SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
VOL. 5, NO. 3, DECEMBER 2022
DOI: 10.35377/saucis.05.03.1191850

Research Article

The Effect of Numerical Mapping Techniques on Performance in Genomic Research

Seda Nur Gülocak[1], Bihter Daş[2]

[1]Department of Software Engineering, Technology Faculty, Firat University; sedanurgulocak@gmail.com;
[2]Corresponding Author; Department of Software Engineering, Technology Faculty, Firat University;
bihterdas@firat.edu.tr;

Received 19 October 2022; Accepted 1 November 2022; Published online 31 December 2022

Abstract

In genomic signal processing applications, digitization of these signals is needed to process and analyze DNA signals. In the digitization process, the mapping technique to be chosen greatly affects the performance of the system for the genomic domain to be studied. The purpose of this review is to analyze how numerical mapping techniques used in digitizing DNA sequences affect performance in genomic studies. For this purpose, all digital coding techniques presented in the literature in the studies conducted in the last 10 years have been examined, and the numerical representations of these techniques are given in a sample DNA sequence. In addition, the frequency of use of these coding techniques in four popular genomic areas such as exon region identification, exon-intron classification, phylogenetic analysis, gene detection, and the min-max range of the performances obtained by using these techniques in that area are also given. This study is thought to be a guide for researchers who want to work in the field of bioinformatics.

Keywords: Numerical mapping techniques, Genomic analysis, DNA encoding schemes, Genomic signal processing, DNA sequence

1. Introduction

Deoxyribose nucleic acid (DNA) is the biological structure that is located inside the cells, which are the building blocks of the human body, and creates the genetic code in which all human characteristics are encoded. DNAs consist of sugar groups, phosphate groups and bases linked by ester bonds. These bases are Adenine, Thymine, Guanine, and Cytosine. In the DNA chain consisting of two long polymers, Adenine pairs with Thymine while Guanine pairs with Cytosine. The codes created by certain combinations of the building blocks called nucleic acids that makeup DNA are called genes. Genes; These are personal codes that determine all the characteristics of the body, such as eye color, height, hairstyle, or susceptibility to genetic diseases. A gene has exons and introns. The intron is the non-amino acid coding portion of a gene. Exons are the protein-coding parts of the gene. The triple arrangement of bases in DNA is referred to as codons and these codons code for the different amino acids that make up proteins. There are 64 possible codons in a DNA. Three of the codons (UAA, UAG, UGA) are termination or stop codons and do not code for any amino acid. Each of the remaining 61 codons codes for an amino acid, but since there are only 20 amino acids used in protein construction, there is more than one codon encoding the same amino acid [1]. Genetic information from DNA is transferred to RNA. This process is called transcription. The translation is the process of translating the code carried by the mRNA into proteins. Figure 1 shows the basic structure of a protein-coding gene related to transcription and translation in a eukaryotic organism.

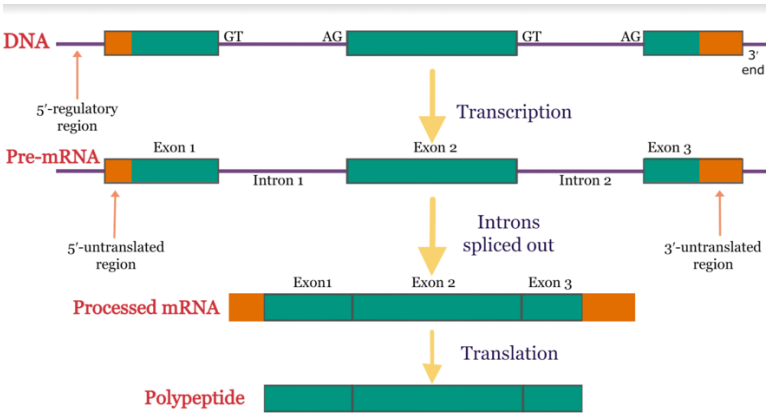


Figure 1 Gene and intergenic regions in a DNA

1.1. Literature Review

In this section, studies that examine the techniques used for digitization of DNA data in various genomic fields such as detection of exon regions from DNA sequences, exon-intron classification, phylogenetic analysis, interspecies similarity/difference, detection of disease, and gene detection are presented. Table 1 lists all studies over the past 15 years examining numerical mapping techniques used by genomic domains.

Table 1 All studies reviewed in the last 15 years

Reference Paper	*Numerical mapping techniques*	*Genomic domain*
Das et al.[3]	1. The Nucleotide Mapping Diploye moment mapping, Minimum entropy mapping, Trigonometric mapping, Variable mapping, Chaos game representation, Gray code mapping, Walsh code mapping, Pseudo-EIIP mapping 2. The Amino acid Mapping Dipola moment and alpha mapping, Binary representation, Genetic code context, EIIP, Complex prime numeric representation, Hyropathy index, P-adic mapping, Ionization-constant, The tetrahedron mapping	Prediction of Exon regions
Wisesty et al. [4]	Voss mapping, Z-curve, tetrahedron, Complex number representation, Integer and real representation, Trigonometric mapping, Paired numeric representation	Diagnosis of breast cancer
Kumar et al. [5]	Position Based Encoding, 2-bit Neural Network based encoding, Hamming Distance Based Encoding, Integer Number Encoding, Trigonometric Encoding, Autocorrelation Based Encoding and proposed Method walsh code	Detection of protein coding regions
Yu et al. [6]	1. Biochemical Properties Atomic number, Electron-Ion Interaction pseudopotential, Molecular mass representation, Thermodynamic properties 2. Primary-Structure Properties Dinucleotide representation, Ring structure, Inter nucleotide distance encoding, Triplet encoding, Frequency of occurence mapping, Minimum entropy mapping 3. Cartesian-Coordinate Properties Integer and real number, Complex number, QPSK/PAM, DNA walk and paired numeric Method 4. Binary and Information Encoding Voss representation, Galois field, Error-Correction code, Ching representation	Genomic signal processing

	5. Graphical Representation CGR and CGR-walk, Tetrahedron, SOM based approach, Quaternion, H-curve and Z-curve	
Kumari et al. [7]	1.Fixed Mapping Voss, Tetrahedron, Complex, Integer, Real, Quaternion, QPSK 2.Cariable Mapping Complex representation of nucleotides by twiddle factor 3.Physico Chemical Property Based Mapping Atomic, Paired, DNA walk, Z-curve, EIIP, Pseudo-EIIP	Genomic signal processing
Das et al. [8]	Voss, Integer, Complex, Real, EIIP, Atomic Number, Paired Numeric, DNA Walk, Molecular Mass, Trigonometric, Entropy, Z- curve, Tetrahedron	Predicting of protein coding regions
Ahmad et al. [9]	Tetrahedron, 4-bit binary coding, Binary coding, Molecular mass, Z-curve, Pathogenity island coding, Entropic segmenttion coding, Paired nucleotide representation, Integer number, Autoregressive coding, Gradient source Localization, EIIP, Paired nucleotide atomic number, Complex number	Genomic signal processing
Jin et al. [10]	Methods based on graphical representation; 2-3-4-5-6 dimensional graphical methods, Chaos game representation, quaternion. Matrix mapping, Coding based on chemical properties, Codons, frequency values, Position statistics, Huffman coding, Euclidean distance coding	Detection of similarity between species
Mendizabal-Ruiz et al. [11]	Integer, Real, EIIP, Atomic number, Paired numeric, Voss, Tetrahedron, Z-curve, DNA walk	Identification of similarity of DNA sequences
Saini et al. [12]	Voss, Tetrahedron, Complex, EIIp, DNA walk, Integer number, Real number, Binary representation, 4-bit binary encoding, Paired nucleotide representation, Quaternion, Inter-numcleotide distance	Genomic signal processing
Mabrouk et al. [13]	Genetic code context, Frequency of nucleotide occurence, Atomic number, 2-bit binary, EIIP	Identification of protein coding regions
Das et al. [14]	Voss mapping, Integer mapping, Complex Mapping, Real Mapping, EIIP Mapping, Atomic Number Mapping, Paired Numeric Mapping, DNA Walk Mapping, The Modecular Mass Mapping	Identification of exon regions
Das et al. [15]	Integer Mapping, Reel Mapping, Atomic Mapping, Molecular Mass Mass Mapping, DNA Walk Mapping, Paired Numeric Mapping, Complex Digital Mapping, EIIP	Classification of exon and intron
Abo-Zahhad et al. [16]	Atomic number, Integer number, Real number, EIIP, Paired Numeric, DNA Walk	Prediction of donor ve acceptor in exon region
Abo-Zahhad et al. [17]	1. Fixed Mapping 2. Voss, Tetrahedron, Complex, Integer, Real, Quaternion 3. Physico Chemical Property Based Mapping EIIP, Paired numeric, DNA-walk, Z-curve	Classification of exon and intron
Kwan et al. [18]	Integer number, Single Galois Indicator, Paired nucleotide atomic umber, Atomic number, Molecular Mass, EIIP, Paired Numeric, Real Number, Complex Number, K-twin pair code, K-bipolar pair code, K-quaternion	Classification of exon and intron
Sharma et al. [19]	Voss, Tetrahedron, Z-curve, Complex, EIIP, Paired numeric, DNA walk, Frequency Nucleotide Occurence, Atomic number, Real number	Identification of exon region
Akalin et al. [20]	Real mapping, Moleculer Mass, EIIP, Shannon Entropy, Paired digital mapping technique	Classification of exon and intron
Akalin et al. [21]	Real mapping, Moleculer Mass, EIIP, Shannon Entropy, Paired digital mapping technique	Prediction of leukaemia
Akhtar et al. [22]	Voss, Tetrahedron, Z-curve, Complex, Queternion, EIIP, QPSK-PAM, Paired numeric	Prediction of exon regions

In this study, all numerical mapping techniques developed in the last 15 years in the literature and used to digitize DNA sequences were examined and the benefits and shortcomings of these numerical techniques in genomic study areas such as exon region detection, exon-intron classification, phylogenetic analysis. Also, disease-causing gene detection were emphasized. Digitization of DNA sequences is extremely important in order to achieve targeted high-performance accuracy in genomic studies such as detection of exon regions, exon-intron classification, disease-causing gene detection,

phylogenetic analysis. Therefore, in this study, all the digital mapping techniques of the last 15 years were introduced in detail and a review study presenting all the techniques was actualized.

1.2 Motivation

Technological developments in biology and computers have advanced rapidly, and thus the emerging branch of bioinformatics has taken the lead among the most popular academic and industrial sectors today. Genome analysis is one of the most studied subjects in the field of bioinformatics, which is the synthesis of mathematics, statistics, computer science, molecular biology, and genetics. Although genomic studies seem to be aimed at basic scientific research, they will be indispensable for clinical informatics in the coming years. Our motivation for this review study is to analyze the effect of digital mapping techniques on the performance of the system in the most popular bioinformatics and genomic fields of study. In addition, while converting DNA analog signals into digital signals that can be understood by the computer in artificial intelligence applications, it is to guide researchers in choosing the correct digital coding technique that can best reflect the structure of DNA.

The remainder of this paper is organized as follows. In section 2, all numerical mapping techniques introduced in the literature by other authors in the last 5 years are searched and listed for this survey article. Section 3 highlights the frequency of use, performance, advantages, and drawbacks of numerical mapping techniques by genomic domains. In addition, mapping techniques that researchers can use according to genomic domains will be recommended along with their reasons. Finally, in Section 4 we conclude our survey with a brief summary.

2. DNA Numerical Mapping Techniques

In this section, the coding techniques developed for the digitization of DNA sequences are comprehensively examined under five main headings. In the literature, coding techniques are also called different names as digital mapping techniques, numerical methods, and coding schemes. However, all the nomenclatures mean the same. 50 digital mapping techniques developed in the last 5 years are classified into five groups according to their general characteristics. These groups are cartesian coordinate coding techniques, biochemical and physicochemical coding techniques, binary and information coding, primary structure coding techniques, and graphically represented coding techniques. At the end of each of these five groups, there are collective digital signal plot graphs and tables of the digital coding techniques examined in that group. Graphs of digitized DNA signals using coding techniques include representations of the DNA sequence digitized by each coding technique applied to the DNA Fasta format dataset with reference number NR_131216.1 from the NCBI database. Since the graphical value ranges of some mapping techniques are different, the numerical representation of these techniques is given in separate figures. In the general tables at the end of the examined groups, there is a brief explanation of the digitization technique in that group, the coding scheme, and the numerical version of this coding technique applied to a sample DNA sequence. Figure 2 shows the hierarchical scheme of all DNA mapping techniques.

