AppliedMath

MDPI

Article

Models of Low-Dimensional Vector-Fuzzy Representations of Genetic Sequences and Amino Acids

Fotini Sereti [1], Dimitrios Georgiou [2] and Theodoros Karakasidis [3,*]

[1] Department of Chemical Engineering, University of Western Macedonia, 50100 Kozani, Greece; aff01310@uowm.gr
[2] Department of Mathematics, University of Patras, 26504 Patra, Greece; georgiou@math.upatas.gr
[3] Condensed Matter Physics Laboratory, Department of Physics, University of Thessaly, 35100 Lamia, Greece
* Correspondence: thkarak@uth.gr

Abstract

Genetic sequences play a central role in biological and medical research, and mathematics provides powerful means for their representation and analysis. Conventional approaches, such as the fuzzy polynucleotide space $[0,\ 1]^{12}$, model codons as 12-dimensional vectors, but this comes at the cost of high dimensionality. In this study, we introduce two new models, Vector-Fuzzy-I and Vector-Fuzzy-II, that map codons and genetic sequences into the 4-dimensional Euclidean space $\mathbb{R}^4$ using vector algebra and fuzzy set theory. In the first model, sequence structure is represented by successive vector addition, while in the second, it is represented by positional frequencies normalized by nucleotide locations. These low-dimensional representations are unique, preserve sequence order, and allow effective measurement of similarity and difference via Euclidean metrics. Compared with the fuzzy polynucleotide space, the proposed models achieve dimensionality reduction while enhancing the resolution of sequence differentiation. Our approach offers new mathematical perspectives for sequence analysis in theoretical biology.

Keywords: DNA; RNA; amino acid; pseudo amino acid composition; polynucleotides; vectors; fuzzy sets; Euclidean metric

1. Introduction

It is commonly accepted that the science of bioinformatics is a constantly developing field in which mathematics play an important role in the analysis of genetic sequences (see for example [1,2]) in order to study biological data in a different manner. We understand that the genetic material of living organisms consist of nucleic acids, DNA and RNA, and that their investigation is a fundamental task in the science of medicine, ranging from understanding disease mechanisms to studying evolution.

The "language" of DNA and RNA consists of four nucleotides, U, C, A, G in the case of DNA, and T, C, A, G in the case of RNA (A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil) [3]. Particularly, DNA and RNA are made of triplets, XYZ, of three nucleic acids (that is, they form an amino acid or a codon) each of which have the possibility to be one of the above "letters". Although over 500 amino acids exist in nature, by far the most important are the 20 amino acids incorporated into proteins and appear with their genetic code (see Table 1).

Check for updates

Received: 11 January 2026
Revised: 15 February 2026
Accepted: 26 February 2026
Published: 4 March 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Table 1. The 20 amino acids.

Amino Acid	3-Letter Code	Name	Reverse Codon
1.	Ala	Alanine	GCU, GCC, GCA, GCG
2.	Cys	Cysteine	UGU, UGC
3.	Asp	Aspartic acid	GAU, GAC
4.	Glu	Glutamic acid	GAA, GAG
5.	Phe	Phenylalanine	UUU, UUC
6.	Gly	Glycine	GGU, GGC, GGA, GGG
7.	His	Histidine	CAU, CAC
8.	Ile	Isoleucine	AUU, AUC, AUA
9.	Lys	Lysine	AAA, AAG
10.	Leu	Leucine	UUA, UUG, CUU, CUC, CUA, CUG
11.	Met	Methionine	AUG
12.	Asn	Asparagine	AAU, AAC
13.	Pro	Proline	CCU, CCC, CCA, CCG
14.	Gln	Glutamine	CAA, CAG
15.	Arg	Arginine	CGU, CGC, CGA, CGG, AGA, AGG
16.	Ser	Serine	UCU, UCC, UCA, UCG, AGU, AGC
17.	Thr	Threonine	ACU, ACC, ACA, ACG
18.	Val	Valine	GUU, GUC, GUA, GUG
19.	Trp	Tryptophane	UGG
20.	Tyr	Tyrosine	UAU, UAC

Related studies refer to amino acids, their biological distance and the concept of pseudo amino acid (pseAA) composition proposed in [4]. Several investigations have been made in order to use various numbers to represent the amino acids through the vehicle of pseAA composition (see for example [5–14]). Thus, it is absolutely natural to give special attention to amino acids, pseudo-amino acid composition and genetic sequences in general. There are several methods that are used in order to extract the characteristics of genomes and find some common characteristics among their constituents. Sadegh-Zadeh [15] showed that a genetic code can be represented in a 12-dimensional space, because a triplet codon XYZ has a $3 \times 4 = 12$ dimensional fuzzy code $(a_1, \cdots, a_{12})$ and is a point in the 12-dimensional fuzzy polynucleotide space $I^{12} = [0,1]^{12}$ (denote by $I = [0,1]$). Additionally, in the same paper, the authors introduced the fuzzy polynucleotide space (FPS) based on the principle of the fuzzy hypercube [16]. In this notation a polynucleotide consisting of a sequence of k triplets XYZ is a point in a I^{12}.

Torres and Nieto [17] mapped a polynucleotide on an I^{12} space by considering the frequencies of presence of the nucleotides at the three base sites of a codon in the coding sequence. In that work, referring to the metric of [15,18,19], they calculated the distances between nucleotides. Further work has been performed using the idea of [20], in which the influences of several metrics have been examined. We note that, in related studies, it is also very important to be in a position to determine how close two genetic sequences are, as there are many important biological and medical implications (see for example [21–23]).

Another line of research involves time-series methods, where physicochemical indices of 20 amino acids are combined with the Hungarian algorithm to map each amino acid into a vector. In this way, a protein sequence can be transformed into a time series within an eleven-dimensional space [24]. Other studies, like that of [25], use systematic experiments with synthetic sequences to explore how CNN architecture—particularly convolutional filter size and max-pooling—affects the ability of first-layer filters to learn sequence motif

representations. Their findings suggest that CNNs designed to promote hierarchical representation learning capture sequence motifs more effectively.

There is also a growing interest in incorporating contextual features. By representing similarity to homologous proteins both within and across species, the prediction of shared paralog functions can be significantly improved. Overall, these results demonstrate that alternative similarity metrics capture complementary aspects of functional similarity that go beyond sequence identity alone [26]. Comparative methods such as sequence analysis are commonly used to detect differences between sequences, whether at the single nucleotide level or as the result of larger phenomena like recombination or deletion. Identifying these variations is crucial for biological and medical insights; however, the massive scale and complexity of genomic data demand considerable computational resources [27].

As the applications of mathematics develop new directions in biological and medical studies (see for example [28–30]), the mathematical branch of vector algebra comes to introduce a new chapter in the direction of mathematical biology. Vector algebra has many applications in medicine such as gene therapy and data analysis of gene expression. Vectors, in this field, can represent genetic material for therapeutic delivery and sets of variables for analysis (see for example [31,32]).

1.1. Description of the Contributions of the Paper

Amino acids are essential components used to synthesize proteins in the human body. In general, protein sequences are the linear orders of amino acids within a polypeptide chain. A schematic representation of such a chain is given in Figure 1, where each amino acid of the chain is presented in a different color.

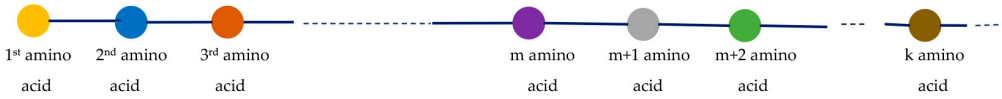


Figure 1. A schematic representation of a genetic sequence.

