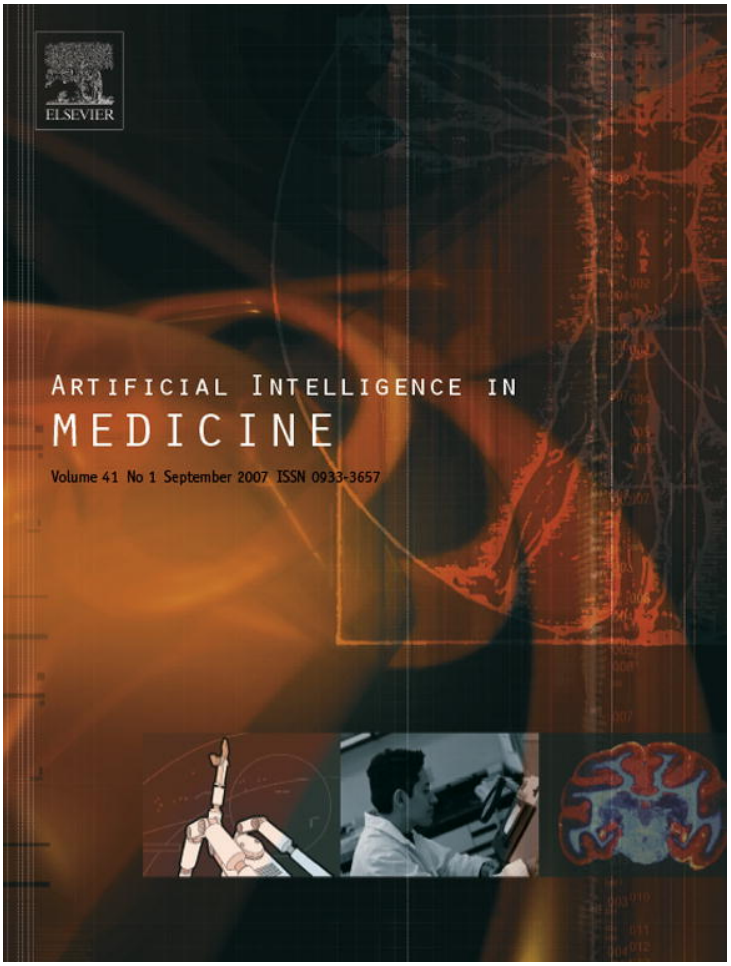

Artificial Intelligence in Medicine (2007) 41, 69—80

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

The fuzzy polynucleotide space revisited

Kazem Sadegh-Zadeh *

University of Münster Clinicum, Münster, Germany

Received 20 November 2006; received in revised form 2 April 2007; accepted 13 April 2007

KEYWORDS
Polynucleotides;
Fuzzy polynucleotides;
Fuzzy polynucleotide space;
Base profile

Summary

Objective: A theory of fuzzy polynucleotides, including an *n*-dimensional metric fuzzy polynucleotide space, has been previously introduced by the present author for fuzzy-theoretical analysis of nucleic acids [Sadegh-Zadeh K. Fuzzy genomes. Artif Intell Med 2000;18:1—28; Sadegh-Zadeh K. Ein Verfahren zur Fuzzydecodierung und Fuzzydechiffrierung von Informationen. *Offenlegungsschrift DE 199 36 925 A 1*. Deutsches Patent- und Markenamt; 2001]. The conceptual framework of that theory has been used by Nieto et al. [Nieto JJ, Torres A, Vázquez-Trasande MM. A metric space to study differences between polynucleotides. Appl Math Lett 2003;16:1289—94; Nieto JJ, Torres A, Georgiou DN, Karakasidis TE. Fuzzy polynucleotide spaces and metrics. Bull Math Biol 2006;68:703—25] and Torres et al. [Torres A, Nieto JJ. The fuzzy polynucleotide space: basic properties. Bioinformatics 2003;19:587—92; Torres A, Nieto JJ. Fuzzy logic in medicine and bioinformatics. J Biomed Biotechnol 2006;1—7 [Article ID 91908]] to create a completely different, 12-dimensional metric space which they have also called 'the fuzzy polynucleotide space'. In the present paper both spaces are compared.
Material and method: Both metric spaces are briefly outlined. Similarity and dissimilarity relationships between polynucleotide strings are measured in both spaces to compare their performance.
Results: Nieto et al.'s and Torres et al.'s metric space measures the relationships between polynucleotide chains incorrectly. Structurally highly different polynucleotide sequences are misclassified as highly similar ones, and completely different sequences are misclassified as identical ones. For this reason their construct is to be considered as a device of misdiagnosis that bears ''fuzzy polynucleotide space'' as a misnomer.

1. Introduction

Sequence analysis and sequence comparison are two basic methods in genetics, genomics, and other areas of inquiry to explore and compare the genetic

* Correspondence address: Am Steinkamp 20, D-49545 Tecklenburg, Germany. Tel.: +49 5482 1604.
E-mail address: ksz-2@medizintheorie.de.

0933-3657/$ — see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2007.04.006

material of different organisms and viruses. In a research program started in the 1990s the present author tried to render fuzzy theory directly accessible to sequence analysis and comparison. To this end he fuzzified the concept of *sequence* and used biopolymers, especially the nucleic acids DNA and RNA, as examples [1,2,7]. His basic step consisted in transforming polynucleotide chains, i.e. nucleic acids, to ordered fuzzy sets. He could in this way demonstrate that a polynucleotide molecule is representable as a point in an *n*-dimensional unit hypercube and constructed a framework for fuzzy-theoretical analysis of polynucleotides. This approach enabled quantitative studies such as the measurement of distances, similarities, and dissimilarities between, and an abstract geometry of, polynucleotide sequences. The *n*-dimensional unit hypercube enriched by a distance function *d*, i.e. $\langle [0, 1]^n, d\rangle$, that he suggested as a metric space for use in such inquiries he named the *fuzzy polynucleotide space*. In a series of publications since 2003 Nieto, Torres et al. [3–6] have fairly disfigured this metric space in that they have reduced it to a 12-dimensional one which they have also termed 'the fuzzy polynucleotide space'. As can be expected from this reduction, however, their construct is irremediably flawed. Here are two simple examples. First, the RNA sequences UACAGU and UGUUAC are diagnosed to be similar to the extent 0.833 and different to the extent 0.167, whereas quite the opposite is true because they have only one nitrogenous base in common at their initial sites. Second, UACAGU and AGUUAC are diagnosed to be identical sequences, whereas again the opposite is true because they have no single base in common at their corresponding sites. To analyze the fault in detail and to explain its genesis causally we first will recall our concept of fuzzy polynucleotides in Sections 2–3. In Sections 4–5 both metric spaces are briefly outlined so we may compare and evaluate them in Sections 6–7. Section 8 closes our comments on Nieto, Torres et al.'s system with a concise causal explanation of its incorrigible defectiveness.

2. Polynucleotides

A polynucleotide such as a DNA or RNA molecule is a linear polymer that consists of many smaller units called its building blocks or monomers (from Greek 'meros' = *part*). As is usual, we will here formally represent the monomers of a polynucleotide by their nitrogenous bases, and thus a polynucleotide itself by its base sequence such as GTTTACGAA. (for more details, see Appendix.) We may in this way conceive a polynucleotide as a sequence of letters, i.e. as a word, of length $m \geq 1$ over a particular alphabet of length $n \geq 1$. An alphabet is an ordered set of $n \geq 1$ signs, characters or letters, $\langle L_1, \ldots, L_n\rangle$, with *n* being its length. We distinguish between:

DNA alphabet $= \langle T, C, A, G\rangle$ and

RNA alphabet $= \langle U, C, A, G\rangle$

each of length 4. Their letters are the initials of the names of nitrogenous bases (Thymine, Cytosine, Adenine, Guanine, Uracil) contained in the five different monomers of polynucleotides (see Appendix). For example, the sequence GTTTACGAA is a word of length 9 over the DNA alphabet ⟨T, C, A, G⟩, while the sequence UGGAAC is a word of length 6 over the RNA alphabet ⟨U, C, A, G⟩. We will in this brief note use RNA words only. The terms 'word' and 'sequence' are used interchangeably. For detailed inquiries, see [1].

