

# DNA Sequence and Polynomial Representations

Rosalio G. Artes, Jr. and Lady Lee L. Lusanta

Department of Mathematics and Statistics  
College of Science and Mathematics  
MSU - Iligan Institute of Technology  
9200 Iligan City, Philippines

Copyright © 2015 Lady Lee L. Lusanta and Rosalio G. Artes Jr. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

## Abstract

A deoxyribonucleic acid (DNA) is a self-replicating material present in nearly all living organisms as the main constituent of chromosomes. It is a nucleic acid that carries the genetic information in cells and some viruses, consisting of two long chains of nucleotides twisted into a double helix and joined by hydrogen bonds between the complementary bases adenine and thymine or cytosine and guanine. Comprehending the DNA sequence has also become indispensable in various fields, particularly in mathematics.

Using binary coding techniques, we established a correspondence between the set of DNA sequences and the ring of polynomials in indeterminate  $x$  with coefficients in  $\mathbb{Z}_2 = \{0, 1\}$ , which is denoted by  $\mathbb{Z}_2[x]$ . Given a DNA sequence  $S = S_0S_1S_2 \cdots S_k$ , where  $S_i \in \{A, C, G, T\}$ , for  $i = 1, 2, \dots, k$ , with  $A$  stands for adenine,  $C$  for cytosine,  $G$  for guanine, and  $T$  for thymine, we model a DNA sequence using a polynomial in  $\mathbb{Z}_2[x]$  via binary codes. We have shown that the sum of the polynomial representations of complementary DNA sequence in a double helix structure is a geometric series of common ratio  $x$  with degree equal to  $2k$ . Finally, we have shown that for every polynomial  $f \in \mathbb{Z}_2[x]$ , there exists a unique DNA sequence whose polynomial representation is exactly equal to  $f(x)$ .

**Subject Classification:** 92B05

**Keywords:** DNA sequence, polynomial representation, binary code

## 1 Introduction

Recent advances in molecular biology significantly affect our lives in many ways. Molecular biology which aims to study DNA's structure and functions, has motivated research in various scientific disciplines, and mathematics is one of them. Over the last few decades, the study of DNA sequence has become one of the most important researches in biomathematics.

In most living cells, deoxyribonucleic acid (DNA) is known to have the famous double-helix structure. The information in DNA is stored as a code made up of four nucleotide bases: adenine, cytosine, guanine, and thymine. Within the helix, an adenine in one polynucleotide is always adjacent to a thymine in the other strand, and similarly guanine is always adjacent to cytosine. This is called base pairing and involves the formation of hydrogen bonds between an adenine and a thymine, or between a cytosine and a guanine [18]. The order or sequence of the nucleotide bases uncovers the information obtained for building and sustaining an organism. It is analogous to the way in which letters of the English alphabet appear in a definite order to produce words and sentences.

Obtaining reliable systems for characterizing DNA sequences is the main phase for creating mathematical descriptors. We focus only on the polynomial representation as the mathematical descriptor in characterizing sequences of DNA. Mathematical modeling then plays an important task not only in capturing and analyzing complex networks governing biological processes, but also in designing efficient experiments to further understand the system. In this study, we model a DNA sequence using a polynomial with binary coefficients using binary coding technique.

## 2 Preliminary Notes

A *deoxyribonucleic acid* (DNA) is a double-helix - two strands wound around each other. DNA has only four different bases: adenine (A), cytosine (C), guanine (G) and thymine (T). A *DNA sequence* is a sequence consisting of four bases.

Let  $R$  be a ring and let  $R[x]$  denote the set of all sequences of elements of  $R(a_0, a_1, \dots)$  such that  $a_i = 0$  for all but a finite number of indices  $i$ . The ring  $R[x]$  is called the *ring of polynomials* over  $R$ .

A *binary word* of length  $n$  over  $\mathbb{Z}_2$  is a sequence  $u = u_1u_2 \cdots u_n$  with each  $u_i \in \mathbb{Z}_2$  for all  $i$ .

### 3 Results and Discussion

Let  $B$  be a finite set of binary strings. Define  $\varphi : B \rightarrow \mathbb{Z}_2[x]$  by

$$\varphi(b) = f_b(x) = \sum_{i=0}^k b_i x^i, b = b_0 b_1 \cdots b_k.$$

The polynomial  $f_b(x)$  is the *polynomial corresponding to*  $b \in B$ .

Given a DNA sequence  $S = S_0 S_1 \cdots S_k$ ,  $S_i \in \{A, C, G, T\}$  we consider the following binary coding:  $[A] = 00$ ,  $[C] = 01$ ,  $[G] = 10$ , and  $[T] = 11$ . Then the *binary representation of*  $S$  is given by

$$b_S = [S_0][S_1] \cdots [S_k].$$

The *polynomial representation corresponding to the DNA sequence*  $S$  is given by

$$f_S(x) = f_{b_S}(x) = \sum_{i=0}^k b_{S_i} x^i.$$

**Theorem 3.1** *Let  $S_1$  and  $S_2$  be the complementary DNA strands of length  $k$ . Then*

$$f_{S_1}(x) + f_{S_2}(x) = \frac{1 - x^{2k+2}}{1 - x}$$

*Proof:* Let  $f_{S_1}(x)$  and  $f_{S_2}(x)$  be the polynomial representations of  $S_1$  and  $S_2$ , respectively. Suppose  $S_1 = D_0 D_1 \cdots D_k$  where  $D_i \in \{A, C, G, T\}$ . Then  $S_2 = D_0^C D_1^C \cdots D_k^C$  where  $D_i^C \in \{A, C, G, T\}$ . Since  $A$  and  $T$  are complementary, as well as  $C$  and  $G$ , it follows that

