Similarity and Dissimilarity of Whole Genomes using Intuitionistic Fuzzy Logic

Subhram Das, Debanjan De, D. K. Bhattacharya

Abstract—**Whole genomes in general and poly-nucleotides, in particular, have an interesting representation in twelve dimensional hypercube I^{12} based on fuzzy set theory, but it has some limitations and drawbacks. With a view to removing such drawbacks of the representation, the present paper derives some new representation of whole genomes and poly-nucleotides based on Intuitionistic Fuzzy set theory and shows that that such a representation is free from any such limitation as mentioned above. Finally it applies the new representation in testing similarity/ dissimilarities of whole genomes and poly-nucleotides.**

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. It is known that 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 $I^{12\times k}$ hypercube. For example if we have the polynucleotide described by the sequence UACUGU (tyrosin/cysteine), it is a point in $I^{2\times 12} = I^{24}$. Similarly if there are four thousand codons in a polynucleotide, then such representation is possible in a 12×4000 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 unmanageable; 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, had the representation been 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 of Kosko, (1992) [9]. 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 used the metric as introduced in (2000) [8]. With the help of

Subhram Das was with Calcutta University. He is now Computer Science & Engineering Department, Narula Institute of Technology Kolkata 700109, India(subhram@gmail.com)

Debanjan De is now Quality Control Officer, Pest Control India Kolkata, India(debanjande85@gmail.com)

D. K. Bhattacharya was with Calcutta University. After that Emaritus Professor, Rabindra Bharati University Kolkata, India (dkb_math@yahoo.com)

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

this metric they could differentiate polynucleotides and some whole genomes. Later on, in (2006) [2] different types of metric were used for comparison of polynucleotides and whole genomes. They could show that in all cases the metrics behaved similarly; this was quite expected as the metrics being defined on a finite dimensional space were all equivalent. In [10], the present authors showed some examples of whole genomes, where all the metrics did not behave similarly. This is illustrated in brief in section 3. The possible reason is that whenever we understand the frequencies of polynucleotide and whole genome on the unit 12 dimensional hypercube, the information is not complete, as it does not consider hesitation factor, which is always present in real situation. In fact Intuitionistic Fuzzy set (IFS) concept is more robust than that of Fuzzy set; because it always accommodates some kind of hesitancies. In the present paper we use the concept of Intuitionistic Fuzzy set theory to develop Intuitionistic Fuzzy representation of polynucleotide and whole genome. We show that no anomaly occurs in the process of representation and finding out similarity/ dissimilarity even in whole genomes.

II. Some Known Result

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

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

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

$$d_2(p,q)=\left(\sqrt{\sum_{i=1}^{12}(p_i-q_i)^2}\right)/\sqrt{12}$$

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

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

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

p = (p1, p2, p3… p12), q = (q1, q2, q3… q12) $\in I^{12}$ are two different points.

B. *Fuzzy representation of polyneucleotides and the role of different types of metric*

Example: UACUGU tyrosine / cysteine

No. of Nucleotides, Total Nucleotides and Fraction of Nucleotides are given below,

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

1) 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] authors show that the role of different metrics remains the same in cases of complete genomes also. They consider 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*.

2) Remark 2:

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

D. *Some Counter Example*

These examples show that Remark 2 is not true in general.

1) Example:

(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.

2) 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.[10]

III. Intuitionistic fuzzy set

Intuitionistic Fuzzy Sets [4, 5] are generalization of Fuzzy sets [6] in which non-membership values are not obtainable from the membership values, rather both of them have to be specified separately.

A. Definition

Let X is a non empty set. An Intuitionistic fuzzy set A on X is defined as $A = \{< x, \mu_A(x), \nu_A(x) >, x \in X\}$, where the functions $\mu_A : X \to [0,1]$ and $\nu_A : X \to [0,1]$ define respectively the degree of membership and the degree of non-membership of the element x in X to the set A, and $0 \le \mu_A(X) + \nu_A(X) \le 1$ for each x in X. Obviously an ordinary fuzzy set can be written as $\{< x, \mu_A(x), 1 - \mu_A(x) >, x \in X\}$.

