RESIDUE NUMBER SYSTEM (RNS) SEQUENCE COMPARISON AND MUTATION (ERROR) ANALYSIS

Ing. Joshua Apigagua Akanbasiam[1], Kwame Osei Boateng[2] and Matthew Glover[2] Addo

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

ABSTRACT

The genetic code may be generated using the Residue Number System (RNS). This achievement demonstrates the viability of protein synthesis. The genetic codes occupy RNS digits and enable sequence comparison as well as further analysis of biological flaws known as mutations. In this research work two mRNA reading frames (strings) are considered as strings of RNS digits. Triplets of these residue number digits constitutes an amino acids and chains of these triplets make up polypeptides, proteins. These digits are compared and single digits errors introduced, these errors mimic biological errors (mutations). The types of errors (mutations) that are manifested between the two reading frames, original and mutant, are detected and reported.

KEYWORDS

Residue Number System (RNS), DNA, RNA, mRNA, Error, Mutation, Sequence comparison, RNS-Genetic code.

1. INTRODUCTION

The human system or biological systems are not spared of errors, mutations, and their inherent catastrophic nature just as digital systems. It has been determined that 1 in 10,000 bases are incorrect, the rate of mutation is significantly improved to 1 in 1,000,000 bases after proofreading[1].Hence, the need to better understand these errors by exploring mathematical and computerized approach in sensing and adapting to them.Natural selection is based on the idea that while mutations might lead to some diseases, like cancer, they can also improve a species' acclimatization to its environment[2]. Understanding life requires the understanding of proteins. They are the most prevalent organic molecules in living things and the most significant biochemical active ingredients are proteins; virtually none of the metabolic processes necessary for life would occur without it[3]. The gene responsible for producing each protein determines its specific sequence. Amino acids are the monomers (or basic units) of a protein that are arranged in a linear sequence[4]. Genomics and bioinformatics are revolutionizing health care systems through personalized and customized medicine. Unearthing a fast and better genomi sequencing approach is much desired hence the scouting of approaches that will allow quick sequencing of patient's genome and easy detection of potential harmful mutations[5, p. 25]. This will aid in early diagnosis and effective treatment of diseases. There have been intense works on the development of comparison methods to compare protein structure. The most important findings of proteins are not from theories but the comparison of sequences and observation of structures. The fundamental goal of comparison methods is to compute some quantitative

measure of similarities or difference and compare protein of an unknown sequence structure and that of a known structure. The binary number system is the traditional method used in implementing these comparisons and computations because it is the most widely used system in modern computing [6]. This research work seeks to generate sequence of amino acids (proteins) as triplets of residue number system digits and compare these strings of amino acid chains to extrapolate findings on similarities and differences (mutation) [7]. The RNS-genetic code is the Rosetta stone that aid in decoding these strings of RNS-digits.

Residue Number System was rediscovered in the 1950's and interesting findings were made to support its implementation in fast arithmetic and fault tolerant computing. Researchers later focused on division algorithms, sign representation and efficient converters, which are some of the major hindrance to the realization of RNS-based application processors[8]. Some progress has been made and lately the interest have been on the application of RNS to other fields of study, notably; Bioinformatics, Forensics, Fast Fourier transforms and error control coding[9], [10][11], [12]. The mRNA strands can be represented as a string of residue number digits and triplets of these digits constitute amino acids, chains of these amino acids constitute proteins. This presents an easier number system, quaternary, alternative to comparing strings of mRNA frames or proteins and offer a better digital architecture. If the nucleotide sequence of the gene's coding region changes, new amino acids might be added to the expanding polypeptide chain, changing the structure and function of the protein [5]. The amino acid sequence of the protein encoded by a gene may change when a mutation occurs in a protein-coding region. The genetic code is degenerate, thus not every DNA mutation significantly alters the sequence of proteins. Synonymous substitution is the name given to a mutation that does not change an amino acid. while a non-synonymous substitution is one that actually results in an amino acid change[13], Nonsynonymous changes can also be categorised as nonsense mutations, which happen when a stop codon is inserted in the middle of the sequence, and missense mutations, which happen when one amino acid is changed by another[14].

2. DEOXYRIBONUCLEIC ACID (DNA) AND RIBONUCLEIC ACID (RNA)

DNA is a molecule that houses the genetic instructions necessary for all known living species to grow, develop, function, and reproduce. [15]**.** One of the two antiparallel strands of DNA runs from sugars at position 5 to sugars at position 3 and vice versa. DNA essentially serves as a database for biological information; when the two strands separate, this data is duplicated. Smaller monomer units called nucleotides make up the two strands. The four nucleobases that contain nitrogen and make up each nucleotide are cytosine (C), guanine (G), adenine (A), and thymine (T)[16]. Deoxyribose, a kind of sugar, and a phosphate group make up these nucleobases. The parental molecules' two strands separate (replicate) to form RNA, with each acting as a template for a fresh complementary strand [17]. The two main differences between RNA and DNA are that the former is single-stranded and the latter employs uracil (U) rather than thymine (T) as its base. There are three types of RNA involved in protein synthesis: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA).An RNA strand that is translated from a protein-coding region of DNA is known as a messenger RNA. mRNA acts as a blueprint for the production of proteins during translation. The basic base-pairs goes with the Chargaff's rule[18], [19].Mutagens can cause DNA damage, or the damage might occur naturally as a result of regular biological processes that alter the DNA sequence. Some of these damages can be rectified, while others may persist despite remedial efforts. These mutations that occur can cause abnormal replication of cells – cancer or can produce organisms with highly adaptable features – the drive for natural selection.

