SSRG International Journal of Pharmacy and Biomedical Engineering
ISSN: 2394 – 2576 / https://doi.org/10.14445/24547484/IJPBE-V12I1P101

Volume 12 Issue 1, 1-7, Jan-Apr 2025

Original Article

Modelling Matured mRNA for Gene Expression Using Residue Number System

Joshua Apigagua Akanbasiam[1], Kwame Osei Boateng[2], Daniel Kuyoli Ngala[3]

[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.*
[3]*Department of Telecommunications Engineering, Ghana Communication Technology University, Kumasi, Ghana.*

[1]*Corresponding Author : ja.akanbasiam@gmail.com*

Received: 09 January 2025 Revised: 18 February 2025 Accepted: 04 March 2025 Published: 15 March 2025

Abstract - *The recent computing power and developments in Artificial Intelligence (AI) and machine learning have shifted research attention to the relevance of ascribing digital and computational ability to bioinformatics. Prior to developing into a mature messenger RNA (mRNA) that controls protein synthesis, the first ribonucleic acid (RNA) transcribed from a gene's deoxyribonucleic acid (DNA) template in most eukaryotes and some prokaryotic species must undergo processing. While processing mRNA, certain non-coding sections, introns, are excised, and the final, known as exons or coding regions, remain and are spliced to form mature mRNA. Number systems form the foundation for digital applications and there has been no known digital or computational model that has focused on RNS' ability to model the matured messenger RNA train. The binary number system has served as the foundation for several digital and bioinformatics models. As there are only four (4) nitrogenous bases, applications based on a quaternary number system are desired for molecular biology. The excision of introns occurs at conserved sequences and the splicing of exons at exon junctions. The conserved sequences are modelled as sequences of RNS digits. In the lariat formation, the conserved sequence links up with a branch point to form a stop codon at the neck of the lariat. During splicing, the exon junctions of the preceding exons are spliced with the start base of the succeeding exon, and these are modelled as di-bases. This enables an easy and simple RNS modelling of intron excision and exon splicing processes. This technique is well-structured algorithmically within the RNS space, reinforcing experts' claim for a quaternary number system for molecular biological applications.*

Keywords - *Conserved sequences, Exons, Intron, Residue number system, Splicing.*

1. Introduction

Personalized medicine and altering genes and proteins are on the rise lately, and this has aided treatment and produced organisms with some high comparative advantage – natural selection. Altering genes and proteins could be spontaneous in an organism or deliberate in the laboratory. The biochemical processes that underlie RNA splicing, a precursor to protein synthesis, have been well examined and described. Modelling these processes to take advantage of mathematical and computational opportunities has relatively little work. Transcriptional and translational processes are the two primary steps in protein synthesis. Pre-mRNA and matured mRNA are produced from DNA through the process of transcription. Translation decodes a matured mRNA into amino acids and, subsequently, proteins. [1] Proteins are important constituents of eukaryotes, and modelling them requires understanding the mRNA train. The mRNA comprises numerous coding sequences, exons, that are differentiated from one another by non-coding areas termed introns. Introns are typically substantially longer than exons and are excised, whereas exons are spliced, forming the matured mRNA. [2] Exon splicing and intron excision are mostly dependent on two mechanisms. The first is that spliceosomes readily recognise conserved sequences in introns, and the second is that a branch point aids in producing a lariat (loop) and is removed during splicing [3]. Splice sites are conserved sequences where introns from main transcripts are snipped. [4] The codon and the di-base tables are the two dictionaries that aid in digitally modelling intron excision and exon splicing. A binary model is non-quaternary, does not span the entire number system space, lacks flexibility, and does not give a wide space for characterizing splicing. An RNS approach is considered in this research since the di-base and codon tables are quaternary and span the entire number system space. They are highly flexible since the codon and di-base tables depend on the moduli sets chosen. The cleaved RNA sequence typically begins at the 5′ end with the dinucleotide GU, or donor site, and ends at the 3′ end with the dinucleotide AG, or acceptor site. Sequences that begin with the dinucleotide AU and end with AC are uncommon

examples of alternate splice site sequences; their splicing mechanism is comparable to that of GU and AG. The branch point is home to another crucial sequence that catalysis the development of a lariat (loop) for intron excision. The branch point always contains an adenine (A) that is loosely conserved. These splice sites allow the intron-exon relation to be modelled as RNS di-bases. The conserved sequences are modelled as di-bases, and the neck of the lariat is modelled as a codon, specifically a stop codon, in both the dominant and the rare cases. Thus, this model's identification of conserved sequence is referenced to the di-base table and is therefore viewed as residue digits or di-base colours. In addition to being simple, the architecture is flexible and spans the entire number system space compared to the binary system. The process of exon splicing and intron excision results in a mature mRNA used as a template for protein synthesis, as shown in Figure 1.

