Some New Approaches towards Analyzing Genome Sequences

Thesis submitted by

Subhram Das

Doctor of Philosophy (Engineering)

School of Bio-Science & Engineering
Faculty Council of Engineering & Technology
Jadavpur University
Kolkata, India

2019

JADAVPUR UNIVERSITY
KOLKATA – 700 032

INDEX NO.: 67 $\frac{15}{19}$ / E

1. **Title of the thesis:** Some New Approaches towards Analyzing Genome Sequences

2. **Name, Designation & Institution of the supervisor:**

   Dr. Dewaki Nandan Tibarewala
   Former Professor
   School of Bioscience & Engineering,
   Jadavpur University, Kolkata 700032, India

   Dr. D. K. Bhattacharya
   Professor Emeritus
   Rabindra Bharati University, Kolkata 700007, India

3. **List of Publications:**

I. **Journal Publications:**

1. **Subhram Das**, Debanjan De, Anilesh Dey, D. K. Bhattacharya, Some anomalies in the analysis of whole genome sequence on the basis of Fuzzy set theory, International Journal of Artificial Intelligence and Neural Networks. 2013 Volume 3: Issue 2, PP. 38-41, ISSN 2250-3749.
2. **Subhram Das**, Debanjan De & D. K. Bhattacharya, Similarity and Dissimilarity of Whole Genomes using Intuitionistic Fuzzy Logic, Notes on Intuitionistic Fuzzy Sets, 2015, Volume 21, Issue 3, PP. 48-53, ISSN 1310-4926.
3. **Subhram Das**, Jayanta Pal, D. K. Bhattacharya, Geometrical method of exhibiting similarity/dissimilarity under new 3D classification curves and establishing significance difference of different parameters of estimation, International Journal of Advanced Research in Computer Science and Software Engineering, 2015, Volume 5, Issue 5, PP. 279-287, ISSN 2277-128X.

4. **Subhram Das**, Tamal Deb, Nilanjan Dey, Amira S. Ashourd, D.K. Bhattacharya, D.N. Tibarewala, Optimal choice of k-mer in composition vector method for genome sequence comparison, Genomics, Elsevier, 2018, 110, 263-273, ISSN 0888-7543
5. **Subhram Das**, Arijit Das, Bingshati Mondal, Nilanjan Dey, D. K. Bhattacharya, D. N. Tibarewala, Genome Sequence comparison under a new form of tri-nucleotide representation based on bio-chemical properties of nucleotides, Accepted in Gene, Elsevier, 2019

II. Book Chapter

1. **Subhram Das**, Soumen Ghosh, Jayanta Pal and Dilip K. Bhattacharya, Use of Fuzzy Set Theory in DNA Sequence Comparison and Amino Acid Classification, 2017, Emerging Research on Applied Fuzzy Sets and Intuitionistic Fuzzy Matrices. IGI Global, 235-253, DOI: 10.4018/978-1-5225-0914-1.ch010

4. List of Patents: Nil

5. List of National and International Conference Publications:

1. **Subhram Das**, Subhra Palit, Anindya Raj Mahalanabish and Nobhonil Roy Choudhury, A New Way to Find Similarity/Dissimilarity of DNA Sequences on the Basis of Di-nucleotides Representation. 2015, Lect. Notes Electrical *Engg*., Vol. 335, Computational Advancement in Communication Circuits and Systems, ISBN 978-81-322-2273-6.
2. **Subhram Das**, Tamal Deb, D. K. Bhattacharya, D. N. Tibarewala, A probabilistic way of comparing DNA sequences, 2016, 5th International Conference on 'Computing, Communication and Sensor Network', CCSN2016, Kolkata, ISBN 81-85824-46-0.
3. **Subhram Das**, Nobhonil Roy Choudhury, D.N. Tibarewala and D.K. Bhattacharya, Application of Chaos Game in Tri-Nucleotide Representation for the Comparison of Coding Sequences of β-Globin Gene, 2018, Industry Interactive Innovations in Science, Engineering and Technology, Lecture Notes in Networks and Systems, vol. 11, Springer, ISBN 978-981-10-3952-2

CERTIFICATE FROM THE SUPERVISORS

This is to certify that the thesis entitled "Some New Approaches towards Analyzing Genome Sequences" Submitted by Shri Subhram Das, who got his name registered on 10th September, 2015 for the award of Ph.D. (Engineering) degree of Jadavpur University is absolutely based upon his own work under the supervision of Dr. D. N Tibarewala and Dr. D. K. Bhattacharya and that neither his thesis nor any part of the thesis has been submitted for any degree/diploma or any other academic award anywhere before.

1.---------------------------------------

Signature of the supervisor and
Date with official seal

2.---------------------------------------

Signature of the supervisor and
Date with official seal

Dedicated to my Parents

Acknowledgement

Although a thesis is known by the name of its author, there are many without whose direct or indirect help, a thesis does not materialize. It is time to thank them.

First of all I like to express my profound gratitude and deep regards to Prof. (Dr.) D. K. Bhattacharya and Prof. (Dr.) D. N. Tibarewala for their exemplary guidance and monitoring throughout the course of this thesis. Their constant supervision, advice, and guidance from the very early stage of this research have given me extraordinary experiences throughout the work. Their creative ideas exceptionally inspired and enriched my growth as a student and a researcher. Above all, they have provided me tremendous encouragement and support in various ways. I am really indebted to them. The blessing, help and guidance given by them from time to time shall carry me a long way in the journey of life on which I am about to embark. I would like to thank them for giving me the opportunity to work under their supervision.

I am grateful to Late Juthika Bhattacharya for giving me plenty of love and blessings.

I would acknowledge Dr. Anilesh Dey for his constant support and assistance, which were very fruitful for shaping up my research.

I wish to express my heartfelt thanks to all of the co-authors in the manuscripts in this thesis. Especially Mr. Tamal Deb, Nilanjan Dey and Mr. Arijit Das for the numerous technical discussions we had.

I convey my sincere regards to Mr. Taranjit Singh the Managing Director, JIS Group, Kolkata, for his kind approval in pursuing out my research work.

I also humbly acknowledge the present and ex-Principals of Narula Institute of Technology, for their consent to carry out my research work.

I would like to express my gratitude to Mr. Jayanta Pal, who has always been a great support at my work place. My sincere thanks also go to my colleagues in general and specially to Mr. Soumen Ghosh at the Narula Institute of Technology for a congenial working atmosphere and for helping me with all the big and small problems that are an unavoidable part of each research work.

It is with immense gratitude that I acknowledge the support of all my teachers who taught me something or the other throughout my journey of learning and acquiring academic excellence.

Last but not the least; I am grateful to my family. Now when my years of work are about to yield results, words fail as I try to express their role in my life. My father Mr. Dilip Kumar Das and mother Mrs. Sandhya Nath Das deserve special mention for their love, best wishes and prayers. I take this opportunity to provide my deepest gratitude to my caring, loving, and supportive wife, Mrs. Shampa Sarkar Das. I can’t articulate how she has helped me throughout; keeping up with my hectic schedule these long years. My daughter Sanghavi & Sannidhyi has always inspired me for my work. Their energy and freshness has never failed to rob me off my tiredness.

So finally, I would wind up by thanking everybody who was important to the successful realization of this thesis, as well as expressing my apology that I could not mention personally everybody one by one.

Subhram Das

School of Bio-Science & Engineering
Jadavpur University
Kolkata, India

CONTENTS

CHAPTER 1 1

1. INTRODUCTION 1

1.1 MONO-NUCLEOTIDE REPRESENTATION OF DNA SEQUENCES 2

1.1.1 2D GRAPHICAL REPRESENTATIONS 2

1.1.2 3D GRAPHICAL REPRESENTATIONS 3

1.1.3 4D GRAPHICAL REPRESENTATION 3

1.1.4 OTHER TYPES OF REPRESENTATIONS 3

1.2 DI-NUCLEOTIDE REPRESENTATION OF DNA SEQUENCES 5

1.3 TRI- NUCLEOTIDE REPRESENTATIONS 7

1.4 DESCRIPTORS AND DISTANCE MEASURES 8

1.4.1 GEOMETRICAL DESCRIPTOR 8

1.4.2 MATRIX FORM OF DESCRIPTOR 8

1.4.3 PROBABILISTIC DESCRIPTORS 10

1.5 MAIN FOCUS AREA 11

1.5.1 REPRESENTATION BASED ON THE THREE PAIR OF CLASSIFIED CURVES 11

1.5.2 FUZZY REPRESENTATION ON A TWELVE DIMENSIONAL HYPERCUBE 12

1.5.3 K-MER REPRESENTATIONS 14

CHAPTER 2 19

2. GEOMETRICAL METHOD OF COMPARISON UNDER NEW 3D CLASSIFICATION CURVES 19

2.1 OUTLINE OF THE EXISTING METHOD AND ITS LIMITATION 19

2.2 3D GENERALIZATIONS 20

2.3 NEW 3D COMPONENT DESCRIPTORS 21

2.4 RESULT AND DISCUSSION 21

2.5 COMPARISON USING STATISTICAL HYPOTHESIS TESTING 25

2.6 CONCLUSION 26

CHAPTER 3 29

3. SOME PROBLEMS OF FUZZY POLYNUCLEOTIDE METRIC SPACE AND NECESSARY MODIFICATIONS 29

3.1 SECTION 1 29

3.1.1 FUZZY POLYNUCLEOTIDE SPACE 29

3.1.2 FUZZY POLYNUCLEOTIDE METRIC SPACE 30

3.1.3 RESULTS ON SEQUENCE COMPARISON 30

3.1.4 LIMITATIONS OF THE RESULTS 31

3.1.5 PROBLEMS WITH FUZZY POLYNUCLEOTIDE METRIC SPACE 31

3.2 SECTION 2 33

3.2.1 INTUITIONISTIC FUZZY SET 33

3.2.2 INTUITIONISTIC FUZZY REPRESENTATION OF POLYNUCLEOTIDE ON A TRIPLET OF I^{36} 34

3.2.3 RESULTS ON SEQUENCE COMPARISON 34

3.2.4 DISCUSSION 36

3.2.5 CONCLUSION 36

CHAPTER 4 37

4. COMPOSITION VECTOR METHOD UNDER OPTIMAL CHOICE OF *K*-MER 37

4.1. OUR METHODOLOGY 38

4.1.1 STEP 1: FREQUENCY OR RANK VECTOR 38

4.1.2 STEP 2: PROBABILITY VECTOR 38

4.1.3 STEP 3: COMPOSITION VECTOR 39

4.1.4 STEP 4: DISTANCE MEASURE 39

4.2 RESULTS AND DISCUSSIONS 40

4.2.1 COMPARATIVE STUDY OF THE PROPOSED APPROACH AGAINST OTHER INACCURATE METHODS 40

4.2.2 COMPARATIVE STUDY OF THE PROPOSED APPROACH VERSUS OTHER AGREEING METHODS 53

4.2.3 COMPOSITION VECTOR METHOD USING *K*-MER AND DISTANCE MEASURE OTHER THAN THE SIMILARITY INDEX 60

4.2.4 Composition Vector Method Using Similarity Index as a Distance Measure 61
4.3 Conclusion 63

CHAPTER 5 65

5. A New Form of Tri-Nucleotide Representation Based On Bio-Chemical Properties of Nucleotides........ 65
5.1 Methodology 65
5.1.1 Di-Nucleotide Groups........ 65
5.1.2 Di-Nucleotide Representation 66
5.1.3 3D Representation of A, C, T & G 66
5.1.4 Tri-Nucleotides Representation of DNA Sequence 66
5.1.5 A Sample of 10-Dimensional Vector Representations of Tri-Nucleotide Using Frequencies 69
5.1.6 Generation of Distance Matrix from 10-Dimentional Tri-Nucleotide Representation........ 71
5.1.7 Construction of Phylogenetic Tree 71
5.2 Result and Discussion 71
5.3 Conclusion........ 86

CHAPTER 6 87

6. Future Scope 87

References........ 89

List of Tables

Table 1.1 A 4 × 4 matrix using pairs of nucleotides 6

Table 1.2 No. of Nucleotides, Total Nucleotides and Fraction of Nucleotides 13

Table 2.1 The coding sequences of the first exon of *β-globin* gene of eleven different species 22

Table 2.2 The graph radii associated with three different patterns of the classification curves 22

Table 2.3 The average angles associated with three different patterns of the classification curves 23

Table 2.4 The relative departures associated with three different patterns of the classification curves 23

Table 2.5 Similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the normalized graph radii 23

Table 2.6 Similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the average angles 24

Table 2.7 Similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the relative departures 24

Table 2.8 Comparison the values calculated by our method (A,B,C) with other methods (A1,B1,C1) 25

Table 3.1 The detailed calculations of distances under different metrics 32

Table 3.2 Intuitionistic Fuzzy Representation of Whole Genome (a), (b), (c) & (d) 35

Table3.3 Distance Measure of Intuitionistic Fuzzy Representation of Whole Genome (a),(b),(c) & (d) 35

Table 3.4 Similarity Measure of Intuitionistic Fuzzy Representation of Whole Genome (a),(b),(c) & (d) 35

Table 4.1 Description of mitochondrial genome of 31 mammalians 41

Table 4.2 TYLCD-causing virus sequences used in this study 44

Table 4.3 The list of 21 genomes used for comparisons 48

Table 4.4 The S segments of 19 HV strains 53

Table 4.5 The 48 hepatitis E viruses 56

Table 4.6 Database of Influenza viruses 62

Table 5.1 9-dimentional representation of 64 tri- nucleotides 67

Table 5.2 64× 10 matrix representation of "AACCTGGATCAAGTCCTTAACGTGGAACCT" 69

Table 5.3 Complete coding sequence of *β-globin* gene of 10 different species 72

Table 5.4 Distance matrix of 10 different species using our method 72

Table 5.5 31 mammalian mitochondrial whole genome 73

Table 5.6 TYLCD-causing virus sequences used in this study 76

Table 5.7 Information of the 59 bacterial genomes from 15 families 80

List of Figures

Fig 1.1 2D vector representation of bases A, C, T, G 4

Fig. 1.2 Distribution of Di-nucleotide in Cartesian 2D coordinates 6

Fig. 2.1 3D Graphical representation of classification curve of the DNA sequences of the first exon of *β-globin* gene of Human (W-S Curve) 20

Fig. 4.1(a) Phylogenetic tree using the Probabilistic method 42

Fig. 4.1(b) Phylogenetic tree obtained using the proposed method 43

Fig. 4.2(a) Phylogenetic tree using Probabilistic method 46

Fig. 4.2(b) Phylogenetic tree using the proposed method 47

Fig. 4.3(a) Phylogenetic trees by maximum likelihood method 49

Fig. 4.3(b) Phylogenetic trees by neighbour joining method with Kimura-2 50

Fig. 4.3(c) Phylogenetic trees by neighbour joining method with Jukes-Cantor Model 51

Fig. 4.3(d) Phylogenetic tree obtained by the proposed method for 21 viruses 52

Fig. 4.4(a) Phylogenetic tree of 19 Hantaviruses by word and rough set theory 54

Fig. 4.4(b) Phylogenetic tree of 19 Hantaviruses obtained by the proposed method 55

Fig. 4.5(a) Phylogenetic tree of HEV genomes by word and rough set theory 58

Fig. 4.5(b) Phylogenetic tree of HEV genomes by the proposed method 59

Fig. 4.6 Phylogenetic tree of 31 mammalian mitochondrial genomes by the *k*-mer method with *k*=5 61

Fig. 4.7(a) Phylogenetic tree under *k*-mer method (*k*=4) with similarity index as the distance measure 62

Fig. 4.7(b) Phylogenetic tree is obtained by the proposed method 63

Fig. 5.1 Phylogenetic tree of complete coding sequence of *β-globin* gene of 10 different species 71

Fig. 5.2(a) Phylogenetic tree of 31 mammalian species using the Probabilistic method 74

Fig. 5.2(b) Phylogenetic tree of complete mitochondrial genome sequence of 31 mammalian species by our method 75

Fig. 5.3(a) Phylogenetic tree using Probabilistic method 78

Fig. 5.3(b) Phylogenetic tree using our method 79

Fig. 5.4(a) The phylogenetic tree of 59 bacteria from 15 families based on our method 83

Fig. 5.4(b) The UPGMA phylogenetic tree of 59 bacteria from 15 families based on multiple encoding vector method 84

Fig. 5.4(c) The UPGMA phylogenetic tree of 59 bacteria from 15 families based on feature frequency profiles (FFP) method using 9-mer 85

CHAPTER 1

1. Introduction

DNA is usually presumed to be the critical macromolecular target for carcinogenesis and mutagenesis. To predict sequence changes induced by different agents, it is imperative to have quantitative measures to compare and contrast the different DNA sequences. In addition, the very rapid rise in available DNA sequence data has also made the problem more emerging and interesting too. This is why over the last few years several authors have presented various methods to assign mathematical descriptors to DNA sequences in order to quantitatively compare the sequences and determine similarities and dissimilarities amongst them. Main difficulties crop up as the sequences are very long and are of unequal lengths too. These make the problem of comparison more complex. Again as the methods are different and there are many such methods, so it becomes necessary to compare these methods and determine which one(s), if any, is best in characterizing DNA sequences. If not, better numerical representation and characterization are to be developed. The aim of the present thesis is to critically analyze the roles of different methods of DNA sequence comparison, which are normally used and to modify them accordingly so that better characterization of DNA sequences may be achieved.

So far as DNA sequence analysis is concerned, there are mainly two different types of methods known so far. One is called Alignment based and other one is called Alignment free. The latter method is preferred to the earlier one, due to the following difficulties associated with the first methods. The difficulties may be mentioned formally as below:

Alignment-based approaches generally give excellent results when the sequences under study are closely related so that the sequences can be reliably aligned, but when the sequences are divergent, a reliable alignment cannot be obtained and hence the applications of sequence alignment become limited. Another limitation of alignment-based approaches is that it has more computational complexity, and it is more time-consuming.

So far as Alignment free methods are concerned with regard to DNA sequence analysis, the details may be given sequentially as below:

Numerical characterization can play an important role in the identification of coding segments in newly emerging sequences, or prediction of functions from sequences. The primary step in creating a mathematical descriptor is to develop reliable techniques for characterizing DNA/RNA sequences numerically. While algorithms can be constructed to generate mathematical representations directly from DNA primary sequences, it is intuitively more appealing to represent a long DNA sequence in the form of a graph and visually identify regions of interest or the distribution of bases along the sequence. Most of the methods that have been proposed in the literature to numerically characterize DNA sequences are based on one or more graphical representations of such sequences, and several applications have been made using these techniques. After numerical representation DNA sequence is first embedded in a graph of finite dimension and then quantification is made with coordinates of the data points of the graph. There have been several approaches to graphical representations of DNA sequences. An advantage of a graphical representation lies in the fact that relevant bits of information can be quickly obtained by visual inspection of the plot of a DNA sequence.

1.1 Mono-nucleotide Representation of DNA Sequences

In fact, researchers have outlined different graphical representations of DNA sequences. These are two dimensional, three dimensional, four dimensional and even more dimensional. Some of the different dimension of graphical representations is given below:

1.1.1 2D Graphical Representations

Representations based on two dimensional Cartesian coordinates remain the staple form of graphical methods for their simplicity and intuitive feel.

We first mention the 2D representation of DNA sequences which was initiated by Gates [1] and which was subsequently modified in [2, 3]. Basically it is a random walk of points moving along or parallel of coordinate axes. Another interesting initial attempt of 2D graphical representation of DNA sequences includes the work of [4, 5, 6, 7]. In all this work the basic difference lies in the fact that all the nucleotides are plotted not along the axes but always along a vector lying in one or more quadrants of the Euclidian plan.

Now we mention 2D graphical representation based on six nucleotide pair giving three characteristics groups of nucleotides [8]. The representation is very simple in the sense that if for example *purine* (R=A,G) of *purine/pyrimidine* group is taken then the nucleotide A is always lies on a straight line parallel to X-axis at a distance three from the origin, C lies on similar line at a distance one from the origin; the rest two nucleotides C and T lie on similar line at a distance two from the origin. As there are three different groups of nucleotides *purine* (R=A,G), *pyrimidine* (Y=C,T), *amino* (M=A,C), *keto* (K=T,G), and weak *H-bond* (W=A,T), and strong *H-bond* (S=C,G), so six such different representation are possible corresponding to three different groups of three characteristic groups of nucleotides. Other similar representations with slight modification are found in [9-12]. Some interesting binary representation of DNA sequences is given in [13-16].

1.1.2 3D Graphical Representations

It may be mentioned that so far as 3D graphical representation of DNA sequences is concerned, the first idea came in [17]. Later on many such 3D representations were obtained. Mention may be made of the papers found in [18-25].

1.1.3 4D Graphical Representation

The only paper we like to mention in connection with 4D graphical representation of DNA sequences is that given in [26]. No doubt that the representation is not good for visualization as it is made in four dimension space. But the representation is very much useful, as sit is binary in nature.

1.1.4 Other Types of Representations

Now we like to mention some of the most important and interesting graphical representation of DNA sequences.

Z-Curve

Z-curve [27] is a special type of curve, whose points are in one-one correspondence with the DNA sequences of any length. This curve is suitable for visualization. Different geometrical representations sufficient to analyze DNA sequences may be unified in the representation in Z-curve.

Real-Numbers Representation

Now we consider real-number representation [28, 29, 30] where the four nucleotides bases are assigned four different real numbers arbitrarily. Analysis is done on the represented real number sequence obtained from the corresponding DNA sequences.

Complex Representation

In [31, 32, 33] the authors consider representation of four nucleotides in the four quadrants of the complex plane by assigning the complex numbers 1+i1, 1-i1, -1+i1, -1-i1. Naturally two nucleotides are mirror image of real axis or the mirror image of the imaginary axis.

Quaternion Representation

The method of complex representation is extended to Quaternion [34]. This is also called the hyper-complex numbers representation of DNA sequences by using Quaternion of the form a+ib+jc+kd where $i^2+j^2+k^2=1$, i.j=0; j.k=0; k.i=0.

Probabilistic Representation [35]

2D graphical representation of DNA sequences is taken as in Fig. 1.1.

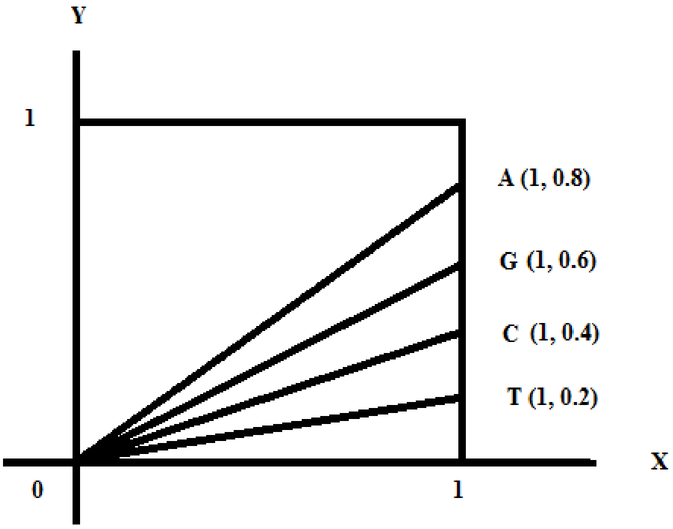

Fig1.1 2D vector representation of bases A, C, T, G

Next, the probability vector corresponding to the above 2D representation values for DNA sequence of length n, is given by $(p_1, p_2, p_3, p_4, p_5)\ where\ p_i = \frac{\boldsymbol{x_i - \vec{y}_i}}{\boldsymbol{\frac{1}{2}n(n-1)y_n}}$, (x_i, y_i) represents the position of the ith nucleotide in the DNA graphical curve, $\vec{y}_i$ represents the choice of y-coordinate value at the ith nucleotide in the DNA graphical curve according to Fig. 1.1.

The representation may be understood from the following sample sequence (ATGGT),

$\vec{y}_1$=0.8, $\vec{y}_2$=0.2, $\vec{y}_3$=0.6, $\vec{y}_4$=0.6, $\vec{y}_5$=0.2, y_5=2.4;

$$(p_1, p_2, p_3, p_4, p_5)$$
$$= (\frac{1-0.8}{\frac{1}{2}\cdot 5\cdot 6-2.4}, \frac{2-0.2}{\frac{1}{2}\cdot 5\cdot 6-2.4}, \frac{3-0.6}{\frac{1}{2}\cdot 5\cdot 6-2.4}, \frac{4-0.6}{\frac{1}{2}\cdot 5\cdot 6-2.4}, \frac{5-0.2}{\frac{1}{2}\cdot 5\cdot 6-2.4})$$

= (0.0159, 0.1429, 0:1905, 0.2698, 0.3810)

1.2 Di-Nucleotide Representation of DNA Sequences

Mononucleotide representation has some limitations in its applications; that is why, it has become necessary to consider di-nucleotide and tri-nucleotide representation. To highlight the necessity of introducing di-nucleotide representation, we look back to 1980, when people started to assess the possible biological significance of a computed structure. This included comparing the energy of folding of a natural single-stranded RNA sequence with the energies of several versions of the same sequence produced by shuffling base order. Later on many developments were made on di-nucleotide representation of DNA sequences. But as we are concerned with genome sequence comparison, so we restrict to the developments in this area only.

Now we come to applications of di-nucleotide representation in genome sequence comparison. We start with the paper of Zhao-Hui Qi and Tong-Rang Fan [36]. They consider a novel 3D graphical representation of DNA sequences based on the pairs of nucleotides (PNs). The model avoids loss of information as evidenced from the earlier work. Given a DNA primary sequence, there are 16 kinds of pairs of nucleotides. For notational convenience, PN denotes a pair of nucleotides. Now the rows and columns of a 4 × 4 matrix are assigned to pairs of nucleotides as given below:

Table 1.1 A 4 × 4 matrix using pairs of nucleotides

	A	**T**	**G**	**C**
A	AA	AT	AG	AC
T	TA	TT	TG	TC
G	GA	GT	GG	GC
C	CA	CT	CG	CC

The matrix is somewhat analogous to the reduced (4 × 4) matrix in [18].

In this matrix each entry is associated with a pair of labels. Thus in the first row we have entries AA, AT, AG and AC. Now one pair of labels is assigned as follows:

(1, 0, 0) ⟶AA, (2, 0, 0) ⟶AT, (3, 0, 0) ⟶ AG, (4, 0, 0) ⟶ AC, (5, 0, 0) ⟶ TA, (6, 0, 0) ⟶ TT, (7, 0, 0) ⟶ TG, (8, 0, 0) ⟶TC,(9, 0, 0) ⟶GA, (10, 0, 0) ⟶GT, (11, 0, 0) ⟶ GG, (12, 0, 0) ⟶GC,(13, 0, 0) ⟶ CA, (14, 0, 0) ⟶CT, (15, 0, 0) ⟶CG, (16, 0, 0) ⟶ CC.

The corresponding curves extend along z axes.

Now we mention another interesting di-nucleotide representation of DNA sequences in [37]. The representation shown in Fig. 1.2.

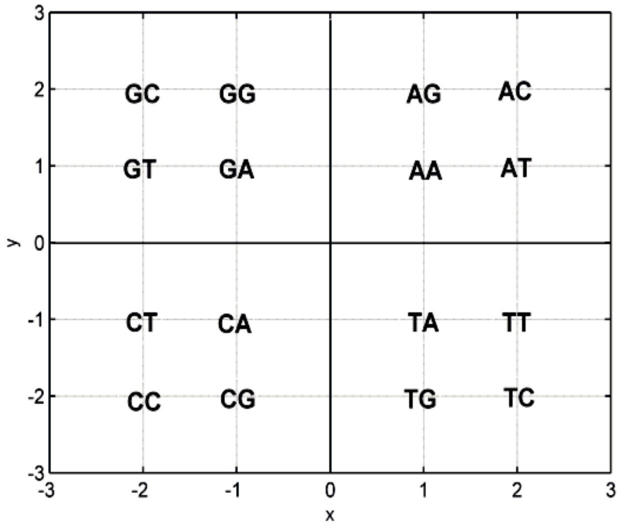

Fig. 1.2 Distribution of Di-nucleotide in Cartesian 2D coordinates

The rule behind this representation is the following:

Each kind of nucleotide is represented by (x, y), in all the four quadrants. The signs of (x, y) is decided by the occurrence of the first base according to the rule A⟶(+, +), G⟶(−, +), C⟶(−, −) and T⟶(+, −). The absolute value of decided by the occurrence of the base at the second site according as $(|x| = 1, |y| = 1) \longrightarrow A, (|x| = 1, |y| = 2) \longrightarrow G, (|x| = 2, |y| = 2) \longrightarrow C, (|x| = 2, |y| = 1) \longrightarrow T$.