3. THE RNS-GENETIC CODE

The RNS-Genetic code is designed with the concept of RNS and number trees. When a forest of RNS is constructed with appropriate roots, nodes, and leaves – the design uniquely categorizes the residue number system digits into blocks of data that is exploited to generate the genetic code. If the moduli sets of RNS are chosen such that there are at least three (3) moduli sets - m_1, m_2 and m_3 and their dynamic range greater than or equal to 64 thus $m_1 \times m_2 \times m_3 \geq 64$ and each of m_1, m_2 and m_3 greater than or equal to four (4). A forest or blocks of residue numbers can be constructed such that with appropriate truncation and digit-base assignment; the genetic code or amino acid table can be constructed from the block (tree) of residue digits. The number of roots of the tree is determined by the first moduli set m_1, the number of nodes representing the second moduli set, m_2 and the leaf nodes conclude the codon formation as the third moduli set m_3. Since there are four nitrogenous bases each modulus is truncated to 4 digits, thus 0, 1, 2 and 3. The RNS-Genetic code table is completed with each residue digits assigned appropriate nitrogenous bases –thus Thymine (T) is assigned residue digit zero (0), Cytosine (C) to residue digit one (1), Adenine (A) is consigned residue digit two (2) and Guanine (G) to residue digit three (3). In the case of an RNA-genetic code Thymine (T) is replaced with Uracil (U) and assigned residue digit zero (0).

Table 1: The RNS based genetic based genetic code

RNS –GENETIC CODE								
0 0 0	0 1 0	0 2 0	0 3 0		UUU	UCU	UAU	UGU
0 0 1	0 1 1	0 2 1	0 3 1		UUC	UCC	UAC	UGC
0 0 2	0 1 2	0 2 2	0 3 2		UUA	UCA	UAA	UGA
0 0 3	0 1 3	0 2 3	0 3 3		UUG	UCG	UAG	UGG
1 0 0	1 1 0	1 2 0	1 3 0		CUU	CCU	CAU	CGU
1 0 1	1 1 1	1 2 1	1 3 1		CUC	CCC	CAC	CGC
1 0 2	1 1 2	1 2 2	1 3 2		CUA	CCA	CAA	CGA
1 0 3	1 1 3	1 2 3	1 3 3		CUG	CCG	CAG	CGG
2 0 0	2 1 0	2 2 0	2 3 0		AUU	ACU	AAU	AGU
2 0 1	2 1 1	2 2 1	2 3 1		AUC	ACC	AAC	AGC
2 0 2	2 1 2	2 2 2	2 3 2		AUA	ACA	AAA	AGA
2 0 3	2 1 3	2 2 3	2 3 3		AUG	ACG	AAG	AGG
3 0 0	3 1 0	3 2 0	3 3 0		GUU	GCU	GAU	GGU
3 0 1	3 1 1	3 2 1	3 3 1		GUC	GCC	GAC	GGC
3 0 2	3 1 2	3 2 2	3 3 2		GUA	GCA	GAA	GGA

4. AMINO ACIDS AND PROTEINS

Proteins' basic units are amino acids. They join to create polypeptides, which are longer chains polymer, or peptides, which are shorter chains. [20]. These polymers are linear and unbranched since each amino acid in the chain is connected to the two amino acids immediately adjacent to it.In order to create proteins, a ribosome must add one amino acid at a time to a building protein chain. This process is known as translation[21]. Reading the genetic code from the mRNA template, which is an RNA copy of one of the organism's genes, determines the order of amino acids that are added.Just twenty (20) of the twenty-two (22) proteinogenic or natural amino acids—which make up the monomer units of proteins—are encoded by the universal genetic code. These amino acids are spontaneously incorporated into polypeptides. When the carboxyl group of one amino acid is joined with the amino group of another amino acid, proteins are

produced. The nucleotides in the gene that codes for a particular protein determine the precise amino acid sequence and makeup of that protein. How the amino acids in a protein chain are arranged is determined by an appropriate sequence of genetic triplets. Three adjacent nitrogenous bases that are arranged in a block along a DNA or RNA filament are known as a triplet (codon).Every protein contains a specific triplet sequence that codes for the amino acid sequence (a 3n-multiplet is a term used to describe such a sequence of n triplets).An example of the sequence of amino acid sequence is shown in figure 1

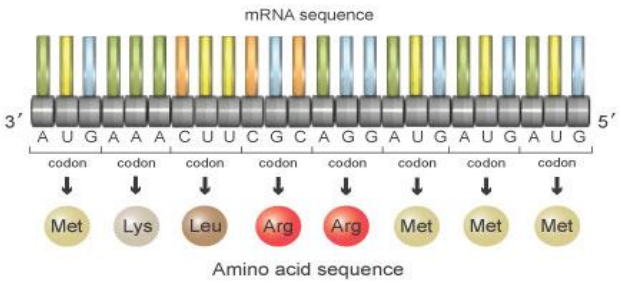

Figure 1: Amino-acid sequence [22]

5. PROTEIN SYNTHESIS

The DNA nucleotide sequence, which is particular to each person, contains instructions for the synthesis of every protein. The two processes of synthesising proteins are transcription and translation [23]. The DNA provides instructions for the formation of a messenger RNA (mRNA) molecule during transcription, and the mRNA also provides instructions for the formation of a polypeptide chain during translation. Triplets of nucleotides are arranged in long sequences called genes. For each of the 20 amino acids used to form proteins, a different arrangement or DNA triplet is used to encode it. An mRNA codon is created during transcription by a DNA triplet, and an amino acid is created during translation by a codon. When RNA polymerase connects to a gene's promoter region, it begins to make mRNA molecules until it encounters a termination sequence or "mark" in the gene.RNA is synthesized according to a base paring rule – Chargaff's rule[19]. A representation of DNA base and its RNS representation based on the moduli sets can be seen in Figure 2.