A chain of single-letter residue codes contains significant information. As protein sequences number many amino acids in their structure forms and are connected with the physicochemical properties of every amino acid or part of this linear order in the view of genetic sequences, we are motivated to study new mathematical representations of amino acids and genetic sequences.

Until now, we have understood the corresponding representation of such biological data in the space $[0,\ 1]^{12}$ as a subspace of $\mathbb{R}^{12}$. However, it is believed that the procedure of reducing amino acid composition serves not only as a key method for analyzing protein structure and function but also expands opportunities within the broader field of machine learning (see for example [33] where the authors, taking into consideration this necessity of reduction, reviewed the strategy and method studies of the reduced amino acid alphabets). Thus, in our investigation, in order to facilitate simplicity of implementation (with less coordinates in the amino acids' compositions) and faster differentiation between genetic sequences, we reduce the dimension of the space where such biological data can be "embedded" and represented in different vector forms.

More precisely, we follow the directions of vector algebra and fuzzy set theory, combine them with the field of representations of biological data and develop new mathematical models for studying amino acids and genetic sequences in general. We pay attention to the following axes:

- We introduce two new representations of genetic sequences in $\mathbb{R}^4$, using vectors and fuzzy sets. In the new representations, we succeed in representing nucleotides, triplets and general genetic sequences in the 4-dimensional Euclidean space $\mathbb{R}^4$, assigning

unit vectors of $\mathbb{R}^4$ to such sequences. Thus, we avoid restricting ourselves to only the 12-dimensional space $\mathbb{R}^{12}$ and investigate the geometrical image in the Euclidean space $\mathbb{R}^4$, significantly decreasing the number of coordinates of the corresponding biological data: amino acids and larger genetic sequences.

- Having these new representations, we investigate the similarity, difference and Euclidean distances between genetic sequences, studying the influence of the new methodologies. For that, we compare the new models with known models representing genetic sequences in the fuzzy polynucleotide space $[0, 1]^{12}$. Our study proves that there is a better differentiation of the sequences in the new models and that the criteria of similarity, difference and Euclidean distance provide better qualitative and quantitative data.

1.2. Outline of the Paper

The paper is organized as follows. In Section 2, we present the basic notions related to fuzzy set theory, vector algebra, metrics and similarity/difference. In Section 3, we insert the first new model of representation for genetic sequences, giving its various applications for known amino acids and genetic sequences. In Section 4, we present the second model, providing its applications for known amino acids and genetic sequences. Finally, in Section 5, we investigate the influence of the new methodologies, comparing then with the known methodologies representing genetic sequences in the fuzzy polynucleotide space $[0, 1]^{12}$. The characteristics of Euclidean distance and similarity/difference are studied.

2. Preliminaries

In this section we give some basic notions which are useful for the rest of our study. In particular, we present three subsections: the first for fuzzy sets, the second for the fuzzy polynucleotide space (FPS) and the third for vector algebra.

2.1. Basic Notions for Fuzzy Sets

We begin our study reminding the reader of the basic notions of fuzzy set theory. For more details we refer to [34–37]. Fuzzy sets were invented in 1965 by Lotfi Zadeh [34] and they replace the notion of an element's participation to a set (being or not being in a set) with the notion of the element's participation in the set with the degree of membership (see [35–37]). There are several modifications of fuzzy sets and, recently, in [38], the authors inserted the notion of $\mathbb{Q}(\varepsilon)$-fuzzy sets: sets with values that may be infinitesimal or infinitely close to some rational number from the interval [0, 1]. However, in our investigation we always take into consideration the standard notion of fuzzy sets as follows:

Definition 1 ([34]). *Let X be a set. A subset A of X is called fuzzy subset of X if there is a function* $\boldsymbol{\mu_A}$ *such that*

(1) μ_A*: X* $\to$ *[0, 1] and*

(2) *A = {(x,* $\mu_A(x)$*): x* $\in$ *X}, that is A is the set of all pairs (x,* $\mu_A(x)$*) such that x* $\in$ *X and* $\mu_A(x)$ *is the degree of its membership in A.*

The nearer the value of $\boldsymbol{\mu_A}(x)$ to unity, the higher the grade of membership of x in A. That is, for x $\in$ X, $\boldsymbol{\mu_A(x) \in [0, 1]}$ gives the so-called *degree of participation* of x in A. In what follows if $\mathbf{X = \{x_1, x_2, \cdots, x_n\}}$ and $\mathbf{A = \{(x_1, \mu_A(x_1), \cdots, (x_n, \mu_A(x_n))\}}$, then we simply write $\mathbf{A = (\mu_A(x_1), \cdots, \mu_A(x_n))}$ or $\mathbf{A = (a_1, \cdots, a_n)}$, where $\mathbf{\mu_A(x_i) = a_i}$ and $\mathbf{a_i \in [0, 1]}$, for each $\mathbf{i = 1, 2, \cdots, n}$.

Definition 2 ([34]). *Let A and B be two fuzzy subsets of a set X. Then*

(1) by $A \wedge B$ we define the fuzzy set for which the membership function $\mu_{A\wedge B}$: $X \to [0, 1]$ is given as

$$\mu_{A\wedge B}(x) = \min\{\mu_A(x), \mu_B(x)\}, \text{ for every } x \in X \quad (1)$$

(2) by $A \vee B$ we define the fuzzy set for which the membership function $\mu_{A\vee B}$: $X \to [0, 1]$ is given as

$$\mu_{A\vee B}(x) = \max\{\mu_A(x), \mu_B(x)\}, \text{ for every } x \in X \quad (2)$$

Additionally, for a set $X = \{x_1, x_2, \cdots, x_n\}$ and a fuzzy subset $A = (a_1, \cdots, a_n)$ of X, where $a_i \in [0, 1]$, by c(A) (see [15,19]) we denote the number

$$c(A) = \sum_{i=1}^{n} a_i \quad (3)$$

Definition 3 ([39]). *Let $X = \{x_1, x_2, \cdots, x_n\}$ be a set and $A = (a_1, \cdots, a_n)$, $B = (b_1, \cdots, b_n)$, where a_i, $b_i \in [0, 1]$, be two fuzzy subsets of X.*

(1) The degree of similarity between A and B, denoted by sim(A,B), is defined to be the number

$$sim(A, B) = \frac{c(A \wedge B)}{c(C)} \quad (4)$$

where C is the fuzzy set $\left(\frac{a_1+b_1}{2}, \cdots, \frac{a_n+b_n}{2}\right)$, that is C is the canonical midpoint between A and B.

(2) The degree of difference between A and B, denoted by dif(A,B), is defined to be the number

$$dif(A, B) = 1 - sim(A, B). \quad (5)$$

2.2. Fuzzy Polynucleotide Space (FPS)

Let us describe how amino acids and genetic sequences can be "embedded" in the hypercube I^{12}. A codon corresponds to a corner of the 12-dimensional unit hypercube I^{12}, called *fuzzy polynucleotide space* (*in short, FPS*). In particular, any element of I^{12} may be viewed as a fuzzy codon. In the case of DNA, we make the following correspondence: U = (1, 0, 0, 0), C = (0, 1, 0, 0), A = (0, 0, 1, 0) and G = (0, 0, 0, 1). As a consequence, if we have a nucleotide described by the codon UCG (serine) this would be written in the I^{12} as

$$UGG = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1). \quad (6)$$

When we have a polynucleotide which is a sequence of m triplets, one would need an $m \times 12$ hyperspace. For example, if we have the polynucleotide described by the sequence s_1 = UAC-UGU (tyrosine/cysteine), it is a point in $I^{2\times 12} = I^{24}$ and is represented by

$$s_1 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0). \quad (7)$$

However, if one considers the frequencies of the nucleotides of the alphabet at the three base sites of a codon (in the coding sequence or any other region) it may be viewed as a point in the hypercube I^{12} as given in the following FPS representation (see Table 2).