Next we consider another di-nucleotide representation which gives a PNN curve [38].

Another interesting di-nucleotide representation may be mentioned [39]. In this paper the authors introduced a novel 2D graphical representation of DNA sequences based on the magic circle, which correspond to 16 dual nucleotides. So, we can reduce a DNA sequence into a 2D plot set as given by the points $\theta_i = \frac{2i\pi}{16} for\ i = 1,2,3, \dots 15,16$.

There are other interesting di-nucleotide representations. But as we are interested in the tri-nucleotide representations, we do not make the list longer.

1.3 Tri- Nucleotide Representations

Again, as tri-nucleotide gives more biological information than mono and di-nucleotides; so tri-nucleotide representation of a genome sequence is a comparatively better choice for sequence comparison. This is why focus has also been given on the genome sequence comparison under tri-nucleotide representations [40-41]. The reason that tri-nucleotide representations are fewer in number is because of limitations of visualizations.

The tri-nucleotide representation as taken up in [41] is as follows: the first base of the tri-nucleotide is assigned as 1⟶A, 2⟶G, 3⟶C, 4⟶T. It is noted that in a tri-nucleotide the second base is always associated with *hydrophobic/hydrophilic* property of the amino acid. So, the second base is determined by the signs of the first and third base. The rule is A⟶(+,+), G⟶(-,+), C⟶(-,-), T⟶(+,-). In [40] 2D graphical representation of tri-nucleotides is determined by the interior of a square; whereas 3D representations of tri-nucleotides are obtained by using the interior of a tetrahedron. Very recently such tri-nucleotide representation, in the name of 3-mer representation, is considered in [42]. In [40] first of all 2D tri-nucleotide representation is obtained the rule of chaos game. But they have not tried for any DNA sequence comparison based on such representation**.** In [43], 2D tri-nucleotide representation of the above papers has been generalized to 3D tri-nucleotide representation in an alternative simple way.

1.4 Descriptors and Distance Measures

There are three types of descriptors for comparison of DNA sequences. The first one is called Geometrical Descriptor and the second one is called Matrix Descriptor and third one is called Probabilistic descriptor.

1.4.1 Geometrical Descriptor

Geometric descriptors are obtained directly from the data points of the graph. In geometrical descriptor, some geometrical invariants of the embedded curve are taken as mathematical descriptors. They may be first order moments (μ_x, μ_y), a graph radius g_R and angle θ subtending with the x-axis defined for each sequence by the formulae $\mu_x = \frac{\sum x_i}{N}, \mu_y = \frac{\sum y_i}{N}, g_R = ({\mu_x}^2 + {\mu_y}^2)^{1/2}$, $tan\, \theta = \frac{\mu_y}{\mu_x}$ where (x_i, y_i) represent the co-ordinates of points on the plot and N is the total number of bases in the segment [44]. Here g_R represents the Base Distribution index and is critically dependent on the position of each base in the sequence. The descriptor may also be relative departure ρ given by $\rho = \frac{1}{N}\sum_1^N \sqrt{\left(x_i - \frac{i}{2}\right)^2 + \left(y_i - \frac{i}{2}\right)^2} = \frac{\sqrt{2}}{2N}\sum_1^N |x_i - y_i| = \frac{\sqrt{2}}{2N}\sum_1^N |(A_i + T_i) - (G_i + C_i)|$

$where\ i = 1,2, \dots, N$

Distance Measure

The distance measure used is Euclidean. In fact, with the help of such distance measure graph similarity/ dissimilarity index $\Delta g_R = [(\mu_{1x} - \mu_{2x})^2 + (\mu_{1y} - \mu_{2y})^2]^{1/2}$ is calculated where μ_1, μ_2 refer to two different DNA sequences.

1.4.2 Matrix Form of Descriptor

In matrix association method a matrix is associated with the coordinates of the points of the graph. Then the descriptors are obtained from the matrices chosen above. The matrices are of the following types D/D, L/L, M/M, J/J and their higher orders [45, 18, 10, 11, 43, 46, 36].

D/D Matrix [45, 18]

A graph theoretic distance matrix D can be formulated as $D_{ij} = (d_{ij})$, where d_{ij} is the number of edges between vertices i and j in the embedded graph. A large number of graph invariants have been formulated based on different types of matrices [45, 18]. One particular matrix, the D/D matrix and its leading eigen value, have been used to quantify shapes of graphs [45]. The elements of the D_E/D_G matrix is $(d_E/d_G)_{ij}$, where d_E represents the Euclidean distance between vertices i and j, whereas d_G is the graph theoretical (topological) distance between the vertex pair (i, j). Such distance/distance (D_E/D_G, or D/D for short) matrices could be directly computed for their eigen values. However, because Euclidean distances are always equal to or less than the graph-theoretical distances by construction, the matrix elements were raised to high powers until all elements <1 vanished leaving only the unit ratios from which the leading eigen values could be easily computed. Randic, Vracko, Nandy and Basak [18] showing the applicability of this technique, leading eigen values of the D/D and associated matrices have been considered to be good descriptors of DNA sequences.

L/L Matrix [10, 11, 36]

The length/length (L/L) matrix also is symmetric matrix whose off-diagonal elements are given as a quotient of the Euclidean distance between a pair of vertices and the sum of geometrical lengths of edges between the same pair of vertices. The $^kL/^kL$ matrix is constructed from the L/L matrix by raising its individual matrix elements to the k^th^ power.

M/M Matrix [36, 43]

The M/M matrix is symmetric matrix whose off-diagonal elements are given as a quotient of the Euclidean distance between two vertices and graph theoretical distance between the two vertices. The entries on the main diagonal are defined by zero.

The M/M matrix is calculated as:

$M_{ij} = \frac{\sqrt{(x_i-x_j)^2+(y_i-y_j)^2+(z_i-z_j)^2}}{|x_i-x_j|+|y_i-y_j|+|z_i-z_j|}$ $where\ (x_i, y_i, z_i)\ and\ (x_j, y_j, z_j)$ represent the co-ordinates of points on the plot for two different species.

Matrix forms of descriptors are also found with di-nucleotide representation [36]**.** They consider matrix method for sequence comparison using D/D, M/M and L/L form of matrices. The utility of the method is illustrated by the examination of similarities/dissimilarities among the complete coding sequence part of *β-globin* gene of several different species. In [36] first of all covariance matrix is determined and then eigen values of this matrix are taken as descriptors.

J/J Matrix [46]

J/J matrix is similar to M/M matrix, the only difference is that in M/M matrix all the differences of (x, y, z) are taken simultaneously, where as in J/J matrix the corresponding differences of the variables x and y are consider first; the difference of z values are considered finally.

The J/J matrix is calculated as:

$$J_{ij} = \frac{\sqrt{(x_i - x_j)^2 + \left(y_i - y_j\right)^2 +}}{\left|x_i - x_j\right| + \left|y_i - y_j\right|} + \frac{\sqrt{(z_i - z_j)^2}}{\left|z_i + z_j\right|}$$

It is found that J/J matrix is most useful out of all such matrices [43].

Distance Measures

They are mostly Euclidean. Sometimes other distance measure like Person's Correlation coefficient is also used [49].

1.4.3 Probabilistic Descriptors

The probability vectors obtained from the graphical representations are the probabilistic descriptors [35].

Distance measures

Kullback–Leibler Divergence (KLD)

In such cases the distance measures are suitable divergence measures. In [35], the distance measure was taken as the symmetric form J of Kullback–Leibler divergence (KLD), where the non-symmetric form I of Kullback-Leibler divergence I is given by

$$I(P,Q) = \sum_{x \in X} P(x)\, log \frac{P(x)}{Q(x)} = \sum_{i=1}^{n} P_i(x)\, log \frac{P_i(x)}{Q_i(x)}$$

P and *Q* being two discrete probability distributions on a universe *X*,

The corresponding symmetric form J of Kullback–Leibler divergence is given by

$J(P,Q) = \frac{I(P.Q)+I(Q,P)}{2}$ where $I(P,Q) = \sum_{x \in X} P(x)\, log \frac{P(x)}{Q(x)} = \sum_{i=1}^{n} P_i(x)\, log \frac{P_i(x)}{Q_i(x)}$

1.5 Main Focus Area

Up to this we have described, in brief, all common type of nucleotide representations of DNA sequences and their descriptors, in general.

We now describe, in details, some special type of mono-nucleotide and tri-nucleotide representations, their descriptors, and their distance measures, which have motivated the work pursued in the thesis

1.5.1 Representation Based on the Three Pair of Classified Curves

2D representation based on pair of classified curves is one of the most important types; some such representations are following,-

G, C, A, T represents the nucleotide bases, whereas G_i, C_i, A_i, T_i are the cumulative occurrence numbers of G, C, A and T respectively, in the subsequence from the first base to the ith base in the sequence. With these notations the following allied works may be cited.

$x_i = C_i - G_i, y_i = T_i - A_i$ in [1],
$x_i = G_i - A_i, y_i = C_i - T_i$ in [2],
$x_i = A_i - C_i, y_i = T_i - G_i$ in [3],
$x_i = (A_i + T_i) - (G_i + C_i),\ \ y_i = (G_i + T_i) - (A_i + C_i),\quad z_i = (C_i + T_i) - (A_i + G_i)$ in [18],
$x_i = (A_i + G_i) - (C_i + T_i),\ \ y_i = (A_i + C_i) - (G_i + T_i),\ \ z_i = (A_i + T_i) - (G_i + C_i)$ in [23]
$\phi(s_i) = (G_i + C_i, A_i + T_i), S = s_1 s_2 s_3 \ldots s_n$ in [44]
In [44] the 2D represented points on the curves are given by $P_i = (x_i, y_i)$ where $x_i = G_i + C_i, y_i = A_i + T_i$;$x_i = G_i + T_i, y_i = A_i + C_i$; $x_i = C_i + T_i, y_i = A_i + G_i$; for W-S curve, W = {A, T}, S = {G, C}; M-K curve, M = {A, C}; K = {G, T}; R-Y curve, R = {A, G}, Y = {C, T} respectively. Further G_i, C_i, A_i, T_i are the cumulative occurrence numbers of G, C, A and T respectively, in the subsequence from the first base to the ith base in the sequence. It is found that in this paper summation is used in place of difference as used in the earlier similar papers. So the authors claim that the representation is non-degenerate.

Descriptors

Descriptors used are first order moments (μ_x, μ_y), a graph radius g_R and angle θ subtending with the x-axis defined for each sequence by the formulae $\mu_x = \frac{\sum x_i}{N}, \mu_y = \frac{\sum y_i}{N}, g_R = (\mu_x{}^2 + \mu_y{}^2)^{1/2}$, $tan\,\theta = \frac{\mu_y}{\mu_x}$ where (x_i, y_i) represent the co-ordinates of points on the plot and N is the total number of bases in the segment [44]. Here g_R represents the Base Distribution index and is critically dependent on the position of each base in the sequence. The descriptor may also be relative departure ρ given by $\rho = \frac{1}{N}\sum_1^N \sqrt{\left(x_i - \frac{i}{2}\right)^2 + \left(y_i - \frac{i}{2}\right)^2} = \frac{\sqrt{2}}{2N}\sum_1^N |x_i - y_i| = \frac{\sqrt{2}}{2N}\sum_1^N |(A_i + T_i) - (G_i + C_i)|$

Distance Measure

The distance measure used is Euclidean. In fact, with the help of such distance measure graph similarity/ dissimilarity index $\Delta g_R = [(\mu_{1x} - \mu_{2x})^2 + (\mu_{1y} - \mu_{2y})^2]^{1/2}$ is calculated where μ_1, μ_2 refer to two different DNA sequences.

Main Result:

The representation is non-degenerate and under the above descriptors, *β-globin* genes of 10 species can be compared effectively.

1.5.2 Fuzzy Representation on a Twelve Dimensional Hypercube

Representation of Fuzzy Polynucleotide Space of Torres and Nieto [47]

Fuzzy set theory is realized in the process of representing a polynucleotide consisting of finite number of codons on a single hypercube I^{12}. This is the background of fuzzy polynucleotide space as introduced by Torres and Nieto (2003) [47]. 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. 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.

Example: S1=UACUGU *tyrosine / cysteine*

Table 1.2 No. of Nucleotides, Total Nucleotides and Fraction of Nucleotides

	No. of Nucleotides				**Total Nucleotides**	**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	0.5	0.5
3rd base	1	1	0	0	2	0.5	0.5	0	0

From Table 1.2 fuzzy representation of S1= UACUGU *tyrosine/cysteine* is (1,0,0,0,0,0,.5,.5,.5,.5,0,0),

Descriptors

To compare genome sequences, descriptors are taken as the 12 component fuzzy vectors generated by each sequence.

Distance Measures [48]

The distance measures are the NTV metric given by

$$d(p,q) = \frac{\sum_{i=1}^{12} |p_i - q_i|}{\sum_{i=1}^{12} max\{p_i, q_i\}} \qquad 1.1$$

and its equivalent metrics given by

$$d_1(p,q) = \frac{d(p,q)}{1 + d(p,q)} \qquad 1.2$$

$$d_2(p,q) = \frac{\sqrt{\sum_{i=1}^{12} (p_i - q_i)^2}}{\sqrt{12}} \qquad 1.3$$

$d_3(p,q) = \frac{d_2(p,q)}{1+d_2(p,q)}$ 1.4

$d_4(p,q) = \frac{\sum_{i=1}^{12}|p_i - q_i|}{12}$ 1.5

Main Result

Comparison based on such fuzzy representation was done on three whole genome sequences and the results of comparison show identical behavior under all distance measures Eq. 1.1 – Eq. 1.5.

1.5.3 *K*-mer representations

This form of representation is a generalization of tri-nucleotide representation. It coincides with a tri-nuclear representation when *k* = 3. It is also called a composition vector method. *k*-mer is a probabilistic approach.

Descriptors

Descriptors are the probability vectors obtained as follows:

Step 1: Frequency or Rank vector

In a molecular sequence (nucleotide or amino acid sequence) of length *N*, any consecutive *k* molecules is called a *k*-string or a *k*-tuple or a *k*-mer, where$1 \leq k \leq N$. Computationally, the *k*-mers are collected by using a sliding window of length *k*. It slides through the sequence by shifting one position at a time. Thereby, *N*-*k*+1 number of such overlapping strings are obtained. The *k*-strings are denoted by the variable *u*. *g(u)* represents the frequency vector or the rank vector of the k-strings. This calculates how many times each *k*-string appears in the sequence

Step 2: Probability Vector:

The probability vector is given by: $f(u) = \frac{g(u)}{N-k+1}$

This formula is applicable when the whole genome sequence is considered as a single entity. However, if only protein coding DNA sequences from the whole genome sequence are considered, then this formula is modified as $f(u) = \frac{\sum_{j=1}^{m} g_j(u)}{\sum_{j=1}^{m}(N_j - k+1)}$ where m is the number of protein-coding DNA sequences from the whole genome, $g_j(u)$ is the number of times that *u* appears in the j^{th} DNA sequence and N_j is the length of j^{th} DNA sequence.

The modified formula avoids the problems, which might occur from the gene order and gene content in a genome sequence.

Step 3: Composition vector:

Generally, the biological data are often obscured by noise and bias. Thus, a signal de-noising process is performed before obtaining the composition vector *h(u)* from the probability vector *f(u)*. Specifically, for each f(u), the estimated noise q(u) is calculated. Afterward, the composition vector h(u) is determined by calculating the signal-to-noise-ratio given by: $h(u) = \frac{f(u)-q(u)}{q(u)}$

Distance Measures

The distance between every pair of composition vectors is calculated by applying some distance measures, which are either angle based or which use divergence measures. Each of the used distance measures *d* between any two vectors *a* and *b* satisfies the following properties (metric):

$d(a,b) \geq 0,$

$d(a,b) = 0 \;\; if\; and\; only\; if\; a = b,$

$$d(a,b) = d(b,a) \qquad (1.6)$$

However, none of them satisfies the triangular inequality property *a*, *b* and *c*, which is given by:

$$d(a,b) \leq d(a,c) + d(c,b) \qquad (1.7)$$

Still they are taken as distance measures, where the measures are as follows.

i) Angle based measure, where the distance measures are expressed as given in [50, 51] and [52, 53]; respectively as:

$$d^{Stuart}(a,b) = -\,log(\frac{1+cos\,\theta}{2}) \qquad (1.8)$$

where the *Cosine* of the angle between two vectors *a* and *b* with dot product *a.b* is given by: $cos\,\theta = \frac{a.b}{\|a\|\|b\|}$, where θ is the angle and $\|.\|$ denotes the Euclidean norm.

ii) Information based measures, where the divergence can be given by:

(a) Kullback-Leibler divergence [54, 55]

$$KL(a,b) = \sum_{i=1}^{n} a_i \, log \frac{a_i}{b_i} \tag{1.9}$$

a = ai, b = bi are the distribution vectors with i = 1,2,...,n.

(b) Jensen-Shannon divergence [55]

$$d^{JS}(a,b) = \sum_{i=1}^{n} [a_i \, log(\, a_i) + b_i \, log(\, b_i) - (a_i + b_i) \, log(\frac{a_i+b_i}{2})] \tag{1.10}$$

iii) Information based similarity index

The information based similarity index (D_k) using *k*-tuple nucleotides between two sequences S_1 and S_2 is given by:

$$D_k(S_1, S_2) = \frac{1}{4^k - 1} \sum_{i=1}^{4^k} |R_1(w_i) - R_2(w_i)| \frac{H_1(w_i) + H_2(w_i)}{\sum_{i=1}^{4^k} [H_1(w_i) + H_2(w_i)]} \tag{1.11}$$

where $R_1(W_i)$, R_2 (W_i) and $H_1(W_i)$, H_2 (W_i) represent the rank and Shannon entropy of a specific *k*-tuple w_i in the sequences S_1 and S_2 respectively.

Main Results

Composition method is an effective method in DNA sequence comparison. But sometimes the results of comparison in the form of phylogenetic trees are not satisfactory. So authors of [50] considered improvements of the CV method in the form of CCV (complete Composition Vector) method and ICV (improved composition vector) method. Till then they could not determine exactly the string length *k* under which always satisfactory results would come. In this connection the observation of the authors in [56] is very important. In fact, they obtained ICV trees for *k* = 3, 4, 5, 6, 7, 8, 9, 10. It is shown that as *k* increases from 3 to 5, the supporting values significantly increase. However this trend decreases as *k* varies from 6 to 10. Thus cut off value may be taken as 5 or 6. The general conclusion is that increasing *k* after a certain number may not certainly improve the result.

From the above literature survey we think that the following questions are pertinent

1. Degeneracy in the representation is a major issue. In this connection, 2D representation of [44] based on the three pair of classified curves is claimed to be the only one, which is non-degenerate amongst all such similar earlier representations, as it uses summation in place of difference. But as the representation depends on the cumulative occurrence numbers of some i^{th} base in the subsequence from the first base, so even the present representation may be degenerate when compared with rearranged sequence. If so, the problem is how to ensure non-degeneracy?

2. Fuzzy representation of [47,48] is no doubt a very useful representation, Further the results of comparison of genome sequences based on all different types of metric on the fuzzy polynucleotide space makes the representation more powerful, as all the results show similar behavior. But as the result holds only for three genomes, so the question remains to see whether result holds in general!. If not, can there be a generalization of fuzzy polynucleotide space to settle the issue?

3. Composition vector methods using *k*-mer involve tri-nucleotide representation. These are applied under choice of different values of *k* and under choice of different distance measures. As it is found that for some values of *k*, and for some choice of distance measure the results are satisfactory, but for other cases results are unsatisfactory, so is it possible to find an optimal value of *k* and a suitable distance measure to get uniform results?

4. In *k*-mer method Information based similarity index is used as a good distance measure. But this is a probabilistic measure. Further for this measure, triangular inequality has not yet been proved; so this measure cannot be a called a 'distance' in the proper sense of the term. Hence a natural query is to see whether it is possible to have an alternative tri-nuclear representation, which is not probabilistic, but can be non-degenerate and can perform equally well as the *k*-mer method under a standard distance measure?

The present thesis attempts to answer these questions as far as possible in four different chapters.

CHAPTER 2

2. Geometrical Method of Comparison Under New 3D Classification Curves

In this chapter first of all degeneracy of the 2D representation of [44] is established through an example. Next corresponding non-degenerate 3D representations are obtained. Finally new 3D descriptors are introduced and DNA sequence comparison is made using such new descriptors. The results are also compared with the earlier ones obtained on the same sequences. By statistical analysis, it is shown that our results are significantly different from the rest.

2.1 Outline of the Existing Method and its Limitation

The 2D represented points on the curves are given by $P_i = (x_i, y_i)$where$x_i = G_i + C_i, y_i = A_i + T_i$;$x_i = G_i + T_i, y_i = A_i + C_i; x_i = C_i + T_i, y_i = A_i + G_i;$ for W-S curve, W = {A, T}, S = {G, C}; M-K curve, M = {A, C}; K = {G, T}; R-Y curve, R = {A, G}, Y = {C, T} respectively. Further G_i, C_i, A_i, T_i are the cumulative occurrence numbers of G, C, A and T respectively, in the subsequence from the first base to the ith base in the sequence.

To show the degeneracy of the method [44], we consider 2D graphical representation of the Sequence S = ATGGTGCACCTGACTCCTGA of the first 20 nucleotides of first exon of B-globin gene of Human as taken up in [44]. For W-S curve, obviously the points of representation are sequentially (0+0, 1+0) = (0,1), (0+0,1+1) = (0,2), (0+1,1+1)=(1,2), (1+1,1+1) = (2,2), (1+1,1+2) = (2,3) etc.

Now let us see what happens if we change the first five nucleotides as TAGGA and keep others unaltered? It is seen that the same points of representation are obtained for a different sequence also. This is true for change of other parts also. The reason is that it is the sum which matters, and not the individuals. The same observation may be made for other types of curves. So we note that the aforesaid 2D representation is degenerate, although apparently it looks non-degenerate.

2.2 3D Generalizations

We note that the degeneracy could be avoided by considering the frequencies of the nucleotides and putting them in the third coordinates of each point. For example for W-S 3D curve of S = ATGGTGCACCTGACTCCTGA, the pints of representations are (0,1,1), (0,2,1), (1,2,1), (2,2,2), (2,3,2) and so on. Obviously it is the inclusion of the frequency, which makes the 3D representation non-degenerate. Fig 2.1 describe 3D Graphical representation of classification curve of the DNA sequences of the first exon of *β-globin* gene of Human (W-S Curve)

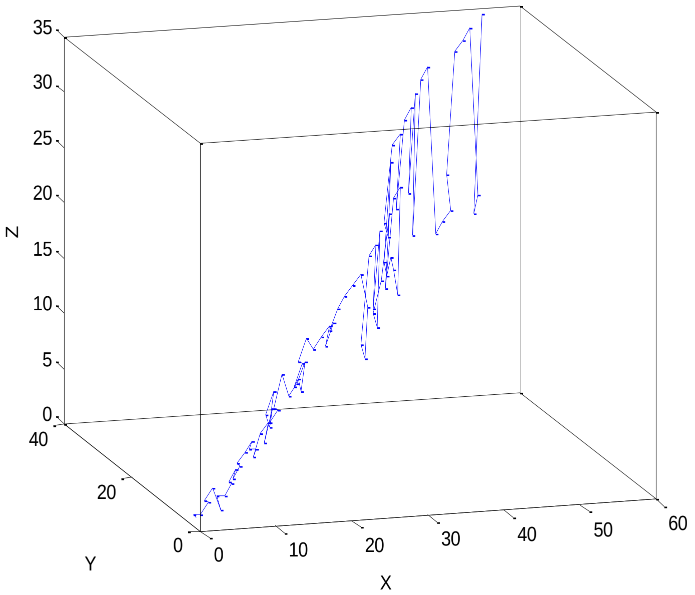

Fig. 2.1 3D Graphical representation of classification curve of the DNA sequences of the first exon of *β-globin* gene of Human (W-S Curve)

2.3 New 3D Component Descriptors

i) Graphical Radii

The Graph radius

$$g_{R1} = (\mu_{x1} + \mu_{y1} + \mu_{z1})/N, \text{where } \mu_{x1} = \sum_{j=1}^{N} x_{1j}\,, \mu_{y1} = \sum_{j=1}^{N} y_{1j}\,, \mu_{z1} = \sum_{j=1}^{N} z_{1j},$$

$$g_{R2} = (\mu_{x2} + \mu_{y2} + \mu_{z2})/N, \text{where } \mu_{x2} = \sum_{j=1}^{N} x_{2j}\,, \mu_{y2} = \sum_{j=1}^{N} y_{2j}\,, \mu_{z2} = \sum_{j=1}^{N} z_{2j},$$

$$g_{R3} = (\mu_{x3} + \mu_{y3} + \mu_{z3})/N, \text{where } \mu_{x3} = \sum_{j=1}^{N} x_{3j}\,, \mu_{y3} = \sum_{j=1}^{N} y_{3j}\,, \mu_{z3} = \sum_{j=1}^{N} z_{3j}$$

where (x_i, y_i, z_i) represent the co-ordinates of points on the plot and N is the total number of bases in the sequence.

ii) Average Angle Measure

$$\Phi = \frac{\frac{\mu_{x1}}{g_{R1}} + \frac{\mu_{y1}}{g_{R1}} + \frac{\mu_{z1}}{g_{R1}}}{3},$$

$$\Psi = \frac{\frac{\mu_{x2}}{g_{R2}} + \frac{\mu_{y2}}{g_{R2}} + \frac{\mu_{z2}}{g_{R2}}}{3},$$

$$\text{Ж} = \frac{\frac{\mu_{x3}}{g_{R3}} + \frac{\mu_{y3}}{g_{R3}} + \frac{\mu_{z3}}{g_{R3}}}{3}$$

iii) Relative Departure

$$\rho_i = \frac{1}{N_i\sqrt{3}}\left[3\sum_{j=1}^{N_i}({x_{ij}}^2 + {y_{ij}}^2 + {z_{ij}}^2) - \sum_{j=1}^{N_i}(x_{ij} + y_{ij} + z_{ij})^2\right]^{1/2}$$

$$i = 1,2,3$$

2.4 Result and Discussion

For our experiment, we have considered the coding sequence of first exon of *β-globin* gene of eleven different species shown in Table 2.1.

