2015

Phylogenetic Tree Construction for Starfish and Primate Genomes via Alignment Free Methods

Ambujam Krishnan
Louisiana State University and Agricultural and Mechanical College

PHYLOGENETIC TREE CONSTRUCTION FOR STARFISH AND PRIMATE GENOMES VIA ALIGNMENT FREE METHODS

A Thesis

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Master of Science

in

The Division of Computer Science and Engineering

by
Ambujam Krishnan
M.Sc., Amrita Vishwa Vidhyapeetham, 2012
August 2015

To my husband, I couldn't have done this without you.

Thank you for all your support along the way.

Acknowledgments

I owe my sincere gratitude for the guidance and the help of several individuals, who believed in me, instilled in me the courage to move forward and lent a supportive shoulder in times of uncertainty. My sincere thanks to my advisor, Dr. Rahul Shah for giving me an opportunity to work with assistantship was a dream come true. Of utmost importance is the help by Dr. David W Foltz, both for his assitantship and advise whenever I needed it. His constant encouragement and patience is duly appreciated.

I would like to thank Dr. David W Foltz, and Dr. Jay Park their willingness to be in my thesis commitee and providing valuable feedback.

A special token of appreciation to my seniors Dr. Manish Patil, Dr. Ajay Panyala and Mr. Sudip Biswas for guiding me with coursework and helping with the implementations. This work would not have been possible without the discussions and help that they provided. This acknowledgment will be incomplete without mentioning my husband Dr. Sharma Thankachan for his support and constant encouragement. There are not enough words to phrase his help and guidance.

I am overwhelmed thinking about the love I have received from my friends Dr. Anand Nair, Dr. Nimesh Poddar, Ms. Catherine Poddar, Mr. Rony Thomas, Mr. Jesil James D'Silva, Mr. George Idicula and Ms. Neha Clare Jose. I pray that the Almighty God will bless every one of them beyond their imaginations. Finally, I would like to express gratitude to my family for their continued emotional support. On a different note, many people have been a part of my graduate education and I am highly grateful to all of them.

Table of Contents

Abstract

A phylogenetic tree is a tree like diagram showing the evolutionary relationship among various species based on their differences or similarity in their physical or genetic makeup.The similarity in their genetic makeup is traditionally measured based on pairwise distance between their gene sequences using sequence alignment methods. Due to the advancement in next generation sequencing technologies there is a huge amount of datasets available for partially or completely sequenced genomes. These massive datasets requires a faster comparison methods other than the traditional alignment-based approaches. Therefore, alignment free approaches are gaining popularity in recent years.

In this thesis, we compare alignment-based and various alignment free methods for phylogenetic tree construction. The alignment free methods we study are based on *k*-mer frequency, *Average Common Substring* (ACS) and ACS with position restrictions and mismatches. The position restricted ACS is a novel contribution of this thesis. To evaluate performance of the alignment free approaches we applied it to phylogeny reconstruction using DNA (27 primate mitochondrial genomes) and protein (Starfish RNA-seq) sequence sets. The phylogenetic trees are constructed using *Neighbor joining* to the distance matrices obtained with the above mentioned alignment-free methods. The resulting phylogenetic trees are then compared with the reference tree using *Branch Score Distance measure*. Both the *Neighbor joining* and the *Branch Score Distance Measure* are calculated by using the programs *neighbor* and *treedist* from the *PHYLIP package*.

Chapter 1
Introduction

A technological boom has occurred in the field of genomics since the first breakthrough research about human genome sequence. As the years passed, the amount of massive data available through next generation sequencing and other new technologies has led to an urgent need for developing new means to find the metadata of huge set of sequences [1]. These massive flows of data have raised many fundamental and challenging questions to the field of modern biology. One such question is the evolutionary history of different species.

A phylogenetic tree or an evolutionary tree is a tree like diagram showing the evolutionary relationship among various biological species based on their physical or genetic characteristics. These are mainly of two types *rooted tree* or an *unrooted tree*. In rooted tree each leaf represents each species and the edges represents the evolutionary time estimates.An unrooted tree doesn't illustrate the time estimate but instead gives an overview about the relatedness between the species.

The construction of phylogenetic tree from molecular data can be broadly classified into two categories *Exhaustive search* and *Step by Step method*. The former method is in which all wide range of possible trees are created and the best one under certain criterion is selected e.g.: *Maximum parsimony*, *Maximum likelihood* and *Fitch-Margoliash* (FM) methods , the latter method is the one in which local relationship is analyzed first and the best tree is build step by step ref e.g: *Neighbor-joining* (NJ) method and many other distance methods [2]. Many methods like the NJ and FM requires a distance matrix.

A distance matrix is evolved from the concept that species which looks similar (either physically or genetically) should be evolutionary more related i.e the number of mutation that needs to change one species to another for a related species should be very less. Thus in a distance matrix every row i and column j value represents the mutation distance from species i to species j. Earlier approach to the creation of distance matrix was through traditional alignment methods like the *Pairwise alignment* or the *Multiple sequence alignment*. These traditional approaches are becoming less popular due to it is computational complexity and less meaningfulness for whole genome comparisons [3]. This thesis revolves around the alignment free approaches, which can satisfy the need of the ever-increasing genome data.

In this thesis, we compare alignment-based and various alignment free methods for phylogenetic tree construction. The alignment free methods we study are based on k-mer frequency, Average Common Substring (ACS) and ACS with position restrictions. The position restricted ACS is a novel contribution of this thesis. To evaluate performance of the alignment free approaches we applied it to phylogeny reconstruction using DNA (27 primate mitochondrial genomes) and protein (Starfish RNA-seq) sequence sets. The phylogenetic trees are constructed using Neighbor joining to the distance matrices obtained with the above mentioned alignment-free methods. The resulting phylogenetic trees are then compared with the reference tree using Branch Score Distance measure. Both the Neighbor joining and the Branch Score Distance Measure are calculated by using the programs neighbor and treedist from the PHYLIP package.

1.1 Related Work

There exists various alignment-free methods in literature and can be broadly classified into the following four types:

1. k-mer/word frequency based methods
2. Methods based on substring
3. Information theory based method
4. Methods based on graphical representation.

Some of the popular methods in the first category are feature frequency profile (FFP) [4, 5], Composition vector (CV) [5, 6], Return time distribution (RTD) [7, 8], frequency chaos game representation (FCGR) [9] and Spaced Words [10]. In this thesis, we will be presenting a detailed experimental study using FFP method. Among methods based on substrings, the most popular method is the Average Common Substring (ACS) based method [11]. An extension of this method called Average Common Substring with k mismatches (ACS-k) is recently introduced and is shown to have better performance than the original ACS method [12]. However, authors were not able to prove an efficient algorithm for its computation, instead they gave a fast heuristic that approximate the distance. In a very recent paper, an efficient algorithm for computing ACS-k for any constant k is introduced. Another popular substring based method is called mutation distance [13]. Based on information theory, there are three methods known: Base base correlation [14, 15, 16], Information correlation and partial information correlation (IC-PIC) [17] and Lempel-Ziv compress [18]. The iterated

maps is another method, and is based on graphical representation [19]. For detailed description of this methods, we refer readers to the following review articles [20, 21, 22, 23].

1.2 The Datasets

To evaluate the performance of the alignment free approaches, we applied it to phylogeny reconstruction using DNA as well as Protein datasets.

Protein dataset is the Starfish RNA-Seq Dataset which is based upon work that was supported by the National Science Foundation under Award Nos. DEB 1036416 to *Dr. Daniel A. Janies* (Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte), DEB 1036358 to *Dr. David W. Foltz* (Department of Biological Sciences, Louisiana State University, Baton Rouge) and *Dr. Christopher L. Mah*, DEB 1036368 to *Dr. Gregory W. Rouse* (Marine Biology Research Division, University of California-San Diego, La Jolla) and DEB 1036366 to *Dr. Gregory A. Wray* (Department of Biology, Duke University, Durham, NC). DNA dataset is the standard 27 primate mitochondrial genomes from Genbank.

1.2.1 Starfish RNA-Seq Dataset

This dataset is the largest assembled dataset for the starfish phylogeny. It contains 8329 OrthoMCL clusters. OrthoMCL [24] is a software which provides method for building othologous groups across difference taxa, using a Markov Cluster algorithm to group otohrologs and parallels. Putative homologs from the Patiria pectinifera cDNA library was aligned locally using tblastn (e-score $<$ 1e-10) to each of the 8329 OrthoMCL clusters and added to the dataset. RNA-Seq data of four ophiuroid taxa (Ophiocoma wendtii, Ophiothrix spiculata, Astrophyton muricatum, and Ophioderma brevispinum) is also added to this data which is to be used as outgroups.This is done by matching OrthoMCL clusters. The starfish species and the order which belongs to is summarized in the table: 1.1.The sequences have a total length of 108 *kb*. Since it is a protein dataset, the character size is 20.

1.2.2 Primate Mitochondrial Genomes Dataset

The primate mitochondrial genomes were downloaded from GenBank (Nucleotide). The table 1.2 shows the Genbank Accession Id's of all primate genomes downloaded. These sequences have a total length of 446 *kb* and since they are nucleotide sequences their character set is 4.

TABLE 1.1. Starfish Species and their order

S.No	Species Name	Order
1	*Remaster*	Velatida
2	*Lophaster*	Velatida
3	*Peribolaster*	Velatida
4	*Pteraster*	Velatida
5	*Odinella*	Forcipulatida
6	*Pisaster*	Forcipulatida
7	*Labidiaster*	Forcipulatida
8	*P.pectinefra*	Valvatida
9	*Asteropsis*	Valvatida
10	*Porania*	Valvatida
11	*Echinaster*	Spinulosida
12	*Henricia*	Spinulosida
11	*Cheiraster*	Notomyotida
11	*Luidia*	Paxillosida
11	*Psilaster*	Paxillosida
11	*Astropecten*	Paxillosida
11	*Ophiothrix*	Ophiuroidea
11	*Ophioderma*	Ophiuroidea
11	*Astrophyton*	Ophiuroidea
11	*Ophiocoma*	Ophiuroidea

TABLE 1.2. 27 Primate Mitchondrial genomes with Acession ID.