Table 2. FPS representation of genetic sequences.

Genetic Sequence	Symbol	Genetic Code	FPS Representation
tyrosine/cysteine	s_1	UAC-UGU	(1, 0, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0)
histidine/cysteine	s_2	CAC-UGU	(0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0)
leucine/cysteine	s_3	CUC-UGU	(0.5, 0.5, 0, 0, 0.5, 0, 0, 0.5, 0.5, 0.5, 0, 0)
histidine/cysteine	s_4	CAU-UGU	(0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 1, 0, 0, 0)
glutamine/cysteine	s_5	CAG-UGU	(0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0.5)
glutamine/cysteine	s_6	CAA-UGU	(0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0, 0.5, 0)

2.3. Basic Notions for Vectors

When dealing with genetic sequences it is of interest to be able to describe how different two sequences are. For this reason, the notions of Euclidean distance, cosine similarity (i.e., cosine of angle), similarity and difference are used (see [15,20,39]). One of the most well-known mathematical spaces is the Euclidean space $\mathbb{R}^n$, $n = 1, 2, \cdots$, where the Euclidean distance and the norm are defined as follows, respectively:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{8}$$

for every $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, $\mathbf{y} = (y_1, y_2, \cdots, y_n) \in \mathbb{R}^n$ and

$$||\mathbf{x}||_d = \sqrt{\sum_{i=1}^{n} x_i^2} \tag{9}$$

for every $\mathbf{x} = (x_1, x_2, \cdots, x_n) \in \mathbb{R}^n$ (subsequently, following the usual notations of vectors, we denote by $\mathbf{x}$ (in bold) any vector). For more details on vector and linear algebra we refer to [40]. Several other metrics and applications are also presented in [41–43]. However, in this paper we only focus on the Euclidean space $\mathbb{R}^4$.

From a geometric point of view, the sum of two vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ and $\mathbf{y} = (y_1, y_2, \cdots, y_n)$, belonging to $\mathbb{R}^n$, can been seen as the diagonal of a parallelogram formed by those two vectors (see Figure 2).

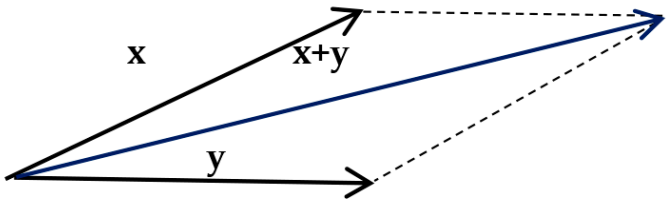


Figure 2. Schematic representation of the vectors **x** and **y** (in black) and their sum **x** + **y** (in blue), drawing the corresponding parallelogram and its diagonal.

Furthermore, having two vectors **x** and **y**, the so-called *cosine similarity* is a metric used to measure how similar given documents are irrespective of their size. In a pure mathematical view, it measures the cosine of the angle, say θ, between the vectors projected in a multi-dimensional space (see Figure 3). The smaller the angle, the higher the cosine similarity [43]. In particular, having two vectors $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, $\mathbf{y} = (y_1, y_2, \cdots, y_n)$ in $\mathbb{R}^n$, $n = 1, 2, \cdots$, different to $\mathbf{0} = (0, 0, \cdots, 0) \in \mathbb{R}^n$, then the *angle* θ between **x** and **y** is defined as follows:

(1) $0 \leq \theta \leq \pi$ and
(2) $\cos\theta = \frac{1}{\|x\|_d \cdot \|y\|_d} \sum_{i=1}^{n} x_i \cdot y_i$.

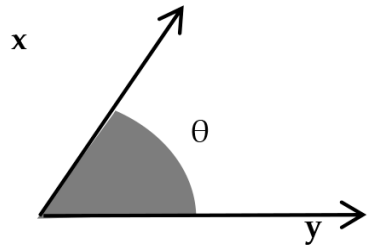


Figure 3. Schematic representation of the angle θ between the vectors **x** and **y**.

3. Vector-Fuzzy-I Representation of Genetic Sequences

In this section, we present the first model of vector representation of the biological data of amino acids and genetic sequences.

Let $(\mathbb{R}^4,+,)$ be the Euclidean vector space and let $\boldsymbol{r_1}$ = (1, 0, 0, 0), $\boldsymbol{r_2}$ = (0, 1, 0, 0), $\boldsymbol{r_3}$ = (0, 0, 1, 0) and $\boldsymbol{r_4}$ = (0, 0, 0, 1) be the linearly independent unit vectors of the space $\mathbb{R}^4$. The first model to provide the representation in $\mathbb{R}^4$ of a given genetic sequence

$$s = \boldsymbol{u_1u_2u_3} - \boldsymbol{u_4u_5u_6} - \boldsymbol{u_7u_8u_9} - \cdots - \boldsymbol{u_{k-5}u_{k-4}u_{k-3}} - \boldsymbol{u_{k-2}u_{k-1}u_k}. \tag{10}$$

which consists of k nucleotides, where the elements of the set $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_k}\}$ are elements of the set {U, C, A, G} and which is described in the following rule (note that, since we have triplets XYZ of nucleotides, the above k is a natural number divided by 3 and these k nucleotides are placed at $\frac{k}{3}$-triplets).

New Representation
[called *Vector-Fuzzy-I (in short VF-I) representation of s*]

Step 1. We make the correspondence of unit vectors to each codon as follows:

$$\boldsymbol{r_1} = (1,\ 0,\ 0,\ 0) = U,\ \boldsymbol{r_2} = (0,\ 1,\ 0,\ 0) = C,\ \boldsymbol{r_3} = (0,\ 0,\ 1,\ 0) = A,\ \boldsymbol{r_4} = (0,\ 0,\ 0,\ 1) = G. \tag{11}$$

As each $\boldsymbol{u_i}$, i = 1, 2, ⋯, k, of the above sequence s is one of the elements U, C, A, G, we find that each $\boldsymbol{u_i}$, i = 1, 2, ⋯, k, is one of the vectors $\boldsymbol{r_1}$, $\boldsymbol{r_2}$, $\boldsymbol{r_3}$, $\boldsymbol{r_4}$. Go to Step 2.

Step 2. Find the vector:

$$\boldsymbol{w_1} = \frac{\boldsymbol{u_1} + \boldsymbol{u_2}}{\|\boldsymbol{u_1} + \boldsymbol{u_2}\|_d} \tag{12}$$

and go to Step 3.

Step 3. Find the vector:

$$\boldsymbol{w_2} = \frac{\boldsymbol{w_1} + \boldsymbol{u_3}}{\|\boldsymbol{w_1} + \boldsymbol{u_3}\|_d} \tag{13}$$

and go to Step 4.

Step 4. Find the vector:

$$\mathbf{w_3} = \frac{\mathbf{w_2} + \mathbf{u_4}}{\|\mathbf{w_2} + \boldsymbol{u_4}\|_{\mathrm{d}}} \tag{14}$$

and go to Step 5.

Step k − 1. Find the vector:

$$\mathbf{w_{k-2}} = \frac{\mathbf{w_{k-3}} + \mathbf{u_{k-1}}}{\|\mathbf{w_{k-3}} + \boldsymbol{u_{k-1}}\|_{\mathrm{d}}} \tag{15}$$

and go to Step k.

