BIG DATA MINING AND ANALYTICS
ISSN 2096-0654 02/06 pp191–210
Volume 1, Number 3, September 2018
DOI: 10.26599/BDMA.2018.9020018

Survey on Encoding Schemes for Genomic Data Representation and Feature Learning—From Signal Processing to Machine Learning

Ning Yu, Zhihua Li, and Zeng Yu*

Abstract: Data-driven machine learning, especially deep learning technology, is becoming an important tool for handling big data issues in bioinformatics. In machine learning, DNA sequences are often converted to numerical values for data representation and feature learning in various applications. Similar conversion occurs in Genomic Signal Processing (GSP), where genome sequences are transformed into numerical sequences for signal extraction and recognition. This kind of conversion is also called encoding scheme. The diverse encoding schemes can greatly affect the performance of GSP applications and machine learning models. This paper aims to collect, analyze, discuss, and summarize the existing encoding schemes of genome sequence particularly in GSP as well as other genome analysis applications to provide a comprehensive reference for the genomic data representation and feature learning in machine learning.

Key words: encoding scheme; data representation; feature learning; deep learning; genomic signal processing; machine learning; genome analysis

1 Introduction

Since the first DNA genome sequence was sequenced in the 1970s[1], four characters, A, T/U, C, and G, representing the nucleotides of Adenine, Thymine/Uracil, Cytosine, and Guanine, respectively, have been remained as the mainstream representation form in genome analysis. The four letters are combined and permuted to denote various biology markers, such as sequences, genes, proteins, RNAs, and DNAs. They are readable, understandable, and convenient for sequence analysis[2]. In the 21st century, especially with the advent of next-generation sequencing technologies, genomic data representation has become more important in many artificial intelligence technologies, such as machine learning and knowledge discovery. Data representation challenges can be alleviated by adopting alternative data preprocessing methods such as feature selection, normalization, and regularization; however, effective methods of representing data and building models for feature learning remain uncertain.

In Digital Signal Processing (DSP), the first step toward feature extraction is converting the original information into sequential digital values. These values are called signals. Digital signal processing has a set of advanced methods to process data, many of which can be useful for knowledge-based discovery in bioinformatics. An example is Genomic Signal Processing (GSP), which is an interdisciplinary method that integrates DSP, pattern recognition, control theory, dynamic system, information theory, communication theory, network modeling, mathematics, and statistics into bioinformatics[3]. The goal of GSP is to discover hidden genomic and proteomic characteristics and

• Ning Yu is with the Department of Computing Sciences, College at Brockport, State University of New York, Brockport, NY 14422, USA. E-mail: nyu@brockport.edu.
• Zhihua Li is with the Department of Computer Science and Technology at Jiangnan University, Wuxi 214122, China. E-mail: zhli@jiangnan.edu.cn.
• Zeng Yu is with the School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China. E-mail: yuzeng2005@163.com.
∗ To whom correspondence should be addressed.
Manuscript received: 2018-01-21; accepted: 2018-01-24

understand the mechanisms of disease and biological system regulations by exerting extant signal processing methods into bioinformatics. In recent decades, scientists have utilized GSP in genome analysis and data processing in various subjects, such as gene detection and prediction, dynamical modeling of the genetic network, sequence analysis, evolutionary analysis, RNA prediction, and so forth.

GSP can provide a good reference for machine learning because they both rely on the intensive numeric computing. The first step in GSP is to convert character data or text in DNA genomes into numerical sequences, which is identical for both DSP and machine learning. This conversion can be called numeric representation[4], encoding scheme[5], or symbolic-to-digital mapping[6], which are rather equivalent.

An encoding scheme determines the steps for signal processing and pattern recognition and determines how far genomic properties can be used to detect the characteristics of particular regions. Thus, diverse encoding schemes are designed for different applications. For example, the most commonly used encoding scheme in GSP is Voss representation[7] because it only contains 0 s and 1 s in sequences that can be easily used for Discrete Fourier Transform (DFT)-related applications. Reports claim that most GSP methods are focused on detecting coding regions[8]. However, this study disputes that. Encoding schemes are applied in many areas, such as detecting 3-periodicity coding region[8], prediction of repeats in genomic sequences[9], sequence alignments[10, 11], phylogenetic tree[12, 13], dynamic genetic network modeling[3, 14], sequence comparisons[8], correlation and fractal analysis[4, 15], motif detection[16], and so forth.

The encoding schemes scrutinized in this paper include atomic representation, Chaos Game Representation (CGR), Electron-Ion Interaction Pseudopotential (EIIP), molecular mass, thermodynamics, three-group classification, and dinucleotide representations among others. These representations can be classified into fixed mapping, which maps single or multiple characters to one number, and flexible mapping[17], which uses a more flexible approach where nucleotide sequences can be encoded based on variant properties or statistics[18]. Representations can be non-graphical or graphical[19] in 1D, 2D, 3D, 4D, 5D, and even 6D transformations[20].

This study comprehensively investigated encoding schemes of genome sequences. However, some schemes may not be covered. In attempt to include all schemes, we further classify them in terms of the following properties: (1) biochemical or biophysical properties, (2) computing/mathematical properties. The former makes the encoding scheme scientifically meaningful, and the latter ensures its computing merits. Furthermore, they are summarized into five categories based on following perspectives: biochemical properties, primary-structure properties, Cartesian-coordinate properties, binary and information encoding, and graphical representation. A summary of the encoding schemes is shown in Fig. 1.

This paper is organized according to the above categories. Sections 2–6 demonstrate the different encoding schemes following the above perspectives, and in Section 7 we compare performance, analyze applications, and discuss regularization methods, such as data normalization in deep learning, which can alleviate the negative impact of a poor encoding scheme. In Section 8, we conclude with future prospects.

2 Biochemical Properties

2.1 Atomic Number

Atomic representation refers to that each nucleobase is assigned its atomic number as an indicator to convert the nucleotide sequence into a series of numerical atomic indicators. $C = 58$, $T = 66$, $A = 70$, $G = 78$. This type of direct mapping was used to measure the fractal dimension difference between sequences of human and chimpanzee[4]. It also gave a set of comparisons to show the diverse results when different encoding schemes of numerical representations were used.

Cervantes-De la Torre et al.[15] adopted atomic representation and obtained a fractal dimension of the 118 bp HAR1 nucleotide sequence for human and chimpanzee which was about 2.02 and 1.96, respectively, with a difference of about 0.06. When adopting the scheme of purine atomic number=62 and pyrimidine atomic number= 42, the difference between human and chimpanzee based on the fractal dimension of the HAR1 gene sequence was 0.07[4]. In a scheme where arbitrary values of $A = 1$, $T = 2$, $C = 3$, and $G = 4$ were assigned, the difference was about 0.03[4]. Thus, it is seen that various encoding schemes in data representation can give different results, which

Name	Scheme	Feature
Atomic Number	*C*=58, *T*=66,A=70, *G*=78	Atomic number of nucleotide
EIIP	*C*=0.1340, *T*=0.1335, *A*=0.1260, *G*=0.0806	Distribution of the free electron pseudopotential energy
Molecular Mass	*C*=111.1, *T*=112.1, *A*=135.13, *G*=151.13 or *C*=110, *T*=125, *A*=134, *G*=150	Molecule mass
Thermodynamics	*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, *CC*=11.0, *GG*=11.0, *GC*=11.1, *CG*=11.9	Encoded according to the enthalpy values of thermodynamic interactions between two molecules
Three-group classification	*(1) R*={*A, G*}, *Y*={*C, T*} *(2) M*={*A, C*}, *K*={*G, T*} *(3) W*={*A, T*}, *S*={*G, C*}	Each group is assigned the same number.
Dinucleotide	Sixteen dinucleotides are mapped to a unit circle.	Encoding the neighboring nucleotides to a 2D plot
Ring Structure	*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).	Extension of dinucleotide encoding; ring structures; molecular weight
Inter-nucleotide Distance	If the same nucleotides are located at the positions of i, $i+k_1$, $i+k_2$, $i+k_3$,..., one encodes $S(i)$, $S(i+k_1)$, $S(i+k_2)$,... as k_1, k_2, k_3,....	Primary structure; calculate the distance between the same nucleotides
Triplet	$<T_1, \sigma_1>$,$<T_2, \sigma_2>$, ..., $<T_{64}, \sigma_{64}>$, T is the triplet of 64 codons, σ is triplet repeat function	The triplet encoding depends on the weight of condons
Frequency-of-occurrence	*C*=0.27215, *T*=0.20576, *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	Single nucleotide frequency or dinucleotide frequency of occurrence
Minimum Entropy Mapping	$H_x(M)$=-$\sum_{k=0}^{N/2} p_x[k;M] \log p_x[k;M]$	Not fixed mapping; Calculate the power spectrum of DNA sequence
Integer Number	$x \leftrightarrow n$, $x \in \{C, T, A, G\}$, $n \in \{0, 1, 2, 3\}$ or $n \in \{-1, -2, 1, 2\}$	C,T,A,G are assigned an integer number
Real Number	$x \leftrightarrow n$, $x \in \{C, T, A, G\}$, n is a real number	EIIP is also one of real number encoding schemes.
Complex Number	*C*=-1+*j*, *T*=1-*j*, *A*=1+*j*, *G*=-1-*j* or *C*=-*j*, *T*=1, *A*=-1, *G*=j	A broad category like integer number and real number
QPSK	*C*=-1-*j*, *T*=1-*j*, *A*=1+*j*, *G*=-1+*j*	Constellation for QPSK scheme in a 2D plane
PAM	*C*=0.5, *T*=1.5, *A*=-1.5, *G*=-0.5	1D encoding scheme
DNA Walk/Paired Numeric	(*C* or *T*)=+1, (*A* or *G*)=-1 or (*C* or *T*)=-1, (*A* or *G*)=+1	Visualizing the cumulative change for pyrimidine (C or T) and purine (A or G)
Voss	*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]	Four sequences of 0 and 1 are formed to represent the genome sequence
Galois Field	α=1↔1↔*C*, α^1=α↔2↔*T*, α^2=α+1 ↔3↔*G*, 0=0↔0↔*A*	A fundamental scheme for signal and information processing.
I Ching Representation	Binary codes for 64 codons	Ancient encoding scheme; binary codes for modern computing
Chaos Game Representation (CGR)	$X_i = 0.5(X_{i-1} + g_{ix})$ $Y_i = 0.5(Y_{i-1} + g_{iy})$ *A*: (0, 0), *T*(1, 0), *G*(1, 1), *C*(0, 1)	Graphical representation; (g_{ix}, g_{iy}) is the corresponding vertex; initiated in a unit square
CGR Walk	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)	Combine CGR with DNA Walk; three-group properties
Tetrahedron	$A = k,\ C = -\frac{2\sqrt{2}}{3}i + \frac{\sqrt{6}}{3}j - \frac{1}{3}k,\ G = -\frac{2\sqrt{2}}{3}i - \frac{\sqrt{6}}{3}j - \frac{1}{3}k,\ T = \frac{2\sqrt{2}}{3}i - \frac{1}{3}k$	Properties of tetrahedron used for representing 4 nucleotides and codons
Self-Organized Map (SOM)	*A*: (0, 0, 0), *T*: (0.289, 0.5, 0.816), *C*: (0.866, 0.5, 0), *G*: (0, 1, 0)	3D coordinate; midpoints between two bases can reflect 3-group properties.
Quaternion	*A* = *i* + *j* + *k*, *C* = - *i* + *j* - *k*, *G* = - *i* - *j* + *k*, *T* = *i* - *j* - *k*	Scale 3D coordinate to 4D quaternion; symmetry of quaternion
H-curve	*A*= *i* + *j* - *k*, *T* = *i* - *j* - *k*, *C* = - *i* - *j* - *k*, *G*= - *i* + *j* – *k*	3D curve
Z-curve	$A_n + C_n + G_n + T_n = n$ $\begin{vmatrix} A_n \\ C_n \\ G_n \\ T_n \end{vmatrix} = \frac{n}{4} \begin{vmatrix} 1 \\ 1 \\ 1 \\ 1 \end{vmatrix} + \frac{1}{4} \begin{vmatrix} +1 & +1 & +1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \\ -1 & -1 & +1 \end{vmatrix} \begin{vmatrix} x_n \\ y_n \\ z_n \end{vmatrix}$	Four faces of tetrahedron are the directions of cumulating nucleotides in DNA sequence, that match the geometric properties of tetrahedron.