S.No	Scientific Name	Common Name	Genbank Accession ID
1	*Cebus albifrons*	White-fronted capuchin	NC_002763.1
2	*Chlorocebus aethiops*	African green monkey	NC_007009.1
3	*Chlorocebus pygerythrus*	Green monkey	NC_009747.1
4	*Chlorocebus sabaeus*	Green monkey	NC_008066.1
5	*Colobus guereza*	Guereza	NC_006901.1
6	*Chlorocebus tantalus*	Green monkey	NC_009748.1
7	*Cynocephalus variegatus*	Sunda flying lemur	NC_004031.1
8	*Gorilla gorilla*	Western Gorilla	NC_001645.1
9	*Homo sapiens*	Human	NC_001807.4
10	*Hylobates lar*	Common gibbon	NC_002082.1
11	*Lemur catta*	Ring-tailed lemur	NC_004025.1
12	*Macaca mulatta*	Rhesus monkey	NC_005943.1
13	*Macaca sylvanus*	Barbary ape	NC_002764.1
14	*Nasalis larvatus*	Proboscis monkey	NC_008216.1
15	*Nycticebus coucang*	Slow loris	NC_002765.1
16	*Pan paniscus*	Pygmy chimpanzee	NC_001644.1
17	*Pan troglodytes*	Chimpanzee	NC_001643.1
18	*Papio hamadryas*	Hamadryas baboon	NC_001992.1
19	*Pongo pygmaeus*	Bornean orangutan	NC_001646.1
20	*Pongo pygmaeus abelii*	Sumatran orangutan	NC_002083.1
21	*Presbytis melalophos*	Mitred leaf monkey	NC_008217.1
22	*Piliocolobus badius*	Western red colobus	NC_008219.1
23	*Pygathrix nemaeus*	Douc langur	NC_008220.1
24	*Pygathrix roxellana*	Golden snub-nosed monkey	NC_008218.1
25	*Semnopithecus entellus*	Hanuman langur	NC_008215.1
26	*Tarsius bancanus*	Horsfield's tarsier	NC_002811.1
27	*Trachypithecus obscurus*	Dusky leaf monkey	NC_006900.1

Chapter 2
Sequence Analysis Methods

Phylogenetic tree construction heavily relies on sequence alignment as the tool to evaluate sequence relatedness to one another. In general if two sequences share high sequence similarity then it implies that they have a recent common ancestor down the evolutionary line than those with low sequence identity. In this section, we briefly describes the two main types of sequence analysis methods used in phylogenetics. The methods used in the current experimental study are explained in detail below.

2.1 Multiple Sequence Alignment

Multiple Sequence Alignment has its roots in pairwise alignment with the difference that more than two sequences can be aligned at a time.When a set of sequences of different species are given it will align all the sequences at one go. This method is highly used in phylogenetics as it has the ability to identify the conserved sequence regions which is an important piece of information when building an evolutionary tree.The main disadvantage in using Multiple Sequence Alignment is that it is computationally very heavy.

2.1.1 Aligning with *MUSCLE*

MUSCLE (Multiple Sequence Alignment with Log Expectation) [25] was founded by Robert C. Edgar in 2004. MUSCLE alogrithm works in three different stages, Progressive stage, Improved progressive stage and refinement stages. The detailed explanation of MUSCLE Algorithm is explained in Algorithm: 2. The main advantage of using MUSCLE Alignment is that this programs performs consistently better in terms of accuracy and speed when compared to other methods like MAFFT, T-COFFEE, CLUSTALW etc [26].

Algorithm 1 MUSCLE Algorithm

1: Calculate the *k*-mer distance $d_{A,B}$ of all pair (A, B) of sequences, where

$$d_{A,B} = \frac{\sum_{\tau} \min(n_A(\tau), n_B(\tau)}{\min(|A|, |B|) - k + 1}$$

Here τ is a k-mer and $n_A(\tau)$ (resp., $n_B(\tau)$) is the number of occurrences of τ in A (resp., B).

2: Using the distance matrix from step 1 create a guide tree using *Neighbor joining*.

3: Align the multiple sequences by progressive method using the guide tree from the previous step.

4: Repeat the above mentioned steps again to create a second-stage guide tree and then using this tree to make a new second multiple alignment.

5: Finally it refines the second multiple alignment and outputs it.

During the progressive alignment, the algorithm adds more gap to the sequences to match the number of characters in sequence [27]. A positive score is given to those that has a match and the score of the whole alignment is the sum of these individual scores. The MUSCLE algorithm is implemented in a very popular freely available software MEGA, Molecular Evolutionary Genetics Analysis [28]. We used MUSCLE implemented in MEGA version 6 [29] to create the reference tree.

2.2 Alignment Free Methods

The data from sequences is flowing at an exponential rate due to the advancement in technologies like the next-generation sequencing technologies. The traditional approaches on sequence alignment were global or local, pairwise or multiple sequence alignments. These traditional approaches are accurate when the sequences are close enough and can be reliably aligned but if the sequences are too divergent and when a reliable alignment cannot be obtained then these alignments lack accuracy. Another main disadvantage of these traditional approaches are their computational complexity. Thus alignment free approaches are gaining popularity as they provides another less computationally intensive alternative when compared to these methods. Alignment free methods can be classified into four main categories they are :

- *k*- mer or word frequency methods
- Substring based methods
- Information theory based methods
- graphical representation based methods

This thesis limits to the implementation and experiments relating to the *k*- mer method and Substring based method.

2.2.1 *k*-mer based methods

It is also known as word frequency or fixed length based methods. In this study, we use one of the k-mer based method called feature frequency profile (FFP) [4, 5]. Let $A[1 \dots |\mathsf{A}|]$ and $B[1 \dots |B|]$ be the two sequences to be compared. A k-mer is a string of length k. The number of distinct k-mers over an alphabet set Σ is $|\Sigma|^k$. Let S be the set of all distinct k-mers and $occ(P, T)$ be the number of occurrences of a k-mer P in a sequence T, then the cosine similarity and distance between A and B are defined as follows:

$$Similarity(A,B) = \frac{\sum_{P \in S} occ(P,A) \times occ(P,B)}{\sqrt{\sum_{P \in S}(occ(P,A))^2 \sum_{P \in S}(occ(P,B))^2}}$$

$$Distance(A,B) = 1/Similarity(A,B)$$

In order to compute the above measures, we created a generalized suffix tree (GST) using Succinct Data Structure Library [30] of A and B. Notice that each node in GST, whose string depth is $\geq k$, but its parents string depth is $< k$ corresponds to a unique k-mer in A or B. Therefore, we shall use the following algorithm to compute the similarity between A and B.

1. Initialize three variables $sumA, sumB$ and $sumAB$ to 0.

2. Visit the nodes in GST, and for each node whose string depth is $\geq k$, but its parents string depth is $< k$, find the number (say a) of leaves in its subtree that corresponds to suffixes of A and the number (say b) of leaves in its subtree that corresponds to suffixes of B. Then, update

$$sumA = sumA + a^2, sumB = sumB + b^2, sumAB = sumAB + ab$$

3. Return $Similarity(A,B) = sumAB/\sqrt{sumA \times sumB}$.

The run time for the execution of the algorithm is $O(n)$, where n is the total length of A and B.

2.2.2 Average Common Substring (ACS) Method

The ACS method is based on the fact that, two similar strings will have many common substrings. Let A_i be the suffix of A staring at position i. Similarly B_j be the suffix of B staring at position j. Also, let $LCP(A_i, B_j)$ be the longest common prefix of A_i and B_j and $|LCP(A_i, B_j)|$ be its length. Then,

$$ACS(A,B) = \frac{1}{|A|} \sum_{i=1}^{|A|} \max_j |LCP(A_i, B_j)|$$

The distance between A and B is given by [11],

$$Distance(A,B) = \frac{\log |B|}{ACS(A,B)} + \frac{\log |A|}{ACS(B,A)}$$

The above distance measure can also be computed in $O(n)$ time using GST. Notice that $\max_j |LCP(A_i, B_j)|$ can be computed in constant time as follows (i) find the leaf node ℓ in GST corresponding to A_i (ii) find the two closest leaves (say B_l and B_r) corresponds to the suffixes of B towards the left and the right side of ℓ. Then $\max_j |LCP(A_i, B_j)|$ is the maximum of $|LCP(A_i, B_l)|$ and $|LCP(A_i, B_r)|$. Therefore, $ACS(A, B)$,$ACS(B, A)$ and finally $Distance(A, B)$ can be computed in $O(n)$ time.

2.2.3 *ACS* with Position restriction

The ACS with position restriction is a novel contribution of this thesis. This is based on a hypothesis that, for two similar sequences A and B, the j that maximizes $|LCP(A_i, B_j)|$ is not too far from i for most of i's. The experiments on real data shows that this is the case for most of the sequences. Based on this, we define our distance measure as follows: for some parameter k, let $t = |A|/k$, partition A into $(k-1)$ overlapping substrings $A^1 = A[1, 2t], A^2 = A[1+t, 3t], A^3 = A[1+2t, 4t], \ldots, A^{k-1} = A[1+(k-2)t, |A|]$. Similarly, let $s = |B|/k$, partition B into $(k-1)$ overlapping substrings $B^1 = B[1, 2s], B^2 = B[1+s, 3s], B^3 = B[1+2s, 4s], \ldots, B^{k-1} = B[1+(k-2)s, |B|]$. Then,

$$Distance(A, B) = \sum_{x=1}^{k-1} Distance(A^x, B^x)$$

Here $Distance(A^x, B^x)$ is the ACS based distance between A^x and B^x. When sequences A and B are long, creating a GST might be hard. However, new methods allows to partition the input sequences into smaller sequences are process them in parts. Interestingly, our experimental results on chapter: 4 shows that for typical values of k, ACS with position restriction outperforms the original ACS measure in terms of quality.

Chapter 3
Tree Construction Methods

The phylogenetic tree construction methods are mainly divided into two types

- Algorithmic methods e.g.: *Neighbor Joining,UPGMA*
- Character-based methods e.g.: *Parsimony,Maximum Likelihood*

The main difference between two types is that the former uses a distance matrix which calculates the score of distances between pair of species and the later directly compares the each column or sites in the multiple sequence alignment. The scope of this thesis limits to the Algorithmic methods such as Neighbor joining and UPGMA.

