DNA Sequence Evolution through Integral Value Transformations

Sk. Sarif Hassan[1,2], Pabitra Pal Choudhury[1], Ranita Guha[1], Shantanav Chakraborty[1], Arunava Goswami[2]

[1]Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

[2]Biological Sciences Division, Indian Statistical Institute, Kolkata, India

Emails (According to order of Authors above): sarimif@isical.ac.in, pabitra@isical.ac.in, ranita1990@gmail.com, shantanav89@gmail.com, agoswami@isical.ac.in

Correspondence to be made to sarimif@isical.ac.in

Abstract

In deciphering the DNA structures, evolutions and functions, Cellular Automata (CA) do have a significant role. A DNA can be thought of as a one-dimensional multi-state CA, more precisely four states of CA namely A, T, C, and G which can be taken as numerals 0, 1, 2 and 3. Earlier, G.Ch. Sirakoulis et al reported the DNA structure, evolution and function through quaternary logic one dimensional CA and the authors have found the simulation results of the DNA evolutions with the help of only four linear CA rules. The DNA sequences which are produced through the CA evolutions, however, are seen by our research team not to exist in the established databases of various genomes although the initial seed (initial global state of CA) was taken from the database. This problem motivated us to study the DNA evolutions from more fundamental point of view. Parallel to CA paradigm we have devised an enriched set of discrete transformations which have been named as Integral Value Transformations (IVT). Interestingly, on applying the IVT systematically, we have been able to show that each of the DNA sequence at various discrete time instances in IVT evolutions can be directly mapped to a specific DNA sequence existing in the database. This has been possible through our efforts of getting quantitative mathematical parameters of the DNA sequences involving Fractals. Thus we have at our disposal some transformational mechanism between one DNA to another.

Key word: Integral Value Transformations (IVT), Olfactory Receptors (ORs), Fractals, Mathematical Morphology, Cellular Automata (CA).

1. ***Introduction:*** Here, we consider a DNA sequence as a one dimensional, one neighborhood, four states CA where each nucleotide A, T, C and G is replaced by 0, 1, 2 and 3 respectively. Earlier, in deciphering DNA structure, evolution and function, quaternary logic one dimensional CA was used by G.Ch. Sirakoulis et al [1]. They used only linear CA rules for time state evolution. It is worth noting that there are only four linear CA rules namely f_0, f_{27}, f_{34} and f_{57} which were applied for generating sequences over the time state evolutions. Out of them only f_{27} and f_{57} are bijective. Also they used uniform CA rules over the entire DNA sequence for CA evolutions.

Consequently, they did not find any significant results in existing gene database. So we follow a different path on using Integral value transformations (IVT) [2, 3, 4] instead of CA. The sequence thus obtained from an initial sequence is broken down into blocks of a fixed length and bijective rules of $IVT^{4,1}{}_{\#}$ are applied in a systematic manner iteratively as described in section 3. The sequence at each stage is classified according to the parameters as proposed earlier [5] and is also blasted in the NCBI database [6]. It is worth noting that most of the $IVT^{4,1}{}_{\#}$ rules applied in obtaining the sequences are non-linear as well as bijective. Thereby, the authors are confirmed that by systematic application of IVT as proposed, DNA sequences that exist in the database can be generated unlike the one proposed in [1].

2. Some Basics on Integral Value Transformations (IVT)

2.1 Definition of IVT:

Integral Value Transformations (IVT) $IVT^{p,k}{}_{\#}$ from $\mathbb{N}_0{}^k$ to $\mathbb{N}_0$ is defined where p denotes the p-adic number, k denotes dimension of the domain and # represents the transformation index [2,3,4]. It is worth noting that these IVTs' correspond to each of the multistate Cellular Automata.

Let us define the IVT in $\mathbb{N}_0$ in 4-adic number systems. There are 256 (4^{4^1}) one variable four state CA rules. Corresponding to each of those CA rules there are 256 IVTs are there in 4 adic system in one dimension.

