Human Family Tree
Mitochondrial Eve

Gopal Menon
Computer Science Department
Utah State University
Logan, Utah 84322
Email: gopal.menon@aggiemail.usu.edu

***Abstract*—Cann, Stoneking and Wilson as described in their paper [2] on Mitochondrial DNA and human evolution used 147 human mitochondrial DNA (mtDNA) samples, and performed restriction mapping on each sample. The restriction sites were mapped using known human mtDNA sequences. The number of restriction site differences between each pair of individuals gave them the extent of the nucleotide sequence divergence. mtDNA is inherited from the mother and unlike nuclear DNA, does not go through combination of DNA from both the parents. All living human beings are thought to have descended from one common ancestor [3]. This is not to be thought of as a specific person, since whenever an ancient mtDNA branch dies out, the most recent common ancestor (MRCA) moves to a more recent female ancestor. The restriction site differences between all the pairs of individuals were used to induce a genealogical tree shown in figure 1. This tree suggested that human beings originated in Africa and then settled the rest of the planet as shown in figure 3. The aim of the project is to start with mtDNA samples from a varied group of people and compute the alignment distance matrix for all pairs of samples. Using this matrix, an evolutionary tree and hierarchical clustering algorithm needs to be implemented in order to induce a tree similar to the one shown in figure 1.**

I. Biological problem and its significance

mtDNA is comprised of 16,569 base pairs and is present only in cells in the animal kingdom. Figure 2 shows a graphic of mtDNA within a cell. It is contained in almost every cell of the body and is inherited only from the mother. In this regard, it is different from nuclear DNA that consists of genetic material from both parents. mtDNA is thought to be a bacterium like organism that had been captured by a cell. It does not have the ability for error checking during replication and is more susceptible to mutations than nuclear DNA. Restriction site differences or the edit distance between two samples can be used as an estimate on how many years ago was the common ancestor alive. Individuals with low edit distances between their mtDNA will have a more recent common ancestor. When humans migrate to a particular region, after a period of time, that region will have many humans with one of the migrated human females as a common ancestor. This common ancestor will be be more recent than the common ancestor for a descendant of one of the migrated humans and a descendant from the original population that did not migrate. mtDNA edit distance can be thus used to induce genealogical trees of the form shown in figure 1. This can lead to an understanding of the origins of populations of certain geographic regions and human migration paths. The paper written by Cann et. al. suggests

