Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 1773 – 1777

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Modeling Optimization and Computing (ICMOC-2012)

A Coding Theoretic Model for Error-detecting in DNA Sequences

Prajna Paramita Debata[a], Debahuti Mishra[b], Kailash Shaw[c], Sashikala Mishra[d]

[a,b,d]*Institute of Technical Education and Research, Siksha O Anusandhan Deemed to be University, Bhubaneswar, Odisha, India*
[c]*Gandhi Engineering College, Bhubaneswar, Odisha, India*

Abstract

A major problem in communication engineering system is the transmitting of information from source to receiver over a noisy channel. To check the error in information digits many error detecting and correcting codes have been developed. The main aim of these error correcting codes is to encode the information digits and decode these digits to detect and correct the common errors in transmission. This information theory concept helps to study the information transmission in biological systems and extend the field of coding theory into the biological domain. In the cellular level, the information in DNA is transformed into proteins**.** The sequence of bases like Adenine (A), Thymine (T), Guanine (G) and Cytosine (C) in DNA may be considered as digital codes which transmit genetic information. This paper shows the existence of any form error detecting code in the DNA structure, by encoding the DNA sequences using Hamming code.

Keywords: Error detecting and correcting codes; Information transmission; Biological systems; Coding theory; Encoding DNA sequences; Hamming code

1. Introduction

The need for coding theory and its techniques stems from the need for error control mechanisms in a communication system. In an engineering communication system; digitized information is encoded by the channel encoder and prepared for transmission (modulation). The encoded stream is 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 (demodulation) and then decoded by the channel decoder. The decoding process involves removal and possibly correction of errors introduced during transmission [1-3]. The decoding of received bit streams is fairly straightforward when the channel encoding algorithms are efficient and known. What if the encoding scheme is unknown or part of the data is missing? How would one design a viable decoder for the received transmission? Communication engineers may not frequently encounter this situation, but for computational biology this is the immediate challenge and barrier to understanding the vast amount of sequence data produced by genome sequencing projects. Genetic information is encoded by a sequence of digits, each with one of four possible values: adenine, cytosine, guanine, and thymine (A, C, G, and T). This digital code is translated into an analog code in the shape of proteins that carry out the functions of living cells [2]. The base sequence in DNA may contain a digital error correcting codes which correct the error during transmission [4]. Here we have proposed a model for the biological encoding and decoding system similar to that of a digital

* Tel.: +91-9439440195; fax: 91-674-2351880
E-mail address: prajnaparamitaa@gmail.com

1877-7058 © 2012 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
doi:10.1016/j.proeng.2012.06.216

communication system. The rest of the paper is organized as follows: in section 2 the related work on error correction for DNA sequences is explained, section 3 gives a closer look at our proposed model. In section 4, preliminary concepts related to our paper is given, in section 5 the proposed algorithm is given, section 6 explains the experimental evaluation finally section 7 concludes the paper and gives a future direction to our proposed work.

2. Related Work

May *et al.* [1] proposed the use of block and convolution codes in the translation initialization process in prokaryotic organisms. Schneider *et al.* [3] proposed some algorithmic procedures to determine the coding and non-coding regions in the DNA structure. Under the coding theory point of view, Liebovitch *et al*. [4] proposed a procedure capable of determining whether a type of error-correcting code is or is not present in the DNA sequence. Rosen [5] proposed a method for detecting linear block codes and so explaining the insertions and deletions in the DNA sequences. Battail [6] argued on the existence of nested codes in the DNA. Yockey [7] proposed one of the first models for gene expression using encoding/decoding concepts from communication theory. Andrea and R. Palazzo *et al.* [8] proposed a model which consists of an encoder (a mapper and a BCH code over Z_4) and a modulator (genetic code).

3. Schematic Representation of Proposed Model

In this model the *source* is the original *DNA sequences*. The *encoder* maps the genetic code alphabet (A, T, G, C) into binary bits and encodes. Then the decoder uses the *Hamming's decoding algorithm* to detect the error. The codeword at the decoder output is directly related to the mature *mRNA*. The genetic code may be viewed as a signal constellation, where each codon is considered as a signal in the signal constellation, the *tRNA* promotes the *matched mapping* (MM) [8], whereas the *ribosome* behaves as a digital signal processor. The final output is related to *protein*. The total procedure is illustrated in fig. 1.

