Interval Entropy Profile Method Based on Maximum Entropy Principle for Estimating Evolutionary Information Content in Genetic Sequences for Their Comparison

Uddalak Mitra
Department of Computer & System Sciences
Visva Bharati University
Santiniketan, Pin:731235, India
Email: uddalakmitra@gmail.com

***Abstract*—Two classes of computational algorithms have been applied to solve the most fundamental problem in Bioinformatics-sequence comparison. Due to computational limitation in alignment methods, alignment-free methods become a predominant approach for devising sequence similarity. Of the alignment-free methods, return time distribution(RTD) based approaches are widely applied to answer various biological queries. However, existence of sequence noise becomes a principle bottle-neck to achieve precision in results for RTD approaches. Additionally, use of single length for k-mer strings often neglects evolutionary information that can be extracted using k-mers of other lengths. I inspect these flaws and propose a new method that considers multiple lengths of k-mer strings for generating features. Markovian dependency is assumed to estimate expected distribution for k-mer strings and to reduce sequence noise. Additionally, a criteria based on uniform independent model of sequence composition for selecting maximum k value is used.**
Experiments on benchmark datasets of 18s-rRNA sequence and whole genome mitochondrial(mtDNA) sequence of eutherians orders show superiority of the proposed method over all existing RTD and two start-of-the-art non-RTD methods.

I. Introduction

Return time distribution(RTD)[1], [2] is an alignment-free approach for sequence comparison. Due to it's simplicity when compared to multiple-sequence-alignment, methods based on RTD have been widely applied for virus classification, subtyping and for phylogeny reconstruction; some formulas for parameterizing RTD viz., standard deviation along with mean[3], [4], [5] and information entropy[6] have been proposed. However, the existing RTD methods suffer certain statistical problems, thereby underestimate the amount of evolutionary information contained in genetic sequences. In this paper, I inspect two of such problems and suggest corresponding solution to circumvent the situation.

Being an exact string matching approach RTD methods suffer from sequence noise due to effect of neutral mutation, thus reduction of noise is an essential step that is to be incorporated for parameterizing a distribution. I enrich the formula of entropy estimation[6] of RTDs, based on principle of maximum entropy, which can quantify nonrandom occurrence pattern of k-mer strings over the sequence. Throughout the paper I will use the term interval entropy to denote the parameter that represent information entropy of a RTD. More precisely, I derive the expected interval entropy of RTD for k-mer strings using principle of maximum entropy and the deviation of expected interval entropy from its observed entropy is used to quantify information entropy of RTD for a k-mer.

Another problem lays on the use of a single length k-mers. By using a fixed-length k-mer strings the methods usually neglect information content of the sequence that can be extracted using k-mers of other lengths, hence underestimates amount of information in the sequence. Thus incorporating multiple lengths for k-mer are likely to extract complete information of a sequence and I termed it as Interval Entropy Profile(IEP). Additionally, I suggest a practical approach under the assumption of a uniform independent model[7] to choose a maximum(optimal[8]) k-mer length to consider for estimating sequence information contributing to precise sequence comparison.

Experiments on natural sequence sets are used to test effectiveness of the proposed method over existing sequence comparison approaches. Use of standard benchmark dataset of 18s-rRNA and mitochondrial genomes(mtDNA) show superiority of the proposed method over all RTD methods(use of parameter standard deviation with mean[3] as well as information entropy[6]) and other two state-of-the-art non-RTD methods[9], [10] in terms of accuracy and biological consistency.

II. Materials and Method

A. Reiterate interval entropy

Define S as a DNA sequence composed of N repeated occurrences of the four nucleotides {A, C, G and T}. Let h(α^k) be the observed interval entropy of a k-mer α^k; where α^k is made by k successive repetition of $\alpha \in$ {A, C, G and T} and ($1 \leq k \leq N$) as defined in our initial work[6], let brief the steps of the method,

1) Computation of successive interval for a k-mer

$$Interval_i = \begin{cases} L_{i+1} - L_i & i < m \\ N - L_m + L_1 & i = m \end{cases} \quad (1)$$

L_i denotes starting location of a k-mer where as m indicates its frequency of occurrence over the sequence.

2) Constructing interval distribution

$$p_j = \frac{f(Interval_j) * Interval_j}{\sum_{i=1}^{m} Interval_i} \quad and \quad \sum_{j=1}^{\beta} p_j = 1 \quad (2)$$

β is the number of unique interval value, p_j denotes appearing probability of the particular interval value = $Interval_j$ and f($Interval_j$) denotes is frequency of occurrence.