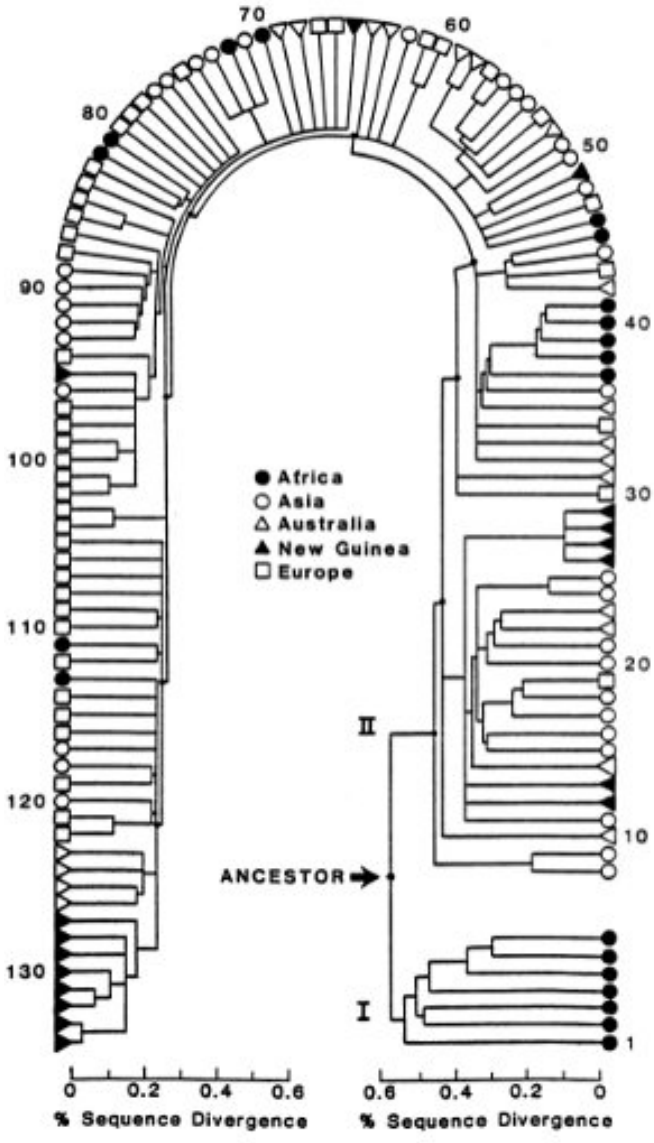

Fig. 1. Human geneological tree based on mtDNA samples [2]

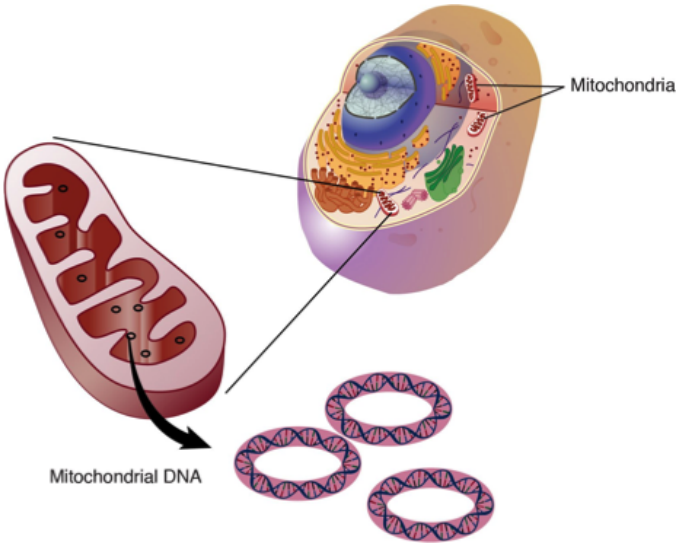

Fig. 2. Mitochondrial DNA[1]

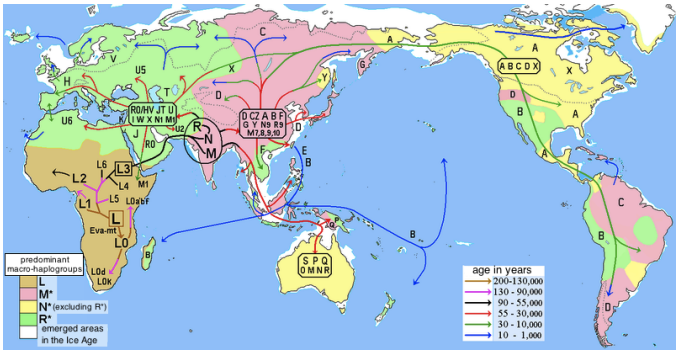

Fig. 3. Human migrations and mitochondrial haplogroups. The letters stand for haplogroups, which are groups of people who share a common ancestor [6].

that human beings originated in Africa since the common ancestor shown in figure 1 would need to be African in order to minimize the number of intercontinental migrations needed to account for the geographic distribution of mtDNA types.

The human mtDNA molecular clock is the rate at which mutations have been accumulating in the mitochondrial genome. The mutation rate differs between coding and non-coding regions of the mtDNA. Coding region mutations may be fatal or may lead to other complications and due to this reason, mutations in this region will lead to purifying selections. Some studies avoid coding region mutations, while other studies consider both regions and apply a correction factor for selection in coding region.

