The Journal of Social Sciences Research
ISSN(e): 2411-9458, ISSN(p): 2413-6670
Special Issue. 6, pp: 670-675, 2018
URL: https://arpgweb.com/journal/journal/7/special_issue
DOI: https://doi.org/10.32861/jssr.spi6.670.675

Academic Research Publishing Group

Original Research Open Access

Analysis of Polynucleotide Sequences for Fuzzy Genomes: A Juxtaposition of Two Fuzzy Approaches

Teh Raihana Nazirah Roslan*
School of Quantitative Sciences, College of Arts and Sciences, Universiti Utara Malaysia, 06010 Sintok, Kedah, Malaysia

Mohd Salmi Md Noorani
School of Mathematical Sciences, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

Abstract

The Human Genome Project is the resplendence of the bioinformatics field, especially in health and medicine. It involves research regarding complete nucleotide sequences of the deoxyribonucleic acid (DNA) in human's chromosome. The primary structure of DNA and ribonucleic acid (RNA) consist of nucleotide construction which became polynucleotide when combined. In reality, genetic research field requires huge biological data, and most of the data are vague with various characteristics. Most of them are incomplete and complex from evolutional, functional, adaptability and other traits. The theory of fuzzy sets and fuzzy logic offers modelling methods in uncertainties and various computational techniques for decision making. This research aims to find similarity, difference, equality and identity between polynucleotide sequences using the concept of fuzzy metric space and fuzzy set theory. The Sadegh-Zadeh fuzzy polynucleotide space (RSZ) is being compared with the Torres and Nieto fuzzy polynucleotide space (RTN) in search of the best approach to analyse polynucleotide sequences. Research methods involve data collection of complete genome sequences for homologous species pairs, construction of the RSZ and RTN models, and data analysis. Outputs from RSZ and RTN are then compared with outputs from the Basic Local Alignment Search Tool (BLAST) for validation purposes from the bioinformatics field. Results show that outputs from both approaches are against each other, and RTN executes outputs that are nearest to the outputs from BLAST. Thus, RTN is the best fuzzy approach to compare complete genome sequences for species pairs.

Keywords: Polynucleotide sequences; Fuzzy genomes; Fuzzy metric space; Fuzzy set theory; Bioinformatics.

1. Introduction

The importance of genetic engineering and science is increasingly recognized through the 21st century. One of the applications is in the field of medicine that is now undergoing transformation from a patient-only profession, to a branch of biotechnology. The Human Genome Project is one of the hallmarks of bioinformatics today. However, this field is very challenging as well as requiring lots of biological data, since most of the data are not clear and varied. In fact, some of them are incomplete and complex in terms of evolution, function, coordination and others. Xu et al. (Xu *et al.*, 2008) explained that there are three situations where elements of ambiguity need to be considered. First, as most of the biological processes of the reality are more blurry than predefined. Secondly, biological objects have various tasks that lead to vague membership for every task, and third; difficulty in classifying biological concepts.

Today, genetics once again undergo a phase of concept change, and raises a debate over the abandonment of the overall gene concept. Microsoft chairman Bill Gates was quoted as saying that the gene is one of the most sophisticated programs now (Limberg, 2007). The question of how to compare two genomes has also been playing in the mind of modern scientists. In fact, it achieved first place in two recent lists of major open issues in bioinformatics (Koonin, 1999; Wooley, 1999). These three statements symbolized the importance of research on genetic material as an information carrier. In addition, according to Ernst

Peter Fischer, the concept of the gene is fuzzy (Limberg, 2007). This highlights the importance of fuzzy set and fuzzy logic in providing a better mechanism for decision making of genetics in general.