3.1 Unweighted Pair-Group Method with Arithmetic Mean

UPGMA approach is based on clustering methods and is on the assumption that the tree is additive. Additive approach is likely to be incorrect because it defines that the distance of all species from root is the same. Due to this incorrect assumption UPGMA is rarely used in phylogenetic [27].The algorithm for UPGMA is as described below

Algorithm 2 UPGMA Algorithm

1: Find pair of taxa A and B with the lowest distance value x from the distance matrix
2: Draw a branch between A and B with branch length as x / 2.
3: Rewrite the distance matrix with A and B as one cluster AB. The distance between all other pair of taxa C and D remains the same in the new matrix. However, the distance between C and AB is taken as the average of the distance between C and A and the distance between C and B.
4: Redo the steps until number of entries in the distance matrix reduces to one.

Since it is rarely used in phylogenetics, this method is not used for comparisons but it is implemented as the method to visualize tree from distance matrix.

We implemented UPGMA in C++ that writes a tree in newick format [31] example:

((((((((((T.obscurus,(P.roxellan,N.larvatus)),P.melaloph), P.nemaeus),S.entellus), P.badius),C.guereza), ((P.hamadrya(M.sylvanus, M.mulatta)), (((C.tantalus,C.aethiops), C.pygeryth),C.sabaeus))), (((((P.troglody,P.paniscus), H.sapiens),G.gorilla), (P.pygmaeus,P.abelii)),H.lar)), C.albifron),(((T.bancanus,N.coucang),L.catta),C.variegat));

This newick formatted tree was then visualized using the software Dendroscope.

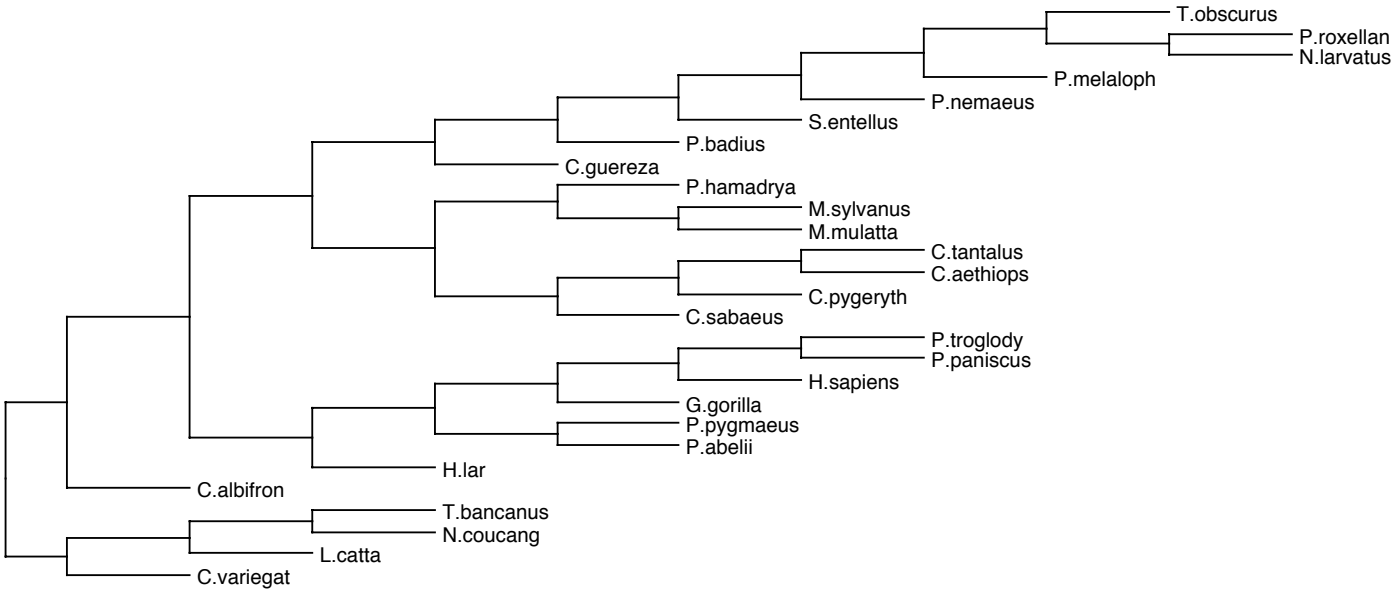


FIGURE 3.1. Tree constructed using UPGMA method

3.2 Neighbor joining

Neighbor joining is a bottom up tree construction method for constructing the phylogenetic trees. It uses the agglomerative clustering method. Neighbor join was created by Naruya Saitou and Masatoshi Nei in 1987 [32]. In this thesis we use neighbor join implemented in the PHYLIP package by Felsenstein/Kuhner lab, University of Washington [33].

Chapter 4
Experimental Methods and Results

4.1 Starfish Benchmark Tree

The Benchmark Tree was created using the following methods and softwares

1. The OrthoMCL clusters [24] were first aligned using MAFFT Version 7.0130 [34]. MAFFT is a multiple sequence alignment software that is much faster and accurate when compared to methods like CLUSTALW. The advantage of using MAFFT is that it gives a good tradeoff between accuracy and computational costs.

2. The Alignments were trimmed to remove the "ragged edges" using program TrimAL [35].TrimAL is used mainly for removing the spurious sequences or regions of poorly aligned. In this step the program is used to trim the left and right boundaries based on it gap statistics.

3. The output of the previous step was then again trimmed using TrimAL's - automated1 heuristic trimming method based on similarity statistics.

4. Phylogenetic trees were constructed for each alignment using RaXML with the following settings: algorithm= rapid/standard Bootstrap analysis number of alternative runs on distinct starting trees=100 amino acid substitution model=PROTCATJTT RAxML (Randomized Accelerated Maximum Likelihood) [36] is a program for creating phylogenetic tree from large datasets using maximum likelihood method.

5. To remove likely paralogous sequences from each locus, the program PhyloTreePruner [37] was used. Removal or paralogous sequences should be removed since orthologs that are result of speciation event are needed for species tree reconstruction than paralogs which are formed due to gene duplication events.

6. The resulting paralogy controlled sequences were again aligned and trimmed using the same software MAFFT and TrimAL.

7. The program FASconCAT [38] was used to concatenate the sequences generated from the previous step to a super matrix for further analysis.

8. To find the best evolutionary model for phylogenetic analysis, the matrix was input to Partitionfinder [39].

9. The best model for each locus from Partitionfinder and the supermatrix were then input to RAxML for maximum likelihood tree. The below showed is the final tree constructed from the above mentioned steps.

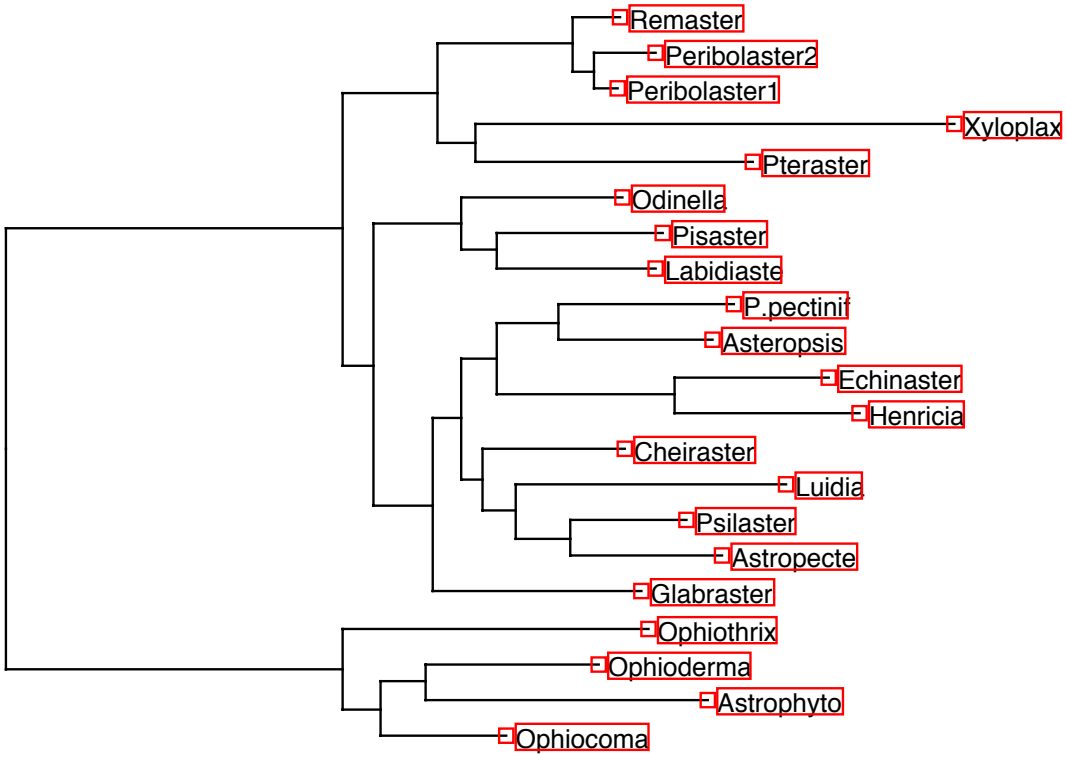


FIGURE 4.1. The Reference tree for Starfish genomes.

