SSRG International Journal of Computer Science and Engineering
ISSN: 2348–8387 / https://doi.org/10.14445/23488387/IJCSE-V11I7P105

Volume 11 Issue 7, 28-32, July 2024

Original Article

Residue Number System (RNS) Di-base Table for SOLiD Sequencing

Joshua Apigagua Akanbasiam[1], Kwame Osei Boateng[2]

[1]*Department of Electrical/Electronics Engineering, Dr Hilla Limann Technical University, Wa, Ghana.*
[2]*Department of Computer Engineering, Kwame Nkrumah University of Science and Technology Kumasi, Ghana.*

Corresponding Author : ja.akanbasiam@dhltu.edu.gh

Received: 20 May 2024 Revised: 30 June 2024 Accepted: 16 July 2024 Published: 31 July 2024

Abstract *- A two-moduli set Residue Number System (RNS) di-base is designed using a grid approach. The grid approach allows for flexibility in the choice of digit-base substitution in generating the di-base table. The entire design is flexible, answers the quest for a quaternary number system for molecular biological considerations, and is also well structured over the entire residue number system space. The initial approach of giving digital connotations to the di-base table is designed in the well-known binary system - which is the most considered number system for contemporary digital applications. This approach is static and also not well structured over the entire binary number system space. RNS has emerged lately as a number system that has numerous advantages over traditional number systems including fast processing, reduced power consumption, and increased resistance to errors. Any set of two (2) moduli RNS can generate the di-base table for SOLiD sequencing. This design is similar to the canonical di-base table designed by Applied Biosystems Instruments (ABI) for SOLiD sequencing. Thus, the design presents similar moduli digits for each cell given any set of moduli but different decimal values. It also suits the quest for a quaternary number system for molecular biological or bioinformatics considerations – since the number of known nitrogenous bases is four (4).*

Keywords *- Residue Number System, di-base table, RNS di-base table, Sanger sequencing, Next Generation Sequencing, SOLiD Sequencing.*

1. Introduction

The process of ascertaining the nucleotide order in a DNA molecule is known as "DNA sequencing." The first complete genome to be sequenced is the bacteriophage φX174. Understanding our DNA can provide explanations for the causes of various diseases and future diseases organisms might develop [1]. The development of the Sanger, Maxam, and Gilbert sequencing techniques—also referred to as first-generation sequencing technologies—was prompted by the significance of DNA sequencing [2]. The future of DNA sequencing suggests that Sanger's sequencing was too slow for today's demand for high speed, accuracy, and a relatively low cost per base sequencing. Next Generation Sequencing (NGS) has emerged to meet the demand for faster, more accurate, and cost-effective sequencing. There are a number of NGS algorithms, but SOLiD stands out for its high accuracy and throughput, and this is attributed to its reliance on a rather unique decoding method based on the di-base table [4]. The digital realization of the di-base table developed by Applied Biosystems Instruments (ABI) for SOLiD sequencing is static and not well structured effectively in the widely held binary number system framework and, as such, lacks flexibility. An innovative RNS di-base table for SOLiD sequencing is designed using 2 moduli set RNS and a grid algorithm. This design maintains the necessary properties for successful SOLiD sequence decoding and also suites the demand for a quaternary number system for ease of molecular biological applications. A code table based on RNS is envisaged to leverage the properties of the number system to give strong support for SOLiD sequencing.

The mathematical principles governing this design are rooted in Gamow's postulations of a ($4^3 = 64$) genetic code table. Thus, a ($4^2 = 16$) will suffice in generating the di-base table. Where the base four (4) represents the four known nitrogenous bases and the index two (2) represents the two (2) moduli sets required to generate the di-bases. The resultant value sixteen (16), represents the number of di-bases for the di-base table for SOLiD sequencing. The di-base table is made of four dyes – blue, green, yellow, and red – and is utilised in SOLiD sequencing to encode all sixteen (16) potential two-base pairs. SOLiD, with its two-base encoding approach, achieves a relatively high accuracy of about 99.9999% and can also differentiate between true genetic variations, Single Nucleotide Polymorphism (SNPs) and

measurement errors. Just as the Rosetta stone was the key to deciphering Egyptian hieroglyphs [5] [6] and the genetic code the key to DNA sequencing, so is the di-base matrix or table a requirement for successful SOLiD sequencing.