Furthermore, various questions arising on genetic materials are answered by comparison techniques between genes sequences. In the United States, the National Center for Biotechnology Information (NCBI) is responsible for storing collections of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) sequences that can be used for genetic purposes. Indeed, this task is not easy and involves many methods and techniques. Examples of such methods are the Pearson correlation coefficient, the average log-probability ratio, and the Chi-Square Test (Garcia *et al.*, 2009). Generally, the genes stored in the database are vulnerable to some problems; such as complex definitions of a particular gene. In fact, sets of non-fuzzy gene sequences are sometimes insufficient to classify the gene in detail. Hence, fuzzy sets and fuzzy theory will be applied in this study. Recent studies which proposed fuzzy sets theory in bioinformatics include distinguishing polynucleotides according to amino acids (Georgiou *et al.*, 2015), encoding sequence of RNA molecule of species in phylogenetic trees (Saw *et al.*, 2017), selecting relevant genes in cancer cells (Murthy and Varma, 2015) and improving the accuracy of fetal status assessments (Lu *et al.*, 2016).

This study aims to diagnose structural relationship of genetic materials; in particular to analyze the similarities, differences, equalities and identities between polynucleotides using fuzzy metric spaces. Specifically, we make

*Corresponding Author

comparisons between Torres and Nieto fuzzy polynucleotide space (Torres and Nieto, 2003) and Sadegh-Zadeh fuzzy polynucleotide space (Sadegh, 2000), by applying their methods on three different pairs of homologous species. There exists a debate between these two authors on the relevance of each other methods, refer to (Sadegh-Zadeh, 2007; Torres and Nieto, 2003). In fact, different data samples were used by the pioneer of both approaches in their respective studies. Thus, we explored these two methods on the same set of data to ensure more efficient juxtaposition. Consecutively, we validate the results via output comparisons with the Basic Local Alignment Search Tool (BLAST) which represents the bioinformatics perspectives. This is conducted since both approaches did not provide any verification of their methods within the bioinformatics view. The fuzzy polynucleotide spaces highlighted in this research can also be used for other purposes such as comparison between DNA motives, and to study the level of illness that a person possesses.

2. Polynucleotides as Genetic Materials

Genetic research is the study of inheritance. An individual will be characterized by certain characteristics of the previous generations. Genomes consist of basic units called genes. Genes are the basic units that determine the characteristics of a biological organism. It is located on a chromosome consisting of DNA, the place of gene formation. The basic structure of DNA and RNA comprises of nucleotide buildup in molecular chain. Each nucleotide molecule is constructed of three major elements; one group of phosphates, one group of pentose sugar and one nitrogenous base. The combination of a long chain of nucleotides is called polynucleotides, connected by a phosphodiester bond.

The DNA structure is constructed of two polypeptide straps that coil in the opposite direction to produce a double helical shape, with a base in the center. The nitrogenous bases for DNA are adenine (A), cytosine (C), guanine (G), and thymine (T). This baseline sequence determines the sequence of amino acids connected to form proteins. A will always pair with T, whereas C will always match G. Through Erwin Chargaff's research on DNA compositions, the number of molecules of A bases is always the same as the number of molecules of the base T. The same goes for the number of molecules of the base C which is also equal to the number of molecules of G. As for RNA, it has the same structure as the nucleotide polymer of DNA, but it has only one coil which can be formed according to various types of proteins. In fact, it also helps to carry out functions in the cell and transfer information between the DNA and the proteins involved. RNA is found in cytoplasm, ribosome and slightly in the nucleus. In addition, its nitrogenous bases differ from the one in DNA structure, where thymine (T) base is replaced by uracil (U). The third difference is the sugar in its structure is ribose sugar, in contrast to the deoxyribose sugar in DNA structure.

There are two techniques to study genetic materials, namely sequence analysis and sequence comparison. Sequence analysis is used to determine the building unit for the nucleic acid which is the nucleotide and its arrangement in the acid molecular chain. In contrast, sequence comparison is a taxonomic task and a diagnosis to determine structural relationships such as similarities, and differences between the nucleic acid chains (Sadegh, 2000). Sequence comparison techniques will be applied in this study.