3) Entropy of interval distribution

$$h = \sum_{j=1}^{\beta} p_j * log_2 \frac{1}{p_j} \quad (3)$$

4) Interval entropies of all possible k-mers of length k represent the feature vector as

$$H^k = \{h(\alpha_1^k), h(\alpha_2^k), \cdots, h(\alpha_i^k), h(\alpha_{i+1}^k), \cdots, h(\alpha_{4^k}^k)\} \quad (4)$$

where h(α_i^k) = h_i / $\sum_{i=1}^{4^k} h_i$, 1≤ i ≤ 4^k and h_i is the entropy of interval distribution of i^{th} k-mer that can be defined by Eq3. In what follows I derive the expected interval entropy for each h(α_i^k), let denote it by $\mathbf{E}(h(\alpha_i^k))$, and propose a criteria to choose maximum k value, k', use for constructing the composite feature vector, Interval Entropy Profile (IEP). I will denote IEP as $\{H^3, H^4, \cdots, H^{k'}\}$, each H^i denotes feature profile at k=i(by Eq 4). It is to be note that, as I a assuming (k-2) order Markovian dependency to derive expected interval entropy, hence it is require to exclude k-mer strings of length 1 and 2 in IEP.

B. Deriving expected interval entropy of a k-mer

Let denote the frequency of a (k-1)-mer α^{k-1} by f(α^{k-1}) where $\alpha \in$ {A, C, G and T}, its trivial to observe that the following relation holds,

$$\begin{aligned} &\mathbf{E}(\alpha^{k-1}A) + \mathbf{E}(\alpha^{k-1}C) + \mathbf{E}(\alpha^{k-1}G) + \mathbf{E}(\alpha^{k-1}T) = f(\alpha^{k-1}) \\ &\mathbf{E}(A\alpha^{k-1}) + \mathbf{E}(C\alpha^{k-1}) + \mathbf{E}(G\alpha^{k-1}) + \mathbf{E}(T\alpha^{k-1}) = f(\alpha^{k-1}) \end{aligned} \quad (5)$$

Clearly, there are four quantities in L.H.S of each equalities, representing possible right and left α expansion of the (k-1)-mer α^{k-1} while $\mathbf{E}(\alpha^{k-1}\alpha)$ and $\mathbf{E}(\alpha\alpha^{k-1})$ denotes expected frequency of the right and left α expansion of the (k-1)-mer respectively, $\alpha \in$ {A, C, G and T}. For interval entropy we can not definitely assume Eq5 holds directly, thus we need to derive the relation between observed interval entropy of a (k-1)-mer and its expansions of length k. If, $\mathbf{E}(h(\alpha_i^k))$ denotes expected interval entropy of the i^{th} k-mer α_i^k, then

$$\begin{aligned} &\mathbf{E}(h(\alpha_i^{k-1}A)) + \mathbf{E}(h(\alpha_i^{k-1}C)) + \\ &\quad \mathbf{E}(h(\alpha_i^{k-1}G)) + \mathbf{E}(h(\alpha_i^{k-1}T)) = h_R(\alpha_i^{k-1}) \\ &\mathbf{E}(h(A\alpha_i^{k-1})) + \mathbf{E}(h(C\alpha_i^{k-1})) + \\ &\quad \mathbf{E}(h(G\alpha_i^{k-1})) + \mathbf{E}(h(T\alpha_i^{k-1})) = h_L(\alpha_i^{k-1}) \end{aligned} \quad (6)$$

$h_R(\alpha_i^{k-1})$ and $h_L(\alpha_i^{k-1})$ are numerical constant that can be calculated from h(α_i^{k-1}). Clearly the system of equations do not have unique solution, because Eq6 is under-determined. Then for an estimated solution, we can select a combination of $\mathbf{E}(.)$ that maximizes their entropy -$\sum_{i=1}^{4^k} \mathbf{E}(h(\alpha_i^k)) * log_2(\mathbf{E}(h(\alpha_i^k)))$ under some constraints.

To select one of the may solutions we can employ maximum entropy principle and frame an optimization problem as:

$$\begin{aligned} &maximize \sum_{i=1}^{4^k} \mathbf{E}(h(\alpha_i^k)) * log(\mathbf{E}(h(\alpha_i^k)) \\ &subject\ to\ \mathbf{E}(h(\alpha_i^k))\ satisfying\ constraints \\ &\qquad\qquad\qquad\qquad in\ Eq6 \end{aligned} \quad (7)$$

To solve the optimization problem, we need to identify the unique set of constraints that are capable to decoupling a k-mer from other. Let, write the i^{th} k-mer α_i^k using its left and right expansions as $\alpha^L\alpha_i^{k-2}\alpha^R$, where α^L and α^R are 1-mer and α_i^{k-2} is the (k-2)-mer which appears in the i^{th} k-mer α_i^k at its index 2 to (k-1). Lets rewrite the set of constraints as

