Evolution and phylogeny of hominoids inferred from mitochondrial DNA sequences

Rumi Kondo

Doctor of Philosophy

Department of Genetics
School of Life Science
The Graduate University
for Advanced Studies

1992

ABSTRACT

This dissertation addresses the 4.9 kb (kilobases) nucleotide sequences of mitochondrial (mt) DNAs from five hominoid species (common and pygmy chimpanzees, gorilla, orangutan and simang), and presents their detailed analyses, together with the known human whole sequence, to assess the tempo and mode of hominoid mtDNA evolution. Particular attention was paid to the rate of synonymous substitutions in protein coding region as well as of silent substitutions in other regions. This work was further extended to the whole mitochondrial genomes of four hominoid species (human, common chimpanzee, gorilla and orangutan) with additionally determined 10 to 12 kb mtDNAs from common chimpanzee, gorilla and orangutan. These hominoid mtDNAs revealed several functionally and evolutionarily characteristic features and provided useful information on the history of hominoid species.

Most significant observations drawn from the present data are summarized as follows. First, comparison of the base compositions in any specified region of hominoid mtDNAs showed a strong base composition bias, as observed in other vertebrate mtDNAs. The L-strand of hominoid mtDNAs is rich in A (adenine) and C (cytosine) contents, but low in G (guanine) content. Base composition biases are strongest at the third codon positions and are evident along the whole genome, independent of the genomic regions. Both codon usage and amino acid preference of mitochondrial protein genes are in agreement with the base composition biases. These observations suggested that there is a biased mutation pressure in mtDNA. A possible cause may be differential deaminations of C residues owing to the asymmetric replication of both L- and H-strands of mtDNA. It is possible that differential deamination has resulted in the reduced number of C residues in the H-strand, although there has been no clear evidence for this possibility in hominoid mtDNAs.

Second, there exist functionally important nucleotide sites over the genome. Together with information on tertiary structures of proteins, as well as on secondary structures of transfer (t) RNAs, ribosomal (r) RNA genes and noncoding regions, the distribution of variable sites among hominoid mtDNAs suggested that some nucleotide sites have been playing important roles in peptide folding, assembly of proteins, or interaction to some other proteins and regulatory elements. Noteworthy are two functionally distinct regions in the major noncoding region (D-loop). One is concerned with promoter sequences for transcription and the other is with three conserved blocks. Oranguan mtDNA sequence revealed unusual substitutions at both of these regions. This suggested that the replication and transcription machinery in orangutan mtDNA may differ from that of other hominoid mtDNAs.

Third, comparison of nucleotide differences observed among closely related hominoids revealed a remarkably biased mode of changes. Between human and chimpanzee, 70% of the observed nucleotide differences are silent changes that occur mostly in the small noncoding regions or at the third codon positions of protein genes. Extensive deletions and additions are observed, but they are found only in the noncoding regions. Such observations suggested a conserved mode of the evolution of hominoid mtDNA genomes. There is also a strong preference to transitions over transversions. Out of 852 variable third positions of codons between the human and common chimpanzee mtDNAs, 93% account for transitions of which 66% are TC transitions (in the L-strand). Within the remaining 7% transversions, CA differences are most frequent while GT are least. These substitution biases correlate well with biased base compositions, particularly the low G content of the L-strand.

Fourth, owing to the outnumbered transitions and strong biases in the base compositions, synonymous substitutions reach rapidly a rather low saturation level. AG transitions attain a saturation level lower than TC transitions (in the L-strand), and such a low ceiling is observed even between the human and chimpanzee pair that diverged around five million years ago. At present, it seems inevitable to select appropriate regions that have experienced theoretically tractable numbers of substitutions. In the case of hominoid mtDNAs, candidates are all types of changes in the tRNA and rRNA regions, transversions in the noncoding regions, and nonsynonymous changes and synonymous transversions in the protein coding regions.

Fifth, rapidly evolving mtDNAs are potentially useful for addressing classical issues in taxonomy, provided that each nucletide site has not undergone extensive multiple-hit substitutions. From the whole 16209 sites of mtDNAs compared among the four hominoid species, it appears that 12137 such sites are suitable to phylogenetic use. The analysis strengthened the pattern and dating in hominoid diversification inferred from the previous analysis of 4.9 kb region in six hominoid species (among African apes, gorilla diverged first about 7.7 million years ago and then chimpanzee and human became distinct about 4.7 million years ago).

Finally, the synonymous and nonsynonymous substitution rates were examined under the assumption of the gorilla divergence being 7.7 million years ago. The extent of the compositional biases differs from gene to gene. Such differences in base compositions, even if small, can bring about considerable variations in observed synonymous differences, and may result in the region-dependent estimate of the synonymous substitution rate. A care should be taken for heterogeneous transition and base composition biases as well as different saturation levels of transition changes. The synonymous substitution rate estimated with this caution showed the uniformity over genes ($2.37 \pm 0.11 \times 10^{-8}$ per site per year) and the high transition rate, about 17 times faster than the transversion rate. These synonymous and transition rates are comparable to the silent substitution rate in the noncoding segments dispersed between genes. On the other hand, the rate of nonsynonymous substitutions differs considerably from gene to gene as expected under the neutral theory of molecular evolution. The average differences in the gorilla - human and gorilla - chimpanzee comparisons indicated that the lowest rate is 0.7×10^{-9} per site per year for *COI* and that the highest rate is 5.7×10^{-9} for *ATPase 8*. The degree of functional constraints (measured by the ratio of the nonsynonymous to the synonymous substitution rate) is 0.03 for *COI* and 0.24 for *ATPase 8*. tRNA genes also showed variability in the base content and thus in the extent of nucleotide differences as well. The substitution rate averaged over 22 *tRNAs* is 5.6 x 10^{-9} per site per year. The rate for *12S rRNA* and *16S rRNA* is 4.1 x 10^{-9} and 6.9 x 10^{-9} per site per year, respectively. All of these observations strongly suggested that mutations themselves occur more or less with the same rate and compositional biases.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to Dr. Satoshi Horai for his invaluable guidance throughout this work. Gratitude is also expressed to Prof. Naoyuki Takahata and Dr.Yoko Satta who were so helpful in the computer programming and spared me much time for stimulating discussions. I owe thanks to Prof. Sadao I. Chigusa and many colleagues of the Human Genetics laboratory, in particular, Prof. Takashi Imamura, Ms. Kumiko Murayama, Dr. Yuko Nakagawa-Hattori and Mr. Seiji Hayashi, for their help and encouragement during this study. Grateful acknowledgement is extended to Profs Toshimichi Ikemura, Kazuo Moriwaki, Masami Hasegawa, Takashi Gojobori, and Dr. Naruya Saitou for serving on my advisory committee. I also thank Drs. Osamu Takenaka and Shintaro Ueda for providing genomic DNA samples for pygmy chimpanzee and gorilla, Dr. Takafumi Ishida for providing the cell lines for common chimpanzee, orangutan and siamang, and Drs. Tadashi Inoue and Kenji Hayasaka for their help at the initial stage of this work. Finally, I would like to express my warmest thanks to my parents Isao and Masako Kondo, and my fiancé Kazuhisa Tsukamoto. They gave me this opportunity, and their continuous love and support have been the motives to finish this piece of work. This work was supported by a Pre-doctoral Fellowship of the Japan Society for the Promotion of Science for Japanese Junior Scientists (0191) to the author.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER ONE

INTRODUCTION

Ever since Darwin, *man's place in nature*, from either a zoocentric or anthropocentric perspective, has been a cardinal question in building comprehensive systems in biology (Darwin 1859; Huxley 1894; Gould 1980). With contributions from both biochemistry and molecular biology and with the discovery of *molecular clocks* which allows to date molecular events in geological time scales, it has become possible to trace with greater assurance the history of hominoids (Zuckerkandl and Pauling 1965; Sarich and Wilson 1967). The idea of using molecules as clocks for phylogenetic analysis rests on an assumption: Once a species diverges and becomes two separate lineages, they accumulate changes or mutations independently. The longer the separation time, the larger number of accumulated mutations. If the rate of accumulation remains steady through time and if this rate is inferred, the biochemical differences between the two species can be converted into physical time since they diverged from a common ancestor.

A great advance in understanding the vertebrate mitochondrial evolution first came with direct sequence comparisons of hominoid mitochondrial (mt) DNAs (Brown et al. 1982). From the analysis of portions of two protein genes and three transfer (t) RNAs, there were three important observations. First, the rate of evolution was about five to 10 times higher than that of nuclear DNA. Second, the ratio of nonsynonymous changes (changes that result in amino acid replacements) to synonymous changes (changes that do not result in amino acid replacements) was small, implying that the nonsynonymous sites of mtDNA are evolutionarily conserved by natural selection. Third, mtDNA has evolved with a bias toward transitions (substitutions between C and T, and A and G). Subsequently, these observations have been confirmed from various parts of primate mtDNAs (Hixson and Brown 1986; Foran et al. 1988; Hayasaka et al. 1988; Kocher and Wilson 1991; Ruvolo et al. 1991). Unfortunately, most studies were based on only 5% of the total mtDNA length, and nevertheless these have long been regarded as representatives of the tempo and mode of mtDNA evolution. There was a clear need to obtain more accurate information. One way to achieve this aim is to quantitatively analyze a long stretch of mtDNA sequences from closely related species that have different divergence times. For this reason, I have focused on the evolution of mtDNAs taken from human, chimpanzees, gorilla, and orangutan and tried to obtain more complete information on the sequences.

My approach was to determine and compare the whole mtDNA genome of the four hominoid species (human, common chimpanzee, gorilla, and orangutan) and to analyze 4.9 kilobases (kb) mtDNA segments in detail for these species as well as pygmy chimpanzee and siamang (Horai et al. 1992). A special attention was paid to the nucleotide changes on the gene-by-gene basis and in the categories of different types of substitutions, such as transversions, transitions, synonymous, nonsynonymous, and codon positions. The detailed examination on the evolutionary characteristics of mtDNAs allowed to point out several new results and cautions in using mtDNA for evolutionary studies. Such gene-by-gene analysis has not been possible with the short portions of mtDNA sequences that had been available before the present data. Owing to a large data set that I used, it was possible to obtain many informative changes. I have reexamined the pattern and dating in hominoid diversification. Finally, considering the characteristics of mtDNA substitutions, I have estimated the rate of nucleotide substitutions, in particular the synonymous and/or silent substitution rate. Before going on to the subject, I would like to briefly introduce the background of my thesis.

Mitochondrion

A mitochondrion is a small (0.5-1.0 μm by 5-10 μm) round-shaped, cytoplasmic organelle found only in eukaryotes (Alberts et al. 1989). Hundreds of these selfreplicating organella may be found in a single mammalian cell. The functional role of the mitochondria is to generate energy necessary for all cells to function efficiently producing ATP (adenosine triphosphate) mediated through the process of oxidative phosphorylation. Animal mitochondria possess a distinct double-stranded circular genome — mtDNA — that is independent of the nuclear genome and has proven to be of great utility in evolutionary studies. Each mitochondrion contains many mtDNA molecules. Hence, there are hundreds to thousand mtDNAs in each mammalian cell, and this facilitates isolation and examination of these molecules.

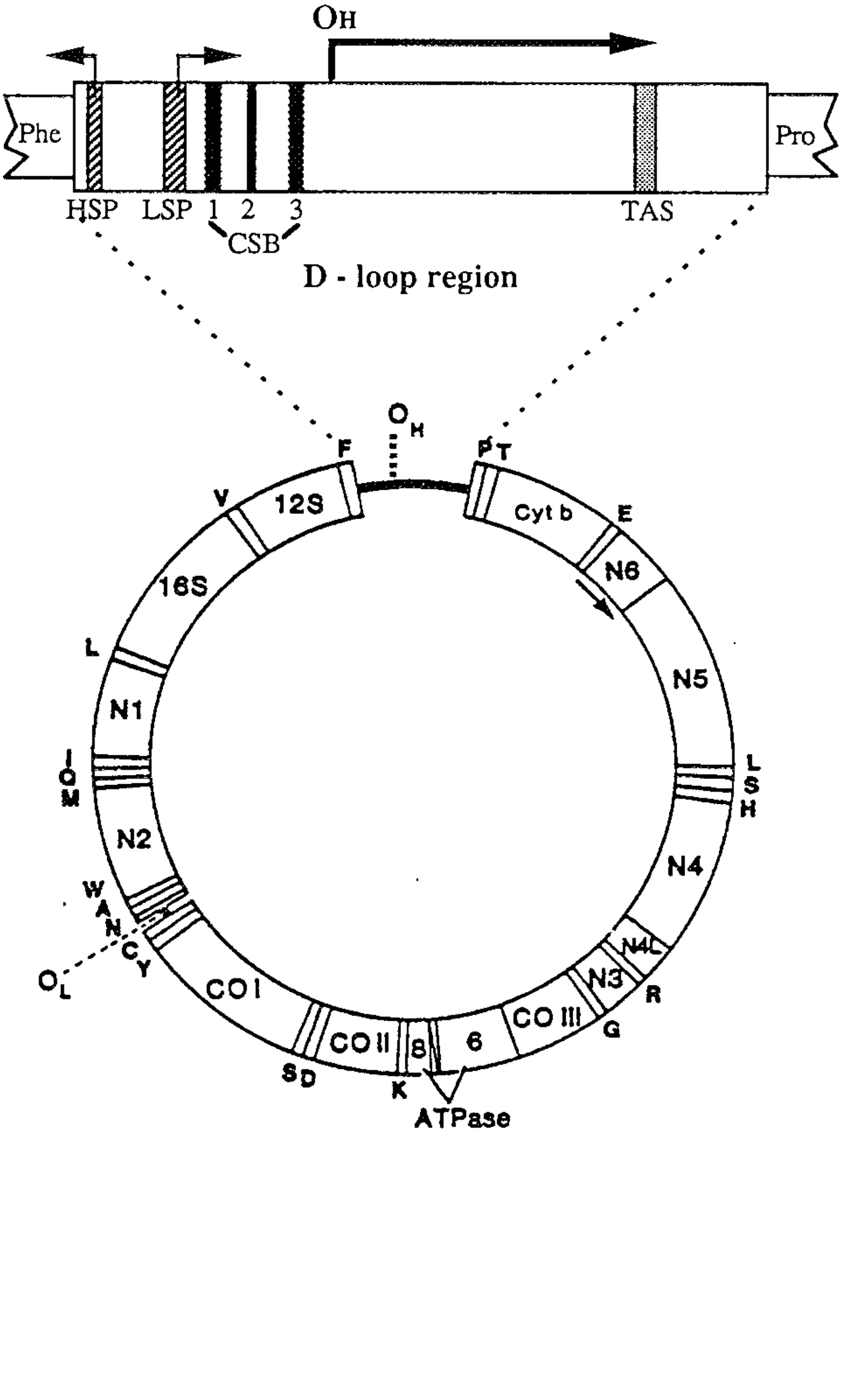
Organization of mammalian mtDNA

A great deal of information on the structure and organization of animal mtDNA was provided from the following complete nucleotide sequences: human (Anderson et al. 1981), mouse (Bibb et al. 1981), rat (Gadaleta et al. 1989), cow (Anderson et al. 1982), fin whale (Árnason et al. 1991), seal (Árnason and Johnsson 1992), chicken (Desjardins and Morais, 1990), frog (Roe et al. 1985), fly (Clary and Wolstenholm 1985), and sea urchins (Jacobs et al. 1989; Cantatore et al. 1989). Partial mtDNA sequence data from various organisms have also been accumulating rapidly, owing to the advent of molecular techniques. One of most convenient techniques recently developed is to directly sequence DNA segments amplified by the polymerase chain reaction (PCR) (Kocher et al. 1989; Irwin et al. 1991). The complete sequence of human mtDNA is 16,569 base pairs (bp) long (Anderson et al. 1981). The molecule encodes 13 proteins, 22 tRNAs, two ribosomal (r) RNAs and is diagrammed in Figure 1.1 (Anderson et al. 1981; Chomyn et al. 1985).

All mitochondrial proteins are involved in electron transport and intracellular respiration and belong to five respiratory complexes (Table 1.1). Although the vast majority of proteins required for mitochondrial function are nuclear-coded and imported from the cytoplasm, the proper functioning of mitochondrial genes is essential to life.

Figure 1.1 Organization of human mitochondrial DNA.

The organization of human mtDNA genome (Anderson et al. 1981). Abbreviations for the genes are as follows: two ribosomal RNA genes (12S and 16S); seven genes for NADH (nicotinamide adenine dinucleotide dehydrogenase) subunits (N1 to N6, N4L); three genes for cytochrome oxidase subunits (COI to COIII); two genes for ATPase subunits (6 and 8); the gene for cytochrome *b* (Cyt *b*); single letter codes for the 22 transfer RNA genes. The origin of replication of heavy (O_H) and light (O_L) strands are designated within the solid bars representing noncoding regions. Relative location of the functional segments in the D-loop region is also indicated: promoters for transcription from the light (LSP) and heavy (HSP) strands, the conserved sequence blocks (CSB-1, CSB-2 and CSB-3), termination associated sequences (TAS). Small arrows indicate the directions of transcription and the large arrow shows the direction of H-strand replication. All 13 protein subunits take part in respiratory complexes embedded in the mitochondrial innermembrane (See Table 4.1).


OH
Phe
Pro
HSP
LSP
1
2
3
CSB
TAS
D - loop region
OH
F
P
T
12S
V
Cyt b
E
N6
16S
L
N5
N1
I
Q
M
L
S
H
N2
N4
W
A
N
C
Y
OL
N4L
N3
R
CO I
CO III
G
CO II
8
6
S
D
K
ATPase

The genetic content and organization of mtDNA have been reviewed recently by Gray (1989).

Figure 1.1 reveals that the genes are organized in an extremely compact form. There are no introns in the coding regions. Except in the displacement-loop (D-loop) region, usually there are less than 10 bp noncoding DNA in the coding regions. Most of the genes are transcribed from the leading strand, termed as the heavy (H)-strand, because of its greater buoyant density in alkaline cesium chloride gradients as a consequence of a positive G + T bias in its base compositions. The genes transcribed from the H-strand are the two ribosomal (r) RNA genes, 14 transfer (t) RNA genes, and 12 protein-coding genes (Ojala et al. 1981). Genes transcribed from the opposite strand, the light (L)-strand, are eight tRNA genes and a protein gene (*ND6*).

Although the D-loop region apparently does not contain structural genes, it too has essential functions. The origins of transcription for both mtDNA strands are found in this segment in addition to the origin of H-strand replication (OH). Other functional elements include promoters for both H and L-strand transcription (HSP and LSP in Figure 1.1; Chang and Clayton 1984; Hixson and Clayton 1985), three transcription factor binding sites (Fisher et al. 1987), and the conserved sequence blocks (CSBs 1, 2, 3) that are associated with the start of DNA synthesis (Walberg et al. 1981). The sites which are related to these functional elements in the control region are generally conserved, only with some exceptions in limited numbers of species (Walberg et al. 1981; Hixson and Clayton 1985; Brown et al. 1986; King and Low 1987).

Transcription and replication of mammalian mtDNA

The mitochondrial function has been comprehensively reviewed by Anderson et al. (1981), Clayton (1982; 1984; 1988) and Attardi (1985). Replication of the H- and L-strands of the mitochondrial genome initiates from separate origins that differ in primary sequence and factors involved in replication (Clayton 1982). The OH is located in the D-loop whereas the origin of L-strand replication (OL) is nested within a cluster of five tRNA genes. A commitment to mtDNA replication begins by initiation of H-strand synthesis that results in strand elongation over the entire genome. Initiation of the L-strand synthesis only occurs after OL is exposed as a single-stranded template. Consequently, when replicating, the H- and L-strands experience the single-stranded state for different periods of time. For H-strand, nucleotide positions proximal to the OL spend least time at the single stranded state. Nucleotide positions distal from the OL in the direction of the L-strand synthesis spend increasing time at the single stranded state.

Each strand of the mtDNA genome is transcribed from a single major promoter in the D-loop region. The two transcription start sites are apart by about 150 bp and the promoters (HSP and LSP) do not overlap, thus functioning as independent entities. Transcription of each strand occurs polycistronicly (Montoya et al. 1982; Chang and Clayton 1984; Bogenhagen et al. 1984). tRNA sequences interspersed between rRNA and protein coding sequences are thought to be recognized as processing signals. After precise cleavage from the primary transcripts, rRNAs are oligoadenylated, mRNAs are polyadenylated, and CCA 3' terminus is added to the tRNAs. Generally, mitochondria require a larger amount of rRNA relative to mRNA or tRNAs. A regulatory role for the transcription termination is known for the 13 bp sequence located within the gene for tRNA$^{\mathrm{LeuUUR}}$ (Christianson and Clayton 1988; Kruse et al. 1989). Since the same transcription units contain genes for tRNAs, rRNAs and mRNAs, differential expression of mammalian mtDNA must be largely controlled by post-transcriptional mechanisms.

Hominoid phylogeny

Our human beings, *Homo sapiens sapiens*, are taxonomically classified with the apes (chimpanzees, gorillas, orangutans, and gibbons) in the order Primates-suborder Anthropoidea-superfamily Hominoidea. The place of humans in the Hominoidea has been highly controversial (Goodman et al. 1983; Foran et al. 1988; Djian and Green 1989; Gibbons 1990). Lack of fossils that allow us to directly trace the evolution of Hominoidea and lack of objective interpretation of small pieces of fossils make it difficult to discern the precise history of hominoids. The observations from the molecules, however, consistently indicate that gibbons diverged first, followed by orangutans, and much later the gorilla-chimpanzee-human divergence occurred (Goodman 1963; Kohne et al. 1972; Sibley and Ahlquist 1984; Ferris et al. 1981). With molecular evidence and the discovery of a facial fossil of *Sivapithecus, Ramapithecus* (a fossil record of which showed that it lived at least 13-16 million years ago) dislodged from the position as the putative first hominid to the hominoid ancestral to all living great apes and humans (Andrew and Cronin 1982; Andrews 1986). It has become widely accepted that human and the African apes share a Pliocene ancestor, much more recent than previously thought (Pilbeam 1984; Mellars and Stringer 1989; Stringer 1990).

The determination of branching order of human, chimpanzee and gorilla (*trichotomy*) is a target of hominoid phylogeny. Several portions of mitochondrial and nuclear DNA of hominoids have been sequenced (e.g. Brown et al. 1982; Hixson and Brown 1986; Koop et al. 1986; Miyamoto et al. 1987; Maeda et al. 1988; Ueda et al. 1989), and analyzed by different statistical methods (e.g. Nei et al. 1985; Hasegawa et al. 1985; 1987; Saitou and Nei 1986). However, the branching order and the divergence times among the three species remained in dispute (Holmquist et al. 1988; Hasegawa 1990; Saitou 1991). This controversy reflects the stochastic nature of the molecular clock and the fact that human, chimpanzee and gorilla might have diverged within a short period of evolutionary time (Saitou and Nei 1986). Mitochondrial DNA which is known to evolve much more rapidly than nuclear DNA (Brown et al. 1982) is useful to this end. In fact, a recent analysis of six hominoid mtDNAs of ca. 5 kb length appears to have resolved the trichotomy (as chimpanzees being closer to human than gorilla) and to have given fairly accurate dating of their divergences (Horai et al. 1992).

Evolutionary studies on mtDNA

There are several other advantages that make the mtDNA to be an intriguing genetic material to study molecular evolution: According to the endosymbiont hypothesis

(Margulis 1981), mitochondria originated from aerobic bacteria that were endocytosed in primitive eucaryotic cells about three billion years ago and now have become virtually indispensable for most eukaryotic cells. Most, if not all, of the genes now found in the mtDNA are thus considered to represent genetic information retained from the original endosymbiont. It is believed that there was massive transfer to the host nucleus and loss of genetic information from the symbiont genome in the course of evolution. The coordinated contribution of two genetic systems — nucleus and mitochondria — suggests that the two genetic systems have started to coevolve since somewhere in the past. Interaction of mitochondria and nucleus is also an interesting evolutionary problem.

Some dramatic reorganization of mitochondrial genomes must have occurred in different organisms (reviewed in Brown 1985; Moritz et al 1987; Gray 1989). In animal mtDNA, the genome size varies from 14.3 kb in *Ascaris suum* (Wolstenholm et al. 1987) to 39.3 kb in *Plactopecten magellanicus* (sea scallop) (Snyder et al. 1987). The small size of *Ascaris* mtDNA reflects the absence of *ATPase 8* found in the vertebrate mtDNA, while the large size of mtDNAs found in lizards, fish and nematodes is due to localized sequence amplification, resulting in direct tandem duplications (0.8-8.0 kb) of both coding and noncoding portions of the genome (Moritz and Brown 1987; Bentzen et al. 1988; Hyman et al. 1988). Different arrangements in gene order or exchange in the coding strands have been found in each animal phylum studied to date (Clary and Wolstenholm 1985; Cantatore et al. 1987; Wolstenholm et al. 1987; Yang and Zhou 1988; Jacobs et al. 1989; Smith et al. 1989, 1990; Pääbo et al. 1991). Different replication, transcription and translation machinery (Desjardins and Morais 1990), variability in GC content and codon usage (Jukes and Osawa 1990) have also been noted. However, the mechanisms that yield such variations in contemporary mitochondrial genome are not known.

There is growing evidence that mtDNA is strongly related to neuromuscular diseases and many other phenomena such as ageing. In order to verify which mutation is critical, we need to pursue comparative studies on the molecule and gain a deeper understanding of the evolutionary aspects of mtDNA.

From a technical point of view, the mode of maternal inheritance and the uniclonal nature eliminate the genetic complexities that accompany biparental transmission (Hutchison et al. 1974; Potter et al. 1975; Giles et al. 1980), although a low level of paternal leakage of mtDNA has been reported in mouse and Drosophila (Kondo et al. 1990; Gyllensten et al. 1991; Kondo et al. 1992). In animal mtDNAs, the size is generally small, about 16-20 kb (25,000 times smaller than the smallest nuclear genome), and sequence rearrangements are very rare due to lack of introns and very few spacer sequences between genes. The rate of evolution is approximately 20 times higher than that of nuclear DNA in mammals (Satta et al. 1991). These characteristics have promoted studies of mtDNA, and it has now become one of the best studied systems concerning replication, transcription and evolution.

Questions to be addressed

Mitochondrial genome constitutes from functionally different regions or genes, implying that evolution of mtDNA genome as a whole is a compound product of various evolutionary processes. Accordingly, each of them must be considered separately. The goal of my thesis is to characterize the molecular clock of hominoid mtDNA, particularly about the synonymous substitution rate, in a statistically adequate manner. The main problem will be multiple hit substitutions in mtDNA caused by compound effect of high mutation rate, transition and base composition biases, and strong functional constraints. How I pursued the goal is the subject of the remaining portions of the thesis.

CHAPTER TWO

MATERIALS AND METHODS

Abbreviations

A list of abbreviations used in this chapter is in Table 2.1.

Table 2.1 Abbreviations used in this chapter

APS	ammoium persulfate
BIS	N, N^1-methylene-bis-acrylamide
BPB	bromophenol blue
BRL	Bethesda Research Laboratories
DTT	dithiothreitol
dNTPs	deoxynucleotide triphosphates
EDTA	ethylenediamine tetraacetic acid, disodium salt
EtOH	ethanol
IPTG	Isoplopylthio-β-D-galactoside
PCR	polymerase chain reaction
PEG	polyethylene glycol
SDS	sodium dodecyl sulfate
TAE	40 mM Tris-acetate, 1 mM EDTA
TBE	89 mM Tris-borate, 89 mM boric acid, 2 mM EDTA
TE	10 mM Tris-Cl (pH 8.0), 1 mM EDTA (pH 8.0)
TY	Tyrpton-Yeast medium: 1 liter of the medium contain 8 g of bacto-trypton, 5 g of bacto-yeast extract, and 2.5 g of NaCl.
TEMED	N, N, N^1, N^1, -tetramethylethylenediamine
UV	ultraviolet
Xgal	5-Bromo-4-chloro-3-indolyl-β-D-galactoside

Sample sources

Some sequence data were obtained from the published studies: The whole human (*Homo s. sapiens*) mtDNA sequences (Anderson et al. 1981); the 4.9 kb region (hatched bar in Figure 2.1) in common chimpanzee (*Pan troglodytes*), pygmy chimpanzee (*Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*) and siamang (*Hylobates syndactylus*) (Horai et al. 1992); 12S rRNA (dashed bar in Figure 2.1) in common chimpanzee and gorilla (Hixson and Brown 1986); and the D-loop noncoding region (open bar in Figure 2.1) in common chimpanzee, pygmy chimpanzee and gorilla (Foran et al. 1988). Other portions of mtDNA for common chimpanzee, gorilla and orangutan were newly determined (shown as solid bar in Figure 2.1) from the same individual samples used in Horai et al. (1992). Genomic DNA of gorilla was provided by Dr. Shintarou Ueda in Tokyo University, Japan. EB virus transformed B cell lines for common chimpanzee (named "Gon") and orangutan (Bornean orangutan named "PopE$_3$"; Ishida and Yamamoto 1987) were provided by Dr. Takafumi Ishida in Tokyo University, Japan.

Reagents

Analytical grade reagents were used to prepare buffers and reagents. TEMED, acrylamide, agarose, and SDS were from Bio-Rad laboratories. Ammonium persulfate was from Sigma. Enzyme grade DTT, Ultrapure dNTPs, and T$_4$ polynucleotide kinase were from Takara Biomedicals. λ exonuclease was from GIBCO BRL. Taq polymerase was from Perkins Elmer-Cetus. Sequenase 7-deaza DNA sequence kit was from United States Biochemical. α-^{32}P-dCTP (400 ci/mmol, or 1 : 5 diluted 3000 ci/mmol) was from Amersham. Oligonucleotide primers were synthesized using DNA synthesizer B (Applied Biosystems) or were orderly made by Tanehashi Co., and are listed in Table 2.2.

Extraction and Cloning of mtDNA

Mitochondrial DNAs were purified from cultured cells of a common chimpanzee (*Pan troglodytes*) and an Bornean orangutan (*Pongo pygmaeus*). As shown in Figure 2.1, mtDNA clones covering the regions C_1 (common chimpanzee), and O_1 and O_2 (orangutan) were obtained. Other parts of mtDNA were amplified by means of the PCR using mtDNA as templates (Saiki et al. 1988). Relevant segments of mtDNA were also amplified from the total DNAs of gorilla (*Gorilla gorilla*). PCR amplified fragments recovered from the gels were either subcloned in the *Sma*I cleaved vector M13mp10 after *Hae*III and/or *Alu*I digestion, or subjected to subsequent PCR cycles to prepare single-stranded template DNAs for sequencing.

Amplification of mtDNA segments

One µl of the total DNA or mtDNA (ca. 50 pg) was subjected to 30 cycles of amplification in a 50 µl reaction volume with 2 units of Taq polymerase. The procedure for setting up a PCR was as follows. 1. Addition of 1 µl of template DNA to 0.5 ml microcentifuge tubes. 2. Addition of 49 µl reaction cocktail consisting of : 5 µl 10 x PCR buffer (described in RECIPES section), 8 µl of dNTP mix (see RECIPES), 5 µl each of two 2 µM primers, 0.2 µl of Taq polymerase, and 25.8 µl sterile double-distilled water. 3. Addition of 2 droplets of light mineral oil to hinder evaporation. Each amplification cycle consisted of denaturation at 94 °C for 10 to 15 sec, annealing at 45 °C for 10 to 15 sec, and extension at 72 °C for 15 sec to 2 min, depending on the size of DNA fragment to amplify. After completion of the PCR, 5 µl of the product was mixed with 2 µl of bromophenol blue loading buffer, and electrophoresed through a 1 to 1.5% agarose gel in 0.5 x TAE buffer containing 0.1 µg/ml ethidium bromide. Electrophoresis was done in a minigel apparatus (Mupid) at 100 V for 25 min. For size standard marker, Marker II (λ / *Hind* III + *Eco* RI double digest) or Marker IV (ϕX174 / *Hae* III digest) (Wako) was used. The gel was photographed on a UV transilluminator. The rest of the sample was stored at 4 °C.

Figure 2.1 Map of mtDNA and location of sequenced clones

The 16,569 bp circular genome has been drawn as a line starting at the heavy strand origin of replication (O_H). The arrow of the L strand (L) and H strand (H) points to the direction of 3' end. The location and the sense strand for each rRNA, tRNA and protein genes are indicated by the names just above or below the strand. Twenty-two tRNA genes are represented by a single letter. There are two tRNAs for leucine (Lu and Lc) and serine (Su and Sa). The numbering follows that of human mtDNA (Anderson et al. 1981). The location of mtDNA clones for common chimpanzee (C_1) and orangutan (O_1, O_2) are shown by arrow bars below the map. The size and location of the published sequence data are shown by hatched (bp 4,121 to bp 9,020 from Horai et al. 1992), dashed (Hixson and Brown 1986; Brown et al. 1982), and open (Foran et al. 1988) bars. The size and location of nucleotide sequences newly determined in the present study are shown by solid bar. Other abbreviations: O_L—the light strand origin of replication; restriction enzyme recognition site—*R* (*Eco*RI), *H* (*Hind*III), and *P* (*Pst*I).

