A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets
D.N. Georgiou, T.E. Karakasidis, Juan J. Nieto, A. Torres

To cite this version:

HAL Id: hal-00627147
https://hal.archives-ouvertes.fr/hal-00627147
Submitted on 28 Sep 2011
A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets

D.N. Georgiou, T.E. Karakasidis, Juan J. Nieto, A. Torres

PII: S0022-5193(10)00419-4
DOI: doi:10.1016/j.jtbi.2010.08.010
Reference: YJTBI6114

To appear in: Journal of Theoretical Biology

Received date: 21 January 2010
Revised date: 22 July 2010
Accepted date: 6 August 2010

Cite this article as: D.N. Georgiou, T.E. Karakasidis, Juan J. Nieto and A. Torres, A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets, Journal of Theoretical Biology, doi:10.1016/j.jtbi.2010.08.010

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting galley proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.


A study of entropy/clarity of genetic sequences using Metric Spaces and Fuzzy Sets

D. N. Georgiou\textsuperscript{a}, T. E. Karakasidis\textsuperscript{b}, Juan J. Nieto\textsuperscript{c},
A. Torres\textsuperscript{d}

\textsuperscript{a}University of Patras, Department of Mathematics, 265 00 Patras, Greece
\textsuperscript{b}University of Thessaly, Department of Civil Engineering, 383 34 Volos, Greece,
\textsuperscript{c}Departamento de Análisis Matemático, Facultad de Matemáticas, Universidad de
Santiago de Compostela, 15782 Spain
\textsuperscript{d}Departamento de Psiquiatría Radiología y Salud Pública, Facultad de Medicina,
Universidad de Santiago de Compostela, 15782 Spain

Abstract

The study of genetic sequences is of great importance in biology and medicine. Sequence analysis and taxonomy are two major fields of application of bioinformatics. In the present paper we extend the notion of entropy and clarity to the use of different metrics and apply them in the case of the Fuzzy Polynucleotide Space (FPS). Applications of these notions on selected polynucleotides and complete genomes both in the $I^{12\times k}$ space, but also using their representation in FPS are presented. Our results show that the values of fuzzy entropy/clarity are indicative of the degree of complexity necessary for the description of the polynucleotides in the FPS, although in the latter case the interpretation is slightly different than in the case of the $I^{12\times k}$ hypercube. Fuzzy entropy/clarity along with the use of appropriate metrics can contribute to sequence analysis and taxonomy.

Key words: DNA, RNA, Polynucleotides, Fuzzy sets, Metric spaces, Entropy, Clarity.
1991 MSC: 03E72, 92B05, 54E35, 92C05

Email addresses: georgiou@math.upatras.gr (D. N. Georgiou), thkarak@uth.gr
(T. E. Karakasidis), annieto@usc.es (Juan J. Nieto), mrتورس@usc.es (A. Torres).

Preprint submitted to Elsevier Preprint 22 July 2010
1 Introduction

Bioinformatics is a relatively new discipline (see Jamshidi N. et al. (2001), Morgenstern B. (2002), Paun Gh. et al. (1998), Percus J. (2002) and Tang B. (2000)) where Mathematics play an important role in the analysis of genetic sequences. The genetic material of living organisms consist of nucleic acids DNA and RNA. The analysis of the genetic material is of great importance for diagnosis and taxonomy reasons. In this course there are two basic strategies that are commonly used: a) sequence analysis, i.e. determination of the building blocks of a nucleic acid (nucleotides) and their order in the molecular chain, and b) sequence comparison used to identify the degree of difference/similarity between polynucleotides, e.g. in order to identify similarity with known viruses.

DNA and RNA are made of triplets \( XYZ \) of codons each of them having the possibility to be one of four nucleotides \( \{ U, C, A, G \} \) in the case of DNA and \( \{ T, C, A, G \} \) in the case of RNA (A=Adenine, C=Cytosine, G=Guanine, T=Thymine, U=Uracil). Sadegh-Zadeh (see Sadegh-Zadeh K. (2000)) showed that the genetic code can be represented in a 12-dimensional space because a triplet codon \( XYZ \) has a \( 3 \times 4 = 12 \) dimensional fuzzy code \( (a_1, ... , a_{12}) \) and it is a point in the 12-dimensional fuzzy polynucleotide space \([0,1]^{12}\) as a subspace of the real space \([0,\infty]^{12}\). Sadegh-Zadeh (see Sadegh-Zadeh K. (2000)) introduced the Fuzzy Polynucleotide Space (FPS) based on the principle of the fuzzy hypercube Kosko B. (1992). In this notation a polynucleotide consisting of a sequence of \( k \) triplets \( XYZ \) is a point in a \( I^{12 \times k} \) space. However, Torres and Nieto (see Torres A. et al. (2003)) mapped a polynucleotide on a \( I^{12} \) space by considering the frequencies of the nucleotides at the three base sites of a codon in the coding sequence. In that work using a metric motivated by publications of Lin Lin C.T. (1997) and Sadegh-Zadeh (see Sadegh-Zadeh K. (2000)), they calculated distances between nucleotides. They also applied their algorithm for the comparison of complete genomes (for example \textit{M.tuberculosis} and \textit{E.coli}). Further work has been recently performed using the idea of Nieto et al. (see Nieto J.J. et al. (2006)) in which the influence of several metrics have been examined. The advantages of this methodology are:

a) one can compare polynucleotides of very big length in a very efficient computationally way and

b) one can apply the algorithm in order to compare polynucleotides of different length as it is the case for genomes of different organisms.