3. Fuzzy polynucleotides

A fuzzy polynucleotide is a polynucleotide sequence represented as a *fuzzy sequence*. To demonstrate, we first introduce the notion of a fuzzy sequence.

If Ω is a non-empty set of any objects, *A* is a *fuzzy set* in, or over, Ω iff there is a function μ_A such that μ_A: $\Omega \rightarrow [0, 1]$ and $A = \{(x, \mu_A(x))|x \in \Omega\}$. That is, A is the set of all pairs $(x, \mu_A(x))$ such that to each member *x* of Ω is attached a real number $\mu_A(x) \in [0, 1]$ indicating the degree of its membership in *A*. Set Ω is referred to as the universe of discourse or the ground set, and the function μ_A from Ω to unit interval [0, 1] constituting the fuzzy set *A* is the membership function of *A*. For example, if $\Omega = \{x, y, z\}$, then $A = \{(x, 0.5), (y, 1), (z, 0.2)\}$ is a fuzzy set over Ω. Another fuzzy set over Ω is $B = \{(x, 0.8), (y, 0), (z, 1)\}$, and so on. There are infinitely many fuzzy sets over a ground set Ω because the number of mappings from Ω to [0, 1] is infinite. This infinite set of all fuzzy sets over Ω is referred to as the fuzzy powerset of Ω and written $F(2^{\Omega})$.

A fuzzy sequence is simply an *ordered fuzzy set*, i.e. a fuzzy set over an ordered ground set $\Omega = \langle x_1, \ldots, x_n\rangle$, for example, $\langle (x_1, 0.8), (x_2, 1), \ldots, (x_n, 0.4)\rangle$. A polynucleotide is representable as such an ordered fuzzy set. To this end we consider the alphabet, over which it is a word, as the ground set Ω and fuzzify this ground set as above. We thereby obtain fuzzy letters of which the polynucleotide can be reconstructed as a fuzzy word. For example, let μ_U, μ_C, μ_A, and μ_G be four different functions

each of which maps the RNA alphabet ⟨U, C, A, G⟩ = Ω as our ground set to unit interval [0, 1]:

$$\mu_U : \langle U, C, A, G\rangle \to [0, 1]$$
$$\mu_C : \langle U, C, A, G\rangle \to [0, 1]$$
$$\mu_A : \langle U, C, A, G\rangle \to [0, 1]$$
$$\mu_G : \langle U, C, A, G\rangle \to [0, 1]$$

to yield the following four different fuzzy letters as ordered fuzzy sets:

$$\text{Letter U} = \langle (U, 1), (C, 0), (A, 0), (G, 0)\rangle \quad \text{read 'U'}$$
$$\text{Letter C} = \langle (U, 0), (C, 1), (A, 0), (G, 0)\rangle \quad \text{read 'C'}$$
$$\text{Letter A} = \langle (U, 0), (C, 0), (A, 1), (G, 0)\rangle \quad \text{read 'A'}$$
$$\text{Letter G} = \langle (U, 0), (C, 0), (A, 0), (G, 1)\rangle \quad \text{read 'G'}$$

By means of these fuzzy letters we may reconstruct, for example, the RNA sequence UGG as the ordered fuzzy set ⟨Letter U, Letter G, Letter G⟩, that is:

$$\langle (U, 1), (C, 0), (A, 0), (G, 0), (U, 0), (C, 0), (A, 0), (G, 1), (U, 0), (C, 0), (A, 0), (G, 1)\rangle.$$

This ordered fuzzy set is a *fuzzy polynucleotide*, i.e. our RNA triplet UGG in a fuzzified form. To simplify the representation of the four fuzzy letters above we may use only their membership degrees. Thus, we obtain the following four simplified fuzzy letters in vector notation:

$$\text{Letter U} = (1, 0, 0, 0),$$
$$\text{Letter C} = (0, 1, 0, 0),$$
$$\text{Letter A} = (0, 0, 1, 0),$$
$$\text{Letter G} = (0, 0, 0, 1).$$

By the use of this vector notation also our RNA sequence UGG above is simplified thus:
$(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1)$.

This 12-dimensional vector represents our example, i.e. the triplet codon UGG. Analogously, the following 24-dimensional vector is the polynucleotide UGGAAC consisting of two triplet codons, i.e. UGG and AAC:

$$(1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0).$$

These examples demonstrate that a polynucleotide of length $n \geq 1$ is a fuzzy polynucleotide of length $4n$. It has a vector of length $4n$ because its alphabet is of length 4. The vectors of our example fuzzy polynucleotides used until now had components only from the bivalent set {0, 1}. This is not necessary, however, for the membership function of a fuzzy set has the entire unit interval [0, 1] as its range. Thus, the set of all fuzzy sets over an alphabet such as ⟨U, C, A, G⟩, i.e. the fuzzy powerset $F(2^{\langle U,\ C,\ A,\ G\rangle})$, is infinite such that any element of $F(2^{\langle U,\ C,\ A,\ G\rangle})$ is a fuzzy letter, for example, the following sequence:

$\langle (U, 0.4), (C, 0.2), (A, 0), (G, 0.8)\rangle$,

that is:

$(0.4, 0.2, 0, 0.8)$.

There may exist circumstances, e.g. an experiment or a genetic examination, in which we are not certain whether a particular triplet such as UGX bears a U, a C, an A or a G at its third site that we have here marked with 'X'. In such cases we may suppose a degree of *possibility* to which any of these four bases may be present at that site to obtain, for example, the following vector:

$(1, 0, 0, 0, 0, 0, 0, 1, 0.4, 0.2, 0, 0.8)$.

This vector says that:

(1) the first base is a U to the extent 1, a C to the extent 0, an A to the extent 0, and a G to the extent 0,
(2) the second base is a U to the extent 0, a C to the extent 0, an A to the extent 0, and a G to the extent 1,
(3) the third base is a U to the extent 0.4, a C to the extent 0.2, an A to the extent 0, and a G to the extent 0.8.

In any case, a polynucleotide of length $n \geq 1$ is a fuzzy polynucleotide of length $4n$, and thus, representable by a real vector $(x_1, x_2, \ldots, x_{4n})$ of length $4n$ such that each component x_i of the vector is an element of [0, 1]. For details of this theory, see [1].

4. The fuzzy polynucleotide space suggested by the present author

As early as 1971 Lotfi A. Zadeh introduced a geometric interpretation of fuzzy sets by stating that they can be represented as points in unit hypercubes (see [8], p. 486). Many years later his idea was taken up by Bart Kosko, who built a promising fuzzy-theoretical framework and geometry thereon [9,10]. This geometry of fuzzy sets we have used in [1] to develop our fuzzy polynucleotide space. For example, the fuzzy nucleotide ⟨(U, 0.4), (C, 0.2), (A, 0), (G, 0.8)⟩ mentioned above is the 4-dimensional vector (0.4, 0.2, 0, 0.8). Thus, it is a point in the 4-dimensional unit hypercube $[0, 1]^4$. Since every triplet codon XYZ has a $3 \times 4 = 12$-dimensional vector, the genetic code comprising 64 single codons can be accommodated in $[0, 1]^{12}$. A polynucleotide of length 2500 is a point in the cube $[0, 1]^{10,000}$. In general, a polynucleotide of length n is a fuzzy polynucleotide of length $4n$, and thus a point in $[0, 1]^{4n}$.