DNA	A	T	G	C	A	G	T	T	C	A	C	T	C	T	A	T	T	G	G	A	A	C	T	G	T	T	C	T	C	T	G	G	A	T	C	G
moduli	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3
moduli_digits	2	0	3	1	2	3	0	0	1	2	1	0	1	0	2	0	0	3	3	2	2	1	0	3	0	0	1	0	1	0	3	3	2	0	1	3
TRANSCRIPTION																																				
mRNA	A	U	G	C	A	G	U	U	C	A	C	U	C	U	A	U	U	G	G	A	A	C	U	G	U	U	C	U	C	U	G	G	A	U	C	G
moduli	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3	m_1	m_2	m_3
moduli_digits	2	0	3	1	2	3	0	0	1	2	1	0	1	0	2	0	0	3	3	2	2	1	0	3	0	0	1	0	1	0	3	3	2	0	1	3
TRANSLATION																																				
Amino_Acid	Met			Glu			Phe			Thr			Lue			Lue			Glut			Lue			Phe			Ser			Gly			Ser		

Figure 2: Residue Number System representation of the Central Dogma of Molecular Biology

The translation process is initiated with a start codon, (AUG) also known as methionine. As more amino acids are added to the expanding chain, the polypeptide chain expands. The mRNA molecule is terminated by a stop codon. Methionine will always be represented by residue digits [2 0 3] irrespective of the moduli sets chosen for the genetic code and this applies to all residue digits representing each amino acid. The growing string of polypeptides is represented as a growing string of residue digits, and these digits are the same for any of the amino acids irrespective of the moduli set chosen. The stop codons [TAA, TAG and TGA] that signal for the

end of a protein formation also have unique residue representations as [022, 023, and 032] respectively. These string of residue digits representing a protein can be resolved into its constituent amino-acids considering triplet of digits and the RNS-Genetic code as a look-up table. Also this can allow for different strings of proteins to be compared and spots of variance, mutations, and the type of variations identified.

6. DNA SEQUENCING

DNA sequencing is the process of determining the exact placement of nucleotides within a DNA molecule. Adenine (A), Guanine (G), Cytosine (C), and Thymine (T) are the four bases, and any technique or technology for figuring out their order is covered (T)[24]. Research and discoveries in biological and medical fields have greatly advanced, thanks to rapid DNA sequencing techniques [25]. There have been intense works on the development of comparison methods to compare protein structure. The most important findings of proteins are not from theories but the comparison of sequences and observation of structures. Hence the intense interest in the development of comparison methods to compare protein structures. This is pushing for more novel approaches to aid in discovering fundamental similarities among proteins and the constancy and adaptability of these proteins structure under greater degree of deviations. The fundamental goal of any comparison method is to compute some quantitative measure of similarities or difference and compare protein of an unknown sequence structure and that of a known structure. Since it is the system that is utilized the most in contemporary computing, the majority of these tasks (comparisons and computations) rely on the binary number system. This research work therefore propose an RNS approach to sequence comparison and the detection of sequence errors (mutations) in a polypeptide or protein chain. The sequence of amino acids (proteins) are generated as triplets of residue number system digits, this is further used to compare these strings of amino acid chains to extrapolate findings on similarities and differences (mutation)[26].

7. ERRORS IN LIVING ORGANISMS – MUTATIONS

Basically, a gene mutation is a change in the DNA's nucleotide sequence.[2].A single nucleotide pair or larger gene regions on a chromosome may be altered by this change. Mutations change the genetic code, leading to genetic variety and the possibility of disease or resistance, such as the production of antibodies [7]. Mutations are brought on by incorrect DNA replication or by mutagens. Silent mutations are those that have no effect; missense mutations and nonsense mutations, on the other hand, can range from having little to no effect to having a significant impact. There are two types of gene alterations that can take place: base-pair insertions or deletions, indels, and point mutations. Silent, missense, and nonsense mutations are the three types of point mutations.[27].A transition is a mutation that happens between two purines or pyrimidines, whereas a transversion is a mutation that happens when a purine is transformed into a pyrimidine or vice versa[28].Synonymous substitution is a type of mutation that does not result in an amino acid change and is categorized as a silent mutation. Nonsynonymous substitutions modify the amino acid and are classified as missense mutations (in which one amino acid is swapped out for another) or nonsense mutations (in which a stop codon is added into the middle of the sequence)[29]. Aside substitutions; insertions and deletions can also occur-indels. Thus base pairs are added or deleted from the original gene sequence, resulting in indels. These mutations are hazardous because they change the template (mRNA) that the amino acids are read from. As a result, a stop codon may be coded for either too early or too late in the translation process. The resulting proteins will either be rather short or very lengthy, and they will almost always be defective.

