SCIENTIFIC OASIS

Decision Making: Applications in Management and Engineering

Journal homepage: www.dmame-journal.org
ISSN: 2560-6018, eISSN: 2620-0104

Digitizing DNA Sequences Using Multiset-Based Nucleotide Frequencies for Machine Learning-Based Mutation Detection

Sanaa Anjum[1], Sajida Kousar[1], Nasreen Kausar[2], Nezir Aydin[3,4], Oludolapo Akanni Olanrewaju[5], Bongumenzi Mncwango[5,*]

[1] Department of Mathematics and Statistics, International Islamic University, Islamabad, Pakistan
[2] Department of Mathematics, Faculty of Arts and Sciences, Yildiz Technical University, Esenler, 34220, Istanbul, Türkiye
[3] College of Science and Engineering, Hamad Bin Khalifa University, 34110 Doha, Qatar
[4] Department of Industrial Engineering, Yildiz Technical University, Besiktas, 34349 Istanbul, Turkey
[5] Department of Industrial Engineering, Durban University of Technology, Durban 4001, South Africa

ARTICLE INFO

Article history:
Received 1 January 2024
Received in revised form 19 June 2024
Accepted 11 July 2024
Available online 25 August 2024

Keywords:
Multiset DNA structure; Multiset average frequency; Recurrent neural network; Gene mutations.

ABSTRACT

Investigating algebraic structures in a non-conventional framework supplements mathematics for hard-nosed practical applications to the fields of theoretical biology and computer science. One such algebraic structure is multigroup whose underlying set is a multiset. The genome is the entire set of DNA instructions found within a cell which contains all the information needed for an individual to develop and function. DNA and RNA are the hereditary materials that play a vital role in the metabolism process of living things, especially protein synthesis. In genomic database DNA sequences are stored in the form of string or text data types. The only data that works with machine learning algorithms is numerical. Thus, it is necessary to transform DNA sequence strings to numerical values. This article is organized in the following manner. Firstly, we prove that standard genetic code is a multigroup and genome architecture of the whole population can be interpreted as the sum of multisets. Next, it is described how a numerical representation of DNA bases relates to its algebraic representation. We further employed Gated Recurrent Unit, Long Short-Term Memory, and Bidirectional Long Short-Term Memory to identify changes between the DNA sequences. Experimental results show that GRU with multiset-based numerical values for DNA bases offers 98% accuracy on testing data. This novel technique will aid in the detection of mutations in complex diseases.

1. Introduction

In conventional set theory, repeating elements within a set is unacceptable. However, in real-world situations, the recurrence of objects in a set cannot be neglected. For example, prime factorization of natural numbers, consideration of repeated roots of polynomial equations, frequent observations in statistical samples, and occurrence of hydrogen atoms within a water molecule. The development of multiset theory started at the beginning of the 1970s. Several

* *Corresponding author.*
E-mail address: bongumenzim@dut.ac.za

https://doi.org/10.31181/dmame7220241213

mathematicians presented various terms (list, bag, heap, bunch, sample, occurrence set, weighted set, and fireset) in different contexts carrying synonymity with multiset by several authors [1]. The term multiset noted by Knuth [2] was first suggested by de Bruijn [3].

The theory of multigroup via multiset was discussed by many researchers [4], but the most acceptable concept of multigroups was given by Nazmul *et al.,* [5] because it follows the non-conventional groups mentioned before in the literature. Additional studies on the theory of multigroups were discussed from time to time. The idea of submultigroup was elaborated in [6]. A complete account on the premise of homomorphism and some homomorphic properties of multigroups and factor multigroups were explained in [7,8]. Some results on normal submultigroups of a multigroup were explicated in [9]. Direct Product of multigroups and its generalization were established in [10]. Multigroup action on multiset was also discussed in a previous study [11]. Some properties of multigroups were analyzed by Ejegwa and Ibrahim [12].

A major challenge for today's and tomorrow's genomics is determining the genome architecture (GA). Recent studies in genomics propose that certain mathematical biophysics rules must be obeyed by GAs. The discovery of the double helix molecular structure of DNA by Watson and Crick [13] in 1953, illustrated that genetic information in the form of sequences of nucleotides is stored in DNA. A nucleotide is a key unit that tightens together to make nucleic acids. A nucleotide is made up of a five-carbon sugar molecule binding to a phosphate group and a nitrogen base. DNA and RNA are the main types of nucleic acids, which are made up of long chains of repeating nucleotides. The base thymine (T) (DNA bases are A, C, G, T) in RNA is replaced with uracil (U) as shown in Figure 1.

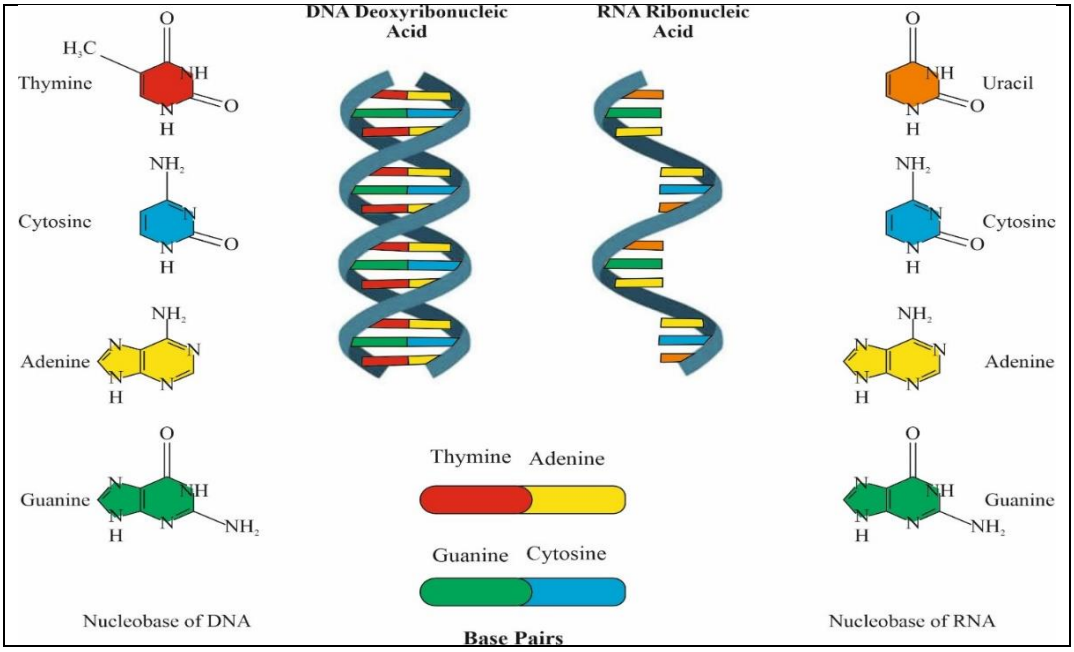

Fig. 1. DNA and RNA Structure

The first major task of molecular biology was to find out how the GC steers the synthesis of proteins. A DNA molecule is translated into RNA through the process of transcription, which is involved in protein synthesis. As a result of this whole process, the codons (triplets) are encoded. Messenger RNA (mRNA) is a single-stranded nucleotide sequence that brings genetic information from the DNA master molecule in the form of triplet codons. mRNA is about 5% of the total RNA in the cell and is more heterogeneous in its coding region than all other RNA. Coding regions are the

