L-Systems: A Mathematical Paradigm for Designing Genes and Genome with Morphological flavor

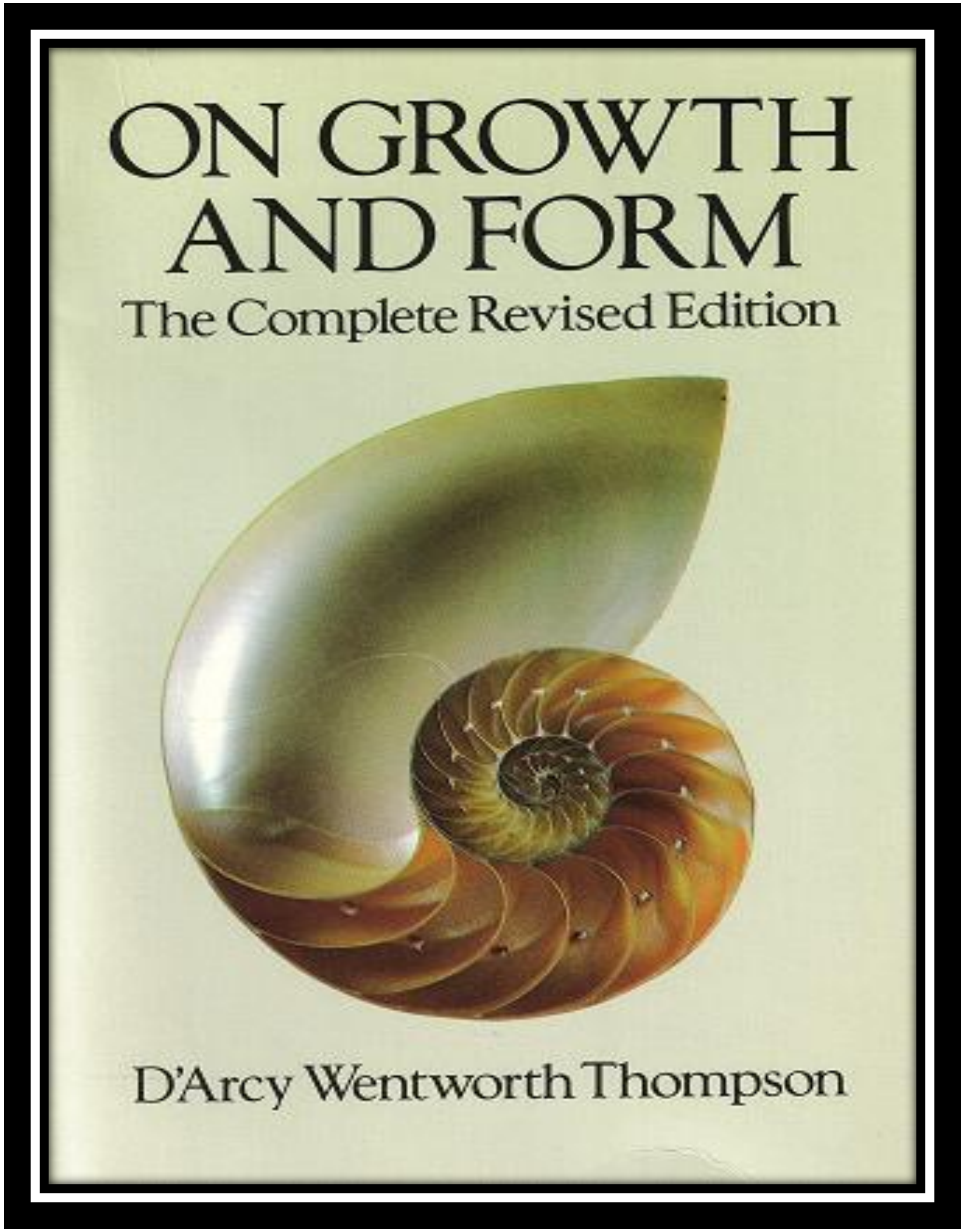
Workshop Honouring Prof. Jean Serra
26th October 2010

Pabitra Pal Choudhury
Applied Statistics Unit
Indian Statistical Institute
Kolkata

The beauty of symmetric and asymmetric pattern in the living world, as in flower or a nautilus shell has been noticed by people since the earliest times.

In 1916, D’arcy Thomson, a biologist initiated his classic work, On Growth and Form, the revised edition of which was published in 1942 and in more than a 1000 pages he treated symmetry and also ordered asymmetry in animals and plants like leaves, flower, shells, horns etc.

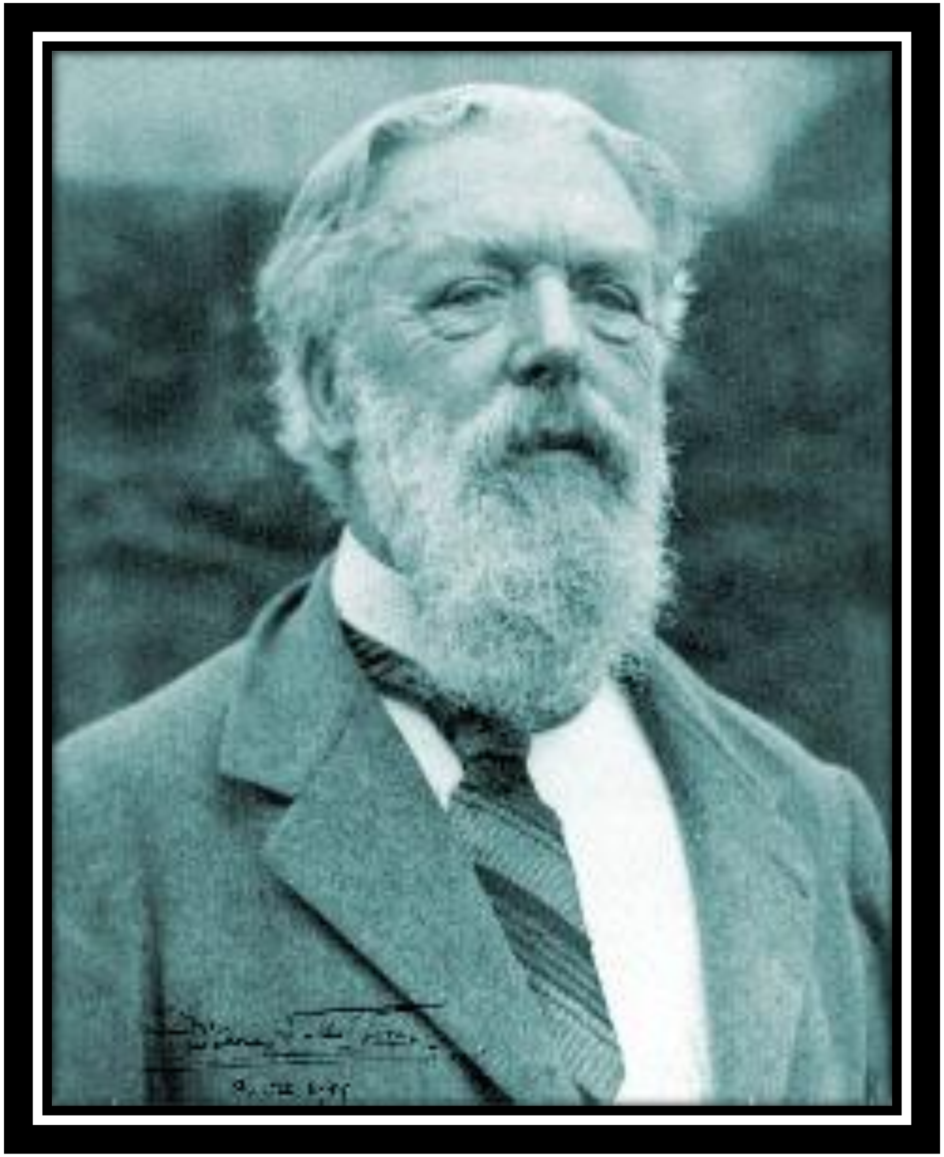
Thompson said, ---” It is not the biologist with an inking of mathematics but the skilled & learned mathematicians who must ultimately deal with such problems as are sketched and adumbrated here”.



ON GROWTH AND FORM

The Complete Revised Edition

D'Arcy Wentworth Thompson



D'Arcy Wentworth Thompson CB FRS FRSE (2 May 1860, Edinburgh – 21 June 1948, St Andrews) was a Scottish biologist, mathematician, and classics scholar.

Nautilus Shell

Nautilus is a marine animal (mollusc) distantly related to octopus. Unlike the latter which can be found in shallow water and is very familiar, Nautilus inhabits the deep seas. Generally we notice the shells of dead Nautilus cast on the sea beach.

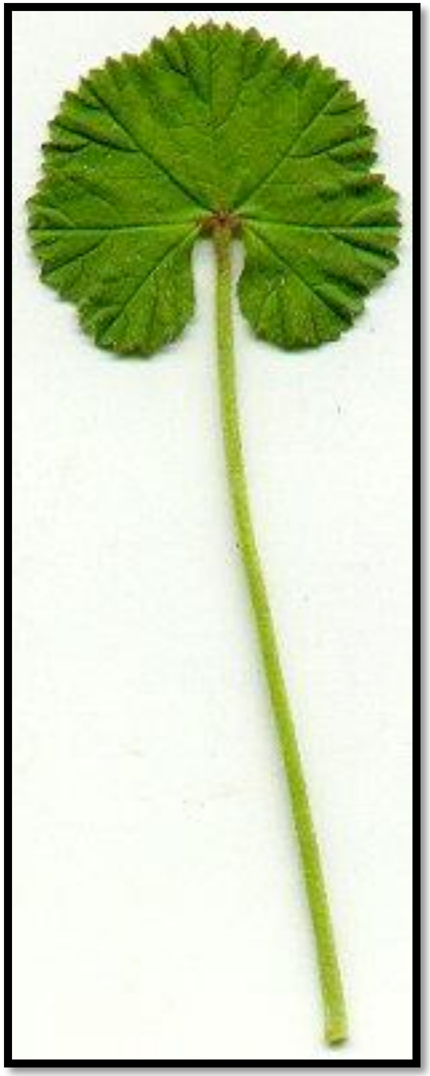
$$r = ae^{b\theta} \; where \; a, b \; are \; parameters.$$

With progressive age each chamber becomes larger than the previous one, always bearing a constant proportion to the former.

Interesting fact is the proportion is basically the **Golden ratio**

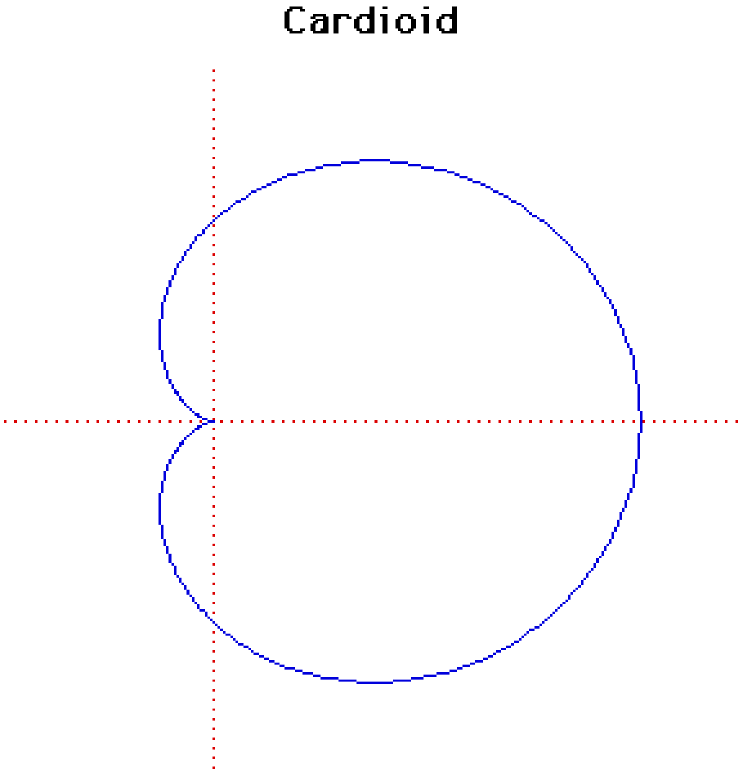
Likewise, Thompson described a curve resembling the outline of a reniform leaf. Variations of the formula $r=\sin(\phi/2)$ can generate various shapes.

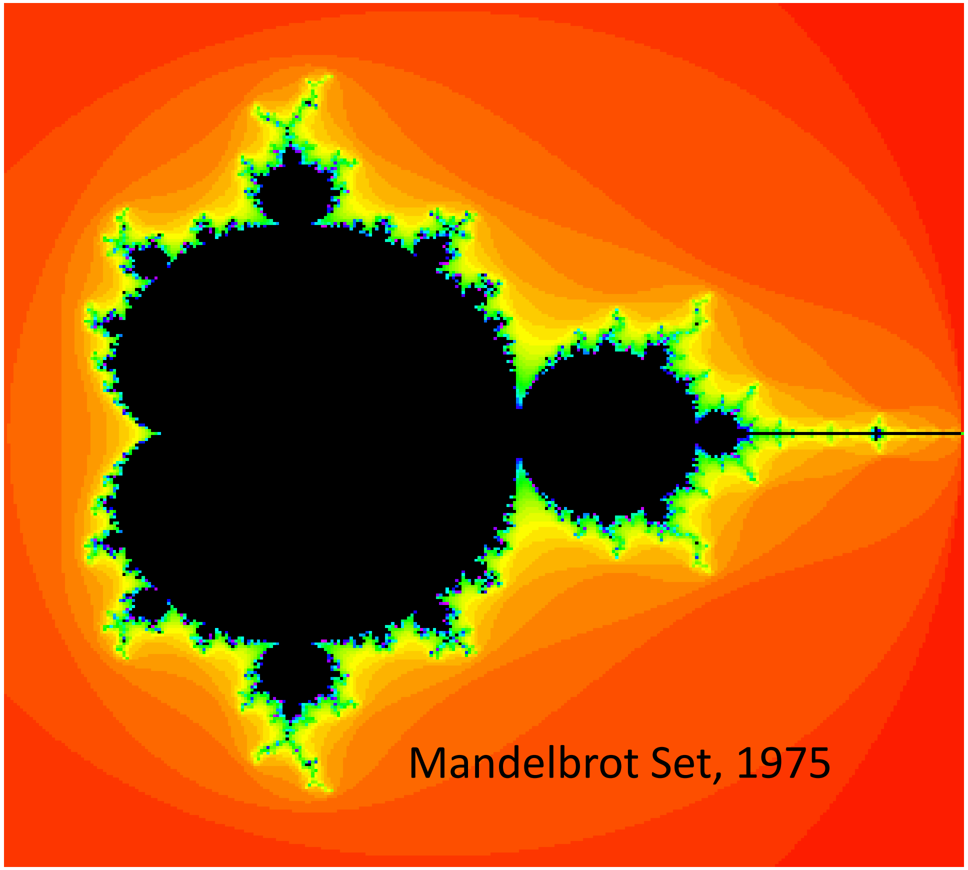
The late T. A. Davis, ISI, Kolkata
worked on Sunflower head.

