

Linear Prediction of Nucleotides in a Genome Sequence



Venkateswarlu Pedakolmi, E. G. Rajan

Abstract: Nucleotides are organic molecules, which are monomer units that form polymers of nucleic acid ‘deoxyribonucleic acid (DNA)’ and ‘ribonucleic acid (RNA)’. The four nucleotides A, T, G and C get connected by phosphodiester bonds to form strands. Strand formation depends on innumerable factors related to inter and intra cellular parameters and functions. One cannot precisely say that a particular strand gets formed using such and such rules. The infinite possibilities of strand formation cannot be determined experimentally or in the framework of classical genetics. One can alternatively formulate a notion of the “Language of Genomes” wherein one can finitely specify infinite strands. This paper introduces a novel prediction algorithm, which generates possible strands based on available nucleotides statistics.

Keywords: *Linear Prediction, Genome Sequences.*

I. INTRODUCTION

Nucleotides have either a purine or a pyrimidine base. The nitrogenous base atoms are known as ribonucleotides if the sugar is ribose, or as deoxyribonucleotides if the sugar is deoxyribose. It is the phosphate atom that associates sugar-ring particles of two nucleotide monomers. Along these lines, a linked long chain of nucleotide monomers structures nucleic corrosive. The associated chain of sugar and phosphate particles frames the 'spine' for a solitary strand or twofold strand nucleic corrosive. The four nucleotides A, T, G and C get associated by phosphodiester bonds to shape strands. Strand development relies upon countless elements identified with bury and intra cell parameters and capacities. One cannot absolutely say that a specific strand gets framed utilizing such and such runs the show. The unending potential outcomes of strand arrangement cannot be resolved tentatively or in the system of traditional hereditary qualities. In a twofold strand helix of a nucleic corrosive, one strand has the bearing of 5 prime-end to 3 prime-end and the integral strand of 3 prime-end to 5 prime-end. Adenine and Thymine structure a base pair and they are associated by a feeble twofold hydrogen bond. So also, Guanine and Cytosine structure a base pair and they are associated by a frail triple hydrogen bond.

A. Linear Prediction of Nucleotides

Singleton nucleotides are additionally called 1-codons. As referenced before, the octet straight forecast calculation is connected to the *Brucella Suis* 1330 genome of length 5806.

Revised Manuscript Received on October 30, 2019.

* Correspondence Author

Venkateswarlu Pedakolmi*, Research Scholar, Dept. of Computer Science, MG-NIRSA, Affiliated to University of Mysore, Manasagangotri, Mysore, Karnataka, India. E-Mail: venkat123.pedakolmi@gmail.com.

E. G. Rajan, Director, MG-NIRSA; Director, PRC Global Technologies Inc., Ontario, Canada. E-mail: rajaneg@yahoo.co.in.

© The Authors. Published by Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP). This is an [open access](https://creativecommons.org/licenses/by-nc-nd/4.0/) article under the CC-BY-NC-ND license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

For comfort, the long succession of length 1330 is isolated into 12 subsequences (i) 1 to 500, (ii) 501 to 1000, (iii) 501

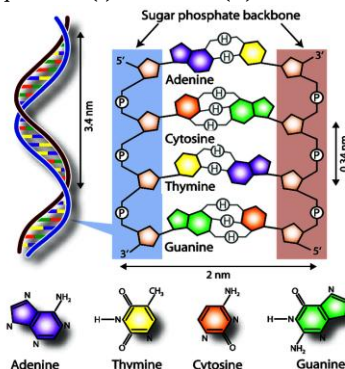


Fig. 1: Chemical bonds between purines and pyrimidines

to 1500, (iv) 1501 to 2000, (v) 2001 to 2500, (vi) 2501 to 3000, (vii) 3001 to 3500, (viii) 3501 to 4000, (ix) 4001 to 4500, (x) 4501 to 5000, (xi) 5001 to 5500 and (xii) 5501 to 5805. It is to be noted here that the 5806th nucleotide is not considered here in light of the fact that that last one is to be anticipated. The consequences of applying the calculation to the primary section, that is, the succession from the main nucleotide to 500th nucleotide are displayed underneath. It ought to be noted here that the initial 8 nucleotides structure the primary information sub arrangement to the calculation with the goal that the ninth nucleotide is anticipated. Moreover, the sub grouping from 5798 to 5805 is given as the last information arrangement with the goal that the 5806th nucleotide is anticipated. This adds up to stating that the anticipated yield begins from the ninth nucleotide to 5806th nucleotide. Presently, the arrangement of *Brucella* Suis 1330 genome of length 5806 is demonstrated as follows.

[illegible]

Fig. 2: Sample Brucella Suis 1330 genome sequence (Part)

II. NUMERICAL REPRESENTATION OF NUCLEOTIDES

Numerical representation of A, T, G and C is given by the map shown below, followed by the numerical string corresponding to the *Brucella Suis* 1330 genome of length 5806.