Step k. Find the vector:

$$\mathbf{w_{k-1}} = \frac{\mathbf{w_{k-2}} + \mathbf{u_k}}{\|\mathbf{w_{k-2}} + \boldsymbol{u_k}\|_{\mathrm{d}}} \tag{16}$$

and go to Step k + 1.

Step k + 1. Assign the genetic sequence s to the vector $\mathbf{w_{k-1}}$.

Clearly, the vectors $\mathbf{w_1}, \mathbf{w_2}, \cdots \mathbf{w_{k-1}}$ are elements of the hypercube I^4 and the location of the nucleotides U, C, A, G ensures the uniqueness of each representation, getting the following remark.

Remark 1. *The VF-I representation of genetic sequences satisfies the following:*

(1) The final vector $\boldsymbol{w}$ *of the VF-I representation of a genetic sequence s is unique.*

(2) If s_1 *and* s_2 *are two different genetic sequences and* $\boldsymbol{w_1}$, $\boldsymbol{w_2}$ *are their VF-I representations, respectively, then* $\boldsymbol{w_1} \neq \boldsymbol{w_2}$.

In the following examples we shall apply the VF-I representation for amino acids and genetic sequences (using their genetic code and Chou's pseudo-amino acid composition) in order to succeed their vector-image in the Euclidean space $\mathbb{R}^4$.

Example 1. *We shall find the VF-I representation of some known nucleotides as follows.*

(1) For CAU (histidine) we have $C = \boldsymbol{r_2}$, $A = \boldsymbol{r_3}$ *and* $U = \boldsymbol{r_1}$. *Thus, the corresponding vectors of VF-I representation are as follows:*

Step 1. *We add the first two vectors corresponding to C and A and construct the corresponding unit vector, as follows:*

$$\boldsymbol{w_1} = \frac{\boldsymbol{r_2} + \boldsymbol{r_3}}{\|\boldsymbol{r_2} + \boldsymbol{r_3}\|_d} = \frac{(0,\ 1,\ 0,\ 0) + (0,\ 0,\ 1,\ 0)}{\|(0,\ 1,\ 0,\ 0) + (0,\ 0,\ 1,\ 0)\|_d} = \frac{(0,\ 1,\ 1,\ 0)}{\sqrt{2}} = \left(0,\ \frac{\sqrt{2}}{2},\ \frac{\sqrt{2}}{2},\ 0\right) \tag{17}$$

Step 2. *We add to the previous unit vector* $\boldsymbol{w_1}$ *the next vector corresponding to U, as follows:*

$$\boldsymbol{w_2} = \frac{\boldsymbol{w_1} + \boldsymbol{r_1}}{\|\boldsymbol{w_1} + \boldsymbol{r_1}\|_d} = \frac{\left(0,\ \frac{\sqrt{2}}{2},\ \frac{\sqrt{2}}{2},\ 0\right) + (1,\ 0,\ 0,\ 0)}{\|\left(0,\ \frac{\sqrt{2}}{2},\ \frac{\sqrt{2}}{2},\ 0\right) + (1,\ 0,\ 0,\ 0)\|_d} = \frac{\left(1,\ \frac{\sqrt{2}}{2},\ \frac{\sqrt{2}}{2},\ 0\right)}{\sqrt{2}} = \left(\frac{\sqrt{2}}{2},\ \frac{1}{2},\ \frac{1}{2},\ 0\right) \tag{18}$$

Thus, CAU corresponds to the following unit vector:

$$CAU = \left(\frac{\sqrt{2}}{2},\ \frac{1}{2},\ \frac{1}{2},\ 0\right). \tag{19}$$

A schematic geometrical image of CAU is given in Figure 4.

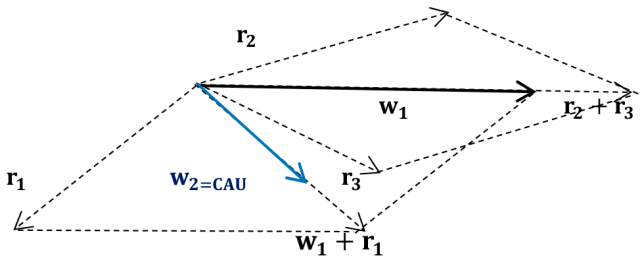


Figure 4. Schematic geometric representation of CAU (in blue) as a vector in the frame of VF-I representation.

In Figure 4, we firstly consider the sum $\boldsymbol{r_2} + \boldsymbol{r_3}$, *getting the corresponding parallelogram and then the vector* $\boldsymbol{w_1}$ *will be the unit part of the diagonal of this parallelogram. Next, we consider the sum* $\boldsymbol{w_1} + \boldsymbol{r_1}$, *getting the corresponding parallelogram and then the final vector* $\boldsymbol{w_2} =$ ***CAU (in blue)*** *will be the unit part of the diagonal of this parallelogram.*

(2) *For CCG (proline) we have C = $\boldsymbol{r_2}$, C = $\boldsymbol{r_2}$ and G = $\boldsymbol{r_4}$. Thus, the corresponding vectors of VF-I representation are as follows:*

Step 1. *We add the first two vectors corresponding to C and C and construct the corresponding unit vector, as follows:*

$$\boldsymbol{w_1} = \frac{\boldsymbol{r_2} + \boldsymbol{r_2}}{\|\boldsymbol{r_2} + \boldsymbol{r_2}\|_d} = \frac{(0,\ 1,\ 0,\ 0) + (0,\ 1,\ 0,\ 0)}{\|(0,\ 1,\ 0,\ 0) + (0,\ 1,\ 0,\ 0)\|_d} = \frac{(0,\ 2,\ 0,\ 0)}{2} = (0,\ 1,\ 0,\ 0) = \boldsymbol{r_2} \quad (20)$$

Step 2. *We add to the resulting unit vector $\boldsymbol{w_1}$ the unit vector corresponding to G, as follows:*

$$\boldsymbol{w_2} = \frac{\boldsymbol{w_1} + \boldsymbol{r_4}}{\|\boldsymbol{w_1} + \boldsymbol{r_1}\|_d} = \frac{(0\ ,1,\ 0,\ 0) + (0\ ,0,\ 0,\ 1)}{\|(0,\ 1,\ 0,\ 0) + (0,\ 0,\ 0,\ 1)\|_d} = \frac{(0,\ 1,\ 0,\ 1)}{\sqrt{2}} = \left(0,\ \frac{\sqrt{2}}{2},\ 0,\ \frac{\sqrt{2}}{2}\right) \quad (21)$$

Thus, CCG corresponds to the following unit vector:

$$CCG = \left(0,\ \frac{\sqrt{2}}{2},\ 0,\ \frac{\sqrt{2}}{2}\right). \quad (22)$$

A schematic geometrical image of CCG is given in Figure 5.

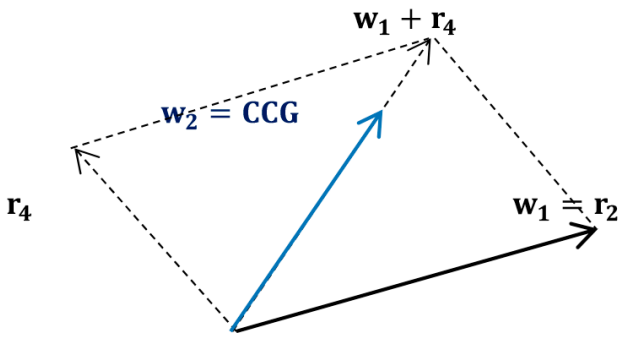


Figure 5. Schematic geometric representation of CCG (in blue) as a vector in the frame of VF-I representation. $\mathbf{w}_1 + \mathbf{r}_4$.