triplet codons, which are translated by the ribosomes (the protein factory) into one or many proteins in eukaryotes and prokaryotes respectively. In coding regions, there is always a start codon like AUG triplet in the beginning of sequence and UAA, UAG, or UGA at the end as stop codons discussed by Goss *et al.,* [32].

Crick *et al.,* [14] in 1961 proposed that the GC is a set of rules for the translation of triplet code (codon) into one amino acid. These triplet codes are called standard genetic code (SGC) [15]. This fact is summarized in a table called the genetic code table (GCT) (See Table 4). For biology, the importance of GCT cannot be overlooked. There are 64 codons (61 meaningful codons, i.e., codons encoding amino acids, and three stop-codons) and only 20 amino acids (21 if including the stop signal) come across in all living creatures. Different codons stand for the same amino acid. This reality is pertained to degeneracy, which is an inexorable feature of the GCT. Degeneracy is correlated with and an outcome of symmetry that behaves as an organizing rule in which genetic information is stored and the way, it controls the protein synthesis process. This idea is the soul of the algebraic approach to interpreting the structure that arises from the GC [16].

Genetic code (GC) has been represented by a variety of algebraic structures, which help look at the consequence of the noteworthy linkage between protein-coding regions among the coding apparatus and the mutational process [17]. From a mathematical perspective, a GC resembles a 3D cube, following steady phylogenetic analyses of protein-coding regions of DNA. The importance of a suitable algebraic structure of GC cannot be ignored because it is useful to understand the semantic properties of codes and can help us explain the gene evolution process.

In the late 19th century, various mathematical models were proposed to interpret the genetic code (GC) of DNA bases in binary form. These binary representations necessitate the existence of a partial order on the set of sixty-four codons. These partial orders are centered on the chemical types of bases (purine and pyrimidine) and their hydrogen bond numbers. By assuming that DNA bases with the same hydrogen bond number but different chemical types are complementary elements, a Boolean lattice for DNA bases is constructed. This complementary behavior led Sánchez *et al*., [18,19] and Sánchez and Grau [20] to propose two Boolean lattices for the GC that are dual to each other.

Grau *et al.,* [21] constructed a GC ring that isomorphic to the ring of integers modulo 64 using these codon properties for GC analysis. They also suggested that endomorphisms and automorphisms could describe gene mutation pathways. Moreover, Sánchez *et al*., [22] defined the Galois field of order 64 for the codons. Then they constructed a finite-dimensional vector space over their proposed field by taking the cross-product of a finite number of copies of the field into consideration. Vector space is a replication of DNA sequences. Linear operators defined in this vector space indicate gene mutation in wild-type genes. Later, Sánchez and Grau [17] proposed that the primary divisions of the genetic code table (GCT) could be developed as equivalence classes from the quotient GC vector spaces over the Galois field of the four bases. This newly established algebraic structure focuses on significant connections between algebraic patterns, codon assignments, and the physicochemical properties of amino acids. Additionally, Sánchez [23] further developed the symmetric group (CG, ∘) related to the genetic code cubes . Sánchez and Grau [24] showed that the present GC structure could be obtained from a former coding architecture by using the additional letter D in the four DNA alphabets for predicting quantitatively the relatedness of GA from the same population or closely related species.

Aisah *et al*., [25] suggested together with three letters D, O, P into the four DNA alphabets that are A, C, G, and T, and exploring the algebraic structure of the Abelian group C_{343} under addition which forms a vector space of dimension one over $GF(7^3)$). Sanchez and Barreto initiated that GC

can be represented as a direct sum of homocyclic abelian groups and proposed this result can be extended to the whole genome defined on the GC. Similar canonical decomposition into p-groups can be considered to the alike species of population's GA [26]. Riaz *et al*., [27] designed codes over lattice-valued intuitionistic fuzzy sets to analyze complex DNA structures.

Mutations are random changes, which affect the DNA sequence of living creatures. Mutations can result from structural changes in a gene. Hereditary disorders and even cancer can result from harmful mutations. The current need is to investigate the genetic mutations that raise the risk of developing disorders in a person on a genetic basis.

To analyze DNA sequences through ML algorithms, the DNA sequences should be converted into numeric sequences. For this purpose, the Voss representation mapping system is a popular choice. However, a lot of other techniques has also been introduced including, the tetrahedron, the quaternion, the integers, the real numbers, and the complex numbers [28]. The previously discussed numerical representations are ad hoc codifications, not algebraic representations. In recent years, genetic code algebraic structures have been introduced that give numerical values to DNA bases referring to their algebraic representation. Various related references are previously cited in this article.

Various recent works based on deep learning (DL) techniques analyze changes and mutation of the DNA sequences in one of the works the authors [29], presented RNN technique to discover mutant sequences from metagenomes. A deep learning technique method for encoding meaningful nucleotide sequences and an attention-based long short-term memory (LSTM) network was discovered by Liu *et al*., [30]. A convolutional neural network (CNN)-based technique was developed by Tampuu *et al*., [31] that accepts raw DNA sequences and outputs the probability that the input sequence is viral.

In our recent study, we have introduced multigroup structure for DNA sequences and proposed that the whole genome can be represented as a sum of multiset. Then we represent a novel numerical mapping technique by representing DNA sequence to multiset-based n_i and average frequency. Further, the extracted frequency for DNA sequences is fed into the Gated Recurrent Unit (GRU), Long Short-Term Memory (LSTM), and bidirectional LSTM models to detect changes in DNA sequence. In the end, visual representations of the spectrograms using both mapping techniques were added, which provided a way to compare the two DNA sequences.

