008) [1] four DNA cassettes of 6-kb (which could be had from Yeast genome itself) are needed. But what governs the whole yeast genome? In this paper, we are trying to explore a possible underlying mathematical principle for making whole genome.

We claim that the proposed methodology could be used in an automated system to seamlessly construct synthetic and natural genes, for modulation of genetic pathways and finally entire genomes, and could be a useful chromosome engineering tool.

2. Designing of Human OR and Mitochondrial Genome

L-System is a deterministic string rewriting system proposed by a biologist Lindenmayer in 1968. P. Prusinkiewicz & A. Lindenmayer (1990) used this system to study symmetry in the plant world [3]. The simplest L-System has been defined as the following:

An L-system is a formal grammar consisting of 4 parts:
A set of *variables*: symbols that can be replaced by production rules.
An *axiom*, which is a string, composed of some number of variables and/or constants. The axiom is the initial state of the system.
A set of *production rules* defining the way variables can be replaced with combinations of constants and other variables. A production consists of two strings - the predecessor and the successor.

(2A) ***L-System Derived Human OR1D2****:* Assuming that four variables to be A, T, C and G, we define a particular L-System as follows:

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

Using this L-System a sequence of 243 bp length can be generated in the 5^{th} iteration. With a proposed algorithm [4] this can lead to the formation of the following sequence (ii).

^{1}**A**TGGATGGAGCCAACCAGAGTGAGTCCTCACAGTTCCTTCTCCTGGGGATGTCAGAGAGTCCTGAGCAG
CAGCAGATCCTGTTTTGGATGTTCCTGTCCATGTACCTGGTCACGGTGCTGGGAAATGTGCTCATCATCC
TGGCCATCAGCTCTGATTCCCCCCTGCACACCCCCGTGTACTTCTTCCTGGCCAACCTCTCCTTCACTGA
CCTCTTCTTTGTCACCAACACAATCCCCAAGATGCTGGTGAACCTCCAGTCCCAGAACAAAGCCATCTCC
TATGCAGGGTGTCTGACACAGCTCTACTTCCTGGTCTCCTTGGTGACCCTGGACAACCTCATCCTGGCCG
TGATGGCCTATGATCGCTATGTGGCCAGCTGCTGCCCCCTCCACTACGCCACAGCCATGAGCCCTGCGCT
CTGTCTCTTCCTCCTGTCCTTGTGTTGGGCGCTGTCAGTCCTCTATGGCCTCCTGCCCACCGTCCTCATG
ACCAGCGTGACCTTCTGTGGGCCTCGAGACATCCACTACGTCTTCTGTGACATGTACCTGGTGCTGCGGT
TGGCATGTTCCAACAGCCACATGAATCACACAGCGCTGATTGCCACGGGCTGCTTCATCTTCCTCACTCC
CTTGGGATTCCTGACCAGGTCCTATGTCCCCATTGTCAGACCCATCCTGGGAATACCCTCCGCCTCTAAG
AAATACAAAGCCTTCTCCACCTGTGCCTCCCATTTGGGTGGAGTCTCCCTCTTATATGGGACCCTTCCTA
TGGTTTACCTGGAGCCCCTCCATACCTACTCCCTGAAGGACTCAGTAGCCACAGTGATGTATGCTGTGGT
GACACCCATGATGAACCCGTTCATCTACAGCCTGAGGAACAAGGACATGCATGGGGCTCAGGGAAGACTC
CTACGCAGACCCTTTGAGAGGCAAACA^{936} (ii)

With the help of BLAST search this is revealed to be the almost actual sequences of OR1D2. This consists of only an exon and the real problem is-how to generate a gene like that of human mitochondrial DNA which has sequence length of more than 15,000 and contains stretches of introns in between.

(**2B**) ***L-System Derived Human Mitochondrial Genome:***

As mentioned above this is a formidable problem and apparently no single L-System can solve it. However, as shown below a novel approach consisting of a set of L-Systems and taking into consideration the fact that this mathematical principle must allow for editing or correcting DNA sequences locally, can solve the problem.

The whole human mitochondrial DNA is equivalent to any bacterial DNA for the purpose of designing/constructing with the help of a set of L-systems. We have a strong belief that nature might use one nucleotide to start with in order to construct the whole chromosome and finally the genome. On the basis of this conjecture, we have been motivated to pick up the L- system methodology. How we are going to select the set of L systems is a relevant question; let us briefly describe the corresponding algorithm as follows:

The design of L systems are as follows where axiom (starting symbol) for the L system is A.
The nucleotide A produces first four consecutive base pairs of the mitochondrial DNA and C, T and G produce next consecutive four base pairs respectively.

Now if the number of mismatches of the DNA sequence is less than three then an L system could be chosen as: the nucleotide A produces the remaining mismatches and C, T and G all produce A.
But if the numbers of mismatches occur in between two and fifteen then an L-system could be chosen as follows: the nucleotide C produces first one third of the remaining mismatches, T produces next one third, G produces the remaining and finally A produces the CTG to achieve the remaining mismatches in the 2nd iteration of the L-system.
Based on the proposed methodology as stated above, we go on generating the L-system iterations until it crosses the given mitochondrial length. Now we compare the generated sequence with the given mitochondrial sequence and mismatching portions are again tried by another set of L-systems following the above pick up policy. And the same process is continued until the whole mitochondrial sequence is matched.