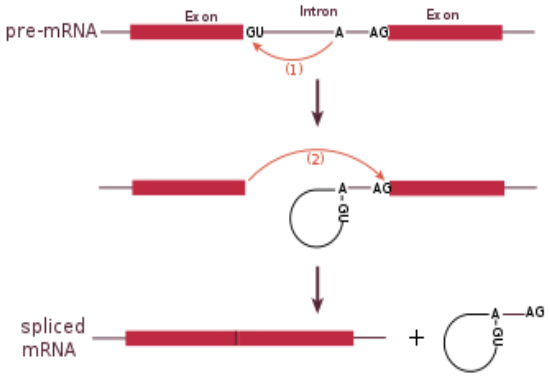


Fig. 1 Intron excision and Exon splicing [3]

2. Intron-Exons

The knowledge of RNA splicing is new, and scientists are beginning to discover its nature and role. However, it is known that introns and splicing have played significant roles in evolution. Every gene has many expressed parts, or coding sequences, called exons, and non-coding sequences, or intervening regions, called introns, between them. [3] Introns are a big biological mystery and were once—and for the most part, still are—considered "junk DNA." [5] They are present in eukaryotes but absent in prokaryotes. Although introns vary in length and type among genes, they are not involved in forming proteins. Recent research has been of interest since it has been revealed that they are significant and significantly impact gene expression and regulation. Exons are the regions that are expressed, which collectively make up mature mRNA and initiate translation. Exon reordering, duplications, and deletions increase variance and produce new gene variants. Plus, alternate splicing is possible with Exons. A single gene can encode several proteins due to the various ways that exons can be constructed. Computational and mathematical modelling of these important processes will leverage recent developments in machine learning and artificial intelligence (AI) to model various proteins for various applications. Most digital models have relied on the binary number system, which is static and not well suited for representing the four nitrogenous bases: Adenine (A), Guanine (G), Cytosine and Thymine (T). An RNS approach is presented in this research work that is flexible and well-suited for the call for a quaternary number system to represent the nitrogenous bases. An overview of RNS is provided in the following sections.

3. Fundamentals of RNS

The main characteristic of the residue number system, which is an integer number system, is that additions, subtraction, and multiplications are always carry-free. Unfortunately, certain procedures, such as comparison, division, and sign detection, are laborious and time-consuming. Also, residue number systems are not convenient when representing fractions. [6] The residue number system can be particularly appealing for special-purpose applications, such as a multitude of digital filters, where the number of adds and multiplications is significantly higher than invocations of magnitude comparison, overflow detection, division, and similar operations. As a consideration in molecular biological applications and bioinformatics, it suits the much-desired quaternary number system proposed by experts. Systems of residue numbers are founded on the congruence relation: When m divides the difference between c and d exactly, two integers, b and c, are said to be congruent modulo m. Put as follows: $b \equiv c \pmod{m}$. As a modulus or base, m excludes 1, which results in trivial congruence. [7] A residue number system is distinguished by a base consisting of a tuple of integers rather than a single radix. By definition, $b \equiv r \pmod{m}$ if r and q, respectively, are the remainder and quotient of the integer division of b by m, that is, $b = qm + r$.

The residue of b with respect to m is measured as r, which is typically represented by the notation $r = |b|m$. The set of least positive residues modulo m refers to the set of m values $\{0; 1; 2;...; m - 1\}$ that the residue may assume. This number system has found niche applications in other research areas and has received considerable research attention in other disciplines. Notable amongst these, in recent times, are bioinformatics and molecular biology. [8] It has been applied in generating the genetic code, representing the mRNA train, and the di-base table for SOLiD sequencing [9], [10]. This research looks at the application of RNS in modelling the matured mRNA from a pre-mRNA.

