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 [1] 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 [2]. 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 2. 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 1.**

I. Biological problem and its significance

Human mtDNA is comprised of 16,569 base pairs. It is contained in 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 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 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

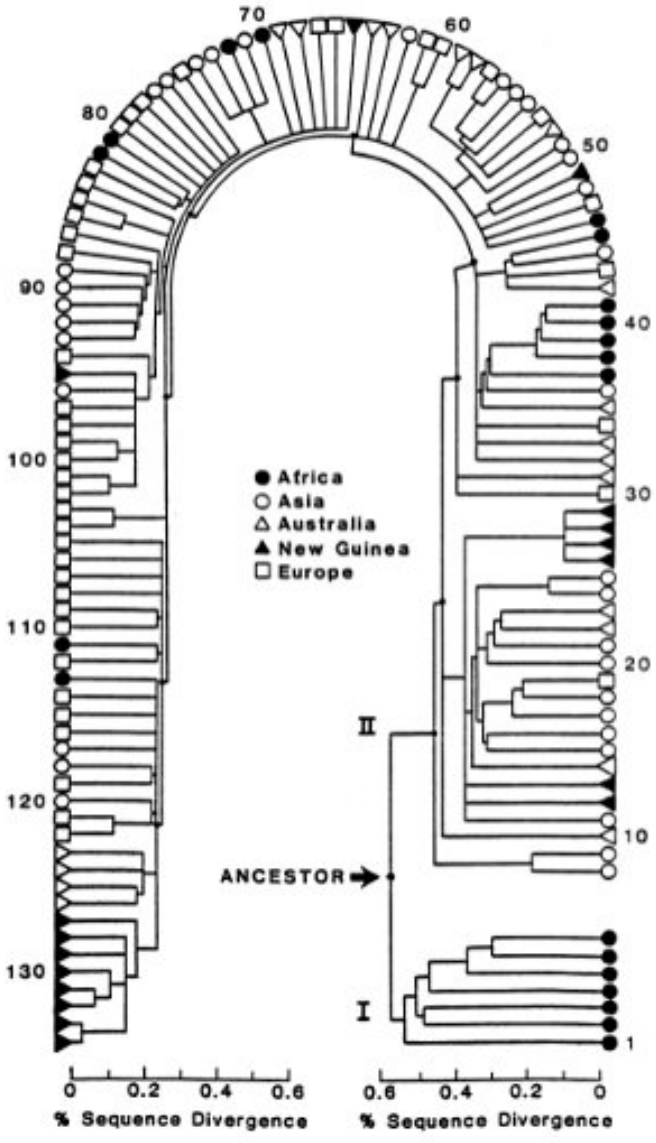

Fig. 1. Human geneological tree based on mtDNA samples

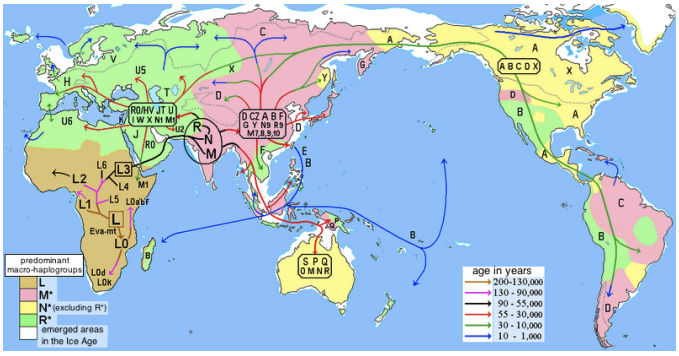

Fig. 2. Human migrations and mitochondrial haplogroups. The letters stand for haplogroups, which are groups of people who share a common ancestor.

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. 3. Number of mtDNA samples from each geographic regions

types.

II. 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 3), 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, either the number of samples would need to be reduced or the project would need to be executed on a parallel computation environment using a Map-Reduce or equivalent framework.

B. Clustering

The evolutionary tree would need to be 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.

III. Project Plan

A. Download mtDNA sequences

2699 mtDNA sequences will be downloaded into text files from the Human Mitochondrial Genome Database [3].

B. Compute Edit Distance Matrix

The edit distance matrix will be computed by finding out the edit distance using all possible pairs of mtDNA. As of this point of time, the computation involved is very large and either a subset of the database will be used or access to a parallel computing environment will need to be obtained.

C. Clustering and Evolutionary Tree

Clustering will be done using the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) clustering algorithm as described in [4] in order to create an evolutionary tree.

D. Evolutionary Tree Illustration

Once the data for the evolutionary tree is available, Java graphical user interface components will be used for showing the tree.

IV. Schedule

Work Item	Target Completion
Complete project proposal presentation	July 25, 2016
Finalize computing platform	July 26, 2016
Download mtDNA sequences	July 27, 2016
Complete project proposal report	July 27, 2016
Compute Edit Distance Matrix	July 31, 2016
Compute Evolutionary Tree	August 7, 2016
Evolutionary Tree GUI	August 9, 2016
Final report and presentation	August 11, 2016

References

[1] *Nature.com*. Nature Publishing Group, n.d. Web. 25 July 2016.
[2] "Mitochondrial Eve." *Wikipedia*. Wikimedia Foundation, n.d. Web. 24 July 2016.
[3] Ingman, M. & Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res 34, D749-D751 (2006).
[4] Jones, Neil C., and Pavel Pevzner. *An Introduction to Bioinformatics Algorithms*. Cambridge, MA: MIT, 2004. Print.