4.3 Sequence Evolution of Mitochondrial DNA in Humans and Chimpanzees: Control Region and a Protein-Coding Region

Thomas D. Kocher\textsuperscript{1,2} and Allan C. Wilson\textsuperscript{2}

Summary. Complete primary structures for the major non-coding region of 13 human and two chimpanzee mitochondrial DNAs (mtDNAs) were determined by direct sequencing via the polymerase chain reaction and compared to published sequences for one other human and two other chimpanzees. The human mtDNAs were chosen to represent the deepest branches found in a genealogical tree relating the restriction maps of 182 types of mtDNA. With the four chimpanzee sequences as outgroups, it was possible to place a root on the tree relating the human sequences. This root is consistent with the idea of an African origin for human mtDNA but does not rule out alternative hypotheses. Our sequences confirm a previous finding that the probability of substitution varies greatly among sites in the control region, some sites being so variable that they have probably changed many times since chimpanzees and humans had a common ancestor. In addition, our results show that the pattern of substitution in the control region has diverged since chimpanzees and humans had a common ancestor. These two observations may help to explain why it is hard to root the human tree with chimpanzee sequences as well as to determine the time of common ancestry for humans using the control region. Therefore, sequences were also obtained for a more slowly-evolving part of mtDNA, viz an 896-bp segment which includes parts of the genes for NADH dehydrogenase subunits 4 and 5. They show that the extent of divergence among humans in this segment of mtDNA is less than 3/79 of that between humans and chimpanzees. This result reinforces estimates based on restriction mapping that the last common ancestor of the humans sampled existed less than 200,000 years ago.

\textsuperscript{1}Department of Zoology, University of New Hampshire, Durham, NH 03824, USA
\textsuperscript{2}Division of Biochemistry and Molecular Biology, University of California, Berkeley, CA 94720, USA
Introduction

Much discussion has accompanied the idea, based on restriction maps of human mtDNA, that all the maternal lineages in our species stem from one woman who probably lived in Africa roughly 200,000 years ago [1]. The possibility of a recent human origin for the human gene pool has faced support from nuclear gene stud-
ies [2,3] and partial sequences of the mitochondrial control region [4,5], as well as morphological analysis of the fossil record [6]. However, the importance of the principle is that, for any finite set of populations, all of the maternal lineages ultimately trace back through female lines of descent. At one time one or more of these maternal lines has been reconstructed [7]. The existence of a unique mtDNA ancestor does not imply that these sequences represent a population bottleneck consisting of a single female. The population to which this ancestral female belonging must have been small, and it is thus been inappropriately. In fact, the population size at the time when our last common maternal ancestor lived may have been large [8,9]. The nuclear gene pool of our species is of course likely to be element from many of the individuals who were more closely related to this maternal ancestor.

The African origin hypothesis for mtDNA was based in part on the method that Cann et al. [10] used for the mounting the genealogical tree. Their method, known as statistical analysis, assumes a uniform rate of evolutionary substitution. Cann et al. [10] pointed the need for using an average rate to trace the tree because the latter method does not assume a constant rate of evolution. The African apes mtDNA lineages are the appropriate group for the human cluster of mtDNA lineages was excluded long ago from conventional (low-resolution) restriction mapping [10,12]. However, it was not possible to apply this method to the tree based on high-resolution restriction maps because high-resolution maps are not available for our sister taxa, the African apes. Without such maps, the identification of hominid restriction sites is difficult.

Unlike restriction maps, nucleotide sequences are easily aligned, and homologous nucleotide positions are easily counted. Taking advantage of the existence of a parallel sequence for the basal region of human mtDNA and one of the restriction sequences for the basal region of mitDNA [12,13], we used the approach suggested in greater detail on the basal lineages of the mitochondrial control region from four humans. By sequence analysis of the entire mitochondrial control region and the 596 additional pairs of mtDNA we made it possible to conduct a more detailed analysis of the mitochondrial relationships among 14 hu-
man and five chimpanzee. A preliminary amount of some genealogical implicative of a small part of these results has appeared [13].

