Chapter 1

INTRODUCTION

1.1. The Mitochondrion

The mitochondrion (Alberts *et. al.*, 2007) is one of the most important organelle in a cell. Its size ranges from 0.5µm to 1.0µm in diameter, making it one of the three largest organelle in a cell (together with nucleus and vacuole – largest in plant cells). Similar to the structure of a nucleus (normally eukaryotes), the mitochondrion has multiple internal compartments that carry out specialized functions. The compartments are outer membrane, intermembrane space, inner membrane, cristae space and the.

The outer mitochondrial membrane (Alberts *et. al.*, 2007) encapsulates the whole organelle, preventing external molecules from penetrating into the mitochondrion. Large number of integral protein and porins provide channels for molecules 5000 Daltons or less to diffuse into the organelle. The larger molecules or protein can enter via the binding to the large multisubunit protein called translocase of the outer membrane. The intermembrane space resides between the outer and inner mitochondrion membrane that deals with the protein composition in itself. Protein such as the protein cytochrome c will pass through this space or reside here.

The inner membrane (Alberts *et. al.*, 2007) houses multiple proteins that can be grouped into five groups; redox reactions of oxidative phosphorylation, adenosine triphosphate (ATP) synthase, specific transport proteins which regulate metabolite passage into and out of the matrix, protein import machinery and mitochondrion fusion and fission protein.

The cristae (Alberts *et. al.*, 2007) expands the surface area of the inner membrane through the folding or invagination of the membrane, providing further compartmentalization that enhance the ATP production. The matrix contains the majority of the proteins of the mitochondrion, together with enzymes, ribosome, tRNA and its own genome.

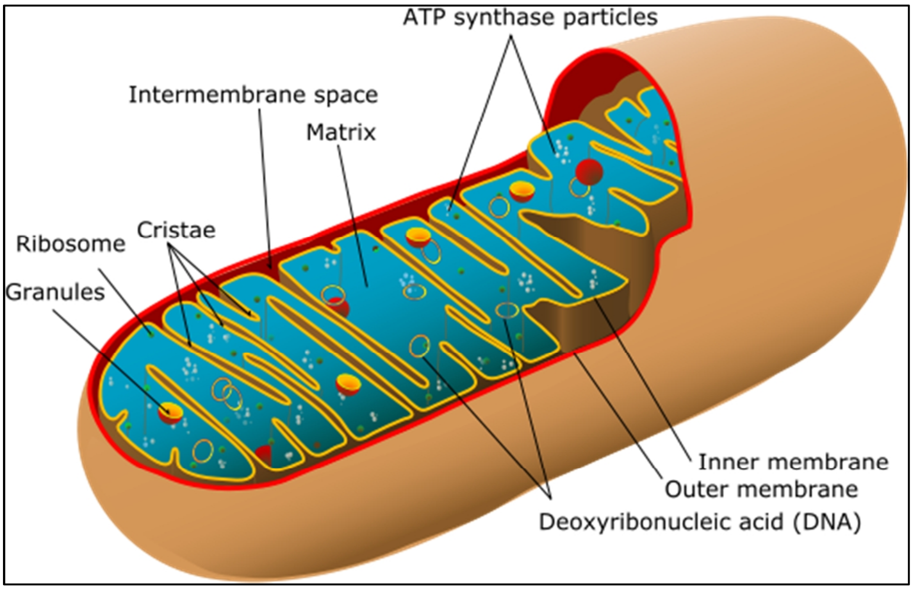


Figure 1.1 Structure of mitochondrion

(Source: http://en.wikipedia.org/wiki/Mitochondrion)

The mitochondrion (Alberts *et. al.*, 2007) has a central role in energy transduction in form of ATP which is used as a source of chemical energy in cells. Besides being the power generator of the cell, the mitochondrion is also involved in other important functions, including ion homeostasis, intermediary metabolism, signaling, cellular differentiation and apoptosis or cell death, as well as the control of the cell cycle and cell growth. All this depends on the molecules reside or produced inside the mitochondrion.

According the endosymbiotic theory (Burger *et. al.*, 2003), the mitochondrion DNA (mtDNA) has separate evolutionary origin. It is thought that the genome of a prokaryotic cell was engulfed by eukaryotes cells. An indirect evidence is that some circular shape of the mitochondrial genome is the similar to those found in prokaryotic cells such as bacteria.

A typical metazoan (e.g. *Homo sapiens*) mtDNA is around 16 kbp long, encoding for 37 genes (Boore, 1999). There are three types of genes: protein coding genes (13), tRNA genes (22) and rRNA genes (2). These genes are well-annotated and rarely found to differ within species in terms of gene order arrangement. The mtDNA genes are ideal for phylogenetic tree building as gene order can also be used for closely related species to find distantly related species because their gene order will differ.

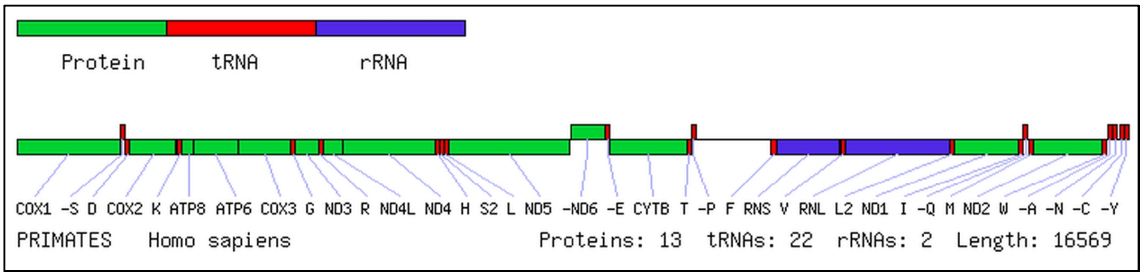

Protein
tRNA
rRNA
COX1 -S D COX2 K ATP8 ATP6 COX3 G ND3 R ND4L ND4 H S2 L ND5 -ND6 -E CYTB T -P F RNS V RNL L2 ND1 I -Q M ND2 W -A -N -C -Y
PRIMATES Homo sapiens
Proteins: 13 tRNAs: 22 rRNAs: 2 Length: 16569

Figure 1.2 Mitochondrial Genome of *Homo sapiens*

(Source: OGRe Genome Viewer)

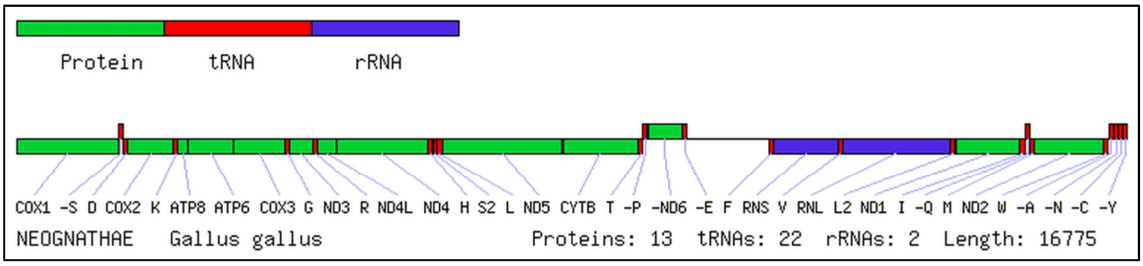

Protein
tRNA
rRNA
COX1 -S D COX2 K ATP8 ATP6 COX3 G ND3 R ND4L ND4 H S2 L ND5 CYTB T -P -ND6 -E F RNS V RNL L2 ND1 I -Q M ND2 W -A -N -C -Y
NEOGNATHAE Gallus gallus
Proteins: 13 tRNAs: 22 rRNAs: 2 Length: 16775

Figure 1.3 Mitochondrial Genome of *Gallus gallus*

(Source: OGRe Genome Viewer)

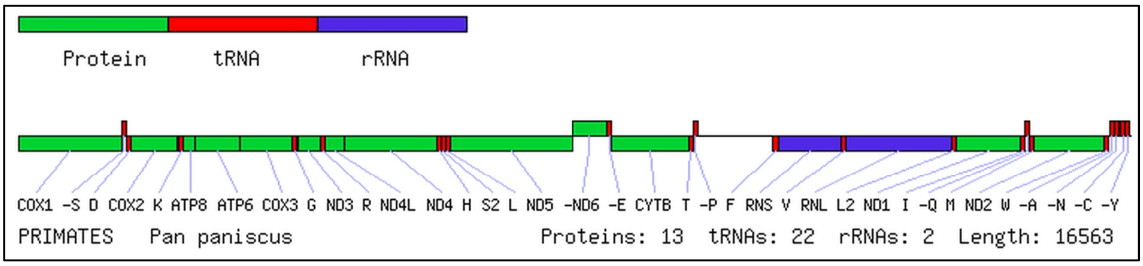

Protein
tRNA
rRNA
COX1 -S D COX2 K ATP8 ATP6 COX3 G ND3 R ND4L ND4 H S2 L ND5 -ND6 -E CYTB T -P F RNS V RNL L2 ND1 I -Q M ND2 W -A -N -C -Y
PRIMATES
Pan paniscus
Proteins: 13
tRNAs: 22
rRNAs: 2
Length: 16563

Figure 1.4 Mitochondrial Genome of *Pan paniscus*

(Source: OGRe Genome Viewer)

1.2. Bootstrapping

Bootstrapping (Efron *et. al.*, 1993) is used as means of approximating the distribution of an estimator of a population parameter using pseudo-samples obtained by resampling from the observed data. The standard error (SE) of the estimator can then be found by checking the standard deviation (SD) of the bootstrap distribution and its 95% confidence interval (CI) is found by taking the 2.5th and 97.5th quantiles.