Fig. 1 Summary of encoding schemes.

can lead to non-uniformity in research.

2.2 Electron-Ion Interaction Pseudopotential

A numerical scheme based on EIIP of nucleotides C, T, A, and G was proposed to replace Voss representation[21], which maps the nucleotides into four binary indicator sequences[7] and is commonly employed in many GSP applications based on DFT. The energies of delocalized electrons in amino acids and nucleotides have been calculated as EIIP. These values have been used in Resonant Recognition Models (RRM) as a substitute for the corresponding amino acid in protein sequences[22]. The EIIP values for the four nucleotides are $C = 0.1340$, $T = 0.1335$, $A = 0.1260$, and $G = 0.0806$. When the EIIP values are substituted to a DNA sequence, it becomes a numerical sequence that denotes the distribution of the free electrons' pseudopotential energies along the DNA sequence.

EIIP uses real numbers, which facilitate scientific computing. As such, it has been adopted in diverse fields such as neural network, wavelet transform, and GSP to reflect the pseudopotential feature of nucleotide sequences[23,24].

2.3 Molecular Mass representation

The molecular mass of nucleotides is used as a numeric scheme to correlate some physical quantities. Molecular mass representation has been utilized in mapping DNA sequences into a multi-dimensional space, where C, T, A, and G were encoded as 110, 125, 134, and 150, respectively[25,26].

In one study, ascending molecular masses of 111.1, 112.1, 135.13, and 151.13 were respectively used for C, T, A, and G[27], and the nucleobases were paired based on these masses in ascending or descending order. The deviation from the actual mass values was little and was probably caused by measurement techniques.

2.4 Thermodynamic properties

According to the thermodynamic properties of neighboring nucleotide interactions[28], the enthalpies of combined nucleotides are shown in Fig. 2. The enthalpy values are $TC = 5.6$, $GA = 5.6$, $CA = 5.8$, $TG = 5.8$, $TA = 6.0$, $AC = 6.5$, $GT = 6.5$, $CT = 7.8$, $AG = 7.8$, $AT = 8.6$, $TT = 9.1$, $AA = 9.1$, $CC = 11.0$, $GG = 11.0$, $GC = 11.1$, and $CG = 11.9$. The unit of these enthalpy values is kcal/mol. Figure 2 also shows the symmetric patterns when combining two nucleotides, which have been discussed in a previous research[29].

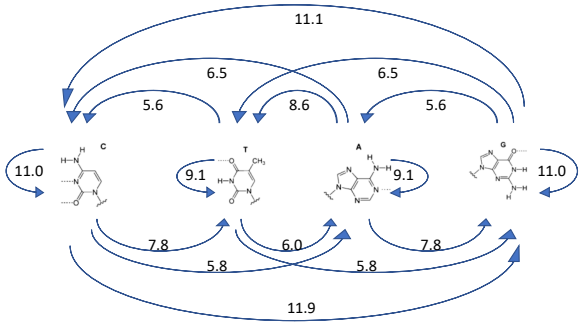

Fig. 2 Enthalpy values of thermodynamic interactions between two molecules. The unit of measurement is kcal/mol[29] (1 cal=4.18 J).

Biochemical properties are reflected in the three-group classification of the four nucleotides[30]. These three groups integrate the thermodynamic properties of DNA nucleotides and are as follows: (1) purine $R = \{A$ or $G\}$ and pyrimidine $Y = \{C$ or $T\}$; (2) amino group $M = \{A$ or $C\}$ and keto group $K = \{G$ or $T\}$; and (3) weak H-bonds $W = \{A$ or $T\}$ and strong H-bonds $S = \{G$ or $C\}$. Many encoding schemes are based on this three-group classfication including CGR-walk[31–33].

In addition, mutations between the nucleotides can be transition or transversion. Transition (A-G, C-T) occurs more frequently than transversion (C-G, T-A). According to the enthalpy values in Fig. 2, weak bonds exist between pairs of transitions compared with those of transversions[29]. Encoding schemes can be designed based on these biochemical features. For example, in mapping using the binary codes, Hamming distance and Euclidean distance were employed to reflect the differences between transition and transversion in light and dark colors, respectively, as shown in Fig. 3[2, 34, 35].

3 Primary-Structure Properties

The properties of a genome primary structure are based on the structure of its DNA sequences, including dinucleotide sets and triplets, as well as related information such as the statistics at each position. These can be considered when designing encoding schemes.

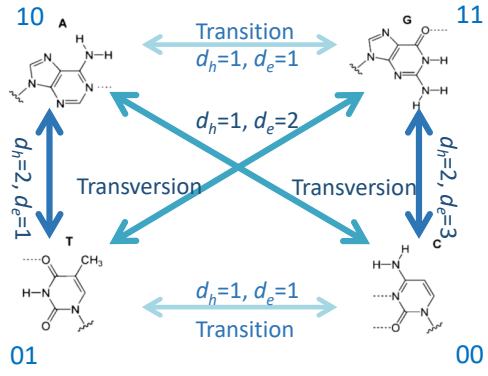

Fig. 3 Difference of transition and transversion between molecules measured by Hamming distance and Euclidean distance[29].

3.1 Dinucleotide representation

Combining two neighboring bases, dinucleotide sets are encoded as values[27, 36]. For example, sixteen points can be put on the circumference of a unit circle as shown in Fig. 4, where each point can be encoded into a polar angle following the below equation,

$$\theta_i = 2i\pi/16, i = 1, 2, ..., 16 \tag{1}$$

Since each dinucleotide is encoded to the coordinates (x_i, y_i), $i = 1, 2, ..., N-1$ in a 2D plane, a DNA sequence can be encoded into a series of codes and its geometrical center can be calculated as follows:

$$\overline{x} = \sum_{i=1}^{N-1} x_i/(N-1), \ \overline{y} = \sum_{i=1}^{N-1} y_i/(N-1) \tag{2}$$

The dinucleotide representation can be applied for measuring the distance between two sequences[36], since it plots any pair of bases into the 2D plot. A Covariance Matrix (CM) can be determined as follows:

$$\mathrm{CM} = \begin{pmatrix} \mathrm{CM}_{xx} & \mathrm{CM}_{xy} \\ \mathrm{CM}_{yx} & \mathrm{CM}_{yy} \end{pmatrix} \tag{3}$$

where

$$\begin{aligned} \mathrm{CM}_{xx} &= \frac{\sum_{i=1}^{N-1}(x_i-\overline{x})(x_i-\overline{x})}{\left(\sqrt{\sum_{i=1}^{N-1}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{N-1}(x_i-\overline{x})^2}\right)}, \\ \mathrm{CM}_{xy} = \mathrm{CM}_{yx} &= \frac{\sum_{i=1}^{N-1}(x_i-\overline{x})(y_i-\overline{y})}{\left(\sqrt{\sum_{i=1}^{N-1}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{N-1}(y_i-\overline{y})^2}\right)}, \\ \mathrm{CM}_{yy} &= \frac{\sum_{i=1}^{N-1}(y_i-\overline{y})(y_i-\overline{y})}{\left(\sqrt{\sum_{i=1}^{N-1}(y_i-\overline{y})^2}\sqrt{\sum_{i=1}^{N-1}(y_i-\overline{y})^2}\right)} \end{aligned} \tag{4}$$

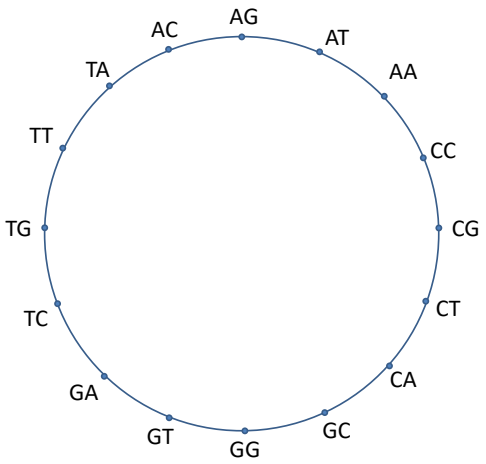

Fig. 4 Dinucleotides placed in a unit circle.

Using two eigenvalues λ_1 and λ_2 for the matrix CM, the distance between two sequences i and j can be calculated as

$$d_{ij} = \sqrt{(\lambda_1^i - \lambda_1^j)^2 + (\lambda_2^i - \lambda_2^j)^2} \tag{5}$$

3.2 Ring structure

Ring structure is an extension of dinucleotide encoding, and it utilizes the ring structures of DNA bases and their corresponding molecular masses. In Section 2.4, we grouped the nucleotides into three sets and have established that they can be encoded in terms of the molecular weight in ascending or descending order. Here, the three nucleotides groups are further classified into six classes: purine, pyrimidine, amino, keto, strong hydrogen bond, and weak hydrogen bond[27]. Subsequently, they are plotted into a 2D plane of position coordinates. The six classes constitute the six vertices of a hexagon. There are six combinations of the different hexagon representations as shown in Fig. 5. Each plot corresponds to an encoding system. For example, in the first hexagon, each dinucleotide is encoded into a two-dimensional vector: 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), and CC: (−0.5, 0).

Thus, for a sequence $S = \{s_1, s_2, ..., s_N\}, s_i \in \{C, T, A, G\}$ and $i = \{1, 2, 3, ..., N\}$, S can be mapped into a series of points P.

$$P_i = \varphi(s_i s_{i+1}, i) = \varphi(x_i, y_i, i) = \varphi(x_{s_i s_{i+1}}, y_{s_i s_{i+1}}, i) \tag{6}$$

where $x_{s_i s_{i+1}}$ and $y_{s_i s_{i+1}}$ represent the encoded values in a corresponding plot and i denotes the z coordinate.

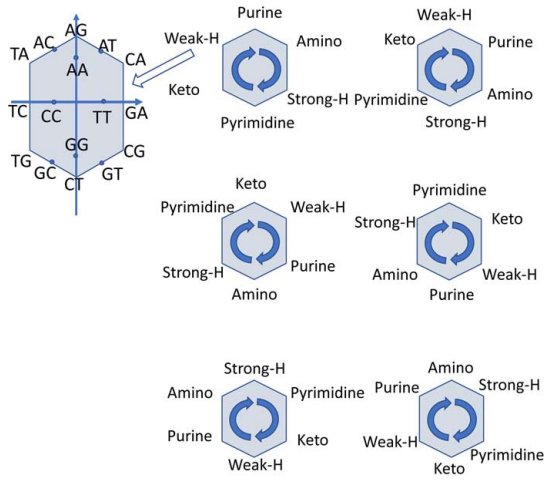

Fig. 5 Six hexagons.

Similar to the calculation of dinucleotide introduced in the previous subsection, the geometric center for a sequence can be calculated with the following equation:

$$u_x = \frac{1}{N}\sum_{i=1}^{N} x_i, u_y = \frac{1}{N}\sum_{i=1}^{N} y_i, u_z = \frac{1}{N}\sum_{i=1}^{N} z_i \tag{7}$$

Euclidian distance can be used to measure the similarity/dissimilarity matrix.

$$\rho = \sqrt{u_x^2 + u_y^2 + u_z^2} \tag{8}$$

3.3 Inter-nucleotide distance encoding

Encoding scheme based on inter-nucleotide distance considers the property of the primary structure of a DNA sequence that can be used for GSP. Nair and Mahalakshmi[37] first adopted this encoding scheme, where the distance between two same nucleotides was encoded as the numerical representation for the corresponding nucleotide. Distance-based methods were adopted and further developed for detecting CpG island[38,39]. For a DNA sequence S, assuming that the same nucleotides are located at positions of i, $i + k_1$, $i + k_2$, $i + k_3$, ..., that is, $S(i) = S(i + k_1) = S(i + k_2) = S(i + k_3) = ...$, then $S(i)$, $S(i + k_1)$, $S(i + k_2)$, ..., are encoded as k_1, k_2, k_3, ...