We point that metrics play an important role on computational biology. Different metrics have been used to study secondary structures (see V. Moulton et al. (2000)) or biopolymer contact structures (see M. Liabres et al. (2004)).
It is very important to be in a position to determine how close two genetic sequences are since there are many important biological and medical implications (see DasGupta B. et al. (1998), Foster M. et al. (1999), Gusev V. D. (1999), Jiang T. et al. (2002), Liben-Nowell D. (2001) and Li M. et al. (2001)). The biological distance among the 20 amino acids can be calculated according to their classification results. Since the concept of pseudo amino acid composition was proposed by Chou (Chou K. C. (2001)), many efforts have been made trying to use various quantities to represent the 20 native amino acids in order to better reflect the sequence-order effects through the vehicle of pseudo amino acid composition (PseAA), along with work in order to choose effective properties for such procedures (Trinquere and Sanejouand (1998)). In an earlier paper (Chou K. C. (2000)), the physicochemical distance among the 20 amino acids (Schneider G., Wrede P., (1994)) was adopted to define PseAA. Subsequently, some investigators used complexity measure factor (Xiao et al., (2005)), some used the values derived from the cellular automata (Xiao et al., (2005b), Xiao et al., (2005c), Xiao et al., (2006), Xiao et al., (2006b)), some used hydrophobic and/or hydrophilic values (Chou (2005), Feng (2002), Wang et al. (2006), Wang et al. (2004); Gao et al. (2005), Chen et al. (2006), and some were through Fourier transform (Liu et al. (2005), Perez-Montoto et al. (2009)), as well as through cellular automaton approach (Xiao et al., (2009b)). The pseudo amino acid composition was originally introduced to improve the prediction quality for protein subcellular localization and membrane protein type (Chou K. C. (2001)), as well as for enzyme functional class (Chou (2005)). Work using pseudo amino acid composition has also been performed (?, Xiao et al., (2008b), Xiao et al., (2009a)). The pseudo amino acid composition can be used to represent a protein sequence with a discrete model yet without completely losing its sequence-order information (Chou and Shen (2007a)), and hence is particularly useful for analyzing a large amount of complicated protein sequences by means of the taxonomic approach. Actually, it has been widely used to study various protein attributes, such as protein structural class (Chen et al. (2006a), Chen et al. (2006b), Lin and Li (2007a), Ding et al. (2007), Gu and Chen (2009)), protein subcellular localization (Chou and Shen (2008), Chou and Shen (2007a), Shen and Chou (2007a), Chou and Shen (2007b)), protein subnuclear localization (Shen and Chou (2005), Mundra et al. (2007)), protein submitochondria localization (Du and Li (2006)), protein oligomer type (Chou and Cai (2003)), conotoxin superfamily classification (Mondal (2006), Lin and Li (2007b)) membrane protein type (Lin et al. (2005), Shen and Chou (2005), Wang et al. (2006), Shen et al. (2006), Chou and Shen (2007b)) apoptosis protein subcellular localization (Chen and Li (2007a), Chen and Li (2007b)) enzyme functional classification (Chou K. C., (2005), Chou and Cai (2004), Zhou et al. (2007), Shen and Chou (2007b)) protein fold pattern (Shen and Chou (2006)), and signal peptide ((Chou and Shen (2007c), Shen and Chou (2007c)). Recent research works on the extension of these kind of parameters in the form of Markov Chain invariants of 2D graph
or networks representation of aminoacid, DNA, and RNA sequences to codify pseudo-aminoacid and pseudo-nucleotide bases composition (Aguero-Chapin et al. (2008), Gonzalez-Diaz et al. (2007a), Gonzalez-Diaz et al. (2007b), Aguero-Chapin et al. (2006), Vilar et al. (2009)) as well as more complex work such as Xiao et al, (20009c), Xiao et al, (2010)). The reader can also consult some recent reviews which made a discussion of many of these previous results (Gonzalez-Diaz et al. (2008), Chou (2009), Lin et al. (2009)).

In the present paper we present some new results concerning the notions of entropy and clarity of a nucleotide that can be used in order to estimate the fuzziness of a polynucleotide. We compare the results with that obtained in Sadegh-Zadeh K. (2000). We note that it is possible to compare sequences using a minimum entropy principle (Sadovsky M.G. (2003)). More precisely we focus on the use of different metrics in the calculation of the entropy and clarity of a polynucleotide in conjunction with the use of FPS which can be used in order to reduce the information necessary for the representation of large polynucleotides.

The structure of the paper is as follows. In section 2 we present the notion of the Fuzzy Polynucleotide Space (FPS) and the entropy concept and give some applications on polynucleotides and selected genomes. We compare some of the results using our entropy definitions with results obtained in Giulia Menconi (2005) where the notion of computable complexity of several complete genomes is analyzed and compared with the classical entropy results. In section 3, clarity of a polynucleotide is considered and results on several polynucleotides are presented. Finally in section 4 the conclusions of the present work are summarized.

2 Entropy and fuzzy polynucleotide space

2a) Fuzzy sets and fuzzy hypercube

Let $X$ be a set. $A$ is a fuzzy subset of $X$ if there is a function $\mu_A$ such that

1) $\mu_A : X \rightarrow [0,1].$

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

In what follows if $X = \{x_1, x_2, ..., x_n\}$ and

$$A = \{(x_1, \mu_A(x_1)), ..., (x_n, \mu_A(x_n))\},$$
then we write

\[ A = (\mu_A(x_1), ..., \mu_A(x_n)). \]

Let \( A \) and \( B \) two fuzzy sets of a set \( X \).

Then by \( A \land B \) we denote the fuzzy set for which the membership function

\[ \mu_{A \land B} : X \rightarrow [0, 1] \]

is defined as following

\[ \mu_{A \land B}(x) = \min\{\mu_A(x), \mu_B(x)\}, \]

for every \( x \in X \).

Also by \( A \lor B \) we denote the fuzzy set for which the membership function

\[ \mu_{A \lor B} : X \rightarrow [0, 1] \]

is defined as following

\[ \mu_{A \lor B}(x) = \max\{\mu_A(x), \mu_B(x)\}, \]

for every \( x \in X \).

For \( A \) a fuzzy set, the fuzzy complement \( A^c \) is defined by \( A^c(x) = 1 - A(x) \), \( x \in X \).

Kosko Kosko B. (1992) introduced a geometrical interpretation of fuzzy sets as points in a hypercube. Indeed, for a given set \( X = \{x_1, x_2, ..., x_n\} \), the set of all fuzzy subsets (of \( X \)) is precisely the unit hypercube

\[ I^n = [0, 1]^n, \]

since any fuzzy subset \( A \) determines a point \( P \in I^n \) given by

\[ P = (\mu_A(x_1), ..., \mu_A(x_n)). \]

Reciprocally, any point \( P = (a_1, ..., a_n) \in I^n \) generates a fuzzy subset \( A \) of \( X \) defined by the map \( \mu_A : X \rightarrow [0, 1] \) such that \( \mu_A(x_i) = a_i, i = 1, 2, ..., n. \)

Nonfuzzy or crisp subsets of \( X = \{x_1, ..., x_n\} \) are given by mappings

\[ \mu : X \rightarrow \{0, 1\} \]

from the set \( X \) into the set \( \{0, 1\} \) and they are located at the \( 2^n \) corners of the \( n \)-dimensional unit hypercube \( I^n \). So, the ground set \( X = \{x_1, ..., x_n\} \) is itself the fuzzy set \( (1, 1, ..., 1) \in I^n \). Also, the empty fuzzy set is the fuzzy set \( (0, 0, ..., 0) \in I^n \), denoted by \( \emptyset \).

Hypercubical calculus is developed in Zaus M. (1999), and some applications of the fuzzy unit hypercube are given in Nieto J.J. et al. (2003), Sadegh-Zadeh K. (1999) and Hegelson C.M. et al. (1998). In this context a codon
corresponds to a corner of the 12-dimensional unit hypercube $I^{12}$. Any element of $I^{12}$ may be viewed as a fuzzy codon.

DNA and RNA can be treated as a language written using an alphabet of strings. The role of strings is played by several chemical compounds. In fact the alphabet for DNA is $\{T, C, A, G\}$ while for RNA $\{U, C, A, G\}$ where A,C,G,T and U stand for Adenine, Cytosine, Guanine, Thymine and Uracil respectively. In this context in the case of RNA alphabet if U is the first letter of this alphabet one codes it as $(1,0,0,0)$ because the first letter U is present, 0 since the second letter does not appear, 0 since the third letter is not present and 0 since the fourth letter G does not appear. In a similar way C is represented as $(0,1,0,0)$, A as $(0,0,1,0)$ and G as $(0,0,0,1)$. So if we have a nucleotide described by the codon UCG (serine) this would be written in the $I^{12}$ hypercube as

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

There are cases where the exact chemical structure of the sequence is not known for the complete sequence. In this case some components of its fuzzy code being neither 0 or 1 but a value in the interval $(0, 1)$ and are sequences not necessarily at a corner of the hypercube. If for example we have a codon

$$(0.3, 0.4, 0.2, 0.1, 0, 0, 1, 0, 1, 0, 0, 0)$$

This stands for XAU. The first letter X is unknown and corresponds to: U to extent 0.3, C to extent 0.4, A to extent 0.2 and G to extend 0.1

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

$$s_1 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0)$$

However if one considers the frequencies of the nucleotides of the alphabet at the three base sites of a codon in the coding sequence it may be viewed as a point in the hypercube $I^{12}$.