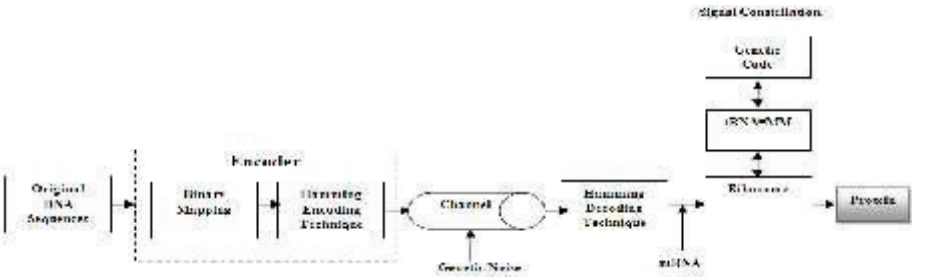

Fig.1 Encoding and decoding of DNA sequences for error correction

4. Preliminary Concepts

In biological communication system, after transcription and translation phase the information in *DNA* is transformed into *proteins* [9]. In transcription, the double stranded *DNA* molecule is used to synthesize a new single stranded molecule called *messenger RNA* (mRNA). The *RNA polymerase* binds to a specific region in the DNA in order to separate the two strands. Once they are separated, one of the strands serves as a template for the creation of the mRNA. The resulting mRNA is consequently spliced to remove the *introns* (non-coding regions) which results in a sequence of pure *exons* called mature mRNA. The mature mRNA travels in the cell until the ribosome binds to it at a specific region in order to start the process of translation [9]. Once the ribosome binds properly (translation initiation), it starts processing triplets of bases (also called codons) of the mature mRNA to produce amino acids. The ribosome serves as a platform for the *transfer RNA* (tRNA) molecule which holds the amino acids. A tRNA molecule connects using its *anti-codon* end with codons found in the mature mRNA until a sequence of amino acids is

chained. The formed chain then folds to finally produce a protein. The DNA can be modelled as an encoded information source that is processed in several steps to produce proteins [9]. During these processing steps, the processed *DNA* is subjected to *genetic error* which results in several types of mutations.

Causes of Genetic error in DNA Sequences

Genetic error in DNA sequences generally occur due to environmental agent and DNA replication. The environmental agent like ultraviolet light, nuclear radiation and certain chemicals can damage DNA by altering nucleotide bases so that they look like other nucleotide bases. When the DNA strands are separated and copied, the altered base will pair with an incorrect base and cause a mutation [10]. DNA replication begins when a protein called DNA helicase separates the DNA molecule into two strands. Next, a protein called DNA polymerase copies each strand of DNA to create two double-stranded DNA molecules. Error occurs when the DNA polymerase makes a mistake, which happens about once every 100,000,000 bases [10]. So we can use various error detecting and correcting codes to detect the presence of errors caused by noise or other impairments or mutations during transmission from the transmitter/nucleus to the receiver/organelle.

5. Algorithm for encoding and decoding of the generated DNA sequence using Hamming code

Input - Original *DNA sequences* from NCBI [11], start codons and stop codons.
Output- Error free sequences.

Step1- Generate the nucleotide sequences from the original *DNA sequences* by matching start and stop codons.

Step2- Select any one sequence from the generated *DNA sequences*.

Step3- Map the sequence using binary mapping. In binary mapping the nucleotides (A, T, G, C) can be mapped as (00, 01, 10, 11) respectively.

Step4- Now the DNA sequence is converted into binary data bits of length k.

Step5- Decide the number of parity bits (p) to be added with the data bits. The parity bits must follow $2^p \geq k + p - 1$[12] where p = number of parity bits and k = number of data bits.

Step6- Represent each data bit with a column vector.

Step7- Represent each parity bit with a column vector containing 1 in the row corresponding to each data bit included in the computation and a zero in all other rows.

Step8- Create a generator matrix, [G], by arranging the column vectors from the step 6 and 7 into a $k \times n$ matrix such that the columns are ordered to match their corresponding bits in a code word. Here $n = k + p$, the total number of code words to be transmitted.

For example – To create a generator matrix that produces code words with the bits ordered p_1, p_2, p_3, d_1, d_2, d_3, d_4 (3 parity bits followed by 4 data bits) use the vectors created in step 6 and 7 and arrange them into a 4×7 matrix.

$$G = \begin{array}{ccccccc} p1 & p2 & p3 & d1 & d2 & d3 & d4 \\ 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 1 \end{array}$$

Step 9- Multiply data bits with generator matrix [G] to produce an encoded DNA sequence.

For Example-If we encode the data value 1010 using Hamming code defined by the matrix *G*, then the encoded bits are: [1 0 1 0]× [*G*] = [1 0 1 1 0 1 0], so 1010 encodes to 1011010.