By supplementing the cube $[0, 1]^n$ with a distance measure d over $[0, 1]^n$ we obtain an n-dimensional metric space $\langle [0, 1]^n, d\rangle$. In this space we can determine how distant from, and thus how different from or similar to, one another two polynucleotide sequences are. We have therefore called that space the *fuzzy polynucleotide space*, FPNS for short (see [1]). We now will briefly outline the FPNS to tackle the task of this paper in the sequel. To this end we need a distance function. A function called 'differ' that measures the difference between two polynucleotide sequences will serve as our distance function. To introduce it we need two auxiliary notions, 'the union of two fuzzy sets' and 'the count of a fuzzy set':

The *union* $A \cup B$ of two fuzzy sets A and B over a ground set Ω is defined by means of the membership function $\mu_{A \cup B}(x)$ thus: $A \cup B = \{(x, \mu_{A \cup B}(x)) | x \in \Omega$ and $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))\}$. For example, if $\{(x, 0.2), (y, 1)\}$ and $\{(x, 0.8), (y, 0.6)\}$ are two fuzzy sets over our ground set $\{x, y\}$, then we have $\{(x, 0.2), (y, 1)\} \cup \{(x, 0.8), (y, 0.6)\} = \{(x, 0.8), (y, 1)\}$. The size or *count* of a fuzzy set A, written $c(A)$, is simply the arithmetic sum of the membership degrees of its members. For instance, $c(\{(x, 0.2), (y, 1)\}) = 0.2 + 1 = 1.2$.[1]

Definition 1. If $A = \{(x_1, a_1), \ldots, (x_n, a_n)\}$ and $B = \{(x_1, b_1), \ldots, (x_n, b_n)\}$ are two fuzzy sets, then the difference between A and B, written differ(A, B), is:

$$\text{differ}(A, B) = \frac{\sum_i |a_i - b_i|}{c(A \cup B)}.$$

As an example we will compute the differences between the following four polynucleotide sequences s_1–s_4:

$s_1 \equiv$ UGGAAC	codes for :	tryptophan–asparagine
$s_2 \equiv$ UACUGG		tyrosine–tryptophan
$s_3 \equiv$ AACUGG		asparagine–tryptophan
$s_4 \equiv$ CUCUGG		leucine–tryptophan

Their fuzzy vectors are:

$s_1 \equiv (1,0,0,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,1,0,0,1,0,0)$,

$s_2 \equiv (1,0,0,0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,1,0,0,0,1)$,

$s_3 \equiv (0,0,1,0,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0,1,0,0,0,1)$,

$s_4 \equiv (0,1,0,0,1,0,0,0,0,1,0,0,1,0,0,0,0,0,0,1,0,0,0,1)$.

We first will compute the differences between some of these sequences:

$$\text{differ}(s_1, s_2) = \frac{10}{11} = 0.91, \qquad \text{differ}(s_1, s_3) = \frac{12}{12} = 1,$$

$$\text{differ}(s_2, s_3) = \frac{2}{7} = 0.285, \qquad \text{differ}(s_2, s_4) = \frac{4}{8} = 0.5,$$

$$\text{differ}(s_3, s_4) = \frac{4}{8} = 0.5.$$

Note that sequences s_1 and s_2 have only one single base in common at their initial sites; sequences s_2 and s_3 have only one base difference; sequences s_1 and s_3 have nothing in common at their corresponding sites; and sequences s_2 and s_4 as well as s_3 and s_4 have four bases in common.

The *fuzzy polynucleotide space* that we have originally introduced and discussed at length in [1] is just the metric space $\langle [0, 1]^n$, differ$\rangle$. In this metric space we may measure not only the difference, but also the degree of similarity and identity between polynucleotides in the following way. The functions similar(A, B) and equal(A, B) read, respectively, "the degree of similarity between fuzzy set A and fuzzy set B" and "the degree of equality between fuzzy set A and fuzzy set B".

Definition 2. similar(A, B) = 1 − differ(A, B).

For example, the degrees of similarity between some of the above-mentioned polynucleotide sequences are:

$\text{similar}(s_1, s_2) = 1 - 0.91 = 0.09$,

$\text{similar}(s_1, s_3) = 1 - 1 = 0$,

$\text{similar}(s_2, s_3) = 1 - 0.285 = 0.715$,

$\text{similar}(s_2, s_4) = 1 - 0.5 = 0.5$,

$\text{similar}(s_3, s_4) = 1 - 0.5 = 0.5$.

Definition 3.

(1) equal(A, B) = similar(A, B).
(2) A and B are *identical* iff equal(A, B) = 1.

Two examples may illustrate:

equal(s_1, s_2) = similar(s_1, s_2) = 0.09
equal(s_1, s_3) = similar(s_1, s_3) = 0.

Definitions 1–3 imply the following theorem that we will refer to in Section 5 below. Let A and B be any fuzzy sets, then:

Theorem 1.

1. A and B are *identical* iff similar(A, B) = 1.
2. A and B are *identical* iff differ(A, B) = 0.

[1] Definition 1 is another, logically equivalent formulation of the same concept of fuzzy set difference introduced by Lin [11] and Sadegh-Zadeh [1].

Fuzzy set identity is thus the maximum degree of similarity and the minimum degree of difference between two fuzzy sets. An example is provided by the following two polynucleotide sequences:

$$s_1 \equiv (1,0,0,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,1,0,0,1,0,0),$$
$$s_5 \equiv (1,0,0,0,0,0,0,1,0,0,0,1,0,0,1,0,0,0,1,0,0,1,0,0).$$

According to Theorem 1 these two sequences are identical because differ(s_1, s_5) = 0 and thus similar(s_1, s_5) = 1.

We may distinguish between fuzzy polynucleotides whose vectors have components only from {0, 1} like those used above, and those whose vectors have at least one component from [0, 1] such as (1, 0, 0, 0, 0, 0, 0, 1, 0.4, 0.2, 0, 0.8). The former ones we call *crisp*. The latter ones are genuinely fuzzy representing states of our uncertainty about their true chemical compounds. A crisp polynucleotide such as any of the 64 usual triplet codons of the genetic code is located at a *corner* of the respective hypercube $[0, 1]^n$, whereas the genuinely fuzzy ones are points on a cube side or within the cube. In any case, by supplementing the cube $[0, 1]^n$ with a distance measure over $[0, 1]^n$, e.g. differ, we obtain an n-dimensional metric space. The n-dimensional fuzzy polynucleotide space $\langle [0, 1]^n$, differ$\rangle$ sketched above will be conveniently denoted $\mathrm{FPNS}_{\text{Sadegh-Z}}$.

5. The fuzzy polynucleotide space suggested by Nieto et al. and Torres et al.

Nieto et al. [3,4] and Torres et al. [5] have utilized the conceptual framework upon which $\mathrm{FPNS}_{\text{Sadegh-Z}}$ is based to suggest an alternative, 12-dimensional 'fuzzy polynucleotide space' that we will refer to as $\mathrm{FPNS}_{\text{Nieto}-\text{Torres}}$. We now will briefly describe it to compare both spaces in the sequel.