In reality non-membership is always associated with some sort of hesitancy. If we fix a fraction θ of membership value as the value of hesitancy, then it is given by $\nu_A(X) = \theta\mu_A(X)$; so non-membership value equals to $\pi_A(X) = 1 - (1+\theta)\mu_A(X)$. Hence an Intuitionistic fuzzy set can be written as $\{< x, \mu_A(x), \nu_A(x), \pi_A(x) >, x \in X\}$,

where $\nu_A(X) = \theta\mu_A(X)$, $\pi_A(X) = 1 - (1+\theta)\mu_A(X)$

B. Distance measure on Intuitionistic fuzzy set

The normalized hamming distance D_{IFS} proposed for IFS by Szmidt and Kacprzyk [7] is given by

$$D_{IFS}(A,B) = \sum_{i=1}^{n} (|\mu_A(x_i) - \mu_B(x_i)| + |\nu_A(x_i) - \nu_B(x_i)| + |\pi_A(x_i) - \pi_B(x_i)|)$$

where A and B are two IFS in X= $\{x_1, x_2, ..., x_n\}$.

Obviously the general form of distance measure would be

$$D^{\alpha}_{IFS}(A,B) = [\sum_{i=1}^{n} (|\mu_A(x_i) - \mu_B(x_i)|^{\alpha} + |\nu_A(x_i) - \nu_B(x_i)|^{\alpha} + |\pi_A(x_i) - \pi_B(x_i)|^{\alpha})]^{\frac{1}{\alpha}}, \alpha \text{ is a natural number}$$

C. Similarity measures on Intuitionistic Fuzzy set [7]

$$S(A,B) = 1 - [1/2n \sum_{j=1}^{n} (\left|\mu_A(x_j) - \mu_B(x_j)\right|)^{\alpha} + (\left|\nu_A(x_j) - \nu_B(x_j)\right|)^{\alpha} + (\left|\pi_A(x_j) - \pi_B(x_j)\right|)^{\alpha}]^{1/\alpha}, \alpha > 0$$

D. Formula of Intuitionistic Fuzzy representation of polynucleotide on a triplet of I^{12}

Suppose fractions of nucleotide at a point on I^{12} be given by $(x_1, x_2, x_3, x_4, y_1, y_2, y_3, y_4, z_1, z_2, z_3, z_4)$. Then the Intuitionistic fuzzy representation of the polynucleotide A is $\{< x, \mu_A(x), \nu_A(x), \pi_A(x) >, x \in X\}$, where

$$\mu_A(x) = (x_1, x_2, x_3, x_4, y_1, y_2, y_3, y_4, z_1, z_2, z_3, z_4)$$

$$\nu_A(x) = (\theta x_1, \theta x_2, \theta x_3, \theta x_4, \theta y_1, \theta y_2, \theta y_3, \theta y_4, \theta z_1, \theta z_2, \theta z_3, \theta z_4)$$

$$\pi_A(x) =$$

$$[\{1-(1+\theta)(x_1)\}, \{1-(1+\theta)(x_2)\}, \{1-(1+\theta)(x_3)\}, \{1-(1+\theta)(x_4)\},$$
$$[\{1-(1+\theta)(y_1)\}, \{1-(1+\theta)(y_2)\}, \{1-(1+\theta)(y_3)\}, \{1-(1+\theta)(y_4)\},$$
$$[\{1-(1+\theta)(z_1)\}, \{1-(1+\theta)(z_2)\}, \{1-(1+\theta)(z_3)\}, \{1-(1+\theta)(z_4)\}]$$