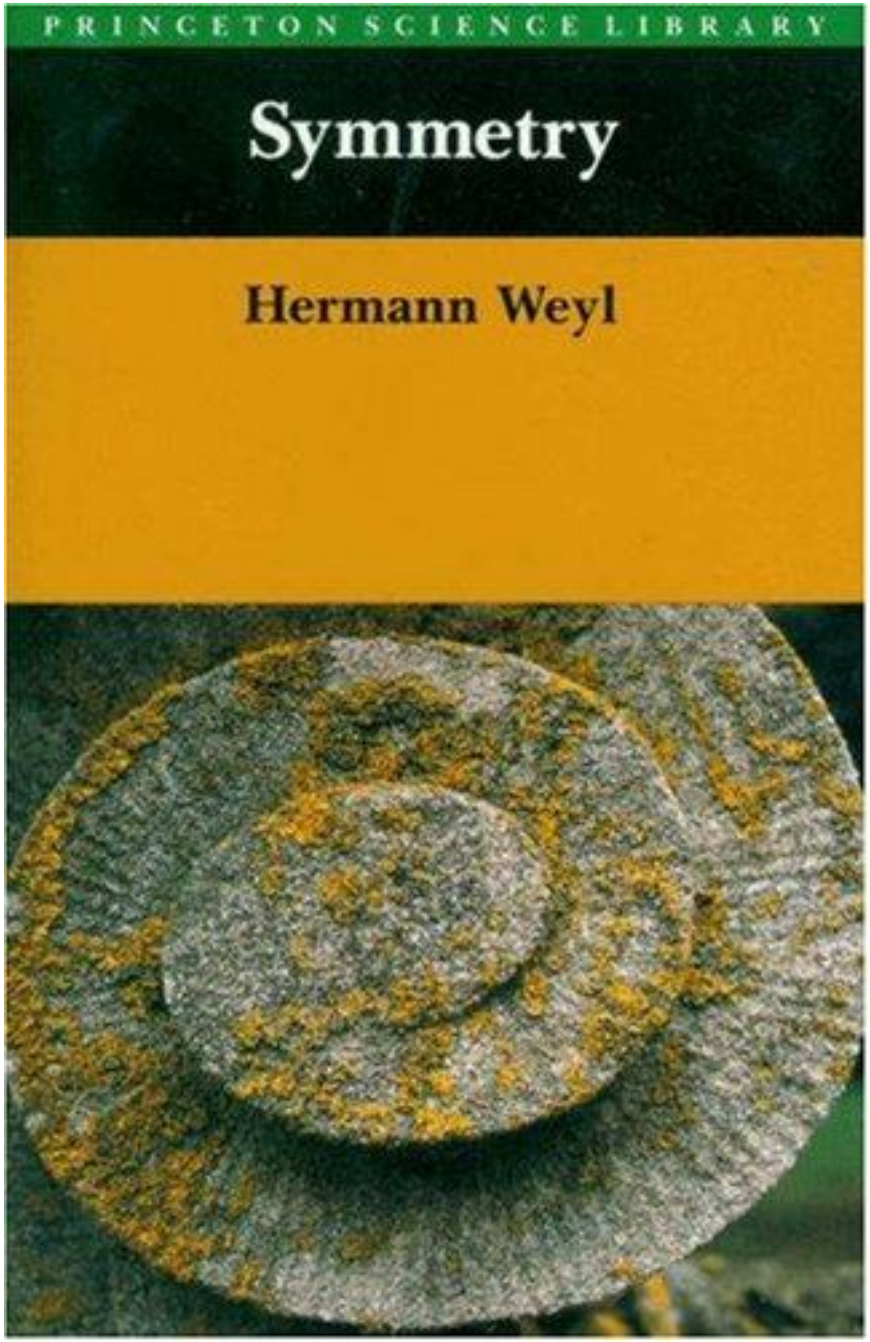


The symmetry pattern in the spirals of the sunflower head was noticed by Thompson. Around 1970, *T. A. Davis of the Indian Statistical Institute* worked on this aspect. Some details are sketched below:

Thompson points out that as early as in 1728 Grandi developed a class of mathematical curves and pointed out the botanical analogies.



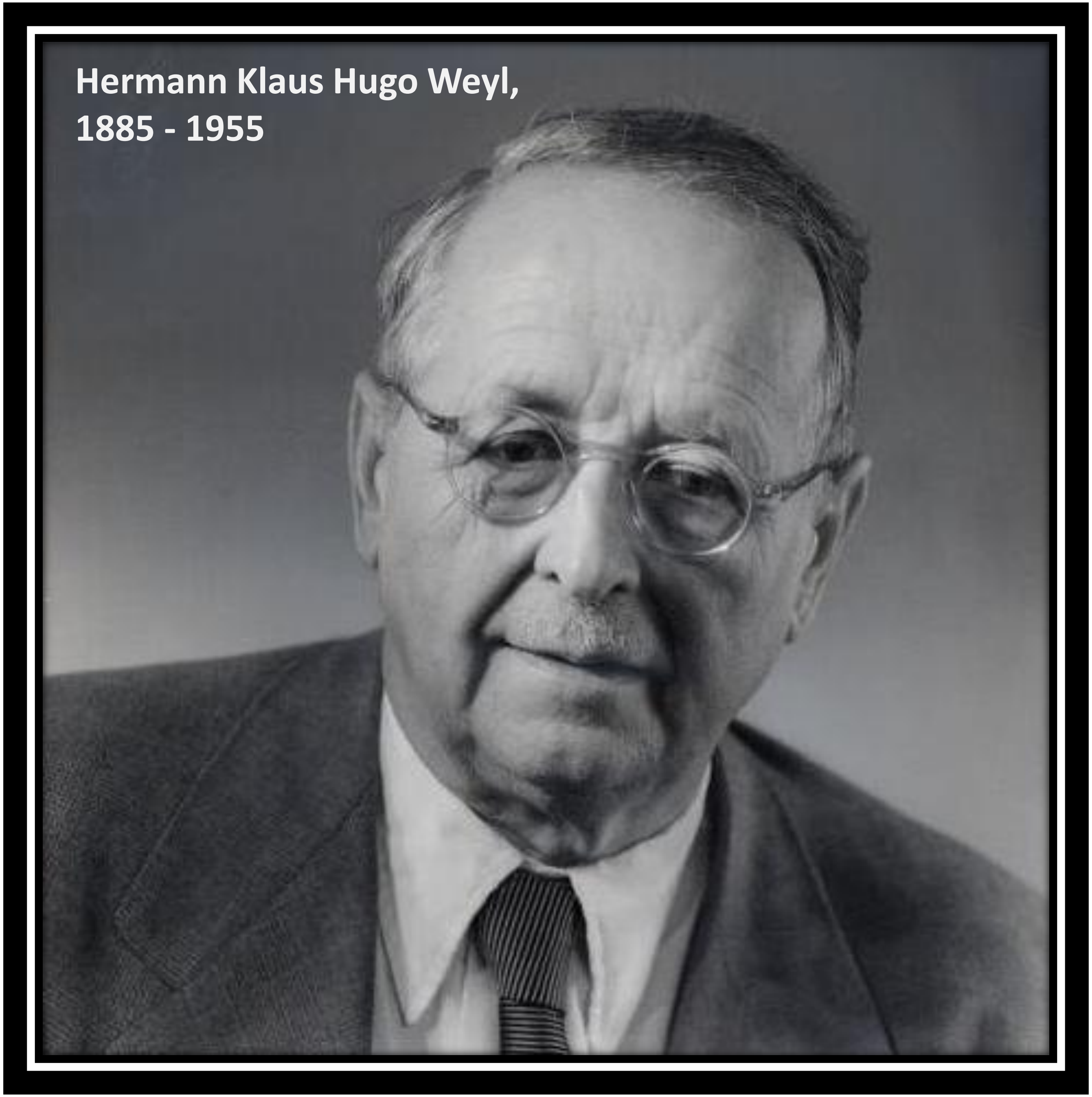




PRINCETON SCIENCE LIBRARY

Symmetry

Hermann Weyl



Hermann Klaus Hugo Weyl,
1885 - 1955

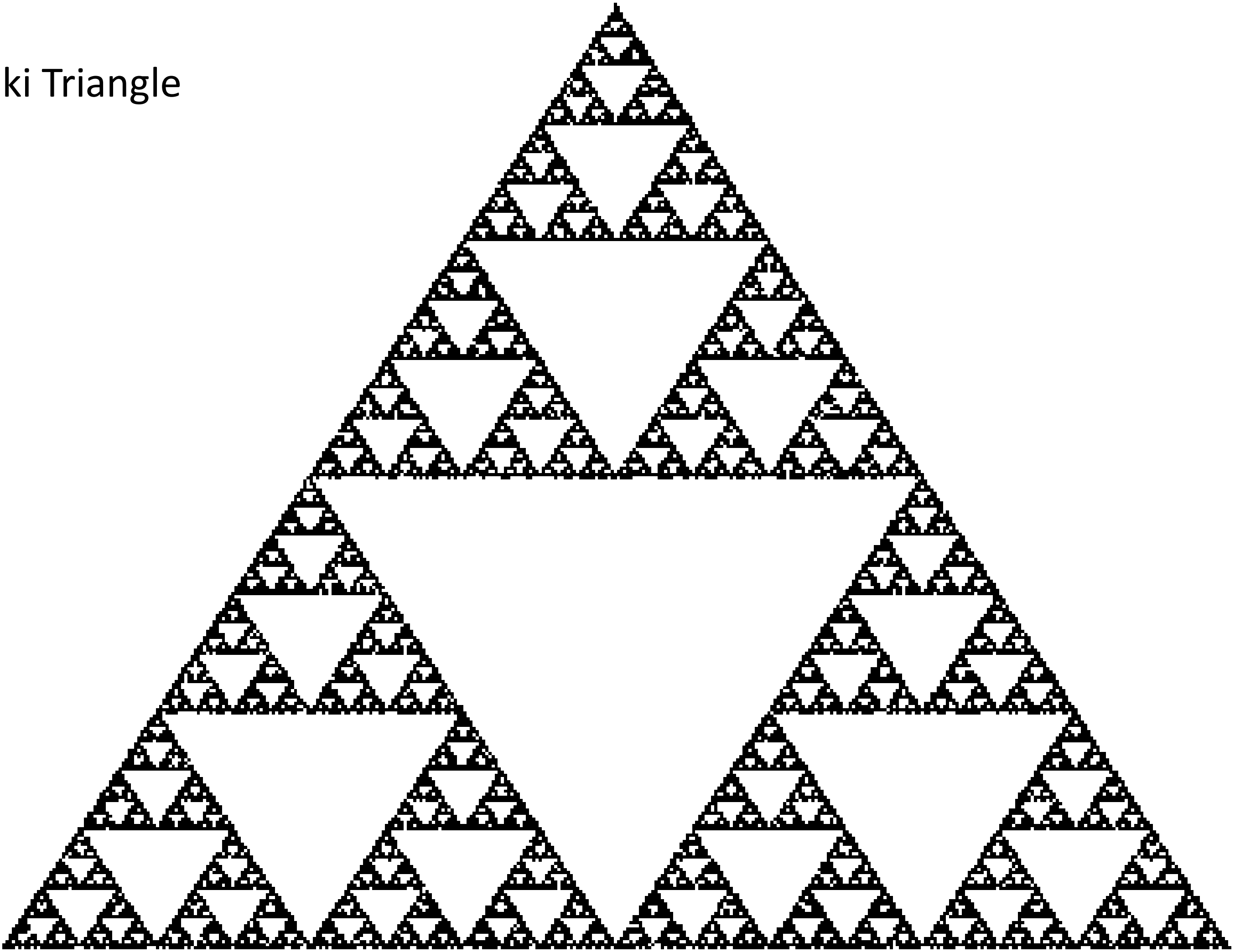
In 1951, **Hermann Klaus Hugo Weyl**, a great early exponent of Relativity theory and, Group Theory & Quantum mechanics and also wrote a book, SYMMETRY.

In this work, Weyl refers to D'arcy Thompson and draws from many other sources such as Haeckel's collection of drawings on diatoms (unicellular organism with beautiful silica shell), brittle stars (echinodermata) and medusae. Many of these have bilateral, radial and/or spherical symmetry. He also emphasized the surprisingly symmetrical figures of ice crystals.

Mathematicians like B. Mandelbrot (1975) have been interested in FRACTALS, i.e. pattern of self similarity which was also pointed out by Weyl.

Self Symmetry

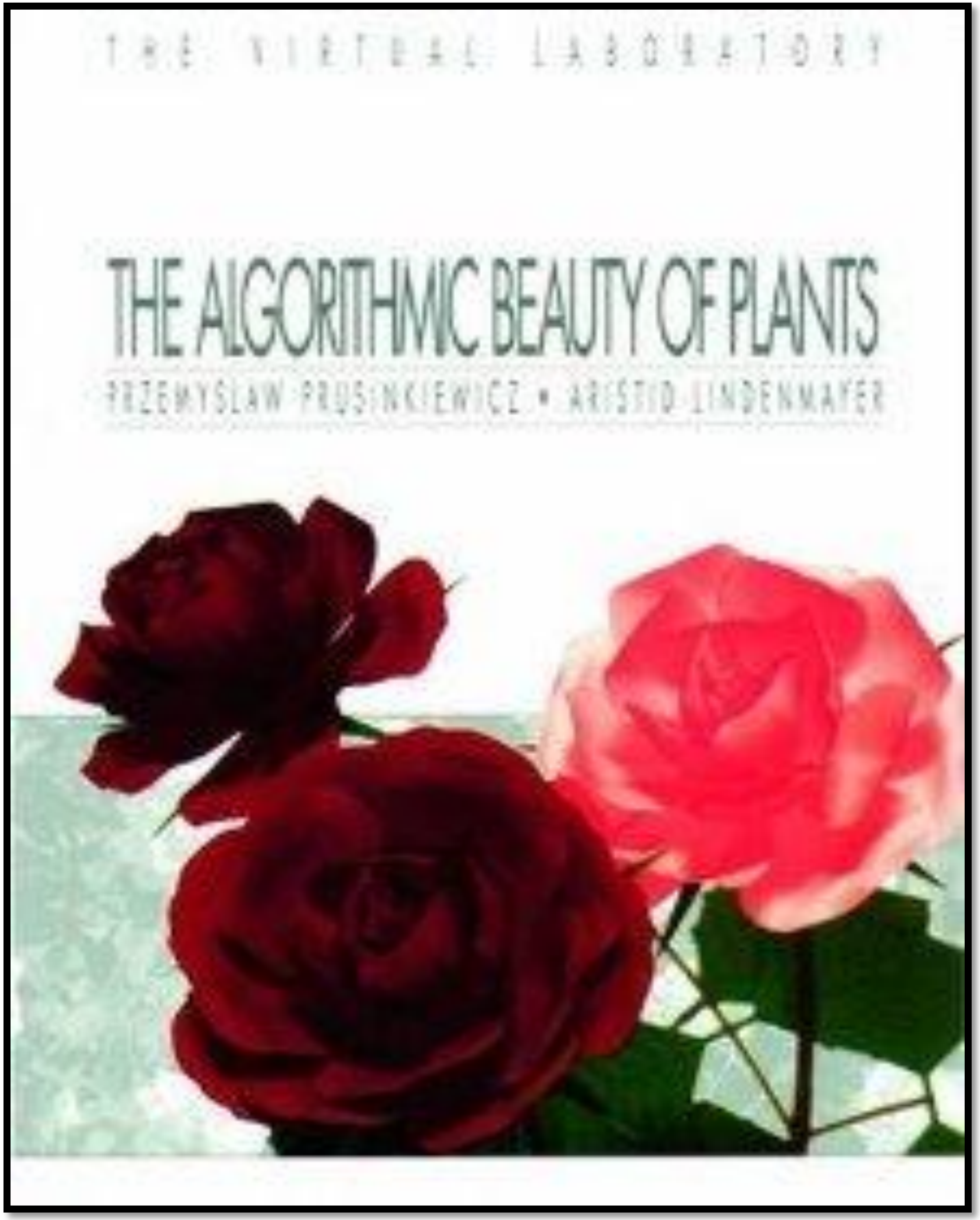
Sierpinski Triangle



In metameric structural patterns in animals such as in the centipede shown here the different segments of the body are nearly of the same size. Likewise in plants buds or branchlets may appear at nearly regular distances in the branches. Weyl points out that such regular intervals translated into one dimensional time would be equivalent to a rhythm in music.

The mathematics of music is not yet developed and he says it is no wonder because the mathematics of symmetry group was developed some four thousand years after such symmetry was depicted in Egyptian Art.

More relevant for our purpose, namely to study the underlying mathematical principle of genome sequences, is the Lindenmayer System. In short, L-System.



THE VIRTUAL LABORATORY

THE ALGORITHMIC BEAUTY OF PLANTS

PRZEMYSLAW PRUSINKIEWICZ • ARISTID LINDENMAYER

Aristid Lindenmayer (November 17, 1925 – October 30, 1989) was a Hungarian biologist

Lindenmayer Systems (L-systems)

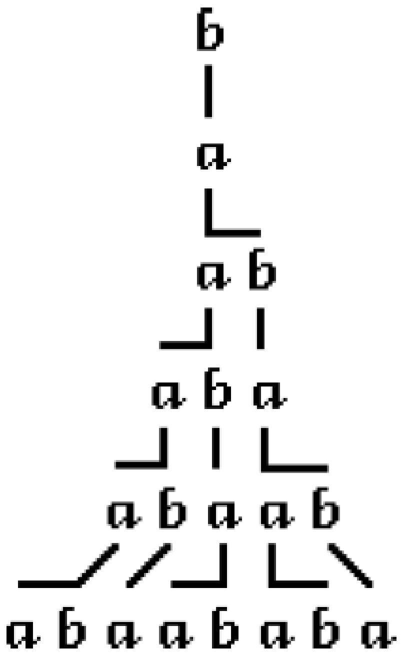
The central concept of L-systems is that of *Rewriting*. In general, *Rewriting* is a technique for defining complex objects by successively replacing parts of a simple initial object using a set of *rewriting rules or productions.*

Consider strings built of two letters *a and b,*

Production Rules:

$a \rightarrow ab$

$b \rightarrow a$



Example of a derivation in a DOL-system

We have started our journey with Human Olfactory Receptor OR1D2.