As pointed out in the last section, the genetic code is 12-dimensional. A triplet codon of the form XYZ has a $3 \times 4 = 12$-dimensional vector of the form $(x_1, x_2, \ldots, x_{12})$ such that each $x_i \in \{0, 1\}$. Thus, any of the 64 codons of the genetic code is located at one of the $2^{12} = 4096$ corners of the unit hypercube $[0, 1]^{12}$ (see [1], pp. 12 f.). The n-dimensional space $\mathrm{FPNS}_{\text{Sadegh-Z}}$ includes the restricted space $\langle [0, 1]^{12}$, differ$\rangle$ as a subspace. It is obvious that in this latter, small space only triplet codons can be dealt with. Polynucleotides of length $n > 3$ require higher-dimensional cubes.

However, Nieto et al. and Torres et al. seem to have been fascinated by the elegant, 12-dimensional space $\langle [0, 1]^{12}$, differ$\rangle$ of the *genetic code* to the effect that they have chosen just this confined space to accommodate therein all polynucleotides of arbitrary length. How is it possible to measure, for example, the difference between two polynucleotides of length 10,000 in a 12-dimensional space, even though they are not points in this space? Judged from a formal and mathematical point of view, Nieto et al. and Torres et al. have had an ingenious idea to solve this problem and to suggest their $\mathrm{FPNS}_{\text{Nieto}-\text{Torres}}$ [4,5]. Unfortunately, however, their idea has fatal consequences that render $\mathrm{FPNS}_{\text{Nieto}-\text{Torres}}$ useless. This claim will be justified below. To this end we introduce the auxiliary notion of the 'base profile of a polynucleotide'. It is the pivotal concept of their system and the device by which they fuzzify a crisp, i.e. non-fuzzy, polynucleotide of arbitrary length. We must clearly understand it because it is the cause of the trouble. It is in fact a position specific relative frequency matrix that reflects the relative frequencies of the 4 nitrogenous bases at each position $i \in \{1, 2, 3\}$ of the triplets of a sequence ($4 \times 3 = 12$ dimensions). We will explain it in three steps: first, by the term 'local *absolute* base frequency' we understand the absolute frequency of a nitrogenous base at site $i \in \{1, 2, 3\}$ of all triplets of a polynucleotide. For example, base U appears 2 times at site 1 of the three triplets of the polynucleotide sequence UGGAACUCU. Thus, the *local absolute frequency* of U at site 1 of the triplets of that molecule is 2.

Second, by the term 'local *relative* base frequency' we understand the relative frequency of a nitrogenous base at site $i \in \{1, 2, 3\}$ of all triplets of a polynucleotide sequence, i.e. its local absolute frequency at site $i \in \{1, 2, 3\}$ divided by the total frequency of all four bases at this site $i \in \{1, 2, 3\}$ in the whole molecule, i.e. divided by the number of triplets constituting the molecule. For example, in the above molecule UGGAACUCU the total frequency of all four bases at site 1 of the triplets is 3. Thus, the *local relative frequency* of U at site 1 of the molecule's triplets is $2/3 = 0.66$.

Third, the ordered set of local relative frequencies of all four bases U, C, A, and G at all three sites 1, 2, and 3 of triplets of a polynucleotide molecule we call the *base profile* of the molecule. Examples will be given in the tables below. Note that the base profile of a polynucleotide yields a 12-dimensional

real vector $(x_1, \ldots, x_{12})$ of the following form with each $x_i \in [0, 1]$ because $4 \times 3 = 12$:

$$(\text{U at site 1},\ \text{C at site 1},\ \text{A at site 1},\ \text{G at site 1},\ \text{U at site 2},\ \text{C at site 2},\ \text{A at site 2},\ \text{G at site 2},\ \text{U at site 3},\ \text{C at site 3},\ \text{A at site 3},\ \text{G at site 3}). \quad (1)$$

We will partition the class of polynucleotides, say RNA, into the category of *single* triplets XYZ of which there are 64 different types, and the category of *compound* sequences. The latter ones are concatenations of $n > 1$ single triplets such as UGGAACUCU and larger molecules. The ingenious idea of Nieto et al. and Torres et al. referred to above may be described as follows:

To treat a polynucleotide of arbitrary length as a point in the small space $[0, 1]^{12}$, determine the *base profile* of the polynucleotide molecule! That is, count the local frequency of each one of the four nitrogenous bases U, C, A, and G separately at each site $i \in \{1, 2, 3\}$ of the triplets XYZ of the polynucleotide, be it a single triplet or a compound polynucleotide. Because $4 \times 3 = 12$ you will obtain 12 natural numbers indicating local absolute frequencies of the four nitrogenous bases. Divide each of these single-base specific 12 numbers by the number of the whole molecule's triplets. You will get 12 local, i.e. position specific, relative frequencies ordered as a real vector of the form $(x_1, x_2, \ldots, x_{12})$ with $x_i \in [0, 1]$ like (1) above. This vector represents the base profile searched for. Thus, you can transform any polynucleotide of arbitrary length to a point in $[0, 1]^{12}$ enabling fuzzy-theoretical analyses in a 12-dimensional space. Nieto et al. and Torres et al. have allegedly been able even to compare whole genomes of different organisms with one another, e.g. the genome of Mycobacterium tuberculosis with that of *Escherichia coli* and others (see [4–6]). To explain why their efforts are all in vain nonetheless, we must go into detail. The procedure may be demonstrated by a simple example. Let us compare the two sequences s_1 and s_2, i.e. UGGAAC and UACUGG, which we have also used in $\text{FPNS}_{\text{Sadegh-Z}}$ in the last section. To prevent confusion between both systems, in Nieto et al.'s system we will signify a sequence by "seq" instead of "s":

$$\text{seq}_1 \equiv \text{UGGAAC}, \qquad \text{seq}_2 \equiv \text{UACUGG}.$$

They are compound sequences of length 6 each and have 24-dimensional vectors that cannot be dealt with in $[0, 1]^{12}$. To do so, their base profiles are determined according to the algorithm introduced above (see Table 1).

From the relative frequencies on the right-hand side of Table 1 we obtain the base profiles of both sequences in the form of the following 12-dimensional membership vectors:

$$\text{seq}_1 \equiv (0.5, 0, 0.5, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5),$$
$$\text{seq}_2 \equiv (1, 0, 0, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5).$$

Note that in the same fashion every polynucleotide of arbitrary length may be transformed to a 12-dimensional vector. It thereby becomes a point in $[0, 1]^{12}$. What is now needed to obtain a metric space is a distance function. Nieto et al. and Torres et al. use and analyze a host of distance functions. But they favor the same distance function *differ* that we have introduced in Definition 1 above, on the one hand, and call it 'd' (cf. [3,5]); and another distance function that they call *dif*, on the other (see [4], p. 716). We will use *dif* that is their most recent device. It is defined indirectly by their similarity function, *sim*, that we must introduce first (see [4], p. 714). An auxiliary notion we need is '*the canonical midpoint*' of two fuzzy sets introduced by Nieto and Torres in ([12], p. 85).

Definition 4. If $A = \{(x_1, a_1), \ldots, (x_n, a_n)\}$ and $B = \{(x_1, b_1), \ldots, (x_n, b_n)\}$ are two fuzzy sets, then $C(A, B)$ is the *canonical midpoint* of A and B iff $C(A, B) = \{(x_1, a_1 + b_1/2), \ldots, (x_n, a_n + b_n/2)\}$.

