

4.0 Results

4.1 GC Content Sliding Window Plot

In this project the alignment-based and alignment-free methods were applied to complete mitochondrion of 9 species which were, *Homo sapiens*, *Pongo pygmaeus*, *Hylobates lar*, *Gorilla gorilla*, *Pan troglodytes*, *Struthio camelus*, *Dromaius novaehollandiae*, *Gallus gallus* and *Terebratulina retusa*. Here, I came up with these GC content plots of the species (figure 4.11) to summarize the genomic characteristics of these species. The size of the sliding window in these graphs is 2000.

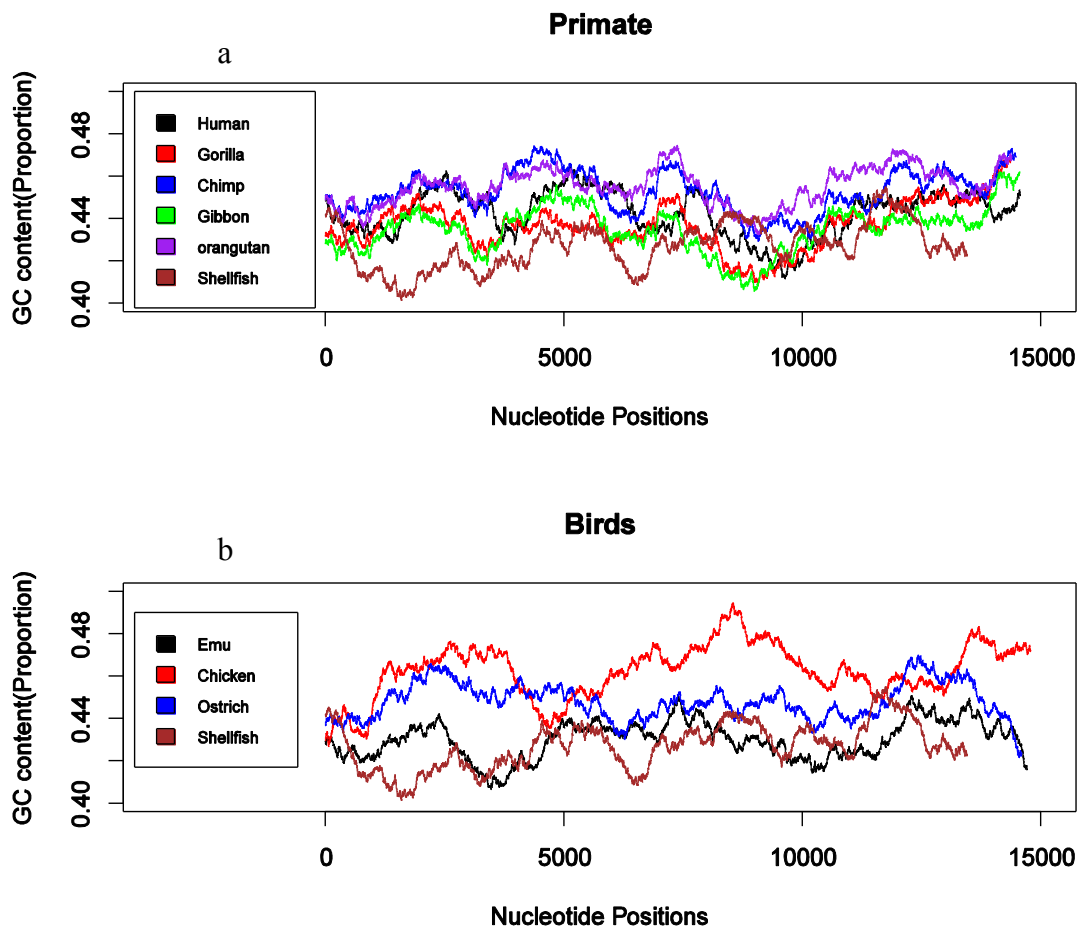


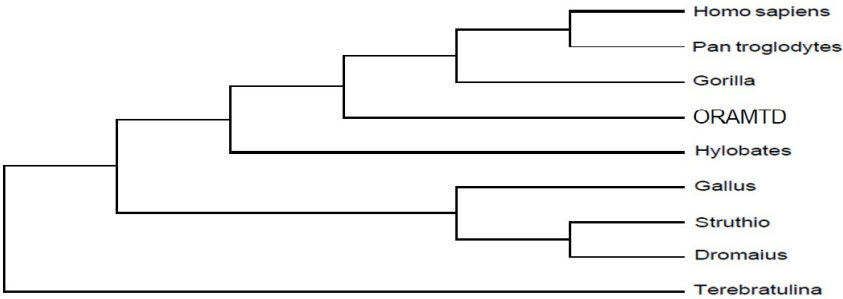
Figure 4.11. GC content slide window pot. Part a is GC content comparison among primates and brachiopoda, part b is GC content comparison among birds and brachiopoda. The size of sliding window is 2000.

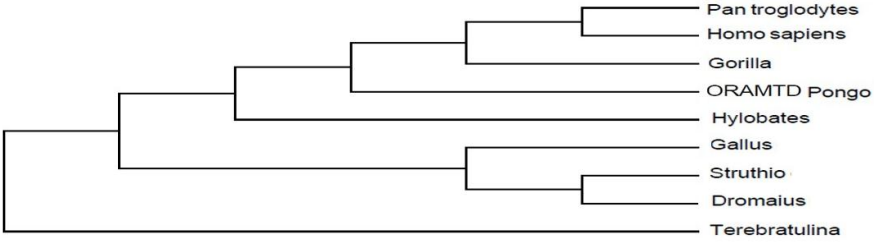
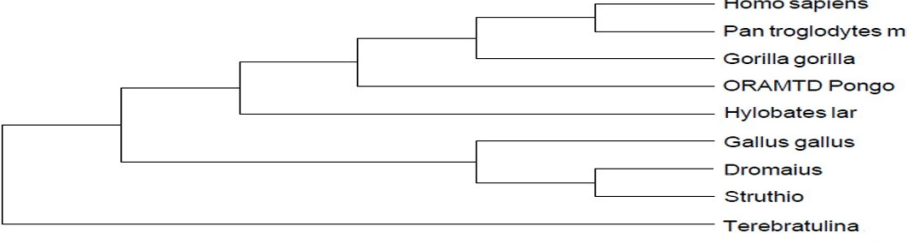
In figure 4.11 it is noticeable that GC content graphs in primates have similar profile. But by comparing primates with birds the noteworthy part is in between nucleotide positions from 5000 to 10000, which in bird are fairly high. Also by comparing primate and bird profiles with brachiopod`s, we can see brachiopod profile starts with higher GC ratio. But in overall the GC ratio in brachiopod is lower comparing to primates and birds.

4.2 Phylogenetic trees

An evolutionary or phylogenetic tree demonstrates the evolutionary relations among groups of organisms. In the trees below, the relationships among different 9 species are represented. These nine species are of three different clades which are primates, birds and brachiopods. All trees in table 1 are showing relations among these three clades but the different demonstrations are the result of different methods. In table 4.1, phylogeny trees which are result of alignment-based methods are similar. It means ClustalW and Muscle`s trees are compatible with our benchmark, T- Coffee. Although, alignment-free methods are generally much faster compared to alignment-based methods, the resulting phylogenetic tree is not as accurate as those alignment-based methods.