3. Two Fuzzy Polynucleotide Space Approaches

The sequence of bases on RNA or DNA molecules is described as a string of letters $S = S_1, ..., S_n$ in a word with length $n \geq 1$. Fuzzy polynucleotides are polynucleotides represented as a fuzzy set of sets and their membership values in the universal set of X, where $X = \langle x_1, ..., x_n \rangle$. A sequence of polynucleotides can be converted into a sequence of fuzzy polynucleotides using fuzzy codes, via functions that map the letters in RNA or DNA as the set of discourse to a value in $[0, 1]^n$.

In this section, two approaches used to construct fuzzy polynucleotides spaces will be discussed. The fuzzy polynucleotide space proposed by in Sadegh (2000) will be referred to as R_{SZ}; while R_{TN} will be used to represent the fuzzy polynucleotide space proposed by Torres and Nieto (2003).

3.1. Sadegh-Zadeh Fuzzy Polynucleotide Space (R_{SZ})

Sadegh-Zadeh constructed a fuzzy metric that contains all the bases on the sequence of RNA and DNA along with their respective functions. They conducted experiments on short protein sequence with the length of 6, for example tyrosine and histidine. There are two important aspects, namely the position for each base, and how the membership value is assigned to each base. For example, for the UAC sequence consisting of three bases; it can be represented by the following fuzzy metric:

Fuzzy metric (UAC) =
< (U in 1,1), (C in 1,0), (A in 1,0), (G in 1,0)
(U in 2,0), (C in 2,0), (A in 2,1), (G in 2,0)
(U in 3,0), (C in 3,1), (A in 3,0), (G in 3,0) >

Referring to the UAC fuzzy metric above; in the first line, U is in the first position of the sequence. Thus U is given the value of 1 and written (U in 1,1). Given that there is no C in the first position of the sequence, then C is assigned a value of 0 and is written (C in 1,0). The same is repeated for all bases in the sequence until the completion of the fuzzy metric. This (mxn)-metric allows the construction of (mxn)-vector space with length n, and thus forming a n-dimensional vector.

By applying unit hypercube to the fuzzy set theory introduced by Kosko and Burgess (1992), a fuzzy polynucleotide space can be constructed. Given the universal set $X = \{x_1, ..., x_n$ with $n \geq 1$; its power set, $F(2^X)$ will

form a *n*-dimensional hypercube with 2^n edges. Every member in $F(2^X)$ is a fuzzy set and represents a point in the hypercube. For a fuzzy set $A = \{(x_1, a_1),...., (x_n, a_n)\}$; it is represented by a *n*-dimensional vector $(a_1, ..., a_n)$ in the interval [0,1]. For *n* single-beings $\{x_i\}$ which is a non-fuzzy set in 2^X, these points are located on the cube coordinates, whereas empty sets lie on the cube's origin.

To compare polynucleotide sequences in terms of similarity, equality, difference and identity; Sadegh-Zadeh incorporated the cube $[0,1]^n$ with the size of distance *d* to form the fuzzy metric space $<[0, 1]^n, d>$. It can be observed that the original fuzzy sequence has been extended to a fuzzy metric space. The involved definitions and theorem are as follows:

Definition 1 If $A = \{(x_1, a_1), ..., (x_n, a_n)\}$ and $B = \{(x_1, b_1), ..., (x_n, b_n)\}$ are two fuzzy sets, the difference between *A* and *B*, written as the *differ* (*A*, *B*) is

$$differ(A,B) = \frac{\sum i\,|ai-bi|}{c\,(A \cup B)} \quad (1)$$

where the function *c* is the summation of all membership values of the elements in its set.

Definition 2

$$similar\,(A,B) = 1 - differ\,(A,B) \quad (2)$$

Definition 3

1. $equal\,(A,B) = similar\,(A, B)$ (3)
2. *A* and *B* are *identical* if and only if $equal\,(A, B) = 1$ (4)

In addition, for any set fuzzy set *A* and *B*, we can refer as below:

Theorem 1

1. *A* and *B* are *identical* if and only if $similar\,(A, B) = 1$ (5)
2. *A* and *B* are *identical* if and only if $differ\,(A, B) = 0$ (6)

3.2. Torres and Nieto Fuzzy Polynucleotide Space (R_{TN})

Torres and Nieto adopted the same conceptual framework as Sadegh-Zadeh by using the hypercube space and sequence letters for comparing polynucleotide sequences. However, the main difference was Torres and Nieto did not expand the hypercube space to a *n*-dimensional space, but limit it to a 12-dimensional hypercube only.