In Figure 5, we firstly consider the sum $\boldsymbol{r_2} + \boldsymbol{r_2}$ and then the vector $\boldsymbol{w_1}$ will be the unit part of this vector, equal to $\boldsymbol{r_2}$. Afterwards, we consider the sum $\boldsymbol{w_1} + \boldsymbol{r_4}$ getting the corresponding parallelogram and then the final vector $\boldsymbol{w_2} = \boldsymbol{CCG}$ (in blue) will be the unit part of the diagonal of this parallelogram.

Regarding the cosine of the angle θ of the above vector representations of

$$CAU = \left(\frac{\sqrt{2}}{2},\ \frac{1}{2},\ \frac{1}{2},\ 0\right) \text{ and } CCG = \left(0,\ \frac{\sqrt{2}}{2},\ 0,\ \frac{\sqrt{2}}{2}\right) \quad (23)$$

applying Definition 3, we have $cos\theta = \frac{\sqrt{2}}{4}$. Thus, $\theta \approx 69°$.

(3) *Similarly, in Table 3 we can see the VF-I representations of the genetic sequences s_1, s_2, s_3, s_4, s_5 and s_6 (approximately, two decimal digits), the genetic codes of which are given in Table 2.*

(4) *In Table 4 we can see the VF-I representations of the genetic sequences with three triplets (approximately, two decimal digits).*

Table 3. VF-I representation of genetic sequences with 2 amino acids.

Genetic Sequence	Symbol	Genetic Code	VF-I Representation
tyrosine/cysteine	s_1	UAC-UGU	(0.90, 0.16, 0.12, 0.39)
histidine/cysteine	s_2	CAC-UGU	(0.87, 0.27, 0.11, 0.41)
leucine/cysteine	s_3	CUC-UGU	(0.89, 0.22, 0.00, 0.40)
histidine/cysteine	s_4	CAU-UGU	(0.91, 0.10, 0.10, 0.39)
glutamine/cysteine	s_5	CAG-UGU	(0.84, 0.12, 0.12, 0.52)
glutamine/cysteine	s_6	CAA-UGU	(0.87, 0.11, 0.27, 0.41)

Table 4. VF-I representation of genetic sequences with 3 amino acids.

Genetic Sequence	Symbol	Genetic Code	VF-I Representation
leucine/asparagine/serine	Σ_1	UUA-AAU-UCU	(0.91, 0.39, 0.13, 0.00)
alanine/glutamic acid/phenylalanine	Σ_2	GCU-GAG-UUU	(0.98, 0.02, 0.07, 0.16)
valine/tyrosine/arginine	Σ_3	GUG-UAU-AGG	(0.19, 0.00, 0.30, 0.94)

4. Vector-Fuzzy-II Representation of Genetic Sequences

In this section, we present the second model of vector representation of the biological data of amino acids and genetic sequences.

Let ($\mathbb{R}^4$,+,) be the Euclidean vector space and the following genetic sequence

$$s = \mathbf{u_1u_2u_3} - \mathbf{u_4u_5u_6} - \mathbf{u_7u_8u_9} - \cdots - \mathbf{u_{k-5}u_{k-4}u_{k-3}} - \mathbf{u_{k-2}u_{k-1}u_k}, \tag{24}$$

which consists of k nucleotides and is in $\frac{k}{3}$-triplets (as in Section 3), where k is a natural number divided with 3 and the elements of the set $\{\boldsymbol{u_1}, \boldsymbol{u_2}, \cdots, \boldsymbol{u_k}\}$ are elements of the set {U, C, A, G}. For the new model, we consider the following meanings.

Definition 4. *Let s be the above genetic sequence (24) and N be any of the nucleotides U, C, A, G. The frequency of N, denoted by F(N), is defined to be a natural number of the set {0, 1, 2, ⋯ } which refers to how many times the nucleotide N appears in the sequence s.*

As each genetic sequence consists of triplets of nucleotides, it is necessary to consider the location of each nucleotide U, C, A, G in each triplet. This will give us the uniqueness of the representation of s. Thus, each nucleotide U, C, A, G is located at the 1st-position, 2nd-position or 3rd-position in a triplet, getting the following meanings of partial positional weight and total positional weight.

Definition 5. *Let s be a genetic sequence as in (24), $j = 1, 2, \cdots, \frac{k}{3}$, be any triplet of it, and N be any of the nucleotides U, C, A, G.*

(1) The partial positional weight of N, denoted by $S_i^j(N)$ is defined to be a number of the set {1, 2, 3} which refers to the position i of the nucleotide N in the triple j of s. That is,

$$S_i^j(N) = \begin{cases} i, & \text{if N is located at the } i-\text{position in the } j-\text{triplet} \\ 0, & \text{otherwise} \end{cases} \tag{25}$$

(2) The total positional weight of N, denoted by Σ(N), is defined to be the sum of all $S_i^j(N)$ in the sequence s. That is,

$$\Sigma(N) = \sum_{J=1}^{k/3} \sum_{i=1}^{3} S_i^j(N) \tag{26}$$

Clearly, F(N) = 0 if and only if $\Sigma(N) = 0$.

Definition 6. *Let s be a genetic sequence consisting as in (24) and N be any of the nucleotides U, C, A, G. The positional frequency of N, denoted by PF(N), is defined to be the natural number as follows:*

$$\mathrm{PF(N)} = \begin{cases} 0, & \text{if } \mathrm{F(N)} = 0, \\ \frac{\mathrm{F(N)}}{\Sigma(\mathrm{N})}, & \text{if } \mathrm{F(N)} \neq 0. \end{cases} \tag{27}$$

Example 2. *We shall find the positional frequencies of U, C, A, G in the following nucleotides:*

(1) For UCG (serine) we have the following:

$$PF(U) = \frac{F(U)}{\Sigma(U)} = \frac{1}{1} = 1,\ PF(C) = \frac{F(C)}{\Sigma(C)} = \frac{1}{2},\ PF(A) = \frac{F(A)}{\Sigma(A)} = 0,\ PF(G) = \frac{F(G)}{\Sigma(G)} = \frac{1}{3}. \tag{28}$$

(2) For CGU (arginine) we have the following:

$$PF(U) = \frac{F(U)}{\Sigma(U)} = \frac{1}{3},\ PF(C) = \frac{F(C)}{\Sigma(C)} = \frac{1}{1} = 1,\ PF(A) = \frac{F(A)}{\Sigma(A)} = 0,\ PF(G) = \frac{F(G)}{\Sigma(G)} = \frac{1}{2}. \tag{29}$$

(3) For UCG-CGU we have the following:

$$PF(U) = \tfrac{F(U)}{\Sigma(U)} = \tfrac{2}{1+3} = \tfrac{1}{2},\ PF(C) = \tfrac{F(C)}{\Sigma(C)} = \tfrac{2}{2+1} = \tfrac{2}{3},\ PF(A) = \tfrac{F(A)}{\Sigma(A)} = 0,\ PF(G) = \tfrac{F(G)}{\Sigma(G)} = \tfrac{2}{3+2} = \tfrac{2}{5}. \tag{30}$$

New Representation

[called *Vector-Fuzzy-II (in short VF-II) representation of s*]

Step 1. Find the vector:

$$\mathbf{w} = \frac{(\mathrm{PF(U)}, \mathrm{PF(C)}, \mathrm{PF(A)}, \mathrm{PF(G)})}{\|(\mathrm{PF(U)}, \mathrm{PF(C)}, \mathrm{PF(A)}, \mathrm{PF(G)})\|_{\mathrm{d}}} \tag{31}$$

and go to Step 2.