By resampling columns of a multiple sequence alignment used to construct a tree, and then comparing the bootstrap phylogenies with the observed tree, one could estimate this SD by reporting the bootstrap support (p), which is the proportion of times a node is retained in the replicates. High bootstrap support is equivalent to low SD.

Figure 1.5 shows the bootstrapping of a multiple sequence alignment in action which resample and create n^{th} subsamples that are theoretically assumed to be available in the real world. During resampling step in the bootstrap iteration, one column is randomly selected and replaced. It will be continued to create a bootstrap replicate (subsample). Later a new tree is constructed using the subsample and compared against the original tree. If the clade exists in the new tree it will be scored 1 while a non-existent clade is scored 0. This will continue until all the iteration is over.

However this may not be reflected in an actual sequence as the sequences created may not comply to some conditions such as the availability of sudden gaps and there's no start or stop codons at the beginning and end of sequences.

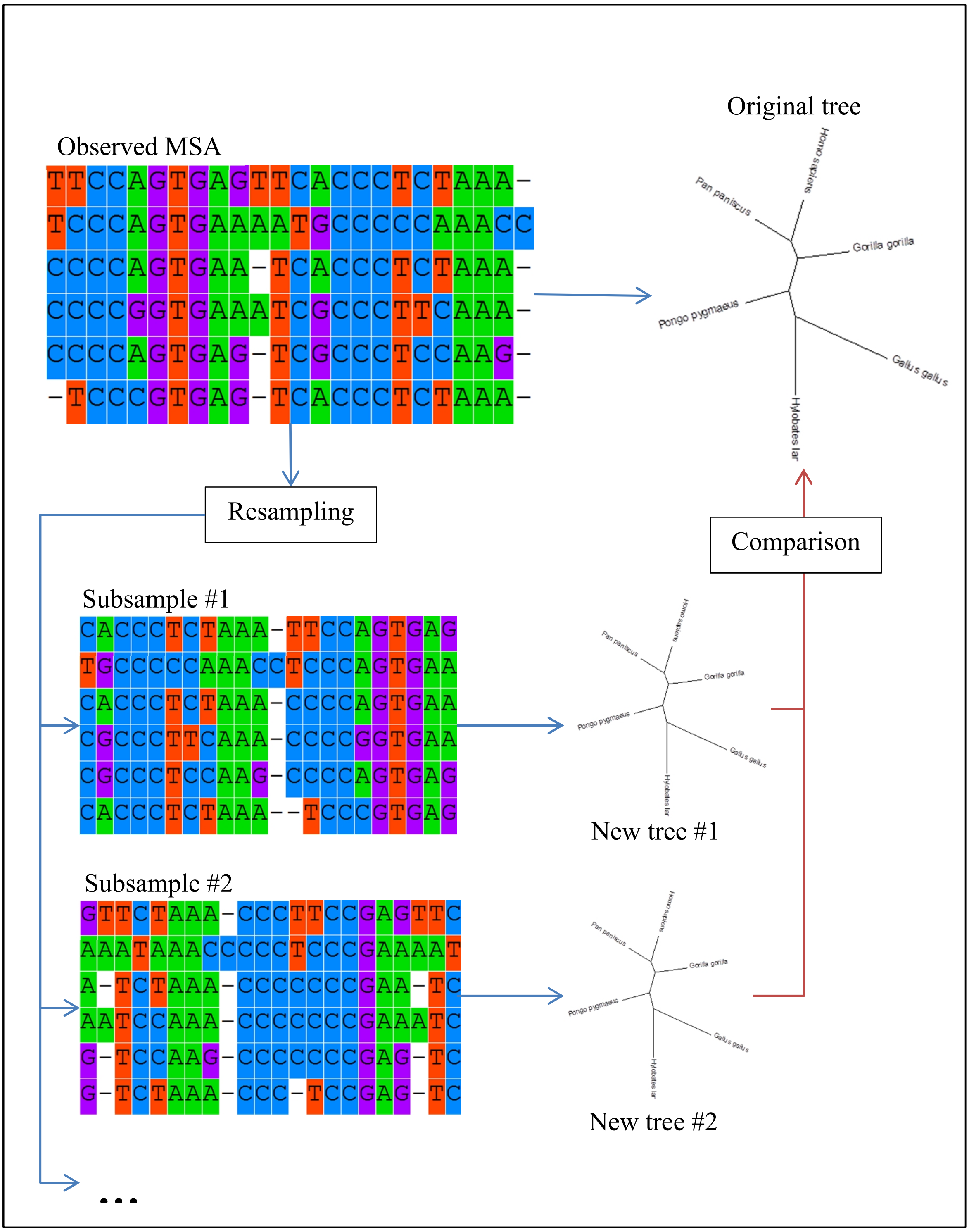


Figure 1.5 Bootstrapping includes column resampling and creating sequences

However when used in phylogenetic trees, bootstrapping assumes that such resampled sequences exist somewhere in the genome and assumes that the real sequence is one of the realization from the subsample. This assumption is unlikely to be true as evolutionary pressure constraints sequence evolution, so certain sequences cannot exist. This represents a significant departure from classical bootstrapping where the bootstrapped values are true realisations from the distribution of the population.

Unfortunately, it appears that bootstrapping phylogenetic trees has developed a kind of necessity among molecular phylogeneticists (Hillis *et. al.*, 1993), possibly because of the appearance of scientific objectivity that it lends to this craft. The only scientifically valid way to assess the value of bootstrapping phylogenetic trees is to empirically determine node support using real sequences.

1.3. Objectives

The objective of this project is to assess the bootstrap method empirically using nodes built from phylogenetics tree from actual genes rather than the artificial resampling of single genes. The assessment is based on the degree of discrepancy in estimated node support between the two methods.

Chapter 2

LITERATURE REVIEW

2.1. Mitochondrion in Evolution

There are several advantages of using mtDNA as a phylogenetic tool (Boore *et. al.*, 1998). All organisms will have mitochondrion (whether encapsulated or not) that have varying and complex states between taxa. Mitochondrion is also unambiguously homologous and also selectively neutral to minimize convergent changes as it is past down maternally (most of the cases). mtDNA is also labile enough to generate a strong signal at many branches but not so labile that subsequent changes would erode it. Other genes may not have all these criteria.

According to Boore *et. al.*, 1998, the mitochondrion gene order can be used as one of the reference point in the inference of a phylogenetic tree. This is because across the metazoan, the gene contents almost all the same and also homologous even across organism – plant, fungi and protists. The great number of potential arrangements makes convergence a rare occurrence, therefore, gene arrangements are likely to be shared only as a result of common ancestry. The gene arrangement is also stable and infrequent rearrangement as there is no genetic recombinant and almost inherited maternally which retain the signal of ancient common ancestry.

The arrangement of genes in mtDNA is highly conserved. However mutation such as nucleotide substitution, point deletion and insertion may can the arrangement to change. For example, a slipped strand mispairing (Boore *et. al.*, 1998) will cause gene duplication within taxa during replication which causes shift in subsequent sequences. Another mechanism happen due to the nature of mitochondrion genome as it is circular and if the origin and termination shifts, some of the genes in between may also be duplicated. The gene order is good for phylogeny between distantly related taxa but poor between closely related taxa as the gene arrangement often remain the same between closer taxa.

The mitochondrion is also smaller in size compared to the whole genome and remains almost the same size across taxa. This smaller size happens due the gene loss and transfer to the nucleus (Adams, 2003). Most of the coding sequence is transferred to nucleus or is loss due to duplicating genes in the cell.

2.2. *Bootstrapping in Phylogenetic*

Efron propose the idea of bootstrapping (Efron, 1982). Later, bootstrapping in phylogenetics was proposed by Felsenstein in 1985. In the journal he shows that how a phylogenetic can be more trusted with addition of bootstrap support value to show its occurrence during bootstrapping.

According to Hillis and Bulls (1993), bootstrapping is a circumstantial method through a series of experiments using simulated samples. Bootstrapping will be affected by multiple factors. The factors are number of characters, taxa, bootstrap iterations performed, rate of change, tree topology, position of the group of interest within the tree, variance of rates of change among lineages, independence of characters, and method of phylogenetic inference. The final bootstrap support value may not reflect the actual phylogenetic accuracy where underestimation or overestimation may occur, depending on the rate of changes in the sequence.

Felsenstein and Kishino (Felsenstein *et. al.*, 1993) point that the result from Hillis and Bulls (1993) relies too much on the P-value instead of the claimed unbiased confidence interval to justify the validity of bootstrap support value. Felsenstein and Kimura think that it this may be an overestimation and overreaction to the reliance to P-value but until a better way of summarizing the result of a tree, the biologist will continue with bootstrapping support.