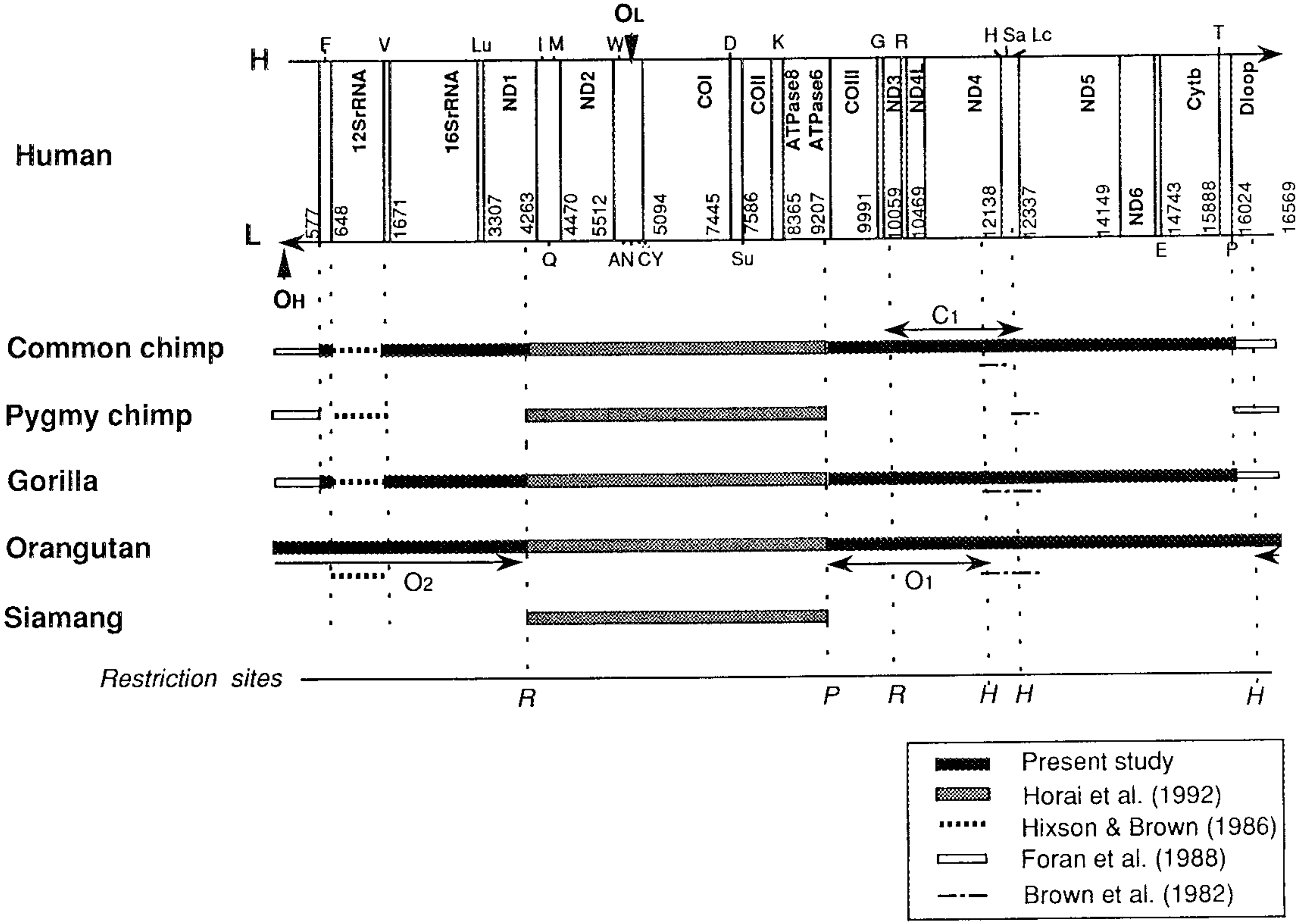

OL
H
F
V
Lu
I M
W
D
K
G R
H Sa Lc
T
12SrRNA
16SrRNA
ND1
ND2
COI
COII
ATPase8
ATPase6
COIII
ND3
ND4L
ND4
ND5
ND6
Cytb
Dloop
Human
577
648
1671
3307
4263
4470
5512
5094
7445
7586
8365
9207
9991
10059
10469
12138
12337
14149
14743
15888
16024
16569
L
Q
AN CY
Su
E
P
OH
C1
Common chimp
Pygmy chimp
Gorilla
Orangutan
O2
O1
Siamang
Restriction sites
R
P
R
H
H
H
Present study
Horai et al. (1992)
Hixson & Brown (1986)
Foran et al. (1988)
Brown et al. (1982)

Table 2.2 Oligonucleotide primers used for PCR and sequencing

ID	Positions*	Sequence
P 2	13211	5'-CCCTTACACAAAATGACATC-3'
P 4	15211R	5'-GAACTAGGTCTGTCCCAATG-3'
P 10	11011	5'-ATCCAGTGAACCACTATCAC-3'
P 14	16403R	5'-ATTGATTTCACGGAGGATGG-3'
P 17	12315R	5'-CTTTTATTTGGAGTTGCACC-3'
P 19	12315	5'-GGTGCAACTCCAAATAAAAG-3'
P 21	1478R	5'-GCGGGTGACGGGCGGTGTGT-3'
P 25	8541	5'-GTTCGCTTCATTCATTGCCC-3'
P 27	14201R	5'-TTAGTAGTAGTTACTGGTTG-3'
P 29	9912R	5'-CAGTATCAGGCGGCGGCTTC-3'
P 32	16190R	5'-CTTGCTTGTAAGCATGGGG-3'
P 38	15375R	5'-CCTAGGGGGTTGTTTGATCC-3'
P 49	3241	5'-AGAGCCCGGTAATCGCATAA-3'
P 53	9209	5'-GACCCACCAATCACATGCCT-3'
P 55	14553R	5'-TAGCGGTGTGGTCGGGTGTG-3'
P 57	11492	5'-CGCCTCACACTCATTCTCAA-3'
P 58	8998	5'-GTACGCCTAACCGCTAACAT-3'
P 59	13051	5'-GGCCCCACCCCAGTCTCAGC-3'
P 60	9912	5'-GAAGCCGCCGCCTGATACTG-3'
P 61	11798	5'-GGACTTCAAACTCTACTCCC-3'
P 62	14844	5'-TCGGCTCACTCCTTGGCGCC-3'
P 63	13910R	5'-GCTAGGGTAGAATCCGAGTA-3'
P 66	8998R	5'-ATGTT(AG)GCGGTTAGGCGTAC-3'
P 70	9510	5'-TACCACTCCAGCCTAGCCCC-3'
P 71	9913R	5'-CAGTATCAGGCGGCGGCTT-3'
P 72	10503R	5'-CCTAGAAGTGAGATGGTAAA-3'
P 73	13913	5'-TCGGATTCTACCCTAGCATC-3'
P 76	520	5'-CACACCGCTGCTAACCCCAT-3'
P 77	707R	5'-GGGTGAACTCACTGGAACGG-3'
P 78	1549	5'-GAGGAGACAAGTCGTAACAT-3'
P 82	15806	5'-GCATCCGTACTATACTTCAC-3'
P 83	3130	5'-AGGACAAGAGAAATAAGGCC-3'
P 84	3404R	5'-CACGTTGGGGCCTTTGCGTA-3'
P 85	4512R	5'-GCTGTGATGAGTGTGCCTGC-3'
P 89	4449R	5'-AGTACGGGAAGGGTATAACC-3'
P 90	14691	5'-CAACCACGACCAATGATATG-3'
P 91	15919R	5'-GTTTTCATCTCCGGTTTACA-3'
P 94	14674	5'-TATTCTCGCACGGACTACGA-3'
P 95	14821R	5'-TTCATCATGCGGAGATGTTG-3'
P 97	11018R	5'-TTTTTTCGTGATAGTGGTTC-3'
P 98	8345R	5'-CATTTCACTGTAAAGAGGTGTGAG-3'
P101	3225	5'-GGTTTGTTAAGATGGCAGAGGCCGG-3'
P108	6619SR	5'-ATATAGACTTCTGGATGACC-3'

(Table 2.2 continued)

ID	Positions*	Sequence
P115	3272R	5'-TAAGAAGAGGAATTGAACCTCTGACCTTAA-3
P116	15806R	5'-GTGAAGTATAGTACGGATGC-3'
P117	13844	5'-ACCTCAACTACCTAACCAAC-3'
P118	13844R	5'-GTTGGTTAGGTAGTTGAGGT-3'
P128	12752PR	5'-CTCTCAGCCGATGAAGAGTT-3'
P129	12070	5'-GTTCATACACCTATCCCCCA-3'
P131	13301R	5'-(GA)TG(CT)AG(GA)AATGCTAGGTGTG-3'
P132	11747	5'-GCAAACTCAAACTACGAACG-3'
P133	15761	5'-GGAGGACAACCAGTAAGCTA-3'
P134	1582R	5'-GTTCGTCCAAGTGCACTTTC-3'
P135	687R	5'-GGATGCTTGCATGTGTAATC-3'
P136	4150R	5'-AGGTGTATGAGTTGGTCGTA-3'
P137	15801	5'-AAGTAGCATCCGTACTATACTT-3'
P150	3454	5'-GCTGACGCCATAAAACTCTT-3'
P151	3943R	5'-GGGCCTGCGGCGTATTCGAT-3'
P152	12906	5'-CCTACACTCCAACTCATGAG-3'
P153	13184	5'-TCACCACTCTGTTCGCAGCA-3'
P154	14432R	5'-ATTGAGGAGTATCCTGAGGC-3'
P155	15830R	5'-GTTGGTATAAGGATTAGGAT-3'
P156	14005R	5'-GGTAATAGCTTTTCTAGTCA-3'
P157	15071	5'-TACTCAGAAACCTGAAACAT-3'
P158	16100R	5'-TCATGGTGGCTGGCAGTAAT-3'
P159	13837	5'-GCCCTAGACCTCAACTACCT-3'
P160	13714R	5'-AATCCTGCGAATAGGCTTCC-3'
P161	13417	5'-GGACTACTCAAAACCATACC-3'
P162	12158	5'-ACATCAGATTGTGAATCTGA-3'
P163	3733R	5'-ATGATGGCTAGGGTGACTTC-3'
P164	15209	5'-TACATTGGGACAGACCTAGT-3'
P168	16411R	5'-TGCGGGATATTGATTTCACG-3'
P169	409	5'-GGCGGTATGCACTTTTAACA-3'
P170	597	5'-CAAAGCAATACACTGAAAAT-3'
P171	1311	5'-CCACGTAAAGACGTTAGGTC-3'
P172	1350R	5'-AATGTAGCCCATTTCTTGCC-3'
P173	2163R	5'-CTTTTAGGCCTACTATGGGT-3'
P174	126	5'-ATCTGTCTTTGATTCCTGCC-3'
P175	414R	5'-GTGCCTGTTGAAAGTGCACA-3'
P178	1908	5'-AACCAGACGAGCTACCTAAG-3'
P179	126R	5'-GGCAGGAATCAAAGACAGAT-3'
P180	15973N	5'-AACTTCACCATCAGCCCCCA-3'

*Numbering of the nucleotide positions refer to Anderson et al. (1981). The numbers without "R" indicate the positions at 5' end of the L-strand primer sequences. The numbers with "R" indicate the positions at 3' end of the H-strand primer sequences. Other alphabets P, S, N indicate that the sequence differs from that of human.

Cloning from the PCR product

PCR amplified fragments were recovered from the gels using GENECLEAN II Kit (BIO 101 Inc.) following the manufactures protocol, and then subjected to digestion with restriction enzyme *Hae*III and/or *Alu*I. After phenol and phenol/chloroform extraction, the digested DNA was precipitated with EtOH and dissolved in sufficient volume of TE for subcloning to the *Sma*I cleaved vector, M13mp10 (0.1 μg/ml). Ligation mixture was prepared in a total volume of 11 μl containing different amounts of insert DNA, 2 μl of vector DNA (0.1 μg/ml of M13mp10), and 1 μl each of 10 x ligation buffer, 10 mM ATP, 0.1 M DTT and T_4 DNA ligase. Ligation reaction was done at 4 °C for over night. Competent cells were prepared freshly for every transformation reaction: 20 ml of 2 x TY was inoculated by 1 ml out of 10 ml over-night culture of XL1Blue in 2 x TY. After shaking for 2 hrs at 37 °C, the culture was transferred to a 50 ml Nalgen tube for 10 min centrifugation in 4N-rotor (TOMY) at 4 °C, 3000 rpm. The pellet was dissolved in 20 ml of ice cold competent cell buffer (50 mM $CaCl_2$ / Tris-HCl, pH 7.4) and incubated on ice for 30 min followed by another 10 min centrifugation at 4 °C, 3000 rpm. The cells were finally resuspended in 2 ml of ice cold competent cell buffer.

For transformation, 0.2 ml aliquots of competent cell suspension were mixed with appropriate amount of DNA ligate in a 15 ml 2059 Falcon tube and stored on ice for 40 min followed by 3 min heating at 42 °C. 0.2 ml of indicator mix and 3 ml of H top agar were mixed in each tube and poured on H plate. After 15 min at room temperature, the plates were incubated at 37 °C for over night.

Isolation of single-stranded DNA from phage

Dispense 1.5 ml of 100-fold diluted over night culture of XL1Blue to 2059 Falcon tubes. Add a single white plaque to each tube and shake at 37 °C for 4.5 hrs. They were then transferred to 1.5 ml centifuge tubes to microcentrifuge at 12000 rpm for 5 min. Supernatants were transferred to new tubes for another centrifugation at 12000 rpm for 5 min. Supernatants were transferred to a new tube and mixed with 0.2 ml of 20% PEG / 2.5 M NaCl. After incubation at room temperature for 20 min, followed by centrifugation at 12000 rpm for 10 min, the supernatants were removed completely and the pellets were dissolved in 100 µl TE. They were extracted once with 50 µl of phenol by voltexing and centrifugation at 12000 rpm for 5 min. The DNA phase was transferred to fresh tubes and DNAs were precipitated by mixing 10 µl of 3 M CH_3COONa and 250 µl of EtOH, followed by incubation at -80 °C for 30 min, and centrifugation at 12000 rpm for 10 min. The DNA pellets were washed once with 75% EtOH, and dried up. They were then dissolved in a final volume of 40 µl of TE of which, 3 µl was mixed with 2 µl of bromophenol blue loading buffer, and electrophoresed through a 1.0% agarose gel in 0.5 x TAE buffer containing 0.1 µg/ml ethidium bromide. Electrophoresis was done in a minigel apparatus (Mupid) at 100 V for 25 min. For size control DNA, single stranded M13 mp10 was used. The gel was photographed on a UV transilluminator. The rest of the samples were stored at 4 °C

Preparing single-stranded template for direct sequencing

Single stranded template DNA for sequencing was prepared from the PCR product in the following two ways.

1. Asymmetrical PCR method

Double stranded template DNA was first amplified as noted above. Then 5 µl of the PCR product was run on a 1.0 % agarose gel with a standard size marker. The aimed DNA fragment is then cut out from the gel to small pieces. Pieces of gels containing the DNA fragment were placed into Suprec™ - 01 (a microcentrifuge tube with filter; Takara Biomedicals) and incubated at -80 °C for 10 min then at 37 °C for 5 min and then microcentrifuged at 10000 rpm (5000 g). The eluted solution obtained was used as template for the subsequent PCR reaction. The reaction condition of the asymmetrical PCR was same as the double stranded PCR, and the reaction cocktail for the asymmetrical PCR differed from that above only in the concentrations of the primers.

The concentrations of the primers used for a 100 μl reaction are 5 μl of the 0.2 μM solution of the limiting primer stock versus 50 μl of the 2 μM solution of the primer in excess. To the PCR cocktail containing 13.5 μl of sterile water, 10 μl of 10 x PCR buffer, 16 μl of dNTPs, a set of primers, and 0.5 μl of Taq polymerase, 5 μl of the template PCR product is added and covered with mineral oil for the PCR thermal cycling. After completion of the PCR, 5 μl of the product was mixed with 2 μl of bromophenol blue loading buffer, and electrophoresed through a 1.5% agarose gel in 0.5 x TAE buffer containing 0.1 μg/ml ethidium bromide. Electrophoresis was done in a minigel apparatus (Mupid) at 100 V for 25 min. For size marker, 2 μl of standard double strand PCR product was used. The gel was photographed on a UV transilluminator. The amplified DNA fragments were purified from the remaining PCR product through filtration by Centricon 30 microconcentrator (Amicon). PCR product and 2 ml of distilled and sterilized water were transferred to the sample reservior, and centrifuged at 5000 rpm for 20 min. This was repeated for two times. Then the concentrated DNA sample was collected to the retenate cup by a centrifugation at 2000 rpm for 5 min. The concentrated DNA sample was transferred to a 1.5 ml centrifuge tube and mixed with 1/10 volume of 3 M NaCl and 2.5 times the volume of EtOH, and incubated at -80 °C for 10 min followed by centrifugation at 12000 rpm for 20 min. The pellet was washed once with 75% EtOH and dried and pelleted by centrifugal concentrator (CC-101, TOMY). DNA was then resuspended in 11 μl of TE, of which 3.5 μl was used as templates, and 0.5 μl of the limiting primer (0.5 μM) was used as primers for each sequencing reaction.

2. λ exonuclease digestion method

This method generates single-stranded template DNAs from PCR products by progressive digestion of one DNA strand with λ exonuclease. Bacteriophage λ exonuclease catalyzes the stepwise release of 5' mononucleotides from the 5' phosphate termini or protruding 5' termini of double-stranded DNA (Little et al. 1967). The enzyme will also work, albeit 100-fold less efficiently, on single-stranded DNA. Because oligonucleotide primers lack the required phosphate residues at 5' terminus, the PCR products obtained by these primers do not serve as substrates. Amplification with one primer containing 5' phosphate terminus and the other without the 5' phosphate residue, however, generates PCR product with only one 5' phosphate terminus. Following digestion with λ exonuclease will thus yield a single-stranded template DNA for sequencing (Higuchi and Ochman 1989).

Phosphorylation of 5' termini of primers were done by T_4 polynucleotide kinase: A total of 60 µl reaction containing 50 µl of 2 µM primer, 6 µl of 10 x kination buffer, 2 µl of 10 mM ATP, and 5 units of T_4 polynucleotide kinase was incubated at 37 °C for 1 hr followed by 70 °C for 30 min. PCR was carried out in a 100 µl reaction as above but using the phosphorylated primer (12 µl) and non-phosphorylated primer (10 µl of 2 µM). The following program of thermal cycler was used: Incubation at 94 °C for 15 sec, 45 °C for 15 sec, 72 °C for 30 to 90 sec, for a total of 30 cycles. The PCR product was extracted with phenol and phenol/chloroform, and after EtOH precipitation, resuspended in 100 µl of λ exonuclease buffer containing 7 units of λ exonuclease. Incubation was done for 1 hour at 37 °C and 30 min at 70 °C. Primers and mononucleotides were removed by precipitation with 60 µl of 20% PEG/ 2.5M NaCl. After 1 hr incubation on ice followed by 30 min microcentrifugation at 12000 rpm, DNA pellet was washed twice with 75 % EtOH and dried and pelleted by centrifugal concentrator (CC-101, TOMY). DNA were then resuspended in 10 µl of sterilized water, of which 3.5 µl was used as templates, and 0.5 µl (0.5 µM) of the sequencing primer (on the same strand as the phospholylated primers) was used as primers for each sequencing reaction.

DNA sequencing

Sequencing reactions were performed by the dideoxynucleotide chain termination method (Sanger et al. 1977) using ^{32}P-dCTP and 7-deaza Sequenase version 2.0. Each reaction was performed using half the volume of that in the manufacture's protocol. I have analyzed the rate of misincorporation in PCR, by subcloning the PCR products of D-loop region of human mtDNA to plasmid vectors, and determining its sequences. The observed rate of misincorporation was one site for every 2500 bps. To avoid the errors due to Taq polymerase, nucleotide sequences were determined either by multiple clones or by direct sequencing of the PCR products.

Sequencing gel electrophoresis

Glass plates of dimensions 55 cm long by 22 cm wide were wiped with EtOH and clipped together with spacers of 0.3 mm thickness. Each gradient gel used was 10 ml of bottom gel and 40 ml of top gel. 65 µl and 20 µl of 25% APS, and 20 µl and 4.5 µl of TEMED were added to the top and bottom gels, respectively. Immediately, the bottom gel and then the top gel were poured down the side of the gel plates using 10 ml and 50 ml disposable syringes. Gels were allowed to polymerize for 30 min to 12 hrs. Sequence reactions were loaded on to the gels and electrophoresed at 2200 V for 3 to 15 hrs. Gels were fixed with 10 % acetic acid / 10 % methanol for 15 min, dried, and exposed to Xray film (AIF RX; FUJI). Fragmental sequences were connected and assembled by GENETYX (Software Development Co., Ltd., Japan).

RECIPES

2% Xgal: Make a stock solution by dissolving 10 mg of X-gal in 0.5 ml of 2% dimethylformamide to make a 20 mg/ml solution. Store at -20 °C.

20% SDS Place 100 g of SDS in an autoclaved bottle containing magnetic stir bar. Add 400 ml sterile distilled water. Stir over low heat. Adjust to pH 7.2 with a few drops of HCl and bring the final volume to 500 ml.

25% APS: Add 2.5 g of ammonium persulfate to 9 ml of distilled water. Store in dark at 4 °C.

40% acrylamide: Dissolve 38 g of acrylamide and 2 g of BIS and adjust to the total volume of 100 ml with distilled water. Store in dark at 4 °C.

0.5M EDTA (pH 8.0): Mix 186.1 g of EDTA with 800 ml of distilled water. While stirring, adjust pH to 8.0 with NaOH pellets. Adjust volume to 1 liter, then autoclave.

1M $CaCl_2$: Dissolve 54 g of $CaCl_2 \cdot 6H_2O$ in 200 ml of pure H_2O (mili Q). Sterilize the solution by passage through a 0.22-micron filter. Store aliquots at -20 °C.

1M DTT: Dissolve 3.09 g of DTT in 20 ml of 0.01 M Sodium acetate (pH 5.2). Filtrate and store at -20 °C.

1M glycine-KOH: Take 50 ml of 2 M glycine and 16.8 ml of 2 M KOH, and adjust to 100 ml with distilled water, autoclave and store at 4 °C.

1M $MgCl_2$: Dissolve 101.7 g of $MgCl_2$ in distilled water and bring to a final volume of 500 ml, then autoclave.

1M Tris-HCl : Mix 121.1 g of Tris base in 800 ml of distilled water. Adjust pH by adding concentrated HCl (To adjust to pH 7.4, 7.6 and 8.0, add 70 ml, 60 ml and 42 ml of HCl, respectively). Adjust volume to 1 liter, then autoclave.

2 x TY: For 1 liter, combine 16 g of bacto-trypton, 10 g of bacto-yeast extract, 5 g of NaCl. Adjust pH to 7.5 with 2 N NaOH and bring to the final volume with distilled water. Autoclave.

2 M glycine: Dissolve 15.1 g of glycine in distilled water and bring to a final volume of 100 ml, autoclave and store at 4 °C.

2 M KOH: Dissolve 11.3 g of KOH in distilled water and bring to a final volume of 100 ml, autoclave.

3 M CH_3COONa: Dissolve 408.1 g of sodium acetate·$3H_2O$ in 800 ml of distilled water. Adjust pH to 5.2 with glacial acetic acid or adjust pH to 7.0 with dilute acetic acid. Adjust the volume to 1 liter with distilled water, and then autoclave.

5M NaCl: Dissolve 292.2 g of NaCl in distilled water and bring to a final volume of 500 ml, autoclave.

10 % acetic acid / 10 % methanol: Bring 200 ml of acetic acid and 200 ml of methanol to a final volume of 2 liters with water.

10 x kination buffer: For 1 ml, take 200 µl of 1M Tris-HCl (pH 8.0), 100 µl of 1M $MgCl_2$, 10 µl of 1M DTT and adjust to the final volume with distilled and sterilized water.

10 x PCR buffer: For 10 ml, take 2.5 ml of 2 M KCl, 1 ml of 1M Tris-HCl (pH 8.3), 150 µl of 1M $MgCl_2$, and 0.5 ml of 2% gelatin, and adjust to the final volume with distilled and sterilized water.

67 µM glycine-KOH: Mix 66.7 ml of 1M glycine-KOH and 33.3 ml of distilled water, autoclave and store at 4 °C.

100 mM IPTG: Dissolve 100 mg of IPTG in 4.2 ml of distilled water. Store at -20 °C.

BPB loading buffer: Adjust with distilled water to 50% glycerol and 0.25% BPB.

competent cell buffer: For 500 ml, take 25 ml of 1M $CaCl_2$ and 5 ml of 1M Tris-HCl and adjust to the final volume with distilled water, then autoclave.

dNTP mix: Take 125 µl each of dATP, dGTP, dCTP, dTTP (100 µmoles; Takara Biomedicals) and 9.5 ml of distilled water. Sterilize by filtration.

H plate: For 1 liter, combine 10 g of Bacto trypton, 8 g of NaCl, and 12 g of Bacto agar. Adjust to the final volume with distilled water. Autoclave. Dispense into petri dishes (20 ml/dish).

H top agar: For 250 ml, combine 2.5 g of Bacto trypton, 2.0 g of NaCl, and 2.0 g of Bacto agar. Adjust to the final volume with distilled water. Autoclave.

indicator mix: Prepare indicator cells by shaking 1/10 diluted overnight XL1Blue culture in 2 x TY at 37 °C. For 15 plates, mix together 450 µl each of 100 mM IPTG and 2% Xgal, and 2.1 ml of indicator cell.

λ exonuclease buffer: Take 100 µl of 67 µM glycine-KOH and 2.5 µl of 1M $MgCl_2$ and adjust to the final volume of 1 ml with distilled and sterilized water. Store in aliquots at -20 °C.

phenol / chloroform: Mix equal volume of phenol and chloroform. Equilibrate the mixture with 0.1M Tris (pH 7.5). Store in dark at 4 °C.

Sequence bottom gel: For 1 gel, dissolve 4.8 g of urea in 1.5 ml of 40% acrylamide stock, 2.5 ml of 10 x TBE buffer, 40% of sucrose and 0.2 ml of 1% BPB.

Sequence top gel: For 1 gel, dissolve 19.7 g of urea in 6 ml of 40% acrylamide stock, 2 ml of 10 x TBE buffer, and distilled water to the final volume of 40 ml.

TAE buffer: For 10 x stock solution, dissolve 48.44 g of Tris base and 27.22 g of $CH_3COONa{\cdot}3H_2O$, 15 ml of CH_3COOH, and 3.72 g of EDTA·Na and bring to a final volume of 1 liter with distilled water. $CH_3COONa{\cdot}3H_2O$ and CH_3COOH are occasionally substituted by 16.41 g of CH_3COONa.

TBE buffer: For 10 x stock solution, dissolve 108 g of Tris base, 55 g of Boric acid and 40 ml of 0.5M EDTA (pH 8.0) in a final volume of 1 liter.

TE (pH 8.0): For 500 ml, take 5 ml of 1M Tris (pH 8.0) and 1 ml of 0.5M EDTA (pH 8.0), bring to the final volume with distilled water, and autoclave.

Tris-saturated phenol: Dissolve phenol by placing in 65 °C waterbath. In a sterile container, add 500 ml of phenol add 250 ml of sterile 0.1M Tris (pH 7.5). Shake vigorously. Add 200 mg of 8-hydroxyquinoline and shake. Store in dark at 4 °C.

CHAPTER THREE

BASE COMPOSITIONS AND REPLICATION IN MITOCHONDRIAL DNA

Base composition biases in animals

One of the most puzzling characteristics of the animal mitochondrial genome, is the base compositions. Among the vertebrates there is a pronounced bias in the G + T composition, where it is high in the H-strands but low in the L-strands. The base composition biases are evident at all types of positions, within genes and in noncoding regions. The most biased positions in animal mtDNAs are the third codon positions of protein genes, which are the least constrained by the requirements of the coding function. Among the invertebrate mtDNAs, the fruit fly has the most biased base compositions, where 93% of the third codon positions of the second strand (L-strand in vertebrates) is either A or T (Clary and Wolstenholm 1985). Roundworm mtDNA has similar A + T-richness (Thomas and Wilson 1991). The third codon positions in sea urchins, on the other hand, are rich in A + C content (Jacobs et al. 1988; Cantatore et al. 1989; Asakawa et al. 1991). Among the vertebrates, the chicken has the strongest bias in base compositions, where 84% of the third codon positions of the L-strand are either A or C. The lowest A + C content is observed in frog mtDNA, the value being 66% (Roe et al. 1985). The base compositions of 4 hominoid mtDNAs also show the general feature of the low G+T (hence high A+C) content in the L-strand mtDNA (Table 3.1). Although there are various degrees of base composition biases in each strand, it is not known how and why these strand biases are maintained in animal mtDNAs.

Table 3.1 Base compositions in different regions of L-strand mtDNA

		1st codon				2nd codon				3rd codon			
Region	(bp)	A	T	G	C	A	T	G	C	A	T	G	C
ND1	954												
human	(H)	30.5	19.8	20.1	29.6	17.9	40.3	9.8	32.1	36.8	11.0	5.7	46.5
c. chimp	(C)	31.8	19.8	19.2	29.3	18.2	39.3	9.8	32.7	36.8	16.0	5.4	41.8
gorilla	(G)	31.5	20.1	19.2	29.3	18.2	40.9	9.8	31.1	36.5	15.1	5.7	42.8
orangutan	(O)	30.5	19.2	17.9	32.4	17.6	40.6	10.1	31.8	40.3	8.8	3.8	47.2
ND2	1041												
	H	39.2	19.6	13.5	27.7	17.9	41.2	9.5	31.4	36.6	16.1	5.8	41.5
	C	40.3	19.3	12.4	28.0	17.6	41.8	9.8	30.8	39.2	15.6	2.9	42.4
	G	38.0	18.4	14.4	29.1	17.9	40.9	9.5	31.7	37.5	18.2	3.7	40.6
	O	36.9	18.4	16.1	28.5	18.7	40.6	8.9	31.7	38.6	12.4	3.5	45.5
COI	1539												
	H	26.3	22.4	28.7	22.6	19.1	40.7	14.6	25.5	35.9	16.8	5.3	42.1
	C	26.1	22.6	28.9	22.4	19.1	40.5	14.6	25.7	34.9	18.7	6.2	40.2
	G	26.9	23.2	28.1	21.8	19.3	40.4	14.4	25.9	35.5	22.0	5.1	37.4
	O	26.7	23.6	27.9	21.8	18.9	40.0	14.8	26.3	34.1	18.1	5.8	41.9
COII	681												
	H	28.6	17.6	26.4	27.3	24.2	38.8	11.0	26.0	33.0	18.9	7.0	41.0
	C	27.8	17.2	27.3	27.8	23.8	39.2	11.5	25.6	37.0	22.5	3.5	37.0
	G	28.2	18.5	26.9	26.4	24.7	39.2	11.0	25.1	34.8	18.1	6.2	41.0
	O	27.8	18.5	26.4	27.3	24.2	39.2	11.0	25.6	35.2	13.7	6.2	44.9
ATP8	159												
	H	47.2	13.2	3.8	35.8	30.2	34.0	3.8	32.1	45.3	13.2	5.7	35.8
	C	41.5	15.1	7.5	35.8	30.2	34.0	3.8	32.1	49.1	13.2	3.8	34.0
	G	43.4	15.0	7.5	34.0	26.4	37.7	5.7	30.2	45.3	15.1	5.7	34.0
	O	37.7	13.2	7.5	41.5	24.5	32.1	3.8	39.6	43.4	9.4	5.7	41.5
ATP6	630												
	H	38.6	12.9	16.7	31.9	16.1	45.2	8.1	30.5	36.2	18.1	5.7	40.0
	C	36.2	12.9	18.1	32.9	16.2	45.2	8.6	30.0	37.6	21.0	5.2	36.2
	G	36.7	13.8	18.1	31.4	16.2	43.3	8.6	31.9	37.1	17.6	6.7	38.6
	O	37.6	13.3	15.7	33.3	15.2	45.2	9.0	30.5	34.8	17.6	5.7	41.9
COIII	783												
	H	23.8	25.7	21.8	28.7	21.5	36.4	16.1	26.1	35.2	17.6	6.9	40.2
	C	23.4	26.1	21.8	28.7	21.5	36.0	15.7	26.8	38.3	20.3	3.8	37.5
	G	23.8	25.3	22.2	28.7	21.5	37.5	15.3	25.7	37.2	17.6	5.0	40.2
	O	23.4	24.9	23.4	28.4	21.1	36.4	16.1	26.4	36.8	13.0	6.1	44.1
ND3	345												
	H	27.8	27.0	19.1	26.1	18.3	48.7	8.7	24.3	42.6	14.8	4.3	38.3
	C	27.8	25.2	19.1	27.8	18.3	47.0	8.7	26.1	42.6	12.2	3.5	41.7
	G	30.4	22.6	17.4	29.6	19.1	47.8	8.7	24.3	41.7	13.9	6.1	38.3
	O	27.0	21.7	19.1	32.2	20.0	44.3	7.8	27.8	44.3	13.9	3.5	38.3

(Table 3.1 Continue)

		1st codon				2nd codon				3 rd codon			
		A	T	G	C	A	T	G	C	A	T	G	C
ND4L	288												
	H	29.2	18.8	22.9	29.2	16.7	51.0	7.3	25.0	37.5	16.7	6.3	39.6
	C	30.2	19.8	21.9	28.1	16.7	51.0	7.3	25.0	43.8	20.8	2.1	33.3
	G	30.2	19.8	21.9	28.1	16.7	52.1	7.3	24.0	41.7	20.8	3.1	34.4
	O	33.3	21.9	19.8	25.0	16.7	50.0	7.3	26.0	35.4	7.3	6.3	51.0
ND4	1371												
	H	34.4	19.9	14.9	30.9	17.9	42.2	11.4	28.4	38.3	14.4	3.5	43.8
	C	34.6	21.2	14.2	30.0	17.7	41.8	11.6	28.9	36.3	15.3	5.5	42.9
	G	34.1	20.6	15.1	30.2	18.2	42.2	10.7	28.9	37.9	16.2	4.4	41.6
	O	35.0	18.2	14.7	32.2	17.7	42.0	11.6	28.7	36.8	12.0	2.8	48.4
ND5	1812												
	H	35.1	19.5	17.6	27.8	20.7	39.2	10.8	29.3	35.4	15.2	3.6	45.7
	C	35.1	20.2	17.4	27.3	20.9	39.6	10.8	28.8	35.3	17.7	3.3	43.7
	G	33.8	20.2	17.9	28.2	20.9	40.2	10.9	28.0	35.1	18.4	3.6	42.9
	O	35.6	17.9	16.9	29.6	20.0	39.2	10.6	30.1	33.9	12.6	4.0	49.5
*ND6	525												
	H	21.1	26.3	46.3	6.3	17.1	45.7	25.1	12.0	20.0	41.1	33.7	5.1
	C	22.9	25.7	45.1	6.3	17.1	44.0	25.7	13.1	21.1	42.3	33.1	3.4
	G	21.7	27.4	46.3	4.6	17.1	44.6	25.7	12.6	14.3	43.4	39.4	2.9
	O	20.6	26.3	47.4	5.7	16.6	43.4	25.7	14.3	13.7	41.7	41.7	2.9
Cytb	1140												
	H	29.5	23.4	19.5	27.6	20.0	40.0	12.9	27.1	36.3	12.1	3.7	47.9
	C	30.8	22.6	18.4	28.2	16.7	40.3	13.2	26.8	36.3	12.1	3.7	47.9
	G	30.0	23.2	18.7	28.2	20.0	39.7	13.2	27.1	36.8	11.3	3.2	48.7
	O	31.3	21.8	18.2	28.7	19.7	39.7	13.2	27.4	36.6	13.9	3.2	46.3

* For ND6, base compositions of H-strand mtDNA was calculated.

