Efficacy of the *fuzzy polynucleotide space* in Phylogenetic Tree construction

Awanti Sambarey [1*], Ashok Deshpande[2]

[1] Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.
awanti@mbu.iisc.ernet.in
[2]Berkeley Initiative in Soft Computing(BISC)-Special Interest Group- (SIG)-Environment Management Systems (EMS) at University of California at Berkeley, USA; College of Engineering Pune
ashok_deshpande@hotmail.com

Abstract.

The study of evolutionary relationships is an important endeavor in the field of Bioinformatics. The fuzzification of genomes led to the introduction of a "*fuzzy polynucleotide space*", which has been successfully used in classification and clustering of amino acids, thereby suggesting a possible application in phylogeny. As phylogenetic trees illustrate similarities and evolutionary relationships among different taxa, through this study we attempt to determine the efficacy of the fuzzy polynucleotide space in phylogenetic tree reconstruction, and discuss its implications in evolutionary biology.

Keywords: Fuzzy polynucleotide space, NTV metric, phylogenetic tree construction.

1. INTRODUCTION

Sequence analysis and comparative genomics play a central role in Bioinformatics. Phylogenetic relationships among organisms are established on the basis of molecular sequences, in order to understand their course of evolution and ancestry. Molecular phylogeny involves building of a relationship tree that shows the probable evolution of various organisms. The conventional tree building approaches are broadly divided into: a) **Distance based approaches**- which take into account the evolutionary distances between all taxa, where the distance represents the number of nucleotide or amino acid changes between sequences. These include methods such as Neighbour Joining, Unweighted Pair Group Mean Average (UPGMA), Minimum Evolution, and alike. b) **Character based approaches**- these include methods such as Maximum Parsimony, Maximum likelihood and Bayesian sequence analysis.

Statistical techniques have played, and will continue to play a pivotal role in sequence analysis. The past decade has witnessed several applications of fuzzy sets and fuzzy logic in bioinformatics, with its successful use in sequence alignment, DNA sequencing, clustering and classification [1,2,3,4]. Fuzzy set theory was first rendered directly accessible to sequence comparisons in the works of Sadegh-Zadeh. He introduced the concept of *Fuzzy Polynucleotides* [5], by transforming nucleic acid sequences into ordered fuzzy sets. The author showed that the genetic code can be considered as a 12 –dimensional code, with each triplet codon XYZ having a $3 \times 4 = 12$ dimensional fuzzy code, and thus falling as a point in what the author termed as the 12- dimensional *fuzzy polynucleotide space* $I=[0,1]^{12}$, where $I \in R$.

Torres and Nieto [6] redefined the Fuzzy Polynucleotide Space, based on the fuzzy hypercube concept proposed by Bart Kosko [7]. Taking into account the frequencies of the nucleotides at the three base sites of a codon in the coding sequence, the authors mapped a given polynucleotide on an I^{12} space which they termed as *Fuzzy Polynucleotide Space (FPNS)*. A sequence of any length could thus be mapped on a 12- dimensional vector, facilitating comparison between sequences of varying lengths. A distance metric *d* that determined distances between the fuzzy vectors of any two polynucleotides, was proposed.

Given the fuzzy polynucleotide space for two sequences p and q, where p = (p1, p2, . . pn), q = (q1, q2, . . . , qn) $\in$ In, n=12, the difference between p and q was calculated as:

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

The distance metric as defined in Equation (1) is termed as the NTV metric. The authors computed the fuzzy polynucleotide space for two genomes of E. coli and M. tuberculosis, considering only the coding regions of these genomes, and the distance between them was calculated. The approach was further extended and distances between other genomes were computed [8].

The NTV metric has also been used for the classification of amino acids via fuzzy equivalence relation [9]. In their research study, the authors used two different distance functions viz. the Minkowski distance function and the NTV metric. The clusters obtained using the NTV metric were the same as that obtained using the Minkowski distance metric for high values of the similarity degree. Nieto and Torres [10] have suggested the possible use of NTV metric in phylogenetic analysis. With this backdrop, we have, in this sequel, made an attempt to study the efficacy of the NTV metric in phylogenetic reconstruction.

2 METHODS

The structured approach is divided into three parts. Section 2.1 deals with data collection, while section 2.2 describes the salient features of sequence analysis, and the results and discussion on phylogenetic tree analysis are presented in section 2.3:

2.1 Data collection

A total of nine datasets were considered for the detailed study. However, the discussion on the results of three major datasets was considered sufficient to test the hypothesis and draw meaningful conclusions. The other datasets are available on request.

Dataset 1 comprises of polyprotein-coding regions of Dengue type 3 viruses. The viral isolates were chosen from different regions of the world. Dataset 2 represents gyrase B gene sequences from members of the genus Microbacterium. Dataset 3 includes vertebrate mitochondrial cytochrome b sequences. The cyt b genes were taken from representative members of the six classes viz. Mammalia, Reptilia, Amphibia, Aves, Chondrichthyes and Osteichthyes of Sub-Phylum Vertebrata. The other datasets, considered in the detailed analysis include gyrase B gene sequences from Burkholderia, ompA gene sequences from the genus Rickettsia, chloroplast matK gene sequences from the family Tillandsioideae, VLTF-1 genes from Penguin-pox virus, low-molecular weight glutenin subunit genes from tall wheatgrass, and mitochondrial genes from the hawkmoth genus Hyles.