$IVT^{4,1}{}_{\#}$ is mapping a non-negative integer to a non-negative integer.

$$IVT^{4,1}{}_{\#}(a) = \left((f_{\#}(a_n)f_{\#}(a_{n-1}) \dots f_{\#}(a_1)\right)_4 = b$$

Where 'a' is a non-negative integer and $a = (a_n a_{n-1} \dots a_1)_4$ and 'b' is the decimal value corresponding to the 4-adic number.

For an example, let us consider $a = 225 = (3201)_4$ and $\# = 120$ so $f_{\#}(0) = 0;\ f_{\#}(1) = 2;\ f_{\#}(2) = 3\ and\ f_{\#}(3) = 1$

Therefore, $IVT^{4,1}{}_{120}(225) = (f_{120}(3)(f_{120}(2)(f_{120}(0)(f_{120}(1))_4 = (1302)_4 = 114.$

Consequently, $IVT^{4,1}{}_{120}(225) = 114.$

Let us denote $\mathfrak{T}^{4,1}{}_{\#}$ as set of all $IVT^{p,k}{}_{\#}$ transformations. It is worth nothing that there are 4! = 24 number of Bijective functions are there in $\mathfrak{T}^{4,1}{}_{\#}$. So of the 256 ($4^{4^1}$) transformations in $\mathfrak{T}^{4,1}{}_{\#}$ four are linear and rest are nonlinear.

Let us warm up little basics on Cellular Automata in the following section.

2.2 Some Basics on Cellular Automata

The concept of Cellular Automata is introduced by Neumann [7] and Ulam [8] as a possible idealization of biological systems, with the particular purpose of modeling biological self-reproduction. Later in 1983, Wolfram et al. [9] studied one-dimensional CA with the help of polynomial algebra precisely Boolean function algebra. CA has been extensively used in deciphering the mathematical formalization/idealization of physical systems in which space and time are discrete.

A one-dimensional CA consists of a regular uniform lattice, which may be infinite in size and expands in a one-dimensional space. Each site of this lattice is called *cell*. At each cell a variable takes values from a discrete set. The value of this variable is the state of the cell. An initial global 4-states of one dimensional CA is shown below:

Figure-I: Initial global 4-states of CA

The CA is known to be a *Discrete Dynamical System* reliant of discrete time and states. The CA evolves in discrete time steps and its evolution is manifested by the change of its cell states with time. Each cell would change itself according to the CA rules. There are 252 non-linear one dimensional four state CA rules as we have seen in the Integral Value Transformations. IVTs are typically special type of CA rules if we restrain the domain $\mathbb{N}_0$ as $\{0, 1, 2, 3\}$.

3. Methods and Results:

3.1 Methods:

Without loss of generality, we have considered DNA sequences of Olfactory Receptors (ORs) namely OR1D2, OR1D3P, OR1D4 and OR1D5. It is worth notifying that OR1D3P, OR1D4 and OR1D5 are most similar sequences to OR1D2. All four fit in to a specific subfamily 'D' under the family '1' as per HORDE classification [10]. We then transform the DNA sequence in terms of numerals by a simple mapping *f* as defined below:

$$f: \{A, T, C, G\} \to \{0,1,2,3\}\ as\ f(A) = 0; F(C) = 1, f(T) = 2\ and\ f(G) = 3$$

Therefore, a DNA sequence is now simply a string of four variables namely 0, 1, 2 and 3 as per coding scheme *f*. Now we apply Integral Value Transformations ($IVT^{4,1}{}_{\#}$) systematically:-

Firstly, we segmented the whole one dimensional initial sheet of DNA of length of ***n*** and divided it into ***r*** multiple blocks. We designate the DNA string as $S(t_0)$.

Secondly, we apply bijective transformations (need not to be all distinct) taken from $\mathfrak{T}^{4,1}{}_{\#}$ over each of the ***r*** different blocks of $S(t_0)$. Thereby, we call this case as ***Hybrid Application of IVTs.*** In other words, we are getting $S(t_1)$ from $S(t_0)$ through hybrid application of IVTs.

Lastly, we apply a unique bijective transformation taken from $\mathfrak{T}^{4,1}{}_{\#}$ over each of the ***r*** blocks of $S(t_1)$ to obtain $S(t_2)$. We call this case ***Uniform Application of IVT.***

Next, we follow the second and last steps successively.

