Proc. of the Intl. Conf. on Recent Trends In Computing and Communication Engineering -- RTCCE 2013
Copyright © Institute of Research Engineers and Doctors. All rights reserved.
ISBN: 978-981-07-6184-4

Some anomalies in the analysis of whole genome sequence on the basis of Fuzzy set theory

Subhram Das[1], Debanjan De[2], Anilesh Dey[3] and D. K. Bhattyachrya[4]

Abstract— **The purpose of the paper is to create some counter examples in order to find out some of the anomalies and inadequacies in the analysis of polynucleotide and whole genomes on the basis of fuzzy set theory.**

***Keywords*— Fuzzy Polynucleotide space, Different types of metric**

I. Introduction

Nucleic acids DNA and RNA are the genetic material of living organisms. There are two basic techniques used in the analysis of genetic material with applications in diagnosis and taxonomy: (a) sequence analysis which is used to determine the building blocks of a nucleic acid, called nucleotides and their order in the molecular chain, and (b) sequence comparison used to identify the degree of difference/similarity between polynucleotides to identify similarity with known viruses. We stick to the discussion of second technique only. DNA and RNA are made of codons, each of which is a triplet of nucleotides, having the possibility to be one of four nucleotides {T, C, A, G} in the case of DNA and {U, C, A, G} in the case of RNA (A: adenine; C: cytosine; G: guanine; T: thymine; U: uracil). So far as representation of a codon, either of DNA or RNA in concerned, it is a problem of representing three nucleotides out of four. For example when we say that we understand U fully in a RNA codon, we mean that we do not understand C, A and G at all. So we represent U as (1, 0, 0, 0). Similarly C, A and G are represented respectively as (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1). Obviously a codon is represented on a twelve dimensional hypercube I^{12}. For example, CAG is represented on I^{12} as (0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1). Also we note that each nucleotide occurs at one of the corners of the hypercube. Normally there is no problem in representation if a single codon like CAG is chosen. But there are cases where the exact chemical structure of the sequence is not known. For example for the codon XAU, where X = (0.2, 0.4, 0.2, 0.1, 0, 0, 1, 0, 1, 0, 0, 0), the first letter X is unknown and corresponds to U to extent 0.2, C to extent 0.4, A to extent 0.2 and G to extend 0.1. Thus one may deal with base sequences not necessarily at a corner of the hypercube. In this case some components of its code is neither 0 nor 1 but a value in the interval (0, 1). Hence in such cases crisp representation of codon in I^{12} fails. The problem becomes more prominent if we like to represent a polynucleotide consisting of finitely many codons or a whole genome consisting of infinitely many such codons. When one takes a polynucleotide, which is a sequence of k triplets, one would need a I^{12xk} hypercube. For example if we have the polynucleotide described by the sequence UACUGU (tyrosin/cysteine), it is a point in $I^{2\times12} = I^{24}$. Similarly if there are four thousand codons in a polynucleotide, then such representation is possible in a 12x4000 dimensional hypercube. Obviously the size of the hypercube is very large and it becomes larger and larger as the number of codons in the polynucleotide increases. The process becomes unmanagable; this is definitely a drawback in the representation. The second and most important difficulty arises when we try to compare two polynucleotides of different lengths. Obviously both types of difficulties could be avoided, if the representation could be made on a single I^{12}. In fact this is the reason why, for representation of a polynucleotide a hypercube I^{12} is chosen. As a matter of fact, necessity of introducing fuzzy set theory is realized in the process of representing a polynucleotide consisting of finite number of codons, n say, on a single hypercube I^{12}. This is the background of fuzzy polynucleotide space as introduced by Torres and Nieto (2003) [1]. Thus the codon XYZ representing the polynucleotide indicates that neither of X or Y or Z is fully understood. By X we understand $.\alpha,.\beta,.\gamma,.\delta$ of U, C, A and G respectively, where $.\alpha+.\beta+.\gamma+.\delta=1$. Similar results hold for Y and Z. On the basis of this assumption, Torres and Nieto (2003) [3] introduced the notion of fuzzy polynucleotide space based on the principle of the fuzzy hypercube Kosko, (1992) [6]. Thus a polynucleotide is represented by the frequencies of the nucleotides at the three base sites of a codon in the coding sequence. The idea of differentiating polynucleotide and whole genomes on the basis of fuzzy set theory is well understood from the work of Angela Torres and Juan J. Nieto in (2003) [1], where they have used the metric as introduced