Table 2.1 The coding sequences of the first exon of *β-globin* gene of eleven different species

Species	Coding sequence
Human	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATTAAGTTGGTGGTGAGGCCCTGGGCAG
Chimpanzee	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAG
Bovine	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGCCTTTTGGGGCAAGGTGAAA
Gorilla	ATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Rat	ATGGTGCACCTAACTGATGCTGAGAAGGCTACTGTTAGTGGCCTGTGGGGAAAGGTGAACCCTGATAATG TTGGCGCTGAGGCCCTGGGCAG
Rabbit	ATGGTGCATCTGTCCAGTGAGGAGAAGTCTGCGGTCCTGCCCTGTGGGGCAAGGTGAATGTGGAAGAAGTTGGTGGTGAGGCCCTGGGC
Mouse	ATGGTTGCACCTGACTGATGCTGAGAAGTCTGCTGTCTCTTGCCTGTGGGCAAAGGTGAACCCCGATGAAGTTGGTGGTGAGGCCCTGGGCAGG
Lemur	ATGACTTTGCTGAGTGCTGAGGAGAATGCTCATGTCACCTCTCTGTGGGGCAAGGTGGATGTAGAGAAAGTTGGTGGCGAGGCCTTGGGCAG
Gallus	ATGGTGCACTGGACTGCTGAGGAGAAGCAGCTCATCACCGGCCTCTGGGGCAAGGTCAATGTGGCCGAATGTGGGGCCGAAGCCCTGGCCAG
Opossum	ATGGTGCACTTGACTTCTGAGGAGAAGAACTGCATCACTACCATCTGGTCTAAGGTGCAGGTTGACCAGACTGGTGGTGAGGCCCTTGGCAG
Goat	ATGCTGACTGCTGAGGAGAAGGCTGCCGTCACCGGCTTCTGGGGCAAGGTGAAAGTGGATGAAGTTGGTGCTGAGGCCCTGGGCAG

Table 2.2 The graph radii associated with three different patterns of the classification curves

Curves	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
W-S	35.676	33.705	35.102	35.969	35.438	36.125	35.050	35.278	36.224	33.591	40.806
M-K	35.736	33.676	35.179	35.494	36.262	36.639	35.611	35.618	36.230	33.712	40.901
R-Y	35.456	33.702	35.172	35.759	35.586	36.063	35.178	35.549	35.976	33.616	40.546

Table 2.3 The average angles associated with three different patterns of the classification curves

Curves	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
W-S	0.5567	0.5554	0.5581	0.5523	0.5617	0.5594	0.5597	0.5598	0.5568	0.5574	0.5559
M-K	0.5594	0.5559	0.5569	0.5597	0.5490	0.5516	0.5509	0.5545	0.5567	0.5554	0.5546
R-Y	0.5601	0.5555	0.5570	0.5555	0.5594	0.5604	0.5577	0.5556	0.5606	0.5570	0.5595

Table 2.4 The relative departures associated with three different patterns of the classification curves

Curves	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
W-S	0.1570	0.1664	0.1578	0.2130	0.1154	0.1544	0.2916	0.1386	0.1668	0.1519	0.1453
M-K	0.1821	0.1664	0.1693	0.1161	0.2087	0.1881	0.3237	0.1686	0.1801	0.1811	0.1906
R-Y	0.1343	0.1664	0.1629	0.1514	0.1570	0.1280	0.2987	0.1578	0.1413	0.1664	0.1453

Then we calculate graph radii, average angles and relative departures associated with three different patterns of the classification curves which are shown in Table 2.2, Table 2.3 and Table 2.4 respectively. We now construct similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the normalized graph radii, average angles and relative departures which are shown in Table 2.5, Table 2.6 and Table 2.7 respectively.

Table 2.5 Similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the normalized graph radii

	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.0083	0.0093	0.0052	0.0064	0.0042	0.0092	0.0046	0.0025	0.0071	0.0016
Goat		0	0.0169	0.0067	0.0088	0.0114	0.0049	0.0110	0.0059	0.0017	0.0069
Opossum			0	0.0119	0.0131	0.0080	0.0177	0.0066	0.0116	0.0158	0.0108
Gallus				0	0.0103	0.0092	0.0102	0.0080	0.0044	0.0066	0.0051
Lemur					0	0.0055	0.0061	0.0072	0.0063	0.0071	0.0058
Mouse						0	0.0106	0.0039	0.0061	0.0098	0.0050
Rabbit							0	0.0113	0.0073	0.0039	0.0078
Rat								0	0.0065	0.0097	0.0057
Gorilla									0	0.0048	0.0011
Bovine										0	0.0057
Chimpanzee											0

Table 2.6 Similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the average angles

	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.00067	0.00005	0.00007	0.00013	0.00027	0.00020	0.00008	0.00013	0.00069	0.00134
Goat		0	0.00070	0.00073	0.00072	0.00093	0.00049	0.00071	0.00080	0.00003	0.00201
Opossum			0	0.00007	0.00010	0.00024	0.00022	0.00004	0.00011	0.00072	0.00131
Gallus				0	0.00016	0.00023	0.00027	0.00010	0.00010	0.00075	0.00128
Lemur					0	0.00022	0.00022	0.00008	0.00013	0.00074	0.00130
Mouse						0	0.00044	0.00022	0.00014	0.00095	0.00108
Rabbit							0	0.00023	0.00032	0.00051	0.00152
Rat								0	0.00011	0.00073	0.00130
Gorilla									0	0.00082	0.00121
Bovine										0	0.00203
Chimpanzee											0

Table 2.7 Similarity/dissimilarity matrix based on the Euclidean distances between the end points of the 3-component vectors of the relative departures

	Human	Goat	Opossum	Gallus	Lemur	Mouse	Rabbit	Rat	Gorilla	Bovine	Chimpanzee
Human	0	0.00053	0.00034	0.00089	0.00059	0.00012	0.00290	0.00036	0.00011	0.00049	0.00037
Goat		0	0.00029	0.00083	0.00079	0.00065	0.00253	0.00049	0.00044	0.00024	0.00078
Opossum			0	0.00084	0.00063	0.00045	0.00279	0.00022	0.00028	0.00032	0.00051
Gallus				0	0.00146	0.00104	0.00302	0.00099	0.00086	0.00105	0.00112
Lemur					0	0.00058	0.00288	0.00050	0.00066	0.00058	0.00057
Mouse						0	0.00299	0.00041	0.00023	0.00060	0.00032
Rabbit							0	0.00295	0.00284	0.00251	0.00322
Rat								0	0.00036	0.00044	0.00035
Gorilla									0	0.00045	0.00045
Bovine										0	0.00073
Chimpanzee											0

We now like to compare our results with the earlier ones obtained under different methods. In the earlier work the results are known for 10 species except goat. So for the sake of comparison we have also chosen the results for the same 10 species only. For convenience, we have used A, B, C to denote the similarity/dissimilarity of the coding sequences of the first exon of the human β-globin gene based on the 3-component vectors of normalized graph radii,the 3-component vectors of average angles, and the 3-component vectors of normalized relative departures respectively. A_1 corresponds to 3-

component graph radius of [11], B_1 corresponds to 3-component vector angles of [14], C_1 corresponds to 3-component relative departures [57]. All are given in the Table 2.8.

Table 2.8 Comparison the values calculated by our method (A, B, C) with other methods (A_1, B_1, C_1)

	A	B	C	A_1	B_1	C_1
Human	0.00827	0.00067	0.00053	0.00567	0.2896	0.0259
Opossum	0.01685	0.0007	0.00029	0.00877	0.4667	0.0485
Gallus	0.00667	0.00073	0.00083	0.00462	0.0235	0.1558
Lemur	0.00881	0.00072	0.00079	0.00822	0.3671	0.0401
Mouse	0.01136	0.00093	0.00065	0.00664	0.2902	0.0315
Rabbit	0.00489	0.00049	0.00253	0.00529	0.1369	0.0212
Rat	0.01102	0.00071	0.00049	0.00634	0.2726	0.036
Gorilla	0.00595	0.0008	0.00044	0.00514	0.254	0.0215
Bovine	0.00171	0.00003	0.00024	0.00191	0.0678	0.008
Chimpanzee	0.00691	0.00201	0.00078	0.00509	0.2189	0.0185

2.5 Comparison using Statistical Hypothesis Testing

Obviously when we compare results of two column vectors, A and A_1, say, each of size 10, we look for equality of means of their populations. This leads to the following hypothesis testing under .05 level of significance error:

H_0: There is no significance difference between the means of the populations of A and A_1.

H_1: There is a significance difference between the means of the populations of A and A_1.

Again H_0 is effective, if the calculated value of ‘t’ statistics is less than the prescribed value, and H_1 is effective, if the calculated value of ‘t’ statistics exceeds the prescribed value.

Now $t = \frac{\bar{x}-\bar{y}}{S.E.(\bar{x}-\bar{y})}$, where $\bar{x}, \bar{y}$ are the sample means and $S.E.(\bar{x}-\bar{y})$ represents the standard error of $(\bar{x}-\bar{y})$. Again $S.E.(\bar{x}-\bar{y})$ for samples of sizes n_1 and n_2 is determined by two formulae, according as (i) the sample variances can be assumed to be the same and (ii) the sample variances cannot be assumed to be the same. Again the condition of equality of sample variances is determined by F-tests (Fisher's F-statistics). In this case also, two hypothesis are taken, viz., H_0: There is no difference between the sample variances and H_A: There is a difference between the sample variances as determined by F-tests (Fisher's F-statistics). F-statistics is given by $F = s_1^2/s_2^2$ where $s_1^2 = \frac{n_1}{n_1-1}S_1^2$ and $s_2^2 = \frac{n_2}{n_2-1}S_2^2$ are the unbiased estimators of population variances; S_1^2, S_2^2 are the sample variances. Value of F is to be compared with the prescribed value at (n_1-1, n_2 -1) degrees of freedom. Now if the value of the F statistics is less than the prescribed value, H_0holds, otherwise H_A holds. Thus first of all by F-test, it is to be decided, which of the above two cases (i) or (ii) is to be considered.

By actual calculation we obtain F=4.647, this value is greater than the standard value of $F_{.05,9,9} = 3.18$. So the t test is to be applied in the second case. This type of t' test is called Welch's approximation 't' test [58]. In this case $t = \frac{\bar{x}-\bar{y}}{S.E.(\bar{x}-\bar{y})}$, where $S.E(\bar{x}-\bar{y}) = \sqrt{(s_1^2/n_1) + (s_2^2/n_2)}$

and the degree of freedom is given by $\nu' = \frac{(s_1^2/n_1+s_2^2/n_2)^2}{((s_1^2/n_1)^2/(n_1-1))+((s_2^2/n_2)^2/(n_2-1))}$

On calculation the value of $\nu'$is found to be 0.0000000034. As it is not an integer so we take the next least integral for ν'. This gives us ν'= 1. On calculation the value of't' is found to be t=443.53. As this value of 't' is greater than the prescribed value of $t_{(2),.05,\nu'}$ =6.31, so we conclude that the results of A and A_1 significantly differ. Similarly we have proved that results of B and B_1; C and C_1 also differed significantly.

2.6 Conclusion

In this paper, we have outlined a 3D graphical representation of DNA sequences based on three types of classification curves, and presented a variant of a mathematical representation for DNA sequences in 3D graphs. Some advantages of classification curves are as follows:

1. The distributions of bases of different types are strictly displayed in these three characteristic curves of the corresponding DNA sequence.
2. In comparison to previous works that used combinations of sums and subtractions, and even that, which involves only sum [44], it properly eliminates plot degeneracy and

can completely avoid loss of information in the transfer of data from a DNA sequence to its mathematical representation.

3. The three characteristics used are graph radii, average angles and relative departures, which are extracted from the 3D graphs and then used to calculate distance values among sequences. Comparison of the results of the examination of similarities/dissimilarities among the coding sequences of the first exon of *β-globin* gene of 10 species with other geometrical methods illustrates the utility of the approach.

4. Although our results look similar to the results obtained by other methods, but statistical analysis shows that they are significantly different. This proves the necessity of introducing such new measures to compare DNA sequences. .

5. Lastly the method is sound and one can find that the computational complexity is only *O(N)*; this greatly reduces the computational time.

CHAPTER 3

3. Some Problems of Fuzzy Polynucleotide Metric Space and Necessary Modifications

This chapter may be subdivided in two sections;

The first section critically examines the results obtained in the comparison of whole genome sequences on the fuzzy metric polynucleotide space [48] and highlights some associated problems in connection with fuzzy representation, which challenges the very construction of fuzzy polynucleotide space. The second section develops some alternative form of general representation, which is free from such allied problems.

Section 3.1

3.1.1 Fuzzy Polynucleotide Space [47]

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. 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. 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, 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. This is the background of fuzzy polynucleotide space as introduced by Torres and Nieto (2003) [47].

3.1.2 Fuzzy Polynucleotide Metric Space [48]

Later on, in (2006) [48] the authors used different types of metric for comparison of polynucleotides and whole genomes. The metrics are of following types:

$$d(p,q) = \frac{\sum_{i=1}^{12}|p_i - q_i|}{\sum_{i=1}^{12} max\{p_i, q_i\}} \quad 3.1$$

$$d_1(p,q) = \frac{d(p,q)}{1+d(p,q)} \quad 3.2$$

$$d_2(p,q) = \frac{\sqrt{\sum_{i=1}^{12}(p_i - q_i)^2}}{\sqrt{12}} \quad 3.3$$

$$d_3(p,q) = \frac{d_2(p,q)}{1+d_2(p,q)} \quad 3.4$$

$$d_4(p,q) = \frac{\sum_{i=1}^{12}|p_i - q_i|}{12} \quad 3.5$$

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

3.1.3 Results on Sequence Comparison

The authors of Nieto et al. (2006) [48] also show that the role of different metrics remains the same in cases of complete genomes also. They consider fuzzy sets of frequencies of the genomes of *M. tuberculosis*, *E. coli* and *A. Aeolicus*. Using the various metrics they compute the distance between *M. tuberculosis* and *E. coli*, the distance between *M. tuberculosis* and *A. Aeolicus* and also the distance between *E.coli* and *A. Aerolicus*. These results show that for complete genomes, the role of different metrics remains the same. In fact, those whole genomes that are known to be biologically similar from their emboss values, are also found to be nearer under the measure of all such metrics.

3.1.4 Limitations of the Results

The results are verified only for three particular types of whole genomes. So from these three results it cannot be concluded that all the metrics behave similarly for all different genome sequences. In support of this we first obtain some counter examples:-

3.1.5 Problems with Fuzzy Polynucleotide Metric Space

Some counter examples

(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.

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)

Table 3.1 The detailed calculations of distances under different metrics

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

Comparison of the results of Table 3.1

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*)

Proposition follows from these results.

Remark

Net conclusion is that all the metrics do not behave similarly for whole genomes, in general. This is quite unexpected as the metrics are defined on finite dimensional spaces. This challenges the very construction of the Fuzzy polynucleotide space. So we introduce the new concept of Intuitionistic fuzzy polynucleotide metric space and carry on investigation of comparison of whole genome sequences on this space in order to show that no unexpected result occurs any more.

Section 3.2

3.2.1 Intuitionistic Fuzzy Set

Intuitionistic Fuzzy Sets [59, 60] are generalization of Fuzzy sets [61] in which non-membership values are not obtainable from the membership values, rather both of them have to be specified separately.

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 $0 \leq \mu_A(x) + \nu_A(x) \leq 1$,for each x in X. Obviously an ordinary fuzzy set can be written as $\{\langle x, \mu_A(x), 1 - \mu_A(x)\rangle, x \in X\}$.

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 is written as

$\{< x, \mu_A(x), \nu_A(x), \pi_A(x) >, x \in X\}$.

Distance Measure on Intuitionistic Fuzzy Set

The normalized hamming distance DIFS proposed for IFS by [62] is given by

$D_{\text{IFS}}(\text{A,B}) = \sum_{i=1}^{n}(|\mu_A\,(x_i) - \mu_B(x_i)| + |\nu_A(x_i) - \nu_B(x_i)| + |\pi_A(x_i) - \pi_B(x_i)|)$ where A and B are two IFS in $X = \{x_1, x_2, \ldots, x_n\}$. Obviously the general form of distance measure would be $D_{\text{IFS}}^{\alpha}(\text{A,B}) = [\sum_{i=1}^{n}\{|\mu_A(x_i) - \mu_B(x_i)|^{\alpha} + |\nu_A(x_i) - \nu_B(x_i)|^{\alpha} + |\pi_A(x_i) - \pi_B(x_i)|^{\alpha}\}]^{\frac{1}{\alpha}}, \alpha$ is a normal number.

Similarity Measures on Intuitionistic Fuzzy Set

$$S(\text{A,B}) = 1 - [1/n\sum_{j=1}^{n}\{(|\mu_A(x_j) - \mu_B(x_j)|)^{\alpha} + (|\nu_A(x_j) - \nu_B(x_j)|)^{\alpha} + (|\prod_A(x_j) - \prod_B(x_j)|)^{\alpha}\}]^{\frac{1}{\alpha}}, \alpha{>}0$$

3.2.2 Intuitionistic Fuzzy representation of polynucleotide on a triplet of I^{36}

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

$$\begin{aligned}\pi_A(x) = [&\{1-(1+\theta)(.x_1)\},\{1-(1+\theta)(.x_2)\},\{1-(1+\theta)(.x_3)\},\{1-(1+\theta)(.x_4)\},\\ &\{1-(1+\theta)(.y_1)\},\{1-(1+\theta)(.y_2)\},\{1-(1+\theta)(.y_3)\},\{1-(1+\theta)(.y_4)\},\\ &\{1-(1+\theta)(.z_1)\},\{1-(1+\theta)(.z_2)\},\{1-(1+\theta)(.z_3)\},\{1-(1+\theta)(.z_4)\}]\end{aligned}$$

3.2.3 Results on Sequence Comparison

Table 3.2, Table 3.3 and Table 3.4 describes Intuitionistic Fuzzy representation, Distance Measure of Intuitionistic Fuzzy Representationand Similarity Measure of Intuitionistic Fuzzy Representation of whole genome (a), (b), (c) & (d) respectively. For simplification of calculation we take $\theta = 0.1$.

Table 3.2 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	0.0266
	0.7437	0.7063	0.7437	0.7063	0.7437	0.7085	0.7437	0.7041	0.7448	0.703	0.7448	0.7074
(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	0.0192
	0.6579	0.7921	0.659	0.791	0.659	0.7899	0.6612	0.7899	0.6623	0.7888	0.6601	0.7888
(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	0.0343
	0.8196	0.6282	0.8218	0.6304	0.8251	0.6249	0.8229	0.6271	0.8262	0.6249	0.8262	0.6227
(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	0.0277
	0.7272	0.7272	0.7492	0.6964	0.7261	0.7272	0.7525	0.6942	0.7294	0.7217	0.7536	0.6953

Table 3.3 Distance Measure of Intuitionistic Fuzzy Representation of Whole Genome (a),(b),(c) & (d)

Distance Between	α=1	α=2	α=3	α=4	α=5	α=6	α=7	α=8	α=9	α=10
(a) & (b)	2.0196	0.155964	0.012543	0.001015	8.24E-05	6.7E-06	5.47E-07	4.47E-08	3.66E-09	3.01E-10
(a) & (c)	1.9096	0.139554	0.01063	0.000815	6.28E-05	4.85E-06	3.76E-07	2.93E-08	2.28E-09	1.79E-10
(a) & (d)	0.3256	0.004555	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) & (c)	3.9292	0.590183	0.092292	0.01452	0.00229	0.000362	5.74E-05	9.11E-06	1.45E-06	2.31E-07
(b) & (d)	1.914	0.144056	0.011584	0.000959	8.11E-05	6.98E-06	6.09E-07	5.38E-08	4.79E-09	4.3E-10
(c) & (d)	2.0152	0.159569	0.013492	0.001174	0.000104	9.44E-06	8.66E-07	8.05E-08	7.55E-09	7.13E-10

Table 3.4 Similarity Measure of Intuitionistic Fuzzy Representation of Whole Genome (a),(b),(c) & (d)

Similarity Between	α=1	α=2	α=3	α=4	α=5	α=6	α=7	α=8	α=9	α=10
(a) & (b)	0.8317	0.9671	0.9806	0.9851	0.9873	0.98856	0.98938	0.989951	0.990373	0.990696
(a) & (c)	0.8409	0.9689	0.9817	0.9859	0.9880	0.98916	0.98993	0.99047	0.990866	0.991168
(a) & (d)	0.9729	0.9944	0.9965	0.9972	0.9976	0.99777	0.99789	0.997969	0.998029	0.998073
(b) & (c)	0.6726	0.9360	0.9623	0.9711	0.9753	0.97775	0.97935	0.980467	0.981289	0.981919
(b) & (d)	0.8405	0.9684	0.9811	0.9853	0.9873	0.98848	0.98921	0.989717	0.990082	0.990358
(c) & (d)	0.8321	0.9667	0.9802	0.9846	0.9867	0.98789	0.98866	0.989185	0.989568	0.989857

3.2.4 Discussion

(i) Distance measures and similarity measures for different values of α show uniform results for whole genomes. Actually this has happened in our case for each value of α. As α increases, distance measure decreases and similarity measures increases. This suggests that better is the result, larger is the value of α .

(ii) The same four genomes as in (a), (b), (c) and (d) mentioned first part of the chapter are chosen in this paper. This is only to show that the anomalies do not occur any more if Intuitionistic Fuzzy representation is used in place of Fuzzy representation of the genome sequences.

(iii) Even we have been able to show that similar behavior of the metrics under different values of α show similar trend, which was not possible for fuzzy representation of the same sequences, still it cannot be concluded that this is a general trend. This is because; the results are verified for only four such sequences.

(iv) The result has to be verified on a larger number of genomes in order to claim that the conclusion is general. But as some value of the parameter θ (hesitancy factor) is always involved in the calculations, so if some contradictory result appears at all, it is only apparent. It can be adjusted by choice of suitable θ.

3.2.5 Conclusion

Thus it can be definitely concluded that the Intuitionistic Fuzzy Set is one of the best tools in analyzing similarity/dissimilarities of complete genomes. It reduces comparing whole genome sequences of any length, equal or unequal to comparison of 36 component intuitionistic fuzzy vectors and further the intuitionistic fuzzy standard distance measure can be conveniently applied to have final comparison of such genome sequences under their Intuitionistic fuzzy representations.

CHAPTER 4

4. Composition Vector Method under Optimal Choice of *k*-mer

The composition vector method is a most significant method that can use the nucleotides strings of length *k* or *k*-mer arrangement of nucleotides. Such composition vector methods are differing only in the distance measures selection, which are angle-based or information based in nature [63]. It is a probabilistic method. In our method, we apply composition vector based on 3-mer representation and similarity based index using Shannon Entropy, as the distance measure. So our method considers tri-nucleotide representations of nucleotides. The advantage of 3-mer representation is that strings of 3-mer are only of length 64, and each component consists of amino acids, which are the building blocks of proteins. The final aim is to obtain phylogenetic trees for sequence comparison. We also compare phylogenetic trees of the present method with those obtained by other methods, which are of the following types:

Neighbor-Joining Method

Neighbor-joining method is a bottom-up clustering method. The algorithm requires knowledge of the distance between each pair of taxa to form the tree.

Jukes and Cantor Model

In the Jukes and Cantor model, the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. The multiple hit correction equation for this model produces a maximum likelihood estimate of the number of nucleotide substitutions between two sequences. It assumes an equality of substitution rates among sites, equal nucleotide frequencies, and it does not correct for higher rate of transitional substitutions as compared to transversion substitutions.

Probabilistic Method

In Probabilistic method, probability distribution of DNA sequences is constructed by using a graphical representation. These probability distributions are then used to make similarity studies by using the symmetries Kullback–Leibler divergence.

Maximum Likelihood Method

Maximum likelihood is a general statistical method for estimating unknown parameters of a probability model.

K-word Frequency Method

The model of *k*-word frequencies is a well-developed one. However, most existing word-based methods neglect relationships among *k*-word frequencies, while a few others focus on the correlation of *k*-words but ignore the word frequency itself.

4.1. Our Methodology

One of the alignment-free techniques is the string composition vector (CV), which employ the frequencies of strings of nucleotide to signify the sequence information. In genome sequence comparison, the CV method provides promising results. Our methodology is at par with the standard CV method with certain modifications and changes.

So we first state the general methodology of CV method:

4.1.1 Step 1: Frequency or Rank Vector

In a molecular sequence (nucleotide or amino acid sequence) of length *N*, any consecutive *k* molecules is called a *k*-string or a *k*-tuple or a *k*-mer, where$1 \leq k \leq N$. Computationally, the *k*-mers are collected by using a sliding window of length *k*. It slides through the sequence by shifting one position at a time. Thereby, *N*-*k*+1 number of such overlapping strings are obtained. The *k*-strings are denoted by the variable *u*. *g(u)* represents the frequency vector or the rank vector of the *k*-strings. This calculates how many times each *k*-string appears in the sequence

4.1.2 Step 2: Probability Vector

The probability vector is given by:

$$f(u) = \frac{g(u)}{N-k+1} \qquad (4.1)$$

This formula is applicable when the whole genome sequence is considered as a single entity. However, if only protein coding DNA sequences from the whole genome sequence are considered, then this formula is modified as follows:

$$f(u) = \frac{\sum_{j=1}^{m} g_j(u)}{\sum_{j=1}^{m}(N_j - k + 1)} \qquad (4.2)$$

where *m* is the number of protein-coding DNA sequences from the whole genome, $g_j(u)$ is the number of times that *u* appears in the j^{th} DNA sequence and N_j is the length of j^{th} DNA sequence. The modified formula in (4.2) avoids the problems, which might occur from the gene order and gene content in a genome sequence.

4.1.3 Step 3: Composition Vector

Generally, the biological data are often obscured by noise and bias [64]. Thus, a signal de-noising process is performed before obtaining the composition vector *h(u)* from the probability vector *f(u)*. Specifically, for each *f(u)*, the estimated noise *q(u)* is calculated as in [52]. Afterward, the composition vector *h(u)* is determined by calculating the signal-to-noise-ratio given by:

$$h(u) = \frac{f(u) - q(u)}{q(u)} \qquad (4.3)$$

4.1.4 Step 4: Distance Measure

Information Based Similarity Index

The information based similarity index (D_k) using *k*-tuple nucleotides between two sequences S_1 and S_2 is given by:

$$D_k(S_1, S_2) = \frac{1}{4^k - 1} \sum_{i=1}^{4^k} |R_1(w_i) - R_2(w_i)| \frac{H_1(w_i) + H_2(w_i)}{\sum_{i=1}^{4^k}[H_1(w_i) + H_2(w_i)]} \qquad (4.4)$$

where $R_1(W_i)$, $R_2\ (W_i)$ and $H_1(W_i)$, $H_2\ (W_i)$ represent the rank and Shannon entropy of a specific *k*-tuple w_i in the sequences S_1 and S_2 respectively.

Actually the absolute difference $|R_1(W_i) - R_2(W_i)|$ of ranks is proportional to the Euclidean distance from a given point to the diagonal line. This is multiplied by the weighted sum of the Shannon entropies of the sequences to give the final expression for the similarity index.

Typically, Shannon entropy measures the information richness of each *k*-tuple in both the sequences, so the increase in the frequency of *k*-tuple occurrences leads to an increase in the similarity among the genetic sequences. Further, in this case, the noise reduction is avoided due to the normalization by the total Shannon entropies. Consequently, the information based similarity index is a good choice for comparing the genome sequences. But a limitation still exists, as this information based similarity index does not readily satisfy the property of triangular inequality of a metric [35]. Thus, in the present work, in order to declare that equation (4.4) is a proper distance measure, the triangular inequality has to be verified. Nonetheless, what we find is that there is practically no violation of the triangular inequality when applied to actual genome sequences. So, we accept equation (4.4) as a proper distance measure.

Accordingly, the methodology of the proposed work specifically consists of the following steps:

1. Calculate the probability vector given by equation (4.1) having 64 components (3-mer),
2. Apply the information based similarity index given by equation (4.4) as the distance measure on the vectors of equation (4.1) to calculate the distance matrix,
3. Draw the phylogenetic tree by applying UPGMA on the above distance matrix.

4.2 Results and Discussions

For the sake of convenience, the results include the following parts: i) comparative study of the proposed work with inaccurate results of other methods, ii) results of other methods that agree with that obtained by the proposed approach, iii) method based on word and rough set theory, iv) the CV method using *k*-mer and distance measure other than similarity index and v) the CV method using *k*-mer and similarity index as the distance measure.

4.2.1 Comparative Study of the Proposed Approach Against Other Inaccurate Methods

a. A dataset of mitochondrial genomes of 31 Mammalians is given in Table 4.1.