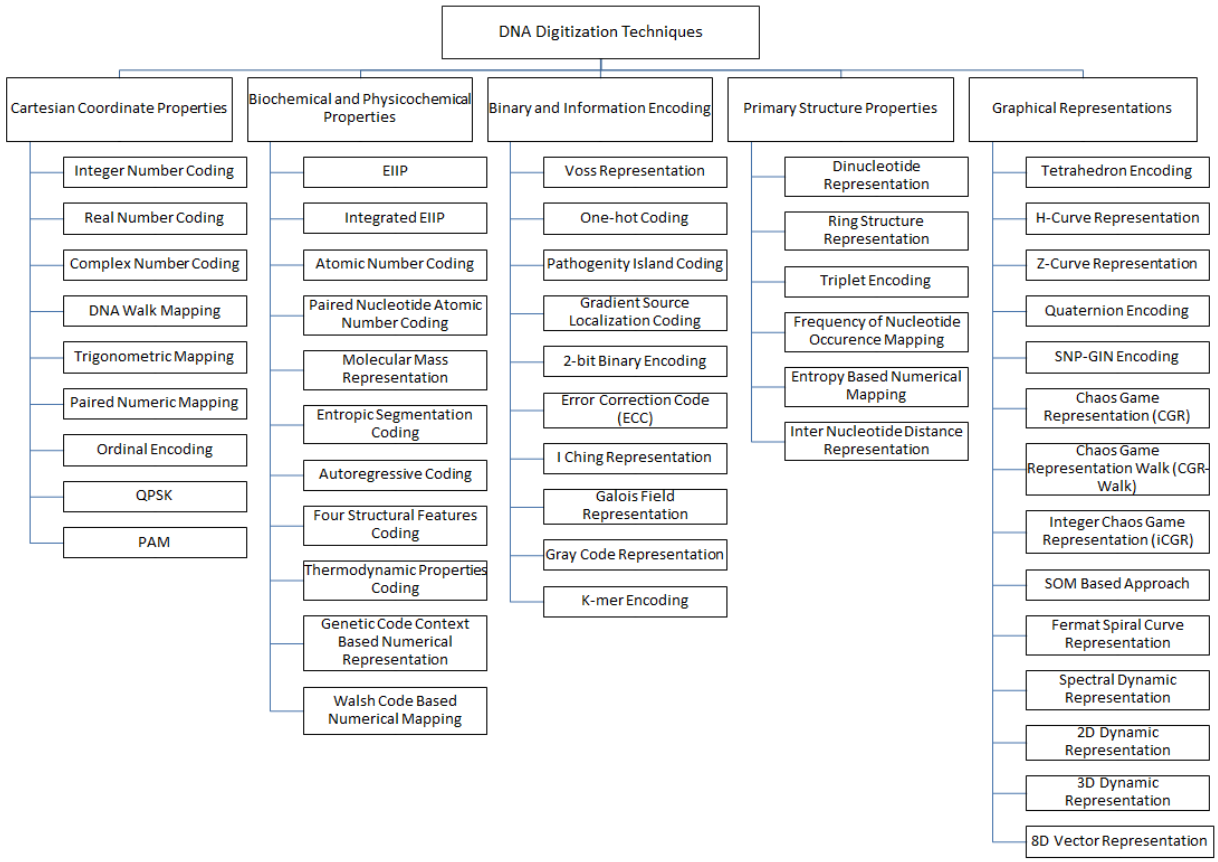


Figure 2 The hierarchical scheme of all DNA mapping techniques

2.1 Cartesian-Coordinate Properties Group

The first group of DNA numerical mapping techniques is cartesian coordinate properties (CCP) digitization techniques. Within this group, there are nine numerical coding techniques namely Integer Number Coding, Real Number Coding, Complex Number Coding, DNA Walk Mapping, Trigonometric Mapping, Paired Numeric Coding, Ordinal Encoding, QPSK (Quadrature Phase Shift Keying), PAM (Pulse Amplitude Modulation) are examined. Table 2 provides a brief summary of all the mapping techniques in the CCP group.

Table 2 The summary of all numerical coding techniques in cartesian-coordinate properties group

The name of technique	Coding Scheme	Numerical Representation	Definition
Integer Number Coding [6,15]	If Purin>Pirimidin T=0, C=1, A=2, G=3 T>A ve G>C ise A=0, C=1, T=2, G=3	X=[AGCTACCGTG] $\hat{X}$=[2, 3, 1, 0, 2, 1, 1, 3, 0, 3]	Nucleotides are represented by integers.
Real Number Coding [6,15]	A=-1.5, T=1.5, C=0.5, G=-0.5	X=[AGCTACCGTG] $\hat{X}$=[-1.5, -0.5, 0.5, 1.5, -1.5, 0.5, 0.5, -0.5, 1.5, -0.5]	Nucleotides are represented by real numbers.
Complex Number Coding [9]	A= -1, C= -j, G= j, T= 1	X=[AGCTACCGTG] $\hat{X}(i)$ = [*1, –1, j, j, 1, –j, –j, –1, j, –1*]	Nucleotides are represented by complex numbers.
DNA Walk Coding [6]	Integer temsil; A= -1, C= 1 G=- 1, T= 1 Complex temsil;	X=[AGCTACCGTG] $\hat{X}(i)$ = [-1, -2, -1, 0, -1, 0, 1, 0, 1, 0]	Nucleotides are encoded by assigning integer or complex numbers and summing their

	A=1, C= -j G= -1, T= j		values along the DNA sequence.
Trigonometric Mapping Coding [3,22]	A= cos(θ) + j ×sin(θ) C= -cos(θ) - j ×sin(θ) G= -cos(θ) + j ×sin(θ) T= cos(θ) - j ×sin(θ)	X=[AGCTACCGTG] $\hat{X}(i)$ = [0.5+0.8660i, -0.5+0.8660i, -0.5+0.8660i,0.5+0.8660i, 0.5+0.8660i, -0.5+0.8660i, -0.5+0.8660i, -0.5+0.8660i 0.5+0.8660i, -0.5+0.8660i]	Nucleotides are encoded by assigning trigonometric equations.
Paired Numeric Coding [15]	Purin(A&G)= 1 Pirimidin(C&T)= -1	X=[AGCTACCGTG] $\hat{X}$ = [1, 1, -1, -1, 1, -1, -1, 1, -1, 1]	Nucleotides are encoded by assigning values according to their structural properties.
Ordinal Encoding [23]	A= 0.25, C= 0.50 G= 0.75, T= 1.00	X=[AGCTACCGTG] $\hat{X}(i)$ = [0.25, 0.75, 0.50, 1.00, 0.25, 0.50, 0.50, 0.75, 1.00, 0.50]	Nucleotides are assigned sequential, linear values.
QPSK [24]	A= 1+j, G= -1+j C= -1-j, T= 1-j	X=[AGCTACCGTG] $\hat{X}(i)$ = [1+j, -1+j, -1-j, 1-j, 1+j, -1-j, -1-j, -1+j, 1-j, -1+j]	2D QPSK constellation complex number values are assigned according to the complementary property of DNA.
PAM [6,25]	A= -1.5, G= -0.5 C= 0.5, T= 0.5	X=[AGCTACCGTG] $\hat{X}(i)$ = [-1.5, -0.5, 0.5, 0.5, -1.5, 0.5, 0.5, -0.5, 0.5, -0.5]	Nucleotides are represented by 1D real numbers.

DNA sample datasets in Genbanks are available in Fasta format and are analog signals. Digitized signal representations of the first 100 bases of the sequence with reference number NR_131216.1 retrieved from the NCBI database, with coding techniques in the "Cartesian-Coordinate Properties (CCP)" group are shown in Figure 3.

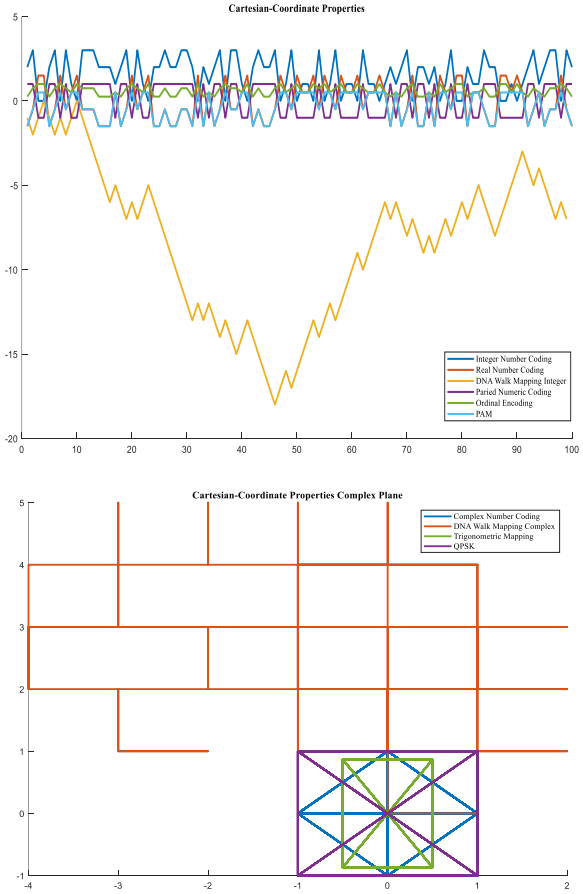


Figure 3 Numerical representations of CCP group techniques

2.2 Biochemical and Physicochemical Properties Group

The second group of DNA numerical mapping techniques is the biochemical and physicochemical (BPP) numerical techniques. Within this group, eleven coding schemes such as EIIP, Integrated EIIP, Atomic Number Coding, Paired Nucleotide Atomic Number Coding, Molecular Mass Representation, Entropic Segmentation Coding, Autoregressive Coding, Four Structural Features Coding, Thermodynamic Properties Coding, Genetic Code Context-Based Numerical coding, Walsh Code Based Numerical Mapping are examined. Table 3 provides a brief summary of all the mapping techniques in the BPP group.

Table 3 The summary of all numerical coding techniques in biochemical and physicochemical properties group

The name of technique	Coding Scheme	Numerical Representation	Definition
EIIP coding [3,6,9]	C=0.1340, T=0.1335, A=0.1260, G=0.0806	X=[AGCTACCGTG] $\hat{X}$ = [0.1260, 0.0806, 0.1340, 0.1335, 0.1260, 0.1340, 0.1340, 0.0806, 0.1335, 0.0806]	Energy values are assigned to the nucleotides
Integrated EIIP coding [26]	EIIP codes for 64 codons	X=[AGCTACCGTG] $\hat{X}$ = [0.3406, 0.3481, 0.3935, 0.3935, 0.3940, 0.3486, 0.3935, 0.2947]	EIIP energy values are assigned to DNA codons.
Atomic number coding [6]	C= 58, T= 66, A= 70, G= 78	X=[AGCTACCGTG] $\hat{X}$ = [70, 78, 58, 66, 70, 58, 58, 78, 66, 78]	Atomic numbers are assigned to nucleotides.
Paired nucleotide atomic coding [9]	A&G= 62, C&T= 42	X=[AGCTACCGTG] $\hat{X}(i)$ = [62, 62, 42, 42, 62, 42, 42, 62, 42, 62]	Atomic numbers are assigned to paired nucleotides.
Molecular mass coding [9,15]	C= 110, G= 150, A= 134, T= 125	X=[AGCTACCGTG] $\hat{X}(i)$ = [134, 150, 110, 125, 134, 110, 110, 150, 125, 150]	Molecular mass values are assigned to nucleotides.
Entropic segmentation coding [9,27,28]	12-Symbol alphabet A_1, A_2, A_3, C_1, C_2,C_3, G_1, G_2, G_3, T_1, T_2, T_3 Calculates entropy by array	X=[AGTTAGTGCT] $\hat{X}(i)$ = [A_1 G_2 S_3 T_3 S_1 T_1 A_2 G_3 T_1 G_2 C_3]	DNA segments and stop codons are represented by the 18-symbol alphabet.
Autoregressive coding [9,29]	Propeller Twist ve DNA Bending Stiffness values for dinükleotides	X=[AGCTACCGTG] Propeller Twist $\hat{X}(i)$ = [-14.00, -11.08, -14.00, -11.85, -13.10, -8.10, -10.03, -13.10, -9.45] Bending Stiffness $\hat{X}(i)$ = [60, 85, 60, 20, 60, 130, 85, 60, 60]	According to the structural properties of DNA, propeller twist and DNA bending stiffness values and dinucleotides are coded.
Four Structural features coding [30]	DNA Bending Stiffness, Dublex Disrupt Energy, Dublex Free Energy, Propeller Twist values for dinükleotides	X=[AGCTACCGTG] DNA bending stiffness $\tilde{x}_\alpha(n)$ = 60, 60, 60, 85, 60 Dublex disrupt energy $\tilde{x}_\beta(n)$ = 16, 16, 13, 36, 19 Dublex free energy $\tilde{x}_\gamma(n)$ = -15, -15, -15, -28, -17 Propeller twist $\tilde{x}_\delta(n)$ = -1400, -1400, -1310, -1003, -94	According to the four physical properties of DNA, the coding for the dinucleotide is performed according to the propeller twist value, DNA bending stiffness, duplex disrupts energy, and duplex free energy values.
Thermodynamic properties coding [6,31]	TC=5.6, GA=5.6, CA=5.8, TG=5.8, TA=6.0, AC=6.5, GT=6.5, CT=7.8, AG=7.8, AT=8.6, TT=9.1, AA=9.1,	X=[AGCTACCGTG] $\hat{X}(i)$ = [7.8, 11.1, 7.8, 6.0, 6.5, 11.0, 11.9, 6.5, 5.8]	Coding is performed by assigning enthalpy values of thermodynamic interactions of nucleotides.

