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

www.elsevier.com/locate/yjtbi

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: YJTBI 6114

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

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

D. N. Georgiou[a] , T. E. Karakasidis[b] , Juan J. Nieto[c] ,
A. Torres[d]

[a] *University of Patras, Department of Mathematics, 265 00 Patras, Greece*

[b] *University of Thessaly, Department of Civil Engineering, 383 34 Volos, Greece,*

[c] *Departamento de Análisis Matemático, Facultad de Matemáticas, Universidad de Santiago de Compostela, 15782 Spain*

[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 Polynuclotide 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 poynucleotides 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), amnieto@usc.es (Juan J. Nieto), mrtorres@usc.es (A. Torres).

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 polynuclotides, 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 *M.tuberculosis* and *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 biopolyment 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 (Trinquier 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 trough cellurar 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 (Liu 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 psuedo-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 calulation 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 polynucloetides.

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 μ_A such that

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

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 $A \wedge B$ we denote the fuzzy set for which the membership function $\mu_{A\wedge B} : X \to [0, 1]$ is defined as following

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

for every $x \in X$.

Also by $A \vee B$ we denote the fuzzy set for which the membership function $\mu_{A\vee B} : X \to [0, 1]$ is defined as following

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

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 $\mu_A : X \to [0, 1]$ 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

$$\mu : X \to \{0, 1\}$$

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 Hegalson 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 $(1, 0, 0, 0)$: 1 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 polynuclotide 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

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

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.

	U	C	A	G	total
First base	2	0	0	0	2
Second base	0	0	1	1	2
Third base	1	1	0	0	2

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

	U	C	A	G
First base	1	0	0	0
Second base	0	0	0.5	0.5
Third base	0.5	0.5	0	0

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

$$f(s_1) = (1, 0, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0).$$

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) *Euclidean distance*

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

2) *Hamming distance*

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

3) *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 $p = q = \emptyset = (0, ..., 0)$, 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 *NTV metric*.

2c) 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

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

$$entropy_{l^1}(A) = \frac{c(C) - l^1(A,C)}{c(C)} = \frac{n - 2l^1(A,C)}{n},$$

$$entropy_{l^2}(A) = \frac{\sqrt{n} - 2l^2(A,C)}{\sqrt{n}},$$

and

$$entropy_l(A) = 1 - l(A,C),$$

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

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

$$\begin{aligned} entropy_{l^1}(A) &= 1 - \frac{l^1(C,A)}{l^1(C,\emptyset)} \\ &= 1 - \frac{l^1(C,A)}{c(C)} \\ &= \frac{c(C) - l^1(A,C)}{c(C)}. \end{aligned}$$

Obviously,

$$c(C) = 0.5 + 0.5 + ... + 0.5 = n \cdot 0.5 = \frac{n}{2},$$

and we have:

$$\begin{aligned} entropy_{l^1}(A) &= \frac{c(C) - l^1(A,C)}{c(C)} \\ &= \frac{\frac{n}{2} - l^1(A,C)}{\frac{n}{2}} \\ &= \frac{n - 2l^1(A,C)}{n}. \end{aligned}$$

Also

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

and

$$\begin{aligned} entropy_{l^2}(A) &= 1 - \frac{l^2(C,A)}{l^2(C,\emptyset)} \\ &= 1 - \frac{l^2(C,A)}{\frac{1}{2}\cdot\sqrt{n}} \\ &= \frac{\sqrt{n} - 2l(A,C)}{\sqrt{n}}. \end{aligned}$$

Now, for the NTV metric we have:

$$l(C,\emptyset) = \frac{n\cdot 0.5}{n\cdot 0.5} = 1.$$

Thus

$$entropy_l(A) = 1 - \frac{l(A,C)}{l(C,\emptyset)} = 1 - l(A,C).$$

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)):

$$ent(A) = \frac{3}{7} = 0.4286.$$

Using the above definition we have:

$$\begin{aligned} entropy_{l^1}(A) &= \frac{2 - 2l^1(A,C)}{2} \\ &= \frac{2 - 2(|0.4 - 0.5| + |0.8 - 0.5|)}{2} \\ &= \frac{2 - 0.8}{2} = \frac{1.2}{2} = 0.6. \end{aligned}$$

We thus observe that the entropy of Sadegh-Zadeh does not coincide with the Hamming entropy

$$ent(A) = 0.4286 \neq entropy_{l^1}(A) = 0.6.$$

Remark. According to the Definition 1 and Theorem 1 we have a geometrical intepretation of entropy is illustrated in Figure 1. The $entropy_{l^1}$ of a fuzzy set A is 1 minus the Euclidean distance $a = l^2(A, C)$ divided by the Euclidean distance $b = l^2(C, \emptyset)$, that is