Table 4.1 Description of mitochondrial genome of 31 mammalians

Number	Genome name on the tree	GenBank ID
1	Human	V00662
2	Pigmy chimpanzee	D38116
3	Common chimpanzee	D38113
4	Gibbon	X99256
5	Baboon	Y18001
6	Vervet Monkey	AY863426
7	MacacaThibetana	NC002764
8	Bornean Orangutan	D38115
9	Sumatran Orangutan	NC002083
10	Gorilla	D38114
11	Cat	U20753
12	Dog	U96639
13	Pig	AJ002189
14	Sheep	AF010406
15	Goat	AF533441
16	Cow	V00654
17	Buffalo	AY488491
18	Wolf	EU442884
19	Tiger	EF551003
20	Leopard	EF551002
21	Indian Rhinoceros	X97336
22	White Rhinoceros	Y07726
23	Black Bear	DQ402478
24	Brown Bear	AF303110
25	Polar Bear	AF303111
26	Giant Panda	EF212882
27	Rabbit	AJ001588
28	Hedgehog	X88898
29	Dormouse	AJ001562
30	Squirrel	AJ238588
31	Blue Whale	X72204

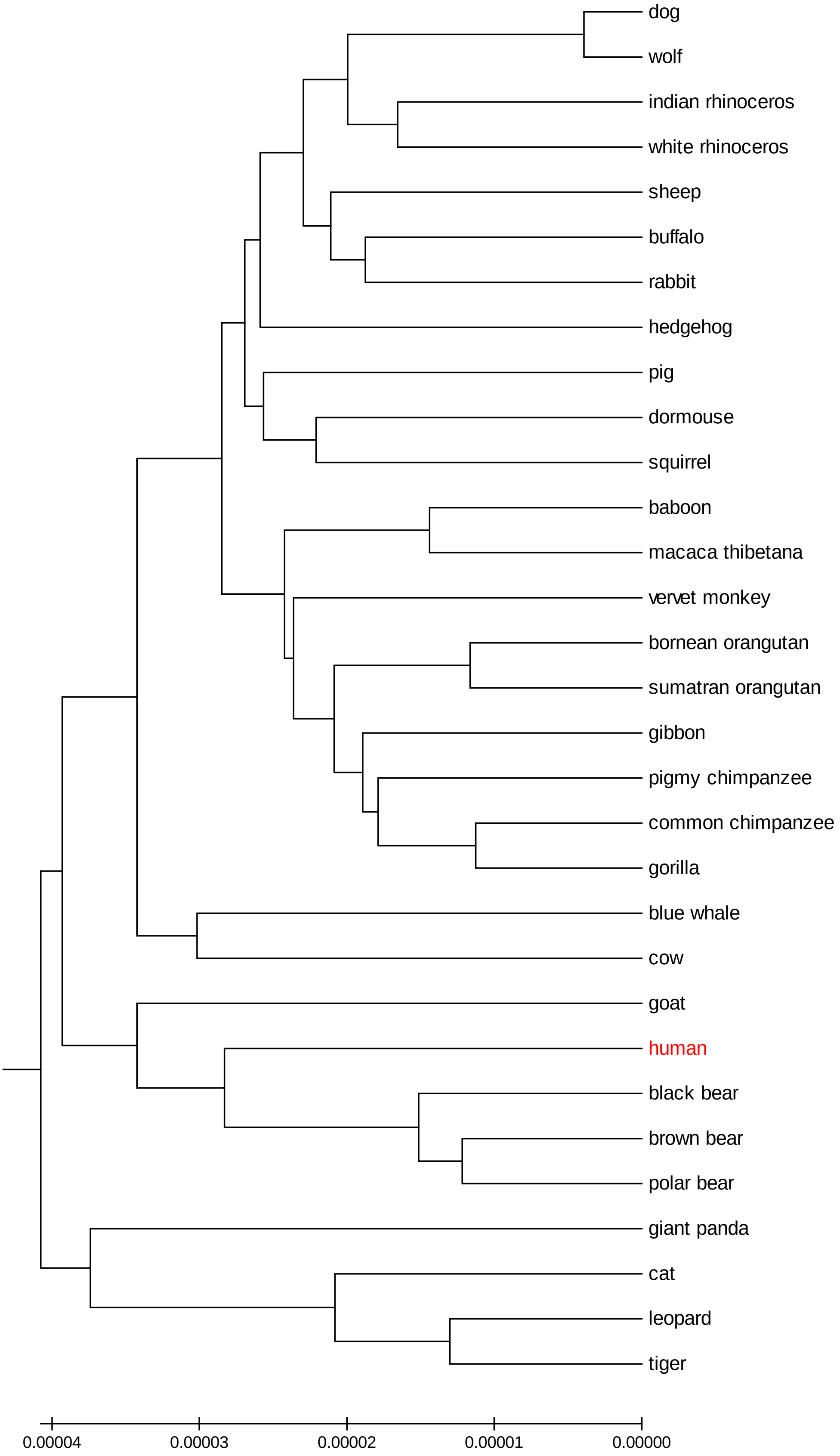

Fig. 4.1(a) Phylogenetic tree using the Probabilistic method

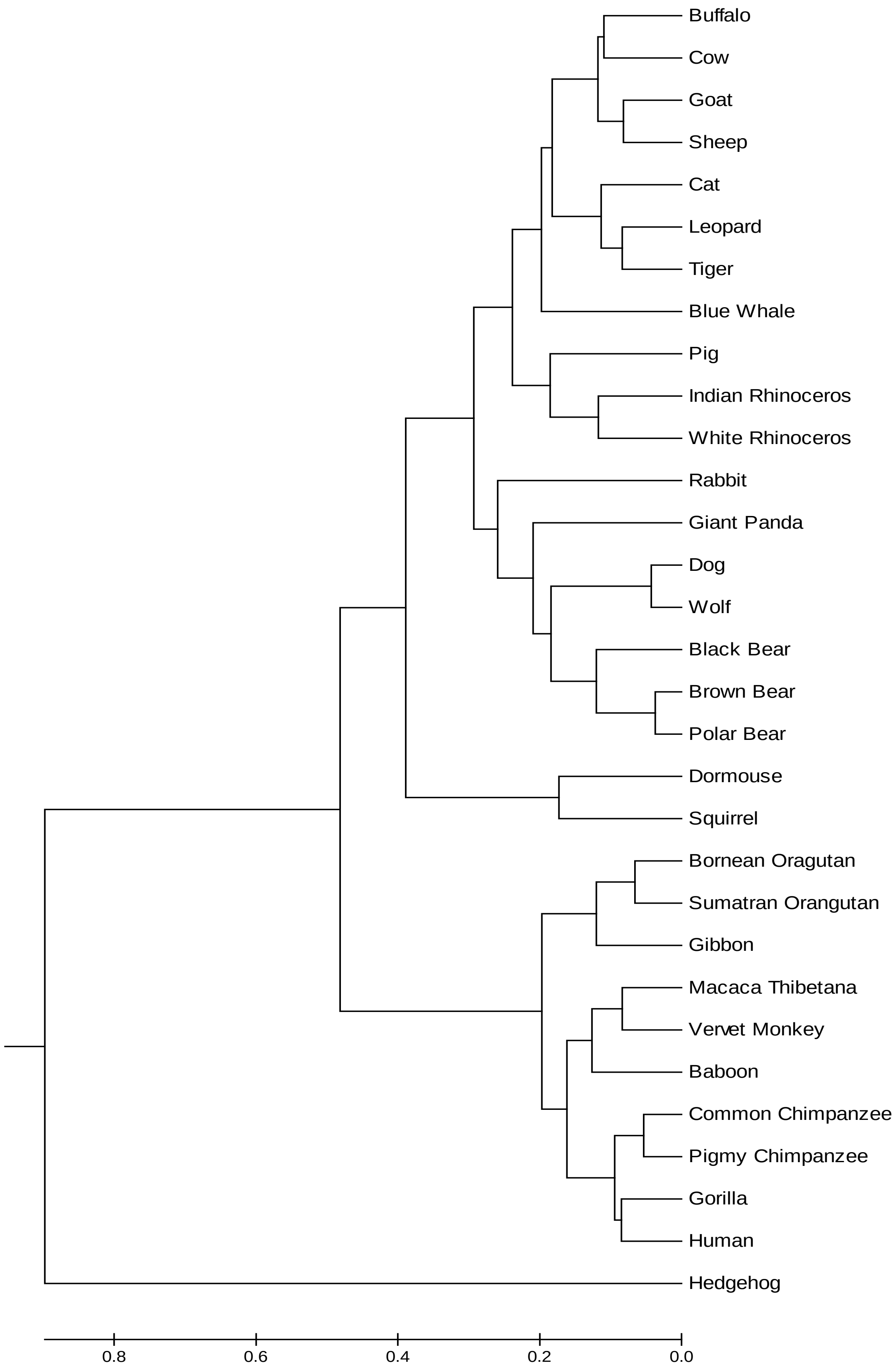

Fig. 4.1(b) Phylogenetic tree obtained using the proposed method

The Phylogenetic tree of this dataset using the probabilistic method [35] is given in Fig. 4.1(a). This result is compared to the proposed method results in Fig. 4.1(b). A comparative study of the results obtained in Fig. 4.1(a) and Fig. 4.1(b) depicts that Human is classified incorrectly in the group of Goat and Bear when using the Probabilistic method as illustrated in Fig. 4.1(a). However, the results of the proposed approach in Fig. 4.1(b) show correct classification. This proves the superiority of the proposed approach over the probabilistic method.

b. A dataset set of 53 complete genome sequences of TYLCV (Tomato Yellow Leaf Curl Virus) is given in Table 4.2.

Table 4.2 TYLCD-causing virus sequences used in this study

Isolate	**Accession No.**	**Length**
TYLCV_IL	X15656	2787
TYLCV_DO	AF024715	2781
TYLCV_CU	AJ223505	2781
TYLCV_Flo	AY530931	2781
TYLCV_Omu	AB116630	2774
TYLCV_Alm	AJ489258	2781
TYLCV_Mis	AB116631	2774
TYLCV_EG_Ism	AY594174	2781
TYLCV_Miy	AB116629	2774
TYLCV_PR	AY134494	2781
TYLCV_MA	EF060196	2781
TYLCV_TR_Mer1_04	AJ812277	2781
TYLCV_Tosa_H	AB192966	2781
TYLCV_Tosa	AB192965	2781
TYLCV_RE4	AM409201	2781
TYLCV_Sic	DQ144621	2781
TYLCV_TN	EF101929	2781
TYLCV_JO	EF054893	2781
TYLCV_MX_Cul	DQ631892	2781
TYLCV_Mld_PT	AF105975	2793

TYLCV_Mld_Aic	AB014347	2787
TYLCV_Mld_Shi	AB014346	2791
TYLCV_Mld_ES7297	AF071228	2791
TYLCV_Mld_ES	AJ519441	2790
TYLCV_Mld_Sz_Yai	AB116632	2791
TYLCV_Mld_Atu	AB116633	2787
TYLCV_Mld_Kis	AB116634	2787
TYLCV_Mld_Sz_Dai	AB116635	2787
TYLCV_Mld_Sz_Osu	AB116636	2787
TYLCV_Mld_RE	AJ865337	2791
TYLCV_Mld_JO	EF054894	2791
TYLCAxV_Alg	AY227892	2772
TYLCMalV	AF271234	2782
TYLCMLV	AY502934	2794
TYLCMLV_ET	DQ358913	2785
TYLCSV	X61153	2773
TYLCSV_Sic	Z28390	2773
TYLCSV_ES1	Z25751	2777
TYLCSV_ES2	L27708	2777
TYLCSV_MA	AY702650	2777
TYLCSV_TN	AY736854	2772
TYLCCNV	AF311734	2734
TYLCCNV_Tb_Y25	AJ457985	2738
TYLCCNV_YM	DQ256460	2731
TYLCKaV_TH_Kan1	AF511529	2752
TYLCKaV_TH_Kan2	AF511530	2752
TYLCKaV_VN	DQ169054	2751
TYLCTHV	X63015	2743
TYLCTHV_MM	AF206674	2746
TYLCTHV_Y72	AJ495812	2748
TYLCTHV_ChMai	AY514630	2747
TYLCTHV_NoK	AY514631	2744
TYLCTHV_SaNa	AY514632	2747

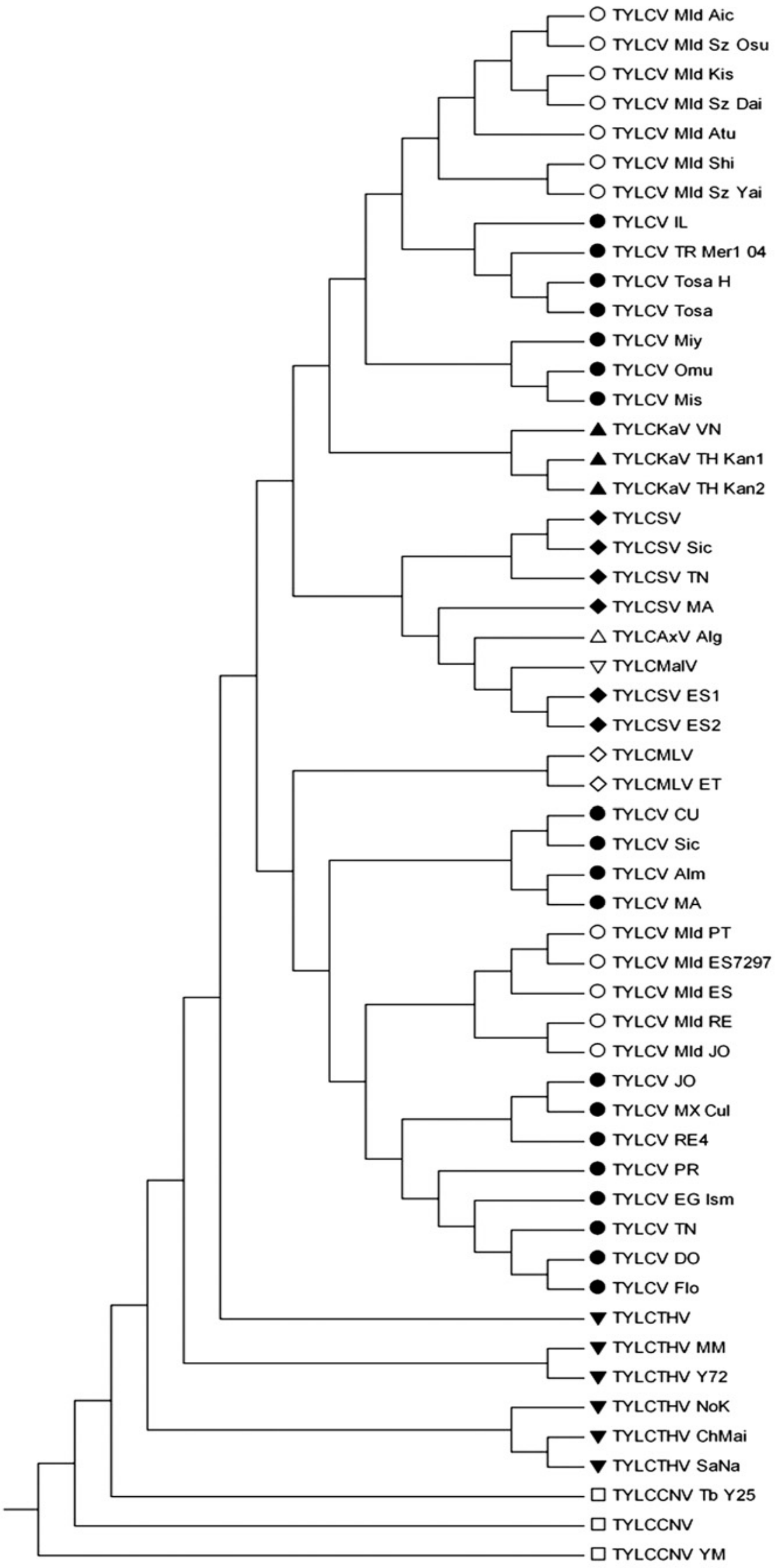

Fig. 4.2(a) Phylogenetic tree using Probabilistic method

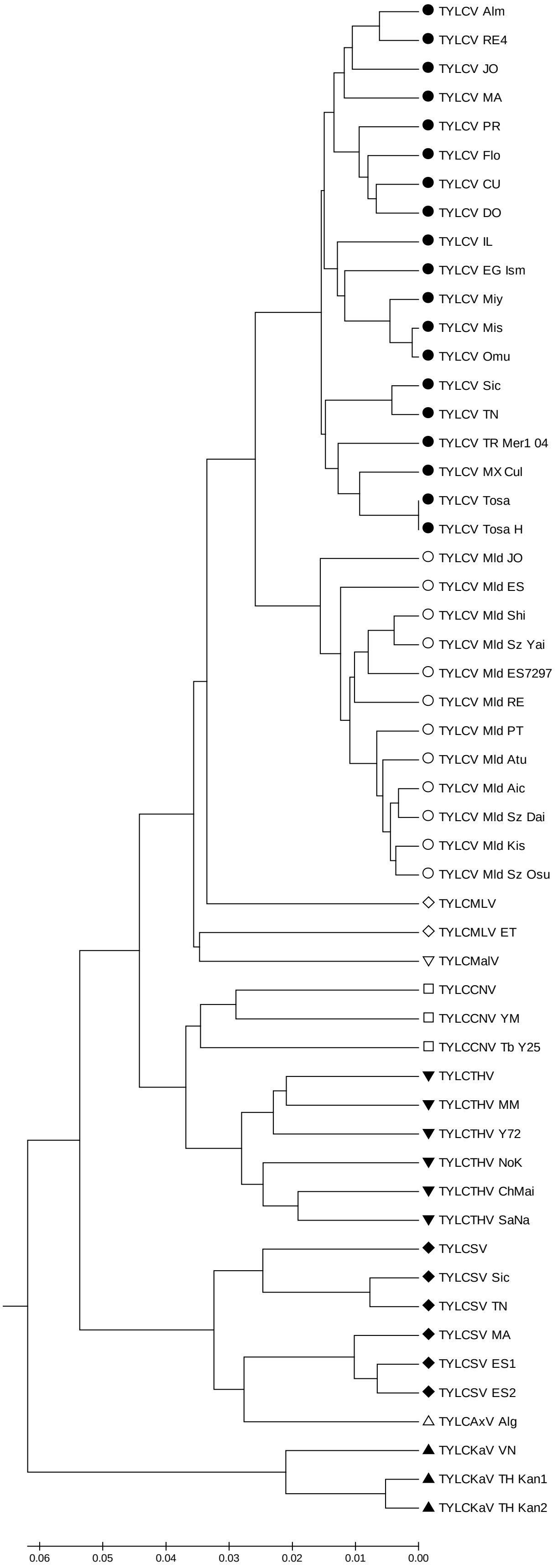

Fig. 4.2(b) Phylogenetic tree using the proposed method

The Phylogenetic tree using the probabilistic method [35] is given in Fig 4.2(a). This result is compared to that of the proposed method in Fig 4.2(b). A comparative study of the results obtained in Fig. 4.2(a) and Fig. 4.2(b) depicts that some of the TYLCV of *severe Phenotypes* are classified in the cluster of *Mild Phenotype* and similarly some of the *Mild Phenotypes* are considered in the cluster of TYLCV *Severe Phenotype* when using the Probabilistic method as illustrated in Fig. 4.2(a). However, the results of the proposed approach in Fig. 4.2(b) show correct classification. This proves the superiority of the proposed approach over the probabilistic method [35] with this second dataset.

c. A dataset of 21 genomes (6 influenza A (H1N1) genomes, 6 swine flu virus genomes, 6 avian virus genomes and 3 human seasonal flu virus genomes) is given in Table 4.3.

Table 4.3 The list of 21 genomes used for comparisons

A(H1N1)-1	Influenza A virus(A/New York/18/2009(H1N1))
A(H1N1)-2	Influenza A virus(A/Canada-ON/RV1527/2009(H1N1))
A(H1N1)-3	Influenza A virus(A/Mexico/InDRE4487/2009(H1N1))
A(H1N1)-4	Influenza A virus(A/Texas/09/2009(H1N1))
A(H1N1)-5	Influenza A virus(A/California/14/2009(H1N1))
A(H1N1)-6	Influenza A Virus(A/New York/1669/2009(H1N1))
swine1	Influenza A virus (A/swine/Alberta/56626/03(H1N1))
swine2	Influenza A virus (A/swine/California/T9001707/1991(H1N1))
swine3	Influenza A virus (A/swine/Nebraska/123/1977(H1N1))
swine4	Influenza A virus (A/swine/Memphis/1/1990(H1N1))
swine5	Influenza A virus (A/swine/Iowa/31483/1988(H1N1))
swine6	Influenza A virus (A/swine/Ontario/55383/04(H1N2))
seasonal1	Influenza A virus (A/Puerto Rico/8/34(H1N1))
seasonal2	Influenza A virus (A/New Caledonia/20/1999(H1N1))
seasonal3	Influenza A virus (A/Wisconsin/67/2005(H3N2))
avian1	Influenza A virus (A/duck/NJ/7717-70/1995(H1N1))
avian2	Influenza A virus (A/blue winged teal/TX/27/2002(H1N1))
avian3	Influenza A virus (A/mallard/Maryland/42/2003(H1N1))
avian4	Influenza A virus (A/mallard/MN/330/1999(H3N1))
avian5	Influenza A virus (A/blue-winged teal/Ohio/1864/2006(H3N8))
avian6	Influenza A virus (A/Duck/NY/185502/2002(H5N2))

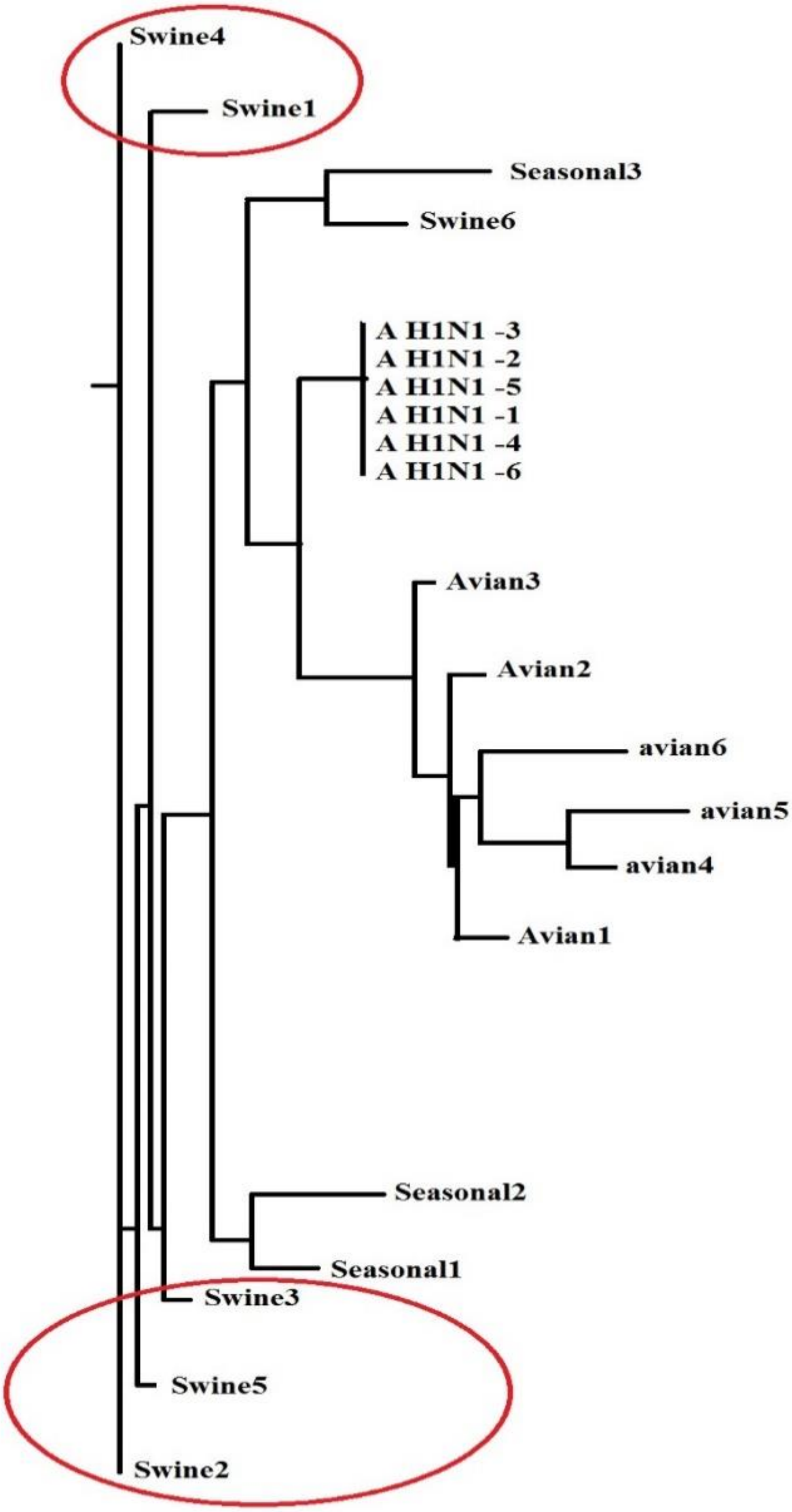

Fig. 4.3(a) Phylogenetic trees by maximum likelihood method

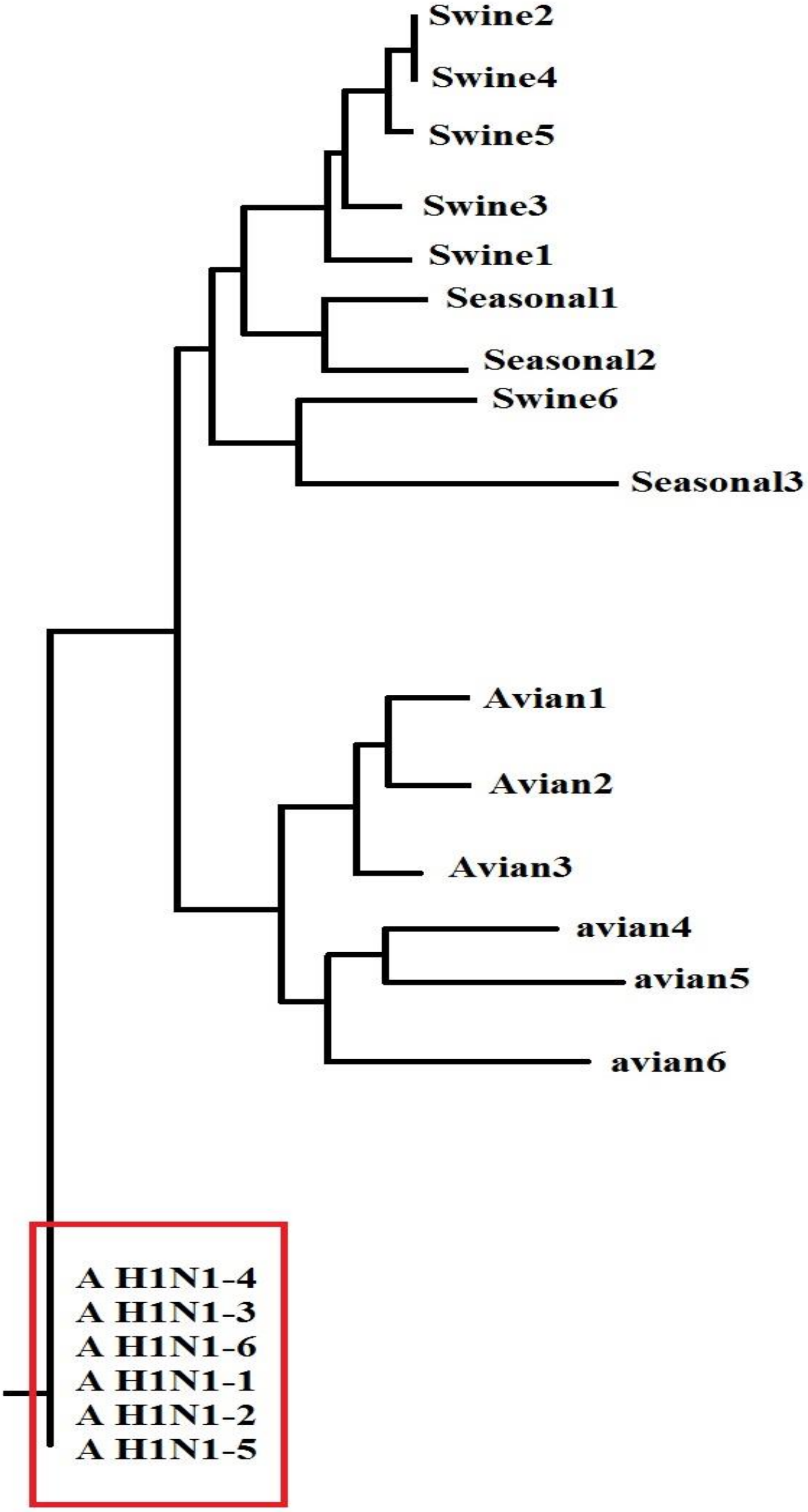

Fig. 4.3(b) Phylogenetic trees by neighbour joining method with Kimura-2

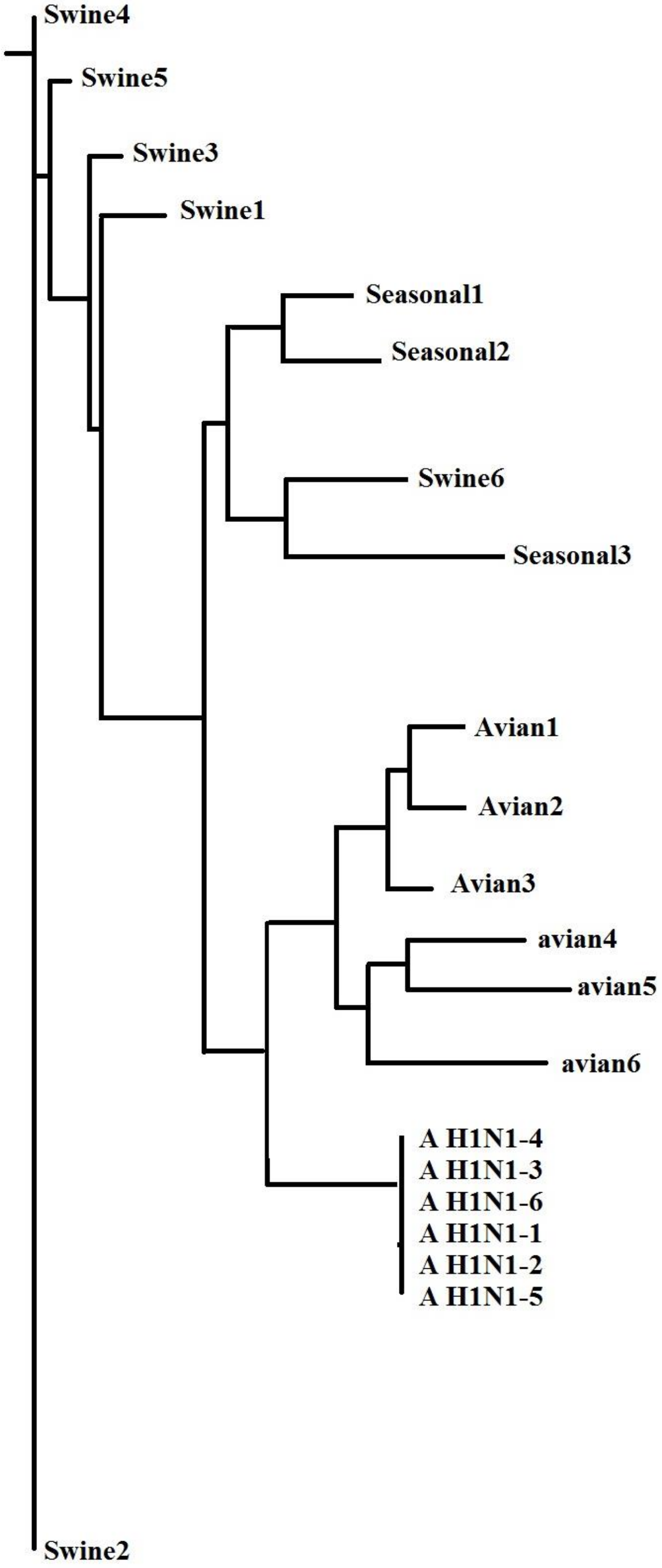

Fig. 4.3(c) Phylogenetic trees by neighbour joining method with Jukes-Cantor Model