	CC=11.0,GG=11.0, GC=11.1, CG=11.9		
Genetic code context (GCC) based numerical coding [13]	Assignment of GCC-based complex number representations to amino acids (Table 8)	X=[AGCTACCGTG] Birinci çerçeve AGC TAC CGT İkinci çerçeve GCT ACC GTG $\hat{X}(i)$ = [0.05 + 88.7i, 0.6 + 88.3i, 1.88 + 193i, 0.06 + 125.1i, 0.60 + 181.2i, 1.32 + 141.4i]	The DNA sequence is read with a reading frame as triplet codons. The sequence is digitized by assigning complex number values to amino acids.
Walsh Code Based coding [5]	A=W_A=0000 T=W_T=0011 G=W_G=0101 C=W_C=0110	X=[AGCTACCGTG] $\hat{X}(i)$= [00000101 011000110000 011001100101 00110101]	The fourth-order Walsh codes are assigned to nucleotides

Digitized signal representations of the sample DNA sequence (NR_131216.1) with EIIP and Integrated EIIP techniques are shown in Figure 4.

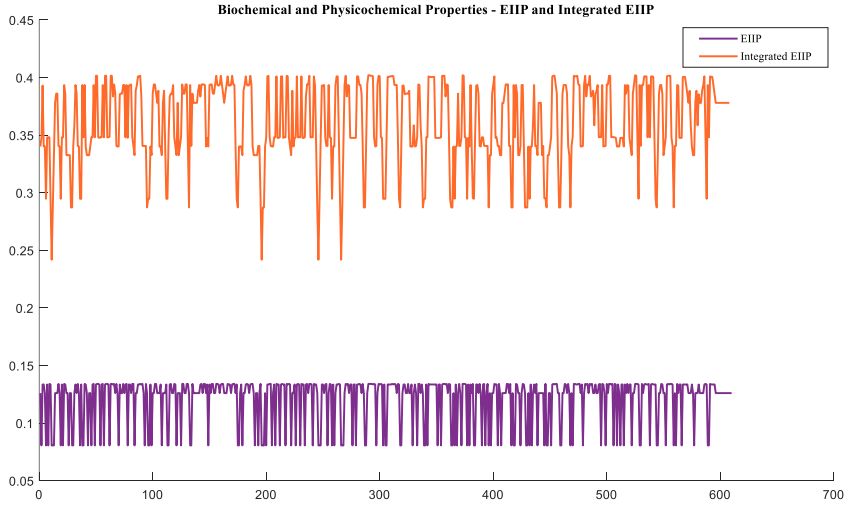


Figure 4 Numerical representations of EIIP and Integrated EIIP techniques

Digitized signal representations with the four structural features coding technique are shown in Figure 5.

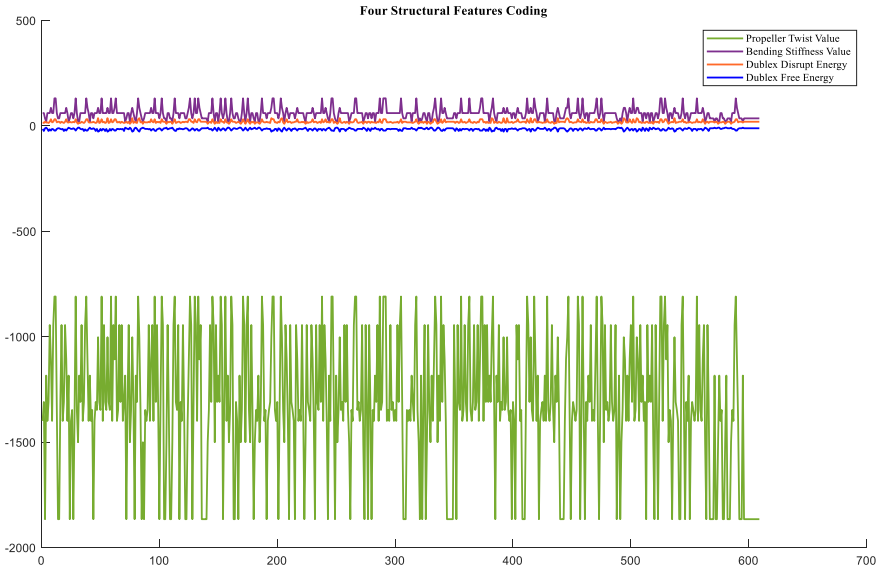


Figure 5 Numerical representations of the four structural features coding

Figure 6 gives the digitized signal plot of the reference sequence NR_131216.1 using the GCC-based coding technique.

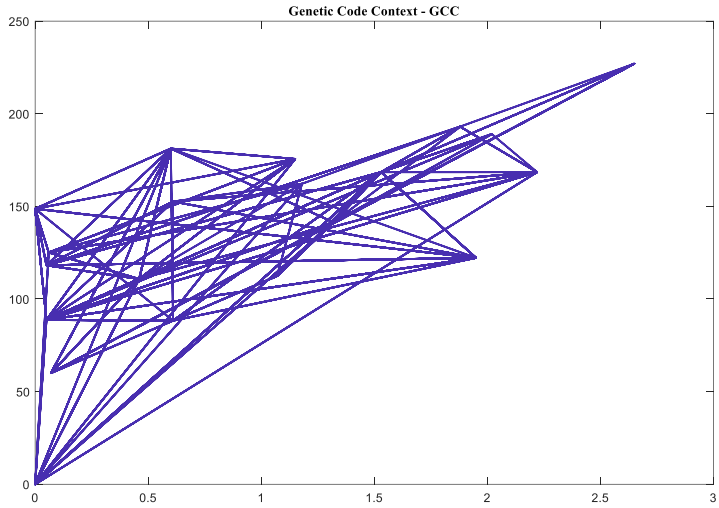


Figure 6 Numerical representations of the GCC coding

Digitized signal representations of the first 100 bases of the NR_131216.1 reference numbered sequences by coding techniques in the "Biochemical and Physicochemical Properties (BPP)" group are shown in Figure 7.

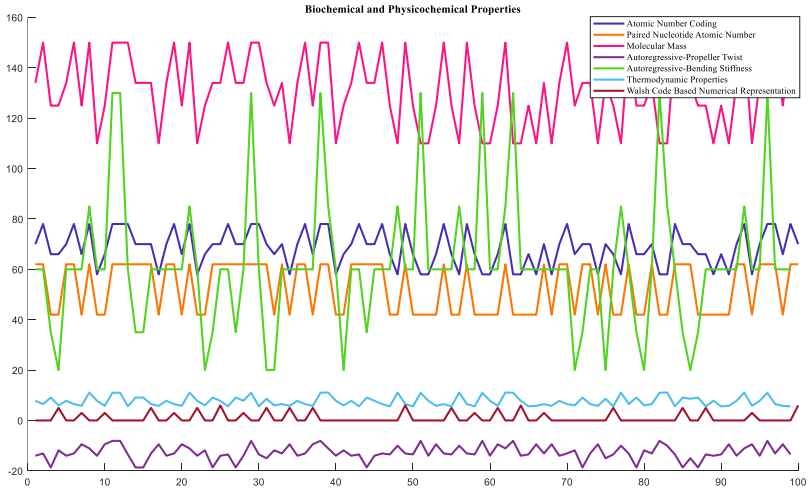


Figure 7 Numerical representations of BPP group techniques

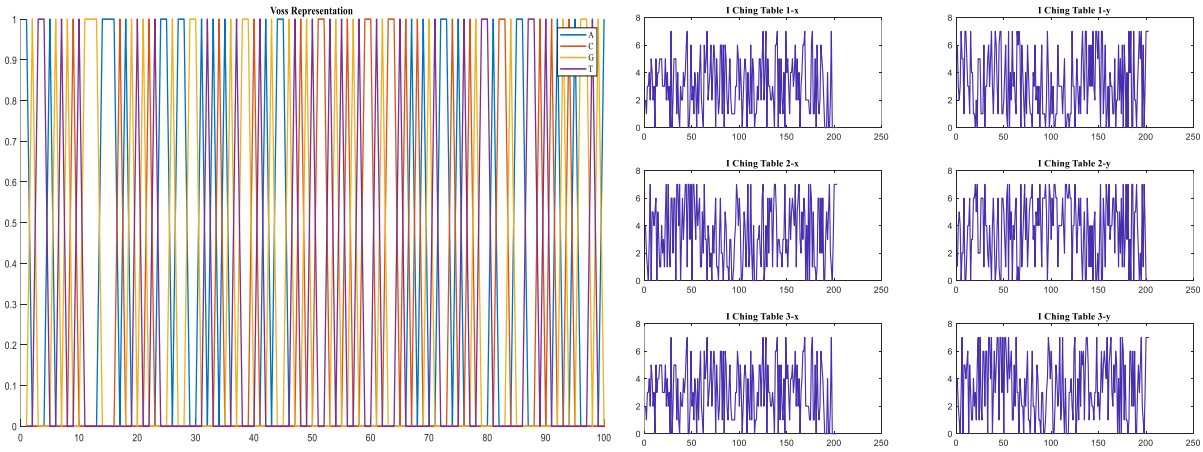
2.3 Binary and Information Encoding Group

The third group of DNA numerical mapping techniques is cartesian coordinate (CCP) digitization techniques. Within this group, ten coding techniques as Voss Representation, One-Hot Coding, Pathogenicity Island Coding, Gradient Source Localization Coding, 2-bit Binary Encoding, Error Correction Code (ECC), I Ching Representation, Galois Field Representation, Gray Code Representation, K-mer Encoding are examined. Table 4 provides a brief summary of all the mapping techniques in the "Binary and Information Encoding" group.