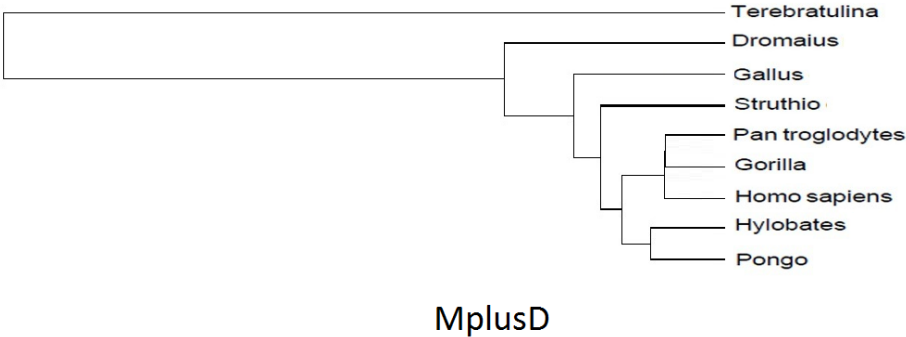

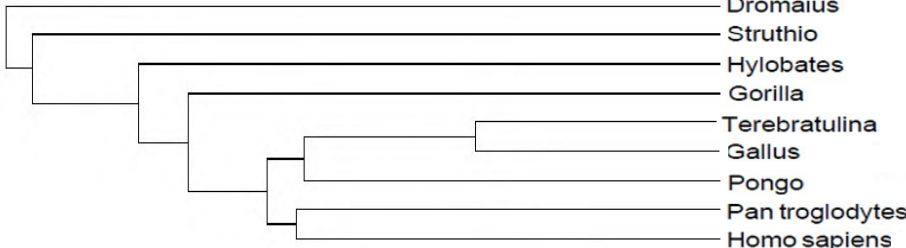
Table 4.1. Phylogeny trees built from alignment-based methods

Inferred tree	Observations
 <p style="text-align: center;">T-coffee</p>	<ul style="list-style-type: none"> - Reference tree - Out group is correctly placed - Clades are correctly placed

 <p style="text-align: center;">Clustalw</p>	<ul style="list-style-type: none"> - Out group is correctly placed - Clades are correctly placed - The phylogeny tree completely matches with the benchmark
 <p style="text-align: center;">Muscle</p>	<ul style="list-style-type: none"> - Out group is correctly placed - Clades are correctly placed - The phylogeny tree completely matches with the benchmark

For the tree built using the MplusD algorithm, the outgroup and clades are correctly placed. However, the species within the primate clade are misgrouped. Four changes have to take place in MplusD phylogeny tree so it will completely matches with the benchmark phylogeny tree. Also in both trees resulted from D2z and Kr algorithms the outgroup and clades are correctly placed. The only incorrect placement in Kr is between *Hylobates lar* and *Pongo pygmaeus* and the only incorrect placement; in D2z it is between *Gallus gallus* and *Struthio camelus*. The D2 algorithm produced a phylogeny three which both outgroup and clades are wrongly placed. When outgroup and clades are not correct, it is complicated to correct the places of wrong placements.

Table 4.2. Phylogeny trees built from alignment-free methods

Inferred tree	Observations
 <p style="text-align: center;">MplusD</p>	<ul style="list-style-type: none"> - Out group is correctly placed - Clades are correctly placed - There are 4 wrong placements
 <p style="text-align: center;">D2z</p>	<ul style="list-style-type: none"> - Out group is correctly placed - Clades are correctly placed - There is 1 wrong placement
 <p style="text-align: center;">D2</p>	<ul style="list-style-type: none"> - Out group is not correctly placed - Clades are not correctly placed