3. RNS DI-Base Table

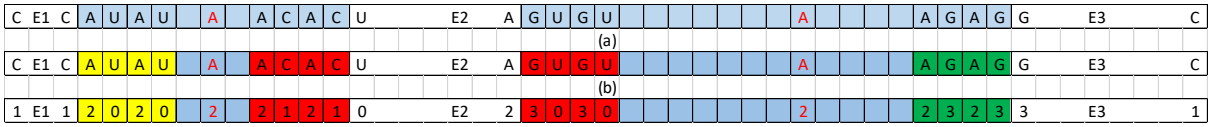
The primary purpose of the di-base table [11] is for Sequencing by Oligonucleotide Ligation and Detection (SOLiD) sequencing. This allowed SOLiD sequencing to outperform its competitors (other next-generation sequencing algorithms) in terms of accuracy and throughput. The Applied Biosystems Instruments (ABI) canonical di-base table lacks

flexibility since it is not well-structured within a number system framework. [12] An innovative approach has been designed, which involves introducing an RNS or RRNS di-base table for SOLiD sequencing. This RNS di-base table permits the creation of the genetic code and di-base table in a single step while preserving the attributes required for effective SOLiD sequence decoding. This makes it possible to create tools that combine the features dependent on the database table and the genetic code. Thus, the use of RNS in exon splicing and intron excision is explored in this study. The process of transcription entails changing DNA from pre-mRNA to mature mRNA. This is achieved through intron excision and exon splicing. Conserved sequences of introns are represented as di-bases from the RNS di-base table, facilitating the characterisation of splicing events. When compared to its sister system, binary digits, the binary combinations at each junction restrict the range of possible descriptions for these conserved sequences.

The exon junctions are viewed as a combination of di-bases and colour codes, referenced from the di-base table. RNS also uses important characteristics, like conserved intron sequences, to provide exon splicing and pre-mRNA excision to generate matured mRNA. Numerous coding regions called exons make up each gene, and non-coding elements called introns keep them apart. It is common practice to eliminate the amino acid AG at the 3' end of the RNA sequence and the dinucleotide GU at the 5' end. Other splicing sites are sometimes found at the dinucleotides AU and AC. Similar to a di-base table, its conserved sequence is considered in colour space or di-base. A lariat is formed when the di-nucleotides GU and AU, which are conserved sequences, combine with the branch point A. At the neck of the lariat, the stop codons UGA and UAA are generated. The RNS database and genetic code tables offer splicing and excision. This process is made easier by the di-base table's essential feature: the di-base and its reverse must have the same colour. The matured mRNA, the actual codon chain for protein synthesis, is generated by splicing the exons after the introns are eliminated. The next section discusses the procedures of exon splicing and intron excision.

4. Intron Excision and EXON Splicing

As stated earlier, transcription is not a direct process in eukaryotes; thus, mRNA undergoes further modifications after being synthesised. Primary transcript, sometimes referred to as pre-mRNA, is the result that comes after transcription. Post-transcription modification of mRNA produces matured mRNA, the actual gene template for protein synthesis. Shortening mRNA involves chopping off parts and joining the remaining segments back together before the mRNA leaves the nucleus. The altered mRNA that results from this procedure, called RNA splice, is referred to as matured mRNA. Exons are mRNA fragments that are re-spliced together after exiting the nucleus. Every gene is broken into a number of coding or expressed regions called exons, which are separated by intervening regions called introns or non-coding sequences. The highly conserved GU region sequence is found at the donor site at the 5' end of introns, while the highly conserved AG sequence is found at the acceptor site at the 3' end. Spliceosomes can identify the introns' beginning and ending locations. Splice sites, conserved sequences, are the locations of cleavage where introns are cut off parent transcripts. An additional significant sequence can be found at the "branch point," which is upstream from an intron's 3' end. The branch point is always an adenine (A), and because a di-base meets up with an adenine (A) at the branch point, it helps build a lariat (loop), whose neck is a codon. The process of intron excision begins with a break of the dinucleotide or conserved sequence at the 5’ end from the mRNA strand. The GU-conserved sequence bends in and joins with the branch point, Adenine (A). The resultant codon formed at the neck of the lariat is UGA, a stop codon. The lariat and the remaining part of the intron conserved sequence attached at the 3’ end are subsequently removed. The remaining exons are then spliced back, thus forming the matured mRNA template, which directs protein synthesis. Alternate splicing site sequences start with the dinucleotide AU and end with AC. The excision and splicing process is similar to the dominant case. The conserved sequences are viewed as strings of di-bases and, in this model, as moduli digits and colour spaces from the di-base table, as shown in Figures 2(a) to 2(c).