Table 4 The summary of all numerical coding techniques in binary and information encoding group

The name of technique	Coding Scheme	Numerical Representation	Definition
Voss Coding [6]	S=[C, G, A, T], Cn=[1,0,0,0], Gn=[0, 1, 0, 0], An=[0, 0, 1, 0], Tn=[0, 0, 0, 1]	X=[AGCTACCGTG] A_{20} = [1000100000] G_{20} = [0100000101] C_{20} = [0010011000] T_{20} = [0001000010]	By creating four sequences, binary values are assigned according to the presence or absence of each base.
One-hot Coding [32]	A:1,0,0,0 T:0,1,0,0 C:0,0,1,0 G:0,0,0,1	X=[AGCTACCGTG] $\hat{X}(i)$ = [1000, 0001, 0010, 0100, 1000, 0010, 0010, 0001, 0100, 0001]	Nucleotides are encoded with four-bit binary values.
Pathonetity Island Coding [9,33]	C&G= 1, A&T= 0	X=[AGCTACCGTG] $\hat{X}(i)$ = [0110011101]	Binary values are assigned according to the presence and absence of pathogenicity islands.
Gradient Source Localization Coding [9]	A= 0, C=1, G=3, T= 2	X=[AGCTACCGTG] $\hat{X}(i)$ = [0, 3, 1, 2, 0, 1, 1, 3, 2, 3]	Integer values are assigned to nucleotides based on gradient source localization
2-bit Binary Encoding [13]	A= 00, G= 10, T= 01, C= 11	X=[AGCTACCGTG] $\hat{X}(i)$ = [00, 10, 11, 01, 00, 11, 11, 10, 01, 10]	Two-bit binary values are assigned to the nucleotides
Error Correction Coding [6,34]	x, y coordinates values according to group codons like Purine-Pyrimidine, Weak-Strong H bond, Amino-Keto	X=[AGCTACCGTG] Purine-Prymidine $\hat{X}(i)$ = [(1,1), (2,6), (3,5), (4,3), (5,6), (6,5), (7,3), (8,7)] Weak-Strong H bond $\hat{X}(i)$ = [(1,1), (2,2), (3,4), (4,7), (5,1), (6,3), (7,2), (8,5)] Amino-Keto $\hat{X}(i)$ = [(1,3), (2,7), (3,3), (4,6), (5,4), (6,5), (7,2), (8,0)]	Nükleotidlere biyokimyasal özelliklerine göre x ve y koordinat değerleri atanır
IChing Coding [6,35]	Binary coding with 3 different I Ching tables according to amino acids	X=[AGCTACCGTG] $\hat{X}(i)$ = [110 100 001 010 100 001 010 101]	Coding is performed with I Ching tables created according to the three biochemical properties of nucleic acids.
Galois Field Coding [36]	0=0 ⇔ 0 ⇔A, x^0=1 ⇔1⇔C, x^1=x ⇔ 2 ⇔T, x^2=x+1 ⇔ 3 ⇔ G	X=[AGCTACCGTG] $\hat{X}(i)$ = [0, 3, 1, 2, 0, 1, 1, 3, 2, 3]	Nucleotides are assigned numerical values corresponding to their quadratic polynomial representation.
Gray Code Coding [36]	A= 00, T=01, C=10, G= 11	X=[AGCTACCGTG] $\hat{X}(i)$ = [0010011100 0101101110]	Two-bit binary codes are assigned to nucleotides by ex-or operation.
K-mer Encoding [37,38]	1-mer coding A → [1,0,0,0] C → [0,1,0,0] G → [0,0,1,0] T → [0,0,0,1]	X=[AGCTACCGTG] $\hat{X}(i)$ = [1000, 0010, 0100, 0001, 1000, 0100, 0100, 0010, 0001, 0010]	The DNA sequence is split into k-mer degments and coded with zeros and ones.

Figure 7(a) gives the digitized signal plot of the sample sequence using the Voss coding technique and Figure 7(b) IChing coding technique.



(a) The Voss (b) IChing

Figure 7 Numerical representations of Voss and IChing coding techniques

Figure 8 gives the digitized signal representations of the first 100 bases of the NR_131216.1 reference numbered sequences with the coding techniques in the “Binary and Information Encoding” group.

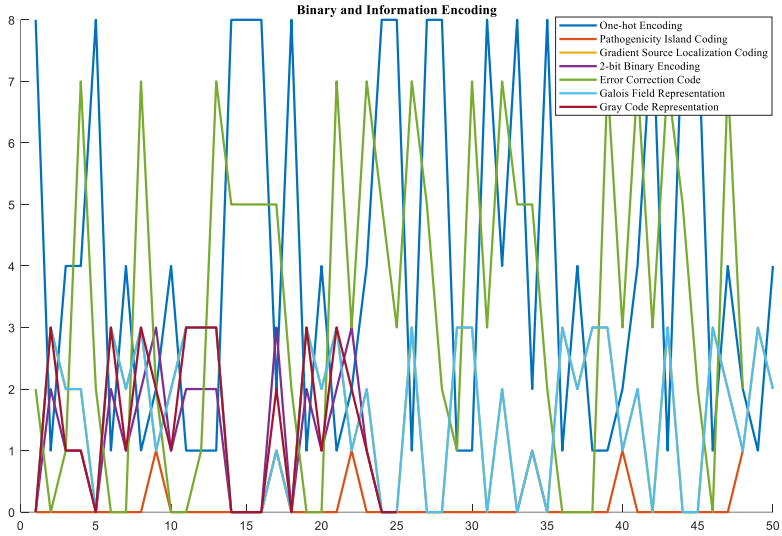


Figure 8 Numerical representations of BIE group techniques

2.4 Primary Structure Properties Group

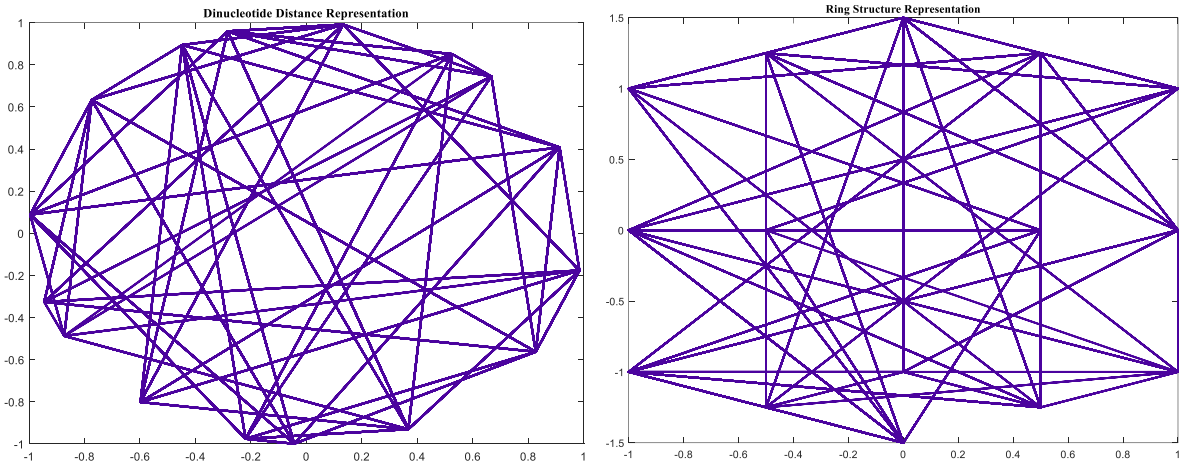
The fourth group of DNA numerical mapping techniques are primary structure feature (PSP) digitization techniques. In this group, six coding techniques are examined, namely Dinucleotide Representation, Ring Structure Representation, Triplet Encoding, Frequency of Nucleotide Occurence Mapping, Entropy Based Numerical Mapping, Inter Nucletide Distance Representation. Table 5 provides a brief summary of all the mapping techniques in the "Primary Structure Properties" group.

Table 5 The summary of all numerical coding techniques in Primary Structure Properties group

The name of technique	Coding Scheme	Numerical Representation	Definition
Dinucleotide Representation [39,40]	16 dinucleotides are placed on the unit circle and coded according to their positions.	X=[AGCTACCGTG] $\hat{X}(i)$ = [(cos(π/2), sin(π/2)), (Cos(13π/4), sin(13π/4)), (Cos(15π/8), sin(15π/8)), (Cos(3π/4), sin(3π/4)), (Cos(5π/8), sin(5π/8)),	Dinucleotides are distributed evenly around a circle and coded with their coordinate values.

		(Cos(π/8), sin(π/8)), (cos(2π), Sin(2π)), (cos(11π/8), Sin(11π/8)), (cos(π), sin(π))]	
Ring Structure Representation [6,41]	AG: (0, 1.5), CT: (0, -1.5), CA:(1,1), TG: (-1, -1), CG: (1, -1), TA: (-1, 1), GA: (1, 0),GT(0.5, -1.25), GC: (-0.5, 1.25), TC: (-1, 0), AC: (-0.5, 1.25), AT: (0.5, 1.25), AA: (0, 1), TT: (0.5, 0), GG: (0,1), CC: (-0.5, 0).	X=[AGCTACCGTG] $\hat{X}(i)$ = [(0, 1.5), (0,-1.5), (-0.5, 1.25), (1, -1), (-1, -1)]	Dinucleotides are placed at the corners of the hexagon according to the six groups they are divided into according to their biochemical properties, and six coding schemes are obtained with six different combinations and coded with the corner coordinates of the hexagon.
Triplet Encoding [6,42]	64 codons are encoded by weights	X=[AGCTACCGTG] $\hat{X}(i)$ = [15.6, 1.1, 11.3, 18.2, 16.4, 14.4, 2.1, 19.4]	Nucleotide triplets and amino acid codons are quantified by weight.
Frequency of Occurence Mapping [6,13]	C=0.27215, T=0.2056, A=0.24300, G=0.27909 or CG:0.01, GC: 0.043, CC: 0.047, GT:0 .049, GG: 0.050, AC: 0.054, TC: 0.057, GA: 0.061, TA:0.067, AG: 0.070, CT: 0.071, TG: 0.074, CA: 0.074, AT: 0.081, AA: 0.097, TT: 0 .097	X=[AGCTACCGTG] $\hat{X}(i)$ = [0.070, 0.071, 0.054, 0.01, 0.074]	Nucleotides or dinucleotides are coded with frequency values according to their frequency of occurrence.
Entropy Based Numerical Mapping [43]	Entropy values calculated according to the new formulas are assigned to 64 codons.	X=[AGCTACCGTG] $\hat{X}(i)$ = [0.7222, 0.8331, 0.9086, 0.8331, 0.8118, 0.5363, 0.6259, 0.9818, 0.9998, 0.9954]	The codons are coded by calculating the entropy values of the modified and fractional new equation of Shannon's entropy equation.
Inter Nucleotide Distance Representation [44]	Each base is encoded with the value of the base distance between the next itself and the same base.	X=[AGCTACCGTG] $\hat{X}(i)$ = [4, 6, 3, 5, 5, 1, 3, 2, 1, 0]	Each base in the DNA sequence is encoded with the base distance value between it and the same base that follows it.

Figure 9(a) gives the digitized signal plot of the reference sequence NR_131216.1 using the Dinucleotide distance coding technique and Figure 9(b) The Ring Structure technique.



(a) Dinucleotide Distance (b) Ring Structure

Figure 9 Numerical representations of Dinucleotide distance coding and the Ring Structure techniques

Figure 10 gives the digitized signal plot of the sample sequence using the Frequency of Nucleotide Occurrence coding technique.

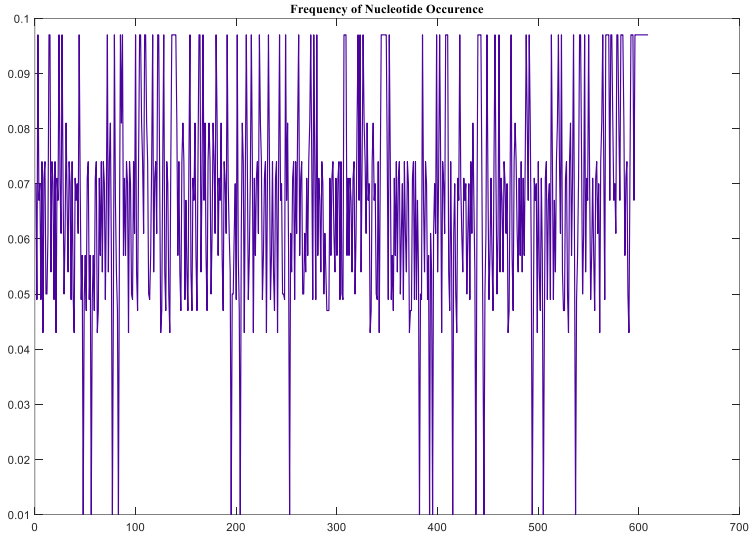


Figure 10 Numerical representation of Frequency of Nucleotide Occurrence Coding

Figure 11 gives the digitized signal representations of the first 100 bases of the NR_131216.1 reference numbered sequences with the coding techniques in the “Primary Structure Properties” group.

