Original Research Article

AN APPLICATION OF RESIDUE NUMBER SYSTEM (RNS) TO NEXT-GENERATION SEQUENCING (SOLiD)

ABSTRACT

Aims: This research work leverages the possibility and potential of an RNS-dibase table to generate the sequence primer and colour space for successful SOLiD sequencing. This design is flexible as compared with its binary counterpart and also presents a quaternary approach to SOLiD sequencing.

Study design: RNS sequence primer and colour space are generated resulting in a successful RNS-SOLiD Sequencing.

Methodology:

One of the most accurate Next Generation Sequencing (NGS) methods currently in use is Sequencing by Oligonucleotide Ligation and Detection (SOLiD). It combines ligation-base chemistry with a di-base labelled probe to produce an accuracy rate of about 99.9999%. RNS has the potential of generating the di-base table which is the Rosetta stone for SOLiD sequencing. Leveraging this possibility, the sequence primer and colour space which are requirements for a successful SOLiD sequencing are generated in RNS space. Following this, SOLiD sequencing is therefore designed using RNS.

Results: An RNS di-base table is presented and this serves as a look-up table for the generation of RNS sequence primer and colour space for successful SOLiD sequencing. A platform-independent algorithm is also developed that effectively illustrates SOLiD sequencing in RNS space.

Conclusion: This lays the groundwork for the incorporation of RNS into SOLiD sequencing. This design is flexible and buttresses the quest for a quaternary number system for molecular biological design and analysis.

Keywords: RNS, SOLiD, Sequence Primer, Colour Space, Single Nucleotide Polymorphism, di-base table, measurement errors, mismatched leading base

1.0 INTRODUCTION

The novel virus (COVID-19) had woken the world to the relevance of DNA sequencing and the need to improve on the tools that enhance sequencing processes. Finding the precise placement of nucleotides within a DNA molecule is known as DNA sequencing, and has become an indispensable part of modern molecular biology and applied fields such as diagnostics and forensics [5]. The Sanger technique also referred to as first-generation sequencing had lower throughput, consumed

more time, and had greater sequencing costs. The Next Generation Sequencing (NGS) technologies were created as a result of these constraints. They substantially reduce the sequencing time and cost and also improve the accuracy of reads. Some of these NGS techniques are SOLiD, Illumina, and Roche 454. Each has its advantages but SOLiD offers a completely drastic departure from the other next-generation sequencing technologies. SOLiD is unique in that the synthesis of DNA is driven by a ligase and the fluorophore is correlated to a dinucleotide and not a single base. Thus it is tugged as the only NGS system to employ ligation-based chemistry with a di-base probe. The digital realization of these technologies hinges on number systems. 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 [8]. It has been given some considerable research attention in fields like Digital Signal Processing (DSP), error control coding, and quite lately bioinformatics [11]. SOLiD sequencing is explained using the concept of Residue Number System (RNS) and an algorithm that further verifies the success of RNS-based SOLiD sequencing is developed. The RNS di-base table offers a successful sequence primer generation in the RNS space and a further colour space generation. And with a known first base, the sequence is generated. This work lays the groundwork for the incorporation of RNS into SOLiD sequencing techniques.

The remainder of this publication is structured as follows: sections two and three examine the fundamental knowledge of number system (binary) and RNS. Section four considers the genesis of DNA sequencing. And sections five and six looks at the overview of NGS and SOLiD sequencing respectively. Sections seven and eight describe the methodology and the design algorithms that support these methods. Section nine discusses the results of the model and section ten draws conclusions on the work.

2.0 NUMBER SYSTEM – BINARY

The primary number system humans are accustomed to is the decimal number system, but computers and other technological advancements fuelled the need for a more sophisticated number system, a binary number system [13]. This number system is used in every digital computing application. The discovery and application of binary number system to digital applications marked an exciting development in digital computing, electronics and telecommunications. Leibniz is credited for giving some meaning to the concept of binary number system in his paper titled “Explanation of Binary Arithmetic”, he supported his work with the I Cheng which dates from the 9th Century BC in China. Leibniz interpreted the hexagrams of the I Cheng as evidence of binary calculus [14]. The positional nature of this number system tends out to have some limiting effect on some digital applications' qualities like speed, energy consumption and error control coding. Residue Number System (RNS) has emerged as a number system that would curb these challenges of the famously known 0’s and 1’s number system. Lately RNS has seen some applications in molecular biology and bioinformatics [reference 2023] with its known attractive features.