2. Preliminaries

A set-like structure in which well-defined elements can present multiple times is called a multiset e.g. $M = \{a, a, b, c, c, c, d\}$ is a multiset. It is customary to denote multiset M as $\{2/a, 1/b, 3/c, 1/d\}$. Let X be a set of elements, then a multiset M over X is a function $C_M : X \to N$ where $N = \{0,1,2,\dots\}$. For each $x \in X, C_M(x)$ denotes the number of times x present in M . A multiset M is an ordinary set if $C_M(x) = 0$ or $1 \quad \forall \quad x \in X$ [2].

The collection of all multisets over X such that no element in the multiset occurs more than n times is called multiset space and is denoted by X^n . Formally, if $X = \{x_1, x_2, \dots, x_k\}$ then $X^n = \{\{n_1/x_1, n_2/x_2, \dots, n_k/x_k\} \quad | \quad i = 1,2,\dots,k \ ; \ n_i \in \{0,1,2,\dots,n\}\}$. The set X^∞ is the set of all multisets over X such that there is no limit on the occurrences of an element in a multiset [3].

If M and N are multisets drawn from a set, X then M and N are equal if and only if $C_M(x) = C_N(x) \quad \forall x \in X$. If M and N are multisets drawn from a set X , then the sum and subtraction of M and N denoted by $M + N$ and $M - N$ is defined as $C_{M+N}(x) = C_M(x) + C_N(x)$ and $C_{M-N}(x) = max\{C_M(x) - C_N(x)\})$ respectively. If M is a multiset drawn from a set,

X then $2M$ and $3M$ represent the sum $M + M$ and $M + M + M$ respectively. In general, kM represent the sum of k $M's$ [27].

Multiset sum operation satisfies commutativity (i.e., $+N = N + M$) and associativity (i.e., $(M + N) + P = M + (N + P)$) [28]. The union and intersection of two multisets M and N drawn from a set X is denoted by $M \cup N$ and $M \cap N$ is defined by $C_{M\cup N}(x) = max\{C_M(x), C_N(x)\}$ and $C_{M\cap N}(x) = min\{C_M(x), C_N(x)\}$ respectively. Let M , $N \in X^m$. Then M is called a submultiset of N $(M \subseteq N)$ if $C_M(x) \leq C_N(x)$ for all $x \in X$. M is a proper submultiset of N $(M \subset N)$ if $C_M(x) \leq C_N(x)$ $\forall$ $x \in X$ and there exists at least one $x \in X$ such that $C_M(x) < C_N(x)$ [3].

Suppose that M is a multiset, then the cardinality of M denoted by $card(M)$, is defined as $card(M) = \sum_{x\in X} C_M(x)$. If M is a submultiset of N , then $card(N) \geq card(M)$ [28].

Let X^n and Y^m are multiset spaces over X and Y respectively and f : $X \to Y$ be a map, then an image $f(M)$ and preimage $f^{-1}(N)$ of multisets M $\in X^n$ and $N \in Y^m$ are defined as:

$$C_{f(M)}(v) = \begin{cases} \bigvee_{u=f^{-1}(v)} C_M(u) & \text{if } f^{-1}(v) \neq \varphi \\ 0 & \text{otherwise} \end{cases} \forall v \in Y.$$

$$\text{And } C_{f^{-1}(N)}(u) = C_N(f(u)) \; \forall u \in X.$$

Let X be a group. A multiset M over X is called a multigroup over X if the count function C_M : $X \to N$, satisfies:

$C_M(xy) \leq C_M(x) \wedge C_M(y), \forall$ $x, y \in X$,

$C_M(x^{-1}) \leq C_M(x), \forall$ $x \in X$.

It follows that,

$C_M(x^{-1}) = C_M(x), \forall$ $x \in X$ since $C_M(x) = C_M\left((x^{-1})\right)^{-1} = C_M(x^{=1}), \forall$ $x \in X$ [3].

3. Multiset Representation of Genetic Code

Let X be a non-empty set of bit strings $\{00,01,10,11\}$. X forms a group under the bitwise binary operation XOR which gives result 0 if both inputs are the same but give 1 if both inputs are different with each other as shown in the see Table 1.

Table 1
Group of bit strings

XOR	00	01	10	11
00	00	01	10	11
01	01	00	11	10
10	10	11	00	01
11	11	10	01	00

DNA basis $\{G, A, T, C\}$ can be classified by considering strong $(S = G, C)$ or weak $(W = A, T)$ number of hydrogen bonds, purine $(R = A, G)$ or pyrimidine $(Y = T, C)$ chemical type and amino $(M = A, C)$ or keto $(K = G, T)$ chemical groups for the four DNA bases. According to the above classification criterion, there are 24 ordered sets for DNA basis. There is a freedom of choice in labelling 2-bit identifier to each of the four DNA basis. The first and second bits can be 0 or 1 for S, W, Y, R, M and K. Table 2 shows that there are 24 ways to label 2-bit identifiers to the four DNA bases.

Table 2
2-bit identifier labeling to the DNA basis

First Bit	Binary Labelling	Second bit	Binary Labelling
S (G, C)/ W (A, T)	G, C=1, A, T=0	Y (T, C)/ R (A, G)	T, C=0, A, G=1
			T, C=1, A, G=0
		M (A, C)/ K (G, T)	A, C=0, G, T=1
			A, C=1, G, T=0
	G, C=0, A, T=1	Y (T, C)/ R (A, G)	T, C=0, A, G=1
			T, C=1, A, G=0
		M (A, C)/K (G, T)	A, C=0, G, T=1
			A, C=1, G, T=0
Y (T, C)/R (A, G)	T, C=0, A, G=1	S (G, C)/ W (A, T)	G, C=1, A, T=0
			G, C=0, A, T=1
		M (A, C)/K (G, T)	A, C =0, G, T=1
			A, C=1, G, T=0
	T, C=1, A, G=0	S (G, C)/ W (A, T)	G, C=1, A, T=0
			G, C=0, A, T=1
		M (A, C)/ K (G, T)	A, C=0, G, T=1
			A, C=1, G, T=0
M (A, C)/K (G, T)	A, C=0, G, T=1	S (G, C)/ W (A, T)	G, C=1, A, T=0
			G, C=0, A, T=1
		Y (T, C)/R (A, G)	T, C=0, A, G=1
			T, C=1, A, G=0
	A, C=1, G, T=0	S (G, C)/ W (A, T)	G, C=1, A, T=0
			G, C=0, A, T=1
		Y (T, C)/R (A, G)	T, C=0, A, G=1
			T, C=1, A, G=0

For Example, let $G \leftrightarrow 00, A \leftrightarrow 01, T \leftrightarrow 10, C \leftrightarrow 11$, which shows a binary representation of DNA bases $\{G, A, T, C\}$ denoted as $S(X)$. The order of DNA base is considered by using several hydrogen bonds (strong/weak) and the first bit is 1 for Y (T, C) and 0 for R (A, G) and the second bit is 1 for M (A, C) and 0 for K (G, T). Table 3 shows that $S(X)$ form group of DNA basis.