**Table 1a.** Number of nucleotides at the three base sites of a codon in the $s_1$ sequence.
<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First base</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Second base</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Third base</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

Table 1b. Fractions of nucleotides at the three base sites of a codon in the coding sequence of $s_1$.

<table>
<thead>
<tr>
<th></th>
<th>U</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>First base</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Second base</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>Third base</td>
<td>0.5</td>
<td>0.5</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Taking into account the number of nucleotides at the three base sites of a codon in the $s_1$ sequence (see Table 1a) as well as the fractions of nucleotides at the three base sites of a codon in the coding sequence of $s_1$ (see Table 1b) sequence $s_1$ can be written in the $I^{12}$ space as

$$f(s_1) = (1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0).$$

In the case of complete genome the frequencies of the nucleotides at the three base sites of a codon in the codon sequence are considered. This can be viewed as point in the hypercube $I^{12}$.

This idea has been applied to the genomes of *M.tuberculosis*, *E.coli*, and *A.aeolicus* to obtain their fuzzy set of frequencies and calculate their corresponding distance in the Fuzzy Polynucleotide Space (FPS) (see Torres A. et al. (2003) and Nieto J.J. et al. (2006)).

When dealing with genetic sequences it is of interest:

i) to be able to describe how different two sequences are. For this reason the notion of distance is used (see Nieto J.J. et al. (2003), Nieto J.J. et al. (2003) and Nieto J.J. et al. (2006)), and

ii) to know how much ordered the sequence is. In this direction the notion of entropy and clarity is employed (see, for example, Sadegh-Zadeh K. (2000)). Since the notion of entropy is related also to the calculation of distances between points in the FPS, we describe briefly in the next section the notion of distance and then pass to the concepts of entropy and clarity.

2b) Metrics - Distances
Consider the $n$-dimensional unit hypercube $I^n$.

If $p = (p_1, ..., p_n), q = (q_1, ..., q_n) \in I^n$ are two different fuzzy polynucleotides, then we consider the following distances between the elements $p$ and $q$:

1) \textit{Euclidean distance}

$$l^2(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}.$$ 

2) \textit{Hamming distance}

$$l^1(p, q) = \sum_{i=1}^{n} |p_i - q_i|.$$ 

3) \textit{Nieto - Torres Vazquez-Trasande (NTV) metric}

$$l(p, q) = \frac{\sum_{i=1}^{n} |p_i - q_i|}{\sum_{i=1}^{n} \max \{p_i, q_i\}}.$$ 

Also, if $p = q = \emptyset = (0, \ldots, 0)$, then $l(\emptyset, \emptyset) = 0$ (see Nieto J.J. et al. (2003)).

The distance $l$ is motivated by publications Lin C.T. (1997) and Sadegh-Zadeh K. (2000). We know that $l$ is a metric Nieto J.J. et al. (2003) and has already been employed in Torres A. et al. (2003) and Nieto J.J. et al. (2006). In A. Dress and T. Lokot (2003) (see, also A. Dress et al. (2004)) it is proposed to call this metric as the \textit{NTV metric}.

2c) \textbf{The entropy of a polynucleotide}

Let $X = \{x_1, ..., x_n\}$ be a set and

$$A = \{(x_1, \mu_A(x_1) = a_1), ..., (x_n, \mu_A(x_n) = a_n)\} \equiv (a_1, ..., a_n),$$

where $a_i \in [0, 1]$, a fuzzy set of X. Then by $c(A)$ (see, for example, Sadegh-Zadeh K. (2000)) we denote the number

$$\sum_{i=1}^{n} \mu_A(x_i) = \sum_{i=1}^{n} a_i.$$ 

In the crisp case, $c(A)$ is the cardinality of $A$.

In the following we propose a new definition of entropy

\textbf{Definition 1.} Let $(I^n, d)$ be a metric space (see, for example, Engelking R. (1977)), $X = \{x_1, ..., x_n\}$ a set, $C = (0.5, ..., 0.5) \in I^n$, $\emptyset = (0, ..., 0) \in I^n$, and...
$F(2^X)$ the fuzzy power set of $X$. The map

$$entropy_d : F(2^X) \to [0, 1],$$

$$entropy_d(A) = 1 - \frac{d(A, C)}{d(C, \emptyset)},$$

for every $A = (a_1, ..., a_n) \in I^n$, is called fuzzy entropy map with respect to the metric $d$.

**Remark.** De Luca and Termini A. De Luca and S. Termini (1972) first axiomatized nonprobabilistic entropy in the setting of fuzzy sets theory (see also Jiu-Liun Fan et al. (2002)). We adopt them here. Let $X$ be a set and let $E$ be a set-to-point map

$$E : F(2^X) \to [0, 1],$$

where $F(2^X)$ is the fuzzy power set of $X$. Hence $E$ is a fuzzy set defined on fuzzy sets of $X$. $E$ is an entropy measure if it satisfies the four De Luca-Termini axioms:

(DT1) $E(A) = 0$ if $A \in 2^X$ ($A$ non-fuzzy), where $2^X$ is the power set of $X$.

(DT2) $E(A) = 1$ if $A(x) = 0.5$, for every $x \in X$.

(DT3) $E(A) \leq E(B)$ if $A(x) \leq B(x)$ when $B(x) \leq 0.5$ and $B(x) \leq A(x)$ when $B(x) \geq 0.5$.

(DT4) $E(A) = e(A^c)$.

It is possible that for some metric $d$ the fuzzy entropy map $entropy_d$ with respect to the metric $d$ to satisfy the De Luca and Termini axioms. So, for example for the metrics $l^1$ and $l^2$ it is clear that the maps

$$E_1(A) \equiv entropy_{l^1}(A) = 1 - \frac{l^1(A, C)}{l^1(C, \emptyset)}$$

and

$$E_2(A) \equiv entropy_{l^2}(A) = 1 - \frac{l^2(A, C)}{l^2(C, \emptyset)}$$

satisfy the above four De Luca-Termini axioms.

Also, it is possible that for some metric $d$ the map $entropy_d$ does not satisfy the De Luca and Termini axioms. So, for example for the NTV metric $l$ the map

$$E_0(A) \equiv entropy_{l}(A) = 1 - \frac{l(A, C)}{l(C, \emptyset)}$$

does not satisfy the above four De Luca-Termini axioms (see Example 2 below).
In the following we present some theorems concerning the specification of the formulas used to calculate entropies in the FPS space for specific metrics.

