ELSEVIER

Artificial Intelligence in Medicine 18 (2000) 1–28

Artificial Intelligence in Medicine

www.elsevier.com/locate/artmed

Fuzzy genomes

Kazem Sadegh-Zadeh *

Theory of Medicine Department, University of Münster Medical Institutions, Waldeyer Street 27, Münster 48149, Germany

Accepted 6 August 1999

Abstract

A metric space — dubbed the *fuzzy polynucleotide space* — is presented for diagnostic purposes in the widest sense to measure the degree of difference and similarity between sequences of nucleic acids. To this end, these acids are transformed to ordered fuzzy sets. They thus become representable as points in n-dimensional unit hypercubes that may be endowed with various metrics. In this way, genetic information in particular and genetics in general become amenable to fuzzy theory, geometry, and topology.

Keywords: Genome; Nucleic acids; Base sequence; Genetic diagnosis; Polynucleotides; Fuzzy hypercube; Fuzzy polynucleotide space; Similarity; Dissimilarity; Polymers

1. Introduction

Medicine at the turn of the century is characterized by the deepest change it has ever been subject to in its history, i.e. its transformation from a healing profession to a branch of biotechnology. Viewed from an evolutionary perspective, this transformation appears as an aspect of a Darwin–Lamarckian autoevolution of life on earth [14]. The nucleic acids DNA and RNA as the genetic material of living things and viruses play the pivotal role in this arena. Necessary techniques in dealing with this material are sequence analysis and sequence comparison.

Sequence analysis or sequencing aims at determining the building blocks of a nucleic acid, i.e. its monomeric units — nucleotides — and their order in the

* Tel.: +49-251-8355291; fax: +49-251-8355339.

E-mail address: zadeh@uni-muenster.de (K. Sadegh-Zadeh)

0933-3657/00/$ - see front matter

PII: S0933-3657(99)00032-9

molecular chain of the acid. It commenced in the early 1960s with the deciphering of the genetic code. *Sequence comparison* is, by contrast, a taxonomic and diagnostic task to determine the structural relationships such as identity, difference, and similarity between chains of nucleic acids whose sequences have already been analyzed and are known. It deals with questions such as, for example, 'is this piece of RNA before my eyes an HIV or something else?'. To answer questions of this type requires reliable techniques of sequence comparison between the unknown and the known. We will in the following be concerned with this problem and will present a novel methodology based on fuzzy theory.

In Section 2, nucleic acids (DNA and RNA) are transformed to ordered fuzzy sets. In so doing our primary aim is to make genes, genomes, and genetics directly amenable to fuzzy theory. A first step in this direction is taken by investigating the unit hypercube geometry of nucleic acids in Section 3. Measures of identity, difference, and similarity for genetic material are provided that contribute to the enhancement of taxonomic and diagnostic accuracy and computation in genetics, microbiology, biochemistry, and biotechnology. Two interesting by-products of our analysis are the recognition that the genetic code is 12-dimensional, and the view that genes and genomes are fuzzy entities. Although our methodology in this paper is applied only to nucleic acids, it is general enough to cover all polymers [15].

2. DNA and RNA as ordered fuzzy sets

In this section, we will represent the nucleic acids DNA and RNA as ordered fuzzy sets to inquire into the ontology and geometry of genetic information in the next section. Since our presentation is intended to be self-contained, a few terminological arrangements on nucleic acids and fuzzy sets may be in order (see Appendix A).

2.1. Polynucleotide sequences formalized

DNA and RNA are linear polymers of nucleotides and are therefore called polynucleotides. We will formalize polynucleotide sequences and will then show, first, that a polynucleotide is an ordered fuzzy set and as such it has, second, an unequivocal *fuzzy code*.

An *ordered fuzzy* set is simply a fuzzy set over an ordered ground set $\Omega = \langle x_1, x_2, x_3, \ldots \rangle$. For example, let the tuple $\langle 0, 1, 2, 3, 4 \rangle$ be the ordered set of the first five natural numbers. The set of prime numbers contained therein yields the following ordered fuzzy set when its elements are supplemented by the degree of their primeness: $\langle (0, 0), (1, 0), (2, 1), (3, 1), (4, 0) \rangle$.

As usual, we will identify a polynucleotide with the sequence of its nitrogenous bases. It is this base sequence — as a linear proxy of the polynucleotide — that we will transform to an ordered fuzzy set.

We consider a base sequence of a DNA or RNA molecule as a string $\mathfrak{s} = s_1 \ldots s_m$ composed of $m \geq 1$ linearly ordered signs $s_1, \ldots, s_m$ that are placed next to each

other in juxtaposition such as, for example, CCGAGTACC as a short segment of a single-strand DNA. We will in this paper be dealing only with single-strand polynucleotides.

Definition 1.

1. If $\langle S_1, \ldots, S_n \rangle$ is the alphabet of a language with $n \geq 1$ signs $S_1, \ldots, S_n$, an instance of a sign $S_j \in \langle S_1, \ldots, S_n \rangle$ is called a string or a *sequence* over $\langle S_1, \ldots, S_n \rangle$ of length 1;
2. If $\mathfrak{s}_1$ and $\mathfrak{s}_2$ are sequences over $\langle S_1, \ldots, S_n \rangle$ of length p and q, respectively, then their concatenation $\mathfrak{s}_1\mathfrak{s}_2$ is a sequence over $\langle S_1, \ldots, S_n \rangle$ of length $p + q$.

For example, the phrase 'GENE' is a sequence of length 4 over the Latin alphabet $\langle$A, B, C, …, Z$\rangle$, whereas the phrase

UACUGU

is a sequence over the RNA alphabet $\langle$U, C, A, G$\rangle$ of length 6 consisting of two codons, UAC for amino acid tyrosine, and UGU for amino acid cysteine. As this latter example demonstrates, a polynucleotide is considered as a sequence of length $m \geq 1$ over a particular alphabet. We distinguish between

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

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

as introduced in Appendix A. With regard to a sequence $\mathfrak{s} = \mathrm{s}_1 \ldots \mathrm{s}_m$ of length m, the m-tuple (1, …, m) is referred to as the *position numbers* of its signs where s_i is its i-th sign with $1 \leq i \leq m$. For instance, in the RNA sequence UACUGU above, s_5 is a G$\in\langle$U, C, A, G$\rangle$.

2.2. An intuitive illustration of the fuzzy code

The fuzzy code of a sequence we are searching for may be intuitively illustrated by a simple example. We will now transform our above-mentioned RNA sequence:

UACUGU

to an informationally equivalent bit sequence, i.e. a sequence that consists only of binary digits 0 and 1, and represents the source sequence. To this end, we represent in a sequence any sign S_i of the RNA alphabet $\langle$U, C, A, G$\rangle$ by the *entire alphabet* in that a sign $S_i \in \langle$U, C, A, G$\rangle$ is represented by 1 if it is present in the sequence, and by 0, else. We thus have:

U$\in\langle$U, C, A, G$\rangle$ in a sequence	is	$\langle 1, 0, 0, 0 \rangle$	or simply:	1000
C$\in\langle$U, C, A, G$\rangle$ in a sequence	is	$\langle 0, 1, 0, 0 \rangle$		0100
A$\in\langle$U, C, A, G$\rangle$ in a sequence	is	$\langle 0, 0, 1, 0 \rangle$		0010
G$\in\langle$U, C, A, G$\rangle$ in a sequence	is	$\langle 0, 0, 0, 1 \rangle$		0001

By this transformation, the above RNA sequence UACUGU turns out to be the following bit sequence of length 24:

100000100100100000011000

Written in vector notation, it is the vector:

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

This 24-dimensional bit vector represents the *fuzzy code* of our example sequence UACUGU. As we will see below, the vector is a point in a 24-dimensional unit hypercube (see Section 3.1). By generalizing this method, we will become able to represent any polynucleotide as a point in a unit hypercube.

The above illustration was meant only as an intuitive and provisional sketch and should not be mistaken for the very idea we now turn to.

2.3. The fuzzy decoding of genetic information

The 24-dimensional example vector we arrived at above had components only from among the bivalent set {0, 1}. Such a restriction is unnatural, however. As we will see below, a polynucleotide base sequence may have a genuinely fuzzy code such as:

(0.2, 0.5, 0.3, …, 0, 0, 1)

in that its components may belong to the real interval [0, 1]. Important and interesting consequences are associated with this fact. To uncover them let us first introduce the notion of a *fuzzy alphabet*.

The notion of 'alphabet' is used here in the broadest sense of the word. An alphabet $\langle S_1, \ldots, S_n \rangle$ is any collection of prototype signs $S_1, \ldots, S_n$ according to which any individual sequence $s_1 \ldots s_m$ may be formed by concatenating concrete copies of the prototypes. Examples are the Morse Code, the Latin alphabet, chemical elements of which chemicals are composed, the elementary phonemes of human speech, etc.

The physical appearance of a sign s_i occurring in an individual sequence $s_1 \ldots s_m$ over an alphabet $\langle S_1, \ldots, S_n \rangle$ may more or less deviate from its prototype $S_j \in \langle S_1, \ldots, S_n \rangle$. For example, '$\mathscr{A}$' in the word '$\mathscr{A}$lphabet' is still an A, although $\mathscr{A} \neq$ A. A prototype sign $S_j \in \langle S_1, \ldots, S_n \rangle$ such as 'A' therefore may be viewed as the name of a fuzzy set such that an individual copy s_i in a sequence may only to a particular extent be a member of that set, i.e. a sign of the type S_j (see Fig. 1).

A fuzzy alphabet is an alphabet comprising fuzzy prototype signs $\langle S_1, \ldots, S_n \rangle$ of the type characterized above. It is interesting to note that all natural language alphabets are fuzzy alphabets. A convincing evidence is provided by considering the multitude of different typographies for all you want to write or to print. All of them are sequences over the same Latin alphabet nonetheless.

