Prediction of Genome Sequences in Terms of Cellular Automata Expansion of Rule Based Logics

Rama Naga Kiran Kumar. K, Ramesh Babu. I

Abstract: This paper proposes a novel concept called “Percentage Nucleotide Concentration of genomes” in terms of cellular automata evolutions of adjoints of Adenine, Thymine, Guanine, and Cytosine. The adjoints of the given a genome sequence are the characteristic binary string sequences. For example, the adjoint of Adenine of a given genome sequence is a binary string consisting of 0’s and 1’s where 1’s corresponds to the presence of Adenine in the genome sequence. So, one can have four adjoint sequences of Adenine, Thymine, Guanine, and Cytosine corresponding to a given genome sequence. One-dimensional three neighborhood binary value cellular automata rules could be applied to an adjoint sequence and the desired number of evolutions obtained. These rules are defined by linear Boolean functions and one can have 256 such linear Boolean functions. The analysis of genome sequences with predictive analytics gives a scope of getting the inherent properties of the genome. The predictive model suits the Nucleotide concentration and is computed for an adjoint sequence and its variation evaluated for its successive evolutions based on a cellular automaton rule.

Keywords: Cellular Automata, Evolutions of Adjoints, Linear Boolean functions, Nucleotide concentration in a genome.

I. INTRODUCTION

The purpose of the research carried out and reported in this paper is that whether it is possible to categorize a set of genomes like the human genome repository. The usage of predictive analytics such as recommender systems provides the possible sequence of genome and leads to the evolution of similar models in the processing of genomes. The concept of “% nucleotide concentration” introduced in this paper seems to show a way to accomplish this task. The genesis of the formulation of this concept originates from chemistry, wherein the quantification notion of percentage ionic concentration of hydrogen (pH value) is used to categorize solutions into three (i) water, whose pH value is 7, (ii) acidic solutions whose pH values are less than 7 and (iii) alkaline solutions whose pH values are greater than 7. On the same lines, an effort was made to categorize genome sets based on four values (i) % nucleotide concentration of Adenine (pA), (ii) % nucleotide concentration of Thymine (pT), (iii) % nucleotide concentration of Guanine (pG) and (iv) % nucleotide concentration of Cytosine (pC). It is reasonable to surmise that these values, possibly their compositions would categorize a given set of genomes. The formulation of the concept is briefly explained below. Section 2 of this paper describes the concept formulation. Section 3 of this paper describes the fundamental notions of adjoints of a genome and their evolution using one-dimensional cellular automata rules defined by linear Boolean functions. Section 4 provides experimental results of a case study pertaining to evaluation of Concentration of Nucleotides in terms of Adjoints of Brucella Suis 1330 Genome Sequence.

II. CONCEPT FORMULATION

Analogous to the notion of pH value of a solution, the values of pA, pT, pG and pC of a genome sequence and possibly composition of these values like the proportion pA:pT:pG:pC seems to pave a way to classify and characterize genome sets. The definition of “Percentage Nucleotide Concentration” of a genome sequence is given below.

Definition
Given a genome sequence, the number of a particular nucleotide, say A, present in that genome sequence is counted and the sum is divided by the total number of nucleotides in that genome sequence. The fraction when multiplied by 100 yields the “Percentage Concentration of Adenine pA”. Similarly, one can evaluate pT, pG and pC.

III. ONE-DIMENSIONAL THREE NEIGHBORHOOD CELLULAR AUTOMATA EVOLUTIONS OF ADJOINTS OF A GENOME SEQUENCE

Adjoint of a particular nucleotide in a genome sequence is the binary sequence obtained by substituting the particular nucleotides in the genome sequence by 1’s and the others by 0’s. For example, let us consider a sample sequence of Brucella Suis 1330 for a case study. The actual length of the genome sequence of Brucella Suis 1330 is 5806. A cellular automaton is an idealized parallel processing system consisting of an array of numbers (1-D, 2-D and more) realized using updating rules based on certain neighborhood. For example, a one dimensional cellular automaton would consist of a finite length array as shown below.

Consider an i-th cell in the array. This cell has a neighbor i-1 on its left and another i+1 on its right. All three put together is called a three neighborhood. One can assign a site (cell) variable xi, ρi, and xi+1 to the three neighborhood cells. At a particular instant of time, these variables take on numerical values, say either 0 or 1. In such a case, the variables are denoted as ρi, ρi, and ρi+1. The value of the i-th cell at the next instant of time is evaluated using an updating rule that involves the present values of the i-th, (i-1)th and (i+1)th cells. This updating rule is basically a linear Boolean function of three variables.
One can construct 256 linear Boolean functions as updating rules of one dimensional three neighborhood binary valued cellular automata. Each rule defines an automaton by itself. So, one dimensional binary valued three neighborhood cellular automata (123CA) rules could be used to model adjoints of a genome sequence. The first twenty linear Boolean functions of cellular automata 123CA are listed below with their decimal equivalents.