$$entropy_{l^2}(A) = 1 - \frac{a}{b}.$$

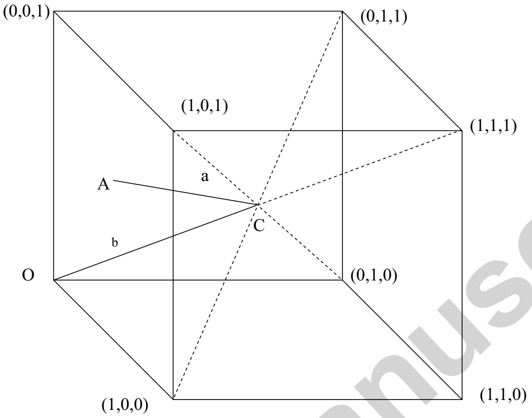


Figure 1

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:

$$entropy_{l^1}(A) = \frac{c(A \wedge A^c)}{c(C)} = \frac{2 \cdot c(A \wedge A^c)}{n} = \frac{2 \cdot l^1(A \wedge A^c, \emptyset)}{n},$$

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:

$$\begin{aligned} A \wedge A^c &= (\min\{a_1, 1 - a_1\}, ..., \min\{a_{n-1}1, 1 - a_{n-1}\}, \min\{a_n, 1 - a_n\}) \\ &= (a_1, ..., a_{n-1}, 1 - a_n). \end{aligned}$$

By the above we have

$$\begin{aligned} entropy_{l^1}(A) &= 1 - \frac{l^1(A,C)}{l^1(C,\emptyset)} \\ &= 1 - \frac{\sum_{i=1}^{n} |0.5 - a_i|}{n \cdot 0.5} \\ &= 1 - \frac{\sum_{i=1}^{n-1} |0.5 - a_i| + |0.5 - a_n|}{n \cdot 0.5} \\ &= 1 - \frac{\sum_{i=1}^{n-1} 0.5 - a_i + a_n - 0.5}{n \cdot 0.5} \\ &= 1 - \frac{n \cdot 0.5 - (a_1 + a_2 + ... + a_{n-1}) + a_n - 1}{n \cdot 0.5} \\ &= \frac{n \cdot 0.5 - n \cdot 0.5 + (a_1 + a_2 + ... + a_{n-1}) - a_n + 1}{n \cdot 0.5} \\ &= \frac{a_1 + a_2 + ... + a_{n-1} + 1 - a_n}{n \cdot 0.5} \\ &= \frac{c(A \wedge A^c)}{c(C)}. \end{aligned}$$

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

$$entropy_{l^1}(A) = \frac{c(A \wedge A^c)}{c(C)} = \frac{2 \cdot c(A \wedge A^c)}{n}.$$

Finally, by the fact that

$$c(A \wedge A^c) = l^1(A \wedge A^c, \emptyset).$$

it follows that

$$entropy_{l^1}(A) = \frac{2 \cdot l^1(A \wedge A^c, \emptyset)}{n}.$$

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:

$$entropy_d(A) = 1 - \frac{d(C,C)}{d(C,\emptyset)} = 1 - 0 = 1.$$

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:

$$entropy_d(A) = 1 - \frac{d(C,A)}{d(C,\emptyset)} = 1 - 1 = 0,$$

where $d = l^1$ or $d = l^2$.

Similarly

$$entropy_d(B) = entropy_d(D) = 0,$$

where $d = l^1$ or $d = l^2$.

For the NTV metric l we have

$$entropy_l(A) = 1 - \frac{l(C,A)}{l(C,\emptyset)} = 1 - \frac{2}{3} = \frac{1}{3},$$

$$entropy_l(B) = \frac{1}{3},$$

and

$$entropy_l(D) = 1 - \frac{l(C,D)}{l(C,\emptyset)} = 1 - \frac{1}{2} = \frac{1}{2}.$$

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=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

$$entropy_{l^1}(s_1) = entropy_{l^2}(s_1) = 0$$

and

$$entropy_l(s_1) = 0.2.$$

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

$$entropy_{l^2}(s_1) = 1 - \frac{l^2(s_1, C)}{l^2(C, \emptyset)} = 1 - \frac{\sqrt{2}}{\sqrt{3}} \approx 0.183503,$$

Hamming metric,

$$entropy_{l^1}(s_1) = 1 - \frac{l^1(s_1, C)}{l^1(C, \emptyset)} = \frac{1}{3} \approx 0.333333,$$

and NTV metric

$$entropy_l(s_1) = 1 - \frac{l(s_1, C)}{l(C, \emptyset)} = \frac{5}{13} \approx 0.384615.$$

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=CACUGU histidine/cysteine

s_3=CUCUGU leucine/cysteine

s_4=CAUUGU histidine/cysteine

s_5=CAGUGU glutamine/cysteine

s_6=CAAUGU 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)):

$$\begin{aligned} s_2 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0.5, 0, 0), \\ s_3 &= (0.5, 0.5, 0, 0, 0.5, 0, 0, 0.5, 0.5, 0.5, 0, 0), \\ s_4 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 1, 0, 0, 0), \\ s_5 &= (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0, 0, 0.5) \end{aligned}$$

and

$$s_6 = (0.5, 0.5, 0, 0, 0, 0, 0.5, 0.5, 0.5, 0, 0.5, 0).$$

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.