2. Residue Number System (RNS)

Residue Number System (RNS) is an integer number system famously attributed to Sun Tzu that speeds up arithmetic computations by splitting them into smaller parts, making each part independent of the other [7]. It is a good alternative to conventional arithmetic, based on a weighted number system. The main cause for performance degradation in arithmetic circuits is the carry propagation scheme, which is avoided using RNS. RNS has transitioned from moduli consideration through effective converter designs and lately, the interest is its application in other fields of study [8]. Traditionally, RNS has applications in Digital Signal Processing, data communications networks, digital image and video processing, and cryptography [9]. Quite lately RNS is considered in areas like the Internet of Things, Cloud storage, artificial intelligence, and bioinformatics.

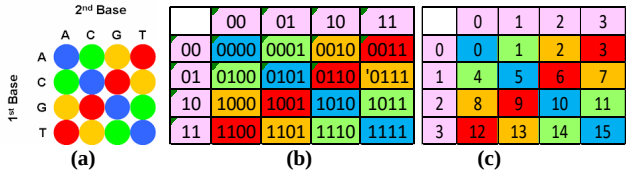
If q and r are the quotient and remainder, respectively, of the integer division of a by m, that is, $a = qm + r$ then, by definition, we have $a \equiv r\ (mod\ m)$. The number r is said to be the residue of a- with respect to m, and this is denoted by $r = |a|m$. The set of m values {0; 1;…; m – 1} that the residue may assume is called the set of least positive residues modulo m. For the efficiency of moduli, the radix-2 moduli are considered, and for higher dynamic range, higher moduli sets like 4, 5, and 6 moduli sets are proposed. The most common moduli set used in RNS applications is the traditional three-moduli set, $\{2^n - 1, 2^n, 2^n + 1\}$. Only two (2) moduli sets are required for the design of an RNS di-base table. Thus, any combinations of these effective moduli can be considered for the design of the di-base table for SOLiD sequencing. $\{2^n - 1, 2^n\}$ or $\{2^n - 1, 2^n + 1\}$ or $\{2^n, 2^n + 1\}$. Some features of RNS that are undesirable and have thwarted the efforts of RNS are, magnitude comparison, overflow detection, division, and the unique representation of numbers is limited to the dynamic range beyond which numbers repeat. Since the RNS di-base table will not necessarily require the invocation of these undesirable features, the RNS di-base table can be very attractive in the design and implementation of an RNS di-base table using a grid approach for SOLiD sequencing.

3. SOLiD Sequencing

There are severe limitations associated with the first-generation sequencing algorithms, speed, accuracy, and cost, leading to the emergence of Next-Generation Sequencing (NGS) methods. Some of these NGS methods are Applied Biosystems Instruments (ABI) - SOLiD sequencing technology, Illumina sequencing technology, Helicos system, Pacific Bioscience technology, and Roche 454 sequencing technologies [1][10]. The advantages of these technologies over each other are solely based on the technology adopted in sequencing [4] [11]. SOLiD sequencing is reliant on the di-base for decoding, and the decoding process assays two base positions at a time, forming the basis for the innovativeness of this sequencing method. The technique encodes each of the sixteen possible two-base combinations using four fluorescent dyes: red, blue, green, and yellow. The benefit of this technique is that it produces sequencing data with a relatively high accuracy that is higher than that of other ways of sequencing. It has the major problem of dealing with palindromic sequences [12]. The sequencer uses a two-base sequencing technique based on ligation sequencing, and during sequencing, each base is examined twice due to the two-base encoding, which provides incredibly high confidence in the detection of rare SNPs. A digital realization is achieved with each of the four bases for each row and column assigned binary digits. This limits the binary design to the numbers zero to fifteen for a sixteen di-base table, making the binary design static. Also, since the nitrogenous bases are four, the quest for a structured quaternary number system for molecular biological designs remains unresolved. The two-modulus set RNS grid design for the di-base table offers a quaternary approach and a more flexible di-base table for SOLiD sequencing. This two-base encoding enables researchers to shift their attention from low-quality data to the biological importance of the data. The major drawbacks are the long run times and the short read length; however, the short read length of the SOLiD sequencing algorithm is usurped by the lowest error rates due to the 2-base encoding technique used [13].