Step 2. Assign the genetic sequence s of (24) to the vector $\mathbf{w}$.

Clearly, the vector $\mathbf{w}$ is an element of the hypercube I^4 and the location of the nucleotides U, C, A, G ensures the uniqueness of each representation, getting the following remark.

Remark 2. *The VF-II representation of genetic sequences satisfies the following:*

(1) The vector $\boldsymbol{w}$ *of the VF-II representation of a genetic sequence s is unique.*

(2) If s_1 *and* s_2 *are two different genetic sequences and* $\boldsymbol{w_1}$, $\boldsymbol{w_2}$ *are their VF-II representations, respectively, then* $\boldsymbol{w_1} \neq \boldsymbol{w_2}$.

In the following examples we shall apply the VF-II representation for amino acids and genetic sequences (using their genetic code and Chou's pseudo-amino acid composition) in order to succeed their vector image in the Euclidean space $\mathbb{R}^4$.

Example 3. *We shall find the VF-II representation of some known nucleotides as follows.*

(1) For UGU (cysteine) we have $PF(U) = \frac{2}{1+3} = \frac{1}{2}$, *as U is present 2 times at positions 1 and 3,* $PF(C) = PF(A) = 0$ *and* $PF(G) = \frac{1}{2}$, *as G is present only once at position 2. Thus, the corresponding vector of VF-II is as follows:*

$$\boldsymbol{w} = \frac{(PF(U),\ PF(C),\ PF(A),\ PF(G))}{\|(PF(U),\ PF(C),\ PF(A),\ PF(G))\|_d} = \frac{\left(\frac{1}{2},\ 0,\ 0,\ \frac{1}{2}\right)}{\|\left(\frac{1}{2},\ 0,\ 0,\ \frac{1}{2}\right)\|_d} = \frac{2}{\sqrt{2}}\left(\frac{1}{2},\ 0,\ 0,\ \frac{1}{2}\right) = \left(\frac{\sqrt{2}}{2},\ 0,\ 0,\ \frac{\sqrt{2}}{2}\right) \tag{32}$$

Thus, UGU corresponds to the following unit vector:

$$UGU = \left(\frac{\sqrt{2}}{2},\ 0,\ 0,\ \frac{\sqrt{2}}{2}\right). \tag{33}$$

(2) For AUG (methionine) we have $PF(U) = \frac{1}{2}$, *as U is present only once at position 2,* $PF(C) = 0$, $PF(A) = \frac{1}{1} = 1$, *as A is located once at position 1, and* $PF(G) = \frac{1}{3}$, *as G is present only once at position 3. Thus, the corresponding vector of VF-II is as follows:*

$$\boldsymbol{w} = \frac{(PF(U),\ PF(C),\ PF(A),\ PF(G))}{\|(PF(U),\ PF(C),\ PF(A),\ PF(G))\|_d} = \frac{\left(\frac{1}{2},\ 0,\ 1,\ \frac{1}{3}\right)}{\|\left(\frac{1}{2},\ 0,\ 1,\ \frac{1}{3}\right)\|_d} = \frac{6}{7}\left(\frac{1}{2},\ 0,\ 1,\ \frac{1}{3}\right) = \left(\frac{3}{7},\ 0,\ \frac{6}{7},\ \frac{2}{7}\right) \tag{34}$$

Thus, AUG corresponds to the following unit vector:

$$AUG = \left(\frac{3}{7},\ 0,\ \frac{6}{7},\ \frac{2}{7}\right). \tag{35}$$

Additionally, regarding the cosine of the angle θ of the above vector representations of

$$UGU = \left(\frac{\sqrt{2}}{2},\ 0,\ 0,\ \frac{\sqrt{2}}{2}\right) \text{ and } AUG = \left(\frac{3}{7},\ 0,\ \frac{6}{7},\ \frac{2}{7}\right) \tag{36}$$

and applying Definition 3, we have $\cos\theta = \frac{5\sqrt{2}}{14}$. *Thus,* $\theta \approx 60^\circ$.

(3) Similarly, in Tables 5 and 6 we can see the VF-II representations of the genetic sequences s_1, s_2, s_3, s_4, s_5, s_6; Σ_1, Σ_2; *and* Σ_3 *(approximately, two decimal digits), the genetic codes of which are given in Tables 2–4.*

Table 5. VF-II representation of genetic sequences with 2 amino acids.

Genetic Sequence	Symbol	Genetic Code	VF-II Representation
tyrosine/cysteine	s_1	UAC-UGU	(0.61, 0.34, 0.51, 0.51)
histidine/cysteine	s_2	CAC-UGU	(0.50, 0.50, 0.50, 0.50)
leucine/cysteine	s_3	CUC-UGU	(0.57, 0.57, 0.00, 0.57)
histidine/cysteine	s_4	CAU-UGU	(0.33, 0.77, 0.38, 0.38)
glutamine/cysteine	s_5	CAG-UGU	(0.39, 0.78, 0.39, 0.31)
glutamine/cysteine	s_6	CAA-UGU	(0.39, 0.78, 0.31, 0.39)

Here, it is interesting to see the cosine similarity of the sequences s_5 *and* s_6, *as we can see a similar representation in Table 5. Then, applying Definition 3 we get* $\cos\theta \approx 1$. *Hence, the corresponding vectors of* s_5 *and* s_6 *tend to be nearly parallel and present very large similarity with an angle close to* $\theta = 0^\circ$ *between them.*

Table 6. VF-II representation of genetic sequences with 3 amino acids.

Genetic Sequence	Symbol	Genetic Code	VF-II Representation
leucine/asparagine/serine	Σ_1	UUA-AAU-UCU	(0.58, 0.58, 0.58, 0.00)
alanine/glutamic acid/phenylalanine	Σ_2	GCU-GAG-UUU	(0.43, 0.49, 0.49, 0.58)
valine/tyrosine/arginine	Σ_3	GUG-UAU-AGG	(0.53, 0.00, 0.71, 0.46)

5. Characteristics of Vector-Fuzzy Representation

The advantage and, simultaneously, the essential characteristic of the new vector-fuzzy models are based on the fact that we can represent the biological data of amino acids and bigger genetic sequences as vectors in $\mathbb{R}^4$, giving a pure mathematical approach. Additionally, the new models provide a way to investigate "measurements" between genetic sequences. The distance $d(s_i, s_j)$ between two genetic sequences s_i and s_j is determined by the corresponding Euclidean distance $d(\mathbf{w_i}, \mathbf{w_j})$ of the unit vectors $\mathbf{w_i}$ and $\mathbf{w_j}$ that are assigned to s_i and s_{jj}, respectively, and the similarity $sim(s_i, s_j)$ (resp. $dif(s_i, s_j)$) is determined by the corresponding similarity $sim(\mathbf{w_i}, \mathbf{w_j})$ (resp. $dif(\mathbf{w_i}, \mathbf{w_j})$), where $\mathbf{w_i}$ and $\mathbf{w_j}$ are given in the Vector-Fuzzy I and Vector-Fuzzy II representations.

In Sections 3 and 4 we have seen the new models of vector fuzzy representations (VF-I and VF-II) for the genetic sequences s_1, s_2, s_3, s_4, s_5 and s_6. According to the known FPS representation of these sequences in I^{12}, we can compare these two models via the characteristics of Euclidean distance, similarity and difference.