II. Other Similar Studies

A. Neanderthal DNA Sequences and Origin of Modern Humans

As part of the study [8], DNA material was extracted from a Neanderthal specimen found in 1856 in western Germany. A hitherto unknown mitochondrial mtDNA sequence was found in the specimen. Sequence comparisons with human mtDNA sequences, as well as phylogenetic analyses, showed that the Neanderthal sequence fell outside the variation of modern humans. Furthermore, the age of the common ancestor of the Neanderthal and modern human mtDNAs was estimated to be four times greater than that of the common ancestor of human mtDNAs. This suggested that Neanderthals went extinct without contributing mtDNA to modern humans as an ancestor.

B. An early modern human from Romania with a recent Neanderthal ancestor

Neanderthals are thought to have disappeared in Europe 39,000-41,000 years ago but they have contributed one to three percent of the DNA of present-day people in Eurasia. In the study [9], they analyzed DNA from a 37,000-42,000-year-old modern human from Pestera cu Oase, Romania. Although the specimen contained small amounts of human DNA, they used an enrichment strategy to isolate sites that were informative about its relationship to Neanderthals and present-day humans. They found that on the order of six to nine percent of the genome of the Oase individual was derived from Neanderthals, more than any other modern human sequenced to date. Three chromosomal segments of Neanderthal ancestry were over 50 centimorgans in size, indicating that this individual had a Neanderthal ancestor as recently as four to six generations back. However, the Oase individual did not share more alleles with later Europeans than with East Asians, suggesting that the Oase population did not contribute substantially to later humans in Europe.

III. Computer Science Problem

A. Edit Distance

The edit distance between two mtDNA sequences can be found using Dynamic Programming. Since only the edit distance score is needed, the computation can be done using an array of two columns with the number of rows equal to the length of a mtDNA sequence. For 2699 samples (see figure 4), 3.6 million edit distance computations would need to be done. Even with C++, this amount of computation would take too long to execute. In order to complete the project, a subset of the samples was used. A penalty of 1 was used for insertion, deletion and substitution during edit distance computation.

B. Clustering

The evolutionary tree was found out by running clustering on the mtDNA samples using the edit distance as the input to clustering. mtDNA samples with smaller edit distances would form clusters. These would be clusters corresponding to leaves of the evolutionary tree with a common ancestor. The clusters would then be combined based on the distance between the cluster centroids. This would give us common ancestors of the initial clusters. The process would be repeated till we find a common ancestor for all the mtDNA samples.

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm [5] was used for heirarchical clustering of mtDNA samples. The length of an edge between two clusters will be the difference in heights between them and will be a measure of their separation in years. Given two clusters C_1 and C_2, the UPGMA algorithm defines the distance between them to be the average pairwise distance:

$$D(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i \epsilon C_1} \sum_{j \epsilon C_2} D(i, j)$$

Algorithm 1 shows the details of the algorithm and it is called with parameters D, the edit distance matrix and n the number of elements.

Algorithm 1 UPGMA Algorithm

procedure UPGMAALGORITHM.$UPGMA(D, n)$
  Form n clusters, each with a single element
  Construct graph T by assigning isolated vertex to each cluster
  Assign height $h(v) = 0$ to every vertex v in graph
  while there is more than one cluster **do**
    Find two closest clusters C_1 and C_2
    Merge C_1 and C_2 into a new cluster C with $|C_1|+|C_2|$ elements
    for every cluster $C^* \neq C$ **do**
      $D(C, C^*) = \frac{1}{|C||C^*|} \sum_{i \epsilon C} \sum_{j \epsilon C^*} D(i, j)$
    end for
    Add a new vertex C to T and connect to vertices C_1 and C_2
    Vertex height $h(C) = \frac{D(C_1, C_2)}{2}$
    Assign length $h(C) - h(C_1)$ to the edge (C_1, C)
    Assign length $h(C) - h(C_2)$ to the edge (C_2, C)
    Remove rows and columns of D corresponding to C_1 and C_2
    Add a row and column to D for the new cluster C
  end while
  return T
end procedure

IV. Tasks

A. Download mtDNA sequences

A subset of the 2699 mtDNA sequences was downloaded into text files from the Human Mitochondrial Genome Database [4]. Furthermore, clustering was run on a further subset. The subset details can be seen by referring to the phylogenetic trees in this report.