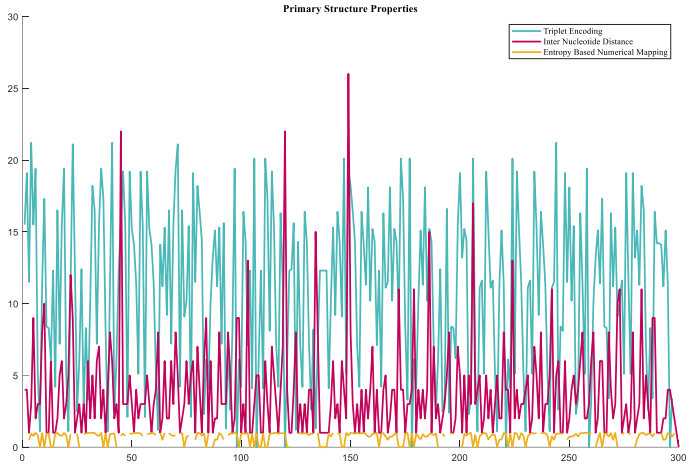


Figure 11 Numerical representation of Primary Structure Properties Group

2.5 Graphical Representation Group

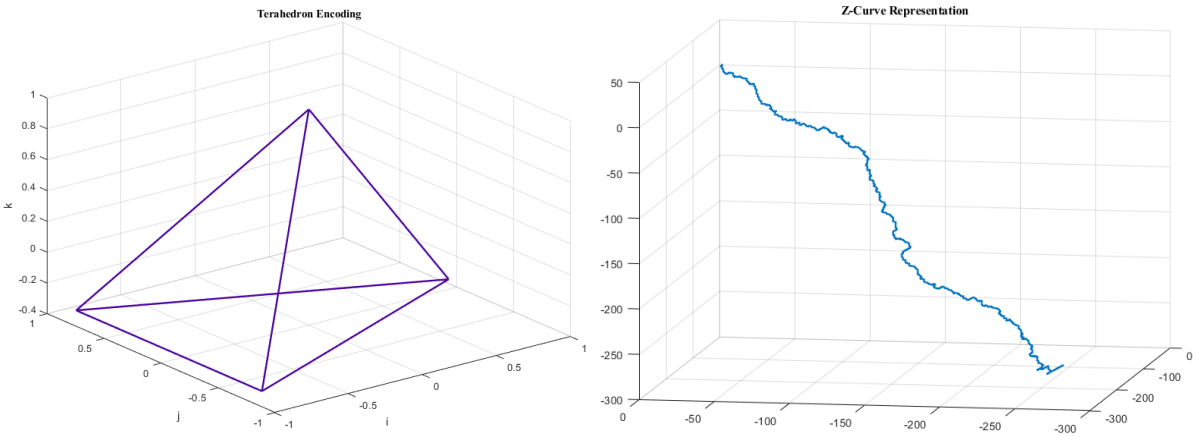
The fifth group of DNA numerical mapping techniques is Graphical Representation (GR) digitization techniques. Within this group, Tetrahedron Encoding, H-Curve Representation, Z-Curve Representation, Quaternion Encoding, SNP-GIN Encoding, Chaos Game Representation, (CGR), Chaos Game Representation Walk (CGR-Walk), Integer Chaos Game Representation (iCGR) Fourteen coding techniques are examined, namely, Som Based Approach, Fermat Spiral Curve Representation, Spectral Dynamic Representation, 2D Dynamic Representation, 3D Dynamic Representation, 8D Dynamic Representation. Table 6 provides a brief summary of all the mapping techniques in the "Graphical Representation" group.

Table 6 The summary of all numerical coding techniques in Graphical representation group

The name of technique	Coding Scheme	Numerical Representation	Definition
Tetrahedron Encoding [6,45]	A= k, G= $-\frac{2\sqrt{2}}{3}i-\frac{\sqrt{6}}{3}j-\frac{1}{3}k$, C= $-\frac{2\sqrt{2}}{3}i+\frac{\sqrt{6}}{3}j-\frac{1}{3}k$, T= $\frac{2\sqrt{2}}{3}i-\frac{1}{3}k$	*X*=[AGCTACCGTG] $\hat{X}_1$= $\left[0,-\frac{\sqrt{2}}{3},-\frac{\sqrt{2}}{3},\frac{2\sqrt{2}}{3},0,-\frac{\sqrt{2}}{3},-\frac{\sqrt{2}}{3},-\frac{\sqrt{2}}{3},\right.$ $\left.\frac{2\sqrt{2}}{3},-\frac{\sqrt{2}}{3}\right]$, $\hat{X}_2$ = $\left[0,-\frac{\sqrt{6}}{3},\frac{\sqrt{6}}{3},0,0,\frac{\sqrt{6}}{3},\frac{\sqrt{6}}{3},-\frac{\sqrt{6}}{3},0,\frac{\sqrt{6}}{3}\right]$, $\hat{X}_3$ = $\left[1,-\frac{1}{3},-\frac{1}{3},-\frac{1}{3},1,-\frac{1}{3},-\frac{1}{3},-\frac{1}{3},-\frac{1}{3},-\frac{1}{3}\right]$	Nucleotides are placed at the four corners of the tetrahedron and coded by the numerical equations of the corners.
H-Curve Representation [6,46]	A = $\frac{1}{2}i-\frac{\sqrt{3}}{2}j$, T = $\frac{1}{2}i+\frac{\sqrt{3}}{2}j$, C = $\frac{\sqrt{3}}{2}i+\frac{1}{2}j$, G = $\frac{\sqrt{3}}{2}i-\frac{1}{2}j$	*X*=[AGCTACCGTG] $\hat{X}_1$ = [-0.3660i, 0.3660i, 1.3660i, 1.3660i, -0.3660i, 1.3660i, 1.3660i, 0.3660i, 1.3660i, 0.3660i]	Nucleotides are encoded with functions created by vectors i, and j on the x, and y axes.
Z-Curve Representation [6,47]	$\hat{X}_1(i)\begin{cases}X(i-1)+1 & if X(i)=\\ X(i-1)+(-1) & other\end{cases}$ $\hat{X}_2(i)\begin{cases}X(i-1)+1 & if X(i)=\\ X(i-1)+(-1) & other\end{cases}$ $\hat{X}_3(i)\begin{cases}X(i-1)+1 & if X(i)=\\ X(i-1)+(-1) & other\end{cases}$	*X*=[AGCTACCGTG] $\hat{X}_1$ = [-1, 0, -1, 0, -1, -2, -3, -2, -1, 0] $\hat{X}_2$ = [1, 0, 1, 0, 1, 2, 3, 2, 1, 0] $\hat{X}_3$ = [1, 0, -1, 0, 1, 0, -1, -2, -1, -2]	Nucleotides are encoded by sets of vectors towards the four faces of the tetrahedron.
Quaternion Encoding [6,18]	A= i+j+k, C=-i+j-k, G=-i-j+k, T=i-j-k	*X*=[AGCTACCGTG] $\hat{X}(i)$= [*i+j+k, -i-j+k, -i+j-k, i-j-k, i+j+k, -i+j-k, -i+j-k, -i-j+k, i-j-k, -i-j+k*]	Nucleotides are represented by 4D quaternion equations.
SNP GIN Encoding [49]	1000 → A|A veya A|- 0100 → C|C veya C|- 0010 → G|G veya G|- 0001 → T|T veya T|- 1100 → A|C, 1010 → A|G 1001 → A|T, 0110 → C|G 0101 → C|T, 0011 → G|T	*X*=[AGCTACCGTG] $\hat{X}(i)$ =>Genotypes → A|G C|T A|C C|G T|G Nucleotides → AGCTACCGTG Binary Encoding → 10100101110001100000 Hexadecimal → A5C60	GINs are created by assigning four pairs of SNPs to nucleotides.
Chaos Game Representation (CGR) [6,50]	A: (0, 0), T: (1, 0), G: (1, 1), C: (0, 1)	*X*=[AGCTACCGTG] $\hat{X}_1$ = [(0,0), (0,0), (0.5, 0.5), (0.25, 0.75), (0.625, 0.375), (0.3125, 0.1875), (0.1563, 0.5938), (0.0781, 0.7969), (0.5391, 0.8984), (0.7695, 0.4492)]	The DNA sequence is mapped according to the coordinate values within the unit square.
Chaos Game Representation Walk (CGR-Walk) [51,52]	CGR_{RY}: A(0, 0),T(1, 0),C(0, 1),G(1, 1) CGR_{MK}: A(0, 0),T(1, 0),G(0, 1),C(1, 1) CGR_{WS}: A(0, 0),G(1, 0),C(0, 1),T(1, 1)	*X*=[AGCTACCGTG] Purine-Prymidine $\hat{X}_1$ = [[(0,0), (0,0), (0.5, 0.5), (0.25, 0.75), (0.625, 0.375), (0.3125, 0.1875), (0.1563, 0.5938), (0.0781, 0.7969), (0.5391, 0.8984), (0.7695, 0.4492)]] Amino-Keto $\hat{X}_1$ = [(0,0), (0,0), (0, 0.5), (0.5, 0.75), (0.75, 0.375), (0.375, 0.1875), 0.6875, 0.5938), 0.8438, 0.7969), (0.4219, 0.8984), 0.7109, 0.4492)] Weak-Strong $\hat{X}_1$ = [(0,0), (0,0), (0.5, 0), (0.25, 0.5), (0.625, 0.75), (0.3125, 0.375), (0.1563, 0.6875),(0.0781, 0.8438), (0.5391, 0.4219), (0.7695, 0.7109)]	The chaos game is performed in the form of a DNA walk, taking into account the thermodynamic properties of representative DNA.

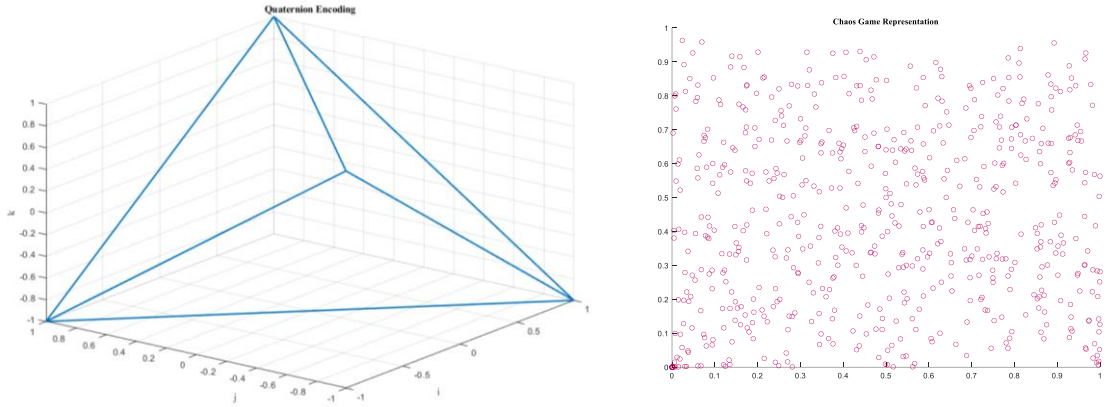
Integer Chaos Game Representation (iCGR) [53]	A=(1,1), T=(-1,1), C=(-1,-1), G=(1,-1)	*X*=[AGCTACCGTG] $\hat{X}_1$ = [(1,1), (3, -1), (-1,-5), (-9,3), (7,19), (-25,-13), (-89, -77), (39,-205), (-217, 51), (295, -461)]	The Chaos game representation is performed with integers rather than floating-point numbers.
SOM Based Approach [6,54]	A: (0, 0, 0), T: (0.289, 0.5, 0.816), C: (0.866, 0.5, 0), G: (0, 1, 0)	*X*=[AGCTACCGTG] $\hat{X}_1$ = [(0,0,0), (0,1,0), (0.866, 0.5, 0), (0.289, 0.5, 0.816), (0,0,0), (0.866, 0.5, 0), (0.866, 0.5, 0), (0,1,0), (0.289, 0.5, 0.816), (0,1,0)]	Nucleotides are paired with all four corners. Coding is performed with the distance values between the AG and CT vertices.
Fermat Spiral Curve Representation [55]	Representation of the four sub-strings formed according to the positions A, T, C, and G in the Fermat spiral	*X*=[AGCTACCGTG] $\hat{X}_1$ (Aseq)= [1,0,0,0,5,0,0,0,0,0] $\hat{X}_1$ (Gseq)= [0,2,0,0,0,6,0,8,0,0] $\hat{X}_1$ (Cseq)= [0,0,0,0,0,0,0,0,9,0] $\hat{X}_1$ (Tseq)= [0,0,3,4,0,0,7,0,0,10]	Global and local location information of nucleotides in DNA is mapped on the Fermat spiral.
Spectral Dynamic Representation [56]	Representation of the effusions of each base by a series of lines	*X*=[AGCTACCGTG] $\hat{X}_1$ (A)= [1,0,0,0,1,0,0,0,0,0] $\hat{X}_1$ (G)= [0,1,0,0,0,1,0,1,0,0] $\hat{X}_1$ (C)= [0,0,0,0,0,0,0,0,1,0] $\hat{X}_1$ (T)= [0,0,1,1,0,0,1,0,0,1]	The distributions of DNA nucleotides are represented by four separate split line plots.
2D Dynamic Representation [57]	A=(-1,0), G=(1, 0), C=(0, 1), T=(0,-1)	*X*=[AGCTACCGTG] $\hat{X}_1$ = [(-1,0), (0,0), (0,1), (0,0), (-1,0), (-1,1), (-1,2), (0,2), (0,1) (1,1)]	DNA sequences are represented by point masses in the 2D Euclidean space.
3D Dynamic Representation [57,58]	A=(-1, 0, 1), G=(1, 0, 1), C=(0, 1, 1), T=(0, -1, 1)	*X*=[AGCTACCGTG] $\hat{X}_1$ = [(-1,0,1), (0,0,2), (0,1,3), (0,0,4), (-1,0,5), (-1,1,6), (-1,2,7), (0,2,8), (0,1,9), (1,1,10)]	DNA/RNA sequences are represented in the 3D plane.
8D Vector Representation [59,60]	A=(1, 0.2), T=(1, -0.2), C=(1, 0.3), G=(1, -0.3) $z_i = y_i / i$ K =(m_z, v_z) $m_z = \frac{1}{n}\Sigma_{i=1}^{n}$ $v_z = \frac{1}{n}\Sigma_{i=1}^{n}(z_i - m_z)^2$	*X*=[AGCTACCGTG] $\hat{X}_1$ = [(1, 0.2), (2, -0.2), (3, 0.2), (4,0), (5, 0.2), (6, 0.5), (7, 0.8), (8, 0.5), (9, 0.3), (10, 0)] Slope=1.5370e-04 Variance=0.0182+0.0200i	8D vectors are formed with mean, variance values from a zigzag plot of DNA/RNA sequences