A → 1
T → 2
G → 3
C → 4

3,1,1,2,3,1,2,2,1,4,4,1,1,3,3,4,4,1,1,3,4,2,4,4,2,2,4,2,2,3,3,4,2,3,1,3,4,2,2,2,3,4,4,2,2,2,3,4,1
1,2,4,3,2,4,1,1,3,2,2,2,2,1,3,4,4,1,2,2,3,2,1,4,3,1,1,3,1,4,4,1,3,4,4,1,3,4,4,1,3,2,1,2,2
4,2,3,2,4,2,3,4,2,3,1,3,4,4,2,2,1,3,2,4,4,1,3,1,2,4,3,4,1,3,4,4,4,2,3,1,2,3,4,2,4,3,4,2,4,2,2,3,4,2
4,4,3,1,3,4,2,4,2,4,2,3,4,2,4,1,3,4,4,2,4,4,2,3,1,4,4,2,4,3,4,4,4,2,2,3,4,2,4,1,2,4,3,4,4,2,4,2,2,2,4,2
3,4,4,4,2,4,4,4,1,3,3,3,4,1,4,4,2,4,2,3,2,3,1,4,3,4,3,4,4,4,1,3,3,1,4,4,2,3,3,2,2,2,3,2,3,1,4,1
3,4,2,4,3,4,4,1,3,4,3,2,3,4,3,4,4,1,3,2,3,3,1,3,2,2,2,3,1,3,1,3,3,2,3,1,1,3,3,2,2,4,4,2,3,4,4,1,3,3,2,4,1
2,4,3,1,3,2,4,4,2,3,3,1,4,3,2,3,3,3,4,4,4,1,1,2,3,4,4,1,4,4,3,4,3,2,3,3,4,4,2,3,3,2,4,1,1,4,2,1,4,3,4,4,1,3,4,3
4,4,3,2,3,1,1,3,4,1,1,3,3,1,3,2,2,4,4,4,3,4,2,3,4,3,3,4,4,1,1,3,3,4,4,3,4,4,2,3,4,4,1,1,3,3,4,4,3,4,3
2,3,4,3,4,4,3,4,1,2,4,1,3,4,4,1,4,2,3,2,4,4,1,4,3,3,3,3,1,4,4,1,2,3,1,4,3,3,4,4,2,3,3,4,4,1,2,4,4,1,3,2,2,2,3,4,4,1
2,4,1,4,1,3,3,3,4,4,2,2,4,1,3,2,3,1,1,3,4,4,3,1,3,3,3,4,3,2,4,3,4,4,1,3,3,2,4,4,4,4,3,1,4,1,2,4,1,3,4,1,1,3,3
2,3,4,3,2,3,4,4,4,3,4,4,2,3,4,2,3,3,2,2,4,4,3,4,2,3,2,2,3,4,4,4,1,4,2,2,3,2,3,4,2,1,1,3,1,1,4,2,2,2
3,4,4,3,3,4,1,4,3,4,2,2,2,2,3,2,2,2,4,4,4,3,4,1,4,1,4,4,4,4,4,3,1,2,3,3,4,4,3,2,2,2,2,1,2,4,2,2,3,3,3,1,4,4
1,3,1,4,4,4,1,1,4,2,1,1,3,1,3,1,1,4,3,1,4,3,3,4,2,3,1,4,3,4,2,3,3,1,2,4,3,1,1,4,4,4,4,2,2,2,1,4,4,4,1,4,4,2,4,4
3,1,4,4,4,1,1,3,4,4,4,2,4,1,4,1,1,2,3,3,3,3,3,1,4,3,1,2,3,2,2,4,1,3,3,3,3,2,2,3,1,4,4,2,3,3,4,4,4,4,3
4,4,4,3,2,4,2,3,1,3,4,2,4,4,2,4,3,1,1,4,3,4,1,1,4,3,4,3,1,3,3,3,3,2,3,4,1,3,1,2,3,4,2,2,1,2,3,2,3,1,2,3,2,4,1,3,3
3,3,4,3,2,3,3,1,4,1,4,2,3,4,3,4,4,3,4,4,2,3,2,4,4,1,3,4,4,4,4,2,1,1,3,4,4,4,4,4,1,4,4,1,4,4,1,3,4,2,3,2