Step10 - Repeat step 3 to step 8 for all the generated sequences.

Step11- For decoding a parity check matrix [*H*] will be constructed.

For example: A 3×7 parity check matrix [*H*] may be constructed such that row 1 contains 1s in the position of the first parity bit and all the data bits that are included in its parity calculation. Row 2 contains 1s in the position of the second parity bit and all the data bits that are included in its parity calculation. Row 3 contains 1s in the position of the third parity bit and all the data bits that are included in its parity calculation. So the matrix [*H*] may be defined as follows:

$$H = \begin{matrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{matrix}$$

Step12- The syndrome vector(*S*) will be calculated as $S=Hr^T$, where *r* = received bits.

Step13- If *s* == 0, then assumes that there are no errors and stores the sequence for generating amino acid sequences.

Step14- If *s*≠0, then discard the sequence.

6. Experimental Evaluation

We have generated the 155 *DNA sequences* form the original *DNA sequence* collected from the NCBI with GI no. ACU08131 [11] by using the above mentioned algorithm. Table 1 shows the comparison with only one of the generated sequence in error and without error. More sequences can be given but it has been avoided here due to limit of space.

Table 1 Error detection in generated DNA sequence using Hamming code

DNA seq. GI no.	Generated DNA sequence	Binary Mapping A=00,T=01 G=10,C=11	Length of data bits (*k*)	Parity Bits (*p*)	Encoded Code words	Received bits(*r*)	Calculate Syndrome (*S*)	Error Detected
ACU08131	CTGGGCTAA	1101101011010000	18	5	0010011 0110101 0110100 00	001001 101101 010110 10000	*S*=0	No
ACU08131	CTGGGCTAA	1101101011010000	18	5	0010011 0110101 0110100 00	001001 101101 010110 110010	*S*≠0	YES

Observations

- In this paper, every generated *DNA* sequence is mapped into binary bits. These bits are considered as data bits which can be encoded and decoded using the Hamming code [12-13]. In the table 1 the nucleotide sequences CTGGGCTTAA are encoded as 0010011011010101011010000.

- If there will be change of one nucleotide (A (00) → G (10)) then the error is detected by decoding received bits. From biological point of view, this observation shows *silent mutation*. *Silent mutation* shows that the single nucleotide will often not change the resulting amino acid rather lead to an error.
- In this observation, the amino acids generated from the sequence are not changed after mutation. But in case of some other sequences, the change in single nucleotide may change the amino acid. In biological context this mismatch is known as a *single nucleotide polymorphism* (SNP).
- So the error in generated *DNA* sequence should be detected first. Further it can be corrected by any of the error correcting codes.

7. Conclusion and Future Work

This paper shows that the DNA sequence can be encoded and decoded by Hamming code to detect the single nucleotide polymorphism (SNP). As a summary, this work will help in stimulating the interdisciplinary research efforts to apply techniques from the field of communication theory to other problems from the field of genetics. This work can be extended to not only detect the errors but also to modify the found errors.

References

[1] E. May, M. Vouk, D. Bitzer and D. Rosnick, An error-correcting code framework for genetic sequence analysis. *Journal of the 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] L.S. Liebovitch, Y. Tao, A.T. Todorov, and L. Levine. Is there an error correcting code in the base sequence in DNA? Biophysical Journal, vol. 71, pp. 1539-1544, 1996.
[5] G.L. Rosen. Examining coding structure and redundancy in DNA. *IEEE Engineering in Medicine and Biology*. 2006; 25: 62-68.
[6] G. Battail. Information theory and error correcting codes in genetics and biological evolution. *Introduction to Bio-semiotic*, Springer, November 2006.
[7] H. Yockey, Information Theory and Molecular Biology, *Cambridge University Press*: Cambridge, 1992.
[8] 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
[9] Z. Dawy, P. Hanus, J. Weindl, J. Dingel, and F. Morcos. On genomic coding theory. *European Transactions on Telecommunications*. 2008; 18: 873-879.
[10] http://learn.genetics.utah.edu/archive/sloozeworm/mutationbg.html.
[11] http://www.ncbi.nlm.nih.gov/nuccore.
[12] S. J. Chung. Network Architecture: Hamming Codes and Cyclic Redundancy for Transmission Error Correction. ACM. 2001; 33:4.
[13] Chugo Fujihashi, All Error Detecting Modified Hamming Codes for Ideal Optical Channel and Application to Practical System. *Academic Reports Fac. Eng. Tokyo Polytech. Univ.* 2008; 3: 1.