8. EFFECTS OF MUTATIONS

The genetic differences we observe between people are caused by modifications to the DNA's nucleotide sequence. Genetic changes at the molecular level can be linked to several kinds of alterations. Mutations can range from having no effect through to having small effect or a large effect. Sickle cell anaemia, an inherited condition that damages red blood cells, is connected to a single amino acid alteration in humans. [30]. The sequence of one of the polypeptide chains that makes up hemoglobin, the protein that carries oxygen in the blood, has been slightly altered. As a result, glutamic acid, which is typically the chain's sixth amino acid, has been replaced with valine. The substitution of valine for glutamic acid in the amino acid chain results in the formation of long filaments of haemoglobin, which transforms the normally disc-shaped red blood cells into abnormal crescent shapes. Hence, only 2 amino acids separate a normal haemoglobin molecule from a sickle cell molecule out of the approximately 600 amino acids. The types of chloroplasts within leaf cells can explain the phenotypic of leaves in plants[31]. The presence of normal chloroplast, which produces green pigment, accounts for the green phenotype. A gene in the chloroplast DNA that controls the synthesis of green pigment is mutated, leading to the white phenotype.

9. RESIDUE NUMBER SYSTEM(RNS)

Recently, a lot of scientific interest has been focused on the peculiar non-weighted number system known as the Residue Number System. The main goal of contemporary research in this topic is to design new and more efficient RNS-based building blocks that increase the performance of digital signal processing (DSP) applications. The fundamental idea behind the RNS is to uniquely represent big integers using a collection of smaller residues, leading to carry-free, fast, and parallel arithmetic [6]. This system uses the modulus operation as its foundation, with modulo serving as the divider and residue serving as the final product of the division process[8]**.** The research community is quite interested in RNS, and lately one of the key areas of study has been the use of RNS in other disciplines to increase processing speed and decrease the digital implementation footprint.RNS has found niche applications in bioinformatics, error detection, error correction, and fault tolerance and some arithmetic-intensive digital signal processing (DSP). Since the nature of biological errors mimics those of digital errors we seek to apply RNS in biological error (mutation) detection and correction by comparing polypeptide chains. Some work have been done on the application of RNS to Smith Waterman algorithm or accelerator [10]. Similar to how the index of a binary number system determines the range of data representation and the number of digits that may be expressed, the number of moduli sets in the case of RNS does the same.$2^3 = 8$; Thus the number of bits that can represent this in the binary number system is three (3) and the range of digits is 8;

$2^n => 2^3 = 8$
$hence\ 0 \rightarrow 2^n - 1;\ 0 \rightarrow 7$
$thus\ 0 \rightarrow 111$

In the case of RNS, given that the moduli sets (m_1, m_2 and m_3) are (2, 3, and 5) respectively then the number of digits that can be represented are;

$(m_1 \times m_2 \times m_3) = M; => (2 \times 3 \times 5) = 30;$
$hence\ 0 \rightarrow M - 1; 0 \rightarrow 29;$
$000 \rightarrow (m_1 - 1, m_2 - 1, m_3 - 1)$
$thus\ 000 \rightarrow 124.$

RNS is a concatenation of digits from all moduli sets represented – thus for two (2) moduli sets; m_1 and m_2, two (2) digits are required and for three (3) moduli sets; m_1, m_2 and m_3, three (3) digits are required and for four (4) moduli sets; m_1, m_2, m_3 and m_4, four digit concatenations is required and so on. Since DNA has four (4) bases we chose a quaternary RNS system where higher modulus (> 3) of each moduli concatenation is truncated. Also an amino-acid is made of triplets of bases and as such a three (3) moduli set system is chosen – so that each three (3) digit concatenation is an amino-acid, this can be decoded by the RNS-genetic code.

10. RNS DATA CONVERSION

One of the major issues in efficient RNS systems design is data conversion. The main types of conversion are; forward conversion and reverse conversion. Forward conversion is the transformation of a weighted binary number to the residue representation, and reverse conversion is the transformation of a residue represented number to its corresponding weighted binary number. The reverse conversion is the most challenging part of most architectures[32]. Reverse conversion is accomplished using the Chinese Remainder Theorem (CRT), Mixed-Radix Conversion (MRC),the New Chinese Remainder Theorems (New CRTs) and other methods[33][34]. Both the CRT and MRC have their issues; the CRT requires huge modulo M operations (where M is the dynamic range of the number system), whereas the MRC is slow due to its sequential nature. If RNS conversion process can be fast, then RNS would equally compete as one of the number systems of choice for many modern day digital applications.