Torres and Nieto compared the genome sequences of *Mycobacterium tuberculosis* and *Escherichia coli*. To display a polynucleotide sequence with an arbitrary length as a point in space $[0,1]^{12}$, the nucleotide number on each site of one codon in the sequence of the genome will be calculated. This means that the number of occurrences of each of the four nitrogenous bases is calculated separately on each site $i \in \{1,2,3\}$ in the XYZ tri-codon. Since 4 x 3 = 12; we will get 12 original numbers involving the number of each base for each base site. Next, divide each of these 12 numbers by the total number of nucleotides per site. This gives the fraction of each base on each base site for a codon. Finally, vector construction takes place from the breakdown of each base of each site in the form $(x_1, x_2, ..., x_{12})$ with $x_i \in [0,1]$. To compare polynucleotide sequences in terms of equality, difference, similarity and identity; Torres and Nieto introduced the following terms:

Definition 4 If $A = \{(x_1,a_1), …., (x_n,a_n)\}$ and $B = \{(x_1,b_1), …., (x_n,b_n)\}$ are two fuzzy sets, then *C(A,B)* is the *middle canonical point* for *A* and *B*, if and only if

$$C(A,B) = \{(x_1,(a_1+b_1/2)), …., (x_n, (a_n+b_n/2))\} \quad (7)$$

Definition 5

1. $similarity\,(A,B) = c(A \cap B)/c(C(A, B))$ (8)
2. $difference\,(A,B) = 1 - similarity\,(A,B)$ (9)

Definition 6

1. $equality\,(A,B) = similarity\,(A,B)$ (10)
2. *A* and *B* are *idt* if and only if $equality\,(A, B) = 1$ (11)

Theorem 2

1. *A* and *B* are *idt* if and only if $similarity\,(A, B) = 1$ (12)
2. *A* and *B* are *idt* if and only if $difference\,(A, B) = 0$ (13)

4. Results and Discussion

It is important to note that different data samples were used by the pioneer of both approaches in their respective studies. Hence, this section will apply both approaches to complete genome pair of the same species so that comparisons between these two approaches can be done efficiently. Also included in this section is an analysis of the complete genome pairs using the BLAST software for the purpose of validating the outputs from a bioinformatics perspective.

4.1. Data Collection

The collected data were pairs of complete genome sequences from species *Cryphonectria parasitica pleC9* and *Cryphonectria parasitica strain KFC9-J2.31*; *Schistosoma mansoni DIF_7* and *Schistosoma mansoni expressed protein Smp_002740.1*; as well as *Caenorhabditis remanei hypothetical protein CRE_20589* and *Caenorhabditis remanei hypothetical protein CRE_04390*. These data were obtained from the National Center for Biotechnology Information (NCBI) website, http://www.ncbi.nlm.nih.gov. The genome pair *Cryphonectria parasitica pleC9* and

Cryptonectria parasitica strain KFC9-J2.31 consist of 1364 bases and 1295 bases; while *Schistosoma mansoni pair DIF_7* and *Schistosoma mansoni expressed protein Smp_002740.1* consist of 441 bases and 570 bases respectively. The genome pair *Caenorhabditis remanei hypothetical protein CRE_20589* and *Caenorhabditis remanei hypothetical protein CRE_04390* each consists of 540 bases and 450 bases.

These pairs of sequences were chosen because they were all pairs of homologous polynucleotide sequences, meaning each pair of sequences share the same parent, and have a high level of equality between each other. This feature is very important to ensure the validity of the output of this study when compared to the output from the BLAST software. BLAST will analyze these homologous sequences by identifying the short match between sequences. Next, the alignment of these sequences will be performed.