Only protein-coding genes were considered, and the coding sequences were extracted from National Centre for Biotechnology information (NCBI). All the datasets comprised of experimentally validated, non-redundant sequences. For majority of the datasets, the phylogenetic relationships have been well established. Generic and species information were obtained from taxonomy database of NCBI.

2.2 Sequence Analysis

Multiple sequence alignment was performed using ClustalW [11] . The sequence data was used to determine distances using DNADIST program of the Phylogeny Inference Package (PHYLIP)[12]. The Jukes-Cantor distance parameter was selected for determining evolutionary distances. Each sequence for all the datasets was mapped onto a 12-dimensional fuzzy vector i.e., each sequence was represented in terms of its fuzzy polynucleotide space. Distance matrices were computed using the NTV distance metric for the same sequences.

2.3 Phylogenetic Analysis

Neighbour Joining (NJ) method, one of the most effective distance based methods, was used for phylogenetic tree construction. The distance matrices generated through DNADIST and NTV metric served as input for the NEIGHBOR program of PHYLIP. Bootstrap values were set to 1000 for all trees.

Table 1. Different strains of Dengue type 3 used in this study

Strain/Isolate	Country*	Identifier	Accession Number
BID V1015	VNM	DEN_VNM1	EU482459
BID V1017	VNM	DEN_VNM2	EU482461
PF92/2986	FRN	DEN_FRN1	AY744683
PF89/320219	FRN	DEN_FRN2	AY744678
PF89/27643	FRN	DEN_FRN3	AY744677
"ThD3_1283_98"	THN	DEN_THN1	AY676349
"C0360/94"	THN	DEN_THN2	AY923865
"ThD3_0104_93"	THN	DEN_THN3	AY676350
"ThD3_0055_93"	THN	DEN_THN4	AY676351
"C0331/94"	THN	DEN_THN5	AY876494
"BR DEN3 RO1-02"	BRZ	DEN_BRZ1	EF629370
“BR DEN3 290-02"	BRZ	DEN_BRZ2	EF629369
"BR DEN3 95-04"	BRZ	DEN_BRZ3	EF629366
"BR DEN3 97-04"	BRZ	DEN_BRZ4	EF629367
"BR74886/02"	BRZ	DEN_BRZ5	AY679147
DENV-3/VE/BID-V2484/2007	VZL	DEN_VZL1	FJ850111
DENV-3/VE/BID-V2480/2007	VZL	DEN_VZL3	FJ850109
DENV-3/VE/BID-V2455/2001	VZL	DEN_VZL4	FJ850098
DENV-3/VE/BID-V2452/2001	VZL	DEN_VZL5	FJ850097
DENV-3/KH/BID-V2089/2006	CBD	DEN_CBD1	FJ639729
DENV-3/KH/BID-V2088/2005	CBD	DEN_CBD2	FJ639728
DENV-3/KH/BID-V2086/2005	CBD	DEN_CBD3	FJ639727
DENV-3/KH/BID-V2083/2004	CBD	DEN_CBD4	FJ639726
DENV-3/KH/BID-V2081/2003	CBD	DEN_CBD5	FJ639724
DENV-3/US/BID-V2119/2002	USA	DEN_USA1	FJ547082
DENV-3/US/BID-V2118/2001	USA	DEN_USA2	FJ547081
BDH02-7	BAN	DEN_BAN1	AY496877
BDH02-4	BAN	DEN_BAN2	AY496874
BDH02-3	BAN	DEN_BAN3	AY496873

*Countries are abbreviated as follows: Vietnam=VNM, France=FRN, Thailand=THN, Brazil=BRZ, Cambodia=CBD, United Stated of America=USA, Bangladesh=BAN, Venezuela=VZL.

Table 2. : Members of the genus *Microbacterium* whose gyrB sequences were considered in this study

Species	**Accession no**
M.aerolatum	AM181475
M.arborescens	AM181476
M.aurantiacum	AM181477
M.aurum	AM181478
M.chocolatum	AM181479
M.dextranolyticum	AM181480
M.esteraromaticum	AM181481
M.flavescens	AM181482
M.foliorum	AM181483
M.hominis	AM181484
M.imperiale	AJ784798
M.keratanolyticum	AM181485
M.ketosireducens	AM181486
M.kitamiense	AM181487
M.lacticum	AM181488
M.liquefaciens	AM181489
M.laevaniformans	AM181490
M.luteolum	AM181491
M.maritypicum	AM181492
M.oxydans	AM181493
M.phyllosphaerae	AM181494
M.resistens	AM181495
M.saperdae	AM181496
M.schleiferi	AM181497
M.terregens	AM181498
M.testaceum	AM181499
M.thalassium	AM181500
Agromyces albus	AM181501