Table 1 Local absolute and relative frequencies of nitrogenous bases at the three base sites of all triplet codons in UGGAAC and UACUGG

	Absolute frequencies					Relative frequencies			
	U	C	A	G	Total	U	C	A	G
UGGAAC (seq_1)									
First site	1	0	1	0	2	0.5	0	0.5	0
Second site	0	0	1	1	2	0	0	0.5	0.5
Third site	0	1	0	1	2	0	0.5	0	0.5
UACUGG (seq_2)									
First site	2	0	0	0	2	1	0	0	0
Second site	0	0	1	1	2	0	0	0.5	0.5
Third site	0	1	0	1	2	0	0.5	0	0.5

For example, the canonical midpoint of $A = \{(x, 0.2), (y, 1)\}$ and $B = \{(x, 0.8), (y, 0.6)\}$ is the fuzzy set $C(A, B) = \{(x, 0.5), (y, 0.8)\}$. We recall an additional auxiliary that we need, i.e. the notion of 'the intersection of two fuzzy sets':

The intersection $A \cap B$ of two fuzzy sets A and B over a ground set Ω is defined by means of the membership function $\mu_{A \cap B}(x)$ thus: $A \cap B = \{(x, \mu_{A \cap B}(x)) | x \in \Omega$ and $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))\}$. For example, if $\{(x, 0.2), (y, 1)\}$ and $\{(x, 0.8), (y, 0.6)\}$ are two fuzzy sets over the ground set $\{x, y\}$, then we have $\{(x, 0.2), (y, 1)\} \cap \{(x, 0.8), (y, 0.6)\} = \{(x, 0.2), (y, 0.6)\}$.

In the following definition ([4], p. 714), the term "sim(A, B)" stands for "the degree of similarity between fuzzy set A and fuzzy set B"; and the term "dif(A, B)" stands for "the degree of difference between fuzzy set A and fuzzy set B".

Definition 5.

(1) $\mathrm{sim}(A, B) = c(A \cap B)/c(C(A, B))$.
(2) $\mathrm{dif}(A, B) = 1 - \mathrm{sim}(A, B)$.

The recent version of Nieto et al.'s fuzzy polynucleotide space, $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$, is the metric space $\langle [0, 1]^{12}, \mathrm{dif} \rangle$ (cf. [4], pp. 714 ff.). In this space we obtain the following similarity and dissimilarity values for our above-mentioned sequences seq_1 and seq_2:

$$\begin{aligned}
&\mathrm{sim}(\mathrm{seq}_1, \mathrm{seq}_2) \\
&\quad = c(\mathrm{seq}_1 \cap \mathrm{seq}_2)/c(C(\mathrm{seq}_1, \mathrm{seq}_2)) \\
&\quad = c(0.5, 0, 0, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5) \\
&\qquad /c(0.75, 0, 0.25, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5) \\
&\quad = 2.5/3 = 0.833, \\
&\mathrm{dif}(\mathrm{seq}_1, \mathrm{seq}_2) = 1 - 0.833 = 0.167.
\end{aligned}$$

Recall that in $\mathrm{FPNS}_{\mathrm{Sadegh\text{-}Z}}$ we obtained for the same sequences the following values:

$$\begin{aligned}
&\mathrm{similar}(\mathrm{UGGAAC}, \mathrm{UACUGG}) = 0.09, \\
&\mathrm{differ}(\mathrm{UGGAAC}, \mathrm{UACUGG}) = 0.91.
\end{aligned}$$

Thus, $\mathrm{sim} \neq \mathrm{similar}$ and $\mathrm{dif} \neq \mathrm{differ}$. These remarkably large differences between the values obtained in both spaces might be due to the fuzzification method used by Nieto et al. and Torres et al. that consists in the determination of the base profile of a polynucleotide (see Table 1). For the moment this aspect may be ignored because it is immaterial. The inadequacy of $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$ is another matter and, as we will show in the next section, something unpleasant. To prepare the proof of this claim we need two additional notions. In an analogous fashion, as we did for our $\mathrm{FPNS}_{\mathrm{Sadegh\text{-}Z}}$, Nieto et al. ([4], p. 720) also introduce *equality* and *identity* notions for fuzzy sets thus:

Definition 6.

(1) $\mathrm{eq}(A, B) = \mathrm{sim}(A, B)$.
(2) A and B are *identical* iff $\mathrm{eq}(A, B) = 1$.

Definitions 5–6 imply the following theorem that is an analogue of Theorem 1 stated in Section 4:

Theorem 2.

(1) A and B are *identical* iff $\mathrm{sim}(A, B) = 1$.
(2) A and B are *identical* iff $\mathrm{dif}(A, B) = 0$.

6. $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$ is irremediably faulty

Before we pinpoint the reason why $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$ is objectionable, let us first present an intuitive example that may foreshadow our causal diagnosis below. We will see that in the metric space $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$ it does not make any difference whether you use in a text a word such as, say JUANNIETO, or any permutation of its triplets, say ETONNIJUA. Both words are considered identical in $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$ to the effect that no practicable and sensible semantics will be possible. This peculiarity of Nieto et al.'s system is not only strongly counter-intuitive, but also leads to absurd results. Let us reconsider, for example, our two polynucleotide sequences used in the last section, i.e.:

$$\begin{aligned}
&\mathrm{seq}_1 \equiv \mathrm{UGGAAC} \quad \text{codes for}: \text{tryptophan}-\text{asparagine} \\
&\mathrm{seq}_2 \equiv \mathrm{UACUGG} \qquad\quad \text{tyrosine}-\text{tryptophan}
\end{aligned}$$

These sequences have only one single base in common, i.e. base U at their initial sites. The remainder of the sequences, 83.3%, is completely different. Their congruence amounts merely to 16.7%. We have seen, however, that in $\mathrm{FPNS}_{\mathrm{Nieto\text{-}et\text{-}al}}$ they appear highly similar and only slightly different:

$$\begin{aligned}
&\mathrm{sim}(\mathrm{seq}_1, \mathrm{seq}_2) = 0.833, \\
&\mathrm{dif}(\mathrm{seq}_1, \mathrm{seq}_2) = 1 - 0.833 = 0.167.
\end{aligned}$$

This result is unacceptable because the reality is the other way around. We may witness an even more strange performance of $\mathrm{FPNS}_{\mathrm{Nieto\text{-}et\text{-}al}}$ by rearranging the triplets of the first sequence above, seq_1. This sequence is a concatenation of the triplets UGG and AAC. A permutation of its triplets yields:

$$\mathrm{seq}_3 \equiv \mathrm{AACUGG} \quad \text{codes for}: \text{asparagine}-\text{tryptophan}$$

Note that seq_1 and seq_3 are compounds of the same triplet codons AAC and UGG, although they are two completely different polynucleotide *molecules*; and as indicated by the peptid strings they code for, their causal effects are different as well. Due to these molecular-biological differences no biochemist and no geneticist or genetic engineer would consider seq_1 and seq_3 to be identical polynucleotide sequences. Correspondingly, no such expert would approve gene-technically substituting in a chromosome the gene AACUGG for the gene UGGAAC on the grounds that they were identical. They are not. However, it is exactly the supposition of this objectively non-existent identity that renders $FPNS_{Nieto-Torres}$ unacceptable. The reason is this. Although $seq_1 \neq seq_3$, Table 2 demonstrates that seq_3 has the same base profile as seq_1 that was displayed in Table 1 in the last section.

