Iranian Journal of Fuzzy Systems
Volume 16, Number 6, (2019), pp. 33-41

A computational method to analyze the similarity of biological sequences under uncertainty

A. Khastan[1] and L. Hooshyar[2]

[1,2] *Department of Mathematics, Institute for Advanced Studies in Basic Sciences (IASBS), Zanjan 45137-66731, Iran.*

khastan@iasbs.ac.ir, lida.hooshyar@iasbs.ac.ir

Abstract

In this paper, we propose a new method to analyze the difference and similarity of biological sequences, based on the fuzzy sets theory. Considering the sequence order and some chemical and structural properties, we present a computational method to cluster the biological sequences. By some examples, we show that the new method is relatively easy and we are able to compare the sequences of arbitrary lengths.

Keywords: Similarity of biological sequences, fuzzy polynucleotide space, fuzzy clustering, unit hypercube, fuzzy similarity matrix.

1 Introduction

Genetic code of every organism can be represented as a sequence of alphabets, such as twenty amino acids of protein or four base pairs of RNA and DNA. Since all living organism cell are composed of DNA molecules, some living organisms are biologically similar and some of them are distinct. One of the goals of bioinformatics is to align a large number of sequences and study their evolutionary relationships through comparative sequence analysis. In sequence alignment, sequences with high degree of similarity have similar function and structure, and such sequences help in deriving phylogenetic or evolutionary relationships among organisms [9].

Throughout the history of science, there has been a need to manage and model uncertainty in the real world phenomena. In bioinformatics, the variability exhibited in the nature of genome requires computational and theoretical models to be flexible enough to capture the main aspects without seeing every deviation as something completely new. For instances, in considering a new gene sequence, it is important to know how similar it is to a particular sequence; if it is less similar, it may produce a different effect, if it is very similar, it probably has the same structure and function. It is not so much a question of whether or not the two genes are the same, as it is a question of how much this particular instance of the new gene resembles a prototype. In these situations, alternate methodologies should be used to aid us in making automated evaluations. Other sources of uncertainty that need to be considered include lack of expressiveness or faithfulness of some features that we extract, incompleteness in the data extracted from actual samples, lack of clear boundaries between classes of proteins, proteins that are members of more than one class, etc. Theory of fuzzy set and fuzzy logic provides a different way to view the problem of modeling uncertainty and offer a wide range of computational tools to aid decision making, see [21, 23] and references therein.

In 1990, Sadegh-Zadeh tried to render fuzzy theory accessible to sequence comparison and analysis [16]. To this end, the author fuzzified the concept of sequence and used biopolymers, especially the nucleic acids RNA and DNA, as examples. He demonstrated that a polynucleotide molecule is representable as a point in an $n-$dimensional unit hypercube and constructed a framework for fuzzy theoretical analysis of polynucleotides. The $n-$dimensional unit hypercube is enriched by a distance function d, i.e. $([0,1]^n, d)$, that he suggested as a metric space, namely the fuzzy polynucleotide space. In this metric space, the author presented a similarity and difference analysis to compare polynucleotide sequences. The difference between two polynucleotides is reconstructed as a particular geometric distance between two points in the

Corresponding Author: A. Khastan
Received: June 2018; Revised: December 2018; Accepted: April 2019.

fuzzy polynucleotide hypercube $[0,1]^n$. The less they differ from one another, the closer in the cube they reside [16]. Once a polynucleotide sequences has been transformed to an ordered fuzzy set, it can through its fuzzy code be represented as a point in a unit hypercube. A genetic code may be viewed as 12−dimensional, because a triplet codon XYZ has a $3 \times 4 = 12$ dimensional fuzzy code $(a_1, ..., a_{12})$ and therefore it is a point in the 12−dimensional fuzzy polynucleotide space $[0,1]^{12}$. It is obvious that in this latter, small space only triplet codons can be dealt with. Polynucleotides of length $n > 3$ require higher dimensional cubes. For instance, given two polynucleotide sequences such as UACUGU and CACUGU, each of them is located at a particular point of a 24−dimensional unit cube. Then a polynucleotide of length $n \geq 1$ is a fuzzy polynucleotide of length $4n$, and thus, representable by a real vector $(x_1, x_2, ..., x_{4n})$ of length $4n$ such that each component x_i of the vector is an element of $[0,1]$. Therefore, a polynuclotide consisting of a sequence of k triplets would be a point in a $[0,1]^{12\times k}$ space [16].