For example, for a short sequence of AGTTCTACCAGC, the first and second A's are encoded as 6 and 3, respectively, because of the distances between the first two A's and the next two A's. Similarly, G, C, and T are encoded. Thus, the sequence is encoded as $\{6, 9, 1, 2, 3, 6, 3, 1, 3, 2, 1, 0\}$[37]. Furthermore, the encoding scheme was slightly modified into a cyclic structure[40], where the last nucleotide in the sequence is connected to the first. Thus, the total lengths of the four inter-nucleotide distance sequences denoted as N^{A}, N^{T}, N^{G}, and N^{C} are equal to the sequence length N.

An inter-nucleotide distance-based genome analysis was conducted[40] where the distribution of inter-nucleotide distance for each nuclotide shows a power law behavior. Kullback-Leibler distance, Kolmogorov-Smirnov distance, and correlation coefficient can be adopted to measure the difference between this distribution and a reference distribution. The relative error expressed in Eq. (9) can be used to compare various distance distributions based on their inter-nucleotide distance.

$$r(k) = \frac{f_0(k) - f(k)}{f_0(k)} \tag{9}$$

where $f_0(k)$ is the observed relative frequency

distribution of distance k, and $f(k)$ is the reference relative frequency distribution[40,41].

Inter-nucleotide or inter-dinucleotide distance-based methods have been applied to analyze the distribution kernel of genome patterns for a whole DNA genome[42,43], and similar research has been conducted in recent years[44].

Studies on constructing phylogenetic trees using inter-nucleotide distance have stimulated the development of inter-nucleotide distance-based techniques[41,45]. The algorithms have become more efficient because of the use of k-word distance, which count the distance between k-tuple, $2 \leqslant k \leqslant 9$[45,46]. Extensive studies on phylogenetic tree construction have used inter-amino-acid distance to measure the distance of amino acids in protein sequence[47].

3.4 Triplet encoding

A weight-based numerical representation was proposed based on the properties of nucleotide triplets and codons of amino acid[48]. This encoding scheme was used to measure the distance between two sequences. For two pairs of triplets (X_1, Y_1) and (X_2, Y_2), if the corresponding codons of X_1 and Y_1 are encoded into the same amino acid, while those of X_2 and Y_2 are encoded into another amino acid, the distance between the pairs can be expressed as

$$|\psi(X_1) - \psi(Y_1)| < |\psi(X_2) - \psi(Y_2)| \quad (10)$$

where ψ is a mapping from triplet to weight. The weight consists of two parts: the amino acid and the codon, which are its integer and fractional parts, respectively. For example, the first codon (GCT) of alanine has a weight of 1.1 and its second codon (GCC) has a weight of 1.2.

For a DNA sequence $G = g_1, g_2, g_3, ..., g_N$, $g \in \{C, T, A, G\}$, its triplet sequence is $G = t_1, t_2, t_3, ..., t_M$ where $M = [N/3]$ and t_i is a triplet. A mapping Θ is illustrated as

$$\Theta(G) = \{(1, \psi(t_1)), (2, \psi(t_2)), ..., (M, \psi(t_M))\} \quad (11)$$

Furthermore, a triple-repeat function δ is defined to represent the occurrence of triplet in a sequence. Given two coding sequences, A and B, the triplet-repeat model set for a sequence G is $G = < T_1, \delta_1 >, < T_2, \delta_2 >, ..., < T_{64}, \delta_{64} >$, where T is the triplet of 64 codons. A weight deviation between the two sequences is shown in Eq. (12). It can be used to measure the similarity between A and B.

$$\mathrm{WD}(A, B) = \frac{\sum_{i=1}^{64} |\delta_A(i) - \delta_B(i)| \cdot \psi(T_i)}{64} \quad (12)$$

3.5 Frequency-of-occurrence mapping

The occurrences of DNA nucleotide differ in various regions[49,50], such as intron and exon. On the basis of the frequencies of nucleotide occurrence, the fractional occurrence can be statistically calculated and used as a key parameter in detecting these regions. Thus, nucleotides are represented by their fractional occurrences. For example, C, T, A, and G in exons are encoded as 0.272 15, 0.205 76, 0.243 00, and 0.279 09, respectively, following the statistics of their occurrences.

Position count function uses the binary count for each position to generate the frequency of nucleotide occurence in that position[51]. DNA sequence has been modeled with a random process that assigns values according to a probability distribution on the alphabet (C, T, A, G)[52].

Besides the frequency of single nucleotides, the frequencies of dinucleotides and triplets have also been considered[9,50,53,54]. They show different frequencies among species, indicating the different statistics on DNA genome of various species. Thus, this encoding scheme can vary in different DNA genomes. For human genome, the below frequencies of dinucleotide were used as an encoding scheme[5]: 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.

3.6 Minimum entropy mapping

To reduce the noise of DNA sequences and concentrate on the relevant information after mapping, Minimum Entropy Mapping (MEM) was designed[55]. MEM is an encoding scheme that maps by minimizing the spectral entropy of a DNA sequence rather than fixed mapping. The search of encoding scheme is an iterative process following an exhaustive search algorithm.

Each iteration will assign the vector of C, T, A, and G by choosing the increment of fixed Δh. Once the values of C, T, A, and G are calculated, it converts the sequences into numeric sequences and calculates the Fourier spectra for C, T, A, and G sequences. Subsequently, the energy is computed in terms of the Fourier transform. The spectrum P can be expressed in

$$P[k] = |X_A[k]|^2 + |X_C[k]|^2 + |X_G[k]|^2 + |X_T[k]|^2 \tag{13}$$

where

$$X[k] = \sum_{n=0}^{N-1} x[n] \mathrm{e}^{-\mathrm{i}\frac{2\pi}{N}nk}, k = 0, 1, ..., N-1 \tag{14}$$

The entropy H_x of a sequence is defined as

$$H_x(M) = -\sum_{k=0}^{N/2} p_x[k; M] \log p_x[k; M] \tag{15}$$

where

$$p_x[k; M] = P_x[k; M]/\varepsilon_P(M) \tag{16}$$

and

$$\varepsilon_P(M) = \sum_{k=0}^{N/2} p_x[k; M] \tag{17}$$

The final encoding scheme meets that

$$\bar{M} = \arg \min_{A,C,G,T \in \mathbf{R}} H_x(M) \tag{18}$$

Under this MEM encoding scheme, the sequences' spectrum via Fourier transform can be expressed as

$$X_s[k] = \sum_{n=0}^{N-1} x[n; \bar{M}] \mathrm{e}^{-\mathrm{i}\frac{2\pi}{N}nk}, k = 0, 1, ..., N-1 \tag{19}$$

This method was validated as an effective method to reduce noise and enhance concentration on signals[55], which can be applied in detecting periodicity in DNA sequences. The entropic information can also be applied to other applications[56,57], such as DNA sequence analysis, coding/non-coding detection, and so forth.

4 Cartesian-Coordinate Properties

4.1 Integer and real number

Integer and real number representations are direct and are commonly used encoding schemes[58,59]. In many scenarios of mapping DNA nucleotides, especially in the early stage of genomic studies, C, T, A, and G have been arbitrarily assigned to an integer or a real number, such as 0, 1, 2, and 3. For example, DNA barcode on the mitochondrial gene cytochrome c oxidase I (COI) is a core global bioidentification system for animals. DNA sequences were encoded to four integers ($A = 1, G = 2, C = 3$, and $T = 4$)[60,61].

This broad encoding scheme has been applied to many other applications[59,62]. Kent et al.[63] used 2-bit format to compress and store the DNA sequences in a compact randomly-accessible format, which uses a 16-byte header to contain the encoding information and pack each DNA nucleotide to two bits per base, T: 00, C: 01, A: 10, and G: 11. However, this type of arbitrary assignment is criticized for its inability to provide real signals to aid understanding in biological research[64]. Instead, a weight-based assignment is employed to the spectral transformation where the weight coefficients are derived from the enthalpy analysis of each nucleotide pair[64]. This numerical representation method is one of the supportive evidence for a novel encoding method[64]. Moreover, EIIP is also a set of real numbers which has many applications in genomic sequence analysis[23].

A complementary encoding scheme using an integer or a real number is popular in neural network community because its mean is zero and the deviations are symmetric. Such symmetric and complementary properties are beneficial to data training and feature learning. For example, the code book {C: −1, T: −2, A: 2, G: 1} has showed its importance in supervised deep learning networks in recent studies[5].

4.2 Complex number

The complex representation reflects the complementary nature of AT and CG pairs as $A = 1 + j$, $C = -1 + j$, $G = -1 - j$, and $T = 1 - j$[17,65–68], which better translates the features of nucleotides into mathematical properties. Complex number representation is a 2D numerical mapping. By placing the nucleotides on different vertices on a two-dimensional Cartesian-coordinate plane, the encoding values for C, T, A, and G are different. The complex number representation is regarded as a dimensionality reduction technique since 3D projection can be reduced to 2D[6]. It leads to two types of mapping methods by changing the projection planes: $A = 1 + j$, $C = -1 - j$, $G = -1 + j$, $T = 1 - j$; and $A = -1 + j$, $C = -1 - j$, $G = 1 + j$, $T = 1 - j$. For the former, the pairs of nucleotides CG and AT are in mathematics complex conjugates, while purines and pyrimidines have equal imaginary parts and real parts with opposite signs. For the latter, the two complementary strands of a DNA molecule correspond to digital signals of equal absolute values and opposite signs so that their algebraic sum is zero, which benefits computing[6].

Furthermore, if the Cartesian coordinate is rotated by 45 degrees, the complex numbers for nucleotides are encoded as: $A = -1$, $C = -j$, $G = j$, and $T = 1$. In a quaternion representation of DNA bases, pure quaternions are assigned to each base: $A = i + j + k$,

$C = i - j - k$, $G = -i - j + k$, and $T = -i + j - k$.

Since the complex representation is viewed as 2D mapping, real-number representation[23,69,70] can be viewed as a 1D Cartesian-coordinate mapping[71]. A typical representation for the 1D mapping is $A = 1.5$, $C = 0.5$, $G = -0.5$, and $T = -1.5$, where AT and CG are complementary[49]. Another alternative real-number mapping is $A = -1.5$, $T = 1.5$, $C = 0.5$ and $G = -0.5$[69]. The vectors connecting the origin to four points, (1, 1), (−1, 1), (−1, −1) and (1, −1), have rotational angles of $\pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$ with the x-axis[70]. Bases C, G, A, and T are accordingly defined as 1, 3, 5, and 7.

4.3 Quadrature Phase Shift Keying (QPSK)/Pulse Amplitude Modulation (PAM) schemes

The QPSK scheme shows constellations in a complex plane, whereas PAM scheme shows the real representation[72]. The complex, real, and integer representations can be regarded as constellation diagrams, which are widely applied in digital communications[55,69]. Figure 6 displays the QPSK/PAM schemes for the real and complex representations.

In the QPSK scheme, complex numbers represent the bases[73]: $A = 1 + j$, $G = -1 + j$, $C = -1 - j$, and $T = 1 - j$. In PAM scheme, real numbers denote the bases[73]: $A = -1.5$, $G = -0.5$, $C = 0.5$, and $T = 0.5$. Thus, QPSK/PAM schemes uniformly represent data on 2D planes and ensure symmetry of genetic codes.