metric	s_2	s_3	s_4	s_5	s_6
l^2	0.292893	0.292893	0.183503	0.292893	0.292893
l^1	0.5	0.5	0.333333	0.5	0.5
l	0.5	0.5	0.384615	0.5	0.5

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=UACUAC

s_8=UAGUAU

s_9=UACUCG

which correspond in the FPS respectively to

$$s_7 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0)$$

$$s_8 = (1, 0, 0, 0, 0, 1, 0, 0, 0.5, 0, 0, 0.5),$$

$$s_9 = (1, 0, 0, 0, 0, 0.5, 0.5, 0, 0, 0, 0.5, 0.5).$$

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.

metric	s_7	s_8	s_9
l^2	0	0.087129	0.183503
l^1	0	0.166667	0.333333
l	0.2	0.285714	0.384615

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

$$r_1 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0)$$

$$r_2 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1)$$

$$r_3 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1).$$

The entropy in the I^{36} space is

$$entropy_d(r_1) = entropy_d(r_2) = entropy_d(r_3) = 0,$$

where $d = l^1$ or $d = l^2$, and

$$entropy_l(r_1) = entropy_l(r_2) = entropy_l(r_3) = 0.2.$$

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:

$$r_1 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0)$$

$$r_2 = (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 2/3, 1/3)$$

$$r_3 = (2/3, 1/3, 0, 0, 0, 1, 0, 0, 0, 1/3, 1/3, 1/3)$$

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.

metric	r_1	r_2	r_3
l^2	0	0.077042	0.206508
l^1	0	0.111111	0.277778
l	0.2	0.255814	0.35

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

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

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

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

and

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

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

$$
\begin{aligned}
entropy_{l^2}(A) &= 1 - \frac{l^2(C,A)}{l^2(C,\emptyset)} \\
&= 1 - \frac{\sqrt{(a_1-0.5)^2+...+(a_{12}-0.5)^2}}{\sqrt{(0.5-0)^2+...+(0.5-0)^2}} \\
&= 1 - \frac{\sqrt{(a_1-0.5)^2+...+(a_{12}-0.5)^2}}{\sqrt{3}}.
\end{aligned}
$$

In FPS we have that

$$a_1 + ... + a_{12} = 3.$$

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

$$entropy_{l^2}(A) = 1 - \frac{\sqrt{(a_1-0.5)^2+...+(a_{12}-0.5)^2}}{\sqrt{3}}$$

with the restriction

$$a_1 + ... + a_{12} = 3.$$

we have a maximum of the entropy at the point E.

As above, we can see that

$$entropy_{l^1}(A) = 1 - \frac{l^1(C,A)}{l^1(C,\emptyset)} = 1 - \frac{|a_1-0.5|+...+|a_{12}-0.5|}{6}$$

with the restriction

$$a_1 + ... + a_{12} = 3.$$

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 *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 *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.*

	T	C	A	G
First base	48995	42525	78622	70293
Second base	73438	46554	86585	33858
Third base	77233	54942	62523	45737

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

	T	C	A	G
First base	0.2038	0.1769	0.327	0.2923
Second base	0.3054	0.1936	0.3601	0.1408
Third base	0.3212	0.2285	0.26	0.1902

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.

metric	*M.tuberculosis*	*E.coli*	*A.aeolicus*	*M.pneumoniae*
l^2	0.475876	0.488814	0.478769	0.482055
l^1	0.500033	0.499967	0.499917	0.499967
l	0.500033	0.499967	0.499917	0.499967

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_1 (for more details on the notions of K and H_1 see Giulia Menconi (2005)). What is of interest is that in the case of $entropy_{l^1}$ and $entropy_l$ results are practially identical, while $entropy_{l^2}$ results in a small but identifiable difference between the two genomes in the same sense like K and H_1.

Table 7. Comparison of results obtained using $entropy_l$, $entropy_{l^1}$, $entropy_{l^2}$ with the results of complexity K and probabilistic entropy H_1 of Giulia Menconi (2005) in the case of A.aeolicus and E.coli.

Genome	K	H_1	$entropy_{l^2}$	$entropy_{l^1}$	$entropy_l$
A.aeolicus	1.883	1.976	0.478	0.499	0.499
E.Coli	1.893	1.987	0.489	0.499	0.499

3 Clarity and Fuzzy Polynucleotide Space

Definition 2. Let d be a metric in I^n, $X = \{x_1, ..., x_n\}$ a set and $A = (a_1, ..., a_n)$, where $a_i \in [0, 1]$, a fuzzy set of X. *The clarity of A, with respect to the metric d*, denoted by $clarity_d(A)$ is defined to be the number

$$1 - entropy_d(A),$$

that is

$$clarity_d(A) = 1 - entropy_d(A).$$

Example 11. Let $X = \{x_1, x_2\}$ be a set, $A = (0.4, 0.8)$ a fuzzy set of X and consider the metric space (I^2, l^1). Then, we have

$$clarity_{l^1}(A) = 1 - entropy_{l^1}(A) = 1 - 0.6 = 0.4.$$

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

$$clar(A) = 1 - ent(A) = 1 - \frac{3}{7} = \frac{4}{7} = 0.5714.$$

We observe that

$$clarity_{l^1}(A) = 0.4 \neq clar(A) = 0.5714.$$

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

1) $entropy_d(A), clarity_d(A) \in [0, 1]$.