$$\begin{aligned} &\mathbf{E}(h(A\alpha_i^{k-2}A)) + \mathbf{E}(h(A\alpha_i^{k-2}C)) + \\ &\quad \mathbf{E}(h(A\alpha_i^{k-2}G)) + \mathbf{E}(h(A\alpha_i^{k-2}T)) = h_R(A\alpha_i^{k-2}) \\ &\mathbf{E}(h(A\alpha_i^{k-2}A)) + \mathbf{E}(h(C\alpha_i^{k-2}A)) + \\ &\quad \mathbf{E}(h(G\alpha_i^{k-2}A)) + \mathbf{E}(h(T\alpha_i^{k-2}A)) = h_L(\alpha_i^{k-2}A) \\ &\mathbf{E}(h(C\alpha_i^{k-2}A)) + \mathbf{E}(h(C\alpha_i^{k-2}C)) + \\ &\quad \mathbf{E}(h(C\alpha_i^{k-2}G)) + \mathbf{E}(h(C\alpha_i^{k-2}T)) = h_R(C\alpha_i^{k-2}) \\ &\mathbf{E}(h(A\alpha_i^{k-2}C)) + \mathbf{E}(h(C\alpha_i^{k-2}C)) + \\ &\quad \mathbf{E}(h(G\alpha_i^{k-2}C)) + \mathbf{E}(h(T\alpha_i^{k-2}C)) = h_L(\alpha_i^{k-2}C) \\ &\mathbf{E}(h(G\alpha_i^{k-2}A)) + \mathbf{E}(h(G\alpha_i^{k-2}C)) + \\ &\quad \mathbf{E}(h(G\alpha_i^{k-2}G)) + \mathbf{E}(h(G\alpha_i^{k-2}T)) = h_R(G\alpha_i^{k-2}) \\ &\mathbf{E}(h(A\alpha_i^{k-2}G)) + \mathbf{E}(h(C\alpha_i^{k-2}G)) + \\ &\quad \mathbf{E}(h(G\alpha_i^{k-2}G)) + \mathbf{E}(h(T\alpha_i^{k-2}G)) = h_L(\alpha_i^{k-2}G) \\ &\mathbf{E}(h(T\alpha_i^{k-2}A)) + \mathbf{E}(h(T\alpha_i^{k-2}C)) + \\ &\quad \mathbf{E}(h(T\alpha_i^{k-2}G)) + \mathbf{E}(h(T\alpha_i^{k-2}T)) = h_R(T\alpha_i^{k-2}) \\ &\mathbf{E}(h(A\alpha_i^{k-2}T)) + \mathbf{E}(h(C\alpha_i^{k-2}T)) + \\ &\quad \mathbf{E}(h(G\alpha_i^{k-2}T)) + \mathbf{E}(h(T\alpha_i^{k-2}T)) = h_L(\alpha_i^{k-2}T) \end{aligned} \quad (8)$$

$h_L(\alpha^L\alpha_i^{k-2})$ and $h_R(\alpha_i^{k-2}\alpha^R)$ can be computable from $h(\alpha_i^{k-1})$, α^L and α^R are 1-mer, hence we are assuming existence of (k-2) order Markovian dependency for a k-mer. Particularly, we are interested in deriving relation between $\mathbf{E}(h(\alpha^L\alpha_i^{k-2}\alpha^R))$ with $h_L(\alpha^L\alpha_i^{k-2})$ and $h_R(\alpha_i^{k-2}\alpha^R)$.

To simply the notations, let, $h_L(\alpha_i^{k-2}A)$ = γ_1, $h_L(\alpha_i^{k-2}C)$ = γ_2, $h_L(\alpha_i^{k-2}G)$ = γ_3, $h_L(\alpha_i^{k-2}T)$ = γ_4, $h_R(A\alpha_i^{k-2})$ =

λ_1, $h_R(C\alpha_i^{k-2}) = \lambda_2$, $h_R(G\alpha_i^{k-2}) = \lambda_3$ and $h_R(T\alpha_i^{k-2}) = \lambda_4$. Further, $\mathbf{E}(h(A\alpha_i^{k-2}A)) = h_{11}$, $\mathbf{E}(h(C\alpha_i^{k-2}A)) = h_{21}$, $\mathbf{E}(h(G\alpha_i^{k-2}A)) = h_{31}$ and $\mathbf{E}(h(T\alpha_i^{k-2}A)) = h_{41}$, thus $h_{11} + h_{21} + h_{31} + h_{41} = \gamma_1$, other constraints can similarly be simplified. Let the Lagrange function is defined as

$$F = \sum_{i,j=1}^{4} h_{ij} * log_2(h_{ij}) + \sum_{i=1}^{4} \kappa_i(\gamma_i - h_{i1} - h_{i2} - h_{i3} - h_{i4}) + \sum_{j=1}^{4} \zeta_j(\lambda_j - h_{1j} - h_{2j} - h_{3j} - h_{4j}) \tag{9}$$

κ_i and ζ_j are Lagrange multiplier for the equations that involve γ_i and λ_j. Taking $\delta F/\delta h_{ij}$ =0, we have $log_2 ij$ + 1 + κ_i + ζ_j =0. Finally we can have

$$h_{ij} = 2^{1+\kappa_i+\zeta_j} \tag{10}$$

Using Eq10 and 8 we have,

$$\begin{aligned} 2^{-(\kappa_i+1)}(2^{-\zeta_1} + 2^{-\zeta_2} + 2^{-\zeta_3} + 2^{-\zeta_4}) = \gamma_i \\ 2^{-(\zeta_j+1)}(2^{-\kappa_1} + 2^{-\kappa_2} + 2^{-\kappa_3} + 2^{-\kappa_4}) = \lambda_i \end{aligned} \tag{11}$$