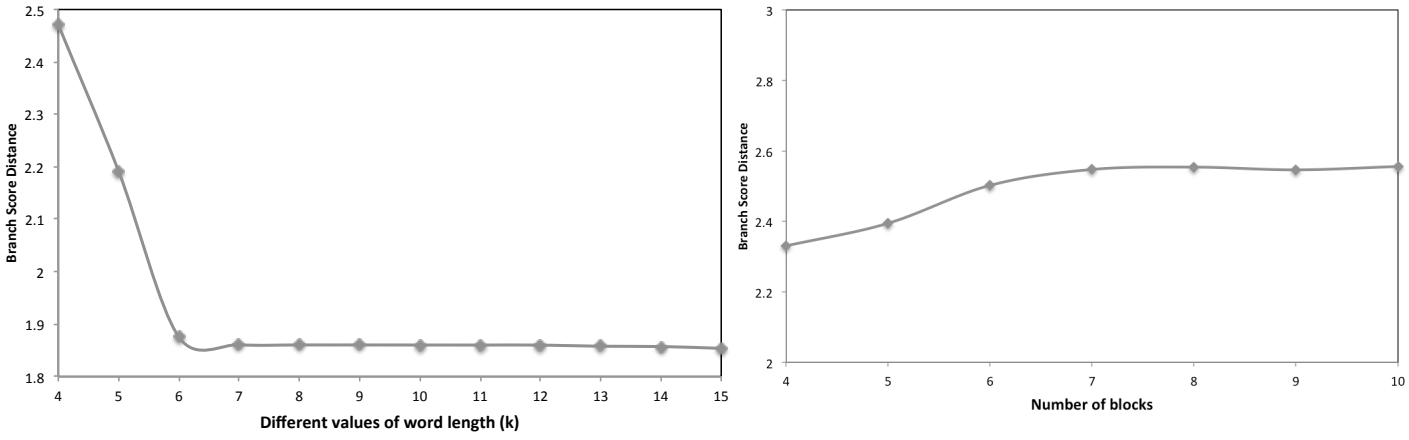
Performances of all the alignment free methods are compared using this tree as the benchmark. The creation of this benchmark tree was time consuming and it was done in high performance computing resources provided by LSU. The steps involving Partitionfinder and RAxML was computationally intensive. It was made to run on 4 nodes and 16 processors for each node. Total processing time took almost 72 hours.

4.2 Performance Evaluation of Starfish Tree with Alignment-free Methods

The different alignment free methods performed are *k*-mer, Average Common Substring and Average Common Substring with position restriction being the novel idea. The distance matrix constructed using the above said alignment free methods were then used to create tree using Neighbor joining method. The resulting trees are then compared to the benchmark tree using branch score distance measure. Both the Neighbor joining and the Branch score distances are calculated by using the programs neighbor and treedist from the PHYLIP package [33]. To visualize the differences between the trees tanglegram Algorithm from the Dendroscope software [40] was used.

The Branch Score Distance can be defined as the sum of squared errors of two tree which is calculated by the differences in tree topology and branch lengths. The major limitation of using this Branch Score Distance is that we cannot conclude any immediate interpretations such as larger distance is significantly larger than the smaller one .Even with the major limitation this distance is widely used because it is calculated by considering all the possible branches between two trees [33]. If the compared tree is more close to the reference tree it will have a lower Branch Score Distance. Therefore lower the score more close it is to the reference tree.

In starfish dataset, ACS and ACS with position restriction shows a slightly better tree than *k*-mer method when compared to the benchmark tree. The differences in the Alignment free method trees and the benchmark tree is explainable as starfish dataset is RNA-Seq dataset and the benchmark tree was constructed by finding best evolutionary model for each locus.



(a) *k*-mer method (b) ACS with Position restriction method

FIGURE 4.2. Graph showing Branch Score distance vs Alignment free methods on Starfish dataset

4.2.1 *k*-mer Method

As mentioned in 2.2.1, this method basically divides the input sequences into fixed length blocks/words of size *k* and then calculate the relative frequencies of such *k* length blocks. Experimental studies have found that *k* value is best when it ranges from 4-10. As with the same reason we have tried to fix the value of *k* in our study from 4 - 15.As shown in the figure 5.1a the Branch Score Distance is shown minimum of 1.860754 when the word length if fixed to a size of 7. The tree for which the branch score distance is minimum is shown below.

The tree obtained using *k*- mer method was very poor that it showed a star like phylogeny when drawn using phylogram as the branching diagram. *k*-mer method for this particular dataset was not able to capture the amount of inferred evolutionary changes (branch length). Thus to get a better estimate of the common ancestry a cladogram is used as the branching diagram.

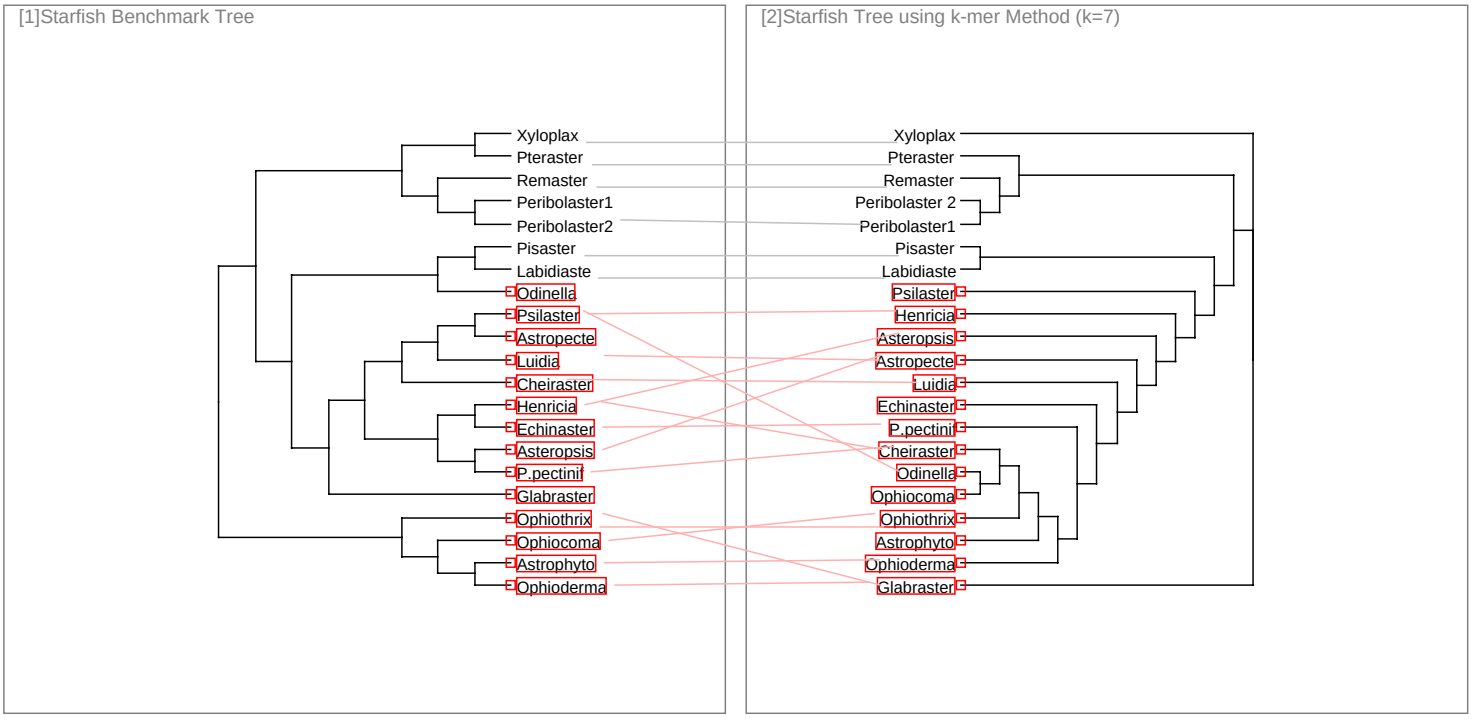


FIGURE 4.3. Comparison of Starfish Benchmark Tree with *k*-mer method shown as rectangular *cladogram*

The order Velatida which contains species *Remaster, Lophaster, Peribolaster* is retained in the tree using *k*- mer method when compared to the benchmark tree. Xyloplax, the only known genus within the

Concentricycloidea was within the Velatida in the original tree but it is not recovered as the same in the tree drawn using *k*- mer method. Another clade which was recovered is the Forcipulatacea order which contains *Pisaster, Labidiaster and Odinella* but Odinella which was sister to the Pisaster and Labidiaster was not recovered as the same using *k*-mer.

4.2.2 ACS Method

Average Common Substring method as descibed in 2.2.2 is calculated by word matching with variable length. This method is showing better result when compared to the *k*-mer method. The Branch Score distance of 2.31 is obtained. This value cannot be compared to the Branch Score Distance of the *k*-mer method because Branch Score Distance is defined as the measure with no immediate statistical interpretation that is one cannot say whether a larger distance is significantly larger than a smaller one.