*** Why Human?**

****Why Olfactory Receptor?**

*****Why OR1D2?**

Why Human?

- Human genome has been sequenced.

Why we start our journey with Olfaction

We took Olfactory Receptors (ORs) because of their unique features…

- ORs loci in human genome occur in clusters and are unevenly spread over 21 chromosomes.
- Human ORs are free from any stop codon (TAG, TGA, TAA). Such receptors are known to the biologist as exons.
- OR sequence length is relatively small, almost 1000bp.

Before going to ‘Why OR1D2’ let us warm up with OR1D2

- OR denotes Olfactory Receptors.
- 1 denotes family name.
- D denotes subfamily name.
- 2 denotes the member name.

WHY OR1D2?

- **Ligands for only Twelve human olfactory receptors are known.**

- One of them, OR1D2, binds to Bourgeonal, a volatile chemical constituent of the fragrance of **Lily of the valley** or **Our Lady's tears**, **Convallaria majalis** (also the **national flower of Finland). Picture of the flower should be attached**

>Human Olfactory Receptor (**OR1D2**): Full length gene sequence (Exon)= 936 bp (A,T,G,C)

**ATGGATGGAGGCAACCAGAGTGAAGGTTCAGAGTTCCTTCTCCTGGGGATGTCAGAGAGTC
CTGAGCAGCAGCGGATCCTGTTTTGGATGTTCCTGTCCATGTACCTGGTCACGGTGGTGGG
AAATGTGCTCATCATCCTGGCCATCAGCTCTGATTCCCGCCTGCACACCCCCGTGTACTTC
TTCCTGGCCAACCTCTCCTTCACTGACCTCTTCTTTGTCACCAACACAATCCCCAAGATGC
TGGTGAACCTCCAGTCCCATAACAAAGCCATCTCCTATGCAGGGTGTCTGACACAGCTCTA
CTTCCTGGTCTCCTTGGTGGCCCTGGACAACCTCATCCTGGCTGTGATGGCATATGACCGC
TATGTGGCCATCTGCTGCCCCCTCCACTACACCACAGCCATGAGCCCTAAGCTCTGTATCT
TACTCCTTTCCTTGTGTTGGGTCCTATCCGTCCTCTATGGCCTCATACACACCCTCCTCAT
GACCAGAGTGACCTTCTGTGGGTCACGAAAAATCCACTACATCTTCTGTGAGATGTATGTA
TTGCTGAGGATGGCATGTTCCAACATTCAGATTAATCACACAGTGCTGATTGCCACAGGCT
GCTTCATCTTCCTCATTCCCTTTGGATTCGTGATCATTTCCTATGTGCTGATTATCAGAGC
CATCCTCAGAATACCCTCAGTCTCTAAGAAATACAAAGCCTTCTCCACCTGTGCCTCCCAT
TTGGGTGCAGTCTCCCTCTTCTATGGGACACTTTGTATGGTATACCTAAAGCCCCTCCATA
CCTACTCTGTGAAGGACTCAGTAGCCACAGTGATGTATGCTGTGGTGACACCCATGATGAA
TCCCTTCATCTACAGCCTGAGGAACAAGGACATGCATGGGGCTCTGGGAAGACTCCTAGAT
AAACACTTTAAGAGGCTGACA**

*** Sequence is collected from HORDE
http://genome.weizmann.ac.il/horde/

An Invitation To The Problem

It appears that OR1D2 is a string of A, T, C, and G RANDOMLY inserted one after another.

In realty, this may not be the case!!!!

Had it been so, then neither we would be able to smell the fragrance of rose properly nor distinguish it from that of Lilly.

So, There Should Be A Beautiful Organization in The Sequence.

Origin of work

Nature could make 4**936 OR sequences, instead nature has selected only 300 (Approx) as functional OR in human and 300 are considered to be non-functional (called pseudo genes).

Now, what governs this selection process?

What are the selection keys?

OR1D2, OR1D4 and OR1D5

- **OR1D2 was used as query in BLASTn search for similar ORs in HORDE and OR1D4 and OR1D5 with more than 80% identical sequences were found in the same genomic loci and the above three ORs have 108 base pair mismatches among them.**
- **As per HORDE, ORs with >80% nucleotide match are called members of the same subfamily of OR. >60% : OR Family.**

Methodology: L-System

- 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 a close relative of the full length olfactory receptors of the same subfamily can be generated.

The 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

The L-System Generated Sequence

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

Conclusion in Research

We claim In two ways:

1. ***If The above sequence functions*** **as OR1D2, then we could say that 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.**

2. ***If the above sequence is not at all functioning*** **as OR1D2, then we could be able to find out the selection key (the functioning part for the olfaction purpose).**

A Big Question to The Biologists!!

- **Since already we know the functioning of OR1D2, our research outcome has to be experimentally confirmed about the dilemma as posed in the previous slide.**
- **That is, whether the created DNA sequence is at all functioning or not functioning.**

Another Approach

- Now you may have the following question…
- We have designed a variable region of ORs. Why not the full receptors?
- To answer this question we are ready with a set of L-systems to capture the whole receptor even genome.
- The procedure is shown in the next few slides…

A New principle

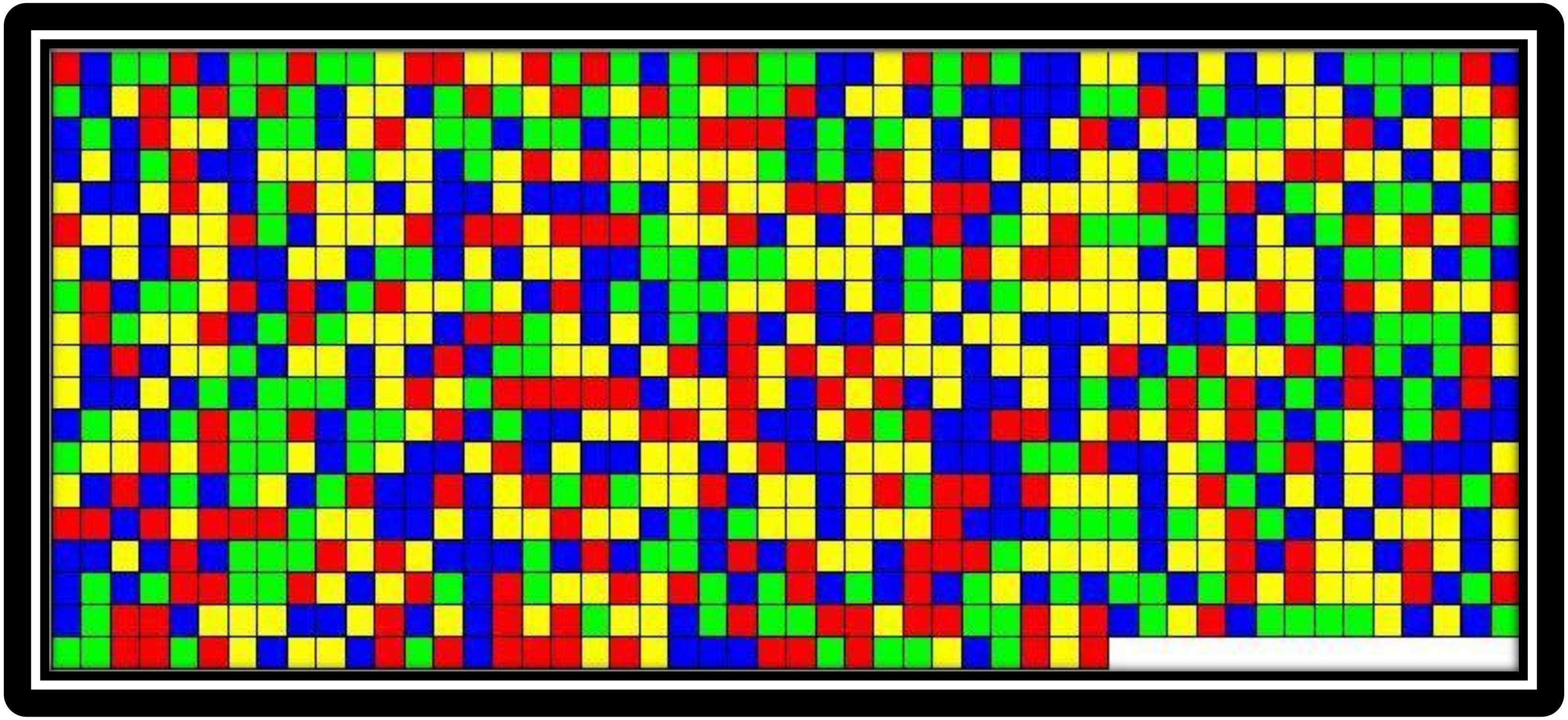
But, Is the Procedure have some novelty?

The beauty of the proposed methodology would enlighten us to answers of some unanswered questions in "**Evolutionary Science**".

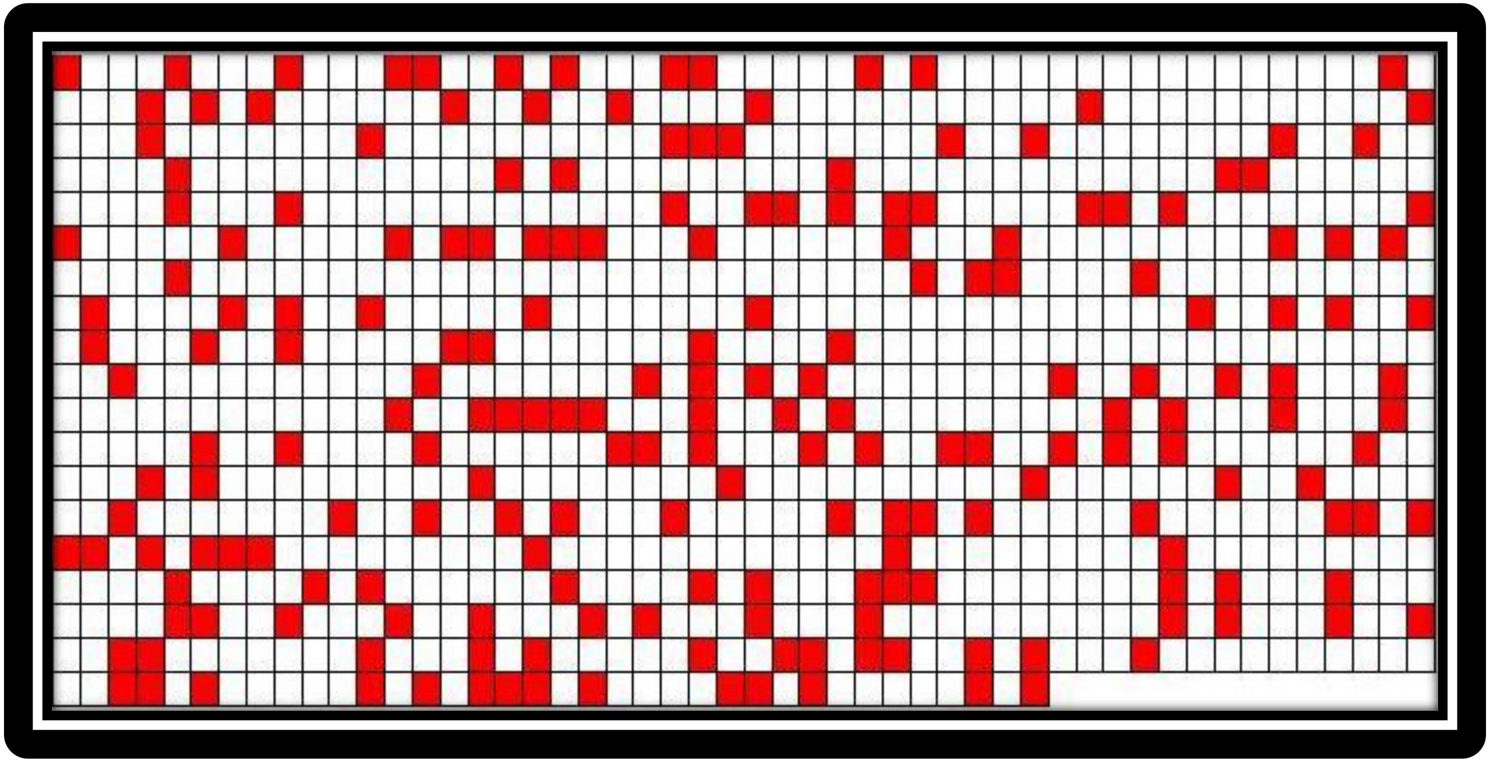
OUR IMMEDIATE RESEARCH ENDEAVORS –SELECTION RATHER THAN CONSTRUCTION

Fractals and Mathematical Morphology **in Deciphering The Quantitative Content in DNA sequences.**