(c)

Fig. 2 A model of intron conserve sequences and branch point

5. Methodology

RNS also uses key features, such as conserved intron sequences, to facilitate exon splicing and pre-mRNA excision to produce matured mRNA. Each gene is made up of numerous non-coding segments called introns that are positioned between the numerous coding sequences called exons. The dinucleotide GU at the 5' end of the RNA sequence and the di-base AG at the 3' end are the main deletion targets. Occasionally, dinucleotides AU and AC are the locations of alternate splicing sites. Di-bases are conserved sequences found at intron junctions that can be resolved using the di-base table in the RNS space. GU and AG di-bases are, therefore, composed of the residue digits [3 0] and [2 3], while AU and AC are composed of the residue digits [2 0] and [2 1],

respectively. On the di-base table, the AU-conserved sequence is yellow, whereas the GU-conserved sequence is red. Figure 3 shows the GU conserved sequence's branch point and predominant case. When the branch nucleotide A is used to make the lariat, the conserved sequences GU and AU bend in and become UG and UA, respectively. This results in residue digits [0, 3] and [0, 2], where the original colours, red and yellow, respectively, are maintained. Figures 3 and 4 describe converting a GU-conserved sequence into an intron lariat. This is easier since a di-base and its reverse share a similar colour.

Therefore, the colour coding is conserved if the conserved sequences create a lariat by reversing the locations of their nucleotides. Another noteworthy characteristic is that during lariat formation, both UG and UA attach to the branch nucleotide A to generate stop codons UGA and UAA with RNS residue digits 032 and 022, respectively. Any combination of moduli will produce an identical RNS codon and di-base digits, making the design universal. The design's algorithm or computational practicality is examined in the next section.

6. Design Flow and Algorithm

The research explores the application of RNS in intron excision and exon splicing. RNS is generated as a tree of numbers, which presents the residue digits as blocks of numbers. These building elements make it simple to create the RNS genetic code table and the di-base table. This allows for tools that combine the functions reliant on the genetic code with those of the database table. Thus, these tables serve as the reference for modelling conserved sequences as di-bases and the neck of the lariat as codons. Conserved sequences of introns are represented as di-bases from the RNS di-base table, facilitating the characterization of splicing events. This serves as the basis for RNS intron excision and exon splicing. The result is a matured mRNA which directs the production of proteins. Figure 3 illustrates the design flow for an RNS intron excision and exon splicing.