4. Di-Base Table

The di-base table is made of four colours: blue, green, yellow, and red. Each colour is an intersection of two nitrogenous bases in a grid. The row and column of the nitrogenous bases intersect in a cell, which represents the di-base. These di-bases are assigned their respective di-base colours in conformity with the canonical di-base table for SOLiD next-generation sequencing. Each cell is, therefore, a concatenation of two bases (di-base). The colours for each cell are represented following the rules that each column or row has each of the four colours. The leading diagonal has one unique colour: blue. The trailing diagonal also has one unique colour: red. A base and its reverse should have the same colour [12]. Figure 1 below captures the canonical di-base table and its binary and corresponding decimal digit representation.



(a) (b) (c)

Fig. 1 The di-base table (a) Canonical [14] (b) Binary digits (c) Binary-Decimal values

If each row or column of the grid of bases is represented by the number 0 – 3, it implies the di-base table is constructed with a two-bit binary representation. This binary representation is tied to each cell and it is static and also has static decimal values. This table is observed as a four (4) bit binary representation, from $0 - 2^n - 1$, where n is the number of bits, binary zero (0000) to fifteen (1111). Per this representation, the colour blue will always be represented in binary as 0000, 0101, 1010 and 1111, with the decimal values as 0, 5, 10 and 15, respectively. Also, the colour red, made of the trailing diagonal, will always have the binary digits 0011, 0110, 1001 and 1100 and the decimal values 3, 6, 9 and 12, respectively. Thus the di-base table in this form is static and limited to the digits 0 to 15 for binary and decimal representation and does not suit the entire binary number system space.

5. Methodology

The di-base table is generated as a two-moduli RNS. Additionally, a grid approach is adopted in generating the di-base table. Which is flexible compared with its static binary counterpart. The RNS-based approach to generating the di-base is universal and spans the entire RNS space, implying that any chosen set of moduli can be used for its generation. The designing of an RNS di-base code for successful SOLiD sequencing requires two (2) moduli set RNS. The two moduli sets are considered as m_1 and m_2. These two moduli are arranged into a grid of m_1 rows of blocks and m_2 columns of blocks. The intersection of any two digits of the two moduli occurs in a cell that constitutes an RNS representation. These combinations of digits in the cell are replaced with nitrogenous bases. The order of the digit-base substitution is 0 –> A, 1–> C, 2–> G, 3–>T, where A is Adenine, C is cytosine, G is guanine, and T is Thymine. The colour codes are assigned to each combination of bases in the grid, and this conforms to the basic rules set by the canonical di-base table for SOLiD sequencing. The table generated is similar to the canonical di-base designed by Applied Biosystems Instruments (ABI) and also obeys all the rules set out for successful SOLiD sequencing. The concatenated two digits of RNS have a decimal value - thus, each di-base has its decimal value. Irrespective of the moduli set chosen, the sorted RNS digits that make up the di-bases in each cell do not change, but the decimal values change. This characteristic can be exploited for further molecular biological analysis.