In this way we have the following twenty four L-systems' covering the mitochondrial sequence (NCBI database (NC_012920)) as shown in the as given below:

L-System for iteration number 1:
A --> GATC
C --> ACAG
T --> GTCT
G --> ATCA

L-System for iteration number 2:
A --> ACAG
C --> GTCT
T --> ATCA
G --> CCTA

L-System for iteration number 3:
A --> TCAT
C --> AACC
T --> TCAC
G --> GGGA

L-System for iteration number 4:
A --> TCGG
C --> GAGT
T --> CAGG
G --> TATT

L-System for iteration number 5:
A --> CGGG
C --> ATCG
T --> GTTT
G --> GTGG

L-System for iteration number 6:
A --> ACGG
C --> TTCC
T --> ACGC
G --> GACT

L-System for iteration number 7:
A --> ACGC
C --> TGTT
T --> TTAC
G --> AAAA

L-System for iteration number 8:
A --> TTCA
C --> AAAA
T --> GGGA
G --> ATAG

L-System for iteration number 9:
A --> TCAA
C --> AGGG
T --> AAAG
G --> CAAG

L-System for iteration number 10:
A --> TCAA
C --> AGCA
T --> GGCG
G --> ACCT

L-System for iteration number 11:
A --> TAAG
C --> AGGA
T --> ACGG
G --> TACC

L-System for iteration number 12:
A --> ACGG

C --> TTCT
T --> CCCA
G --> GTCC

L-System for iteration number 13:
A --> CCAG
C --> GATC
T --> GCCG
G --> GACA

L-System for iteration number 14:
A --> CCAG
C --> GATC
T --> CCGG
G --> ACAT

L-System for iteration number 15:
A --> CCAG
C --> GTGA
T --> GGAA
G --> TCCA

L-System for iteration number 16:
A --> CCGG
C --> ATCA
T --> CGAC
G --> TTAT

L-System for iteration number 17:
A --> CCGG
C --> CGCG
T --> TTCG
G --> TGGC

L-System for iteration number 18:
A --> CGCC
C --> GTTC
T --> TTCC
G --> ATTT

L-System for iteration number 19:
A --> GTAT
C --> TATA
T --> CGGT
G --> GAAC

L-System for iteration number 20:
A --> TTTA
C --> TACG
T --> GGAC
G --> GCAT

L-System for iteration number 21:
A --> TTTA
C --> TACG
T --> GGAC
G --> GCAT

L-System for iteration number 22:
A --> TTTA
C --> TACG
T --> GGAC
G --> GCAT

L-System for iteration number 23:
A --> TTTAT
C --> ACGGG
T --> ACGCA
G --> TAGGA

L-System for iteration number 24:
A --> CTG
C --> ACCTA
T --> ACGGG
G --> CACG

We have also observed in the publicly available database (NCBI: NC_012920) the human mitochondrial DNA sequence consist one ‘n’. Just before this ‘n’ there is a nucleotide C which allows replacing the ‘n’ in (3061 position) by any one of four nucleotides. The interesting fact is that when ‘n’ is replaced by A or T to cover up the whole mitochondrial sequence, it takes only 22 L-systems whereas 24 L-systems if ‘n’ is replaced by C or G. This mathematical principle, therefore, would be very useful in finding the mismatches in making full proof genome databases of all organisms.

3. Concluding Remarks:

In summary, in this paper, we report relatively simple models of context free L-system for making variable region of the OR and a set of context free L-Systems to design human mitochondrial genome. We claim that the proposed methodology could be used in an automated system to seamlessly construct

synthetic and natural genes, for modulation of genetic pathways and finally entire genomes, and could be a useful chromosome engineering tool. Results from the above mentioned studies clearly show that coding regions of the genes in the genome is made from the specificities of the particular L-system and the large amount of non-coding region is necessary for giving the stability of the strings of DNA on thermodynamic scale. Another important point emanated from these studies is that application of L-systems in a particular combination allows repairing of sudden change in genome locally and thereby helps to keep the string of chromosomal DNA intact. This result can be further extended to study of the nuclear chromosomal DNA. Also results from this study show clearly that this kind of approach can be very useful for correcting the ambiguity of the DNA sequences present in published and genomic sequences of various genbank databases.

Appendix-I

Supplementary materials:

A video is attached herewith (File: Sarif-et-al.pps). This video shows how L-system works for human mitochondrial DNA sequence. We have showed first three L-system iterations.

Acknowledgements

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. The author Dr. Arunava Goswami is thankful to 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.

References

1. Gibson DG et al. (2008) One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic *Mycoplasma genitalium* genome, *Proc. Natl. Acad. Sci U. S. A.* 105 (51): 20404-20409.

2. Daniel G Gibson et al. (2009), Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nature Methods* 6, 343-345

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

4. Hassan Sk. Sarif et al. (2010), Designing exon for human olfactory receptor gene subfamilies using a mathematical paradigm, (communicated to Journal of Bioscience, Springer).