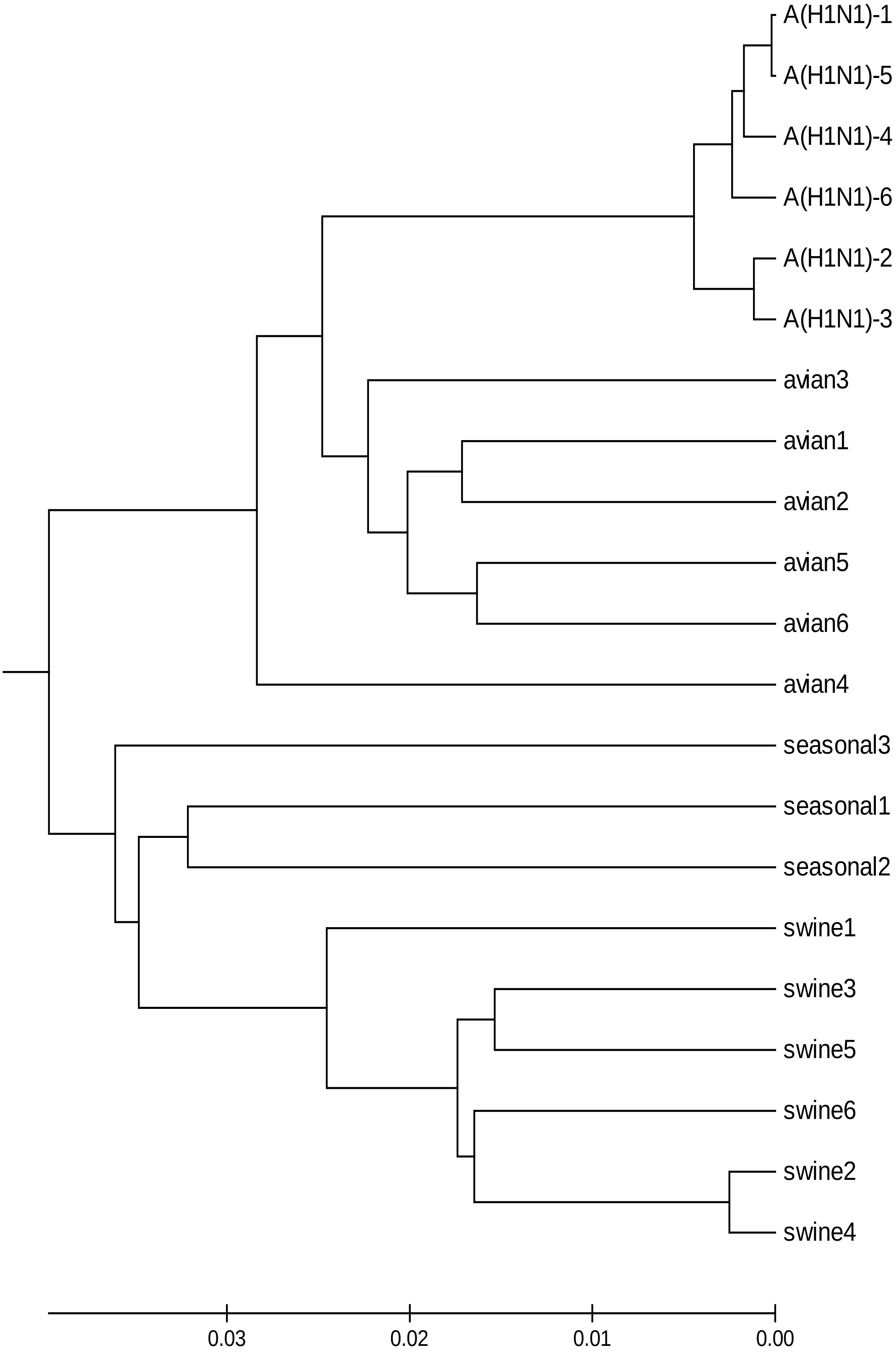

Fig. 4.3(d) Phylogenetic tree obtained by the proposed method for 21 viruses

Results of comparative study in applying the maximum likelihood method, the Neighbour Joining method with Kimura-2 parameter model, the Neighbour Joining method with Jukes-Cantor Model with the proposed approach are demonstrated in Fig. 4.3(a), Fig. 4.3(b), Fig. 4.3(c) and Fig. 4.3(d); respectively. The phylogenetic result obtained by using maximum likelihood method (Fig. 4.3(a)) shows that the swine flu virus genomes are not clustered correctly. The result obtained by using the neighbour-joining method (NJ) with Kimura model (Fig. 4.3(b)) do not obviously show the origin of A (H1N1) genomes, because all A(H1N1) genomes are very far away from other genomes. Besides, the result obtained from NJ method with Jukes-Cantor model (Fig. 4.3(c)) also fails to cluster the swine flu viruses correctly. The NJ method with the Kimura and Jukes-Cantor models yields totally different phylogenetic trees. Thus, these three methods provide unsatisfactory results. On the contrary, the proposed approach in (Fig. 4.3(d)) shows no discrepancy appears in clustering different types of viruses. In fact, 6 Influenza A(N1H1) viruses are now occupying the top position of the phylogenetic tree and they are clustered together. Similarly, the 6 swine flu virus genomes, 6 avian virus genomes and 3 human seasonal flu virus genomes are grouped together in the respective clusters. Thus, the proposed approach provides superior results compared to the maximum likelihood method and also all types of neighbour joining methods.

4.2.2 Comparative Study of the Proposed Approach Versus Other Agreeing Methods

The comparative study of the proposed approach versus the method of word and rough set theory with different datasets is as follows.

a. Data set of 19 HV strains is given in Table 4.4.

Table 4.4 The S segments of 19 HV strains

No.	Strain	Type	AC(GenBank)	Region
1	Z10	HTN	AF184987	Shengzhou
2	Z5	HTN	EF103195	Shengzhou
3	Z251	HTN	EF595840	Longquan
4	ZLS6-11	HTN	FJ753396	Lishui
5	ZLS-12	HTN	FJ753398	Lishui
6	76–188	HTN	M14626	Korea
7	Gou3	SEO	AF184988	Jiande

No.	Strain	Type	AC(GenBank)	Region
8	ZJ5	SEO	FJ753400	Jiande
9	K24-v2	SEO	AF288655	Xinchang
10	K24-e7	SEO	AF288653	Xinchang
11	Z37	SEO	AF187082	Wenzhou
12	ZT71	SEO	AY750171	Tiantai
13	ZT10	SEO	AY766368	Tiantai
14	ZY27	SEO	AF406965	Heilongjiang
15	PF26	SEO	AY006465	Heilongjiang
16	SR-11	SEO	M34881	Japan
17	80–39	SEO	AY273791	Korea
18	R22	SEO	AF288295	Henan
19	L99	SEO	AF288299	Jiangxi

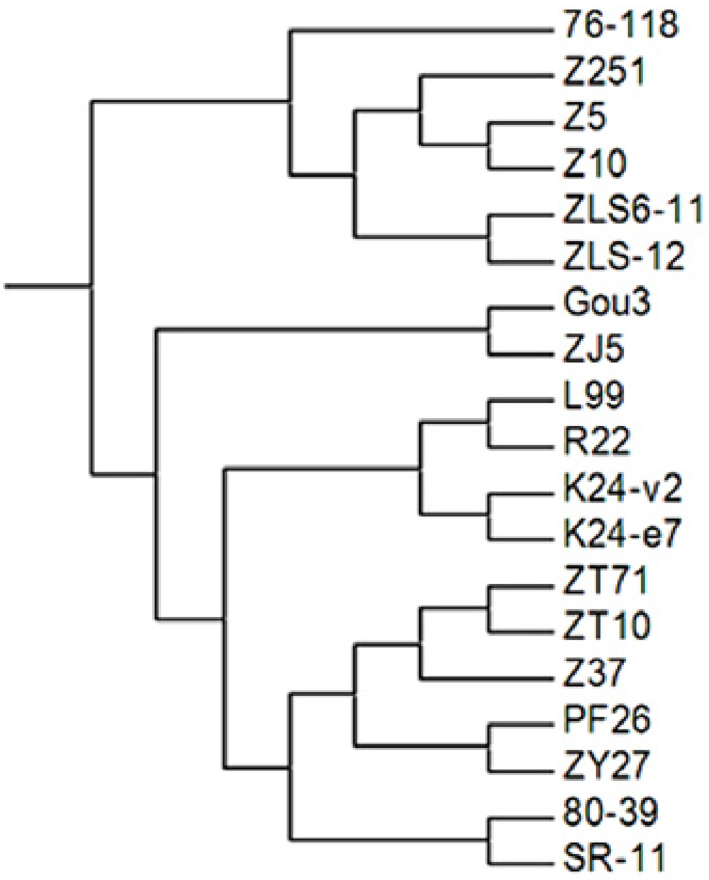

Fig. 4.4(a) Phylogenetic tree of 19 Hantaviruses by word and rough set theory

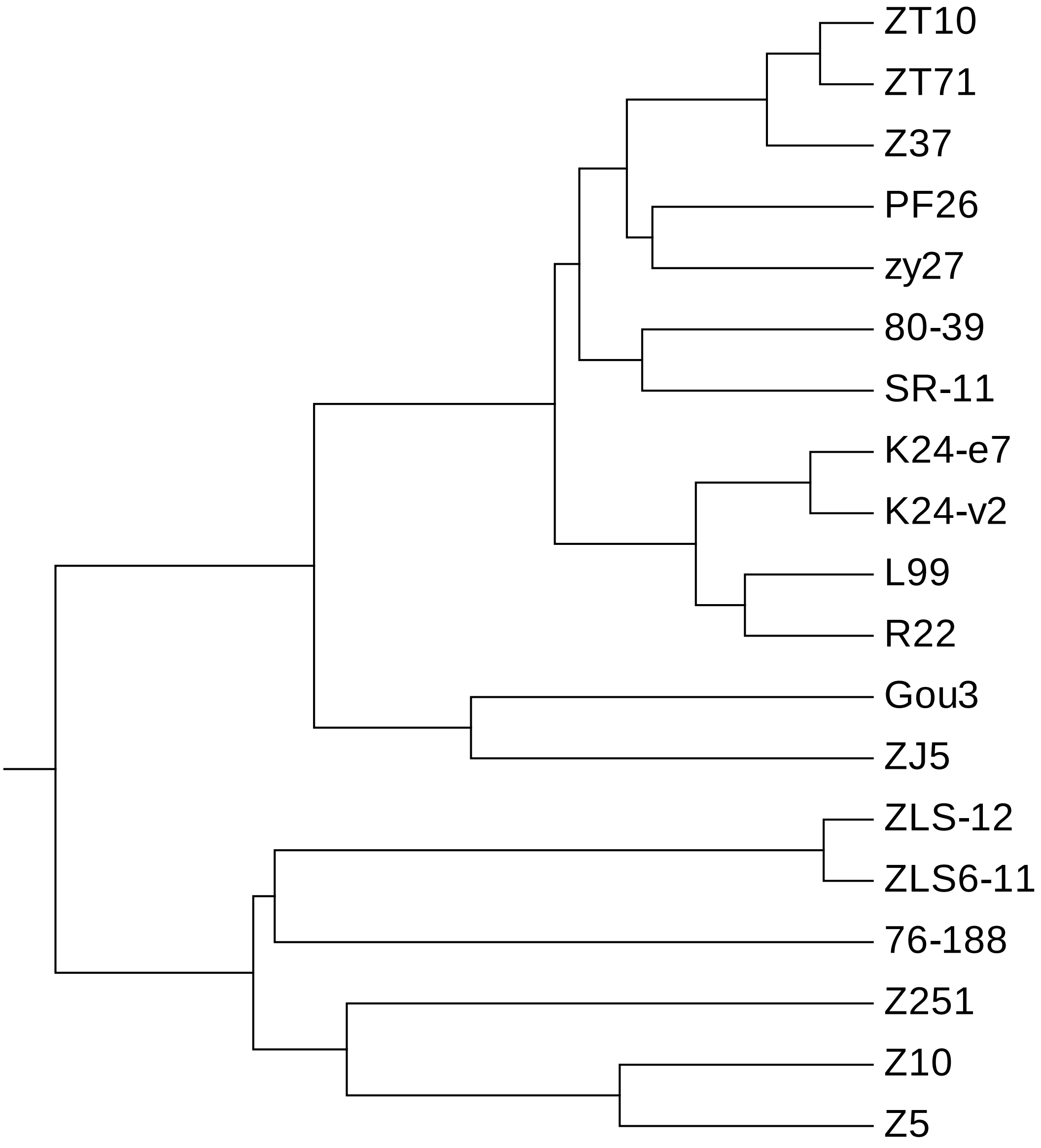

Fig. 4.4(b) Phylogenetic tree of 19 Hantaviruses obtained by the proposed method

The phylogenetic tree using the Method of word and rough set theory [65] is given in Fig 4.4(a). The result is obtained by the proposed method is given in Fig 4.4(b). This comparative result establishes that Fig. 4.4(b) resembles Fig. 4.4(a), where the clustering remains the same, only their positions are interchanged. Thus, both the proposed approach and the Method of word and rough set theory achieve correct clustering except for the positions of the clusters.

b. The dataset of 48 hepatitis E Viruses is given in Table 4.5.

Table 4.5 The 48 hepatitis E viruses

No.	Strain name	AC	Genotype/ Subtype	Country
1	B1(Bur-82)	M73218	Ia	Burma(Rangoon)
2	B2(Bur-86)	D10330	Ia	Burma(Rangoon)
3	I2(Mad-93)	X99441	Ia	India(Madras)
4	I3	AF076239	Ia	India(Hyderabad)
5	NP1(TK15/92)	AF051830	Ia	Nepal(Kathmandu)
6	P2(Abb-2B)	AF185822	Ia	Pakistan(Abbotabad)
7	Yam-67	AF459438	Ia	India(YamunaNagar)
8	C1(CHT-88)	D11092	Ic	China(Xinjiang,Hetian)
9	C2(KS2-87)	L25595	Ic	China(Xinjiang,Kashi)
10	C3(CHT-87)	L08816	Ic	China(Xinjiang,Hetian)
11	C4(Uigh179)	D11093	Ic	China(Xinjiang,Uighur)
12	ChinaHebei	M94177	Ic	China(Hebei)
13	P1(Sar-55)	M80581	Ic	Pakistan(Rangoon)
14	I1(FHF)	X98292	Ib	India
15	Morocco	AY230202	Id	Morocco
16	T3	AY204877	Ie	Chad
17	M1	M74506	II	Mexico(Telixtac)
18	HE-JA10	AB089824	IIIa	Japan(Tokyo)
19	JKN-Sap	AB074918	IIIa	Japan(Sapporo)
20	JMY-HAW	AB074920	IIIa	Japan(Sapporo)
21	SW-US1	AF082843	IIIa	USA
22	US1	AF060668	IIIa	USA(Minnesota)
23	US2	AF060669	IIIa	USA(Tennessee)
24	ARKELL	AY115488	IIIa	Canada(Ontario,Guelph)
25	JBOAR1-HYO04	AB189070	IIIb	Japan(Hyogo)
26	JDEER-HYO03L	AB189071	IIIb	Japan(Hyogo)
27	JJT-KAN	AB091394	IIIb	Japan(Kanagawa)
28	JMO-HYO03L	AB189072	IIIb	Japan(Hyogo)
29	JRA1	AP003430	IIIb	Japan(Tokyo)

No.	Strain name	AC	Genotype/ Subtype	Country
30	JSO-HYO03L	AB189073	IIIb	Japan(Tokyo)
31	JTH-HYO03L	AB189074	IIIb	Japan(Tokyo)
32	JYO-HYO03L	AB189075	IIIb	Japan(Tokyo)
33	SWJ570	AB073912	IIIb	Japan(Tochigi)
34	KYRGYZ	AF455784	IIIc	Kyrgyzstan
35	HE-JA1	AB097812	IVa	Japan(Hokkaido)
36	HE-JK4	AB099347	IVa	Japan(Tochigi)
37	HE-JI4	AB080575	IVa	Japan(Tochigi)
38	JAK-Sai	AB074915	IVa	Japan(Saitama)
39	JKK-SAP	AB074917	IVa	Japan(Sapporo)
40	JSM-SAP95	AB161717	IVa	Japan(Hokkaido)
41	JSN-SAP-FH	AB091395	IVa	Japan(Hokkaido)
42	JSN-SAP-FH02C	AB200239	IVa	Japan(Hokkaido)
43	JTS-SAP02	AB161718	IVa	Japan(Hokkaido)
44	JYW-SAP02	AB161719	IVa	Japan(Hokkaido)
45	SWJ13-1	AB097811	IVa	China(Uighur)
46	SWCH25	AY594199	IVc	China(Beijing)
47	T1	AJ272108	IVc	USA
48	CCC220	AB108537	IVb	China(Changchun)

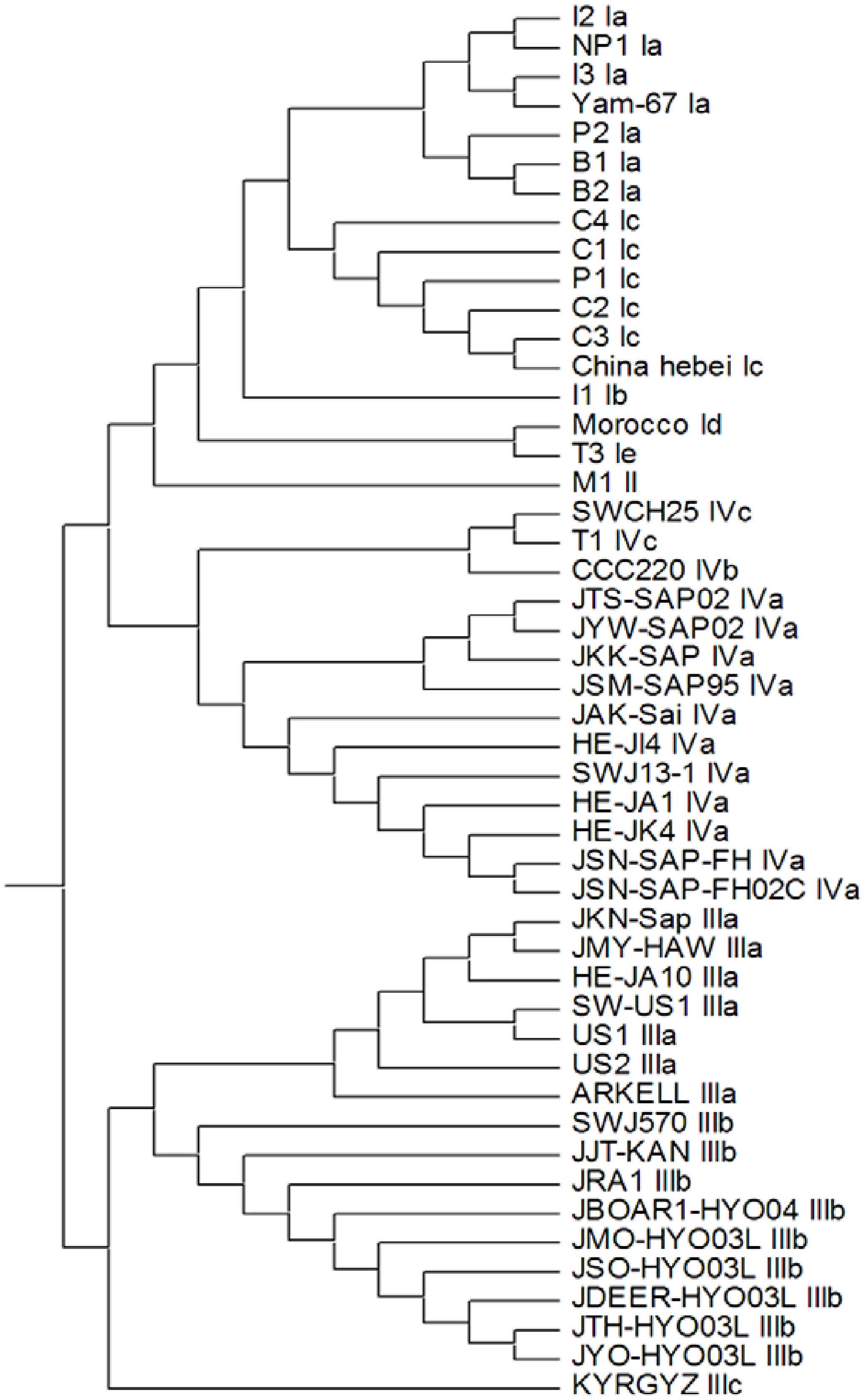

Fig. 4.5(a) Phylogenetic tree of HEV genomes by word and rough set theory

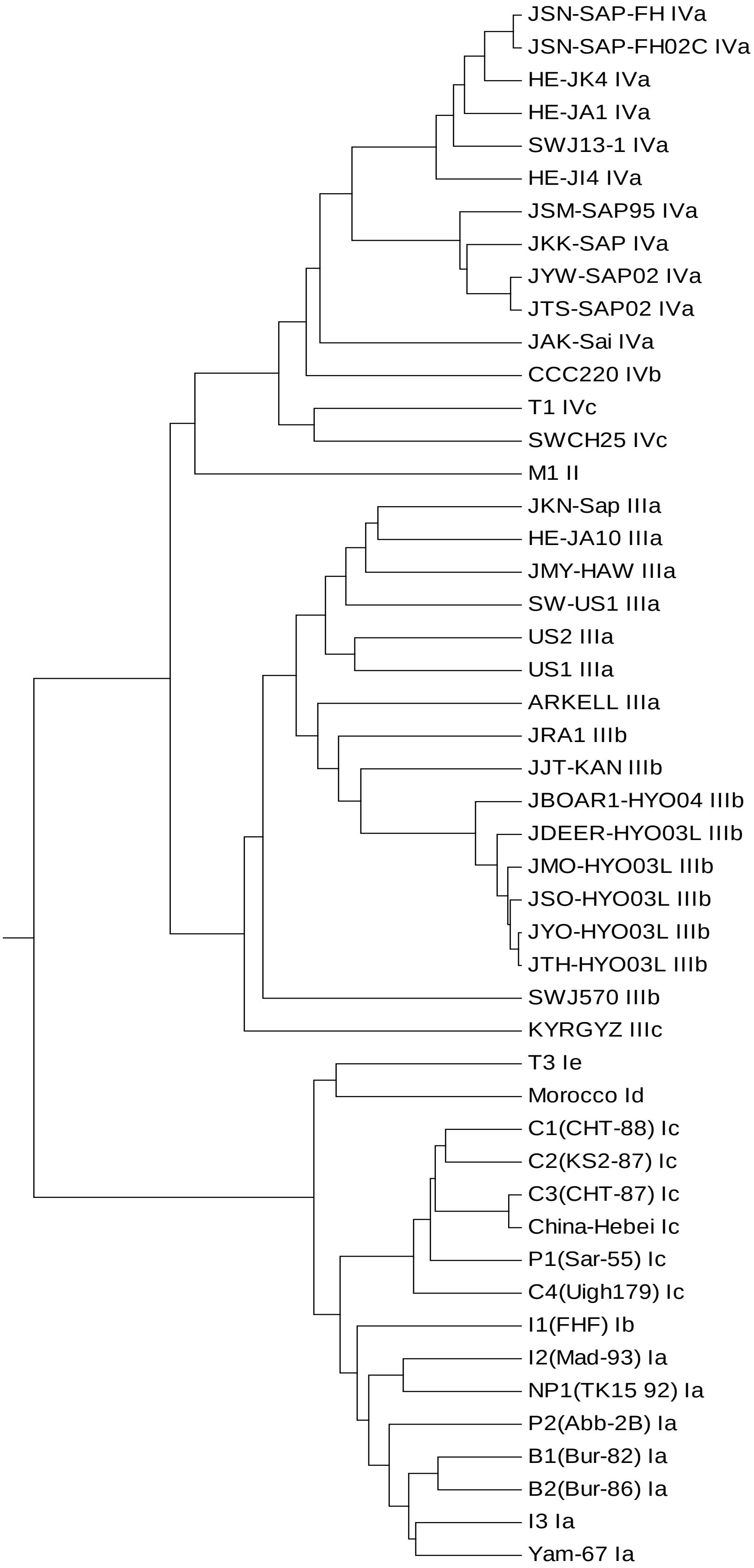

Fig. 4.5(b) Phylogenetic tree of HEV genomes by the proposed method

The Phylogenetic tree based on Method of word and rough set theory and the proposed method are illustrated in Fig. 4.5(a) and Fig. 4.5(b); respectively. Fig. 4.5(a) and Fig. 4.5(b) depict that there is no fundamental difference between the results of proposed approach and those of the Method of word and rough set theory [65].