Nieto et al. [14, 15] and Torres et al. [20] tried to solve this difficulty by minimizing the dimension of fuzzy polynucleotide space for all polynucleotide sequences with different lengths. They accommodated all polynucleotides of arbitrary length in the unit hypercube $[0,1]^{12}$. Considering the frequencies of the nucleotides at the three base sites of a codon in the coding sequence, they mapped a polynyclotide of arbitrary length on the 12-dimensional unit hypercube $[0,1]^{12}$, while in the approach of Sadegh-Zadeh [16], it is only possible to represent a triplet codon in $[0,1]^{12}$. Therefore, using the method presented in [14, 15, 20], it is possible to transform any polynucleotide of arbitrary length to a point in $[0,1]^{12}$, enabling fuzzy theoretical analysis in a 12−dimensional space.

Later in [17], the author studied both metric spaces presented in [16] and [20] and measured dissimilarity and similarity relationships between polynucleotide strings in both spaces to compare their performance. The author showed that in some cases, Torres et al.'s metric space [20] measures the relationships between polynucleotide chains incorrectly [17, 22].

In [10] by taking the average contents of biological sequences and their information entropies as variables, a fuzzy method is used to cluster them. In [25], the authors used the concept of fuzzy integral to analysis of DNA sequences. In [2, 3], the authors studied the number of alignments between two DNA sequences. A multiple sequence alignment algorithm is presented in [9] to measure the similarity of sequences based on fuzzy parameters and a dynamic programming is used to guarantee the optimal alignment of the sequences. For the latest results in the field, see also [1, 4, 7, 18, 24]. Recently in [8], the authors studied a comparison of genetic sequences by an extension of fuzzy topological approach.

In this paper, using the sequence-order and some chemical and structural properties, we transform biological sequences of arbitrary length to the ordered fuzzy sets. Then, they become representable as points in the unit hypercube $[0,1]^{12}$. In this space, we compute the differences and similarities between nucleotides and compare complete genomes. Finally, we employ a fuzzy based method to cluster the biological sequences.

2 A new approach to analyze the similarity

The nucleic acids RNA and DNA play the pivotal role as the genetic material of living things and viruses in production of proteins. RNA and DNA are made of triplet of codons each of them having the possibility to be one of four nucleotides T, C, A, G in the case of RNA and U, C, A, G in the case of DNA (A: adenine; C: cytosine; G: guanine; T: thymine; U: uracil). It is well-known that the biological sequences have some structural and chemical properties. Therefore, these attributes are usually used to study biological sequences. For the DNA sequences, the four elements {A,T,G,C} can be divided into three groups:
(1) purine group (R={A,G})/pyrimidine group (Y={C,T});
(2) ketone group (M={A, C})/amino group (K={G,T}); and
(3) weak hydrogen bonds group (W={A, T})/strong hydrogen bonds group (S={C,G}).

- Purine and pyrimidine group

  Purines and pyrimidines are two kinds of nitrogen-containing bases. They make up the two different kinds of nucleotide bases in RNA and DNA. The two-carbon nitrogen ring bases (guanine and adenine) are purines, while the one-carbon nitrogen ring bases (cytosine and thymine) are pyrimidines.

- Ketone and amino group

  A ketone is an organic compound that contains a carbonyl functional group. Aldehyde contains a hydrogen atom connected to its carbonyl group while ketone does not have a hydrogen atom attached. The amino group is one of several nitrogen containing functional groups found in organic molecules. What distinguishes the amino group is that the nitrogen atom is connected by single bonds to either hydrogen or carbon.

- Weak and strong hydrogen bonds group

A hydrogen bond is the electromagnetic attraction created between a partially positively charged hydrogen atom attached to a highly electronegative atom and another nearby electronegative atom. A hydrogen bond is a type of dipole-dipole interaction; it is not a true chemical bond. These attractions can occur between molecules (intermolecularly) or within different parts of a single molecule (intramolecularly). Hydrogen bonds occur in inorganic molecules, such as water, and organic molecules, such as DNA and proteins. The two complementary strands of DNA are held together by hydrogen bonds between complementary nucleotides (A&T, C&G). A hydrogen bonds which is the attractive force between the hydrogen attached to an electronegative atom of molecule could be strong or weak.

In this study, we propose a method to analyze the similarity of DNA sequences. It is easy also to apply this approach to RNA sequences.

2.1 Method

In this section, similarly to the results of Nieto et. al [14, 15], we transform biological sequences of arbitrary length to the ordered fuzzy sets as points in $[0,1]^{12}$. We consider three different groups of nucleotides in the sequences to obtain $12-$dimensional vectors. In the following, we explain the method in four steps, where in steps (1)-(3), we replace each nucleotide by its corresponding group's alphabets.