Table 3. : Members whose cyt B sequences were considered in this study

Taxa	Species	Identifier
Class Mammalia	*Loxodonta cyclotis*	African forest elephant (f.elephant)
	Loxodonta africana	African Savanna elephant(s.elephant)
	Cynopterus horsfieldi	Bat
	Equus caballus	Horse
	Rhinoceros unicornis	Rhino
	Cavia porcellus	Guinea Pig
	Myoxus glis	Fat Dormouse
	Delphinus delphis	Dolphin
	Kogia breviceps	Sperm Whale
	H.liberiensis	Hippo
	Bos taurus	Cow
	Cervus duvaucelii	Deer
Order Primates		
Hominoidea:Apes	*Homo sapiens*	Human
	Pan paniscus	Pygmy Chimp
	Gorilla gorilla gorilla	Western Gorilla
	Pongo pygmaeus	Orangutan
	Hylobates lar	Common Gibbon(Co-Gibbon)
	Hylobates gabriellae	Red-cheeked Gibbon
	Aotus lemurinus griseimembra	A.lemurinus
	Saimiri boliviensis boliviensis	S.boliviensis
	Eulemur fulvus albifrons	Lemur
Class Aves	*Buteo buteo*	Buzzard
	Phaethon rubricauda	Red-tailed tropic
Class Amphibia	*Xenopus laevis*	X.levis
	Bufo japonicus	Japanese toad
	Pelobates cultripes	Western spadefoot toad
Class Reptilia	*Typhlops reticulatus*	Worm Snake
	Naja naja	Cobra
Class Osteichthyes	*Parargyrops edita*	Parargyrops
	Tribolodon nakamurai	Bony Fish
Class Chondrichthyes	*Chimaera monstrosa*	Rabbit fish

3. RESULTS

For all the datasets, there was a marked difference in the tree topologies for the trees constructed using the Jukes-Cantor distance and the NTV metric. The trees generated employing the Jukes-Cantor distance conformed to the observed phylogenetic relationships for all the datasets, while the NTV metric based trees showed varying results. Figure 1(a) and 1(b) represent the trees generated for dataset 1, employing the Jukes Cantor distance and the NTV metric respectively. As can be observed, the NTV metric fails to show distinct clusters for all the viral isolates from different countries.

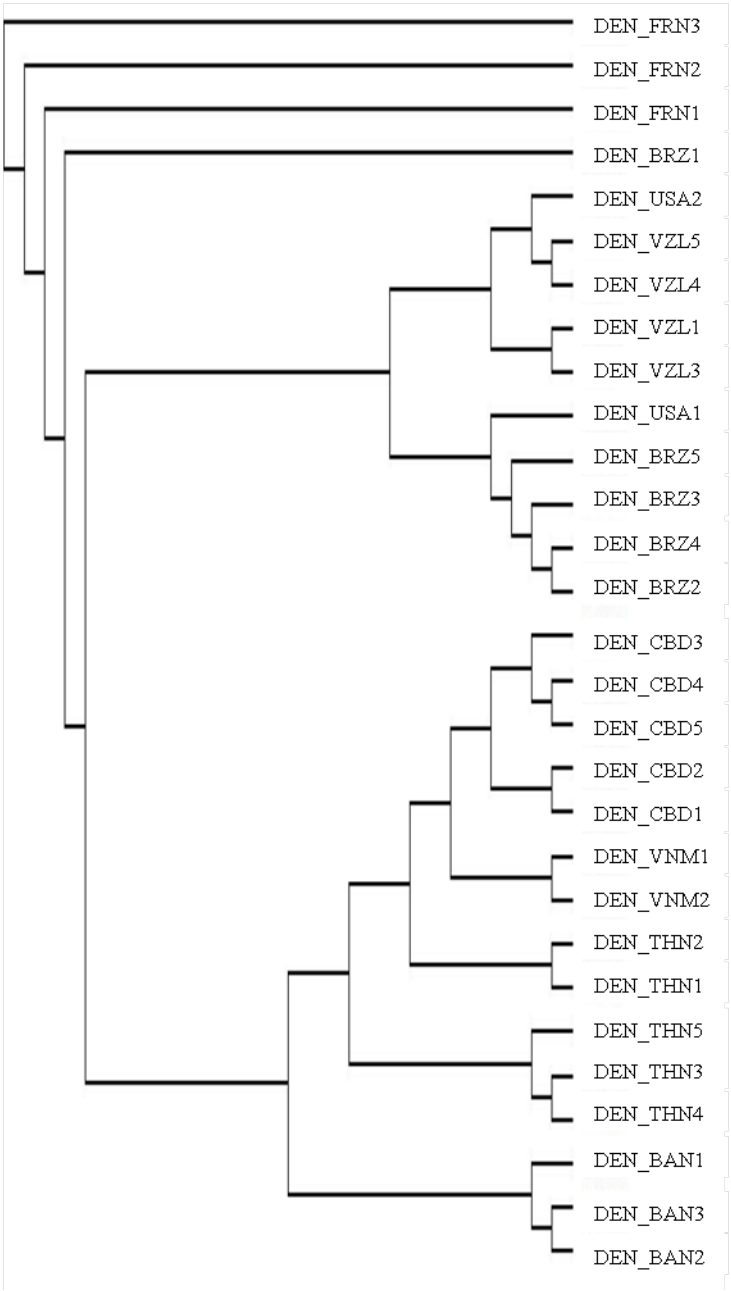

Fig. 1 (a). Tree constructed employing the Jukes-Cantor distance model in DNADIST using the NEIGHBOUR JOINING method for Dataset 1.