4.4 DNA Walk and paired numeric method

DNA Walk model[74] can be represented on a 2D plane, which graphically shows a path along a DNA sequence,

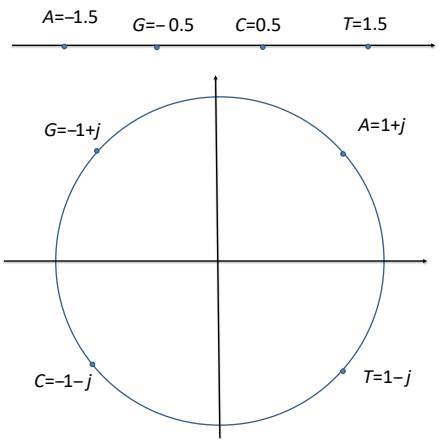

Fig. 6 Constellation for real number and complex number representations.

whose value is upwards (+1) for a pyrimidine (C or T) and downwards (−1) for a purine (A or G). A graph keeps oscillating with the nucleotide value on the x-axis, and the sequence progresses in a cumulative manner. The DNA walk can be used to visualize changes in nucleotide composition, base pair patterns, and evolution along a DNA sequence.

Although DNA walk is a simple numerical representation method, it can show the overall statistical features of DNA sequences. Thus, it forms the basis of many improved methods[31,75].

Similar to DNA walk, paired numeric method encodes the pairs of complementary nucleotides AT and CG as +1 and −1, respectively, so that all strands of DNA helix are identically represented[49]. Nucleotides are commonly paired and complemented in encoding schemes. For example, in the complex number representation of Section 4.2, nucleotides are mapped, paired, and complemented in a 2D symmetry plane, while for real-number representation, they are symmetrically paired, mapped, and complemented in a 1D plot.

5 Binary and Information Encoding

5.1 Voss representation

Voss[7] proposed a method to represent a DNA sequence by four binary sequences and applying long-range fractal correlation. It has since been utilized as a canonical numerical representation for GSP, especially for DFT-based methods. The encoding method is shown in the following matrix.

$$\begin{aligned} S &= [C, G, A, T], \\ C_n &= [1, 0, 0, 0], \\ G_n &= [0, 1, 0, 0], \\ A_n &= [0, 0, 1, 0], \\ T_n &= [0, 0, 0, 1] \end{aligned} \tag{20}$$

Fourier technique uses Voss as the most prominent numerical method to detect short-range correlations such as the 3-periodicity on protein-coding regions. The quantitative measure for detecting the relative strength of this periodicity is based on Fourier transformation[76,77]. The transformation reflects the 3-periodicity because of amino acid and codon usage biases. The unequal frequency of occurrence of the amino acids in proteins results to a 3-base periodicity in the coding regions of a DNA genome.

Through numerical representations of DNA genomes, DSP-based features are extracted, analyzed,

and classified in the spectral domain or spatio-temporal domain.

A binary representation is mostly used to represent genome sequences. DNA sequence containing nucleotides C, T, A, and G are converted into four separate binary sequences, $x_C[n]$, $x_T[n]$, $x_A[n]$, and $x_G[n]$, where 1 and 0 respectively represent the presence and absence of a base in the corresponding positions.

The DFT is shown as follows:

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j(2\pi nk/N)},\ 0 \leqslant k \leqslant N-1 \quad (21)$$

where $x[n]$ is a finite-length sequence of length N. GeneScan calculated the signal-to-noise ratio of the peak at $k = N/3$ as $P = S[N/3]/\hat{S}$, where $S[k] = \sum_m |X_m[k]|^2$, $m = \{C, T, A, G\}$ and $\hat{S}$ is the average spectral content of S[76]. P was assigned a value of 4 to distinguish between coding and non-coding sequences.

Abbasi et al.[78] applied an FIR bandpass filter of order 8 with central frequency of $2\pi/3$ to numerical sequences and multiplied the sequences with an impulse train of periodicity 3 to emphasize the period-3 property in exonic region.

Conventional signal processing methods are often combined with other techniques. For example, a 22-alphabet-based encoding method that considered the distribution of dinucleotides and Jensen-Renyi divergence was proposed to detect the borders between coding and non-coding regions[79]. A technique based on inter-stop distance was used to measure the distance between two stop symbols (amino acid) to identify the coding region in prokaryotes and simple eukaryotes[80].

5.2 Galois field

An encoding scheme was integrated into Galois field by which nucleotides were mapped to Galois field $GF(4)$ and the sequence was transformed into orthogonal (n, k) codes[81]. The labels in Eq. (22) are created for the nucleotide elements with the primitive polynomial for $GF(2)$: $\alpha^2 + \alpha + 1 = 0$. Any $GF(2)$ binary pair corresponds to one of four $GF(4)$ symbols. The polynomial can be manipulated by addition, multiplication, subtraction, and division in $GF(4)$.

$$\begin{aligned} &\alpha^0 = 1 \Leftrightarrow 1 \Leftrightarrow C,\\ &\alpha^1 = \alpha \Leftrightarrow 2 \Leftrightarrow T,\\ &\alpha^2 = \alpha + 1 \Leftrightarrow 3 \Leftrightarrow G,\\ &0 = 0 \Leftrightarrow 0 \Leftrightarrow A \end{aligned} \quad (22)$$

Furthermore, $GF(4)$ encoding scheme method can be applied in other sophisticated encoding methods such as error-correction code, which is an important information coding technique in genome analysis. The binary error-correcting coding structure reflects the nature of genome coding and efficiency in detecting genome redundancy and gene mutations. Similar ideas to Galois Field have been developed. In a research, the vector corresponding to the third base in a codon (the least significant digit) was multiplied with 1, while those corresponding to the second base and first base (the most significant digit) were multipled with 2 and 2^2, respectively[2]. Other encoding schemes have been combined with $GF(4)$ to reflect the features of DNA sequence[6].

5.3 Error-Correction Code

Cyclic block codes are usually known in communication channel as Error-Correction Codes (ECC). In biology, these values are not needed, and a block coding structure is reconstructed to fit biological applications[72]. Currently, many variants of ECC are adopted in DNA computing[82]. DNA sequences are encoded as large ECC to equip them with error-correction mechanisms in accordance with modern communication theory.

The ring of integers modulo 4 was used to construct the algebraic structure of a codeword to verify the presence of an error-correcting code underlying DNA sequences[34]. To verify a match between a given DNA sequence and a codeword, the 24 permutations of C, T, A, and G were considered. A BCH code denoted by (n, k, d) was associated with the DNA sequence structure, where n is the sequence length, k is the length of the input information sequence responsible for creating the DNA sequence, and d is the minimum difference between any two codewords[83].

A codeword was designed for every k-symbol information sequence and the codebook was a set of all codewords made by the encoder[84]. For a received sequence, a decoder gave a method for selecting the codeword to be transmitted. Each symbol is a bit and can be denoted as 0 or 1.

A convolutional encoding method was adopted, where the error-correction coding theory was used to encode and analyze DNA sequences for prokaryotic organisms[35]. The convolutional code model was designed based on degeneracy of codons. The Hamming distance between two codes was calculated,

which is the number of positions with different corresponding symbols. By shifting each position, the distances between adjacent codes were calculated. The average code distances of the selected species outlined their genomic characteristics.

5.4 I Ching representation

A genetic code based on natural patterns of symmetry and periodicity was proposed to show the harmony between the graphical geometry and biological reality[85,86]. Three binary representations of the genetic code according to the ancient I Ching of Fu-Xi were defragged[86], based on three biochemical properties of nucleic acids: H-bonds, purine/pyrimidine rings, and the keto-enol/amino-imino tautomerism, yielding the last pair a 32/32 single-strand self-annealed genetic code and I Ching tables. Twenty amino acids can be directly mapped to the I Ching tables, which can contain $4^3 = 64$ codes. Some codes denote the same amino acid.

6 Graphical Representation

6.1 CGR and CGR-Walk

CGR was proposed as a new method for representing DNA sequences[87]. It generates a genomic signature to characterize genomes by providing compact, lossless, and visual appearance[88] and combining them with distinct sequence statistics.

The CGR encoding scheme maps the DNA sequence in a unit square, whose four vertices are encoded with the nucleotides, $A : (0, 0)$, $T : (1, 0)$, $G : (1, 1)$, and $C : (0, 1)$, and starts the sequence at the center of the square. For a DNA sequence, the first nucleotide is mapped halfway between the starting point and the corresponding vertex for the nucleotide, and the remaining nucleotides of the sequence are plotted halfway between the previous point and the corresponding vertices. Mathematically, a DNA sequence can be represented as

$$\begin{cases} X_i = 0.5(X_{i-1} + g_{ix}), \\ Y_i = 0.5(Y_{i-1} + g_{iy}) \end{cases} \tag{23}$$

where (g_{ix}, g_{iy}) is the corresponding vertex of this nucleotide and (X_i, Y_i) and (X_{i-1}, Y_{i-1}) are the current and previous plotted points on the coordinate, respectively.

In some articles, CGR is widely regarded as a Markov Chain model[16,88–90]; however, some other articles[91] disagreed. It has been proved that the frequencies of nucleotides, dinucleotides or trinucleotides cannot solely determine the patterns in CGRs, whereas frequencies of oligonucleotides of all lengths can solely determine them[88]. The local similarity between sequences can be reflected in the distance between CGR points, since these points aggregate closer to a certain region with the increase of local similarity. Studies have unveiled the dependent relationship between CGR patterns and local similarity[33].

Visual tools for data inspection are important in genome analysis as they facilitate the quantitative analysis and comparison of DNA sequences[27,48,92]. Following the exponential growth of genome data, visual and graphical representations have become more important. In terms of the number of dimensions, visualization methods can be categorized into 2D, 3D, 4D, 5D, and 6D, which are all constructed based on numerical transformation and representation[20,93].

The CGR-walk model was proposed by combining CGR with DNA walk method, which considers the thermodynamic properties of the three groups introduced in Section 2. Local similarity/dissimilarity was further studied[31,32]; three 2D CGR spaces were created, and each vertex was differentially encoded with nucleotides, as shown in Table 1. The basic CGR encoding rules of the three spaces are the same as those of traditional CGR. The CGR-walk model provides the numerical foundation as well as graphical representations for long-range correlation studies and genome analysis.

The CGR-walk model provides a solid numerical representation for studying the relationship between long-range correlation and Hurst exponent[25,31,74,77]. Hurst exponent is a measure for long-range correlation in a DNA sequence that is related to the auto-correlations of time series[94]. A DNA sequence with N elements is transformed into a finite set of numerical values by summing the x and y components of CGR. That is,

$$u_i = x_i + y_i \tag{24}$$

where u_i is the calculated value for the i-th nucleotide in a DNA sequence, and x_i and y_i are the x and y

Table 1 Encoded initial positions of CGR-walk.

Category	Encoded initial position
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)

coordinates of CGR, respectively[94].