3.0 RESIDUE NUMBER SYSTEM (RNS)

The Residue Number System (RNS), a non-positional integer number system, is often credited to Sun Tzu, it speeds up arithmetic computations by splitting them into smaller parts making each part independent of the other and executing them in parallel[15] [16]. RNS is more appealing for digital applications than traditional weighted number systems (binary) since it has carry-free arithmetic and lacks ordered significance. As high-speed signal processing applications become challenged due to some shortfalls of the conventional binary number system; RNS-based implementations are desired. RNS has been given considerable research attention in fields like DSP, error control coding, image processing, and quite lately bioinformatics. Among its good features are carry-free addition, borrow-free subtraction and single-step multiplications. It also promises a smaller digital footprint and generally low power consumption which are highly rated qualities for contemporary digital applications [17]. Data conversion is a challenge for RNS processors but these challenges have received considerable research attention lately due to the desire for the realisation of RNS digital applications. But one of the approaches to a successful RNS implementation is the conversion of operands from the weighted number system to the RNS otherwise known as forward conversion and conversely from the RNS to the weighted numbers system known as reverse conversion. The forward conversion process is much easier than the reverse conversion process. The two primary ways to reverse conversion present various difficulties; the CRT is intricate and laborious due to its massive modulo-M operation, whereas the MRC is a sequential procedure. However the CRT is more desirable because the conversion process can be paralleled [18]. Some other challenges of RNS are operations like division, scaling and magnitude comparison, these are slower and more complicated to implement. The use of RNS in bioinformatics (generating the di-base table for SOLiD sequencing) does not require these operations, hence these limitations do not affect the scope of this research.

4.0 THE GENESIS OF DNA SEQUENCING

Adenine, Guanine, Cytosine, and Thymine are the four nucleotide bases that make up the DNA molecule. DNA sequencing is a biochemical technique for establishing their order. [19]. Reverse genetics, which converts the amino acid sequence of

a gene of interest into a nucleotide sequence based on the proper codons, was the main technique used to get DNA sequences prior to the invention of the first sequencing methods [20][21]. Two methods were developed in the mid-1970s for directly sequencing DNA – the Maxam-Gilbert wandering spot analysis and the Sanger capillary electrophoreses methods. These became known as the first-generation sequencing technologies. Modern molecular biological processes, as well as applied disciplines like biotechnology, diagnostics, and forensics, all depend on DNA sequencing. Although two independent DNA sequencing techniques were developed about the same time, Sanger's chain-termination sequencing technique eventually replaced the Maxam-Gilbert technique as the preferred technique. Sanger's method was considered easier and that of Maxam-Gilbert was seen as dangerous because it used chemicals which were radioactive and toxic[19]. Sanger's sequencing was used in many sequencing projects including the Human Genome Project (HGP). The ability to sequence genomes opens new possibilities for biomedical applications and biological research. The relevance of the HGP on humans shifted the future of sequencing to making the sequencing process cost-effective and timely. For big undertakings like sequencing a full genome or metagenome, this technology was thought to be costly, ineffective and error-prone. The first generation (1G) sequencing was the standard for some decades, but the expense and length of the sequencing process proved a barrier to the advancement of genomics, ushering in the Next Generation Sequencing (NGS) technologies, one of which is SOLiD.