- Step 1.

  In the first step, we consider the purine and pyrimidine groups as a first index and, then, we map all of the nucleotides to this class. In fact, to obtain a new sequence, we replace nucleotides A and G by R and nucleotides C and T by Y. Therefore, all possible occurrence frequencies for a pair of nucleotids in the new sequence are $\{RR, RY, YR, YY\}$.

  Example 2.1. *Consider sequence S as $S = ATGGTGCACCTGACTC$. After transcription, we obtain S_1 as $S_1 = RYRRYRYRYYYRRYYY$.*

- Step 2.

  In this step, we consider the amino and ketone groups as the second index, then, we map all nucleotides to this class. We replace nucleotides A and C by M and nucleotides G and T by K. So, all possible occurrence frequencies for a pair of nucleotids in the new sequence are $\{MM, MK, KM, KK\}$.

  Example 2.2. *Consider the sequence S, as previous example, after transcription we obtain*

  $$S_2 = MKKKKKMMMMKKMMKM.$$

- Step 3.

  As the last index, we consider the weak and strong hydrogens bonds groups, then we map all nucleotides to this class. Therefore, all possible occurrence frequencies in this step is $\{WW, WS, SW, SS\}$.

  Example 2.3. *If we consider sequence S as previous examples, then after transcription, we have*

  $$S_3 = WWSSWSSWSSWSWSWS.$$

- Step 4.

  In this step, for each biological sequence S with arbitrary length, we create a vector with 12 elements. First, we define the 12-dimensional vector B as follows

  $$B = [RR, RY, YR, YY, MM, MK, KM, KK, WW, WS, SW, SS].$$

  To obtain the information vector of the sequence S, we compute relative frequencies of occurrence for all elements of B in the corresponding sequences S_1, S_2 and S_3. Let us denote N_j and L, the occurrence frequencies of $j-$th element of B and the sequence length, respectively. We define the relative frequencies of occurrence as

  $$f_j = \frac{N_j}{L}, \quad j = 1, ..., 12.$$

For instance, f_1 represents the relative frequencies of occurrence of RR in the sequence S_1 and f_6 represents the relative frequencies of occurrence of MK in the sequence S_2. Finally, we define the information vector of the biological sequence S as $V = [f_1, f_2, ..., f_{12}]$. So, any arbitrary sequence is representable as a point in 12-dimential unit hypercube $[0,1]^{12}$. This approach enable quantitative studies such as the measurement of distance, similarities and dissimilarities between nucleotide sequences. Then, similarly to Nieto [20] and Sadegh-Zadeh [16], we consider a distance function d in $12-$dimensional unit hypercube to obtain the fuzzy polynucleotide space i.e., $([0,1]^{12}, d)$. Given $A = (A_1, A_2, ..., A_{12})$ and $B = (B_1, B_2, ..., B_{12})$, not both equal to the empty set $\varnothing = (0, 0, ..., 0)$, the distance between A and B is defined as [15, 16]

$$\text{differ}(A, B) = \frac{\sum_{i=1}^{12}(|A_i - B_i|)}{\sum_{i=1}^{12}\max(A_i, B_i)}. \tag{1}$$

Another metric to measure the similarity between two sequences is defined as [15, 16]

$$\text{similar}(A, B) = 1 - \text{differ}(A, B). \tag{2}$$

Example 2.4. [17] *Let us consider two sequences $S = TGGAAC$ and $S^* = AACTGG$. According to steps (1)-(3), we obtain*

$$S_1 = YRRRRY,\ S_2 = KKKMMM,\ S_3 = WSSWWS,$$
$$S_1^* = RRYYRR,\ S_2^* = MMMKKK,\ S_3^* = WWSWSS.$$

The corresponding information vectors are

$$V = (\frac{3}{6}, \frac{1}{6}, \frac{1}{6}, 0, \frac{2}{6}, 0, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}),\ \text{and}\ V^* = (\frac{2}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, 0, \frac{2}{6}, \frac{1}{6}, \frac{2}{6}, \frac{1}{6}, \frac{1}{6}).$$

Then, we obtain $\text{differ}(V, V^*) = 0.23$ *and* $\text{similar}(V, V^*) = 0.77$. *If we consider the method presented by Sadegh-Zadeh* [17], *we must transform S and S^* to the points in* $[0,1]^{24}$ *to obtain* $\text{differ}(V, V^*) = 1$ *and* $\text{similar}(V, V^*) = 0$. *Also using the approach of Nieto et. al.* [15], *we deduce* $\text{differ}(V, V^*) = 0$ *and* $\text{similar}(V, V^*) = 1$. *Note that S and S^* are compounds of the same triplet codons AAC and UGG, although they are two different polynucleotide molecules. Due to these molecular-biological differences, they could not be identical polynucleotide sequences.*

