Expert Systems with Applications 39 (2012) 89–98

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Clustering based distributed phylogenetic tree construction

Esra Ruzgar *, Kayhan Erciyes

*Computer Eng. Dept., Izmir University, Gursel Aksel Bulvari, 14, Uckuyular 35350, Izmir, Turkey*

ARTICLE INFO

*Keywords:*
Y-DNA Haplogroups
Clustering
Phylogenetic tree
Neighbor-joining

ABSTRACT

Phylogenetic tree construction has received much attention recently due to the availability of vast biological data. In this study, we provide a three step method to build phylogenetic trees. Firstly, a density-based clustering algorithm is used to provide clusters of the population at hand using the distance matrix which shows the distances of the species. Secondly, a phylogenetic tree for each cluster is constructed by using the neighbor-joining (NJ) algorithm and finally, the roots of the small phylogenetic trees are connected again by the NJ algorithm to form one large phylogenetic tree. To our knowledge, this is the first method for building phylogenetic trees that uses clustering prior to forming the tree. As such, it provides independent phylogenetic tree formation within each cluster as the second step, hence is suitable for parallel/distributed processing, enabling fast processing of very large biological data sets.

The proposed method, clustered neighbor-joining (CNJ) is applied to 145 samples from the Y-DNA Haplogroup G. Distances between male samples are the variation in their set of Y-chromosomal short tandem repeat (STR) values. We show that the clustering method we use is superior to other clustering methods as applied to Y-DNA data and also independent, fast distributed construction of phylogenetic trees is possible with this method.

## 1. Introduction

A phylogenetic tree or an evolutionary tree is a graph with no cycles (tree) that shows the evolutionary relationships among various biological species based on their genetic closeness. Three of most widely used methods in the construction of a phylogenetic tree based on some optimization criteria of the tree to be formed are the maximum parsimony (MP), maximum likelihood (ML) and minimum evolution (ME). Maximum parsimony approach (Eck & Dayhoff, 1966; Fitch, 1971) examines all possible topologies or a certain number of topologies that are likely to close true tree and chooses one that shows the smallest amount of total evolutionary change as the final tree. Maximum likelihood approach (Felsenstein, 1981) tries to estimate trees by formulating a probabilistic model of evolution and applying known statistical methods. It involves finding that evolutionary tree which yields the highest probability of evolving the observed data. Minimum evolution approach (Edwards & Cavalli-Sforza, 1963) searches for the tree that minimizes total branch lengths.

The phylogenetic tree construction can further be classified into two groups based on the inputs utilized: matrix methods and sequence methods. The former group, which uses a distance matrix of genetic measure, includes the unweighted pair-group method using arithmetic averages (UPGMA) (Sokal & Michener, 1958), the minimum evolution method of Cavalli-Sforza and Edwards (1967), the distance Wagner method (Farris, 1972), the modified Farris method (Tateno, Nei, & Tajima, 1982), the neighbor-joining (NJ) algorithm of Saitou and Nei (1987), the median-joining (MJ) networks method (Bandelt, Forster, & Rohl, 1999) and others. On the other hand, the latter group, which utilizes amino acid or nucleotide sequences directly, contains maximum parsimony methods of Eck and Dayhoff (1966), the maximum likelihood method of Felsenstein (1981), method of Tateno (1990), the reduced median (RM) networks method (Bandelt, Forster, Sykes, & Richards, 1995).

Male Y-chromosome bears a sequence of patterns of nucleotides called short tandem repeats (STRs) which are passed from father to son and remain unchanged for many generations. The set of STR values that is obtained for Y-chromosome markers is called a *haplotype*. By comparing STR values of two or more males, it is possible to obtain information about genealogical relationship between them. STR analysis is used in forensic science for genetic fingerprinting and also, because of the availability of reasonably priced Y-chromosome testing of STRs, popular recent applications of STR data is to estimate the paternal ancestry and migratory patterns of humans, where lost family relations due to wars or disasters may be found.

In this study, a three-step method to construct phylogenetic trees is proposed and applied to sample data from Y-DNA Haplogroup (Y-DNA Haplogroup G Project, 2010). Having calculated pairwise distances between all samples and generated the distance

* Corresponding author. Tel.: +90 232 2464949; fax: +90 232 2240909.
*E-mail address:* esra.ruzgar@izmir.edu.tr (E. Ruzgar).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.06.049

matrix, firstly, samples are divided into several clusters. Secondly, trees are constructed for each cluster by using neighbor-joining (NJ) algorithm for each cluster independently and finally, the big tree is formed by the NJ algorithm, treating each small phylogenetic tree root as an individual node. To the best of our knowledge, clustered NJ (CNJ) method is the first study that provides building of the clusters prior to the building of the phylogenetic tree. In this sense, CNJ method is suitable for parallel/distributed processing as NJ or any other tree building algorithm may be performed independently within each cluster. Our main contribution is the provision of this asynchronous processing which will allow processing of very large amounts of data over a distributed computer network such as the Grid or any general network. Such large amounts of data are very difficult to handle with the existing methods. Although many methods have been presented for constructing phylogenetic trees and networks from amino acid sequences, nucleotide sequences or gene frequencies, STR data is not widely used for phylogeny. Our second contribution is the application of this method to STR data for building the phylogenetic tree.