Region	(bp)	A	T	G	C	D-loop	(bp)	A	T	G	C
12S rRNA	948	33.9	21.6	19.3	26.1	H	1122	30.2	23.2	13.5	33.1
16S rRNA	1554	35.2	21.8	17.4	25.6	C	1113	30.0	22.2	13.7	34.1
tRNA (L)	949	35.5	27.8	15.3	21.4	P	1121	30.2	23.0	13.3	33.5
tRNA (H)	548	33.4	21.9	14.6	28.2	G	918	29.6	22.3	16.1	31.9
						O	917	27.6	23.4	15.8	33.2

Average of human, common chimpanzee, gorilla and orangutan was calculated for the tRNA and rRNA genes.

General features of base composition biases in hominoids

In protein coding regions, the base compositions for each codon positions are different from gene to gene, but are similar among hominoids. Opposed to the first and second codon positions, whose base compositions rely on the amino acid content of the respective gene, the third codon positions (where about 70% of the sites are synonymous) are distinctly biased in the same direction for all genes. In the L-strand, the A+C content of the third codon positions ranges from 74% in *COII* to 86% in *Cyt b*. The same bias applies even for *ND6*, which is encoded on the opposite strand. In *ND6*, the G+T content at the third codon positions of the H-strand is 78%, and this corresponds to the A+C content of the L-strand. Codon usages show that the codons perfectly pair the anticodons are not always preferred (Table 3.2). Rather, codons ending by both A and C are most often used in the four-codon degenerated families. In *ND6*, codons ending by G or T that do not pair perfectly with the anticodons are mostly preferred (Table 3.2). It seems that the base composition biases seriously affect the codon usages of *ND6*. Hence, whether or not codons match the anticodons of tRNAs with Watson-Crick pairings, is not the direct cause for the base compositional biases.

As noted above, tRNA genes are located on both of the mtDNA strands. One could therefore examine, whether the coding strand of the tRNA genes possess a trend for certain base composition biases. The tRNA coding regions are grouped into two in terms of the sense strand; the 14 tRNA genes that use the H-strand as the template strand are designated as tRNA(L), while the 8 tRNA genes that use the L-strand as the template strand is designated as tRNA(H). Regardless of the coding strand, the L-strand is high in A (ca. 34%) and low in G (ca. 15%) content for both tRNA(L) and tRNA(H), although the relative contents of T and C differ. The base compositions of the two rRNA genes also follow the general biases. For tRNA or rRNA, GC or AT contents in the stem parts of tRNA and rRNA genes will be 50% if all base pairs are by Watson-Crick pairings. The deviation of AC contents from 50% indicates that actual compositional biases are

Table 3.2 Codon usages in 13 protein genes

The abbreviations are: human (H), common chimpanzee (C), gorilla (G), orangutan (O). The codons that pair with the anticodons of the tRNA by complete Watson-Crick pairings are underlined. Trm: termination codon

ND1 Total 318 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:   4  7  5  5 |Ser:   2  3  3  2 |Tyr:   5  4  5  1 |Cys:   0  0  0  0 |U
  |Phe:  12  9 10 12 |Ser:   8  7  8 10 |Tyr:   9 10  8  8 |Cys:   0  0  1  1 |C
  |Leu:   5  5  5  3 |Ser:   8  8  8 10 |Trm:   0  0  0  0 |Trp:   9  9  8  8 |A
  |Leu:   1  0  2  0 |Ser:   0  1  0  0 |Trm:   0  0  0  0 |Trp:   0  0  1  1 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   6  4  2  3 |Pro:   2  3  5  3 |His:   0  0  0  0 |Arg:   0  0  1  0 |U
  |Leu:  19 18 21 19 |Pro:  17 16 13 13 |His:   2  3  3  4 |Arg:   2  2  1  2 |C
  |Leu:  27 29 27 38 |Pro:   2  3  3  6 |Gln:   6  4  6  7 |Arg:   5  5  4  5 |A
  |Leu:   5  4  6  3 |Pro:   1  0  0  0 |Gln:   0  2  0  0 |Arg:   0  0  1  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:  10 10  9  7 |Thr:   1  8  4  3 |Asn:   0  4  7  1 |Ser:   1  1  0  0 |U
  |Ile:  13 14 15 19 |Thr:  20 16 19 17 |Asn:  13 10  8 14 |Ser:   2  2  2  2 |C
  |Met:  13 13 15  9 |Thr:  14 11 10 12 |Lys:   6  7  7  6 |Trm:   0  0  0  0 |A
  |Met:   3  3  4  4 |Thr:   0  2  0  2 |Lys:   1  0  0  1 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   1  1  1  1 |Ala:   3  3  5  2 |Asp:   0  0  0  0 |Gly:   0  3  1  0 |U
  |Val:   5  5  4  2 |Ala:  14 13 13 15 |Asp:   4  3  3  3 |Gly:   8  5  7  9 |C
  |Val:   3  3  4  4 |Ala:  10 10  8  6 |Glu:   8  8  9 10 |Gly:   1  2  2  4 |A
  |Val:   1  0  0  0 |Ala:   0  0  0  0 |Glu:   3  3  2  1 |Gly:   3  2  2  0 |G
--+------------------+------------------+------------------+------------------+--
```

ND2 Total 347 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:   8  6  4  2 |Ser:   4  3  4  5 |Tyr:   2  2  6  3 |Cys:   0  0  0  0 |U
  |Phe:   7  7  9 12 |Ser:  11 13 14 10 |Tyr:   8  8  5  8 |Cys:   0  0  0  0 |C
  |Leu:   8  7  4  5 |Ser:   7  9  6  8 |Trm:   0  0  0  0 |Trp:  10 11 10 11 |A
  |Leu:   1  1  0  0 |Ser:   1  0  1  0 |Trm:   0  0  0  0 |Trp:   1  0  1  0 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   7  8  7  5 |Pro:   5  3  4  3 |His:   1  1  1  0 |Arg:   1  0  1  0 |U
  |Leu:  18 18 20 19 |Pro:  12 14 14 14 |His:   3  3  2  5 |Arg:   3  4  3  3 |C
  |Leu:  26 27 32 26 |Pro:   4  5  4  7 |Gln:   7 10  9  8 |Arg:   0  0  0  0 |A
  |Leu:   4  4  3  8 |Pro:   2  0  0  0 |Gln:   3  0  1  1 |Arg:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:   8 11 11 13 |Thr:   5  7 11  4 |Asn:   6  8  8  3 |Ser:   1  0  0  0 |U
  |Ile:  24 24 20 24 |Thr:  24 23 19 18 |Asn:  14 11 11 16 |Ser:   4  6  5  4 |C
  |Met:  23 24 17 15 |Thr:  11 12 14 18 |Lys:  10 11 10 12 |Trm:   0  0  0  0 |A
  |Met:   1  1  4  1 |Thr:   3  1  0  0 |Lys:   2  1  2  0 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   3  3  2  0 |Ala:   4  2  3  5 |Asp:   0  0  0  0 |Gly:   1  0  1  0 |U
  |Val:   2  1  2  5 |Ala:   7  6  9 10 |Asp:   0  0  1  3 |Gly:   7  9  7  7 |C
  |Val:   3  3  6  6 |Ala:   8  7  7  8 |Glu:   5  6  6  5 |Gly:   5  4  5  5 |A
  |Val:   0  0  1  0 |Ala:   1  2  0  0 |Glu:   1  0  0  1 |Gly:   0  0  0  1 |G
--+------------------+------------------+------------------+------------------+--
```

COI Total 513 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:  12 11 14 12 |Ser:   9  7 10  5 |Tyr:   4  8 11  6 |Cys:   0  0  0  0 |U
  |Phe:  29 30 28 30 |Ser:  10 11 11 12 |Tyr:  18 14 11 15 |Cys:   1  1  1  1 |C
  |Leu:   7  6  9 11 |Ser:   7  8  6 11 |Trm:   0  0  0  0 |Trp:  16 14 14 13 |A
  |Leu:   0  2  0  2 |Ser:   2  2  2  0 |Trm:   0  0  0  0 |Trp:   0  2  2  3 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   5  7  9  2 |Pro:   5  8  5  5 |His:   4  4  5  4 |Arg:   2  2  1  1 |U
  |Leu:  13 11  9 17 |Pro:  18 14 16 14 |His:  14 13 14 15 |Arg:   2  2  2  4 |C
  |Leu:  33 31 28 24 |Pro:   6  7  7  8 |Gln:   5  6  6  6 |Arg:   4  3  4  3 |A
  |Leu:   4  5  6  5 |Pro:   0  0  0  3 |Gln:   1  1  0  1 |Arg:   0  1  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:  12 19 17 21 |Thr:   5  2  5  4 |Asn:   4  5  7  8 |Ser:   0  1  2  0 |U
  |Ile:  26 19 23 19 |Thr:  11 15 13 10 |Asn:  13 12 10  9 |Ser:   4  3  2  5 |C
  |Met:  25 22 27 25 |Thr:  16 17 18 18 |Lys:   9 10  9 10 |Trm:   0  0  0  0 |A
  |Met:   7  9  4  4 |Thr:   2  0  0  4 |Lys:   1  0  1  0 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   4  4  3  2 |Ala:  11 10 11 11 |Asp:   2  3  7  5 |Gly:   7  5  6  7 |U
  |Val:  12 12 11 16 |Ala:  15 17 16 17 |Asp:  13 12  8  9 |Gly:  17 20 17 22 |C
  |Val:  17 18 15 13 |Ala:  13 12 12 13 |Glu:   7  9  8  7 |Gly:  19 16 19 13 |A
  |Val:   3  2  4  2 |Ala:   1  2  1  0 |Glu:   3  1  2  2 |Gly:   3  5  4  4 |G
--+------------------+------------------+------------------+------------------+--
```

COII Total 228 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:   3  7  5  2 |Ser:   1  0  2  3 |Tyr:   1  1  3  4 |Cys:   1  1  1  1 |U
  |Phe:   7  4  6  9 |Ser:   4  4  3  2 |Tyr:   8  8  6  5 |Cys:   2  2  2  2 |C
  |Leu:   4  3  5  2 |Ser:   4  5  4  5 |Trm:   0  0  1  1 |Trp:   3  4  4  3 |A
  |Leu:   1  0  1  3 |Ser:   0  0  0  0 |Trm:   1  1  0  0 |Trp:   1  0  0  1 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   7  5  3  2 |Pro:   1  2  1  3 |His:   2  3  3  1 |Arg:   2  2  1  0 |U
  |Leu:   6  6  7  8 |Pro:   9  8  9  5 |His:   4  3  3  5 |Arg:   1  1  1  3 |C
  |Leu:  12 16 12 16 |Pro:   3  5  5  7 |Gln:   6  6  6  6 |Arg:   3  3  4  3 |A
  |Leu:   3  2  4  2 |Pro:   2  0  0  0 |Gln:   1  1  1  1 |Arg:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:   7  7  4  3 |Thr:   5  7  3  5 |Asn:   2  1  1  1 |Ser:   1  2  1  0 |U
  |Ile:  15 14 19 20 |Thr:   8  5  6  6 |Asn:   5  5  7  6 |Ser:   0  0  0  1 |C
  |Met:   8  9  7  6 |Thr:   8  8  9  7 |Lys:   4  4  3  4 |Trm:   0  0  0  0 |A
  |Met:   2  1  3  3 |Thr:   0  0  0  1 |Lys:   0  0  1  0 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   1  5  3  1 |Ala:   4  2  3  3 |Asp:   3  4  3  1 |Gly:   2  2  4  1 |U
  |Val:   7  6  6  8 |Ala:   5  7  8  7 |Asp:   8  7  8 10 |Gly:   4  4  2  5 |C
  |Val:   5  3  4  4 |Ala:   4  4  3  3 |Glu:   8 10  9  9 |Gly:   3  4  4  5 |A
  |Val:   0  1  0  0 |Ala:   1  1  1  1 |Glu:   3  1  2  2 |Gly:   2  1  1  0 |G
--+------------------+------------------+------------------+------------------+--
```

ATPase 8 Total 53 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:   0  1  1  1 |Ser:   0  0  0  0 |Tyr:   1  0  0  0 |Cys:   0  0  1  0 |U
  |Phe:   1  0  1  1 |Ser:   0  0  0  0 |Tyr:   1  2  1  0 |Cys:   0  0  0  0 |C
  |Leu:   1  1  2  2 |Ser:   1  2  0  1 |Trm:   0  0  0  0 |Trp:   1  2  1  1 |A
  |Leu:   0  0  0  0 |Ser:   0  0  0  0 |Trm:   0  0  0  0 |Trp:   1  0  1  1 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   1  0  0  2 |Pro:   1  0  0  0 |His:   0  1  0  0 |Arg:   0  0  0  0 |U
  |Leu:   2  2  3  3 |Pro:   6  7  5  4 |His:   1  0  1  3 |Arg:   0  0  0  0 |C
  |Leu:   4  4  4  4 |Pro:   2  2  3  4 |Gln:   2  2  1  1 |Arg:   0  0  0  0 |A
  |Leu:   0  1  0  0 |Pro:   0  0  0  0 |Gln:   0  0  1  1 |Arg:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:   1  1  2  0 |Thr:   1  0  1  1 |Asn:   2  4  2  1 |Ser:   0  0  0  0 |U
  |Ile:   1  0  0  2 |Thr:   4  4  4  6 |Asn:   3  1  2  1 |Ser:   0  0  0  0 |C
  |Met:   5  5  4  0 |Thr:   2  1  2  3 |Lys:   4  5  5  5 |Trm:   0  0  0  0 |A
  |Met:   1  1  1  1 |Thr:   0  0  0  0 |Lys:   1  0  0  0 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   0  0  1  0 |Ala:   0  0  0  0 |Asp:   0  0  0  0 |Gly:   0  0  0  0 |U
  |Val:   0  1  0  1 |Ala:   0  1  1  1 |Asp:   0  0  0  0 |Gly:   0  0  0  0 |C
  |Val:   1  1  1  0 |Ala:   0  0  0  1 |Glu:   1  1  1  1 |Gly:   0  0  0  0 |A
  |Val:   0  0  0  0 |Ala:   0  0  0  0 |Glu:   0  0  0  0 |Gly:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
```

ATPase 6 Total 210 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:   4  4  3  4 |Ser:   4  3  1  0 |Tyr:   1  1  1  0 |Cys:   0  0  0  0 |U
  |Phe:   3  4  3  4 |Ser:   2  1  3  4 |Tyr:   2  2  2  3 |Cys:   0  0  0  0 |C
  |Leu:   4  6  7  6 |Ser:   3  3  4  3 |Trm:   0  0  0  0 |Trp:   3  3  4  3 |A
  |Leu:   1  0  1  1 |Ser:   0  0  0  0 |Trm:   0  0  0  0 |Trp:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   5  5  6  3 |Pro:   3  2  3  1 |His:   0  1  1  2 |Arg:   0  1  0  1 |U
  |Leu:   7  6  7 11 |Pro:   7  8  7  9 |His:   6  6  5  4 |Arg:   2  1  2  2 |C
  |Leu:  20 22 17 22 |Pro:   3  2  3  2 |Gln:   7  8  6  5 |Arg:   2  2  2  2 |A
  |Leu:   5  4  5  3 |Pro:   0  1  0  1 |Gln:   0  0  2  2 |Arg:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:  12 12  8 14 |Thr:   4  8  7  4 |Asn:   1  3  2  2 |Ser:   0  0  0  1 |U
  |Ile:  15 12 14 11 |Thr:  12  9 10 14 |Asn:   8  5  7  6 |Ser:   3  4  3  4 |C
  |Met:   8  8 10  7 |Thr:   9  8  9  7 |Lys:   5  4  3  4 |Trm:   0  0  0  0 |A
  |Met:   3  2  2  2 |Thr:   0  0  0  2 |Lys:   1  1  2  1 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   2  2  1  1 |Ala:   2  2  3  3 |Asp:   0  0  0  0 |Gly:   0  0  1  1 |U
  |Val:   1  3  3  3 |Ala:  11 10 12 10 |Asp:   1  1  0  1 |Gly:   4  4  3  2 |C
  |Val:   4  4  3  3 |Ala:   3  6  5  4 |Glu:   2  1  2  2 |Gly:   3  2  3  3 |A
  |Val:   1  1  1  0 |Ala:   1  0  0  0 |Glu:   0  1  1  0 |Gly:   0  1  0  0 |G
--+------------------+------------------+------------------+------------------+--
```

COIII Total 261 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:  13 11  9  6 |Ser:   2  3  2  3 |Tyr:   4  5  4  2 |Cys:   0  0  1  0 |U
  |Phe:  10 11 13 15 |Ser:   7  6  5  4 |Tyr:   7  7  7  9 |Cys:   1  1  0  2 |C
  |Leu:   3  3  4  4 |Ser:   8  7  9  8 |Trm:   0  0  0  0 |Trp:   9 11 10 10 |A
  |Leu:   0  1  0  0 |Ser:   0  1  0  0 |Trm:   0  0  0  0 |Trp:   3  1  2  2 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   3  3  3  2 |Pro:   2  2  2  2 |His:   5  3  3  2 |Arg:   0  0  2  1 |U
  |Leu:   9 10  8 10 |Pro:   6  6  6  7 |His:  12 13 14 15 |Arg:   2  2  1  0 |C
  |Leu:  16 16 19 15 |Pro:   3  5  4  3 |Gln:   9  9  9  9 |Arg:   4  3  2  3 |A
  |Leu:   3  2  2  5 |Pro:   1  0  0  0 |Gln:   0  1  0  0 |Arg:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:   7 11  7  6 |Thr:   2  4  3  2 |Asn:   1  2  2  1 |Ser:   1  0  0  0 |U
  |Ile:   7  3  8  8 |Thr:  10  8  9  8 |Asn:   5  4  4  4 |Ser:   3  3  4  5 |C
  |Met:   8  9  9  8 |Thr:  11 12  9 13 |Lys:   3  3  3  3 |Trm:   0  0  0  0 |A
  |Met:   3  1  3  2 |Thr:   1  1  1  1 |Lys:   0  0  0  0 |Trm:   0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   2  2  1  0 |Ala:   1  1  2  4 |Asp:   2  3  3  2 |Gly:   1  3  2  1 |U
  |Val:   4  5  6  8 |Ala:  11 11 11 10 |Asp:   1  0  0  1 |Gly:  10  8  9  9 |C
  |Val:   6  6  6  4 |Ala:   3  3  4  4 |Glu:   5  6  6  5 |Gly:   4  7  3  7 |A
  |Val:   1  0  0  2 |Ala:   0  0  0  0 |Glu:   2  0  1  2 |Gly:   4  2  4  2 |G
--+------------------+------------------+------------------+------------------+--
```

ND3 Total 115 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:   2  2  3  1 |Ser:   0  1  0  0 |Tyr:   1  0  0  0 |Cys    0  1  1  1 |U
  |Phe:   6  6  5  7 |Ser:   4  3  3  3 |Tyr:   2  3  3  4 |Cys    1  0  0  0 |C
  |Leu:  10  7  5  3 |Ser:   1  1  1  1 |Trm:   0  0  0  0 |Trp    4  4  3  4 |A
  |Leu:   0  1  1  0 |Ser:   0  0  0  1 |Trm:   0  0  0  0 |Trp    0  0  1  0 |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:   1  1  2  3 |Pro:   3  3  1  2 |His:   0  0  0  0 |Arg    0  0  0  0 |U
  |Leu:   4  4  2  2 |Pro:   2  3  4  4 |His:   0  0  0  0 |Arg    1  1  1  1 |C
  |Leu:  11 14 15 16 |Pro:   3  2  2  3 |Gln:   3  3  3  3 |Arg    0  0  0  0 |A
  |Leu:   2  1  3  3 |Pro:   0  0  1  0 |Gln:   0  0  0  0 |Arg    0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 A|Ile:   5  4  5  6 |Thr:   1  1  2  1 |Asn:   0  1  0  0 |Ser    1  0  1  0 |U
  |Ile:   4  3  4  3 |Thr:   4  5  5  5 |Asn:   4  3  5  4 |Ser    0  2  1  0 |C
  |Met:   7  7  6  5 |Thr:   2  2  2  4 |Lys:   3  3  3  3 |Trm    0  0  0  0 |A
  |Met:   1  0  1  0 |Thr:   0  1  0  0 |Lys:   0  0  0  0 |Trm    0  0  0  0 |G
--+------------------+------------------+------------------+------------------+--
 G|Val:   1  0  0  0 |Ala:   1  0  0  1 |Asp:   1  0  0  1 |Gly    0  0  1  0 |U
  |Val:   1  3  1  1 |Ala:   7  8  7  6 |Asp:   2  3  3  3 |Gly    2  1  0  1 |C
  |Val:   1  1  2  1 |Ala:   0  0  0  1 |Glu:   3  5  5  5 |Gly    1  0  1  2 |A
  |Val:   0  0  0  0 |Ala:   0  0  0  0 |Glu:   2  0  0  0 |Gly    0  1  0  0 |G
--+------------------+------------------+------------------+------------------+--
```

ND4L Total 96 codons

```
  |         U          |         C          |         A          |         G          |
  |                    |                    |                    |                    |
--+-------H--C--G--O--+-------H--C--G--O--+-------H--C--G--O--+-------H--C--G--O--+--
 U|Phe:  2  2  2  0  |Ser:  0  2  1  1  |Tyr:  2  3  2  1  |Cys:  0  0  1  0  |U
  |Phe:  1  1  2  3  |Ser:  4  2  2  3  |Tyr:  2  1  2  3  |Cys:  2  2  1  2  |C
  |Leu:  1  2  2  3  |Ser:  3  4  4  4  |Trm:  0  0  0  0  |Trp:  0  0  0  0  |A
  |Leu:  0  0  0  1  |Ser:  1  0  0  0  |Trm:  0  0  0  0  |Trp:  0  0  0  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
 C|Leu:  1  2  3  2  |Pro:  1  0  0  0  |His:  1  1  0  0  |Arg:  0  0  1  0  |U
  |Leu:  6  4  4  5  |Pro:  1  2  2  2  |His:  2  2  2  2  |Arg:  1  1  0  1  |C
  |Leu: 14 14 13 11  |Pro:  0  0  0  0  |Gln:  0  0  1  0  |Arg:  0  0  0  0  |A
  |Leu:  1  1  1  1  |Pro:  0  0  0  0  |Gln:  0  0  0  0  |Arg:  0  0  0  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
 A|Ile:  5  3  4  1  |Thr:  1  1  1  0  |Asn:  2  3  1  1  |Ser:  0  0  0  0  |U
  |Ile:  2  4  3  6  |Thr:  3  4  4  5  |Asn:  4  3  5  6  |Ser:  0  0  0  0  |C
  |Met:  9  9  9  8  |Thr:  1  1  1  2  |Lys:  0  0  0  0  |Trm:  0  0  0  0  |A
  |Met:  1  1  1  3  |Thr:  0  0  0  0  |Lys:  0  0  0  0  |Trm:  0  0  0  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
 G|Val:  0  1  0  0  |Ala:  1  1  3  1  |Asp:  0  0  0  0  |Gly:  0  1  1  0  |U
  |Val:  2  1  2  3  |Ala:  5  3  2  4  |Asp:  1  1  1  1  |Gly:  2  1  1  3  |C
  |Val:  2  4  3  1  |Ala:  2  4  3  3  |Glu:  2  2  2  1  |Gly:  2  2  2  1  |A
  |Val:  2  0  1  0  |Ala:  1  0  0  0  |Glu:  0  0  0  1  |Gly:  0  0  0  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
```

ND4 Total 457 codons

```
  |         U          |         C          |         A          |         G          |
  |                    |                    |                    |                    |
--+-------H--C--G--O--+-------H--C--G--O--+-------H--C--G--O--+-------H--C--G--O--+--
 U|Phe:  9  6  8  4  |Ser:  5  1  4  6  |Tyr:  2  6  5  2  |Cys:  1  1  0  0  |U
  |Phe: 11 15 11 11  |Ser: 17 19 18 18  |Tyr: 11  7  8 12  |Cys:  2  2  1  2  |C
  |Leu:  8 13 15  7  |Ser: 10 11  9  9  |Trm:  0  0  0  0  |Trp: 12 12 12 12  |A
  |Leu:  1  1  0  0  |Ser:  1  2  2  0  |Trm:  0  0  0  0  |Trp:  1  1  1  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
 C|Leu: 10  8 15 12  |Pro:  3  5  4  2  |His:  1  2  2  2  |Arg:  0  1  1  1  |U
  |Leu: 31 33 27 32  |Pro: 14 14 14 19  |His: 12 10 12 13  |Arg:  5  5  3  4  |C
  |Leu: 41 34 36 40  |Pro:  6  4  5  3  |Gln:  9  9  9  9  |Arg:  4  4  4  5  |A
  |Leu:  4  6  3  5  |Pro:  0  1  1  0  |Gln:  1  1  1  0  |Arg:  0  0  1  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
 A|Ile: 16 15 15  9  |Thr:  8  8  3  6  |Asn:  2  6  3  1  |Ser:  2  1  1  2  |U
  |Ile: 23 24 24 30  |Thr: 17 19 21 23  |Asn: 21 17 20 19  |Ser:  8  9  8 10  |C
  |Met: 24 20 25 27  |Thr: 22 21 21 19  |Lys: 10 10 10  9  |Trm:  0  0  0  0  |A
  |Met:  2  6  3  4  |Thr:  1  1  1  0  |Lys:  1  1  1  1  |Trm:  0  0  0  0  |G
--+-------------------+-------------------+-------------------+-------------------+--
 G|Val:  0  0  4  2  |Ala:  6  5  4  3  |Asp:  0  1  2  0  |Gly:  1  4  3  3  |U
  |Val:  4  2  0  2  |Ala: 12 12 15 13  |Asp:  3  2  1  4  |Gly:  9  6  7  9  |C
  |Val:  8  7  6  6  |Ala:  8  8  9 10  |Glu:  9  8  7  8  |Gly:  4  5  5  4  |A
  |Val:  1  1  1  1  |Ala:  0  1  1  0  |Glu:  0  1  2  1  |Gly:  3  2  2  1  |G
--+-------------------+-------------------+-------------------+-------------------+--
```

ND5 Total 604 codons

```
  |        U        |        C        |        A        |        G        |
  |                 |                 |                 |                 |
--+------H--C--G--O-+------H--C--G--O-+------H--C--G--O-+------H--C--G--O-+--
 U|Phe:  5 10 18  6 |Ser:  2  5  6  5 |Tyr:  6  7  3  3 |Cys:  1  0  0  2 |U
  |Phe: 33 27 22 27 |Ser: 20 17 15 15 |Tyr: 10 12 14 12 |Cys:  5  5  5  4 |C
  |Leu:  7 11  9  5 |Ser: 13 13 15 16 |Trm:  1  1  1  1 |Trp: 11 11 11 11 |A
  |Leu:  2  0  1  0 |Ser:  1  2  1  0 |Trm:  0  0  0  0 |Trp:  1  1  1  1 |G
--+-----------------+-----------------+-----------------+-----------------+--
 C|Leu: 11 13 14 17 |Pro:  7  7  9  3 |His:  1  1  4  4 |Arg:  1  1  1  2 |U
  |Leu: 31 32 31 39 |Pro: 16 18 14 20 |His: 13 10 11  9 |Arg:  3  4  5  2 |C
  |Leu: 45 45 43 42 |Pro:  9  6  7 11 |Gln: 17 19 18 18 |Arg:  3  3  3  3 |A
  |Leu:  8  5  8  7 |Pro:  0  0  0  1 |Gln:  3  1  2  1 |Arg:  0  0  0  0 |G
--+-----------------+-----------------+-----------------+-----------------+--
 A|Ile: 18 23 17 15 |Thr: 13 12 11  9 |Asn:  6  8  7  1 |Ser:  3  1  2  2 |U
  |Ile: 36 32 39 41 |Thr: 29 30 22 38 |Asn: 27 24 25 31 |Ser: 10 11 10 12 |C
  |Met: 23 25 24 21 |Thr: 22 22 23 16 |Lys: 20 20 21 20 |Trm:  0  0  0  0 |A
  |Met:  3  2  2  5 |Thr:  1  1  1  2 |Lys:  1  1  0  2 |Trm:  0  0  0  0 |G
--+-----------------+-----------------+-----------------+-----------------+--
 G|Val:  6  4  0  0 |Ala:  5  7 10  5 |Asp:  3  3  2  0 |Gly:  4  5  7  2 |U
  |Val:  4  4  8  4 |Ala: 19 17 18 26 |Asp:  8 10  9  9 |Gly: 12 11 11 10 |C
  |Val:  5  3  5  8 |Ala: 20 17 16 15 |Glu:  8  6  8  7 |Gly: 10 11  8 11 |A
  |Val:  0  3  2  0 |Ala:  0  0  1  0 |Glu:  1  3  1  3 |Gly:  1  1  2  2 |G
--+-----------------+-----------------+-----------------+-----------------+--
```

ND6 Total 175 codons

```
  |        U        |        C        |        A        |        G        |
  |                 |                 |                 |                 |
--+------H--C--G--O-+------H--C--G--O-+------H--C--G--O-+------H--C--G--O-+--
 U|Phe:  8  7  7  7 |Ser:  3  3  3  2 |Tyr: 11 11 11 10 |Cys:  1  1  2  3 |U
  |Phe:  2  0  0  0 |Ser:  0  0  1  1 |Tyr:  0  1  0  0 |Cys:  0  0  0  0 |C
  |Leu:  8 11 12  8 |Ser:  2  1  1  1 |Trm:  0  0  0  0 |Trp:  3  1  2  2 |A
  |Leu:  6  5  6  8 |Ser:  0  0  0  1 |Trm:  0  0  0  0 |Trp:  2  4  3  3 |G
--+-----------------+-----------------+-----------------+-----------------+--
 C|Leu:  0  0  0  0 |Pro:  4  3  4  3 |His:  0  0  0  0 |Arg:  1  1  1  1 |U
  |Leu:  0  0  0  0 |Pro:  0  1  0  1 |His:  0  0  0  0 |Arg:  0  0  0  0 |C
  |Leu:  0  0  0  0 |Pro:  1  1  0  0 |Gln:  0  0  0  0 |Arg:  0  0  0  0 |A
  |Leu:  3  3  1  3 |Pro:  0  0  0  0 |Gln:  0  0  0  0 |Arg:  2  2  2  2 |G
--+-----------------+-----------------+-----------------+-----------------+--
 A|Ile: 12 11 12 11 |Thr:  2  2  2  2 |Asn:  3  4  4  4 |Ser:  3  5  4  3 |U
  |Ile:  0  1  0  0 |Thr:  0  0  0  0 |Asn:  1  0  1  0 |Ser:  2  1  0  2 |C
  |Met:  2  4  0  1 |Thr:  1  2  2  2 |Lys:  1  1  1  0 |Trm:  0  0  0  0 |A
  |Met:  8  6  9  7 |Thr:  0  1  1  1 |Lys:  1  1  1  2 |Trm:  1  1  1  1 |G
--+-----------------+-----------------+-----------------+-----------------+--
 G|Val:  9 12 10 11 |Ala:  4  4  3  5 |Asp:  3  3  3  2 |Gly:  8  7 10  9 |U
  |Val:  4  0  2  0 |Ala:  0  2  1  0 |Asp:  0  0  0  1 |Gly:  0  0  0  0 |C
  |Val:  9  9  3  4 |Ala:  1  1  1  2 |Glu:  1  1  2  3 |Gly:  6  5  1  1 |A
  |Val:  9  8 16 16 |Ala:  3  2  3  4 |Glu:  9  8  7  7 |Gly: 15 17 19 18 |G
--+-----------------+-----------------+-----------------+-----------------+--
```

Cytb Total 380 codons

```
  |        U         |        C         |        A         |        G         |
  |                  |                  |                  |                  |
--+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+-------H--C--G--O-+--
 U|Phe:  7  9  7  7  |Ser:  0  2  3  3  |Tyr:  6  4  3  2  |Cys:  0  0  0  0  |U
  |Phe: 17 16 21 15  |Ser: 11  9  8  8  |Tyr: 11 13 13 14  |Cys:  2  2  2  2  |C
  |Leu:  7  7  6  9  |Ser: 14  9 11 11  |Trm:  0  0  0  0  |Trp: 11 10 12 10  |A
  |Leu:  2  1  2  0  |Ser:  0  2  0  0  |Trm:  0  0  0  0  |Trp:  1  2  0  2  |G
--+------------------+------------------+------------------+------------------+--
 C|Leu:  7  9  3  7  |Pro:  4  2  6  1  |His:  2  1  1  2  |Arg:  0  0  0  0  |U
  |Leu: 21 17 19 18  |Pro:  9  9  8 11  |His: 10 11 12 12  |Arg:  3  3  3  3  |C
  |Leu: 25 28 29 28  |Pro:  9 11  8 12  |Gln:  8  7  9  7  |Arg:  4  4  3  4  |A
  |Leu:  2  3  3  3  |Pro:  1  1  2  0  |Gln:  0  1  0  1  |Arg:  0  0  1  0  |G
--+------------------+------------------+------------------+------------------+--
 A|Ile: 11  9  9 13  |Thr:  3  0  1  7  |Asn:  3  2  1  2  |Ser:  1  0  0  0  |U
  |Ile: 28 31 27 20  |Thr: 13 16 19 17  |Asn: 12 13 13 13  |Ser:  3  4  5  6  |C
  |Met: 12 10 13 15  |Thr: 13 21 15 15  |Lys:  8  8  9  6  |Trm:  0  0  0  0  |A
  |Met:  3  3  1  3  |Thr:  1  0  1  0  |Lys:  1  0  0  2  |Trm:  0  0  0  0  |G
--+------------------+------------------+------------------+------------------+--
 G|Val:  1  0  1  1  |Ala:  0  4  2  4  |Asp:  1  2  3  2  |Gly:  0  2  3  2  |U
  |Val:  3  5  4  5  |Ala: 17 12 12 12  |Asp: 10  8  8  8  |Gly: 12 13 11 12  |C
  |Val:  6  5  6  7  |Ala:  8  4  7  3  |Glu:  4  4  4  4  |Gly:  9 10  8  8  |A
  |Val:  0  0  0  0  |Ala:  0  0  0  0  |Glu:  0  1  0  0  |Gly:  3  0  2  1  |G
--+------------------+------------------+------------------+------------------+--
```

much stronger in the loops of tRNAs and rRNAs. The D-loop noncoding region also shows the general base composition biases. Although it is noncoding, the level of bias is lower than the third codon positions of protein genes. The G content of the D-loop region is ca. 15%, which is higher than that of the third codon positions (which ranges from 2% to 7%), and is close to that of tRNA genes.