Table 3
$S(X)$ Group of DNA basis

	G	A	T	C
G	G	A	T	C
A	A	G	C	T
T	T	C	G	A
C	C	A	T	G

By taking the direct product of 3 copies of $S(X)$ that is $S(X) \times S(X) \times S(X)$ give us a group of 64-codons which is GC as shown in Table 4.

Table 4
The standard genetic code table

First base position	Second base position: G	Second base position: A	Second base position: T	Second base position: C	Third base position
G	GGG	GAG	GTG	GCG	G
G	GGA	GAA	GTA	GCA	A
G	GGT	GAT	GTT	GCT	T
G	GGC	GAC	GTC	GCC	C
A	AGG	AAG	ATG	ACG	G
A	AGA	AAA	ATA	ACA	A
A	AGT	AAT	ATT	ACT	T
A	AGC	AAC	ATC	ACC	C
T	TGG	TAG	TTG	TCG	G
T	TGA	TAA	TTA	TCA	A
T	TGT	TAT	TTT	TCT	T
T	TGC	TAC	TTC	TCC	C
C	CGG	CAG	CTG	CCG	G
C	CGA	CAA	CTA	CCA	A
C	CGT	CAT	CTT	CCT	T
C	CGC	CAC	CTC	CCC	C

Consider a multiset $S = \{48/G, 48/A, 48/T, 48/C\}$ over $S(X)$, where 48 is the count of each DNA base in the standard genetic code table. S forms a multigroup by defining the count function $C_S : S(X) \to N$ and satisfying conditions of multigroup for all $x \in S(X)$. In fact, S is standard genetic code multigroup whatever order and classification of DNA basis have been considered and 2-bit identifier labeling for these bases has been taken out of 24 possibilities. In this scenario, it is evident that a gene of any length has a representation of kS+ submultisets of S, where k is some positive integer. Corresponding to the 24 ordered sets for DNA basis, which isomorphic groups. This result indicates that it is possible to study GA of whole population within the framework of sum of multisets, which provide an error-free and consistent presentation of genome sequencing data. This representation is illustrated in Algorithm 3.1 followed by examples.

3.1 Algorithm

- *Consider DNA sequence string*
- *from collections import Counter*
  dna1 = 'ATG…. TAA
  a = counter (dna1)
  print (a)
  output: Counter $(dna1 = \{'A': n_1, 'T': n_2, 'C': n_3, 'G': n_4\})$
- *Breaking* $n_i = 48 + 48 + \cdots + d$; *such that* $d < 48$ *is some positive integer, where* $i = 1,2,3,4.$
- *DNA Sequence= kS+ Submultisets of S, where k is some positive integer.*

3.1.1 Example

mecA gene (Methicillin resistance gene)

Name: Staphylococcus aureus subsp. Aureus NCTC 8325 chromosome, complete genome
GenBank Accession: NC-007795.1
Sequence length (bp): 720

ATGAGAATAGAACGAGTAGATGATACAACTGTAAAATTGTTTATAACATATAGCGATATCGAGGCCCGTGGA

TTTAGTCGTGAAGATTTATGGACAAATCGCAAACGTGGCGAAGAATTCTTTTGGTCAATGATGGATGAAATT
AACGAAGAAGAAGATTTTGTTGTAGAAGGTCCATTATGGATTCAAGTACATGCCTTTGAAAAAGGTGTCGAA
GTCACAATTTCTAAATCTAAAAATGAAGATATGATGAATATGTCTGATGATGATGCAACTGATCAATTTGATG
AACAAGTTCAAGAATTGTTAGCTCAAACATTAGAAGGTGAAGATCAATTAGAAGAATTATTCGAGCAACGAA
CAAAAGAAAAAGAAGCTCAAGGTTCTAAACGTCAAAAGTCTTCAGCACGTAAAAATACAAGAACAATCATTG
TGAAATTTAACGATTTAGAAGATGTTATTAATTATGCATATCATAGCAATCCAATAACTACAGAGTTTGAAGA
TTTGTTATATATGGTTGATGGTACTTATTATTATGCTGTATATTTTGATAGTCATGTTGATCAAGAAGTCATTA
ATGATAGTTACAGTCAATTGCTTGAATTTGCTTATCCAACAGACAGAACAGAAGTTTATTTAAATGACTATGC
TAAAATAATTATGAGTCATAACGTAACAGCTCAAGTTCGACGTTATTTTCCAGAGACAACTGAATAA

A multiset over the above gene is

$E = \{275/A, 217/T, 138/G, 90/C\}$.

One could see that $E = S + J_1 + 3I_1 + J_2 + I_2 + I_3$ is the sum of submultisets of S where, $I_1 = \{48/A, 48/T, 0/G, 0/C\}, I_2 = \{48/A, 25/T, 0/G, 0/C\}, I_3 = \{35/A, 0/T, 0/G, 0/C\}$, $J_1 = \{48/G, 0/C, 0/A, 0/T\}$, and $J_2 = \{42/G, 42/C, 0/A, 0/T\}$ are submultisets of S.

Note that if $X = \{A, T, G, C\}$ and $Y = \{G, C, A, T\}$ are groups of DNA bases, where $A \leftrightarrow 00, T \leftrightarrow 01, G \leftrightarrow 10, C \leftrightarrow 11$ and $G \leftrightarrow 00, C \leftrightarrow 01, A \leftrightarrow 10, T \leftrightarrow 11$ are their binary representation respectively. And S is standard genetic code multigroup over *X* and *Y*. Then $I_1 = \{48/A, 48/T, 0/G, 0/C\}$, $I_2 = \{48/A, 25/T, 0/G, 0/C\}$, $I_3 = \{35/A, 0/T, 0/G, 0/C\}$ are submultigroups of S corresponding to X and $J_1 = \{48/G, 0/C, 0/A, 0/T\}$, $J_2 = \{42/G, 42/C, 0/A, 0/T\}$ are submultigroups of S corresponding to Y.