We have first sequenced data to examine the human mitochondrial ancestor. Previous approaches to tracing the current human mtDNA has been a recent common maternal ancestor [14,15]. We have sequenced one branch of Chimpanzee by the 3' part of the ND gene and 5' part of the NOS gene. The trees and model of evolution of the current human mtDNA have been a more efficient migration in progress [16,17]. Since the same tree topology was relatively well known [18,19], these

4.3 Evolution of Mitochondrial DNA in Humans and Chimpanzees

sequences allow us to make an extreme estimate of the age of this common ancestor.

Materials and Methods

mtDNA. Thirteen human mtDNA were chosen from the mapping study [20] to represent the most divergent mitochondrial DNA types and those are the most informative for the determination of the root. Eight of them are of African origin, while 4 are Asian and one is Australian. Table1 shows the relationships of these samples to those previously studied. In addition, we sequenced two chimpan-zee mtDNA which were determined to be highly divergent in a restriction site analysis [10].

 Amplification and sequencing. The production of sequences from large number of individuals is refined with classical cloning techniques. Furthermore, se-
quence errors introduced during vector propagation, while negligible for many

Table 1. List of individuals studied, their geographic origins, and their site in earlier studies.

Table 2. Sequences of primers used for amplification and sequencing.

Fig. 1. Sequencing strategy for two parts of mtDNA, ND1-5 and an 806-bp region containing the control region (1969) for identity, location, and behavior with four parts of two noncoding regions. The schematic representation of the human mtDNA control region is in the lower part of the figure. The entire control region is divided into two parts: the restriction endonuclease (1969) and the second part, one of the light strand (1969) and one of the heavy strand (1969).

Results and Discussion

Sequences of Control Regions

Accuracy of sequencing via PCR. Two of the 15 mtDNAs that we characterized (1A and 1B) had been partially sequenced by Groppel et al. [22] after enzymatic cloning in a series of plasmid vectors (Table 1). A comparison of these sequences with ours revealed 3 discrepancies among 30 variable sites (see legend of Fig. 2). We used the consensus of these sequences as our best estimate of

changes during evolution (parentheses) were constructed from the sequences using the computer algorithm from the analysis package programs PAUP [27] and PHYLP [20,29]. Distance trees were built with the neighbor-joining algorithm [30]. Divergence estimates were excluded from all phylogenetic analyses.

939
Table 4. Matrix of differences due to base substitutions in the control region

<table>
<thead>
<tr>
<th>Hx</th>
<th>H1</th>
<th>H2</th>
<th>H3</th>
<th>H4</th>
<th>H5</th>
<th>H6</th>
<th>H7</th>
<th>H8</th>
<th>H9</th>
<th>H10</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H2</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>H3</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H4</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H5</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H6</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H7</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H8</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H9</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>H10</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

*Phylogenetic Partitioning of the Variation

Distinction of human and chimpanzee mtDNA. Considering all variable sites in the control region, we find a complete geographic separation between the 18 mtDNA types in humans and the 3 types in chimpanzees. This separation has long been evident from restriction analysis of the whole mtDNA sequence [18]. The distinction between human and chimpanzee control regions is supported by 100% of the restriction sites used in a parameter analysis [28,29]. Furthermore, the branching order for the chimpanzee samples that based on restriction analysis [18] and evolves among hominoid speciation (see below).

Restriction of the tree relating human mtDNAs. Of the 38 sites in the control region that vary by base substitution among humans, 29 are phylogenetically informative about relationships among humans. The other 9 are uninformative in that the variation at those sites can be explicated equally parsimoniously by any tree. Figure 5a shows a graphical representation, i.e., an incipient tree relating the 18 mtDNAs. No other network explains the variation at the 53 positions more simply than this tree.
To place a root on this network, the 4 chimpanzees were examined next and see only at the same 39 informative sites but also at those positions where only one human type shares a base with one or more chimpanzees. The root falls on that branch in the network which leads to type H2 (Fig. 2b).

The result (Fig. 3b) has two major branches. One leads to H2, while the second leads to the common ancestor of H1, on the one hand, and a cluster of 22 other mDNAs on the other hand. This cluster then divides into subclusters, via the H4-H8 subcluster and the H10-H19 subcluster. Since H2 and H1 are both from Africa and Africans occur in both subclusters, the root tree is easily reconcilable with an African ancestry for human and mDNA.