Base composition biases and asymmetric replication

If the base composition biases is not related to the function of the strands, then we must look at other ways in which the two strands differ. One possibility is in their mode of replication. Replication of the two strands of the mitochondrial genome initiates from separate origins, O_L and O_H. O_L and O_H differ in both primary sequence and some factors involved in replication (Clayton 1982; 1991). It is therefore reasonable to expect some differences in the replication of the two strands, which give rise to a different probability of misincorporation. In such a model, a polymerase complex for the replication of the H-strands prefers not to incorporate Cs. Reversely, an L-strand replication complex prefers not to incorporate Gs.

Another way for a directional mutation pressure to cause the strand biases may be a different potential for damage faced during the replication. As noted, vertebrate mtDNA replication is asymmetric (Figure 3.1). The H-strand synthesis starts first from O_H. The newly made H-strand proceeds, displacing the parental H-strand, to two thirds of the way around the mtDNA and thereby exposes the O_L. During that time, the parental H-strand spends a single-stranded state. On the other hand, the daughter L-strand synthesis begins from O_L, and proceeds using the parental H-strand as template. Therefore, the L-strand never become single-stranded. Among the spontaneous lesions in DNA, deamination of C occur 250 times more frequently in a single-stranded DNA than in double-stranded DNA (Friedberg 1985).

Since deamination of C preferentially causes C to T substitutions in the single-stranded stage of H-strand, this may be another possible molecular mechanism for the base composition biases (Brown and Simpson 1982). If it is so, the base composition bias may be stronger in the region that has a longer single-stranded state. Substitutions that decrease C and increase T in the H-strand will give rise to decreasing of G and increasing of A in the L-strand. I have examined if the time spent at the single stranded state has any relation with the directional mutation pressure by the relative frequency of G to A (G/A) at the third codon positions of the L-strand (Fig 3.2). The decrease in the value G/A indicates a decrease in G and an increase in A in the L-strand, corresponding to a decrease in C and increase in T in the H-strand. For all four hominoids, the value G/A of the L-strand showed a tendency to decrease as the time assumed at the single-stranded state increases. However, the correlation between the duration of single-stranded state in the H-strand, and the decrease of G content in the L-strand was significant only in gorilla and orangutan ($df = 9, P < 0.01$).

If the base composition biases are caused by a directional mutation of a particular type of substitutions, we would expect to observe preponderance in such differences by the comparison of closely related species. To examine more quantitatively whether there is any increase in a particular type of substitutions in relation to the duration of single-stranded state in the H-strand, I have counted the observed synonymous differences between human and chimpanzee (Figure 3.3). It should be noted that the direction of substitutions is ignored in this comparison. Unless obscured by multiple hit substitutions, an increase in number of the G to A substitutions in the L-strand due to C deamination, should appear as an increase in AG differences between human and chimpanzee. However, TC nor AG differences show apparent relation to the duration of the single-stranded state. I admit, however, that the number of differences used in the analysis of Figure 3.3 are not large and may be subjected to large amount of sampling errors. As will be discussed later, synonymous differences between human

Figure 3.1 A) Asymmetric replication of vertebrate mtDNA and B) relationship between the nucleotide positions at the H-strand and the time at the single stranded state.

A) 1)The replication of the H-strand (shown by the dashed arrow H) begins from a unique origin O_H, and displaces the parental H-strand (thick line). 2) The L-strand replication begins from another unique origin O_L when it is exposed by the newly synthesized H-strand. 3) H-strand replication continuously proceeds till complete. Therefore, the parental L-strand is never single-stranded.

B) Time spent in the single stranded state for the H-strand vertebrate mtDNA is inferred from the distance from the origin, a polymerization rate of 270 nucleotides per minute, and the details of replication proposed by Clayton (1982).

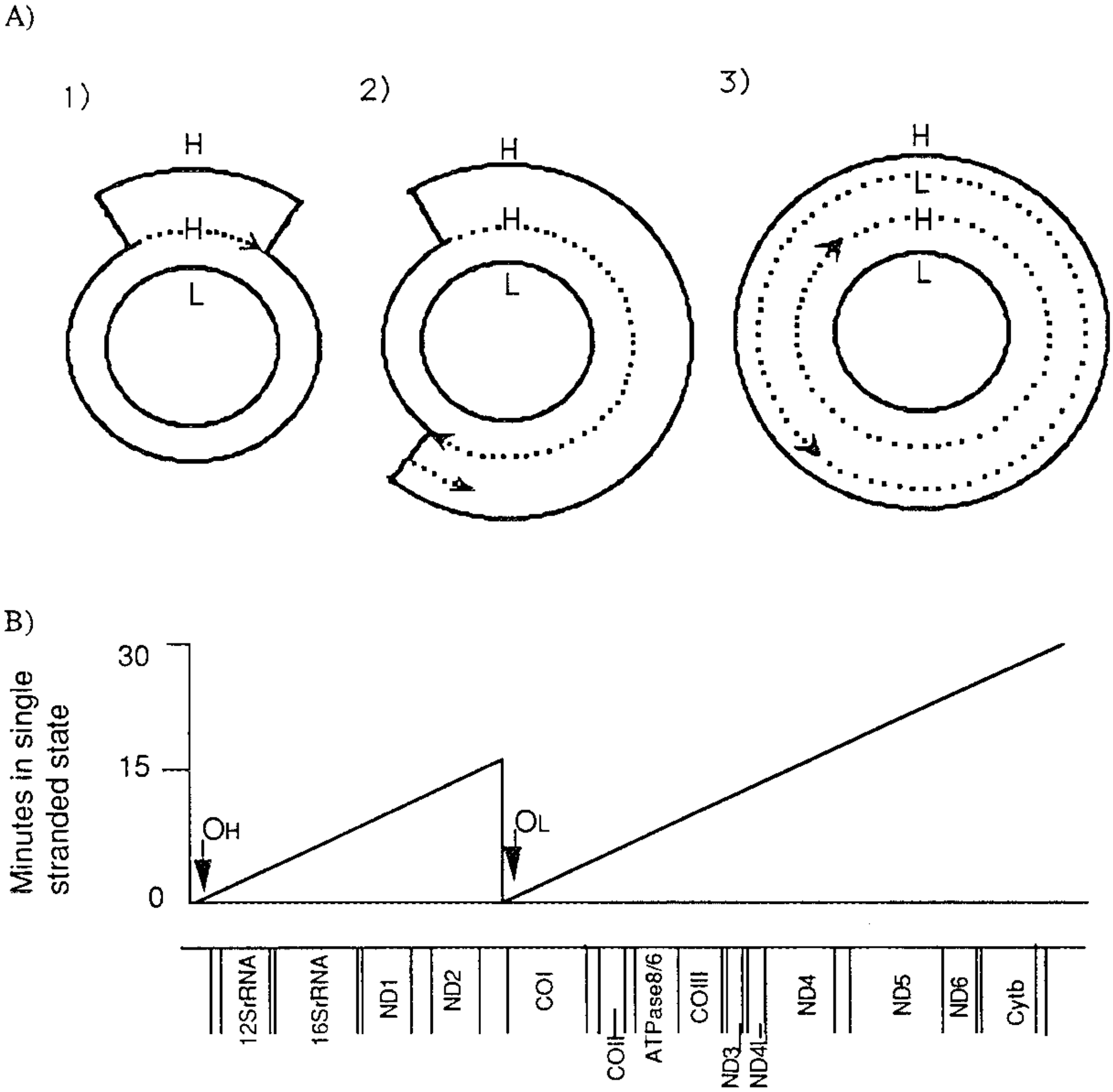

A)
1)
H
H
L
2)
H
H
L
3)
H
L
H
L
B)
30
15
0
Minutes in single stranded state
OH
OL
12SrRNA
16SrRNA
ND1
ND2
COI
COII
ATPase8/6
COIII
ND3
ND4L
ND4
ND5
ND6
Cytb

Fig 3.2 Relationship between the base compositions and the distance from the replication origin.

For the third codon positions of the L-strand, the base content of G over A from 11 protein genes were plotted against the distance from the O_L. Decrease in the G/A indicates decrease of G in the L-strand, hence decrease of C in the H-strand.

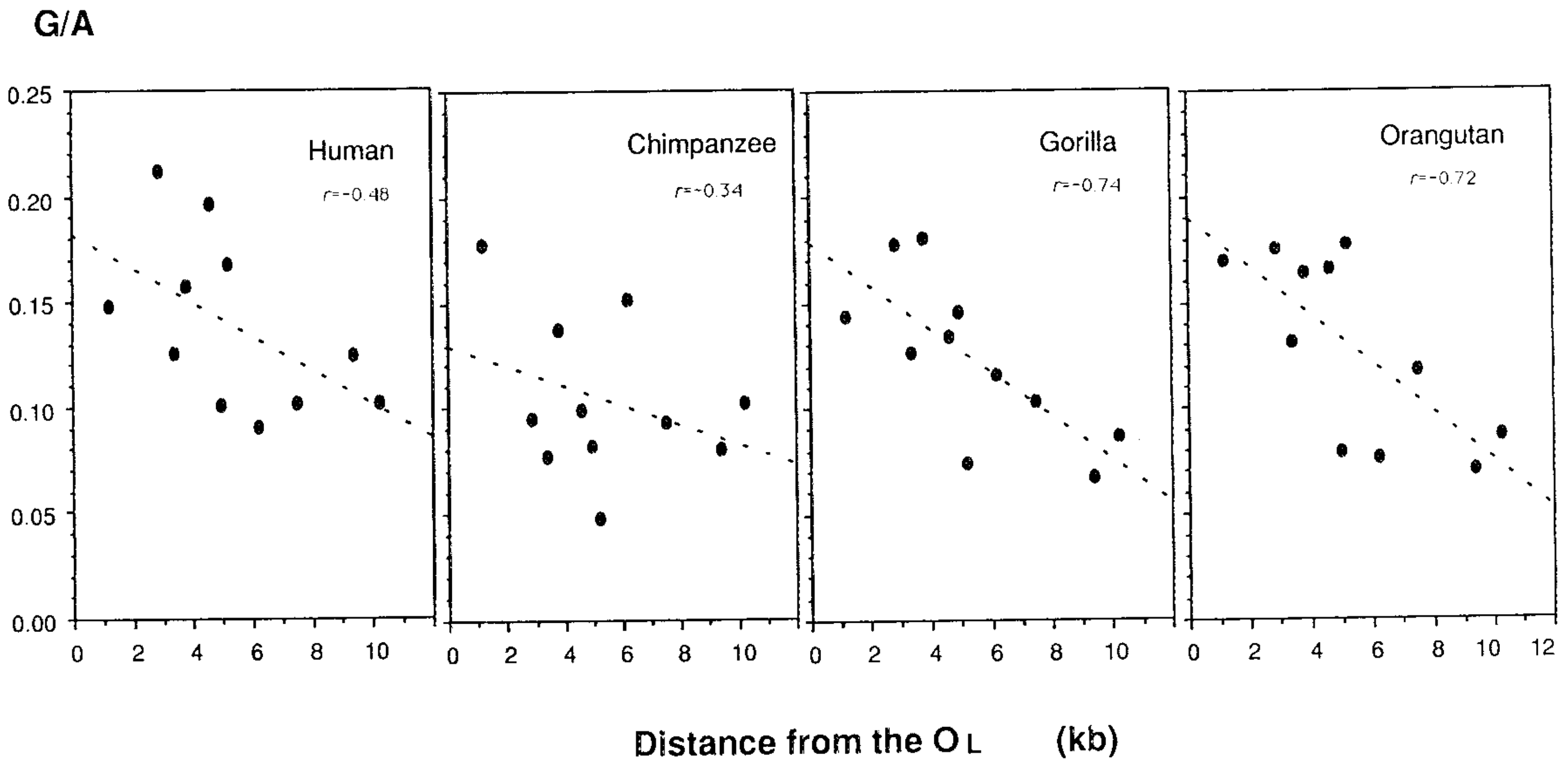
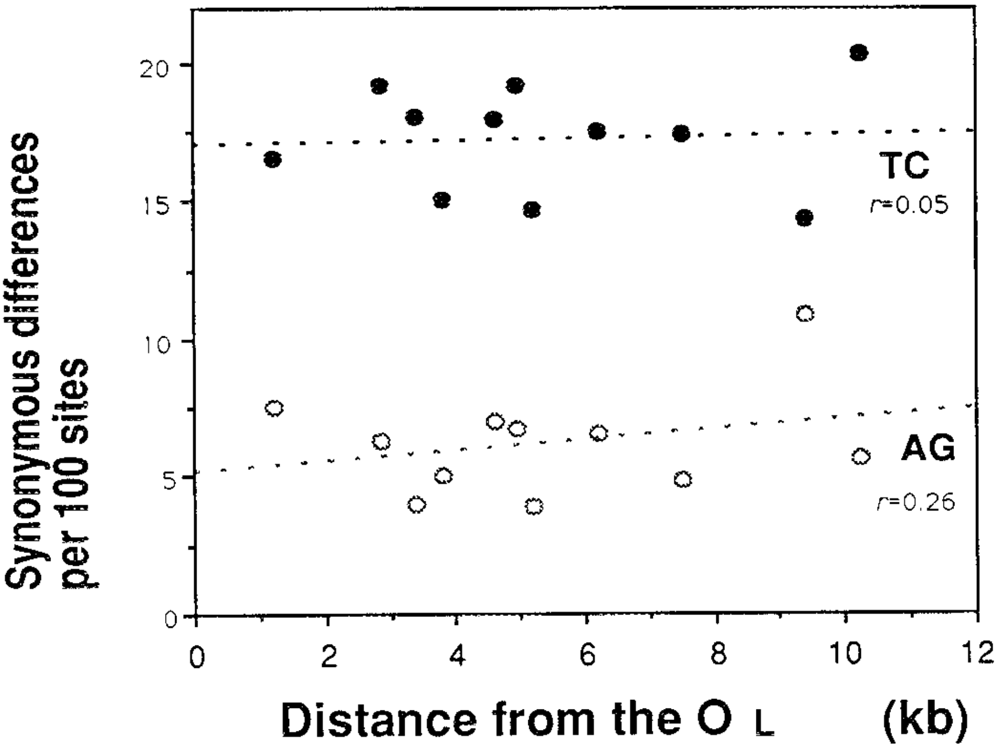

G/A
0.25
0.20
0.15
0.10
0.05
0.00
Human
r=-0.48
Chimpanzee
r=-0.34
Gorilla
r=-0.74
Orangutan
r=-0.72
0
2
4
6
8
10
12
Distance from the OL (kb)

Figure 3.3 Relationship between the synonymous differences and the distance from the replication origin.

Synonymous differences per site between human and common chimpanzee were plotted against the distance from the O_L. The solid dots indicate the TC synonymous differences and white dots represent the AG synonymous differences.


Synonymous differences per 100 sites
0
5
10
15
20
0
2
4
6
8
10
12
TC
r=0.05
AG
r=0.26
Distance from the O L (kb)

and gorilla or chimpanzee and gorilla would be too large and involve serious amount of multiple hit substitutions, so I decided not to use them .

Conclusions

As observed in other vertebrate mtDNAs, hominoid mtDNAs possess a strong bias in the G+T compositions, where it is high in the H-strand and low in the L-strands. The base composition biases are evident along the whole genome. Independent of the coding genes, the L-strand of mtDNA is always rich in A and C, and low in G content. Therefore, the bias in the base compositions is not due to constraints imposed by function. It is likely that the evolution of mtDNA is influenced by directional mutation pressure, as indicated by low GC contents in gram-positive eubacteria (Osawa et al. 1990).

A possible cause for the strand biases may be a different potential for damage faced by replication. Due to the asymmetric replication, the H-strand transiently becomes single-stranded. This single-stranded molecule is subjected to a higher probability of deamination at C, and results in reduced numbers of C and increased numbers of T in the H-strand. The base compositions in the rearranged region in marsupial, and the different base composition biases in the Echinoderms present supportive observations to the model (Thomas and Wilson 1991). This model has also been suggested from the observations in carp, rat and other mammalian mtDNAs (N. Nikou et al. personal communication). However, observations from hominoid mtDNAs suggest that the C deamination may not only be the cause for the base compositional biases. Additional process is also required to explain the bias between A and G compositions in the H-strand. Detailed studies of mitochondrial replication in diverse groups of animals should provide great insights into the relationships between replication and evolution by base substitution. Whatever the molecular mechanisms are, directional mutation pressure appears to operate in animal mtDNAs.

CHAPTER FOUR

STRUCTURAL AND FUNCTIONAL PROPERTIES OF MITOCHONDRIAL GENES

In this chapter I will discuss about the structural and functional characteristics of the mitochondrial protein genes, tRNA genes, rRNA genes and noncoding regions by examining the distribution of variable sites in structure models of these regions.

Protein genes

There are 13 protein genes encoded in mtDNA, which belong to five respiratory complexes (Table 4.1). Of these, *ND6* is the only protein gene that use the L-strand for the template strand (Figure 2.1). There are two cases (*ATPase 8* and *ATPase 6*, and *ND4L* and *ND4*), where two proteins overlap partially in the reading frames. However, tRNA genes between the genes for protein and ribosomal RNAs, seem to serve as punctuation signals for the processing of the primary transcripts (Ojala et al. 1981). Some properties of mitochondrial mRNAs can be given: They lack a 5' noncoding leader sequence and the cap structure which generally participate in the binding of mRNA of protein genes to bacterial and eucaryotic ribosomes, respectively. In most cases, mitochondrial protein genes terminate with an incomplete stop codon (U or A), and the stop codon is created by addition of poly (A) tail at their 3' end that is added post-transcriptionally by a mitochondrial poly-A polymerase (Ojala et al. 1981).

1. Codon strategy

Animal mitochondrial genetic codes differ slightly from the universal (i.e. eukaryotic nuclear and prokaryotic) ones. Variations are also notable within the mitochondrial genetic codes (reviewed in Attardi 1985). This is because mitochondrial codes are translated so as to require fewer tRNA species than those required by the

Table 4.1 Respiratory complexes and mtDNA subunits

	Complex	Polypeptides	mtDNA encoded polypeptides	
I	NADH ubiquinone oxidoreductase	25	7	ND1, ND2, ND3, ND4, ND4L, ND5, ND6
II	Succinate ubiquinone oxidoreductase	5	0	
III	Ubiquinol cytochrome *c* oxidoreductase	9-10	1	Cytochrome *b*
IV	Ferrocytochrome *c* oxygen oxidoreductase	8	3	COI, COII, COIII
V	ATP-synthase	12-14	2	ATPase 6, ATPase 8

eukaryotic nuclear and prokaryotic systems. None of the genetic codes differ radically from one another. This suggests that the mitochondrial codes are altered descendants of the universal codes.

In bacterial or lower-eukaryote genes, codon usage is closely related to the cellular concentration of tRNAs, which may be important in regulating the supply of aminoacyl-tRNAs for efficient translation (Ikemura 1981, 1982). This may not be applicable to the mitochondrial system, because in mitochondria, codon members specifying each amino acid is read by one tRNA species.

Initiation codons

Protein synthesis in mitochondria starts at N-formylmethionine (ATG, ATA) as in bacteria. However, sequences of the 5' end of mRNAs has also revealed that the mature transcripts could start from the base immediately after the tRNA at the same strand and not from the initiation codon (Montoya et al. 1982). Among hominoid mtDNAs, initiation codons vary in *ND1*, *ND3* and *ND5* (Figure 4.1). The *ND1* mRNA has been sequenced in humans: The initiation codon ATA lies after two spacer bases, and after a triplet codon, there is another methionine ATG which assures the reading frame. In common chimpanzee, the initiation codon ATA(Met) in human is substituted by ACA(Thr). This suggests that chimpanzee *ND1* either uses ACA(Thr) as initiator codon or initiate the reading frame from the ATG, two codons down stream. The variability in the reading frame and the initiation codons in *ND1* has been noted in mouse, rat, and Xenopus (Bibb et al. 1981; Roe et al. 1985; Gadeleta et al. 1989). In orangutan *ND1*, the second methionine ATG conserved in other hominoids is replaced by GTA(Val). Then the initiation codon ATA in human is replaced by initiation codon ATG. Such feature is also observed in whale *ND1* (Árnason et al. 1991). This suggests that ATG may possibly be a stronger initiation codon than ATA or other initiation codons.

Figure 4.1 Variations in initiation and termination codons in hominoid mtDNAs.

Initiation and termination codons that are variable among the hominoids are listed in the table, together with the alignment in the vicinity of the initiation or termination codons. The termination or initiation codon is indicated by the bold letter in the alignment. Only the bases different from human is shown. Abbreviations for the hominoid species are: CHIMP-common chimpanzee, GORIL-gorilla, ORANG-orangutan, PYGMY-pygmy chimpanzee, SIAMA-siamang.

Initiation codons

Gene	Human	Chimpanzee	Gorilla	Orangutan
ND1	ATA	ACA*	ATA	ATG
ND3	ATA	ATA	ATA	ATT*
ND5	ATA	ATA	ATA	ACA*

```
ND1 SerUCN←        ND1→
HUMAN    TTCTTAACAACATACCCATGGCCAACCTCCTACTC
CHIMP         G      C       A
GORIL              T   T       T     T
ORANG              A  G  TG AAT         G

ND3                Gly→   ND3 →
HUMAN    ACATTCAAAAAAGAGTAATAAACTTCGCCTTAATTT
CHIMP                                T C
GORIL    GT CC                         C G  C
ORANG       C               T       T C  GC C

ND5            LeuCUN→  ND5  →
HUMAN    CTCCAAATAAAAGTAATAACCATGCACACTACTATA
CHIMP                            T TG    C
GORIL                        T   T  G    C
ORANG                    C G     TT   C  C
```

Termination codons

Gene	Human	Chimp	Pygmy	Gorilla	Orangutan	Siamang
COI	AGA	AGA	AGA	AAA*	GAG*	AGA
COII	TAG	TAG	TAA	TAA	TAA	TAA

```
COI              COI→     SerUCN←
HUMAN    GTATACATAAAATCTAGACAAAAAAGGAAGGAATCG
CHIMP
PYGMY      G
GORIL         T     G    A
ORANG      C  T      C CGAG
SIAMA      C  T      C

COII                  COII→      noncoding
HUMAN    ATAGGGCCCGTATTTACCCTATAGCACCCCCTCTA-C
CHIMP         A        C  T          TT     C
PYGMY         A        C  T     A    TT     C
GORIL         A        CG      AT     TCTC T
ORANG         A        CG TT    A T TT  A CCC
SIAMA            T              A C  G  CTCTG-
```

The initiation codon ATA of human *ND5*, lies immediately after the tRNA$^{Leu(CUN)}$, and after a triplet there is another methionine ATG, which is conserved among the hominoid *ND5*. In orangutan, however, the reading frame initiates from the ACA(Thr), directly after a tRNA, assured by another methionine ATG two codons down stream. The use of ATC initiation codon has been suggested in mouse *ND5*.

ND3 is another case where a transcript initiate immediately after the tRNA. In orangutan, the initiation codon ATA in other hominoids is replaced by ATT (Ile). Such a variation in the ND3 initiation codon is also noted in rat (ATT) and mouse (ATC).

Termination codons

Mammalian mtDNA generally use the termination codons TAA, TAG, AGA and AGG. AGA and AGG are an example for the exchange in genetic codes, where they code for arginine in the universal code, and serine in Drosphila mtDNA. The *COI* and *COII* genes have variable termination codons, although they are the two-best conserved protein coding genes at the amino acid level, as discussed later. The *COI* termination codon is AGA in human, chimpanzees and siamang, but the same codon position is substituted by AAA (Lys) in gorilla and GAG (Asp) in orangutan. Because AAA and GAG are frequently used in the *COI* coding region, an acquisition of such new termination codons seem to be unlikely. It is possible that AGG occurring in down-stream can be used as a termination codon. This down-stream termination codon would cause a 10 bp overlap between *COI* and *tRNA*$^{Ser(UCN)}$ genes. However, this overlap would not give rise to any serious effect on transcription because these two genes are coded on the opposite strands. Furthermore, amino acid replacements in *COI* are more frequent in the 3' end than in other regions, indicating weak functional constraints in this region. Thus elongation of the C-terminal of *COI* may not be functionally defective. Similar instances of *COI* have also been noted in bovine and mouse (Anderson et al. 1982; Bibb et al. 1981). It therefore seems to be a general phenomenon that the 3' end of *COI* is flexible and the gene has been using different termination codons in different positions. The *COII* termination codon is variable among hominoids, although the position is fixed; TAG in common chimpanzee and human, TAA in pygmy chimpanzee, gorilla, orangutan and siamang.