**Theorem 1.** Let \( X = \{x_1, \ldots, x_n\} \) and \( A = (a_1, \ldots, a_n) \) a fuzzy set of \( X \). Then, the following statements are true:

\[
entropy_{h1}(A) = \frac{c(C) - h^1(A, C)}{c(C)} = \frac{n - 2h^1(A, C)}{n},
\]

\[
entropy_{h2}(A) = \frac{\sqrt{n} - 2h^2(A, C)}{\sqrt{n}},
\]

and

\[
entropy_{h3}(A) = 1 - l(A, C),
\]

where \( C \) is the fuzzy set \((0.5, \ldots, 0.5)\) of \( I^n \).

**Proof.** It is known that (see Sadegh-Zadeh K. (2000)) \( c(C) = l^1(C, \emptyset) \). Thus

\[
entropy_{h1}(A) = 1 - \frac{h^1(C, A)}{l^1(C, \emptyset)}
= 1 - \frac{h^1(C, A)}{c(C)}
= \frac{c(C) - h^1(A, C)}{c(C)}.
\]

Obviously,

\[
c(C) = 0.5 + 0.5 + \ldots + 0.5 = n \cdot 0.5 = \frac{n}{2},
\]

and we have:

\[
entropy_{h3}(A) = \frac{c(C) - h^1(A, C)}{c(C)}
= \frac{\frac{n}{2} - h^1(A, C)}{\frac{n}{2}}
= \frac{n - 2h^1(A, C)}{n}.
\]

Also

\[
l^2(C, \emptyset) = \frac{1}{2} \cdot \sqrt{n},
\]

and
\[ \text{entropy}_{H^2}(A) = 1 - \frac{l^2(C, A)}{I^2(C, \emptyset)} = 1 - \frac{\frac{1}{2} \cdot \sqrt{n}}{\frac{1}{2} \cdot \sqrt{n}} = \frac{\sqrt{n} - 2l(A, C)}{\sqrt{n}}. \]

Now, for the NTV metric we have:
\[ l(C, \emptyset) = \frac{n \cdot 0.5}{n \cdot 0.5} = 1. \]

Thus
\[ \text{entropy}_{TV}(A) = 1 - \frac{l(A, C)}{l(C, \emptyset)} = 1 - l(A, C). \]

This complete the proof.

**Example 1.** Let \( X = \{x_1, x_2\} \) be a set and \( A = (0.4, 0.8) \) a fuzzy set of \( X \). We consider the metric space \((I^2, l^1)\).

Using the definition of entropy given in Sadegh-Zadeh K. (2000) we have (see page 23 of Sadegh-Zadeh K. (2000)):
\[ \text{ent}(A) = \frac{3}{2} = 0.4286. \]

Using the above definition we have:
\[ \text{entropy}_{H^2}(A) = \frac{2 - 2l(A, C)}{2} = \frac{2 - 2(|0.4 - 0.5| + |0.8 - 0.5|)}{2} = \frac{2 - 0.8}{2} = \frac{1.2}{2} = 0.6. \]

We thus observe that the entropy of Sadegh-Zadeh does not coincide with the Hamming entropy
\[ \text{ent}(A) = 0.4286 \neq \text{entropy}_{H^2}(A) = 0.6. \]
Remark. According to the Definition 1 and Theorem 1 we have a geometrical interpretation of entropy is illustrated in Figure 1. The entropy of a fuzzy set $A$ is $1$ minus the Euclidean distance $a = I^2(A, C)$ divided by the Euclidean distance $b = I^2(C, \emptyset)$, that is

$$\text{entropy}_2(A) = 1 - \frac{a}{b}.$$

![Figure 1](0,0,1)(1,0,1)(1,1,1)

Theorem 2. Let $X = \{x_1,...,x_n\}$ be a set and $A = (a_1,...,a_n)$, where $a_i \in [0,1]$, a fuzzy set of $X$. Let $A^c = (1-a_1,...,1-a_n) \in I^n$. Then:

$$\text{entropy}_1(A) = \frac{c(A \land A^c)}{c(C)} = \frac{2 \cdot c(A \land A^c)}{n} = \frac{2 \cdot I^1(A \land A^c, \emptyset)}{n},$$

where $C$ is the fuzzy set $C = (0.5,...,0.5) \in I^n$.

Proof. Suppose that $a_i \leq 0.5$ for every $i = 1,2,...,n-1$, and $a_n > 0.5$. Then, we have:

$$A \land A^c = (\min\{a_1, 1-a_1\},...\min\{a_{n-1},1-a_{n-1}\}, \min\{a_n,1-a_n\})$$

$$= (a_1,...,a_{n-1},1-a_n).$$

By the above we have
\[ \text{entropy}_{h1}(A) = 1 - \frac{I^1(A, C)}{I^1(C, \emptyset)} \]

\[ = 1 - \frac{\sum_{i=1}^{n} |0.5 - a_i|}{n \cdot 0.5} \]

\[ = 1 - \frac{\sum_{i=1}^{n-1} |0.5 - a_i| + |0.5 - a_n|}{n \cdot 0.5} \]

\[ = 1 - \frac{\sum_{i=1}^{n-1} 0.5 - a_i + a_n - 0.5}{n \cdot 0.5} \]

\[ = 1 - \frac{n \cdot 0.5 - (a_1 + a_2 + ... + a_{n-1}) + a_n - 1}{n \cdot 0.5} \]

\[ = \frac{a_1 + a_2 + ... + a_{n-1} + 1 - a_n}{n \cdot 0.5} \]

\[ = \frac{c(A \land A^c)}{c(C)} \]

Now \( c(C) = \frac{n}{2} \). Thus,

\[ \text{entropy}_{h1}(A) = \frac{c(A \land A^c)}{c(C)} = \frac{2 \cdot c(A \land A^c)}{n}. \]

Finally, by the fact that

\[ c(A \land A^c) = I^1(A \land A^c, \emptyset). \]

it follows that

\[ \text{entropy}_{h1}(A) = \frac{2 \cdot I^1(A \land A^c, \emptyset)}{n}. \]

When some of the \( a_i \) are less than or equal \( 0.5 \) and others greater than \( 0.5 \), the proof is analogous.

In the following we present some examples of applications of the use of the entropy definitions in various cases of polynucleotides from relatively small ones up to large ones.

**Example 2.** Let \((I^2, d)\) be a metric space, \(X = \{x_1, x_2\}\) a set, and \(A = C = (0.5, 0.5)\) a fuzzy set of \(X\). Then, we have:

\[ \text{entropy}_{d}(A) = 1 - \frac{d(C, C)}{d(C, \emptyset)} = 1 - 0 = 1. \]

**Example 3.** Let \((I^2, d)\) be a metric space, \(X = \{x_1, x_2\}\) a set and \(A = (0, 1)\), \(B = (1, 0)\) and \(D = (1, 1)\) three fuzzy sets of \(X\). Then, we have:
\[ \text{entropy}_{d}(A) = 1 - \frac{d(C, A)}{d(C, \emptyset)} = 1 - 1 = 0, \]

where \( d = l^1 \) or \( d = l^2 \).

Similarly
\[ \text{entropy}_{d}(B) = \text{entropy}_{d}(D) = 0, \]

where \( d = l^1 \) or \( d = l^2 \).