In what follows, the term 'alphabet' is meant as a fuzzy alphabet in this sense. DNA and RNA alphabets are thus considered fuzzy alphabets. That means that a particular RNA base, for example, may be present in a particular position of a

sequence to some degree between 0 and 1. This is the basic thought our framework rests upon. Accordingly, a base sequence of the form:

$$s_1 s_2 \ \ldots \ s_m$$

such as UACUGU will have a real vector:

$$(r_1, \ \ldots, \ r_q)$$

as its fuzzy code with $r_i \in [0, 1]$. The determination of this fuzzy code of a sequence is referred to here as *fuzzy decoding*. It is carried out by applying to a sequence $\mathfrak{s}$ the function *fcode* that yields the fuzzy code of the sequence:

$$fcode(\mathfrak{s}) = (r_1, \ \ldots, \ r_q).$$

This function *fcode* will be a composition:

$$fvector \circ fset \circ gmatr \equiv fcode$$

of three functions:

fvector	reminiscent of:	fuzzy vector
fset		fuzzy set
gmatr		ground matrix

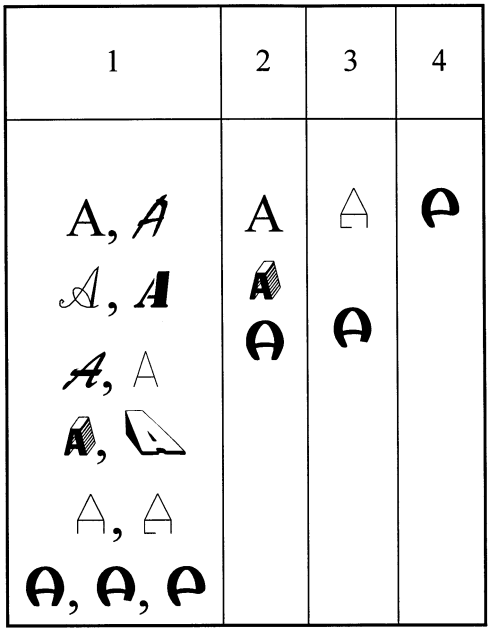

Fig. 1. For instance, given the Latin alphabet ⟨A, B, …, Z⟩, the copies in column 1 of this figure are As to differing degrees. While those in column 2, for example, are As to the extent 1, the copies in column 3 are so to a lesser extent than 1, and the copy in column 4 is an A to the extent 0.

When successively applied to a sequence, they return its fuzzy code:

$$fcode(\mathfrak{s}) = fvector(\,fset(gmatr(\mathfrak{s}))).$$

This formula is the key for the fuzzy decoding we are presently pursuing.

2.3.1. The ontology of a sequence

In order for a base sequence $\mathfrak{s} = s_1 \ldots s_m$ to be transformed to a fuzzy set, a ground set Ω must be identified upon which $\mathfrak{s}$ is such a fuzzy set. Given the alphabet $\langle S_1, \ldots, S_n \rangle$ over which $\mathfrak{s}$ is a sequence, the ground set Ω may be constructed as the ordered set of *all* combinatorially possible occurrences in that sequence of the alphabet signs $S_1, \ldots, S_n$, that is $\langle 1, \ldots, m \rangle \times \langle S_1, \ldots, S_n \rangle$ where the m-tuple $(1, \ldots, m)$ represents the position numbers of the sequence. This ground set Ω is produced by the function *gmatr* in the following way:

Definition 2. Let $\mathfrak{s} = s_1 \ldots s_m$ be a sequence over $\langle S_1, \ldots, S_n \rangle$, then $gmatr(\mathfrak{s}) =$ *ground_matrix* such that

$$\text{ground\_matrix} = \left\{ \begin{array}{c} 1 \\ . \\ . \\ m \end{array} \right\} * (S_1, \ldots, S_n) = \left\{ \begin{array}{c} S_1 1, \ldots, S_n 1 \\ \cdots \\ \cdots \\ S_1 m, \ldots, S_n m \end{array} \right\}$$

This $m \times n$ ground_matrix is the outer product of the column vector $(1, \ldots, m)$ of the sequence's position numbers with the row vector $(S_1, \ldots, S_n)$ of the alphabet. An entry '$S_i j$' in the matrix reads 'sign S_i of the alphabet in position j of the sequence' irrespective of whether or not S_i is actually present in position j of the sequence. The matrix in its rows contains the ordered ground set $\Omega = \langle S_1 1, \ldots, S_i j, \ldots, S_n m \rangle$ we were searching for. For example, given the triplet codon:

UAC

over the RNA alphabet $\langle$U, C, A, G$\rangle$, we obtain its ground set Ω by the (3×4)-matrix:

$$gmatr(\text{UAC}) = \left\{ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \right\} * (\text{U, C, A, G}) = \left\{ \begin{array}{cccc} \text{U in 1}, & \text{C in 1}, & \text{A in 1}, & \text{G in 1} \\ \text{U in 2}, & \text{C in 2}, & \text{A in 2}, & \text{G in 2} \\ \text{U in 3}, & \text{C in 3}, & \text{A in 3}, & \text{G in 3} \end{array} \right\}$$

2.3.2. The sequence as an ordered fuzzy set

The ground_matrix above containing the ordered ground set $\Omega = \langle S_1 1, \ldots, S_i j, \ldots, S_n m \rangle$ allows for the construction of the sequence $\mathfrak{s} = s_1 \ldots s_m$ as an ordered

fuzzy set in a simple way. What is needed is a membership function of the sequence, $\mu_{\mathfrak{s}}$, for this purpose.

A membership function $\mu_{\mathfrak{s}}$ maps the ground_matrix to the unit interval [0, 1] with each $\mu_{\mathfrak{s}}(S_ij)$ being the extent to which a sign S_i of the alphabet is present in position j of the sequence. All ensuing pairs $(S_ij, \mu_{\mathfrak{s}}(S_ij))$ are then collected to yield an ordered fuzzy set $\langle(S_11, \mu_{\mathfrak{s}}(S_11)), \ldots, (S_nm, \mu_{\mathfrak{s}}(S_nm))\rangle$. This ordered fuzzy set is the fuzzified sequence $\mathfrak{s}$. These two steps are accomplished by the following two Definitions 3–4.

Definition 3. $\mu_{\mathfrak{s}}$: ground_matrix $\rightarrow$ [0, 1] such that $\mu_{\mathfrak{s}}(S_ij) = \mu_{Si}(s_j)$ for all $S_ij \in$ ground_matrix.

The membership function $\mu_{\mathfrak{s}}$ of the sequence $\mathfrak{s}$ in this definition determines to what extent $\mu_{\mathfrak{s}}(S_ij)$ a particular sign S_i of the alphabet is a member of the sequence in its position j. For example, in the triplet UAC, we have $\mu_{\mathrm{UAC}}(\mathrm{A\ in\ 2}) = 1$, whereas $\mu_{\mathrm{UAC}}(\mathrm{A\ in\ 3}) = 0$. Note that the membership function $\mu_{\mathfrak{s}}$ of the sequence is itself defined by a second membership function, i.e. by μ_{Si} that determines to what extent a concrete copy s_j actually occurring in position j of the sequence is a member of the fuzzy sign S_i of the fuzzy alphabet. Regarding our example codon UAC over the RNA alphabet $\langle$U, C, A, G$\rangle$, for instance, we have according to Definition 3:

$$
\begin{array}{lll}
\mu_{\mathfrak{s}}(\mathrm{S}_1 1) = \mu_{S1}(s_1) = 1 & \text{i.e.} & \mu_{UAC}(\mathrm{U\ in\ 1}) = \mu_U(s_1) = 1 \\
\mu_{\mathfrak{s}}(\mathrm{S}_2 1) = \mu_{S2}(s_1) = 0 & & \mu_{UAC}(\mathrm{C\ in\ 1}) = \mu_C(s_1) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_3 1) = \mu_{S3}(s_1) = 0 & & \mu_{UAC}(\mathrm{A\ in\ 1}) = \mu_A(s_1) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_4 1) = \mu_{S4}(s_1) = 0 & & \mu_{UAC}(\mathrm{G\ in\ 1}) = \mu_G(s_1) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_1 2) = \mu_{S1}(s_2) = 0 & & \mu_{UAC}(\mathrm{U\ in\ 2}) = \mu_U(s_2) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_2 2) = \mu_{S2}(s_2) = 0 & & \mu_{UAC}(\mathrm{C\ in\ 2}) = \mu_C(s_2) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_3 2) = \mu_{S3}(s_2) = 1 & & \mu_{UAC}(\mathrm{A\ in\ 2}) = \mu_A(s_2) = 1 \\
\mu_{\mathfrak{s}}(\mathrm{S}_4 2) = \mu_{S4}(s_2) = 0 & & \mu_{UAC}(\mathrm{G\ in\ 2}) = \mu_G(s_2) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_1 3) = \mu_{S1}(s_3) = 0 & & \mu_{UAC}(\mathrm{U\ in\ 3}) = \mu_U(s_3) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_2 3) = \mu_{S2}(s_3) = 1 & & \mu_{UAC}(\mathrm{C\ in\ 3}) = \mu_C(s_3) = 1 \\
\mu_{\mathfrak{s}}(\mathrm{S}_3 3) = \mu_{S3}(s_3) = 0 & & \mu_{UAC}(\mathrm{A\ in\ 3}) = \mu_A(s_3) = 0 \\
\mu_{\mathfrak{s}}(\mathrm{S}_4 3) = \mu_{S4}(s_3) = 0 & & \mu_{UAC}(\mathrm{G\ in\ 3}) = \mu_G(s_3) = 0
\end{array}
$$

Thus, the global membership function $\mu_{\mathfrak{s}}$ of the sequence is defined by the local membership values $\mu_{S1}(s_1), \ldots, \mu_{Sn}(s_m)$ of its signs $s_1, \ldots, s_m$. The outcome of the computation we obtain in this way from ground_matrix is a matrix of values:

$$
\left\{
\begin{array}{c}
\mu_{\mathfrak{s}}(S_1 1), \ldots, \mu_{\mathfrak{s}}(S_n 1) \\
\cdots \\
\cdots \\
\mu_{\mathfrak{s}}(S_1 m), \cdots, \mu_{\mathfrak{s}}(S_n m)
\end{array}
\right\}
$$

Each of the values $\mu_{\mathfrak{s}}(S_i j)$ in the matrix is a membership degree according to Definition 3. Now a new function, *fset*, combines any component $S_i j$ of the ground matrix with the corresponding membership degree $\mu_{\mathfrak{s}}(S_i j)$ and returns a pair: $(S_i j, \mu_{\mathfrak{s}}(S_i j))$. A fuzzy matrix ensues:

Definition 4. *fset(ground_matrix)* = *fuzzy_matrix* such that

$$\text{fuzzy\_matrix} = \left\{ \begin{array}{l} (S_1 1, \mu_{\mathfrak{s}}(S_1 1)), \ldots, (S_n 1, \mu_{\mathfrak{s}}(S_n 1)) \\ \cdots \\ \cdots \\ (S_1 m, \mu_{\mathfrak{s}}(S_1 m)), \ldots, (S_n m, \mu_{\mathfrak{s}}(S_n m)) \end{array} \right\}$$

The fuzzy_matrix in its rows contains the $(m \times n)$-element, ordered *fuzzy set*:

$$\langle (S_1 1, \mu_{\mathfrak{s}}(S_1 1)), \ldots, (S_i j, \mu_{\mathfrak{s}}(S_i j)), \ldots, (S_n m, \mu_{\mathfrak{s}}(S_n m)) \rangle \quad (1)$$

This ordered fuzzy set represents our source sequence $\mathfrak{s}$ over the alphabet $\langle S_1, \ldots, S_n \rangle$. We have thus transformed the sequence to an ordered fuzzy set in two steps:

$$fset(gmatr(\mathfrak{s})) = \langle (S_1 1, \mu_{\mathfrak{s}}(S_1 1)), \ldots, (S_i j, \mu_{\mathfrak{s}}(S_i j)), \ldots, (S_n m, \mu_{\mathfrak{s}}(S_n m)) \rangle.$$

This fuzzy set describes to what extent any sign of the alphabet occurs in a position of the base sequence. For our example triplet UAC over the RNA alphabet ⟨U, C, A, G⟩, for instance, we get:

$$\begin{array}{ll} fset(gmatr(\text{UAC})) = & \langle (\text{U in } 1, 1), (\text{C in } 1, 0), (\text{A in } 1, 0), (\text{G in } 1, 0), \\ & (\text{U in } 2, 0), (\text{C in } 2, 0), (\text{A in } 2, 1), (\text{G in } 2, 0), \\ & (\text{U in } 3, 0), (\text{C in } 3, 1), (\text{A in } 3, 0), (\text{G in } 3, 0) \rangle. \end{array}$$

2.3.3. The genetic fuzzy code

In an ordered fuzzy set $\langle (x_1, a_1), \ldots, (x_m, a_m) \rangle$, the m-tuple $(a_1, \ldots, a_m)$ of its membership degrees is referred to as its fuzzy vector. The *fuzzy vector* of the fuzzy_matrix that we have arrived at above is isolated by the function *fvector* in the following way:

Definition 5. *fvector(fuzzy_matrix)* $= \langle \mu_{\mathfrak{s}}(S_1 1), \ldots, \mu_{\mathfrak{s}}(S_n m) \rangle$

Since the fuzzy_matrix is an $(m \times n)$-matrix, we obtain an $(m \times n)$-dimensional vector of the form:

$$(r_1, \ldots, r_{m \times n})$$

with $r_i \in [0, 1]$ where m is the length of the base sequence $\mathfrak{s} = s_1 \ldots s_m$ and n is the length of the alphabet $\langle S_1, \ldots, S_n \rangle$. It is the fuzzy vector of the ordered fuzzy set Eq. (1) above representing our source sequence. In other words, it is the fuzzy code of the polynucleotide sequence $\mathfrak{s}$:

Definition 6. If $\mathfrak{s}$ is a sequence, then $fcode(\mathfrak{s}) = fvector(fset(gmatr(\mathfrak{s})))$.

Any single-strand polynucleotide has a fuzzy code in this sense. Regarding our earlier example UACUGU, for instance, we have:

$$fcode(\text{UACUGU}) = fvector(fset(gmatr(\text{UACUGU})))$$

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

As already pointed out, the fuzzy code of a polynucleotide base sequence may also have real components between 0 and 1. This is the case, for example, when a base in the sequence is defective and may fit more than one fuzzy class, or when a base is not identifiable with certainty or has not yet been identified. In such cases, probabilistic conjectures may yield fuzzy membership degrees needed to fill in the fuzzy code (see below).

3. The geometry of polynucleotides

Through its fuzzy code, a polynucleotide is representable as a point in a unit hypercube. This unit hypercube whose points are polynucleotides is dubbed a *fuzzy polynucleotide space*. It allows for a geometry of polynucleotides that appears a promising approach in genetic taxonomy and diagnosis. In this section we will introduce this geometry. To this end, some terminology on the unit hypercube representability of polynucleotides may be useful (for details, see [11,13]). The geometry of the unit hypercube we use in the following is due to Bart Kosko [8,9].

3.1. The fuzzy polynucleotide space (FPNS)

Given any finite ground set $\Omega = \{x_1, \ldots, x_n\}$ with $n \geq 1$ members, its fuzzy powerset $F(2^{\Omega})$ forms an n-dimensional unit hypercube such that each member of $F(2^{\Omega})$, a fuzzy set, is a point in the cube [18,8]. This basic finding may be explained in the following way:

The unit interval [0, 1] is a *line* of length 1. A coordinate system consisting of two coordinate axes x and y both of which are unit intervals [0, 1] is a unit *square*, written $[0, 1] \times [0, 1]$, or $[0, 1]^2$ for short. A coordinate system consisting of three coordinate axes x, y, and z all of which are unit intervals [0, 1] is a unit *cube*, written $[0, 1] \times [0, 1] \times [0, 1]$, or $[0, 1]^3$ for short. In general, a coordinate system consisting of n coordinate axes $x_1, \ldots, x_n$ all of which are unit intervals [0, 1] is an

n-dimensional unit cube, called a unit hypercube and written $[0, 1] \times \ldots \times [0, 1]$, or $[0, 1]^n$ for short. Thus, an n-dimensional unit cube is:

the unit line between 0 and 1	if $n = 1$
the unit square	if $n = 2$
the ordinary unit cube	if $n = 3$
the unit hypercube $[0, 1]^n$	if $n \geq 1$.

For our discussion below, it is worth mentioning that a hypercube $[0, 1]^n$ has 2^n corners.

A fuzzy set is a *point* in an n-dimensional unit hypercube. This may be intuitively illustrated as follows.

If $\{(x_1, a_1), \ldots, (x_n, a_n)\}$ is a fuzzy set with $n \geq 1$ members, the ordered n-tuple $(a_1, \ldots, a_n)$ of its membership degrees is referred to as its fuzzy vector where an a_i is the degree of membership of object x_i in that set, for $i \geq 1$.

Let $\Omega = \{x_1, \ldots, x_n\}$ be any ground set. We will hold the spelling of Ω in a constant order of n columns $x_1, x_2, \ldots, x_n$. We can thus use for any fuzzy set $\{(x_1, a_1), \ldots, (x_n, a_n)\} = A \in F(2^\Omega)$ the vector notation and represent it by its n-dimensional fuzzy vector $(a_1, \ldots, a_n)$. For instance, if our ground set Ω is $\{x_1, x_2, x_3\}$, we write:

(a_1, a_2, a_3)	for fuzzy set	$\{(x_1, a_1), (x_2, a_2), (x_3, a_3)\}$

such as:

(1, 1, 1)	for fuzzy set	$\{(x_1, 1), (x_2, 1), (x_3, 1)\}$
(0.2, 0.8, 0.6)	for fuzzy set	$\{(x_1, 0.2), (x_2, 0.8), (x_3, 0.6)\}$
(1, 0, 1)	for fuzzy set	$\{(x_1, 1), (x_2, 0), (x_3, 1)\}$

The ith component a_i in column $i \geq 1$ of such a fuzzy set vector $(a_1, \ldots, a_n)$ represents the membership degree $\mu_A(x_i) = a_i$ of the corresponding object x_i. Our three example sets above are three-dimensional vectors. A membership function μ_A thus defines a fuzzy set A as an n-dimensional vector $A = (\mu_A(x_1), \ldots, \mu_A(x_n)) = (a_1, \ldots, a_n)$ with $a_i \in [0, 1]$. Taking into account that geometrically an n-dimensional vector $(a_1, \ldots, a_n)$ of reals with components in $[0, 1]$ defines:

a point on a line	if $n = 1$
a point in a square	if $n = 2$
a point in a cube	if $n = 3$
a point in a hypercube	if $n \geq 1$

we arrive at the above-mentioned geometrical idea: A fuzzy set $\{(x_1, a_1), \ldots, (x_n, a_n)\}$ as an n-dimensional vector $(a_1, \ldots, a_n)$ with components in $[0, 1]$ is a point in an n-dimensional unit hypercube $[0, 1]^n$. Hence, given any ground set Ω with $n \geq 1$ members, its fuzzy powerset $F(2^\Omega)$ forms an n-dimensional unit hypercube. The n singletons $\{x_i\}$ of its ordinary part 2^Ω are allocated to the coordinates of the cube. Thus, the 2^n members of the ordinary powerset 2^Ω inhabit the 2^n corners of the