3.1.2 Example

INS gene

Name: INS gene (Homo Sapiens INS gene, partial)

GenBank Accession: AJ009655.1

Sequence length (bp): 1393

AGCAGGTCTGTTCCAAGGGCCTTTGCGTCAGGTGGGCTCAGGGTTCCAGGGTGGCTGGACCCCAGGCCCCA
GCTGTGCAGCAGGGAGGACGTGGCTGGGCTCGTGAAGCATGTGGGGGTGAGCCCAGGGGCCCCAAGGCA
GGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGCCTGTCTCCCAGATCACTGTCCTTCTGCCATGGCCCTGTGG
ATGCGCCTCCTGCCCCTGCTGGCGCTGCTGGCCCTCTGGGGACCTGACCCAGCCGCAGCCTTTGTGAACCAA
CACCTGTGCGGCTCACACCTGGTGGAAGCTCTCTACCTAGTGTGCGGGGAACGAGCTTCTTCTACACACCCA
AGACCCGCCGGGAGGCAGAGGACCTGCAGGGTGAGCCAACCGCCCATTGCTGCCCCTGGCCGCCCCCAGCC
ACCCCCTGCTCCTGGCGCTCCCACCCAGCATGGGCAGAAGGGGGCAGGAGGCTGCCACCCAGCAGGGGGTC
AGGTGCACTTTTTTAAAAAGAAGTTCTCTTGGTCACGTCCTAAAAGTGACCAGCTCCCTGTGGCCCAGTCAGA
ATCTCAGCCTGAGGACGGTGTTGGCTTCGGCAGCCCCGAGATACATCGAGGGTGGGCACGCTCCTCCCTCCA
CTCGCCCCTCAAACAAATGCCCCGCAGCCCATTTCTCCACCCTCATTTGATGACCGCAGATTCAAGTGTTTTGT
TAAGTAAAGTCCTGGGTGACCTGGGGTCACAGGGTGCCCCACGCTGCCTGCCTCTGGGCGAACACCCCATCA
CGCCCGGAGGAGGGCGTGGCTGCCTGCCTGAGTGGGCCAGACCCCTGTCGCCAGCCTCACGGCAGCTCCAT
AGTCAGGAGATGGGGAAGATGCTGGGGACAGGCCCTGGGGAGAAGTACTGGGATCACCTGTTCAGGCTCC
CACTGTGACGCTGCCCCGGGGCGGGGGAAGGAGGTGGGACATGTGGGCGTTGGGGCCTGTAGGTCCACAC
CCAGTGTGGGTGACCCTCCCTCTAACCTGGGTCCAGCCCGGCTGGAGATGGGTGGGAGTGCGACCTAGGCT
GGCGGGCAGGCGGGCACTGTGTCTCCCTGACTGTGTCCTCCTGTGTCCCTCTGCCTCGCCGCTGTTCCGGAAC
CTGCTCTGCGCGGCACGTCCTGGCAGTGGGGCAGGTGGAGCTGGGCGGGGGCCCTGGTGCAGGCAGCCTG
CAGCCCTTGGCCCTGGAGGGGTCCCTGCAGAAGCGTGGCATTGTGGAACAATGCTGTACCAGCATCTGCTCC
CTCTACCAGCTGGAGAACTACTGCAACTAGACGCAGCCTGCAGGCAGCCCCACACCCGCCGCCTCTGCACCG
AGAGAGATGGAATAAAGCCCTTGAACCAGC

A multiset over above gene is

$$F = \{456/C, 447/G, 255/T, 235/A\}.$$

One can see that $F = 4S + 5K_1 + K_2 + L_1 + L_2$ is the sum of submultisets of S.

Note that if $X = \{C, G, T, A\}$ and $Y = \{T, A, C, G\}$ are groups of DNA bases, where $C \leftrightarrow 00, G \leftrightarrow 01, A \leftrightarrow 10, T \leftrightarrow 11$ and $T \leftrightarrow 00, A \leftrightarrow 01, G \leftrightarrow 10, C \leftrightarrow 11$ are their binary representation respectively. And S is standard genetic code multigroup over *X* and *Y.* Then

$K_1 = \{48/C, 48/G, 0/T, 0/A\}, K_2 = \{24/C, 15/G, 0/T, 0/A\}$ are submultigroups of S corresponding to X and $L_1 = \{48/T, 43/A, 0/G, 0/C\}, L_2 = \{15/T, 0/A, 0/G, 0/C\}$ are submultigroups of S corresponding to Y.

4. Numerical Mapping technique

A new mapping technique which is a multiset-based numerical mapping technique is applied in this work. A multiset over mecA gene of sequence length 720 bp is $E = \{275/A, 217/T, 138/G, 90/C\}$ represent the count of each base in DNA sequence. Algorithm 3.1 is used to take an average of the frequencies of each nucleotide in the DNA sequence. Since it is already proved that every DNA sequence can be written as the sum of multisets of S, an average of multiples of 48 on each nucleotide gives the average frequency for the DNA bases. The Average frequency values of DNA bases for the above DNA sequence are as follows $A = 45.8, T = 43.4, G = 46.0, C = 45.0$. Dataset used for analysis is given in Table 5.

Table 5
Dataset uses for the analysis

Index	Access number
1	KT279557.1
2	KT279556.1

5. Training of Models

Applications of natural language processing (NLP) include chatbots, machine translation, sentiment analysis, speech recognition, and more. To perform these tasks, NLP systems often rely on artificial neural networks (ANNs), which are models that mimic the structure and function of biological neurons. Sequential data is processed by a specific type of ANN called Recurrent neural networks (RNN). These are comprised of feed-forward neural networks, and their behavior is identical to that of human brains. The RNN uses each node as a memory cell to aid this network so that it can be able to remember the sentence's context [32-34].

Since machine learning algorithms only require numbers [35], in our recent experiment, we used multiset-based average frequency and count nucleotide mapping for numerical conversion of DNA sequence. A flow chart of the proposed is provided in the following Figure 2.