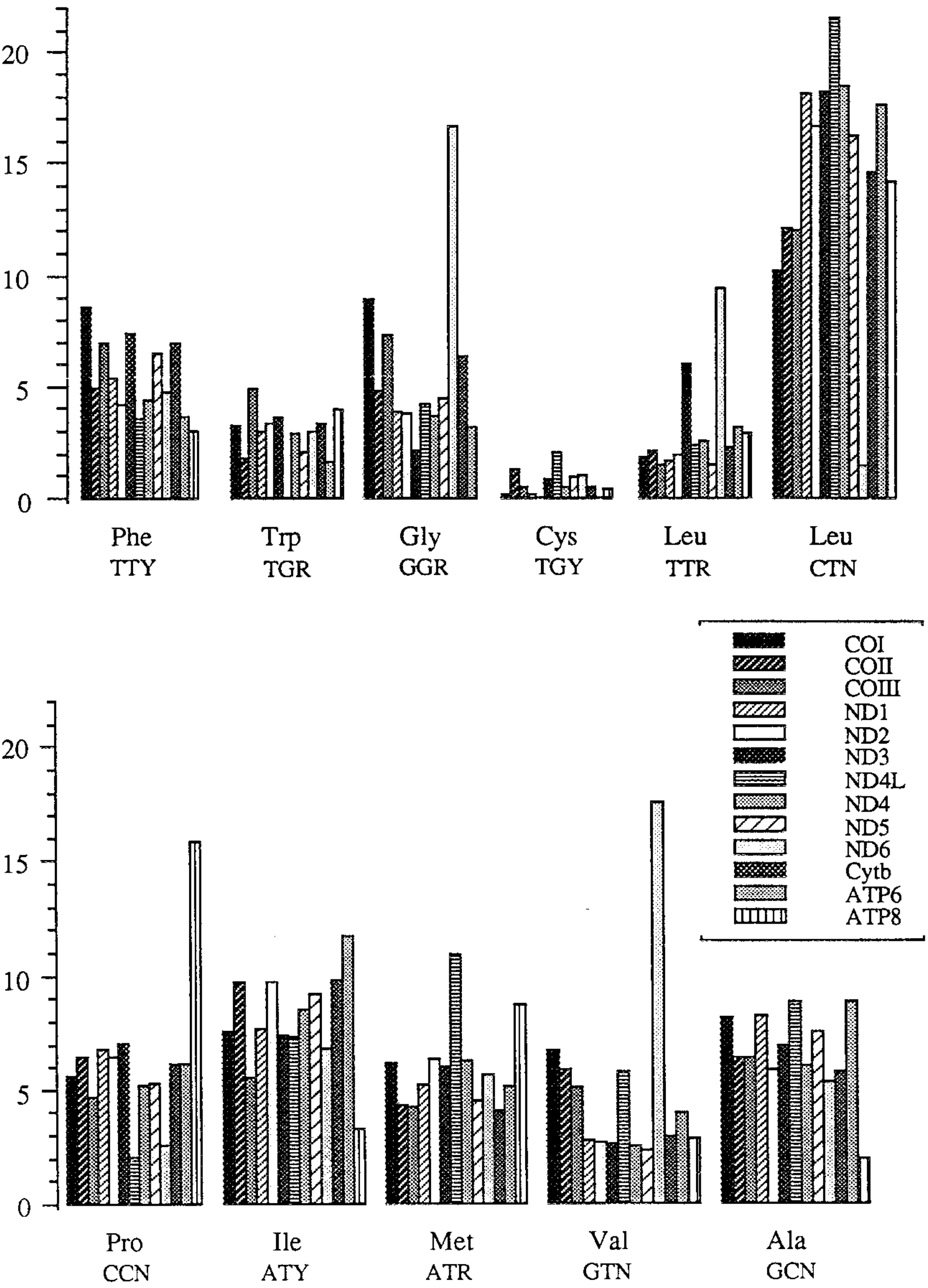
Codon usage for each gene

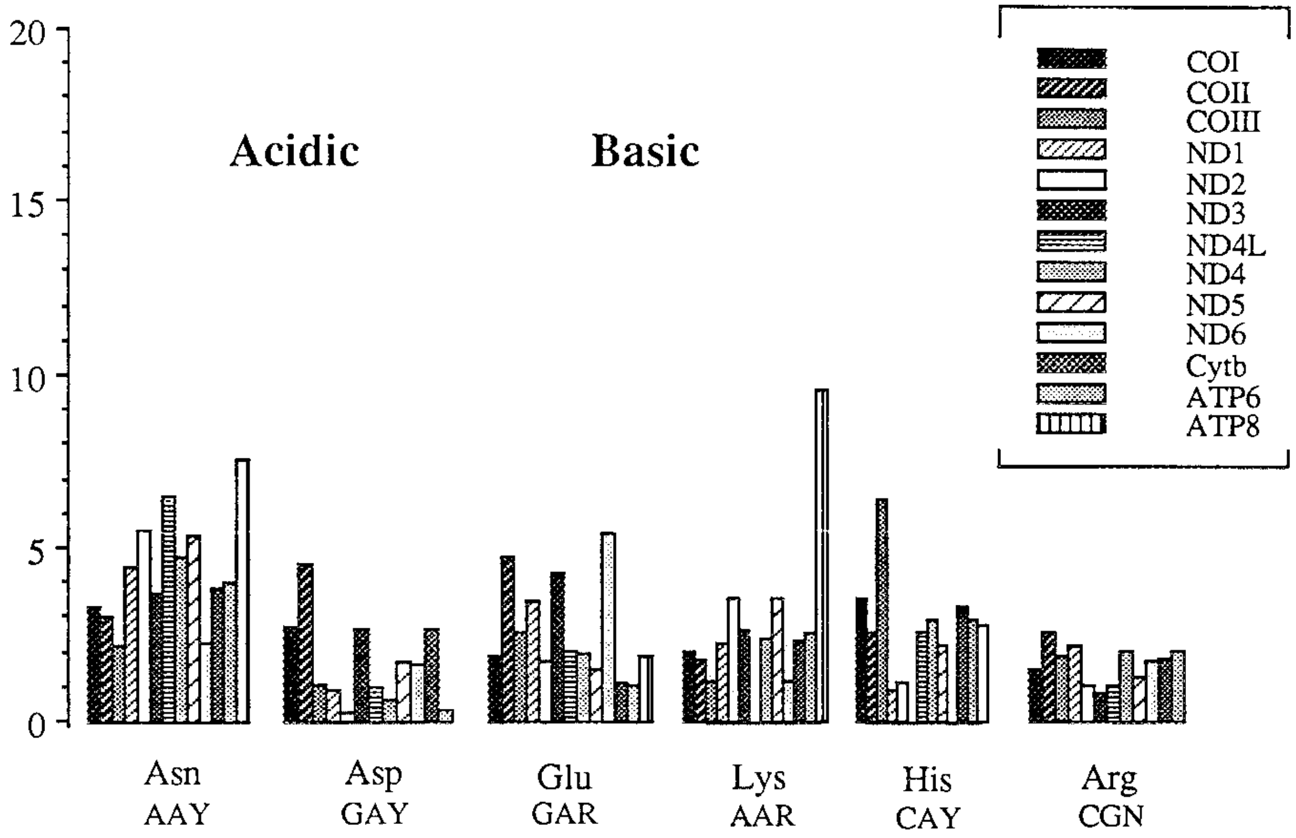
In the mammalian mtDNA codons, one would observe a bias favoring codons ending by A or C. As discussed in the part of base compositions (Chapter three), the codon usage at the third codon positions reflects the directional force that makes up the base compositional biases in each mtDNA strands. Table 3.2 summarizes the codon usage of each gene in 4 hominoid species. Codon usage differs from species to species, but most changes occur between synonymous sites of codons, so that the amino acid usage (sum of the number of codons for each amino acid) are similar among species. The codon read by a tRNA through Watson-Crick pairing is most often used in two-codon groups (underlined in Table 3.2), whereas C residues that do not perfectly match the Watson-Crick pairing are preferred (over A residues) at the third codon positions in four-codon groups. This is the case for Pro and Ala in most genes. The exception in two codon groups is Met, where ATA is preferred over ATG in all genes but *ND6*. This may be because ATG is preferentially used as an initiation codon. *ND6*, which is the only gene that use the H-strand as the sense strand for *ND6* mRNA, prefers T or G over A or C. This can be explained by the directional mutation pressure on the opposite strands. There is, however, an interesting exception. The Leu TTA is preferred over TTG or CTG in *ND6*. Moreover, *ND6* is the only gene that frequently use the TTA, but rarely use CTA or CTG (Figure 4.2). There seems be some unknown reason for the preference of Leu TTA in *ND6* .

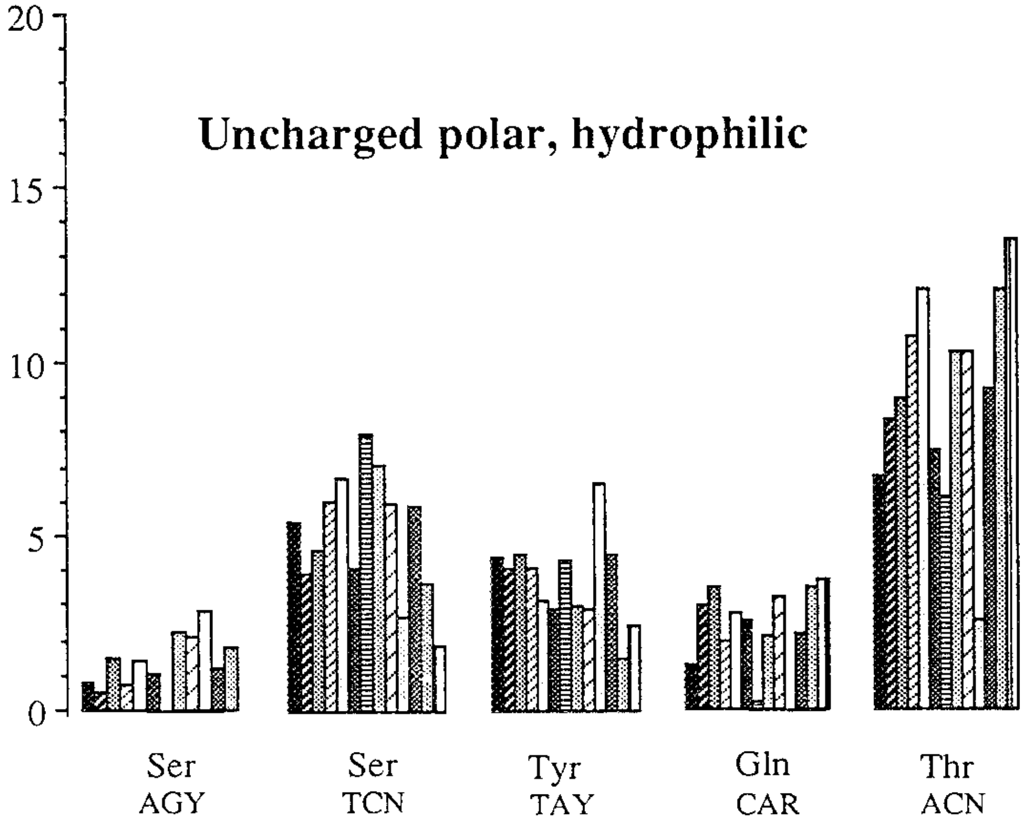
Figure 4.2 Amino acid compositions of 13 protein genes in hominoid mtDNA.

The amino acid composition of the 13 protein genes and the average percentage of the 4 hominoid species (human, common chimpanzee, gorilla and orangutan) were deduced from the codon usage table (Table 3.2), and compared for each amino acid. The codons for leucine (Leu) and serine (Ser) recognized by two different tRNAs, were calculated separately. Amino acids are shown in the three letter code together with their genetic codes. Y stands for T or C, R stands for A or G, and N can be any of the four bases,T, C, A or G.

hydrophobic nonpolar




20
15
10
5
0
Acidic
Basic
COI
COII
COIII
ND1
ND2
ND3
ND4L
ND4
ND5
ND6
Cytb
ATP6
ATP8
Asn
AAY
Asp
GAY
Glu
GAR
Lys
AAR
His
CAY
Arg
CGN


20
15
10
5
0
Uncharged polar, hydrophilic
Ser
AGY
Ser
TCN
Tyr
TAY
Gln
CAR
Thr
ACN

2. Amino acid usages

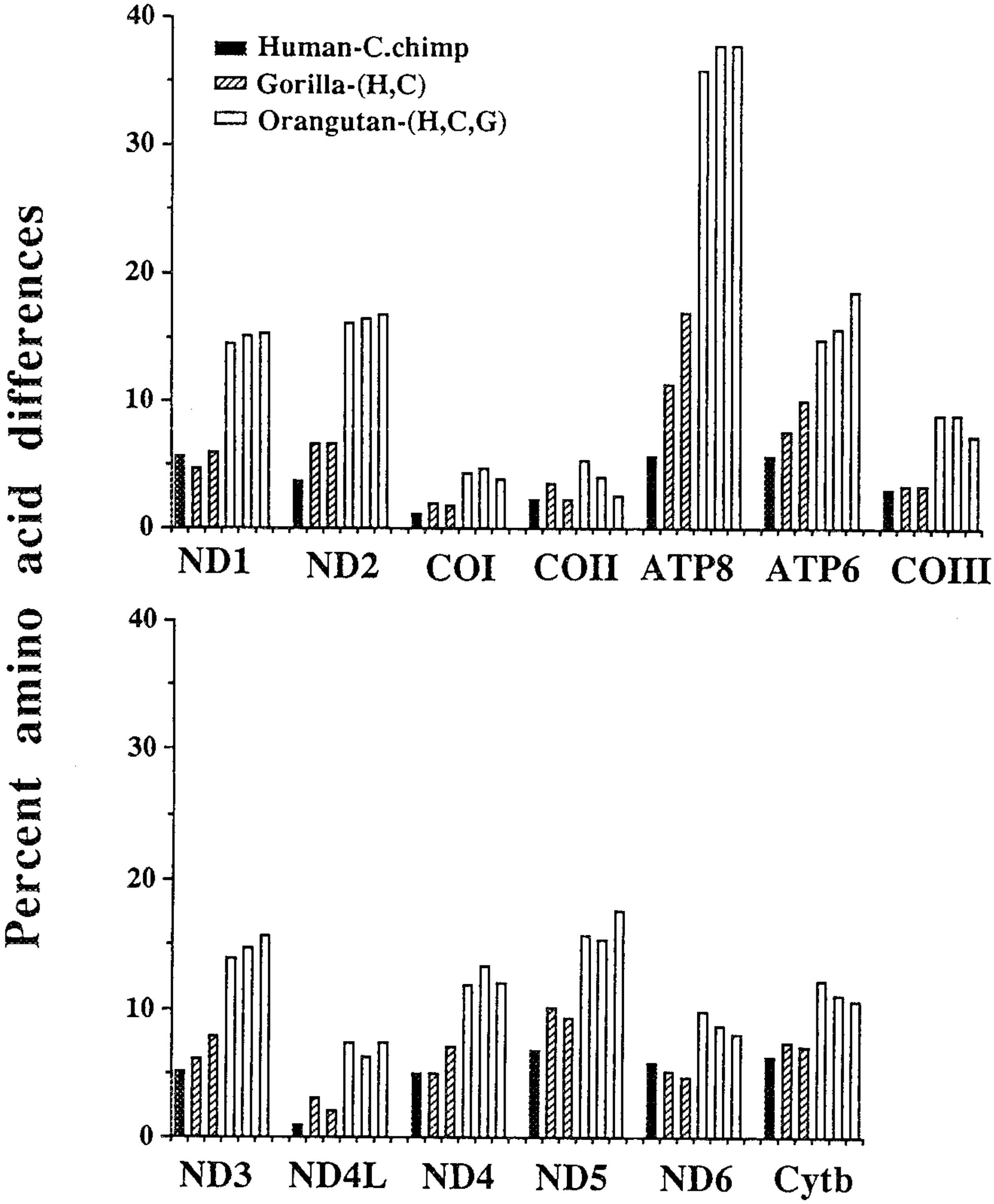
All mitochondrial proteins belong to subunits of the respiratory complexes located in the inner membrane of mitochondria (Table 4.1). Amino acid composition for each mitochondrial protein was deduced from the codon usage table (Table 3.2), and is shown in Figure 4.2. It seems that the mitochondrial proteins generally have similar amino acid compositions: Cys is commonly the least used amino acid. In most genes, both Leu and Thr are frequently used among the hydrophobic nonpolar and hydrophilic amino acids. However, some protein genes use some amino acids more frequent than others: For example, COI and COII frequently use amino acids such as Gly, Val and Asp whose codons begin with G. ATPase 8 dominantly use Pro, Thr, Asn and Lys, whose codons are rich in A or C. On the other hand, ND6 shows frequent occurrence of Val, Gly, and Glu, whose codons are rich in G or T, and rarely uses Leu(CUN). In this sense, the amino acid usage of ND6 is unique compared to other proteins belonging to the respiratory complex I. Such observations suggest that although the amino acid usage is greatly reflected by the nature and function of the respective protein (which may be the possible explanations for COI, COII and ATPase 8), some directional mutation pressure may also influence the amino acid usages as in the case of ND6.

3. Similarity in amino acid sequences

Amino acid sequence differences between different genes or parts of genes most likely reflect different degrees of selective constraints against them (Kimura 1983). To obtain a broad view of the different constraints for each gene, a comparison is made between amino acid sequences for each pair of 4 hominoid species (human, common chimpanzee, gorilla and orangutan). The degree of amino acid sequence differences clearly shows that COI and COII are the most conserved proteins, while ATPase 8 is the least conserved (Figure 4.3). Among the ND subunits in respiratory complex I, ND4L is the most conserved and ND5 is the least. The degree of conservation in Cyt *b* comes between the two extreme ND subunits.

Figure 4.3 Pairwise amino acid differences among hominoid mtDNA genes.

For each gene percent amino acid differences from pairwise comparisons among the 4 hominoid species are shown. From left, the pairs compared are human (H) - common chimpanzee (C), H - gorilla (G), C - G, H - orangutan (O), C - O, and G - O.


Percent amino acid differences
40
30
20
10
0
Human-C.chimp
Gorilla-(H,C)
Orangutan-(H,C,G)
ND1
ND2
COI
COII
ATP8
ATP6
COIII
40
30
20
10
0
ND3
ND4L
ND4
ND5
ND6
Cytb

Noteworthy are the extensive changes accumulated in the orangutan lineage, as observed in ND1, ND2, and ATPase 8. When these amino acid differences that are unique to the orangutan were examined, the frequent changes were involved with Tyr and Leu in ND1, and with Met in ND2 and ATPase 8.

Subunits COI and COII contain hemes and coppers which constitute all the four redox centers of the cytochrome oxidase (complex IV; see Table 4.1). These centers are responsible for the catalytic function of the enzyme (reviewed in Capaldi 1990) so that they may be most functionally important among the mitochondrial proteins. ATP synthase (complex V) has the catalytic part (F_1) and an integral membrane component with proton channel (F_0). ATPase 6 and ATPase 8 participate in the F_0 part of the complex V, but their functional roles are poorly understood. NADH ubiquinone dehydrogenase (complex I) catalyzes both the reduction of Q analogs and proton translocation, and subunits ND1 through ND6 belong to the hydrophobic protein fraction of the enzyme (Hatafi 1985). Involvement of ND1 in ubiquinone binding has been suggested. However, none of the ND subunits encoded by mtDNA contain iron sulfer centers, and their functions are obscure. Ubiquinol cytochrome c oxidoreductase (complex III) catalyzes electron transfer from dihydroubiquinone to cytochrome *c*. Cyt *b* contains histidine pairs that are conserved from yeast to human (Widger et al. 1984). These conserved histidine pairs are considered as the ligand pairs for the hems b562 and b566.

4. Location of the amino acid substitutions in a structure model.

The conservation of the cytochrome oxidase components may well reflect their central role in the catalytic mechanism of the enzyme (Saraste 1990; Capaldi 1990). For other protein subunits, restricted regions of their polypeptide chains may be concerned with catalysis, and amino acids in such regions thereby may be constrained from changing. To look at the different degrees of selective constraints against different parts of genes, I examined the disposition of amino acid replacement sites in the proposed folding models (Figure 4.4). The number of transmembrane helices and the topological orientation relative to the two sides of the innermembrane are based on previously reported studies (Irwin et al. 1991; Poyton et al. 1992; Fearnley and Walker 1992). The boundaries of each helix in the folding model were deduced both from the hydrophobicity (SOAP profile: Kyte and Doolittle 1982) and the prediction of secondary structure (Chou and Fasman 1978).

For COI, it has been proposed that heme a is associated with helices ii and x, and the binuclear reaction center (heme a3-Cu8) is associated with helices vi, vii, viii and x (Babcock and Wikstrom 1992). The amino acid sequences for these regions are highly conserved. For COII, it has been proposed that CuA is located in a cytochrome-c-binding domain, which is within the hydrophilic region at the C-terminal end of the polypeptide and on the cytoplasmic (C) side of the membrane (Wikstrom et al. 1984).

ATPase 8 is the most changeable subunit among hominoids. Amino acid replacements are observed throughout the transmembrane region and the matrix (M) side of the membrane. It is therefore possible that required amino acids at particular sites of these regions can be flexible in their biochemical properties and thus the degree of selective constraints is rather low.

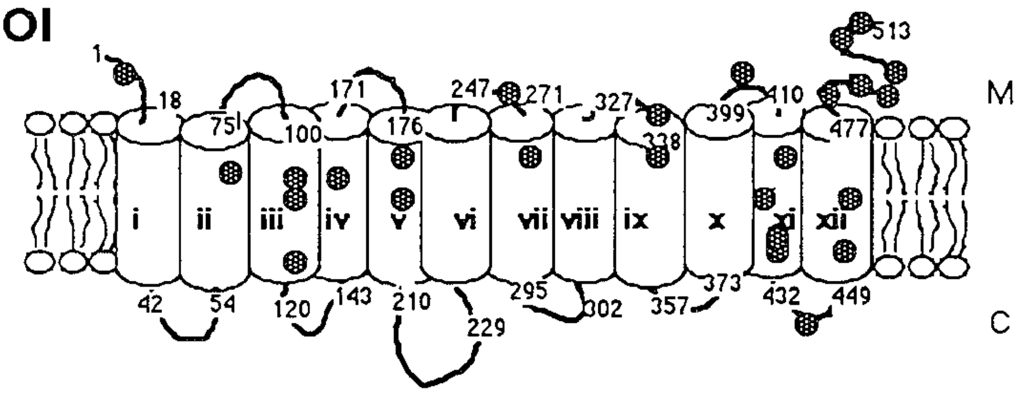
For Cyt *b*, it is suggested that hemes b-562 and b-566 are coordinated to histidine residues in transmembrane helices ii and iv (Widger et al. 1984). The Q_1 redox center involves a short portion of the first transmembrane segment. These regions are relatively conserved in the Cyt *b* polypeptide. Most of the polypeptides on the cytoplasmic side of the membrane are implicated in the Q_0 redox center (Howell 1989). This appears to explain the reduced changes in the cytoplasmic side of Cyt *b* polypeptide.

For ND subunits, the topological orientation relative to the two sides of the innermembrane is not known, however, some conserved regions has been pointed out by extensive amino acid sequence comparisons of both vertebrates and invertebrates (Fearnley and Walker 1992). In ND1, the conserved segments lie in the polar segments

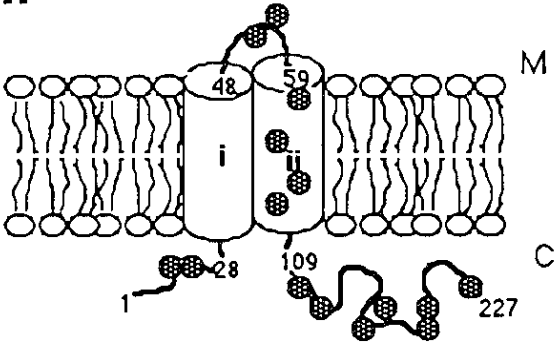
Figure 4.4 Folding models for the disposition of respiratory protein subunits (1) — subunits from complex III, IV and V.

The location of amino acid replacement site observed in any of the four hominoid species (human, common chimpanzee, gorilla and orangutan) is indicated by a dotted circle. These models are based on the algorithms of Eiselberg et al. (1984), Rao and Argos (1986), and Kyte and Doolittle (1982), and the folding models of *S. cerevisiae* mtDNA (Poyton et al. 1992). In each case, a residue 1 designates the N terminus of the primary translation product. The matrix and the cytoplasmic faces of the membrane are designated by M and C, respectively. Predicted transmembrane helices are shown as barrels, and the amino acid residues that are at the boundaries of these helices are indicated. The deduced orientation relative to the two sides of the membrane is based on the following studies: The heme orientation and spacing in COI, the orientation of Cu_A in COII, and the cross linking studies with cytochrome *c* in COIII by Wikstrom et al. (1984), dispositions of ATPase subunits by Nagley (1988) and Cox et al. (1986), and the structure of Cyt *b* by Irwin et al. (1991).

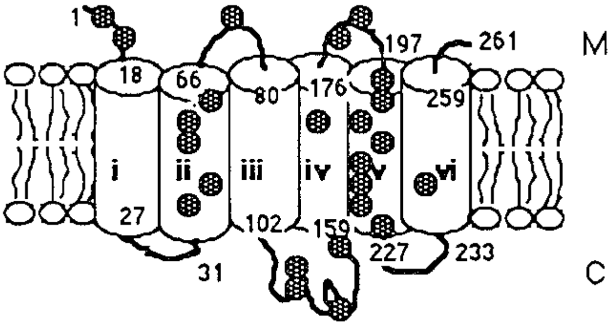
COI



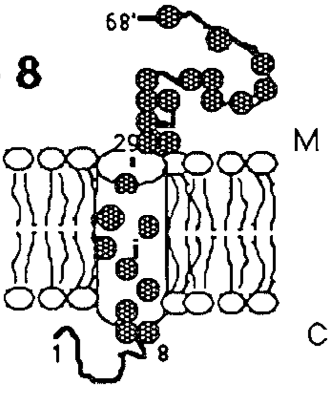
COII



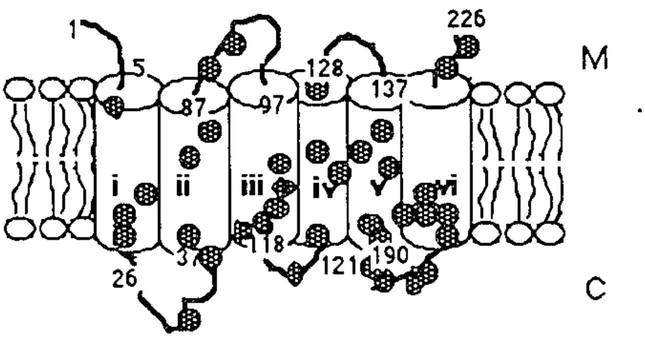
COIII



ATPase 8



ATPase 6



Cyt b

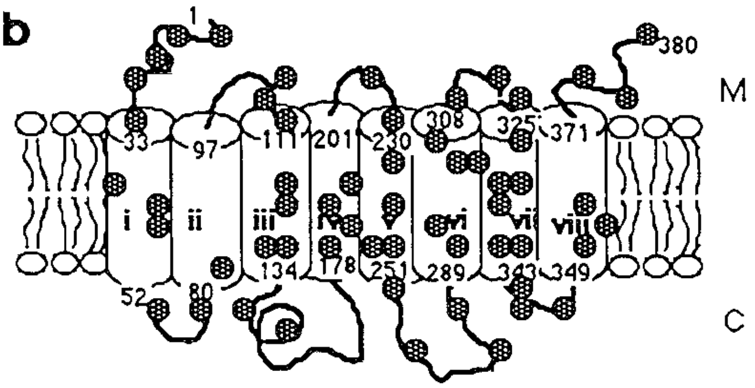
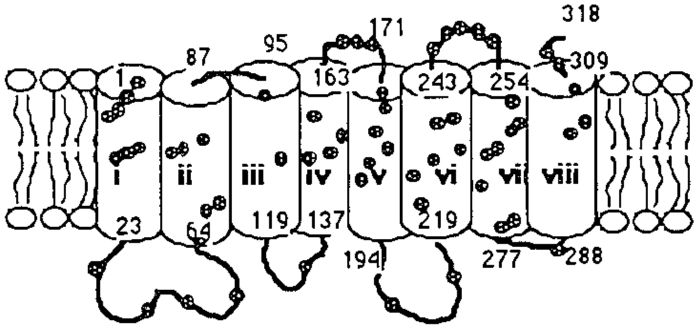


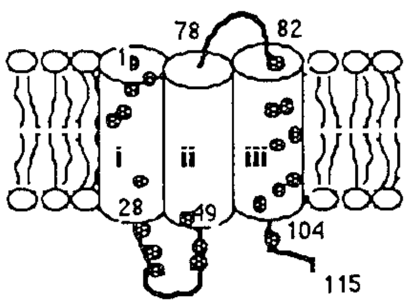
Figure 4.5 Folding models for the disposition of respiratory protein subunits (2) — subunits from complex I.

The location of amino acid replacement site observed in any of the four hominoid species (human, common chimpanzee, gorilla and orangutan) is indicated by a dotted circle. These models are based on the algorithms of Chou and Fasman (1978), and Kyte and Doolittle (1982). The location and the number of helices refer to the comparative studies by Fearnley and Walker (1992). In each case, a residue 1 designates the N terminus of the primary translation product. Predicted transmembrane helices are shown as barrels, and the amino acid residues that are at the boundaries of these helices are indicated. The orientation relative to the two sides of the membrane is not known.

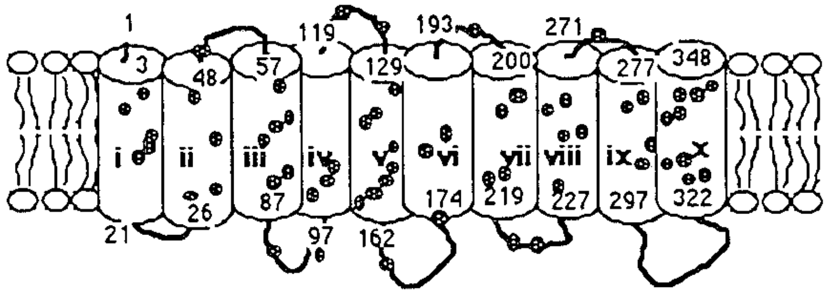
ND1



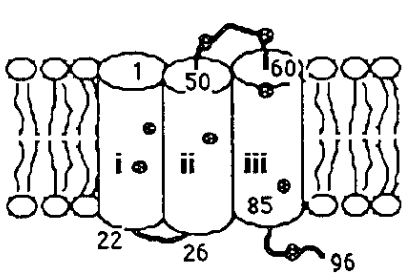
ND3



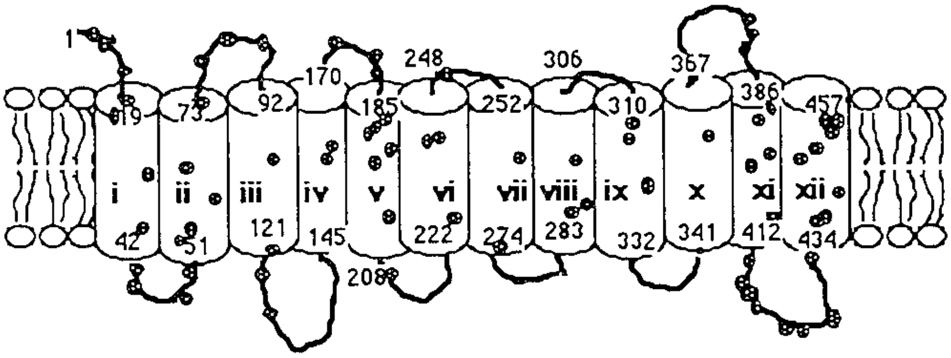
ND2



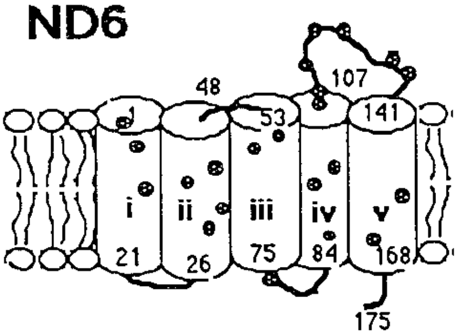
ND4L



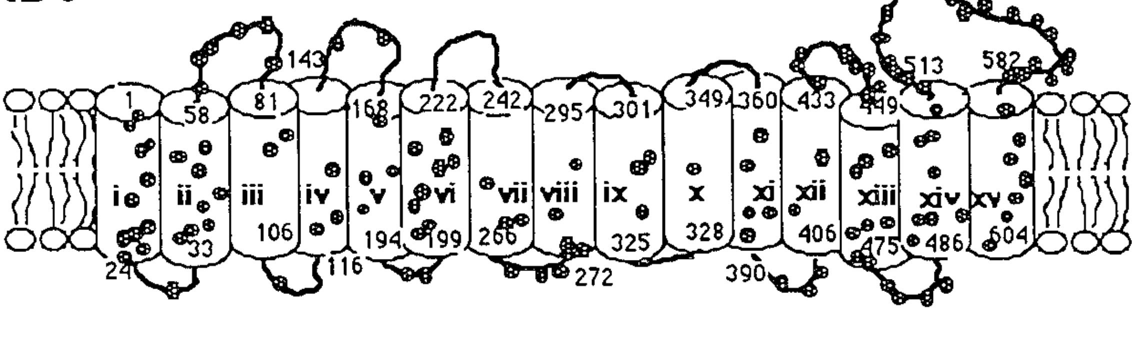
ND4



ND6



ND5



of helices i, ii, v and vi, and the most conserved hydrophilic stretches are on the same side (the lower side of Figure 4.5) of the membrane. In ND2, although it is not clear from the figure, invariant amino acids are confined to a conserved region between helices viii and ix. In ND3, helices ii and the loop between the helices ii and iii are well conserved. However, the residues that are widely conserved among animals are located around the C-terminus and the lower side of the membrane. For example, a conserved Cys residue flanked by invariant Gly residues is located in the extensive loop between the helices i and ii. In ND4, helices v-viii are known to be conserved in animals. Among the hominoids, helices ix-x are also conserved. In ND5, the most conserved regions are located between helices vi and vii, and viii and ix, respectively. A considerable difference between vertebrates and invertebrates is known for ND6, and extensive length variation between the helices iv and v has been observed in other animals. This region is also variable among the hominoids. The helix v and the N-terminus is known to be similar among the vertebrates.

The distribution of the variable sites in the hominoid protein subunits indicate the different degrees of conservation in regional units. Such would be helices constructed in the membrane spanning domain by hydrophobic sequences, and by hydrophilic sequences lying outside the lipid bilayer. The assembly and cooperative function of the protein subunits in each respiratory complex should be understood by a future extensive analysis of both mitochondrial and nuclear coded protein subunits from various organisms.

tRNA genes

There are 22 tRNA genes encoded in the animal mtDNA, sufficient to read all but the termination codons in the mitochondrial genetic code (Table 3.2). Fourteen tRNA genes use H-strand as templates (tRNA(L)), whereas the remaining eight use the L-strand as templates (tRNA(H)). Besides having a structural role, the tRNA genes act as recognition sites for processing enzymes, which cleave the primary transcript at the junction between the protein and tRNA genes (Ojala et al. 1981). Mammalian mitochondrial tRNAs are considered as a separate class of tRNA molecules for the following reasons: They are smaller than the prokaryotic or cytoplasmic counterparts. They lack a number of usually invariant bases; the TψCRA sequence in the TψC loop, and G-G invariant bases in the DHU loop. The most dramatic case of mammalian mitochondrial tRNA is the tRNA$^{Ser(AGY)}$, which completely lacks the entire DHU arm and could not form a clover-leaf structure. The only conserved universal feature is the base U immediately preceding the anticodon and pyrimidine following the anticodon.

1. Distribution of variable sites in tRNAs

The location of variable sites and base mispairs in the secondary structures of the 22 tRNAs show that the number of substitutions differs substantially among various tRNA domains such as stems (helical regions) and loops (Figure 4.6, 4.7). Although the nucleotide substitutions are found in all of the domains, stems are more conserved than loops and gaps. The most conserved domain is the anticodon loop which contains an anticodon of each tRNA. On the other hand, the most variable domain is the TψC loop and the second most is the DHU loop.