B. Compute Edit Distance Matrix

The edit distance matrix for the subset of the samples was computed by finding out the edit distance for all possible pairs of mtDNA.

Area	Samples
Africa	287
North America	11
South America	14
Asia	922
Australia	34
Europe	1192
Melanesia	56
Micronesia	4
Middle East	45
Polynesia	9
South Asia	125
Total	**2699**

Fig. 4. Number of mtDNA samples from each geographic regions

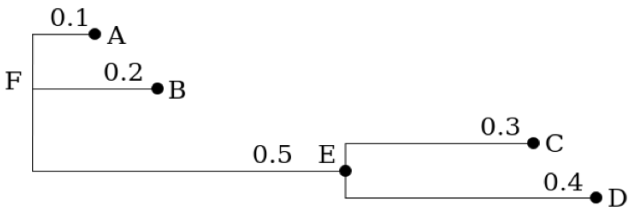

Fig. 5. Newick format illustration [7] .

C. Clustering and Evolutionary Tree

Clustering was done using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering algorithm as described above in order to create an evolutionary tree.

D. Evolutionary Tree Illustration

A script for showing the evolutionary tree was generated during clustering. The script was generated in the Newick format [7]. An example of this format is shown in figure 5. Nested parentheses are used for specifying heirarchy where node name is followed by colon and node height. The Newick format script for the tree shown in figure 5 is $(A : 0.1, B : 0.2, (C : 0.3, D : 0.4) : 0.5)$.

V. Results

A. Africa Evolutionary Tree

The tree is shown in figure 6. The number 9.99 at the bottom left of the tree is the scale and represents the edit distance. We can see that groups with large separation (with a less recent common ancestor) are all from sub-saharan Africa. They appear to be the humans whose common ancestor is Mitochondrial Eve. People who live near to each other appear to have a more recent common ancestor like in the case of Morocco and Canary Islands