For the NTV metric \( l \) we have
\[ \text{entropy}_{l}(A) = 1 - \frac{l(C, A)}{l(C, \emptyset)} = 1 - \frac{2}{3} = \frac{1}{3}, \]

\[ \text{entropy}_{l}(B) = \frac{1}{3}, \]

and
\[ \text{entropy}_{l}(D) = 1 - \frac{l(C, D)}{l(C, \emptyset)} = 1 - \frac{1}{2} = \frac{1}{2}. \]

We see that for points at the corners of \( I^2 \), NTV metric does not result zero values as is the case for metrics \( l^1 \) or \( l^2 \).

**Example 3.** Consider the sequences employed also by Sadegh-Zadeh in Sadegh-Zadeh K. (2000):

\( s_1 = \text{UACUGU tyrosine/cysteine} \)

This point belongs to the 24-dimensional unit cube and it corresponds to a corner in \( I^{24} \). Following the methodology of Torres and Nieto Torres A. et al. (2003) we calculate the frequencies (fractions) of the nucleotide at the three base sites in order to obtain their fuzzy representation in the \( I^{12} \) hyperspace. The corresponding results appear in tables 1a and 1b. Note that the entropy in \( I^{24} \) is
\[ \text{entropy}_{3}(s_1) = \text{entropy}_{2}(s_1) = 0 \]

and
\[ \text{entropy}_{l}(s_1) = 0.2. \]

In the \( I^{12} \) space the frequencies give a point in the \( I^{12} \) space:

\[ f(s_1) = (1, 0, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0). \]

Note that now we identify \( s_1 \) in \( I^{24} \) and \( f(s_1) \) in \( I^{12} \).

If \( C = (0.5, \ldots, 0.5) \in I^{12} \), then, we have the following entropies for the Euclidean metric

14
\[ \text{entropy}_{H}(s_1) = 1 - \frac{P(s_1, C)}{P(C, \emptyset)} = 1 - \frac{\sqrt{2}}{\sqrt{3}} \approx 0.183503, \]

Hamming metric,

\[ \text{entropy}_{H}(s_1) = 1 - \frac{l(s_1, C)}{l(C, \emptyset)} = 1 - \frac{1}{3} \approx 0.333333, \]

and NTV metric

\[ \text{entropy}_{H}(s_1) = 1 - \frac{l(s_1, C)}{l(C, \emptyset)} = \frac{5}{13} \approx 0.384615. \]

We can see that there is a difference in the results when dealing with FPS. This subtlety will be analyzed with further results.

**Example 4.** Consider the sequences:

- \( s_2 = \text{CACUGU histidine/cysteine} \)
- \( s_3 = \text{CUCUGU leucine/cysteine} \)
- \( s_4 = \text{CAUUGU histidine/cysteine} \)
- \( s_5 = \text{CAGUGU glutamine/cysteine} \)
- \( s_6 = \text{CAUGUG glutamine/cysteine} \)

These are points in a 24-dimensional unit cube since they are made of 2 triplets. Following the methodology of Torres A. et al. (2003) we calculated the frequencies (fractions) of the nucleotides at the three base sites in order to obtain their fuzzy representation in the \( I^{12} \) hyperspace. The entropy in the \( I^{24} \) is again zero as there is no uncertainty concerning the chemical composition. However when dealing with FPS results will be different. In the case of FPS zero entropy means maximum order we have the same triplet all along the genetic sequence.

The above sequences are represented in the \( I^{12} \) space as (see Nieto J.J. et al. (2006)):

\[
\begin{align*}
s_2 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0), \\
s_3 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0), \\
s_4 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0, 1, 0, 0), \\
s_5 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0.5) \\
\end{align*}
\]
and 

\[ s_6 = (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0, 0, 0) . \]

The results of entropy using the various metrics are summarized in Table 2.

Table 2. Entropy values for sequences \( s_2, s_3, s_4, s_5, s_6 \) using the metrics \( l, l^1 \) and \( l^2 \) calculated in FPS.

<table>
<thead>
<tr>
<th>metric</th>
<th>( s_2 )</th>
<th>( s_3 )</th>
<th>( s_4 )</th>
<th>( s_5 )</th>
<th>( s_6 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l^2 )</td>
<td>0.292893</td>
<td>0.292893</td>
<td>0.183503</td>
<td>0.292893</td>
<td>0.292893</td>
</tr>
<tr>
<td>( l^1 )</td>
<td>0.5</td>
<td>0.5</td>
<td>0.333333</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>( l )</td>
<td>0.5</td>
<td>0.5</td>
<td>0.384615</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

\( s_2, s_3, s_4, s_5 \) and \( s_6 \) present the same entropy results although the exact value changes depending on the metric used and only \( s_4 \) presents different entropy which is lower. In fact \( s_2, s_3, s_5 \) and \( s_6 \) have the same number of coordinates being equal to 0.5 and all the others 0, while \( s_4 \) has only four coordinates equal to 0.5, one equal to 1 and all the others equal to 0.

Example 5. Now consider the following sequences:

\( s_7 = \text{UACUAC} \)

\( s_8 = \text{UAGUAU} \)

\( s_9 = \text{UACUCG} \)

which correspond in the FPS respectively to

\[ s_7 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0) \]

\[ s_8 = (1, 0, 0, 0, 0, 1, 0, 0, 0.5, 0, 0, 0.5), \]

\[ s_9 = (1, 0, 0, 0, 0, 0.5, 0.5, 0, 0, 0.5, 0.5). \]

The corresponding entropy values appear in Table 3. Note in the first one, \( s_7 \), the same triplet UAC is repeated all along the sequence. In the second one, \( s_8 \), the dinucleotide UA is repeated at the same base positions.

Table 3. Calculated entropy values for sequences \( s_7, s_8 \) and \( s_9 \) using the metrics \( l, l^1 \) and \( l^2 \).
<table>
<thead>
<tr>
<th>metric</th>
<th>$s_7$</th>
<th>$s_8$</th>
<th>$s_9$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$</td>
<td>0</td>
<td>0.087129</td>
<td>0.183503</td>
</tr>
<tr>
<td>$l^1$</td>
<td>0</td>
<td>0.166667</td>
<td>0.333333</td>
</tr>
<tr>
<td>$l$</td>
<td>0.2</td>
<td>0.285714</td>
<td>0.384615</td>
</tr>
</tbody>
</table>

The program to compute the entropy using these three metrics is available on request from the authors.

As we know from physics, entropy is a measure of the order/disorder of the system, which represents the degree of complexity in order to describe the system. In the case where we have repetition of the same triplet (as it is the case for UAC in sequence $s_7$) we have maximum order resulting in zero entropy. In the case where we have repetition of the same dyad like UA in sequence $s_8$ we have a slightly higher entropy and in other cases like sequence $s_9$ entropy increases further.

We remind here that there is a probabilistic definition of entropy for signals related to thermodynamics see Shannon, C.E. (1946)). This classical measure of entropy is defined as

$$H = - \sum p_j \log_2(p_j)$$

where $p_j$ are the non-zero probabilities of a signal to have a given value.