Several clustering algorithms are also compared to find the algorithm that fits best with the biological data. Rest of the paper is organized as follows. Section 2 gives brief background information about human DNA structure and Y-chromosome. In Section 3, all steps and details of phylogenetic tree construction with CNJ method are described. Section 4 concludes with summary and discussions.

## 2. Background

### 2.1. Human DNA structure

Deoxyribonucleic acid, DNA, is the genetic material that we inherit from our parents. The total collection of DNA for a single person or organism is referred to as its genome. DNA is a long string of nucleotide units attached to one another. In a single nucleotide, there exist three components: a sugar molecule, a phosphate group, and a nitrogenous base. There are four different types of bases: Adenine (A), Guanine (G), Cytosine (C) and Thymine (T). In a single nucleotide, the sugar is attached at one end to a phosphate group, and because the sugar of that nucleotide can attach to another phosphate at its other end, many nucleotides in a long chain can be tied together.

DNA has two sides or strands, and these strands are twisted together like a twisted ladder called the double helix. The nitrogenous bases point inward on the ladder and form pairs with bases on the other strand. Each base pair is formed from two complementary nucleotides (purine with pyrimidine) bound together by hydrogen bonds. The base pairs in DNA are Adenine with Thymine and Cytosine with Guanine.

### 2.2. Y-chromosome Haplogroups

A chromosome is an organized structure of DNA and located in cells. The human genome is composed of 23 kinds of chromosomes and every child receives two sets of 23 chromosomes, one from the mother and one from the father, for a total of 46 chromosomes. One pair of chromosomes, called the sex chromosomes, is responsible for determining sex and the remaining 22 pairs are called autosomes. In male genome, sex chromosome pair is composed of one X-chromosome and one Y-chromosome and in female genome, it is composed of two X-chromosomes.

Y-chromosome is passed from father to son almost unchanged where only small changes called mutations occur during generations. Y-DNA STR data is used in genetic genealogy, which tries to find genetic relationships between individuals as well as in forensic science to identify the individuals. The genealogical Y-DNA testing involves looking at STR segments of DNA on the Y-chromosome. The number of repetitions of a sequence differs from one person to another person, particular number of repetitions is known as an *allele* of the marker. The variations in STR segments are caused by mutations that increase or decrease the number of repeats. An STR on the Y-chromosome is designated by a DYS number (DNA Y-chromosome segment number), for example, an allele of DYS393 marker is 12, also called the marker's value. The value 12 means that the DYS393 sequence of nucleotides is repeated 12 times with a DNA sequence of AGAT.

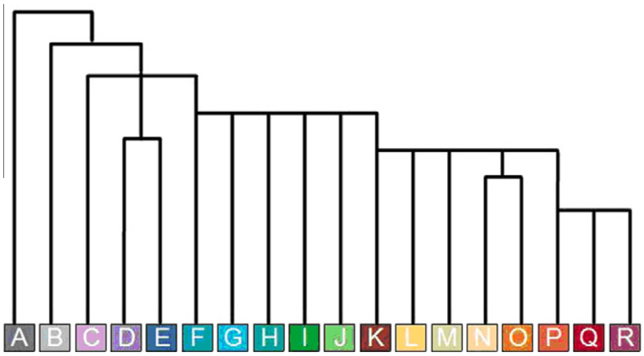

**Fig. 1.** Y-DNA Haplogroup tree.

All combinations of DNA marker values for an individual show his haplotype. Haplogroup is a group of similar haplotypes that share a common ancestor with a single nucleotide polymorphism (SNP) mutation which is a DNA sequence variation occurring when a single nucleotide in the genome differs between members of the species. For instance, two DNA sequences AAGC *C* TA and AAGC *T* TA contain a difference in a single nucleotide. Haplogroups are used to define genetic populations of the world as shown in Fig. 1 where all major Haplogroups that exist in the world are depicted in a major phylogenetic tree.

### 2.3. Clustering

Clustering is a process of partitioning a set of data into a set of meaningful subclasses, called clusters. Clustering helps us to understand genetic relationship between samples more easily. In the clustering process, intra-cluster distances should be minimized and inter-cluster distances should be maximized. Clustering methods can be divided into different groups such as hierarchical, partitioning-based, and density/neighborhood-based (Han & Kamber, 2001).

*Partitioning algorithms* construct a partition of a database *D* of *n* objects into a set of *k* clusters. The partitioning algorithm typically starts with an initial partition of *D* and then uses an iterative control strategy to optimize an objective function. This two-step procedure consists of, firstly, determining *k* representatives minimizing the objective function and secondly, assigning each object to the cluster with its representative closest to the considered object. The second step implies that a partition is equivalent to a Voronoi diagram and each cluster is contained in one of the Voronoi cells (Ester, Kriegel, Sander, & Xu, 1996). Some examples of this method are k-means (MacQueen, 1967), k-medoids (Vinod, 1969) and fuzzy c-means (FCM) (Bezdek, 1981; Dunn, 1973).

*Hierarchical algorithms* create a hierarchical decomposition of *D*. The hierarchical decomposition is represented by a dendrogram, a tree that iteratively splits *D* into smaller subsets until each subset consists of only one object. In such a hierarchy, each node of the tree represents a cluster of *D*. The dendrogram can either be