4.2. Results and Analysis

The results of the analysis on pairs of complete genome sequences for species in the study will be presented in this subsection. The outputs for R_{SZ} consist of *similar*, *differ*, *equal* and *identical*; whereas the outputs for R_{TN} are *similarity*, *difference*, *equality* and *idt*. The outputs from BLAST are denoted as *sim*, *diff* and *eql*. Table 1 shows the comparison between outputs from R_{SZ} and R_{TN} analysis; with outputs obtained from BLAST analysis for S_1≡*Cryphonectria parasitica pleC9* and S_2≡ *Cryphonectria parasitica strain KFC9-J2* sequences.

Table-1. Comparison between outputs obtained for S_1 and S_2 sequences

	R_{SZ}	R_{TN}	BLAST	ERROR R_{SZ}	ERROR R_{TN}
differ(S_1,S_2) /*difference*(S_1,S_2)/*diff*(S_1,S_2)	0.8154	0.0221	0.0100	0.8054	0.0121
equal(S_1,S_2)/*equality*(S_1,S_2)/*eql*(S_1,S_2)	0.1846	0.9779	0.9900	0.8054	0.0121
similar(S_1,S_2)/*similarity*(S_1,S_2)/*sim*(S_1,S_2)	0.1846	0.9779	0.9900	0.8054	0.0121

Subsequently, the outputs of BLAST obtained for S_3≡*Schistosoma mansoni expressed protein Smp_002740.1* and S_4≡*Schistosoma mansoni DIF_7* will be compared with outputs from R_{SZ} and R_{TN}. Table 2 shows the comparison between output from R_{SZ} and R_{TN} with output from BLAST analysis for S_3 and S_4 sequences.

Table-2. Comparison between outputs obtained for S_3 and S_4 sequences

	R_{SZ}	R_{TN}	BLAST	ERROR R_{SZ}	ERROR R_{TN}
differ(S_3,S_4) /*difference*(S_3,S_4)/*diff*(S_3,S_4)	0.7586	0.0303	0.0000	0.7586	0.0303
equal(S_3,S_4)/*equality*(S_3,S_4)/*eql*(S_3,S_4)	0.2414	0.9697	1.0000	0.7586	0.0303
similar(S_3,S_4)/*similarity*(S_3,S_4)/*sim*(S_3,S_4)	0.2414	0.9697	1.0000	0.7586	0.0303

Finally, an analysis was carried out on pairs of complete genome sequences for S_5≡*Caenorhabditis remanei hypothetical protein CRE_20589* and S_6≡*Caenorhabditis remanei hypothetical protein CRE_04390*. The outputs obtained can be seen in the table below.

Table-3. Comparison between outputs obtained for S_5 and S_6 sequences

	R_{SZ}	R_{TN}	BLAST	ERROR R_{SZ}	ERROR R_{TN}
differ(S_5,S_6) /*difference*(S_5,S_6)/*diff*(S_5,S_6)	0.8594	0.1957	0.1000	0.7594	0.0957
equal(S_5,S_6)/*equality*(S_5,S_6)/*eql*(S_5,S_6)	0.1406	0.8043	0.9000	0.7594	0.0957
similar(S_5,S_6)/*similarity*(S_5,S_6)/*sim*(S_5,S_6)	0.1406	0.8043	0.9000	0.7594	0.0957

4.3. Discussion

The outputs obtained from both approaches will be compared with the outputs from the BLAST analysis using the same data. The aim is to see which approach produces the output closest to the BLAST output which represents the bioinformatics perspective. This study uses *blastn* algorithm under the basic function of BLAST. Sequence inputs in BLAST are in FASTA or GenBank format, while the output is in HTML format. BLAST will seek out a sequential sequence by finding a short match between these sequences. Then the alignment will be done between the sequences until the complete sequence is complete. Among the best features of BLAST is its speed in controlling database containing large numbers of genomes like GenBank. In fact, it promises precise information.