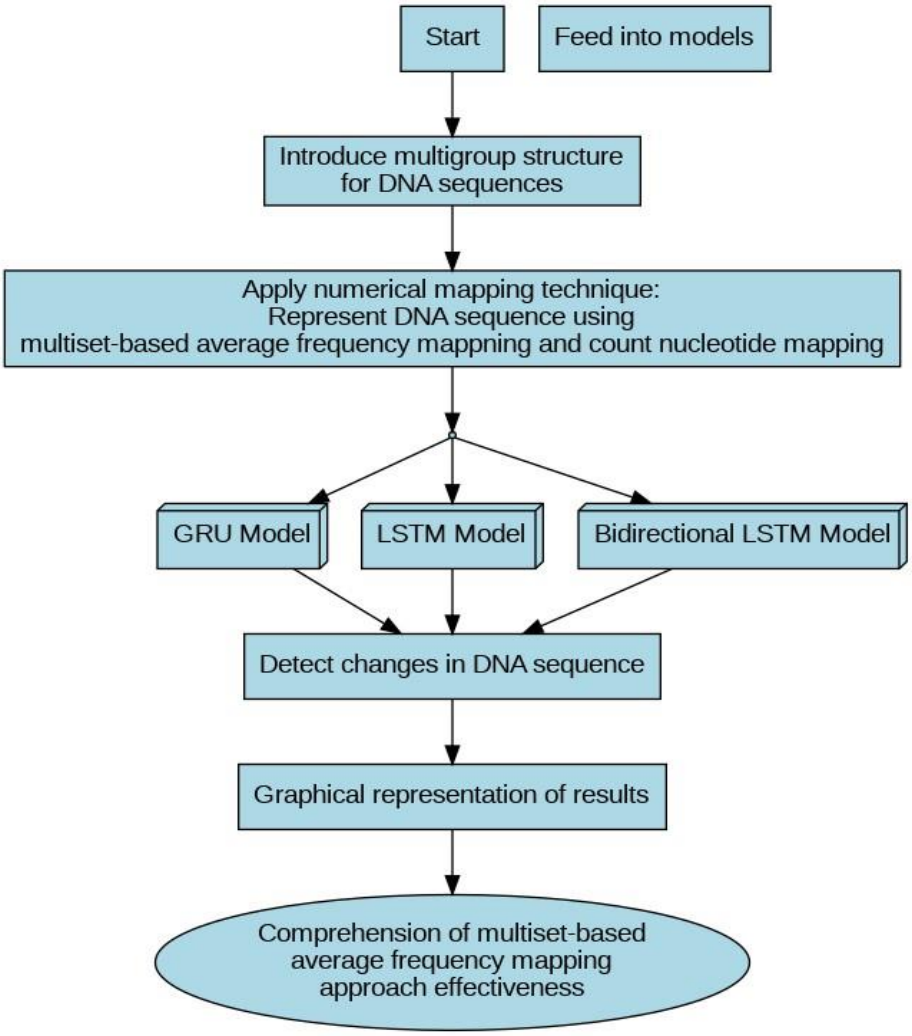

Fig. 2. Flow chart of the proposed method

We trained the LSTM, GRU and bidirectional LSTM model on a subsequence of our reference gene. For this experiment, a subsection of the gene is used. Three Machine Learning models LSTM, GRU, and Bidirectional LSTM were used for learning. TensorFlow library is used in Jupyter notebook. First, convert nucleotide sequence to a numerical index, and a sequence length of 50 is used to construct the array for learning. Two numerical mapping techniques are used to digitize the sequential data, frequency of each DNA base, and multiset-bases average frequency of each DNA base in a DNA sequence. The models are tested for both mapping techniques. The training and validation accuracy of GRU model on 50 epochs achieved an accuracy of 98.50 % using multiset-based n_i average frequency of DNA bases. Figure 3 provides the accuracy comparison for GRU model between the count nucleotides mapping and multiset-based n_i average mapping.

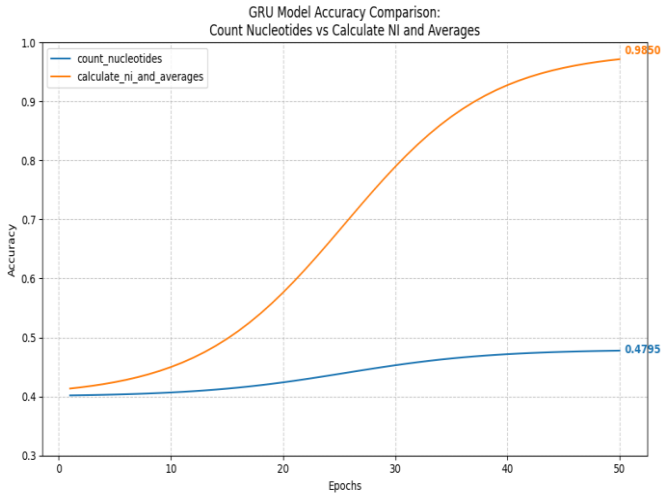

Fig. 3. Accuracy comparison for GRU model

Insufficient information is provided by count nucleotide mapping to enable the models to discriminate between changed and unchanged sequences. LSTM and BLSTM seem less appropriate for this task than GRU. However, numeric values through count nucleotide frequency of DNA bases do not get the required level of accuracy which is given in Table 6.

Table 6. Accuracy, Sensitivity and Precision results

Numerical mapping techniques	Models	Accuracy (%)	Sensitivity (%)	Precision (%)
Count nucleotide mapping	LSTM	47.95	0.00	0.00
	BLSTM	41	0.00	0.00
	GRU	47.95	0.00	0.00
multiset-based n_i-average frequency of nucleotide	LSTM	49.42	0.00	0.00
	BLSTM	47.95	0.00	0.00
	GRU	98.50	97.31	99.80

6. Spectrogram Analysis

The spectrograms visualisations offer a means of contrasting the two DNA sequences and the two distinct mapping techniques. These graphics aid in our comprehension of why the multiset-bases n_i -average frequency mapping approach would have yielded better results in the machine learning models, especially for the GRU model. The four nucleotides (A, T, C, and G) are represented by the y-axis, and the position along the sequence is indicated by the x-axis. With brighter colours (yellow) denoting higher counts and darker colours (blacker) denoting lower counts, the colour intensity represents the number of each nucleotide in a specific window.

Multiset-bases n_i average frequency offers a normalized view of the data, which may facilitate the models' ability to identify subtle variations amongst sequences whereas the count nucleotides mapping provides a direct representation of the data. Sequence 1 Sequences 1 and 2 exhibit similar patterns, indicating that these sequences are highly identical. The patterns we observe may be indicative of many DNA functional domains, including structural motifs, regulatory elements, and

coding sections. The sequences' strong similarity suggested the possibility that they are closely related or identical alleles of the same gene (see Figure 4).