E. Intuitionistic Fuzzy representation, difference and similarity of polynucleotides & whole genome

For simplification of calculation we take $\theta = 0.1$

TABLE I
Intuitionistic Fuzzy Representation of Polynucleotides S1

S1	1	0	0	0	0	0	.5	.5	.5	.5	0	0
	0	0	0	0	0	0	.05	.05	.05	.05	0	0
	0	1	1	1	1	1	.45	.45	.45	.45	1	1

TABLE 2
Intuitionistic Fuzzy Representation of Polynucleotides S2

S2	.5	.5	0	0	0	0	.5	.5	.5	.5	0	0
	.05	.05	0	0	0	0	.05	.05	.05	.05	0	0
	.45	.45	1	1	1	1	.45	.45	.45	.45	1	1

TABLE 3
Intuitionistic Fuzzy Representation of Polynucleotides S3

S3	.5	.5	0	0	.5	0	0	.5	.5	.5	0	0
	.05	.05	0	0	.05	0	0	.05	.05	.05	0	0
	.45	.45	1	1	.45	1	1	.45	,45	.45	1	1

TABLE 4: DISTANCE MEASURE OF INTUITIONISTIC FUZZY REPRESENTATION OF POLYNUCLEOTIDES S1, S2 & S3

Distance	α=1	α=2	α=3	α=4	α=5	α=6	α=7	α=8	α=9	α=10
S1 vs S2	2.1	1.01	0.50775	0.257525	0.131282	0.067234	0.034586	0.017867	0.009268	0.004827
S1 vs S3	4.3	2.12	1.09075	0.56555	0.294439	0.153846	0.08066	0.042427	0.022385	0.011846
S2 vs S3	2.2	1.11	0.583	0.308025	0.163158	0.086611	0.046074	0.024559	0.013117	0.007019

TABLE 5: SIMILARITY MEASURE OF INTUITIONISTIC FUZZY REPRESENTATION OF POLYNUCLEOTIDES S1, S2 & S3

Similarity	α=1	α=2	α=3	α=4	α=5	α=6	α=7	α=8	α=9	α=10
S1 vs S2	0.825	0.91625	0.93351	0.940636	0.94447	0.94686	0.94846	0.94961	0.950463	0.95111
S1 vs S3	0.64166	0.87866	0.9142	0.927734	0.93474	0.939	0.94184	0.94386	0.945364	0.94652
S2 vs S3	0.81666	0.91220	0.93038	0.937918	0.94201	0.94457	0.94631	0.94756	0.948514	0.94924

TABLE 6: INTUITIONISTIC FUZZY REPRESENTATION OF WHOLE GENOME (A),(B),(C) & (D)

(a)	0.233	0.267	0.233	0.267	0.233	0.265	0.233	0.269	0.232	0.27	0.232	0.266
	0.0233	0.0267	0.0233	0.0267	0.0233	0.0265	0.0233	0.0269	0.0232	0.027	0.0232	0.026
	0.7437	0.7063	0.7437	0.7063	0.7437	0.7085	0.7437	0.7041	0.7448	0.703	0.7448	0.707
(b)	0.311	0.189	0.31	0.19	0.31	0.191	0.308	0.191	0.307	0.192	0.309	0.192
	0.0311	0.0189	0.031	0.019	0.031	0.0191	0.0308	0.0191	0.0307	0.0192	0.0309	0.019
	0.6579	0.7921	0.659	0.791	0.659	0.7899	0.6612	0.7899	0.6623	0.7888	0.6601	0.788
(c)	0.164	0.338	0.162	0.336	0.159	0.341	0.161	0.339	0.158	0.341	0.158	0.343
	0.0164	0.0338	0.0162	0.0336	0.0159	0.0341	0.0161	0.0339	0.0158	0.0341	0.0158	0.034
	0.8196	0.6282	0.8218	0.6304	0.8251	0.6249	0.8229	0.6271	0.8262	0.6249	0.8262	0.622
(d)	0.248	0.248	0.228	0.276	0.249	0.248	0.225	0.278	0.246	0.253	0.224	0.277
	0.0248	0.0248	0.0228	0.0276	0.0249	0.0248	0.0225	0.0278	0.0246	0.0253	0.0224	0.027
	0.7272	0.7272	0.7492	0.6964	0.7261	0.7272	0.7525	0.6942	0.7294	0.7217	0.7536	0.695

TABLE 7: DISTANCE MEASURE OF INTUITIONISTIC FUZZY REPRESENTATION OF WHOLE GENOME (A),(B),(C) & (D)

