Constructing Phylogenetic Trees Based on Intra-group Analysis of Human Mitochondrial DNA

Ivan Vogel[1], František Zedek[2], and Pavel Očenášek[1]

[1] Faculty of Information Technology,
Brno University of Technology,
Božetěchova 2, 612 66 Brno, Czech Republic
xvogel01@stud.fit.vutbr.cz, ocenaspa@fit.vutbr.cz
http://www.fit.vutbr.cz/research/groups/nes@fit/index.php.en
[2] Department of Botany and Zoology,
Masaryk University,
Kotlářská 2, 611 37 Brno, Czech Republic
fzedek@gmail.com
http://www.sci.muni.cz/botzool/?lang=en

Abstract. This paper describes a modified algorithm for inferring phylogenetic trees based on distance techniques. The input of the algorithm consists of predefined clusters of data. It uses a usual agglomerative approach, however it involves a novel technique for distance matrix creation as the task of clustering predefined groups of human mitochondrial DNA sequences should be fulfilled.

Keywords: intra-group analysis, substitution model, position-specific clustering vector, mtDNA, population divergence, phylogenetic tree, neighbor joining.

1 Introduction

Phylogenetic tree inference is a very common method for visualising evolutionary relationships among species. Furthermore, thanks to the significant progress in the field of molecular biology, we are now able to work with organisms on the molecular level and thereby analyse nucleotide sequences, which can potentially bring us much more information about the species comparing with phenotype phylogenetic methods. The human mitochondrial genome is commonly used for studying the origins and migrations of different human populations. Mitochondrial DNA is intended for this purpose because of the relatively high mutation rate in comparison to the corresponding nuclear DNA. Moreover, forensic laboratories occasionally use an mtDNA comparison to identify human remains. In our paper, we are going to apply a novel technique of clustering of predefined clusters to identification of whole populations of human individuals.

M.J. Smith, G. Salvendy (Eds.): Human Interface, Part I, HCII 2011, LNCS 6771, pp. 165–169, 2011.

2 Problem Definition

Every distance method for constructing phylogenetic trees uses a single biological sequence as its particular input unit. That means that every leaf node in the result tree matches exactly to one input sequence. By designing our new algorithm, we work with another assumption. Let's suppose that we have a set of DNA sequences that can be classified into disjunct groups/clusters with high membership probability. We indeed assume that sequences from one group are closely related and their intra-group evolution distance is smaller in comparison to distances of sequences from other groups. In some cases there might be greater intra-group average distance which means that not every sequence of a group would be present in the same subtree of standard phylogenetic analysis. In this instance, there must be an effort made to estimate the probable position of the aggregated node with high accuracy according to the elements of the group.

The goal is to find a proper representation for every predefined group. One possible solution is to randomly choose a representative sequence for each group. Another one is to build a consensus sequence for each cluster. There is, however, a certain loss of information in both cases. We therefore present another solution using frequence analysis of predefined clusters (see subsection 2.2).

2.1 General Distance-Algorithm Template

A phylogenetic analysis of any set of biomolecular sequences based on distance metrics uses the following algorithm template:

Input: set of unaligned biomolecular sequences
Output: bootstrapped result tree

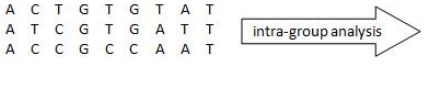
1. Application of multiple alignment on input sequences
2. Phylogenetic distance estimation and distance matrix creation
3. Application of appropriate distance method (mostly neighbor-joining [3])
4. Statistical evaluation of tree topology (mostly bootstrapping [4])

For our algorithm modification, points two and four are crucial.

2.2 Intra-group Analysis

Let's assume we have a predefined cluster. There is the need to find a proper representation of this data structure for future distance estimation among other clusters. We perform an intra-group frequence analysis of every single apriori cluster. The situation is depicted in figure 1.

We count for every single column of the group $\mathbb{X}$ (in fig.1 repesented by three sequences) so-called position-specific clustering vector (hereinafter PSCV), which contains the relative occurence of nucleotides (T, C, G, and A) in a concrete position. We thereby receive a sort of representative sequence in the form of a simple table, on which the probabilities of nucleotide occurences for every sequence position are depicted. It is straightforward that the sum of elements of every single PSCV must be 1 (as the sum of probabilities of nucleotide states in the site has to be exactly 1).



N/P	1	2	3	4	5	6	7	8	9
A	1	0	0	0	0	0	0,67	0,67	0
T	0	0,33	0,33	0	0,67	0	0,33	0,33	1
C	0	0,67	0,67	0	0,33	0,33	0	0	0
G	0	0	0	1	0	0,67	0	0	0

Fig. 1. Cluster data transformation to frequence analysis table

2.3 Distance Estimation between Two Distinct Clusters

The task in this subsection is to estimate the number of substitutions between two distinct clusters, that is, how many substitutions do we need to perform to get from one cluster to another cluster. Let's assume we have cluster $\mathbb{A}$ and cluster $\mathbb{B}$ and their PSCVs on position n, that is $v_{\mathbb{A}}[n]$ and $v_{\mathbb{B}}[n]$, respectively.

$$v_{\mathbb{A}}[n] = \begin{pmatrix} p_{\mathbb{A}}^{T} \\ p_{\mathbb{A}}^{C} \\ p_{\mathbb{A}}^{A} \\ p_{\mathbb{A}}^{G} \end{pmatrix}, v_{\mathbb{B}}[n] = \begin{pmatrix} p_{\mathbb{B}}^{T} \\ p_{\mathbb{B}}^{C} \\ p_{\mathbb{B}}^{A} \\ p_{\mathbb{B}}^{G} \end{pmatrix} \tag{1}$$

To attain the probability of substitution from nucleotide T to T (which means, both clusters contain nucleotide T at this position), we simply multiply $p_{\mathbb{A}}^{T}$ by $p_{\mathbb{B}}^{T}$, which goes for the three remaining nucleotides, as well. That is, to get probability N_n, that no substitution occurs at position n, we simply perform a dot product of $v_{\mathbb{A}}[n]$ and $v_{\mathbb{B}}[n]$. The probability of substitution at this position is therefore $N_s = 1 - N_n$.