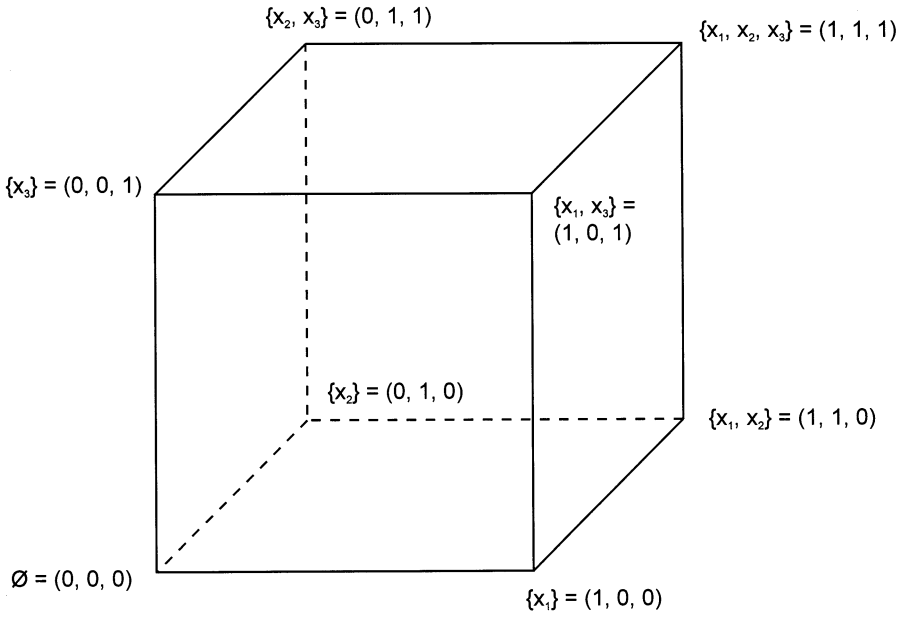

Fig. 2. Since more than three dimensions are not graphically representable, this illustration may be viewed as a proxy for all n-dimensional unit hypercubes $[0, 1]^n$. We have a three-element ground set $\Omega = \{x_1, x_2, x_3\}$. The coordinate axes of the hypercube are allocated to the singletons $\{x_1\}$, $\{x_2\}$, $\{x_3\}$. The $2^n = 8$ vertices of the cube represent the eight fuzzified elements of the ordinary powerset 2^{Ω}. The entire fuzzy powerset $F(2^{\Omega})$ forms the unit cube. The fuzzy set $A = \{(x_1, 0.5), (x_2, 0.4), (x_3, 0.7)\}$ is exemplified as a point in the cube in Fig. 3.

cube, with the empty set $\varnothing$ residing at the cube origin. The rest of the fuzzy powerset $F(2^{\Omega})$ fills in the lattice to produce the solid cube. The cube $[0, 1]^n$ therefore may be termed a fuzzy hypercube. See Figs. 2 and 3.

Once a polynucleotide sequence $\mathfrak{s}$ has been transformed to an ordered fuzzy set $\langle(S_1 1, \mu_{\mathfrak{s}}(S_1 1)), \ldots, (S_i j, \mu_{\mathfrak{s}}(S_i j)), \ldots, (S_m n, \mu_{\mathfrak{s}}(S_m n))\rangle$, it can through its fuzzy code:

$$\langle \mu_{\mathfrak{s}}(S_1 1), \ldots, \mu_{\mathfrak{s}}(S_m n)\rangle$$

be represented as a point in a unit hypercube. A sequence of real length $n \geq 1$ is a point in a $4n$-dimensional hypercube because $m = 4$ is the number of nitrogenous bases in its alphabet. For instance, our example mRNA sequence UACUGU used in Section 2.2 above with its fuzzy code:

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

is a point in a 24-dimensional unit hypercube. An HIV with $\sim 10\,000$ nucleotides is a point in a 40 000-dimensional unit hypercube.

The unit hypercube representation of polynucleotides yields a space that is dubbed a *fuzzy polynucleotide space*, FPNS. The space is attained in the following manner:

Given a base sequence $\mathfrak{s}$ of length $n \geq 1$, the ground set of which it is a fuzzy subset is $\Omega = \langle S_1 1, \ldots, S_4 n \rangle$. This set has $4n$ elements (see Section 2.3.2). An element $S_i j$ says 'nitrogenous base S_i in position j of the sequence'. Allocate now the coordinate axes of a $4n$-dimensional unit hypercube to the elements of the ordered ground set Ω. The 2^{Ω} ordinary members of the fuzzy powerset $F(2^{\Omega})$ reside at the corners of the cube, while the solid cube houses the remainder of $F(2^{\Omega})$.

There are of course two different types of fuzzy polynucleotide spaces, a fuzzy DNA space over the alphabet ⟨T, C, A, G⟩ and a fuzzy RNA space over the alphabet ⟨U, C, A, G⟩. But we will not enter into the subtleties of this differentiation here (see [11]).

3.2. The genetic code is 12-dimensional

The physical space can be, and is, treated as an interpretation of the three-dimensional real space $[0, \propto]^3$ and is therefore considered three-dimensional. By adding the time as a fourth dimension Einstein's four-dimensional universe is obtained as an interpretation of the four-dimensional real space $[0, \propto]^4$. The objects that are dealt with in the former space are three-dimensional *because* they are points of a three-dimensional space. The objects that are dealt with in the latter space are four-dimensional *because* they are points of a four-dimensional space.

By analogy, the genetic code may be viewed as 12-dimensional because a triplet codon XYZ has a $3 \times 4 = 12$-dimensional fuzzy code $(a_1, \ldots, a_{12})$ and is thus a point in the 12-dimensional fuzzy polynucleotide space $[0, 1]^{12}$ as a subspace of the real

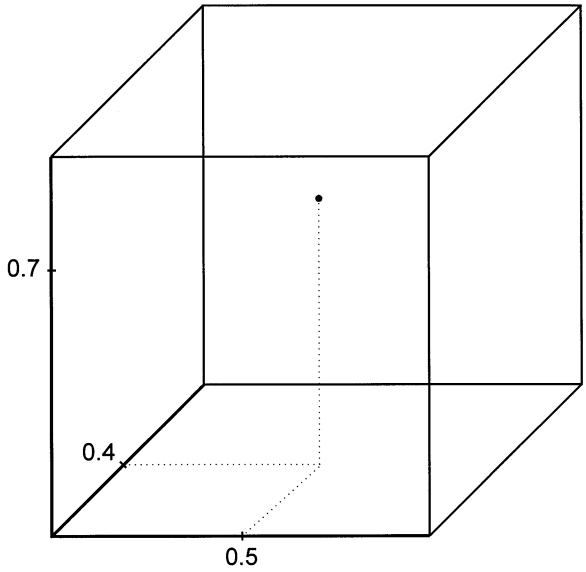

Fig. 3. The same hypercube as in Fig. 2. The dot within the cube is the fuzzy set $A = \{(x_1, 0.5), (x_2, 0.4), (x_3, 0.7)\}$ with its fuzzy vector (0.5, 0.4, 0.7). For details, see [8,12].

space $[0, \infty]^{12}$. Any of the 64 codons of the genetic code is located at one of the $2^{12} = 4096$ corners of this 12-dimensional unit cube. This may be illustrated by a few codons. The amino acids they code for are also listed:

Codon	its fuzzy code	the coded amino acid
UUU	(1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0)	phenylalanine
UCG	(1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1)	serine
CCG	(0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1)	proline
CAU	(0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0)	histidine

A base sequence consisting of only one single base X over the alphabet $\langle$U, C, A, G$\rangle$ is four-dimensional. Any additional base adds four dimensions. Thus, a base sequence $\mathfrak{s} = X_1 \ \ldots \ X_n$ of length n is a $4n$-dimensional object of the fuzzy polynucleotide space. In the next section we will be concerned with the geometry of this space. We should therefore be aware that a base sequence need not necessarily reside at a corner of a hypercube. It can reside within the cube as well when the components of its fuzzy code are not confined to 0 and 1 as above, but also include membership degrees between 0 and 1 such as, e.g.

$$(0.3, 0.4, 0.1, 0.2, 1, 0, 0, 0, 1, 0, 0, 0) \tag{2}$$

This is the vector of a mutant of the above triplet UUU and differs from UUU in that in its position 1 it contains:

U to the extent 0.3
C to the extent 0.4
A to the extent 0.1
G to the extent 0.2.

How is this possible? This four-dimensional possibility reflects states of uncertainty where no sufficient knowledge about the chemical structure of a sequence is available. Probabilistic predictions in experiments of the outcome of replications may be considered as examples. In an experiment of this kind the vector (Eq. (2)) may predict the copy of the segment UUU of a replicating virus. This hypothetical triplet (Eq. (2)) is not located at a corner of the cube. It is a point on one of the cube's sides. As our information about the experiment changes, the vector (Eq. (2)) also changes due to the fluctuating probability distribution. A vector of the form:

$$(r_1, \ldots, r_{12})$$

may thus change into:

$$(r'_1, \ldots, r'_{12})$$

such that $r_i \neq r'_i$. Temporal fluctuations of these vectors represent the trajectory of a point in the fuzzy polynucleotide space $[0, 1]^{12}$. Suppose now that in our experiment regarding the triplet UUU above:

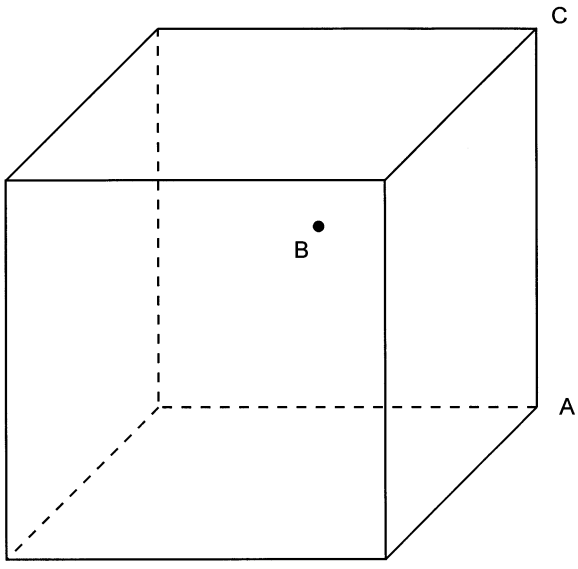

Fig. 4. For illustration purposes a three-dimensional hypercube is used instead of a 12-dimensional one because the latter one is graphically not representable. The initial triplet codon UUU may reside at the corner A. Point B is the predicted state of its mutant. The actually emerging mutant resides at the corner C. What is the distance between A and B and between B and C ? How close to the target C was the prediction of point B ? Questions of this type are dealt with in genetic geometry in Section 3.3.

the initial triplet UUU is the point A
the predicted, hypothetical triplet (Eq. (2)) is the point B
and the actually emerging copy is the point C

of the cube (Fig. 4). What geometrical relationships do exist between these three points? Is it possible to conclude from the distance between the final point C and the hypothesized point B how accurate our prediction has been? We now turn to problems of this type.

3.3. Sequence comparison

The n-dimensional fuzzy polynucleotide space $[0, 1]^n$ we have constructed thus far may be endowed with any metric to become a metric space. In this metric space we will do difference and similarity analyses to compare polynucleotide sequences with one another.[1]

The similarity concept that we need for sequence comparison is built upon the notion of fuzzy set difference. We will introduce this notion for polynucleotides since they have become fuzzy sets, and as such, are points of a polynucleotide space. For details of the concepts and relationships used in this section, see [13,12,8–10].