2) $entropy_d(A) = 1 - clarity_d(A)$.

3)

$$clarity_{l^1}(A) = \frac{l^1(A, C)}{c(C)}.$$

4)

$$clarity_{l^1}(A) = \frac{2l^1(A, C)}{n}.$$

5)

$$clarity_{l^1}(A) = \frac{c(C) - c(A \wedge A^c)}{c(C)}.$$

6)

$$clarity_l(A) = \frac{2l(A, C)}{\sqrt{n}}.$$

7)

$$clarity_l(A) = l(A, C).$$

Proof. Follows by Theorems 1 and 2 and by fact that $clarity_d(A) = 1 - entropy_d(A)$.

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 $a = l^2(A, C)$ divided by the distance $b = l^2(C, \emptyset)$, that is

$$clarity_{l^1}(A, B) = \frac{a}{b}.$$

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

$$entropy_d(A) = entropy_d(B) = 0$$

and

$$clarity_d(A) = clarity_d(B) = 1,$$

where $d = l^1$ or $d = l^2$. Also,

$$entropy_l(A) = entropy_l(B) = \frac{1}{3}$$

and

$$clarity_l(A) = clarity_l(B) = 1 - \frac{1}{3} = \frac{2}{3}.$$

(2) Let $X = \{x_1, ..., x_n\}$ be a set and A a fuzzy set of X such that $A = C$. Then $entropy_d(A) = 1$ and $clarity_d(A) = 0$.

(3) We consider the following polynucleotide sequences:

s_1=UACUGU (tyrosine/cysteine)

s_2=CACUGU (histidine/cysteine)

s_3=CUCUGU (leucine/cysteine)

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

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

$s_2 = (0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0)$

$s_3 = (0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0)$

Using the distances l^2, 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.

metric	s_1	s_2	s_3
l^2	1	1	1
l^1	1	1	1
l	0.8	0.8	0.8

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.

metric	s_1	s_2	s_3
l^2	0.816497	0.707107	0.707107
l^1	0.66667	0.5	0.5
l	0.67385	0.5	0.5

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 *M.tuberculosis*, *E.coli*, *A.aeolicus*. and *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 *M.tuberculosis*, *E.coli*, *A.aeolicus* and *M.pneumoniae*.

metric	M.tuberculosis	E.coli	A.aeolicus	M.pneumoniae
l^2	0.524124	0.511186	0.521231	0.57945
l^1	0.499967	0.500033	0.500083	0.500033
l	0.499967	0.500033	0.500083	0.50

4 Conclusions

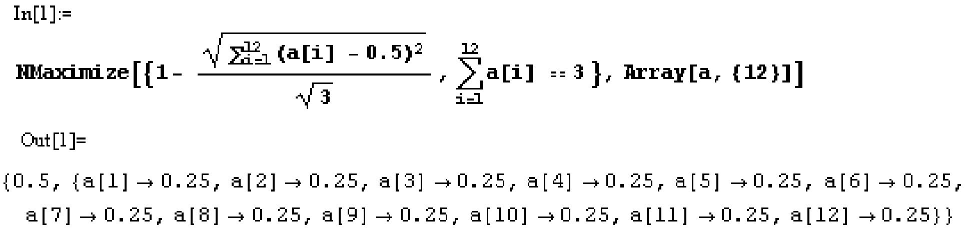
We present results concerning the notions of fuzzy entropy and clarity of a polynucleotide. We propose a new definition of entropy and we consider several applications using different metrics in the case of $I^{12\times k}$ space, where k is the number of codons of the polynucleotide. We also examine the behavior of these notions when those polynucleotides are projected in the I^{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]:=

$$\texttt{NMaximize}\left[\left\{1-\frac{\sqrt{\sum_{i=1}^{12}(\texttt{a[i]}-0.5)^2}}{\sqrt{3}},\ \sum_{i=1}^{12}\texttt{a[i]}==3\right\},\ \texttt{Array[a, \{12\}]}\right]$$

Out[1]=

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

References

Aguero-Chapin G., Gonzalez-Diaz H., Molina R., Varona-Santos J., Uriarte E., Gonzalez-Diaz Y., 2006, Novel 2D maps and coupling numbers for protein sequences. The first QSAR study of polygalacturonases; isolation and prediction of a novel sequence from Psidium guajava L. FEBS Lett., Feb 6;580(3):723-30.