11. METHODOLOGY

The main objective is to successfully compare two polypeptides or proteins and report on the similarity or otherwise between the two strings. The successful generation of the genetic code with RNS, RNS-genetic code, is the Rosetta stone for the successfully resolution of residue number digits to amino-acids and then proteins. A three(3) moduli RNS is chosen such that a concatenation of the three (3) digits of each modulus constitutes an amino-acid (codon). The growing chain of sequence of bases that make up a polypeptide or protein is seen as a growing chain of RNS digits. The RNS-genetic code generated, serves as Rosetta stone to resolve each codon (triplet of bases) to its RNS-digits and further convert these RNS-digits to their decimal values. Each codon has a unique decimal value based on the moduli sets chosen but its RNS-digits are the same irrespective of the moduli set chosen. Methionine has RNS-digits [2 0 3] irrespective of the moduli sets chosen. With this feature, two or more polypeptides or chains of amino acids (proteins) can be compared. A comparison of the two strings, reference string and compared string of polypeptide or proteins allows for various analysis – the types of errors (mutations) in the compared string can be observed and reported. Suppose we seek to perform a pairwise comparison of polypeptide or protein X and Y, where X is the reference string and Y is the compared string. The strings X and Y can be presented as a string of RNS digits where every set of three (3) digits in the string represent a codon or an amino acid. Each triplet (codon) has a unique decimal value based on the moduli sets chosen and the same RNS-digits irrespective of the moduli sets chosen. An erroneous codon in the compared string (Y) results in a difference in the bases (residue-digits) and the decimal value at this location. Since each triplet in the chain corresponds to one and only one particular amino-acid in the RNS-genetic code. The similarity point produces a zero (0) syndrome when the two strings are compared and the erroneous point(s) produce a non-zero syndrome. The non-zero values signifies a variance in the amino-acid at this location and hence an error (mutation). The algorithm designed detects the type of error (mutation) as silent, missense or nonsense and reports it as such.

11.1. Design Flow – Algorithm

```
CompareAminoAcid(reference, compared)
Input: 2 aminoacid strings ( reference string and compared)
Output: Error indicator
1  error_0 ← No Error
2  error_1 ← Deletion Error
3  error_2 ← Insertion Error
4  error_3 ← Misense Error
5  error_4 ← Silence Error
6  error_5 ← Nonsense Error
7  stopCodon ← UUA,UGA,UAG
8  referenceLength ← length of the Reference Protein
9  comparedLength ← length of the Compared Protein
10 codonCount ← referenceLength / 3
11 if ( comparedLength<referenceLength)
12     errorCode ← error_1
13 if (comparedLength>referenceLength)
14     errorCode ← error_2
15 if (comparedLength = referenceLength)
16     count ← 0
17     for i ← 0 to codonCount - 1
18         referenceCodon ← Substring(reference, count, 3)
19         comparedCodon ← Substring(compared, count, 3)
20         referenceCodonValue ← CodonValue(referenceCodon)
21         comparedCodonValue ← CodonValue(comparedCodon)
22         if (comparedCodon = stopCodon)
23             if (count < (comparedLenth - 3))
24                 errorCode ← error_5
25         if ( comparedCodonValue !=
           referenceCodonValue)
26             comparedCodonName ← CodonName(comparedCodon)
27             referenceCodonName ← CodonName(referenceCodon)
28             if (comparedCodonName != referenceCodonName)
29                 errorCode ← error_3
30             if (comparedCodonName = referenceCodonName)
31                 errorCode ← error_4
32         if (comparedCodonValue = referenceCodonValue)
33             errorCode ← error_0
34         count ← count + 1
35 return errorCode
```