4.2.3 Composition vector method using *k*-mer and distance measure other than the similarity index [63]

In this comparative study, a dataset of 31 Mammalian mitochondrial genomes is given in Table 4.1. The phylogenetic tree of 31 mammalian mitochondrial genomes by the *k*-mer method, where *k*=5 is given in Fig 4.6. The phylogenetic tree obtained by the proposed method for the same dataset has already been given in Fig. 4.1(b). From the comparative study of Fig. 4.6 and Fig. 4.1(b), it is reported that all the 31 mammalian species are clustered properly. There is no fundamental difference between these two phylogenetic trees.

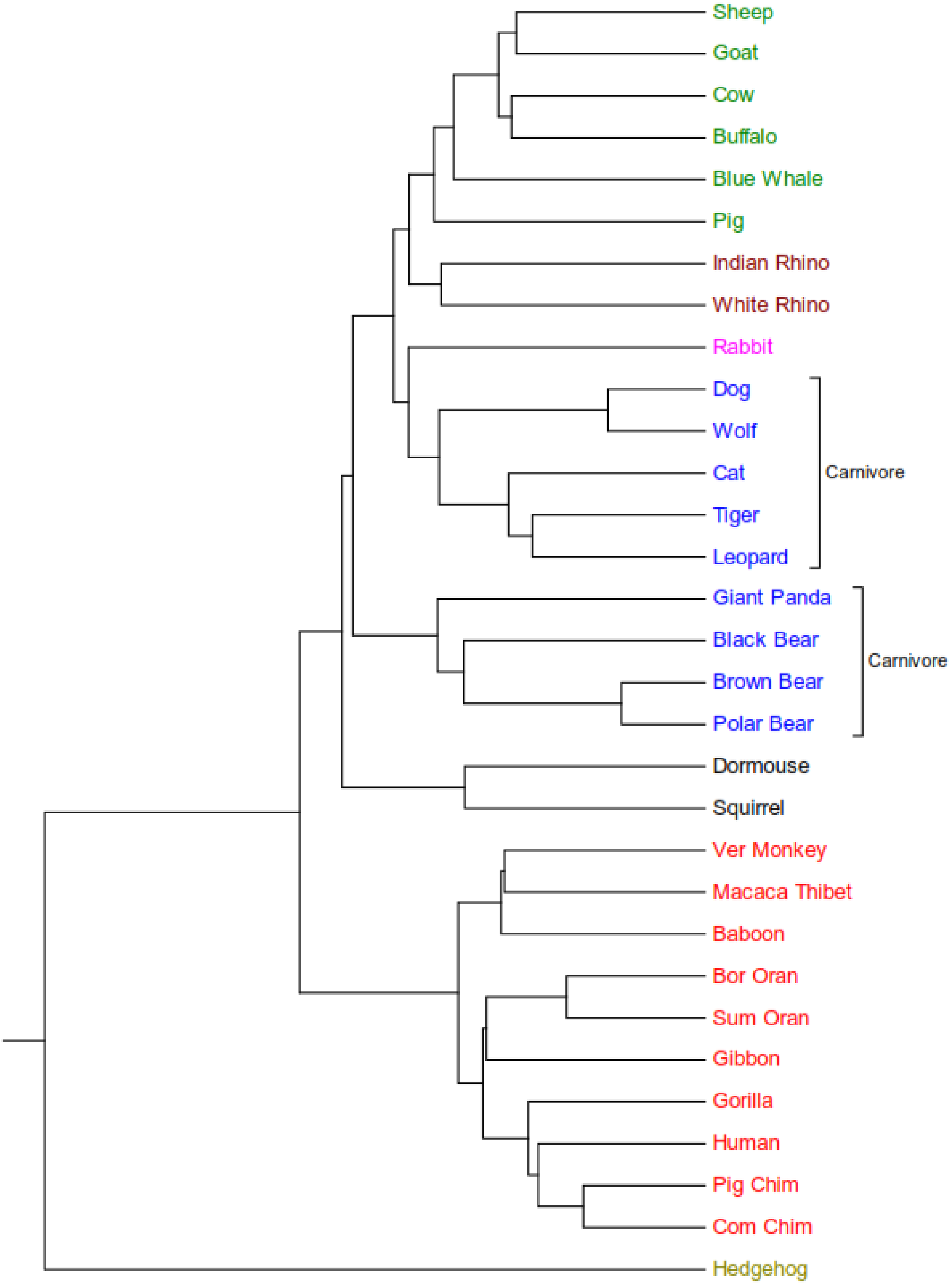

Fig. 4.6 Phylogenetic tree of 31 mammalian mitochondrial genomes by the *k*-mer method with *k*=5

4.2.4 Composition Vector Method Using Similarity Index as a Distance Measure [63]

In this comparative study, a dataset Influenza Virus is given in Table 4.6.

Table 4.6 Database of Influenza viruses

Year	Name	Acc. No.
1933	WSN (H1N1)	M12597
1934	Puerto Rico (H1N1)	J02150
1942	Bellamy (H1N1)	M12596
1947	Fort Monmouth (H1N1)	K00577
1950	Fort Warren (H1N1)	K00576
1957	Denver (H1N1)	M12592
1960	Ann Arbor (H2N2)	M12591
1968	Berkeley (H2N2)	M12590
1972	Victoria (H3N2)	AY210316
1977	Alaska (H3N2)	K01332
1977	USSR (H1N1)	K00578
1980	Maryland (H1N1)	M12595
1984	Houston (H1N1)	M12594
1985	Houston (H1N1)	M12593
1985	Houston (H3N2)	M17699
1997	Hong Kong (H3N2)	AF256183

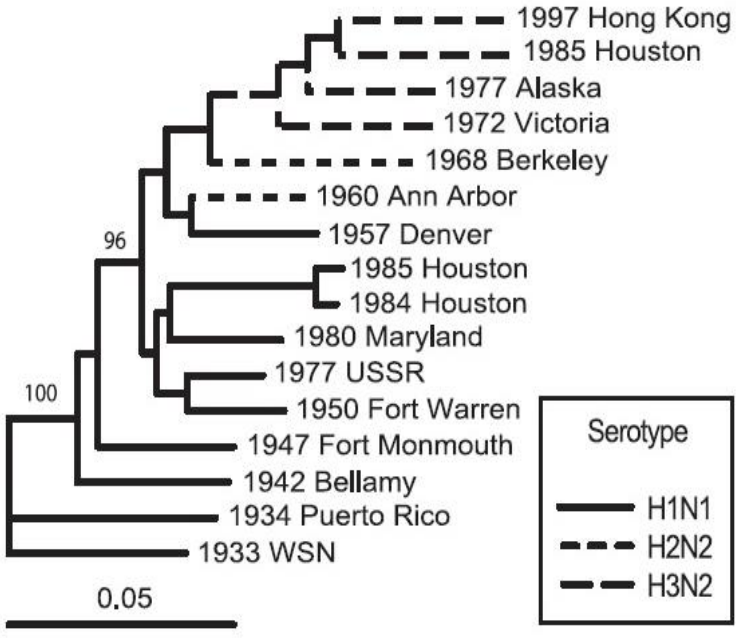

Fig. 4.7(a) Phylogenetic tree under *k*-mer method (*k*=4) with similarity index as the distance measure

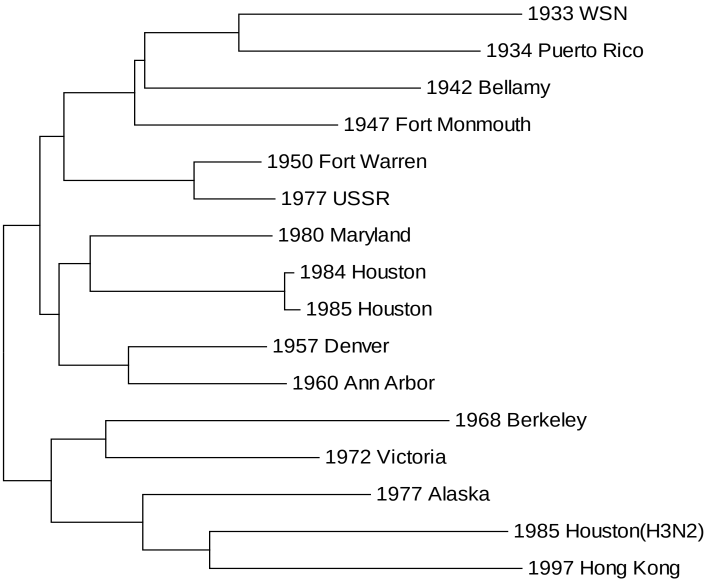

Fig. 4.7(b) Phylogenetic tree is obtained by the proposed method

The phylogenetic tree under the CV method with similarity index as the distance measure is given in Fig. 4.7(a). The phylogenetic tree is obtained by the proposed method is given in Fig. 4.7(b). The phylogenetic tree for Influenza virus as exhibited by Fig 4.7(a) and Fig 4.7(b) are the same. There is no difference in the organization of the trees.

The preceding results depict that the composition vector method using k-mer$(k \neq 3)$ has also been applied under the similarity index as the distance measure. Interestingly, in this case also 3-mer is an optimal choice. Thus, it is proved that choice of 3-mer and choice of similarity index as the distance measure provides a sort of unified approach towards comparison of whole genome sequences by composition vector method.

4.3 Conclusion

The present work established that the use of 3-mer as the string length and information based similarity measure using the similarity index as the distance measure provided an average successful composition vector method in almost all genome sequence comparison. The proposed method specially provided outstanding results compared to the existing methods of genome sequences comparison in case of mitochondrial genome of 31 mammalians, 53 complete genome sequences of TYLCV (Tomato Yellow Leaf Curl Virus) and 21 genomes (6 influenza A (H1N1), 6 swine flu virus, 6 avian virus genomes and 3 human seasonal flu virus).

CHAPTER 5

5. A New Form of Tri-Nucleotide Representation Based On Bio-Chemical Properties of Nucleotides

Genetic sequence analysis, classification of genome sequence and evolutionary relationship between species using their biological sequences, are the emerging research domain in Bioinformatics. Several methods have already been applied to DNA sequence comparison under tri-nucleotide representation. In this chapter, a new form of tri-nucleotide representation is proposed for sequence comparison. The comparison does not depend on the alignment of the sequences. In this representation, the bio-chemical properties of the nucleotides are considered. The novelty of this method is that the sequences of unequal lengths are represented by vectors of the same length and each of the tri-nucleotide formed out of the given sequence has its unique representation. To validate the proposed method, it is verified on several data sets related to mammalians, viruses and bacteria. The results of this method are further compared with those obtained by methods such as probabilistic method, natural vector method, Fourier power spectrum method, multiple encoding vector method, and feature frequency profiles method. Moreover, this method produces accurate phylogeny in all the cases. It is also proved that the time complexity of the present method is less.

5.1 Methodology

5.1.1 Di-Nucleotide Groups

The four bases A, C, G, T of primary DNA sequences, can be classified in two different ways:

Chemical Structures

Two classes:-

i) *Purine* group R = (A, G) and *Pyrimidine* group Y = (C, T).

ii) *Amino* group M = (A,C) and *Keto* group K = (G, T).

Strength of the Hydrogen Bond

One class:-

Weak *H–bonds* W = (A, T) and Strong *H–bonds* S = (G, C).

5.1.2 Di-Nucleotide Representation

Let $S = a_1 a_2 a_3 a_4 \ldots . a_n$ be a DNA primary sequence. Using the above classifications, we assign numbers from 1 to 6 to $a_i a_{i+1}$,*(i=1,2,.....n-1)* as follows:

$$a_i a_{i+1} = \begin{cases} 1 \; if \; a_i a_{i+1} = AT \\ 2 \; if \; a_i a_{i+1} = CG \\ 3 \; if \; a_i a_{i+1} = AC \\ 4 \; if \; a_i a_{i+1} = GT \\ 5 \; if \; a_i a_{i+1} = AG \\ 6 \; if \; a_i a_{i+1} = CT \end{cases}$$

5.1.3 3D Representation of A, C, T & G

Since each nucleotide A, C, G, T occurs in three groups, so each nucleotide is represented by a triplet of real numbersas given below:

$$A = (1,3,5)$$
$$C = (2,3,6)$$
$$T = (1,4,6)$$
$$G = (2,4,5)$$

5.1.4 Tri-Nucleotides Representation of DNA Sequence

The tri-nucleotides of the DNA sequence of length *n* are *n-2* in number. The first one starts with the first nucleotide of the sequence, the next tri-nucleotide starts from second nucleotide of the sequence and the process continues.The 9–dimensional tri-nucleotide representation is shown in Table 5.1.

Table 5.1 9-dimensional representation of 64 tri- nucleotides

Serial Number	Tri-nucleotide	9-dimention Representation								
1	AAA	1	3	5	1	3	5	1	3	5
2	AAC	1	3	5	1	3	5	2	3	6
3	AAT	1	3	5	1	3	5	1	4	6
4	AAG	1	3	5	1	3	5	2	4	5
5	CAA	2	3	6	1	3	5	1	3	5
6	CAC	2	3	6	1	3	5	2	3	6
7	CAT	2	3	6	1	3	5	1	4	6
8	CAG	2	3	6	1	3	5	2	4	5
9	TAA	1	4	6	1	3	5	1	3	5
10	TAC	1	4	6	1	3	5	2	3	6
11	TAT	1	4	6	1	3	5	1	4	6
12	TAG	1	4	6	1	3	5	2	4	5
13	GAA	2	4	5	1	3	5	1	3	5
14	GAC	2	4	5	1	3	5	2	3	6
15	GAT	2	4	5	1	3	5	1	4	6
16	GAG	2	4	5	1	3	5	2	4	5
17	ACA	1	3	5	2	3	6	1	3	5
18	ACC	1	3	5	2	3	6	2	3	6
19	ACT	1	3	5	2	3	6	1	4	6
20	ACG	1	3	5	2	3	6	2	4	5
21	CCA	2	3	6	2	3	6	1	3	5
22	CCC	2	3	6	2	3	6	2	3	6
23	CCT	2	3	6	2	3	6	1	4	6
24	CCG	2	3	6	2	3	6	2	4	5
25	TCA	1	4	6	2	3	6	1	3	5
26	TCC	1	4	6	2	3	6	2	3	6
27	TCT	1	4	6	2	3	6	1	4	6
28	TCG	1	4	6	2	3	6	2	4	5
29	GCA	2	4	5	2	3	6	1	3	5
30	GCC	2	4	5	2	3	6	2	3	6

Serial Number	Tri-nucleotide	9-dimention Representation								
31	GCT	2	4	5	2	3	6	1	4	6
32	GCG	2	4	5	2	3	6	2	4	5
33	ATA	1	3	5	1	4	6	1	3	5
34	ATC	1	3	5	1	4	6	2	3	6
35	ATT	1	3	5	1	4	6	1	4	6
36	ATG	1	3	5	1	4	6	2	4	5
37	CTA	2	3	6	1	4	6	1	3	5
38	CTC	2	3	6	1	4	6	2	3	6
39	CTT	2	3	6	1	4	6	1	4	6
40	CTG	2	3	6	1	4	6	2	4	5
41	TTA	1	4	6	1	4	6	1	3	5
42	TTC	1	4	6	1	4	6	2	3	6
43	TTT	1	4	6	1	4	6	1	4	6
44	TTG	1	4	6	1	4	6	2	4	5
45	GTA	2	4	5	1	4	6	1	3	5
46	GTC	2	4	5	1	4	6	2	3	6
47	GTT	2	4	5	1	4	6	1	4	6
48	GTG	2	4	5	1	4	6	2	4	5
49	AGA	1	3	5	2	4	5	1	3	5
50	AGC	1	3	5	2	4	5	2	3	6
51	AGT	1	3	5	2	4	5	1	4	6
52	AGG	1	3	5	2	4	5	2	4	5
53	CGA	2	3	6	2	4	5	1	3	5
54	CGC	2	3	6	2	4	5	2	3	6
55	CGT	2	3	6	2	4	5	1	4	6
56	CGG	2	3	6	2	4	5	2	4	5
57	TGA	1	4	6	2	4	5	1	3	5
58	TGC	1	4	6	2	4	5	2	3	6
59	TGT	1	4	6	2	4	5	1	4	6
60	TGG	1	4	6	2	4	5	2	4	5
61	GGA	2	4	5	2	4	5	1	3	5
62	GGC	2	4	5	2	4	5	2	3	6
63	GGT	2	4	5	2	4	5	1	4	6
64	GGG	2	4	5	2	4	5	2	4	5

5.1.5 A Sample of 10-Dimensional Vector Representations of Tri-Nucleotide Using Frequencies

As it is noted that there are always 64 different types of tri-nucleotides present in the sequence of length *n*; so in order to accommodate (*n-2*) numbers of tri-nucleotides from the said 64 varieties, the frequencies of the tri-nucleotides are to be considered. For the sake of convenience, we add the frequencies as an additional component to the 9-component representation of tri-nucleotides. As a result each tri-nucleotide is represented by a 10 component vector; the number of different tri-nucleotides remains the same as 64 in each sequence. For illustration, we consider the following sample: “AACCTGGATCAAGTCCTTAACGTGGAACCT”. Here total number of bases is 30. Therefore, 30-2=28 number of tri-nucleotides is represented in 64×10 dimensional matrices in Table 5.2.

Table 5.2 64× 10 matrix representation of “AACCTGGATCAAGTCCTTAACGTGGAACCT”

Tri-nucleotide	64×10 matrix									
AAA	1	3	5	1	3	5	1	3	5	0
AAC	1	3	5	1	3	5	2	3	6	3
AAT	1	3	5	1	3	5	1	4	6	0
AAG	1	3	5	1	3	5	2	4	5	1
CAA	2	3	6	1	3	5	1	3	5	1
CAC	2	3	6	1	3	5	2	3	6	0
CAT	2	3	6	1	3	5	1	4	6	0
CAG	2	3	6	1	3	5	2	4	5	0
TAA	1	4	6	1	3	5	1	3	5	1
TAC	1	4	6	1	3	5	2	3	6	0
TAT	1	4	6	1	3	5	1	4	6	0
TAG	1	4	6	1	3	5	2	4	5	0
GAA	2	4	5	1	3	5	1	3	5	1
GAC	2	4	5	1	3	5	2	3	6	0
GAT	2	4	5	1	3	5	1	4	6	1
GAG	2	4	5	1	3	5	2	4	5	0
ACA	1	3	5	2	3	6	1	3	5	0
ACC	1	3	5	2	3	6	2	3	6	2
ACT	1	3	5	2	3	6	1	4	6	0
ACG	1	3	5	2	3	6	2	4	5	1

Tri-nucleotide	64×10 matrix									
CCA	2	3	6	2	3	6	1	3	5	0
CCC	2	3	6	2	3	6	2	3	6	0
CCT	2	3	6	2	3	6	1	4	6	3
CCG	2	3	6	2	3	6	2	4	5	0
TCA	1	4	6	2	3	6	1	3	5	1
TCC	1	4	6	2	3	6	2	3	6	1
TCT	1	4	6	2	3	6	1	4	6	0
TCG	1	4	6	2	3	6	2	4	5	0
GCA	2	4	5	2	3	6	1	3	5	0
GCC	2	4	5	2	3	6	2	3	6	0
GCT	2	4	5	2	3	6	1	4	6	0
GCG	2	4	5	2	3	6	2	4	5	0
ATA	1	3	5	1	4	6	1	3	5	0
ATC	1	3	5	1	4	6	2	3	6	1
ATT	1	3	5	1	4	6	1	4	6	0
ATG	1	3	5	1	4	6	2	4	5	0
CTA	2	3	6	1	4	6	1	3	5	0
CTC	2	3	6	1	4	6	2	3	6	0
CTT	2	3	6	1	4	6	1	4	6	1
CTG	2	3	6	1	4	6	2	4	5	1
TTA	1	4	6	1	4	6	1	3	5	1
TTC	1	4	6	1	4	6	2	3	6	0
TTT	1	4	6	1	4	6	1	4	6	0
TTG	1	4	6	1	4	6	2	4	5	0
GTA	2	4	5	1	4	6	1	3	5	0
GTC	2	4	5	1	4	6	2	3	6	1
GTT	2	4	5	1	4	6	1	4	6	0
GTG	2	4	5	1	4	6	2	4	5	1
AGA	1	3	5	2	4	5	1	3	5	0
AGC	1	3	5	2	4	5	2	3	6	0
AGT	1	3	5	2	4	5	1	4	6	1
AGG	1	3	5	2	4	5	2	4	5	0
CGA	2	3	6	2	4	5	1	3	5	0
CGC	2	3	6	2	4	5	2	3	6	0
CGT	2	3	6	2	4	5	1	4	6	1
CGG	2	3	6	2	4	5	2	4	5	0
TGA	1	4	6	2	4	5	1	3	5	0
TGC	1	4	6	2	4	5	2	3	6	0
TGT	1	4	6	2	4	5	1	4	6	0
TGG	1	4	6	2	4	5	2	4	5	2
GGA	2	4	5	2	4	5	1	3	5	2
GGC	2	4	5	2	4	5	2	3	6	0
GGT	2	4	5	2	4	5	1	4	6	0
GGG	2	4	5	2	4	5	2	4	5	0

5.1.6 Generation of Distance Matrix from 10-Dimentional Tri-Nucleotide Representation

From each pair of genome sequence represented by a *64×10* matrices, the distance matrix is obtained by applying Euclidean distance as the distance measure.

5.1.7 Construction of Phylogenetic Tree

Phylogenetic tree for the given taxonomy of DNA sequences is obtained by applying UPGMA on the distance matrix.

5.2 Result and Discussion

In this paper, first of all, we consider 9-component vector of tri-nucleotide based on different classifications curve of the four bases, according to their chemical structure and the strength of the hydrogen bond to differentiate DNA sequences of different species. Next frequencies of the tri-nucleotides are taken as an additional component to make the representation a 10 component one. The inclusion of this additional component has made the graphical representation of the DNA sequence non-degenerate. Further this has reduced the comparison of every pair of DNA sequences, whatever may be their lengths, to a comparison of sequences of effective length of 64 components only.

Proposed approach is used to differentiate similarity and dissimilarity of the first exon of *β-globin* gene, whole genome sequences of 31 mammalian mitochondrial genes and Whole genome sequences of 53 Tomato Yellow Leaf Curl Virus (TYLCV). Finally the results of the present paper are compared with those obtained earlier by other methods.

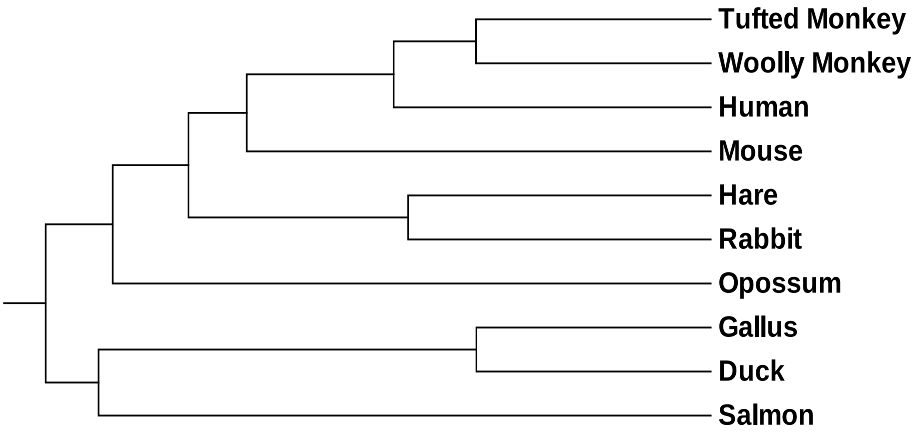

Fig. 5.1 Phylogenetic tree of complete coding sequence of *β-globin* gene of 10 different species

a. Dataset of DNA sequences of complete coding sequence of *β-globin* gene of 10 different species, considered for our experiment are given in Table 5.3. The distance matrix is given in Table 5.4. Finally MEGA tools are used to obtain the phylogenetic tree. It is shown in Fig.5.1.

Table 5.3 Complete coding sequence of *β-globin* gene of 10 different species

Serial Number	No. of the Species	GenBank ID
1	Hare	Y00347.1
2	Rabbit	V00882.1
3	Human	U01317.1
4	Mouse	V00722.1
5	Tufted Monkey	AY279115.1
6	Salmon	NM_001123672.1
7	Duck	X15739.1
8	Gallus	V00409.1
9	Woolly Monkey	AY279114.1
10	Opossum	J03643.1

Table 5.4 Distance matrix of 10 different species using our method

	Hare	Rabbit	Human	Mouse	Tufted Monkey	Salmon	Duck	Gallus	Woolly Monkey	Opossum
Hare		18.276	32.741	29.950	27.857	37.523	39.547	37.443	30.968	33.466
Rabbit			32.894	29.614	32.80	41.617	44.114	41.617	35.539	33.675
Human				25.632	20.736	43.267	42.048	40.620	17.578	38.859
Mouse					27.659	37.296	38.013	36.483	30.725	35.341
Tufted Monkey						39.497	39.749	37.630	14.177	38.026
Salmon							37.148	36.797	42.579	36.606
Duck								14.142	42.509	43.589
Gallus									40.902	40.546
Woolly Monkey										37.323
Opossum										

b. Table 5.5 exhibits data set of whole genome sequences of 31 mammalian mitochondrial genes. Fig. 5.2(a) and Fig. 5.2(b) shows the phylogenetic trees obtained by the probabilistic method [35] and by our present method respectively. In Fig. 5.2(a), it is found that human is misclassified with the group of Goat and Bear. But no such misclassification results from our present method. This shows that the present method is superior to the probabilistic method [35].