[1] Subhram Das
Computer Science & Engineering, Narula Institute of Technology, Kolkata
India

[2] Debanjan De
Quality Control Officer, Pest Control, Kolkata
India

[3] Anilesh Dey
Electronics & Communication Engineering, Narula Institute of Technology, Kolkata
India

[4] D. K. Bhattyachrya
Emaritus Professor, Rabindra Bharati University, Kolkata
India

in (2000) [5]. With the help of this metric they could differentiate polynucleotides and some whole genomes. Later on, in (2006) [2] the authors used different types of metric for comparison of polynucleotides and whole genomes. They could show that in all cases the metrics behaved similarly; this is quite expected as the metrics being defined on a finite dimensional space are all equivalent. In section II, we show some examples of whole genomes, where all the metrics do not behave similarly.

In fact this challenges the very formation of fuzzy polynucleotide space.

II. Some known results

A. Different types of Metric used on polynucleotide spaces [2006] [2]

$$d\left(p,q\right)=\frac{\sum_{i=1}^{12}\left|p_i-q_i\right|}{\sum_{i=1}^{12}\max\left\{p_i,q_i\right\}}\ ---(1)$$

$$d_1(p,q)=\frac{d(p,q)}{1+d(p,q)}\ ---(2)$$

$$d_2(p,q)=\frac{\sqrt{\sum_{i=1}^{12}(p_i-q_i)^2}}{\sqrt{12}}\ --(3)\qquad ,$$

$$d_3(p,q)=\frac{d_2(p,q)}{1+d_2(p,q)}\ ---(4)$$

$$d_4(p,q)=\frac{\sum_{i=1}^{12}\left|p_i-q_i\right|}{12}\ ---(5)\qquad ,$$

$$d_5(p,q)=\frac{\sum_{i=1}^{12}\left|p_i-q_i\right|}{1+\sum_{i=1}^{12}\left|p_i-q_i\right|}\ ---(6)$$

$p = (p_1, p_2, p_3\ldots\ p_{12})$, $q = (q_1, q_2, q_3\ldots\ q_{12}) \in I^{12}$ are two different points.

B. Fuzzy representation of polynucleotide and the role of different types of metric

Example: UACUGU tyrosine / cysteine

No. of Nucleotides		Total	Fraction of Nucleotides
	U C A G		U C A G
1st base	2 0 0 0	2	1 0 0 0
2nd base	0 0 1 1	2	0 0 .5 .5
3rd base	1 1 0 0	2	.5 .5 0 0

So fuzzy representation of S1= UACUGU tyrosine / cysteine is (1,0,0,0,0,0,.5,.5,.5,.5,0,0)

Similarly fuzzy representation of S2 = CACUGU histidine/cysteine is (.5,.5,0,0,0,0,.5,.5,.5,.5,0,0) and fuzzy representation of S3 = CUCUGU leucine/cysteine is (.5,.5,0,0,.5,0,0,.5,.5,.5,0,0)

Remark 1

The authors of [2006] [2] prove that all the metrics *d*, *d_1*, *d_2*, *d_3*, *d_4*, *d_5* behave identically and establish that X and Y are nearer than X and Z. This is also biologically justified as X and Y differ only in the first base, whereas X and Z differ in the first two bases.