Distance	α=1	α=2	α=3	α=4	α=5	α=6	α=7	α=8	α=9	α=10
(a) vs (b)	2.019	0.15596	0.01254	0.00101	8.24E-05	6.7E-06	5.47E-07	4.47E-08	3.66E-09	3.01E-10
(a) vs (c)	1.909	0.13955	0.01063	0.00081	6.28E-05	4.85E-06	3.76E-07	2.93E-08	2.28E-09	1.79E-10
(a) vs (d)	0.325	0.00455	7.19E-05	1.2E-06	2.09E-08	3.72E-10	6.76E-12	1.24E-13	2.32E-15	4.37E-17
(b) vs (c)	3.929	0.59018	0.09229	0.01452	0.00229	0.00036	5.74E-05	9.11E-06	1.45E-06	2.31E-07
(b) vs (d)	1.914	0.14405	0.01158	0.00095	8.11E-05	6.98E-06	6.09E-07	5.38E-08	4.79E-09	4.3E-10
(c) vs (d)	2.015	0.15956	0.01349	0.00117	0.000104	9.44E-06	8.66E-07	8.05E-08	7.55E-09	7.13E-10

TABLE 8: SIMILARITY MEASURE OF INTUITIONISTIC FUZZY REPRESENTATION OF WHOLE GENOME (A),(B),(C) & (D)

Similarity	α=1	α=2	α=3	α=4	α=5	α=6	α=7	α=8	α=9	α=10
(a) vs (b)	0.8317	0.96709	0.98063	0.98512	0.98729	0.98855	0.98937	0.989951	0.990373	0.9906
(a) vs (c)	0.84086	0.96886	0.98167	0.98591	0.98796	0.98915	0.98993	0.99047	0.990866	0.9916
(a) vs (d)	0.97267	0.99376	0.99535	0.9974	0.99774	0.99765	0.99788	0.997969	0.998029	0.9987
(b) vs (c)	0.67256	0.93598	0.96234	0.97107	0.97529	0.97775	0.97934	0.980467	0.981289	0.9819
(b) vs (d)	0.8405	0.96837	0.98114	0.98533	0.98733	0.98848	0.98921	0.989717	0.990082	0.9035
(c) vs (d)	0.83206	0.96671	0.98016	0.98457	0.98668	0.98788	0.98865	0.989185	0.989568	0.9898

IV. RESULT AND DISCUSSION

Distance and similarity measures are for different values of α show uniform results for polynucleotides and whole genomes. They also do work satisfactorily as is evidenced from the results of S1, S2, and S1, S3. As supported biologically distance between the first pair should be less than the next pair and consequently the similarity of the first pair should be greater than the next pair. Actually this has happened in our case for each value of α .

As α increases, distance measures increase and similarity measures decrease. This suggests that better is the result, larger is the value of α taken.

V. CONCLUSIONS

The results are verified only for the four genomes chosen in this paper. Obviously the conclusion is not true in general. But as some value of the parameter θ is involved in the calculations, so if some contradictory result appears at all, it is only apparent. It can be adjusted by choice of some suitable value of θ . Thus it can be definitely concluded that the Intuitionistic Fuzzy Set is one of the best tools in analyzing similarity of complete genomes.

REFERENCES

[1] Nieto,J.J.,Torres,A,V-T, 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. B.M. Biology 2006 68: 703–725.

[3] T, A, Nieto,J.J,(2003).The fuzzy poly. space: Basic properties 587–592

[4] K. A, Intuitionistic fuzzy sets, Fuzzy Sets and Systems 20(1986)87-96.

[5] K. A, More on intuitionistic fuzzy sets, Fuzzy Sets. (1989)37-46

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

[7] S,J.K,Distances between Intuitionistic fuzzy,System2000 505-518

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

[9] Kosko,B. (1992) Neural networks and fuzzy systems. Prentice-Hall

[10] S. Das, D. D, A. D, D.K.B, "Some anomalies." IJAINN 2013 38–41