3,4,4,2,4,4,2,3,1,3,1,3,3,1,3,4,2,4,1,3,4,4,4,2,4,1,3,2,2,4,4,2,4,1,4,4,2,1,4,1,1,1,2,3,3,3,1,3
4,1,1,4,1,1,3,4,3,4,2,2,4,2,4,1,3,1,3,3,4,4,3,4,1,3,3,4,1,3,3,1,4,2,1,1,1,4,3,1,3,2,2,4,1,2,4,2,3,4,2,3,1,1,3,3,4
3,4,2,4,1,3,4,1,4,1,3,4,3,4,4,2,4,3,3,1,4,4,4,1,1,4,1,3,3,4,4,4,1,2,3,3,1,3,3,4,3,2,2,1,3,4,2,3,1,3,2,2,2,2,2
1,3,2,1,4,3,4,4,2,2,2,3,1,3,3,3,3,3,1,3,3,3,4,2,4,1,3,1,1,4,4,4,1,1,3,4,1,1,2,3,4,4,4,4,1,1,3,2,4,1,4,1,2,3,3
4,2,3,3,4,2,4,4,1,3,3,3,3,3,1,4,2,3,2,2,3,4,1,4,1,3,1,1,4,2,3,4,1,2,4,4,1,2,3,3,3,4,1,1,4,3,2,2,2
4,1,3,3,1,4,2,4,1,3,4,4,3,3,3,4,4,2,3,4,3,1,3,3,4,4,3,2,4,1,3,2,2,4,2,4,2,4,4,4,3,1,3,3,4,2,4,4,4,1,4,4
3,3,4,4,2,4,3,4,4,2,4,4,2,4,4,2,4,3,4,1,3,4,4,4,2,4,3,2,3,3,4,4,2,3,1,3,1,3,2,4,3,4,2,1,2,4,1,2,4,2,3,4,4,4,1,3
4,4,4,4,4,1,3,4,4,4,1,3,3,3,3,1,3,3,2,1,2,2,4,2,2,4,2,2,1,2,4,3,4,4,4,4,4,4,1,4,2,4,2,3,3,2,1,3,3
4,1,3,1,4,3,2,3,4,4,2,1,1,4,3,2,3,1,4,2,3,4,4,1,4,1,4,3,3,1,2,1,1,1,4,4,4,2,3,4,4,1,3,1,3,3,3,2,3,2,1,2,4,4,4,1,2
2,2,2,1,4,1,3,1,2,3,1,3,1,1,4,1,3,3,4,1,4,1,3,3,3,1,3,4,2,2,1,1,3,4,1,1,3,2,2,3,4,4,4,1,3,4,2,1,1,2,1,1,3,2,3,2,1
3,1,1,3,2,2,2,4,4,4,1,4,1,3,1,2,3,4,2,3,3,4,3,2,1,2,1,1,2,4,4,4,1,3,3,4,1,2,2,1,4,1,3,4,4,2,4,4,1,1,3,1,1,4,3,1,4
4,2,3,2,2,3,3,3,2,1,3,4,2,3,4,2,3,2,2,2,2,2,2,3,2,3,3,3,1,4,4,3,2,2,4,1,4,1,3,3,2,2,4,2,3,1,4,3,2,3,1,3,3,3
2,3,1,2,3,2,2,3,2,3,3,3,4,3,1,3,1,3,1,3,3,1,3,1,4,4,2,4,4,1,3,4,1,3,3,2,3,1,4,1,4,2,3,1,3,3,3,2,4,2,3,3,3
3,4,4,2,3,3,4,2,1,3,3,1,4,1,2,4,1,3,3,4,3,2,2,2,1,4,4,2,3,3,1,1,2,2,4,1,3,4,2,3,3,3,3,4,2,4,2,1,4,4,1,3,3,2
3,3,1,2,3,1,1,1,2,3,4,4,3,1,2,2,3,1,1,3,3,3,3,3,1,4,1,3,4,1,3,4,1,4,1,2,3,3,3,1,4,2,3,2,3,3,4
1,4,1,2,4,2,3,3,4,1,3,2,3,1,2,3,2,4,1,4,1,4,1,3,4,4,1,3,3,2,2,2,2,1,3,2,3,4,4,1,3,3,1,3,4,4,4,1,3,3,4,2,2,2,3,1