[1] The inspiration for this idea has come from Manfred Eigen's works [1–5].

3.3.1. The difference between two polynucleotides

The difference between two polynucleotides is reconstructed as a particular geometric distance between two points in the fuzzy polynucleotide hypercube $[0, 1]^n$. Given two polynucleotide sequences such as UACUGU and CACUGU, each of them is located at a particular point of a 24-dimensional unit cube. The less they differ from one another, the closer in the cube they reside. For instance, of the following three sequences:

$\mathfrak{s}_1 = \text{UACUGU}$	codes for:	tyrosine/cysteine
$\mathfrak{s}_2 = \text{CACUGU}$		histidine/cysteine
$\mathfrak{s}_3 = \text{CUCUGU}$		leucine/cysteine

sequence $\mathfrak{s}_2$ in the cube is located closer to sequence $\mathfrak{s}_1$ than sequence $\mathfrak{s}_3$ because there is only one base difference between $\mathfrak{s}_2$ and $\mathfrak{s}_1$, whereas there are two base differences between $\mathfrak{s}_3$ and $\mathfrak{s}_1$. Thus, the difference or dissimilarity between them in the polynucleotide space is reflected by a particular kind of geometric distance that we will develop and measure in this subsection. To this end the polynucleotide space is extended into a metric space.

A *metric space* is defined as a pair $\langle X, d\rangle$ comprising a non-empty set X and a binary function d from $X \times X$ to real numbers such that for all $x, y, z \in X$ we have:

$d(x, y) \geq 0$	*non-negativity*
$d(x, y) = 0$ iff $x = y$	*the identification property*
$d(x, y) = d(y, x)$	*symmetry*
$d(x, y) + d(y, z) \geq d(x, z)$	*the triangle property*

Here, 'iff' stands for 'if and only if'. The function d is called a distance function or a metric over X. For example, let X be the set of all n-dimensional fuzzy codes of polynucleotides with $n \geq 1$. Given two such codes $(a_1, \ldots, a_n) = x$ and $(b_1, \ldots, b_n) = y$ defining the two points x and y in a polynucleotide space $[0, 1]^n$, each element ℓ^p of the Minkowski class of metrics:

$$\ell^p(x, y) = (\Sigma_i |a_i - b_i|^p)^{1/p} \text{ for } 1 \leq i \leq n \text{ and } p \geq 1$$

provides a distance function, denoted by ℓ^p, that renders $\langle X, \ell^p\rangle$ a metric space. For instance, we obtain the Hamming distance if $p = 1$:

$$\ell^1(x, y) = \Sigma_i |a_i - b_i| \text{ for } 1 \leq i \leq n,$$

and the Euclidean distance if $p = 2$:

$$\ell^2(x, y) = (\Sigma_i |a_i - b_i|^2)^{1/2}$$

To give a simple example, the Hamming distance between our two three-dimensional vectors $x = (0.9, 0.2, 0.4)$ and $y = (0.3, 0.5, 1)$ is:

$$\ell^1(x, y) = |0.9 - 0.3| + |0.2 - 0.5| + |0.4 - 1| = 1.5$$

whereas their Euclidean distance is:

$$\ell^2(x, y) = (|0.9 - 0.3|^2 + |0.2 - 0.5|^2 + |0.4 - 1|^2)^{1/2} = (0.81)^{1/2} = 0.9.$$

All Minkowski distances are formally equivalent in that for any pair, $\ell^i(x, y)$ and $\ell^j(x, y)$, there are positive numbers a and b such that $a \cdot \ell^i(x, y) \leq \ell^j(x, y) \leq b \cdot \ell^i(x, y)$. We will use the Hamming distance ℓ^1 because of its simplicity. Thus, our metric space is any n-dimensional polynucleotide space enriched by ℓ^1, that is $\langle [0, 1]^n, \ell^1 \rangle$.

The second notion we need is the count of a fuzzy set A, denoted by c(A). If $A = (\mu_A(x_1), \ldots, \mu_A(x_n))$ is a fuzzy set represented in vector notation, its ordinary size or count, c(A), is simply the sum of its membership values:

Definition 7. $c(A) = \Sigma_i \, \mu_A(x_i)$ for $1 \leq i \leq n$.

From this definition we obtain, in the unit hypercube, the count of a set A as its Hamming distance to the empty set $\varnothing$ at the cube origin, that is $\ell^1(A, \varnothing)$. (See Fig. 5):

Theorem 1. $c(A) = \ell^1(A, \varnothing)$.

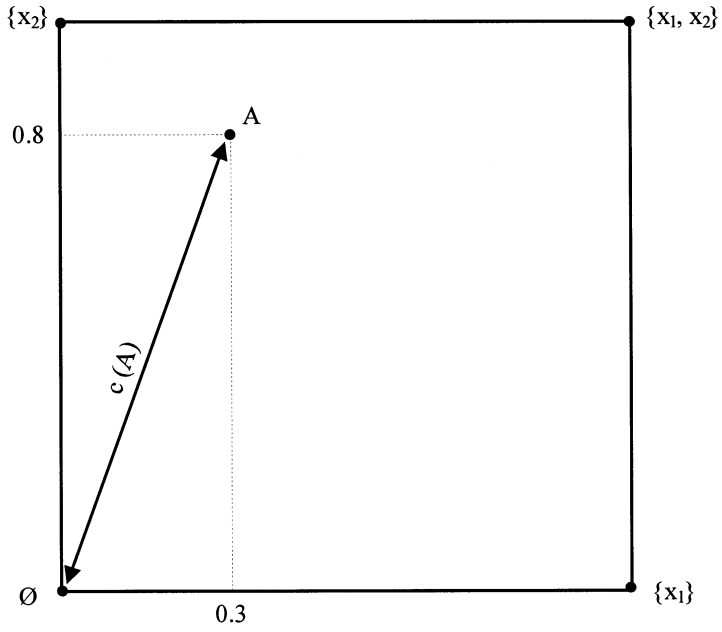

Fig. 5. Geometrical interpretation of the count of a fuzzy set. For simplicity's sake a two-dimensional hypercube is used. Point A in the cube is fuzzy set A = (0.3, 0.8). The count of the set equals the Hamming length of the vector drawn from the origin of the hypercube to the point A. In the present case we have c(A) = 0.3 + 0.8 = 1.1. See Theorem 1.

Proof.

$$\begin{aligned} c(A) &= \Sigma_i \mu_{\mathrm{A}}(x_i) \qquad \text{for } 1 \le i \le n \\ &= \Sigma_i |\mu_{\mathrm{A}}(x_i) - 0| \\ &= \Sigma_i |\mu_{\mathrm{A}}(x_i) - \mu_{\varnothing}(x_i)| \\ &= \ell^1(\mathrm{A}, \varnothing) \end{aligned}$$

The notion of fuzzy set difference we have arrived at is due to Chin-Teng Lin [10]. We will symbolize it by *differ*$(A, B) = r$ and read 'the degree of difference between fuzzy set A and fuzzy set B is r'. It is defined as follows:

Definition 8. $$\textit{differ}(A,B) = \frac{\Sigma_i \max(0, \mu_A(x_i) - \mu_B(x_i)) + \Sigma_i \max(0, \mu_B(x_i) - \mu_A(x_i))}{c(A \cup B)}$$

It denotes the sum of mutual, positive differences between membership degrees:

$$\max(0,\ \mu_{\mathrm{A}}(x_i) - \mu_{\mathrm{B}}(x_i)) + \max(0,\ \mu_{\mathrm{B}}(x_i) - \mu_{\mathrm{A}}(x_i))$$

of all members:

$$\Sigma_i\ \max(0,\ \mu_{\mathrm{A}}(x_i) - \mu_{\mathrm{B}}(x_i)) + \Sigma_i\ \max(0,\ \mu_{\mathrm{B}}(x_i) - \mu_{\mathrm{A}}(x_i))$$

of both sets A and B normalized by their count $c(\mathrm{A} \cup \mathrm{B})$:

$$\Sigma_i\ \max(0,\ \mu_{\mathrm{A}}(x_i) - \mu_{\mathrm{B}}(x_i)) + \Sigma_i\ \max(0,\ \mu_{\mathrm{B}}(x_i) - \mu_{\mathrm{A}}(x_i))/c(\mathrm{A} \cup \mathrm{B})$$

to get the scale [0, 1] for the measure *differ*(A, B). Hence, it is a binary set function that maps the Cartesian product $\mathrm{F}(2^{\Omega}) \times \mathrm{F}(2^{\Omega})$ of the unit hypercube to [0, 1]:

$$\textit{differ}\colon \mathrm{F}(2^{\Omega}) \times \mathrm{F}(2^{\Omega}) \to [0,\ 1].$$

For instance, our three example sequences above:

$\mathfrak{s}_1 = \mathrm{UACUGU}$
$\mathfrak{s}_2 = \mathrm{CACUGU}$
$\mathfrak{s}_3 = \mathrm{CUCUGU}$

with their fuzzy codes:

$\mathfrak{s}_1 = (1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0)$
$\mathfrak{s}_2 = (0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0)$
$\mathfrak{s}_3 = (0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 0,\ 0,\ 0)$

differ from one another to following extents:

differ$(\mathfrak{s}_1,\ \mathfrak{s}_2) = 2/7 = 0.285$
differ$(\mathfrak{s}_1,\ \mathfrak{s}_3) = 4/8 = 0.5.$

A closer look at the numerator of the ratio in Definition 8 reveals that it reflects the Hamming distance between sets A and B. That yields:

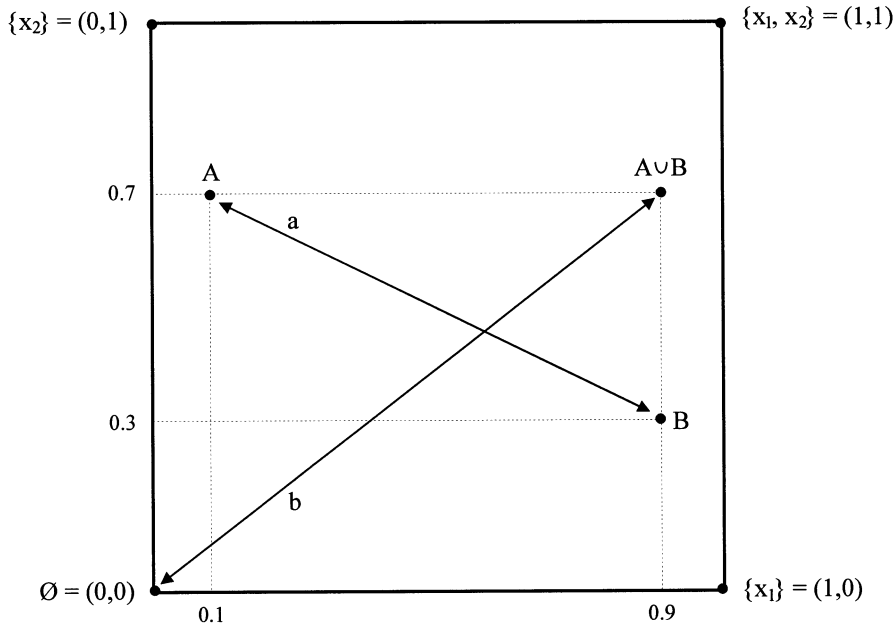

Fig. 6. A simple illustration of the difference relationship in a two-dimensional hypercube. Set A = (0.1, 0.7), set B = (0.9, 0.3). According to Theorem 3, their difference is the Hamming distance *a* divided by the Hamming distance *b*, i.e. differ(A, B) = a/b = 0.75. The greater the distance *a*, the greater the difference between the two sets, and vice versa.

Theorem 2. $differ(A,B) = \dfrac{\ell^1(A, B)}{c(A \cup B)}$

If, due to Theorem 1, the denominator of this fraction is also replaced by the equivalent Hamming distance, we obtain:

Theorem 3. $differ(A, B) = \dfrac{\ell^1(A, B)}{\ell^1(A \cup B, \varnothing)}$

The latter theorem shows that the difference between two polynucleotides *A* and *B* is proportional to their Hamming distance in the cube (see Fig. 6). Since the Hamming distance is a metric, the difference function *differ* turns out to be a metric too. Thus, a fuzzy polynucleotide space $[0, 1]^n$ endowed with *differ*, that is $\langle [0, 1]^n$, differ$\rangle$, is a metric space.

3.3.2. The similarity between two polynucleotides

The less different two entities, the more similar they are. According to this intuitive concept, similarity is the inverse of difference that is reflected by the following definition. The term $similar(A, B)$ therein says 'degree of similarity between A and B'.

Definition 9. $similar(A, B) = 1 - differ(A, B)$.
A theorem that we will not prove here enables convenient computations (see [12]):

Theorem 4.
1. $similar(A, B) = c(A \cap B)/c(A \cup B)$
2. $differ(A, B) = 1 - similar(A, B)$.

Some examples may illustrate the difference and similarity relationships between sequences. Also their Hamming distance is listed to show that this relationship is less informative than similarity and difference. Note that phenomenologically there is only one base difference between two neighbouring sequences in the list. So each sequence is a minimum mutant of the neighbouring ones.

$\mathfrak{s}_1$: AAAGGG	codes for	lysine/glycine
$\mathfrak{s}_2$: *C*AAGGG		glutamine/glycine
$\mathfrak{s}_3$: *CG*AGGG		arginine/glycine
$\mathfrak{s}_4$: *CGU*GGG		arginine/glycine
$\mathfrak{s}_5$: *CGUC*GG		arginine/arginine
$\mathfrak{s}_6$: *CGUCA*G		arginine/glutamine
$\mathfrak{s}_7$: *CGUCAC*		arginine/histidine.

According to Theorem 4, we obtain the following values between sequence $\mathfrak{s}_1$ and the rest:

$similar(\mathfrak{s}_1, \mathfrak{s}_2) = 5/7 = 0.71$	$differ(\mathfrak{s}_1, \mathfrak{s}_2) = 0.28$	$\ell^1(\mathfrak{s}_1, \mathfrak{s}_2) = 2$
$similar(\mathfrak{s}_1, \mathfrak{s}_3) = 4/8 = 0.5$	$differ(\mathfrak{s}_1, \mathfrak{s}_3) = 0.5$	$\ell^1(\mathfrak{s}_1, \mathfrak{s}_3) = 4$
$similar(\mathfrak{s}_1, \mathfrak{s}_4) = 3/9 = 0.33$	$differ(\mathfrak{s}_1, \mathfrak{s}_4) = 0.66$	$\ell^1(\mathfrak{s}_1, \mathfrak{s}_4) = 6$
$similar(\mathfrak{s}_1, \mathfrak{s}_5) = 2/10 = 0.2$	$differ(\mathfrak{s}_1, \mathfrak{s}_5) = 0.8$	$\ell^1(\mathfrak{s}_1, \mathfrak{s}_5) = 8$
$similar(\mathfrak{s}_1, \mathfrak{s}_6) = 1/11 = 0.09$	$differ(\mathfrak{s}_1, \mathfrak{s}_6) = 0.91$	$\ell^1(\mathfrak{s}_1, \mathfrak{s}_6) = 10$
$similar(\mathfrak{s}_1, \mathfrak{s}_7) = 0/12 = 0$	$differ(\mathfrak{s}_1, \mathfrak{s}_7) = 1$	$\ell^1(\mathfrak{s}_1, \mathfrak{s}_7) = 12$

This may be exemplified by the first line of computation:

$$\mathfrak{s}_1 = (0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1)$$
$$\mathfrak{s}_2 = (0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1).$$

Thus, we have:

$$\mathfrak{s}_1 \cap \mathfrak{s}_2 = (0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1)$$
$$\mathfrak{s}_1 \cup \mathfrak{s}_2 = (0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1).$$

$$\begin{aligned} similar(\mathfrak{s}_1, \mathfrak{s}_2) &= c(\mathfrak{s}_1 \cap \mathfrak{s}_2)/c(\mathfrak{s}_1 \cup \mathfrak{s}_2) \\ &= 5/7 \\ &= 0.71 \\ differ(\mathfrak{s}_1, \mathfrak{s}_2) &= 1 - 0.71 = 0.28 \end{aligned}$$

A geometric interpretation of similarity is illustrated in Fig. 7. As it was pointed out in the last section, an n-dimensional fuzzy polynucleotide space $[0, 1]^n$, FPNS, enriched by the difference function *differ* yields a metric space $\langle FPNS, d\rangle$ where d stands for *differ*. We thus have:

$$d\colon FPNS \times FPNS \to [0, 1], \text{ where } d(\mathfrak{s}_i, \mathfrak{s}_j) = differ(\mathfrak{s}_i, \mathfrak{s}_j).$$

The structure $\langle FPNS, \mathcal{O}_d\rangle$ is a topological space if $\mathcal{O}$ is a topology on *FPNS*. Given any sequence $\mathfrak{s}_i \in FPNS$ and any particular degree δ of difference, a ball of radius δ, i.e. a δ-ball, centered at the point $\mathfrak{s}_i$ and denoted by $B_\delta(\mathfrak{s}_i)$, may be defined in the following way:

$$\begin{aligned} B_\delta(\mathfrak{s}_i) &= \{\mathfrak{s}_j \mid differ(\mathfrak{s}_i, \mathfrak{s}_j) < \delta\} && \text{an open } \delta\text{-ball} \\ B_\delta(\mathfrak{s}_i) &= \{\mathfrak{s}_j \mid differ(\mathfrak{s}_i, \mathfrak{s}_j) \leq \delta\} && \text{a closed } \delta\text{-ball.} \end{aligned}$$

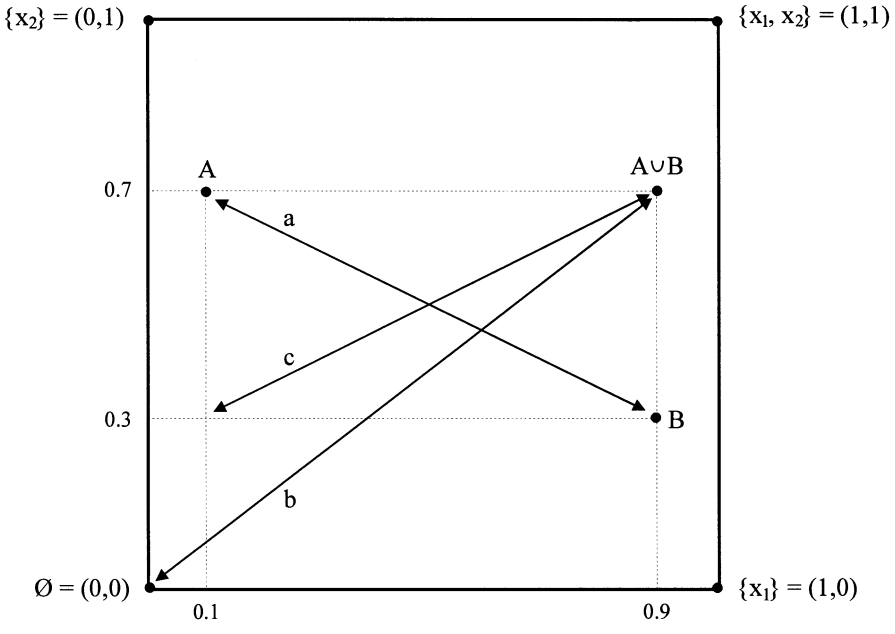

Fig. 7. An amendment to Fig. 6 where it was demonstrated that for fuzzy sets A = (0.1, 0.7) and B = (0.9, 0.3), we have differ(A, B) = a/b = 0.75. Since diagonal c equals diagonal a, we obtain differ(A, B) = c/b. Due to Definition 9, similar(A, B) = 1 − differ(A, B). Hence, similar(A, B) = $1 - c/b = (b - c)/b$. Similarity between nucleic acid sequences is thus a geometric relationship in the fuzzy polynucleotide space.