Table 5.5 31 mammalian mitochondrial whole genome

Serial Number	Name of the Species	Gene Bank ID
1	Human	V00662.1
2	Pigmy Chimpanzee	D38116.1
3	Common Chimpanzee	D38113.1
4	Gibbon	X99256.1
5	Baboon	Y18001.1
6	Vervet Monkey	AY863426.1
7	MacacaThibetana	NC_002764.1
8	Bornean Orangutan	D38115.1
9	Sumatran Orangutan	NC_002083.1
10	Gorilla	D38114.1
11	Cat	U20753.1
12	Dog	U96639.2
13	Pig	AJ002189.1
14	Sheep	AF010406.1
15	Goat	AF533441.1
16	Cow	V00654.1
17	Buffalo	AY488491.1
18	Wolf	EU442884.2
19	Tiger	EF551003.1
20	Leopard	EF551002.1
21	Indian Rhinoceros	X97336.1
22	White Rhinoceros	Y07726.1
23	Black Bear	DQ402478.1
24	Brown Bear	AF303110.1
25	Polar Bear	AF303111.1
26	Giant Panda	EF212882.1
27	Rabbit	AJ001588.1
28	Hedgehog	X88898.2
29	Dormouse	AJ001562.1
30	Squirrel	AJ238588.1
31	Blue Whale	X72204.1

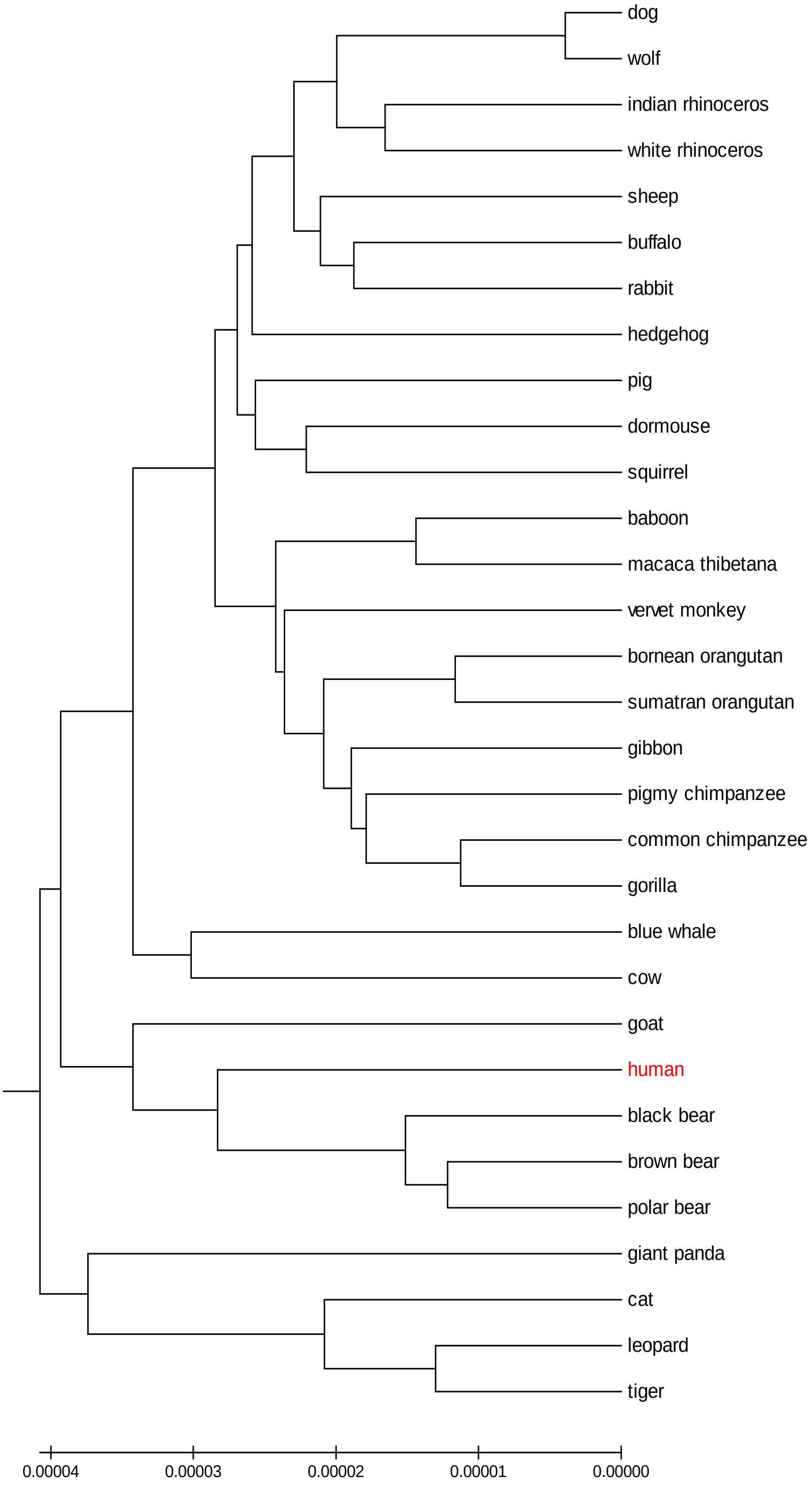

Fig. 5.2(a) Phylogenetic tree of 31 mammalian species using the Probabilistic method

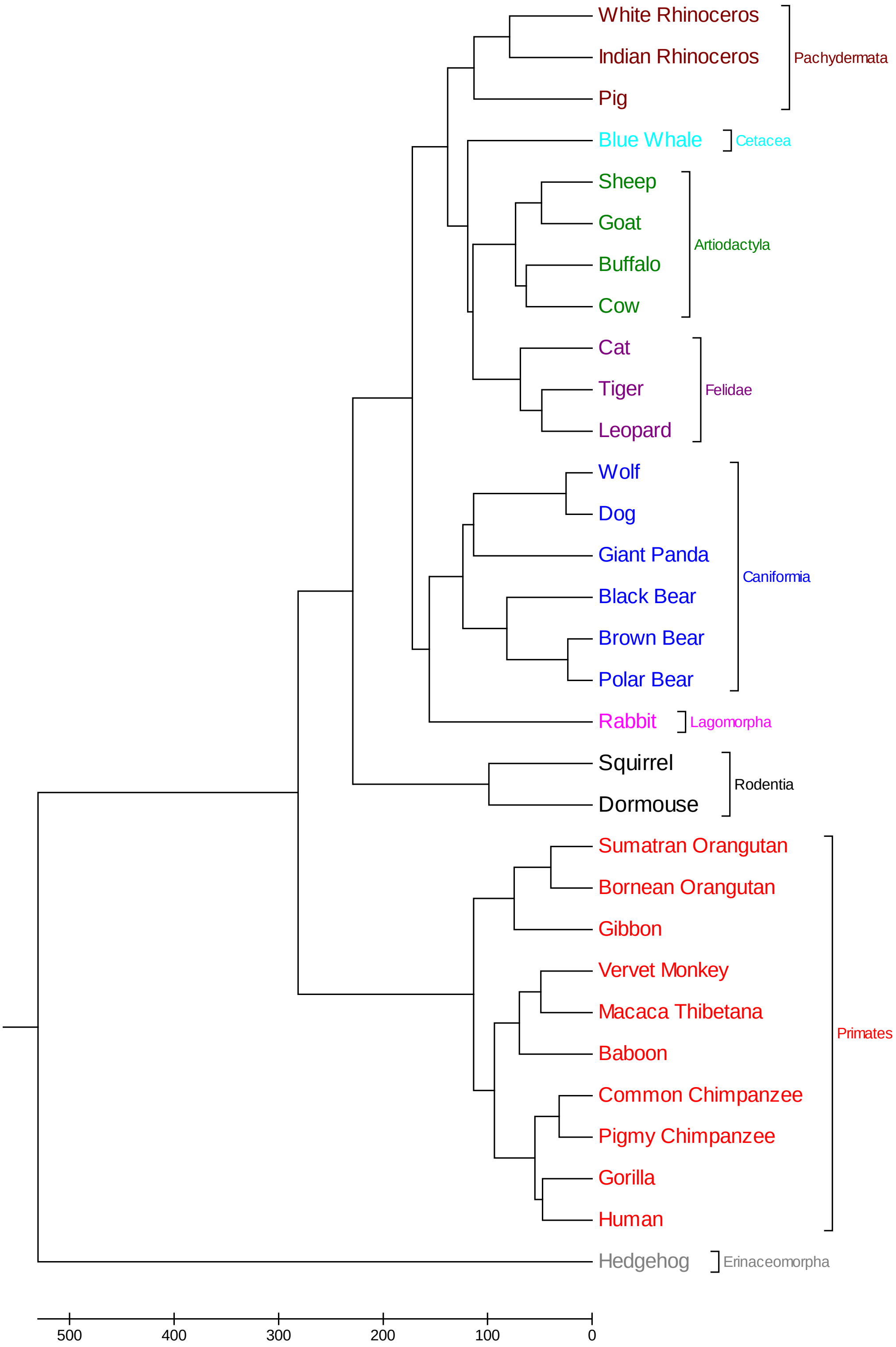

Fig. 5.2(b) Phylogenetic tree of complete mitochondrial genome sequence of 31 mammalian species by our method

c. Table 5.6 exhibits data set of 53 complete genome sequence of TYLCV. The corresponding Phylogenetic trees given by probabilistic tree [35] and our present method are given by Fig. 5.4(a) and Fig. 5.3(b) respectively. It is found that in Fig. 5.3(a) some of the TYLCV of *Severe phenotype* are classified with *Mild Phenotypes*. These misclassifications, no doubt, challenge the soundness of the probability method. But the present method is free from any such misappropriation, as no misclassification occurs in Fig. 5.3(b).

Table 5.6 TYLCD-causing virus sequences used in this study

Serial Number	Isolate	Accession no.	Length
1	TYLCV_IL	X15656	2787
2	TYLCV_DO	AF024715	2781
3	TYLCV_CU	AJ223505	2781
4	TYLCV_Flo	AY530931	2781
5	TYLCV_Omu	AB116630	2774
6	TYLCV_Alm	AJ489258	2781
7	TYLCV_Mis	AB116631	2774
8	TYLCV_EG_Ism	AY594174	2781
9	TYLCV_Miy	AB116629	2774
10	TYLCV_PR	AY134494	2781
11	TYLCV_MA	EF060196	2781
12	TYLCV_TR_Mer1_04	AJ812277	2781
13	TYLCV_Tosa:H	AB192966	2781
14	TYLCV_Tosa	AB192965	2781
15	TYLCV_RE4	AM409201	2781
16	TYLCV_Sic	DQ144621	2781
17	TYLCV_TN	EF101929	2781
18	TYLCV_JO	EF054893	2781
19	TYLCV_MX_Cul	DQ631892	2781
20	TYLCV_Mld_PT	AF105975	2793
21	TYLCV_Mld_Aic	AB014347	2787
22	TYLCV_Mld_Shi	AB014346	2791
23	TYLCV_Mld_ES7297	AF071228	2791
24	TYLCV_Mld_ES	AJ519441	2790
25	TYLCV_Mld_Sz_Yai	AB116632	2791
26	TYLCV_Mld_Atu	AB116633	2787

27	TYLCV_Mld_Kis	AB116634	2787
28	TYLCV_Mld_Sz_Dai	AB116635	2787
29	TYLCV_Mld_Sz_Osu	AB116636	2787
30	TYLCV_Mld_RE	AJ865337	2791
31	TYLCV_Mld_JO	EF054894	2791
32	TYLCAxV_Alg	AY227892	2772
33	TYLCMalV	AF271234	2782
34	TYLCMLV	AY502934	2794
35	TYLCMLV_ET	DQ358913	2785
36	TYLCSV	X61153	2773
37	TYLCSV_Sic	Z28390	2773
38	TYLCSV_ES1	Z25751	2777
39	TYLCSV_ES2	L27708	2777
40	TYLCSV_MA	AY702650	2777
41	TYLCSV_TN	AY736854	2772
42	TYLCCNV	AF311734	2734
43	TYLCCNV_Tb_Y25	AJ457985	2738
44	TYLCCNV_YM	DQ256460	2731
45	TYLCKaV_TH_Kan1	AF511529	2752
46	TYLCKaV_TH_Kan2	AF511530	2752
47	TYLCKaV_VN	DQ169054	2751
48	TYLCTHV	X63015	2743
49	TYLCTHV_MM	AF206674	2746
50	TYLCTHV_Y72	AJ495812	2748
51	TYLCTHV_ChMai	AY514630	2747
52	TYLCTHV_NoK	AY514631	2744
53	TYLCTHV_SaNa	AY514632	2747

We also compare our results with those obtained in the same sequence as given in [42]. We find that the results remain the same. This shows that both these methods are equally important and equally effective.

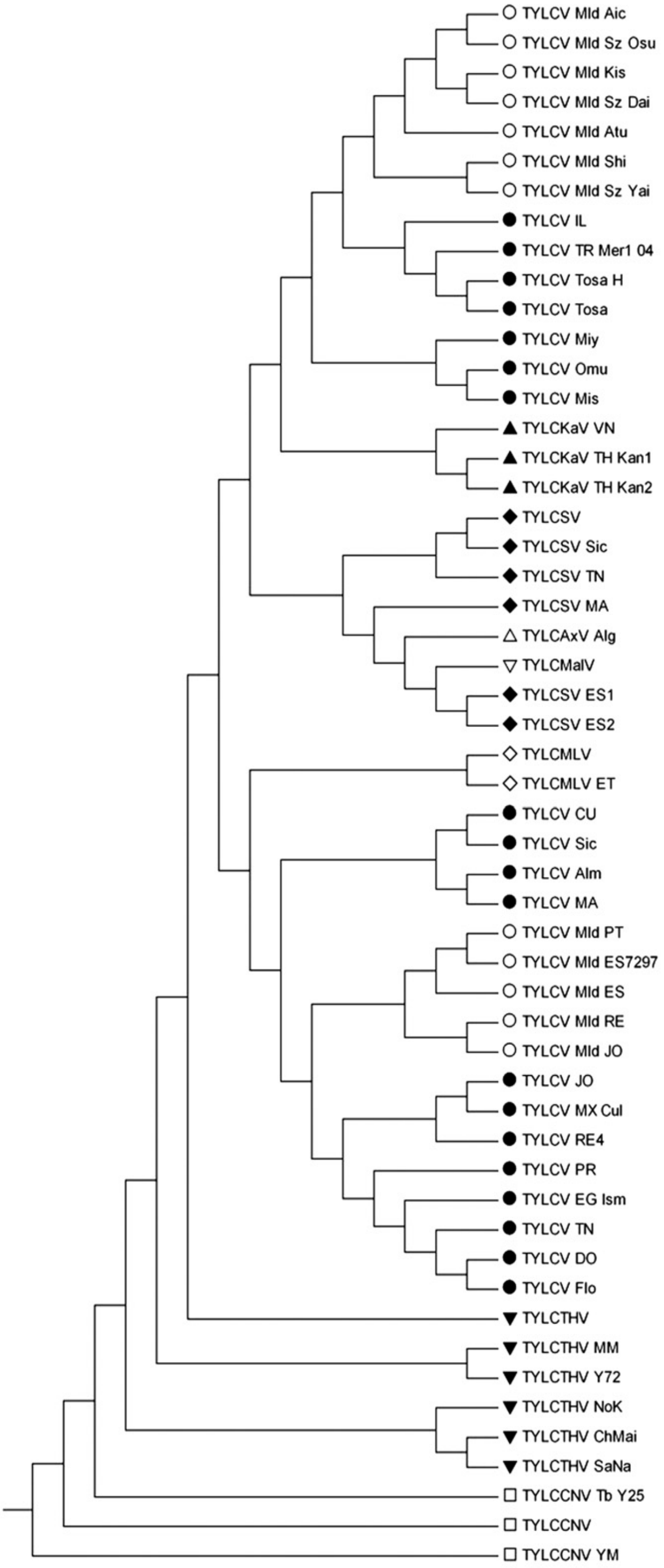

Fig. 5.3(a) Phylogenetic tree of 53 complete genome sequence of TYLCV using Probabilistic method

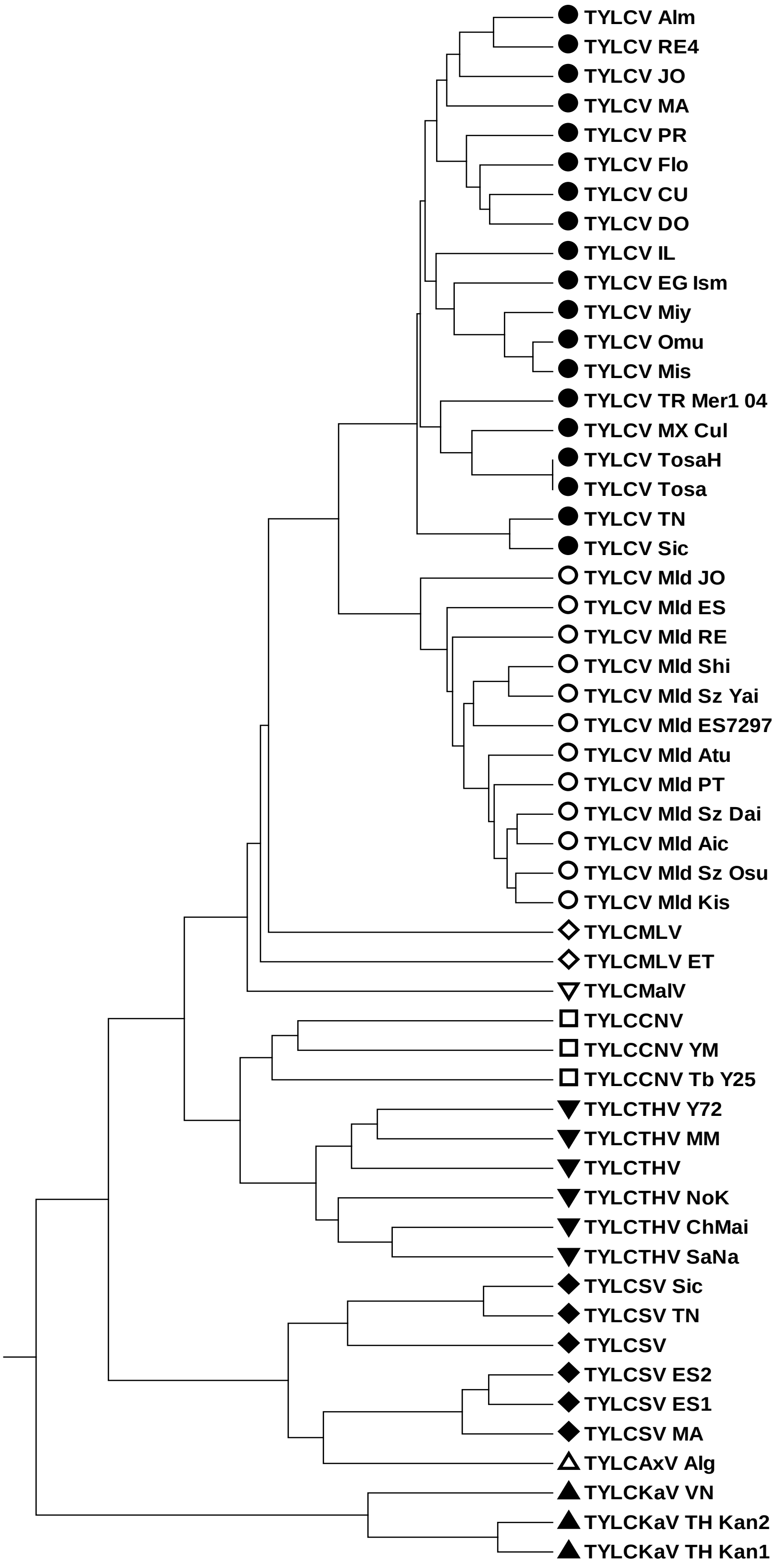

Fig. 5.3(b) Phylogenetic tree of 53 complete genome sequence of TYLCV using our method

d. For our next experiment we consider the data set of 59 bacterial genomes of 15 different families: Aeromonadaceae, Alcaligenaceae, Bacilleceae, Borreliaceae, Burkholderiaceae, Caulobacteraceae, Clostridiaceae, Desulfovibrionaceae, Enterobacteriaceae, Erwiniaceae, Lactobacillaceae, Mycoplasmataceae, Rhodobacteraceae, Staphylococcaceae, Yersiniaceae (Table 5.7). The lengths of these genome sequences are very large, dataset vary from 3 to 10 Mb. Constructed phylogenetic tree using our method is shown in Fig.5.4(a) In Figure Fig. 5.4(a)**,** all the bacteria are clearly grouped into fifteen discrete clusters. It is also shown that the relative position of all closely related bacteria in each family is precise. To verify our results, we compare it with the phylogenetic tree taken from the multiple encoding vector method [66] shown in Fig. 5.4(b). As per Fig. 5.4(b), all the bacteria of Enterobacteriaceae family are not clustered in a single clade; they have made a wrong cluster with Yersiniaceae. Again, we compare our phylogeny with that of feature frequency profiles (FFP) method [67] shown in Fig. 5.4(c). In Fig. 5.4(c)**,** the three families Lactobacillaceae, Clostridiaceae and Staphylococcaceae from PhylumBacilli are not clustered together as per closely related family. But In Fig. 5.4(a), all the bacteria are classified distinctly as per their family. Therefore, for classification, of 59 bacterial genomes of 15 different families, our method produces more accurate result than multiple encoding vector method and FFP method.

Table 5.7 Information of the 59 bacterial genomes from 15 families

Family	Species	Accession number	Genome Length(bp)
Aeromonadaceae	Aeromonas hydrophila strain AHNIH1	NZ_CP016380.1	4,906,118
	Aeromonas hydrophila strain GYK1	NZ_CP016392.1	4,906,118
	Aeromonas hydrophila YL17	CP007518.2	4,796,281
	Aeromonas veronii strain AVNIH1	NZ_CP014774.1	4,756,751
	Aeromonas veronii strain TH0426	NZ_CP012504.1	4,923,009
Bacillaceae	Bacillus anthracis str. A0248	CP001598.1	5,227,419
	Bacillus anthracis str. A16R	CP001974.2	5,228,828
	Bacillus anthracis str. Ames	AE016879.1	5,227,293
	Bacillus anthracis str. CDC 684	CP001215.1	5,230,115
	Bacillus anthracis str. H9401	NC_017729.1	5,218,947

	Bacillus anthracis str. Sterne	AE017225.1	5,228,663
	Bacillus cereus E33L	CP000001.1	5,300,915
Alcaligenaceae	Bordetella bronchialis strain AU17976	NZ_CP016171.1	5,966,919
	Bordetella bronchiseptica 253	NC_019382.1	5,264,383
	Bordetella flabilis strain AU10664	NZ_CP016172.1	5,835,727
Borreliaceae	Borrelia duttonii Ly	CP000976.1	931,674
	Borrelia hermsii DAH	CP000048.1	922,307
	Borrelia recurrentis A1	CP000993.1	930,981
	Borrelia turicatae 91E135	CP000049.1	917,330
Caulobacteraceae	Phenylobacteriumzucineum HLK1	CP000747.1	3,996,255
	Caulobacter crescentus CB15	NC_002696.2	4,016,947
	Caulobacter crescentus NA1000	NC_011916.1	4,042,929
Clostridiaceae	Clostridium perfringens ATCC 13124	CP000246.1	3,256,683
	Clostridium perfringens SM101	CP000312.1	2,897,393
	Clostridium perfringens str. 13 DNA	BA000016.3	3,031,430
Desulfovibrionaceae	Desulfovibrio vulgaris DP4	CP000527.1	3,462,887
	Desulfovibrio vulgaris Hildenborough	AE017285.1	3,570,858
	Desulfovibrio vulgaris RCH1	CP002297.1	3,532,052
Erwiniaceae	Erwinia pyrifoliae DSM 12163	FN392235.1	4,026,286
	Erwinia sp. Ejp617	CP002124.1	3,909,168
	Erwinia tasmaniensis strain ET1-99	CU468135.1	3,883,467
Lactobacillaceae	Lactobacillus acidophilus NCFM	NC_006814.3	1,993,560
	Lactobacillus helveticus DPC 4571	NC_010080.1	2,080,931
	Lactobacillus johnsonii NCC 533	NC_005362.1	1,992,676
Mycoplasmataceae	Mycoplasma agalactiae PG2	CU179680.1	877,438
	Mycoplasma conjunctivae HRC-581T	FM864216.2	846,214
	Mycoplasma fermentans JER	CP001995.1	977,524
Burkholderiaceae	Ralstoniaeutropha H16	AM260480.1	2,912,490
	Ralstoniaeutropha JMP134	CP000091.1	2,726,152
Rhodobacteraceae	Rhodobactersphaeroides 2.4.1	CP000144.2	943,018
	Rhodobactersphaeroides ATCC 17029	CP000578.1	1,219,053
	Rhodobactersphaeroides KD131	CP001151.1	1,297,647
Staphylococcaceae	Staphylococcus carnosus subsp. carnosus TM300	AM295250.1	2,566,424
	Staphylococcus epidermidis ATCC 12228	AE015929.1	2,499,279
	Staphylococcus epidermidis RP62A	CP000029.1	2,616,530

	Staphylococcus haemolyticus JCSC1435 DNA	AP006716.1	2,685,015
	Staphylococcus lugdunensis HKU09-01	CP001837.1	2,658,366
Yersiniaceae	Yersinia pestis Antiqua	CP000308.1	4,702,289
	Yersinia pestis CO92	AL590842.1	4,653,728
	Yersinia pestis D106004	CP001585.1	4,640,720
	Yersinia pestis KIM10+	AE009952.1	4,600,755
	Yersinia pestis Z176003	CP001593.1	4,553,586
Enterobacteriaceae	Escherichia coli ABU 83972	CP001671.1	5,131,397
	Escherichia coli APEC O1	CP000468.1	5,082,025
	Escherichia coli ATCC 8739	CP000946.1	4,746,218
	Escherichia coli BL21	CP001665.1	4,570,938
	Shigella flexneri 2002017	CP001383.1	4,650,856
	Shigella flexneri 2a str. 301	AE005674.2	4,607,202
	Shigella sonnei Ss046	CP000038.1	4,825,265

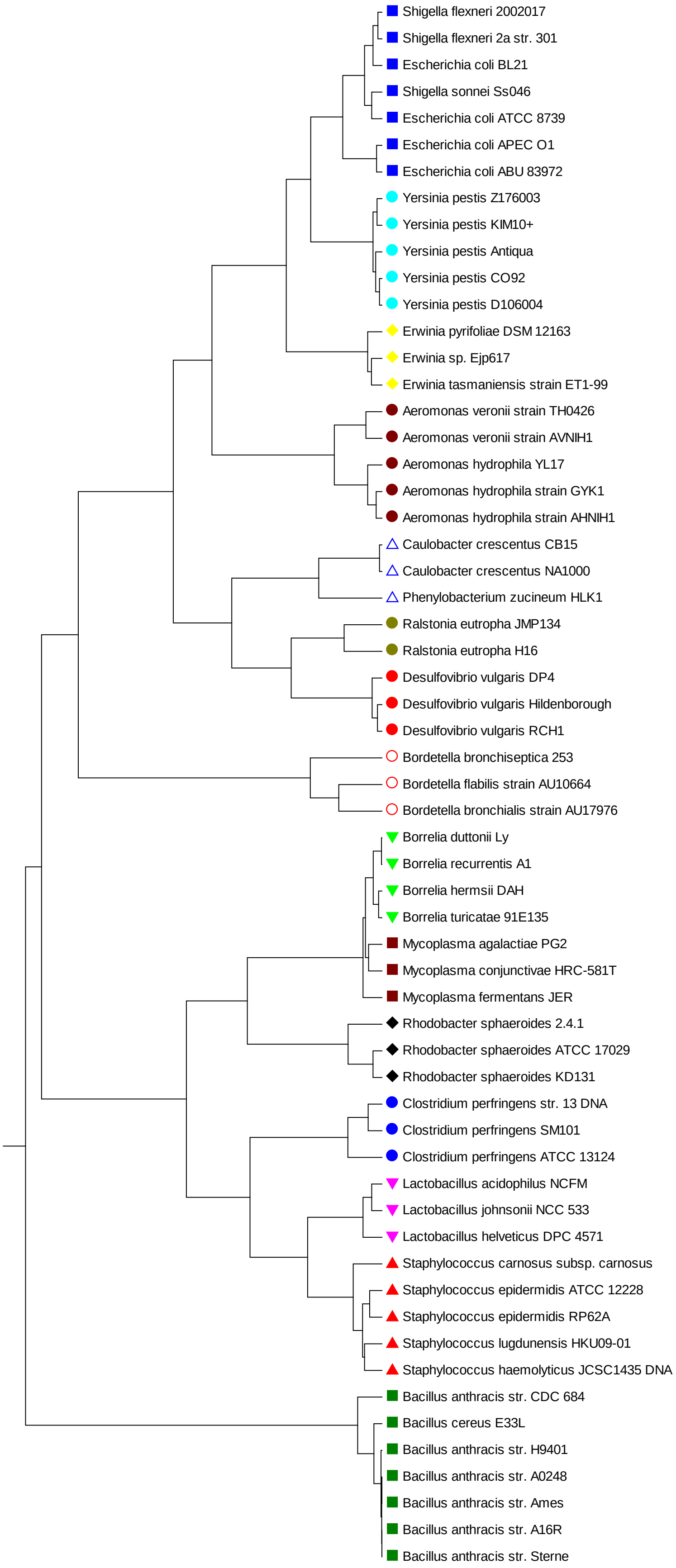
Shigella flexneri 2002017
Shigella flexneri 2a str. 301
Escherichia coli BL21
Shigella sonnei Ss046
Escherichia coli ATCC 8739
Escherichia coli APEC O1
Escherichia coli ABU 83972
Yersinia pestis Z176003
Yersinia pestis KIM10+
Yersinia pestis Antiqua
Yersinia pestis CO92
Yersinia pestis D106004
Erwinia pyrifoliae DSM 12163
Erwinia sp. Ejp617
Erwinia tasmaniensis strain ET1-99
Aeromonas veronii strain TH0426
Aeromonas veronii strain AVNIH1
Aeromonas hydrophila YL17
Aeromonas hydrophila strain GYK1
Aeromonas hydrophila strain AHNIH1
Caulobacter crescentus CB15
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Ralstonia eutropha JMP134
Ralstonia eutropha H16
Desulfovibrio vulgaris DP4
Desulfovibrio vulgaris Hildenborough
Desulfovibrio vulgaris RCH1
Bordetella bronchiseptica 253
Bordetella flabilis strain AU10664
Bordetella bronchialis strain AU17976
Borrelia duttonii Ly
Borrelia recurrentis A1
Borrelia hermsii DAH
Borrelia turicatae 91E135
Mycoplasma agalactiae PG2
Mycoplasma conjunctivae HRC-581T
Mycoplasma fermentans JER
Rhodobacter sphaeroides 2.4.1
Rhodobacter sphaeroides ATCC 17029
Rhodobacter sphaeroides KD131
Clostridium perfringens str. 13 DNA
Clostridium perfringens SM101
Clostridium perfringens ATCC 13124
Lactobacillus acidophilus NCFM
Lactobacillus johnsonii NCC 533
Lactobacillus helveticus DPC 4571
Staphylococcus carnosus subsp. carnosus
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Staphylococcus lugdunensis HKU09-01
Staphylococcus haemolyticus JCSC1435 DNA
Bacillus anthracis str. CDC 684
Bacillus cereus E33L
Bacillus anthracis str. H9401
Bacillus anthracis str. A0248
Bacillus anthracis str. Ames
Bacillus anthracis str. A16R
Bacillus anthracis str. Sterne