However mathematics models are only suitable if genes are independent of each other. However this may not be case in any organism especially complex species. For example, co-dependency can be found between FoxO3 and Pax3/7 to co-ordinately recruit RNA polymerase II and form a pre-initiation complex (PIC) to activate MyoD transcription in myoblasts (Hu *et. al.*, 2008).

2.3. Software

R (R Development Core Team, 2012) is a language and environment for statistical computing which can be further enhanced using additional packages. For building and analysis of phylogenetic tree, the package, Analyses of Phylogenetics and Evolution (APE) and phangorn is essential package in R.

According to Paradis (2004), APE is a package written in the R language for the use in molecular evolution and phylogenetics as well as providing reading and writing data and manipulating phylogenetic trees, with several advanced methods for phylogenetic and evolutionary analysis. There are limitation as in APE which nicely complemented by phangorn (Schliep *et. al.*, 2011) especially in tree building with reconstruction method phylogenies with distance based methods, maximum parsimony or maximum likelihood (ML) and performing Hadamard conjugation with other features as well.

Molecular Evolutionary Genetics Analysis is a freeware tool for mining online databases, building sequence alignments and phylogenetic trees, and using methods of evolutionary bioinformatics in basic biology, biomedicine, and evolution (Kumar *et. al.*, 2008). Currently it is at version 5 with additional features (Tamura *et. al.*, 2011). With additional of a collection of maximum likelihood (ML) analyses for inferring evolutionary trees, selecting best-fit substitution models (nucleotide or amino acid), inferring ancestral states and sequences (along with probabilities), and estimating evolutionary rates site-by-site. Using computer simulation, the result from MEGA5 compares favorably with other software packages (e.g. RaxML7, PhylML3) in terms of efficiency and accuracy of phylogenetic tree. Additionally, MEGA5 provides a user friendly interface for building and analyze phylogenetic tree with other features as well.

Python the COmparative GENomic Toolkit (PyCogent) (Knight *et. al.*, 2007) is a multipurpose framework which can be used for analyses of biological sequences in a novel probabilistic manner, devising workflows, and generating publication quality graphics that relies on connectors to remote databases, generalized probabilistic techniques for working with biological sequences that uses various built-in libraries, and controllers for third-party applications such as MUSCLE and ClustalW. This is very useful when dealing with a very large amount of data as well as a smaller one as it can provide automation.

Chapter 3

METHODOLOGY

There are four major steps in execution of the study as shown in Figure 3.1.

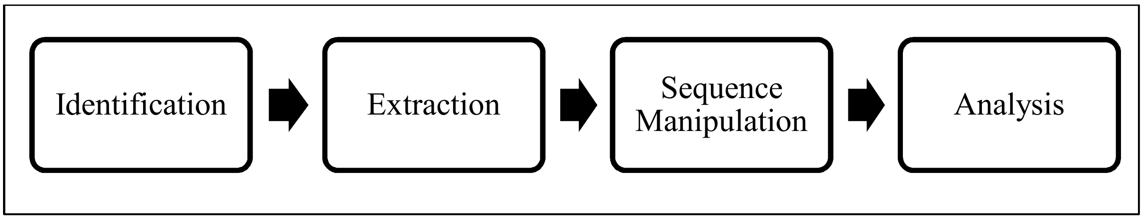


Figure 3.1 Major steps in execution of the study

In identification, a trusted and well curated tree is selected (Perelman *et. al.*, 2011) as the gold standard for the phylogenetic tree. The tree is widely accepted and proven to be true. The taxa include *Pan paniscus* (bonobo or also known as pygmy chimpanzee), *Homo sapiens* (human), *Gorilla gorilla* (western gorilla), *Pongo pygmaeus* (bornean orangutan), *Hylobates lar* (common gibbon) and *Gallus gallus* (red junglefowl or chicken) as the outgroup.

The next step, extraction, the whole genome of mitochondrion of each taxa is identified, selected and extracted from National Center for Biotechnology Information (NCBI) using EFetch() method in PyCogent. The GI chosen are 5835820 (for *Hylobates lar*), 5835163 (for *Pongo pygmaeus*), 5835135 (for *Pan paniscus*), 5835149 (for *Gorilla gorilla*), 5834843 (for *Gallus gallus*) and 251831106 (for *Homo sapiens*).

Encoded into each of the retrieved sequence, there is additional information which tells the position of start and stop of the genes inside the whole mitochondrial genome sequence. The position of the genes is well curated. Using the start and stop information, the position of the genes is extracted into separate fasta files, according to their respective species.

Alignment using MUltiple Sequence Comparison by Log- Expectation (MUSCLE) with gap penalty of -500 is used on all the extracted genes, including the whole mitochondrial genome sequence. A gap penalty of -500 is applied because the default value of the gap penalty of MUSCLE in MEGA is -400. Additional gap penalty is applied so that the aligned sequences will not contain too many unrestricted gaps but not too much until no gaps can be inserted. All of this is automated using the PyCogent from extraction from NCBI to the 37 separated fasta files of each gene of each taxa.

Maximum likelihood (ML) tree is selected (Felsenstein *et. al.*, 1985) and the gold standard is constructed from the aligned mitochondrion genome in both MEGA and R. Both tools will give a slightly different result. Bootstrap support will also be added when building the tree. The pairwise distance uses the Tamura-Nei, 93(TN93) model while the bootstrap iteration is at 1000. The tree building in MEGA involves a friendly user interface while R uses automation in codes. In R, align sequence is read using APE package and a pairwise distance matrix is calculated and is used for building the neighbor-joining trees. Further calculation using phangorn package will produce the ML tree with the bootstrap support (codes available in Appendix F and Appendix G). The trees built are unrooted type and all the layouts are available in the Appendix D and Appendix E.

After all the ML trees using all the genes and the mitochondrial genome are constructed from the sequence manipulation step, the analysis process begins. The layout of each tree is recorded and every frequency of the same node as the gold standard will be recorded. The frequencies represent the empirical estimate of the mitochondrion tree. The recorded data will be analysed and multiple statistical methods are applied to build graphs to enable clarity and understanding.

Chapter 4

RESULTS

The golden standard of the ML tree for both MEGA and R is the same as they have the same topology (almost the same bootstrap value as well)

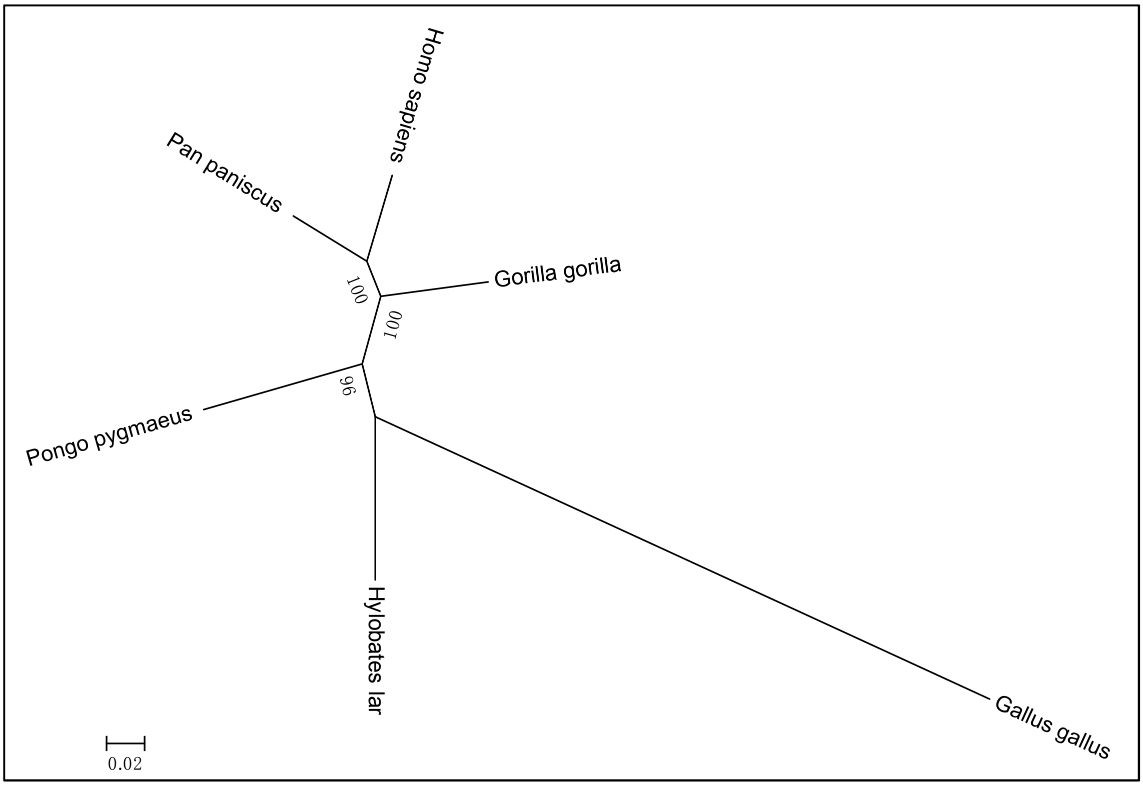


Figure 4.1 Topology of the reference tree

The correct topology is the *Homo sapiens* with *Pan paniscus*, connected to *Gorilla gorilla*, then *Pongo pygmaeus*, to *Hylobates lar* and finally the outgroup, *Gallus gallus*. As seen from the Figure 4.1, the distance of *Gallus gallus* is very definite and this is expected as it is the outgroup..