Figure 12 (a) gives the digitized signal plot of the sample sequence using the tetrahedron encoding technique and (b) Z-Curve representation. Figure 13(a) gives the digitized signal plot of the sample sequence using the quaternion encoding Figure 13(b) Chaos Game representation. Figure 14 provides a digitized signal plot of the sample sequence according to the thermodynamic properties of purine-pyrimidine, amino-keto, and strong-weak H bonds using the Chaos game representation walk technique. Figure 15 gives the digitized signal plot of the sample sequence using the SOM-based coding technique.



(a) Tedrahedron Encoding (b) Z-Curve

Figure 12 Numerical representations of Tedrahedron Encoding and Z-Curve techniques



(a) Quaternion (b) Chaos Game

Figure 13 Numerical representations of Quaternion Encoding and Chaos Game techniques

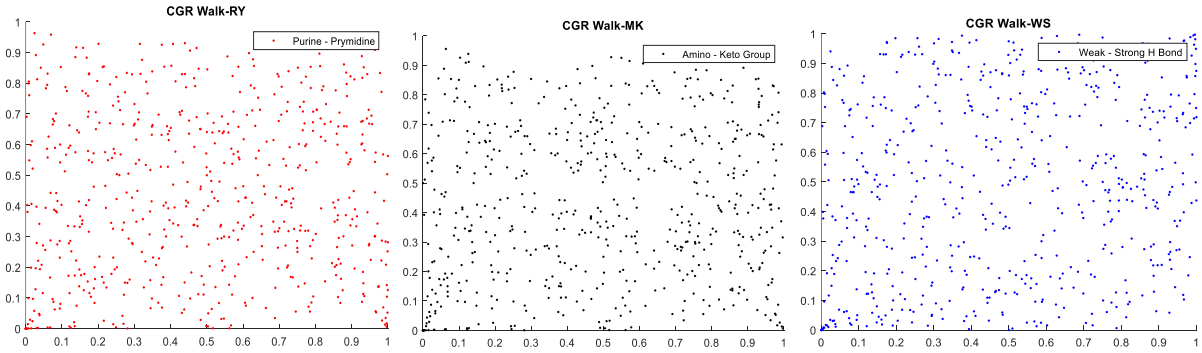


Figure 14 Numerical representation of Chaos Game Representation Walk (RY, MK, WS)

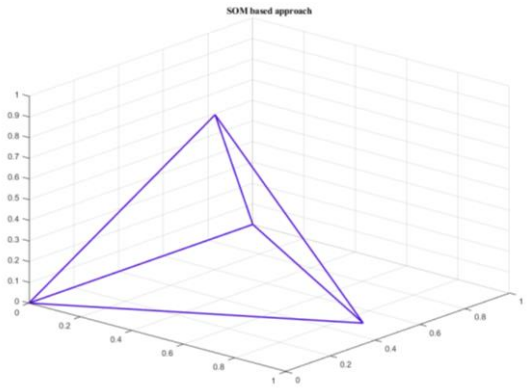


Figure 15 Numerical representation of SOM based coding

Figure 16 gives the digitized signal plot of the sample sequence using the Fermat spiral curve coding technique.

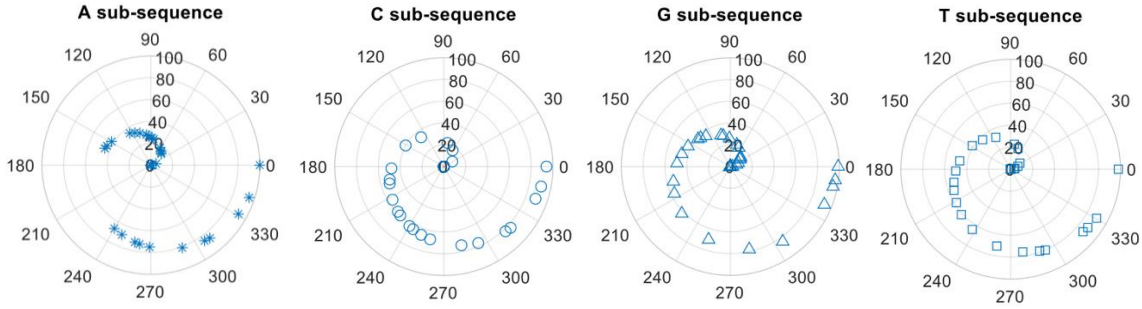


Figure 16 Numerical representation of Fermat spiral curve coding

Figure 17 gives a digitized signal plot of the sample sequence using the Spectral dynamic representation technique.

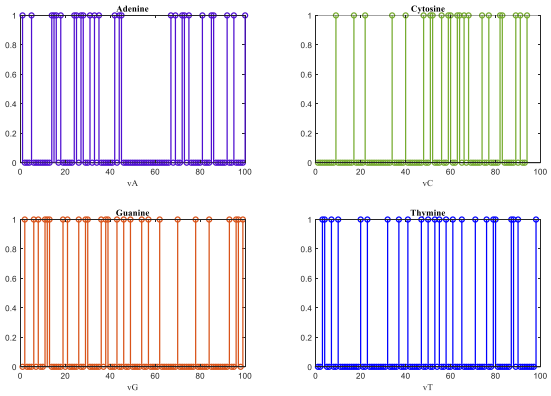
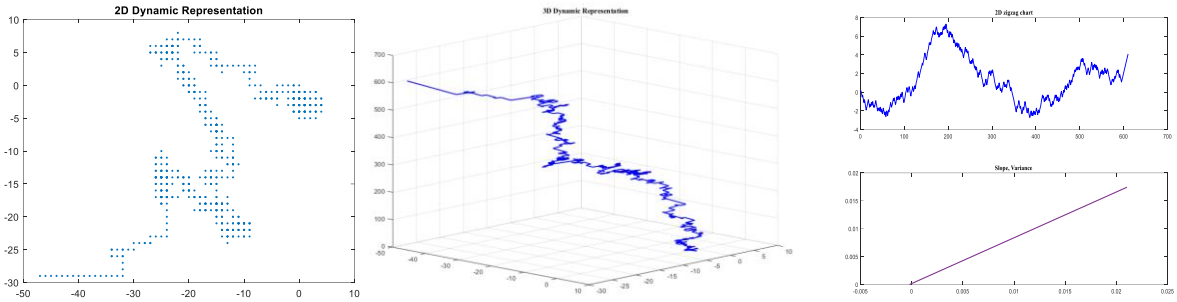


Figure 17 Numerical representation of Spectral dynamic representation

Figure 18(a) gives the digitized signal plot of the sample array using the 2D dynamic representation technique, Figure 18(b) 3D dynamic, Figure 18(c) 8D dynamic.



(a) 2D (b) 3D (c) 8D

Figure 18 Numerical representations of 2D, 3D and 8D dynamic represantations

3.Performance Comparison of Numerical Mapping Techniques in Genomic Fields

The aim of this review is to analyze how the digital mapping techniques (coding scheme -numerical methods), which are used for digitizing DNA sequences in bioinformatics studies and have gained popularity in recent years, affect performance in genomic fields. In this section, the frequency of use of DNA coding techniques for 4 popular genomic fields (identification of exon regions, exon-intron classification, phylogenetic analysis, gene detection) and the min-max range of the performances obtained using these techniques in that field was given. Figure 19 shows the most used numerical coding techniques for the identification of exon regions. Table 7 shows min-max performance intervals of the most used coding techniques for the identification of exon regions.

Figure 19 The most used coding techniques for the identification of exon regions

Table 7 Min-max performance intervals of the coding techniques in the identification of exon regions

Coding Technique	**Min-Max Performance Interval (%)**	**Coding Technique**	**Min-Max Performance Interval**	**Coding Technique**	**Min-Max Performance Interval**
Integer	36-81	EIIP	48-97	Voss	60-92
Real	50-81	IEEP	80-88	One-hot	70-90
Complex	64-75	Autoregressive	48-76	Binary	65-75
DNA-Walk	78-100	Four structural features	70-78	Pathogenity	60-70
Trigonometric	73-87	Thermodynamic	70-80	Gray code	65-80
Paired numeric	47-94	Genetic code context	66-79	K-mer	85-93
Frequency nucleotide occurence	81-100	Walsh code	83-91	8D vector	70-75
Z-curve	83-95	Galois field	65-82	Atomic number	39-86
Inter nucleotide distance	80-85	PAM	58-75	Entropic segmentation	72-86
Tetrahedron	71-79	Entropy-based	92-100	Gradient Source localization	65-80
Quaternion	70-75	Moleculer Mass	51-68	Paire dnucleotide atomic	80-86

As seen in Figure 19, the three most commonly used techniques for the identification of exon regions are EIIP, Paired numeric and Voss techniques, respectively. Looking at Table 7, it is seen that the performance values obtained in detecting exon regions with these three techniques are above 90%. However, although not used as often as these techniques in the literature, maximum 100% performance has been achieved in studies using Entropy-based, Frequency nucleotide occurrence, and DNA-Walk techniques. Therefore, this review is thought to increase the use of these techniques in the introduction of these techniques and in most genomic areas from now on. Figure 20 shows the most used numerical coding techniques for the classification of exon-intron.

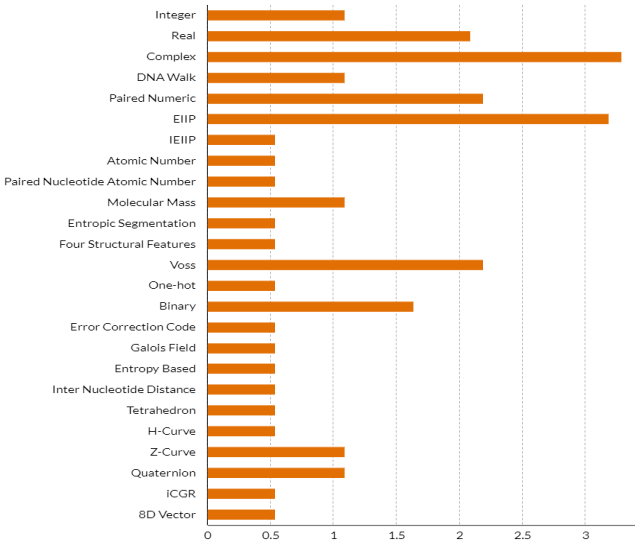


Figure 20 The most used coding techniques for the classification of exon-intron

Table 8 shows min-max performance intervals of the most used coding techniques for the classification of exon-intron.