6.2 Tetrahedron

A tetrahedral numerical representation of C, T, A, and G was projected on a 3D coordinate system to delineate the distance between nucleotides[6,59,85,93,95], as shown in Fig. 7. A typical representation of tetrahedron is shown in Eq. (25). As an application[92], one of the tetrahedron encoding schemes for codons is displayed in Fig. 8.

$$\begin{aligned} A &= k, \\ C &= -\tfrac{2\sqrt{2}}{3}i + \tfrac{\sqrt{6}}{3}j - \tfrac{1}{3}k, \\ G &= -\tfrac{2\sqrt{2}}{3}i - \tfrac{\sqrt{6}}{3}j - \tfrac{1}{3}k, \\ T &= \tfrac{2\sqrt{2}}{3}i - \tfrac{1}{3}k \end{aligned} \tag{25}$$

Furthermore, since the tetrahedron is a subset of a cube, it can be rotated to fit the coordinates of the cube. Thus, the encoding scheme for the nucleotides is as follows:

$$\begin{aligned} A &= i + j + k, \\ C &= -i + j - k, \\ G &= -i - j + k, \\ T &= i - j - k \end{aligned} \tag{26}$$

6.3 SOM-based approach

Kohonen and Somervuo[96,97] proposed the use of Self-Organizing Maps (SOM) for unsupervised training in the artificial neural network. This novel approach encodes the genomic sequences into fixed-size, metric-based vectors[93], as shown in Fig. 9. On an irregular tetrahedron, the distance is set as 1 between AG and CT vertices and 2 for the remaining pairs. Nucleotides C, T, A, and G positions are mapped at the four vertices of the tetrahedron. Symbols for ambiguous nucleotides are encoded using the 3D coordinates corresponding to the midpoint (R, Y, K, M, S, and W) between two bases: the centroid of the plane and the centroid of the tetrahedron (N). In terms of the spatial position of the irregular tetrahedron, nucleotides are encoded as the following numbers: A: $(0, 0, 0)$, T: $(0.289, 0.5, 0.816)$, C: $(0.866, 0.5, 0)$, and G: $(0, 1, 0)$. The Euclidean distance between any two nucleotides can be calculated easily. Moreover, the midpoints (R, Y, K, M, S, and W) and the centroids can be encoded as numerical values according to the positions in the irregular tetrahedron. Based on Self-Organizing Maps, a new means was provided for the phylogenetic analysis of distant species[98].

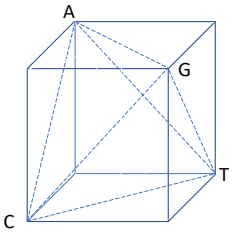

Fig. 7 3-dimensional tetrahedron in a cube.

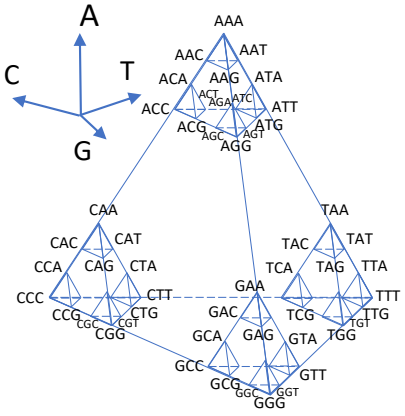

Fig. 8 Tetrahedron encoding scheme for codons.

6.4 Quaternion

The quaternion approach[71] is a 4D hypercomplex representation derived from Eq. (27).

$$h = a_0 i_0 + \cdots + a_j i_j + \cdots + a_N i_N,\ N \in \mathbf{Z}^+ \cup \{0\} \tag{27}$$

where $a \in \mathbf{R}$, $0 \leqslant j < N$, $i_0 = 1$ and i_j, $0 < j \leqslant N$, are imaginary units. When $j = 1$, it is a complex number; when $j = 2$, it can be reflected on a 3D plane; when $j = 3$, it is a 4D quaternion[71]. Seven variants of quaternion encoding were researched for detecting the protein-coding regions[66]. These variants are verified as different complex-number encoding schemes instead of quaternion encoding since the complex and quaternion encoding can be unified to an

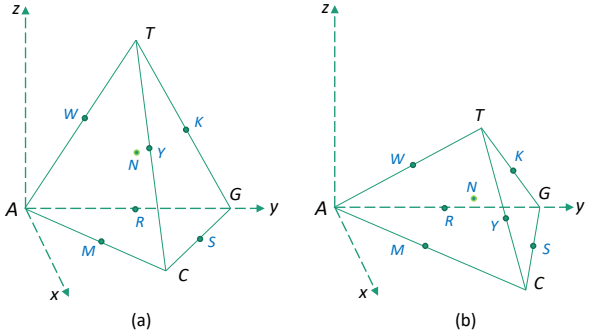

Fig. 9 Encoding methods based on (a) a regular tetrahedron and (b) an irregular tetrahedron.

N-dimensional hypercomplex number encoding. One of the encoding assignments can be

$$\begin{aligned} A &= i + j + k, \\ C &= -i + j - k, \\ G &= -i - j + k, \\ T &= i - j - k \end{aligned} \tag{28}$$

This approach is equivalent to the rotated tetrahedron (with 3 imaginary units i, j, and k and no real number) described in Section 6.2.

Addtionally, the quaternion is proposed to replace the complex number representation to diminish the symbol challenges of complex numbers, which could lead to some undected significant patterns. According to an article[71], the symmetry of the quaternions results to symbol permutation invariance, when compared with complex number representation, which lacks symmetry.

6.5 H-curve and Z-curve

In the 1980s, scientists tried to create alternatives to letter-series representations of nucleotide sequences[99]. H-curve was presented as an alternative to represent a DNA sequence as a 3D curve, in which each nucleotide is encoded as a vector with the below functions:

$$\begin{aligned} g_w(A) &= i + j - k, \\ g_w(T) &= i - j - k, \\ g_w(C) &= -i - j - k, \\ g_w(G) &= -i + j - k \end{aligned} \tag{29}$$

where i, j, and k are unit vectors pointing to the Cartesian x, y, and z axes, respectively, and w is the loci of this nucleotide. Thus, H-curve is defined as

$$h(z) = \sum_{w=1}^{n} g_w(z) \tag{30}$$

where $z \in \{C, T, A, G\}$ and n is the sequence length. The end points of an H-curve describe the nucleotide composition of the sequence[99]. Vector k value is the total number of nucleotides in the sequence. The j value is the accumulated count of purine (A, G) versus pyrimidine (C, T), which increases or decreases by one unit on encountering a purine or a pyrimidine, respectively. The i value is the accumulated count of C+G content versus A+T content, which increases or decreases by one unit on encountering (A or T) or (C or G), respectively.

The H-curve representation provides a simple way to view, sort, and compare various gene structures. However, its coordinates are rather sophisticated to compute. H-curve was compressed into a 2D plane[100]. However, it has a drawback of possibly forming a loop or circuit, and the sequences are not uniquely described, which could result in information degeneracy. It was later improved by compressing 2D Cartesian to the two quadrants of a Cartesian plane to avoid degeneracy[101]. The encoding system is as follows:

$$\begin{aligned} g_w(A) &= \tfrac{1}{2}i - \tfrac{\sqrt{3}}{2}j, \\ g_w(T) &= \tfrac{1}{2}i + \tfrac{\sqrt{3}}{2}j, \\ g_w(C) &= \tfrac{\sqrt{3}}{2}i + \tfrac{1}{2}j, \\ g_w(G) &= \tfrac{\sqrt{3}}{2}i - \tfrac{1}{2}j \end{aligned} \tag{31}$$

where w is the location of a nucleotide in the sequence, and i and j are the vectors in a Cartesian plane. The uniqueness of a sequence is proved by following this system[101].

Similar to H-curve, the Z-curve is a 3-D curve that provides a unique set of vectors for visualizing and analyzing DNA sequences[102–104]. The three components of the Z-curve, $\{x_n, y_n, z_n\}$, represent three independent nucleotide coordinates that denote a DNA sequence in a 3D coordinate system. However, the vectors move only in four directions, which are A-face, T-face, C-face, and G-face, as shown in Fig. 10. These directions (A_n, C_n, G_n, and T_n) can be converted into coordinate vectors (x_n, y_n, z_n). Assuming that each move counts a nucleotide, the summation of the four moves equals the number of nucleotide because of the geometric properties of a tetrahedron. The mathematical expression[102] is as follows:

$$A_n + C_n + G_n + T_n = n \tag{32}$$

The conversion between (x_n, y_n, z_n) and (A_n, C_n, G_n, T_n) can be represented as

$$\begin{vmatrix} A_n \\ C_n \\ G_n \\ T_n \end{vmatrix} = \frac{n}{4} \begin{vmatrix} +1 \\ +1 \\ +1 \\ +1 \end{vmatrix} + \frac{1}{4} \begin{vmatrix} +1 & +1 & +1 \\ -1 & +1 & -1 \\ +1 & -1 & -1 \\ -1 & -1 & +1 \end{vmatrix} \cdot \begin{vmatrix} x_n \\ y_n \\ z_n \end{vmatrix} \tag{33}$$

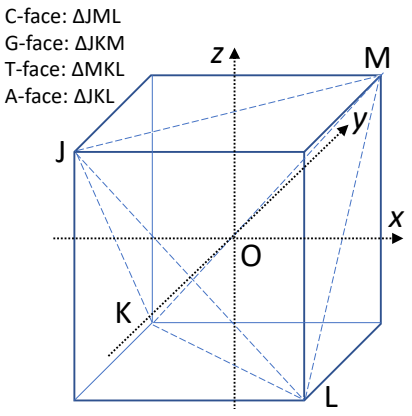

Fig. 10 Tetrahedron-based coordinate system in Z-curve.

When compared with H-curve, Z-curve can be regarded as a symmetry projection on 3D coordinate system, whereas, H-curve can be viewed as a spatial projection of the Z-curve by rotating the coordinate system. Both of them are tetrahedron-based encoding schemes.

7 Analysis and Discussion

In designing an encoding scheme, few considerations have to be made, such as the number of biological properties to be reflected and the measure in which the schemes can improve data processing in feature extraction. Relatively, spectral analysis of DNA sequences is often significant in GSP for detecting feature signals. However, data modeling and training are more important in deep learning. Thus, choosing the best encoding scheme is often application-specific.

7.1 Genomic signal processing

Sixteen numerical codes were discussed in details when searching for the period-3 signals from genome sequence[66]. Eight of them were grouped under complex number encoding, including quaternary codes that were derived from 3D DNA walk for graphical representation[67]. A comparison showed efficiency of the quaternary code ($C = -1$, $T = j$, $A = 1$, $G = -j$). In a period-3 spectral classification of exon and intron sequences with two data sets, the quaternary codes were more efficient in predicting and computing.

Similar assessments have been conducted[105], and more than ten encoding schemes, including several real-number representations, frequency-of-occurrence, paired numeric, complex quaternion, and inter-nucleotide distance, were compared for protein-coding detection. The results showed that paired numeric encoding scheme was the most accurate representation[49]. Paired numeric scheme exploits the property that AT and CG pairs are rich in introns and exons, respectively. They are paired in a complementary manner and values of $+1$ and -1 are utilized to distinguish them.

GSP methods are also used for computing the distance between pairs of DNA sequences without a need for alignment[8], which is useful where similarity distances are needed. Similarity/dissimilarity can be quantified based on the distance between different pairs[27,31,36,48] and represented numerically. In addition to a good transformation of DNA sequences, a clear numerical representation is very vital in measuring the distance between pairs. Thus, the alignment-free method is preferable to the conventional alignment-based methods, because the bulky data that develops from the latter result to complicated computation for post-genomic studies. The previously introduced numerical representations for distance measurement, including dinucleotide representation[36], ring structure-based encoding[27], I Ching representation, and weight encoding[48], perform differently, and better performance is expected in those models based on biological principles. For example, dinucleotide representation ignores the order the nucleotides, which can hide some biological characteristics, resulting in an inaccurate estimation of similarity[36]. A similar conclusion has been reported[106], where nine encoding schemes were compared for measuring the similarity of DNA sequences. Biological-based mappings with mathematical merits, such as Voss and Tetrahedron representations, performed better than other encoding schemes.

7.2 Neural network

Four types of numerical representations were used in a four-classifier neural network system for gene identification[107] and prediction of gene regulatory networks[108,109]. Numerical methods can predict more accurate results, compared with feature-based approaches that use distinct markers to differentiate DNA sequences from non-promoter sequences such as CpG islands, TATA box, and CAAT box, but these numerical methods work inefficiently for eukaryotic genomes due to the diverse patterns of promoters. The numerical representations include (1) $C = 1$, $T = -1$, $A = -2$, $G = 2$, (2) $C = 1$, $T = -1$, $A = -1$, $G = 1$, (3) $C = (10)_2$, $T = (01)_2$, $A = (00)_2$, $G = (11)_2$, (4) $C = (0010)_2$, $T = (0100)_2$, $A = (1000)_2$, $G = (0001)_2$.