Reliability of the topology. Determination of the root of the human tree could be a challenging problem because of the possibility that the earlier population grew rapidly and expanded rapidly across the Old World, if so were the case, the length of time during which founding lineage remained in the ancestral geographic area may have been so short as to allow accumulation of a significant number of mutations. Our aim in the next 4 paragraphs is to demonstrate that the topology shown in Fig. 5b is robust to different rooting algorithms and that the major features of the tree are consistent both among analyses and with previous studies. We offer these demonstrations of the correctness of our rooted tree.

Statistical rooting. A measure of the internal consistency of the data is provided by the bootstrap algorithm [9]. In this method the variable sites are randomly remapped (with replacement) to produce new data sets for phylogenetic analysis. The consensus tree for 100 such bootstrap samples has essentially the same topology as in Fig. 5b. The African rooting is supported, although the relative positions of the deep African branches (H12, H3) are different. The remaining taxa commonly fall into the two subclusters (leading to H4-H8 and so H1-H10) found in the shortest tree (Fig. 5).

Distance analysis gives consistent topology. The raw count of the number of positions different among each pair of human sequences (Table 2) was used to approximate the true evolutionary distance. These values were not corrected by any formula because the individual humans are very closely related. Application of the neighbor-joining algorithm [30] produces a tree with a topology nearly identical to the parsimony consensus trees. Only the relative position of one Asian (H12) changes; this is a slight change, moving H12 from H1 to H3.

Topologies with an Asian root. Trees in which the root lies between lineages leading to African mDNAs are in all these cases more parsimonious than those shown in Fig. 5b. In the most parsimonious of these "Asian" trees, the lineage leading to H8 (Asiatic) and H11 (African) can be on one side, while the remaining lineages lie on the other side. This tree requires 5 more mutational steps than does the tree in Fig. 5b.

When the SIM5 sequences available for 7 of the mDNAs are considered together with the control region data as the 39 informative sites, a parsimony tree
is obtained that links PH and HV very tightly (bootstrap value = 90%). This result makes implausible the suggestion that PH could be an outlier for other human mtDNA. This suggestion was raised because it seems to be rare in the tree based on high-resolution mtDNA sequences [11,29]. PH is a member of the deepest non-African branch. Wilson et al. [13] noted the implausibility of PH being an outlier based on the waving tree test.

Time scale based on ND-4.5. In this case it is of considerable interest to relate the topology of this tree to a time scale so that conclusions can be made as to when the different clades might have diverged. If the age of the oldest mtDNA common ancestor is 150,000 years ago (Table 2), the ages of the remaining lineages can be estimated quite precisely. The STR rate is 0.7·10·0.0001/73,000 which is 0.00001 year/age, which is sufficient to estimate the common ancestors of diverse populations.

The ND-5 sequences for human populations presented here make a new, independent estimate of how long ago the oldest common mitochondrial ancestor lived (Table 5). In making this estimate we use the following rationale and assumptions:

1. Because mtDNA evolves very rapidly, multiple his occur often and must be corrected in calculations of divergence.
2. The last common mtDNA ancestor of humans and chimpanzees lived 4·5 million years ago [14,15].

To estimate 4, we assume that the variance is very large. This is weaker than to assume that the variance is very small. For a normal distribution, functional significance may therefore be reached only when complete mtDNA sequence data are available.

Conclusions

African origin. The analysis based on the central reversal and ND-4.5 sequences reported here supports the hypothesis of an African origin of human mtDNA, as suggested by Cavalli-Sforza et al. [11], Vignal et al. [4], and Horai and Hasegawa [5]. Despite the large amount of sequence data presented, however, that suggestion is not necessarily correct. Functional significance may therefore be reached only when complete mtDNA sequences are available.

There is a further means for testing an African origin for the mtDNA of modern humans—one that does not depend on phylogenetic analysis. First, according to recent estimates of world population size, the mtDNA population in Africa is many times larger than the rest of the world’s population [13,32]. Hence the mtDNA diversity is due predominantly to natural mutations (43,44), and we would expect the oldest populations to be the most diverse. If the populations were equally old, one would expect the oldest populations to be the most diverse. If the populations were equally old, one