There are some important terms in BLAST which are being used in this study. *Query coverage* is the percentage of coverage between inputs of species sequences, and sequences that have a meaningful alignment with it. *E value* is the comparison between the alignment of the species' input sequence and the sequences that have a meaningful alignment with it, with an alignment value of expectations having points equal or more than it. The expectation alignment value is available through the search for any sequence in the database having a medium of the same sequence size. The lower the value of *E value*, the more equal the sequence of species input and the sequence of findings were. This study focuses on the value of *E value = 0* since it indicates the most appropriate match. In addition, *Max ident* shows the percentage of similarity among the sequences compared.

The analysis was conducted on the complete genomic pair *S_1≡Cryphonectria parasitica pleC9* and S_2≡*Cryphonectria parasitica strain KFC9-J2.31*. These sequences had *query coverage* of 94%, so the comparison of similarities and differences between them can be accomplished. R_{SZ} produced *similar* (S_1, S_2) = 0.1846, *differ* (S_1, S_2) = 0.8154 and *equal* (S_1, S_2) = 0.1846. Both S_1 and S_2 were interpreted as non-identical sequences. R_{TN} recorded *similarity* (S_1, S_2) = 0.9779, *difference* (S_1, S_2) = 0.0221 and *equality* (S_1, S_2) = 0.9779. Same as R_{SZ}, S_1 and S_2 were

also interpreted as not identical to each other. BLAST recorded *sim* (S_1, S_2) = 0.9900, *diff* (S_1, S_2) = 0.0100 and *eql* (S_1, S_2) = 0.9900 for *E value* = 0. Three different *E values* of 0, 2e-161 and 7e-52 were also produced together with *sim* and *diff* values for each. The values for these *Alignments* vary according to the base position of S_1 and S_2. It can be seen that *sim* values range from 96% to 99%, while *diff* values range from 0% to 2%. Additionally, R_{SZ} generated a huge error of 0.8054, while R_{TN} recorded a very small error of 0.0121.

Subsequently, the analysis was carried out on the complete genomics pair S_3≡*Schistosoma mansoni expressed protein Smp_002740.1* and S_4≡*Schistosoma mansoni DIF_7*. S_3 was chosen to be compared with S_4 because only this sequence had *query coverage* of 72%. Other sequences in the list record 100% of *query coverage*, which will make it difficult to compare the difference. R_{SZ} produced *similar* (S_3, S_4) = 0.2414, *differ* (S_3, S_4) = 0.7586 and *equal* (S_3, S_4) = 0.2414. Both S_3 and S_4 were interpreted as non-identical sequences. R_{TN} recorded *similarity* (S_3, S_4) = 0.9697, *difference* (S_3, S_4) = 0.0303 and *equality* (S_3, S_4) = 0.9697. Following R_{SZ}, S_3 and S_4 were also interpreted as not identical to each other. For the BLAST outputs, *sim* (S_3, S_4) = 1.0000, *diff* (S_3, S_4) = 0.0000 and *eql* (S_3, S_4) = 1.000 were recorded for *E value* = 0. Only one *E value* was produced along with *sim* value (S3, S4) = 1,000 and *diff* (S3, S4) = 0.0000. For the errors generated, R_{SZ} produced a huge error of 0.7586, while R_{TN} showed a very low error of 0.0303.

Finally, we compared the complete genome pair of S_5≡*Caenorhabditis remanei hypothetical protein CRE_20589* and S_6≡*Caenorhabditis remanei hypothetical protein CRE_04390*. S_5 and S_6 had *query coverage* of 98%, so the comparisons of similarity and the difference between them can be done. R_{SZ} produced *similar* (S_5, S_6) = 0.1406, *differ* (S_5, S_6) = 0.8594 and *equal* (S_5, S_6) = 0.1406. Both S_5 and S_6 were interpreted as non-identical sequences. R_{TN} recorded *similarity* (S_5, S_6) = 0.8043, *difference* (S_5, S_6) = 0.1957 and *equality* (S_5, S_6) = 0.8043. Again here as in R_{SZ}, S_5 and S_6 were also interpreted as not identical to each other. Observations made on the BLAST outputs gave us *sim* (S_5, S_6) = 0.9000, *diff* (S_5, S_6) = 0.1000 and *eql* (S_5, S_6) = 0.9000 for *E value* = 3e-160. 11 different types of *E value* were generated along with *sim* and *diff* values for each. These values could compare the similarities and differences according to the base position on S_5 and S_6 more clearly. The *sim* values range from 90% to 91%, while *diff* values are all 0%. Comparison made on the errors revealed that R_{SZ} resulted in a big error of 0.7594, while R_{TN} gave a small error of 0.0957.