Thus,

$$\begin{aligned} (2^{-\zeta_1} + 2^{-\zeta_2} + 2^{-\zeta_3} + 2^{-\zeta_4}) * \\ (2^{-\kappa_1} + 2^{-\kappa_2} + 2^{-\kappa_3} + 2^{-\kappa_4}) \\ = 2 * (\gamma_1 + \gamma_2 + \gamma_3 + \gamma_4) = 2 * \rho \end{aligned} \tag{12}$$

From Eq10 and 12 we have $2^{-(\kappa_i+\zeta_j)} = 2 * \gamma_i\lambda_j$ / ρ. Then, by Eq10 we have $h_{ij} = \gamma_i\lambda_j$ / ρ which can be used to normalize observed interval entropy, hence de-noising.

C. Choosing maximum k value

To determine maximum k value for a set of sequences, I constructed a combined sequence generated by concatenating all the sequences of a given sequence set. Let denote the combined sequence by S, that provides empirical nucleotide distribution for the sequence set, the representation vector H^k of the sequence S is then computed. Let for a random sequence(by random sequence I mean a sequence at which probability of appearing all the four bases are equal), of length same as with S, the representation vector is H'^k. The difference between these two representation vector can be obtained by Kullback-Leibler distance as

$$D(H^k, H'^k) = \sum_{i=1}^{4^k} H^k(i) * log_2(H^k(i)/H'^k(i)) \tag{13}$$

The distance is small if two distribution are close to each other, which indicates H^k does not contain rich evolutionary information, hence shows similarity with random sequence.

To construct IEP for a sequence, the ratio R = $\sum \frac{H^k}{H'^k}$ is taken for k = 3 to 10; Figures 1 and 2 shows R values against different k values for 18s-rRNA and mtDNA sequence set. The magnitude of R becomes close to 1 at k = 5 for 18s-rRNA sequence set, which indicates k values greater than 5 does not contains much evolutionary information and hence should be discarded from IEP. Similarly, for mtDNA sequence set, k values greater than 9 should be discarded.

III. Experiment

Experiments are conducted on a Laptop PC with Intel-Corei5 processor and 4 GB RAM. 64-bit version of Matlab-R2014a(8.3.0.532) software platform is used for all implementation. Built-in function 'seqneighjoin()', is used to construct phylogeny trees, which in turn implement Neighbor-Join[14] algorithm.

A. Datasets

The benchmark datasets used in this study are detailed in Table I and II.

TABLE I: 18s-rRNA sequence set

Sequence label	Accession No.	Sequence label	Accession No.
D1	AF173614	D21	AF173623
D2	AF173610	D22	AF173631
D3	AF173626	D23	AF173627
D4	AF173622	D24	AF173609
D5	AF173638	D25	AF173605
D6	AF173637	D26	AF115860
D7	AF173617	D27	X00686
D8	AF173630	D28	X82564
D9	AF173625	D29	K01593
D10	AF173628	D30	M11188
D11	AF173624	D31	V01270
D12	AF173613	D32	X06778
D13	AF173611	D33	K03432
D14	AF173612	D34	M10098
D15	AF173632	D35	U13369
D16	AF173618	D36	X03205
D17	AF173615	D37	X02995
D18	AF173616	D38	X04025
D19	AF173636	D39	AJ279506
D20	AF173619	-	-

TABLE II: mtDNA sequence set

Sequence label	Accession No.	Sequence label	Accession No
D1	V00662	D11	X72004
D2	D38116	D12	U20753
D3	D38113	D13	X61145
D4	D38114	D14	X72204
D5	D38115	D15	V00654
D6	X99256	D16	X14848
D7	Y18001	D17	V00711
D8	X79547	D18	Z29573
D9	Y07726	D19	Y10524
D10	X63726	D20	X83427

B. Case study 1: 18s-rRNA

Relationship among eukaryotic tetrapod species is a widely discussed area in phylogeny and evolution. An important research topic is whether birds are more closely related to reptiles or mammals. We use the dataset, described in Xia et al[11], as our eukaryotic dataset to study the phylogeny relationship among these species using their 18s-rRNA gene sequences for a scope to compare with biological findings.