Aguero-Chapin G., Gonzalez-Diaz H., Riva G. D., Rodriguez E., Sanchez-Rodriguez A., Podda G., Vazquez-Padron R. I., 2008, MMM-QSAR Recognition of Ribonucleases without Alignment: Comparison with an HMM Model and Isolation from Schizosaccharomyces pombe, Prediction, and Experimental Assay of a New Sequence. J Chem Inf Model., Feb 25;48(2):434-448.

Bardossy A. and Duckstein L., 1995, Fuzzy Rule-Based Modeling with Applications to Geophysical, Biological and Engineering Systems, CRC Press, Boca Raton.

Bezdek J.C., 1981, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York.

Chen C., Zhou X., Tian Y., Zou X., Cai P., 2006, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network, Anal Biochem 357, 116-121.

Chen, C., Tian, Y. X., Zou, X. Y., Cai, P. X. and Mo, J. Y., 2006b, Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol, 243, 444-448.

Chen, C., Zhou, X., Tian, Y., Zou, X. and Cai, P., 2006a, Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem, 357, 116-121.

Chen, Y. L. and Li, Q. Z., 2007a, Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. Journal of Theoretical Biology, 248, 377-381.

Chen, Y. L. and Li, Q. Z., 2007b, Prediction of the subcellular location of apoptosis proteins. Journal of Theoretical Biology, 245, 775-783.

Chou K. C., 1995, A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space, Proteins: Structure, Function & Genetics 21, 319-344.

Chou K. C., 2000b, Review: Prediction of protein structural classes and subcellular locations, Current Protein and Peptide Science 1, 171-208.

Chou K. C., Prediction of protein cellular attributes using pseudo amino acid composition, PROTEINS: Structure, Function, and Genetics (Erratum: ibid., 2001, Vol.44, 60) 43 (2001), 246-255.

Chou K. C., Prediction of protein subcellular locations by incorporating quasi-sequence-order effect, Biochemical & Biophysical Research Communications 278 (2000), 477-483.

Chou K. C., Prediction of G-protein-coupled receptor classes, Journal of Proteome Research 4 (2005), 1413-1418.

Chou, K.-C., 2009, Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology, Current Proteomics 6, 262-274.

Chou, K. C. and Cai, Y.D., 2003, Predicting protein quaternary structure by pseudo amino acid composition. PROTEINS: Structure, Function, and Genetics, 53, 282-289.

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

Chou K. C., Cai Y. D., Prediction of membrane protein types by incorporating amphipathic effects, Journal of Chemical Information and Modeling 45 (2005), 407-413.

Chou K. C., Cai Y. D., Predicting protein-protein interactions from sequences in a hybridization space, Journal of Proteome Research 5, (2006) 316-322.

Chou K. C., Cai Y. D., Zhong W. Z., 2006b, Predicting networking couples for metabolic pathways of Arabidopsis, EXCLI Journal 5, 55-65.

Chou K. C., D. W. Elrod, 2002, Bioinformatical analysis of G-protein-coupled receptors, Journal of Proteome Research 1, 429-433.

Chou K. C., Elrod D. W., 1999, Protein subcellular location prediction, Protein Engineering 12, 107-118.

Chou K. C., Elrod D. W., 1999b, Prediction of membrane protein types and subcellular locations, PROTEINS: Structure, Function, and Genetics 34, 137-153.

Chou K. C., Elrod D. W., 2003, Prediction of enzyme family classes, Journal of Proteome Research 2 (2003) 183-190.

Chou, K. C. and Shen, H. B., 2007a, Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. Journal of Proteome Research, 6, 1728-1734.

Chou, K. C. and Shen, H. B., 2007a, Review: Recent progresses in protein subcellular location prediction. Analytical Biochemistry, 370, 1-16.

Chou, K. C. and Shen, H. B., 2007b, Large-scale plant protein subcellular location prediction. Journal of Cellular Biochemistry, 100, 665-678.

Chou, K. C. and Shen, H. B., 2007bb, MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm, 360, 339-345.

Chou, K. C. and Shen, H. B., 2007c, Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm, 357, 633-640.

Chou, K. C. and Shen, H. B., 2008, Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms. Nature Protocols, 3, 153-162.

Chou, K. C., 2005, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics, 21, 10-19.

Chechetkin V. R., 2003, Block structure and stability of the genetic code. J. Theoretical Biology 222, 177-188

DasGupta B., Jiang T., Kannan S. and Sweedyk E., On the complexity and approximation of syntenic distance, Discrete Applied mathematics 88 (1998), 59-82.

Ding, Y. S., Zhang, T. L. and Chou, K. C., 2007, Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein & Peptide Letters, 14, 811-815.

Du, P. and Li, Y., 2006, Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics, 7, 518.

A. Dress and T. Lokot, A simple proof of the triangle inequality for the NTV metric, Applied Mathematics Letters 16 (2003), 809-813.

A. Dress, T. Lokot, and L. D. Pustyl'nikov, A new scale-invariant Geometry of L_1 space, Applied Mathematics Letters 17 (2004), 815-820.