5.0 OVERVIEW – NEXT GENERATION SEQUENCING

Several Next Generation Sequencing algorithms make up for the challenges of the earlier known (first-generation) sequencing technologies. The greatest motivations are that these technologies offer economical sequencing, high fidelity, and greater sequencing throughput. These are achieved through the varied principles employed by each technology. NGS technologies have enhanced our fundamental biological knowledge, widened the scope of metagenomics analysis and enabled novel applications [12]. These technologies are each based on a specific method; the first commercially available sequencer, the Roche/454, is based on pyrosequencing. The Illumine (Solexa) genome analyzer is based on sequencing by synthesis whereas the SOLiD sequencer from Applied Biosystems is based on sequencing by ligation [22]. These principles account for the unique advantage of each technology over the other. The emphasis of this research is SOLiD, which among the NGS technologies uses DNA ligase and a distinctive method to sequence the amplified fragments. Enzymology, high-resolution optics, chemistry, software and hardware engineering work in concert to produce sequencing in most cases. The field of computers focuses on the software and hardware aspects of these systems – analysis, base-calling, assembler algorithms and software [23]. The cardinal steps for SOLiD sequencing are the di-base table, the sequence primer, and the colour space. All these steps are designed in the RNS space.

6.0 SOLID SEQUENCING

SOLiD sequencing invented by Applied Biosystems (Life Technologies), is distinctive and provides sequencing that radically departs from traditional sequencing technologies. It is the only NGS system to employ ligation-based chemistry with a di-base labelled probe [23]. SOLiD depicts a DNA fragment as a starting base, followed by a series of overlapping dimers (adjacent pairs of bases). Higher precision (up to 99.9999%), error-correcting, and differentiation between real polymorphism and measurement errors are made possible by the dimer overlapping properties and the structure of the colour code. This has the best raw accuracy among its peers or other commercially known NGS systems. It also allows researchers to focus on the biological significance of data rather than poor-quality data. In contrast to sequencing errors, which only impact one colour, single nucleotide polymorphisms alter two of the colours in the colour space (measured twice). Compared with others the SOLiD technology is unique in that the synthesis of DNA is not driven by a DNA polymerase, but a ligase and also the fluorophores are correlated to dinucleotide and not a single base. The sequencer uses two-base sequencing technology that is based on ligation sequencing. Each round of the five-cycle process involves multiple ligations between the 16 different di-base probes, each of which has four fluorescent dyes attached to it. A di-base colour coding scheme and sequence alignment are used to determine and cross-check the sequence [24]. The developed algorithms are used to demonstrate the feasibility of RNS-based SOLiD sequencing.

7.0 METHODOLOGY

SOLiD sequencing depends on the di-base table as the look-up table to construct the sequence primer leading to the generation of the colour space and with a known leading base, the sequence is determined. The RNS di-base in the figure below forms the dictionary for RNS SOLiD sequencing.

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

Figure 1: An RNS di-base table

Utilising particular 8-mer probes, the ligation is carried out, Figure 2 (a): These probes have a cleavage site between the fifth and sixth nucleotides and are eight bases long. The first two bases complement the sequence of nucleotides and also indicate the luminous dye. The first two bases assume residue digits, which reflect the luminous dye. Bases 3 through 5 which are degenerate, NNN, can couple with any nucleotide on the template sequence. Bases 6 through 8 are similarly degenerate, ZZZ, but as the reaction progresses, they are cleaved off along with the luminous dye as can be seen in Figure 2 (c). Further ligation is possible due to the fluorescent dye and bases 6–8 being cleaved. This is observed in Figure 2 (d).