Node A, B and C are shown in Figure 4.2. The node A (circled in red) is between *Homo sapiens* and *Pan* paniscus. The node C (circled in green), is between *Pongo pygmaeus* and *Hylobates lar* with *Gallus gallus* while node B (circled in blue) is *Gorilla gorilla* which is between node A and B.

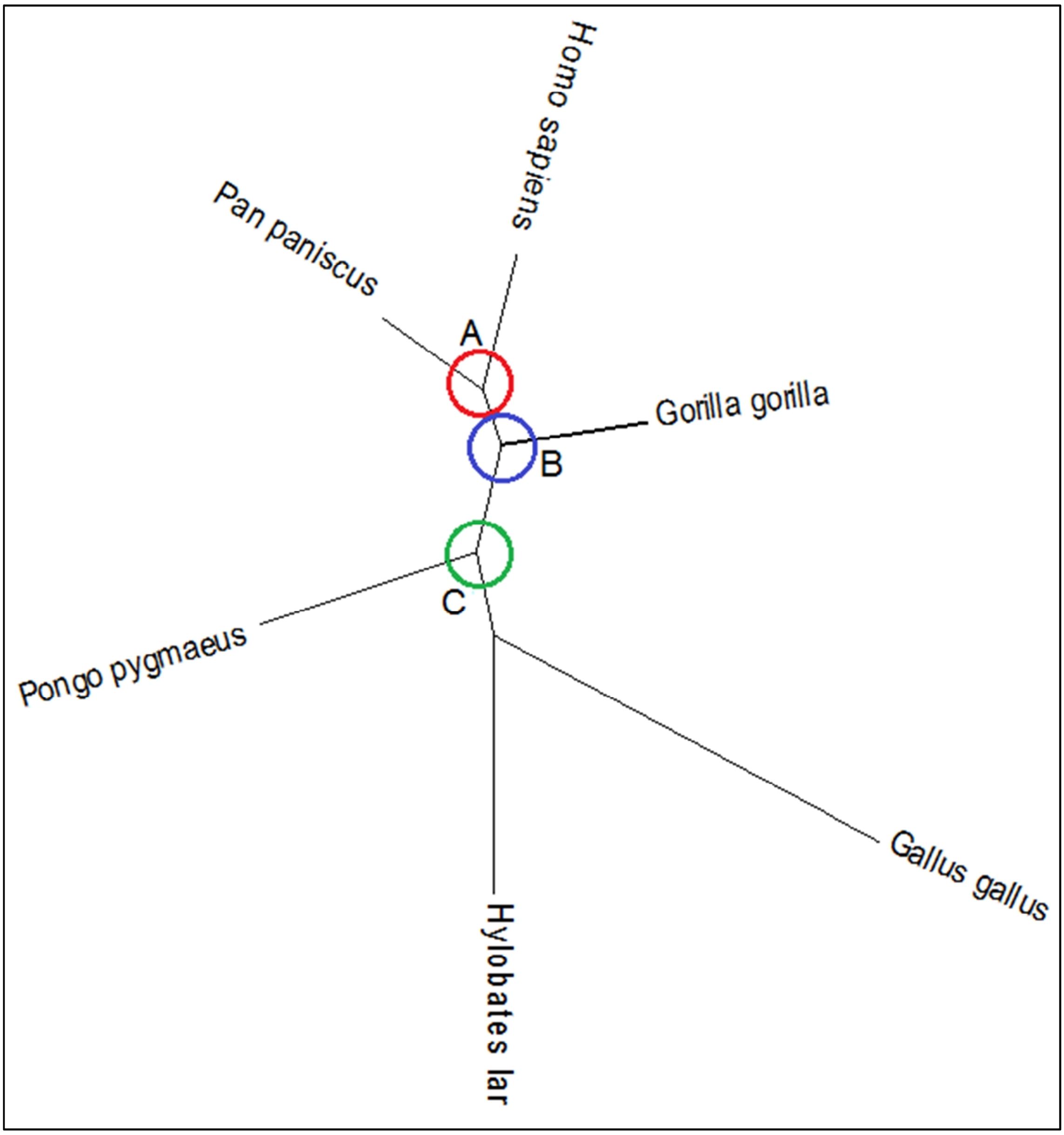


Figure 4.2	Position of node A, B and C on the topology of the reference tree

The frequencies of the appearance of each node in the topology of the built trees from each gene in mtDNA are recorded into tables. There are some differences in the frequencies of the node between the tree generated by MEGA and R are captured into tables. Each tables record for different type of genes in mtDNA and can be seen in Appendix A (protein coding sequence), Appendix B (Ribosomal RNA Sequence) and Appendix C (Transfer RNA Sequence). The total is summarized in in Table 4.1.

Table 4.1 Total Frequency of Node Appearance

Type	***MEGA***			***R***		
	A	**B**	**C**	**A**	**B**	**C**
Protein Coding	11	10	7	10	10	4
rRNA	0	0	2	0	0	2
tRNA	11	6	6	11	7	6
TOTAL	22	16	15	21	17	12

From the frequency we can construct an empirical estimate which can be calculated by finding the percentage of the node over total gene (37) as show in Table 4.2.

Table 4.2 Empirical Estimate for Each Node

Type	***MEGA (%)***			***R (%)***		
	A	**B**	**C**	**A**	**B**	**C**
Protein Coding	30	27	19	27	27	11
rRNA	0	0	5	0	0	5
tRNA	30	16	16	30	19	16
TOTAL	60	43	40	57	46	32

Figure 4.3 and 4.4 show the different empirical estimate calculated from the frequency of node between MEGA and R.