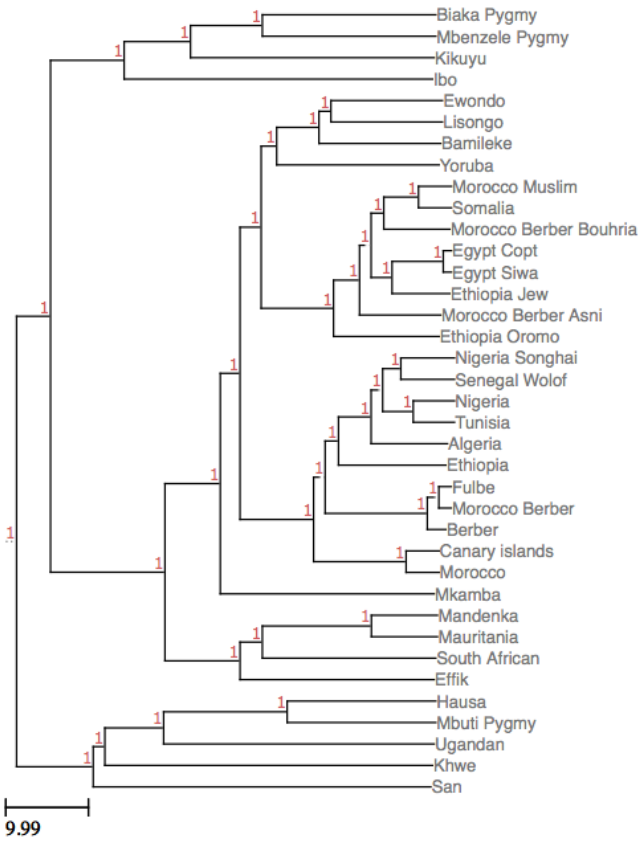

Fig. 6. Africa Phylogenetic Tree

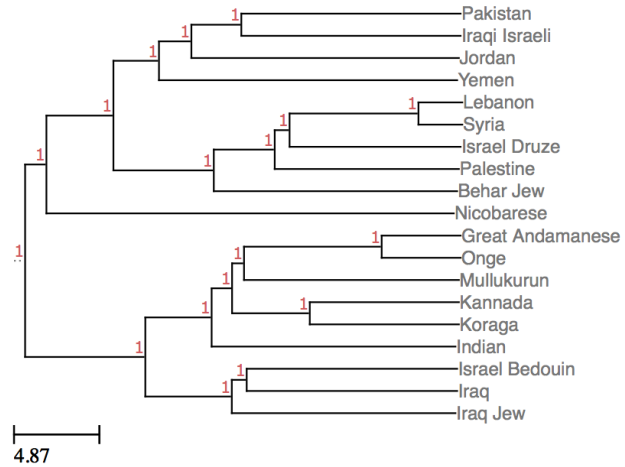

Fig. 7. South Asia and Middle East Phylogenetic Tree

B. South Asia and Middle East Evolutionary Tree

The tree is shown in figure 7. Mainland Indians (figure 8) are closer to Onges (figure 9) than they are to Iraqis (figure 10). Outward features are thought to be a more recent development from around 50,000 years ago [5] and these cannot be used a guide to gauge the closeness of relations between peoples.

Fig. 8. Kal Penn - Actor of Gujrati Indian ethnicity

Fig. 9. Woman of Onge Indian ethnicity

Fig. 10. Nouri Al-Maliki - ex Prime Minister of Iraq

C. Asia Evolutionary Tree

The tree is shown in figure 11. Japanese and Han Chinese are very closely related. However their culture and customs are very different. Khirgiz and Korean mtDNA samples were

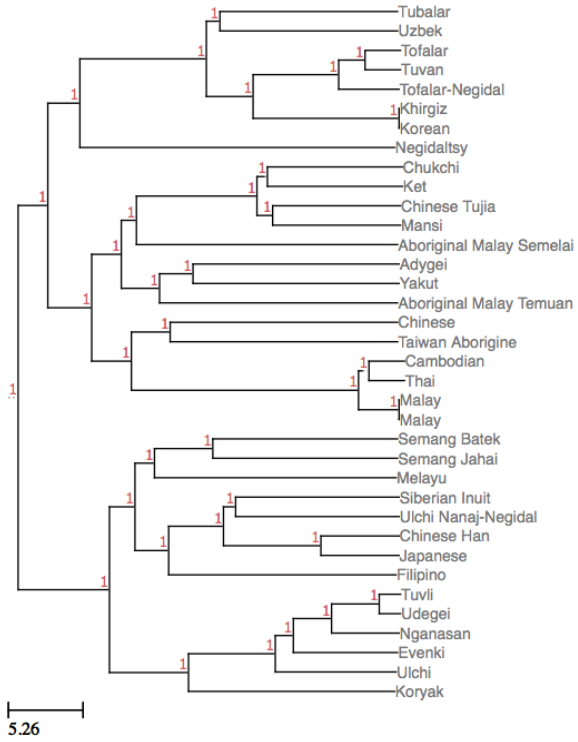

Fig. 11. Asia Phylogenetic Tree

identical and this is the reason these two sample show a very recent common ancestor. This was a problem with the data.

D. Europe Evolutionary Tree

The tree is shown in figure 12. People from Spain are not very close in terms of a recent common ancestor. This can be seen in Spanish samples from Andalusia, Leon, Galicia and Maragato. Similar is the case with Italians. French people are closer to Crimean Tatars than to Europeans from closer regions.The European American seems to have a matrilineal ancestor from Southern Italy.