As noted by many authors (e.g., Jukes 1969; Kimura 1983), the stems accumulate compensatory substitutions which restore Watson-Crick base pairings. Compensatory substitutions are found in the acceptor stems of the gorilla $tRNA^{Thr}$, orangutan $tRNA^{Thr}$, $tRNA^{Lys}$ and $tRNA^{Asp}$, and siamang $tRNA^{Gln}$, DHU stem of orangutan $tRNA^{Thr}$, the anticodon stem of human $tRNA^{Asn}$, orangutan $tRNA^{Val}$ and also in the TψC stem of human $tRNA^{Asp}$, gorilla $tRNA^{Gly}$, $tRNA^{Thr}$, and orangutan $tRNA^{Thr}$, $tRNA^{Glu}$.

2. Codon usage and tRNA variability

The proportion of variable sites in the aligned tRNA genes shows a strong heterogeneity in the nucleotide substitutions per site (Figure 4.8): The proportion ranges from 4.2% in $tRNA^{Leu(CUN)}$ and 4.4% in $tRNA^{Met}$ to 34.8% in $tRNA^{Thr}$. The low

Figure 4.6 Distribution of variable sites in the putative secondary structure of hominoid mitochondrial tRNAs (1).

Each tRNA consists of 14 structurally distinct regions: 5'-acceptor (A) stem - gap - dihydrouridine (DHU) stem - DHU loop - DHU stem - gap - anticodon (AC) stem - AC loop - AC stem - variable loop - TψC stem - TψC loop - TψC stem - A stem- 3'. Among the hominoid tRNAs, size variation occur only at DHU loop, variable loop and TψC loop. Hatched areas indicate positions that base differences are observed among the comparison of 6 hominoids (human, common chimpanzee, pygmy chimpanzee, gorilla, orangutan, and siamang). A solid line between nucleotides (boxes) in stem parts indicates a Watson-Crick pair in all the species, while a dot indicates a non Watson-Crick pair in some species.

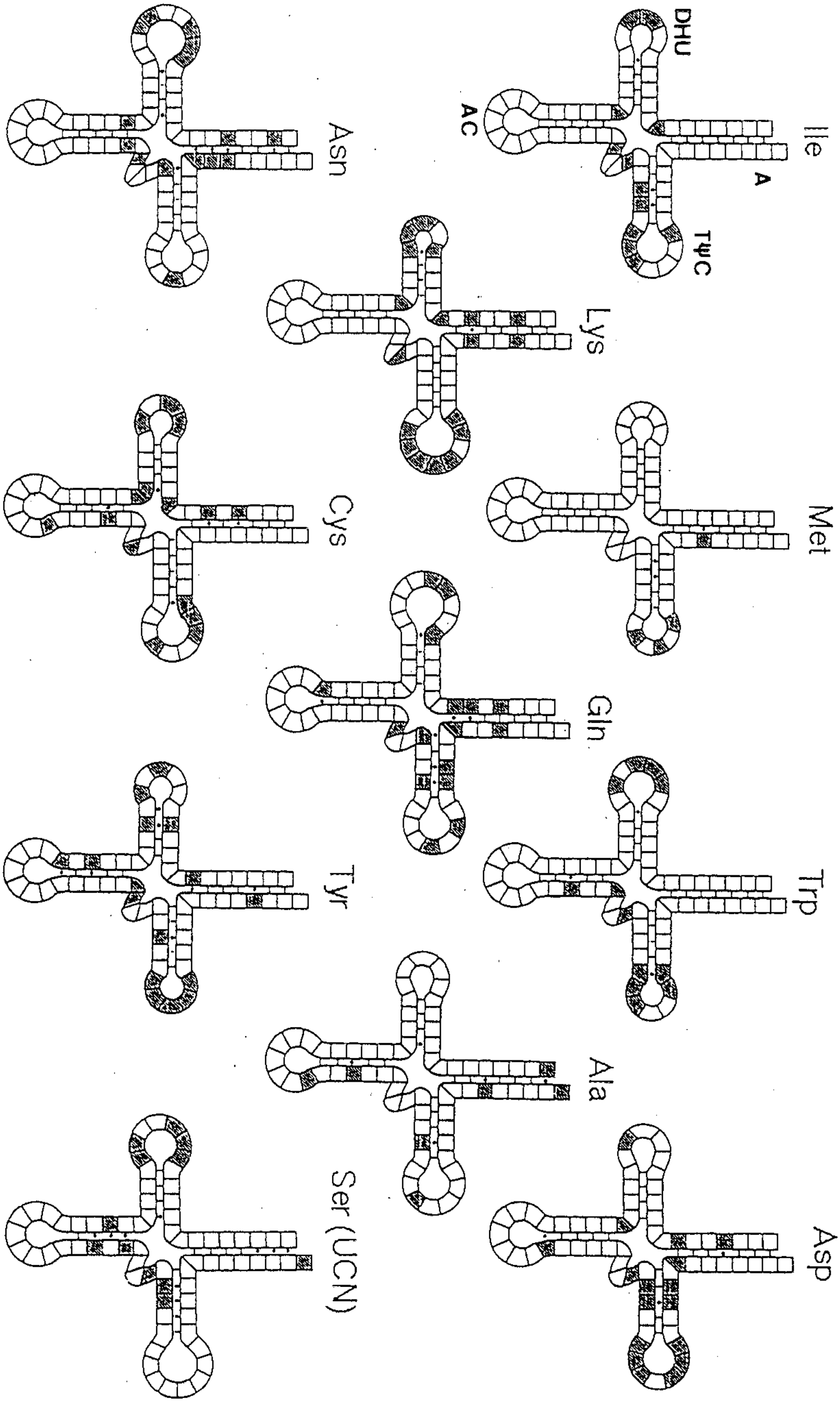

Ile
DHU
AC
A
TΨC
Lys
Met
Asn
Cys
Gln
Trp
Tyr
Ala
Asp
Ser (UCN)

Figure 4.7 Distribution of variable sites in the putative secondary structure of hominoid mitochondrial tRNAs (2)

Hatched areas indicate positions that base substitutions have taken place among the comparison of 4 hominoids (human, common chimpanzee, gorilla, and orangutan). A solid line between nucleotides (boxes) in stem parts indicates a Watson-Crick pair in all the species, while a dot indicates a non Watson-Crick pair in some species. The location of the conserved 13 bp sequence in $\mathrm{tRNA}^{\mathrm{Leu(UUR)}}$ is shown by an aligned bar.

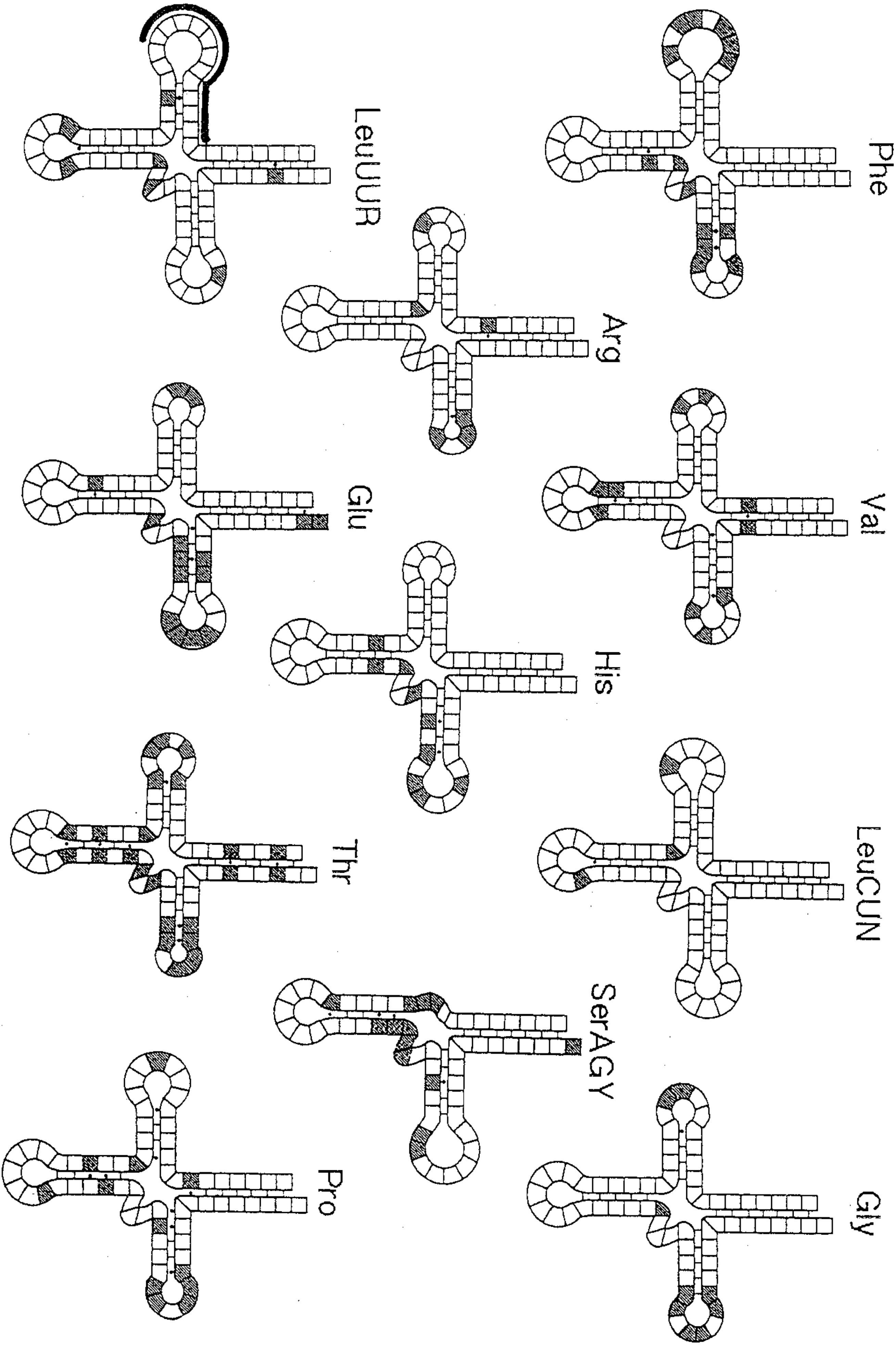
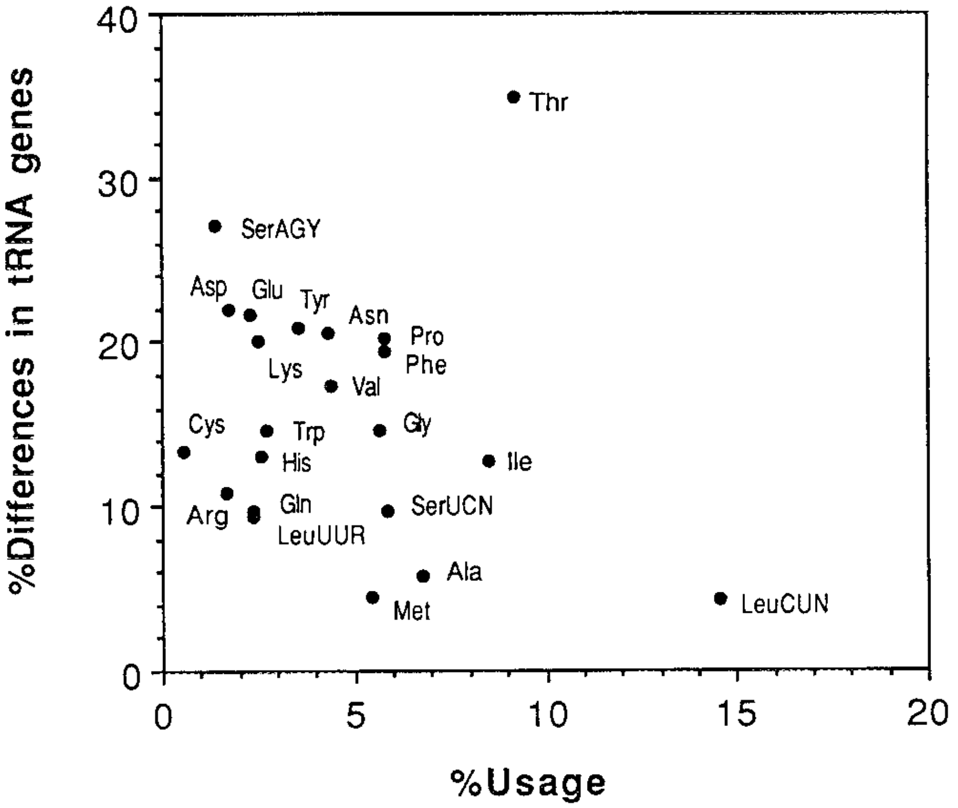

Phe
LeuUUR
Arg
Val
Glu
His
LeuCUN
Thr
SerAGY
Gly
Pro

Figure 4.8 Relationship between codon usage and observed differences in tRNA.

The percent usage of the tRNAs were deduced from the codon usages in the human mtDNA genome. The percent nucleotide changes for each tRNA were calculated from the size of the tRNA gene and the number of nucleotide differences observed in a phylogenetic tree of 4 hominoids (human, common chimpanzee, gorilla, and orangutan). The name of tRNA for each plot is designated by a three letter amino acid code.


%Differences in tRNA genes
0
10
20
30
40
0
5
10
15
20
%Usage
Thr
SerAGY
Asp
Glu
Tyr
Asn
Pro
Phe
Lys
Val
Cys
Trp
Gly
His
Ile
Arg
Gln
LeuUUR
SerUCN
Ala
Met
LeuCUN

number of substitutions in $tRNA^{Met}$ is due probably to its important role in the transcription initiation. The high number of substitutions in the $tRNA^{Thr}$ is due to the many compensatory substitutions. Since it was suggested by the sequence analysis of $tRNA^{His}$, $tRNA^{Ser(AGY)}$ and $tRNA^{Leu(CUN)}$ that the number of substitutions in tRNA varies inversely with the frequency of the corresponding codons used in the mtDNA (Brown et al. 1982), I examined this rule for all tRNA genes (Figure 4.8). The correlation is not apparent in this extensive study: The two disparate tRNA genes are $tRNA^{Leu(CUN)}$ and $tRNA^{Thr}$, both are frequently used in the mitochondrial protein genes. However, $tRNA^{Leu(CUN)}$ is one of the least variable tRNA gene, and $tRNA^{Thr}$ the most variable among the hominoids. For the two tRNA genes, the number of sites that are variable among the hominoid species differ by 20 sites. The rate of tRNA evolution appear to be partly determined by both the function of the tRNA gene itself and the codon usage patterns. For example, a 13 bp sequence in the DHU stem and DHU loop of $tRNA^{LeuUUR}$ has a regulatory role for the transcription termination (Christianson, Clayton 1988; Kruse et al. 1989). A point mutation within this conserved 13 bp sequence is associated with a genetic defect in mitochondrial encephalomyopathies (MELAS) that cause severe human neuromuscular disorders (Goto et al. 1990). Several other point mutations in tRNA genes that are associated in human genetic disorder are reported (Silvestri et al. 1992). For example, a point mutation in TψC stem of $tRNA^{Lys}$ is frequently found in patients with Myocloneus epilopsy ragged red fibers (MERRF). It has been demonstrated that the same mutation impaired protein synthesis and cellular oxidation of cultured cells (King and Attardi 1989). Curiously, however, this position is variable among the hominoids. Further comparative studies should lead to the understanding for the function and evolution of the tRNA genes.

rRNA genes

Mammalian mitochondrial ribosomes belong to a distinct class of ribosomes, because of dramatical differences in size. Secondary structure models of large and small rRNAs from prokaryotes, eukaryotes, chloroplasts and mitochondria, have been proposed both from comparative sequencing studies and structural analyses (Maly and Brimacobe 1983; Hixson and Brown 1986). The major structural features of these rRNAs are, however, similar despite the vast evolutionary distances among them. The structural similarity and differences of rRNAs may provide understanding of the functional role of the conserved sequences and evolution of the rRNA molecules. Below, I will examine the proportion of variable sites among the four hominoids for each rRNA domains. The domain structures are defined by long range interactions which are among the elements better conserved in the structure (Cantatore and Saccone 1987). The source of sequence data is as follows: human (Anderson et al. 1981); 12SrRNA of common chimpanzee and gorilla (Hixson and Brown 1986); 12SrRNA of orangutan, and 16SrRNA of common chimpanzee, gorilla and orangutan (present study).

The secondary structure model of hominoid 12S rRNA have been proposed (Hixson and Brown 1986), and is organized into three domains (Cantatore and Saccone 1987). From the 5' end, domain 1, 2 and 3 represent sites 1 to 283, 284 to 546, and 547 to 956, respectively. The proportion of variable sites observed among the four hominoids is 16.6% in domain 1, and 11.8% and 10.5% in domain 2 and 3, respectively. There is a hairpin loop structure comprised by the 3' terminal nucleotides, which is highly conserved among mammalian 12S rRNA. This structure may be important for the binding of the two ribosomal subunits through interaction with the 16S rRNA (Azad 1979).

The secondary structure of 16S rRNA is organized into six domains (Cantatore and Saccone 1987). The first domain (sites 1 to 187) has a low number of secondary structure interactions. The proportion of variable sites among the 4 hominoids is 23.2%. The area between the end of the second domain (sites 190 to 601) and the beginning of the fourth domain (sites 781 to 1023) is highly variable and also displays a low primary and secondary structure homology among animal 16S rRNAs. The proportion of variable sites in domain 3 is 22.9%. The fourth and fifth (sites 1024 to 1448) domains are very conserved, the proportion of variable sites are 6.2% and 16.9%, respectively.

Mitochondrial rRNA genes change rapidly compared to the nuclear counterpart, although they are the most conserved coding regions in the mtDNA. Comparison of the two rRNA genes indidcate that, as a whole, 12S rRNA are more conservative than 16S rRNA.

Control regions

There are two major control regions in the mtDNA. One is the D-loop region, which spans between the genes for *tRNA^{Phe}* and *tRNA^{Pro}*. This region contains most of the regulatory elements for mtDNA transcription and replication. The other is the O_L sequence, located on the H-strand between the *tRNA^{Asn}* and *tRNA^{Cys}*. In spite of these regulatory roles in the control region, substitutions, additions and deletions accumulate more rapidly in these regions than in other mtDNA regions. Extensive level of divergences are observed in the interspecies comparisons. Particular mutational events that are probably due to pausing of the polymerase enzymes at the secondary structure level, recombination, or replication slippage mechanisms may occur in the control regions.

The transcription of human mtDNA L- and H-strand, starts at promoter LSP and HSP, respectively. Nucleotides that are critical for the accurate promoter activity have been identified by site directed mutagenesis analyses (Hixson and Clayton 1985). Figure 4.9 summarizes the corresponding nucleotide sequences in the hominoids; the orangutan sequence determined in the present study, and other reported sequences (Anderson et al. 1981; Foran et al. 1988). It is noteworthy that orangutan mtDNA possess nucleotide changes that may cause reduction in the human promoter activity.

In several vertebrate species, three conserved sequence blocks (CSB's) are present upstream from the 5' termini. There are evidence from mouse and human that a switch from RNA to DNA synthesis can occur at CSB-1 (Chang and Clayton 1984, Chang et al. 1985). No function, however, is known for either CSB-2 and CSB-3. Alignment of the orangutan CSB-1 sequence with other hominoids shows that four substitutions and a deletion occur in the orangutan CSB-1 (78% in homology). Alignment of hominoid CSB-2 and CSB-3 sequences shows that part of CSB-2 and CSB-3 is deleted in gorilla, and all of CSB-2 and part of CSB-3 are deleted from orangutan mtDNA. It is also known that all of CSB-3 has been deleted in cow mtDNA. Such marked divergences of CSB-2 and CSB-3 in gorilla and orangutan argue against their involvement in D-loop DNA initiation.

A small region surrounding the O_L, differs from other intergenic sequences by its greater size and by the conservation of several features (Figure 4.10). There are two 11 bp stretch that are completely conserved among the hominoids. This region is proposed to construct the stem of O_L structure (Brown 1985). Also conserved is a tandem T repeat of six to seven nucleotides located in a 12 bp sequence (13 bp in orangutan) that corresponds to the proposed loop sequence of the O_L structure.

Figure 4.9 Alignment of the transcription promoter sequences of hominoid mtDNA.

The nucleotides of the heavy strand promoter (HSP) and light strand promoter (LSP) have been aligned with the corresponding site in the 5 hominoids. Only the nucleotide sites different from human are shown for the 4 hominoids. The start site of transcription is indicated by the arrows. Nucleotides written in italics represent positions where induced mutations cause marked reduction in promoter activity in human. Nucleotides with asterisks indicate the positions that are absolutely required for promoter function in human (Hixson and Clayton 1985). Orangutan sequence is newly determined in this study. Human sequence is taken from Anderson et al. (1981). Nucleotide sequences of the remaining hominoids are from Foran et al. (1988). Species abbreviations are: COMCH - common chimpanzee, PYGCH - pygmy chimpanzee, ORANG - orangutan.

HSP

```
                          * * *  |---->
HUMAN           C C A A A C C C C A A A G A C
COMCH
PYGCH                                       T
GORILLA
ORANG               C           A
```

LSP

```
                             *    |---->
HUMAN           A T A C C G C C A A A G A T
COMCH                           T
PYGCH
GORILLA
ORANG             C   T A     T
```

Figure 4.10 Alignment of nucleotide sequences of L-strand replication origin in 6 hominoid mtDNAs.

The template (i.e. H-strand) sequences of human mtDNA (Anderson et al. 1981) are shown in 5' to 3' orientation. For other hominoids (orangutan - present study; others - Foran et al. 1988), only the nucleotide sites different from human are shown. Note that the two stem sequences make perfect Watson-Crick pairings. The high proportions of T's occurring as a tandem repeat in the loop sequence has been conserved, and may be of functional significance. Species abbreviations are: COMCH - common chimpanzee, PYGCH - pygmy chimpanzee, ORANG - orangutan.

Proposed L-strand origin structure

tRNA^{Cys} - Stem Loop stem -- tRNA^{Asn}

```
HUMAN          CTTCTCCCGCC   -TTTTTTCCCGGC    GGCGGGCGCCG
COMCH                        -      T TT
PYGCH                        -      T TT
GORILLA                      -      T TT
ORANG                        T      A  C
SIAMANG                      -      T  C
```

Conclusions

The distribution of variable sites among the hominoid mtDNAs give some inferences to the functional constraints of each region and the mechanisms of evolution in the mtDNA. The region which is suggested to have some important functional role is always conserved among the hominoids. Therefore, further comparative analysis from variety of mtDNAs should point out more clearly the regions of the genome that are functionary invariable. Such sites may involve with the conformation or assembly of gene products, or with interactions to some proteins or regulatory elements.

Another important factor to consider is the pressure of mutation that constructs the base composition biases of the H- and L-strands. For the protein genes that use the H-strand as the template, codons or amino acids using C or A are preferred. It is interesting to note that some protein genes which use codons involving C or A more frequent than other protein genes (*ATPase 8*), show relatively large amount of amino acid replacement changes. On the other hand, some protein genes (*COI* and *COII*) that use codons involving G more frequent than other protein genes are relatively conservative in amino acid sequences.

CHAPTER FIVE

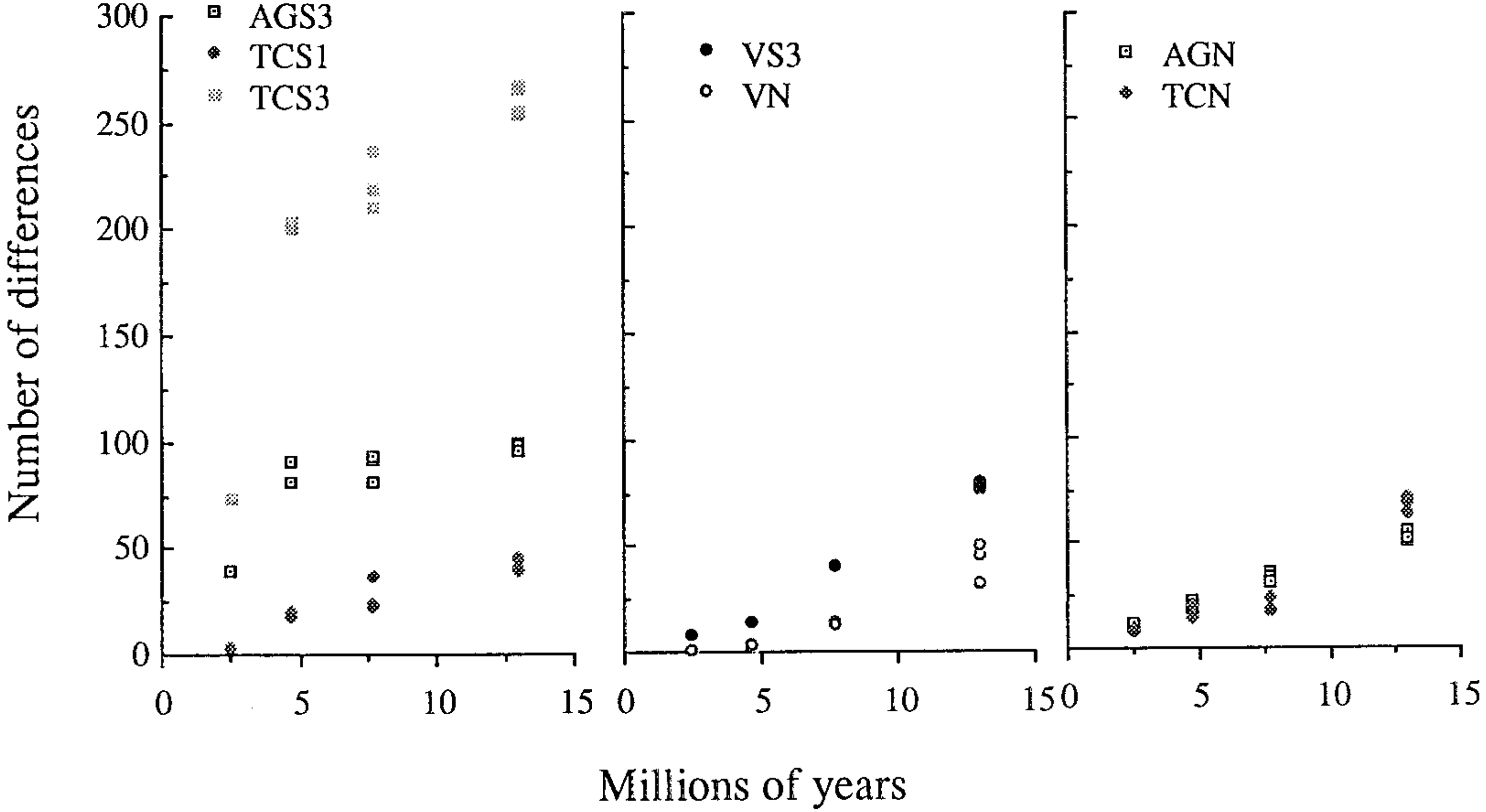
NUCLEOTIDE DIFFERENCES AND MODE OF EVOLUTION

Several features were illustrated from the comparison of portions of mtDNAs. The most intriguing observation is the excess of transitions over transversions and the decrease in the observed rate of transitions with increasing divergence time (Brown et al. 1982). Although questions still remain for the elevated transition rate, the decrease in the percentage of transitions relative to divergence time was explained by the erase of the record of transition and accumulation of transversion over a large period of time, due to multiple substitution at the same site (Brown et al. 1982; De Salle et al. 1987). A wider and deeper understanding of the process of nucleotide substitutions in the mitochondrial genome is necessary to infer a reliable rate of nucleotide substitutions and the phylogenetic relationship of hominoids.

In this chapter I will discuss the characteristics of the nucleotide substitutions in hominoid mtDNAs. Excluding small overlapping parts, and noncoding parts in which insertion/deletion is frequent, I classified the nucleotide substitutions into transitions (AG/TC) and transversions (V). Transitions were divided into two types, substitutions between A and G (AG), and substitutions between T and C (TC). Such classifications were made in order to examine the actual mode of nucleotide substitutions very carefully. As mentioned before, it is known that transitions predominate transversions. It is also reasonable to discriminate between AG and TC substitutions because their mode of substitutions are likely to be different due the extreme biases in the base compositions. Consequently, nucleotide substitutions in tRNA and rRNA regions were classified into three different categories; AG, TC and V. In the protein coding region, I further treated three codon positions separately and classified them into synonymous (S) and nonsynonymous (N). At first positions, there are AG nonsynonymous (AGN), TC

Figure 5.1 Accumulation of nucleotide differences in the protein coding region.

Nucleotide differences in protein coding region (4008 sites) included in the homologous 4.9 kb region (hatched bar in Figure 2.1) are examined. Observed number of nucleotide differences for a pair of hominoid species is plotted against estimated divergence times (Horai et al. 1992). The differences are examined in the categories of TC transitions, AG transitions, and transversions (V) that are synonymous (S) and nonsynonymous (N). Most synonymous substitutions occur at the third codon positions except for the first codon positions of leucine codons (TCS_1). Differences of synonymous TC transition (TCS) are divided further into TCS_1 and the TCS at third codon positions (TCS_3). The pairs compared from the left are: common chimpanzee (CC) - pygmy chimpanzee (PC) at 2.5 million years ago (mya); human (H) - CC and H - PC at 4.7 mya; gorilla (G) - CC, G - PC and G - H at 7.7 mya; and orangutan (O) - CC, O - PC, O - H and O - G at 13 mya.


Number of differences
0
50
100
150
200
250
300
AGS3
TCS1
TCS3
VS3
VN
AGN
TCN
0
5
10
15
Millions of years

nonsynonymous (TCN), TC synonymous (TCS), and transversional nonsynonymous (VN) substitutions. Similarly, at the second positions, there are AGN, TCN, and VN substitutions, and at the third positions, there are AGS, TCS, VS and VN substitutions. For convenience, these symbols are used together with a subscript when coding positions are specified; for example, TCS$_1$ means TC synonymous substitutions at the first codon positions. To clearly distinguish between the *observed* differences and the *inferred* (or actual) number of substitutions, I will use the word *differences* for the *observed* changes, and the word *substitutions* for the *inferred* changes.

Nucleotide differences in the protein genes

The mode of accumulation of the observed nucleotide differences in the protein coding region is shown in Figure 5.1. The high substitution rate of mtDNA raises several cautions. Synonymous transitions such as AGS$_3$, TCS$_1$ and TCS$_3$ level off rapidly and are in some cases saturated even between human and chimpanzees. This is consistent with well-known high transition rates in mammalian mtDNA, the ratio of AG/TC changes to V changes being about 10 (Brown et al. 1982). However, the kinetic behaviors of various types of synonymous changes (AGS$_3$, TCS$_1$, TCS$_3$ and VS$_3$) differ from one another. Such differences may be accounted for by their different saturation levels. Usually, the content of G residues at the third codon positions is extremely low (Table 3.1), so the saturation level of AGS$_3$ must be low and attained rapidly. The slower levelling-off, as observed in TCS$_3$ and particularly in VS$_3$, is due to a relative abundance of A, C, and T residues. This situation is similar for TCS$_1$ in which *Leu* codons (UUR and CUR) are involved.

The mode of accumulation of the observed nucleotide differences for each gene is shown in Figure 5.2. Clearly, synonymous transitions (AGS and TCS) attain the saturation level for all genes. Different genes show different behavior in the level and time of saturation. *ND3*, *Cyt b* and *ATPase 8* are the rapidly saturating genes, whereas slower levelling-off is seen in *ND4L*, *COI*, and *COII*. The influence of the base compositions in the saturation levels is clear from the accumulation of observed nucleotide differences in *ND6*. Being encoded on the opposite strand, the third codon positions in *ND6* is high in G and T contents (Table 3.1). Thus the synonymous transitions (AGS and TCS) in *ND6* show a clear contrast to the other genes in the level of saturation. In this case, the content of C residues at the third codon positions is extremely low (3.6% in average; Table 3.1). Consequently, the saturation level of TCS is lower and attained rapidly than that of AGS3.

None of the protein genes show saturation in VS3. Since transversions are infrequent, the number of VS3 differences are small irrespective of being synonymous. For example, the observed percent VS3 differences in the pairwise comparison involving orangutan are only 2%. The stochastic fluctuation must therefore be large in some genes, however, a relatively linear and similar level of accumulation is observed.