Remark 2.5. *Using the new approach, by considering more structural and chemical properties, we map all sequences with different lengths to the points in $[0,1]^{12}$. This method enables quantitative studies such as the measurement of distance, similarities and dissimilarities between nucleotide sequences with different lengths. Here, the order of nucleotides plays an important role to obtain information vector, where, in the method presented by Nieto et. al. [15], only the frequencies of polynucleotides at the three base sites of a codon in the coding sequence is important.*

Remark 2.6. *Let S and S^* be two sequences with different lengths L and L^*, respectively. Then, their information vectors are different too, i.e., $V \neq V^*$. Indeed, suppose $V = V^*$. It is easy to check that*

$$\sum_{j=1}^{4} N_j = \sum_{j=5}^{8} N_j = \sum_{j=9}^{12} N_j = L - 1.$$

Therefore, we have $\frac{N_j}{L} = \frac{N_j^*}{L^*},\ \ j = 1, 2, ..., 12$, *or, equivalently,* $N_j = \frac{L}{L^*}N_j^*$. *Therefore, we have*

$$L - 1 = \sum_{j=1}^{4} N_j = \frac{L}{L^*}\sum_{j=1}^{4} N_j^* = \frac{L}{L^*}(L^* - 1) = L - \frac{L}{L^*},$$

which is a contradiction with the assumption $L \neq L^*$. Therefore, we have $V \neq V^*$.

Example 2.7. *Let us consider two biological sequences S and S^* with different lengths, where*

$$S = ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTG,$$

is a part of Human DNA sequence with $L = 40$ and $S^* = ATGCTGACTGCTGAGGAGAA$, is a part of Goat DNA sequence with $L = 20$ (see Table 1). After transcription according to steps (1)-(3), for the sequence S, we have

$$S_1 = RYRRYRYRYYYRRYYYYYRRRRRRRRRYYYRYYRYYRYYR, \tag{3}$$
$$S_2 = MKKKKKMMMMKKMMKMMKKMKKMKMMKKMKKMMKKKMMKK, \tag{4}$$
$$S_3 = WWSSWSSWSSWSWSWSSWSWSSWSWWSWSWSSSSWWWSWS, \tag{5}$$

and, for the sequence S^*, we obtain

$$S_1^* = RYRYYRRYYRYYRRRRRRRR, \quad (6)$$
$$S_2^* = MKKMKKMMKKMKKMKKMKMM, \quad (7)$$
$$S_3^* = WWSSWSWSWSSWSWSSWSWW, \quad (8)$$

where (3),(6) show the purin/pyrimidine sequences and (4),(7) show the amino/ketone sequences and (5),(8) show the weak/strong hydrogen sequences. So, the information vectors for S and S^* are as follows

$$V = \{\frac{10}{40}, \frac{9}{40}, \frac{9}{40}, \frac{11}{40}, \frac{8}{40}, \frac{10}{40}, \frac{9}{40}, \frac{12}{40}, \frac{4}{40}, \frac{14}{40}, \frac{13}{40}, \frac{8}{40}\},$$
$$V^* = \{\frac{8}{20}, \frac{4}{20}, \frac{4}{20}, \frac{3}{20}, \frac{2}{20}, \frac{6}{20}, \frac{6}{20}, \frac{5}{20}, \frac{2}{20}, \frac{7}{20}, \frac{7}{20}, \frac{3}{20}\}.$$

Finally, we use differ *and* similar *functions (1)-(2) to find their difference and similarity. So, we obtain*

$$\text{differ}(S, S^*) = \frac{\sum_{i=1}^{12}(|V_i - V_i^*|)}{\sum_{i=1}^{12}(\max(V_i, V_i^*)} = 0.1811,$$

$$\text{similar}(S, S^*) = 1 - \text{differ}(S, S^*) = 0.8189.$$

Example 2.8. *Consider the biological sequences of Human, Bovine and Gallus as Table 1. Let us denote the information vectors of these sequences by V_1, V_2 and V_3, respectively. By direct calculation, it is easy to obtain*