From the relative frequencies on the right-hand side of Table 2 we obtain for sequence seq_3 the following membership vector reduced to 12-dimensions:

$$seq_3 \equiv (0.5, 0, 0.5, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5) \equiv \text{AACUGG}.$$

Recall that for seq_1 we had obtained the following vector that is identical with seq_3 above:

$$seq_1 \equiv (0.5, 0, 0.5, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5) \equiv \text{UGGAAC}$$

It is therefore not surprising that the awkwardness for $FPNS_{Nieto-Torres}$ presents itself convincingly right now. Although the degree of molecular congruence between AACUGG and UGGAAC is 0, one may easily compute:

$$\begin{aligned} &\text{sim}(seq_1, seq_3) \\ &\quad = c(seq_1 \cap seq_3)/c(C(seq_1, seq_3)) \\ &\quad = c(0.5, 0, 0.5, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5) \\ &\qquad /c(0.5, 0, 0.5, 0, 0, 0, 0.5, 0.5, 0, 0.5, 0, 0.5) \\ &\quad = 3/3 = 1, \\ &\text{dif}(seq_1, seq_3) = 1 - 1 = 0, \\ &\text{eq}(seq_1, seq_3) = 1 \quad \text{i.e.}, \\ &\text{eq}(\text{UGGAAC}, \text{AACUGG}) = 1. \end{aligned}$$

These results, in conjunction with Theorem 2 stated in the last section, imply that in $FPNS_{Nieto-Torres}$:

UGGAAC and AACUGG are identical.

This consequence also follows both from the various difference and similarity concepts Nieto et al. introduce and analyze in [4], and from the premise that the space $FPNS_{Nieto-Torres}$, i.e. $\langle [0, 1]^{12}, \text{dif} \rangle$, is allegedly a metric space. The definition of the term 'metric space' implies for any two points $x, y \in [0, 1]^{12}$ that the following statement is true:

$$\text{dif}(x, y) = 0 \quad \text{iff} \quad x = y.$$

Hence, in $FPNS_{Nieto-Torres}$ sequences seq_1 and seq_3 are identical polynucleotides. But are they really identical? They have no single base in common at their corresponding sites. And as we have pointed out above, they code for completely different peptid strings. To call them identical would be like claiming that a six-digit natural number such as 831456 is identical with its permutation 456831. In contrast, we obtain the following realistic results in $FPNS_{Sadegh\text{-}z}$:

- differ(UGGAAC, AACUGG) = 1
- similar(UGGAAC, AACUGG) = 0
- UGGAAC and AACUGG are not identical.

It is worth noting that in $FPNS_{Nieto-Torres}$ even polynucleotide sequences of unequal size may turn out 'identical' when they have the same base profile, and thus, identical vectors. Here is a simple example:

$seq_4 \equiv$ UGGAAC codes for : tryptophan−asparagine
$seq_5 \equiv$ AACUGGAACUGG asparagine−tryptophan−asparagine−tryptophan

with:

$$\text{sim}(s_4, s_5) = 1, \quad \text{dif}(s_4, s_5) = 0, \quad \text{eq}(s_4, s_5) = 1.$$

So, seq_4 and seq_5 are also identical. However, it will be impossible to convince any student of biosciences or medicine or the present reader of this paper that things are so as $FPNS_{Nieto-Torres}$ maintains.

7. Results

It has been shown in the preceding sections that $FPNS_{Nieto-Torres}$, proposed and extensively analyzed by Nieto et al. [3,4] and Torres et al. [5,6], is counter-intuitive. In their so-called "fuzzy polynucleotide space" two structurally and functionally completely different polynucleotide chains turn out highly similar and even identical. On this account it is to be considered as defective.

Table 2 Local absolute and relative frequencies of nitrogenous bases at the three base sites of all triplet codons in AACUGG (compare with UGGAAC in Table 1. They are identical!)

	Absolute frequencies					Relative frequencies			
	U	C	A	G	Total	U	C	A	G
AACUGG (seq_3)									
First site	1	0	1	0	2	0.5	0	0.5	0
Second site	0	0	1	1	2	0	0	0.5	0.5
Third site	0	1	0	1	2	0	0.5	0	0.5

The cause of its defectiveness is easily identified. It consists in the use of the purely statistical *base profiles* to transform any polynucleotide of arbitrary length to a 12-dimensional vector. The authors seem to have overlooked that 'the base profile of *x* is *y*' is not a one-to-one, but a many-to-one function that ignores the molecule-specific order of triplets in a molecule. As we have seen above, two or more different polynucleotide sequences such as:

UGGAAC
AACUGG

will therefore occupy one and the same point in $[0, 1]^{12}$ when they have the same base profile. Thus, the determination of the base profile of a polynucleotide chain destroys the order in which single triplets are linearly linked to form what has come to be called a polynucleotide *chain* or *sequence*. By ignoring the order of the triplet codons in a sequence Nieto et al.'s conception treats the ordered triplets of a polynucleotide chain as an unordered *heap* of triplets. This is demonstrated by the tables of base frequencies and profiles in the present paper and in [4,5]. These statistics disregard the relational location of a *triplet* in a sequence, and hence, the sequential nature of a polynucleotide.

The disregard of $FPNS_{Nieto-Torres}$ for order in a polynucleotide sequence is characterized by identifying such a *sequence* of $n > 1$ triplet codons with any other one that possesses the same base profile. However, base profile is a statistical 'average out' attribute and is defined by the proportion of nitrogenous bases and the three triplet sites {1, 2, 3}. As we have exemplified above by the following two sequences with 'identical' base profile:

$seq_4 \equiv$ UGGAAC codes for : tryptophan−asparagine
$seq_5 \equiv$ AACUGGAACUGG asparagine−tryptophan−asparagine−tryptophan,

identical base profiles are ubiquitous to the effect that a polynucleotide of length 2^{58} may possess the same base profile as a polynucleotide of length 6. It is irrational to call them identical nonetheless because in the same fashion the human genome could turn out identical with that of Drosophila melanogaster. But Nieto et al. would have to explain how it is possible that these two identical genomes bring about two species of such huge difference that exists between Homo sapiens and Drosophila melanogaster.

From what has just been stated we can conclude that in $FPNS_{Nieto-Torres}$ a polynucleotide sequence has a potentially infinite number of pseudo-doubles which fake similarity and identity with the original. A salient subset of these pseudo-doubles is the tripletwise permutation set of a polynucleotide. To define the term 'tripletwise permutation' we may consider any polynucleotide molecule that is composed of $n \geq 1$ different *tri*plets as a sequence of the form $TRI_1TR_2 \ldots TRI_n$ whose units are its triplet codons TRI_i with $1 \leq i \leq n$. For instance, the RNA molecule:

$seq_6 \equiv$ CUCAGGUCACAC

of length 12 comprises the following four triplets: CUC, AGG, UCA, CAC. It is thus a concatenation thereof. Any re-ordering of the triplet sequence in a molecule of the form $TRI_1TR_2 \ldots TRI_n$ is a tripletwise permutation of the molecule. For instance, our polynucleotide chain seq_6 concatenating 4 triplets has 4! = 24 tripletwise permutations.

Any polynucleotide sequence composed of $n \geq 1$ different triplets $TRI_1, \ldots, TRI_n$ has $n!$ tripletwise permutations. Unfortunately, in $FPNS_{Nieto-Torres}$ all $n!$ tripletwise permutations of such a polynucleotide turn out to be identical molecules because they have the same base profile and for that reason the same 12-dimensional vector. For instance, regarding the above example with $n = 4$ triplets we present only one of its 24 tripletwise permutations, i.e. sequence seq_7, that in $FPNS_{Nieto-Torres}$ is identical with seq_6:

$seq_6 \equiv$ CUC−AGG−UCA−CAC codes for leucine−arginine−serine−histidine
$seq_7 \equiv$ AGG−CUC−CAC−UCA codes for arginine−leucine−histidine−serine.