Fig. 3: Numerical Representation of Brucell Suis 1330 (Part)

III. OCTET LINEAR PREDICTION ALGORITHM

Let us consider the sequence $x(n)$ of length 52: $x(n) - 0, 52, 46, 50, 53, 35, 18, 25, 40, 44, 60, 83, 73, 66, 37, 50, 72, 38, 67, 55, 75, 87, 65, 68, 77, 72, 77, 76, 74, 77, 76, 82, 58, 0, 26, 20, 54, 56, 92, 59, 73, 42, 46, 114, 206, 162, 134, 175, 163, 77, 75, 70$. Assume that all these 52 values are measured values of a parameter associated with a linear dynamical system. These values are treated as velocities of the system. Acceleration is difference between subsequent velocity values.

Algorithm Formulation

Prediction of 9th velocity value $\tilde{V}_9$ just after the 8-point sequence $x_1(n)$ is carried out in two phases: (i) Estimation of 9th value $\tilde{V}_9$ using a linear regression formula $\tilde{V}_9 = a + 9b$ and (ii) Estimation of 8th acceleration value $\tilde{A}_8$ using a linear regression formula $\tilde{A}_8 = c + 8d$ and subtracting $\tilde{A}_8$ from $\tilde{V}_9$ in order to get the predicted velocity value $\hat{V}_9$.

Basic Formula:

9th predicted velocity =

9th estimated velocity – 8th estimated acceleration

$$\hat{V}_9 = \tilde{V}_9 - \tilde{A}_8$$

Phase #1

$$\tilde{V}_9 = a + 9b$$

$\tilde{V}_9$ is the 9th estimated value of velocity

$$a = \bar{V} - b\bar{x}$$

$$b = \frac{\sum_{i=1}^8 [(x_i - \bar{x})(V_i - \bar{V})]}{\sum_{i=1}^8 (x_i - \bar{x})^2}$$

$$\bar{V} = \sum_{j=1}^8 V_j = \text{mean of 8 samples of velocity}$$

$$\bar{x} = \sum_{i=1}^8 x_i = \text{mean of 8 index values (1 to 8) = 4.5 always}$$

Then,

$$a = \bar{V} - \left[\frac{\sum_{i=1}^8 (x_i - \bar{x})(V_i - \bar{V})}{\sum_{i=1}^8 (x_i - \bar{x})^2} \right] \bar{x}$$

In other words,

$$a = \bar{V} - 4.5 \left[\frac{\sum_{i=1}^8 (x_i - 4.5)(V_i - \bar{V})}{\sum_{i=1}^8 (x_i - 4.5)^2} \right]$$

$$\tilde{V}_9 = \bar{V} - 4.5 \left[\frac{\sum_{i=1}^8 (x_i - 4.5)(V_i - \bar{V})}{\sum_{i=1}^8 (x_i - 4.5)^2} \right] + 9 \left[\frac{\sum_{i=1}^8 (x_i - 4.5)(V_i - \bar{V})}{\sum_{i=1}^8 (x_i - 4.5)^2} \right]$$

$$\tilde{V}_9 = \bar{V} + 4.5 \left[\frac{\sum_{i=1}^8 (x_i - 4.5)(V_i - \bar{V})}{\sum_{i=1}^8 (x_i - 4.5)^2} \right] \text{ if } \tilde{V}_9 \geq 0;$$

$$= 0 = \tilde{V}_9 = \bar{V}_9 \text{ if } \tilde{V}_9 < 0 \text{ (no further computation)}$$

Phase #2

Now

$$A_1 = V_1 - V_2, A_2 = V_2 - V_3, A_3 = V_3 - V_4, A_4 = V_4 - V_5$$

$$A_5 = V_5 - V_6, A_6 = V_6 - V_7, A_7 = V_7 - V_8$$

$\tilde{A}_8 = 8^{\text{th}}$ estimated value of acceleration obtained using A_1 to A_7

This is evaluated using the equation

$$\tilde{A}_8 = c + 8d$$

$\tilde{A}_8$ is the 8th estimated value of acceleration

$$c = \bar{A} - d\bar{x}$$

$$d = \frac{\sum_{i=1}^7 [(x_i - \bar{x})(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - \bar{x})^2}$$

$$\bar{A} = \sum_{j=1}^7 A_j = \text{mean of 7 samples of acceleration}$$

$$\bar{x} = \sum_{i=1}^7 x_i = \text{mean of 7 index values (1 to 7) = 4 always}$$

then

$$c = \bar{A} - \left[\frac{\sum_{i=1}^7 [(x_i - 4)(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - 4)^2} \right] \bar{x}$$

In other words,

$$c = \bar{A} - 4 \left[\frac{\sum_{i=1}^7 [(x_i - 4)(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - 4)^2} \right]$$

$$\tilde{A}_8 = \bar{A} - 4 \left[\frac{\sum_{i=1}^7 [(x_i - 4)(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - 4)^2} \right] + 8 \left[\frac{\sum_{i=1}^7 [(x_i - 4)(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - 4)^2} \right]$$

$$\tilde{A}_8 = \bar{A} + 4 \left[\frac{\sum_{i=1}^7 [(x_i - 4)(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - 4)^2} \right]$$

$$\text{Let } \eta = \bar{V} + 4.5 \left[\frac{\sum_{i=1}^8 (x_i - 4.5)(V_i - \bar{V})}{\sum_{i=1}^8 (x_i - 4.5)^2} \right] - \bar{A} + 4 \left[\frac{\sum_{i=1}^7 [(x_i - 4)(A_i - \bar{A})]}{\sum_{i=1}^7 (x_i - 4)^2} \right]$$

$$\lambda = \tilde{V}_9 - \tilde{A}_8$$

Now, the final predicted 9th value $\hat{V}_9$ is given by:

$$\hat{V}_9 = \lambda; \text{ when } \lambda \geq 0 \\ = 0; \text{ when } \lambda < 0$$

Now, the final predicted 9th value $\hat{V}_9$ is given by:

$$\hat{V}_9 = \lambda; \text{ when } \lambda \geq 0 \\ = 0; \text{ when } \lambda < 0$$

In order to explain this concept, the first 16 velocities of $x(n)$ are considered here.