The four encoding schemes convert the inputs of four neural networks from characters to numerical sequences. Schemes (1) and (2) are integer representations. Scheme (1) utilizes the unequal distances among the nucleotides, namely, $G > C$ and $T > A$, introducing undesirable characteristics to the resultant numerical sequences, and thus, features in the promoter regions are affected[108]. By assigning the same value to two nucleotides, $A = T = -1$, $C = G = 1$, scheme (2) ignores the difference between A and T or C and G. Scheme (3) is a promising numerical representation that has

been effective in gene/exon prediction[17], and it is also used for multi-classifier neural networks[108]. Scheme (4) uses binary numbers that are orthonormal to each other and have identical Hamming distance between any two of them. A similar test examined two different numerical representations that preserve different biological properties for multi-classifier neural networks[108].

7.3 Deep learning

Deep learning method has emerged as an advanced technique for genomic sequence analysis[110]. Deep Neural Network (DNN) is one of the implementations in deep learning, and it generally refers to methods that automatically learn complex functions[111] and map data through multiple levels of the feed-forward neural network to reveal some intractable and non-linear relationships between input data and hidden factors.

Several common encoding schemes, including EIIP, Galois field GF(4), Binary, Enthalpy, Entropy, and so forth, were assessed and their impact on DNA annotation for the deep auto-encoder network was evaluated[5]. The standard benchmarks were adopted on human gene splicing sites[112], which contained real and fake splice sites and were divided into training and validation sets.

It was shown that Complementary scheme ($C = -1$, $T = -2$, $A = 2$, $G = 1$) performs more effectively in a data set with a large number of features[5], and schemes such as Binary and DAX perform more effectively in a data set with fewer features. This also demonstrates that the performance of encoding schemes is basically application-specific.

Five encoding popular encoding schemes, DAX, EIIP, Complementary, Enthalpy, and Galois, were selected in an experiment to encode the lincRNA sequences to train and find lincRNA features[113]. The Complementary scheme ($C = -1, T = -2, A = 2, G = 1$) had optimum performance on deep neural network, and results showed that the encoding scheme that reflects biochemical properties and has a symmetric structure may perform better over a particular application.

7.4 Normalization/regularization in deep learning

New techniques are evolving in machine learning and deep learning to improve data representation methods because of the importance of data representation in data mining and pattern recognition. Normalization/regularization enhances data representation and compensates for the inadequacies of poor encoding schemes. Likewise, it could eclipse the effectiveness of a good encoding scheme.

The neural network has been known to converge faster if the input training data are normalized[114]. Analytical results have shown that a non-zero mean of features is disadvantageous to the optimization[115]. In this context, novel optmization algorithms were invented in DNN to reduce the internal covariate shift by whitening[115, 116]. However, whitening each layer is costly; therefore, it is easier to independently normalize each scalar feature by making it have a zero mean and unit variance[114, 117]. In a layer of d-dimensional input where $x = (x_1, x_2, ..., x_d)$, the normalization equation is

$$\widehat{x_k} = \frac{x_k - E(x_k)}{\sqrt{\mathrm{Var}(x_k)}} \tag{34}$$

where $E(x_k)$ and $\mathrm{Var}(x_k)$ can be calculated over the training data set. Normalization techniques can improve the data representation of DNN as seen in ImageNet classification and mini-batch procedures where a better performance was achieved with less training steps[117].

7.5 Complex number in deep learning

Although complex number representation is not commonly used in deep learning applications, some innovatary research has shown it to be impactful[118–123]. It was thought to be unfeasible in artificial neural network compared with real-number representations, which are good for differential equations and linear algebra. However, recent research[118, 119] found that complex number representation can reduce matrix size and memory usage. It improved robustness when used in the Recurrent Neural Network (RNN)[120, 121]. Furthermore, convolutionary neural network can also adopt complex number representation to enhance the predictive capability in image recognition[122]. Complex number was also explored in Generative Adversarial Networks (GAN) for a new Jacobian algorithm, and a better convergence on GAN architectures was achieved[123].

Because of the importance of complex values in quantum mechanics, complex number representation could be commonly adopted for data representation and encoding genome sequences in deep learning and feature learning.

8 Conclusion

GSP and machine learning share some similarities such as encoding scheme and data representation because they all involve intensive numeric computing. For further processing, encoding schemes convert original data into appropriate representations recognizable by a machine system, and this is the fundamental step in data representation and feature learning. Encoding schemes for data representation rely on the combination of different properties on a genome sequence, such as biological, chemical, physical, mathematical, computational, and graphical properties for optimum performance in scientific and computing applications.

In this paper, we cover over 25 significant encoding schemes to provide a comprehensive reference for applications in GSP, machine learning, and deep learning. These encoding schemes are introduced and categorized according to different perspectives, including biochemical, primary-structure, Cartesian-coordinate, binary, graphical, and information encoding perspectives. To determine the best encoding scheme for an application, we scrutinized and analyzed the typical applications including GSP and machine learning. Encoding schemes were found to be dependent upon specific needs, as seen in Section 7 examples. However, most efficient encoding schemes are symmetrical, have a mean of 0 and structures suitable for computing/math, and can reflect some biochemical or biophysical properties. For examples, the complimentary encoding scheme ($C = -1$, $T = -2$, $A = 2$, $G = 1$) and the quaternary code ($C = -1$, $T = j$, $A = 1$, $G = -j$) have shown the best performance in DNN and GSP for genome annotation and pattern recognition, respectively.

These common features of efficient encoding schemes, particularly symmetry, and a zero mean, coincide with the merits of normalization/regularization that ensure a normal distribution with unit variance[117]. Thus, normalization/regularization are also important techniques to achieve an efficient encoding scheme. Their capacity to alleviate the negative impact of poor encoding schemes is very important to data representation in genomic analysis applications.

As observed in the literature, it is inadvisable to arbitrarily assign values to DNA genome data, as it could lead to failure of feature learning or necessitate normalization/regularization procedures to ensure a normal distribution of data. However, normalization/regularization method does not always guarantee a compensation for the lapses in data representation. An appropriate encoding scheme that ensures normal distribution is always best to boost the performance of feature learning, and this has solely been adopted before the advent of normalization/regularization methods. Nonetheless, additional normalization/regularization procedures can further adjust the data distribution and make it more feasible to feature learning. A flow chart of the positions of encoding scheme is illustrated in the path B of Fig. 11.

Compared with protein sequence or other biological sequences, DNA genome sequence is relatively simple, and it is a good context to study the effects of encoding schemes and illustrate the usefulness of those schemes that were often skipped by scientists. We collated those existing encoding schemes applied in DNA sequence analysis, particularly in GSP, artificial intelligence, and emerging deep learning applications, and analyzed them from the perspectives of biochemical, primary structure, mathematical properties, computing properties, and visualization properties. By their performance in different applications, we find that no encoding scheme is most suitable for all applications and the encoding scheme design has to adapt the context

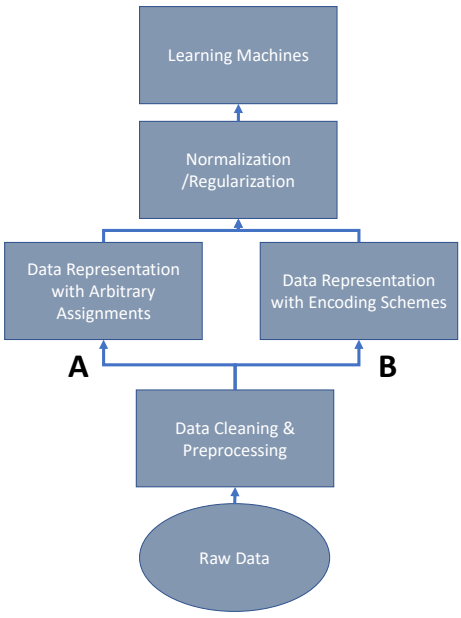

Fig. 11 Flow chart on the position of encoding scheme in feature learning.

of specific applications. However, good schemes have characteristics such as mathematical symmetry and they properly reflect biochemical properties. From literature, we speculate that some encoding schemes are overlooked often because of the adjustment effects of normalization/regularization. However, normalization/ regularization cannot guarantee efficiency in data representation and feature learning. Thus, they would work best when coupled with a good encoding scheme. This paper also provides a reference for scientists on encoding schemes employed in bioinformatics, especially for genomic sequence analysis.

Similarly, for other areas, such as protein sequence analysis and text mining, character-to-numeric conversion is also very important, and if lagging, could result in problems in data representation and feature learning. These applications could be more complicated where more features are contained, and each of which may have more representations. Therefore, elaborate encoding schemes are needed in those areas and may be a major survey in the future.

Acknowledgment

We appreciate the supports from the Department of Computing Sciences, State University of New York College at Brockport.

References

[1] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison III, P. M. Slocombe, and M. Smith, Nucleotide sequence of bacteriophage ϕX174 DNA, *Nature*, vol. 265, no. 5596, pp. 687–695, 1977.

[2] N. Yu, X. Guo, F. Gu, and Y. Pan, Signalign: An ontology of DNA as signal for comparative gene structure prediction using information-coding-and-processing techniques, *IEEE Trans. NanoBioscience*, vol. 15, no. 2, pp. 119–130, 2016.

[3] D. Anastassiou, Genomic signal processing, *IEEE Signal Process. Mag.*, vol. 18, no. 4, pp. 8–20, 2001.

[4] T. Holden, R. Subramaniam, R. Sullivan, E. Cheung, C. Schneider, G. Jr. Tremberger, A. Flamholz, D. H. Lieberman, and T. D. Cheung, ATCG nucleotide fluctuation of Deinococcus radiodurans radiation genes, in *Proc. Instruments, Methods, and Missions for Astrobiology X*, San Diego, CA, USA, 2007, p. 669417.

[5] N. Yu, Z. Yu, F. Gu, and Y. Pan, Evaluating the impact of encoding schemes on deep auto- encoders for DNA annotation, in *Bioinformatics Research and Applications*, Z. Cai, O. Daescu, and M. Li, eds. Springer International Publishing, 2017, pp. 390–395.

[6] P. D. Cristea, Conversion of nucleotides sequences into genomic signals, *J. Cell. Mol. Med.*, vol. 6, no. 2, pp. 279–303, 2002.

[7] R. F. Voss, Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, *Phys. Rev. Lett.*, vol. 68, no. 25, pp. 3805–3808, 1992.

[8] E. Borrayo, E. G. Mendizabal-Ruiz, H. Vlez-Pérez, R. Romo-Vázquez, A. P. Mendizabal, and J. A. Morales, Genomic signal processing methods for computation of alignment-free distances from DNA sequences, *PLoS One*, vol. 9, no. 11, p. e110954, 2014.

[9] B. Hutter, V. Helms, and M. Paulsen, Tandem repeats in the CpG islands of imprinted genes, *Genomics*, vol. 88, no. 3, pp. 323–332, 2006.

[10] Z. M. Ning, A. J. Cox, and J. C. Mullikin, SSAHA: A fast search method for large DNA databases, *Genome Res.*, vol. 11, no. 10, pp. 1725–1729, 2001.

[11] K. Katoh, K. Misawa, K. I. Kuma, and T. Miyata, MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform, *Nucleic Acids Res.*, vol. 30, no. 14, pp. 3059–3066, 2002.

[12] B. R. King, M. Aburdene, A. Thompson, and Z. Warres, Application of discrete Fourier inter-coefficient difference for assessing genetic sequence similarity, *EURASIP J. Bioinform. Syst. Biol.*, vol. 2014, no. 1, p. 8, 2014.

[13] T. Hoang, C. C. Yin, H. Zheng, C. L. Yu, R. L. He, and S. S. T. Yau, A new method to cluster DNA sequences using Fourier power spectrum, *J. Theor. Biol.*, vol. 372, pp. 135–145, 2015.

[14] W. Peng, J. X. Wang, B. H. Zhao, and L. S. Wang, Identification of protein complexes using weighted PageRank-nibble algorithm and core-attachment structure, *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 12, no. 1, pp. 179–192, 2015.