Table 8 Min-max performance intervals of the coding techniques in the classification of exon-intron

Coding Technique	**Min-Max Performance Interval (%)**	**Coding Technique**	**Min-Max Performance Interval**	**Coding Technique**	**Min-Max Performance Interval**
Complex	60-80	DNA-Walk	67-95	Error correction code	60-70
Entropic segmentation	65-80	Integer	65-96	Entropy-based	92-96
8D vector	70-80	Real	38-61	Galois field	70-75
Z-curve	80-86	Binary	70-88	Paired nucleotide atomic	60-75
Voss	75-88	Paired-numeric	75-95	Moleculer mass	59-65
Tedrahedron	60-75	Quaternion	66-95	IEIIP	80-92
EIIP	85-95	Inter nucleotide atomic	75-85	Atomic number	58-76
Four structure featues	75-82	One-hot	64-80	H-curve	60-76

As seen in Figure 20, the three most commonly used techniques for the identification of exon regions are Complex, EIIP and Voss techniques, respectively. Looking at Table 8, while the performances of Complex and Voss techniques were above 80%, the performance increased up to 95% in studies conducted with the EIIP technique. Apart from these, Entropy-based, Integer, and EIIP techniques were used in studies with the highest performance in exon-intron classification. Figure 21 shows the most used numerical coding techniques for the phylogenetic analysis.

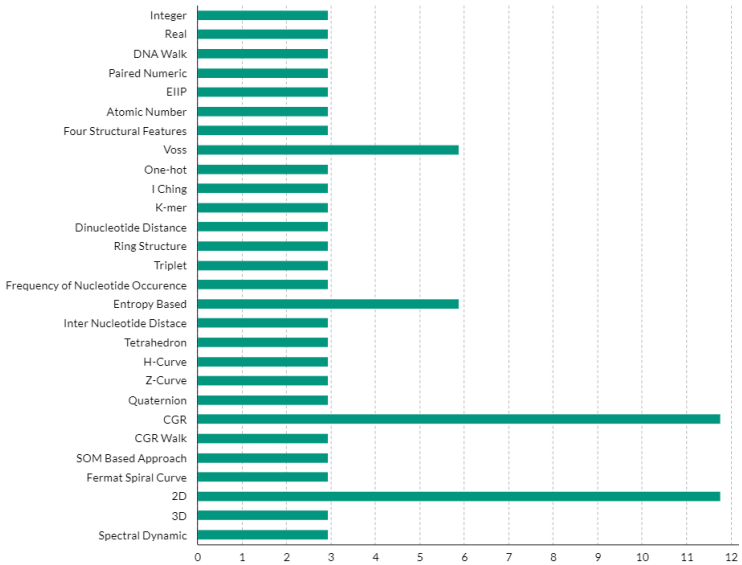


Figure 21 The most used coding techniques for the phylogenetic analysis

Table 9 shows min-max performance intervals of the most used coding techniques for the phylogenetic analysis

Table 9 Min-max performance intervals of the coding techniques in the phylogenetic analysis.

Coding Technique	**Min-Max Performance Interval (%)**	**Coding Technique**	**Min-Max Performance Interval**	**Coding Technique**	**Min-Max Performance Interval**
CGR	80-90	Voss	42-75	Quaternion	65-72
2D	78-90	Tetrahedron	36-70	Entropy	82-90
Fermat spiral	70-80	Z-curve	70-100	Four structural features	60-68
Integer	75-84	H-curve	70-80	CGR Walk	50-60
Real	80-100	DNA Walk	75-88	Sprectral Dynamic	55-65
EIIP	86-98	Dinucleotide	66-85	One-hot	75-80
Atomic number	80-98	Inter nucleotide distance	75-80	Inter nucleotide distance	55-72
Paired numeric	80-100	3D	80-90	Triplet encoding	75-83

As seen in Figure 21, the two most commonly used techniques for the phylogenetic analysis are CGR ve 2D techniques. Voss and Entropy-based techniques are the second most frequently used techniques after these. Looking at Table 9, Real and Z-Curve techniques were used in the highest performing studies for phylogenetic analysis. After these, EIIP and atomic number techniques were used in the studies with the highest performance. Figure 22 shows the most used numerical coding techniques for the detection of gene.

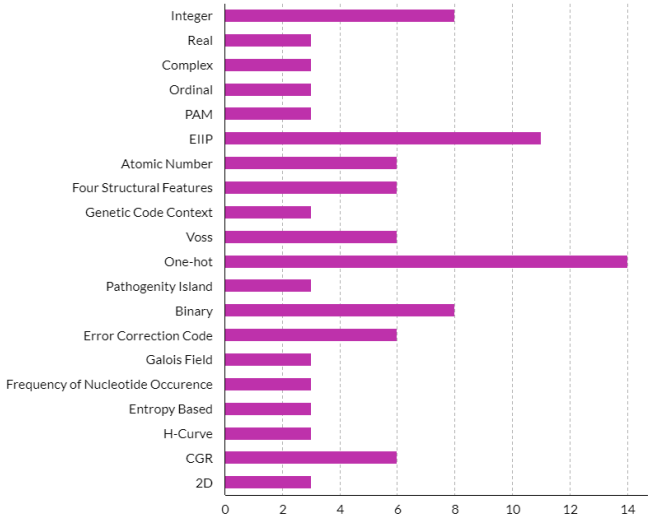


Figure 22 The most used coding techniques for the detection of gene

Table 10 shows min-max performance intervals of the most used coding techniques for the detection of gene.

Table 10 Min-max performance intervals of the coding techniques in the detection of gene.

Coding Technique	**Min-Max Performance Interval (%)**	**Coding Technique**	**Min-Max Performance Interval**	**Coding Technique**	**Min-Max Performance Interval**
CGR	70-83	Complex	65-75	One-hot	78-96
Four structural	70-75	Integer	63-99	Voss	80-98
2D	70-75	Real	80-97	Pathogenity Island	65-75
Error correction code	72-78	Binary	91-100	EIIP	61-100
H-curve	70-80	Galois Field	70-75	Atomic number	65-97
Entropy	80-100	Genetic code context	62-79	Frequency of nucleotide	72-100

As seen in Figure 23, the three most commonly used techniques for gene detection are One-hot, EIIP, and Integer techniques, respectively. Looking at Table 10, success performances of 96%, 100%, and 99%, respectively, were obtained in gene screening studies using these techniques. Apart from these, 100% maximum success performance has been achieved in studies using the frequency of nucleotide technique and the Entropy-based technique, although it is not used very often.

4. Conclusion

This study is an attempt to review the DNA numerical mapping techniques used in the analysis of DNA sequences and to present the advantages and disadvantages of each technique to researchers. Each coding technique is exemplified in a DNA sequence, showing how that DNA sequence is digitized. Then, the frequencsy of use of these coding techniques in the 4 most popular study areas in the last 10 years and the max-min range of the performances obtained using these coding techniques were analyzed. This review will guide researchers in developing new coding techniques, and will facilitate previous researchers to improve their work. It will also guide researchers in discovering new techniques using innovative ideas.

References

[1] R. H. Thomas. “Molecular Evolution and Phylogenetics. Masatoshi Nei and Sudhir Kumar. Oxford University Press, Oxford. 2000. pp. 333. Price £65.00, hardback. ISBN 0 19 513584 9.,” *Heredity,* vol. 86, no. 3, pp. 385–385, 2001, doi: 10.1046/j.1365-2540.2001.0923a.x.

[2] M. Akhtar, J. Epps, and E. Ambikairajah. “Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction”, *IEEE J. Sel. Top. Signal Process.*, vol. 2, no 3, pp 310-321, Jun. 2008, doi: 10.1109/JSTSP.2008.923854.

[3] L. Das, J. K. Das, S. Mohapatra and S. Nanda. “DNA numerical encoding schemes for exon prediction: a recent history”, *Nucleosides, Nucleotides & Nucleic Acids*, vol.40, no 10, pp. 985-1017, Oct. 2021, doi: 10.1080/15257770.2021.1966797.

[4] U. N. Wisesty, T. R. Mengko and A. Purwarianti. “Gene mutation detection for breast cancer disease: A review”, *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 830, no 3, pp. 032051, Apr. 2020, doi: 10.1088/1757-899X/830/3/032051.

[5] M. Raman Kumar and N. K. Vaegae. “A new numerical approach for DNA representation using modified Gabor wavelet transform for the identification of protein coding regions”, *Biocybernetics and Biomedical Engineering*, vol. 40, no 2, pp. 836-848, Apr. 2020, doi: 10.1016/j.bbe.2020.03.007.

[6] N. Yu, Z. Li, and Z. Yu. "Survey on encoding schemes for genomic data representation and feature learning—from signal processing to machine learning", *Big Data Mining and Analytics*, vol. 1, no 3, pp. 191-210, Sep. 2018, doi: 10.26599/BDMA.2018.9020018.

[7] P. K. Kumari, "A Survey on Numerical Representation of DNA Sequences", *Asian Journal For Convergence In Technology (AJCT) ISSN -2350-1146*, Apr. 2018.

[8] L. Das, J. K. Das, S. Nanda and S. Mohapatra. "DNA Coding Sequence Prediction: A Review", içinde *2018 International Conference on Applied Electromagnetics, Signal Processing and Communication (AESPC)*, Oct. 2018, vol. 1, pp. 1-6. doi: 10.1109/AESPC44649.2018.9033278.

[9] M. Ahmad, L. T. Jung and A.-A. Bhuiyan. "From DNA to protein: Why genetic code context of nucleotides for DNA signal processing? A review", *Biomedical Signal Processing and Control*, vol. 34, pp. 44-63, Apr. 2017, doi: 10.1016/j.bspc.2017.01.004.

[10] X. Jin *et al.* "Similarity/dissimilarity calculation methods of DNA sequences: A survey", *Journal of Molecular Graphics and Modelling*, vol. 76, pp. 342-355, Sep. 2017, doi: 10.1016/j.jmgm.2017.07.019.

[11] G. Mendizabal-Ruiz, I. Román-Godínez, S. Torres-Ramos, R. A. Salido-Ruiz and J. A. Morales, "On DNA numerical representations for genomic similarity computation", *PLOS ONE*, vol. 12, no 3, p. e0173288, Mar. 2017, doi: 10.1371/journal.pone.0173288.

[12] S. Saini and L. Dewan. "Comparison of Numerical Representations of Genomic Sequences: Choosing the Best Mapping for Wavelet Analysis", *Int. J. Appl. Comput. Math*, vol. 3, no 4, pp. 2943-2958, Dec. 2017, doi: 10.1007/s40819-016-0277-1.

[13] Mabrouk, M.S. Advanced Genomic Signal Processing Methods in DNA Mapping Schemes for Gene Prediction Using Digital Filters --Gene prediction, Digital filters, 3- Base periodicity, Exon, Intron, Bioinformatics, Genomic signal processing", *American Journal of Signal Processing*, p. 13, 2017.

[14] L. Das, J. K. Das and S. Nanda, "Identification of exon location applying kaiser window and DFT techniques", içinde *2017 2nd International Conference for Convergence in Technology (I2CT)*, Apr. 2017, pp. 211-216. doi: 10.1109/I2CT.2017.8226123.

[15] B. Das and I. Turkoglu. "Classification of DNA sequences using numerical mapping techniques and Fourier transformation, Journal of the Faculty of Engineering and Arcitecture of Gazi University, 2016, doi: 10.17341/gazimmfd.278447.

[16] M. Abo-Zahhad, S. M. Ahmed and S. A. Abd-Elrahman. "Integrated Model of DNA Sequence Numerical Representation and Artificial Neural Network for Human Donor and Acceptor Sites Prediction", *IJITCS*, vol. 6, no 8, pp. 51-57, July. 2014, doi: 10.5815/ijitcs.2014.08.07.