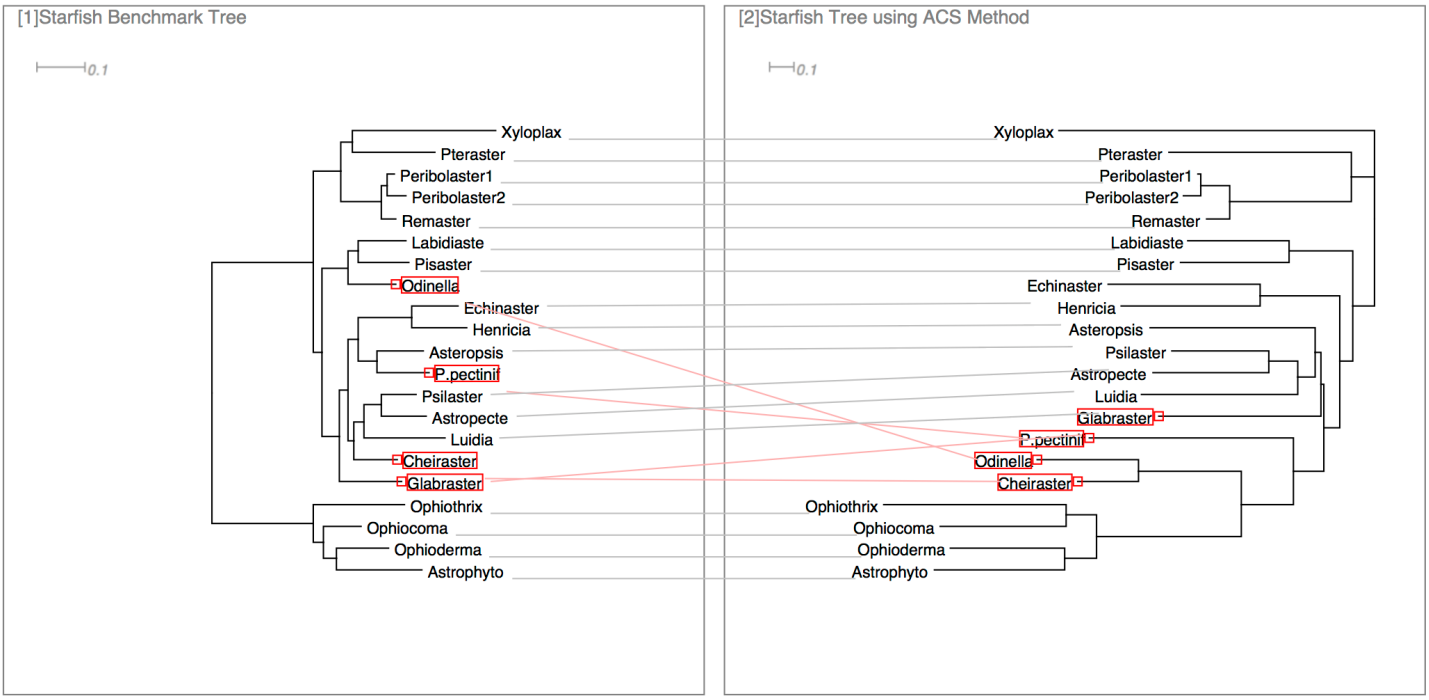


FIGURE 4.4. Comparison of Starfish Benchmark Tree with *ACS* method

The results of *k*-mer method was not good enough to be compared with the branch length, though the result of ACS method was comparable to the benchmark tree but the branch length was about 10 times the branch length of the original. Among the order Velatida, species Remaster, Lophaster and Peribolaster relationship was recovered. Most importantly the sister relationship of Pteraster to the other Velatidan was also seen in the tree using ACS method. Xyloplax which was placed within the Velatida in the original tree is seen basal to

the Velatidans in ACS method. From the Forcipulatacea order Pisaster and Labiadster relationship is reflected in the new tree but Odinella was seen out from the Forcipulatacea clade. Paxillosidans was recovered as the same from the benchmark tree in ACS. Ophiuroidea which was the closest sister taxon with the asteroids was completely misplaced in the tree with ACS method with a marked difference that the Ophioderma, Astrophyton relationship was the same with the benchmark tree.

4.2.3 ACS with position restriction

As mentioned in 2.2.3 the experiment is conducted with different values of blocks ranging from 4-15. The best value of Branch Score Distance of 2.33 was obtained at block of length 4. Eventhough Branch Score Distance cannot be used for immediate statistical interpretation, this value was comparable to the Branch Score Distance obtained using ACS method. A similar result to that of the ACS method was obtained in

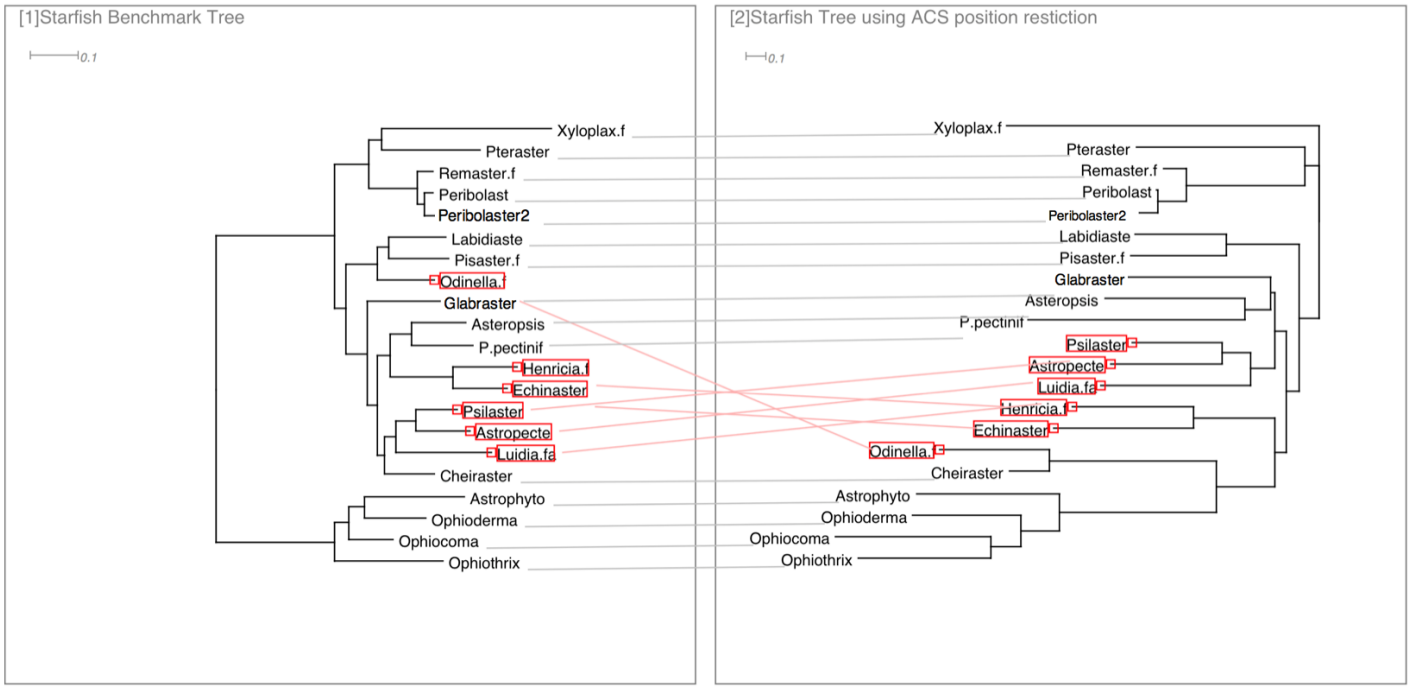


FIGURE 4.5. Comparison of Starfish Benchmark Tree with *ACS* position restriction method

terms of the branch length. The branch length of the tree obtained from ACS with position restriction is about 10 times to that of the original tree. The topological order of spinulosidans and paxillosidans was obtained as the same when compared to the benchmark tree. The valatidans shows a similar classification to the original tree with a notable difference that the porania is also classified with the same clade as Patria pectinefra and Asteropsis. In the Forcipulatacean order only Odinella was misplaced but the sister relationship of Pisaster

and Labidiaster was retained. Tree using ACS position restriction also potrayed the same behavior of xyloplax as with the other methods like *k*-mer and ACS that it is seen basal to the velatidans.

4.3 27 Primates Michondrial Benchmark Tree

The benchmark tree was created using MUSCLE algorithm as described in Algorithm 2. MUSCLE Alignment was selected as this program performs better in terms of accuracy and speed for large scale data like the primates data. Once the sequences were aligned it was then given as the input to construct the maximum likelihood tree. Both multiple sequence alignment as well as construction of the tree using maximum likelihood was done using the software MEGA 6.06.

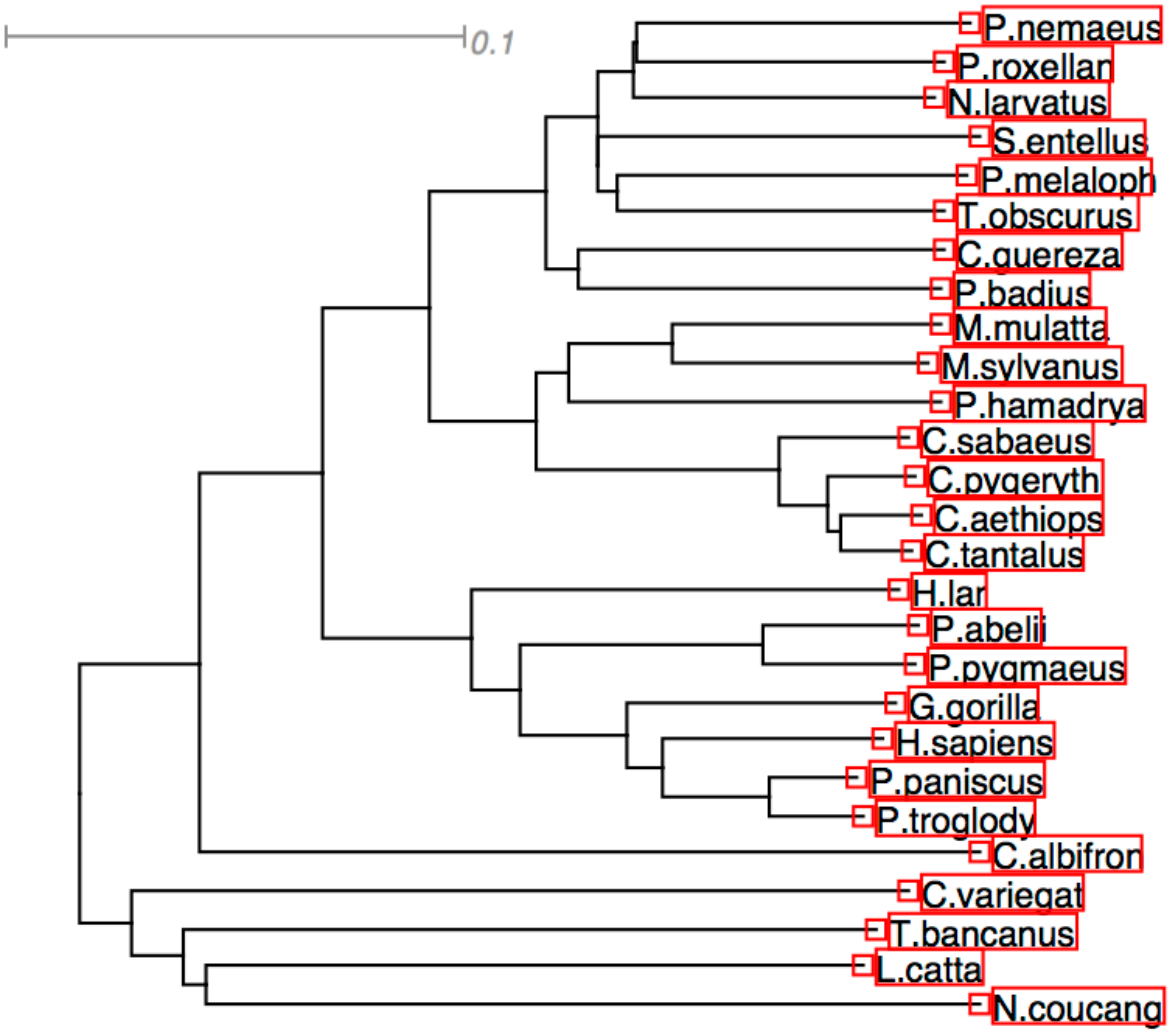
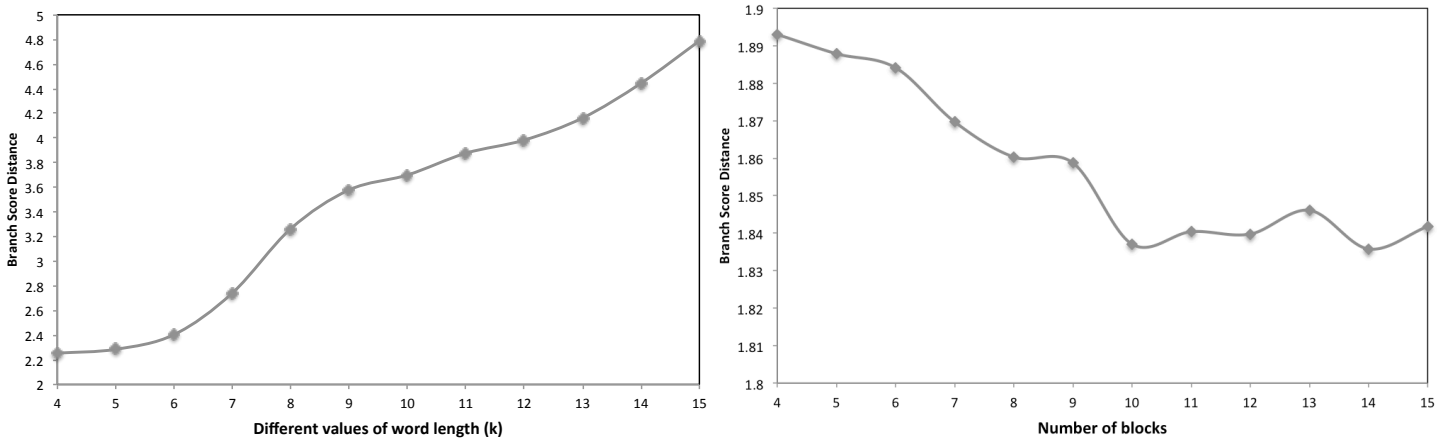