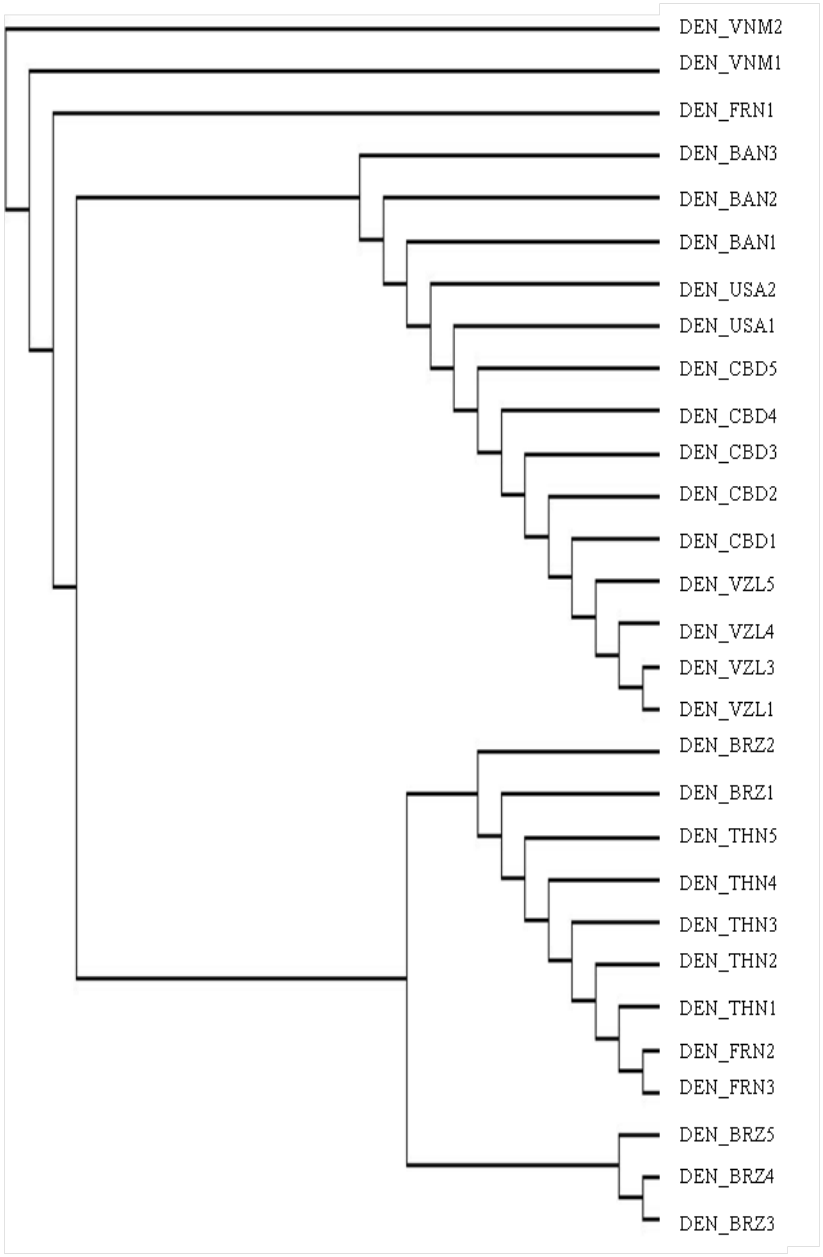

Fig. 1(b). Tree constructed employing the NTV distance metric and the NEIGHBOUR JOINING method for Dataset 1

Figures 2(a) and 2(b) reflect the difference in tree topologies for the trees generated using the Jukes Cantor distance and the NTV metric for Dataset 2. 2(a) conforms to established phylogeny o*f Microbacterium* [13] , however the NTV based phylogenetic tree shows starkly contrasting results, and does not agree with the known phylogenetic relationship of the family. For example, the NTV metric classifies *M.arborescens* with *M.aerolatum*, while it is known to be evolutionarily closer to *M. imperiale* instead, as reflected by the Jukes-Cantor distance in 2(a). The established phylogeny of the Microbacterium genus follows distinct clusters, while the NTV-based tree shows incorrect and fewer clades of taxa.