12. RESULTS AND DISCUSSION OF RESULTS

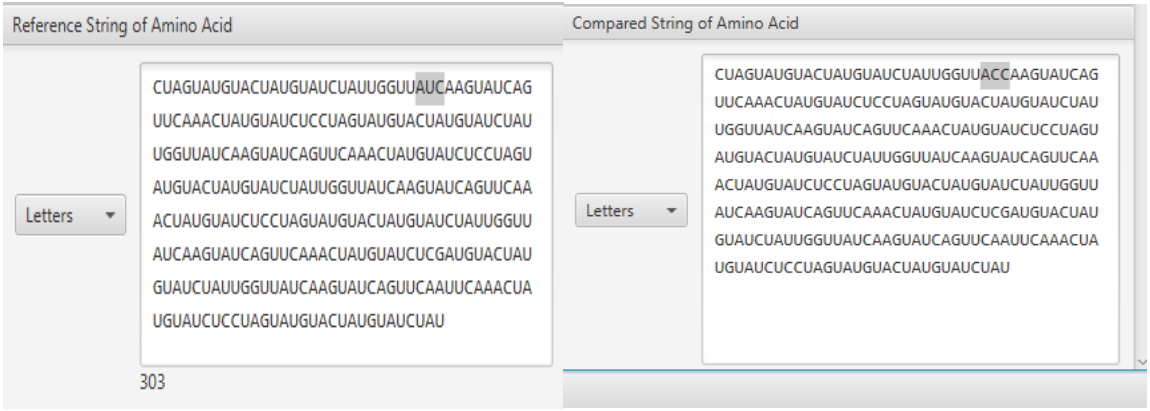

Figure 3: Reference and Compared Sequence of Amino Acids – RNA

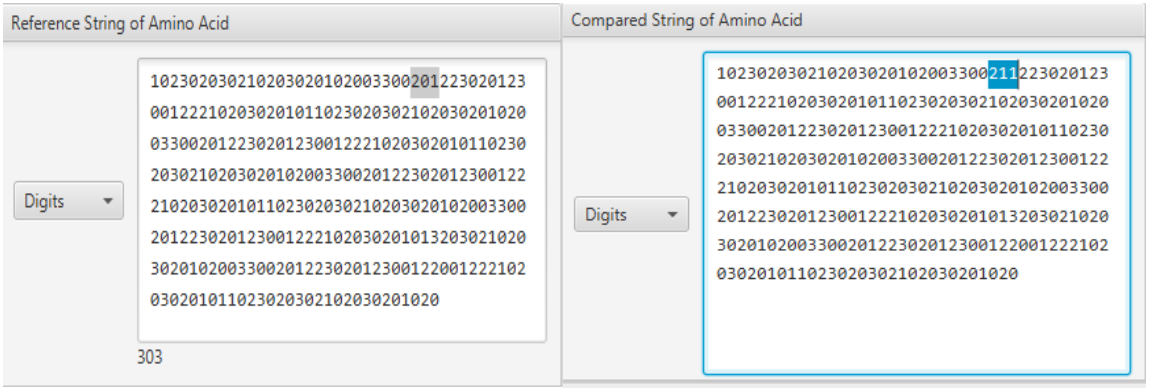

Figure 4: Reference and Compared Sequence of Amino Acids – RNS

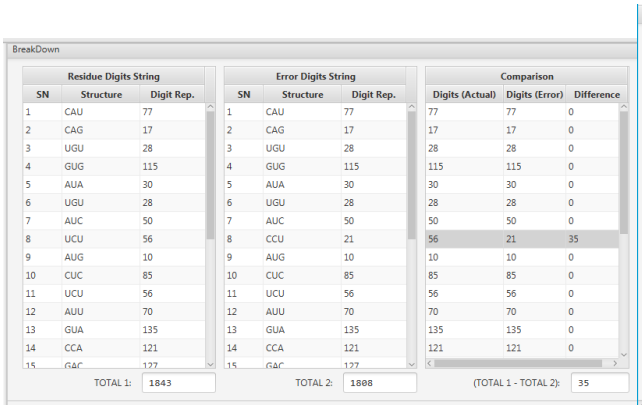

Figure 5: Error Syndrome Detection

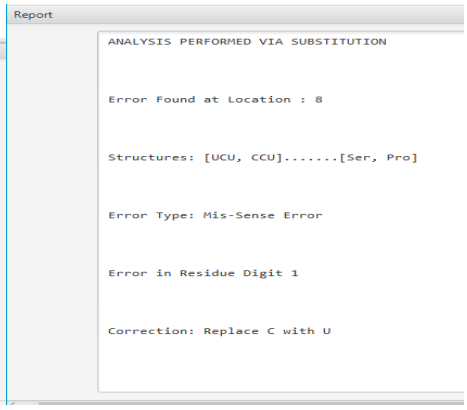

Figure 6: Type and Location of mutation

The Java software developed allow for generation of the genetic code and pairwise comparison of two polypeptides. The protein chain can be entered in RNS digits – Figure 4. The RNS digits protein chain can further be converted to the known DNA or RNA bases as seen in Figures 3. The reference string and the compared strings are analysed and the syndrome determined. The codon or set of residue digits have unique decimal values based on the moduli set chosen, these values assist in determining the syndrome at each compared location. If the syndrome is a non-zero then

an error is recorded and if it's a zero, then there is no error. The error location is highlighted in the entries of the reference and compared strings and also the comparison (syndrome detection) form as seen in figures 4 and 5. The type and location of the errors is presented in figure 6. The errors (mutations) reported in this work are those of substitution, deletion and insertion.

13. OBSERVATIONS

The reading frame or mRNA chain in figures 3, and 4 are presented in the form of DNA, or RNA bases and as RNS digits respectively. A summary of how the syndrome is calculated and how errors in RNS representation of the reading frame (mRNA) is shown in figures 5. The error location is highlighted indicating the non-zero syndrome value(s). Figure 6 gives a summary of the error analysis performed and the type of error detected. In this case a substitution error was observed, and as a result a serine (ser) with triplets UCU has become a proline (pro) with triplets CCU, this indicates a missense error. The additional information presented are that the error is as a result of a change in the residue digit of the first modulus of the amino acid and the suggested correction to this error is to replace the C in residue digit one (1) of the serine (CCU) with a U.

14. CONCLUSION

In this paper, we proposed an Open Reading Frame (ORF) as a string of residue number digits. A global sequence comparison using residue number system is performed. We further considered single digit or base errors, mutations, in pairwise globally compared sequences. Different analysis are performed by deleting, substituting and inserting residue digits (bases). The errors that are manifested are seen as mutations and are classified based on the mutation type – silence, missense and nonsense errors.

15. RECOMMENDATIONS AND FUTURE DIRECTION

This research work can be further applied to the concept sequence alignment by allowing gap penalties and including scoring matrix. A hardware implementation can also be performed to compare some other digital performance parameters such as; digital footprint, timing diagrams and logic gate size with already existing models.

REFERENCES

[1] "Chapter 12: DNA Damage and Repair – Chemistry." https://wou.edu/chemistry/courses/online-chemistry-textbooks/ch450-and-ch451-biochemistry-defining-life-at-the-molecular-level/chapter-12-dna-damage-repair-and-mutations/ (accessed Jan. 05, 2021).

[2] B. N. Al-nuaimi, "Gene Mutations," *Diagnostic Pathol. Mol. Oncol.*, no. March, pp. 1-10-1–13, 2016, doi: 10.1016/b978-0-323-37678-5.50010-4.

[3] B. Fenderson, *Molecular Biology of the Cell,5th Edition*, vol. 40, no. 9. 2008.

[4] S. Azad, "Amino acids: Its types and uses," *Int. J. Clin. Diagnostic Pathol.*, vol. 1, no. 1, pp. 13–16, 2018, [Online]. Available: www.patholjournal.com.