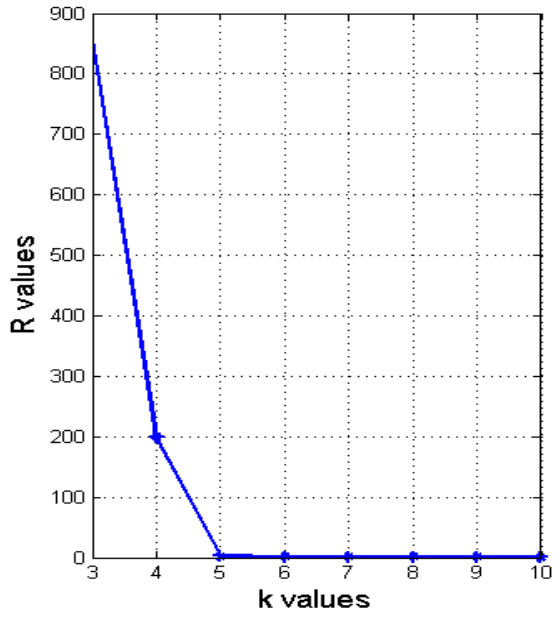

Fig. 1: R values at k = 1 to 10 for 18s-rRNA sequence set

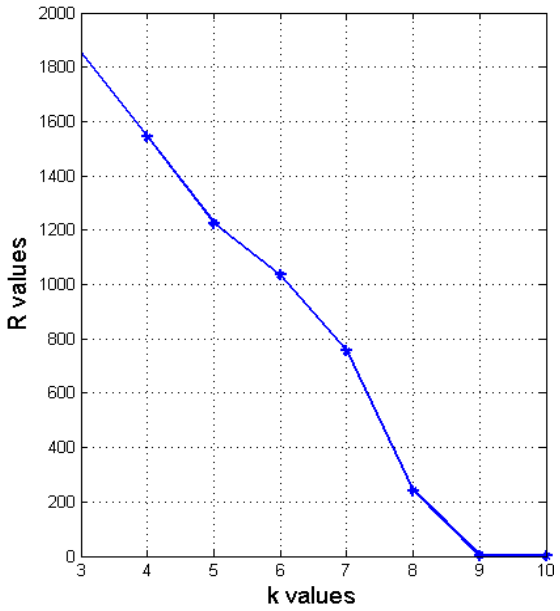

Fig. 2: R values at k = 1 to 10 for mtDNA sequence set

18s-rRNA genes are widely used in molecular analysis to reconstruct the evolutionary history of organisms, especially in vertebrates, as its slow evolutionary rate makes it suitable to reconstruct ancient divergences. The benchmark dataset of 18s-rRNA contains 24 species from Birds(sequence D1 to D24), 2 reptiles(sequence D25 and D26), 10 mammals(sequence D27 to D36) and 3 Amphibians(sequence D37 to D39). According to the benchmark report, birds are closely related with reptiles and mammals are excluded from their group. Additionally, Amphibians are placed as out-group species.

Phylogeny for the sequence set are drawn by different methods and shown in Figures 3 to 7. Both the RTD[3] method and FSWM[9] show some misclassification among the species; These methods are even unable to perform correct segregation of the four groups of the benchmark species; additionally benchmark groping is not consistent with the grouping achieved with these methods. However, Clustal Omega[10] and RTD with interval entropy[6] show clear segregation of species, although benchmark grouping is not achieved. Clustal Omega placed mammal with the groups of reptiles and amphibians, while birds are isolated from all these species. On the other hand RTD with interval entropy grouped birds with mammals, while reptiles and amphibians are grouped together. The proposed method IEP shows consistent results with benchmark groupings. Birds are grouped with reptiles, then mammals are grouped with their common clade, amphibians are completely isolated as out-group.

C. Case study 2: Mitochondrial genome set

The mitochondrial dataset consists of 7 Primates(D1-D7), 8 Ferungulates(D8-D15), 2 Rodents(D16-D17), and 3 out-groups(D18-D20) species. Previous reports [12], [13] show grouping of Primates with Ferungulates. Rodents species are excluded from the group of Primates and Ferungulates, based on analysis of whole genome mtDNA sequences. On the contrary, the NADH dehydrogenase 1(ND1) data[12] strongly suggest that Primates and Rodents clade exclude Ferungulates, which is in contradiction with the overall mtDNA evidence. Like the previous experiment I derive phylogeny trees for the sequence set using all the methods considered in this study. Figures 7 to 12 show the resulting phylogeny.

RTD method[3] shows highly mis-classified grouping of the species, on the other hand FSWM shows complete biological consistency as reported by the benchmark dataset. In case of Clustal Omega[10] Primates are grouped with Rodents, while Ferungulates are excluded from their clade, which is not supported by whole genome grouping. Similar misclassification are also sown with the RTD method using interval entropy[6]. The proposed method gropued Primates with Ferungulates and Rodents are exclude from their groups, as well as out-grouped species are isolated.