[15] F. Cervantes-De la Torre, J. I. González-Trejo, C. A. Real-Ramírez, and L. F. Hoyos-Reyes, Fractal dimension algorithms and their application to time series associated with natural phenomena, *J. Phys. Conf. Ser.*, vol. 475, no. 1, p. 012002, 2013.

[16] S. Vinga, A. M. Carvalho, A. P. Francisco, L. M. Russo, and J. S. Almeida, Pattern matching through chaos game representation: Bridging numerical and discrete data structures for biological sequence analysis, *Algorithms Mol. Biol.*, vol. 7, no. 1, p. 10, 2012.

[17] H. K. Kwan and S. B. Arniker, Numerical representation of DNA sequences, in *Proc. 2009 IEEE International Conf. Electro/Information Technology*, Windsor, ON, Canada, 2009, pp. 307–310.

[18] S. bai Arniker and H. K. Kwan, Advanced numerical representation of DNA sequences, in *Proc. 2012 Int. Conf. Bioscience, Biochemistry and Bioinformatices*, Singapore, 2012, pp. 1–5.

[19] D. Bielinska-Waz, Graphical and numerical representations of DNA sequences: statistical aspects of similarity, *J. Math. Chem.*, vol. 49, no. 10, pp. 2345–2407, 2011.

[20] A. Roy, C. Raychaudhury, and A. Nandy, Novel techniques of graphical representation and analysis of DNA sequences—A review, *J. Biosci.*, vol. 23, no. 1, pp. 55–71, 1998.

[21] I. Cosic, Macromolecular bioactivity: Is it resonant interaction between macromolecules?—Theory and applications, *IEEE Trans. Biomed. Eng.*, vol. 41, no. 12, pp. 1101–1114, 1994.

[22] E. Pirogova and I. Cosic, Examination of amino acid indexes within the resonant recognition model, in *Proc. 2^{nd} Conf. Victorian Chapter of the IEEE EMBS*, Melbourne, Australia, 2001, pp. 1–4.

[23] J. Ning, C. N. Moore, and J. C. Nelson, Preliminary wavelet analysis of genomic sequences, in *Proc. 2003 IEEE Bioinformatics Conf. Computational Systems Bioinformatics*, Stanford, CA, USA, 2003, pp. 509–510.

[24] A. Nair and S. P. Sreenadhan, A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), *Bioinformation*, vol. 1, no. 6, pp. 197–202, 2006.

[25] H. E. Stanley, S. V. Buldyrev, A. L. Goldberger, Z. D. Goldberger, S. Havlin, R. N. Mantegna, S. M. Ossadnik, C. K. Peng, and M. Simons, Statistical mechanics in biology: How ubiquitous are long-range correlations? *Phys. A*, vol. 205, nos. 1–3, pp. 214–253, 1994.

[26] W. Li and K. Kaneko, Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence, *EPL*, vol. 17, no. 7, p. 655, 1992.

[27] A. T. M. G. Bari, M. R. Reaz, A. K. M. T. Islam, H. J. Choi, and B. S. Jeong, Effective encoding for DNA sequence visualization based on nucleotide's ring structure, *Evol. Bioinform.*, vol. 9, pp. 251–261, 2013.

[28] K. J. Breslauer, R. Frank, H. Blcker, and L. A. Marky, Predicting DNA duplex stability from the base sequence, *Proc. Natl. Acad. Sci. USA*, vol. 83, no. 11, pp. 3746–3750, 1986.

[29] N. Yu, X. Guo, F. Gu, and Y. Pan, DNA AS X: An information-coding-based model to improve the sensitivity in comparative gene analysis, in *Bioinformatics Research and Applications*, R. Harrison, Y. H. Li, and I. Mandoiu, eds. Springer International Publishing, 2015, pp. 366–377.

[30] M. H. Garzon and R. J. Deaton, Codeword design and information encoding in DNA ensembles, *Nat. Comput.*, vol. 3, no. 3, pp. 253–292, 2004.

[31] W. Deng and Y. H. Luan, Analysis of similarity/ dissimilarity of DNA sequences based on chaos game representation, *Abstr. Appl. Anal.*, vol. 2013, p. 926519, 2013.

[32] J. Gao and Z. Y. Xu, Chaos game representation (CGR)-walk model for DNA sequences, *Chin. Phys. B*, vol. 18, no. 1, pp. 370–376, 2009.

[33] J. S. Almeida, J. A. Carriço, A. Maretzek, P. A. Noble, and M. Fletcher, Analysis of genomic sequences by chaos game representation, *Bioinformatics*, vol. 17, no. 5, pp. 429–437, 2001.

[34] L. C. B. Faria, A. S. L. Rocha, J. H. Kleinschmidt, M. C. Silva-Filho, E. Bim, R. H. Herai, M. E. B. Yamagishi, and R. Jr. Palazzo, Is a genome a codeword of an error-correcting code? *PLoS One*, vol. 7, no. 5, p. e36644, 2012.

[35] X. Liu and X. L. Geng, A convolutional code-based sequence analysis model and its application, *Int. J. Mol. Sci.*, vol. 14, no. 4, pp. 8393–8405, 2013.

[36] Z. B. Liu, B. Liao, W. Zhu, and G. H. Huang, A 2D graphical representation of DNA sequence based on dual nucleotides and its application, *Int. J. Quantum Chem.*, vol. 109, no. 5, pp. 948–958, 2009.

[37] A. S. S. Nair and T. Mahalakshmi, Visualization of genomic data using inter-nucleotide distance signals, in *Proc. IEEE Genomic Signal Processing*, Bucharest, Romania, 2005.

[38] M. Hackenberg, C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martínez-Aroza, and J. L. Oliver, *CpGcluster*: A distance-based algorithm for CpG-island detection, *BMC Bioinf.*, vol. 7, p. 446, 2006.

[39] N. Yu, X. Guo, A. Zelikovsky, and Y. Pan, GaussianCpG: A Gaussian model for detection of human CpG island, in *Proc. 5^{th} Int. Conf. Computational Advances in Bio and Medical Sciences*, Miami, FL, USA, 2015, p. 1.

[40] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, and P. J. S. G. Ferreira, Genome analysis with inter-nucleotide distances, *Bioinformatics*, vol. 25, no. 23, pp. 3064–3070, 2009.

[41] L. Q. Zhou, R. Li, and G. S. Han, A method based on the improved inter-nucleotide distances of genomes to construct vertebrates phylogeny tree, in *Proc. 7^{th} Int. Conf. Biomedical Engineering and Informatics*, Dalian, China, 2014, pp. 776–780.

[42] C. A. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. Rodrigues, and P. J. Ferreira, Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions, *J. Integr. Bioinform.*, vol. 8, no. 3, p. 172, 2011.

[43] Mujiono, I. Wasito, and I. Veritawati, Fractal dimension approach for clustering of DNA sequences based on internucleotide distance, in *Proc. 2013 Int. Conf. Information and Communication Technology*, Bandung, Indonesia, 2013, pp. 82–87.

[44] C. A. C. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. M. O. S. Rodrigues, and P. J. S. G. Ferreira, Distances between dinucleotides in the human genome, in *Proc. 5^{th} Int. Conf. Practical Applications of Computational Biology & Bioinformatics*, 2011, pp. 205–211.

[45] S. Y. Ding, Y. Li, X. W. Yang, and T. M. Wang, A simple k-word interval method for phylogenetic analysis of DNA sequences, *J. Theor. Biol.*, vol. 317, pp. 192–199, 2013.

[46] J. Tang, K. R. Hua, M. Y. Chen, R. M. Zhang, and X. L. Xie, A novel k-word relative measure for sequence comparison, *Comput. Biol. Chem.*, vol. 53, pp. 331–338, 2014.

[47] X. H. Xie, Z. G. Yu, G. S. Han, W. F. Yang, and V. Anh, Whole-proteome based phylogenetic tree construction with inter-amino-acid distances and the conditional geometric distribution profiles, *Mol. Phylogenet. Evol.*, vol. 89, pp. 37–45, 2015.

[48] S. Zou, L. Wang, and J. F. Wang, A 2D graphical representation of the sequences of DNA based on triplets and its application, *EURASIP J. Bioinform. Syst. Biol.*, vol. 2014, no. 1, p. 1, 2014.

[49] M. Akhtar, J. Epps, and E. Ambikairajah, On DNA numerical representations for period-3 based exon prediction, in *Proc. 2007 IEEE Int. Workshop on Genomic Signal Processing and Statistics*, Tuusula, Finland, 2007, pp. 1–4.

[50] K. Jabbari and G. Bernardi, Cytosine methylation and CpG, TpG (CpA) and TpA frequencies, *Gene*, vol. 333, pp. 143–149, 2004.

[51] S. Datta and A. Asif, A fast DFT based gene prediction algorithm for identification of protein coding regions, in *Proc. 2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, pp. 653–656.

[52] A. S. Motahari, G. Bresler, and D. N. C. Tse, Information theory of DNA shotgun sequencing, *IEEE Trans. Inf. Theory*, vol. 59, no. 10, pp. 6273–6289, 2013.

[53] M. W. Simmen, Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals, *Genomics*, vol. 92, no. 1, pp. 33–40, 2008.

[54] J. Tuqan and A. Rushdi, A DSP approach for finding the codon bias in DNA sequences, *IEEE J. Sel. Top. Signal Process.*, vol. 2, no. 3, pp. 343–356, 2008.

[55] L. Galleani and R. Garello, The minimum entropy mapping spectrum of a DNA sequence, *IEEE Trans. Inf. Theory*, vol. 56, no. 2, pp. 771–783, 2010.

[56] R. Román-Roldán, P. Bernaola-Galván, and J. Oliver, Application of information theory to DNA sequence analysis: A review, *Pattern Recognition*, vol. 29, no. 7, pp. 1187–1194, 1996.

[57] 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, 2000.

[58] P. Dan Cristea, Genetic signal representation and analysis, in *Proc. Functional Monitoring and Drug-Tissue Interaction*, San Jose, CA, USA, 2002, pp. 77–84.

[59] P. Cristea, Genetic signal analysis, in *Proc. 6^{th} Int. Symp. Signal Processing and Its Applications*, Kuala Lumpur, Malaysia, 2001, pp. 703–706.

[60] P. D. N. Hebert, A. Cywinska, S. L. Ball, and J. R. deWaard, Biological identifications through DNA barcodes, *Proc. Roy. Soc. B Biol. Sci.*, vol. 270, no. 1512, pp. 313–321, 2003.

[61] S. Ratnasingham and P. D. N. Hebert, Bold: The barcode of life data system, *Mol. Ecol. Notes*, vol. 7, no. 3, pp. 355–364, 2007.

[62] V. Afreixo, C. A. C. Bastos, A. J. Pinho, S. P. Garcia, and P. J. S. G. Ferreira, Genome analysis with distance to the nearest dissimilar nucleotide, *J. Theor. Biol.*, vol. 275, no. 1, pp. 52–58, 2011.

[63] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler, The human genome browser at UCSC, *Genome Res.*, vol. 12, no. 6, pp. 996–1006, 2002.

[64] G. Kauer and H. Blöcker, Applying signal theory to the analysis of biomolecules, *Bioinformatics*, vol. 19, no. 16, pp. 2016–2021, 2003.

[65] E. A. Cheever, D. B. Searls, W. Karunaratne, and G. C. Overton, Using signal processing techniques for DNA sequence comparison, in *Proc. 15^{th} Annu. Northeast Bioengineering Conference*, Boston, MA, USA, 1989, pp. 173–174.

[66] H. K. Kwan, B. Y. M. Kwan, and J. Y. Y. Kwan, Novel methodologies for spectral classification of exon and intron sequences, *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, p. 50, 2012.

[67] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, *New Approaches to Genome Sequence Analysis Based on Digital Signal Processing*. University of California, CA, USA, 2002.