[5] J. Pavsner, *Bioinformatics and Functional Genomics 2nd Edition*, 2nd ed. .

[6] O. Amos and B. Premkumar, *Residue Number System Theory and Implementation*. 2007.

[7] I. M. Balan and I. Bara, "Genetic Mutations and Natural Selection–Steps on the Path of Evolution," *Analele Stiint. ale Univ. Alexandru Ioan Cuza" din Iasi Sec. II a. Genet. si Biol. Mol.*, vol. 8, no. 1, pp. 219–222, 2007, [Online]. Available: http://85.122.19.9/index.php/gbm/article/viewFile/412/400.

[8] K. A. Gbolagade, *Effective Reverse Conversion in Residue Number System Processors*. 2010.

[9] K. O. B. and K. A. G. E.Y. Baagyere, "Bioinformatics: An Importanc Application of Residue Number system," *Bioinformatics: An Importanc Application of Residue Number system*, 2011. http://library1.nida.ac.th/termpaper6/sd/2554/19755.pdf (accessed Dec. 30, 2020).

[10] K. O. Boateng and E. Y. Baagyere, "A Smith-Waterman Algorithm Accelerator Based on Residue

Number System," *Int. J. Electron. Commun. Eng.*, vol. 5, no. 1, pp. 99–112, 2012, [Online]. Available: http://www.irphouse.com.

[11] B. D. Tseng, G. A. Jullien, and W. C. Miller, "Implementation of FFT Structures Using the Residue Number System," *IEEE Trans. Comput.*, vol. C–28, no. 11, pp. 831–845, 1979, doi: 10.1109/TC.1979.1675263.

[12] N. Z. Haron and S. Hamdioui, "Redundant residue number system code for fault-tolerant hybrid memories," *ACM J. Emerg. Technol. Comput. Syst.*, vol. 7, no. 1, pp. 1–19, 2011, doi: 10.1145/1899390.1899394.

[13] J. M. Nicholson, "Will we cure cancer by sequencing thousands of genomes?," *Mol. Cytogenet.*, vol. 6, no. 1, pp. 2–5, 2013, doi: 10.1186/1755-8166-6-57.

[14] "Missense mutation - Wikipedia." https://en.wikipedia.org/wiki/Missense_mutation (accessed Mar. 01, 2021).

[15] J. D. W. with A. Berry, *DNA - The secret life*. 2003.

[16] L. A. Pray and A. A. Aa, "Discovery of DNA Structure and Function: Watson and Crick The First Piece of the Puzzle: Miescher Discovers DNA Citation: Pray, L. (2008) Discovery of DNA structure and function: Watson and Crick. Nature Education 1(1):100," 2008, Accessed: Oct. 12, 2021. [Online]. Available: https://www.nature.com/scitable/topicpage/discovery-of-dna-structure-and-function-watson-397.

[17] M. Morange, "The Central Dogma of molecular biology," *Resonance*, vol. 14, no. 3, pp. 236–247, 2009, doi: 10.1007/s12045-009-0024-6.

[18] M. E. B. Yamagishi and R. H. Herai, "Chargaff's 'Grammar of Biology': New Fractal-like Rules," no. December, 2011, [Online]. Available: http://arxiv.org/abs/1112.1528.

[19] S. Jain, N. Raviv, and J. Bruck, "Attaining the 2nd Chargaff Rule by Tandem Duplications," *IEEE Int. Symp. Inf. Theory - Proc.*, vol. 2018-June, pp. 2241–2245, 2018, doi: 10.1109/ISIT.2018.8437526.

[20] "Protein structure: Primary, secondary, tertiary & quatrenary (article) | Khan Academy." https://www.khanacademy.org/science/biology/macromolecules/proteins-and-amino-acids/a/orders-of-protein-structure (accessed Dec. 31, 2020).

[21] F. H. C. Crick, "ON PROTEIN SYNTHESIS."

[22] "The Information in DNA Determines Cellular Function via Translation | Learn Science at Scitable." https://www.nature.com/scitable/topicpage/the-information-in-dna-determines-cellular-function-6523228/ (accessed Nov. 20, 2021).

[23] E. Ros, A. G. Torres, and L. Ribas de Pouplana, "Learning from Nature to Expand the Genetic Code," *Trends Biotechnol.*, pp. 1–14, 2020, doi: 10.1016/j.tibtech.2020.08.003.

[24] S. S. Bisht and A. K. Panda, *DNA sequencing: Methods and applications*, vol. 9788132215. 2013.

[25] A. E. Men, P. Wilson, K. Siemering, and S. Forrest, "Sanger DNA Sequencing," *Next Gener. Genome Seq. Towar. Pers. Med.*, pp. 1–11, 2008, doi: 10.1002/9783527625130.ch1.

[26] W. R. Taylor, A. C. W. May, N. P. Brown, and A. Aszódi, "Protein structure: Geometry, topology and classification," *Reports Prog. Phys.*, vol. 64, no. 4, pp. 517–590, 2001, doi: 10.1088/0034-4885/64/4/203.

[27] E. V. Koonin and A. S. Novozhilov, "Origin and evolution of the genetic code: The universal enigma," *IUBMB Life*, vol. 61, no. 2, pp. 99–111, 2009, doi: 10.1002/iub.146.