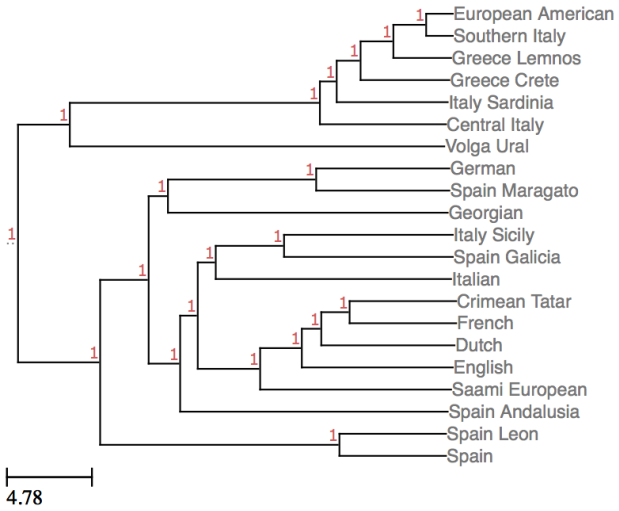

Fig. 12. Europe Phylogenetic Tree

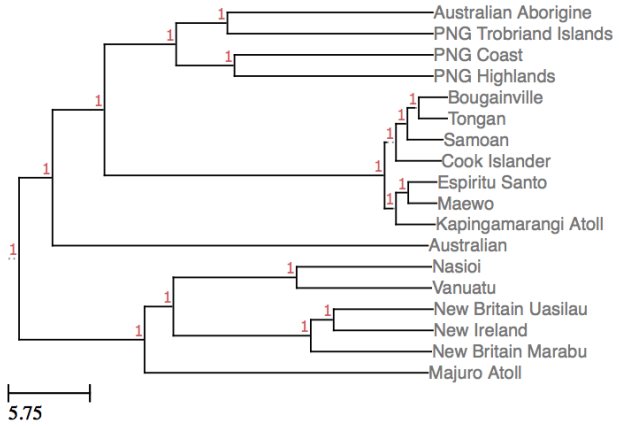

Fig. 13. Oceania and Pacific Islands Evolutionary Tree

E. Oceania and Pacific Islands Evolutionary Tree

The tree is shown in figure 13. Papua New Guinea (PNG) people seem to be further apart from each other than other groups in the tree with a common ancestor. PNG is on the eastern side of one island and is not a very large country.

F. North and South America Evolutionary Tree

The tree shown in figure 14. North American and Guarini (from different continents) samples are closer than North American and Navajo (from the same continent).