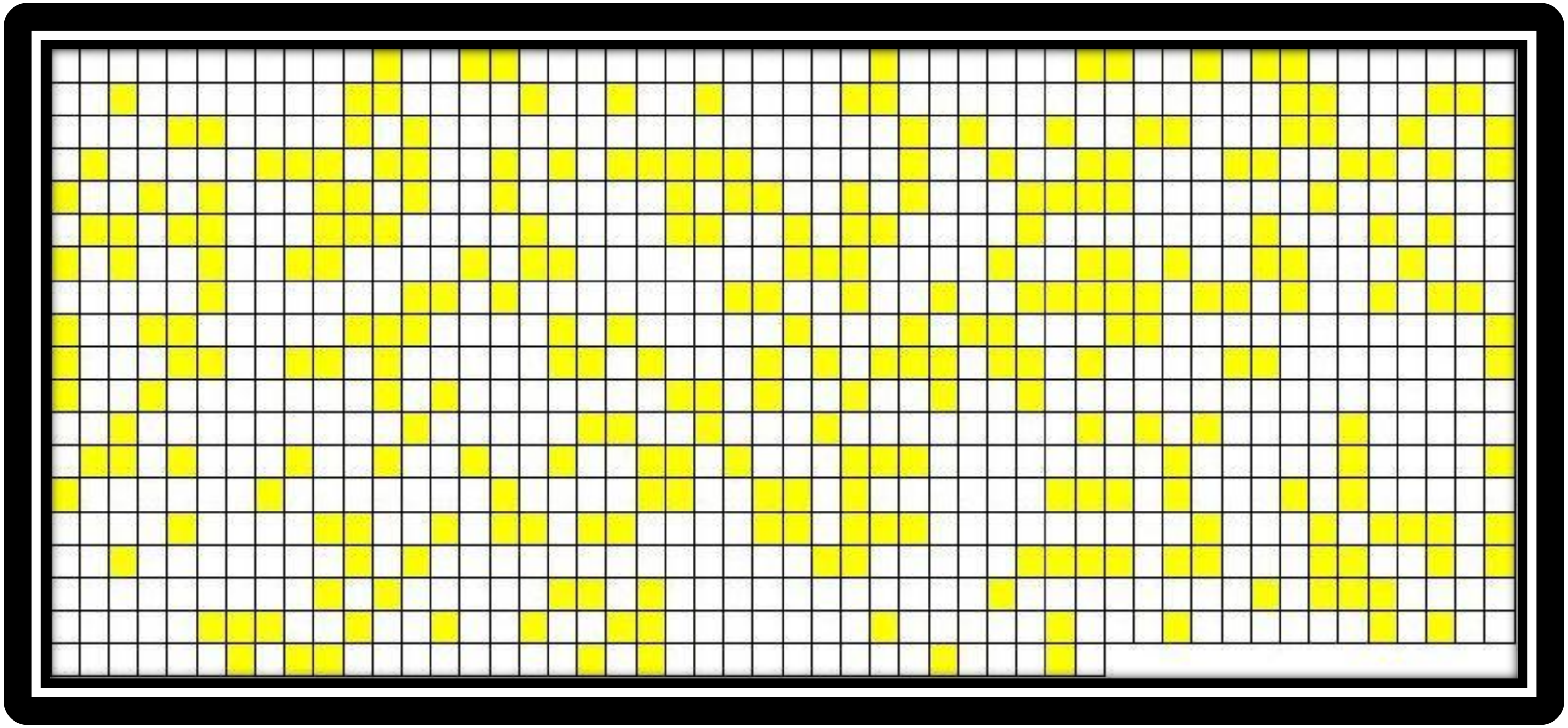
Color Template of DNA, OR1D2



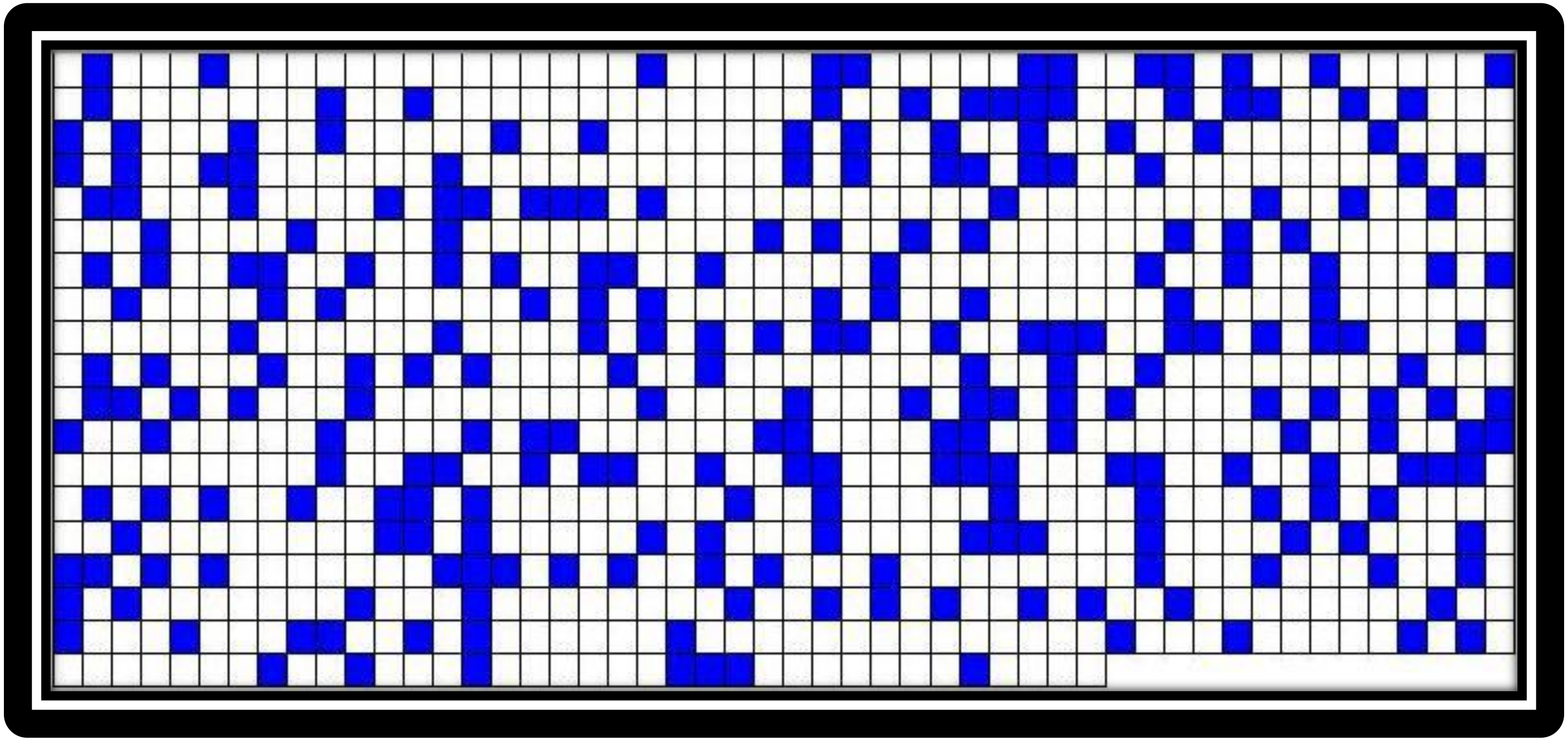
Color Template of ‘A’, OR1D2



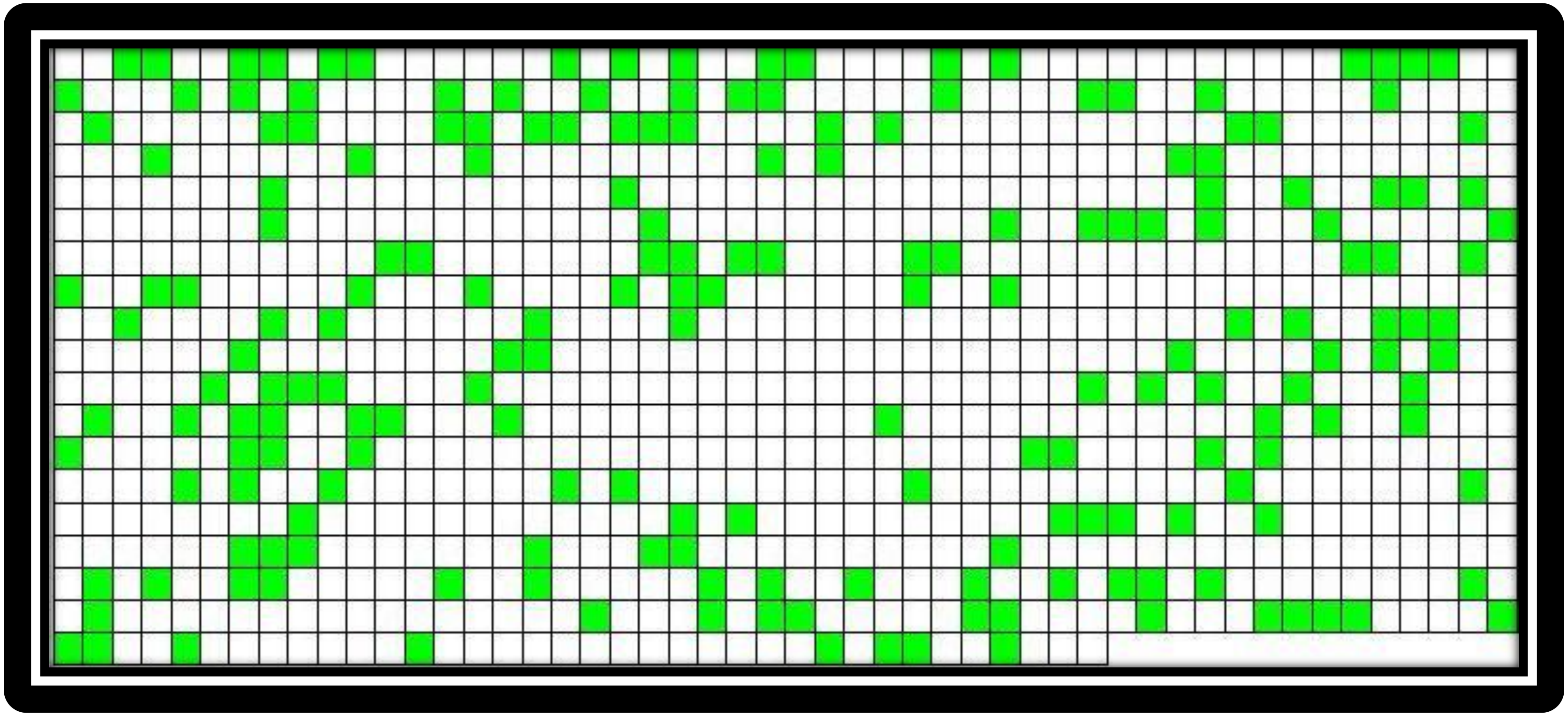
Color Template ‘T’, OR1D2



Color Template of ‘C’, OR1D2



Color Template ‘G’, OR1D2



Succolarity: A Measure for Continuous Density

Succolarity of The Template of A: 0.001026
Succolarity of The Template of T: 0.001690
Succolarity of The Template of G: 0.000522
Succolarity of The Template of C: 0.001482

DNA Sequence: A Multi-Fractal

The Texture of A, T, C, and G generate Fractals.

Collatz function like Integral Value Transformations and subsequent morphological analysis

Collatz Conjecture

A function T is defined on $\mathbb{N}\ to\ \mathbb{N}$. T is defined as follows,

$$T(n) = 3n + 1; if\ n\ is\ odd$$

$$T(n) = \frac{n}{2}\ ; if\ n\ is\ even$$

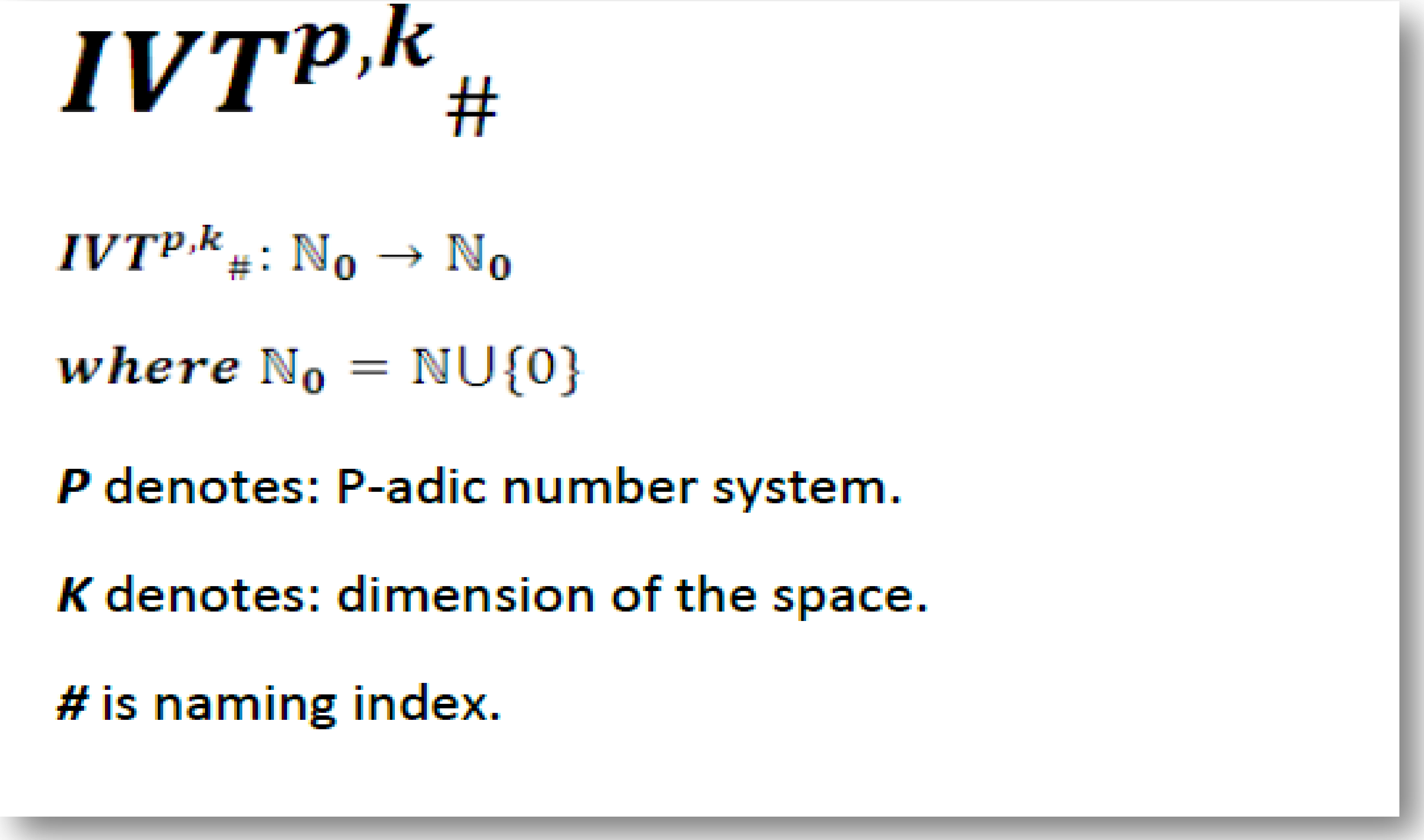
The iterative scheme is defined as $X_{n+1} = T(X_n)$

There is a natural number i such that $X_i = 1$ for all initial value X_0. This is what is known as *Collatz Conjecture*.

Collatz like function

*We have defined a class of transformations named as Integral Value Transformations (IVT) where we have had a Collatz Like function (s) where we have been able to make **proof for Collatz like conjecture** in our context.*

Integral Value Transformations (IVT)



$$IVT^{p,k}{}_{\#}$$

$IVT^{p,k}{}_{\#}: \mathbb{N}_0 \rightarrow \mathbb{N}_0$

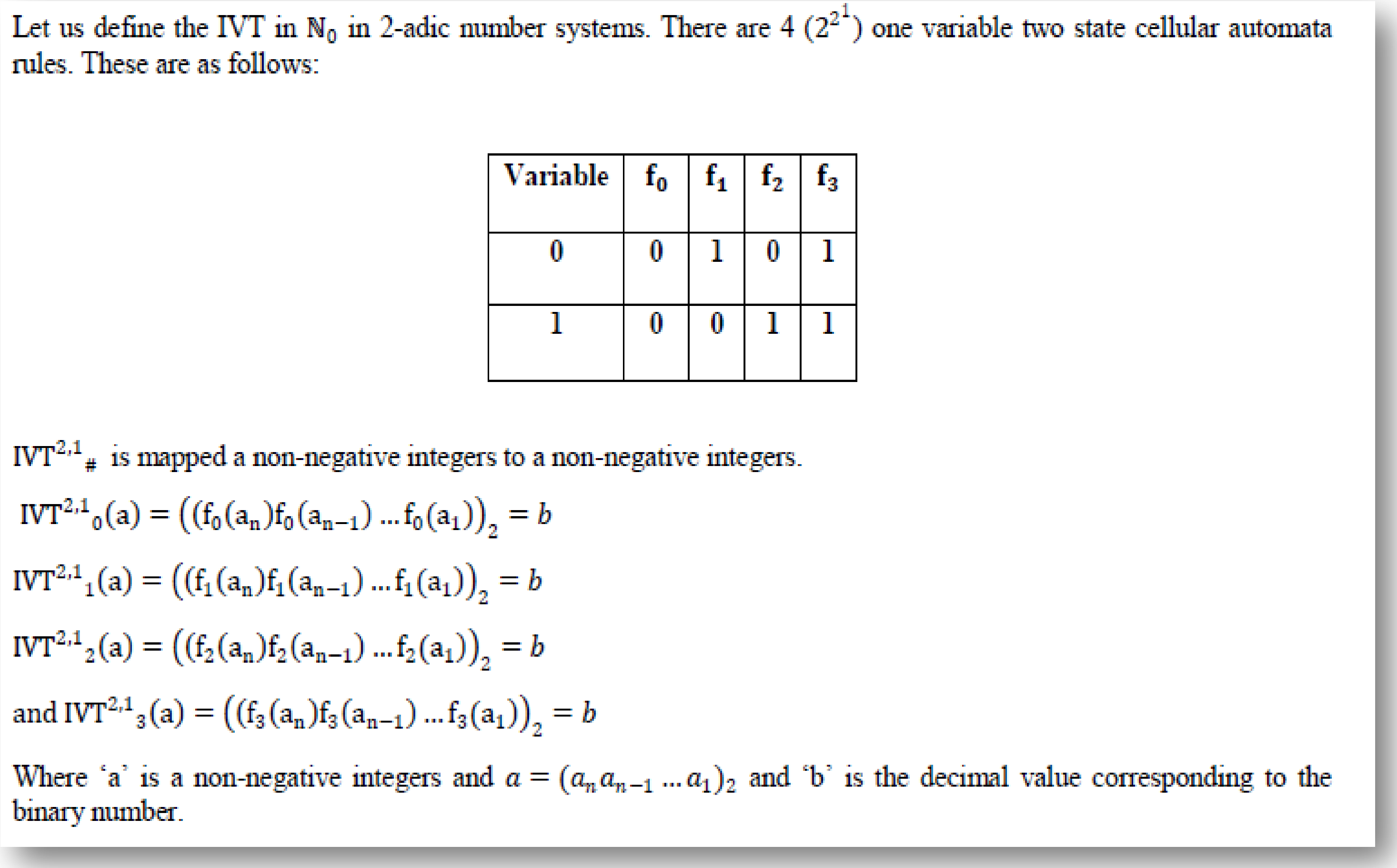
$where\ \mathbb{N}_0 = \mathbb{N} \cup \{0\}$

P denotes: P-adic number system.