Applying this definition in the case of selected polynucleotides as represented in the FPS the role of $p_j$ is played by the non-zero coordinates of the polynucleotide. In this case for the entropy of $s_7$, $s_8$ and $s_9$ we have

$$H(s_7) = 0,$$

$$H(s_8) = 1$$

and

$$H(s_9) = 2.$$  

What is of interest is that sequence $s_7$ which corresponds to the most ordered sequence results in zero entropy. In the other two sequences we have increasing entropy as in the case of definitions based on metrics given above. It is also remarkable that the ratio of entropies of sequences $s_8$ and $s_9$ equals 2 in both cases: probabilistic entropy and entropy based on metrics.

**Example 6.** Consider the following sequences with three triplets

$$r_1 = UACUACUAC$$
\[ r_2 = UACUACUAG \]

\[ r_3 = UACCAAUAG \]

represented in the \( I^{3 \times 12} = I^{36} \) space as

\[
\begin{align*}
    r_1 &= (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0) \\
    r_2 &= (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1) \\
    r_3 &= (1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1).
\end{align*}
\]

The entropy in the \( I^{36} \) space is

\[
\text{entropy}_d(r_1) = \text{entropy}_d(r_2) = \text{entropy}_d(r_3) = 0.
\]

where \( d = l^1 \) or \( d = l^2 \), and

\[
\text{entropy}_d(r_1) = \text{entropy}_d(r_2) = \text{entropy}_d(r_3) = 0.2.
\]

However in the FPS representation the entropy of the sequences would not result zero values. In fact zero values would be reproduced only for the sequences where the same triplet is repeated all along the sequence, for the second where we have repetition of the dyad \( UA \) at the same base positions entropy increases and is higher in the third case. Following the methodology of Torres A. et al. (2003) we calculated the frequencies (fractions) of the nucleotides at the three base sites in order to obtain their fuzzy representation in the \( I^{12} \) hyperspace

\[
\begin{align*}
    r_1 &= (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0) \\
    r_2 &= (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2/3, 1/3) \\
    r_3 &= (2/3, 1/3, 0, 0, 0, 1, 0, 0, 0, 1/3, 1/3, 1/3)
\end{align*}
\]

The corresponding entropy values are summarized in Table 4.

**Table 4.** Calculated entropy values for sequences \( r_1, r_2 \) and \( r_3 \) using the metrics \( l, l^1 \) and \( l^2 \).
<table>
<thead>
<tr>
<th>metric</th>
<th>$r_1$</th>
<th>$r_2$</th>
<th>$r_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$</td>
<td>0</td>
<td>0.077042</td>
<td>0.206508</td>
</tr>
<tr>
<td>$l^1$</td>
<td>0</td>
<td>0.111111</td>
<td>0.277778</td>
</tr>
<tr>
<td>$l$</td>
<td>0.2</td>
<td>0.255814</td>
<td>0.35</td>
</tr>
</tbody>
</table>

**Remarks.** (1) In the case of large sequences the maximum entropy would correspond to the characteristic case in which we have equiprobable distribution of all alphabet letters at all three bases. This would correspond to the point:

$$E = (0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25, 0.25)$$

in the FPS representation. Its entropy in the metric $l^2$ is

$$\text{entropy}_{l^2}(E) = 1 - \frac{l^2(E, C)}{l^2(C, \emptyset)} = 1 - \frac{\sqrt{12} \cdot 0.25^2}{\sqrt{12} \cdot 0.5^2} = 0.5.$$ 

The entropies obtained by the use of metrics $l^1$ and $l$ are:

$$\text{entropy}_{l^1}(E) = 1 - \frac{l^1(E, C)}{l^1(C, \emptyset)} = 0.5$$

and

$$\text{entropy}_l(E) = 1 - \frac{l(E, C)}{l(C, \emptyset)} = 0.5.$$ 

A point in the FPS corresponds to a corner of the hypercube when we have the same triplet all along the sequence. If we have maximum order, the point occupies a corner of the hypercube. The bigger the distance from the corners, the bigger the entropy, and thus the bigger the complexity to describe the sequence. In the probabilistic definition of entropy, for the point $E$ we have:

$$H(E) = 6$$

which is the maximum possible value in $I^{12}$.

(2) If $A = (a_1, ..., a_{12}) \in I^{12}$, $C = (0.5, ..., 0.5) \in I^{12}$, and $\emptyset = (0, ..., 0) \in I^{12}$, then
\[
entropy_{H^2}(A) = 1 - \frac{i^2(C, A)}{i^2(C, \emptyset)} \\
= 1 - \frac{\sqrt{(a_1 - 0.5)^2 + \ldots + (a_{12} - 0.5)^2}}{\sqrt{(0.5 - 0)^2 + \ldots + (0.5 - 0)^2}} \\
= 1 - \frac{\sqrt{(a_1 - 0.5)^2 + \ldots + (a_{12} - 0.5)^2}}{\sqrt{3}}
\]

In FPS we have that 
\[
a_1 + \ldots + a_{12} = 3.
\]

This comes out from the fact that each \(a_i\) corresponds to the frequency of appearance of a letter of the DNA (or RNA) alphabet at each triplet base (first, second, third). For each base these probabilities which correspond to a quadruplet of \(a_i\) sums to 1, so for the three bases this results in 3.

Using a simple program of Mathematica (see Appendix 1) we see that the map
\[
entropy_{H^2}(A) = 1 - \frac{\sqrt{(a_1 - 0.5)^2 + \ldots + (a_{12} - 0.5)^2}}{\sqrt{3}}
\]
with the restriction
\[
a_1 + \ldots + a_{12} = 3,
\]
we have a maximum of the entropy at the point \(E\).

As above, we can see that
\[
entropy_{H^1}(A) = 1 - \frac{i^1(C, A)}{i^1(C, \emptyset)} = 1 - \frac{|a_1 - 0.5| + \ldots + |a_{12} - 0.5|}{6}
\]
with the restriction
\[
a_1 + \ldots + a_{12} = 3.
\]

have a maximum (of entropy) at the point \(E\).

Example 7. Following the methodology of Torres A. et al. (2003), we consider the point
\[
(0.1632, 0.3089, 0.1724, 0.3556, 0.2036, 0.3145, 0.1763, 0.3056, 0.1645, 0.3461, 0.1593, 0.3302) \in I^{12}.
\]
which corresponds to the fuzzy set of frequencies of the genome of \textit{M.tuberculosis} (see Torres A. et al. (2003)), the point
\[
(0.1605, 0.2420, 0.2600, 0.3374, 0.3116, 0.2286, 0.2846, 0.1752, 0.2619, 0.2568, 0.1831, 0.2981) \in I^{12}.
\]
which corresponds to the fuzzy set of frequencies of the genome of \textit{E.coli}, and the point
\[
(0.1706, 0.1605, 0.3241, 0.3446, 0.3282, 0.1735, 0.3478, 0.1504, 0.2139, 0.2455, 0.3052, 0.2352) \in I^{12}
\]
which corresponds to the fuzzy set of frequencies of the genome of *A.aeolicus* (see Nieto J.J. et al. (2006)).

We also compare the entropies of the *Mycoplasma Pneumoniae* using our definitions of entropy. In tables 5a and 5b we present the results concerning the representation of *Mycoplasma Pneumoniae* in FPS.