A. De Luca and S. Termini, A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory, Inform. and Control 20 (1972), 301-312.

Engelking R., General Topology, Warszawa 1977.

Foster M., Heath A. and Afzal M., Application of distance geometry to 3D visualization of sequence relation-ships, Bionformatics 15 (1999), 89-90.

Feng Z. P., 2002, An overview on predicting the subcellular location of a protein, In Silico Biol 2, 291-303.

Freeland S.J. and Hurst L.D., 1998, The Genetic Code is one in a million, Journal of Molecular Evolution, 47, 238-248.

Gusev V.D., Nemytikova L.A. and Chuzhanova N.A., On the complexity measures of genetic sequences, Bioinformatics 15(1999), 994-999.

Giulia Menconi, Sublinear growth of information in DNA sequences, Bulletin of Mathematical Biology, 67 (2005), 737-759.

Georgiou D.N., Karakasidis T.E., Nieto J.J., and Torres A., A study of genetic sequences using Metric Spaces and Fuzzy Sets, Preprint. SOS

Gonzalez-Diaz H., Aguero-Chapin G., Varona J., Molina R., Delogu G., Santana L.,Uriarte E., Podda G., 2007b, 2D-RNA-coupling numbers: a new computational chemistry approach to link secondary structure topology with biological function. J Comput Chem., Apr 30;28(6):1049-56.

Gonzalez-Diaz H., Gonzalez-Diaz Y., Santana L., Ubeira F. M., Uriarte E., 2008, Proteomics, networks and connectivity indices. Proteomics., Feb;8(4):750-78.

Gonzalez-Diaz H., Perez-Castillo Y., Podda G. 2007a, Uriarte E.Computational chemistry comparison of stable/nonstable protein mutants classification models based on 3D and topological indices.J Comput Chem., Sep;28(12):1990-5.

Gonzalez-Diaz H., Vilar S., Santana L., Uriarte E. 2007 Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. Curr Top Med Chem., 7(10):1015-29.

Guo Y. Z., Li M., Lu M., Wen Z., Wang K., Li G., Wu J., 2006, Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform, Amino Acids 30, 397-402.

Gao Y., Shao S. H., Xiao X., Ding Y. S., Huang Y. S., Huang Z. D., Chou K. C., 2005, Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter, Amino Acids 28, 373-376.

Gu, F., Chen, H. 2009, Evaluating long-term relationship of protein sequence by use of D-interval conditional probability and its impact on protein structural class prediction. Protein and Peptide Letters 16 (10), 1267-1276.

Hegalson C.M. and Jobe T.H., The fuzzy cube and causal efficacy: Representation of concomitant mechanisms in stroke, Neural Networks 11(1998), 549-555.

Hashimoto H., 1983, Szpilrajn's theorem on fuzzy orderings, Fuzzy Sets and Systems, 10, 101-108.

Homaeian L., Kurgan L. A., Cios K. J., Ruan J, Chen K., 2007, Prediction of Protein Secondary Structure Content for the Twilight Zone Sequences. Proteins, 69(3):486-498

Jiang T., Lin G., Ma B. and Zhang K., A general edit distance between RNA structures, Journal of Computational Biology 9(2002), 371-388.

Jiu-Liun Fan and Yuan-Liang Ma, Some new fuzzy entropy formulas, Fuzzy Sets and Systems 128 (2002), 277-284.

Jamshidi N., Edwards J.S., Fahland T., Church G.M. and Palsson B.O., Dynamic simulation of the human red blood cell matabolic network, Bioinformatics 17(2001), 286-287.

Karakasidis T. E. and Georgiou D. N., 2004, Partitioning elements of the periodic table via fuzzy clustering technique, Soft Computing (Springer-Verlag), 8, 231-236.

Kawashima, S. and Kanehisa, M., 2000, AAindex: amino acid index database. Nucleic Acids Res., 28, 374

Kawashima, S., Ogata, H., and Kanehisa, 1999, M.; AAindex: amino acid index database. Nucleic Acids Res., 27, 368-369

Kedarisetti K, Kurgan L, Dick S, 2006, Classifier Ensembles for Protein Structural Class Prediction with Varying Homology. Biochemical and Biophysical Research Communications, 348(3):981-988

Klir G.J. and Yuan B., 1995, Fuzzy Sets and Fuzzy Logic (Theory and Applications), Prentice Hall PRT New Jersey.

Kurgan L. A., Stach W., Ruan J., 2007, Novel scales based on hydrophobicity indices for secondary protein structure. J. Theoretical Biology 248, 354-366

Kurgan L., Chen K., 2007, Prediction of Protein Structural Class for the Twilight Zone Sequences. Biochemical and Biophysical Research Communications, 357(2):453-460

Kosko B., Neural Networks and Fuzzy Systems, Prenyice-Hall, Englewood Cliffs, NJ, (1992).

Lin C.T., Adaptive subsethood for radial basis fuzzy systems. In Kosko, B. (ed.), Fuzzy Engineering, Prentice-Hall, Upper Saddle River, NJ,(1997), 429-464.