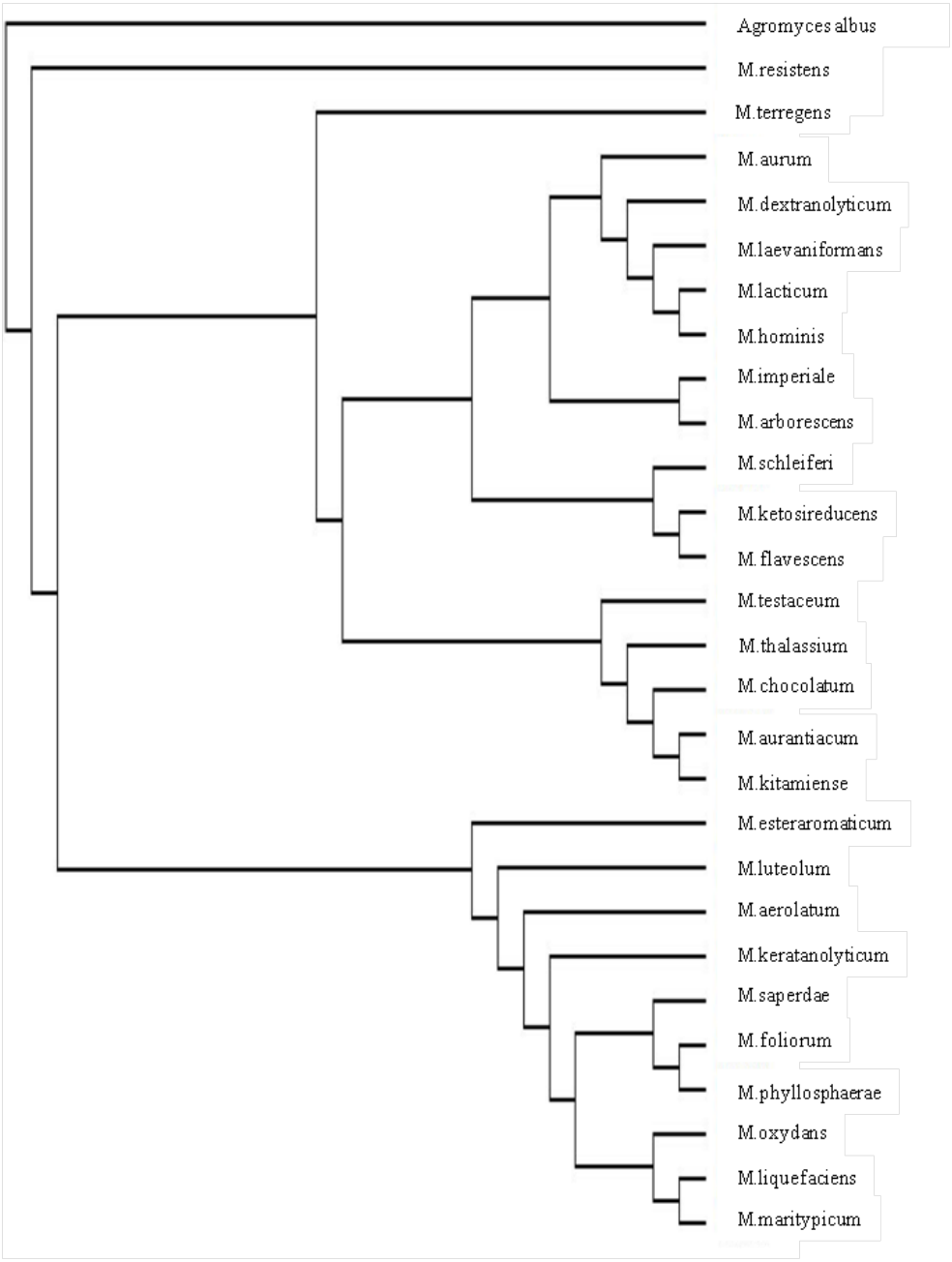

Fig. 2 (b). Tree constructed employing the Jukes-Cantor distance model in DNADIST using the NEIGHBOUR JOINING method for Dataset 2, using *Agromyces* as outgroup.