[68] N. Rao and S. J. Shepherd, Detection of 3- periodicity for small genomic sequences based on AR technique, in *Proc. 2004 Int. Conf. Communications, Circuits and Systems*, Chengdu, China, 2004, pp. 1032–1036.

[69] N. Chakravarthy, A. Spanias, L. D. Iasemidis, and K. Tsakalis, Autoregressive modeling and feature analysis of DNA sequences, *EURASIP J. Appl. Signal Process.*, vol. 2004, p. 952689, 2004.

[70] Z. G. Yu, V. V. Anh, Y. Zhou, and L. Q. Zhou, Numerical sequence representation of DNA sequences and methods to distinguish coding and non-coding sequences in a complete genome, in *Proc. 11^{th} World Multi-Conf. Systemics, Cybernetics and Informatics: WMSCI 2007*, 2007, pp. 171–176.

[71] A. K. Brodzik and O. Peters, Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences, in *Proc. 2005 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 2005, pp. 373–376.

[72] G. Rosen, Examining coding structure and redundancy in DNA, *IEEE Eng. Med. Biol. Mag.*, vol. 25, no. 1, pp. 62–68, 2006.

[73] G. L. Rosen and J. D. Moore, Investigation of coding structure in DNA, in *Proc. 2003 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Hong Kong, China, 2003, p. II–361–4.

[74] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley, Long-range correlations in nucleotide sequences, *Nature*, vol. 356, no. 6365, pp. 168–170, 1992.

[75] J. A. Berger, S. K. Mitra, M. Carli, and A. Neri, Visualization and analysis of DNA sequences using DNA walks, *J. Franklin Inst.*, vol. 341, nos. 1&2, pp. 37–53, 2004.

[76] S. Tiwari, S. Ramachandran, A. Bhattacharya, S. Bhattacharya, and R. Ramaswamy, Prediction of probable genes by Fourier analysis of genomic sequences, *Bioinformatics*, vol. 13, no. 3, pp. 263–270, 1997.

[77] W. T. Li, T. G. Marr, and K. Kaneko, Understanding long-range correlations in DNA sequences, *Phys. D Nonlinear Phenom.*, vol. 75, nos. 1–3, pp. 392–416, 1994.

[78] O. Abbasi, A. Rostami, and G. Karimian, Identification of exonic regions in DNA sequences using cross-correlation and noise suppression by discrete wavelet transform, *BMC Bioinformatics*, vol. 12, p. 430, 2011.

[79] S. P. Deng, Y. X. Shi, L. Y. Yuan, Y. X. Li, and G. H. Ding, Detecting the borders between coding and non-coding DNA regions in prokaryotes based on recursive segmentation and nucleotide doublets statistics, *BMC Genomics*, vol. 13, no. Suppl 8, p. S19, 2012.

[80] C. A. C. Bastos, V. Afreixo, S. P. Garcia, and A. J. Pinho, Inter-stop symbol distances for the identification of coding regions, *J. Integr. Bioinform.*, vol. 10, no. 3, p. 230, 2013.

[81] G. L. Rosen, Signal processing for biologically-inspired gradient source localization and DNA sequence analysis, PhD dissertation, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA, USA, 2006.

[82] D. Limbachiya, B. Rao, and M. K. Gupta, The art of DNA strings: Sixteen years of DNA coding theory, arXiv preprint arXiv: 1607.00266, 2016.

[83] L. C. B. Faria, A. S. L. Rocha, J. H. Kleinschmidt, R. Palazzo, and M. C. Silva-Filho, DNA sequences generated by BCH codes over GF(4), *Electron. Lett.*, vol. 46, no. 3, pp. 203–204, 2010.

[84] L. Zhang, F. C. Tian, S. Y. Wang, and X. Liu, A novel coding method for gene mutation correction during protein translation process, *J. Theor. Biol.*, vol. 296, pp. 33–40, 2012.

[85] F. Castro-Chavez, A tetrahedral representation of the genetic code emphasizing aspects of symmetry, *BIOcomplexity*, vol. 2012, no. 2, pp. 1–6, 2012.

[86] 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. 1, no. 1, p. 3, 2012.

[87] H. J. Jeffrey, Chaos game representation of gene structure, *Nucleic Acids Res.*, vol. 18, no. 8, pp. 2163–2170, 1990.

[88] Y. W. Wang, K. Hill, S. Singh, and L. Kari, The spectrum of genomic signatures: From dinucleotides to chaos game representation, *Gene*, vol. 346, pp. 173–185, 2005.

[89] J. Joseph and R. Sasikumar, Chaos game representation for comparison of whole genomes, *BMC Bioinformatics*, vol. 7, p. 243, 2006.

[90] C. Dutta and J. Das, Mathematical characterization of chaos game representation: New algorithms for nucleotide sequence analysis, *J. Mol. Biol.*, vol. 228, no. 3, pp. 715–719, 1992.

[91] N. Goldman, Nucleotide, dinucleotide and trinucleotide frequencies explain patterns observed in chaos game representations of DNA sequences, *Nucleic Acids Res.*, vol. 21, no. 10, pp. 2487–2491, 1993.

[92] F. Castro-Chavez, Most used codons per amino acid and per genome in the code of man compared to other organisms according to the rotating circular genetic code, *Neuroquantology*, vol. 9, no. 4, p. 500, 2011.

[93] S. Delgado, F. Morán, A. Mora, J. J. Merelo, and C. Briones, A novel representation of genomic sequences for taxonomic clustering and visualization by means of self-organizing maps, *Bioinformatics*, vol. 31, no. 5, pp. 736–744, 2015.

[94] 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, 2001.

[95] H. T. Chang, N. W. Lo, W. C. Lu, and C. J. Kuo, Visualization and comparison of DNA sequences by use of three-dimensional trajectories, in *Proc. 1st Asia-Pacific Bioinformatics Conf. Bioinformatics 2003*, Adelaide, Australia, 2003, pp. 81–85.

[96] T. Kohonen, Self-organized formation of topologically correct feature maps, *Biol. Cybern.*, vol. 43, no. 1, pp. 59–69, 1982.

[97] T. Kohonen and P. Somervuo, How to make large self-organizing maps for nonvectorial data, *Neural Netw.*, vol. 15, nos. 8&9, pp. 945–952, 2002.

[98] A. P. Boyle, C. L. Araya, C. Brdlik, P. Cayting, C. Cheng, Y. Cheng, K. Gardner, L. W. Hillier, J. Janette, L. X. Jiang, D. Kasper, et al., Comparative analysis of regulatory information and circuits across distant species, *Nature*, vol. 512, no. 7515, pp. 453–456, 2014.

[99] E. Hamori and J. Ruskin, H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences, *J. Biol. Chem.*, vol. 258, no. 2, pp. 1318–1327, 1983.

[100] M. A. Gates, Simpler DNA sequence representations, *Nature*, vol. 316, no. 6025, p. 219, 1985.

[101] S. S. T. Yau, J. S. Wang, A. Niknejad, C. X. Lu, N. Jin, and Y. K. Ho, DNA sequence representation without degeneracy, *Nucleic Acids Res.*, vol. 31, no. 12, pp. 3078–3080, 2003.

[102] R. Zhang and C. T. Zhang, Z curves, an intutive tool for visualizing and analyzing the DNA sequences, *J. Biomol. Struct. Dyn.*, vol. 11, no. 4, pp. 767–782, 1994.

[103] H. K. Kwan, R. Atwal, and B. Y. M. Kwan, Wavelet analysis of DNA sequences, in *Proc. 2008 Int. Conf. Communications, Circuits and Systems*, Fujian, China, 2008, pp. 816–820.

[104] C. L. 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, p. e101363, 2014.

[105] 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, 2008.

[106] 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, 2017.

[107] R. Ranawana and V. Palade, A neural network based multi-classifier system for gene identification in DNA sequences, *Neural Comput. Appl.*, vol. 14, no. 2, pp. 122–131, 2005.

[108] S. B. Arniker, H. K. Kwan, N. F. Law, and D. P. K. Lun, DNA numerical representation and neural network based human promoter prediction system, in *Proc. 2011 Annu. IEEE India Conf.*, Hyderabad, India, 2011, pp. 1–4.

[109] X. Xie, S. Wu, K. M. Lam, and H. Yan, Promoterexplorer: An effective promoter identification method based on the AdaBoost algorithm, *Bioinformatics*, vol. 22, no. 22, pp.

2722–2728, 2006.
[110] L. Deng and D. Yu, Deep learning: Methods and applications, Tech. Rep. MSR-TR-2014-21, 2014, http://research.microsoft.com/apps/pubs/default.aspx?id=209355
[111] Y. Bengio, A. Courville, and P. Vincent, Representation learning: A review and new perspectives, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.
[112] M. G. Reese, F. H. Eeckman, D. Kulp, and D. Haussler, Improved splice site detection in genie, *J. Comput. Biol.*, vol. 4, no. 3, pp. 311–323, 1997.
[113] N. Yu, Z. Yu, and Y. Pan, A deep learning method for lincRNA detection using auto-encoder algorithm, *BMC Bioinformatics*, vol. 18, no. Suppl 15, p. 511, 2017.
[114] G. B. Orr and K. R. Müller, *Neural Networks: Tricks of the Trade*. Springer, 1998, p. 1524.
[115] S. Wiesler, A. Richard, R. Schluter, and H. Ney, Mean-normalized stochastic gradient for large-scale deep learning, in *Proc. 2014 IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Florence, Italy, 2014, pp. 180–184.
[116] T. Raiko, H. Valpola, and Y. LeCun, Deep learning made easier by linear transformations in perceptrons, in *Proc. 15th Int. Conf. Artificial Intelligence and Statistics*, La Palma, Canary Islands, 2012, pp. 924–932.
[117] S. Ioffe and C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv: 1502.03167, 2015.
[118] I. Danihelka, G. Wayne, B. Uria, N. Kalchbrenner, and A. Graves, Associative long short-term memory, arXiv preprint arXiv: 1602.03032, 2016.
[119] C. Jose, M. Cisse, and F. Fleuret, Kronecker recurrent units, arXiv preprint arXiv: 1705.10142, 2017.
[120] L. Jing, Ç. Gülçehre, J. Peurifoy, Y. C. Shen, M. Tegmark, M. Soljacic, and Y. Bengio, Gated orthogonal recurrent units: On learning to forget, arXiv preprint arXiv: 1706.02761, 2017.
[121] M. Arjovsky, A. Shah, and Y. Bengio, Unitary evolution recurrent neural networks, arXiv preprint arXiv: 1511.06464, 2015.
[122] C. Trabelsi, O. Bilaniuk, Y. Zhang, D. Serdyuk, S. Subramanian, J. F. Santos, S. Mehri, N. Rostamzadeh, Y. Bengio, and C. J. Pal, Deep complex networks, arXiv preprint arXiv: 1705.09792, 2017.
[123] L. Mescheder, S. Nowozin, and A. Geiger, The numerics of GANs, arXiv preprint arXiv: 1705.10461, 2017.

Ning Yu currently is an assistant professor at the State University of New York College at Brockport. He earned the PhD degree in computer science from Georgia State University in 2016 and has published more than 20 papers in prestigious journals/conferences, such as IEEE Transactions, BMC Bioinformatics, and PLOS One. His current research focuses on big data analytics, deep learning, network and information security, information processing, and high performance computing.

Zhihua Li received the PhD degree in computer science from Jiangnan University, Wuxi, China in 2009. His research interests include network technology, parallel and distributed computing, information security, data mining, and pattern recognition. He is currently a professor of the Department of Computer Science and Technology at Jiangnan University.

Zeng Yu received the BS and MS degrees from China University of Mining and Technology in 2008 and 2011, respectively. He was a visiting scholar with Georgia State University, USA, from 2014 to 2016. He is currently a PhD candidate in the School of Information Science and Technology, Southwest Jiaotong University, China. His research interests include data mining, bioinformatics, deep learning, and cloud computing.