For instance, we have seen above that the sequences $\mathfrak{s}_1, \ldots, \mathfrak{s}_7$ reside in a particular neighbourhood from one another with a distance of 0.28 to 1 between them. A specific distance, e.g. $\delta = 0.8$ would yield the closed δ-ball:

$$B_{0.8}(\mathfrak{s}_1) = \{\mathfrak{s}_j | \mathit{differ}(\mathfrak{s}_1, \mathfrak{s}_j) \leq 0.8\} = \{\mathfrak{s}_1, \ldots, \mathfrak{s}_5\}$$

In this way, relatives and 'gene families' may be defined and identified as δ-balls centered around a particular source gene or genome. This expressive power of our framework also sheds some light on the notion of *species*. 'Species' has traditionally been defined phenotypically as a particular kind of organism whose members share similar anatomical characteristics. We have learned in the meantime, however, that behind the phenotype of an organism is the genotype as its cause, i.e. the genetic makeup of the organism. But different individuals have distinct genomes residing at distinct points of the fuzzy polynucleotide space. So, the genetic material of the species does not merely occupy a single point of this space. It is distributed over a wide region of the hypercube and looks like the burning lights of a big city viewed from a plane in the night. Each point of the region is a mutant of its generator sequence, and through the arrival of such mutants the whole region moves over the cube like a cloud drifts in the sky. A molecular theory of evolution based on the thermodynamic treatment of this 'molecular cloud' may be found in [1–5]. See also [11].

In closing, we may also define the *identity* between polynucleotides in the following way. The term $\mathit{equal}(A, B)$ in the definition says 'degree of equality between A and B'.

Definition 10.

1. $\mathit{equal}(A, B) = \mathit{similar}(A, B)$
2. A and B are *identical* iff $\mathit{equal}(A, B) = 1$.

Thus, identity is the maximum degree of equality between two fuzzy sets. Two polynucleotide sequences $\mathfrak{s}_1$ and $\mathfrak{s}_2$ as fuzzy sets are identical if $\mathfrak{s}_1$ equals $\mathfrak{s}_2$ to the extent 1.

3.3.3. The entropy of a polynucleotide

The amount of vagueness and indeterminacy a set carries within itself is referred to as its fuzziness or *fuzzy entropy*. It is measured with a fuzzy entropy measure, denoted by *ent*, that maps the hypercube to [0, 1]:

$$\mathit{ent}: F(2^{\Omega}) \rightarrow [0, 1].$$

The definition of this function *ent* is based upon the notions of *nearest* and *farthest* ordinary set to be understood in the following way [6,7,13]:

In a unit hypercube comprising the fuzzy powerset $F(2^{\Omega})$ of a ground set Ω, there is always an ordinary set among $2^{\Omega} \subseteq \mathrm{F}(2^{\Omega})$ which is the nearest one to a fuzzy set

$A \in F(2^{\Omega})$, denoted by A_{near}, and another one that is the farthest one to A, denoted by A_{far}. We will define them informally. A_{near} is a vector $(b_1, \ldots, b_n)$ such that when $A = (a_1, \ldots, a_n)$,

$$\begin{aligned} b_i &= 1 \quad \text{if } a_i > 0.5 \\ &= 0 \quad \text{if } a_i < 0.5 \\ &= 0 \text{ or } 1 \quad \text{if } a_i = 0.5. \end{aligned}$$

And A_{far} is a vector $(b_1, \ldots, b_n)$ such that when $A = (a_1, \ldots, a_n)$,

$$\begin{aligned} b_i &= 0 \quad \text{if } a_1 > 0.5 \\ &= 1 \quad \text{if } a_i < 0.5 \\ &= 1 \text{ or } 0 \quad \text{if } a_i = 0.5. \end{aligned}$$

For example, if $A = (0.2, 0.8, 0.6)$, we have $A_{near} = (0, 1, 1)$ and $A_{far} = (1, 0, 0)$. Let $A = (a_1, \ldots, a_n)$ be any fuzzy set in a unit hypercube. We may recall that ordinary sets reside at the cube's 2^n vertices (see Fig. 2 above). Thus, there is among them a *vertex* nearest to A in the cube called A_{near}, and another one farthest to A referred to as A_{far}. The fuzzy entropy of A is defined as the ratio of the Hamming distance from vertex A_{near} to vertex A_{far}:

Definition 11. $$ent(A) = \frac{\ell^1(A, A_{near})}{\ell^1(A, A_{far})}$$

Fig. 8 provides a geometrical illustration. It shows that at the vertices of the hypercube, $ent(A) = 0$ because at a vertex the numerator of the ratio at the right-hand side of the equation in Definition 11 is 0. Hence, there is no fuzzy entropy at a vertex. This reflects the fact that the inhabitants of the cube vertices are members of the classical set 2^{Ω}. Any component of a set membership vector $(a_1, \ldots, a_n) \in 2^{\Omega}$ at a vertex is either 1 or 0 that amounts to the nonfuzzy information that an object x_i definitely is, or is not, a member of the set. By contrast, if the fuzzy set $A = (a_1, \ldots, a_n)$ is the hypercube midpoint, we get according to Definition 11 $ent(A) = 1$ because A at the midpoint is equidistant from all 2^n vertices. For details, see [13].

The opposite of fuzzy entropy is the *clarity* of a set. The clarity of a set A, denoted by *clar(A)*, is the additive inverse of its entropy:

Definition 12. $clar(A) = 1 - ent(A)$.

We have therefore $clar(A) = 1$ at all vertices, whereas $clar(A) = 0$ at the center of the hypercube.

From these conceptual preliminaries, it follows that a real polynucleotide such as UACUGU has an entropy of 0 and hence, a clarity of 1. For the fuzzy code of such

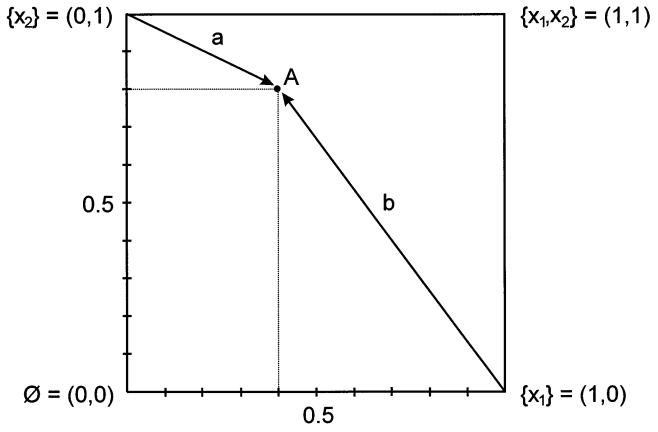

Fig. 8. Illustration of fuzzy entropy in a two-dimensional hypercube. The farthest vertex A_{far} resides opposite the long diagonal from the nearest vertex A_{near}. We have $ent(A) = a/b$ where $a = \ell^1(A, A_{near})$ and $b = \ell^1(A, A_{far})$. Hence, $ent(A) = 0$ at a vertex and $ent(A) = 1$ at the center of the hypercube. Fuzzy entropy smoothly increases as a set point moves from any vertex to the midpoint of the hypercube and thus its distance to its complement decreases. In the present example where set $A = (0.4, 0.8)$, we have $ent(A) = (|0.4 - 0| + |0.8 - 1|)/(|0.4 - 1| + |0.8 - 0|) = 0.6/1.4 = 0.428$.

a polynucleotide is a bit sequence with components from the bivalent set {0, 1}. The base sequence UACUGU is thus an element of the ordinary powerset 2^{Ω} and resides at a vertex of the cube.

4. Conclusion

We have transformed polynucleotide chains to ordered fuzzy sets. The ordered membership vector of such a fuzzy set, termed its fuzzy code, represents a point in an n-dimensional unit hypercube. A polynucleotide thus becomes a unique point in the hypercube. We have therefore dubbed this cube a *fuzzy polynucleotide space*. The geometry, topology and logic that can be done in this space render polynucleotides directly amenable to fuzzy theory. We have demonstrated this approach by difference, similarity and entropy analyses that may be useful in polynucleotide sequence comparison, and thus in genetic taxonomy and diagnosis in the widest sense. Our approach is not confined to nucleic acids, however. It is a general framework for all polymers [15].[2]

Acknowledgements

I thank my son Manuel for drawing the figures for this paper.

[2] The method of fuzzy decoding described in this paper is part of a patent held by the author at the German Patent Office.

Appendix A

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 inherit from their parents. Some other viruses bear RNA as their genetic material.

DNA is present in every cell of an organism. Encoded in its chemical structure is the information that programs all the cell's, and thus all the organism's, activities. However, it is not directly involved in running the operations of the cell and of the organism. DNA is responsible for the structure of the proteins, including enzymes, made by the cell. It directs the synthesis of a type of RNA, called messenger RNA (= mRNA). The mRNA then interacts with the protein-synthesizing machinery of the cell to direct the production of a protein. This production process governs the life and death affairs of cells and organisms because intracellular enzymes — as specific proteins — are responsible for the synthesis and breakdown of practically all the chemicals in a cell.

Proteins are polymers. A polymer is a large macromolecule consisting of many identical or similar building blocks, called its monomers, that are linked by bonds to form a chain. The monomers of a protein are amino acids such as alanine, glycine, serine, etc. Diverse though proteins may be, they are all polymers of the same set of 20 amino acids. The sequence of the amino acids in a protein chain determines the role the protein plays in metabolism. For two protein chains such as serine–alanine–glycine- and serine–glycine–alanine- are distinct. The chemical structure of the mRNA that produces a protein is responsible for its specific amino acid sequence, and thus the responsibility is due ultimately to the chemical structure of the DNA in the well-known double helix of a cell which produces that particular mRNA.

This latter, inner structure of DNA and RNA itself reflects the sequence of their building blocks in their molecular chain. For DNA and RNA are also linear polymers. Their monomeric units are called nucleotides. A large number of nucleotides are linearly linked by bonds to form a *polynucleotide*, a chain of DNA or RNA. For example, the tiny RNA virus HIV consists of $\sim 10\,000$ nucleotide monomers. Any of the 46 human chromosomes in a human cell nucleus is composed of ~ 65 million of them.