FIGURE 4.6. The Reference tree for 27 Primate mitochondrial genomes.

4.4 Performance Evaluation of primate genomes with Alignment-free Methods

The methodology followed for performance evaluation of primate genomes is the same as that for the Starfish genomes. As with the starfish dataset even the primates dataset shows that ACS and ACS with position restriction is slightly better tree than *k*-mer based method.



(a) *k*-mer method (b) ACS with position restriction method

FIGURE 4.7. Graph showing Branch Score distance vs *k*-mer and ACS with position on primates dataset

4.4.1 k-mer Method

The methodology followed for *k*-mer method is exactly the same done for the starfish dataset. The word length value was fixed from 4 - 15. The lowest branch score distance 2.285 was obtained for a word length of 5.

The tree obtained using *k*- mer method was very poor just like in starfish dataset that it showed a star like phylogeny when drawn using phylogram as the branching diagram. Thus to get a better estimate of the common ancestry a calodram is used as the branching diagram.

k-mer method used in these experiments were the cosine *k*-mer distance method. This method might be inefficient to capture the branch length information for larger datasets like the primates and the starfish datasets. 12 out of 27 species were misplaced in the *k*-mer method. Most of the wrongly classified species belonged to the Cercopithecidae (old world monkey family) especially subfamily Colobinae such as the Langur (leaf monkey) group (S. entellus, T.obscurus, P.melaloph), Odd-nosed group (P.roxellan,P.nemacues,N.larvatus), African group (P.badius, T.banacanus) etc. The ape clade which consists of orangutan (P.pygmaeus), gibbon (H.lar), chimpanzee (P.trogodytes), human(H.sapiens) and gorilla(G.gorilla) was correctly classified in

the *k*-mer method. Papio hamadyras in the clade has also correctly clustered with the macaques. These misclassification can be due to relatively small amount of sequence information provided by the mitochondrial genomes *(roughly 16.5 kb)*.

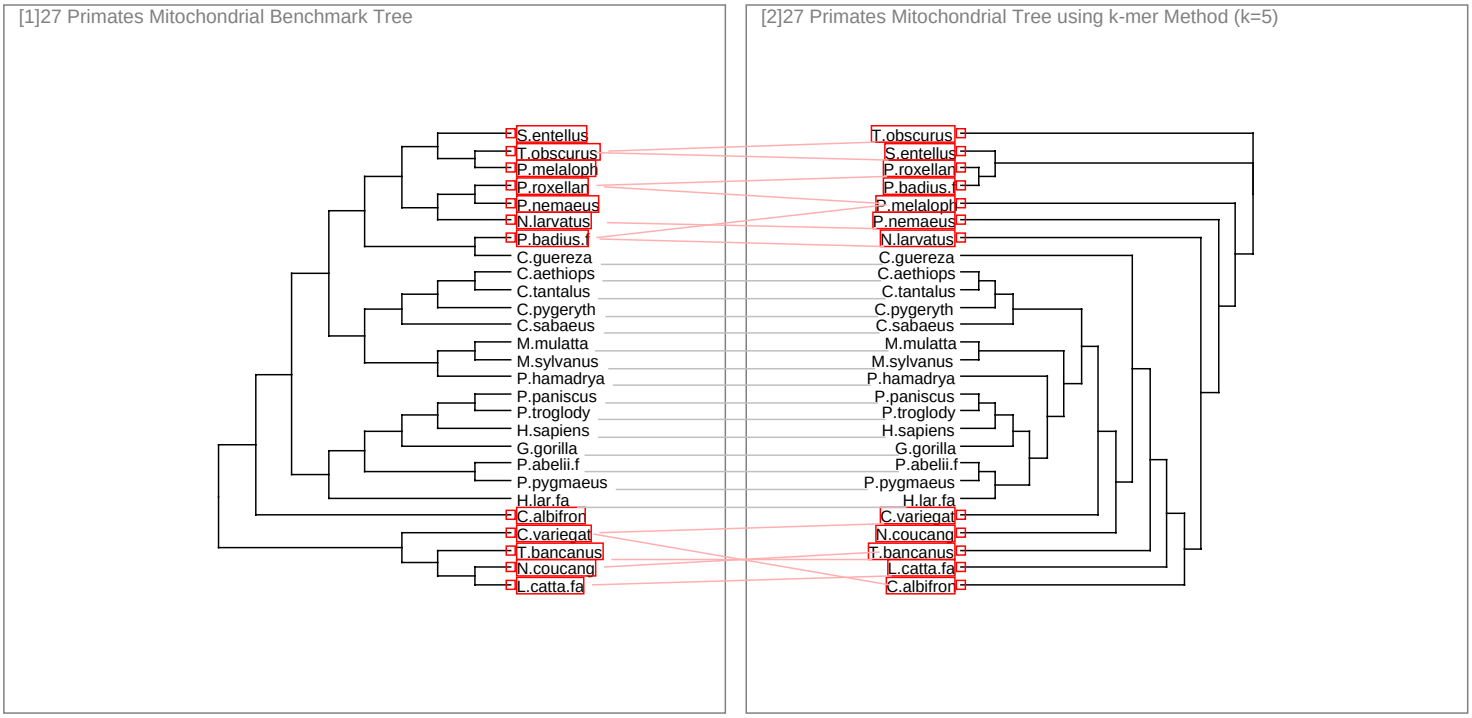


FIGURE 4.8. Comparison of 27 Primates Benchmark Tree with *k*-mer method.

4.4.2 ACS Method

ACS method, like the starfish dataset result, is showing a equal result with the *k*-mer method except with the difference that the branch lengths are comparable and branch score distance of 1.811 is obtained.This distance is lower than the branch score distance obtained by *k*-mer method which is 2.285. Most of the wrongly classified clades were similar to the tree obtained using *k*-mer method. As seen in *k*-mer result, Papio hamadyras species is correctly placed with the macaques i.e Rhesus macaque (Macaca mulatta) and Barbary macaque (M. sylvanus). The wrongly clustered groups include Colobinae such as the Langur (leaf monkey) group (S. entellus, T.obscurus, P.melaloph), Odd-nosed group (P.roxellan, P.nemacues, N.larvatus), African group (P.badius, T.banacanus). The ape clade Orangutan (P.pygmaeus), gibbon (H.lar), chimpanzee (P.trogodytes), human (H.sapiens) and gorilla (G.gorilla) is clustered correctly like in the *k*-mer method. The