[28] L. R. Nemzer, "A binary representation of the genetic code," *BioSystems*, vol. 155, pp. 10–19, 2017, doi: 10.1016/j.biosystems.2017.03.001.

[29] "What is a Silent Mutation? - 99Science." https://www.99science.com/what-is-a-silent-mutation/ (accessed Mar. 01, 2021).

[30] E. Ifeanyi and B. Ogechi, "Sickle Cell Anaemia : A Review Scholars Journal of Applied Medical Sciences (SJAMS) Sickle Cell Anaemia : A Review," *Sch. J. Appl. Med. Sci.*, vol. 3, no. 6B, pp. 2244–2252, 2015, [Online]. Available: https://www.researchgate.net/publication/301327927%0ASickle Cell Anaemia: A Review Article · January 2015%0ACITATIONS.

[31] M. H. Zhao, X. Li, X. X. Zhang, H. Zhang, and X. Y. Zhao, "Mutation mechanism of leaf color in plants: A review," *Forests*, vol. 11, no. 8, pp. 1–19, 2020, doi: 10.3390/F11080851.

[32] A. Persson and L. Bengtsson, "Forward and reverse converters and moduli set selection in signed-digit residue number systems," *J. Signal Process. Syst.*, vol. 56, no. 1, pp. 1–15, 2009, doi: 10.1007/s11265-008-0249-8.

[33] "An efficient architecture for designing reverse converters based on a general three-moduli set -

ScienceDirect."
https://www.sciencedirect.com/science/article/abs/pii/S1383762108000544?via%3Dihub (accessed Jan. 12, 2021).

[34] W. Wang, M. N. S. Swamy, M. O. Ahmad, and Y. Wang, "Applications of the New Chinese Remainder Theorems for three moduli sets," *Can. Conf. Electr. Comput. Eng.*, vol. 1, pp. 571–576, 1999, doi: 10.1109/ccece.1999.807262.

AUTHORS

Ing. Joshua ApigaguaAkanbasiam is currently a PhD. Candidate in the Department of Computer Engineering, KNUST, and he holds a Masters in Telecommunications Engineering and BSc. in Computer Engineering, KNUST. He is a Senior Lecturer and currently the head of the department of Electrical/Electronics Engineering, Dr.HillaLimann Technical University formerly Wa Polytechnic - Ghana. Prior to Lecturing he served as the Maintenance Manager of Mass Telecom Innovation (MTI) a company that maintained Telecom Power for MTN and Airtel cell sites in the three Northern Regions. He is a member of GhIE and IEEE. His research interest is in Broadband services, Computer Arithmetic and Applications, RNS and its application, Digital Signal Processing, and Bioinformatics.

Prof.Ing. Kwame Osei Boateng holds a Doctorate in Systems Engineering from the Graduate School of Engineering and Science of Ehime University, Japan, Masters in Computer Science from the same university and BSc. (Hons.) degree in Electrical/Electronic Engineering from the University of Science and Technology (now Kwame Nkrumah University of Science and Technology), Kumasi Ghana. Since March 2003, Prof.Ing. Boateng has been working for Kwame Nkrumah University of Science and Technology and has risen through the ranks as a Senior Lecturer, an Associate Professor and currently a Professor. He has held several positions as head of department of Computer Engineering, the dean of Faculty of Electrical and Computer Engineering of the College of Engineering and director of the Institute of Distance Learning and ICT consultant. Prof.Ing. Boateng is a member of the GhIE, IEEE and IEEE Computer Society. He has served on technical programme committee for the 11th IEEE (ATS'02) and IEEE VLSI Test Symposium (VTS), ICAST 2012, WTS 2013, and ESTE 2015. His present research interests are in the areas of design, test and diagnosis of logic and VLSI circuits, test of mixed-signal circuits, network security protocols, applications of residual number systems (RNS), image processing and smart metering.

Prof. Matthew Glover Addo; Mathew Glover Addo is a Professor of Microbiology/Molecular Biology at the Department of Theoretical and Applied Biology, Faculty of Biosciences at the Kwame Nkrumah University of Science and Technology (KNUST). Currently, he is the Director of the Institute of Distance Learning, KNUST. Prof Addo holds a BSc in Biological Sciences from KNUST-Ghana, an MSc in Biotechnology from the University of Bergen, Norway and Doctor of Science/Doctor of Philosophy degrees at the Université Paris Sud IX, Paris, France and KNUST, Kumasi. He completed the doctorate programme in May, 2011. During the doctorate research programme, Prof.Addo designed an efficient screening method for the identification of genes involved in the mitochondrial genome stability using Caenorhabditiselegans as a model organism. He worked with the Functions and Dysfunction of Mitochondrial group of the Institute of Genetics and Microbiology where for the first time, he identified four (4) new nuclear genes (Y105E8A.23, dnj-10, atad-3, and phi-37) involved in mitochondrial stability. Prof.Addo has has 20 years of teaching, collaborative research and consultancy experience. His specialisation is in Gene Expression, Clinical Infectious Microorganisms and Food and Water Microbiology. He has served the international and local communities in a number of capacities and has several peered reviewed publications in reputable international journals. He has also assessed a number of MSc/MPhil and PhD theses. Prof.Addo is a visiting Lecturer/Assessor at the ISA-Lille University, Lille, France. Before his appointment as Director of IDL, Prof Addo was the Dean of Faculty of Biosciences.