M. Liabres and F. Rossello, A new family of metrics for biopolymer contact structures, Computational Biology and Chemistry 28 (2004), 21-37.

Liben-Nowell D., On the structure of syntenic distance, Journal of Computational Biology 8(2001), 53-67.

Li M., Badger J.H., Chen X., Kwong S., Kearney P. and Zhang H., An information-based sequence distance and its application to whole mitochondrian phylogeny, Bioinformatics 17 (2001), 149-154.

Lin Z, Pan X. 2001, Accurate prediction of protein secondary structural content. J Protein Chem., 20:217-220.

Lin, H. and Li, Q. Z., 2007a, Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. Journal of Computational Chemistry, 28, 1463-1466.

Lin, H. and Li, Q. Z., 2007b, Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun, 354, 548-551.

Lin, W.-Z., Xiao, X., Chou, K.-C., 2009, GPCR-GIA: A web-server for identifying G-protein coupled receptors and their families with grey incidence analysis, Protein Engineering, Design and Selection 22, 699-705.

Liu H., Wang M., Chou K. C., 2005, Low-frequency Fourier spectrum for predicting membrane protein types, Biochem Biophys Res Commun 336, 737-739.

Morgenstern B., A simple and space-efficient fragment-chaining algorithm for alignment of DNA and protein sequences, Appl. Math. Lett. 15(1)(2002), 11-16.

V. Moulton, M. Zuker, M. Steel, R. Pointon and D. Penny, Metrics on RNA secontary structures, Journal of Computational Biology, Vol. 7, No. 1/2 (2000), 277-292.

Mondal S., Bhavna R., Mohan Babu R., Ramakumar S., 2006, Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification, J Theor. Biol. 243, 252-260.

Mocz G., 1995, Fuzzy cluster analysis of simple physicochemical properties of amino acids for recognizing secondary structure in proteins, Protein Science 4, 1178-1187.

Mundra, P., Kumar, M., Kumar, K.K., Jayaraman, V.K. and Kulkarni, B.D., 2007, Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. Pattern Recognition Letters, 28, 1610-1615.

Nieto J.J., Torres A. and Vazquez-Trasande M.M., A metric space to study differences between polynucleotides, Appl. Math. Lett., 16 (2003), 1289-1294.

Nieto J.J. and Torres A., Midpoints for fuzzy sets and their application in medicine, Artificial Inteligence in Medicine 17(2003), 81-101.

Nieto J.J., Torres A., Georgiou D.N, and Karakasidis T.E., Fuzzy Polynucleotide spaces and Metrics, Bull. Math. Biology 68 (2006), 703-725.

Paun Gh., Rozenberg G. and A. Saloma, DNA Computing: New Computing Paradigms, Springer, Berlin, (1998).

Percus J., Mathematics of Genome Analysis, Gambridge University Press, Cambridge, (2002).

Perez-Montoto, L.G., Santana, L., Gonzalez-Diaz, H., 2009, Scoring function for DNA-drug docking of anticancer and antiparasitic compounds based on spectral moments of 2D lattice graphs for molecular dynamics trajectories, European Journal of Medicinal Chemistry 44 (11), 4461-4469.

Schneider G., Wrede P., The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site, Biophysical Journal 66 (1994), 335-344.

Shen, H. B. and Chou, K. C., 2005, Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochemical & Biophysical Research Communications, 334, 288-292.

Shen, H. B. and Chou, K. C., 2006, Ensemble classifier for protein fold pattern recognition. Bioinformatics, 22, 1717-1722.

Shen, H. B. and Chou, K. C., 2007a, Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun, 355, 1006-1011.

Shen, H. B. and Chou, K. C., 2007b, EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm, 364, 53-59.

Shen, H. B. and Chou, K. C., 2007c, Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm, 363, 297-303.

Shen, H. B., Yang, J. and Chou, K. C., 2006, Fuzzy KNN for predicting membrane protein types from pseudo amino acid composition. Journal of Theoretical Biology, 240, 9-13.

Shen, H.-B., Chou, K.-C., 2009, Gpos-mploc: A top-down approach to improve the quality of predicting subcellular localization of gram-positive bacterial proteins, Protein and Peptide Letters 16,1478-1484.

Shen, H.-B., Chou, K.-C.,2009, A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0, Analytical Biochemistry 394, 269-274.

Shen, H. B. and Chou, K. C., 2005, Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Comm, 337, 752-756.

Sadegh-Zadeh K., Fuzzy genomes, Artificial Intelligence in Medicine 18 (2000), 1-28.

Michael G. Sadovsky, The method to compare nucleotide sequences based on the minimum entropy principle, Bulletin of Mathematical Biology 65 (2003), 309-322.

Shannon C.E., A Mathematical Theory of Communiction. The Bell Systems Technical Journal, 27 (1948), 379-423.

Sadegh-Zadeh K., Fundamentals of clinical methodology: 3. Nosology, Artificial Inteligence in medicine 17(1999), 87-108.