6. Design Algorithm

RNSDibase(m_1, m_2)
Input: Two relatively prime numbers, m_1, m_2
Output: Genetic Dibase Table

```
1.      Array dibaseTable[4, 4]
2.      for i ← 0 to m1 - 1 do
3.              for j ← 0 to m2 - 1 do
4.                      if ((i < 4) && (j < 4))
5.                              dibaseTable[i,j] ← ij
6       return dibaseTable
```

The Residue Number System di-base algorithm takes two relatively prime integer numbers, m_1 and m_2, as inputs. The inputs are processed to generate a permutation of the 4 nitrogenous di-bases for SOLiD sequencing. At the beginning, an m x n array is defined as dibaseTable. Since there are 4 nitrogenous bases for SOLiD sequences, m and n are both assigned 4. The algorithm continues with two nested loops. The outer loop is initialized with a counter **i,** which is assigned 0 and runs to m-1. The inner loop counter variable is defined as j initialized at 0 and runs from 0 to n-1. During each iteration of the inner loop, dibaseTable[i, j] is indexed, the value of the loop counters **ij** if both counters **i** and **j** are less than 4. The algorithm returns the dibaseTable array when the loops complete execution

7. Results and Discussion

4,5	0	1	2	3	4
0	00	01	01	03	04
1	10	11	12	13	14
2	20	21	22	23	24
3	30	31	32	33	34

(a)

4,5	0	1	2	3	4
0	0	16	12	8	4
1	5	1	17	13	9
2	10	6	2	18	14
3	15	11	7	3	19

(b)

Fig. 2 RNS Di-Base Table, moduli [4, 5] (a) Residue digits (b) RNS-Decimal value

5,6	0	1	2	3	4	5
0	00	01	02	03	04	05
1	10	11	12	13	14	15
2	20	21	22	23	24	25
3	30	31	32	33	34	35
4	40	41	42	43	44	45

(a)

5,6	0	1	2	3	4	5
0	0	25	20	15	10	5
1	6	1	26	21	16	11
2	12	7	2	27	22	17
3	18	13	8	3	28	23
4	24	19	14	9	4	29

(b)

Fig. 3 RNS Di-Base Table, Moduli [5,6] (a) Residue digits (c) RNS-Decimal values

4,5	0	1	2	3
0	00	01	02	03
1	10	11	12	13
2	20	21	22	23
3	30	31	32	33

(a)

5,6	0	1	2	3
0	00	01	02	03
1	10	11	12	13
2	20	21	22	23
3	30	31	32	33

(b)

Fig. 4 RNS Di-base table residue digits (a) moduli [4,5] (b) moduli [5,6]

Figure 2 (a) presents the concatenated residue digits for the rows and columns of the RNS di-base table for the moduli set [4, 5]. Figure 2 (b) holds the corresponding decimal values for each cell of the RNS di-base table in Figure 2 (a).

4,5	0	1	2	3
0	0	16	12	8
1	5	1	17	13
2	10	6	2	18
3	15	11	7	3

(a)

5,6	0	1	2	3
0	0	25	20	15
1	6	1	26	21
2	12	7	2	27
3	18	13	8	3

(b)

Fig. 5 RNS Di-base table - Decimal Values (a) Moduli [4,5] (b) Moduli [5,6]

Figure 3 (a) and (b) hold the concatenated residue digits and corresponding decimal representation for the moduli set [5, 6]. Figure 4 presents a truncated RNS di-base with residue digits for moduli sets [4, 5] and [5, 6] in (a) and (b), respectively. It can be observed that the residue digits for any two moduli set selected for the di-base table will always be the same irrespective of the moduli. Also, figure 5 shows a truncated RNS di-bas table with corresponding decimal values for moduli sets [4, 5] and [5, 6]. This further proves that the corresponding decimal values will always be different for different moduli sets, and the di-base table will not be affected by any change in the relatively prime moduli set selected for implementation. Thus, the design can take advantage of any effectively implemented moduli sets. With the use of the RNS grid, each of the digits that make up a particular block of colour will always be in the same position irrespective of the number and type of moduli sets selected. The main difference in choosing different moduli comes from changes in the decimal values of each of the concatenated residue digits that make up each colour.