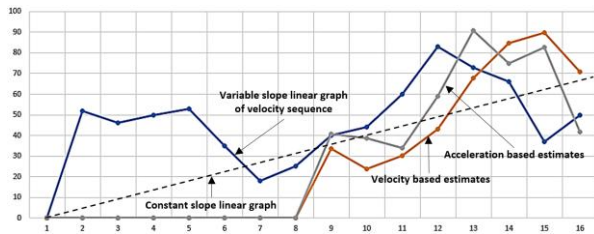


Fig. 4: Velocity, acceleration and estimation of

9th velocity based on velocity and acceleration

From Fig. 4, one can visualize that estimation of 9th velocity based on previous 8 velocities and the 8th acceleration is closer to the actual 9th value when compared to estimation of 9th value exclusively based on previous 8 velocities alone. Fig. 5 shows the graphical representation of the velocity sequence $x(n) = 0, 52, 46, 50, 53, 35, 18, 25, 40, 44, 60, 83, 73, 66, 37, 50, 72, 38, 67, 55, 75, 87, 65, 68, 77, 72, 77, 76, 74, 77, 76, 82, 58, 0, 26, 20, 54, 56, 92, 59, 73, 42, 46, 114, 206, 162, 134, 175, 163, 77, 75, 70$

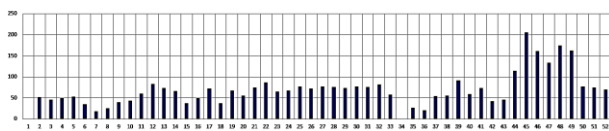


Fig. 5: Graphical representation of $x(n)$

The resulting sequence $\hat{\mathbf{x}}(n)$ of predicted values is given below.

$\hat{x}(\mathbf{n}) = 0, 0, 0, 0, 0, 0, 0, 0, 0, 40.53571429, 38.64285714,$
 $34.03571429, 59.14285714, 90.67857143,$
 $74.7582.75, 41.85714286, 71.92857143,$
 $82.67857143, 9.428571429, 76.03571429,$
 $40.10714286, 86.60714286, 96.39285714,$
 $56.57142857, 78.89285714, 92.64285714,$
 $71.2142857183, 72.92857143, 71.75, 82.46428571,$
 $76.89285714, 85.21428571, 46.25, 0, 43.32142857,$
 $0, 47.39285714, 28.78571429, 96.07142857,$
 $44.28571429, 109.3928571, 44.67857143, 66.75,$
 $147.5357143, 236.3928571, 127.6785714,$
 $158.4642857, 246.4642857, 199.0357143,$
 $66.89285714, 89.42857143.$

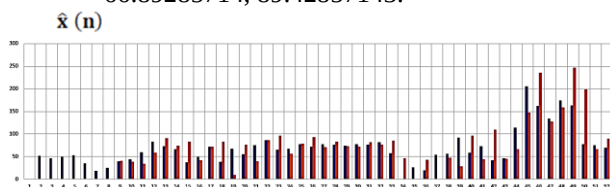


Fig. 6(a) shows $\hat{x}(n)$ along with $x(n)$ as bar chart.
Fig. 6(a): Bar chart representation of $x(n)$ and $\hat{x}(n)$

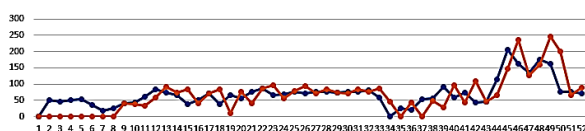


Fig. 6(b) shows $\hat{x}(n)$ along with $x(n)$ as discrete signal form

(b): Discrete signal forms of $x(n]$ and $\hat{x}(n]$

Fig. 6: Graphical representation of $x(n)$ and $\hat{x}(n)$

IV. LINEAR PREDICTION OF NUCLOTIDES OF BRUCELLA SUI 1330 GENOME SEQUENCE

The four nucleotides A, T, G and C get connected by phosphodiester bonds to form strands. Strand formation depends on innumerable factors related to inter and intra cellular parameters and functions. One cannot precisely say that a particular strand gets formed using such and such rules. The infinite possibilities of strand formation cannot be determined experimentally or in the framework of classical genetics. One can alternatively formulate a notion of the “Language of Genomes” wherein one can finitely specify infinite strands, Fig. 7 shows a finitely generated quaternary tree structure of strand formation of nucleic acids.