C. Fuzzy representation of Complete Genomes and the role of different types of metric

In [2006] [2] the authors have shown that the role of different metrics remains the same in cases of complete genomes also. They have considered fuzzy sets of frequencies of the genome of M. tuberculosis, the genome of E. coli. and the genome of A. Aeolicus. Using the various metrics they have computed the distances between M. tuberculosis and E. coli and also the distances between M. tuberculosis and E. coli with A. Aeolicus.

Remark 2

The results obtained indicate that the various metrics employed in this work present similar behavior to the results obtained using the metric used in Torres and Nieto (2003) [1] for these complete genomes also.

D. Some counter examples

In this article we prove that Remark 2 is not true in general. In fact we take some counter examples of whole genomes, and consider the same fuzzy representation and same metrics as in [2], in order to show that the metrics do not behave identically in all cases. It is also shown that some of the metrics become even meaningless for comparison.

a) The complete genome sequence of Corynebacterium diphtheriae NCTC 13129. It is available at http://www.ncbi.nlm.nih.gov. Its accession number is >gi|38231477|emb|BX248353.1|

The genome comprises of 2488679 base pairs.

b) The complete genome sequence of Haemophilus influenzae 86-028NP. It is available at http://www.ncbi.nlm.nih.gov. Its accession number is >gi|156617157|gb|CP000057.2|

The genome comprises of 1914526 base pairs.

c) The complete genome sequence of Halobacterium sp. NRC-1. It is available at

http://www.ncbi.nlm.nih.gov. Its accession number is >gi|12057215|gb|AE004437.1|

The genome comprises of 2014275 base pairs.

d) The complete genome sequence of Xylella fastidiosa 9a5c. It is available at

http://www.ncbi.nlm.nih.gov. Its accession number is>gi|12057211|gb|AE003849.1|

The genome comprises of 2679306 base pairs.

E. *Proposition*

For the Fuzzy polynucleotide of types (a), (b), (c), (d), the metrics d, d_2, d_4 are not at all feasible for comparison; d_1 and d_5 behave identically; d_3 behaves just opposite to both d_1 and d_5.

Proof:

Fuzzy set of frequencies for genome (a) is

(0.233,0.267,0.233,0.267,0.233,0.265,0.233,0.269,0.232,0.270,0.232,0.266)

Fuzzy set of frequencies for genome (b) is

(0.311,0.189,0.310,0.190,0.310,0.191,0.308,0.191,0.307,0.192,0.309,0.192)

Fuzzy set of frequencies for genome (c) is

(0.164,0.338,0.162,0.336,0.159,0.341,0.161,0.339,0.158,0.341,0.158,0.343)

Fuzzy set of frequencies for genome (d) is

(0.248,0.248,0.228,0.276,0.249,0.248,0.225,0.278,0.246,0.253,0.224,0.277)

The detailed calculations of distances under different metrics are given as follows:

Genome	**d**	**d_1**	**d_2**	**d_3**	**d_4**	**d_5**
C.diphtheriae, H.influenzae	0.265	0.210	0.077	0.071	0.076	0.479
Halobacterium.sp, X.fastidiosa	0.265	0.209	0.077	0.072	0.076	0.478

Proposition follows from the following results on comparison

i) d (C.diphtheriae, H.influenzae) =

d(Halobacterium.sp, X.fastidiosa)

ii) d_1 (C.diphtheriae, H.influenzae) >

d_1 (Halobacterium.sp, X.fastidiosa)

iii) d_2 (C.diphtheriae, H.influenzae) =

d_2 (Halobacterium.sp, X.fastidiosa)

iv) d_3 (C.diphtheriae, H.influenzae) <

d_3 (Halobacterium.sp, X.fastidiosa)

v) d_4 (C.diphtheriae, H.influenzae) =

d_4 (Halobacterium.sp, X.fastidiosa)

vi) d_5 (C.diphtheriae, H.influenzae) >

d_5 (Halobacterium.sp, X.fastidiosa).