8. Complexity of Algorithm – Big O

Summary

$t(n) = 1 + (1+m1)+(1+m2) + 3(1+m1)(1+m2) + 2((1+m1)(1+m2)) + 1$

$t(n) = 9 + 6m1 + 6m2 + 3m1m2 + 2m1m2$

as m1, m2 becomes large, m1 = n, m2 = n

$t(n) = 5n^2 + 12n + 9$

as $n \rightarrow \infty$

$t(n) = O(n^2)$

Fig. 6 Big O notation of RNS di-base algorithm

Figure 6 presents the big O notation of the RNS di-base algorithm. This presents an objective manner of comparing the time complexity of the algorithm in a hardware-independent manner. The time complexity is O of n^2 which represents the worst case and is similar in comparison with a binary case. But it remains that the RNS approach has the advantage of flexibility over the entire number system space as compared with binary which is static. Also, the RNS di-base approach suits the quest for a quaternary number system for molecular biological applications.

9. Conclusion

A two-moduli set RNS di-base table is designed. Compared with the canonical di-base table design by Applied Biosystems Instruments (ABI), whose digital realization is static and hence does not span the entire binary number system space. This design is flexible and spans the entire residue number system space. Thus, any two moduli set RNS will successfully generate the di-base table with similar RNS digits in each cell but different decimal values. The design further suits the quest for a quaternary number system for molecular biological applications.

In summary, the RNS di-base table generated is similar to the canonical di-base designed by Applied Biosystems Instrument. It obeys all the rules set out for a successful SOLiD sequencing and offers flexibility over the RNS number system space. It also presents a quaternary approach to the generation of the di-base table, which is a desired consideration by most molecular biologists – since there are four (4) known nitrogenous bases.

References

[1] Chu Cheng, Zhongjie Fei, and Pengfeng Xiao, "Methods to Improve the Accuracy of Next-Generation Sequencing," *Frontiers in Bioengineering and Biotechnology*, vol. 11, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Eva C. Berglund, Anna Kiialainen, and Ann-Christine Syvänen, "Next-Generation Sequencing Technologies and Applications for Human Genetic History And Forensics," *Investigative Genetics*, vol. 2, no. 1, pp. 1-15, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[3] Mehdi Kchouk, Jean-François Gibrat, and Mourad Elloum, "Generations of Sequencing Technologies: From First to Next Generation," *Biology and Medicine*, vol. 9, no. 3, pp. 1-8, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Heena Satam et al., "Next-Generation Sequencing Technology: Current Trends and Advancements," *Biology*, vol. 12, no. 7, pp. 1-25, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] R. Stone, Rosetta Stone, vol. 11, pp. 1-8, 2007.

[6] Koji Tamura, "The Genetic Code: Francis Crick's Legacy and Beyond," *Life*, vol. 6, no. 3, pp. 1-5, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[7] Hassan Kehinde Bello, and Kazeem Alagbe Gbolagade, "Residue Number System: An Important Application in Bioinformatics," *International Journal of Computer Applications*, vol. 179, no. 10, pp. 28-33, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[8] Cansu Demirkiran et al., "Leveraging Residue Number System for Designing High-Precision Analog Deep Neural Network Accelerators," *arXiv*, pp. 1-7, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[9] Akanni Gabriel, Eseyin Joseph, and Kazeem A. Gbolagade, "A Residue Number System and Secret Key Crypto System Review in Cyber Security," *International Journal of Innovative Science and Research Technology*, vol. 7, no. 10, pp. 1896-1901, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[10] Satpal Singh Bisht, and Amrita Kumari Panda, "DNA Sequencing: Methods and Applications," *Advances in Biotechnology*, pp. 11-23, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[11] Jerzy K. Kulski, "Next-Generation Sequencing — An Overview of the History, Tools, and 'Omic' Applications," *Next Generation Sequencing - Advances, Applications and Challenges*, pp. 3-60, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[12] Applied-Biosystems, "Principles of Di-Base Sequencing and the Advantages of Color Space Analysis in the SOLiD System," *Application Note*, pp. 2-5, 2011.

[13] Life Technologies, SOLiD Data - 2 Base Encoding, 2007. [Online]. Available: https://www.lanl.gov/conferences/finishfuture/pdfs/2007%20talks/SOLiD%20Data%20V1.0.8.pdf

[14] 2 Base Encoding – Wikipedia. [Online]. Available: https://en.wikipedia.org/wiki/2_base_encoding