IV. Conclusion

Mutation occurs more or less in random manner over the nucleotide sequences and natural selection shapes the evolution. However, neutral mutation remains as sequence noise that introduces some randomness in genetic sequences. Thus reduction of such randomness is essential for estimating true evolutionary information in sequences. Further, its likely to miss evolutionary information through distribution of single length k-mers(words) because its almost unknown exactly what probability distribution or even any probabilistic model is used to construct a genetic sequence. Thus, application of multiple distribution is more powerful to extract evolutionary information or at least positive estimation. Hence, I integrate multiple word lengths to construct representation vector for

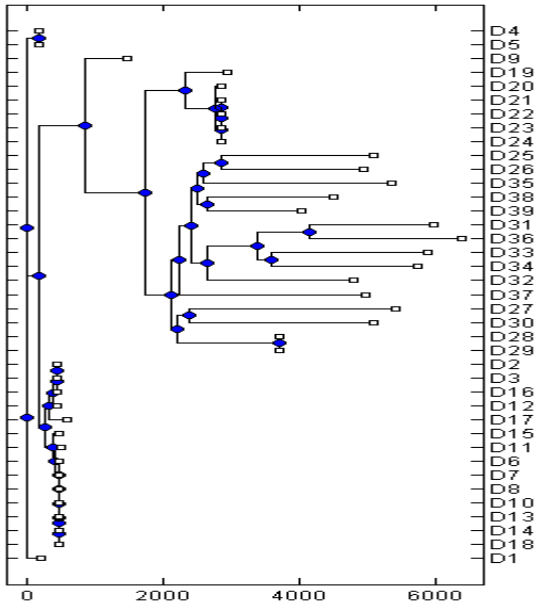

Fig. 3: 18s-rRNA phylogeny using RTD method with standard deviation and mean

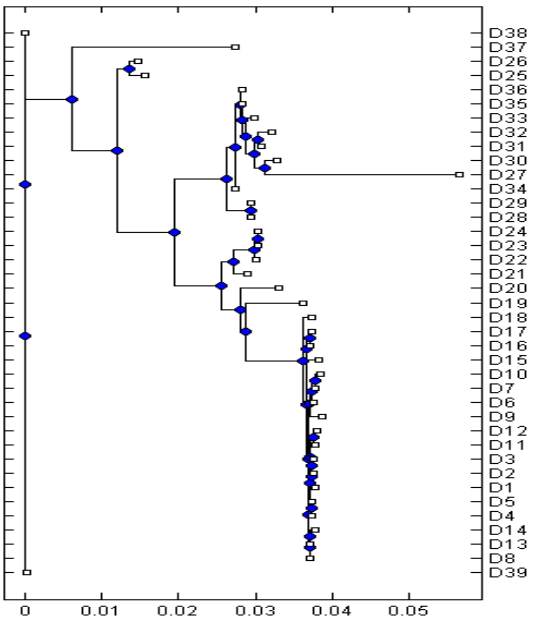

Fig. 4: 18s-rRNA phylogeny using FSWM

a sequence. Experiments are conducted to supports my point view. Results of the experiments show that noise reduction and use of multiple length of k-mer strings to construct features is capable to extracting more information compared to features without considering them and hence highly accurate results are obtained.

REFERENCES

[1] Nair ASS, Mahalakshmi T, Visualization of genomic data using inter-nucleotide distance signals, Proceedings of IEEE Genomic Signal Processing. Romania, 33, 2005.

[2] Vera Afreixo, Carlos A. C. Bastos, Armando J. Pinho, Sara P. Garcia, and Paulo J. S. G. Ferreira, Genome analysis with inter-nucleotide distances, Bioinformatics, 25, (3064 – 3070), 2009.

[3] Kolekar P, Kale M., Kulkarni-Kale U, Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping, Molecular Phylogenetics and Evolution, 65, (510 – 522), 2012.

[4] Kolekar P, Kale M, Kulkarni-Kale U, Genotyping of Mumps viruses based on SH gene: Development of a server using alignment-free and alignment-based methods, Immunome Res., 7, (1 – 7), 2011.

[5] Kolekar P, Vaishali P. Waman, Mohan M. Kale, Urmila Kulkarni-Kale, RV-Typer: A Web Server for Typing of Rhinoviruses Using Alignment-Free Approach, PLOS ONE, 11, 2016.

[6] Uddalak Mitra and Balaram Bhattacharyya, Alignment-Independent Sequence Analysis Based on Interval Distribution: Application to Subtyping and Classification of Viral Sequences, //doi.org/10.1007/978-981-10-3953-9_48,2016.

[7] Guoqing Lu, Shunpu Zhang and Xiang Fang, An improved string composition method for sequence comparison, BMC Bioinformatics, 9, (8329 – 8334), 2011.

[8] Guanghong Zuo, Qiang Li and Bailin Hao, On K-peptide length in composition vector phylogeny of prokaryotes, Computational Biology and Chemistry, 53, (166 – 173), 2014.

[9] C.A. Leimeister, S. Sohrabi-Jahromi, B. Morgenstern, Fast and Accurate Phylogeny Reconstruction using Filtered Spaced-Word Matches, Bioinformatics,33,971 – 979, 2017.

[10] Fabian Sievers, Andreas Wilm, David Dineen, Toby J Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, Hamish McWilliam, Michael Remmert, Johannes Soding, Julie D Thompson, Desmond G Higgins, Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega, Molecular Systems Biology, 539, 2011.

[11] XUHUA XIA, ZHENG XIE AND KARL M.KJER, 18S Ribosomal RNA and Tetrapod Phylogeny, Syst. Biol, 52, 283 – 295, 2003.