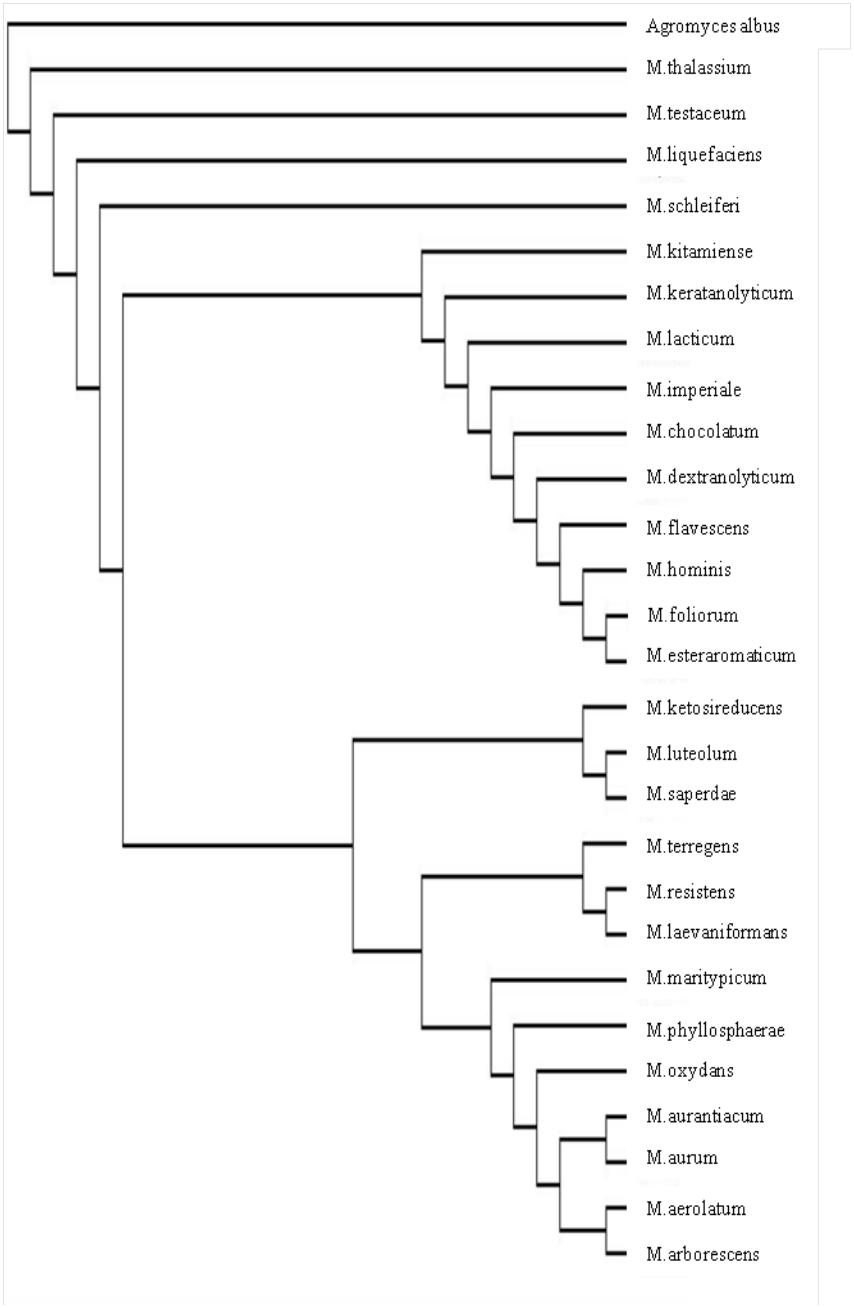

Fig. 2 (b). Tree constructed employing the NTV distance metric and NEIGHBOUR JOINING method for Dataset 2, using Agromyces as outgroup

Figures 3(a) and (b) similarly reflect the differences observed in the two methods of phylogenetic reconstruction for dataset 3. As can be observed, the NTV metric does not give distinct clusters for different members of the vertebrate classes. Also, it misclassifies African Savanna elephant with guinea pig, which otherwise belong to different classes. While the Gibbons are seen to cluster together in a single clade in both trees, the similarity between the generated trees is otherwise limited. These results are contrary to those seen in 3(a), and differ from the already proven taxonomic relationships in vertebrates. Further, cytochrome B sequences are highly conserved among eukaryotes, and are known to conform to different relationships among the data representatives. The NTV-metric could not correctly capture distances even among such well-documented similarities.

