



International Journal of Emerging Technologies in Computational and Applied Sciences (IJETCAS)

www.iasir.net

A Coding Based Framework for Error Control in DNA Sequences

Bhawani Sankar Biswal¹, Anjali Mohapatra²

¹PG Student, Deptt. of CSE, IIIT Bhubaneswar, Odisha, INDIA

²Asst. Professor, Deptt. of CSE, IIIT Bhubaneswar, Odisha, INDIA

Abstract: In the present day scenario, information transmission is a big issue in a biological system. Information regarding an organism or a species or proper cell functioning in molecular biology are the valuable aspects in a biological domain. So the transmission of those information from source end to the receiver end over a noisy channel is one of the biggest challenge. By introducing the concept of coding theory into the biological domain may gain insight into potential error-correcting aspects of genomic sequences and systems. In a cellular level, the information in DNA is transformed into proteins. Considering the DNA sequence bases [Adenine (A), Thymine (T), Guanine (G), Cytosine (C)] as the digital codes that we use in Digital communication system, we can transmit the genetic information over a noisy channel. In this paper we propose a new framework that not only detects the error but also modifies it during the transmission of genetic information in a communication channel using Hamming coding algorithm.

Keywords: Computational Biology, DNA sequencing, Error Correction, Hamming Principle, EDA

I. Introduction

In a communication system the coding theory and its techniques play a vital role in detection of errors and its control. These theoretic methods are also being used in molecular computing and genetic analysis. The availability of the large amount of genome datasets in public databases [4]-[6] helps to explore the error detection and correction capabilities of channel codes. In an engineering communication system we basically use two techniques for error control i.e. modulation and demodulation. The channel encoder encodes the digitized information and Modulation refers to prepare the information ready for transformation and in the other hand, Demodulation deals with the received message, prepares it for decoding and then the message is decoded by the channel decoder. The decoding process involves the removal and the correction of errors introduced during transmission [1]-[3].

In biological communication system, the information in DNA requires two phases (transcription & translation) to be transformed into protein. Mutations in DNA sequences generally occur through one of two processes: DNA replications and some of the environmental agents. When a DNA helicase separates DNA molecule into two strands, the DNA replication takes effect. Next, a protein called DNA polymerase each strand of DNA to create two double stranded DNA molecules. If the DNA polymerase makes any kind of mistake then the error occurs. But here the chances of error are very less i.e. once in 100,000,000 bases [17]. At the other hand, some of the environmental agents like nuclear radiations, ultraviolet light, and certain chemicals also damage the DNA sequences. It can affect the nucleotide bases by altering the base values.

Double stranded DNA Molecule $\Rightarrow$ messenger RNA (mRNA) $\Rightarrow$ chained Amino acids $\Rightarrow$ Proteins.

Schematic Representation DNA transformation into protein:

The encoder processes the digitized information frame by frame. An input frame consists of a fixed number, k ; of information symbols that are presented to the encoder. The output frame, the frame to be transmitted, consists of n (also fixed) output symbols, where the size of k & n are same. If the sequence is

corrupted during transmission, then the output stream mismatches with the input stream and we know that there is an error occurred with our original DNA sequence. Then in the next phase we modify the error with the help of our proposed algorithm and will be able to regenerate the original source at the receiver side. In computational biology the genetic information is encoded by the sequence of digits each with one of the four possible values. And in the decoding part, the digital code is converted into the analog code in protein form that carries the information regarding the function of the living cells [2].

In this paper, a new algorithm is proposed based on the coding framework for this channel encoding and decoding in molecular biology. This model not only detects the error during transmission but also modifies it.

II. Literature Survey

Application of coding theory to genetic data evolved in late 1950s [7-8]. Since then, error correcting (EC) coding methods have been applied to genetic sequence analysis and classification, biological chip design, as well as analysis of genetic regulatory processes. Konopka [9] applied the theory of degenerate coding to analyze redundancy and the error control properties of the genetic code (the codon to amino acid mapping). Sengupta and Tompa [10] approach the problem of oligo array design from a combinatorial design framework and use EC coding methods to increase the fidelity of oligo array construction. Reif and LaBean [11] propose EC coding-based methods for the development of error correction strands for repairing errors in DNA chips.