[12] Cao Y, Janke A, Waddell PJ, Westerman M, Takenaka O, Murata S, Okada N, Paabo S, Hasegawa M, Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders, J Mol Evol, 47, 307 – 322, 1998.

[13] Hasan H. Otu, Khalid Sayood, A new sequence distance measure for phylogenetic tree construction, Bioinformatics, 19, 2122 – 2130, 2003.

[14] Saitou N and Nei M, The neighbor-joining method: a new method for reconstructing phylogenetic trees, Molecular Biology and Evolution, 4,(406 – 425), 1987.

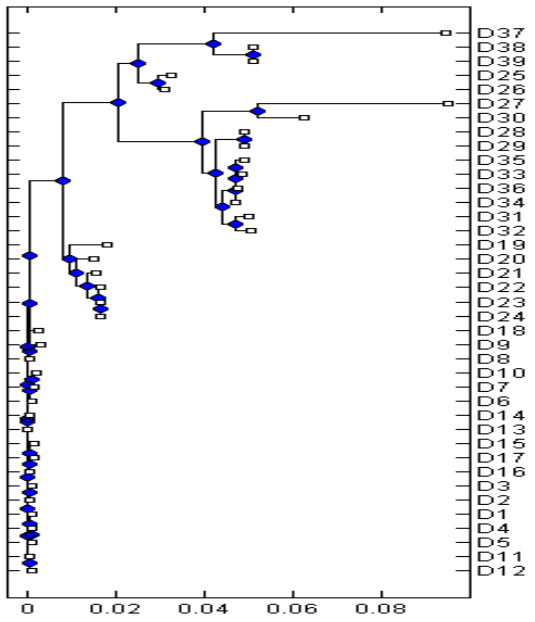

Fig. 5: 18s-rRNA phylogeny using Clustal Omega

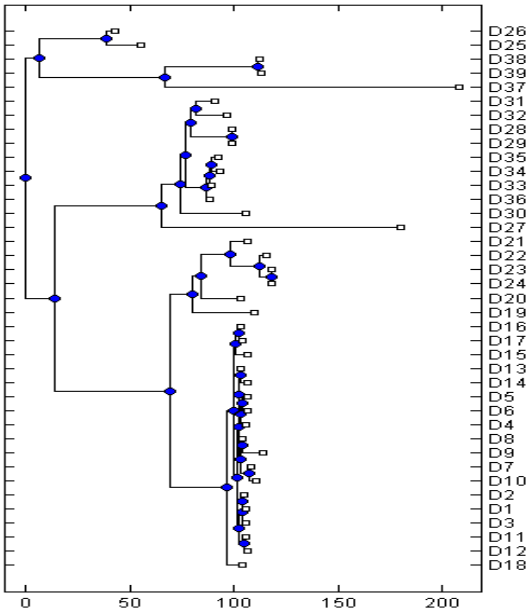