Fig. 5.4(a) The phylogenetic tree of 59 bacteria from 15 families based on our method

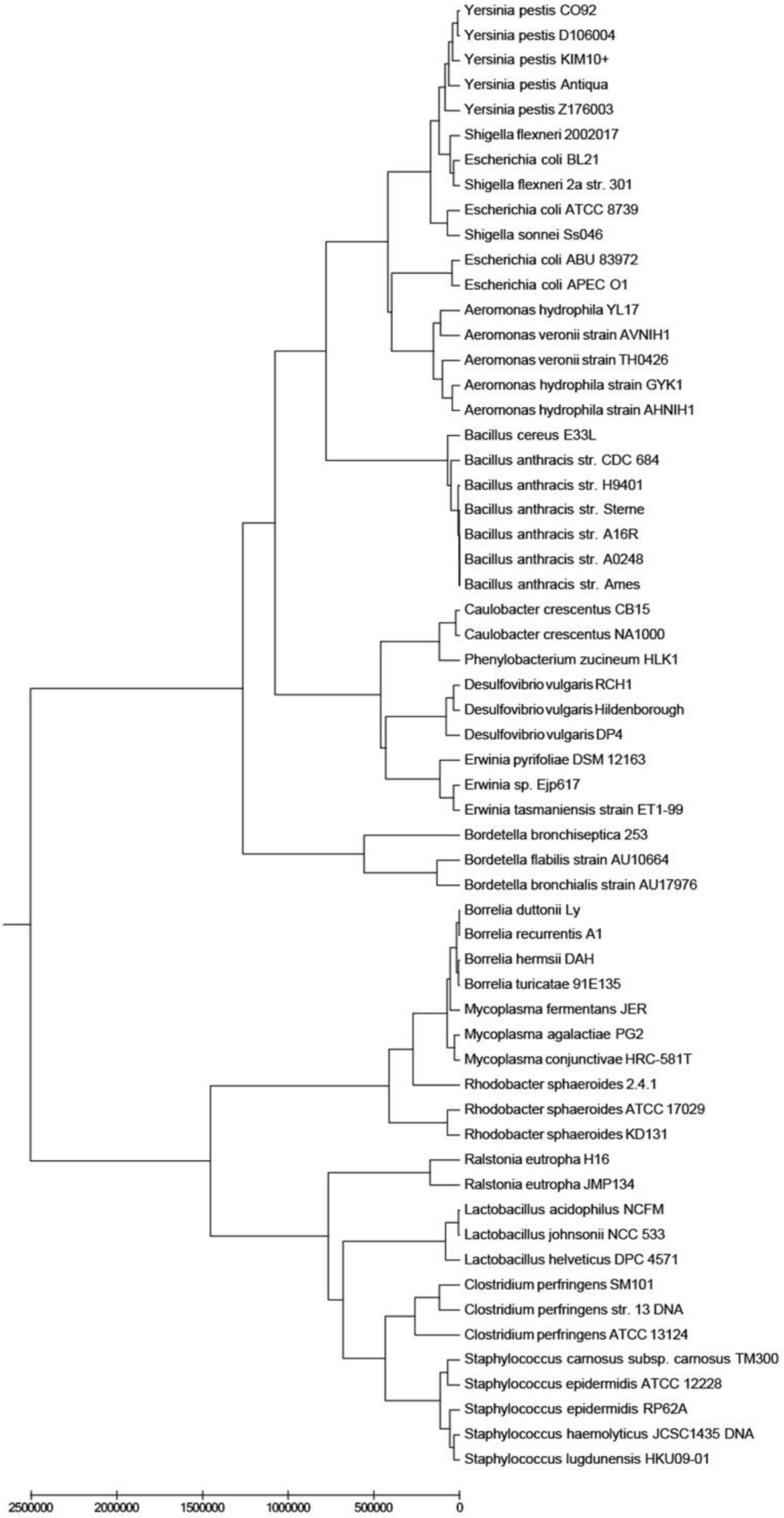
Yersinia pestis CO92
Yersinia pestis D106004
Yersinia pestis KIM10+
Yersinia pestis Antiqua
Yersinia pestis Z176003
Shigella flexneri 2002017
Escherichia coli BL21
Shigella flexneri 2a str. 301
Escherichia coli ATCC 8739
Shigella sonnei Ss046
Escherichia coli ABU 83972
Escherichia coli APEC O1
Aeromonas hydrophila YL17
Aeromonas veronii strain AVNIH1
Aeromonas veronii strain TH0426
Aeromonas hydrophila strain GYK1
Aeromonas hydrophila strain AHNIH1
Bacillus cereus E33L
Bacillus anthracis str. CDC 684
Bacillus anthracis str. H9401
Bacillus anthracis str. Sterne
Bacillus anthracis str. A16R
Bacillus anthracis str. A0248
Bacillus anthracis str. Ames
Caulobacter crescentus CB15
Caulobacter crescentus NA1000
Phenylobacterium zucineum HLK1
Desulfovibrio vulgaris RCH1
Desulfovibrio vulgaris Hildenborough
Desulfovibrio vulgaris DP4
Erwinia pyrifoliae DSM 12163
Erwinia sp. Ejp617
Erwinia tasmaniensis strain ET1-99
Bordetella bronchiseptica 253
Bordetella flabilis strain AU10664
Bordetella bronchialis strain AU17976
Borrelia duttonii Ly
Borrelia recurrentis A1
Borrelia hermsii DAH
Borrelia turicatae 91E135
Mycoplasma fermentans JER
Mycoplasma agalactiae PG2
Mycoplasma conjunctivae HRC-581T
Rhodobacter sphaeroides 2.4.1
Rhodobacter sphaeroides ATCC 17029
Rhodobacter sphaeroides KD131
Ralstonia eutropha H16
Ralstonia eutropha JMP134
Lactobacillus acidophilus NCFM
Lactobacillus johnsonii NCC 533
Lactobacillus helveticus DPC 4571
Clostridium perfringens SM101
Clostridium perfringens str. 13 DNA
Clostridium perfringens ATCC 13124
Staphylococcus carnosus subsp. carnosus TM300
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Staphylococcus haemolyticus JCSC1435 DNA
Staphylococcus lugdunensis HKU09-01
2500000
2000000
1500000
1000000
500000
0

Fig. 5.4(b) The UPGMA phylogenetic tree of 59 bacteria from 15 families based on multiple encoding vector method

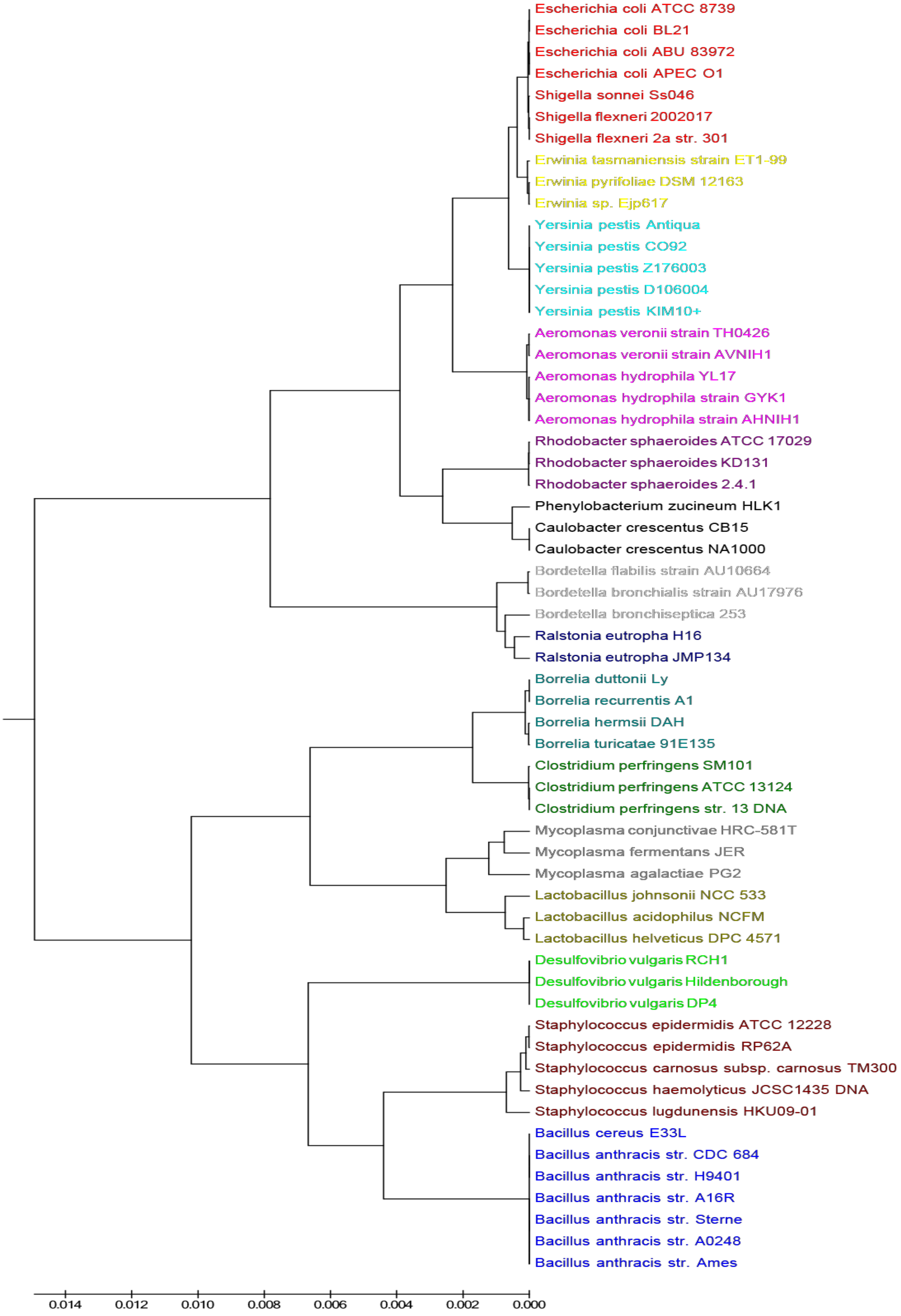
Escherichia coli ATCC 8739
Escherichia coli BL21
Escherichia coli ABU 83972
Escherichia coli APEC O1
Shigella sonnei Ss046
Shigella flexneri 2002017
Shigella flexneri 2a str. 301
Erwinia tasmaniensis strain ET1-99
Erwinia pyrifoliae DSM 12163
Erwinia sp. Ejp617
Yersinia pestis Antiqua
Yersinia pestis CO92
Yersinia pestis Z176003
Yersinia pestis D106004
Yersinia pestis KIM10+
Aeromonas veronii strain TH0426
Aeromonas veronii strain AVNIH1
Aeromonas hydrophila YL17
Aeromonas hydrophila strain GYK1
Aeromonas hydrophila strain AHNIH1
Rhodobacter sphaeroides ATCC 17029
Rhodobacter sphaeroides KD131
Rhodobacter sphaeroides 2.4.1
Phenylobacterium zucineum HLK1
Caulobacter crescentus CB15
Caulobacter crescentus NA1000
Bordetella flabilis strain AU10664
Bordetella bronchialis strain AU17976
Bordetella bronchiseptica 253
Ralstonia eutropha H16
Ralstonia eutropha JMP134
Borrelia duttonii Ly
Borrelia recurrentis A1
Borrelia hermsii DAH
Borrelia turicatae 91E135
Clostridium perfringens SM101
Clostridium perfringens ATCC 13124
Clostridium perfringens str. 13 DNA
Mycoplasma conjunctivae HRC-581T
Mycoplasma fermentans JER
Mycoplasma agalactiae PG2
Lactobacillus johnsonii NCC 533
Lactobacillus acidophilus NCFM
Lactobacillus helveticus DPC 4571
Desulfovibrio vulgaris RCH1
Desulfovibrio vulgaris Hildenborough
Desulfovibrio vulgaris DP4
Staphylococcus epidermidis ATCC 12228
Staphylococcus epidermidis RP62A
Staphylococcus carnosus subsp. carnosus TM300
Staphylococcus haemolyticus JCSC1435 DNA
Staphylococcus lugdunensis HKU09-01
Bacillus cereus E33L
Bacillus anthracis str. CDC 684
Bacillus anthracis str. H9401
Bacillus anthracis str. A16R
Bacillus anthracis str. Sterne
Bacillus anthracis str. A0248
Bacillus anthracis str. Ames
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000

Fig. 5.4(c) The UPGMA phylogenetic tree of 59 bacteria from 15 families based on feature frequency profiles (FFP) method using 9-mer

5.3 Conclusion

Alignment-based sequence comparison methods are standard for their straightforward and clear mechanism. In the past few decades, these methods have been widely used to find evolutionary relationship and they have provided significant results. However, the main challenge of this method is high time complexity. Incidentally, various alignment-free measures have also been proposed that operate with linear time complexity, but still they have not been able to compute large number of sequences in all the cases. The main motivation to choose alignment-free methods is definitely the purpose to compute large-scale datasets with low time complexity. Therefore, the objective of this study is not to counter the existing methods of phylogenetic study, but fairly to provide a new alignment-free technique to investigate the evolutionary relationship of DNA sequences. The advantage of our method is that we consider not only tri-nucleotide to represent the DNA sequences, but generate numerical vector descriptor, which depends on the Bio-Chemical properties of the nucleotides. Present method supplies only real positive values to represent tri-nucleotide. By introducing frequency of such tri-nucleotide, we avoid non-degeneracy in representation of DNA sequences. Further, based on vector descriptor, a distance matrix is generated with the help of a simple Euclidean distance formula. The proposed approach is not only easy to compute, but also no additional parameters are required to construct phylogeny. Finally, the present method is verified on different datasets related to mammalians, viruses and bacteria. Proposed approach not only performs well as other alignment-free methods, but it also shows that the calculated time complexity is significantly less. It is established that this new approach provides satisfactory, even better results compared to any other alignment-free methods. Therefore, our method is a good alternative to the existing alignment-free methods with a linear time complexity.

CHAPTER 6

6. Future Scope

1. To develop fuzzy poly-nucleotide space on I^{16} applying four bases in place of three as I^{12}.

2. To apply *3*-mer representation of nucleotides of chapter 4 in analysis of protein sequences by using the codon representation of the amino acid.

REFERENCES

[1] Gates, M. A., A simple way to look at DNA, Journal of Theoretical Biology, 1986, 119, 319-328

[2] Nandy, A., A New Graphical Representation and Analysis of DNA Sequence Structure: I. Methodology and Application to Globin Genes, Current Science, 1994, 66, 309–314.

[3] Leong, P. M., Morgenthaler, S., Random walk and gap plots of DNA sequences, Comput Appl Biosci. 1995, 11, 503-507.

[4] Guo, X., Randic, M., Basak, S. C., A novel 2-D graphical representation of DNA sequences of low degeneracy, Chemical Physics Letters, 2001, 350, 106-112

[5] Stephen, S., Yau, T., Wang, J.,Niknejad, A., Lu, C.,Jin, N., Ho, Y. K., DNA sequence representation without degeneracy, Nucleic Acids Research, 2003, 31, 3078–3080

[6] Liao, B., A 2D graphical representation of DNA sequence, Chemical Physics Letters, 2005, 401, 196-199

[7] Liao, B, Tan, M., Ding, K., Application of 2-D graphical representation of DNA sequence, Chemical Physics Letters, 2005, 414, 296-300

[8] He, P., Wang, J., Numerical Characterization of DNA Primary Sequence, Internet Electron. J. Mol. Des. 2002, 1, 668-674.

[9] Song, J., Tang, H., A new 2-D graphical representation of DNA sequences and their numerical characterization, Journal of biochemical and biophysical methods, 2005, 63, 228-239

[10] Randic, M., Vracko, M., Lers, N., Plavsic, D., Novel 2-D Graphical representation of DNA sequence and their numerical characterization, Chem. Phys. Lett., 2003, 368, 1–6.

[11] Randic, M., Vracko, M., Lers, N., Plavsic, D., Analysis of similarity/dissimilarity of DNA sequences based on novel 2-D graphical representation, Chemical Physics Letters, 2003, 371, 202-207

[12] Yao, Y., Liao, B., Wang, T., A 2D graphical representation of RNA secondary structures and the analysis of similarity/dissimilarity based on it, Journal of Molecular Structure: THEOCHEM, 2005, 755, 131-136.

[13] Li, C.; Wang, J., Numerical Characterization and Similarity Analysis of DNA Sequences Based on 2-D Graphical Representation of the Characteristic Sequences, Combinatorial Chemistry & High Throughput Screening, 2003, 6, 795-799

[14] Liao, B. and Wang, T., New 2D graphical representation of DNA sequences. J. Comput. Chem., 2004, 25: 1364-1368

[15] Liao, B. and Ding, K., Graphical approach to analyzing DNA sequences. J. Comput. Chem., 2005, 26, 1519-1523.

[16] Wang, J., Zhang, Y., Characterization and similarity analysis of DNA sequences grounded on a 2-D graphical representation, Chemical Physics Letters, 2006, 423, 50-53

[17] Hamori, H., and Ruskin, J., H curves, a novel method of representation of nucleotide series especially suited for long DNA sequences., J. Biol. Chem., 1983, 258, 1318-1327

[18] Randić M, Vračko M, Nandy A, Basak SC. On 3-D Graphical Representation of DNA Primary Sequences and Their Numerical Characterization. J Chem. Inf.Comput. Sci. 2000, 40, 1235–1244

[19] Li, C.; Wang, J. (Review Article) On a 3-D Representation of DNA Primary Sequences, Combinatorial Chemistry & High Throughput Screening, 2004, 7, 23.

[20] Yao, Y., Nan, X., Wang, T., Analysis of similarity/dissimilarity of DNA sequences based on a 3-D graphical representation, Chemical Physics Letters,2005, 411, 248-255

[21] Yuan, C., Liao, B., Wang, T., New 3D graphical representation of DNA sequences and their numerical characterization, Chemical Physics Letters,2003, 379, 412-417

[22] Liao, B., Wang, T., 3-D graphical representation of DNA sequences and their numerical characterization, Journal of Molecular Structure: THEOCHEM, 2004, 681, 209-212.

[23] Liao, B., Zhang, Y., Ding, K., Wang, T., Analysis of similarity/dissimilarity of DNA sequences based on a condensed curve representation, Journal of Molecular Structure: THEOCHEM, 2005, 717, 199-203.

[24] Zhu, W., Liao, B., Ding, K., A condensed 3D graphical representation of RNA secondary structures, Journal of Molecular Structure: THEOCHEM, 2005, 757, 193-198

[25] Bai, F., Zhu, W., Wang, T., Analysis of similarity between RNA secondary structures, Chemical Physics Letters, 2005, 408, 258-263

[26] Chi, R., Ding, K., Novel 4D numerical representation of DNA sequences, Chemical Physics Letters, 2005, 407, 63-67

[27] Zhang R, Zhang CT. A Brief Review: The Z-curve Theory and its Application in Genome Analysis. Curr Genomics. 2014; 15(2):78-94.

[28] Akhtar, M., Epps, J., and Ambikairajah, E., Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction, IEEE Journal of Selected topics in Signal Processing, Vol. 2, No. 3, pp. 310-321, June 2008.

[29] Chakrabarty, N., Spanias, A.,LasemidisL. D. and Tsakalis, K., Autoregressive Modeling and Feature Analysis of DNA Sequences, EURASIP Journal on Applied Signal Processing, Vol.2004, No. 1, pp. 13-28, Jan. 2004.

[30] Zhou, H., and Yan, H., Autoregressive Models for Spectral Analysis of Short Tandem Repeats in DNA Sequences, IEEE International Conference on Systems, Man and Cybernetics, Taipei, Taiwan, Oct.8-11, 2006.

[31] Anastassiou, D., Genomic Signal Processing, IEEE Signal Processing Magazine, Vol. 18, No. 4, pp. 8-20, July 2001.

[32] Cristea, P. D., Genetic Signal Representation and Analysis, SPIE Conference, BIOS’2002- International Biomedical Optics Symposium, Molecular Analysis and Informatics, San Jose USA, B.O.4623-10, pp.77-84, January 21-24, 2002.

[33] Cattani, C., Complex Representation of DNA Sequences, 2nd International Conference on Bioinformatics Research and Development-BIRD, Vol. 13, pp.528-537, Australia, July 07-09, 2008.

[34] A. K. Brodzik, and O. Peters, "Symbol-balanced quaternionic periodicity transform for latent pattern detection in DNA sequences," in Proc. IEEE ICASSP, vol. 5, pp. 373-376, 2005.

[35] Yu, C., Deng, M., Yau, S.S.-T., DNA sequence comparison by a novel probabilistic method. Inf. Sci. 181 (2011) 1484–1492

[36] Qi, Z.-H., Fan, T.-R., PN-curve: a 3D graphical representation of DNA sequences and their numerical characterization. Chemical Physics Letters, 2007; 442:434–440.

[37] J. F. Yu, J. H. Wang, X. Sun, Analysis of similarities/dissimilarities of DNA sequences based on a novel graphical representation, MATCH Commun. Math. Comput. Chem. 2010; 63:493–512.

[38] Xiao Qing Liua, Qi Daib, Zhilong Xiua, Tianming Wang, PNN-curve: A new 2D graphical representation of DNA sequences and its applications -Journal of Theoretical Biology, 2006, 243, 555–561.

[39] Z. B. Liu, B. Liao, W. Zhu, G. H. Huang, A 2–D graphical representation of DNA sequence based on dual nucleotides and its application, Int. J. Quantum Chem. 2009; 109:948–958.

[40] M. Randic, J. Zupan, A.T. Balaban, Unique graphical representation of protein sequences based on nucleotide triplet codons, Chemical Physics Letters, 2004; 397:247–252.

[41] Yu, J.F., Sun, X., Wang, J.H., TN curve: a novel 3D graphical representation of DNA sequence based on trinucleotides and its applications. Journal of Theoretical Biology, 2009; 261: 459-468.

[42] Das, S., Deb, T., Dey, N., Ashour, A. S., Tibarewala, D. N., and Bhattacharya, D. K., Optimal choice of k-mer in composition vector method for genome sequence comparison, Genomics, 2018; 110: 263-273

[43] Das, S., Choudhury, N. R., Tibarewala, D. N., Bhattacharya, D. K., Application of Chaos Game in Tri-Nucleotide Representation for the Comparison of Coding

Sequences of β-Globin Gene, 2018, Industry Interactive Innovations in Science, Engineering and Technology, Lecture Notes in Networks and Systems, vol. 11, Springer

[44] Yu-hua Yao, Xu-ying Nan, Tian-ming Wang, A new 2D graphical representation—Classification curve and the analysis of similarity/dissimilarity of DNA sequences, Journal of Molecular Structure: THEOCHEM, Volume 764, Issues 1–3, 2006, 101-108

[45] Randić, M., Witzmann, F., Vračko, M., & Basak, S. C. (2001). On characterization of proteomics maps and chemically induced changes in proteomes using matrix invariants: Application to peroxisome proliferators. Medicinal Chemistry Research, 10(7-8), 456-479.

[46] J. Luo, J. Guo and Y. Li, A New Graphical Representation and Its Application in Similarity/Dissimilarity Analysis of DNA Sequences, 2010 4th International Conference on Bioinformatics and Biomedical Engineering, Chengdu, 2010, pp. 1-5.

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

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

[49] Das S., Palit S., Mahalanabish A.R., Choudhury N.R. (2015) A New Way to Find Similarity/Dissimilarity of DNA Sequences on the Basis of Dinucleotides Representation. In: Maharatna K., Dalapati G., Banerjee P., Mallick A., Mukherjee M. (eds) Computational Advancement in Communication Circuits and Systems. Lecture Notes in Electrical Engineering, vol 335. Springer, New Delhi

[50] Lu, G., Zhang, S. and Fang, X. (2008) An improved string composition method for sequence comparison. BMC Bioinformatics, 9 (Suppl 6), S15.

[51] Stuart, G.W., Moffett, K. and Baker, S., Integrated gene and species phylogenies from unaligned whole genome protein sequences. Bioinformatics, 62 (2002) 100–108.

[52] Chou, K.C., Shen, H.B., Review: recent advances in developing web-servers for predicting protein attributes. Nat. Sci. 2 (2009) 63–92.

[53] Stuart, G.W., Moffett, K. and Leader, J.J., A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. Molecular Biology and Evolution, 19 (2002) 554–562.

[54] Wang, J. and Zheng, X., WSE, a new sequence distance measure based on word frequencies, Mathematical Biosciences, 215 (2008) 78–83.

[55] Jianhua Lin., Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1) (1991) 145–151.

[56] Xiaomeng Wu, Zhipeng Cai, Xiu-Feng Wan, Tin Hoang, Randy Goebel, Guohui Lin, Nucleotide composition string selection in HIV-1 subtyping using whole genomes, Bioinformatics. 2007; 23(14): 1744–1752.

[57] Liao, B., Tian-ming Wang, Analysis of similarity/dissimilarity of DNA sequences based on 3-D graphical representation, Chemical Physics Letters, 388, 1, 2004, 195-200

[58] Welch, B.L. (1938) The Significance of the Difference between Two Means when the Population Variances are Unequal, Biometrika, 29, 350-362.

[59] Atanassov, K. (1986) Intuitionistic fuzzy sets, Fuzzy Sets and Systems, 20, 87–96.

[60] Atanassov, K. (1989) More on intuitionistic fuzzy sets, Fuzzy Sets and Systems 33, 37–46.

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

[62] Sadegh-Zadeh, K., Fuzzy genomes. Artif. Intell. Med. 18, (2000), 1-28.

[63] Yang AC, Goldberger AL, Peng CK. Genomic classification using an information-based similarity index: application to the SARS coronavirus. J Comput Biol. 12(8) (2005) 1103.

[64] Charlebois, R.L., Beiko, R.G. and Ragan, M.A., Microbial phylogenomics: Branching out. Nature, 421 (2003), 17–217.

[65] Li C., Yang, Y., Jia, M.D., Zhang, Y.Y., Yu, X.Q., Wang, C.Z. Phylogenetic analysis of DNA sequences based on k-word and rough set theory. Physica A 398 (2014) 162–171.

[66] Li, Y., He, L., He, R. L., & Yau, S. S. T. (2017). A novel fast vector method for genetic sequence comparison. Scientific reports, 7(1), 12226.

[67] Sims, G. E., Jun, S. R., Wu, G. A., & Kim, S. H. (2009). Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. *Proceedings of the National Academy of Sciences*, *106*(8), 2677-2682.