$$V_1 = \{\frac{31}{92}, \frac{21}{92}, \frac{21}{92}, \frac{18}{92}, \frac{18}{92}, \frac{17}{92}, \frac{18}{92}, \frac{38}{92}, \frac{9}{92}, \frac{28}{92}, \frac{27}{92}, \frac{27}{92}\},$$
$$V_2 = \{\frac{33}{86}, \frac{18}{86}, \frac{18}{86}, \frac{16}{86}, \frac{15}{86}, \frac{17}{86}, \frac{18}{86}, \frac{35}{86}, \frac{11}{86}, \frac{24}{86}, \frac{23}{86}, \frac{27}{86}\},$$
$$V_3 = \{\frac{30}{92}, \frac{21}{92}, \frac{21}{92}\frac{19}{92}, \frac{22}{92}, \frac{20}{92}, \frac{21}{92}, \frac{28}{92}, \frac{33}{92}, \frac{24}{92}, \frac{25}{92}, \frac{9}{92}\}.$$

So, the difference and similarity between these sequences are as follows

$$\text{differ}(V_1, V_2) = 0.0809, \quad \text{similar}(V_1, V_2) = 0.9191,$$
$$\text{differ}(V_2, V_3) = 0.2354, \quad \text{similar}(V_2, V_3) = 0.7646,$$
$$\text{differ}(V_1, V_3) = 0.2273, \quad \text{similar}(V_1, V_3) = 0.7727.$$

3 Fuzzy clustering

To consider the uncertainty and reflect the fuzziness existing in the clustering analysis, fuzzy set theory is a powerful tool to tackle clustering problem. In this section, using the definition of information vector, we study the clustering of biological sequences by fuzzy methods. Fuzzy clustering is a form of clustering in which each data point can belong to more than one cluster or partition. Cluster analysis involves assigning data points to clusters such that items in the same class or cluster are as similar as possible, while items belonging to different classes are as dissimilar as possible. Clusters are identified via similarity measures. These similarity measures include distance, connectivity, and intensity. Fuzzy clustering algorithms have been studied widely in the last years. The first fuzzy clustering method was introduced in [6]. The most widely used fuzzy clustering algorithms are the fuzzy c-means algorithm and fuzzy equivalence relation clustering method. Bezdek in [5] developed the fuzzy c-means clustering method to classify each object into clusters with different or same degree of membership. Later, Klir and Yuan [12] presented a fuzzy equivalent relation-based hierarchical clustering method to study the cluster problem.

In this study, we employ the second method. A fuzzy equivalence relational clustering algorithm is based on a dissimilarity measure extracted from data. For detailed content for fuzzy clustering, see [10, 13].

The fuzzy clustering has mainly three steps. The first step is creating the data matrix. Let $U = \{V_1, V_2, ..., V_n\}$ be the set of information vectors of biological sequences, where V_i denotes the information vector of $i-$th biological sequence and n

Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCA AGGTGAACGT GGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGA AAGTGGATGA AGTTGGTGCTGAGGCCCTGGGCAG
Opposum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTA AGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCA AGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCA AGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCA AAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCACTGCCCTGTGGGGCA AGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAA AGGTGAACCCTGATAATGTTGGCGCTGAGGCCCTGGGCAG
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCA AGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGA AAGTGGATGAAGTTGGTGGGAGGCCCTGGGCAG
Chimpazee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTCCGTTACTGCCCTGTGGGGCAA GGTGAACGTGGATGAAGTTGGTGGTGAGGCGAGGTTGGTATCAAGG

Table 1: Coding sequences

is the total number of sequences. Each information vector has 12 characters $V_i = [f_{i1}, f_{i2}, ..., f_{im}] (i = 1, 2, ..., n, m = 12)$. Thus, we obtain the data matrix as

$$\begin{pmatrix} f_{11} & f_{12} & \dots & f_{1n} \\ f_{21} & f_{22} & \dots & f_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ f_{m1} & f_{m2} & \dots & f_{mn} \end{pmatrix}.$$

The second step is to setup the fuzzy similarity matrix. Here, we use the similarity measure defined in (2) to get the $n \times n$ fuzzy similarity matrix R, $R_{ij} = \text{similar}(V_i, V_j)$. The third step is to cluster. In this study, we apply the clustering method based on the equivalent matrix. Let R be a fuzzy similarity matrix of the data. Then, we transform the fuzzy similarity matrix into a fuzzy equivalent matrix $R^* = t(R)$, by finding the transmission closure of R. The transitive closure, $t(R)$, of a fuzzy relation R is defined as the relation that is transitive, contains R and has the smallest possible membership grades. The operation of $t(R)$ is the same as that of matrix's principles, except that when multiply is to take maximum and adds is to take minimum. By Theorem 1 in [11], if R is a fuzzy similarity matrix, there is a minimum natural number k, so that the transitive closure, $t(R) = R^k$, for more details see [10].