This example speaks for itself. It demonstrates convincingly the disregard for order that is characteristic of Nieto et al.'s construct. The term "fuzzy polynucleotide space" in "$FPNS_{Nieto-Torres}$" is therefore a misnomer. Their construct does not treat polynucleotides as *ordered* biomolecules to measure relationships such as distances and similarities between them. It is concerned with unordered *heaps* of their triplets whose properties and relationships it measures. It is true that all 24 tripletwise permutations of the RNA chain CUCAGGUCACAC comprise *the same heap* of 4 triplets. That is, their triplet heaps are *identical*. But that does not mean that they are identical *sequences*. For this reason $FPNS_{Nieto-Torres}$ enables inquiries only into unordered heaps, but not into ordered sequences. So, it does not bear any relevance to biosciences and medicine. In contrast to what its authors repeatedly claim in their publications on their construct [3–6], it cannot be reasonably used in sequence analysis, sequence comparison, diagnostics or elsewhere.

8. Conclusion

To summarize, it is not Nieto et al.'s concepts of difference, similarity, equality or identity that would render their system inadequate. Conceptual shortcomings of this type are corrigible. Incorrigibly faulty, however, is the very fundament of $FPNS_{Nieto-Torres}$ consisting in the method the authors use to fuzzify polynucleotide chains so as to transform them to points in $[0, 1]^{12}$, i.e. the concept of base profile. In our concluding analysis of this issue we confine ourselves to RNA. What is said also applies to DNA if one substitutes the word 'DNA' for 'RNA'.

Let *bp* be a unary function termed 'the base profile of' that maps all non-fuzzy RNA sequences to their *base profile* – as introduced in Section 5 above – such that for any RNA molecule y of length $n \geq 3$ we have that $bp(y) = (x_1, \ldots, x_{12})$ with each $x_i \in [0, 1]$. We thus obtain the mapping:

$$bp : \{RNA\} \rightarrow [0,1]^{12}$$

where for clarity's sake '{RNA}' stands as shorthand for '$\{y|y$ is an RNA molecule of length $n \geq 3\}$' representing the set of all RNA sequences comprising at least one triplet. $FPNS_{Nieto-Torres}$ is based on such a mapping to produce the material that it processes. Unfortunately, however, the function bp is not injective and thereby creates an undesirable consequence. To explain, let us first consider its inverse, bp^{-1}. While bp maps a point of its domain to a point of its range, its inverse maps a point of its domain $[0, 1]^{12}$ to a *set* of RNA molecules. Its range is thus a family of sets, i.e. the powerset of {RNA}:

$$bp^{-1} : [0,1]^{12} \rightarrow 2^{\{RNA\}}$$

such that for any $\boldsymbol{x} = (x_1, \ldots, x_{12}) \in [0, 1]^{12}$:

$$bp^{-1}(\boldsymbol{x}) = \{y \in \{RNA\} | bp(y) = \mathbf{x}\} \in 2^{\{RNA\}}. \quad (2)$$

Let us call an object that is identical with a particular object *x*, an *identical* of *x*. The set of all identicals of an object *x*, written id(*x*), is:

$$id(x) = \{y | x \text{ is identical with } y\}. \quad (3)$$

This informal definition (3) and Theorem 2 in Section 5 imply for any RNA sequence s_i:

$$id(s_i) = \{s_j | sim(s_i, s_j) = 1\} = \{s_j | dif(s_i, s_j) = 0\}. \quad (4)$$

The undesirable consequence of bp referred to above may now be described as follows. Given any particular RNA sequence comprising $n \geq 1$ triplets such that its base profile is the point s_i in $[0, 1]^{12}$, we have according to (2)–(4) above the following relationships in $FPNS_{Nieto-Torres}$:

$$\begin{aligned} \{s_j | sim(s_i, s_j) = 1\} &= \{s_j | dif(s_i, s_j) = 0\} \\ &= bp^{-1}(s_i) \in 2^{\{RNA\}} \\ &= \{y \in \{RNA\} | bp(y) = s_i\}. \end{aligned} \quad (5)$$

Statements (4) and (5) imply:

$$id(s_i) = \{y \in \{RNA\} | bp(y) = s_i\}.$$

The resulting set $\{y \in \{RNA\} | bp(y) = s_i\}$ represents the set of all identicals of our RNA sequence s_i. Fatal to $FPNS_{Nieto-Torres}$ is the fact that this set is potentially infinite. That is, according to $FPNS_{Nieto-Torres}$ every RNA molecule consisting of $n \geq 1$ triplets has a potentially infinite set of identicals. The reason is that for any such RNA sequence s_i the following two Inductions 1 and 2 obtain in $FPNS_{Nieto-Torres}$:

Induction 1.

(1) $s_i \in id(s_i)$,
(2) If $s_j \in id(s_i)$, then $s_p \in id(s_i)$ for $\forall s_p \in tp(s_j)$ where $tp(s_j)$ is the tripletwise permutation set of s_j.

Proof. The basis step, step 1, is trivial. The induction step, step 2, is implied by the definition of the operator bp according to which all elements of a tripletwise permutation set of an RNA molecule have the same base profile and thus the same 12-dimensional vector. Therefore, they are the same point in $[0, 1]^{12}$.

Induction 2.

(1) $s_i \in \mathrm{id}(s_i)$,
(2) If $s_j \in \mathrm{id}(s_i)$, then $2s_j \in \mathrm{id}(s_i)$ such that $2s_j$ is a sequence of the same structure and twice as large as s_j.

Proof. Step 1 is trivial. Again, the induction step follows from the definition of the operator bp according to which an RNA sequence s_j and any of its multipliers have the same base profile and thus the same 12-dimensional vector. So, they are the same point in $[0, 1]^{12}$.

Note that the basis step, step 1, in both inductions follows from the predicate-logical axiom of identity ('$x = x$') and is thus a logical truth. On this account a set $\mathrm{id}(s_i)$ is never empty. It contains at least sequence s_i itself. So, the induction step in both inductions fires automatically to the effect that the set of identicals of a polynucleotide, $\mathrm{id}(s_i)$, inductively grows. The system $\mathrm{FPNS}_{\mathrm{Nieto-Torres}}$ must therefore be viewed as *inductively explosive* in that the set of identicals of a polynucleotide molecule unstoppably expands. In this way the system spontaneously generates a potentially infinite set of false identity claims and is therefore useless both in theory and practice. This disaster is due to its fundamental operator bp. In conjunction with Theorem 2, stated in Section 5, bp leads to inductive explosion of the set of identicals of a polynucleotide.

By contrast, the original system $\mathrm{FPNS}_{\mathrm{Sadegh\text{-}Z}}$ suggested by the present author is based on the concept of an ordered fuzzy set. In this system the mapping of polynucleotides to the metric space $\langle [0, 1]^n, \mathrm{differ} \rangle$ is injective. Equivocations and problems as above cannot arise. Moreover, $\mathrm{FPNS}_{\mathrm{Sadegh\text{-}Z}}$ is capable of dealing with genuinely fuzzy polynucleotide sequences whose vectors are of the form $(x_1, \ldots, x_n)$ with each $x_i \in [0, 1]$. Nieto, Torres et al.'s system, however, also lacks this capability because its base profile function bp operating on {RNA} deals only with crisp, non-fuzzy polynucleotide sequences with bit vectors of the form $(y_1, \ldots, y_n)$ such that $y_n \in \{0, 1\}$.

Appendix

Many publications on polynucleotides in computer science journals contain errors regarding the nature, structure, and function of polynucleotides. Their authors confuse, for example, nucleotides with nucleosides, nucleotides with nitrogenous bases or even with nucleic acids, triplet codons with single nucleotide molecules, codons with amino acids they code for, and the like. It may therefore be helpful at this juncture to clarify the terminology so as to prevent misunderstandings.