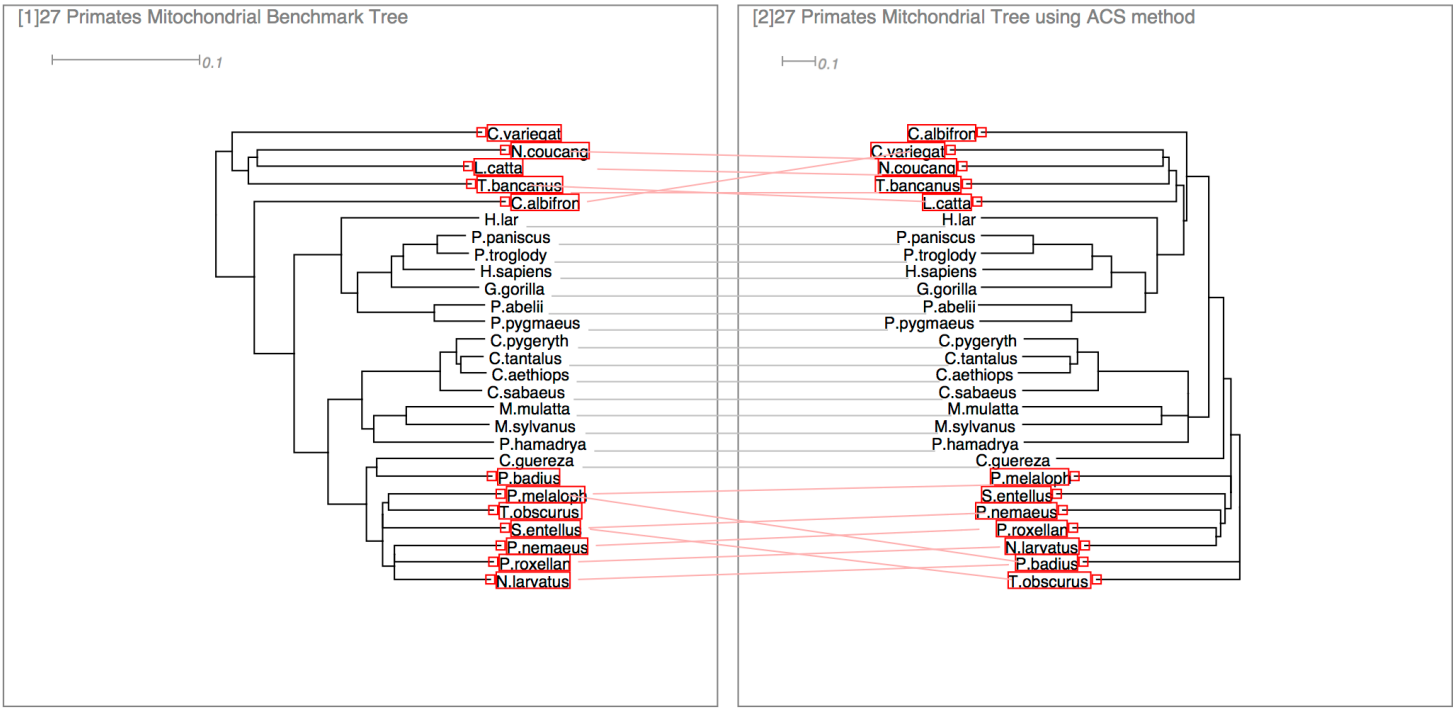


FIGURE 4.9. Comparison of 27 Primates Benchmark Tree with *ACS*-mer method.

Old World monkeys which belongs to the tribe Cercopithecini and genus Chlorocebus like species C.tantalus, C.aethiops, C.pygerythrus, C.sabaeus are consistently clustered correctly in both *k*-mer and ACS method.

4.4.3 ACS with position restriction

ACS with position resctrion is showing a slightly better result in terms of topology and branch length when compared to the methods like *k*-mer and ACS. The branch score distance of 1.83 is obtained at block length of 10.

The consistently misplaced clades which belong to Colobinae family were still clustered wrongly with the ACS with position except with the Odd-nosed group (P.roxellan, N.larvatus) and African group P.badius and Chlorocebus C.guereza.The ape clade i.e Orangutan (P.pygmaeus) ,gibbon (H.lar), chimpanzee (P.trogodytes), human (H.sapiens) and gorilla (G.gorilla) and most of the old world monkeys C.tantalus, C.aethiops, C.pygerythrus, C.sabaeus are correctly placed in tree constructed using ACS with position restriction.

The result with ACS with position restriction which is the novel idea of this is showing a slightly better when compared to methods like *k*-mer and ACS in recovering the branch length as well as the topological relationships.

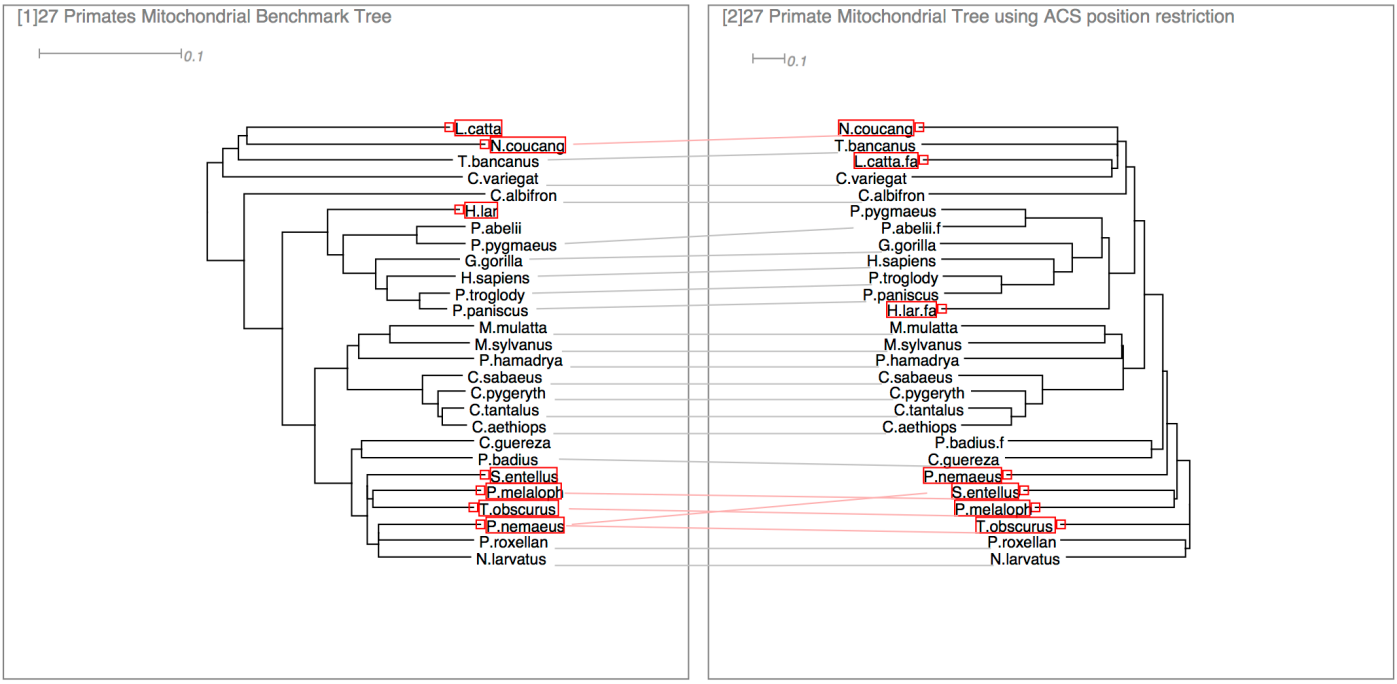
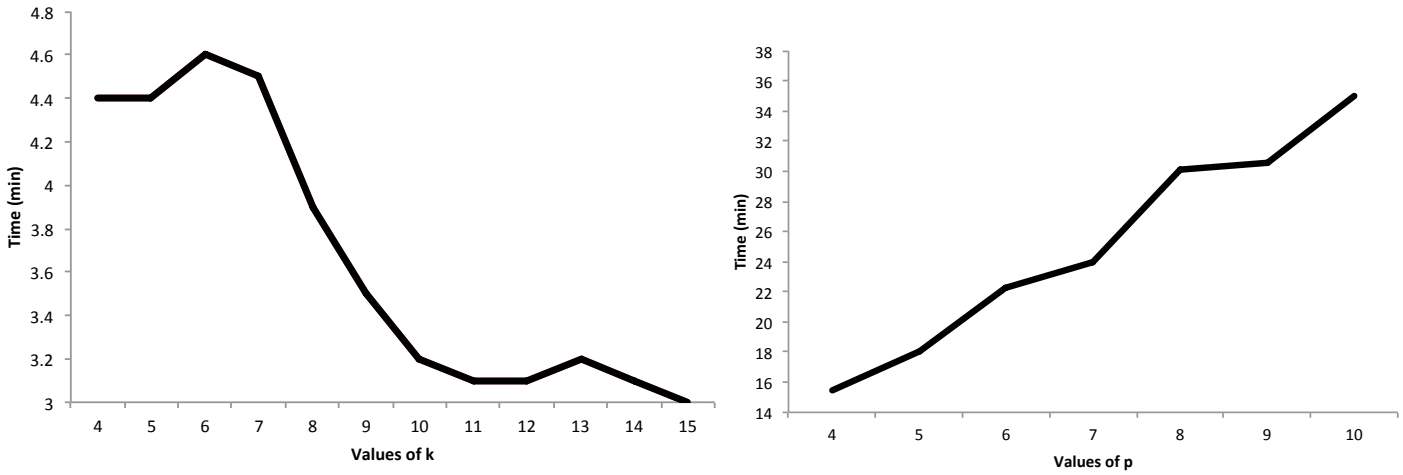


FIGURE 4.10. Comparison of 27 Primates Benchmark Tree with *ACS*-mer method.

Chapter 5
Timing Experiments for Algorithms

In this section, we present the timing results. On 27 primate data, the ACS method took 6.74 mins. The timing results for k-mer method and ACS with position restriction are presented in the following graphs. The experiments on star fish data took considerably more time. Approximately 8-10 mins for k-mer based method and around 33 mins for ACS method.



(a) k-mer method (b) ACS with Position restriction method

FIGURE 5.1. Graph showing timing experiment on Primates dataset

Although our k-mer based method is a linear time algorithm (theoretically), we observe that as the value of k increases, the time required reduces. Where as in ACS with position restriction, the time required increases as the number of partitions (p) increases.

Chapter 6
Conclusions

With the development of new technologies like nextgen sequencing,more and more sequences are generated. This needs a better mode of analysis and methods. In this thesis, we tried to study different methods of alignment free sequence analysis method for phylogenetic tree construction. The methods under study was mainly *k*-mer method, Average Common Substring method and ACS with position restriction being the novel method. These methods were tested on a set of nucleotide *(27 Primated Mitchondrial Dataset)* and a protein *(Starfish Datasets)*.

