MODIFIED METHODS FOR CONSTRUCTING PHYLOGENETIC TREES OF PREDEFINED GROUPS

Ivan Vogel
Master Degree Programme (3), FIT BUT
E-mail: xvogel01@stud.fit.vutbr.cz

Supervised by: Pavel Očenášek
E-mail: ocenaspa@fit.vutbr.cz

Abstract: In our work, we present a modified algorithm for phylogenetic tree inference. The input of this algorithm involves expert preknowledge – preliminary information about membership of certain biological sequences to distinct groups. The main solution works with an algorithmic modification of distance matrix using intra-group analysis combined with the well known neighbor-joining method. Finally, we analyse human mitochondrial DNA with the proposed method and draw a conclusion through comparison with other studies.

Keywords: phylogeny, neighbor-joining, distance matrix, position-specific clustering vector, mtDNA

1 INTRODUCTION

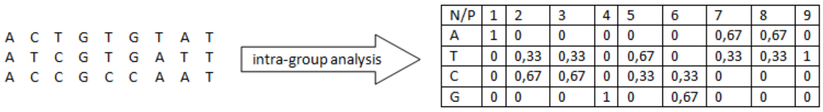
Phylogenetic tree inference is a very common method for visualising evolutionary relationships among species. The human mitochondrial genome is commonly used for studying the origins and migrations of different human populations, because of the relatively high mutation rate in comparison to the corresponding nuclear DNA. The motivation is to reconstruct a phylogenetic tree of different human populations with our algorithmic solution and compare it with a relevant previously published study. The algorithm description as well as obtained results will be presented at the HCI International 2011 Conference [1].

2 PROBLEM DEFINITION

Every distance method for constructing phylogenetic trees uses a single biological sequence as its particular input unit. This means that every leaf node in the result tree matches exactly to one input sequence. 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 average intra-group 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 (see Figure 2a). There is, however, a certain loss of information in both cases. We therefore present another solution using frequency analysis of predefined clusters (see Subsection 2.2).

2.1 General distance-algorithm description

A phylogenetic analysis of any set of biomolecular sequences based on distance metrics consists of application of multiple alignment on input sequences, phylogenetic distance estimation and dis-



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

Figure 1: Cluster data transformation to frequency analysis table

tance matrix creation, application of appropriate distance method (mostly neighbor-joining [1]) and finally a statistical evaluation of tree topology (mostly bootstrapping [2]). In order to take intra-group variability of certain predefined clusters into account, the way of distance matrix creation must be modified.

2.2 Intra-group analysis

Let's assume we have a predefined cluster. We perform an intra-group frequency 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) the 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.

2.3 Distance estimation between two distinct clusters

The main task is to estimate the number of substitutions between two distinct clusters, that is, how many substitutions 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 i, that is $v_{\mathbb{A}}[i] = (p^T_{\mathbb{A}}|p^C_{\mathbb{A}}|p^A_{\mathbb{A}}|p^G_{\mathbb{A}})$ and $v_{\mathbb{B}}[i] = (p^T_{\mathbb{B}}|p^C_{\mathbb{B}}|p^A_{\mathbb{B}}|p^G_{\mathbb{B}})$. To attain the probability of substitution from nucleotide T to T (which means both clusters contain nucleotide T at this position), we simply multiply $p^T_{\mathbb{A}}$ by $p^T_{\mathbb{B}}$, which goes for the three remaining nucleotides, as well. That is, to get the probability P_n, that no substitution occurs at position i, we perform a dot product of $v_{\mathbb{A}}[i]$ and $v_{\mathbb{B}}[i]$. The probability of substitution at this position is therefore $P_s = 1 - P_n$. The evolution distance between two nucleotide sequences can be estimated with the Jukes-Cantor substitution model [2].We extend this model combining with the previously mentioned theoretical explanations and substitute $\hat{p}$ (which stands for the proportion of substitution sites to all sites). The result is depicted in Equation 1.

$$\hat{d} = -\frac{3}{4} ln \left(1 - \frac{4}{3} \frac{\sum_{i=1}^{n} (1 - v_{\mathbb{A}}[i] \cdot v_{\mathbb{B}}[i])}{n} \right) \tag{1}$$

3 CASE STUDY ON HUMAN MITOCHONDRIAL DNA

We applied phylogenetic inference based on intra-group analysis to human mitochondrial DNA. Analysed sequence groups are listed in Table 1. The mitochondrial genome of humans consists of approximately 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, CONCLUSION AND FUTURE WORK

The phylogeny of predefined groups is shown in Figure 2. The modified Jukes-Cantor model as well as p distance metrics (Subsection 2.3) were applied to the predefined clusters for the distance matrix estimation. According to high similarity among the groups p distance seems to be accurate

Data	Predefined group	Number of sequences
European, Sardinian, Italian	Europe	215
Papua New Guinean , Melanesian, Australian	Australia/Oceania	41
Japanese, Chinese	East Asia	720
American	America	5
African	Africa	4

Table 1: List of processed human mitochondrial genome populations along with groups joined into and numbers of sequences worked with

enough for the distance measure (low substitution rate). Figure 2a shows a phylogenetic tree created from consense sequences while in Figure 2b the result of our modified algorithm is presented. Both trees were constructed by the neighbor-joining method. Furthermore, a bootstrap analysis with 500 replications was performed.

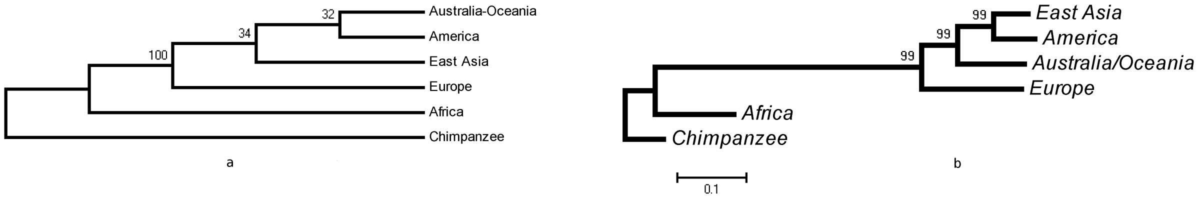


Figure 2: Comparison of two different approaches to predefined group clustering; (b) the scale bar indicates *p* distance of reconstructed branches

All the clades in Figure 2b show strong bootstrap support. 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 [3], The tree built from consensus sequences (cladogram in Figure 2a) by contrast shows lower bootstrap values and the obtained topology seems to be uncorrect (according to [3]). We therefore suggest that our algorithm may be a helpful tool for future phylogenetic analyses. Furthermore, a design of a top down approach solution for phylogenetic tree inference is actually in progress using distance metrics presented in this paper.

ACKNOWLEDGEMENT

This study was supported by the Ministry of Education, Youth and Sports of the Czech Republic (projects MSM0021630528 and MSM0021630503) and projects FEKT/FIT-S-11-2, FIT-S-11-1 and FIT-S-11-2.

REFERENCES

[1] Vogel, I., Zedek, F., Ocenasek, P.: Constructing Phylogenetic Trees Based on Intra-Group Analysis of Human Mitochondrial DNA, HCII 2011 Conference, Lecture Notes in Computer Science, Orlando, FL, USA, 2011

[2] Yang, Z.: Computational Molecular Evolution. Oxford University Press, 2006

[3] Jun Z. Li, et al.: Patterns of Variation Worldwide Human Relationships Inferred from Genome-Wide, Science 319, 2008