The genetic material of biological species known as nucleic acids consists of large sequential molecules. There are two types of nucleic acids, *d*eoxyribo*n*ucleic *a*cid (DNA) and *r*ibo*n*ucleic *a*cid (RNA). DNA is the genetic material that all single-cell and multiple-cell *organisms* and some types of *viruses* (DNA viruses) inherit from their parents (recall 'double helix'). Some other viruses bear RNA as their genetic material and are therefore called RNA viruses. Both DNA and RNA govern, among other things, the production of proteins in organisms and viruses, and thus their life and death affairs. Our knowledge of their structure and function is therefore essential in managing medical and biotechnological problems.

As a linear polymer a DNA and RNA molecule is a sequence of smaller molecules called its monomeric units. Chemically these monomeric units belong in the category of nucleotides. A number of $n > 1$ nucleotides are linearly linked by bonds to form a chain that is called a trinucleotide or *triplet* if $n = 3$, an *oligonucleotide* if 'n is small', and a *polynucleotide* if 'n is large'. A mononucleotide is a single nucleotide molecule. For simplicity's sake we will use the term 'polynucleotide' only to denote all elements of the whole category. For example, the genetic material of the tiny RNA virus HIV consists of about 10,000 nucleotide monomers. Any of the 46 human chromosomes in a human cell nucleus is composed of about sixty-five million of them.

A mononucleotide is itself composed of three smaller molecular building blocks: a five-carbon sugar, a phosphate group, and a nitrogenous base (see Fig. 1a). In a DNA and RNA polynucleotide chain, a nucleotide monomer has its phosphate group bonded to the sugar of the next nucleotide link. So the chain has a regular sugar-phosphate backbone with variable appendages. These appendages are *four* possible nitrogenous bases called:

Adenine = A
Cytosine = C
Guanine = G
Thymine = T

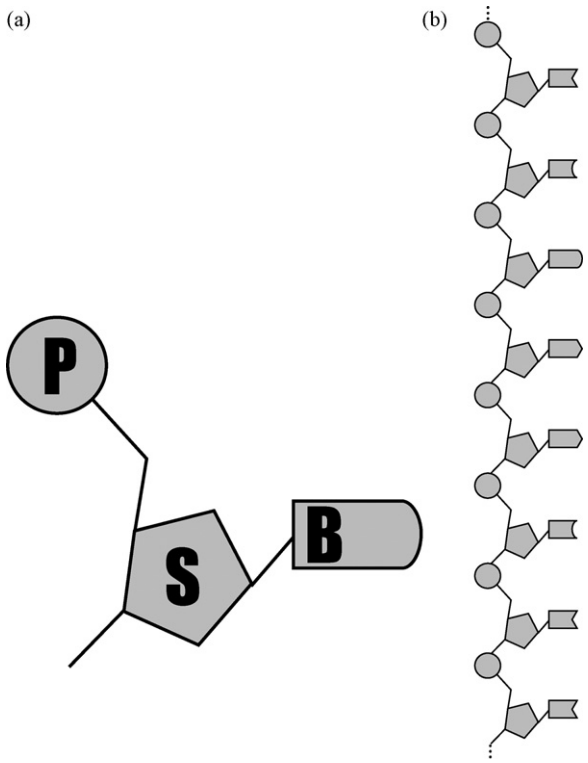

Figure 1 A nucleotide monomer (a) and a single-strand polynucleotide (b). P = phosphate group, S = sugar, and B = nitrogenous base. In DNA, the sugar is deoxyribose, whereas the sugar molecule in RNA is ribose. A DNA chain is usually double-stranded. When synthesizing mRNA and as a cell prepares to divide, the two strands are separated. But we will here not be concerned with these details.

in DNA, but in place of the latter,

Uracil = U

in RNA. The specific sequence of these base appendages in a polynucleotide is characteristic of the molecule and is referred to as its *base sequence* (see Fig. 1b). Whereas a particular polynucleotide may have the base sequence GUAUACUGU..., etc., another one may have the base sequence GTTTACACT..., etc.

In a cell's chain of command, instructions for protein synthesis flow from DNA to RNA (i.e., messenger RNA = mRNA) to protein. In the latter step, the genetic message encoded in an mRNA base sequence such as GUAUACUGU... orders amino acids into a protein of specific amino acid sequence. The mRNA message is read in the cell as a sequence of base triplets XYZ, analogous to three-letter code words. An mRNA base triplet XYZ is therefore called a *codon*. A triplet codon XYZ along an mRNA sequence specifies which one of the 20 amino acids will be inserted in the appropriate site of a protein chain. For example, the codon GUA is responsible for the amino acid valine. Since there are four bases for mRNA, there are $4 \times 4 \times 4 = 64$ such codons making up the dictionary of the *genetic code*. The dictionary is redundant because $64 > 20$. It is not one-to-one, but many-to-one. For instance, four codons GUA, GUC, GUG, and GUU stand for the amino acid valine, and thus you get from the above mRNA segment GUAUACUGU... the protein chain valine-tyrosine-cysteine-..., etc.

Due to these bio-informational facts, the focus of our concern in the main text will be the *base sequence* of polynucleotides in that we will translate it into an ordered fuzzy set. The idea behind this plan is the recognition that by translating a subject into a fuzzy set the constructs of fuzzy theory become accessible to that subject domain.

References

[1] Sadegh-Zadeh K. Fuzzy genomes. Artif Intell Med 2000;18:1—28.
[2] Sadegh-Zadeh K. Ein Verfahren zur Fuzzydecodierung und Fuzzydechiffrierung von Informationen. *Offenlegungsschrift DE 199 36 925 A 1*. Deutsches Patent- und Markenamt, 2001.
[3] Nieto JJ, Torres A, Vázquez-Trasande MM. A metric space to study differences between polynucleotides. Appl Math Lett 2003;16:1289—94.
[4] Nieto JJ, Torres A, Georgiou DN, Karakasidis TE. Fuzzy polynucleotide spaces and metrics. Bull Math Biol 2006;68:703—25.
[5] Torres A, Nieto JJ. The fuzzy polynucleotide space: basic properties. Bioinformatics 2003;19:587—92.
[6] Torres A, Nieto JJ. Fuzzy logic in medicine and bioinformatics. J Biomed Biotechnol 2006;1—7 [Article ID 91908].
[7] Sadegh-Zadeh K. Fuzzy polymers. Research Report 1997—1998, University of Münster. See http://www.uni-muenster.de/Rektorat/Forschungsberichte-1997-1998/fo05ea07.htm (Accessed: 15 March 2007)
[8] Zadeh LA. Towards a theory of fuzzy systems. In: Kalman RE, DeClairis RN, editors. Aspects of networks and systems theory. New York: Holt, Rinehart & Winston; 1971 . pp. 469—90.
[9] Kosko B. Neural Networks and Fuzzy Systems. A dynamical systems approach to machine intelligence. Englewood Cliffs, NJ: Prentice Hall; 1992.
[10] Kosko B. Fuzzy engineering. Upper Saddle River, NJ: Prentice Hall; 1997.
[11] Lin CT. Adaptive subsethood for radial basis fuzzy systems. In: Kosko B, editor. Fuzzy engineering. Upper Saddle River, NJ: Prentice Hall; 1997. pp. 429—64 [Chapter 13].
[12] Nieto JJ, Torres A. Midpoints for fuzzy sets and their application in medicine. Artif Intell Med 2003;27:81—101.