Andrea and R. palazzo *et al.* [12] proposed a model which consists of an encoder (a mapper and a BCH code over Z_4) and modulator (genetic code). Yockey [13] proposed one of the first models for gene expression using encoding and decoding concepts from communication theory. Battail [14] argued on the existence of the nested codes in DNA. Rosen [15] proposed a method for detecting linear block codes and so explaining the insertions and deletions in the DNA sequences. Leibovitch *et al.* proposed a procedure capable of determining whether a type of error-correcting code is or is not present in the DNA sequence. Schneider *et al.* [3] proposed algorithms to determine coding and non coding regions in DNA structure. May *et al.* [1] proposed the use of block and convolution codes in the translation initialization process in prokaryotic organisms. P.P Debata *et al.* [16] in 2012 proposed a model based on hamming coding algorithm that is used to detect the errors during transmission of genomic sequences. But their work was restricted to only the error detection. Modification of those errors is still a challenging task.

III. Proposed Model

A new algorithm, Error Detection Algorithm (EDA) is being proposed here. In which, we take original DNA sequences collected from the large databases as **input** and receive error free DNA sequence as **output**.

The channel encoder maps the base nucleotides of original DNA sequences (A,T,G,C) to binary bits and encodes the whole sequence. The encoded stream is now transmitted through a potentially noisy channel where the sequence can be corrupted in a random fashion. The output of the channel, the received message is prepared for decoding and then decoded by the channel decoder using Hamming Decoding Algorithm. The decoding process involves removal and correction of errors introduced during transmission. The decoding mechanism now can exceed the code's error correction capability in our newly proposed EDA model.

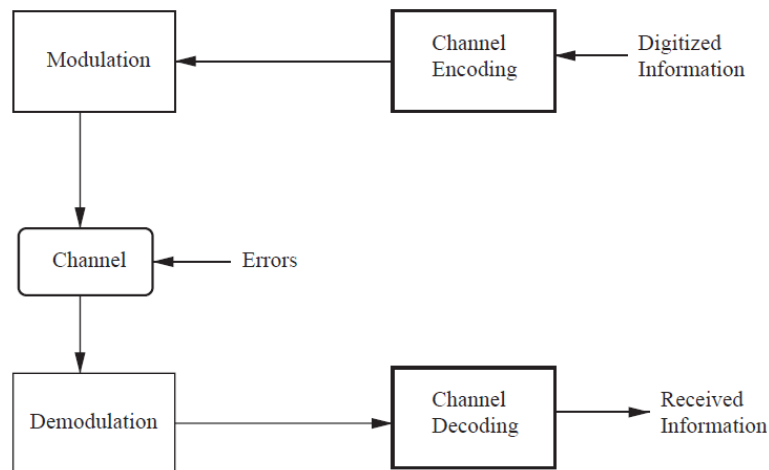


Figure.1: Block Diagram of the EDA model

ERROR DETECTION ALGORITHM (EDA):

- Step 1 : Generate nucleotide sequences from original DNA sequences.
- Step 2 : Select any sequence from the generated DNA sequences.
- Step 3 : Map the sequences by using binary mapping.
- Step 4 : Convert the DNA sequence into binary data bits of length K .

Step 5 : Add the parity bits to the data bits following $2^P \geq K+P-1$. where p = number of parity bits & k =number of data bits.

Step 6 : Represent each data bit & parity bit with a column vector.

Step 7 : Create a hamming matrix(H) by arranging the column vectors.

Step 8 :Repeat step-3 to 7 for all generated DNA sequences

Step 9 :For decoding a parity check matrix same as (H) will be constructed.

Step 10 :The syndrome vector s will be calculated by Checking up to the P^{th} parity relation starting from the first parity.

Step 11 :If $s=0$, assume there is no error, and store the sequence for generating amino acid sequences.

Step 12 :If $s \neq 0$, Capture the Frame no. and the error position of the corresponding frame .

Step 13 :Modify the error bit at the traced location and generate the error free DNA sequence .

Step14: Stop.

IV. Datasets

The EDA's performance was examined on various publically available data sets that are being used for gene structure prediction. i.e. ACU08131, AGGGLINE, AGU04852, ALOEGLOBIM etc. These data sets are available in <http://genome.crg.es/datasets/genomics96/> or directly from NCBI database.

V. Discussion & Results

The EDA was implemented on MATLAB 2008 and run on a Windows 7 machine with a Core i3 processor with 2 GB of RAM. Around 100 DNA sequences had been created from the above said data sets. And tested with the EDA .For example, a selected DNA sequence "CTGGGCTAAAAGGTCC" from the dataset having GI no. ACU08131 is mapped into the binary bits like[A=00,T=01,G=10,C=11].So the corresponding data bit in binary form is"11011010101101000000001010 011111". During transmission, If there is a change/error in the nucleotide bases of the DNA sequences($C = 11 \rightarrow T = 01$) [often called *single nucleotide polymorphism(SNP)*], then the frame no. and it's position in the frame is detected by EDA(Error has been generated randomly).The exact location of the error is now traced & finally rectified with EDA . And the output generated in the decoding side now is an error free DNA sequence.