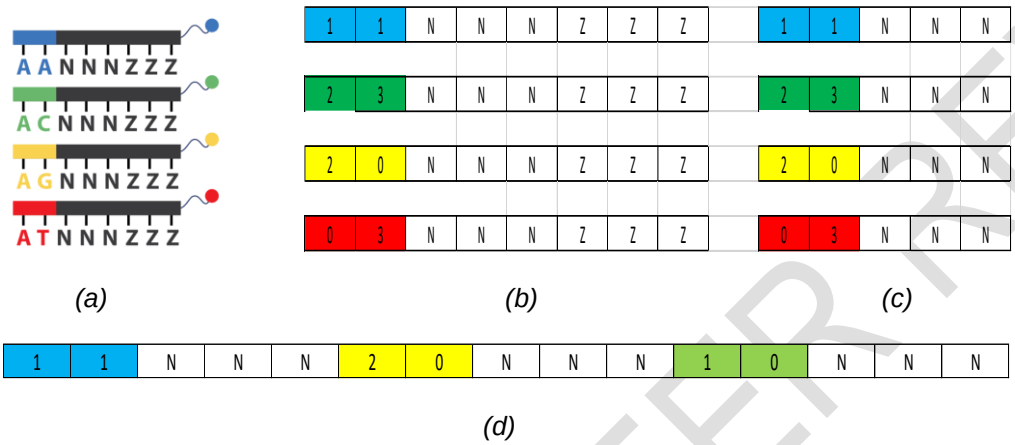

Figure 2: The octomer (8-mers), RNS representation and cleavage, and ligation

The basic structure of the sequencing stage is five rounds, with each round having five to seven cycles. A P1-complementary universal primer is added before each round. The sequence primer begins at base numbers 4 and 5 and this is represented by the moduli set [1,1] which has the fluorescence blue. Bases 6-8 are the three degenerate bases, NNN, which are not cleaved. Whereas the cleaved degenerate bases, ZZZ, are not shown on this diagram. The next is [2,3] which is green and followed by the three degenerate bases at base positions 11-13. The sequence process continues and completes at base positions 24-25 which has residue digits [1, 1] with fluorescence blue, Figure 3. The next sequence primer is offset by one (1) thus n-1. In this process, the leading bases in the previous (n) become the second bases for the current, (n-1). Thus, residue digit 1 at base column 4, residue digit 2 at base column 9, residue digit 2 at base column 14, and so on are repeated for n-1. The residue digit of the first base therefore determines the fluorescence for that di-base. In the first case for offset n-1, the first residue digit is 1 resulting in [1,1] representing the colour blue, the second is 1 representing the colour red and has the di-base [1,2], and this continues till the end of n-1 and the primer is offset by one, resulting n-2. The next which is n-2 has base columns 3, 7, 12, 18, and 23 of n-1 repeated for n-2. The second bases for n-2 are therefore 3, 1, 3, 0 and 0. Once all five (5) rounds of the sequence primers are generated, the colour space is generated. In generating the colour space, this begins at base columns 3-4, row n-1, and tracing up corresponds to the colour blue. Traversing the entire sequence, the next colour space is blue, which sits well for bases columns 4-5, row n. There is a colour space yellow for the next, which aligns well with base columns 5-6, row n-4. This process continues until all colours are determined.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
n					1	1				2	3				2	0				0	3				1	1
n-1				1	1				1	2				2	2				3	0				2	1	
n-2			3	1				1	1				3	2				0	3				0	2		
n-3		3	3				0	1				1	3				1	0				1	0			
n-4	0	3				2	0				1	1				0	1				2	1				
colour space																										

Figure 3: RNS-Sequence Primer and Colour Space Generation

The residue digits (data) are encoded as colours, which are then used to decode the sequence. Knowing just one base (residue digit) in the sequence allows for the decoding of the complete sequence. The reference sequence is transformed into an RNS colour space.

8.0 DESIGN FLOW – ALGORITHM

The algorithms sequence primer generator and colour space generator prescribe the process of generating the sequence primer using the RNS di-base table and subsequently generate the colour space for every sequence primer input. The algorithms to generate the sequence prime and the colour space can be seen below.

```
SequencePrimerGenerator(base1[], base2[])
Input: 2 arrays of nitrogenous bases
output: A sequence primer
1    sequenceprimer[n, m]
2    count <------ 0
3    startcolumnindex <------ 5
4    columnindex <------ startcolumnindex
5    for I <----- 0 to m-1
6        for j <----- columnindex to n-1
7            sequenceprimer[I, j] <----- base1[count]
8            if i = 0
9            sequenceprimer[i, j + 1] <----- base2[count]
10           else
11               sequenceprimer[I, j + 1] <----- sequenceprimer[I - 1, j + 1]
12               count <----- count +1
13               columnindex <----- columnindex + 5
14       startcolumnindex <----- startcolumnindex -1
15       columnindex <----- startcolumnindex
16   return sequenceprimer

ColorSpaceGenerator(sequenceprimer[])
Input: sequenceprimer
output: color space
1    count <----- 0
2    colorspace[]
3    for i <----- 3 to m - 1
4        for j <----- 0 to n - 1
5            baseA <----- sequenceprimer[j, i]
6            baseB <----- sequenceprimer[j, i + 1]
7            if baseA ≠ empty and baseB ≠ empty
8                colorspace[count] <-----GetBaseBaseColor(baseA, baseB)
9            count <----- count + 1
10   return colorspace
```