### Linear Boolean Function

| Linear Boolean Function | Decimal Equivalent |
|------------------------|--------------------|
| 0                      | 0                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 1                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 2                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 3                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 4                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 5                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 6                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 7                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 8                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 9                  |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 10                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 11                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 12                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 13                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 14                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 15                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 16                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 17                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 18                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 19                 |
| $\tilde{\xi}_i\tilde{\xi}_{i+1}$ | 20                 |

For the case study rule number 90 is applied to the adjoints of Brucella Suis 1330 genome sequence and 500 evolutions generated. Rule 90 is shown below.

$\tilde{\xi}_i\tilde{\xi}_{i+1}$ + $\tilde{\xi}_i\tilde{\xi}_{i+1}$  

Since the image of the 500 evolutions of Brucella Suis 1330 is large, a small portion of the images are presented in this paper.

### IV. CONCENTRATION OF NUCLEOTIDES IN ADJOINTS OF BRUCELLA SUIS 1330 GENOME SEQUENCE

The values of pA, pT, pG and pC of the Brucella Suis 1330 genome sequence are computed for the adjoints A(n), T(n), G(n) and C(n) and their 500 evolutions using 123CA rules based one linear Boolean functions. Fig. 1 shows the evolutions of the adjoints of A(n), T(n), G(n) and C(n) using the linear Boolean function rule 90 of 123CA. The values are tabulated and the corresponding graphs shown subsequently. Table 1 shows the pA values of A(n) of Brucella Suis 1330 genome sequence and the 500 generations of A(n) using rule 90 of 123CA. Figs. 2 and 3 shows the graphs of the variations of pA values of all generations. Table 2 shows the pT values of T(n) of Brucella Suis 1330 genome sequence and the 500 generations of T(n) using rule 90 of 123CA. Figs. 4 and 5 shows the graph of the variations of pT values of all generations. Table 3 shows the pG values of G(n) of Brucella Suis 1330 genome sequence and the 500 generations of G(n) using rule 90 of 123CA. Fig. 4 shows the graph of the variations of pG values of all generations. Table 4 shows the pC values of C(n) of Brucella Suis 1330 genome sequence and 500 generations of C(n) using rule 90 of 123CA. Fig. 5 shows the graph of the variations of pC values of all generations.
Rule number 90 is applied to $A(n)$ and its 500 generations. It is observed that the $pA$ value becomes minimum at regular intervals of 1, 2, 4, 8, 16, 32, 64, 128 and 256. This indicates a fractal behavior of the evolution.

$\min(A(n)) = 30.45126$ and $\max(A(n)) = 33.06924$. The deviation is 2.61.

Fig. 3: Minimum $pA$ values of $A(n)$ and its evolutions

---

Rule number 90 is applied to $T(n)$ and its 500 generations. It is observed that the $pT$ value becomes minimum at regular intervals of 1, 2, 4, 8, 16, 32, 64, 128 and 256. This indicates a fractal behavior of the evolution.

$\min(T(n)) = 30.45126$ and $\max(T(n)) = 33.06924$. The deviation is 2.61.

Fig. 4: $pT$ values of $T(n)$ and of its evolutions

---

Rule number 90 is applied to $G(n)$ and its 500 generations. It is observed that the $pG$ value becomes minimum at regular intervals of 1, 2, 4, 8, 16, 32, 64, 128 and 256. This indicates a fractal behavior of the evolution.

$\min(G(n)) = 43.00723$ and $\max(G(n)) = 44.29900$. The deviation is 1.46.

Fig. 5: Minimum $pG$ values of $G(n)$ and of its evolutions

---

Rule number 90 is applied to $C(n)$ and its 500 generations. It is observed that the $pC$ value becomes minimum at regular intervals of 1, 2, 4, 8, 16, 32, 64, 128 and 256. This indicates a fractal behavior of the evolution.

$\min(C(n)) = 30.45126$ and $\max(C(n)) = 33.06924$. The deviation is 2.61.