K denotes: dimension of the space.

is naming index.

IVT in 1 dimension…



Let us define the IVT in $\mathbb{N}_0$ in 2-adic number systems. There are 4 (2^{2^1}) one variable two state cellular automata rules. These are as follows:

Variable	f_0	f_1	f_2	f_3
0	0	1	0	1
1	0	0	1	1

$IVT^{2,1}{}_{\#}$ is mapped a non-negative integers to a non-negative integers.

$$IVT^{2,1}{}_0(a) = \big((f_0(a_n)f_0(a_{n-1})\ldots f_0(a_1)\big)_2 = b$$

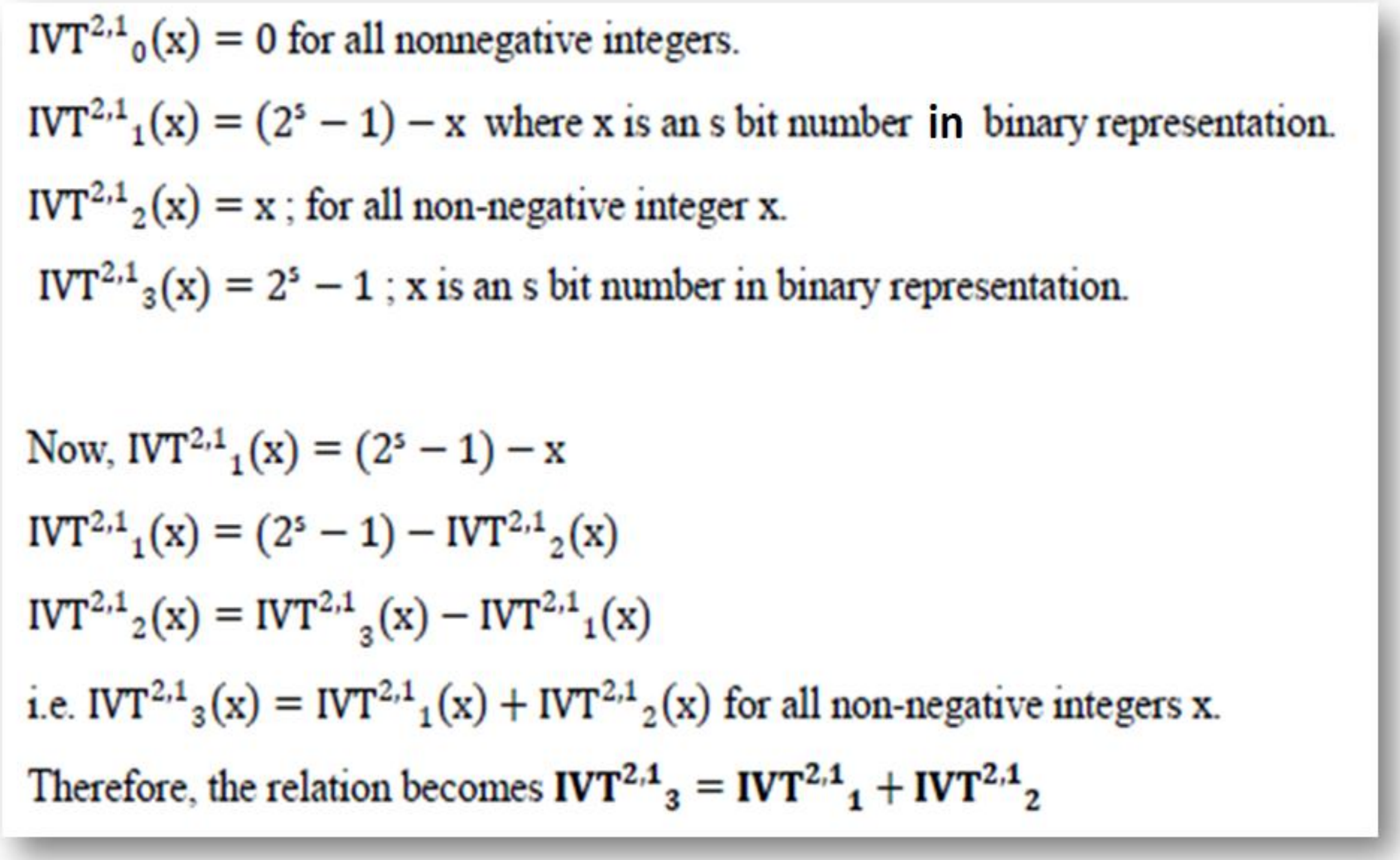
$$IVT^{2,1}{}_1(a) = \big((f_1(a_n)f_1(a_{n-1})\ldots f_1(a_1)\big)_2 = b$$

$$IVT^{2,1}{}_2(a) = \big((f_2(a_n)f_2(a_{n-1})\ldots f_2(a_1)\big)_2 = b$$

and $IVT^{2,1}{}_3(a) = \big((f_3(a_n)f_3(a_{n-1})\ldots f_3(a_1)\big)_2 = b$

Where 'a' is a non-negative integers and $a = (a_n a_{n-1} \ldots a_1)_2$ and 'b' is the decimal value corresponding to the binary number.

Algebraic relation of IVTs



$IVT^{2,1}{}_{0}(x) = 0$ for all nonnegative integers.

$IVT^{2,1}{}_{1}(x) = (2^{s} - 1) - x$ where x is an s bit number in binary representation.

$IVT^{2,1}{}_{2}(x) = x$; for all non-negative integer x.

$IVT^{2,1}{}_{3}(x) = 2^{s} - 1$; x is an s bit number in binary representation.

Now, $IVT^{2,1}{}_{1}(x) = (2^{s} - 1) - x$

$IVT^{2,1}{}_{1}(x) = (2^{s} - 1) - IVT^{2,1}{}_{2}(x)$

$IVT^{2,1}{}_{2}(x) = IVT^{2,1}{}_{3}(x) - IVT^{2,1}{}_{1}(x)$

i.e. $IVT^{2,1}{}_{3}(x) = IVT^{2,1}{}_{1}(x) + IVT^{2,1}{}_{2}(x)$ for all non-negative integers x.

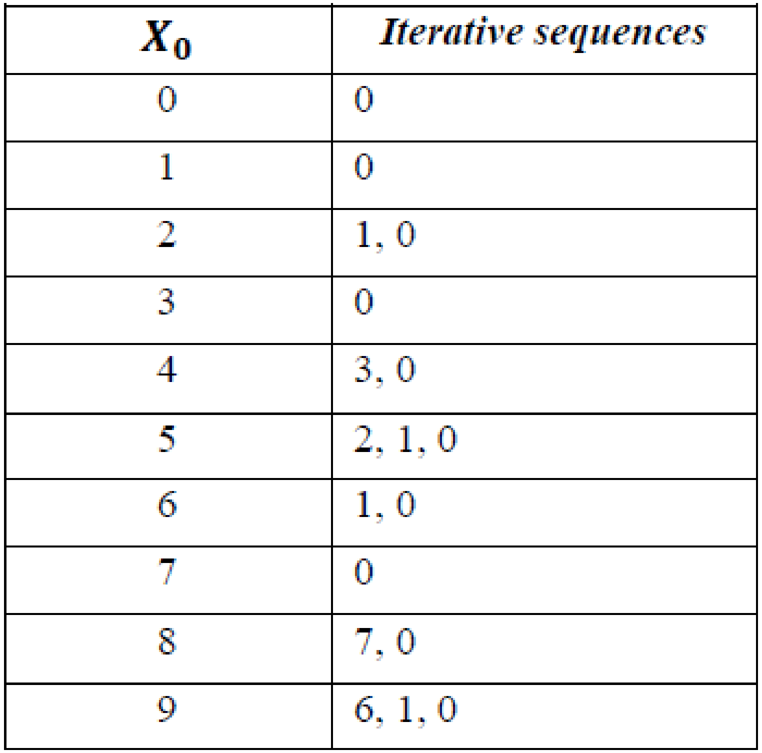
Therefore, the relation becomes $\mathbf{IVT^{2,1}{}_{3} = IVT^{2,1}{}_{1} + IVT^{2,1}{}_{2}}$

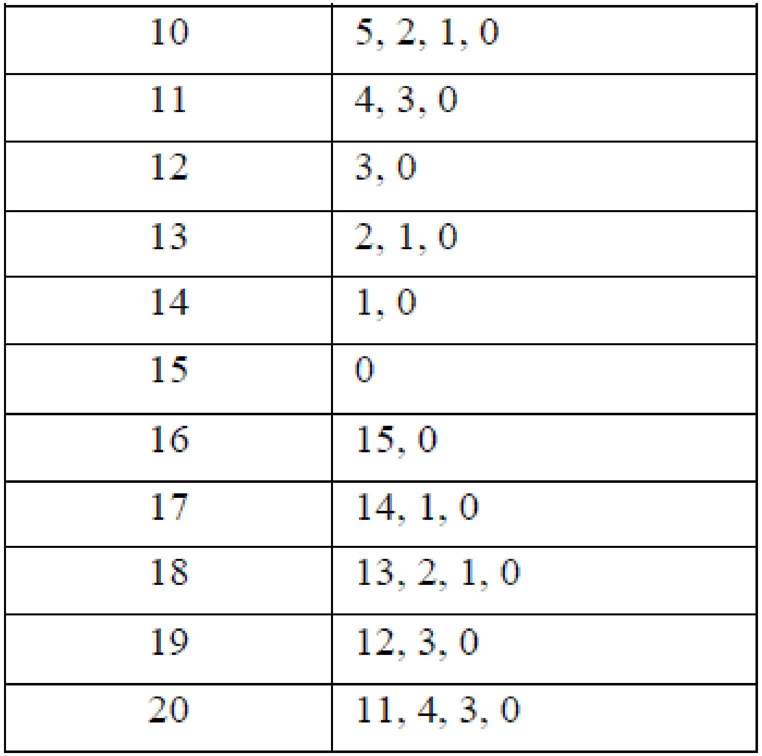
Collatz like Conjecture

Let us consider an iterative scheme as $X_{n+1} = f(X_n)$ where f is a function from $\mathbb{N}_0$ to $\mathbb{N}_0$.

If we consider f as three IVTs' then we have X_{n+1}=0, X_0, $2^s - 1$ for all X_0 (s bit in binary representation) corresponding to $IVT^{2,1}{}_0$, $IVT^{2,1}{}_2$ and $IVT^{2,1}{}_3$ respectively. So, iterative scheme for these three transformations are basically static systems. But the transformation $IVT^{2,1}{}_1$ shows a significant dynamism in the iterative sequence. Let us take one example, let X_0 be 19. 19 (=10011_2) is a five (s) bit number. Therefore 19 maps to $(2^5 - 1) - 19 = 12$.

Iterative convergence





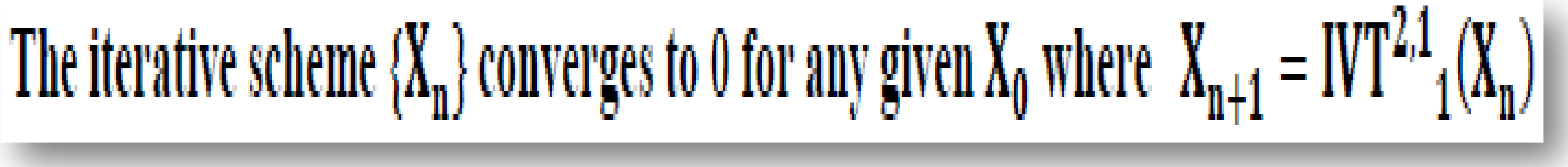
***X*$_0$**	***Iterative sequences***
0	0
1	0
2	1, 0
3	0
4	3, 0
5	2, 1, 0
6	1, 0
7	0
8	7, 0
9	6, 1, 0
10	5, 2, 1, 0
11	4, 3, 0
12	3, 0
13	2, 1, 0
14	1, 0
15	0
16	15, 0
17	14, 1, 0
18	13, 2, 1, 0
19	12, 3, 0
20	11, 4, 3, 0

Lemma

(I) *For any non-negative integer of the form* $X_0 = 2^n + P$, $\mathrm{IVT}^{2,1}{}_1(X_0) = 2^n - (P + 1)$ *for some non-negative integer.*

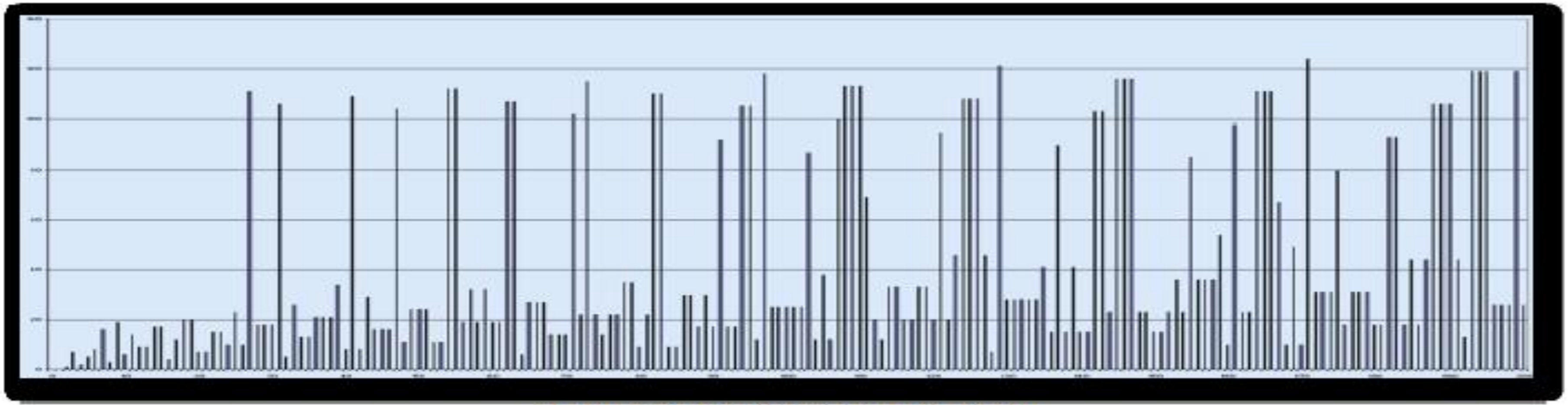
(II) *For any non-negative integer of Merseene form* $X_0 = 2^n - 1$, $\mathrm{IVT}^{2,1}{}_1(X_0) = 0$.