Example 3.1. *Let us consider the DNA sequences of the first exon of $\beta-$globin gene of 11 different species in the Table 1. We present information vectors of the corresponding DNA sequences in Table 2. So, we obtain the similarity matrix by the new method as follows*

$$R = \begin{bmatrix} 1.0000 & 0.9111 & 0.8660 & 0.7680 & 0.8819 & 0.9441 & 0.9403 & 0.9088 & 0.9759 & 0.9256 & 0.9713 \\ 0.9111 & 1.0000 & 0.0833 & 0.7691 & 0.9121 & 0.8829 & 0.9014 & 0.8973 & 0.9201 & 0.9540 & 0.9039 \\ 0.8660 & 0.8335 & 1.0000 & 0.8127 & 0.8789 & 0.9087 & 0.8432 & 0.8664 & 0.8694 & 0.8355 & 0.8683 \\ 0.7680 & 0.7691 & 0.8127 & 1.0000 & 0.7855 & 0.7716 & 0.7329 & 0.8161 & 0.7640 & 0.7648 & 0.7677 \\ 0.8819 & 0.9121 & 0.8789 & 0.7855 & 1.0000 & 0.8897 & 0.8859 & 0.8824 & 0.8874 & 0.8838 & 0.8844 \\ 0.9441 & 0.8829 & 0.9087 & 0.7716 & 0.8897 & 1.0000 & 0.9188 & 0.8993 & 0.9384 & 0.9027 & 0.9416 \\ 0.9403 & 0.9014 & 0.8432 & 0.7329 & 0.8859 & 0.9188 & 1.0000 & 0.8771 & 0.9339 & 0.9109 & 0.9413 \\ 0.9088 & 0.8973 & 0.8664 & 0.8161 & 0.8824 & 0.8993 & 0.8771 & 1.0000 & 0.9110 & 0.9082 & 0.9184 \\ 0.9759 & 0.9201 & 0.8694 & 0.7640 & 0.8874 & 0.9384 & 0.9339 & 0.9110 & 1.0000 & 0.9222 & 0.9658 \\ 0.9256 & 0.9540 & 0.8355 & 0.7648 & 0.8838 & 0.9027 & 0.9109 & 0.9082 & 0.9222 & 1.0000 & 0.9302 \\ 0.9713 & 0.9039 & 0.8683 & 0.7677 & 0.8844 & 0.9416 & 0.9413 & 0.9184 & 0.9658 & 0.9302 & 1.0000 \end{bmatrix}.$$

Human	(31/92, 21/92, 21/92, 18/92, 18/92, 17/92, 18/92, 38/92, 9/92, 28/92, 27/92, 27/92)
Goat	(33/86, 18/86, 18/86, 16/86, 14/86, 20/86, 19/86, 32/86, 9/86, 25/86, 24/86, 27/86)
Opposum	(28/92, 21/92, 21/92, 21/92, 21/92, 18/92, 20/92, 31/92, 12/92, 31/92, 30/92, 18/92)
Gallus	(30/92, 21/92, 21/92, 19/92, 22/92, 20/92, 21/92, 28/92, 33/92, 24/92, 25/92, 9/92)
Lemur	(33/92, 20/92, 20/92, 18/92, 12/92, 22/92, 21/92, 36/92, 13/92, 29/92, 28/92, 21/92)
Mouse	(29/94, 21/94, 21/94, 22/94, 18/94, 18/94, 19/94, 38/94, 11/94, 29/94, 28/94, 25/94)
Rabbit	(33/90, 21/90, 20/90, 15/90, 15/90, 17/90, 17/90, 40/90, 9/90, 28/90, 27/90, 25/90)
Rat	(30/92, 22/92, 22/92, 17/92, 18/92, 19/92, 20/92, 34/92, 16/92, 25/92, 24/92, 26/92)
Gorilla	(32/93, 21/93, 21/93, 18/93, 18/93, 17/93, 18/93, 39/93, 9/93, 28/93, 27/93, 28/93)
Bovine	(33/86, 18/86, 18/86, 16/86, 15/86, 17/86, 18/86, 35/86, 11/86, 24/86, 23/86, 27/86)
Chimpanzee	(36/105, 24/105, 24/105, 20/105, 20/105, 20/105, 19/105, 45/105, 13/105, 31/105, 30/105, 30/105)

Table 2: Information vectors of sequences

We see that the largest entries are associated with the pairs (Human-Gorilla, 0.9759), (Human-Chimpanzee, 0.9713), (Gorilla-Chimpanzee, 0.9658) and (Goat-Bovine, 0.9540). On the other hand, the smallest entries in the similarity values appear in the rows belonging to Opossum (the most remote species from the remaining mammals) and Gallus (the only non-mammalian representative) which is consistent with the known facts of evolution.
In order to see this more clearly, we apply the fuzzy clustering method to obtain the phylogenetic tree in Figure 1. We see that all the species with similar characteristics are clustered into one group and shows they have close evolutionary relationship. For example, Goat and Bovine are in the same group and so are the Human, Gorilla and Chimpanzee. Species Gallus is separated from other species and Opossum is also far from other mammal species.