Fig. 6: $pC$ values of $C(n)$ and of its evolutions

---

Table 2: $pT$ values of $T(n)$ and its 500 evolutions

| $e$  | $pT$    |
|------|---------|
| 1    | 30.45126|
| 2    | 32.15639|
| 4    | 31.94971|
| 8    | 32.3803 |
| 16   | 32.65587|
| 32   | 32.19084|
| 64   | 31.82914|
| 128  | 31.82914|
| 256  | 33.06924|

Table 3: $pG$ values of $G(n)$ and its 500 evolutions

| $e$  | $pG$    |
|------|---------|
| 1    | 43.00723|
| 2    | 43.97175|
| 4    | 43.86841|
| 8    | 43.21347|
| 16   | 43.79952|
| 32   | 42.8350 |
| 64   | 43.74785|
| 128  | 43.57561|
| 256  | 44.29900|

Table 4: $pC$ values of $C(n)$ and its 500 evolutions

| $e$  | $pC$    |
|------|---------|
| 1    | 30.45126|
| 2    | 32.15639|
| 4    | 31.94971|
| 8    | 32.3803 |
| 16   | 32.65587|
| 32   | 32.19084|
| 64   | 31.82914|
| 128  | 31.82914|
| 256  | 33.06924|
V. CONCLUSIONS

This paper proposes a novel concept called “Percentage Nucleotide Concentration of genomes” in terms of cellular automata evolutions of adjoints of Adenine, Thymine, Guanine, and Cytosine. The research carried out and reported in this paper exhibits the possibility to categorize a set of genomes like the human genome repository the predictive method allows us to guess the pattern of sequences which helps the formulation of genome sequences. In short, the concept of “Percentage Nucleotide Concentration (PNC)” introduced in this paper seems to show a way to accomplish this task.

REFERENCES

1. Carey, Nessa (2015). Junk DNA: A Journey Through the Dark Matter of the Genome. Columbia University Press. ISBN 9780231170840.
2. Chris M. Rands, Stephen Meader, Chris P. Ponting and Gerton Lunter (2014). “8.2% of the Human Genome Is Constrained: Variation in Rates of Turnover across Functional Element Classes in the Human Lineage”. PLoS Genet 10 (7): e1004525. doi:10.1371/journal.pgen.1004525. PMC 4109858. PMID 25057982.
3. Costa, Fabrício (2012). "7 Non-coding RNAs, Epigenomics, and Complexity in Human Cells". In Morris, Kevin V. Non-coding RNAs and Epigenetic Regulation of Gene Expression: Drivers of Natural Selection. Caister Academic Press, ISBN 1904455948.
4. Dan Graur, Yichen Zheng, Nicholas Price, Ricardo B. R. Azevedo1, Rebecca A. Zufall and Eran Elhaik (2013). "On the immortality of television sets: "function" in the human genome according to the evolution-free viewpoint of ENCODE" (PDF). Genome Biology and Evolution 5 (3): 578–90. doi:10.1093/gbe/evt028. PMC 3622293. PMID 23431001.
5. Doolittle, W. Ford (2013). "Is junk DNA bunk? A critique of ENCODE", Proc Natl Acad Sci USA110 (14): 5294–5300. Bibcode:2013PNAS..110.5294D. doi:10.1073/pnas.1212176110. PMC 3619371. PMID 23479647.
6. Elgar G, Vavouri T, Vavouri (July 2008). "Tuning in to the signals: noncoding sequence conservation in vertebrate genomes", Trends Genet. 24 (7): 344–52. doi:10.1016/j.tig.2008.04.005. PMID 18514361.
7. Ehret CF, De Haller G, De Haller (1963). "Origin, development, and maturation of organelles and organelle systems of the cell surface in Paramecium", Journal of Ultrastructure Research. 9 Supplement 1: 1, 3–42. doi:10.1016/S0022-166X(63)80088-X. PMID 14073743.
8. Gregory TR, Hebert PD; Hebert (April 1999). "The modulation of DNA content: proximate causes and ultimate consequences". Genome Res. 9 (4): 317–24. doi:10.1101/gr.9.4.317 (inactive 2015-02-01). PMID 10207154.
9. Hawkins, JS.; Kim, H.; Nason, JD.; Wing, RA.; Wendel, JF. (Oct 2006). "Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium", Genome Res 16 (10): 1252–61. doi:10.1101/gr.5282906. PMID 16954538.
10. International Human Genome Sequencing Consortium (February 2001). "Initial sequencing and analysis of the human genome", Nature 409 (6822): 879–888. Bibcode:2001Natur.409..860L. doi:10.1038/35057062. PMID 11237011.

AUTHORS PROFILE

K Rama Naga Kiran Kumar completed his B.E(CSE), M.Tech(CSE) and pursuing Ph.D. (CSE) from Acharya Nagarjuna University. His area of interest is in AI, Data Mining, Machine Learning, Software Engineering.

Dr. Ramesh Babu I, done his B.E (ECE), M.E (Computer Engg), Ph.D. (CSE) in ANU. His area of interest is in Image Processing, Computer Graphics, AI, Data Mining, Machine Learning, Software Engineering.