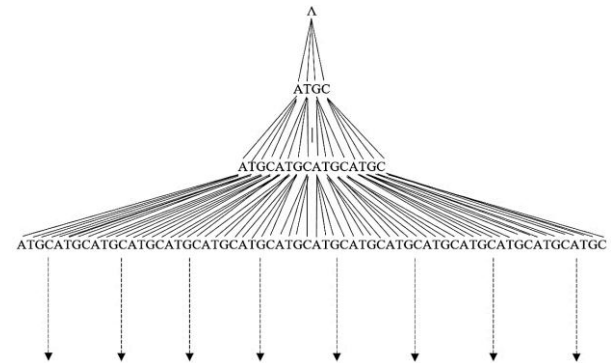


Fig. 7: Quaternary tree structure for strand formation

Fig. 7 shows three levels of nucleotides. One can generate 64 strands of length 3. As the length increases, the number of strands increases as per the formula 4^n , where n is the length of the strand. Strands of length 3 are called triplet codons or 3-tuple codons. Similarly, one can think of n -tuple codons where n is any number. As an example, *Brucella Suis* 1330 genome is considered here. The length of the genome is 5806. In this case, one can decide a set of 4^{5806} strands. The question that arises here is that whether it is possible to decide this set by means of an algorithm and in such a case is it possible to predict the entire sequence, say for example, *Brucella Suis* 1330 from the starting subsequence of length 8 or more. This section provides results of an empirical work carried out by applying Octet Algorithm on the *Brucella Suis* 1330 genome sequence.

Linear Prediction of Nucleotides

Singleton nucleotides are also called 1-codons. As mentioned earlier, the **octet linear prediction algorithm** is applied to the *Brucella Suis* 1330 genome of length 5806. For convenience, the long sequence of length 1330 is divided into 12 subsequences (i) 1 to 500, (ii) 501 to 1000, (iii) 501 to 1500, (iv) 1501 to 2000, (v) 2001 to 2500, (vi) 2501 to 3000, (vii) 3001 to 3500, (viii) 3501 to 4000, (ix) 4001 to 4500, (x) 4501 to 5000, (xi) 5001 to 5500 and (xii) 5501 to 5805. It is to be noted here that the 5806th nucleotide is not considered here because that last one is to be predicted. The results of applying the algorithm to the first segment, that is, the sequence from the first nucleotide to 500th nucleotide are presented below. It should be noted here that the first 8 nucleotides form the first input sub sequence to the algorithm so that the 9th nucleotide is predicted.

Linear Prediction of Nucleotides in a Genome Sequence

Likewise, the sub sequence from 5798 to 5805 is given as the last input sequence so that the 5806th nucleotide is predicted. This amounts to saying that the predicted output starts from the 9th nucleotide to 5806th nucleotide. The octet linear prediction algorithm is applied to this numerical string of *Brucella Suis* 1330 genome given in Fig. 3 and the predicted sequence is given in Fig. 8.

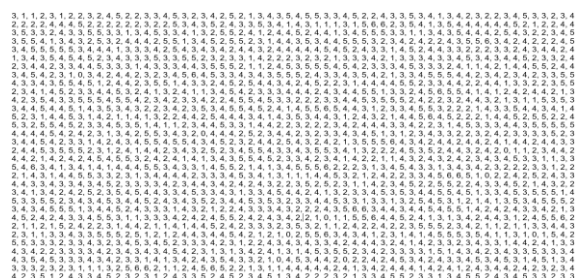


Fig. 8: Predicted sequence of the Brucella Suis 1330 genome

Now the error in predicting nucleotides is evaluated by subtracting actual values from the predicted values and the prediction error sequence is given below in Fig. 9.

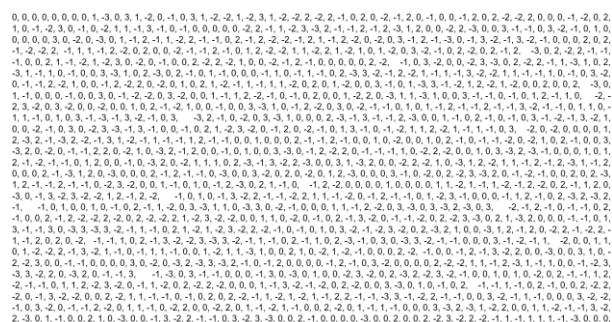


Fig. 9: Prediction Error Sequence

Fig. 10 shows actual numerical values from 1 to 500 corresponding to the genome sequence and the predicted values along with the prediction error.