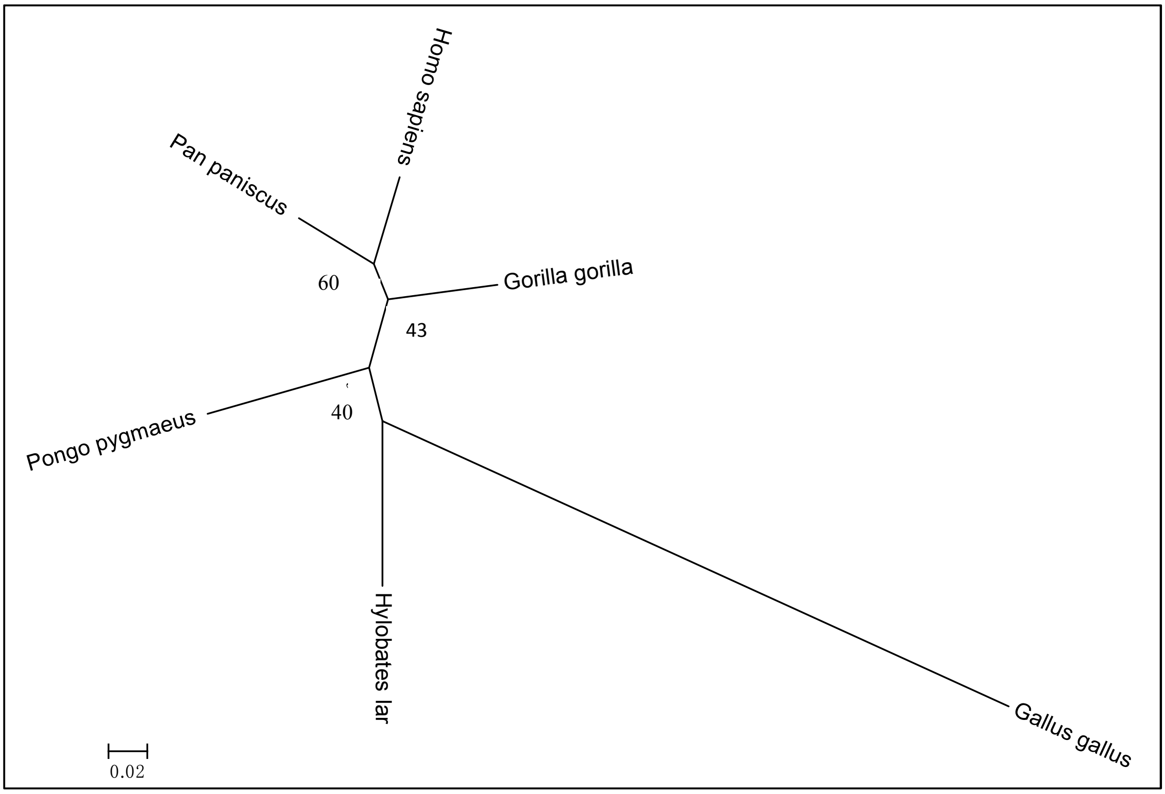


Figure 4.3 Empirical estimate of the reference tree in MEGA

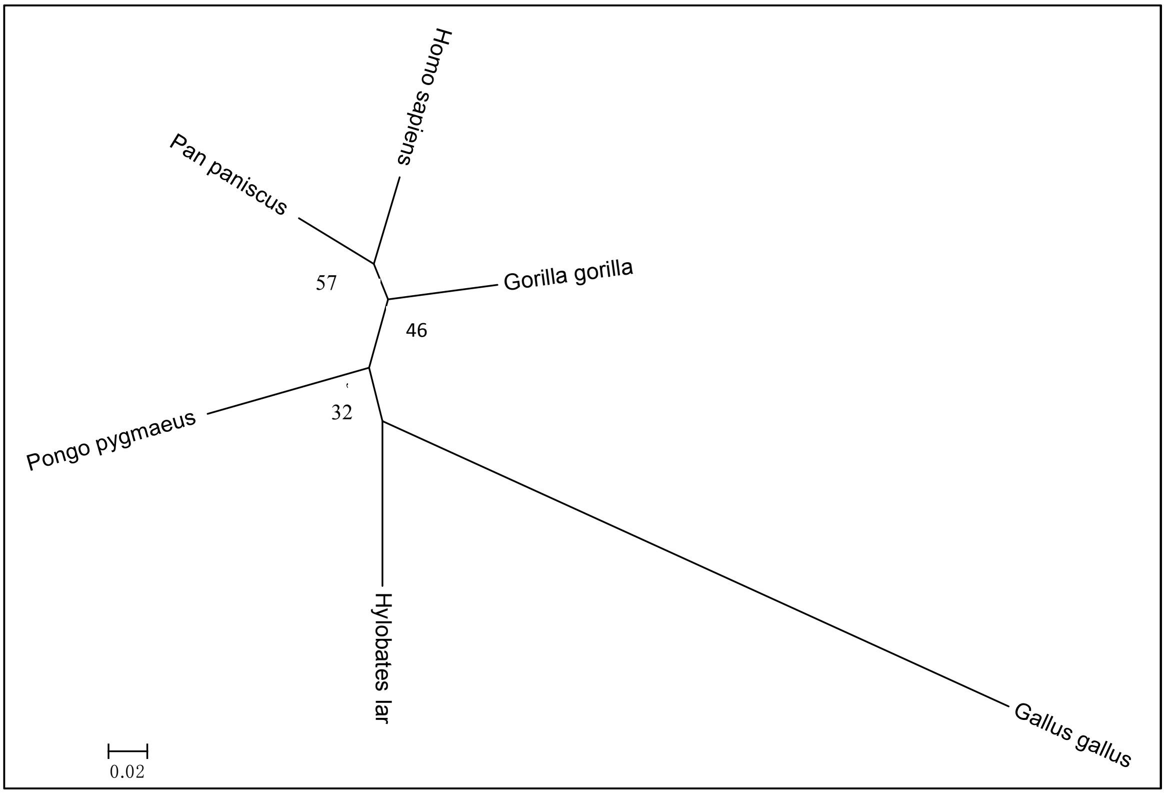


Figure 4.4 Empirical estimate of the reference tree in R

The use of R in the result is not for full comparison but a highlight of different outcome from different tools. Only the result from MEGA5 will be used for the analysis of the empirical estimates for bootstrap support. A comparison is done on the frequency of each of the bootstrap value from the phylogenetic tree built. The tRNA and non-tRNA (rRNA and protein coding) are separated due to the dissimilarity in length of sequences. A summary of the data can be found in Table 4.3.

Table 4.3 Summary of bootstrap values in node A, B and C from MEGA

Type	**Node A**	**Node B**	**Node C**
Min	43	59	49
1st Quarter	67	79	63
Median	84	88	67
Mean	76	86	68
3rd Quarter	90	99	69
Maximum	96	100	91
Variance	305	212	121
Standard Deviation	18	15	11

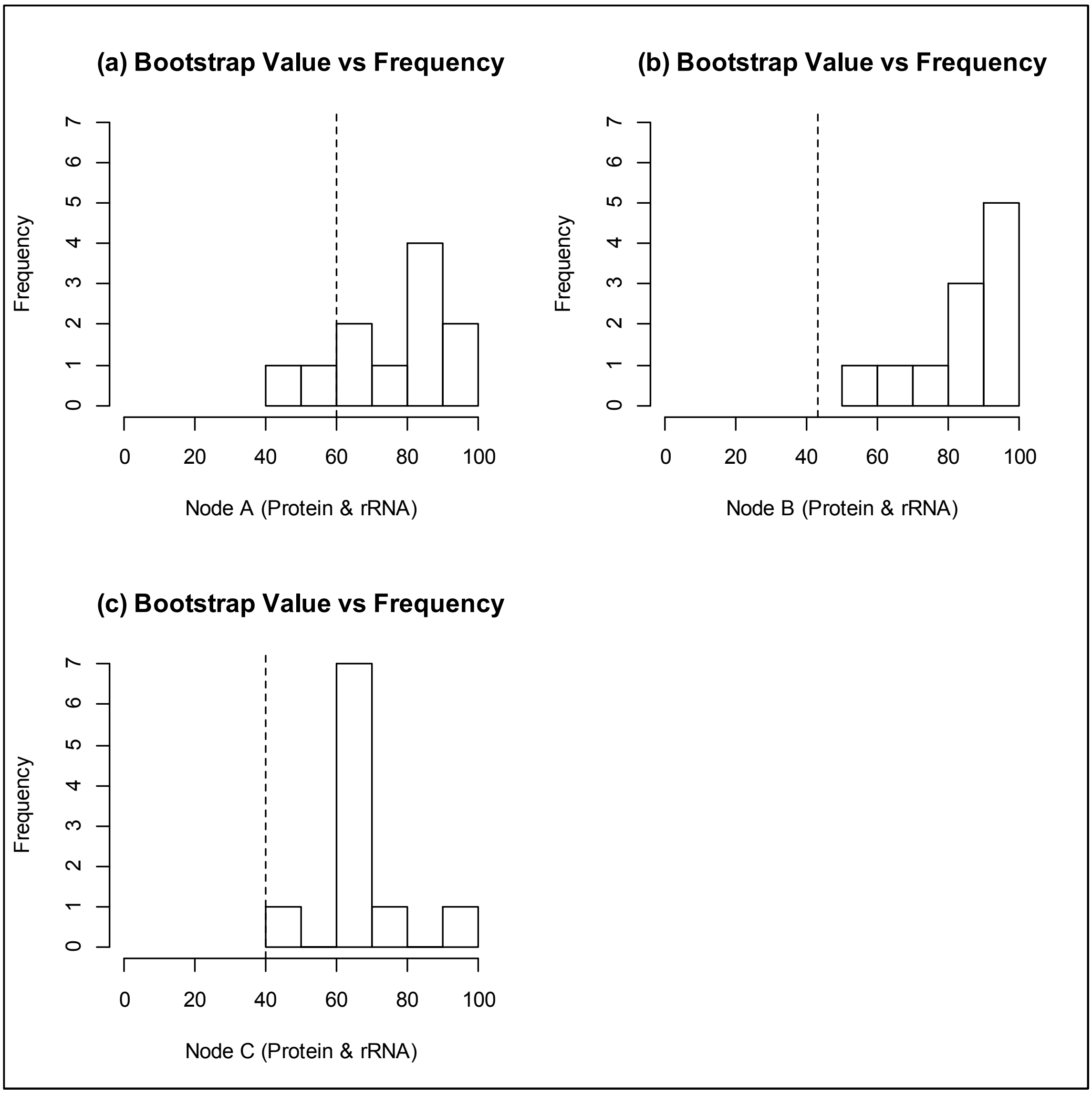


Figure 4.5 Histogram of frequency of bootstrap support for (a) Node A, (b) Node B and (c) Node C for protein and rRNA genes. The dotted line reflect the empirical estimate from all genes

From Figure 4.5, we can see that bootstrap support for the same nodes in different trees tend to be much higher than the empirical node support.