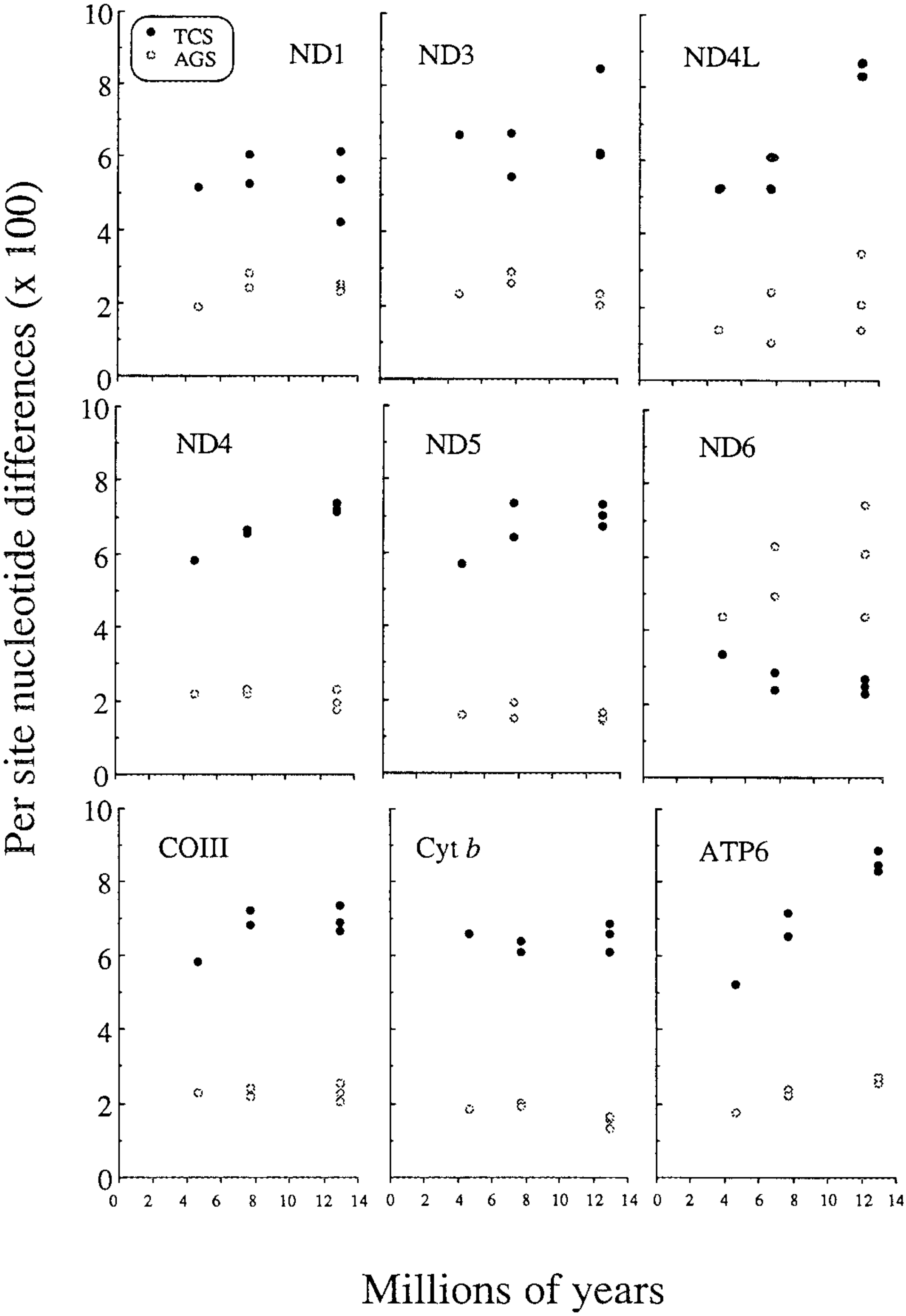
Nonsynonymous differences, against which selective constraints would be stronger than against synonymous ones, also show no evidence of levelling-off among hominoids. However, opposed to VS3 differences, the observed nucleotide differences differ considerably among genes. Noteworthy are elevated changes observed in *ATPase 8* for the comparisons involving orangutan.

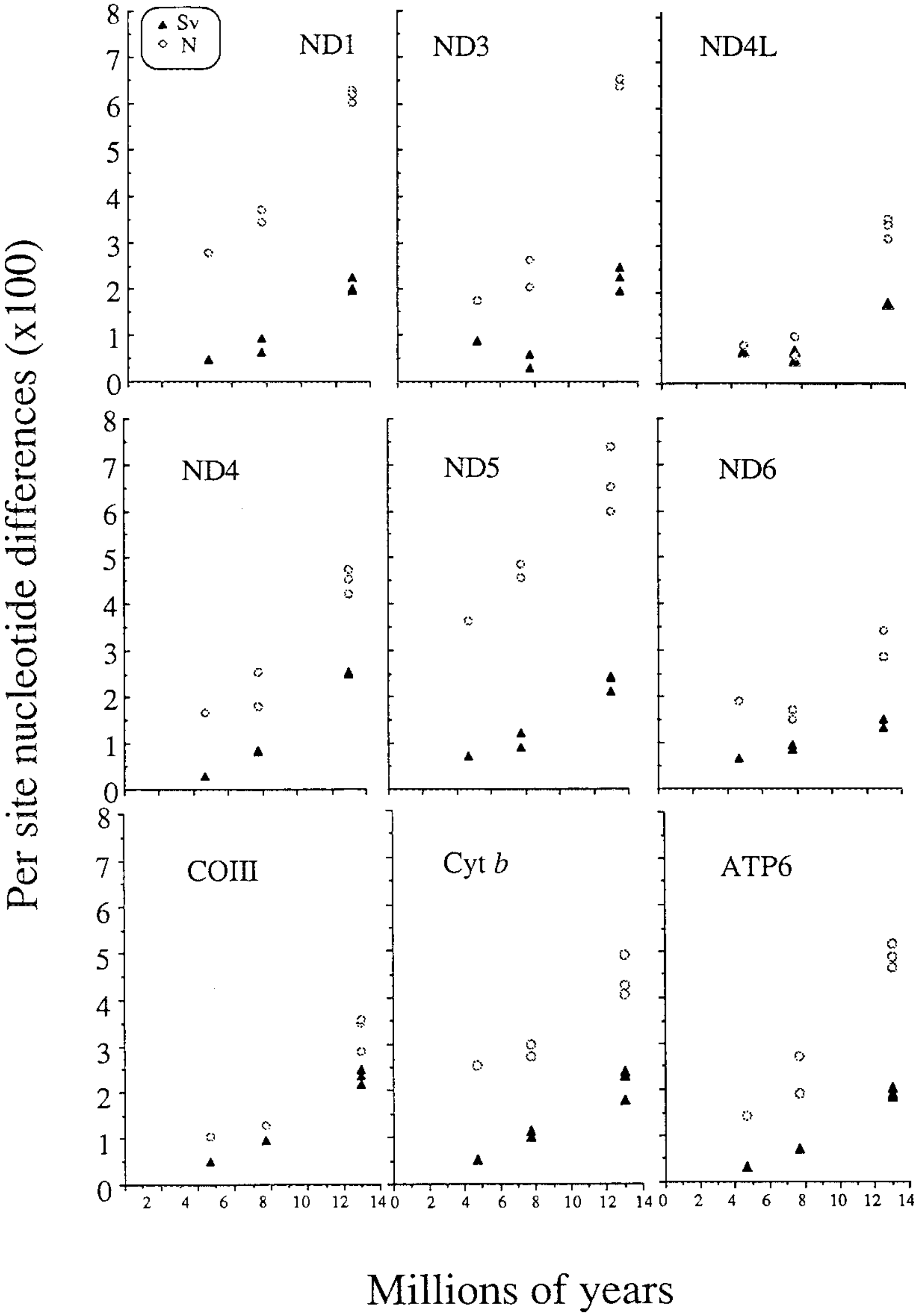
Nucleotide differences in the tRNA and rRNA genes

The differences in the tRNA and rRNA regions are examined in the categories of TC, AG transitions and transversions. tRNA genes are grouped into tRNA(L) and tRNA(H) as noted before. In both groups, the observed sequence differences do not show any evidence of saturations up to the divergence of orangutan (Figure 5.3). The differences in the 11 tRNA genes included in the 4.9 kb region (Bar A in Figure 2.1) did not show any saturation effect up to the comparisons between siamang (Table 1 in Horai et al. 1992). Similarly in both 12S rRNA and 16S rRNA coding regions, the observed sequence differences do not show any evidence for saturations up to the divergence of orangutan (Fig 5.3).

Figure 5.2 Accumulation of nucleotide differences in each of 13 protein genes.

Observed number of nucleotide differences for a pair of hominoid species is plotted against estimated divergence times (Horai et al. 1992). For each gene, the nucleotide differences are examined in the categories of synonymous TC transitions, AG transitions, and transversions (V), and nonsynonymous differences (N). They were divided by the size of each gene and converted into the number of differences per site. The pairs compared from the left are: common chimpanzee (CC) - pygmy chimpanzee (PC) at 2.5 mya; human (H) - CC and H - PC at 4.7 mya; gorilla (G) - CC, G - PC and G - H at 7.7 mya; and orangutan (O) - CC, O - PC, O - H and O - G at 13 mya.


● TCS
○ AGS
ND1
ND3
ND4L
ND4
ND5
ND6
COIII
Cyt *b*
ATP6
Per site nucleotide differences (x 100)
0
2
4
6
8
10
0
2
4
6
8
10
12
14
Millions of years



Sv
N
ND1
ND3
ND4L
ND4
ND5
ND6
COIII
Cyt *b*
ATP6
Per site nucleotide differences (x100)
0
1
2
3
4
5
6
7
8
0
2
4
6
8
10
12
14
Millions of years

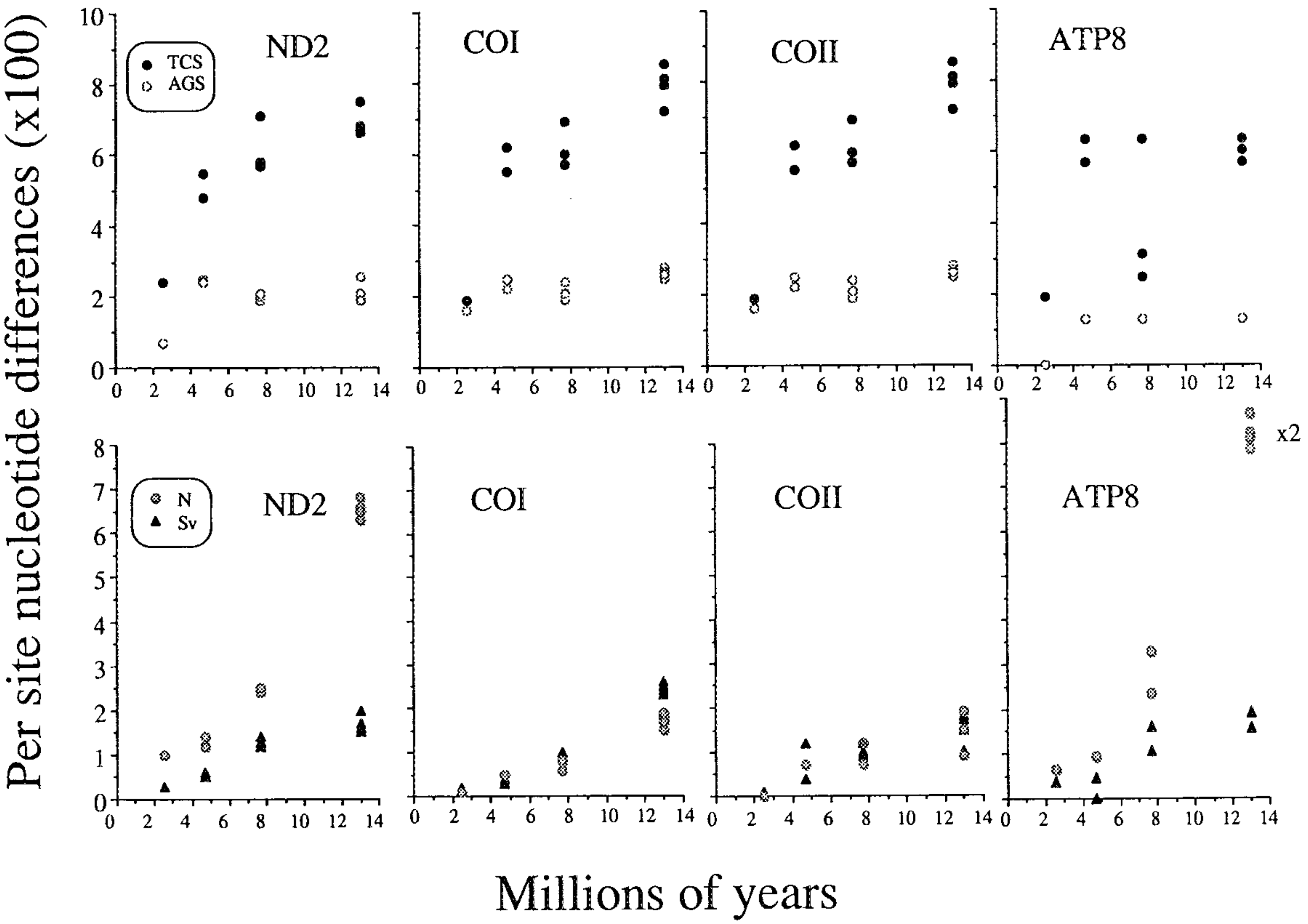

ND2
COI
COII
ATP8
● TCS
○ AGS
ND2
COI
COII
ATP8
▨ N
▲ Sv
x2
Per site nucleotide differences (x100)
Millions of years
0
2
4
6
8
10
0
1
2
3
4
5
6
7
8
0
2
4
6
8
10
12
14

Figure 5.3 Accumulation of nucleotide differences in tRNA, 12S rRNA and 16S rRNA genes.

Observed number of nucleotide differences for a pair of species is plotted against estimated divergence times (Horai et al. 1992). For each gene, the nucleotide differences are examined in the categories of TC transitions, AG transitions, and transversions (V). The pairs compared from the left are: human (H) - CC at 4.7 mya; gorilla (G) - CC and G - H at 7.7 mya; and orangutan (O) - CC, O - H and O - G at 13 mya.

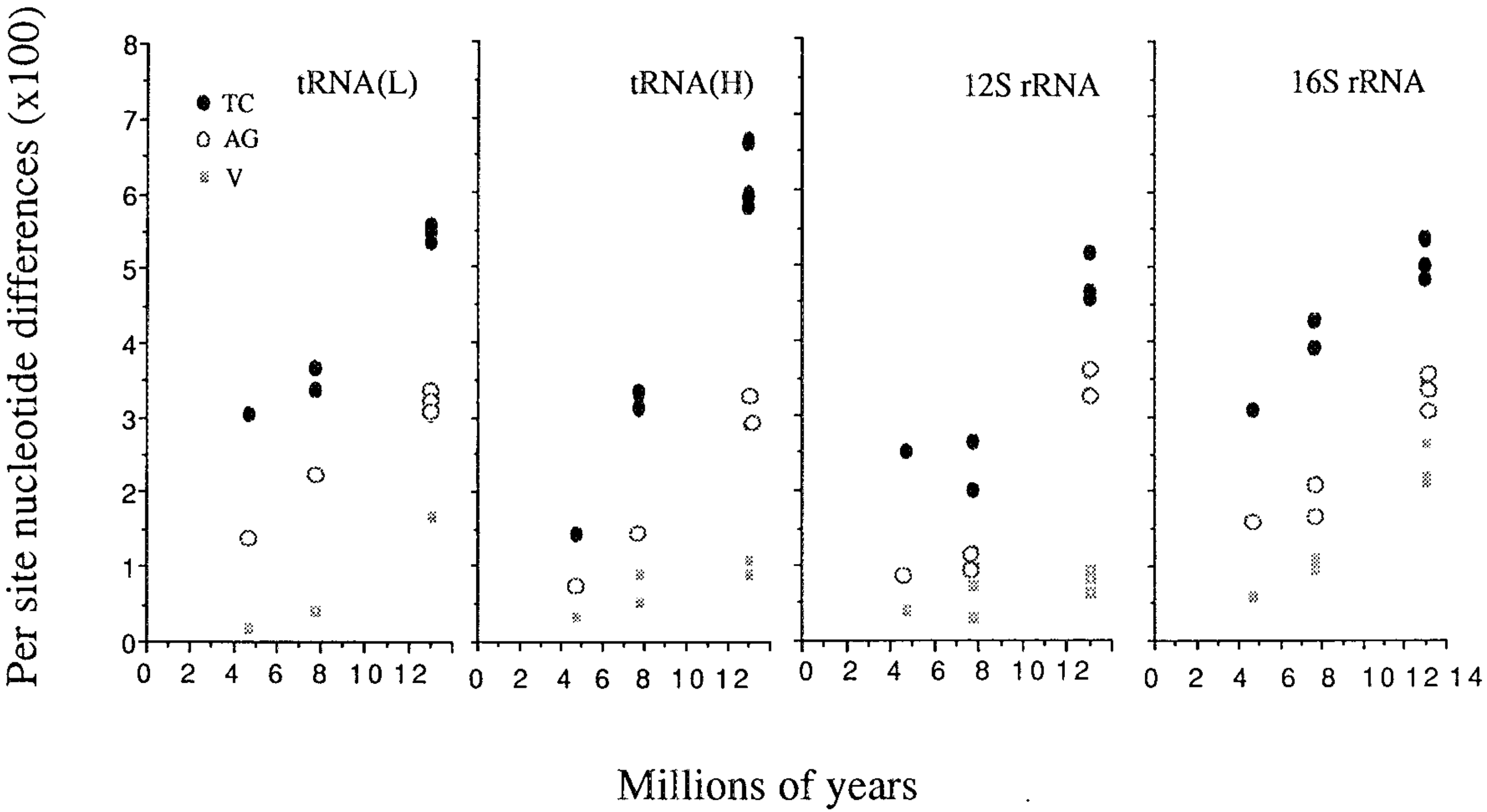
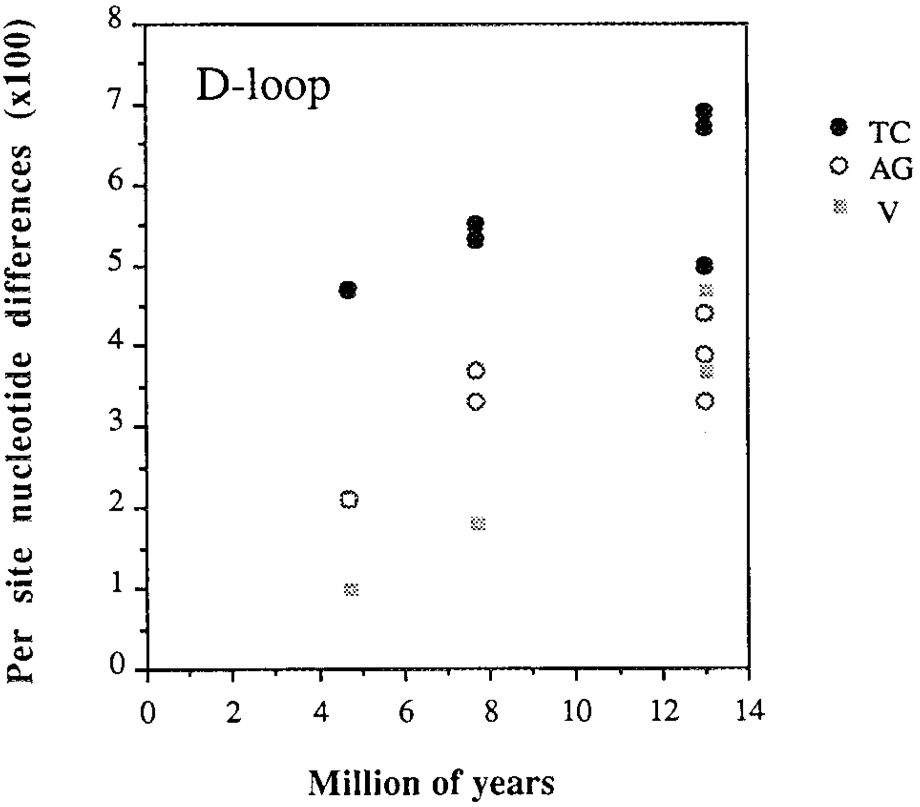

Per site nucleotide differences (x100)
0
1
2
3
4
5
6
7
8
tRNA(L)
● TC
○ AG
▨ V
tRNA(H)
12S rRNA
16S rRNA
0 2 4 6 8 10 12
0 2 4 6 8 10 12
0 2 4 6 8 10 12
0 2 4 6 8 10 12 14
Millions of years

Figure 5.4 Accumulation of nucleotide differences in the major noncoding region (D-loop).

Observed number of nucleotide differences for each pair of species is plotted against estimated divergence times (Horai et al. 1992). Nucleotide differences are examined in the categories of TC transitions, AG transitions, and transversions (V). The pairs compared from the left are: human (H) - CC at 4.7 mya; gorilla (G) - CC and G - H at 7.7 mya; and orangutan (O) - CC, O - H and O - G at 13 mya.


D-loop
Per site nucleotide differences (x100)
0
1
2
3
4
5
6
7
8
0
2
4
6
8
10
12
14
Million of years
● TC
○ AG
▨ V

Nucleotide differences in the noncoding regions

The noncoding region is analyzed for two separate portions. One is an assembly of small sequences dispersed between genes in the 4.9 kb region compared among six hominoids (common chimpanzee, pygmy chimpanzee, human, gorilla, orangutan, and siamang). Of the total 114 noncoding sites, there are 31 variable sites due to single nucleotide changes (or 27%). Deletions and/or insertions occur at 32 sites, the frequency being as high as 28 %. This implies that although single base substitutions in the noncoding region are frequent, deletions and/or insertions are also much more common than in any other regions.

The other portion is the major noncoding region (D-loop). This region was compared among 5 species (common and pygmy chimpanzees, human, gorilla (Foran et al. 1988), and orangutan (newly determined in this study)). Alignment of the orangutan sequence with the other four hominoid sequences was done by hand, taking account of the transition bias in the region (Horai and Hayasaka 1990), so as to minimize the number of transversions in the sequence alignment. This region contains the O_H and promoters for transcription (HSP, LSP). The region is also the target sites for numerous proteins (eg. transcription and regulatory factors) and enzymes (eg. DNA and RNA polymerase). On the other hand, being the only major noncoding region in the mtDNA, this region is estimated to evolve three to five times faster than the remainder of the mtDNA (Aquadro and Greenberg 1983; Cann et al. 1984; Horai and Hayasaka 1990). The D-loop region is, therefore, subjected to various evolutionary pressures, and the occurrence of nucleotide substitutions is significantly non-random (Horai and Hayasaka 1990). When compared with human, gorilla and orangutan have large portions of deletions and are much shorter than the other three species. Consequently, in the aligned 1147 sites, sites shared by the five species only amount to 779 sites (68%). The shared sites are assembly of the relatively conserved portion of the D-loop. Examination of the nucleotide differences in such sites, however, indicates a considerable amount of multiple hit substitutions (Figure 5.4). The proportion of transversion is higher than those in tRNA, rRNA and protein regions. Nevertheless, the observed number of transversions accumulate linear with time.

Conclusions

In this chapter, I have examined the nucleotide differences observed among the closely related hominoid species. It revealed a remarkably biased mode of substitutions in the mtDNAs, which is why the multiple hit substitutions can occur in a relatively short period of time. One is that nucleotide substitutions occur at restricted sites. Between human and chimpanzee, 70% of the observed differences are silent changes that occur mostly in the small noncoding regions or third codon positions of protein genes. Second is the strong preference to transitions. For example, out of 852 third codon positions that differ between human and common chimpanzee, 93% account for transitions and 66% are the TC transitions (in the L-strand). This indicates that a site C almost always change to a T, and a T to C. Likewise, an A almost always change to G, and a G to A. A consequence from this is the relatively constant C+T content observed among different species (Table 3.1). Third is the biased base compositions. The base compositions of hominoid mtDNAs are similar to one another, which suggest a stationary model of substitutions in which C to T changes and T to C changes (or A to G, and G to A changes) accumulate at similar rate. Therefore, when the base composition is biased, substitutions involving the rare base ("G" in the case of the L-strand vertebrate mtDNA) are more likely to make multiple hit substitutions. An extreme case is observed in Drosophila mtDNA, where G+C content at the third codon positions is as low as 3.3% (Satta and Takahata 1990).

Due to the biased mode of substitutions and the elevated rate of substitutions in mtDNA, the saturation of nucleotide differences is observed even between human and chimpanzee. Such differences are not accurate as a measure for molecular clock. In the case of hominoid mtDNA, all changes in the tRNA and rRNA regions can be used as measures for molecular clock, whereas only transversions in the noncoding regions, and nonsynonymous changes and synonymous transversions in the protein coding regions are considered to act as molecular clock.

The idea of using molecular clock for phylogenetic analyses rests on an assumption that the rate of accumulation of nucleotide changes remains steady through time. As I have shown here, majority of the differences in the mtDNA account for multiple hit substitutions. When using mtDNA for phylogenetic analyses, it is therefore essential to pick out sites that sufficiently act as a molecular clock.

CHAPTER SIX

HOMINOID PHYLOGENY

Resolution of trichotomy and the estimation of divergence times

The precise branching pattern (cladogram) and dating in hominoid diversification have been the topics in the hominoid phylogeny (Goodman et al. 1983; Foran et al. 1988; Djian and Green 1989; Gibbons 1990). To resolve the trichotomy problem, it is essential to find a number of nucleotide substitutions that can be assigned to internodal branches in the cladogram of hominoids (Saitou and Nei 1986). The longest DNA sequences available from the nuclear genome (the ψη-globin gene and its flanking region; 11,483 bp) assign about 8 to 14 substitutions that can support the human-chimpanzee clade. The number of substitutions are not large enough, and the likelihood is not significantly higher than that of the human-gorilla or chimpanzee-gorilla clade (Goodman et al. 1989). Recently, comparison of common region of 4,938 bp length for pygmy and common chimpanzees, gorilla, orangutan and siamang (This region is shown by hatched bars in Fig 2.1.) gave a reliable answer to the problem (Horai et al. 1992). The sequence differences clearly indicated that the closest relatives to human are chimpanzees rather than gorilla. With relatively small numbers of nucleotide differences, the same conclusion was drawn by the *COII* gene (Ruvolo et al. 1991), DNA-DNA hybridization (Sibley and Ahlquist 1984; Sibley et al. 1990; Caccone and Powell 1989), the ψη- (Koop et al. 1986; Miyamoto et al. 1987) and ε-globin genes (Koop et al. 1989), the ribosomal RNA gene (Gonzalez et al. 1990), and the immunoglobulin-ε pseudogene (Ueda et al. 1989).

By using only unsaturated parts of sequence differences in which the mtDNA genealogy is not obscured by multiple substitutions, the divergence times of gorilla (T_g), human (T_h), and between common and pygmy chimpanzees (T_c) are estimated to be: T_g

= 7.7 ± 0.7 million years ago (mya), T_h = 4.7 ± 0.5 mya and T_c = 2.5 ± 0.5 mya (Horai et al. 1992). The T_g is similar to the estimates from DNA-DNA hybridization (Sibley and Ahlquist 1984; Sibley et al. 1990; Caccone and Powell 1989), the ψη-globin genes (Goodman et al. 1990; Koop et al. 1986; Miyamoto et al. 1987), and the ribosomal RNA gene (Gonzalez et al. 1990), while the T_h is similar to the estimates from the ribosomal RNA gene (Gonzalez et al. 1990) and immunoglobulin-ε pseudogene (Ueda et al. 1989). The maximum likelihood estimates based on the 896 bp mtDNA (Brown et al. 1982) are much shorter, T_g = 5.1 mya and T_h = 3.9 mya (Hasegawa and Kishino 1991), which is probably due to the relatively small region compared. The difference from the estimate in *COII* gene (Ruvolo et al. 1991) is due to their assumption of T_h at 6 mya, and to their data set which was largely based on the synonymous differences.

Here in this section, I will use only the unsaturated sites from the complete mtDNA sequences of four hominoids (human, common chimpanzee, gorilla and orangutan). From these sites, I will estimate the genetic distances among the four species and construct a phylogenetic tree. The proportion of the branch lengths of this tree will be compared with those obtained from the analyses of 4.9 kb region. The rooting of the orangutan branch will be particularly important because orangutan lineage showed an increased mutation rate (Horai et al. 1992). Since the new data set does not include a proper out-group to root the orangutan branch in a phylogenetic tree, the root of orangutan was deduced in proportion to the orangutan branch in the phylogenetic tree obtained from the 4.9 kb region.

Phylogenetic analysis of the whole mitochondrial genome

To obtain a general view of the hominoid mitochondrial genome, I began by analyzing the observed differences of the whole mtDNA region, ignoring possible heterogeneous substitution rates along DNA sequences. Table 6.1 summarizes the observed number of differences in the pairwise comparisons of the 4 hominoid mtDNAs. Among the 16,209 shared sites for all the four species, differences between human and

Table 6.1 Nucleotide differences in the pairwise comparisons of 4 hominoid mtDNAs

Pairs	Number of differences							Transitions
	Total	TC	AG	CA	TA	CG	TG	(%)
H-C	1388 (8.6)	864	406	52	41	17	8	91.5
H-G	1707 (10.5)	1040	447	102	74	28	16	87.1
C-G	1683 (10.4)	1010	451	104	79	19	20	86.8
H-O	2349 (14.5)	1240	588	271	157	60	33	77.8
C-O	2394 (14.8)	1284	587	282	168	42	31	78.2
G-O	2428 (15.0)	1294	591	272	182	55	34	77.6

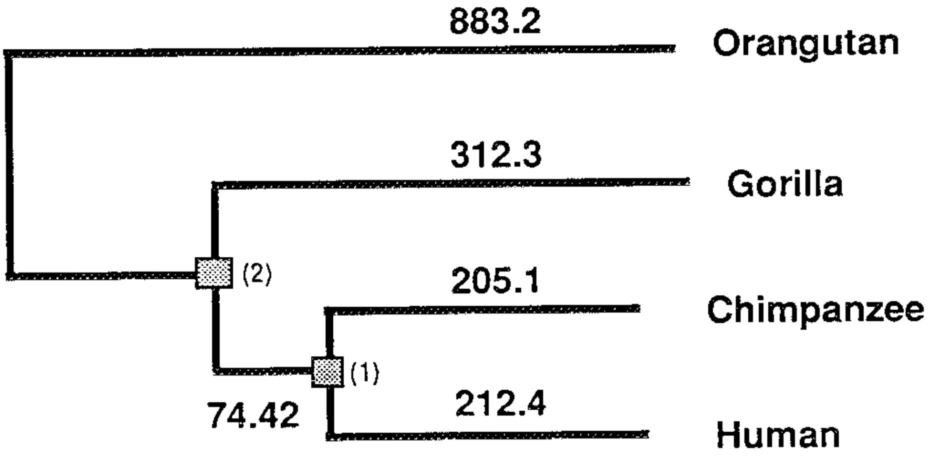
A total of 16,209 shared sites were compared among the 4 hominoid mtDNAs. The numbers in the parenthesis indicate the number of differences per 100 sites. Abbreviations: H (human), C (common chimpanzee), G (gorilla), and O (orangutan). TC indicates the differences between bases T and C in the L-strand. Likewise, AG, CT, CA, TA, CG and TG indicate the differences between the two respective bases in the L-strand.

Figure 6.1 Phylogenetic tree of 4 hominoid mtDNAs.

A phylogenetic tree was constructed by the neighbour joining method (Saitou and Nei 1987) using only the unsaturated differences observed in the whole mitochondrial genome. Included in the total number of 12,137 sites examined are 22 tRNAs, 12S rRNA, 16S rRNA, synonymous transversions and nonsynonymous sites of 13 protein genes. The numbers designated along the branches are the estimated number of differences for each branch. The nodes of the tree are numbered as (1) and (2).

NJ tree 12137 bp

(nonsynonymous + synonymous transversions + rRNAs + tRNAs)



chimpanzee was observed at 1,388 sites, which is 8.6% of the compared sites. In the human-gorilla and chimpanzee-gorilla pairs, differences were observed at 10.4% and 10.5% of the compared sites. The number of differences are larger than the human-chimpanzee pair by 319 and 295 sites, respectively. The differences between orangutan and the other three species are as high as 15% of the compared sites. As a whole, the observed sequence differences undoubtedly support the human-chimpanzee clade, provided that the mtDNA genealogy is topologically identical to the hominoid species tree.

There is a strong bias in the type of substitutions. In the L-strand, the differences in TC transitions constitutes 62% of the differences in the human-chimpanzee pair. Together with AG transitions, differences due to transitions become as high as 92%. The proportion of transitions in the observed differences decrease as the divergence time of a pair increases. Namely, the percent of transitions for the pairs involving orangutan is 78%. As noted in Chapter five, this can be explained by the saturation of the transitional changes due to multiple hit substitutions opposed to the linear accumulation of transversional changes. Within the transversions, differences between C and A is largest and differences between T and G are smallest. This is explained by the biased base composition of the L-strand. The relatively small differences for the changes involving G is due to the low G content in the L-strand.

There are a number of statistical studies for correcting multiple-hit substitutions (Kimura 1983; Nei 1987). For long sequences which include many different genes, however, another complication might occur due to gene-specific evolutionary rates and/or to extensive multiple-hit substitutions in some regions. To avoid these problems and obtain a reliable dating for diversification, I restricted an analysis to relatively conserved regions where one can follow the evolutionary process quite accurately (discussed in Chapter five). For this reason, AGS$_3$, TCS$_3$ and TCS$_1$ differences were excluded.

Table 6.2 Branch length in a phylogenetic tree of 4 hominoids. (b) Estimated branch length based on the tree in Figure 6.1. (a) The previous observation of Horai et al. (1992).