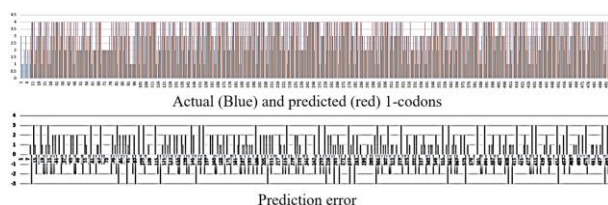


Fig. 10: Subsequence from 1 to 500 (1-codons)

Fig. 11 shows actual numerical values from 501 to 1000 corresponding to the genome sequence and the predicted values along with the prediction error.

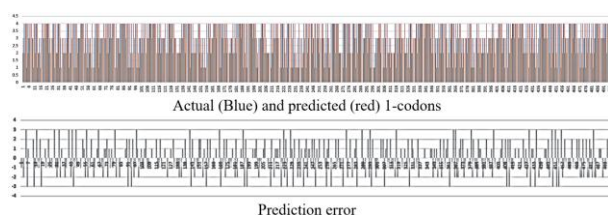


Fig. 11: Subsequence from 501 to 1000 (1-codons)

Fig. 12 shows actual numerical values from 1001 to 1500 corresponding to the genome sequence and the predicted values along with the prediction error.

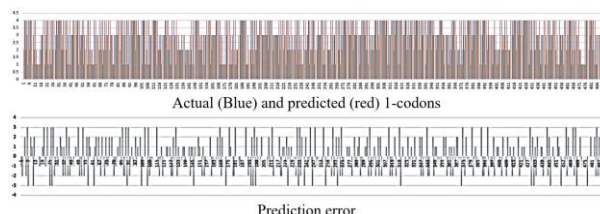


Fig. 12: Subsequence from 1001 to 1500 (1-codons)

Fig. 13 shows actual numerical values from 1501 to 2000 corresponding to the genome sequence and the predicted values along with the prediction error.

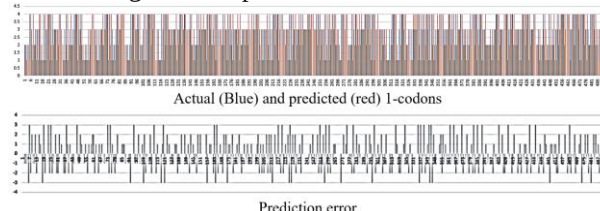


Fig. 13: Subsequence from 1501 to 2000 (1-codons)

Fig. 14 shows actual numerical values from 2001 to 2500 corresponding to the genome sequence and the predicted values along with the prediction error.

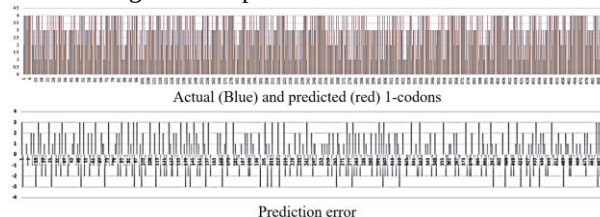


Fig. 14: Subsequence from 2001 to 2500 (1-codons)

Fig. 15 shows actual numerical values from 2501 to 3000 corresponding to the genome sequence and the predicted values along with the prediction error.

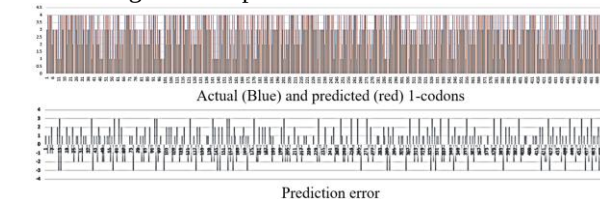


Fig. 15: Subsequence from 2501 to 3000 (1-codons)

Fig. 16 shows actual numerical values from 3001 to 3500 corresponding to the genome sequence and the predicted values along with the prediction error.

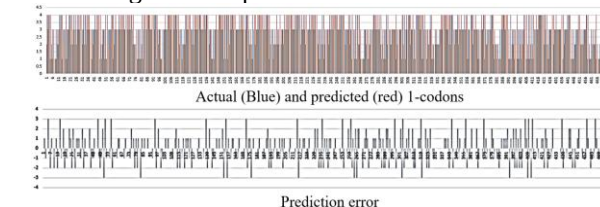


Fig. 16: Subsequence from 3001 to 3500 (1-codons)

Fig. 17 shows actual numerical values from 3501 to 4000 corresponding to the genome sequence and the predicted values along with the prediction error.

