Journal of Theoretical Biology 267 (2010) 95–105

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

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

ARTICLE INFO

Article history:
Received 21 January 2010
Received in revised form
22 July 2010
Accepted 6 August 2010
Available online 11 August 2010

Keywords:
Polynucleotides
Fuzzy sets
Metric spaces
Entropy
Clarity

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 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.

1. Introduction

Bioinformatics is a relatively new discipline (see Jamshidi et al., 2001; Morgenstern, 2002; Paun et al., 1998; Percus, 2002; Tang, 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) (Freland and Hurst, 1998). Sadegh-Zadeh (see Sadegh-Zadeh, 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 $(a_1,\ldots,a_{12})$ 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, 2000) introduced the Fuzzy Polynucleotide Space (FPS) based on the principle of the fuzzy hypercube (Kosko, 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 and Nieto, 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 (1997) and Sadegh-Zadeh (see Sadegh-Zadeh, 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 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 Moulton et al., 2000) or biopolyment contact structures (see Liabres and Rossello, 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 Stephen and Freeland, 2008;

* Corresponding author.
E-mail address: thkarak@uth.gr (T.E. Karakasidis).

0022-5193/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2010.08.010

Kedarisetti et al., 2006; DasGupta et al., 1998; Chechetkin, 2003; Chou, 1995, 2000b; Foster et al., 1999; Gusev et al., 1999; Jiang et al., 2002; Liben-Nowell, 2001; Li et al., 2001; Mocz, 1995). 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 (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 (Kawashima et al., 1999; Kawashima and Kanehisa, 2000; Trinquier and Sanejouand, 1998). In an earlier paper (Chou, 2000a), the physicochemical distance among the 20 amino acids (Schneider and Wrede, 1994) was adopted to define PseAA. Subsequently, some investigators used complexity measure factor (Xiao et al., 2005a), some used the values derived from the cellular automata (Xiao et al., 2005b,c, 2006a,b), some used hydrophobic and/or hydrophilic values (Wolfenden, 2007; Zhang et al., 2001; Chou, 2005b; Feng, 2002; Wang et al., 2006, 2004; Gao et al., 2005; Chen et al., 2006a; Kurgan et al., 2007; Kurgan and Chen, 2007) and some were through Fourier transform (Liu et al., 2005; Perez-Montoto et al., 2009; Guo et al., 2006), as well as trough 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 and Cai, 2005, 2006; Chou, 2001; Chou and Elrod, 1999a, 2002, 2003; Shen and Chou, 2009a, b), as well as for enzyme functional class (Chou, 2005b; Chou and Elrod, 1999b). Work using pseudo amino acid composition has also been performed (Xiao et al., 2008a,b, 2009a; Zhang et al., 2008; Lin and Pan, 2001). 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,b,d), 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,b; Lin and Li, 2007a; Ding et al., 2007; Gu and Chen, 2009; Homaeian et al., 2007), protein subcellular localization (Chou and Shen, 2008, 2007a,b,Shen and Chou, 2007a), protein subnuclear localization (Shen and Chou, 2005a,b; Mundra et al., 2007) protein submitochondria localization (Du and Li, 2006), protein oligomer type (Chou and Cai, 2003), conotoxin superfamily classification (Mondal et al., 2006; Lin and Li, 2007b) membrane protein type (Liu et al., 2005; Shen and Chou, 2005a; Wang et al., 2006; Shen et al., 2006; Chou and Shen, 2007c) apoptosis protein subcellular localization (Chen and Li, 2007a,b) enzyme functional classification (Chou, 2005a; Chou and Cai, 2004; Zhou et al., 2007; Shen and Chou, 2007b) protein fold pattern (Shen and Chou, 2006), and signal peptide (Chou et al., 2006; Chou and Shen, 2007e; 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., 2007c, 2007b, 2007a; Aguero-Chapin et al., 2006; Vilar et al., 2009; Georgiou et al., 2009) as well as more complex work such as Xiao and Lin, 2009, 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 (2000). We note that it is possible to compare sequences using a minimum entropy principle (Sadovsky, 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 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 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

2.1. 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 (for details see Bardossy and Duckstein, 1995; Bezdek, 1981; Klir and Yuan, 1995; Hashimoto, 1983; Terrano et al., 1992; Zimmermann, 1991)

(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 $X=\{x_1,x_2, \ldots,x_n\}$ and

$$A = \{(x_1,\mu_A(x_1)), \ldots, (x_n,\mu_A(x_n))\},$$

then we write

$$A = (\mu_A(x_1), \ldots, \mu_A(x_n)).$$

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 (1992) introduced a geometrical interpretation of fuzzy sets as points in a hypercube. Indeed, for a given set $X=\{x_1,x_2, \ldots,x_n\}$, 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

$$P = (\mu_A(x_1), \ldots, \mu_A(x_n)).$$

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

Nonfuzzy or crisp subsets of $X=\{x_1,\ldots,x_n\}$ 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