Main Result

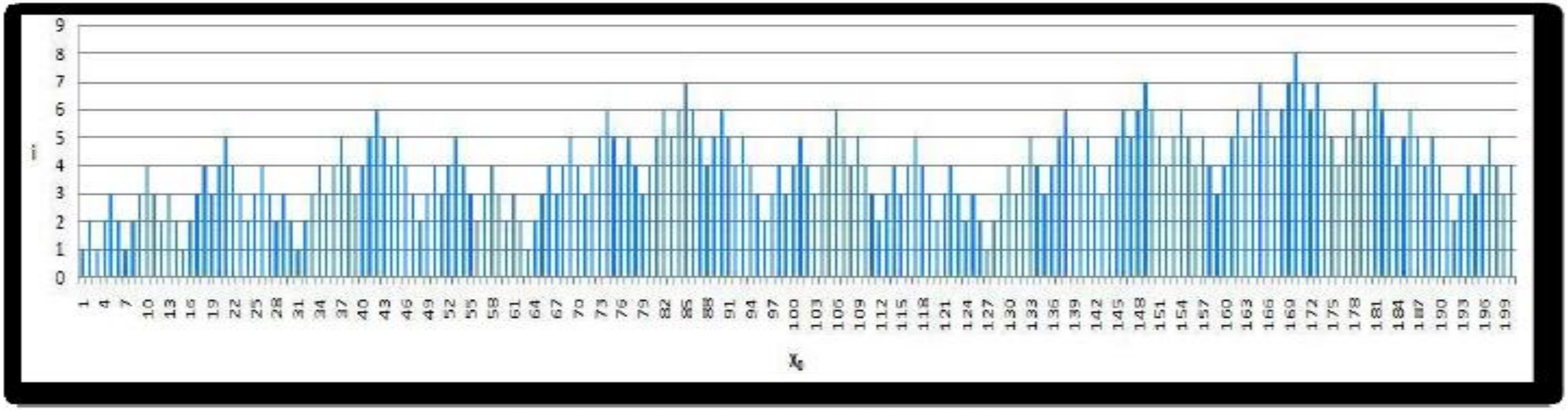


The iterative scheme $\{X_n\}$ converges to 0 for any given X_0 where $X_{n+1} = IVT^{2,1}_{1}(X_n)$

Convergence behavior of Collatz and IVT

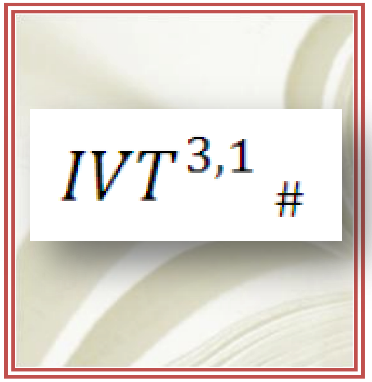


[Figure 1: Collatz Graph (1-200)]



[Figure 2: $IVT^{2,1}{}_1$ Graph (1-200)]

$IVT^{3,1}{}_{\#}$



Clearly there are $3^{3^1} = 27$ integral value transformations.

In this case $\{IVT^{3,1}{}_{1}, IVT^{3,1}{}_{3}, IVT^{3,1}{}_{9}\}$ is basis for the $IVT^{3,1}{}_{\#}$ system.

Algebraic relation

The algebraic formulations of the remaining IVTs can be defined as a linear combination of the above three IVTs. Each $\mathrm{IVT}^{3,1}{}_{\#}$ satisfies the following linear combination.

$$\mathrm{IVT}^{3,1}{}_{\#}=\mathrm{a}.\ \mathrm{IVT}^{3,1}{}_{1}+\mathrm{b}.\ \mathrm{IVT}^{3,1}{}_{3}+\mathrm{c}.\ \mathrm{IVT}^{3,1}{}_{9}$$

$$where\ a, b, c \in \{0,1,2\}$$

satisfies the following relation:

$$\# = a + 3b + 9c$$

Collatz behavioral IVTs

$IVT^{3,1}_{0}$

$IVT^{3,1}_{1}$

$IVT^{3,1}_{2}$

$IVT^{3,1}_{6}$

$IVT^{3,1}_{7}$

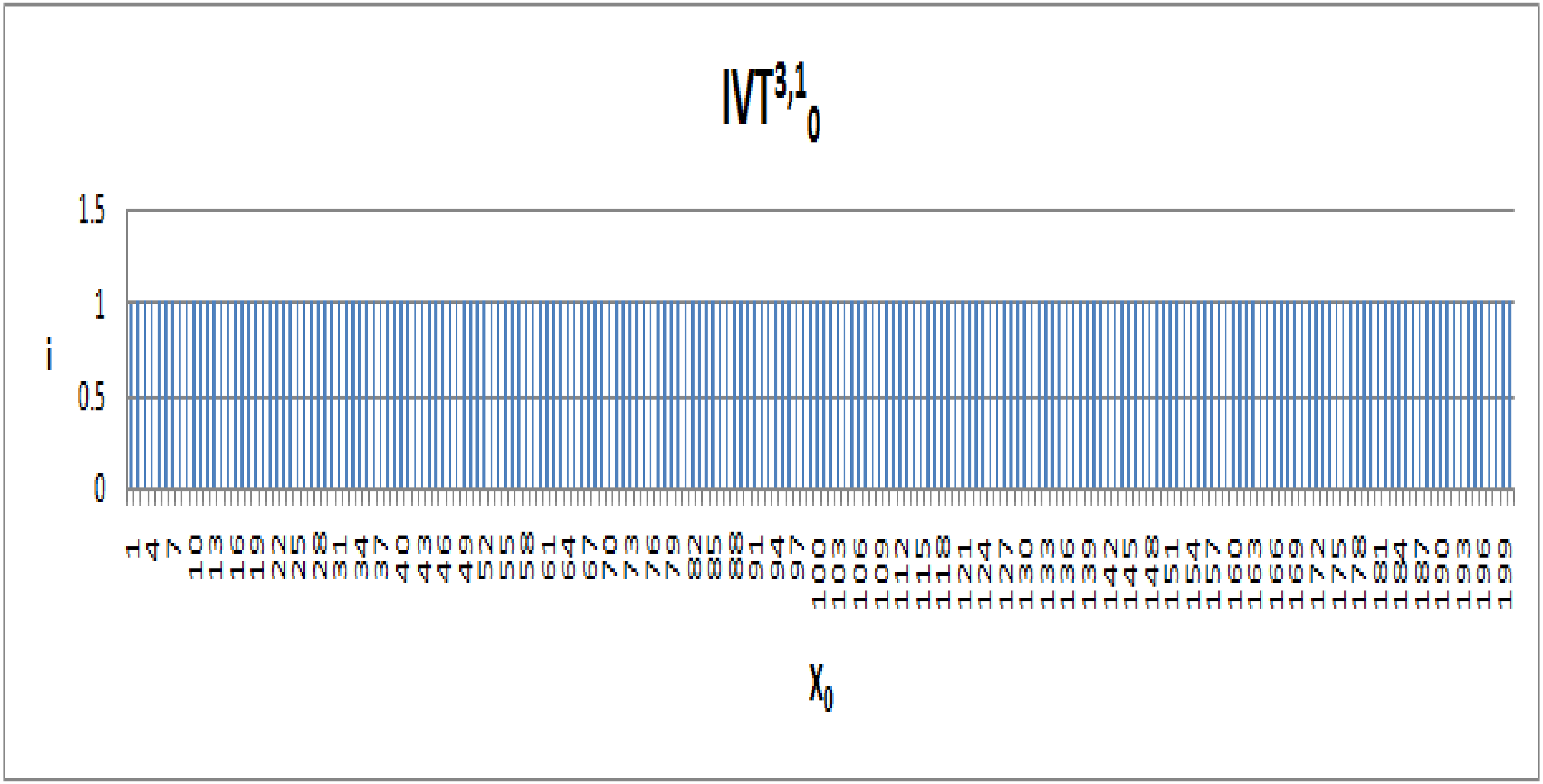
$IVT^{3,1}_{8}$

$IVT^{3,1}_{9}$

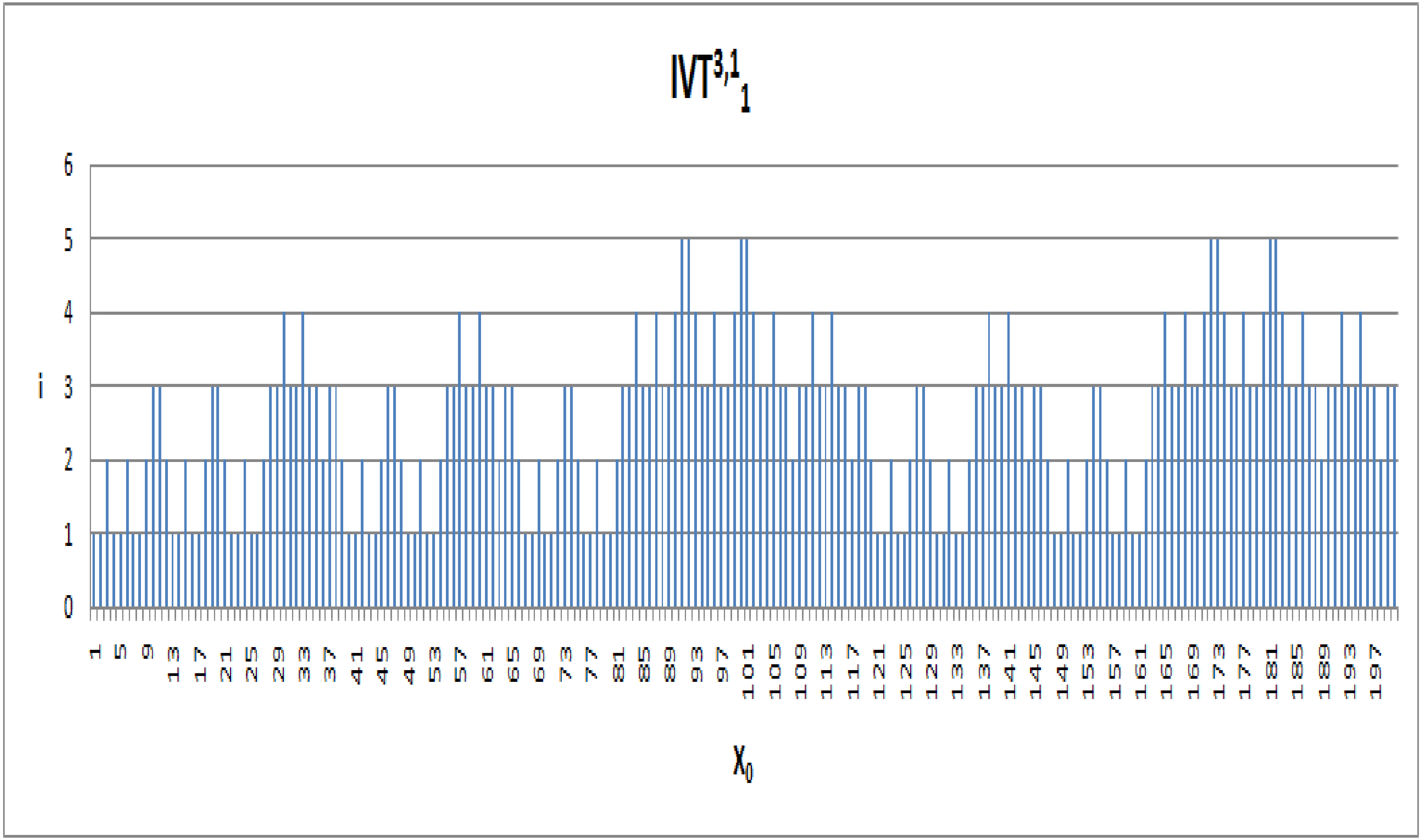
$IVT^{3,1}_{10}$

$IVT^{3,1}_{11}$

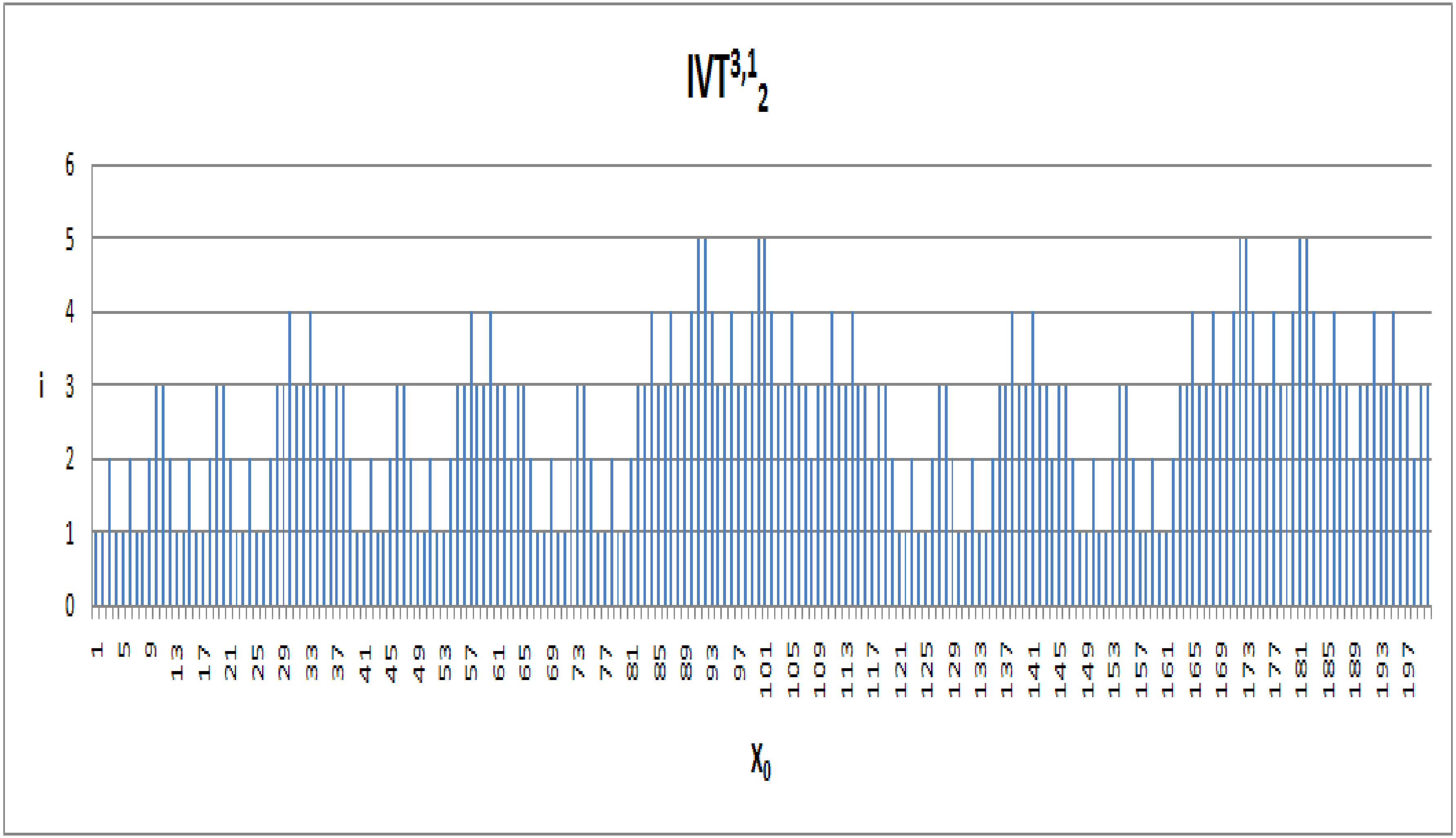
Convergence Graph of $IVT^{3,1}{}_{0}$



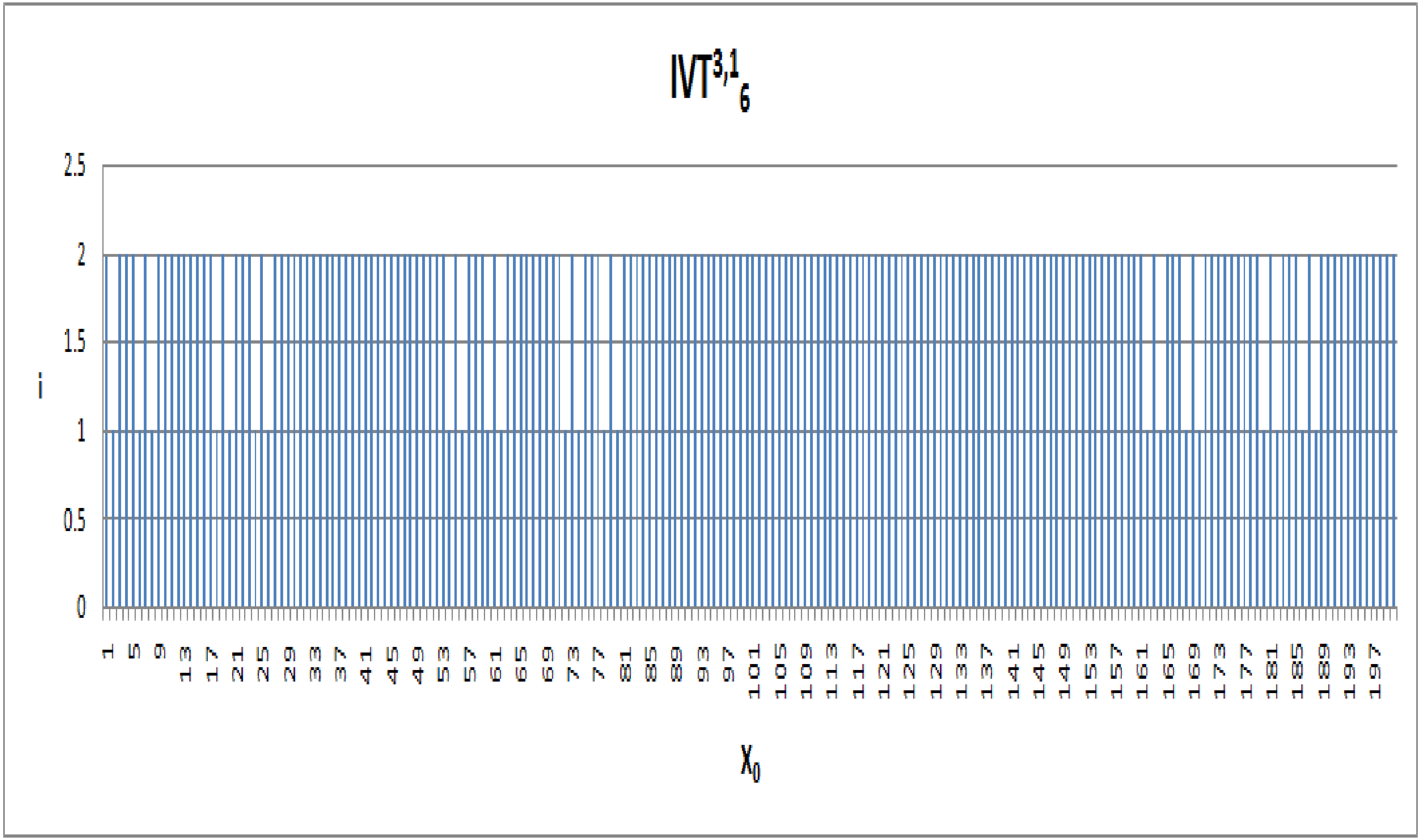
Convergence Graph of $IVT^{3,1}{}_{1}$



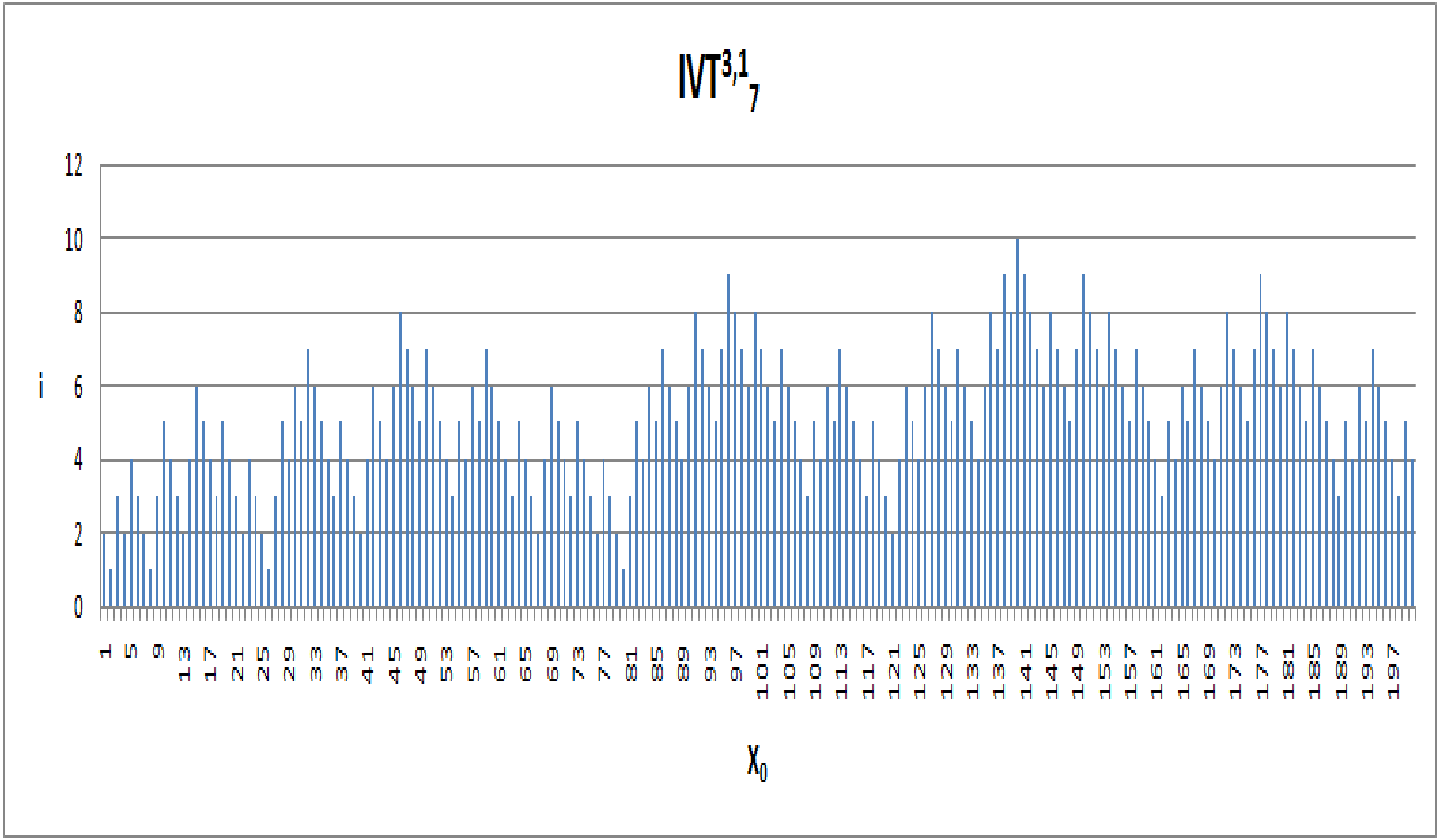
Convergence Graph of $IVT^{3,1}{}_2$



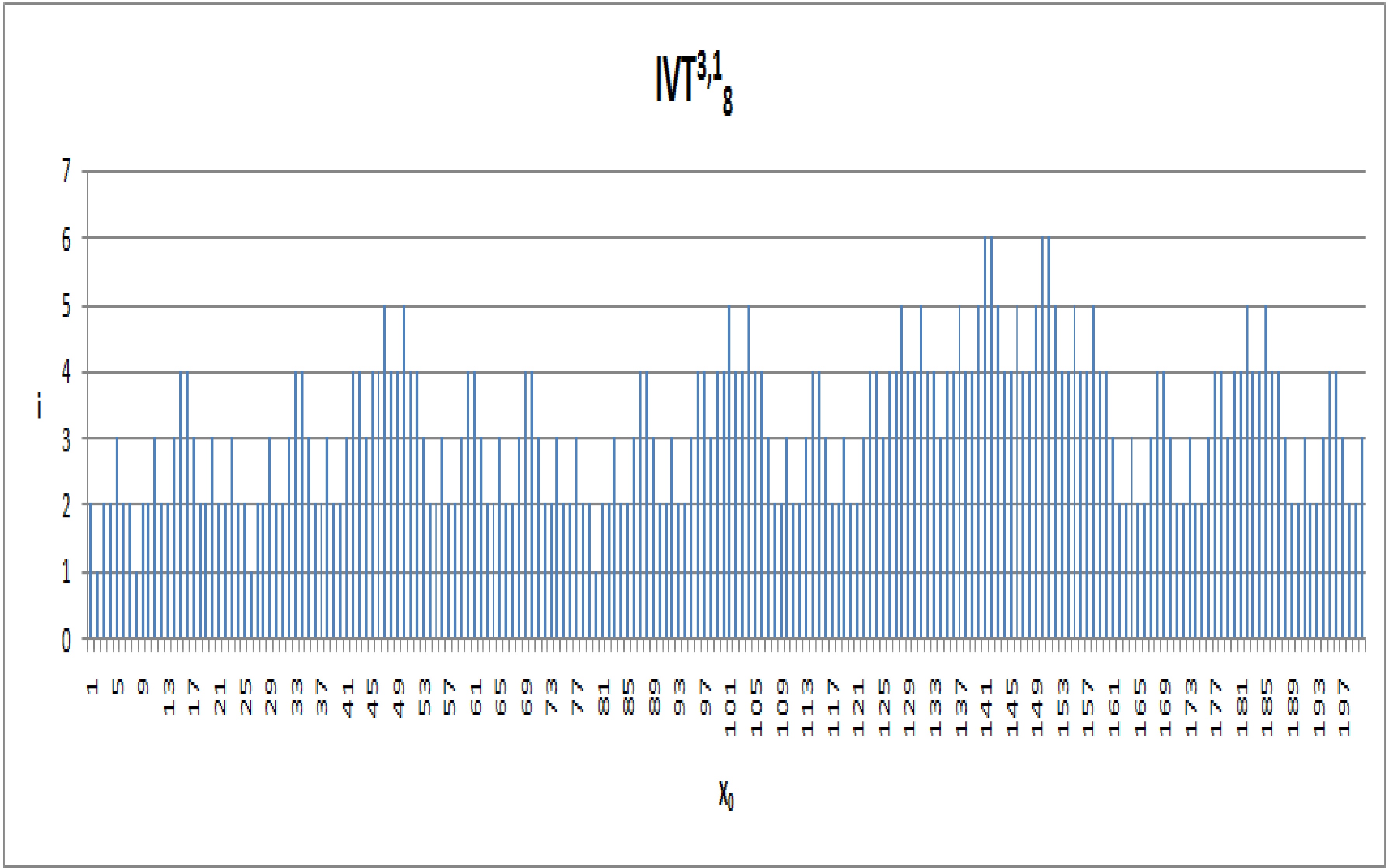
Convergence Graph of $IVT^{3,1}{}_{6}$



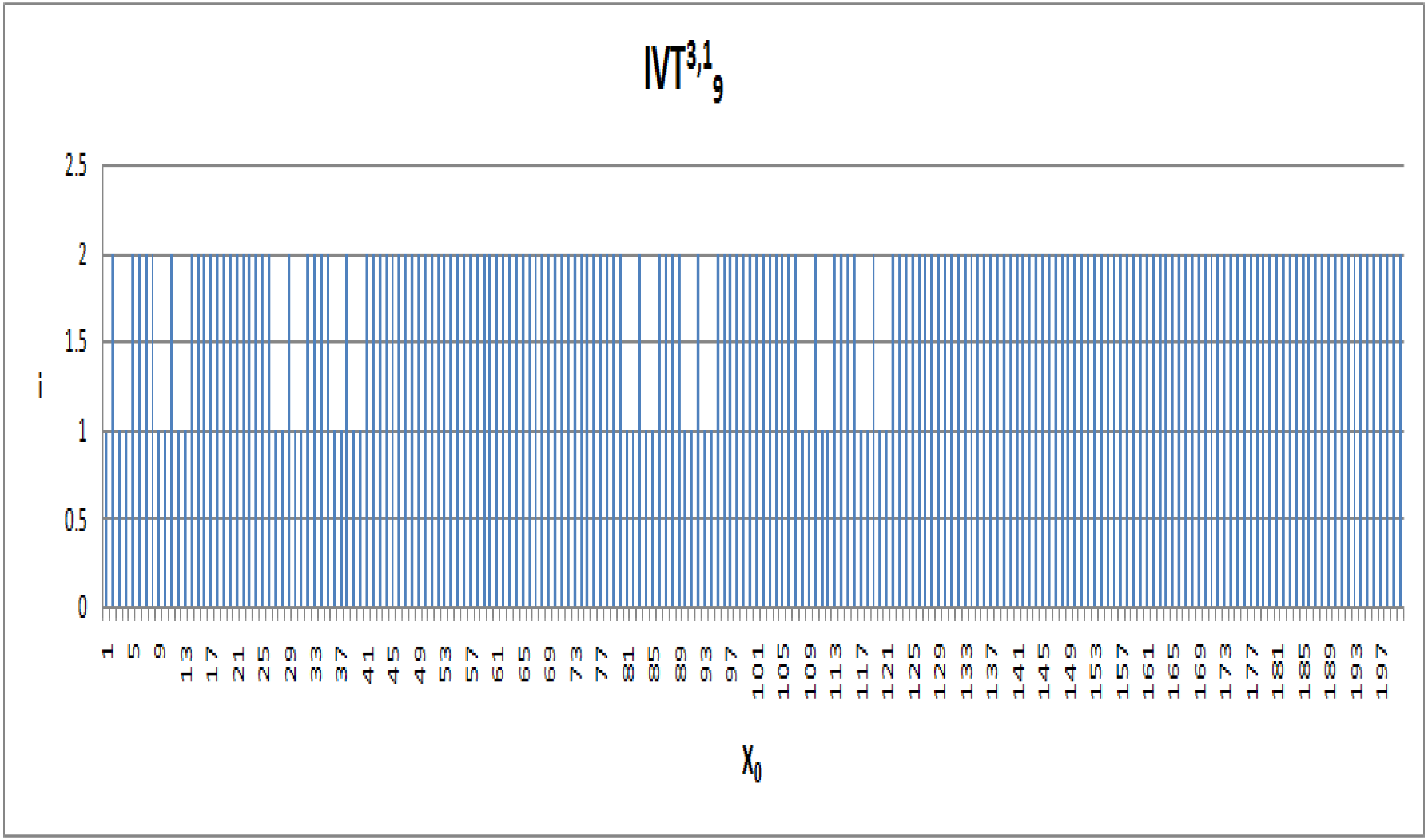
Convergence Graph of $IVT^{3,1}{}_{7}$



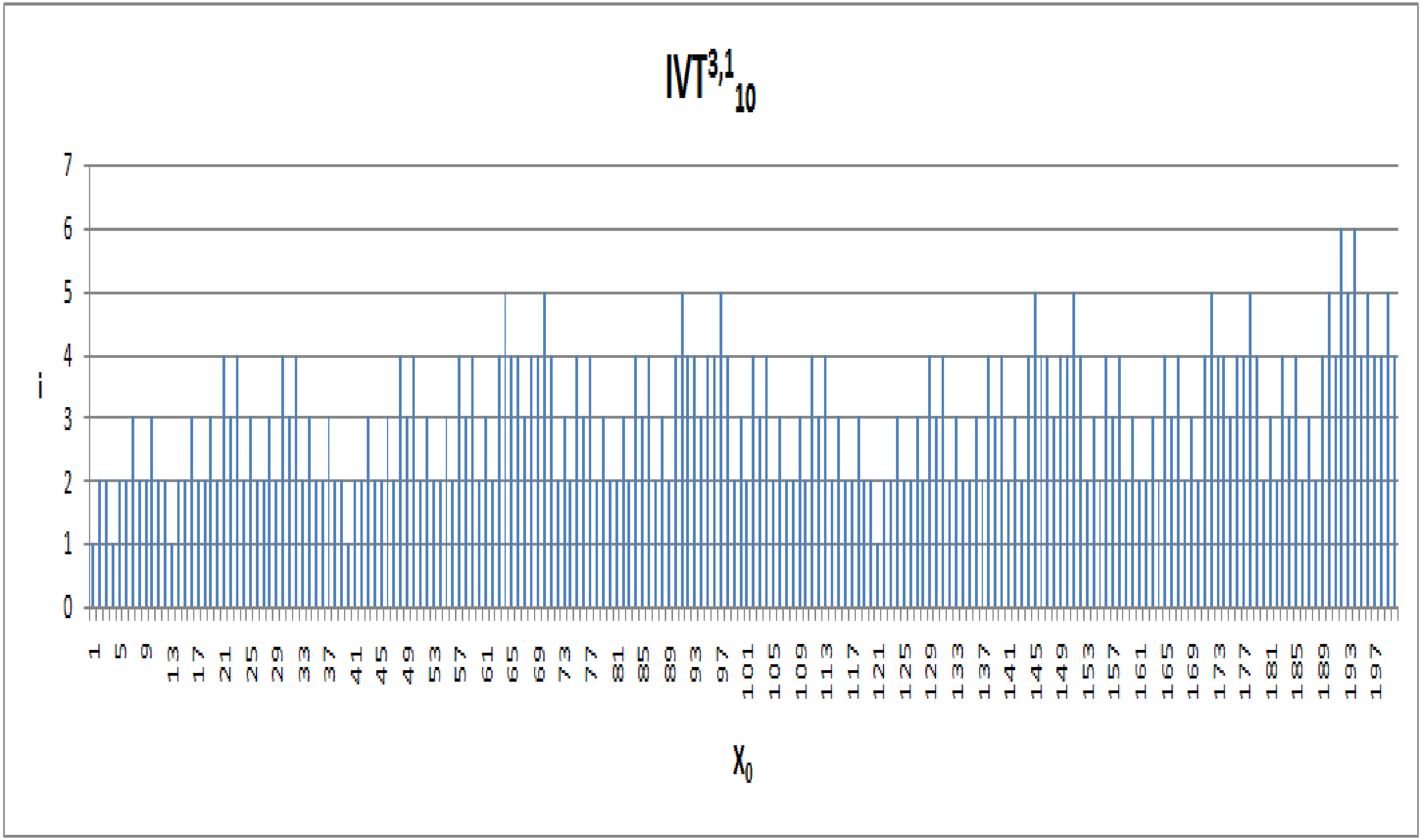
Convergence Graph of $IVT^{3,1}_{8}$



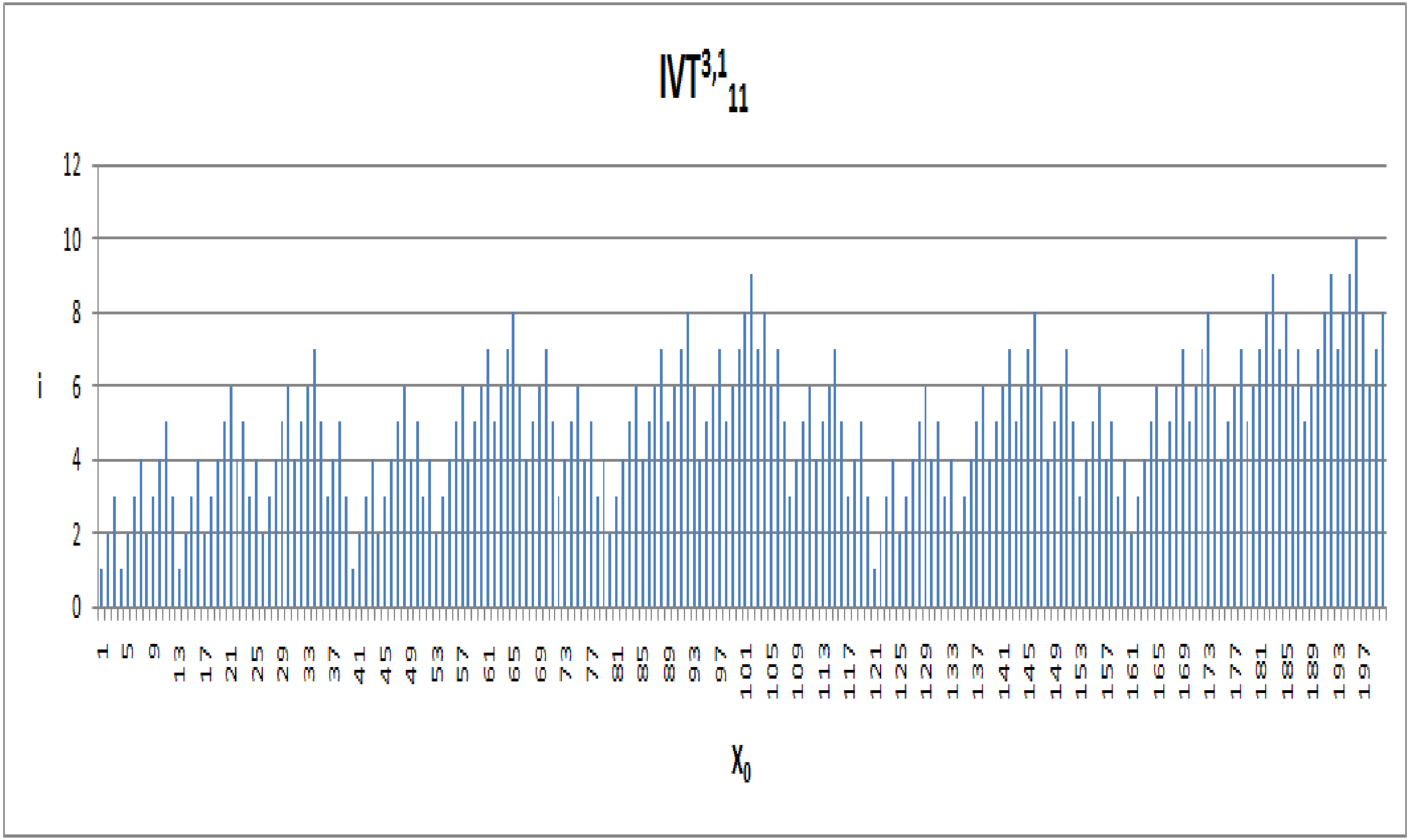
Convergence Graph of $IVT^{3,1}{}_{9}$



Convergence Graph of $IVT^{3,1}_{10}$



Convergence Graph of $IVT^{3,1}{}_{11}$



Morphology can enlighten the convergence behavior of each IVT

What we have observed :
In case of Collatz function, the convergence graph shows very **slow rate of convergence** for different X0’s whereas in **Our IVT domain**, there are Collatz like transformations for which the convergence graph shows very **rapid convergence**.

In our cases, not more than 10 iterations required to converge to 0 for set {0,1 ,2 …200} whereas the number 27 takes 111 iteration for converging to 1 in the Collatz function.

- This kind of convergence behavior we would like to translate in terms of Mathematical morphology.

An honest question:
Does the study enlighten something?

» This is the first report where an analogous Collatz conjecture is defined and proved. Who knows, This may be a clue to prove Collatz Conjecture too.

- Classification of Cellular Automata or IVTs.
- Global bijection between set of multistate CA rules and set of IVTs play the crucial role.

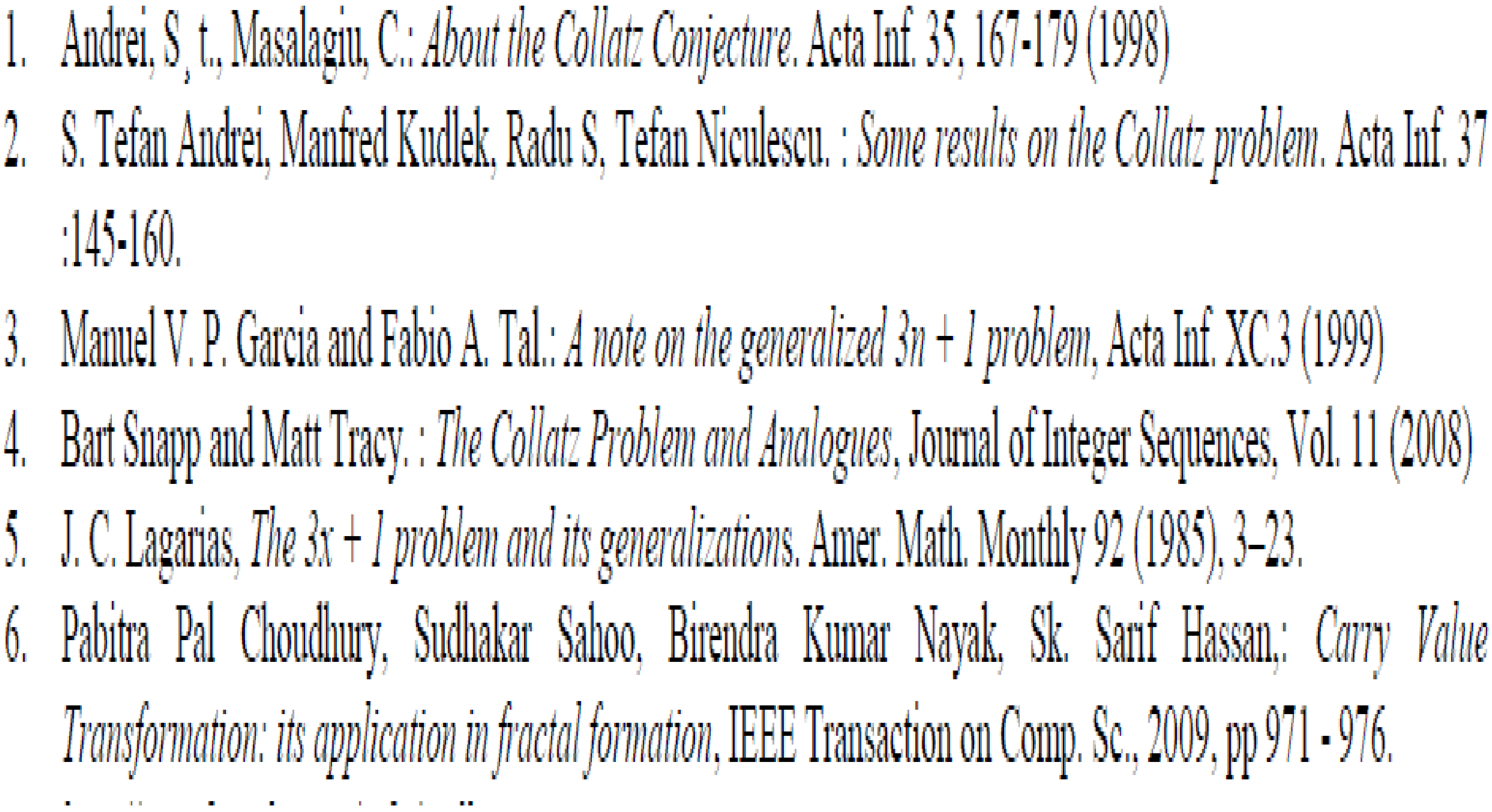
References

- 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.**
- 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.**
- Malnic B, Hirono J, Sato T, Buck L-B (**1999**) **Combinatorial receptor codes for odors. Cell 96: 713-723.**
- SK. S. Hassan, Pabitra Pal Choudhury, Arunava Goswami, R. L. Brahmachary **Designing exons for human olfactory receptor gene subfamilies using a mathematical paradigm** Journal of Bioscience, Springer, (September 2010).