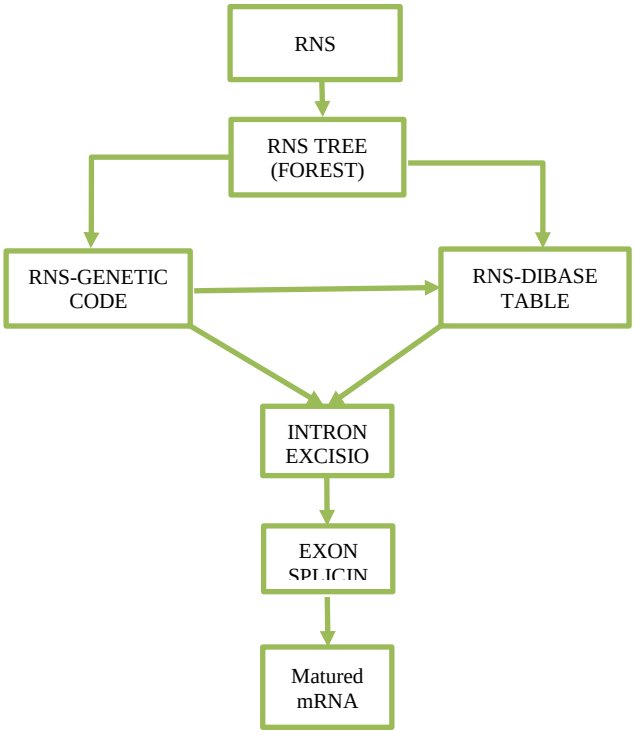


Fig. 3 Block diagram for RNS intron excision and exon splicing

6.1. Algorithm

1	**Algorithm: IntronExisionAndExonSplicing**
2	INPUT: 2 relatively prime numbers (m1, m2), pre_mRNA
3	OUTPUT: matured_mRNA
4	exon
5	exonCount <-- 0
6	exonStartPosition <-- 0
7	exonEndPosition
8	branchPoint
9	searchIntronStart <-- false;
10	searchIntronEnd <-- false;
11	searchBranchPoint <-- false;
12	DibaseTable <-- GenerateRNSDibaseTable
13	DibaseColorTable<-- GenerateDibaseColors(DibaseTable)
14	intronStartSequenceColor <-- DibaseColorTable[intronStartSequenceBase1, intronStartSequenceBase2]
15	intronEndSequenceColor <-- DibaseColorTable[intronEndSequenceBase1, intronEndSequenceBase2]

```
16  for i <-- 0 to length of pre_mRNA - 1
17      if searchIntronStart = true
18          baseA <-- pre_mRNA[i]
19          if introStartSequenceBase = baseA
20              if DibaseColourTable(baseA, pre_mRNA[i+1]) = intronStartSequenceColor
21                  baseB <-- pre_mRNA[i+2]
22                   if baseB = base1
23                  if DibaseColorTable[baseB, Pre_mRNA[i+3]] = intronStartSequenceColor
24                      i <-- i + 4
25                      exonEndPosition <--- i
26                      intronSearch <-- false
27                      searchBranchPoint <-- true;
28      if searchBranchPoint = true
29          if pre_mRNA[i] = branchPoint
30              //Form Lariat
31              searchBranchPoint <-- false
32              searchIntronEnd <-- true
33               i <-- i + 1
34      if searchIntronEnd = true
35          baseA <-- pre_mRNA[i]
36          if intronEndSequenceBase = baseA
37              if DibaseColorTable(baseA, pre_mRNA[i+1]) = intronEndSequenceColor
38                  baseB <-- pre_mRNA[i+2]
39                   if intronEndSequenceBase = baseB
40                      if DibaseColorTable(baseB,pre_mRNA[i+3])= intronEndSequenceColor
41                          searchIntronEnd <-- false;
42                          searchIntronStart <-- true;
43                             i <-- i + 4;
44                          //Complete Intron Exicion
45                          m <-- 0;
46                          for j <-- exonStartPosition to exonEndPosition
47                              tempExon[m] <-- pre_mRNA[exonStartPosition]
48                               m <-- m + 1
49                          exon[exonCount] <-- tempExon
50                          exonCount = exonCount + 1
51                          exonStartPosition <o-- i
52  //Exon Splicing
53  for n <-- 0 to exonCount
54      matured_mRNA[n] < exon[n]
55  return matured_mRNA
```

7. Results and Discussion

In conclusion, a basic pre-mRNA train is shown in Figure 4 as three exons (green) and two introns (sea blue). The graphic also shows the pre-mRNA train as residue digits in the second instance and nitrogenous bases in the first. The introns' start bases, known as conserved sequences, are shown as red in the dominant case and yellow in the rare case. The branch point in the introns, which is shown by the red print "A," is crucial for the development of lariats, leading to intron excision. The lariat formations that result in the intron's excision are explained in Figure 5. UAA or UGA, whose residue representations are 022 and 032 for the rare and dominant intron conserve sequences, respectively, are the consequence of the di-base bending-in to form a stop codon during the lariat formation process. Importantly, the actual protein production codon chain is the mature mRNA, which is produced by splicing the exons left behind after intron removal. This event is illustrated in Figure 6, which shows the splicing of Exons 1 (E1), 2 (E2), and 3 (E3).