[17] M. Abo-Zahhad, S. M. Ahmed and S. A. Abd-Elrahman. "Genomic Analysis and Classification of Exon and Intron Sequences Using DNA Numerical Mapping Techniques", *IJITCS*, vol. 4, no 8, pp. 22-36, July. 2012, doi: 10.5815/ijitcs.2012.08.03.

[18] H. K. Kwan, B. Y. M. Kwan and J. Y. Y. Kwan, "Novel methodologies for spectral classification of exon and intron sequences", *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no 1, p. 50, Feb 2012, doi: 10.1186/1687-6180-2012-50.

[19] S. D. Sharma, K. Shakya and S. N. Sharma, "Evaluation of DNA mapping schemes for exon detection", in *2011 International Conference on Computer, Communication and Electrical Technology (ICCCET)*, Mar. 2011, pp. 71-74. doi: 10.1109/ICCCET.2011.5762441.

[20] F. Akalin and N. Yumusak. "Classification of exon and intron regions obtained using digital signal processing techniques on the DNA genome sequencing with EfficientNetB7 architecture", *GUMMFD*, 37:3 (2022) 1355-1371.

[21] F. Akalin and N. Yumuşak. "Classification of ALL and CML malignancies being among the main types of leukaemia with graph neural networks and fuzzy logic algorithm," *GUMMFD*, Mar. 2022, doi: 10.17341/gazimmfd.1022624.

[22] L. Das, S. Nanda and J. K. Das. "An integrated approach for identification of exon locations using recursive Gauss Newton tuned adaptive Kaiser window", *Genomics*, vol. 111, no 3, pp. 284-296, May. 2019, doi: 10.1016/j.ygeno.2018.10.008.

[23] A. C. H. Choong and N. K. Lee. "Evaluation of convolutionary neural networks modeling of DNA sequences using ordinal versus one-hot encoding method", içinde *2017 International Conference*

on Computer and Drone Applications (IConDA), Nov. 2017, pp. 60-65. doi: 10.1109/ICONDA.2017.8270400.

[24] N. Chakravarthy, A. Spanias, L. D. Iasemidis and K. Tsakalis. "Autoregressive Modeling and Feature Analysis of DNA Sequences", *EURASIP J. Adv. Signal Process.*, vol. 2004, no 1, pp. 952689, Jan. 2004, doi: 10.1155/S111086570430925X.

[25] R. M. Kumar and N. K. Vaegae. "Walsh code based numerical mapping method for the identification of protein coding regions in eukaryotes", *Biomedical Signal Processing and Control*, vol. 58, no. 101859, Ap. 2020, doi: 10.1016/j.bspc.2020.101859.

[26] B. Das, S. Toraman, and I. Turkoğlu. "A novel genome analysis method with the entropy-based numerical technique using pretrained convolutional neural networks," *Turk J Elec Eng & Comp Sci*, vol. 28, no. 4, pp. 1932–1948, Jul. 2020, doi: 10.3906/elk-1909-119.

[27] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán and H. E. Stanley. "Finding Borders between Coding and Noncoding DNA Regions by an Entropic Segmentation Method", *Phys. Rev. Lett.*, vol. 85, no 6, pp. 1342-1345, Aug. 2000, doi: 10.1103/PhysRevLett.85.1342.

[28] D. Nicorici and J. Astola. "Segmentation of DNA into Coding and Noncoding Regions Based on Recursive Entropic Segmentation and Stop-Codon Statistics", *EURASIP J. Adv. Signal Process.*, vol. 2004, no 1, pp. 832471, Dec. 2004, doi: 10.1155/S1110865704309212.

[29] N. Y. Song and H. Yan. "Autoregressive modeling of DNA features for short exon recognition", in *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2010, pp. 450-455. doi: 10.1109/BIBM.2010.5706608.

[30] Q. Zheng, T. Chen, W. Zhou, L. Xie and H. Su. "Gene prediction by the noise-assisted MEMD and wavelet transform for identifying the protein coding regions", *Biocybernetics and Biomedical Engineering*, Vol. 41, no 1, 2021, doi: 10.1016/j.bbe.2020.12.005.

[31] R. Harrison, Y. Li and I. Măndoiu, Ed. *Bioinformatics Research and Applications: 11th International Symposium, ISBRA 2015 Norfolk, USA, June 7-10, 2015 Proceedings*, c. 9096. Cham: Springer International Publishing, 2015. doi: 10.1007/978-3-319-19048-8.

[32] Z. Abbas, H. Tayara and K. T. Chong. "4mCPred-CNN—Prediction of DNA N4-Methylcytosine in the Mouse Genome Using a Convolutional Neural Network", *Genes*, vol. 12, no 2, Feb. 2021, doi: 10.3390/genes12020296.

[33] P. Liò and M. Vannucci. "Finding pathogenicity islands and gene transfer events in genome data", *Bioinformatics*, vol. 16, no 10, pp. 932-940, Oct. 2000, doi: 10.1093/bioinformatics/16.10.932.

[34] L. Zhang, F. Tian, S. Wang and X. Liu. "A novel coding method for gene mutation correction during protein translation process", *Journal of Theoretical Biology*, vol. 296, pp. 33-40, Mar. 2012, doi: 10.1016/j.jtbi.2011.11.031.

[35] F. Castro-Chavez. "Defragged Binary I Ching Genetic Code Chromosomes Compared to Nirenberg's and Transformed into Rotating 2D Circles and Squares and into a 3D 100% Symmetrical Tetrahedron Coupled to a Functional One to Discern Start from Non-Start Methionines through a Stella Octangula", *J Proteome Sci Comput Biol*, vol. 2012, no 1, pp. 3, 2012, doi: 10.7243/2050-2273-1-3.

[36] M. Raman Kumar and V. Naveen Kumar. "A Numerical Representation Method for a DNA Sequence Using Gray Code Method", içinde *Soft Computing for Problem Solving*, Singapore, 2020, pp. 645-654. doi: 10.1007/978-981-15-0184-5_55.

[37] L. Deng, H. Wu, X. Liu and H. Liu. "DeepD2V: A Novel Deep Learning-Based Framework for Predicting Transcription Factor Binding Sites from Combined DNA Sequence", *International Journal of Molecular Sciences*, vol. 22, no 11, Jan. 2021, doi: 10.3390/ijms22115521.

[38] Q. Zhang, Z. Shen and D.-S. Huang, "Modeling in-vivo protein-DNA binding by combining multiple-instance learning with a hybrid deep neural network", *Sci Rep*, vol. 9, no 1, p. 8484, June. 2019, doi: 10.1038/s41598-019-44966-x.

[39] M. Randić, D. Butina and J. Zupan, "Novel 2-D graphical representation of proteins," *Chemical Physics Letters*, vol. 419, no. 4, pp. 528–532, Feb. 2006, doi: 10.1016/j.cplett.2005.11.091.

[40] Z. Liu, B. Liao, W. Zhu and G. Huang, "A 2D graphical representation of DNA sequence based on dual nucleotides and its application", *International Journal of Quantum Chemistry*, vol. 109, no 5, pp. 948-958, 2009, doi: 10.1002/qua.21919.

[41] A. T. M. Bari, M. Reaz, A. T. Islam, H.-J. Choi, and B.-S. Jeong. "Effective Encoding for DNA Sequence Visualization Based on Nucleotide's Ring Structure", *Evolutionary bioinformatics online*, vol. 9, pp. 251-61, July. 2013, doi: 10.4137/EBO.S12160.

[42] S. Zou, L. Wang and J. Wang. "A 2D graphical representation of the sequences of DNA based on triplets and its application", *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2014, no 1, pp. 1, Jan 2014, doi: 10.1186/1687-4153-2014-1.

[43] B. Das and I. Turkoglu. "A novel numerical mapping method based on entropy for digitizing DNA sequences", *Neural Comput & Applic*, vol. 29, 8: 207-215, Apr. 2018, doi: 10.1007/s00521-017-2871-5.

[44] A. Sankar, A. Nair and M. Thiru. "Visualization of genomic data using inter-nucleotide distance signals", Jan. 2005.

[45] Das, B. "A deep learning model for identification of diabetes type 2 based on nucleotide signals". Neural Comput & Applic (2022). https://doi.org/10.1007/s00521-022-07121-8

[46] Das, B. "An implementation of a hybrid method based on machine learning to identify biomarkers in the Covid-19 diagnosis using DNA sequences", Chemometrics and Intelligent Laboratory Systems (2022),v. 230, 104680, ttps://doi.org/10.1016/j.chemolab.2022.104680

[47] C.-T. Zhang and J. Wang. "Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve", *Nucleic Acids Research*, vol. 28, no 14, pp. 2804-2814, Tem. 2000, doi: 10.1093/nar/28.14.2804.

[48] C. Yu, M. Deng, L. Zheng, R. L. He, J. Yang and S. S.-T. Yau, "DFA7, a New Method to Distinguish between Intron-Containing and Intronless Genes", *PLOS ONE*, vol. 9, no 7, pp. e101363, Tem. 2014, doi: 10.1371/journal.pone.0101363.

[49] R. R. Garafutdinov, A. R. Sakhabutdinova, P. A. Slominsky, F. G. Aminev and A. V. Chemeris, "A new digital approach to SNP encoding for DNA identification", *Forensic Science International*, vol. 317, no. 110520, Dec. 2020, doi: 10.1016/j.forsciint.2020.110520.

[50] T. Hoang, C. Yin and S. S.-T. Yau, "Numerical encoding of DNA sequences by chaos game representation with application in similarity comparison", *Genomics*, vol. 108, no 3, pp. 134-142, Oct 2016, doi: 10.1016/j.ygeno.2016.08.002.

[51] W. Deng and Y. Luan, "Analysis of Similarity/Dissimilarity of DNA Sequences Based on Chaos Game Representation", *Abstract and Applied Analysis*, vol. 2013, p. e926519, Mar. 2013, doi: 10.1155/2013/926519.

[52] Z.-G. Yu and V. Anh. "Time series model based on global structure of complete genome", *Chaos, Solitons & Fractals*, vol. 12, no 10, pp. 1827-1834, Aug. 2001, doi: 10.1016/S0960-0779(00)00147-8.

[53] C. Yin, "Encoding and Decoding DNA Sequences by Integer Chaos Game Representation", *Journal of Computational Biology*, vol. 26, no 2, pp. 143-151, Feb. 2019, doi: 10.1089/cmb.2018.0173.

[54] A. P. Boyle *et al.*, "Comparative analysis of regulatory information and circuits across distant species", *Nature*, vol. 512, no 7515, Aug. 2014, doi: 10.1038/nature13668.

[55] Z. Mo *et al.*, "One novel representation of DNA sequence based on the global and local position information", *Sci Rep*, vol. 8, no 1, p. 7592, May. 2018, doi: 10.1038/s41598-018-26005-3.

[56] D. Bielińska-Wąż and P. Wąż, "Spectral-dynamic representation of DNA sequences", *Journal of Biomedical Informatics*, vol. 72, pp. 1-7, Aug. 2017, doi: 10.1016/j.jbi.2017.06.001.

[57] A. Czerniecka, D. Bielińska-Wąż, P. Wąż, and T. Clark, "20D-dynamic representation of protein sequences", *Genomics*, vol. 107, no 1, pp. 16-23, Jan. 2016, doi: 10.1016/j.ygeno.2015.12.003.

[58] D. Zhang. "A New Numerical Method for DNA Sequence Analysis Based on 8-Dimensional Vector Representation", *Journal of Applied Mathematics and Physics*, vol. 7, no 12, Dec. 2019, doi: 10.4236/jamp.2019.712204.

[59] F. Ben Nasr, and A. E. Oueslati, "CNN for human exons and introns classification", içinde *2021 18th International Multi-Conference on Systems, Signals Devices (SSD)*, Mar. 2021, pp. 249-254. doi: 10.1109/SSD52085.2021.9429303.

[60] A. Rokas, "Phylogenetic Analysis of Protein Sequence Data Using the Randomized Axelerated Maximum Likelihood (RAXML) Program", *Current Protocols in Molecular Biology*, vol. 96, no 1, pp. 19.11.1-19.11.14, 2011, doi: 10.1002/0471142727.mb1911s96.