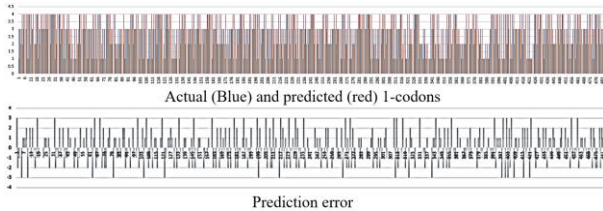


Fig. 17: Subsequence from 3501 to 4000 (1-codons)

Fig. 18 shows actual numerical values from 4001 to 4500 corresponding to the genome sequence and the predicted values along with the prediction error.

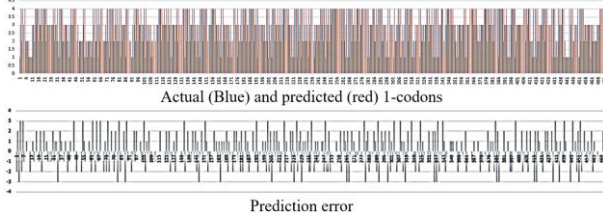


Fig. 18: Subsequence from 4001 to 4500 (1-codons)

Fig. 19 shows actual numerical values from 4501 to 5000 corresponding to the genome sequence and the predicted values along with the prediction error.

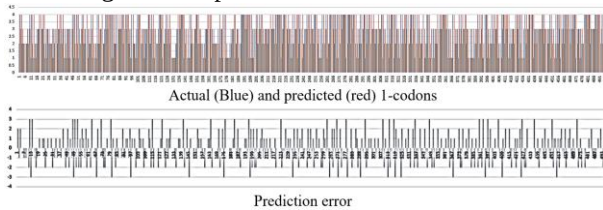


Fig. 19: Subsequence from 4501 to 5000 (1-codons)

Fig. 20 shows actual numerical values from 5501 to 5806 corresponding to the genome sequence and the predicted values along with the prediction error.

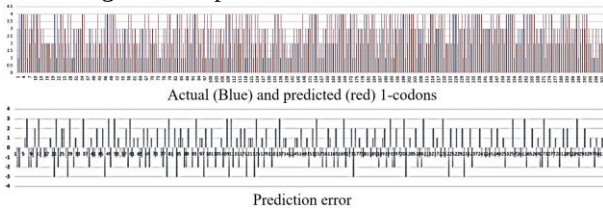


Fig. 20: Subsequence from 5501 to 5806 (1-codons)

The prediction accuracy is calculated using the formula

$$\text{Prediction Accuracy} = \frac{\text{Number of Nucleotides predicted correctly}}{\text{Total Number of nucleotides}} \times 100$$

$$\text{Prediction Accuracy} = \frac{1667}{5806} \times 100 = 28.71662365 \%$$

V. OBSERVATIONS AND CONCLUSIONS

From the above empirical study, it is observed that linear prediction of nucleotides of a genome sequence could be carried out using the novel linear octet prediction algorithm introduced in this paper. The novel technique and results presented in this paper are outcome of a prolonged research carried out in the mathematical modeling of genomes and their evolutions. As a future work, one can as well look into the possibilities of modeling genome evolutions using cellular automata tools.

REFERENCES

1. Sean Eddy (2012) The C-value paradox, junk DNA, and ENCODE, *Curr Biol* 22(21):R898–R899.
2. Doolittle, W. Ford (2013). "Is junk DNA bunk? A critique of ENCODE". *Proc Natl Acad Sci USA* **110** (14): 5294–5300. Bibcode:2013PNAS..110.5294D. doi:10.1073/pnas.1221376110. PMC 3619371. PMID 23479647.
3. Palazzo, Alexander F.; Gregory, T. Ryan (2014). "The Case for Junk DNA". *PLoS Genetics* **10** (5): e1004351. doi:10.1371/journal.pgen.1004351. ISSN 1553-7404.
4. Waterhouse, Peter M.; Hellens, Roger P. (25 March 2015). "Plant biology: Coding in non-coding RNAs". *Nature* **520** (7545): 41–42. doi:10.1038/nature14378.
5. Tutar, Y. (2012). "Pseudogenes". *Comp Funct Genomics* **2012**: 424526. doi:10.1155/2012/424526. PMC 3352212. PMID 22611337.
6. Petrov DA, Hartl DL; Hartl (2000). "Pseudogene evolution and natural selection for a compact genome". *J. Hered.* **91** (3): 221–7. doi:10.1093/jhered/91.3.221. PMID 10833048.
7. Ohno, Susumu (1972). H. H. Smith, ed. So Much "junk" DNA in Our Genome. Gordon and Breach, New York. pp. 366–370. Retrieved 2013-05-15.

AUTHORS PROFILE



P Venkateswarlu completed his M.Tech Computer Science & Engineering and pursuing Ph.D from University of Mysore. His area of interest is in AI, Data Mining, Machine Learning, Software Engineering.



Dr. E G Rajan, done his Ph.D from IIT-Kanpur. His area of interest is in Signal and Image Processing, Data Mining.