VI. Results

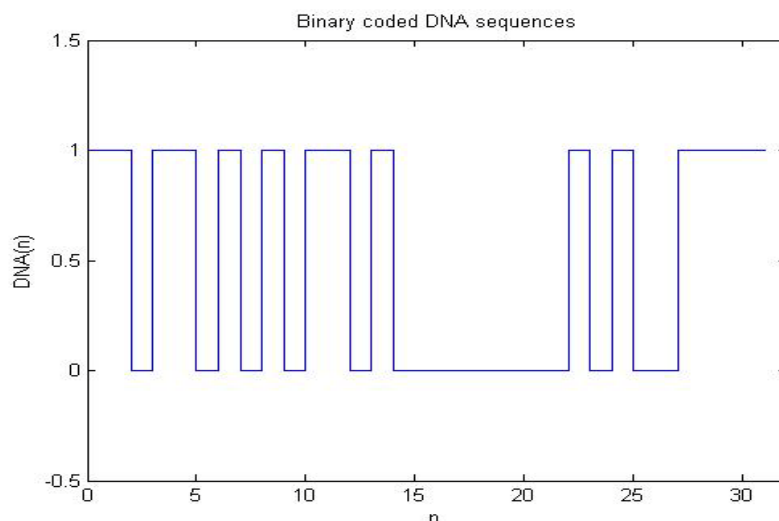


Figure.2: DNA sequence in Binary form.

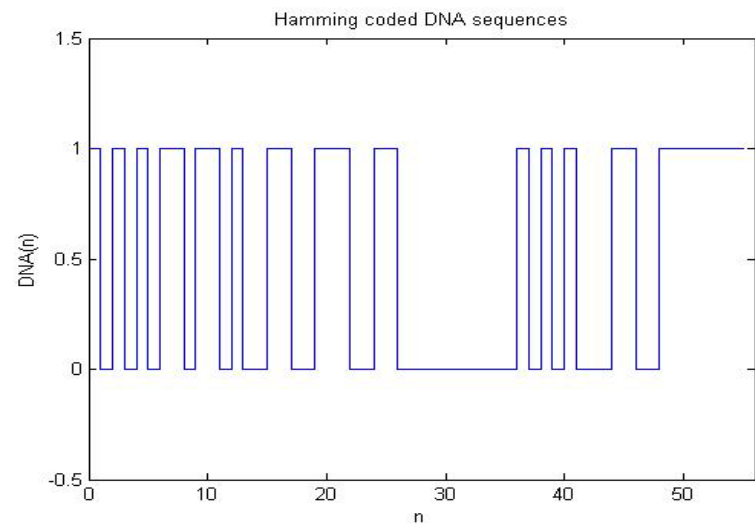


Figure.3: DNA sequence ready for transmission(Hamming Coded).