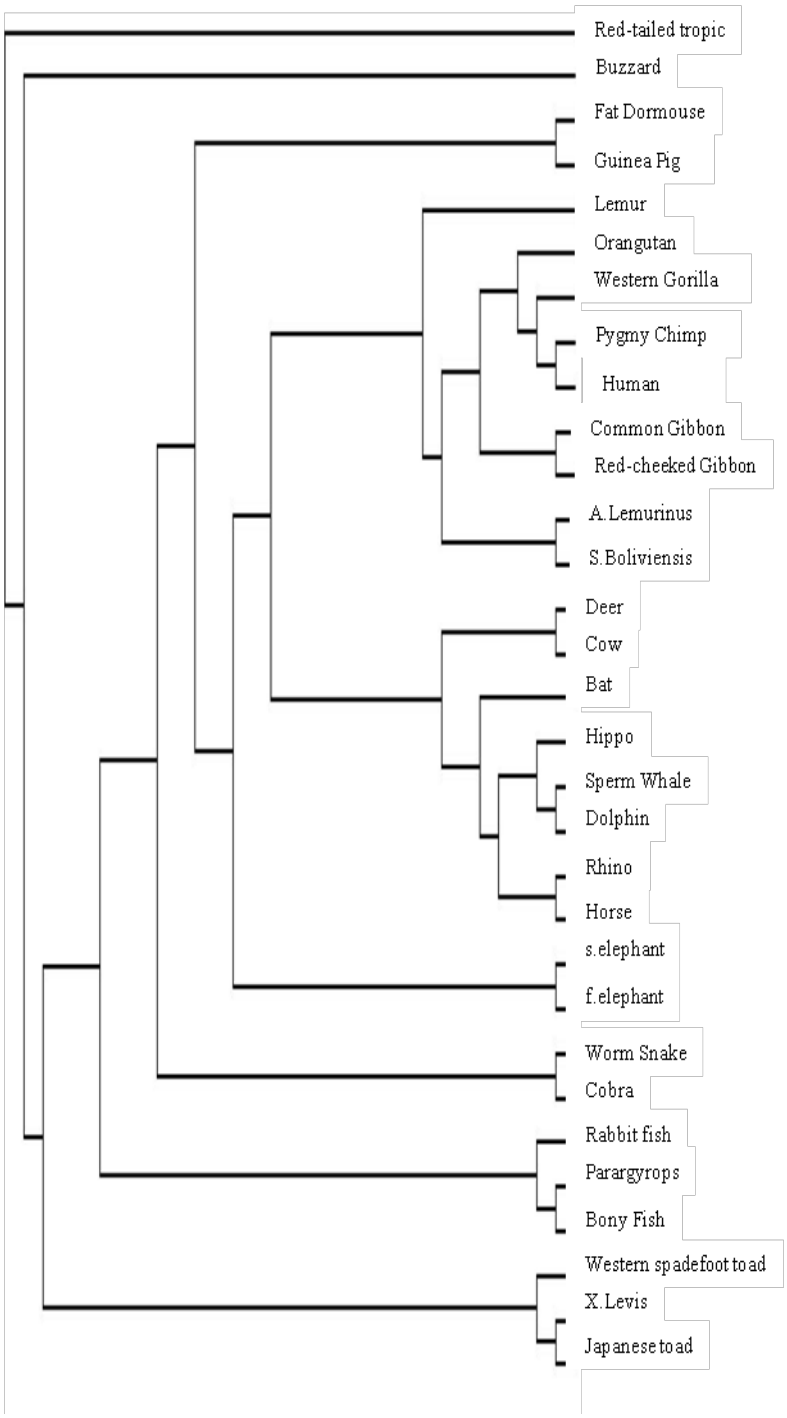

Fig. 3 (a). Tree constructed employing the Jukes-Cantor distance model in DNADIST using the NEIGHBOUR JOINING method for Dataset 3

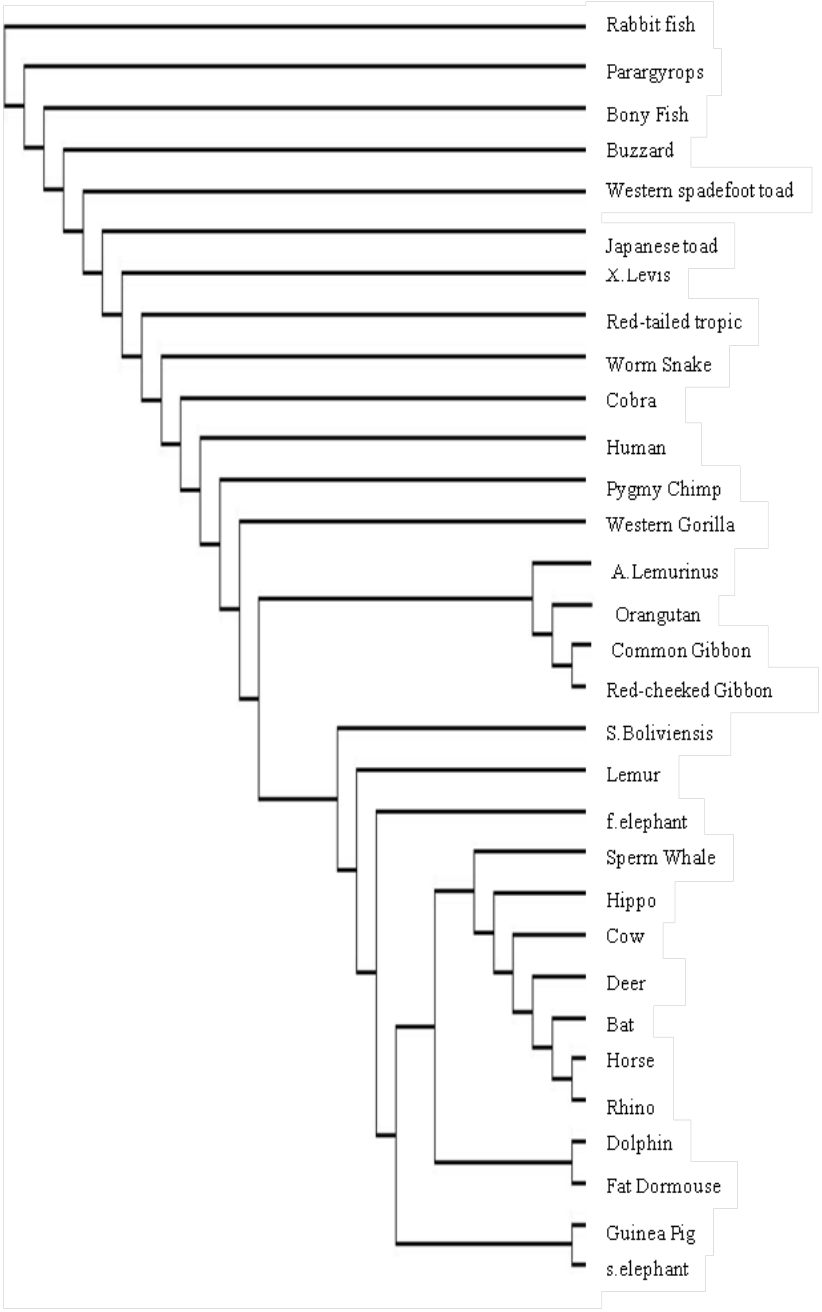

Fig. 3 (b). Tree constructed employing the NTV distance metric and the NEIGHBOUR JOINING method for Dataset 3.