$$b_A + b_T = 00 + 11 = 11$$

and

$$b_C + b_G = 01 + 10 = 11.$$

Thus,

$$\begin{aligned} b_{S_1} + b_{S_2} &= b_{D_0 D_1 \cdots D_k} + b_{D_0^C D_1^C \cdots D_k^C} \\ &= b_{D_0 A A \cdots A} + b_{A D_1 A \cdots A} + \cdots + b_{A A A \cdots D_k} + b_{D_0^C A A \cdots A} \\ &\quad + b_{A D_1^C A \cdots A} + \cdots + b_{A A A \cdots D_k^C} \\ &= \underbrace{11000000 \cdots 00}_{2(k+1)\text{-terms}} + \underbrace{00110000 \cdots 00}_{2(k+1)\text{-terms}} + \cdots + \underbrace{000000 \cdots 11}_{2(k+1)\text{-terms}} \\ &= \underbrace{11111111 \cdots 11}_{2(k+1)\text{-terms}} \end{aligned}$$

Note that the polynomial representation of  $\underbrace{11111111 \cdots 11}_{2(k+1)\text{-terms}}$  is given by

$$f_{S_1}(x) + f_{S_2}(x) = 1 + x + x^2 + \cdots + x^{2k+1}.$$

Since

$$1 + x + x^2 + \cdots + x^{2k+1} = \frac{1 - x^{2k+2}}{1 - x}.$$

Therefore, we have

$$f_{S_1}(x) + f_{S_2}(x) = \frac{1 - x^{2k+2}}{1 - x}.$$

■

For every polynomial with binary coefficients, there exists a DNA sequence with polynomial representation corresponding to it. To see this, we have the following result.

**Theorem 3.2** *For every  $f \in \mathbb{Z}_2[x]$ , there exists a DNA sequence  $S$  such that  $f_S(x) = f(x)$ .*

*Proof:* Let  $f \in \mathbb{Z}_2[x]$ . Consider the following cases:

Case 1:  $\deg f = 2k + 1$

Partition the binary coefficients into  $k$  adjacent pairs. For each pair, there corresponds a code in  $\{A, C, G, T\}$ . The codes corresponds to the polynomial representation corresponding to the DNA sequence.

Case 2:  $\deg f = 2k$

Add the last term  $0x^{2k+1}$  in  $f$ . Then  $\deg f$  is odd. Proceed as Case 1. Thus, there exists a DNA sequence  $S$  such that  $f_S(x) = f(x)$ . ■

## References

- [1] Austin, Andrea. (2007). *The Circuit Partition Polynomial with Applications and Relations to the Tutte and Interlace Polynomials*.
- [2] Bhattacharyya, Debnath and Bandyopadhyay, Samir Kumar. (Feb 2013) *Hiding Secret Data in DNA Sequence*. International Journal of Scientific & Engineering Research. Volume 4, Issue 2. ISSN 2229-5518.
- [3] Brown, Terry. (2011). *Introduction to Genetics: A Molecular Approach*. United States of America: Garland Science, Taylor & Francis Group, LLC.
- [4] Chartrand, G. and Lesniak, L. (1996). *Graphs & Digraphs*. United States of America: Chapman & Hall/ CRC.

- [5] Cooper, Necia Grant. (1994). *The Human Genome Project: Deciphering the Blueprint of Heredity*. University Science Books.
- [6] Dahm, Ralf. (2005). *Friedrich Miescher and the discovery of DNA*. *Developmental Biology*. 274-288.  
<http://dx.doi.org/10.1016/j.ydbio.2004.11.028>
- [7] Gunstream, Stanley (2012). *Explorations in Basic Biology*. California: Pearson Education Inc.
- [8] Jafarzadeh, Nafiseh and Iranmanesh, Ali. (2012). *A Novel Graphical and Numerical Representation for Analyzing DNA Sequences Based on Codons*. MATCH Communications in Mathematical and in Computer Chemistry.
- [9] Hsu, H. Z. and Lee, R. C. T. *DNA Based Encryption Methods*. The 23rd Workshop on Combinatorial Mathematics and Computation Theory.
- [10] Hungerford, Thomas W. (1974). *Algebra*. United States of America: Springer-Verlag New York, Inc.
- [11] Kaptcianos, Jonathan.(2008, April 1). *A Graph Theoretical Approach to DNA Fragment Assembly*. *American Journal of Undergraduate Research*.
- [12] Kaptcianos, Jonathan. (2012, August 23). *A Graph Theory Aiding DNA Fragment Assembly*.
- [13] Munshi, Anjana. (2012). *DNA Sequencing - Methods and Applications*. Croatia: Intech. <http://dx.doi.org/10.5772/2158>
- [14] Patrinos, Ari and Drell, Daniel. *The Human Genome Project: Sequencing the Future*.
- [15] Pevzner, Pavel. (2000). *Computational Molecular Biology: An Algorithmic Approach*. MIT Press.
- [16] Pevzner et al. (2001, June 7). *An Eulerian path approach to DNA fragment assembly*.
- [17] Vance, Eldridge P. (1980). *Modern College Algebra and Trigonometry*. Addison-Wesley Publishing Company.
- [18] Weaver, Robert F. (2002). *Molecular Biology*.

**Received: February 15, 2015; Published: March 14, 2015**