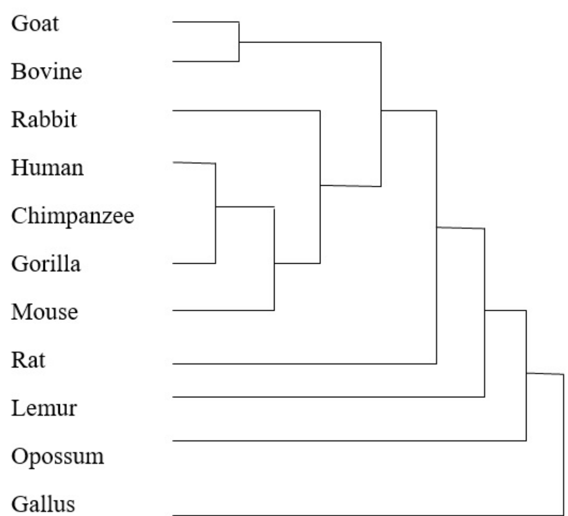

Figure 1: The phylogenetic tree of 11 kinds of species generated by new method.

In [10], to analyze the similarity of DNA sequences, the average content of purine group, ketone group and weak hydrogen considered. Since the corresponding sequences may be different from each other and their average contents may be the same, the information entropy is considered to describe the order of sequences. Then, the authors obtained the $n \times m$ data matrix, where n is the total number of sequences and m is the number of properties that they considered for each sequence. In this paper, considering more chemical and structural properties of biological sequences, we obtain different information vectors for sequences with different lengths, without calculating the information entropy. The clustering methods in [10] and here are based on the fuzzy equivalence relation and they are applicable for sequences with arbitrary lengths. As we can see in Figure 1 and Figure 2, our method is basically consistent with the former results in [10].

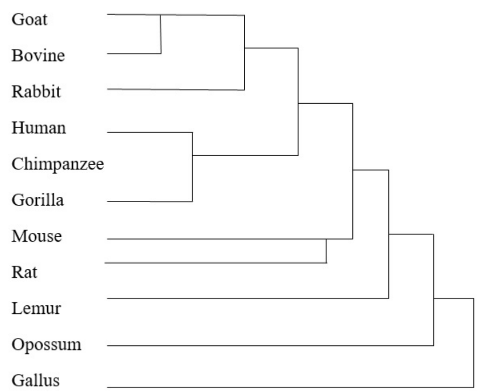

Figure 2: The phylogenetic tree of 11 kinds of species generated by the presented method in [10].

4 Conclusions

In this study, a new computational method to transform the biological sequences to points in the fuzzy unit hypercube is presented. The chemical and structural properties of biological sequences are taken as variables to create the information vectors. Then, using differ and similar functions, the similarity and difference between biological sequences of arbitrary lengths are measured and a method for fuzzy clustering of biological sequences is proposed. Experiments on the coding sequences of the β-globin gene for 11 different species shows that our proposed method is efficient and feasible.

Results here might be used in further research on studying protein secondary structural sequences. However, in this paper, we only considered the most used sequences to verify our model. For future study, we propose to take into account more information of biological sequences and consider more complex sequences.

Acknowledgement

The authors are grateful to the Editor and the anonymous Reviewers for their interesting and valuable comments.

References

[1] M. Ahmad, L. T. Jung, M. A. Bhuiyan, *On fuzzy semantic similarity measure for DNA coding*, Computers in Biology and Medicine, **69** (2016), 144-151.

[2] H. Andrade, J. J. Nieto, A. Torres, *The number of alignments between two DNA sequences*, International Journal of Biomathematics, **9** (2016), 1650053(1-6).

[3] H. Andrade, I. Area, J. J. Nieto, A. Torres, *The number of reduced alignments between two DNA sequences*, BMC Bioinformatics, **15** (2014), 94-98.

[4] S. Bandyopadhyay, *An efficient technique for superfamily classification of amino acid sequences: Feature extraction*, fuzzy clustering and prototype selection, Fuzzy Sets and Systems, **152** (2005), 5-16.