As a result of the analysis, it is found that the outputs produced by R_{SZ} and R_{TN} were very different and contradictory. This means that the two approaches were mutually contradictory to one another. In fact, the claim made by [10], the authors of R_{TN} where *similar* (A, B) ≠ *similarity* (A, B); and *differ* (A, B) ≠ *difference* (A, B) were true. In addition, when the outputs of both approaches were compared with the BLAST output, it was found that R_{TN} produced the least error. This shows that R_{TN} produced the closest output to BLAST, showing that it is the best approach to compare pairs of complete genome sequences for species in the study.

5. Conclusion

The fuzzy polynucleotide spaces presented in this study aims to be the site of a polynucleotide sequence comparison. The polynucleotide sequences were represented as points in dimensional hypercube using the concept of fuzzy metric space. The comparison between polynucleotide sequences involved two approaches namely R_{SZ} by Sadegh-Zadeh, as well as R_{TN} by Torres and Nieto. Each approach was equipped with different fuzzy code building methods as well as different distance functions. The outputs of both approaches were compared with the outputs from the BLAST analysis using the same data of complete genome sequences for homologous species pairs. Our findings revealed that both R_{SZ} and R_{TN} approaches were mutually contradictory. R_{TN} was found to be the best approach in comparing the similarities, differences, equalities and identities between complete genome pairs of species in the study. For future study, it may be worth to consider the results obtained from this study with the theoretical formulations for both approaches respectively.

References

Garcia, F., Lopez, F. J., Cano, C. and Blanco, A. (2009). Study of fuzzy resemblance measures for DNA motifs. *Fuzzy Systems*, (2009): 1175-80.

Georgiou, D. N., Karakasidis, T. E., Megaritis, A. C., Nieto, J. J. and Torres, A. (2015). An extension of fuzzy topological approach for comparison of genetic sequences. *Journal of Intelligent and Fuzzy Systems*, 29(5): 2259-69.

Koonin, E. V. (1999). The emerging paradigm and open problems in comparative genomics. *Bioinformatics*, 15(1999): 265-66.

Kosko, B. and Burgess, J. C. (1992). *Neural Networks and Fuzzy Systems.* Prentice Hall: New Jersey.

Limberg, J. (2007). Fuzzy polynucleotides-two approaches to a theory of fuzzy genomes. *Fuzzy Information Processing Society*, (2007): 639-43.

Lu, Y., You, Q. and Li, X. (2016). Automatical assessment of fetal status based on fuzzy theory and Euclidean distance. *Journal of Biomedical Engineering*, 33(3): 436-41.

Murthy, V. B. and Varma, G. (2015). Bioinformatics, A case study of selection of relevant genes in cancer cell. *International Journal of Advanced Research in Computer Science*, 6(3).

Sadegh-Zadeh, K. (2007). The fuzzy polynucleotide space revisited. *Artificial Intelligence in Medicine*, 41(1): 69-80.

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

Saw, A. K., Nandi, S. and Tripathy, B. C. (2017). Fuzzy code on RNA secondary structure. *International Journal of Pure and Applied Mathematics*, 114(3): 483-501.

Torres, A. and Nieto, J. J. (2003). The fuzzy polynucleotide space, Basic properties. *Bioinformatics,* 19(5): 587-92.

Wooley, J. C. (1999). Trends in computational biology, A summary based on a RECOMB plenary lecture. *Journal of Computational Biology,* 6(1999): 459-74.

Xu, D., Keller, J. M., Popescu, M. and Bondogula, R. (2008). *Applications of fuzzy logic in bioinformatics. Series on advances in bioinformatics and computational Biology.* Imperial College Press: London. 9.