Our results indicates that the Average Common substring with position restriction is performing better than ACS and *k*-mer methods when compared to the Alignment methods. A notable highlight of this study was the result of ACS with position restriction when compared to the Starfish dataset, as this reference tree was constructed using a strict procedure and with a background assumption on the amino acid substitution models. The shortage of this method is its inability to capture the evolutionary timeline. Currently, the branch length obtained from ACS with position restriction is ten times that of the reference tree. However, our experimental study has proven that future works on correcting the branch length using ACS with position restriction can lead to a potential new method that can outlay the traditional alignment methods for phylogenetic tree construction in terms of speed as well as accuracy.

Bibliography

[1] Michael L Metzker. Sequencing technologies: the next generation. *Nature Reviews Genetics*, 11(1):31–46, 2010.

[2] Fiona SL Brinkman and Detlef D Leipe. Phylogenetic analysis. *Bioinformatics: a practical guide to the analysis of genes and proteins*, pages 323–358, 2001.

[3] Shengli Zhang and Tianming Wang. A novel alignment-free method for phylogenetic analysis of protein sequences. In *Proceedings of the 10th WSEAS international conference on Applied computer science*, pages 67–71. World Scientific and Engineering Academy and Society (WSEAS), 2010.

[4] Gregory E Sims, Se-Ran Jun, Guohong Albert Wu, and Sung-Hou Kim. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the National Academy of Sciences*, 106(40):17077–17082, 2009.

[5] Gregory E Sims and Sung-Hou Kim. Whole-genome phylogeny of escherichia coli/shigella group by feature frequency profiles (ffps). *Proceedings of the National Academy of Sciences*, 108(20):8329–8334, 2011.

[6] Hao Wang, Zhao Xu, Lei Gao, and Bailin Hao. A fungal phylogeny based on 82 complete genomes using the composition vector method. *BMC evolutionary biology*, 9(1):195, 2009.

[7] Pandurang Kolekar, Mohan Kale, and Urmila Kulkarni-Kale. Alignment-free distance measure based on return time distribution for sequence analysis: applications to clustering, molecular phylogeny and subtyping. *Molecular phylogenetics and evolution*, 65(2):510–522, 2012.

[8] Pandurang S Kolekar, Mohan M Kale, and Urmila Kulkarni-Kale. Research open access genotyping of mumps viruses based on sh gene: Develop-ment of a server using alignment-free and alignment-based methods. 2011.

[9] Klas Hatje and Martin Kollmar. A phylogenetic analysis of the brassicales clade based on an alignment-free sequence comparison method. *Frontiers in plant science*, 3, 2012.

[10] Chris-Andre Leimeister, Marcus Boden, Sebastian Horwege, Sebastian Lindner, and Burkhard Morgenstern. Fast alignment-free sequence comparison using spaced-word frequencies. *Bioinformatics*, page btu177, 2014.

[11] Igor Ulitsky, David Burstein, Tamir Tuller, and Benny Chor. The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13(2):336–350, 2006.

[12] Chris-Andre Leimeister and Burkhard Morgenstern. kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics*, 30(14):2000–2008, 2014.

[13] Bernhard Haubold, Peter Pfaffelhuber, Mirjana Domazet-Los? o, and Thomas Wiehe. Estimating mutation distances from unaligned genomes. *Journal of Computational Biology*, 16(10):1487–1500, 2009.

[14] Zhihua Liu, Jihong Meng, and Xiao Sun. A novel feature-based method for whole genome phylogenetic analysis without alignment: application to hev genotyping and subtyping. *Biochemical and biophysical research communications*, 368(2):223–230, 2008.

[15] Zhi-Hua Liu and Xiao Sun. Coronavirus phylogeny based on base-base correlation. *International journal of bioinformatics research and applications*, 4(2):211–220, 2008.

[16] Jinkui Cheng, Xu Zeng, Guomin Ren, and Zhihua Liu. Cgap: a new comprehensive platform for the comparative analysis of chloroplast genomes. *BMC bioinformatics*, 14(1):95, 2013.

[17] Yang Gao and Liaofu Luo. Genome-based phylogeny of dsdna viruses by a novel alignment-free method. *Gene*, 492(1):309–314, 2012.

[18] Hasan H Otu and Khalid Sayood. A new sequence distance measure for phylogenetic tree construction. *Bioinformatics*, 19(16):2122–2130, 2003.

[19] Jonas S Almeida. Sequence analysis by iterated maps, a review. *Briefings in bioinformatics*, 15(3):369–375, 2014.

[20] Susana Vinga and Jonas Almeida. Alignment-free sequence comparison?a review. *Bioinformatics*, 19(4):513–523, 2003.

[21] Kai Song, Jie Ren, Gesine Reinert, Minghua Deng, Michael S Waterman, and Fengzhu Sun. New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing. *Briefings in bioinformatics*, page bbt067, 2013.

[22] Bernhard Haubold. Alignment-free phylogenetics and population genetics. *Briefings in bioinformatics*, 15(3):407–418, 2014.

[23] Oliver Bonham-Carter, Joe Steele, and Dhundy Bastola. Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. *Briefings in bioinformatics*, page bbt052, 2013.

[24] Li Li, Christian J Stoeckert, and David S Roos. Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome research*, 13(9):2178–2189, 2003.

[25] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.

[26] Fabiano Sviatopolk-Mirsky Pais, Patrícia de Ruy, Guilherme Oliveira, and Roney Coimbra. Assessing the efficiency of multiple sequence alignment programs. *Algorithms for Molecular Biology*, 9:4, 2014.

[27] Barry G Hall. *Phylogenetic trees made easy: a how-to manual*, volume 547. Sinauer Associates Sunderland, 2004.

[28] Sudhir Kumar, Masatoshi Nei, Joel Dudley, and Koichiro Tamura. Mega: a biologist-centric software for evolutionary analysis of dna and protein sequences. *Briefings in bioinformatics*, 9(4):299–306, 2008.

[29] Koichiro Tamura, Glen Stecher, Daniel Peterson, Alan Filipski, and Sudhir Kumar. Mega6: molecular evolutionary genetics analysis version 6.0. *Molecular biology and evolution*, 30(12):2725–2729, 2013.

[30] Simon Gog, Timo Beller, Alistair Moffat, and Matthias Petri. From theory to practice: Plug and play with succinct data structures. In *13th International Symposium on Experimental Algorithms, (SEA 2014)*, pages 326–337, 2014.

[31] Daniel H Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[32] James A Studier, Karl J Keppler, et al. A note on the neighbor-joining algorithm of saitou and nei. *Molecular biology and evolution*, 5(6):729–731, 1988.

[33] Joseph Felsenstein. {PHYLIP}: phylogenetic inference package, version 3.5 c. 1993.

[34] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[35] Salvador Capella-Gutiérrez, José M Silla-Martínez, and Toni Gabaldón. trimal: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25(15):1972–1973, 2009.

[36] Alexandros Stamatakis. Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313, 2014.

[37] Kevin M Kocot, Mathew R Citarella, Leonid L Moroz, and Kenneth M Halanych. Phylotreepruner: A phylogenetic tree-based approach for selection of orthologous sequences for phylogenomics. *Evolutionary bioinformatics online*, 9:429, 2013.

[38] Patrick Kück and Karen Meusemann. Fasconcat: convenient handling of data matrices. *Molecular phylogenetics and evolution*, 56(3):1115–1118, 2010.

[39] Robert Lanfear, Brett Calcott, Simon YW Ho, and Stephane Guindon. Partitionfinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular biology and evolution*, 29(6):1695–1701, 2012.

[40] Daniel H Huson and Celine Scornavacca. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. *Systematic biology*, page sys062, 2012.

[41] Luca Pinello, Giosuè Lo Bosco, and Guo-Cheng Yuan. Applications of alignment-free methods in epigenomics. *Briefings in bioinformatics*, page bbt078, 2013.

[42] Chuang Peng. Distance based methods in phylogenetic tree construction. *NEURAL PARALLEL AND SCIENTIFIC COMPUTATIIONS*, 15(4):547, 2007.

[43] Michael Höhl, Isidore Rigoutsos, and Mark A Ragan. Pattern-based phylogenetic distance estimation and tree reconstruction. *Evolutionary bioinformatics online*, 2:359, 2006.

Appendix

- Amino acid: an organic compound containing both carboxyl and amino group.
- Clade: a group of organisms that includes an ancestor and all descendants of that ancestor.
- Cladogram: a branching diagram depicting the successive points of species divergence from common ancestral lines without regard to the degree of deviation.
- Dendrogram: a treelike diagram depicting evolutionary changes from ancestral to descendant forms, based on shared characteristics.
- DNA: Deoxy-ribo nucleic acid. It contains the genetic information.
- Gap: a space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another.
- Genbank Acession ID: sequence identification number that represents a single, specific sequence in the GenBank database.
- Homolog: a gene related to a second gene by descent from a common ancestral DNA sequence. The term, homolog, may apply to the relationship between genes separated by the event of speciation (see ortholog) or to the relationship between genes separated by the event of genetic duplication.
- Order: a taxonomic category of related organisms ranking below a suborder and above a family or superfamily.
- Ortholog: they are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.
- Outgroup: species or group of species closely related to but not included within a taxon.

[0]The definitions for the terms are from http://www.biology-online.org/dictionary

- Paralog: they are genes related by duplication within a genome. Orthologs retain the same function in the course of evolution, whereas paralogs evolve new functions, even if these are related to the original one.

- Phylogram: it is a phylogenetic tree that has branch spans proportional to the amount of character change.

- Protein: a molecule composed of polymers of amino acids joined together by peptide bonds.

- RNA-Seq: it is a high-throughput method by which the sequence of each RNA molecule in an organism can be determined.

- Substitution model: it describes the process from which a sequence of characters changes into another set of traits.

- Speciation: Speciation is the origin of a new species capable of making a living in a new way from the species from which it arose. As part of this process it has also aquired some barreir to genetic exchage with the parent species.

[0]The definitions for the terms are from http://www.biology-online.org/dictionary

Vita

Ambujam Krishnan was born in Kerala, India, in 1989. She obtained her Bachelor's degree in Biotechnology in 2010 and Master's degree in Bioinformatics in 2012 from Amrita Vishwa Vidhyapeetham, Kerala. During her Masters degree in Bioinformatics, she was awarded Gold medal for being the university topper.