III. Results and Discussion

Fuzzy representation of whole genome on I^{12} is definitely a useful tool for ultimate comparison of two genomes of any large size and also being of different lengths. The representation being on I^{12} may be convenient for further analysis with the help of Euclidian metrics on I^{12}. I^{12} being a finite dimensional space, it is expected that all the metrics should behave similarly, as the metrics are all equivalent. This was justified in the previous papers for polynucleotides and whole genomes. But the results were verified only for some particular cases. The natural question therefore remains to see whether such results are true in general. The present paper establishes through some counter examples that the results are not true for all genomes. It may be remarked that the present discussion indirectly opens the questions of modification of the idea of fuzzy representation of whole genomes in order to see that in those represented spaces the metrics behave similarly.

Acknowledgment

I do here by acknowledge Dr. Sanjay Kumar Palit for his suggestion for improvement of the paper.

ISBN: 978-981-07-6184-4

References

[1] Nieto, J.J., Torres, A., Vazquez-Trasande, M.M., 2003. A metric space to study differences between polynucleotides. Appl. Math. Lett. 27, 1289–1294.

[2] Nieto, J.J., Torres, A., Georgiou, D.N., Karakasidis,T.E, 2006. Fuzzy Polynucleotide Spaces and Metrics. Bulletin of Mathematical Biology (2006) 68: 703–725.

[3] Torres, A., Nieto, J.J., (2003). The fuzzy polynucleotide space: Basic properties. Bioinformatics 19(5), 587–592.

[4] L.A. Zadeh, Fuzzy sets, Inform. and Control 8 (1965) 338-353.

[5] Sadegh-Zadeh, K., 2000. Fuzzy genomes. Artif. Intell. Med. 18, 1–28.

[6] Kosko,B. (1992) Neural networks and fuzzy systems. Prentice-Hall, Englewood Cliffs, NJ**.**

About Author (s):

Subhram Das was born in West Bengal, India in 1979. He received the B.Tech degree in Information Technology from Kalyani University (India) in 2003 and M.Tech degree in the same discipline from Calcutta University (India) in 2005. He is working as Assistant Professor in Computer Science & Engineering department at Narula Institute of Technology, Kolkata since 2003. Presently, he is doing his research work under the supervision of Prof. D. K. Bhattacharya. His research topics include Bioinformatics & Computational Biology.

Debanjan De was born in West Bengal, India in 1985. He received the M.Sc. (Tech) degree in Bioinformatics from Jadavpur University (India), DOEACC B-Level and B.Sc (Hons.) in Microbiology from Calcutta University. He is working as Quality Control Officer in Pest Control, Kolkata, India. Presently, he is doing his research work in Calcutta University. His research topics include Bioinformatics & Computational Biology.

Anilesh Dey was born in West Bengal, India in 1977. He received the B.E in Electronics from Nagpur University and M.Tech (Gold-Medallist) in Instrumentation and Control Engineering from Calcutta University. He is working as Assistant Professor of Electronics and Communication Engineering at Narula Institute of Technology, Kolkata since 2006. Presently, he is pursuing Ph.D degree at School of Bioscience and Engineering, Jadavpur University (India) under the supervision of Prof. D. K. Bhattacharya and Prof D.N. Tibarewala. He is author or co-author of more than 9 scientific papers in international/national journals and proceedings of the conferences with reviewing committee. His research topics nonlinear time series analysis, time and frequency domain analysis of bio-medical and music signals, effect of music in autonomic and central nervous system.

D. K. Bhattacharya was born in West Bengal, India in 1943. He is a retired Professor and Head in the department of Pure Mathematics University of Calcutta, India. He is presently an UGC Emeritus Fellow; prior to this he was an AICTE Emeritus Fellow of Govt. of India. He had his undergraduate, postgraduate and doctoral duty from the University of Calcutta. He has a long teaching experience of forty six years; he has supervised many Ph.D. students in Pure and Applied Mathematics. He is author or co-author of about 80 scientific papers in international / national journals and proceedings of the conferences with reviewing committee. His expertise is in Mathematical modelling and optimal control. His present interest is in application of Mathematics in Biology and Medicine including Bio-informatics.