The results, on applying the proposed systematic technique of application of IVTs on OR1D2, OR1D3P, OR1D4 and OR1D5, are enumerated in the following section.

3.2 Results:

The IVTs which are used to generate different $S(t_i)$s are listed below in table-I and table-II. The IVTs have been chosen here at random, without loss of any generality. Similarly other IVTs could be used to generate other different sequences and classified, mapped accordingly.

Seq.	***Hybrid IVTs***	***Hybrid IVTs***	***Hybrid IVTs***	***Hybrid IVTs***	***Hybrid IVTs***
$\mathbf{S(t_0)}$	$\mathbf{S(t_1)}$	$\mathbf{S(t_3)}$	$\mathbf{S(t_5)}$	$\mathbf{S(t_7)}$	$\mathbf{S(t_9)}$
OR1D2	114,210,27,156,120,75,78,198,99,180	225,177,27,57,30,147,141,180,54,135	147,30,120,156,201,27,198,216,45,75	225,30,156,114,78,201,39,216,45,177	114,156,225,210,75,30,45,201,141,27
OR1D4	30,78,120,114,210,216,198,135,45,147	201,210,216,27,78,75,108,114,45,156	39,78,216,114,120,45,156,180,210,225	210,225,39,120,45,156,78,114,30,108	78,210,75,27,114,120,180,156,216,39
OR1D5	141,177,180,27,120,156,75,54,225,108	114,99,141,177,78,27,45,201,54,75	225,27,201,99,78,177,147,54,39,45	225,27,30,99,180,108,39,54,210,177	114,135,147,78,99,225,156,108,54,177
OR1D3P	198,45,108,57,99,180,225,114,120,39.	45,114,78,225,120,27,180,177,216,210	177,27,99,78,201,198,180,54,120,45	45,156,54,225,177,99,201,78,108,57	27,39,225,180,78,198,75,99,141,201

Table-I: Hybrid IVTs are used to generate $\mathbf{S(t_i)}$ ***for different odd i's.***

Seq.	***Uniform IVTs***	***Uniform IVTs***	***Uniform IVTs***	***Uniform IVTs***	***Uniform IVTs***
$\mathbf{S(t_1)}$	$\mathbf{S(t_2)}$	$\mathbf{S(t_4)}$	$\mathbf{S(t_6)}$	$\mathbf{S(t_8)}$	$\mathbf{S(t_{10})}$
OR1D2	141	177	120	216	135
OR1D4	216	156	75	141	120
OR1D5	177	198	30	99	198
OR1D3P	225	180	30	99	99

Table-II: Uniform IVTs are used to generate $\mathbf{S(t_i)}$ ***for different even i's.***

In table-III, we are classifying all the $S(t_i)$s corresponding to a particular OR based on two of our proposed parameters and then mapping them to a specific OR based on Hurst exponent as calculated in [5].

Sequences	*Class according to Mean of Poly-String*	*Class according to SD of Poly-String*	*Hurst Exponent*	*Maps to Human OR according to Hurst Exponent*	*Search Result in NCBI Database*
$S(t_0)$= OR1D2	*CGTA*	*CGAT*	*0.598911*	***OR1D2***	Homo sapiens OR, (OR1D2).
$S(t_1)$	TACG	TACG	0.603008	OR2AP1	Do not map in NCBI
$S(t_2)$	GCTA	GCTA	0.633465	OR11M1P	Pongo abelii OR 1D2-like, mRNA.
$S(t_3)$	AGTC	ATCG	0.640663	OR2B3	Do not map in NCBI
$S(t_4)$	CTGA	CGAT	0.617221	OR6R1P	Gorilla gorilla isolate PPOR1D2 OR gene.
$S(t_5)$	TCAG	TCGA	0.580502	OR6K5P	Do not map in NCBI
$S(t_6)$	CGAT	CGTA	0.627906	OR1D4	Homo sapiens FOSMID clone ABC24-1938I17 from chromosome 17, complete Sequence. Pan troglodytes olfactory receptor PTOR1D2 (OR1D2), mRNA.
$S(t_7)$	TGCA	TGCA	0.649803	OR5P3	Do not map in NCBI
$S(t_8)$	TCGA	TCGA	0.650111	OR4C3	Mus musculus OR (Olfr385), mRNA.
$S(t_9)$	TGCA	TGCA	0.577422	------	Do not map in NCBI
$S(t_{10})$	GACT	GACT	0.609444	------	Do not map in NCBI
$S(t_0)$ = OR1D4	CGTA	CGTA	0.626152	**OR1D4**	Homo sapiens OR, (OR1D4).
$S(t_1)$	ATCG	CATG	0.596918	----	Do not map in NCBI
$S(t_2)$	ATGC	GATC	0.597076	OR13C3	Homo sapiens OR, (OR1D5), mRNA Homo sapiens chromosome 17, clone CTD-2309O5
$S(t_3)$	GTAC	ATGC	0.601067	----	Do not map in NCBI
$S(t_4)$	CGAT	AGCT	0.638198	OR5BD1P	Homo sapiens chromosome 17 genomic contig, GRCh37 reference primary Assembly, Pan paniscus isolate PPOR1D5 olfactory receptor gene
$S(t_5)$	TGAC	TAGC	0.614901	----	Do not map in NCBI