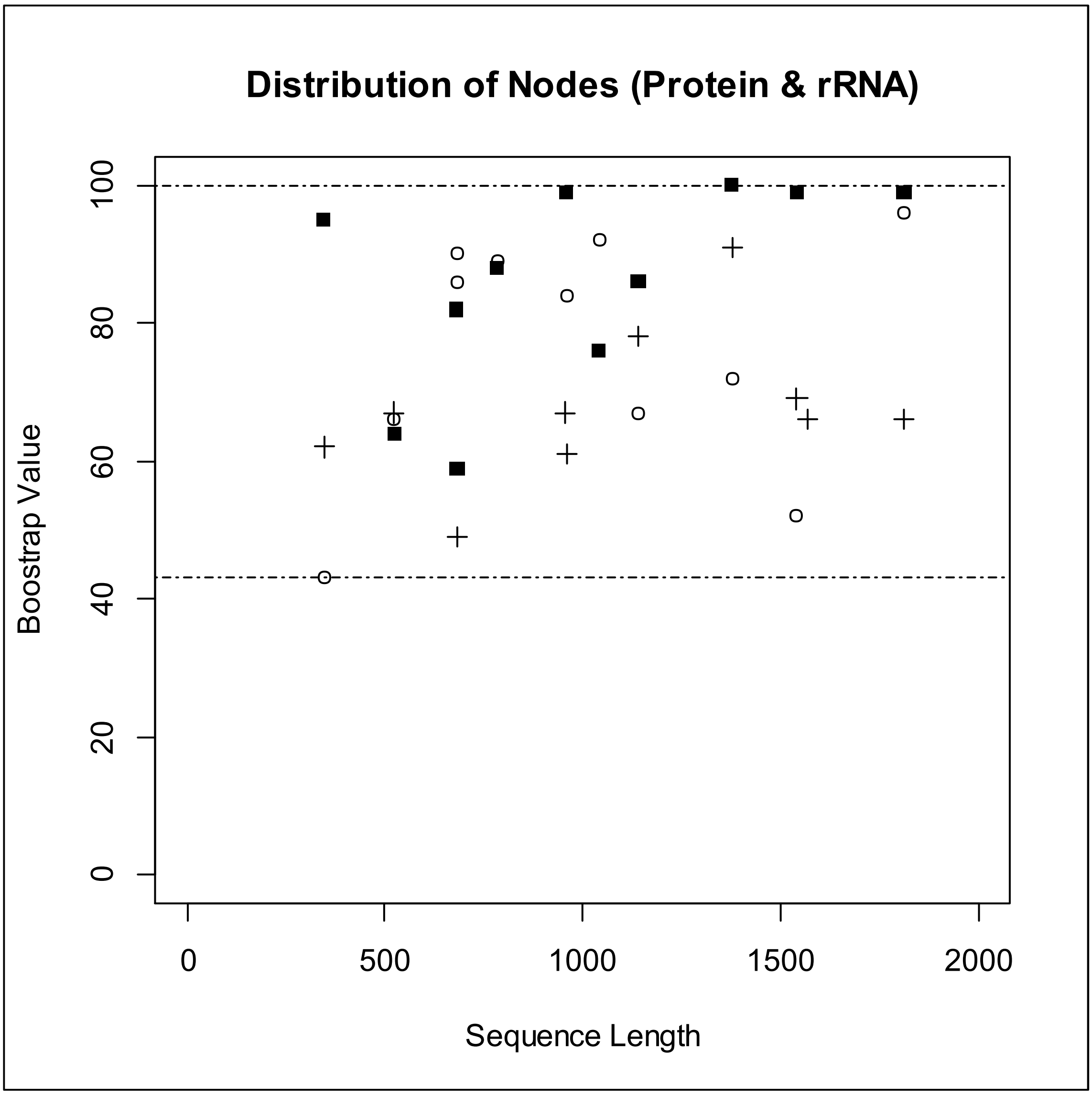


Figure 4.6 The distribution of bootstrap values against sequence length for node A (represented by ○), node B (represented by ■) and node C (represented by +) using protein and rRNA genes.

Figure 4.6 shows the distribution of the bootstrap value against the sequence length. The difference of the highest bootstrap support (100%) with lowest bootstrap value (43%) is 57% thus the variation is quite sizeable.

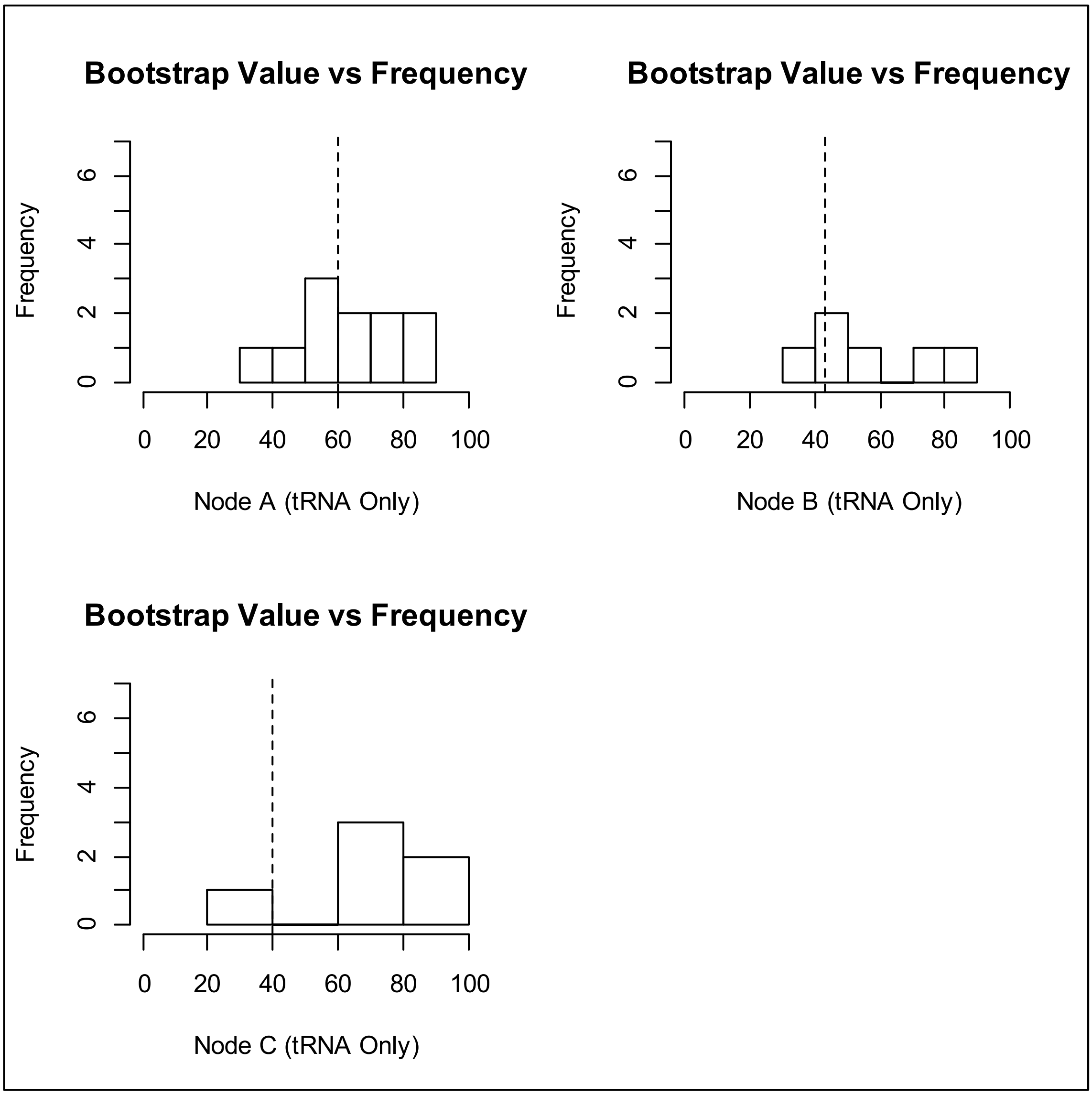


Figure 4.7 Histogram of frequency of bootstrap support for (a) Node A, (b) Node B and (c) Node C for tRNA genes. The dotted line reflect the empirical estimate from all genes

Similarly to the histogram of protein and rRNA, the empirical node support tends to be much smaller than the bootstrap support values.