Samaras P., Kungolos A., Karakasidis T., Georgiou D., Perakis K., 2001, Statistical Evaluation of PCDD/F Emission Data During Solid Waste Combustion by Fuzzy Clustering Techniques, Journal of Environmental Science and Health, Marcel Dekker, Inc.(part A), 36, 153-161.

Stephen Y. L., Freeland J., 2008, A quantitative investigation of the chemical space surrounding amino acid alphabet formation. J. Theoretical Biology 250, 349-361

Tang B., Evaluation of some DNA cloning strategies, Computers Math. Applic. 39(11)(2000), 43-48.

Torres A. and Nieto J.J., The fuzzy polynucleotide space:basic properties, Bioinformatics, Vol. 19, No. 5 (2003), 587-592.

Terano T., Asai K., and Sugeno M., 1992, Fuzzy Systems Theory and its Applications, Academic Press, Harcount Brace Jovanovich Publishers, San Diego, California.

Torres A., and Nieto J. J., 2006, Fuzzy logic in medicine and bioinformatics, Journal of Biomedicine and Biotechnology, Article ID 91908.

Vilar, S., Gonzalez-Diaz, H., Santana, L., Uriarte, E., 2009, A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer, Journal of Theoretical Biology 261 (3), 449-458.

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

Wang M., Yang J., Liu G. P., Xu Z. J., Chou K. C., 2004, Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition, Protein Engineering, Design, and Selection 17, 509-516.

Wang S. Q., Yang J., Chou K. C., 2006, Using stacked generalization to predict membrane protein types based on pseudo amino acid composition, Journal of Theoretical Biology 242, 941-946.

Xiao X., Shao S. H., Huang Z. D., Chou K. C., Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, Journal of Computational Chemistry 27, (2006) 478-482.

Xiao X., Shao S., Ding Y., Huang Z., Chen X., Chou K. C., Using cellular automata to generate Image representation for biological sequences, Amino Acids 28 (2005), 29-35.

Xiao X., Shao S., Ding Y., Huang Z., Chen X., Chou K. C., An Application of Gene Comparative Image for Predicting the Effect on Replication Ratio by HBV Virus Gene Missense Mutation, Journal of Theoretical Biology 235 (2005), 555-565.

Xiao X., Shao S., Ding Y., Huang Z., Huang Y., Chou K. C., 2005, Using complexity measure factor to predict protein subcellular location, Amino Acids 28 (2005), 57-61.

Xiao X., Shao S. H., Ding Y. S., Huang Z. D., Chou K. C., Using cellular automata images and pseudo amino acid composition to predict protein sub-cellular location, Amino Acids 30, (2006) 49-54.

Xiao, X., Lin, W.Z., and Chou, K.C., Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29, (2008) 2018-2024.

Xiao, X., Wang, P., and Chou, K.C.,Predicting protein structural classes with pseudo amino acid composition: an approach using geometric moments of cellular automaton image. J Theor Biol 254,(2008) 691-696.

Xiao, X., Wang, P., and Chou, K.C., Predicting protein quaternary structural attribute by hybridizing functional domain composition and pseudo amino acid composition. J Appl Crystallogr 42, (2009) 169-173.

Xiao, X., Wang, P., Chou, K. C., 2009b. GPCR-CA: A cellular automaton image approach for predicting G-protein-coupled receptor functional classes. Journal of Computational Chemistry. 30 (2009) 1414-1423.

Xiao, X., Lin, W. Z., Application of protein grey incidence degree measure to predict protein quaternary structural types. Amino Acids.37 (2009) 741-749.

Xiao, X., Wang, P., and Chou, K.C., Quat-2L: a web-server for predicting protein quaternary structural attributes. Molecular Diversity, (2010) DOI 10.1007/s11030-010-9227-8.

Zhang T. L., Ding Y. S., Chou K. C., 2008, Prediction protein structural classes with pseudo-amino acid composition: Approximate entropy and hydrophobicity pattern. J. Theoretical Biology 250, 186-193

Zhang Z. D., Sun Z. R., Zhang C. T. 2001, A new approach to predict the helix/strand content of globular proteins. J Theor Biol, 208:65-78.

Zhou, X. B., Chen, C., Li, Z. C. and Zou, X. Y., 2007, Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. Journal of Theoretical Biology, 248, 546-551.

Zimmermann H. J., 1991, Fuzzy Theory and its Applications, Kluwer Academic Publishers, New York.

Zaus M., Crisp and Soft Computing with Hypercubical Calculus, Physica-Verlag, Heideberg, (1999).

Vilar, S., Gonzalez-Diaz, H., Santana, L., Uriarte, E., 2009, A network-QSAR model for prediction of genetic-component biomarkers in human colorectal cancer, Journal of Theoretical Biology 261, 449-458.

Trinquier G. and Sanejouand Y-H.i, 1998, Which effective property of amino acids is best preserved by the genetic code?, Protein Engineering 11, 153169.