[5] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*, Plenum, New York, 1981.

[6] J. C. Dunn, *Will-separated clusters and the optimal fuzzy partition*, Journal of Cybernetics, **4** (1974), 95-104.

[7] D. N. Georgiou, T. E. Karakasidis, A. C. Megaritis, J. J. Nieto, A. Torres, *An extension of fuzzy topological approach for comparison of genetic sequences*, Journal of Intelligent and Fuzzy Systems, **29** (2015), 2259–2269.

[8] D. N. Georgiou, T. E. Karakasidis, J. J. Nieto, A. Torres, *A study of entropy/clarity of genetic sequences using metric spaces and fuzzy sets*, Journal of Theoretical Biology, **267** (2010), 95-105.

[9] N. Gill, S. Singh, *Biological sequence matching using fuzzy logic*, International Journal of Scientific & Engineering Research, **2** (2011), 1-5.

[10] W. Huang, J. Zhang, Y. Wang, D. Huang, *A simple method to analyze the similarity of biological sequences based on the fuzzy theory*, Journal of Theoretical Biology, **265** (2010), 323-328.

[11] S. Hui-Qin, X. Zhang, *Fuzzy cluster based on rough set and result evaluating*, Journal of Fudan University, **43** (2004), 819-822.

[12] G. J. Klir, B. Yuan, *Fuzzy sets and fuzzy logic theory and application*, Prentice Hall PTR, Upper Saddle River, NJ, 1995.

[13] S. Miyamoto, H. Ichihashi, K. Honda, *Algorithms for fuzzy clustering: Methods in c-means clustering with applications*, Springer, Berlin, 2008.

[14] J. J. Nieto, A. Torres, M. M. Vazquez-Trasande, *A metric space to study differences between polynucleotides*, Applied mathematics letters, **16** (2003), 1289-1294.

[15] J. J. Nieto, A. Torres, D. N. Georgiou, T. E. Karakasidis, *Fuzzy polynucleotide spaces and metrics*, Bulletin of Mathematical Biology, **68** (2006), 703-725.

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

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

[18] S. S. Sahu, G. Panda, *A novel feature representation method based on Chou's pseudo amino acid composition for protein structural class prediction*, Computational Biology and Chemistry, **34** (2010), 320–327.

[19] C. Shen, Y. Ding, J. Tang, J. Song, F. Guo, *Identification of DNA-protein binding sites through multi-scale local average blocks on sequence information*, Molecules, **22** (2017), 2079(1-20).

[20] A. Torres, J. J. Nieto, *The fuzzy polynucleotide space: basic properties*, Bioinformatics, **19** (2003), 587–592.

[21] A. Torres, J. J. Nieto, *Fuzzy logic in medicine and bioinformatics*, Journal of Biomedicine and Biotechnology, **2006** (2006), 1-7.

[22] A. Torres, J. J. Nieto, *Comments on the fuzzy polynucleotide space revisited*, Artificial Intelligence in Medicine, **41** (2007), 81-82.

[23] D. Xu, J. M. Keller, M. Popescu, R. Bondugula, *Applications of fuzzy logic in bioinformatics*, Imperial College Press, London, 2008.

[24] L. Yang, W. Zhang, *A multiresolution graphical representation for similarity relationship and multiresolution clustering for biological sequences*, Journal of Computational Biology, **24** (2017), 299-310.

[25] S. Zhang, Y. Zhang, I. Gutman, *Analysis of DNA sequences based on the fuzzy integral*, MATCH Communications in Mathematical and in Computer Chemistry, **70** (2013) 417-430.

Iranian Journal of Fuzzy Systems
Volume 16, Number 6 (2019), pp. 209
University of Sistan and Baluchestan

A computational method to analyze the similarity of biological sequences under uncertainty

A. Khastan and L. Hooshyar

یک روش محاسباتی برای تحلیل شباهت توالی‌های بیولوژیکی تحت عدم قطعیت

چکیده. در این مقاله، یک روش جدید بر اساس نظریه مجموعه‌های فازی برای بررسی تفاوت و شباهت توالی‌های بیولوژیکی، ارائه می‌کنیم. با توجه به ترتیب توالی و برخی از ویژگی‌های شیمیایی و ساختاری، یک روش محاسباتی برای خوشه‌بندی توالی‌های بیولوژیکی معرفی می‌کنیم. با ارائه چندین مثال، نشان می‌دهیم که روش جدید نسبتاً آسان بوده و قادر به مقایسه توالی‌های با طول دلخواه می‌باشد.