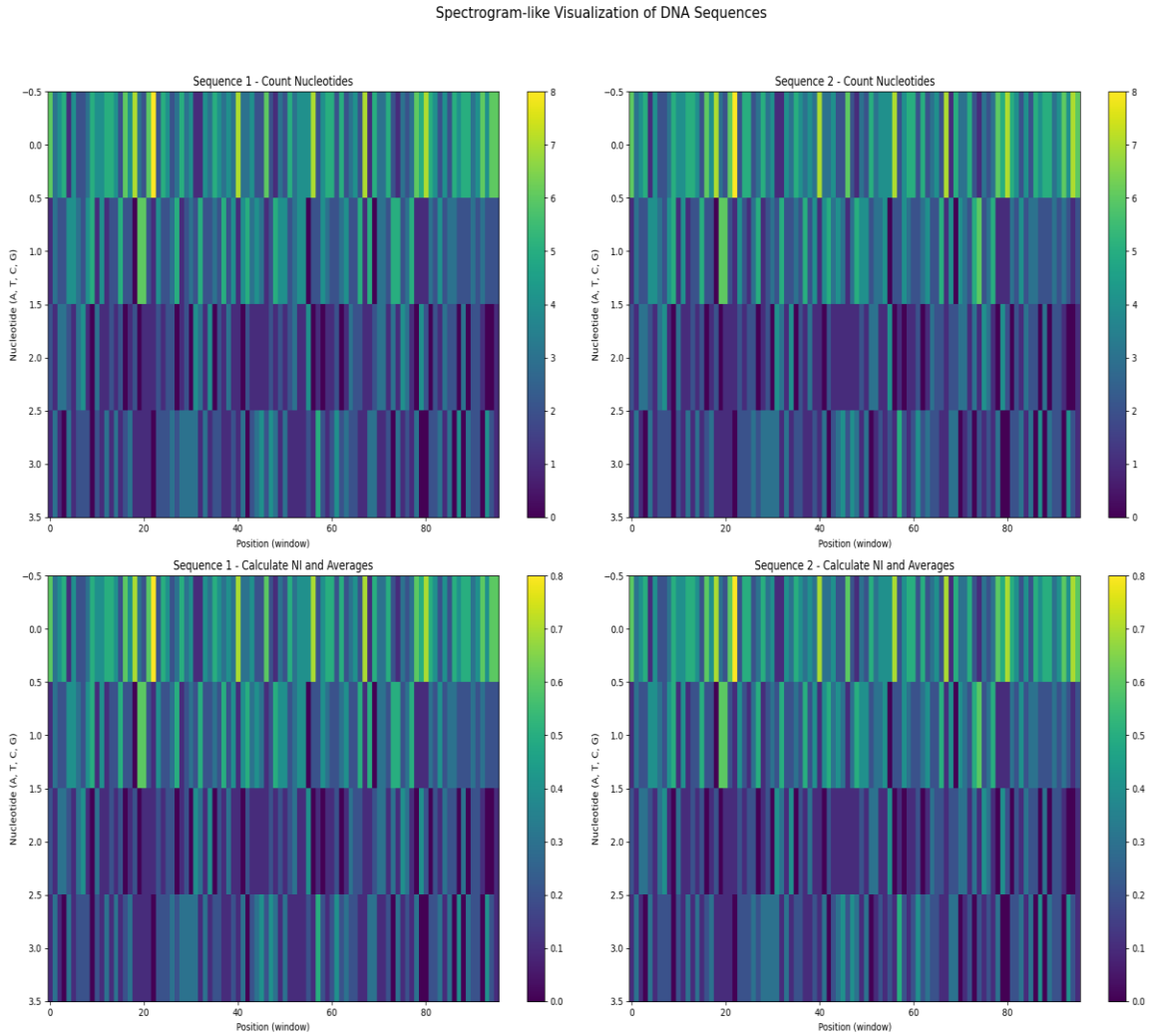

Fig. 4. Visualizations of two sequences using two different mapping methods

7. Conclusions

Multiset theory plays an important role in practical problems. This manuscript depicts that stranded genetic code is a multigroup and genome of any species has a representation of the sum of multisets which presents genome sequencing data in a useful and meaningful manner. We compared two numerical mapping techniques to convert DNA sequence data to numerical values. Additionally, we tested three ML models to detect changes in the DNA sequences. The multiset-based n_i average frequency mapping provides more informative features than count nucleotide mapping, allowing the GRU model to learn effectively. Spectrogram visualization of proposed Mapping demonstrates how accumulating mutations change the global genetic profile. The high similarity between the sequences suggests they might be alleles of the same gene or closely related genes. In the future, this novel technique could be used to detect mutations for early prediction of complex diseases.

Author Contributions

S.A and S.K presented the conceptual framework, designed the methodology, and supervised the research, drafted the original manuscript, N.k, conducted a formal investigation, analysis, Methodology, Conceptualization, Validation, B.M; Review and editing, and funding. O.A.O; Review and editing N.A; Conceptualization, management, review, and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This research was not funded by any grant.

Data Availability Statement

Datasets used for analysis are presented in Table 5.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