A nucleotide is itself composed of three smaller molecular building blocks: a five-carbon sugar, a phosphate group, and a nitrogenous base (see Fig. 9 (a)). 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

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 it and is referred to as its *base sequence* (see Fig. 9(b)). Whereas a particular polynucleotide may have the base sequence GUAUACUGU…, another one may have the base sequence GTTTACACT…

In a cell's chain of command, instructions for protein synthesis flow from DNA to 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 as a sequence of base triplets XYZ, analogous to three-letter code words. An mRNA base triplet XYZ is therefore called a *codon*. A codon XYZ along an mRNA sequence specifies which one of the 20 amino acids will be incorporated at the corresponding position 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-…

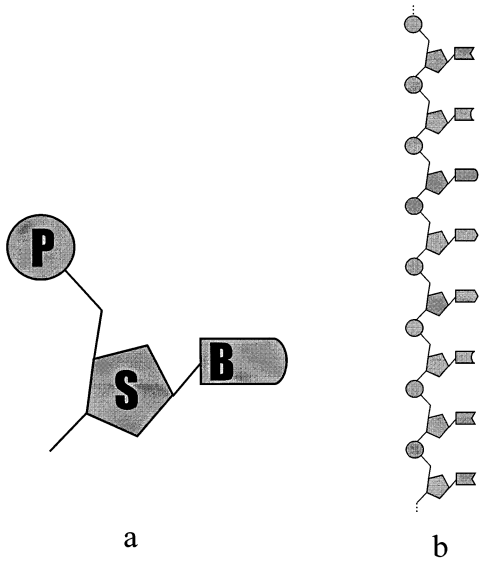

Fig. 9. A nucleotide monomer (a) and a single-strand polynucleotide (b). P, phosphate group; S, sugar; B, nitrogenous base. In DNA, the sugar is deoxyribose, whereas the sugar of RNA is ribose. A DNA chain is usually double-stranded. When synthesizing mRNA and as a cell prepares to divide, the two strands are separated. However, we are not concerned with these details here.

Due to these bio-informational facts, the focus of our concern 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.

The second term we need in this paper is the notion of an 'ordered fuzzy set'. We distinguish between ordinary sets such as $\{x_1, x_2, x_3, \ldots, \}$ and fuzzy sets. In an ordinary set an object either is definitely a member of the set or it is definitely not a member of that set. By contrast, a *fuzzy* set is a collection of objects with grades of membership in that collection [16,17]. Thus, a fuzzy set does not have clear-cut boundaries between members and non-members as an ordinary set does. For example, the set of young people is such a fuzzy set. A person x may be young to a particular extent, whereas another person y may be young to a lesser degree than x is. Thus, both persons are to different degrees members of the same set of young people. The membership degrees of the set smoothly decrease in the direction of zero membership, i.e. non-membership. There is no dividing line between this set and the set of non-young people. Each of the following terms also denotes a fuzzy set: tree, bush, big orange, much larger than 5, healthy, ill, diseased.

Considering a set A as a fuzzy set means that an object x is to some degree a member of that set. Let us express this membership degree by $\mu(x, \mathrm{A})$, to be read as 'the degree of membership of x in set A', conveniently abbreviated to $\mu_{\mathrm{A}}(x)$. The symbol μ_{A} is referred to as the membership function of set A which assigns to an object x its membership degree $\mu_{\mathrm{A}}(x)$.

The membership degree $\mu_{\mathrm{A}}(x)$ of object x in set A is supposed to be a real number in the unit interval [0, 1]. Thus, the expression '$\mu_{\mathrm{young}}(\mathrm{David}) = 0.6$' says that David is to the extent 0.6 a member of the set of young people, or equivalently, that he is young to the extent 0.6. Let the expression '$f{:}X \rightarrow Y$' indicate that f is a function from set X to set Y. A fuzzy set may be defined in the following way. 'Iff' stands for 'if and only if'.

Definition 13. Given any set Ω, A is a *fuzzy subset* of Ω if there is a function μ_{A} such that

1. μ_{A}: $\Omega \rightarrow [0,1]$
2. $\mathrm{A} = \{(x, \mu_{\mathrm{A}}(x)) | x \in \Omega\}$, i.e. A is the set of all pairs $(x, \mu_{\mathrm{A}}(x))$ such that x is a member of Ω and $\mu_{\mathrm{A}}(x)$ is the degree of its membership in A.

Definition 14. *A* is a *fuzzy set*, also called a fuzzy set *in* or *over* Ω, if A is a fuzzy subset of Ω.

In the present context, set Ω is referred to as the *ground set* (instead of 'base set' to prevent equivocation). For example, given the ground set $\{a, b, c, d\}$ of four doctors and a function $\mu_{\text{proficient-doctor}}$ that assigns to any $x \in \{a, b, c, d\}$ the degree of her proficiency, then the following set is a fuzzy subset of our ground set, and thus a fuzzy set: Proficient doctor $= \{(a, 0), (b, 0.4), (c, 0.8), (d, 1)\}$.

Given any ground set $\Omega = \{x_1, x_2, x_3, \ldots, \}$, the set of all of its *ordinary* subsets is known as its powerset. It is conveniently denoted by 2^{Ω} because every n-element set has a powerset of 2^n elements.

On the other side, the set of all *fuzzy* subsets of a ground set Ω is referred to as its *fuzzy powerset* and is denoted by $F(2^{\Omega})$. This powerset is uncountably infinite because every ground set may be mapped to [0, 1] in infinitely different ways to generate infinitely many fuzzy sets. Amongst the elements of the fuzzy powerset $F(2^{\Omega})$ we distinguish five particular ones that will be of special interest below:

The ground set $\Omega = \{x_1, x_2, x_3, \ldots, \}$ is itself the fuzzy set $\{(x_1, 1), (x_2, 1), (x_3, 1), \ldots\}$. The empty fuzzy set is $\{(x_1, 0), (x_2, 0), (x_3, 0), \ldots\}$, denoted by $\varnothing$. The negation of any fuzzy set A, called its *complement* and denoted by A^c or *Not* A, is a fuzzy set that is defined by the following membership function μ_{Ac}:

$$\mu_{Ac}(x) = 1 - \mu_A(x),$$

i.e. $A^c = \{(x, \mu_{Ac}(x)) | x \in \Omega \text{ and } \mu_{Ac}(x) = 1 - \mu_A(x)\}$. For instance, the complement of fuzzy set $A = \{(a, 0), (b, 0.4), (c, 0.8), (d, 1)\}$ is $A^c = \{(a, 1), (b, 0.6), (c, 0.2), (d, 0)\}$.[3]

The *intersection* of two fuzzy sets A and B, denoted by $A \cap B$, is a fuzzy set defined by the following membership function $\mu_{A \cap B}$:

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)).$$

That is: $A \cap B = \{(x, \mu_{A \cap B}(x)) | x \in \Omega \text{ and } \mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))\}$. Their *union*, denoted by $A \cup B$, is a fuzzy set defined by the following membership function $\mu_{A \cup B}$:

$$\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x)),$$

i.e. $A \cup B = \{(x, \mu_{A \cup B}(x)) | x \in \Omega \text{ and } \mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))\}$. For instance, given two fuzzy sets $A = \{(a, 0), (b, 0.4), (c, 0.8), (d, 1)\}$ and $B = \{(a, 0.9), (b, 0.5), (c, 0.3), (d, 0.7)\}$, we have $A \cap B = \{(a, 0), (b, 0.4), (c, 0.3), (d, 0.7)\}$ and $A \cup B = \{(a, 0.9), (b, 0.5), (c, 0.8), (d, 1)\}$.

References

[1] Eigen M. Virus-Quasispezies oder die Büchse der Pandora. Spektrum der Wissenschaft, Dezember 1992, pp. 42–55.
[2] Eigen M, Biebricher CK. Sequence space and quasispecies distribution. In: Domingo E, Holland JJ, Ahlquist P, editors. RNA Genetics, vol. III. Variability of RNA Genomes. Boca Raton, FL: CRC Press, 1988, pp. 211–245.
[3] Eigen M, McCaskill J, Schuster P. The molecular quasi-species. J Phys Chem 1988;92:6881–91.
[4] Eigen M, McCaskill J, Schuster P. The molecular quasi-species. Adv Chem Phys 1989;75:149–263.
[5] Eigen M, Winkler-Oswatitsch R. Steps towards Life. A Perspective on Evolution. Oxford: Oxford University Press, 1996.

[3] The two functions *min* and *max* used below are the usual minimum and maximum operators. Of two numbers m and n, the smaller one is called $min(m, n)$ and the larger one is called $max(m, n)$. For example, $\min(5,3) = 3$ and $\max(5,3) = 5$.

[6] Kaufmann A. Introduction to the Theory of Fuzzy Subsets, vol. I. Fundamental Theoretical Elements. New York: Academic Press, 1975.
[7] Kosko B. Fuzzy entropy and conditioning. Inform Sci 1986;40:165–74.
[8] Kosko B. Neural Networks and Fuzzy Systems. A Dynamical Systems Approach to Machine Intelligence. Englewood Cliffs, NJ: Prentice Hall, 1992.
[9] Kosko B. Fuzzy Engineering. Upper Saddle River, NJ: Prentice Hall, 1997.
[10] Lin CT. Adaptive subsethood for radial basis fuzzy systems. In: Kosko, B, editor. Fuzzy Engineering. Upper Saddle River, NJ: Prentice Hall, 1997, pp. 429–464 [chapter 13].
[11] Sadegh-Zadeh K. The fuzzy logic of the genome. Manuscript (in German), 1998.
[12] Sadegh-Zadeh K. Introduction to Fuzzy Theory. Manuscript (in German), 1998. Submitted.
[13] Sadegh-Zadeh K. Advances in fuzzy theory. Artif Intell Med 1999;15:309–23.
[14] Sadegh-Zadeh K. When Man Forgot Thinking: The Emergence of Machina Sapiens. Manuscript (in German), 1986–1999. Submitted.
[15] Sadegh-Zadeh K. The fuzzy polymer space (in preparation).
[16] Zadeh LA. Fuzzy sets. Inf Control 1965;8:338–53.
[17] Zadeh LA. Fuzzy sets and systems. In: J. Fox, editor. System Theory. Brooklyn, New York: Polytechnic Press, 1965, pp. 29–39.
[18] Zadeh LA. Towards a theory of fuzzy systems. In: Kalman RE, DeClairis RN, editors. Aspects of Networks and Systems Theory. New York: Holt, Reinhart & Winston, 1971, pp. 469–490.