**Table 5a.** The number of nucleotides at the three base sites of a codon in the codon sequence of *Mycoplasma Pneumoniae.*

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>First base</td>
<td>48995</td>
<td>42525</td>
<td>78622</td>
<td>70293</td>
</tr>
<tr>
<td>Second base</td>
<td>73438</td>
<td>46554</td>
<td>86585</td>
<td>33858</td>
</tr>
<tr>
<td>Third base</td>
<td>77233</td>
<td>54942</td>
<td>62523</td>
<td>45737</td>
</tr>
</tbody>
</table>

**Table 5b.** Fractions of nucleotides at the three base sites of a codon in the coding sequences of *Mycoplasma Pneumoniae*:

<table>
<thead>
<tr>
<th></th>
<th>T</th>
<th>C</th>
<th>A</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>First base</td>
<td>0.2038</td>
<td>0.1769</td>
<td>0.327</td>
<td>0.2923</td>
</tr>
<tr>
<td>Second base</td>
<td>0.3054</td>
<td>0.1936</td>
<td>0.3601</td>
<td>0.1408</td>
</tr>
<tr>
<td>Third base</td>
<td>0.3212</td>
<td>0.2285</td>
<td>0.26</td>
<td>0.1902</td>
</tr>
</tbody>
</table>

Thus, the genome of *M.pneumoniae* is represented in the $I^{12}$ hypercube by the point

$(0.2038, 0.1769, 0.327, 0.2923, 0.3054, 0.1936, 0.3601, 0.1408, 0.3212, 0.2285, 0.26, 0.1902) \in I^{12}$.

**Table 6.** Entropy values for sequences *M.tuberculosis*, *E.coli*, *A.aeolicus* and *M.pneumoniae* using the metrics $l$, $l^1$ and $l^2$ calculated in FPS.

<table>
<thead>
<tr>
<th>metric</th>
<th><em>M.tuberculosis</em></th>
<th><em>E.coli</em></th>
<th><em>A.aeolicus</em></th>
<th><em>M.pneumoniae</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$</td>
<td>0.475876</td>
<td>0.488814</td>
<td>0.478769</td>
<td>0.482055</td>
</tr>
<tr>
<td>$l^1$</td>
<td>0.500033</td>
<td>0.499967</td>
<td>0.499917</td>
<td>0.499967</td>
</tr>
<tr>
<td>$l$</td>
<td>0.500033</td>
<td>0.499967</td>
<td>0.499917</td>
<td>0.499967</td>
</tr>
</tbody>
</table>

The corresponding entropies of all long polynucleotides appear in Table 6. We observe that metrics give the same numerical results while the NTV metric gives different results and can differentiate in a more clear way the complexity.
of polynucleotides in their FPS representation.

**Remark.** In table 7 we compare the results obtained using the three above entropy definitions for the *A. aeolicus* and *E. coli* with the results of Menconi Giulia Menconi (2005) where they computed the complexity *K* of a genome and the probabilistic entropy *H*<sub>1</sub> (for more details on the notions of *K* and *H*<sub>1</sub> see Giulia Menconi (2005)). What is of interest is that in the case of *entropy*<sub>2</sub> and *entropy*<sub>3</sub> results are practically identical, while *entropy*<sub>1</sub> results in a small but identifiable difference between the two genomes in the same sense like *K* and *H*<sub>1</sub>.

**Table 7.** Comparison of results obtained using *entropy*<sub>1</sub>, *entropy*<sub>2</sub>, *entropy*<sub>3</sub> with the results of complexity *K* and probabilistic entropy *H*<sub>1</sub> of Giulia Menconi (2005) in the case of *A. aeolicus* and *E. coli*.

<table>
<thead>
<tr>
<th>Genome</th>
<th>K</th>
<th><em>H</em>&lt;sub&gt;1&lt;/sub&gt;</th>
<th><em>entropy</em>&lt;sub&gt;2&lt;/sub&gt;</th>
<th><em>entropy</em>&lt;sub&gt;1&lt;/sub&gt;</th>
<th><em>entropy</em>&lt;sub&gt;3&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>A. aeolicus</em></td>
<td>1.883</td>
<td>1.976</td>
<td>0.478</td>
<td>0.499</td>
<td>0.499</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>1.893</td>
<td>1.987</td>
<td>0.489</td>
<td>0.499</td>
<td>0.499</td>
</tr>
</tbody>
</table>

3 Clarity and Fuzzy Polynucleotide Space

**Definition 2.** Let *d* be a metric in *I*<sup>+</sup>, *X* = \{*x*<sub>1</sub>, ..., *x*<sub>*n*</sub>\} a set and *A* = (*a*<sub>1</sub>, ..., *a*<sub>*n*</sub>), where *a*<sub>*i*</sub> ∈ [0, 1], a fuzzy set of *X*. The *clarity* of *A*, with respect to the metric *d*, denoted by *clarity*<sub>*d*</sub>(*A*), is defined to be the number

\[
1 - \text{entropy}_d(*A*),
\]

that is

\[
\text{clarity}_d(*A*) = 1 - \text{entropy}_d(*A*).
\]

**Example 11.** Let *X* = \{*x*<sub>1</sub>, *x*<sub>2</sub>\} be a set, *A* = (0.4, 0.8) a fuzzy set of *X* and consider the metric space (*I*<sup>2</sup>, *l*<sup>1</sup>). Then, we have

\[
\text{clarity}_{l^1}(*A*) = 1 - \text{entropy}_{l^1}(*A*) = 1 - 0.6 = 0.4.
\]

Also, using the definition of clarity given in Sadegh-Zadeh K. (2000) we have:

\[
\text{clar}(*A*) = 1 - \text{ent}(*A*) = 1 - \frac{3}{7} = \frac{4}{7} = 0.5714.
\]
We observe that

\[ \text{clarity}_{A1}(A) = 0.4 \neq \text{clar}(A) = 0.5714. \]

**Theorem 3.** Let \( X = \{x_1, \ldots, x_n\} \) and \( A = (a_1, \ldots, a_n) \) a fuzzy set of \( X \). Then, the following statements are true:

1) \( \text{entropy}_d(A), \text{clarity}_d(A) \in [0, 1]. \)
2) \( \text{entropy}_d(A) = 1 - \text{clarity}_d(A). \)
3) \[
\text{clarity}_{A1}(A) = \frac{l_1(A, C)}{c(C)}.
\]
4) \[
\text{clarity}_{A1}(A) = \frac{2l_1(A, C)}{n}.
\]
5) \[
\text{clarity}_{A1}(A) = \frac{c(C) - c(A \land \neg A)}{c(C)}.
\]
6) \[
\text{clarity}_d(A) = \frac{2l(A, C)}{\sqrt{n}}.
\]
7) \[
\text{clarity}_d(A) = l(A, C).
\]

**Proof.** Follows by Theorems 1 and 2 and by fact that \( \text{clarity}_d(A) = 1 - \text{entropy}_d(A). \)

**Remark.** According the Definition 2 and Theorem 3 we have a geometrical interpretation of clarity as illustrated in Figure 1. The clarity of a fuzzy set \( A \) is the Euclidean distance \( a = d(A, C) \) divided by the distance \( b = d(C, \emptyset) \), that is

\[ \text{clarity}_{A1}(A, B) = \frac{a}{b}. \]

**Example 12.** (1) Let \( X = \{x_1, x_2\} \) be a set and \( A = (0, 1), B = (1, 0) \) two fuzzy sets of \( X \). Then

\[ \text{entropy}_d(A) = \text{entropy}_d(B) = 0 \]

and

\[ \text{clarity}_d(A) = \text{clarity}_d(B) = 1, \]
where \( d = l^1 \) or \( d = l^2 \). Also,

\[
\text{entropy}_d(A) = \text{entropy}_d(B) = \frac{1}{3}
\]

and

\[
\text{clarity}_d(A) = \text{clarity}_d(B) = 1 - \frac{1}{3} = \frac{2}{3}
\]

(2) Let \( X = \{x_1, \ldots, x_n\} \) be a set and \( A \) a fuzzy set of \( X \) such that \( A = C \). Then \( \text{entropy}_d(A) = 1 \) and \( \text{clarity}_d(A) = 0 \).