[1] Blizard, W. D. (1991). The Development of Multiset Theory. The Review of Modern Logic, 1(4), 319 – 52.
[2] Knuth, D. E. (1982). The art of computer programming. 8th ed. Addison-Wesley.
[3] De Bruijn, N. G. (1983). Denumerations of rooted trees and multisets. Discrete Applied Mathematics, 6(1), 25-33. https://doi.org/10.1016/0166-218X(83)90097-5
[4] Ibrahim, A. M., & Ejegwa, P. A. (2016). A Survey on the Concept of Multigroups. Journal of the Nigerian Association of Mathematical Physics, 38, 1-8.
[5] Nazmul, S., Majumdar, P., & Samanta, S. K. (2013). On multisets and multigroups. Annals of Fuzzy Mathematics and Informatics, 6(3), 643--656.
[6] Ejegwa, P. A., & Ibrahim, A. M. (2017). Characteristics submultigroups of a multigroup. Gulf Journal of Mathematics, 5(4), 1-8. https://doi.org/10.56947/gjom.v5i4.115
[7] Ejegwa, P. A., & Ibrahim, A. M. (2017). Some homomorphic properties of multigroups. Buletinul Academiei de Ştiinţe a Republicii Moldova Matematica, 1(83), 67-76.
[8] Ejegwa, P. A., & Ibrahim, A. M. (2017). On Comultisets and Factor Multigroups. Theory and Applications of Mathematics & Computer Science, 7(2), 124-40.
[9] Ejegwa, P. A., & Ibrahim, A. M. (2017). Normal submultigroups and comultisets of a multigroup. Quasigroups and Related Systems, 2(25), 231-244.
[10] Ejegwa, P. A., & Ibrahim, A. M. (2017). Direct Product of Multigroups and Its Generalization. International Journal of Mathematical Combinatorics, 4(2017), 1-18.
[11] Ibrahim, A. M. & Ejegwa, P. A. (2017). Multigroup actions on multiset. Annals of Fuzzy Mathematics and Informatics, 14(5), 515-526.
[12] Ejegwa, P. A., & Ibrahim, A. M. (2020). Some Properties of Multigroups. Palestine Journal of Mathematics, 9(1), 31-47.
[13] Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. Nature, 171, 737-738. https://doi.org/10.1038/171737a0
[14] Crick, F. H. C., Barnett, L., Brenner, S., & Watts-Tobin R. J. (1961). General Nature of the Genetic Code for Proteins. Nature, 192, 1227-1232. https://doi.org/10.1038/1921227a0
[15] Liczner, C., Duke, K., Juneau, G., Egli, M., & Wilds, C. J. (2021). Beyond ribose and phosphate: Selected nucleic acid modifications for structure-function investigations and therapeutic applications. Beilstein Journal of Organic Chemistry, 17, 908-931. https://doi.org/10.3762/bjoc.17.76

[16] Hornos, J. E. M., Hornos, Y. M. M., & Forger, M. (1999). Symmetry and Symmetry Breaking: An Algebraic Approach to the Genetic Code. *International Journal of Modern Physics B*, 13(23), 2795-2885. https://doi.org/10.1142/S021797929900268X
[17] Sánchez, R., & Grau, R. (2006). A novel algebraic structure of the genetic code over the Galois field of four DNA bases. Acta Biotheoretica, 54(1), 27-42. https://doi.org/10.1007/s10441-006-6192-9
[18] Sánchez, R., Morgado, E., & Grau, R. (2004). The Genetic Code Boolean Lattice. arXiv preprint q-bio/0412034. https://doi.org/10.48550/arXiv.q-bio/0412034
[19] Sánchez, R., Morgado, E., & Grau, R. (2005). A genetic code Boolean structure. I. The meaning of Boolean deductions. Bulletin of Mathematical Biology, 67, 1-14. https://doi.org/10.1016/j.bulm.2004.05.005
[20] Sanchez, R., & Grau, R. (2005). A genetic code Boolean structure. II. The genetic information system as a Boolean information system. Bulletin of Mathematical Biology, 67(5), 1017-1029. https://doi.org/10.1016/j.bulm.2004.12.004
[21] Grau, R., Del C. Chavez, M., Sanchez, R., Morgado, E., Casas, G., & Bonet, I. (2006, May). Boolean algebraic structures of the genetic code: possibilities of applications. In International Workshop on Knowledge Discovery and Emergent Complexity in Bioinformatics (pp. 10-21). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71037-0_2
[22] Sánchez, R., Perfetti, L. A., Grau, R., & Morgado, E. (2004). A new DNA sequences vector space on a genetic code Galois field. arXiv preprint q-bio/0412019. https://doi.org/10.48550/arXiv.q-bio/0412019
[23] Sanchez, R., (2018). Symmetric Group of the Genetic--Code Cubes. Effect of the Genetic--Code Architecture on the Evolutionary Process. Communications in Mathematical and in Computer Chemistry, 79, 527-560.
[24] Sánchez, R., & Grau, R. (2009). An algebraic hypothesis about the primeval genetic code architecture. Mathematical Biosciences, 221(1), 60-76. https://doi.org/10.1016/j.mbs.2009.07.001
[25] Aisah, I., Sayyidatunnisa, N. U., Subartini, B., & Kartiwa, A. (2019, July). Vector space of codons sequence over galois field GF (73). In IOP Conference Series: Materials Science and Engineering (Vol. 567, No. 1, p. 012019). IOP Publishing. https://doi.org/10.1088/1757-899X/567/1/012019
[26] Sanchez, R., & Barreto, J. (2021). Genomic abelian finite groups. bioRxiv, 2021-06. https://doi.org/10.1101/2021.06.01.446543
[27] Riaz, A., Kousar, S., Kausar, N., Pamucar, D., & Addis, G. M. (2022). Codes over Lattice-Valued Intuitionistic Fuzzy Set Type-3 with Application to the Complex DNA Analysis. Complexity, 2022(1), 5288187. https://doi.org/10.1155/2022/5288187
[28] Wildberger, N. J. (2003). A new look at multisets. School of mathematics, UNSW Sydney, 2052, 1-21.
[29] Syropoulos, A. (2001). Mathematics of multisets. In Multiset Processing: Mathematical, Computer Science, and Molecular Computing Points of View 1 (pp. 347-358). Springer Berlin Heidelberg.
[30] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. Procedia CIRP, 99, 650-655. https://doi.org/10.1016/j.procir.2021.03.102
[31] Lugo, L., & Barreto, H. E. (2021). A Recurrent Neural Network approach for whole genome bacteria identification. Applied Artificial Intelligence, 35(9), 642-656. https://doi.org/10.1080/08839514.2021.1937161
[32] Syropoulos, A. (2001). Mathematics of Multisets. In: Calude, C.S., PĂun, G., Rozenberg, G., Salomaa, A. (eds) Multiset Processing. WMC 2000. Lecture Notes in Computer Science, 2235. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45523-X_17
[33] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. Procedia CIRP, 99, 650-655. https://doi.org/10.1016/j.procir.2021.03.088
[34] Lugo, L., & Hernández, E. B. (2021). A Recurrent Neural Network approach for whole genome bacteria identification. Applied Artificial Intelligence, 35(9), 642-656. https://doi.org/10.1080/08839514.2021.1922842
[35] Kalaiarasi, K., Soundaria, R., Kausar, N., Agarwal, P., Aydie, H., & Alsamir, H. (2021). Optimization of the average monthly cost of an EOQ inventory model for deteriorating items in machine learning using PYTHON. Thermal Science, 25(2), 347-358. https://doi.org/10.2298/TSCI21S2347K