The evolution distance between two nucleotide sequences can be estimated with the Jukes-Cantor substitution model [2] (see equation 2). $\hat{p}$ in equation 2 stands for the proportion of substitution sites to all sites in the examined sequences and is also known as p distance [5]. We extend this model combining with the previously mentioned theoretical explanations. That is, to estimate $\hat{p}$ we add N_s values of every nucleotide position together and divide it by the length of the representative sequence (number of rows in table from figure 1).

$$\hat{d} = -\frac{3}{4} ln(1 - \frac{4}{3}\hat{p}) \tag{2}$$

3 Case Study on Human mtDNA

We applied our new method to human mitochondrial DNA analysis. It consists of whole mitochondrial genomes of individuals from all over the world. Analysed sequence groups are listed on table 1.

The data come from [1] except the last two records. We additionaly obtained 4 chimpanzee mitochondrial genome sequences from [9].

The motivation was to reconstruct a phylogenetic tree of different human populations with our algorithmic solution and compare it with a relevant previously published study. We worked with the assumption that intra-group variability of a specific nation is smaller than that between different populations.

Table 1. List of processed human mitochondrial genome populations (and eventually a person who collected the collection in parentheses – if known) along with groups joined into and numbers of sequences worked with

Data	Predefined group	Number of sequences
European (Kivisild) Sardinian (Fraumene) Italian (Achilli)	Europe	215
Papua New Guinean (Ingman) Melanesian (Kivisild) Australian (Ingman)	Australia/Oceania	41
Japanese (Tanaka) Chinese (Kong)	East Asia	720
American [7]	America	5
African [8]	Africa	4

3.1 Data Preparation

The mitochondrial genome of humans consists of approx. 16 kbp. We simply selected highly polymorphic genome sites and their neighborhood, and with this received sequences of approximately 200 nucleotides in length.

4 Results and Conclusion

The phylogeny of predefined groups is shown in figure 2. Chimpanzees as an outgroup is placed at the base of the tree and followed by African human populations outward, from which are branches that subsequently correspond to populations from Europe, Australia/Oceania, the Americas, and East Asia. All the clades show strong bootstrap support (figure 2). The tree topology estimated in our analysis agrees well with the previously published phylogeny of 51 human populations based on 650,000 common, single-nucleotide polymorphism loci [6], suggesting that our algorithm may be a helpful tool for future phylogenetic analyses.

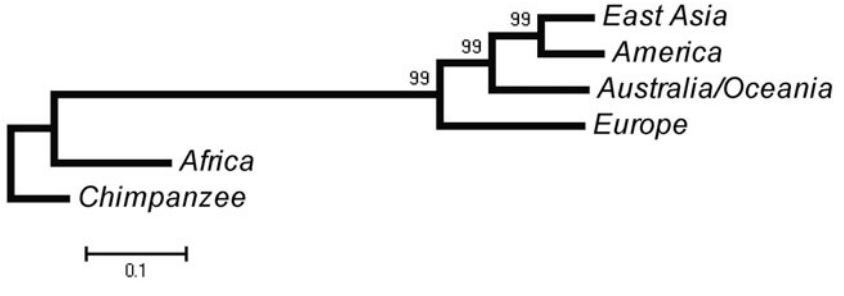


Fig. 2. Constructed phylogenetic tree with 500 bootstrap replications; the scale bar indicates p distance of reconstructed branches

Acknowledgement. This study was supported by the Czech Science Foundation (project GACR 206/09/1405), by the Ministry of Education, Youth and Sports of the Czech Republic (projects LC06073, MSM0021622416, MSM0021630528 and MSM0021630503) and projects FIT-S-11-1, FIT-S-11-2 and FEKT/FIT-S-11-2.

References

1. Ingman, M., Gyllensten, U.: mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Research 34 (suppl. 1)
2. Jukes, T.H., Cantor, C.R.: Evolution of protein molecules. Academic Press, New York (1969)
3. Saitou, N., Nei, M.: The neighbor–joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425 (1987)
4. Yang, Z.: Computational Molecular Evolution. Oxford University Press, Oxford (2006)
5. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, Oxford (2000)
6. Jun, Z.: Li, et al.: Patterns of Variation Worldwide Human Relationships Inferred from Genome-Wide. Science 319 (2008)
7. Forster, P., Harding, R., Torroni, A., Bandelt, H.J.: Origin and evolution of Native American mtDNA variation: a reappraisal. Am. J. Hum. Genet. 59(4), 935–945 (1996)
8. Herrnstadt, et al.: Reduced-Median-Network Analysis of Complete Mitochondrial DNA Coding-Region Sequences for the Major African, Asian, and European Haplogroups. Am. J. Hum. Genet. 70(5), 1152–1171 (2002)
9. Bjork, A., Liu, W., Wertheim, J.O., Hahn, B.H., Worobey, M.: Evolutionary history of chimpanzees inferred from complete mitochondrial genomes. Mol. Biol. E 28(1), 615–623 (2011)