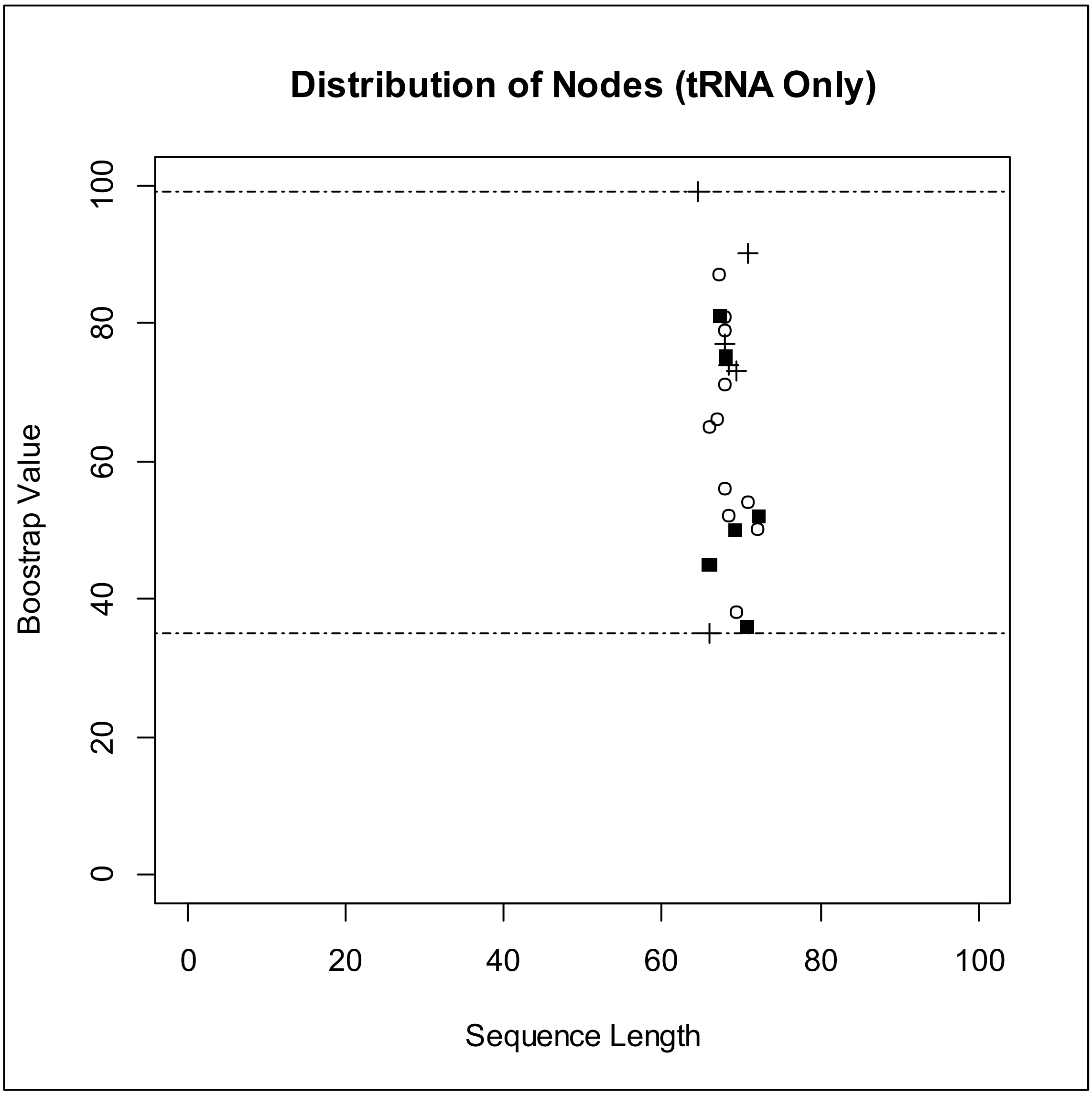


Figure 4.8 The distribution of bootstrap values against sequence length for node A (represented by ○), node B (represented by ■) and node C (represented by +) using tRNA genes only

Figure 4.8 shows that the distribution of bootstrap values for tRNA genes (centered around sequence length of 68) is quite variable. The difference of the highest bootstrap value (99%) with lowest bootstrap value (35%) is 64%, wider than that seen for protein and rRNA genes.

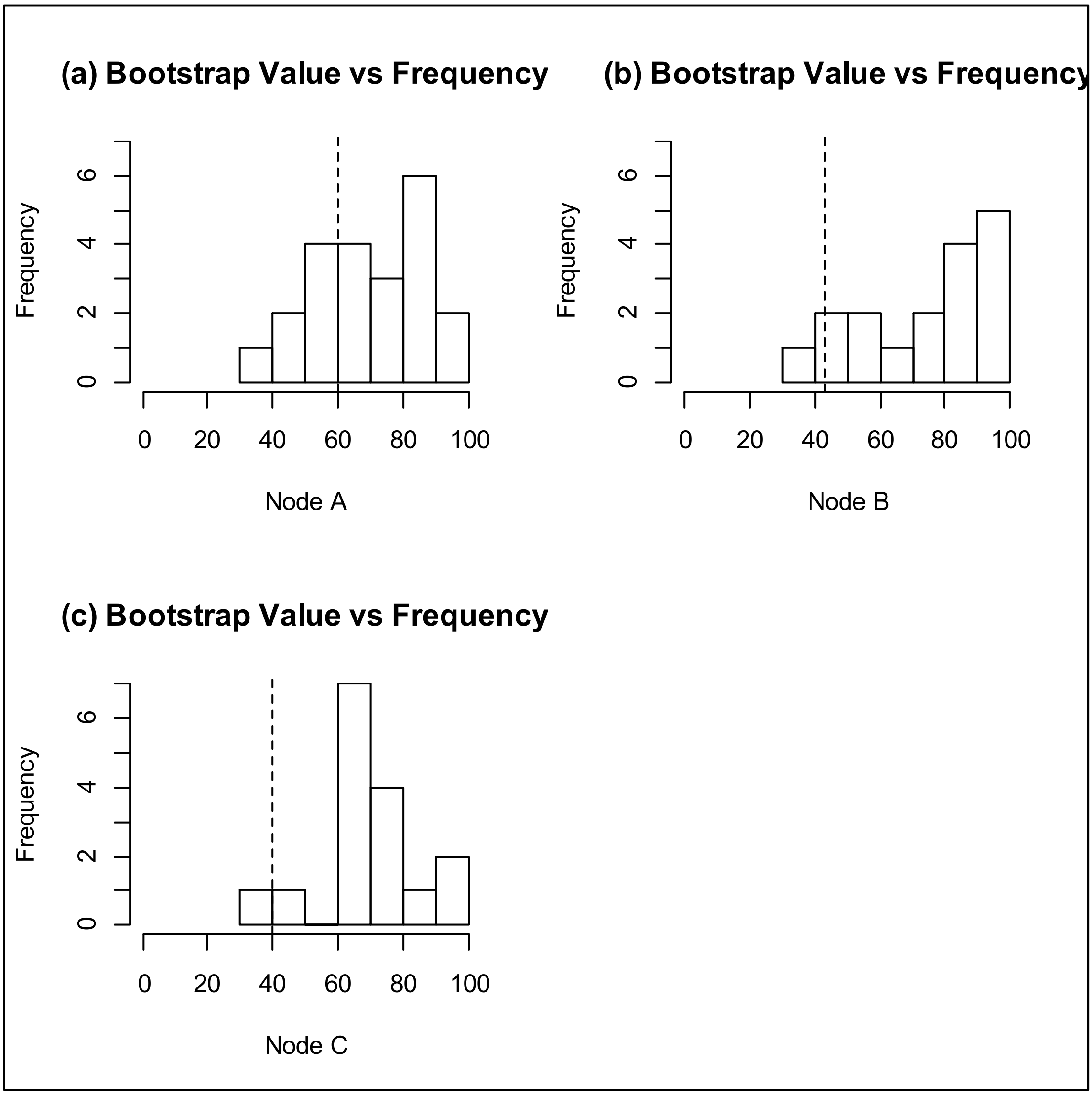


Figure 4.9 Histogram of frequency of bootstrap support for (a) Node A, (b) Node B and (c) Node C for all genes. The dotted line reflect the empirical estimate from all genes

As mentioned, Figure 4.9 also shows a tendency of bootstrap values skewed to the right of the empirical node support.

Further calculation can be done where we can find the percentage of the obtained bootstrap support compared with the empirical estimates. Using the formula as follows, we can calculate the result:

$$\frac{\text{Total number of non-zero bootstrap values} < \text{empirical estimate}}{\text{Total number of non-zero bootstrap values}}$$

The probability of a bootstrap support value falls on or above the empirical estimate for node A is 32% while node B is 6% and node C is 6%. As we move further away from the tip, the percentage of bootstrap values tend to drop and only 6% in the other two nodes, B and C.

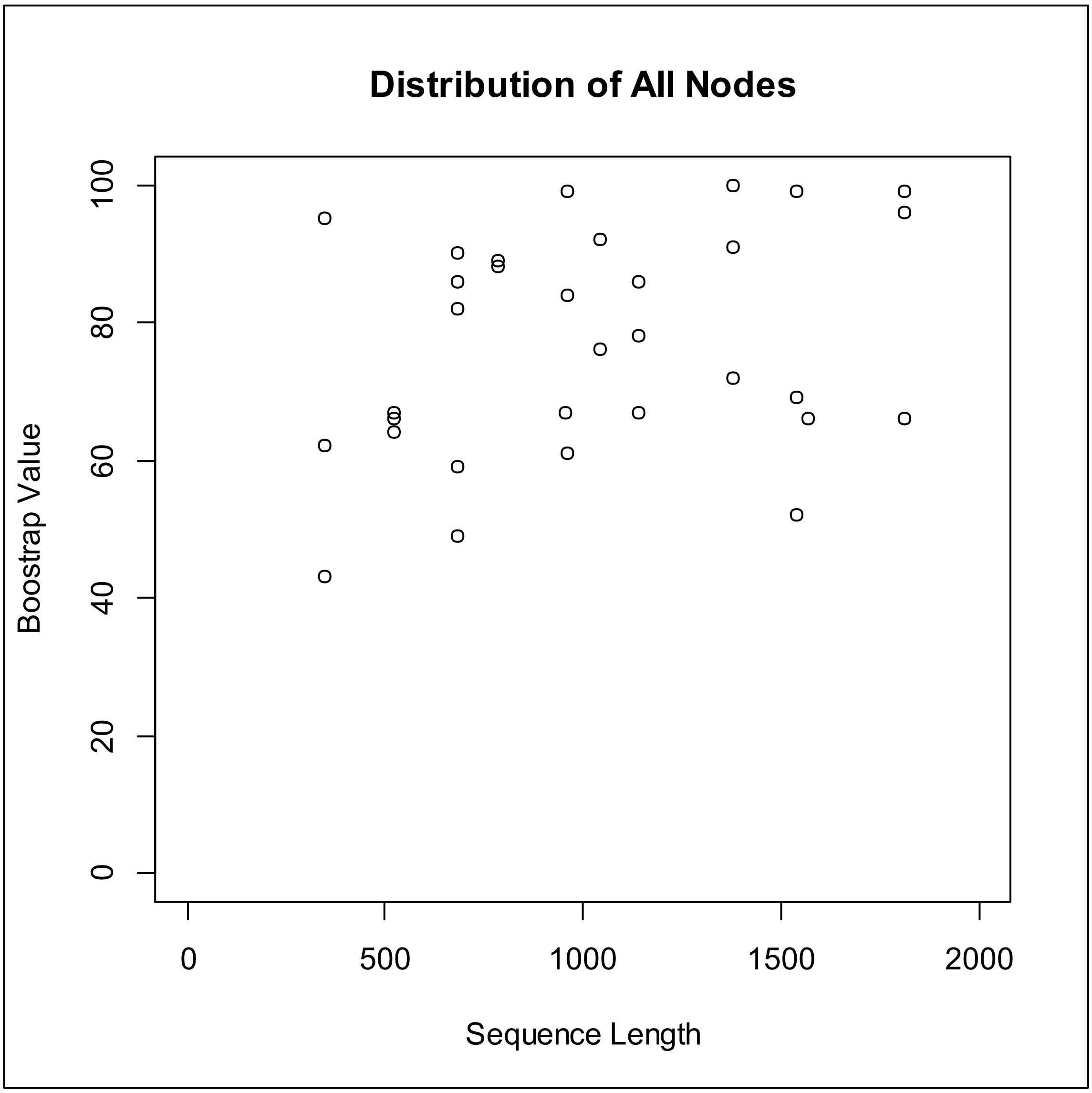


Figure 4.10 The distribution of the nodes from bootstrap value against sequence length for protein and rRNA (represented by ○) and tRNA (represented by ■) using all genes

From the linear regression of bootstrap values against sequence length, the slope coefficient is almost 0, 0.01 (p-value = 0.12) which is statistically significant. Sequence length and bootstrap value seemed and was not to be independent of each other.