$S(t_6)$	CATG	CTAG	0.659798	OR2A13P	Homo sapiens OR (OR1D5), mRNA Cercocebus agilis clone OLG_4 olfactory receptor-like protein gene.
$S(t_7)$	TGAC	TAGC	0.603699	OR8R1P	Do not map in NCBI
$S(t_8)$	GACT	GCAT	0.630871	OR55B1P	Homo sapiens OR (OR1D5), mRNA. Homo sapiens chromosome 17, clone CTD-2309O5.
$S(t_9)$	TACG	TACG	0.594854	OR5BN1P	Do not map in NCBI
$S(t_{10})$	CAGT	CAGT	0.610381	OR9I1	Homo sapiens OR (OR1D5), mRNA.
$S(t_0)$ = OR1D5	CGTA	CGTA	0.61791	**OR1D5**	Homo sapiens OR, (OR1D5),
$S(t_1)$	GTCA	GTAC	0642219	OR11H7P	Do not map in NCBI
$S(t_2)$	TGAC	TGCA	0605042	OR9A1P	Homo sapiens OR1D2, mRNA. Mus musculus OR (Olfr412), mRNA.
$S(t_3)$	CAGT	ACGT	0575205	----	Do not map in NCBI
$S(t_4)$	CGAT	GCAT	0.628091	OR13J1	Homo sapiens OR1D2. Mus musculus OR 412 (Olfr412), mRNA.
$S(t_5)$	TCAG	CTAG	0.606603	OR5AX1	Do not map in NCBI
$S(t_6)$	ATGC	TAGC	0.700373	OR2AL1P	Homo sapiens OR1D2. Mus musculus OR (Olfr412), mRNA Homo sapiens OR1D4.
$S(t_7)$	TAGC	TACG	0.614765	OR4A11P	Do not map in NCBI
$S(t_8)$	CTGA	CTAG	0.598083	OR7E1P	Homo sapiens OR1D2 Mus musculus OR (Olfr412), mR NA.
$S(t_9)$	CGAT	GATC	0.642701	----	Do not map in NCBI
$S(t_{10})$	CAGT	AGTC	0.676853	OR2AS1P	Homo sapiens OR1D4, mRNA.
$S(t_0)$ = OR1D3P	TCAG	CGAT	0.610597	**OR1D3P**	Homo sapiens OR, (OR1D3P).
$S(t_1)$	TGAC	TGAC	0.617120	OR1N2	Do not map in NCBI
$S(t_2)$	TGCA	TGCA	0.662865	OR2L3	Homo sapiens chromosome 17 genomic contig, GRCh37 reference primary assembly.
$S(t_3)$	CTGA	CGTA	0.592167	OR52H2P	Do not map in NCBI