9.0 RESULTS AND DISCUSSION OF MODEL

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
n					1	1				2	3				2	0				0	3				1	1
n-1				1	1				1	2				2	2				3	0				2	1	
n-2			3	1				1	1				3	2				0	3				0	2		
n-3		3	3				0	1				1	3				1	0				1	0			
n-4	0	3				2	0				1	1				0	1				2	1				
colour space																										
SOLiD Sequencing (C)			1	1	1	3	2	2	1	0	0	2	3	3	1	0	1	2	1	2	1	0	2	1	1	
SOLiD Sequencing (T)			0	0	0	2	3	3	0	1	1	3	2	2	0	1	0	3	0	3	0	1	3	0	0	

Figure 4: RNS colour space, sequence primer and SOLiD sequencing.

The di-base table can be generated as RNS digits, this is dynamic compared with the rather static canonical table. The di-base table in RNS digits can be seen in Figure 1. In Figure 3 the colour space from the sequence primer input in the residue number system is generated. When any combination of di-bases is entered into the primer positions the colour is generated automatically from the di-base table as a look-up table and once the last sequence primer is entered the colour space is generated which is unique to the sequence primer entered in n, n-1,…, n-4. SOLiD sequencing is strongly associated with a known leading base or RNS digits in this case. The provision of a leading base – in this case, "C" or "T" (residue digit "1" or "0") generates the sequence which is unique to the colour space and the leading base provided, seen in Figure 3 rows 6 and 7.

10.0 CONCLUSION

The study establishes a foundation for integrating RNS with SOLiD sequencing. An RNS di-base table which is the Rosetta stone to SOLiD sequencing and the octamer probe are modelled in RNS space. The quaternary structure and dynamism of the RNS di-base table is a desirable property for molecular biological designs and analysis. This allowed for the successful generation of the sequence primer and colour space in RNS. The complete sequence is then deciphered once an RNS digit representing a known leading base is supplied. This completes the generation of a SOLiD sequence, which suits any set of moduli chosen.

REFERENCES

[1] *WHO, Genomic sequencing of SARS-CoV-2, no. January. 2021.*

[2] *T. Note, "Sequencing of SARS-CoV-2," no. January 2020.*

[3] *C. M. Morang'a et al., "Genetic diversity of SARS-CoV-2 infections in Ghana from 2020-2021," Nat. Commun., vol. 13, no. 1, pp. 1–11, 2022, doi: 10.1038/s41467-022-30219-5.*

[4] *"University of Ghana scientists sequence genomes of COVID-19 | News Ghana." https://newsghana.com.gh/university-of-ghana-scientists-sequence-genomes-of-covid-19/ (accessed Jan. 02, 2021).*

[5] *L. Cheng, T. Yu, T. Aittokallio, J. Corander, R. Khalitov, and Z. Yang, "Self-supervised learning for DNA sequences with circular dilated convolutional networks," pp. 1–10, 2023.*

[6] *B. Ostash and M. Anisimova, Visualizing Codon Usage Within and Across Genomes: Concepts and Tools, no. May. 2020.*

[7] *S. Schlebusch and N. Illing, "Next generation shotgun sequencing and the challenges of de novo genome assembly," S. Afr. J. Sci., vol. 108, no. 11–12, pp. 1–8, 2012, doi: 10.4102/sajs.v108i11/12.1256.*

[8] *C. Hong, B. Cao, C. Chang, S. Member, T. Srikanthan, and S. Member, "five - moduli set A Residue-to-Binary Converter for a New Five-Moduli Set," 2007.*