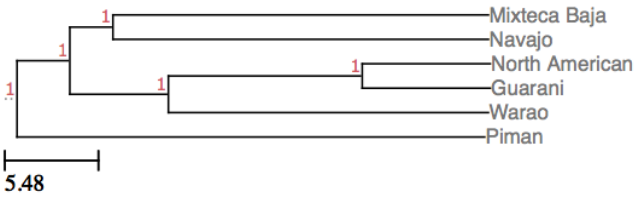

Fig. 14. North and South America Phylogenetic Tree

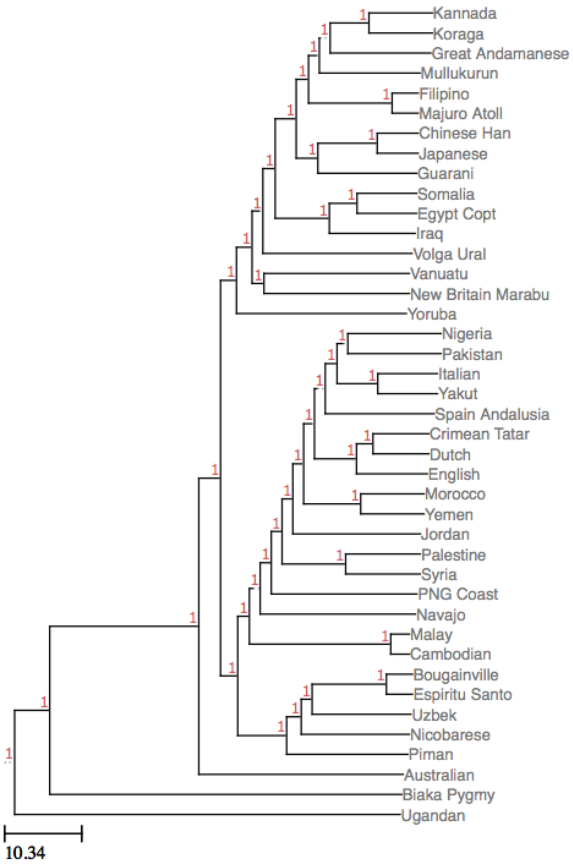

Fig. 15. World Phylogenetic Tree

G. World Evolutionary Tree

The tree is shown in figure 15. People are mostly grouped by geographic region. Some unexplained cases are Guarini (from South America) and Navajo (from North America) are in separate branches altogether. I would have expected them to be closer. Nicobarese (from Indian islands) and Piman (from North America) have a comparatively recent common ancestor even they are on different continents.

VI. Summary and Conclusions

From the phylogenetic trees that were created, we can see that humans from a particular geographic region usually have a more recent common ancestor. However there are cases where mtDNA samples from regions that were far apart, were seen to be closer than the ones where the regions were closer. I am not sure why this is the case. But it is possible that in these cases, the individual had a mixed ancestry and/or lost the details on their matrilineal lineage. In places like Papua New Guinea and Spain, we can see that within a small geographic region, there is a large separation between the samples and their common ancestor. This is possibly due to geographic isolation that prevented the people from interacting with people from other regions. As we saw in the case of the mainland Indians, Onges and Iraqis, outward features for people closer in terms of a recent common ancestor, may be very different. Outward features for people with a more ancient common ancestor may be more similar.

The world family tree in figure 15 did not help me in constructing a human migration as shown in figure 3.

References

[1] "Khan Academy." *Khan Academy*. N.p., n.d. Web. 10 Aug. 2016.
[2] Cann, L. C., M. Stoneking, and A. C. Wilson. "Mitochondrial DNA and human evolution." *Nature* 325 (1987): 1-5.
[3] "Mitochondrial Eve." *Wikipedia*. Wikimedia Foundation, n.d. Web. 24 July 2016.
[4] Ingman, M. & Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34, D749-D751 (2006).
[5] Jones, Neil C., and Pavel Pevzner. *An Introduction to Bioinformatics Algorithms.* Cambridge, MA: MIT, 2004. Print.
[6] "Human Mitochondrial DNA Haplogroup." *Wikipedia*. Wikimedia Foundation, n.d. Web. 25 July 2016.
[7] "Newick Format." *Wikipedia*. Wikimedia Foundation, n.d. Web. 08 Aug. 2016.
[8] Krings, Matthias, et al. "Neandertal DNA sequences and the origin of modern humans." cell 90.1 (1997): 19-30.
[9] Fu, Qiaomei, et al. "An early modern human from Romania with a recent Neanderthal ancestor." *Nature* 524.7564 (2015): 216-219.

VII. Appendix

A. Running the program

The program runs clustering on the mtDNA samples that are listed in the text file *fileListing.txt*. This text file must contain the names of the downloaded mtDNA sample files that are in text format. The mtDNA samples should have nucleotide strings without embedded spaces. They can be separated by line feeds. The input data should be in the format that the human mtDNA samples are in the mtDNA database [4].

B. Obtaining the tree script

The program will create the script in file *WorldTree-Script.txt*. This file can be uploaded to the website http://etetoolkit.org/treeview/ for viewing the tree.