$S(t_4)$	CGTA	CTGA	0.587171	OR1J2	Homo sapiens chromosome 17 genomic contig, GRCh37 reference primary reference.
$S(t_5)$	GATC	TGCA	0.629847	OR4F15	Do not map in NCBI
$S(t_6)$	CGAT	ACTG	0.628202	OR2R1P	Homo sapiens (OR1D5), mRNA Mus musculus OR 412 (Olfr412), mRNA Homo sapiens olfactory receptor, family 1, subfamily D, member 2 pseudogene (LOC100422051) on chromosome 17
$S(t_7)$	TCGA	TCAG	0.613480	OR7E99P	Do not map in NCBI
$S(t_8)$	CAGT	CATG	0.613240	OR9I1	Homo sapiens (OR1D4), mRNA Mus musculus OR (Olfr412), mRNA Homo sapiens OR1D2
$S(t_9)$	CTAG	CATG	0.612120	OR9I1	Do not map in NCBI
$S(t_{10})$	ACTG	ATCG	0.650378	OR4V1P	Homo sapiens OR (OR1D4), mRNA. Mus musculus OR (Olfr412), mRNA Homo sapiens OR1D2.

Also we blasted all $S(t_i)$ in the NCBI database and got the result as tabulated above. Two interesting cases can be observed:

Case I: A sequence $S(t_i)$ maps to a particular known human OR according to Hurst Exponent [5]. However, the same sequence when blasted in NCBI database, maps to a different OR of the same or different species. This is not surprising as the sequence may have similar texture to the mapped OR according to Hurst Exponent but may be structurally and functionally similar to the ORs obtained from the NCBI database.

Case II: A sequence $S(t_j)$ maps to a particular known human OR according to Hurst Exponent. However, the same sequence when blasted in NCBI database does not map to anything at all. This is because, no known OR(s) are similar to $S(t_j)$ at present in the database. It is our strong convictions that, in future it may so happen that the genomic landscape may generate such a gene over biological evolution which, we predict, will be structurally similar to the mapped human OR.

4 .Conclusion and Future endeavors:

In summary, we have devised a methodology by which, starting from a known DNA sequence and by applying non-linear bijective IVT rules on them, we obtain sequences that are similar to that of the existing species. Also, by our proposed methodology we generate sequences that may not be in the genome at present but in future if such sequences are produced by Nature, then their structural similarity with respect to a known sequence has been

predicted. In near future, we are about to explore *Integral Value Transformations* as an alternative to *Cellular Automata*.

Acknowledgement: The authors are grateful to **Prof. Jean Serra,** Emeritus Professor, ESIEE-Engineering, University Paris-Est Former Director of Centre for Mathematical Morphology, France, and **Prof. B. S. Dayasagar**, Indian Statistical Institute, Bangalore for their kind advice and suggestions and also express their earnest gratitude to **Prof. R.L. Brahmachary**, Indian Statistical Institute, Kolkata for his thoughtful research discussions.

References:

[1] ***G.Ch. Sirakoulis*** et al. A cellular automaton model for the study of DNA sequence evolution, *Computers in Biology and Medicine* 33 (2003) 439–453

[2] ***Sk. S. Hassan et al.*** Collatz Function like Integral Value Transformations, Alexandria Journal of Mathematics, Vol 1, No. 2, Nov. 2010. pp-31-35

[3] ***P. Pal Choudhury et al***, Theory of Carry Value Transformation (CVT) and its Application in Fractal formation, Global Journal of Computer Science and Technology, Vol.10 Issue 14 (Ver.1.0) November 2010, pp 89-99.

[4] ***P. Pal Choudhury et al,*** Act of CVT and EVT in the formation of number theoretic fractals, International Journal of Computational Cognition, USA, Vol. 9, No. 1, March 2011.

[5] ***Sk. S. Hassan et al.*** Underlying Mathematics in Diversification of Human Olfactory Receptors in Different Loci. Available from Nature Precedings <http://hdl.handle.net/10101/npre.2010.5475.1> (2010).

[6] http://www.ncbi.nlm.nih.gov/.

[7] ***J. Von Neumann***, Theory of Self-reproducing Automata, University of Illinois Press, Urbana, 1966.

[8] ***S. Ulam***, Some ideas and prospects in biomathematics, Ann. Rev. Bio. 12 (1974) 255.

[9] ***S. Wolfram***, Statistical mechanics of Cellular Automata, *Rev Mod Phys.* 55,601-644 (July 1983).

[10] http://genome.weizmann.ac.il/horde/