Chapter 5

DISCUSSION

A full genome tree is very hard to do as only limited organisms have well-curated and completely sequenced genome available. A fully sequenced genome does not bring much value as we will need it to be fully curated to maximize the number of genes available for analysis within the genome. We only select the overlapping genes from each taxa but the intersection between the organisms will become smaller and eventually becomes null set as the number of taxa increases. Only a well curated genome from a well define phylogeny can be used. The mitochondrial genome was chosen because it is separated from the nuclear genome and it can be found in all organisms. Furthermore the gene content in mitochondrial genome is similar among metazoan (Boore *et. al.*, 1998) thus maximizing the chance of overlapping genes.

Even with automation in PyCogent to extract additional genes from the taxa, the condition of the NCBI database hinders this. There is minimal uniformity in the naming convention such as title and also the parameter such as keyword. The title of same gene or protein differs from one sequence to another. For example the result of searching for “cytochrome c” in the nucleotide database resulted in “Pan paniscus cytochrome c (CYCS) gene, complete cds” and “Chicken cytochrome c gene, allele CC10, complete CDS”. Both titles lack uniformity. Besides that the search mechanism of NCBI depends on both the keywords and title that are not standardized and may return inaccurate result.

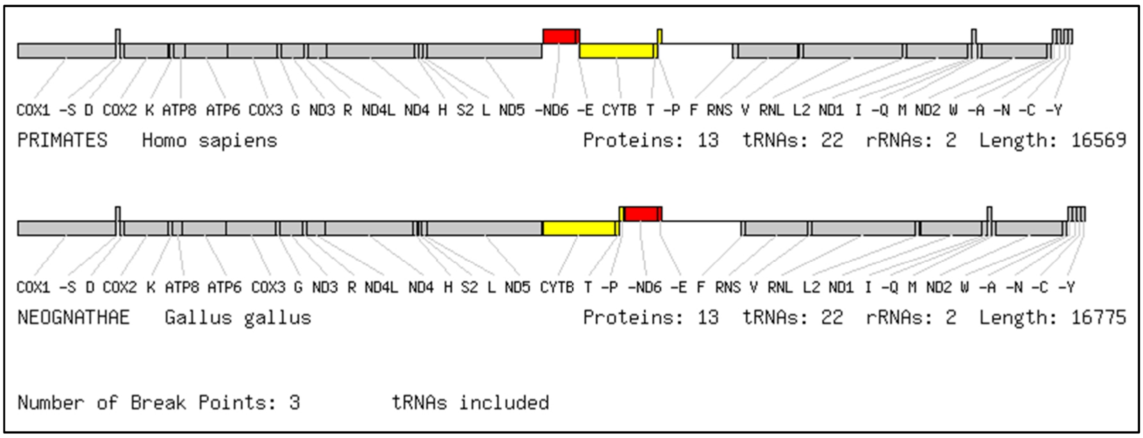
There are many reasons why the frequencies of the node differed between MEGA and R even though the parameters used were the same. In MEGA additional parameters were asked from the user such as the Nearest-Neighbour-Interchange (NNI) and specification of the type of the sequence – vertebrate, plant etc (Tamura *et. al.*, 2011). These small modifications may thus cumulatively affect the outcome of the final tree.

In R the tree is constructed directly through the use of parameters in the method of the APE and phangorn, with limited function arguments. This is because the packages are optional and not bundle into the core of R and it will be difficult to have constant updates. It will be easier in MEGA as it is packaged together as a software bundle and its component can be updated regularly as it shipped together with the core program.

The result of the regression of bootstrap values against sequence length suggest that equal weighting of each length as there seemed to be no biased towards larger bootstrap values for longer sequences.

One of the concerns from the study is that the genes evolution and species evolution may not happen at the same rate (Maddison, 1997). That is why some of the genes tend to have different result compared with the whole mitochondrion genome. However when dealing with more reserved genes which is very crucial for the survival of an organism, we can get a more correct tree. We are supposing that the true species tree can be represented by an average of the evolutionary histories of the gene tree. The frequencies of the faster mutating genes will be cancelled out by the slower mutating genes to obtain the average and the true tree.

As mentioned, the gene order can be used as a phylogeny tool (Boore *et. al.*, 1998). In Figure 5.1, the order of gene Cytochrome b and ND6 gene (along with some tRNAs) is different between *Homo sapiens* and *Gallus gallus*. It is reflected in the mitochondrion genome tree where *Gallus gallus* is shown as an outgroup.



COX1 -S D COX2 K ATP8 ATP6 COX3 G ND3 R ND4L ND4 H S2 L ND5 -ND6 -E CYTB T -P F RNS V RNL L2 ND1 I -Q M ND2 W -A -N -C -Y

PRIMATES Homo sapiens Proteins: 13 tRNAs: 22 rRNAs: 2 Length: 16569

COX1 -S D COX2 K ATP8 ATP6 COX3 G ND3 R ND4L ND4 H S2 L ND5 CYTB T -P -ND6 -E F RNS V RNL L2 ND1 I -Q M ND2 W -A -N -C -Y

NEOGNATHAE Gallus gallus Proteins: 13 tRNAs: 22 rRNAs: 2 Length: 16775

Number of Break Points: 3 tRNAs included

Figure 5.1 Gene order comparison of mitochondrion genome of *Homo sapiens* and *Gallus gallus* (Source: OGRe Genome Viewer)

The result may not reflect the whole story as human themselves contains more than 20,000 genes but we just use 37 from mitochondrion genome only. According to Venter (2001), there are 26,588 protein-encoding transcripts versus the 37 protein coding gens in mtDNA which takes only a mere 0.14%. However this is a very good estimate of the validity of the bootstrap support function in the phylogenetics. This is because we use all the genes from mtDNA which represent the population and not a virtual build-up of the sequences and also the tree used is supported not only by genes but also other traits as well such as morphology and behaviour.

Even though a phylogenetic tree is built, the true relationship and phylogeny of the taxa may not be reflected in the result. This is because in order to deduce the true phylogeny more factors need to be taken into account. These include morphology and also the behaviour (Queiroz *et. al.*, 2009). For example, according to the result, *Homo sapiens* and *Pan paniscus* is the closer compared to *Gorilla gorilla* in the tree. If we compare the morphology between the three species, human and chimpanzee are more similar in both size and overall structure. Gorilla is much more robust and its jaw protrudes more compared to human. In behaviour, both human and chimpanzee use tools regularly to complete tasks. This shows a different degree of intelligence compared to those to do so.

With advancement in technology, the numbers of simultaneous tasks handled by computers far exceed the human capability. This include the computational power of both calculation and retrieval ability of a computer. Additionally we can easily retrieve and align sequence multiple genes from different organism quickly which will generate an ample amount of data. However this may not be feasible in Felsenstein days as during his time in 1990s when the human genome was only officially initiated in 1990 and continued for more than 10 years (Venter *et. al.*, 2001) and the overlapping genes available across taxa were rather limited. Due to the limiting resources, the use of 'virtual' method such as bootstrapping where assumption that it does exist in real world is needed to infer something that may not be found yet. However now that the huge resources from publication, database and annotation are available, the assumption of bootstrapping need to be challenged to bring molecular phylogenetics studies to a higher level.

One of the methods to further justify the empirical estimate is that by using the consensus of the identity value from each gene. This will strengthen and increase the confidence from the constructed trees. This can be covered in further studies.

Chapter 6

CONCLUSIONS

The concept of assessing the robustness of phylogenetic tree using bootstrap support values for nodes appears to rest on shaky foundations. Using the mitochondrial genome, it was found that bootstrap support values were generally much higher than node support determined empirically using all 37 genes from the mitochondrial genome. Even though there are doubts such as the different evolutionary rate in genes compared to species evolutionary rate, we can assume that the average will return the true tree especially on a larger sample size such as a whole nuclear genome which can be focused on future studies. There is a sense of urgency to further explore the dissonance between what bootstrap support purports to offer and what empirically determined results say otherwise.