[9] *C. Demirkiran, R. Agrawal, V. J. Reddi, D. Bunandar, and A. Joshi, "Leveraging Residue Number System for Designing High-Precision Analog Deep Neural Network Accelerators," 2023, [Online]. Available: http://arxiv.org/abs/2306.09481.*

[10] *K. Givaki, A. Khonsari, M. H. Gholamrezaei, S. Gorgin, and M. H. Najafi, "A Generalized Residue Number System Design Approach for Ultra-Low Power Arithmetic Circuits Based on Deterministic Bit-streams," IEEE Trans. Comput. Des. Integr. Circuits Syst., pp. 1–14, 2023, doi: 10.1109/TCAD.2023.3250603.*

[11] *H. Kehinde Bello and K. Alagbe Gbolagade, "Acceleration of Biological Sequence Alignment Using Residue Number System," Asian J. Res. Comput. Sci., vol. 1, no. 2, pp. 1–10, 2018, doi: 10.9734/ajrcos/2018/v1i224735.*

[12] *H. Satam et al., "Next-Generation Sequencing Technology: Current Trends and Advancements," Biology (Basel)., vol. 12, no. 7, pp. 1–25, 2023, doi: 10.3390/biology12070997.*

[13] *M. I. Daabo, "Daabo Work," pp. 458–464, 2018.*

[14] L. Strickland, “Leibniz: Explanation of binary arithmetic (1703),” vol. GM VII, no. 1703, pp. 223–227, 2007, [Online]. Available: http://www.leibniz-translations.com/pdf/binary.pdf.
[15] A. P. Kari, “An e cient image cryptosystem based on the residue number system and hybrid chaotic maps,” pp. 0–22, 2023.
[16] K. O. B. and K. A. G. E.Y. Baagyere, “Bioinformatics: An Important Application Area of Residue Number System.” 2011.
[17] J. Liu, B. Liu, and H. Fu, “Optimizing Residue Number System on FPGA,” Proc. - 2016 IEEE Int. Conf. Internet Things; IEEE Green Comput. Commun. IEEE Cyber, Phys. Soc. Comput. IEEE Smart Data, iThings-GreenCom-CPSCom-Smart Data 2016, pp. 621–624, 2017, doi: 10.1109/iThings-GreenCom-CPSCom-SmartData.2016.137.
[18] N. Habibi and M. R. Salehnamadi, “An improved RNS reverse converter in three-moduli set,” J. Comput. Robot., v ol. 9, no. 2, pp. 27–32, 2016.
[19] S. S. Bisht and A. K. Panda, DNA sequencing: Methods and applications, vol. 9788132215. 2013.
[20] M. Kchouk, J. F. Gibrat, and M. Elloumi, “Generations of Sequencing Technologies: From First to Next Generation,” Biol. Med., vol. 09, no. 03, 2017, doi: 10.4172/0974-8369.1000395.
[21] NCBI), “The principles of DNA Sequencing.”
[22] Applied biosystems, “Applied Biosystems SOLiD™ 3 System Instrument Operation Guide,” pp. 1–8, 2016, [Online]. Available: http://tools.lifetechnologies.com/content/sfs/manuals/4407430b.pdf.
[23] C. Cheng, Z. Fei, and P. Xiao, “Methods to improve the accuracy of next-generation sequencing,” Front. Bioeng. Biotechnol., vol. 11, no. January, pp. 1–13, 2023, doi: 10.3389/fbioe.2023.982111.
[24] Applied-Biosystems, “Principles of Di-Base Sequencing and the Advantages of Color Space Analysis in the SOLiD System,” Appl. Note, pp. 2–5, 2011.

ABBREVIATIONS

RNS	Residue Number System
SOLiD	Sequencing by Oligonucleotide Ligation and Detection
NGS	Next Generation Sequencing
SNP	Single Nucleotide Polymorphism
DNA	Deoxyribonucleic Acid Ribonucleic Acid
RNA	Ribonucleic Acid
DSP	Digital Signal Processing
HGP	Human Genome Project
1G	First Generation Sequencing