Fig. 6: 18s-rRNA phylogeny using RTD method with interval entropy

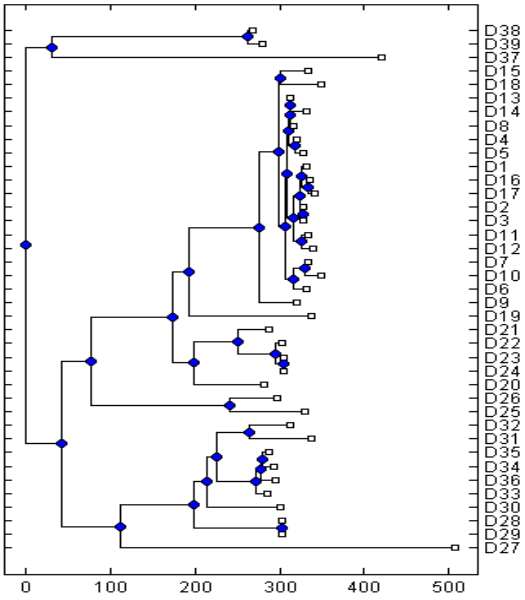

Fig. 7: 18s-rRNA phylogeny using proposed IEP method

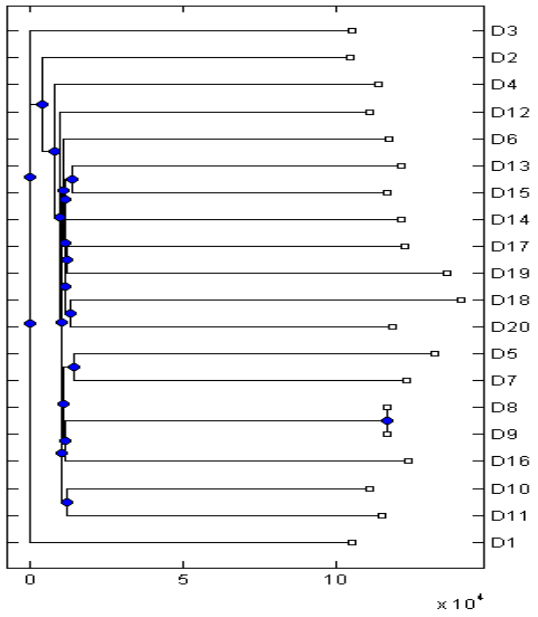

Fig. 8: mtDNA phylogeny using RTD method with standard deviation and mean

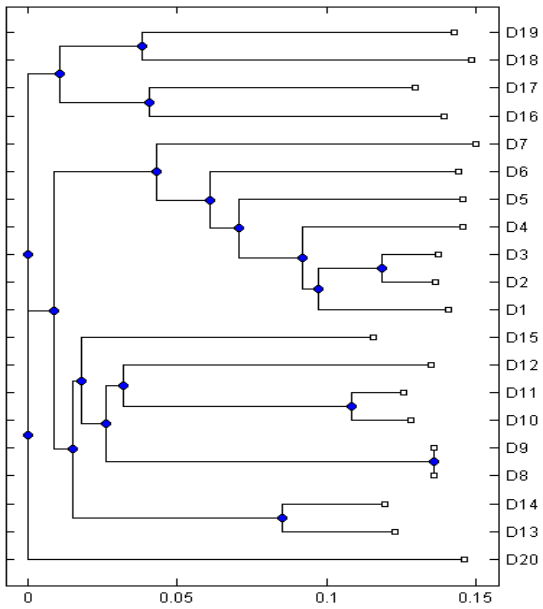

Fig. 9: mtDNA phylogeny using FSWM

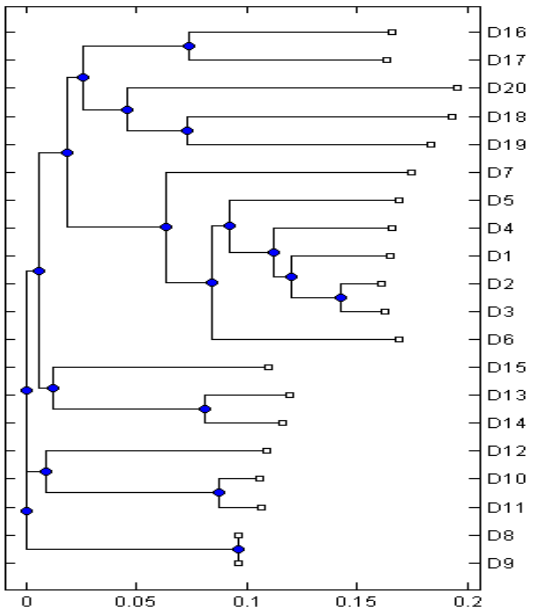

Fig. 10: mtDNA phylogeny using Clustal Omega

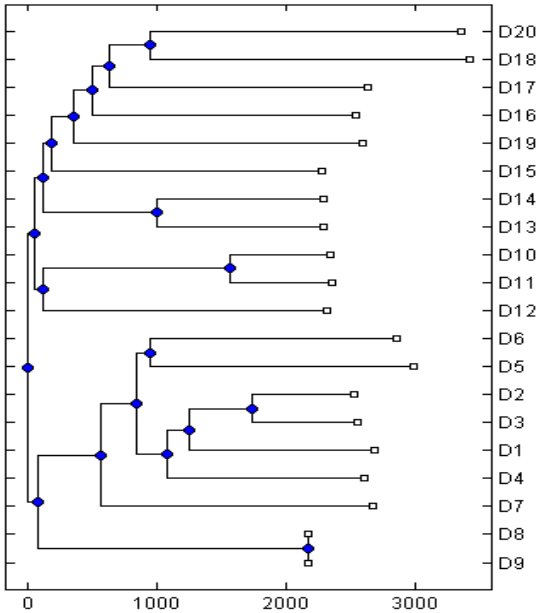

Fig. 11: mtDNA phylogeny using RTD method with interval entropy

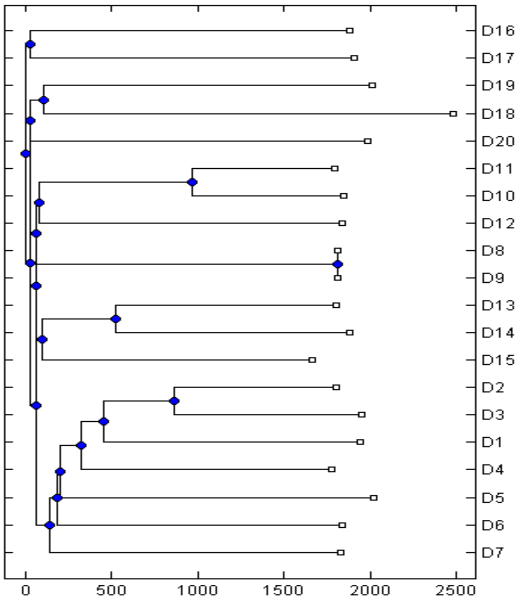

Fig. 12: mtDNA phylogeny using proposed IEP method