- Glusman G, Yanai I, Rubin I, Lancet D (**2001**) **The complete human olfactory subgenome. Genome Research 11: 685-702.**
- Prusinkiewicz P, Lindenmayer A (**1990**) in **The algorithmic beauty of plants, Springer-Verlag ISBN 978-0387972978.**

References…

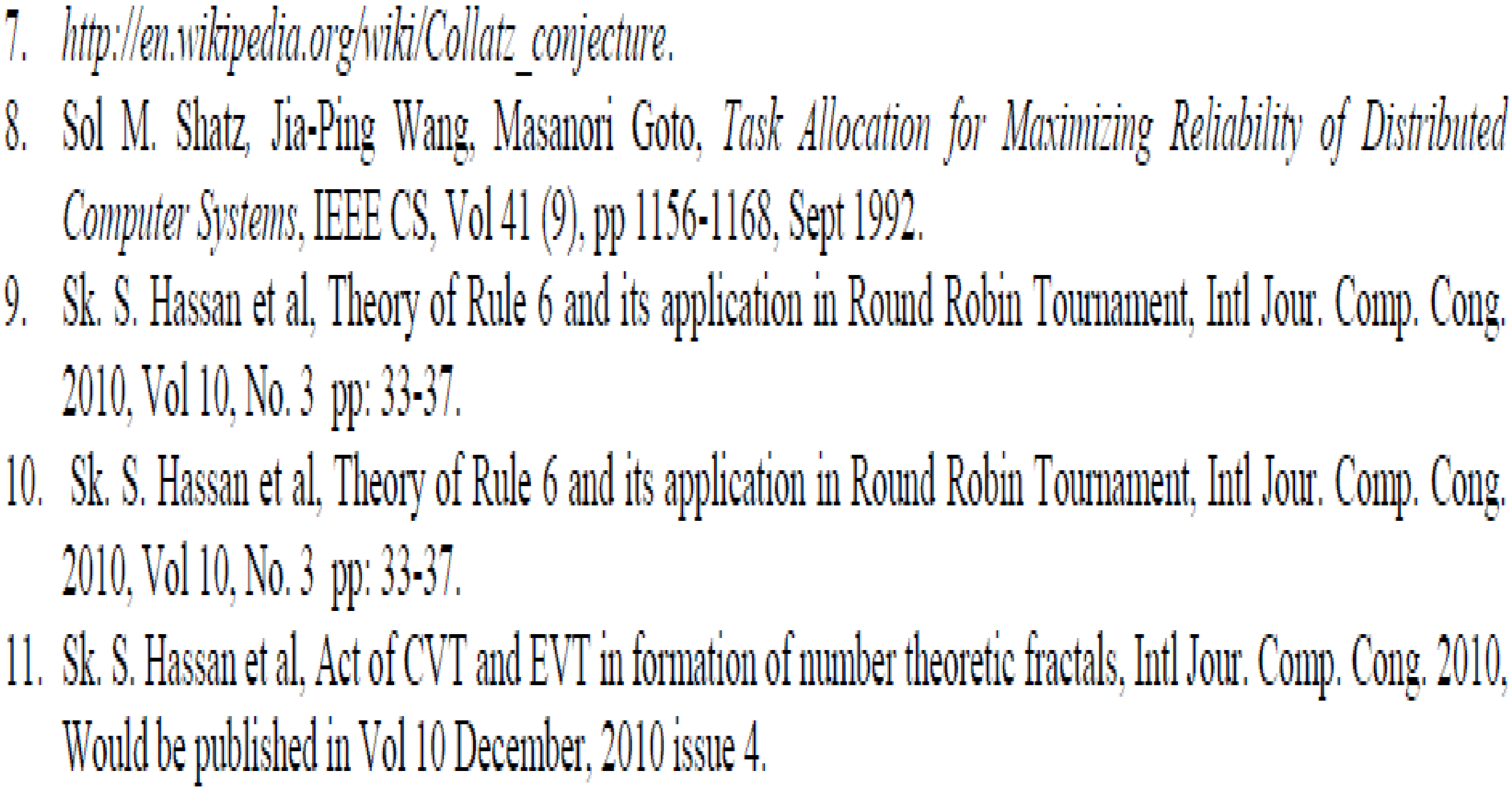
- Yu Zu-Guo et al **2002** **Fractals in DNA sequence analysis** ,Chinese Phys. 11 1313-1318
- Claire Ainsworth et al , **September, 2009**, **Cells go fractal**: Nature
- B. Mandelbrot., The fractal Geometry of Nature, **1975**.
- Bransley, Fractals Everywhere. Springer Verlag.
- **P. P. Choudhury, S. Sahoo, B. K Nayak, and Sk. S. Hassan**, Carry Value Transformation: It's Application in Fractal Formation 2009 IEEE International Advanced Computing Conference (IACC 2009), Patiala, India, 6-7 March, pp 2613-2618, 2009

References



1. Andrei, S¸ t., Masalagiu, C.: *About the Collatz Conjecture*. Acta Inf. 35, 167-179 (1998)
2. S. Tefan Andrei, Manfred Kudlek, Radu S, Tefan Niculescu. : *Some results on the Collatz problem*. Acta Inf. 37 :145-160.
3. Manuel V. P. Garcia and Fabio A. Tal.: *A note on the generalized 3n + 1 problem*, Acta Inf. XC.3 (1999)
4. Bart Snapp and Matt Tracy. : *The Collatz Problem and Analogues*, Journal of Integer Sequences, Vol. 11 (2008)
5. J. C. Lagarias, *The 3x + 1 problem and its generalizations*. Amer. Math. Monthly 92 (1985), 3–23.
6. Pabitra Pal Choudhury, Sudhakar Sahoo, Birendra Kumar Nayak, Sk. Sarif Hassan,: *Carry Value Transformation: its application in fractal formation*, IEEE Transaction on Comp. Sc., 2009, pp 971 - 976.

References…



7. *http://en.wikipedia.org/wiki/Collatz_conjecture.*
8. Sol M. Shatz, Jia-Ping Wang, Masanori Goto, *Task Allocation for Maximizing Reliability of Distributed Computer Systems*, IEEE CS, Vol 41 (9), pp 1156-1168, Sept 1992.
9. Sk. S. Hassan et al, Theory of Rule 6 and its application in Round Robin Tournament, Intl Jour. Comp. Cong. 2010, Vol 10, No. 3 pp: 33-37.

10. Sk. S. Hassan et al, Theory of Rule 6 and its application in Round Robin Tournament, Intl Jour. Comp. Cong. 2010, Vol 10, No. 3 pp: 33-37.

11. Sk. S. Hassan et al, Act of CVT and EVT in formation of number theoretic fractals, Intl Jour. Comp. Cong. 2010, Would be published in Vol 10 December, 2010 issue 4.

ISI Kolkata

Thank you all with a special thanks to **Prof. Jean Serra**