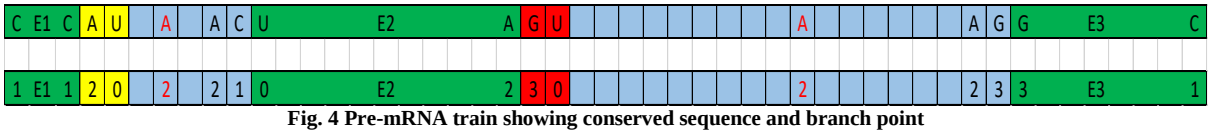


Fig. 4 Pre-mRNA train showing conserved sequence and branch point

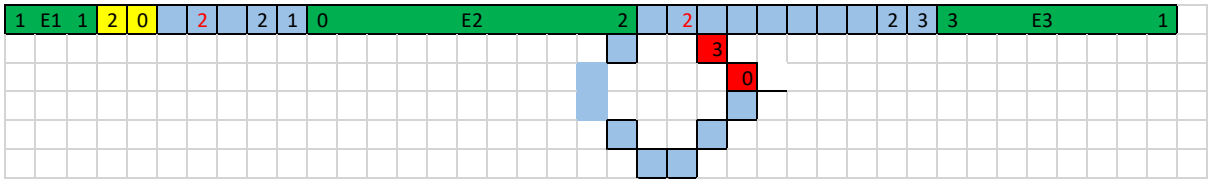


Fig. 5 Intron lariat construction and excision

Fig. 6 Exon Splicing E1, E2 and E3

8. Conclusion

This work advances our knowledge of the role of the Residue Number System in bioinformatics, specifically in relation to the di-base table and genetic code involvement in exon splicing and intron excision. Intron excision, followed by Exon splicing, transforms pre-mRNA into matured mRNA. A simpler approach has been demonstrated to classify conserved intron sequences as residue digits that must be removed. Furthermore, during lariat formation, the conserved di-base (two residue digits) sequence of the RNS bends into a link with the branch point; this joint creates a well-known stop codon (three residue digits) that can be referenced from the codon table. A logical progression from a two-moduli set RNS di-base table or colour coding scheme is an RNS-Genetic codon table with three modules. This model's ability to create the genetic code and the di-base table without ambiguity using any three (3) moduli set of RNS makes it widely suitable. With the exception of the decimal values, which will vary depending on the moduli sets selected, the moduli digits in the RNS tree will always be the same. As a result, the complete process of exon splicing and intron excision is simulated in RNS utilising the RNS-Genetic code and RNS di-base. Splicing and alternative splicing are the foundation of natural selection because genes and proteins are altered on a vast scale, either spontaneously or in laboratory settings, to create organisms with a highly competitive advantage. In the future, this model might be used for splice fault detection or alternative splicing.

Funding Statement

The research is self-funded by the authors

Acknowledgments

We acknowledge Emmanuel Akono Sarsah and Nabson Antuba Akanbasiam for their contributions in assisting in programming and algorithm design, respectively.

References

[1] Sávio Torres de Farias, and Francisco Prosdocimi, “RNP-World: The Ultimate Essence of Life is a Ribonucleoprotein Process,” *Genetics and Molecular Biology*, vol. 45, no. 3, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Amanda S. Solis, Nikki Shariat, and James G. Patton, “Splicing Fidelity, Enhancers, and Disease,” *Frontiers in Bioscience*, pp. 1926-1942, 2008. [CrossRef] [Google Scholar] [Publisher Link]

[3] RNA Splicing - Wikipedia, 2023. [Online]. Available: https://en.wikipedia.org/wiki/RNA_splicing

[4] Donny D. Licatalosi, “Intron Removal by the Spliceosome: A Solo Job or a Team Effort?,” *Molecular Cell*, vol. 81, no. 11, pp. 2275-2277, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[5] Bong-Seok Jo, and Sun Shim Choi, “Introns: The Functional Benefits of Introns in Genomes,” *Genomics Informatics*, vol. 13, no. 4, pp. 112-118, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[6] Harvey L. Garner, "The Residue Number System," *Proceedings of the Western Joint Computer Conference*, San Francisco, California, pp. 146-153, 1959. [CrossRef] [Google Scholar] [Publisher Link]

[7] Amos R. Omondi, and A. Benjamin Premkumar, *Residue Number System Theory and Implementation*, World Scientific Publishing Company, pp. 1-312, 2007. [Google Scholar] [Publisher Link]

[8] Hassan Kehinde Bello, and Kazeem Alagbe Gbolagade, "Acceleration of Biological Sequence Alignment Using Residue Number System," *Asian Journal of Research in Computer Science*, vol. 1, no. 2, pp. 1-10, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[9] Applied Biosystems SOLiD™ 3 System Instrument Operation Guide, Applied Biosystems, pp. 1-8, 2016. [Online]. Available: https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4407430b.pdf

[10] Tech Summary: ABI's SOLiD (Seq. by Oligo Ligation/Detection), UPDATED for v2.0 – SEQanswers, 2021. [Online]. Available: http://seqanswers.com/forums/showthread.php?t=10

[11] ABI Solid Sequencing - Wikipedia, 2021. [Online]. Available: https://en.wikipedia.org/wiki/ABI_Solid_Sequencing

[12] Pavla Hujová et al., "Nucleotides in Both Donor and Acceptor Splice Sites are Responsible for Choice in NAGNAG Tandem Splice Sites," *Cellular and Molecular Life Sciences*, vol. 78, no. 21-22, pp. 6979-6993, 2021. [CrossRef] [Google Scholar] [Publisher Link]