Application 1: Regarding s_1 in combination with s_2 and s_3

With respect to the Euclidean distance and similarity, in the FPS representation we can see that s_1 is closer to s_2 than s_3. This fact remains true in the new representations, which is absolutely expected as, between s_1 and s_3, there are more differences (in two bases) and between s_1 and s_2 there is only one difference (see Tables 7 and 8). In the new models, we have more near distances and larger degrees of similarities between the investigated sequences. We also note that the distance between s_1 and s_3 remains bigger than the distance between s_1 and s_2, and the degree of similarity between s_1 and s_3 remains smaller than the degree of similarity between s_1 and s_2. This fact can be verified by the composition of the sequences, as s_1 and s_3 have a difference of two bases.

Table 7. Comparing s_1, s_2 and s_3 via the Euclidean distance.

Euclidean Distance d			
	FPS Representation Table 1	**VF-I Representation Table 2**	**VF-II Representation Table 4**
$d(s_1, s_2)$	0.7071068	0.1161895	0.1946792
$d(s_1, s_3)$	1	0.1349074	0.5640922

Table 8. Comparing s_1, s_2 and s_3 via the similarity/difference.

Similarity/Difference			
	FPS Representation Table 1	**VF-I Representation Table 2**	**VF-II Representation Table 4**
$sim(s_1, s_2)$	0.8333333	0.9473684	0.9269521
$dif(s_1, s_2)$	0.1666667	0.0526316	0.0730479
$sim(s_1, s_3)$	0.6666667	0.9350649	0.7717391
$dif(s_1, s_3)$	0.3333333	0.0649351	0.2282609

Application 2: Regarding s_2 in combination with s_3, s_4, s_5 and s_6

In the following we calculate the distances and similarities of sequences $\mathbf{s_1}$, $\mathbf{s_3}$, $\mathbf{s_4}$, $\mathbf{s_5}$ and $\mathbf{s_6}$ from $\mathbf{s_2}$. All these sequences have the characteristic in that they differ from $\mathbf{s_2}$ at one base: for example, $\mathbf{s_1}$ and $\mathbf{s_2}$ have the same structure except for the first base of the first triplet where $\mathbf{s_1}$ has a U while $\mathbf{s_2}$ has a C. In the FPS representation, all distances and similarities are equal to each other. Thus, we cannot derive satisfactory results about how different the sequences are. However, we can observe that in vector fuzzy representations the distances differ, giving a satisfactory quantitative criterion for how big the difference is between two polynucleotides (see Tables 9 and 10). Additionally, the fact that, in VF-II, representations of $\mathbf{s_5}$ and $\mathbf{s_6}$ tend to be in parallel (see Section 4) leads to the coincidence of their distances from $\mathbf{s_2}$ as follows:

$$\mathbf{d}(\mathbf{s_2}, \mathbf{s_5}) = \mathbf{d}(\mathbf{s_2}, \mathbf{s_6}) \tag{37}$$

and also, their similarities from $\mathbf{s_2}$, as follows:

$$\mathbf{sim}(\mathbf{s_2}, \mathbf{s_5}) = \mathbf{sim}(\mathbf{s_2}, \mathbf{s_6}). \tag{38}$$

Table 9. Comparing s_2 with s_3, s_4, s_5 and s_6 via the Euclidean distance.

Euclidean Distance d			
	FPS Representation Table 1	**VF-I Representation Table 2**	**VF-II Representation Table 4**
$d(s_1, s_2)$	0.7071068	0.1161895	0.1946792
$d(s_2, s_3)$	0.7071068	0.1228821	0.5144900
$d(s_2, s_4)$	0.7071068	0.1760682	0.3613862
$d(s_2, s_5)$	0.7071068	0.1886796	0.3724245
$d(s_2, s_6)$	0.7071068	0.2262741	0.3724245

Table 10. Comparing s_2 with s_3, s_4, s_5 and s_6 via the similarity/difference.

Similarity/Difference			
	FPS Representation Table 1	**VF-I Representation Table 2**	**VF-II Representation Table 4**
$sim(s_2, s_3)$	0.8333333	0.9400631	0.8086253
$dif(s_2, s_3)$	0.1666667	0.0599369	0.1913747
$sim(s_2, s_4)$	0.8333333	0.9240506	0.8238342
$dif(s_2, s_4)$	0.1666667	0.0759494	0.1761658
$sim(s_2, s_5)$	0.8333333	0.9079755	0.8217054
$dif(s_2, s_5)$	0.1666667	0.0920245	0.1782946
$sim(s_2, s_6)$	0.8333333	0.9036145	0.8217054
$dif(s_2, s_6)$	0.1666667	0.0963855	0.1782946

Consequently, it is believed that the above results are due to the fact that the new methodologies take into account the local structure of the sequences when compared with the approach of FPS, as, in the case of VF-I, the unit vectors are constructed based on the successive addition of vectors corresponding to each nucleotide, and, in the case of VF-II, the frequencies are normalized with the locations of the nucleotides.

6. Discussion

The main tasks in bioinformatics are based on comparison, classification, and phylogenetic analyses of genetic sequences. Their vector representations ensure the extraction of useful and effective conclusions. However, the main problems of such works are the high computational complexity, mainly on high dimensional spaces, and the variety of definitions and approaches, which are developed independently, leading to possible indistinguishable results across related studies. In the present study, we focus on these problems, significantly reducing the dimension of the space where our data can be studied and by using the most known and widely used meanings of the fuzzy sets and vectors, avoiding any extremely new definition that may lead to new confusion.

More precisely, in this paper, we have presented new approaches to the representation of amino acids and genetic sequences, modeling amino acids and genetic sequences in a way that can achieve dimensionality reduction and better differentiation of the sequences. The new models are based on vector algebra and fuzzy set theory and provide representations of genetic sequences as unit vectors in the 4-dimensional space $\mathbb{R}^4$. The location of the amino acids plays an important role in each model as it verifies that each vector is assigned to a unique amino and vice versa. We applied the new models to several known amino acids and genetic sequences, providing examples of their representations in $\mathbb{R}^4$. The new models take into account the local structure of the sequences when compared with the approach of FPS, as, in the case of VF-I, the unit vectors are constructed based on the successive addition of vectors corresponding to each nucleotide, and, in the case of VF-II, the frequencies are normalized with the locations of the nucleotides.

7. Conclusions

Two new models for the vector representations of amino acids and genetic sequences in the space $\mathbb{R}^4$ are given, investigated and applied for several such genetic sequences. Each amino acid and genetic sequence gets a new image representation via vectors in a new mathematical space: the Euclidean space $\mathbb{R}^4$. Compared with the conventional fuzzy polynucleotide space $[0,\ 1]^{12}$ (as a subspace of $\mathbb{R}^{12}$), the proposed models achieve dimensionality reduction while enhancing the resolution of sequence differentiation, thus providing promising tools for sequence analysis in theoretical biology.