The same variation in tree topologies was seen for the phylogenetic trees constructed using the other six datasets for the two methods. NTV-based tree showed diametrically opposite results to the expected tree, based on known relationships. Further, in all cases, the number of distinct clades are significantly lower in the NTV-based trees as opposed to clear clusters observed in the Jukes-Cantor method based trees. This illuminates the limitation of the metric in capturing evolutionary relationships among various taxa. Thus, for all the datasets, the NTV metric failed to correctly represent the phylogenetic relationships among organisms.

4. DISCUSSION

Some of the possible reasons for the failure of the fuzzy polynucleotide space in determining biological distances are as under:

Failure could be due to the observation that Fuzzy Polynucleotide Sequence is same for two different sequences, where one sequence is just a permutation of triplets of the other sequence, as suggested by K. Sadegh Zadeh [14]. The distance between these two sequences would be zero according to the NTV metric, whereas quite the opposite is true.

Another explanation for the limitation of the NTV metric in phylogeny is that phylogeny is a depiction of evolutionary distances between sequences, and takes into account per-site substitutions in a sequence alignment. The conventional distance based approaches used for phylogenetic construction employ distance parameters such as Jukes-Cantor distance, Kimura 2-point correction parameter etc. The Jukes-Cantor substitution model reflects the number of synonymous and non-synonymous substitutions per site of the alignment, and hence is a reflection of the number of changes occurred in DNA over the course of evolution. Since NTV is independent of sequence alignment, but rather depends on the relative base frequencies at each site of the codon, it does not account for evolutionary changes and hence is not an appropriate indicator of distances between biological sequences.

5. CONCLUDING REMARKS

The limited study infers that the fuzzy polynucleotide formalism may not be suitable in the construction of phylogenetic trees, as it is not a true indicator of distances among biological sequences.

6. ACKNOWLEDGEMENTS

The research reported in this sequel was carried out when the authors were with Bioinformatics Centre, Savitribai Phule Pune University, Pune India. The authors specially thank Dr. Urmila Kale for her constant encouragement.

7. REFERENCES

1. Ohlsson M , Schlosshauer M.: A novel approach to local reliability of sequence alignments. *Bioinformatics.* **18**(6):847–854 (2002)
2. Cord´on O, Gomide F, Herrera F, Hoffmann F, Magdalena L.: Ten years of genetic fuzzy systems: current framework and new trends. *Fuzzy Sets and Systems.* **141**(1):5–31(2004)
3. Belacel N, C˘ uperlovic´-Culf M, Laflamme M, Ouellette R.: Fuzzy J-Means and VNS methods for clustering genes frommicroarray data. *Bioinformatics.* **20**(11):1690–1701 (2004)
4. Bandyopadhyay S.: An efficient technique for superfamily classification of amino acid sequences: feature extraction, fuzzy clustering and prototype selection. *Fuzzy Sets and Systems.***152**(1):5–16 (2005)
5. Sadegh-Zadeh, K.: Fuzzy genomes. *Artif. Intell. Med*. **18**, 1–28 (2000)
6. Torres, A., Nieto, J.J.: The fuzzy polynucleotide space: Basic properties. *Bioinformatics* **19**(5), 587–592 (2003)
7. Kosko, B.: *Neural networks and fuzzy systems*. Prentice-Hall, Englewood Cliffs, NJ (1992)
8. Torres, A., Nieto, J.J.: Fuzzy Logic in Medicine and Bioinformatics. *Journal of Biomedicine and Biotechnology* **2006**,1-7 (2006)
9. D.N. Georgiou, T.E.Karakasidis, J.J.Nieto, A.Torres.: Use of fuzzy clustering technique and matrices to classify amino acids and its impact to Chou's pseudo amino acid composition. *Journal of Theoretical Biology*, **257**,17-26 (2009)
10. A. Torres, JJ Nieto.: Comments on ''The fuzzy polynucleotide space revisited'' by Kazem Sadegh-Zadeh, *Artificial Intelligence in Medicine* **41**, 81—82 (2007)
11. Larkin, Mark A., et al.: "Clustal W and Clustal X version 2.0." *Bioinformatics*23.21 **2007**: 2947-2948 (2007)
12. Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2005)
13. Richert K, Brambilla E, Stackebrandt E.: The phylogenetic significance of peptidoglycan types: molecular analysis of the genera Microbacterium and Aureobacterium based upon sequence comparison of gyrB, rpoB, recA and ppk and 16SrRNA genes. *Syst Appl Microbiol* **30**:102–108 (2006)
14. K. Sadegh-Zadeh.: The Fuzzy Polynucleotide Space Revisited. *Artificial Intelligence in Medicine* **41**,69-80 (2007)