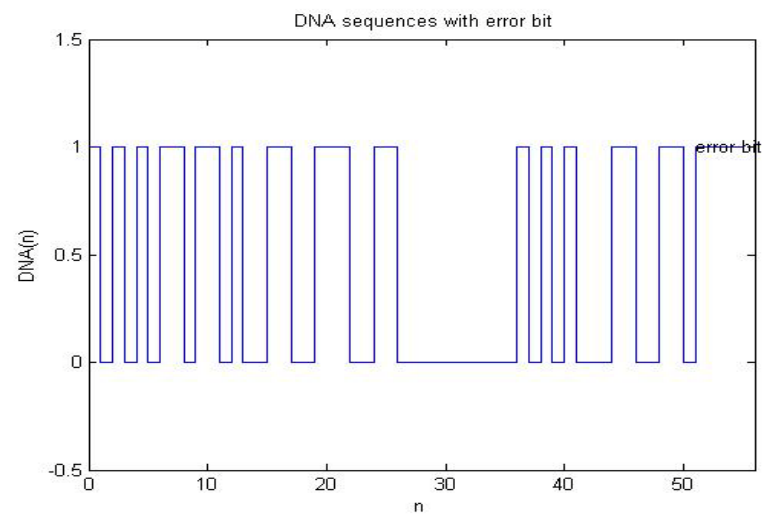


Figure.4: Error detected in particular frame & Position

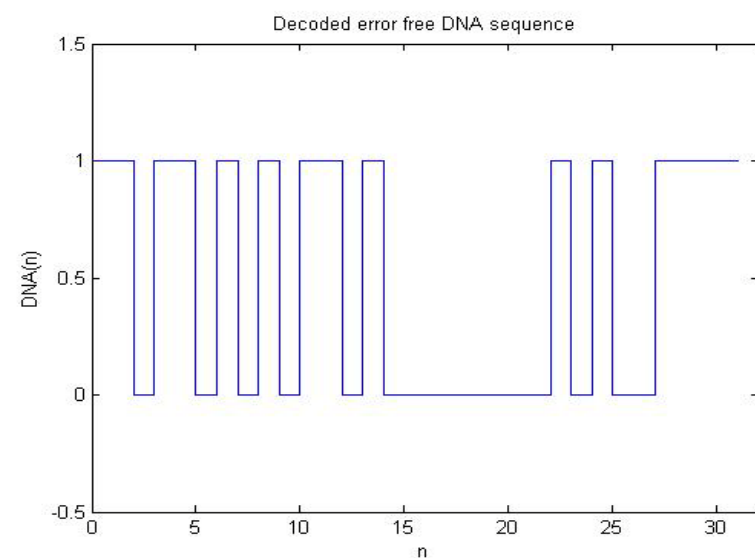


Figure.5: Error free DNA sequence generated by EDA.

VI. Conclusion

This EDA algorithm provides error control mechanism for single bit error correction in a particular frame. This algorithm also detects errors if occurred in different frames at the same point of time. But the EDA fails to check the errors when there are multiple error bits in a single frame. So this work can be extended to detect and modify multiple error bits in a single frame and also to handle the same by using any other coding techniques.

VII. References

1. E.May, M.Vouk, D.Bitzer and D.Rosnick.An error-correcting code framework for genetic sequence analysis.*Journal of Franklin Institute* ,2004; 34: 89-109.
2. R.Dawkins. The Blind Watchmaker, *Longman*,New York,1986.
3. T.D.Schneider ,Information Content of Individual Genetic Sequences. *Journal Of Theoretical Biology*.1997;189:427-441.
4. <http://genome.crg.es/datasets/genomics96/>
5. <http://www.ncbi.nlm.nih.gov/nuccore/>
6. <http://www.ncbi.nlm.nih.gov/protein>
7. B. Hayes, The invention of the genetic code, *Am. Sci.* 86 (1) (1998) 8–14.
8. S.W. Golomb, Efficient coding for the deoxyribonucleic channel, *Proceedings of the Symposia in Applied Mathematics*, New York , NY, USA, *Mathematical Problems in the Biological Sciences*,Vol. 14, American Mathematical Society, Providence, RI, 5–8 April 1961, pp. 87–100.
9. A.K. Konopka, Theory of degenerate coding and informational parameters of protein coding genes, *Biochimie* 67 (1985) 455–468.
10. R. Sengupta, M. Tompa, Quality control in manufacturing oligo arrays: a combinatorial design approach,*J. Comput. Biol.* 9 (1) (2002) 1–22.
11. J. Reif, T. LaBean, Computationally inspired biotechnologies: improved DNA synthesis and associative search using error-correcting codes and vector-quantization, in: A. Condon (Ed.), *DNA Computing: Sixth International Meeting on DNA-Based Computers (DNA6)*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Leiden, The Netherlands, June 2000.
12. A.Andrade,and R.Palazzo Jr.DNA Sequences Generated by Z_4 -linear codes. ISIT 2010,Austin,Texas,U.S.A.,June 13 18,2010.
13. H.Yockey .*Information Theory and Molecular Biology*,*Cambridge University press*:Cambridge,1992.
14. G.Battail . *Information Theory and error correcting codes in genetics and biological evolution.introduction to Bio-semiotic*,Springer,November,2006.
15. G.L.Rosen.Examining Coding Structure and Redundancy in DNA.*IEEE Engineering in Medicine and Biolog.*,2006;25:62-68.
16. P.P.Debata,D.Mishra,K.Shaw,S.Mishra.A coding theoretic model for error-detecting in DNA sequences.ICMOC-2012,ELSEVIER procedia engineering; 38:1773-1777.
17. <http://learn.genetics.utah.edu/archive/sloozeworm/mutationbg.html>.