Author Contributions: Methodology, F.S., D.G., T.K.; investigation, F.S., D.G., T.K.; writing—original draft preparation, F.S., D.G., T.K.; writing—review and editing, F.S., D.G., T.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Paun, G.; Rozenberg, G.; Saloma, A. *DNA Computing: New Computing Paradigms*; Springer: Berlin, Germany, 1998.
2. Percus, J. *Mathematics of Genome Analysis*; Cambridge University Press: Cambridge, UK, 2002.
3. Freeland, S.J.; Hurst, L.D. The genetic code is one in a million. *J. Mol. Evol.* **1998**, *47*, 238–248. [CrossRef]
4. Chou, K.C. Prediction of protein cellular attributes using pseudo amino acid composition. *Proteins-Struct. Funct. Genet.* **2001**, *43*, 246–255. [CrossRef] [PubMed]

5. Chen, C.; Tian, Y.X.; Zou, X.Y.; Cai, P.X.; Mo, J.Y. Using pseudo-amino acid composition and support vector machine to predict proteins tructural class. *J. Theor. Biol.* **2006**, *243*, 444–448. [CrossRef]
6. Chou, K.C.; Cai, Y.D. Predicting protein quaternary structure by pseudo amino acid composition. *Proteins-Struct. Funct. Genet.* **2003**, *53*, 282–289. [CrossRef]
7. Wang, S.; Liu, S. Protein sub-nuclear localization based on effective fusion representations and dimension reduction algorithm LDA. *Int. J. Mol. Sci.* **2015**, *16*, 30343–30361. [CrossRef]
8. Du, P.; Li, Y. Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. *BMC Bioinform.* **2006**, *7*, 518. [CrossRef]
9. Georgiou, D.N.; Karakasidis, T.E.; Nieto, J.J.; Torres, A. A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets. *J. Theor. Biol.* **2010**, *267*, 95–105. [CrossRef] [PubMed]
10. Lin, H.; Li, Q.Z. Using pseudo amino acid composition to predict protein structural class: Approached by incorporating 400 dipeptide components. *J. Comput. Chem.* **2007**, *28*, 1463–1466. [CrossRef]
11. Mondal, S.; Bhavna, R.; MohanBabu, R.; Ramakumar, S. Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. *J. Theor. Biol.* **2006**, *243*, 252–260. [CrossRef]
12. Mundra, P.; Kumar, M.; Kumar, K.K.; Jayaraman, V.K.; Kulkarni, B.D. Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognit. Lett.* **2007**, *28*, 1610–1615. [CrossRef]
13. Xiao, X.; Shao, S.H.; Huang, Z.D.; Chou, K.C. Using pseudo amino acid composition to predict protein structural classes: Approached with complexity measure factor. *J. Comput. Chem.* **2006**, *27*, 478–482. [CrossRef]
14. Zhou, X.B.; Chen, C.; Li, Z.C.; Zou, X.Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *J. Theor. Biol.* **2007**, *248*, 546–551. [CrossRef]
15. Sadegh-Zadeh, K. Fundamentals of clinical methodology: 3. Nosology. *Artif. Intell. Med.* **1999**, *17*, 87–108. [CrossRef]
16. Kosko, B. *Neural Networks and Fuzzy Systems*; Prentice-Hall: Englewood Cliffs, NJ, USA, 1992.
17. Torres, A.; Nieto, J.J. The fuzzy polynucleotide space: Basic properties. *Bioinformatics* **2003**, *19*, 587–592. [CrossRef]
18. Lin, C.T. Adaptive subsethood for radial basis fuzzy systems. In *Fuzzy Engineering*; Kosko, B., Ed.; Prentics-Hall: Upper Saddle River, NJ, USA, 1997; pp. 429–464.
19. Sadegh-Zadeh, K. Fuzzy genomes. *Artif. Intell. Med.* **2000**, *18*, 1–28. [CrossRef]
20. Nieto, J.J.; Torres, A. Midpoints for fuzzy sets and their application in medicine. *Artif. Intell. Med.* **2023**, *17*, 81–101. [CrossRef]
21. Chou, K.C.; Shen, H.B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.* **2007**, *370*, 1–16. [CrossRef]
22. Gusev, V.D.; Nemytikova, L.A.; Chuzhanova, N.A. On the complexity measures of genetic sequences. *Bioinformatics* **1999**, *15*, 994–999. [CrossRef]
23. Jiang, T.; Lin, G.; Ma, B.; Zhang, K. A general edit distance between RNA structures. *J. Comput. Biol.* **2002**, *9*, 371–388. [CrossRef]
24. Li, C.; Dai, Q.; He, P.A. A time series representation of protein sequences for similarity comparison. *J. Theor. Biol.* **2022**, *538*, 111039. [CrossRef]
25. Koo, P.K.; Eddy, S.R. Representation learning of genomic sequence motifs with convolutional neural networks. *PLoS Comput. Biol.* **2019**, *15*, e1007560. [CrossRef]
26. Dennler, O.; Ryan, C.J. Evaluating sequence and structural similarity metrics for predicting shared paralog functions. *NAR Genom. Bioinform.* **2025**, *7*, lqaf051. [CrossRef]
27. Kösoglu-Kind, B.; Loredo, R.; Grossi, M.; Bernecker, C.; Burks, J.M.; Buchkremer, R. A biological sequence comparison algorithm using quantum computers. *Sci. Rep.* **2023**, *13*, 14552. [CrossRef]
28. Lee, Y.C.; Jung, S.H.; Kumar, A.; Shim, I.; Song, M.; Kim, M.S.; Kim, K.; Myung, W.; Park, W.Y.; Won, H.H. ICD2Vec: Mathematical representation of diseases. *J. Biomed. Inform.* **2023**, *141*, 104361. [CrossRef]
29. Liu, Y.; Wu, R.; Yang, A. Research on Medical Problems Based on Mathematical Models. *Mathematics* **2023**, *11*, 2842. [CrossRef]
30. Zayed, A.I. A new perspective on the role of mathematics in medicine. *J. Adv. Res.* **2019**, *17*, 49–54. [CrossRef]
31. Kuruvilla, F.G.; Park, P.J.; Schreiber, S.L. Vector algebra in the analysis of genome-wide expression data. *Genome Biol.* **2002**, *3*, research0011.1–research0011.11. [CrossRef]
32. Liu, E.S.F.; Wu, V.W.C.; Harris, B.; Foote, M.; Lehman, M.; Chan, L.W.C. Vector-model-supported optimization in volumetric-modulated arc stereotactic radiotherapy planning for brain metastasis. *Med. Dosim.* **2017**, *42*, 85–89. [CrossRef]
33. Liang, Y.; Yang, S.; Zheng, L.; Wang, H.; Zhou, J.; Huang, S.; Yang, L.; Zuo, Y. Research progress of reduced amino acid alphabets in protein analysis and prediction. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 3503–3510. [CrossRef]
34. Zadeh, L.A. Fuzzy Sets. *Inf. Control* **1965**, *8*, 338–353. [CrossRef]
35. Klir, G.J.; Yuan, B. *Fuzzy Sets and Fuzzy Logic (Theory and Applications)*; Prentice Hall PRT: Hoboken, NJ, USA, 1995.
36. Terano, T.; Asai, K.; Sugeno, M. *Fuzzy Systems Theory and its Applications*; Academic Press, Harcount Brace Jovanovich Publishers: San Diego, CA, USA, 1992.
37. Zimmermann, H.J. *Fuzzy Theory and Its Applications*; Kluwer Academic Publishers: New York, NY, USA, 1991.

38. Stojanovic, N.; Lakovic, M. $\mathbb{Q}[\varepsilon]$-Fuzzy Sets. *J. Korean Soc. Ind. Appl. Math.* **2024**, *28*, 303–318.
39. Nieto, J.J.; Torres, A.; Georgiou, D.N.; Karakasidis, T.E. Fuzzy polynucleotide spaces and metrics. *Bull. Math. Biol.* **2006**, *68*, 703–725. [CrossRef] [PubMed]
40. Smith, L. *Linear Algebra*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 1984.
41. Engelking, R. *General Topology*; Heldermann Verlag: Berlin, Germany, 1989.
42. Karakasidis, T.E.; Georgiou, D.N. Partitioning elements of the periodic table via fuzzy clustering technique. *Soft Comput.* **2004**, *8*, 231–236.
43. Samaras, P.; Kungolos, A.; Karakasidis, T.; Georgiou, D.; Perakis, K. Statistical evaluation of PCDD/F emission data during solid waste combustion by fuzzy clustering techniques. *J. Environ. Sci. Health Part A* **2001**, *36*, 153–161. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.