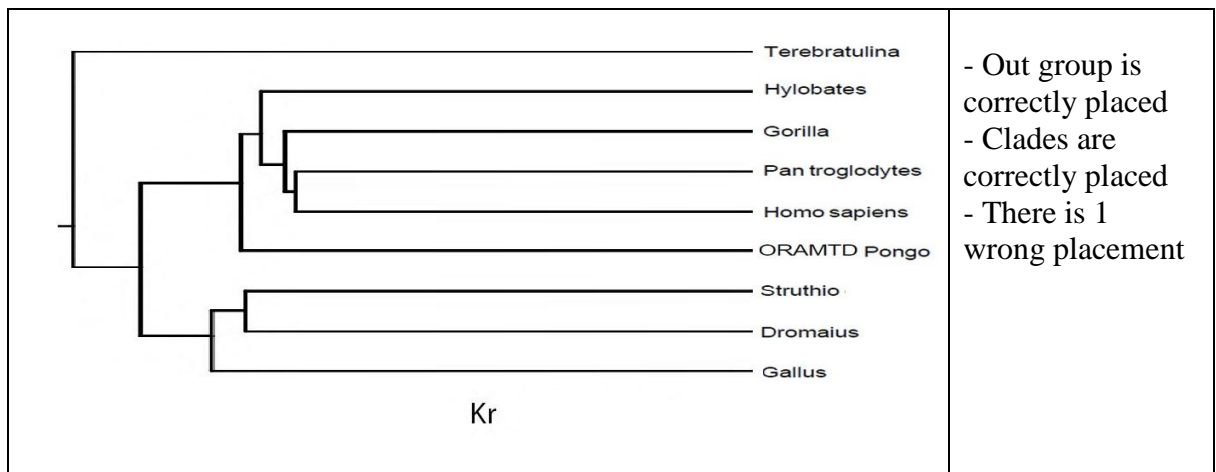


Table 4.3, summarizes the comparison between alignment-based and alignment-free methods. In addition to accuracy, the time in which the phylogeny tree is produced is also important. In general alignment-free methods took far less time to finish compared to alignment-based methods.

Table 4.3. Comparison of phylogeny trees built from alignment-based and alignment-free methods

Method	Algorithm	Outgroup correctly placed?	Clades correctly placed?	Number of changes needed to match reference tree	Processing time (min)
Alignment-free	Kr	Yes	Yes	1	-
	D2	No	No	X	1.8
	D2z	Yes	Yes	1	1.8
	MplusD	Yes	Yes	4	-
Alignment-based	ClustalW	Yes	Yes	0	8
	Muscle	Yes	Yes	0	31
	T-Coffee	Yes	Yes	0	54

5.0 Discussion

With N sequences, the algorithms in MSA method need N -dimensional matrices that have formed in standard pairwise sequence alignment. Therefore, with increasing N sequences, the search space increases exponentially and is also completely dependent on sequence length. The results from MSA methods take $O(N^3L^2)$ time to produce (Notredame, Higgins, & Heringa, 2000). As much as N and L increase, the time for getting output will also increase (Vinga & Almeida, 2003). Among the three alignment-based methods, T-Coffee was the slowest.

In alignment-free methods, the time of getting results depends on the algorithm of each method. In D2z, the time complexity is $O(4^k)$. In this algorithm, the time depends on the number of words with length of k (Kantorovitz, Robinson, & Sinha, 2007). In this project the default value for k was 6. In D2, the computational complexity is $O(4^{2k})$. Therefore compared to D2z, it will take a little more time to complete the work (Reinert, Chew, Sun, & Waterman, 2009). In the Kr method, the complexity of algorithm had been recently improved from $O(N^2L)$ to $O(NL)$. Comparing the algorithm complexities, Kr run time is less than the other algorithms in alignment-free methods (Domazet-Loso & Haubold, 2009).

Considering the tradeoff between speed and accuracy, Kr and D2z methods seem to be reasonable if only rough approximation is needed. Also it is important to notice that the results in this project are limited to only input data. This means that by using other species or by changing the number of sequences the results may be different. For extending the outcome of this project, we need to do more testing using larger data sets (more taxa).

Phylogenetic signals are captured in the form of gene order; within taxa, they are captured in nucleotide substitutions, deletions, insertions and rearrangements. If gene order is an important source of variation in tree topology (usually true when building trees for organisms that have diverged for a long time), then alignment free methods will tend to fail. On the other hand, as long as closely related organisms are compared, satisfactory results using alignment-free methods may be obtained. Alignment-based method can capture phylogenetic signal better than alignment-free methods because of differences in manifest as large blocks of gaps in gene order (Luo *et al*, 2009).

Another limitation of alignment free methods is that they just give distance matrices to summarize the relatedness of the sequences. This means that only distance based trees can be build (e.g. NJ, minimum evolution etc). So if a researcher desires a maximum likelihood or maximum parsimony tree, then alignment free methods cannot be useful in such cases.

Alignment-free methods may be very useful when we try to build phylogenetic tree of large number of closely related bacterial species.

6.0 Conclusions

In this project, the phylogenetic tree inferred using alignment-based methods were ClustalW, the best method with respect to speed and accuracy. The Kr and D2z were most reasonable methods based on the tradeoff between speed and accuracy.

For practical use, it is recommended that alignment based methods be used when constructing phylogenetic trees containing many taxa that have diverged for a long time. However, alignment-free methods should be useful in cases when we wish to infer the phylogeny of large numbers of closely related organisms, since the contribution of gene orders to phylogenetic signal will be small in this case.