(3) We consider the following polynucleotide sequences:

\( s_1 = \text{UACUGU} \) (tyrosine/cysteine)

\( s_2 = \text{CACUGU} \) (histidine/cysteine)

\( s_3 = \text{CUCUGU} \) (leucine/cysteine)

We have the following representations in the \( I^{24} \) space:

\[
s_1 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0)
\]

\[
s_2 = (0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0)
\]

\[
s_3 = (0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0)
\]

Using the distances \( l^1 \), \( l^1 \) and \( l \) for the clarity of \( s_1, s_2, s_3 \in I^{24} \) we obtain the results for entropy values that are summarized in Table 8.

**Table 8.** Calculated clarity values for sequences \( s_1, s_2 \) and \( s_3 \) using the metrics \( l \), \( l^1 \) and \( l^2 \).

<table>
<thead>
<tr>
<th>metric</th>
<th>( s_1 )</th>
<th>( s_2 )</th>
<th>( s_3 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( l^2 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( l^1 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>( l )</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
</tr>
</tbody>
</table>

The results are consistent with the fact that in the case of the \( I^{12 \times k} \) space (in this case of \( I^{24} \)) all above sequences are known with precision. However if we use their representation in the \( I^{12} \) dimensional FPS results the results present some differences.

Using the distances \( l^2 \), \( l^1 \) and \( l \) for the clarity of \( s_1, s_2, s_3 \in I^{12} \) we have the results that appear in Table 9.:
Table 9. Calculated clarity values for sequences $s_1$, $s_2$ and $s_3$ using the metrics $l$, $l^1$ and $l^2$.

<table>
<thead>
<tr>
<th>metric</th>
<th>$s_1$</th>
<th>$s_2$</th>
<th>$s_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$</td>
<td>0.816497</td>
<td>0.707107</td>
<td>0.707107</td>
</tr>
<tr>
<td>$l^1$</td>
<td>0.66667</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>$l$</td>
<td>0.67385</td>
<td>0.5</td>
<td>0.5</td>
</tr>
</tbody>
</table>

Again, results indicate the degree of complexity necessary for the description of the polynucleotides.

(4) We consider the following fuzzy set of frequencies of the genomes of \textit{M.tuberculosis}, \textit{E.coli}, \textit{A.aeolicus}, and \textit{M.pneumoniae}. Using the distances $l^2$, $l^1$ and $l$ for the clarity the corresponding results are summarized in Table 10.

Table 10. Computed clarity for sequences \textit{M.tuberculosis}, \textit{E.coli}, \textit{A.aeolicus} and \textit{M.pneumoniae}.

<table>
<thead>
<tr>
<th>metric</th>
<th>\textit{M.tuberculosis}</th>
<th>\textit{E.coli}</th>
<th>\textit{A.aeolicus}</th>
<th>\textit{M.pneumoniae}</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l^2$</td>
<td>0.524124</td>
<td>0.511186</td>
<td>0.521231</td>
<td>0.57945</td>
</tr>
<tr>
<td>$l^1$</td>
<td>0.499967</td>
<td>0.500033</td>
<td>0.500083</td>
<td>0.500033</td>
</tr>
<tr>
<td>$l$</td>
<td>0.499967</td>
<td>0.500033</td>
<td>0.500083</td>
<td>0.50</td>
</tr>
</tbody>
</table>

4 Conclusions

We present results concerning the notions of fuzzy entropy and clarity of a polynucleotide. We propose a new definition of entropy and we consider several applications using different metrics in the case of $l^{12 \times k}$ space, where $k$ is the number of codons of the polynucleotide. We also examine the behavior of these notions when those polynucleotides are projected in the $l^{12}$ Fuzzy Polynucleotide Space. We observe that in both cases we have a different interpretation of the obtained results for entropy. While in the former case low entropy means that we are close to a corner of the 12xk space in the latter case it means that we have repetition of the same triplet or part of a triplet all along the sequence of the polynucleotide. However, results in both cases show, as expected, that entropy is related to the complexity of description of the sequence. The value of entropy/clarity is representative of the complexity of description of the polynucleotide. Similar entropy means similar degree of complexity. We also apply the definition of probabilistic entropy in the case of selected polynucleotides and we observe that the entropy based on metrics
presents similar behavior with that of probabilistic nature.

Further studies are in progress in order to investigate in more detail the properties of these notions and their biological implications since it seems that the use of FPS space can lead to a reduction of the necessary information and the use of appropriate metrics can be used to differentiate the degree of their complexity.

Acknowledgments

The research of J.J. Nieto has been partially supported by Ministerio de Educacion y Ciencia and FEDER, project MTM2007-61724, and by Xunta de Galicia and FEDER, project PGIDIT05PXIC20702PN.

Appendix 1

In[1]:=

\[ \text{Minimize} \left[ \frac{1}{3} \sum_{i=1}^{12} a[i]^2 \cdot \left( a[i] - 0.5 \right)^2, \sum_{i=1}^{12} a[i] = 3 \right], \text{Array} \{a, \{12\}\} \]

Out[1]=

\{0.5, \{a[1] \rightarrow 0.25, a[2] \rightarrow 0.25, a[3] \rightarrow 0.25, a[4] \rightarrow 0.25, a[5] \rightarrow 0.25, a[6] \rightarrow 0.25, a[7] \rightarrow 0.25, a[8] \rightarrow 0.25, a[9] \rightarrow 0.25, a[10] \rightarrow 0.25, a[11] \rightarrow 0.25, a[12] \rightarrow 0.25\}\}

References


Chou K. C., Cai Y. D., Predicting enzyme family class in a hybridization space, Protein Science 13 (2004), 2857-2863.


receptors, Journal of Proteome Research 1, 429-433.
Chou K. C., Elrod D. W., 1999, Protein subcellular location prediction, Protein Engineering 12, 107-118.
Chechetkin V. R., 2003, Block structure and stability of the genetic code. J. Theoretical Biology 222, 177-188
Foster M., Heath A. and Afzal M., Application of distance geometry to 3D
visualization of sequence relationships, Bionformatics 15 (1999), 89-90.
Nucleic Acids Res., 28, 374


large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun, 355, 1006-1011. 

Wolffenden R., 2007, Experimental measures of amino acid hydrophobicity and the topology of transmembrane and globular proteins. J. Cell Biology 177, i10-i10