(bp)	(H, 1)	(C, 1)	(1, 2)	(G, 2)	(O, 2)
a) 4270	0.0102 ±0.00154	0.0093 ±0.00147	0.0077 ±0.00134	0.0176 ±0.00203	0.0563 ±0.00363
	(0.58)	(0.53)	(0.44)	(1)	(2.83)
b) 12137	0.0175 ±0.00120	0.0169 ±0.00118	0.0061 ±0.00071	0.0257 ±0.00146	0.0728 ±0.00245
	(0.68)	(0.66)	(0.24)	(1)	(3.21)

The numbers given for each branch represent the number of nucleotide differences per site ± 1 standard error. The numbers in the parentheses indicate the branch length relative to the length of gorilla branch.

A phylogenetic tree was constructed by the neighbour joining (NJ) method (Saitou and Nei 1987) (Fig 6.1). Among the 12,137 sites compared, a large number of differences (74.4 sites) was assigned to the internal branch leading to human-chimpanzee clade after the divergence of gorilla. Table 6.2 summarizes the length and the size of sampling errors for each branch. Each branch length of the two trees were compared with respect to its proportion to the gorilla branch. The overall feature of the new tree confirms the branch lengths obtained from the previous analysis of 4.9 kb region. For the internal branch leading to human-chimpanzee clade, the amount of one standard error becames as small as 1/9 the length of the branch. The new data set, however, estimates smaller size of internodal branch relative to the branch of human and chimpanzee. If assuming the length of gorilla branch be 7.7 million years, the branch lengths of human and chimpanzee becomes 5.2 and 5.1 ± 0.2 million years, respectively. This agrees with the divergence time of human and chimpanzee at 4.7 ± 0.5 mya, estimated from the 4.9 kb region (Horai et al. 1992). In the analysis of 4.9 kb region, the root of phylogenetic tree was determined using siamang as an out group species. It was approximately at 1/5 of the orangutan branch (Data 3; 54.9/260.4). When assuming the same proportion to root the phylogenetic tree in Fig. 6.1, and assuming the divergence time of orangutan at 13 mya, the divergence times of gorilla, and chimpanzee-human are calculated at about 7.9 ± 0.3 mya and 5.8 ± 0.2 mya, respectively. These values are also similar to the previous estimates from the 4.9 kb region.

Conclusion

The comparison of the whole mitochondrial genome from four hominoid species confirms the previous study of 4.9 kb region, which suggested that gorilla diverged about 7.7 mya, and chimpanzee and human at about 4.7 mya, assuming the divergence time of orangutan at 13 mya. It should be noted, however, that a gene genealogy (gene tree) based on the molecular clock does not necessarily agree with a species relatedness (species tree), because of the possibilities of ancestral polymorphisms (Nei 1987). The time difference between gorilla and chimpanzee divergences being as long as 3 million years corresponds to approximately 200,000 generations. If the ancestral population size is the same order of magnitude as the present human population size (10^4), the effect of discordance between gene tree and species tree is small (Takahata 1989, Horai et al. 1992). As far as the mitochondrial gene tree is concerned, I would like to conclude that chimpanzees are the closest relatives to human.

CHAPTER SEVEN

CONSIDERATION OF THE CORRECTION METHODS

In the phylogenetic analysis, I have excluded the AGS_3, TCS_3 and TCS_1 differences and used the unsaturated observed differences. However, correction of multiple hit substitutions are necessary for the synonymous substitutions when shorter regions are compared. In this section, I will discuss about the substitution model for the mtDNA. The purpose for this is to find the best way to estimate the synonymous rate for each gene. For the following reasons, I will use the four DNA sequences from common and pygmy chimpanzees, human and gorilla corresponding to the 4.9 kb region (shown by hatched bar in Fig. 2.1): For synonymous transitions, it is important to compare between the closest pairs (i.e. between the two chimpanzee pairs) where saturation effect is not obvious. Without this, it would be difficult to estimate the right substitution model, as suggested from the accumulation of nucleotide differences in Figure 5.2. The orangutan mtDNA is too distantly related from that of chimpanzees and human; about 34% of the third codon positions differ among these species, which is close to the saturation level. Therefore, including the orangutan mtDNA in the following analysis gives rise to a difficult and commonly recognized problem in estimating accurate nucleotide substitution rates. The gorilla mtDNA also differs substantially from the human and chimpanzee mtDNAs. However, if I exclude it from the analysis, the sampling errors become too large. Actually, the observed differences at the third codon positions are about 27%, which seems not too extensive to make reasonable multiple hit corrections. I exclude the *ND1* because it represents only a small portion (14.4%) of the entire gene and is subjected to large sampling errors. This study is based on the analysis of the remaining 5 protein genes; *ND2*, *COI*, *COII*, *ATPase 8*, and *ATPase 6*.

Nucleotide substitutions in a stationary Markov model

There is a substantial heterogeneity from gene to gene in terms of the base compositions at the third codon positions (Table 3.1). Together with extremely high transition rates, this heterogeneity suggests that the dynamics of the nucleotide substitution process at the third codon positions or at the synonymous sites may be different from gene to gene. As discussed previously, it is also important to realize that, because of these biases, the saturation level can differ among genes.

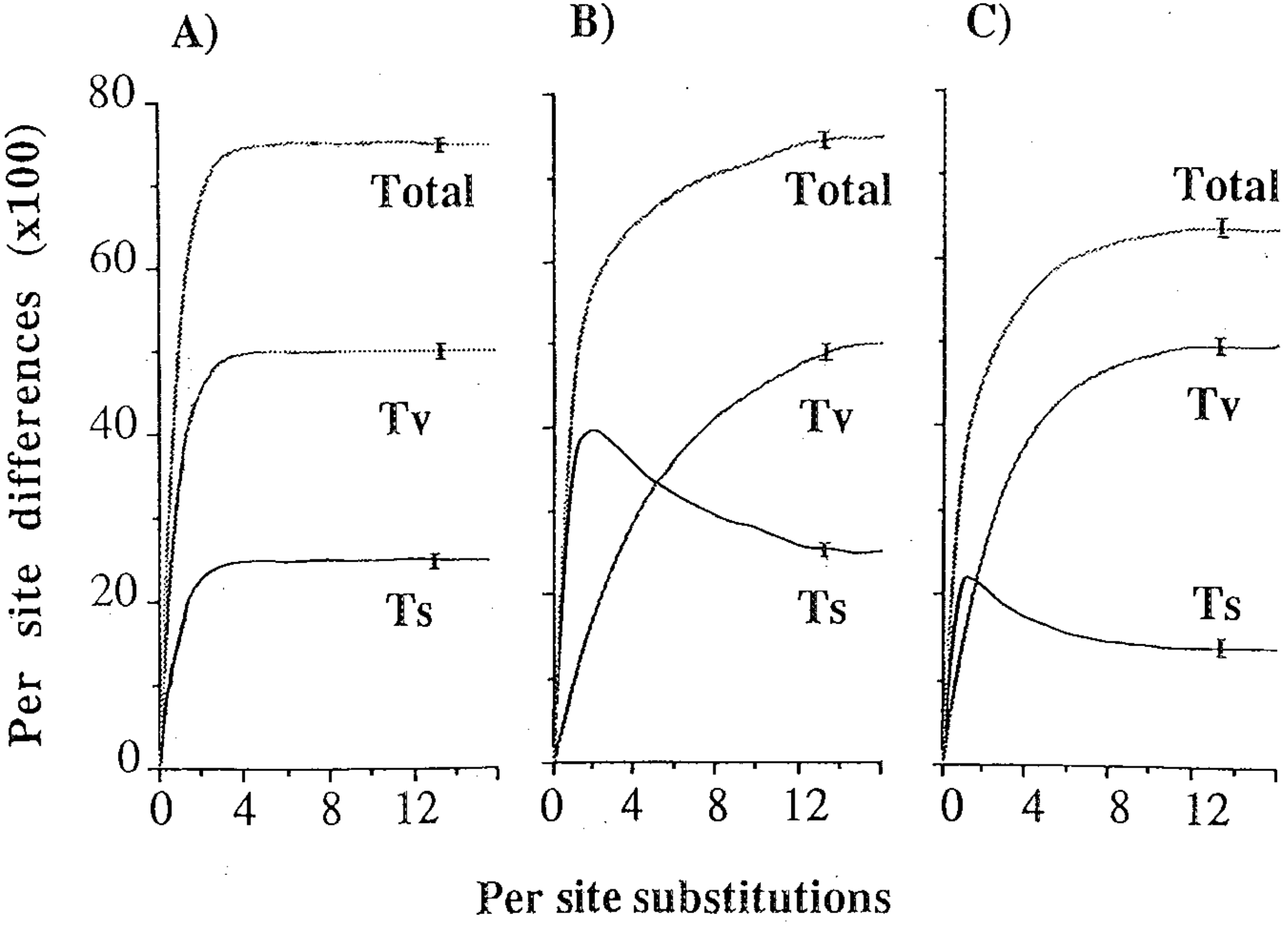
None of the nucleotide substitution models proposed thus far are particularly good at describing empirical mtDNA data (Fitch 1986). The method by Jukes and Cantor (1969) is inappropriate, because it assumes an equal transition-transversion probability and equal frequency among the four nucleotides. For the same reason, Kimura's two-parameter method (1980) may be unsuited, because it ultimately leads to the equal frequency of the four bases despite the assumed different rates between transitions and transversions. Models that consider both different transition-transversion rates and base compositional biases must be used, and some such are developed by Felsenstein (1981), Kimura (1981), Takahata and Kimura (1981), Lanave et al. (1984) and Hasegawa et al. (1985). Under some of these models satisfying reversibility, biased base compositions in a sequence can be kept constant through the nucleotide substitution process. For simplicity, I used the model of Hasegawa et al. (1985) in which the probability of base *i* to be replaced by base *j* is defined by the product of stationary base composition (π_j, j = A, G, C, or T) and the relative transition (α) or transversion rate (β) (For details of the model, see Appendix I). The parameters α and β affect the initial increase of nucleotide differences and the time required to reach the saturation level (A and B in Figure 7.1). On the other hand, the equilibrium frequencies π_j are related to the ultimate saturation level, which is given by $2(\pi_A\pi_G+\pi_C\pi_T)$ for transitions and $2(\pi_A+\pi_G)(\pi_{C+}\pi_T)$ for transversions (B and C in Figure 7.1).

Figure 7.1 The effect of base compositions and transition rates in the nucleotide substitutions.

The effect of π_j, α and β on the saturation level, in which πj is the frequency of base j (j = A, T, G, or C), α and β are relative transition and transversion rates, was examined by a simulation study using a stationary Markov model. The substitution matrix for base i to be replaced by base j is defined as follows (Hasegawa et al. 1985):

$$\begin{array}{c|cccc} i \backslash j & \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \\ \hline \mathrm{T} & 1-(\alpha\pi_C+\beta\pi_A+\beta\pi_G) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\ \mathrm{C} & \alpha\pi_T & 1-(\alpha\pi_T+\beta\pi_A+\beta\pi_G) & \beta\pi_A & \beta\pi_G \\ \mathrm{A} & \beta\pi_T & \beta\pi_C & 1-(\alpha\pi_G+\beta\pi_T+\beta\pi_C) & \alpha\pi_G \\ \mathrm{G} & \beta\pi_T & \beta\pi_C & \alpha\pi_A & 1-(\alpha\pi_A+\beta\pi_T+\beta\pi_C) \end{array}$$

The values of parameters used are: $\alpha=\beta=1$, $\pi_T=\pi_C=\pi_A=\pi_G=0.25$ for A), $\alpha=1$, $\beta=1/17$, $\pi_T=\pi_C=\pi_A=\pi_G=0.25$ for B) and $\alpha=1$, $\beta=1/17$, $\pi_T=0.15$, $\pi_C=0.34$, $\pi_A=0.47$, $\pi_G=0.04$ for C). An ancestral sequence of 1 kb is assumed to have a specified base composition in A), B) or C). According to each matrix, we generated a uniform random number U for each nucleotide site. If U is smaller than a specified probability in the matrix, the site is changed. This process is repeated over the entire sequence and stored for the next round of substitutions. The observed number of differences is counted at given numbers of the total substitutions. For each parameter set, we repeated 1000 times, and computed the average and standard deviation. For the transitions (Ts), the transversions (Tv) and the total changes, the observed number of differences were plotted against the actual number of total substitutions. Graph A) and B) show that the ultimate saturation level is the same even for different values of α and β, as theoretically expected. Graph B) and C) show that the saturation level of the observed differences changes by the base compositions.


A)
B)
C)
Per site differences (x100)
80
60
40
20
0
Total
Tv
Ts
Total
Tv
Ts
Total
Tv
Ts
0
4
8
12
0
4
8
12
0
4
8
12
Per site substitutions

Simulation of nucleotide substitutions in mtDNA

Since most substitutions at the third codon positions are presumably neutral, the base compositions and the transition-transversion rates should reflect the actual mutation process in each gene. For a given gene, the base compositions at the third codon positions are rather similar among chimpanzees, human and gorilla. Importantly, however, there exists a strong bias towards A and C. TG content is extremely low, and the extent of which differs between genes (Table 3.1, Table 7.1). To see the effects of such compositional biases on the nucleotide substitutions, I carried out simulation analysis (Fig 7.2). For each gene, I used the observed base compositions at the third codon positions as equilibrium frequency (π_j, j = A, T, G, or C), but assumed the same rate, irrespective of genes, for transitions and transversions (α=1, β=1/17). The empirical relative values of α=1 and β=1/17 were deduced from the average rate of nucleotide differences among human and chimpanzees. They are also close to the values which were already suggested by many studies (Brown et al. 1982, Hixson and Brown 1986, Hayasaka et al. 1988, Foran et al. 1988, Horai and Hayasaka 1990).

The plot of the observed number of differences per site against actual number of substitutions were made for 3 categories, TC transitions, AG transitions and transversions. The mode of accumulation obtained from the simulation analysis (Figure 7.2) resembles that of the observed differences (Figure 5.2): The saturation level of TC and AG transitions differ. Transversions accumulate linearly with the actual number of substitutions (or time). Importantly, even a small change (< 5%) in the lower base compositions results in a substantial change in the transition differences. For example, genes with lower T content show much lower levels of the CT transition differences, and genes with lower G content show much lower levels of the AG transition differences (Table 7.1, Figure 7.2). This suggests that variation in the base compositions can be an important factor for determining the synonymous differences. Consequently,

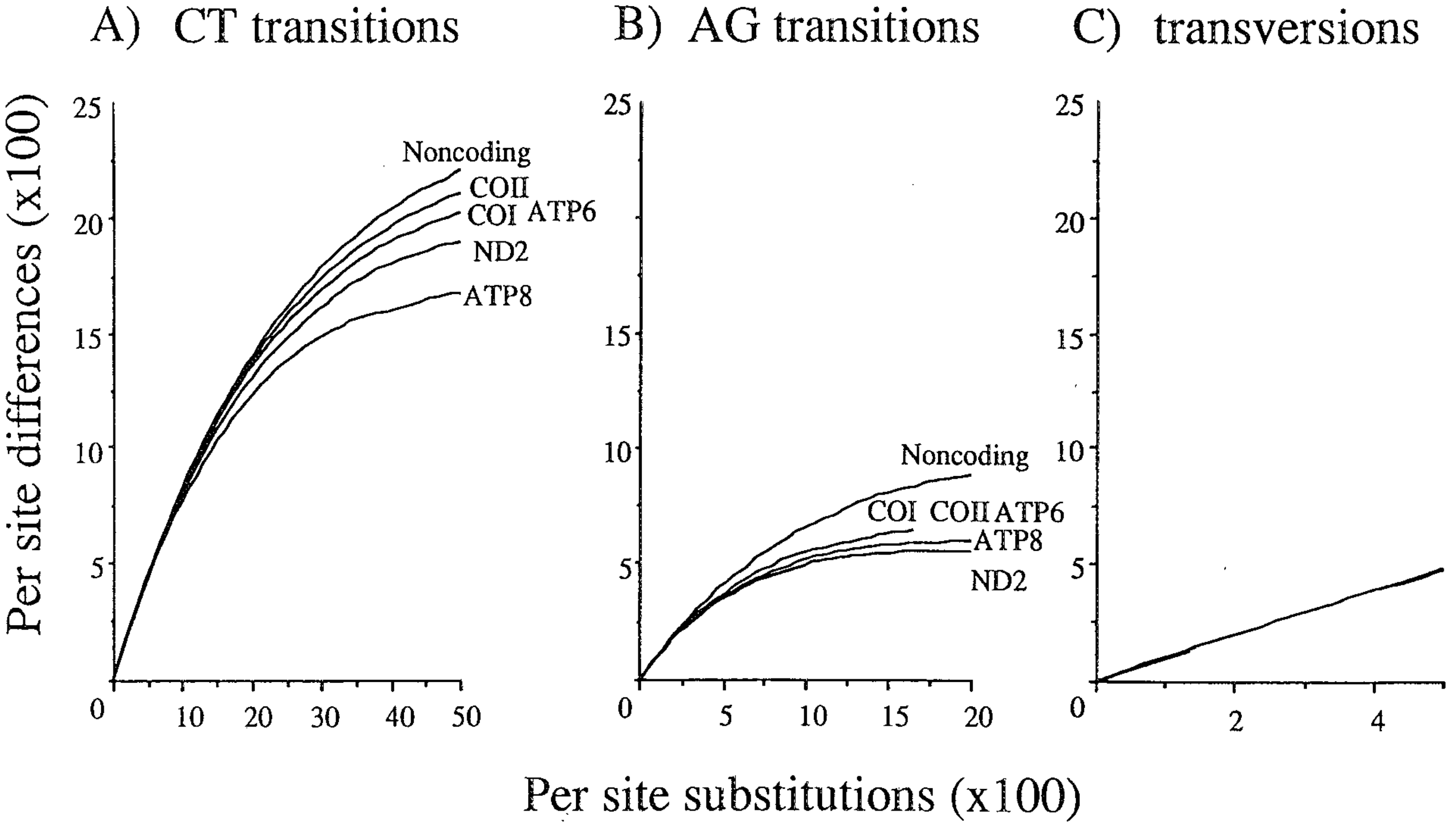
Table 7.1 Base compositions at the third codon positions and noncoding region

Genes	ND2	COI	COII	ATPase 8	ATPase 6	Noncoding
Number of sites	347	513	227	53	150	50
A (%)	38	36	35	47	39	30
C (%)	42	40	40	34	36	39
T (%)	16	19	20	15	20	23
G (%)	4	5	5	4	5	8
TG (%)	20	24	25	19	25	31

Base compositions shown above are the average of common chimpanzee, pygmy chimpanzee, human and gorilla.

Figure 7.2 Simulation of nucleotide substitutions for each gene.

The relationship between the actual number of substitutions and the observed number of differences were examined. The values of π_j (j = A, T, G, or C) are set using the observed base compositions in each gene (Table 7.1) and the relative transition and transversion rates are set as $\alpha=1$ and $\beta=1/17$, respectively (see the legend of Figure 7.1 for details). Graph A) is for TC transitions, B) for AG transitions, and C) for transversions. The actual length of the abscissa in A), B) and C) indicates the same length of time. Note that there is a large difference in the saturation level between TC and AG transitions because of lower AG contents.


A) CT transitions
B) AG transitions
C) transversions
Per site differences (x100)
25
20
15
10
5
0
10
20
30
40
50
Noncoding
COII
COI
ATP6
ND2
ATP8
25
15
10
5
0
5
10
15
20
Noncoding
COI
COII
ATP6
ATP8
ND2
25
20
15
10
5
0
2
4
Per site substitutions (x100)

synonymous substitutions may be inaccurately estimated if we ignore biased transition rates and base compositions.

Estimation of synonymous sites

To deal with the synonymous and nonsynonymous substitutions, one must first determine the numbers of such sites per gene. A usual method for counting the number of synonymous sites per gene assumes equal substitution rates among four nucleotides (e.g. Nei and Gojobori 1986) so that the number of synonymous sites for a two-fold degenerate site is assigned to 1/3. However, this assignment is not valid when transition rates are much faster than transversion rates as in primate mtDNAs. Because there is only a small probability for transversions that result in nonsynonymous changes, the number of synonymous sites for a two-fold degenerate site should be closer to 1 than to 1/3. Clearly, this increases the number of synonymous sites and decreases that of nonsynonymous sites per gene compared to those expected under Jukes and Cantor's model.

I have compared the number of synonymous and nonsynonymous sites that are estimated in two different ways (Table 7.2). Values for *Sng* and *Nng* rely on the assumption that all four nucleotides can mutate equally likely (Nei and Gojobori 1986). The values *S* and *N* are estimated by a model that reflects biased substitutions (Hasegawa et al. 1985; For details of the method, see Appendix I). Empirical relative values of transition (α=1) and transversion (β=1/17), and different base frequencies (π_j, j = A, T, G, or C) at the third codon positions of each gene were taken into account in the estimation. As the result, the values in *S* became larger than *Sng* by 133-144%, and the values in *N* became smaller than *Nng* by 86-90%.

In order to study quantitatively different ways of counting synonymous sites (nonsynonymous sites = the total number of sites – synonymous site), I compared the correlation of the synonymous and nonsynonymous differences per site. The total

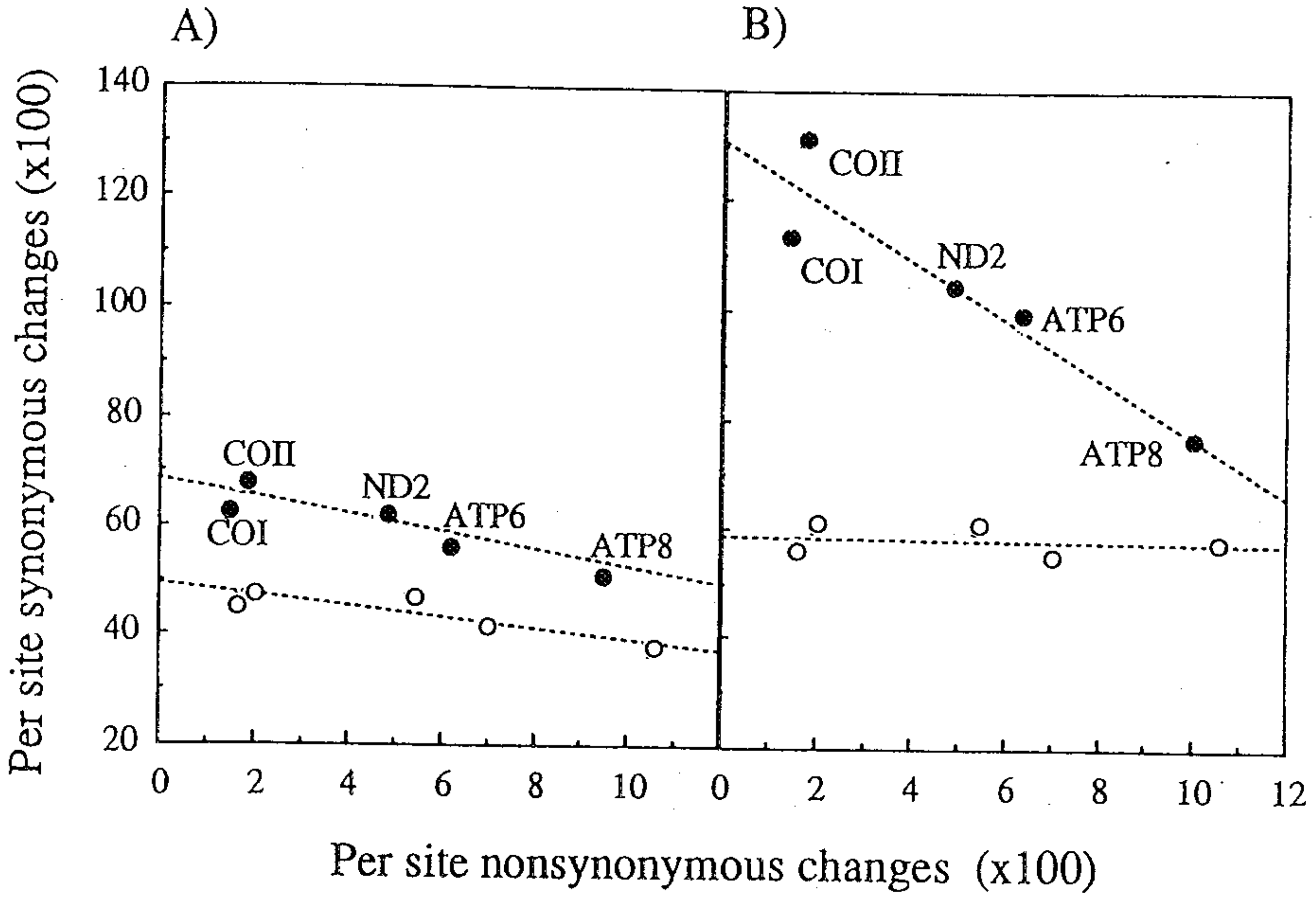
Table 7.2 Estimated number of synonymous and nonsynonymous sites

Genes	*S*	*Sng*	*N*	*Nng*
ND1	323.1	242.9	630.9	711.1
ND2	341.7	257.4	699.3	783.6
COI	507.0	368.0	1032.0	1171.0
COII	224.5	160.7	456.5	520.3
ATP8	50.0	37.0	109.0	122.0
ATP6	220.2	160.7	409.8	469.3
COIII	257.6	179.3	525.4	603.6
ND3	119.5	83.0	225.5	262.0
ND4L	101.7	74.8	186.3	213.2
ND4	456.7	338.1	914.3	1032.9
ND5	592.1	434.1	1219.9	1377.9
ND6	160.8	122.1	364.2	402.9
Cytb	369.4	272.6	770.6	867.5

Synonymous (*S*) and nonsynonymous (*N*) sites were estimated from base contents and transition to transversion ratio of 17 to 1. *Sng* and *Nng* are synonymous and nonsynonymous sites estimated by the Nei and Gojobori's method.

Figure 7.3 Relationship of synonymous and nonsynonymous changes.

The relationship of observed synonymous and nonsynonymous differences (A) and inferred synonymous and nonsynonymous substitutions (B) are shown. The inferred number of synonymous and nonsynonymous sites is calculated by Jukes and Cantor's model which assume equal probability for all nucleotide substitutions (indicated by a solid circle), and by the Hasegawa et al's. model which consider both biases in base composition and transition rate (indicated by the open circle) (see Appendix I). The number of differences for each gene was estimated from the total branch length of a tree of common chimpanzee, pygmy chimpanzee, human and gorilla obtained by the ordinary least squares (see Appendix I). In A), the number of observed synonymous and nonsynonymous differences in each gene is calculated using the observed number of pairwise differences. For the solid circle in B), the number of synonymous and nonsynonymous substitutions is estimated based on Kimura's two parameter model (Kimura 1980), and it was converted into the per-site number of changes using the number of sites inferred from the Jukes and Cantor's model. This is an example of inconsistent use of different substitution models in computing the number of synonymous sites and correcting multiple hit substitutions. For the open circle in B), the number of synonymous and nonsynonymous substitutions is estimated based on Figure 7.2 (see Appendix I), and used the number of sites inferred from the same Hasegawa et al's. model.


A)
B)
Per site synonymous changes (x100)
140
120
100
80
60
40
20
COII
COI
ND2
ATP6
ATP8
COII
COI
ND2
ATP6
ATP8
0
2
4
6
8
10
0
2
4
6
8
10
12
Per site nonsynonymous changes (x100)

number of synonymous and nonsynonymous differences in a tree of four species (common and pygmy chimpanzee, human and gorilla) were estimated by the least squares method. The total number of synonymous and nonsynonymous differences were divided respectively by the number of synonymous (*S* and *Sng*) and nonsynonymous (*N* and *Nng*) sites estimated in Table 7.2. Table 7.2 shows that the number of synonymous sites are underestimated, whereas that of nonsynonymous sites are overestimated under Jukes and Cantor's model. The plot of nonsynonymous differences per site against synonymous differences per site is shown in Figure 7.3A. The distribution of the plots shows that the nucleotide differences per synonymous site are overestimated by 20% and those per nonsynonymous site are underestimated. Consequently, the relationship between the synonymous and nonsynonymous differences based on Jukes and Cantor's model (indicated by a solid circle) shows a stronger negative correlation than that based on Hasegawa et al's. model (indicated by an open circle).

Multiple hit corrections

In order to evaluate the accuracy of correction methods, I applied Kimura's two-parameter model (Kimura 1980) and Hasegawa et al.'s model to the synonymous differences (Figure 7.3B). Correction with Kimura's two-parameter model (indicated by a solid circle in Figure 7.3B) considerably inflated the inferred number of synonymous substitutions, and the negative correlation becomes even stronger than that in Figure 7.3A. On the other hand, the correction based on Hasegawa et al.'s model (see Figure 7.2 and Appendix) gives similar estimates of the synonymous substitutions for all the genes (indicated by an open circle in Figure 7.3B). As mentioned, the latter estimates depend on α and β. In case of *ATPase 8* which shows the lowest saturation level in Figure 7.2, the estimate of synonymous substitutions is substantially influenced by the ratio of transitions to transversions. The relative values of α and β were estimated from the observed numbers of transitions and transversions in the comparisons among common chimpanzee, pygmy chimapnzee, and human. They turned out to be α=1 and β=1/17, which are very close to the maximum likelihood estimates obtained by the

DNAML in PHYLIP (Felsenstein 1990). It is reasonable to expect that the genes on the mtDNA genome have more or less the same mutation rate and therefore similar levels of synonymous substitutions.

Conclusions

The simulation studies showed that the biased transition rates and base compositions can be important factors for determining the synonymous differences, and estimating the number of synonymous sites in mtDNA. Whether or not these factors are considered in the model makes a great difference in the estimates. We therefore need to be cautious in choosing the nucleotide substitution model for analysis. In general, the number of synonymous sites per two-fold degenerate sites must be determined at least in terms of α, β and π_j (j = A, T, G or C). I claim that the number of synonymous sites must be counted consistently with the model of nucleotide substitutions actually used for multiple hit corrections. For example, Nei and Gojobori's method should be used with the Jukes and Cantor's substitution model. When the number of differences is large, and when the substitutions are highly biased as in the case of mammalian mtDNA, an inconsistent use in substitution models may give substantially overestimated or underestimated values. This caution may be applied to many other methods for correcting multiple hit substitutions (e.g. Brown et al. 1982; Li et al. 1985).

I have used a stationary Markov model to simulate the nucleotide substitutions in the mtDNA. Although the model may still be far from reality, the mode of accumulation of nucleotide differences obtained by the model resembles that of the observed differences fairly well. Correction of multiple substitutions by the model suggests that genes on the mtDNA genome have more or less the same mutation rate.

CHAPTER EIGHT

ESTIMATION OF SUBSTITUTION RATES

Substitution rates of protein genes

It is obvious from my analysis that the number of synonymous substitutions per site (distances) is fairly uniform over the genes. This supports the assumption of equal synonymous rates for all mitochondrial genes. I took account the weighted average of the synonymous substitutions over the genes when compared between gorilla and the remaining three species; common chimpanzee, pygmy chimpanzee and human (Appendix I). With the divergence time of gorilla at 7.7 mya which was estimated by Horai et al. (1992), the synonymous rate for the hominoid mtDNA becomes $2.37 \pm 0.11 \times 10^{-8}$ per site per year (Table 8.1). This rate is 5 to 10 times faster than that of nuclear DNAs (Brown et al. 1982) or even faster (20 times) if the synonymous substitution rate for nuclear genes is 1.2×10^{-9} (Satta et al. 1991).

For the nonsynonymous rate, any multiple hit correction is practically unnecessary among the four species. The rate of nonsynonymous substitutions for each gene is estimated from the average of the nonsynonymous differences between gorilla and the remaining two to three species; common chimpanzee and human, or common chimpanzee, pygmy chimpanzee and human (Table 8.1, Appendix I). The nonsynonymous substitution rate per site per year ranges from 0.7×10^{-9} for *COI* to 5.7 $\times 10^{-9}$ for *ATPase 8*, which is at least 5 times lower than that of the synonymous change. The degree of functional constraints (measured by the ratio of the nonsynonymous to the synonymous substitution rate) being 0.03 and 0.24, respectively.

Table 8.1 Rate of nucleotide substitutions

Regions		(x 10^{-9} /site/year)
synonymous		23.7 ± 1.10
noncoding		26.8 ± 5.90
nonsynonymous		
	ND1	2.1 ± 0.48
	ND2	2.4 ± 0.49
	COI	0.7 ± 0.23
	COII	0.9 ± 0.37
	ATPase 8	5.7 ± 1.93
	ATPase 6	3.6 ± 0.79
	COIII	1.4 ± 0.43
	ND3	2.5 ± 0.90
	ND4L	1.2 ± 0.66
	ND4	2.3 ± 0.43
	ND5	4.0 ± 0.49
	ND6	1.7 ± 0.57
	Cyt *b*	3.0 ± 0.53
12S rRNA		4.1 ± 0.56
16S rRNA		6.9 ± 0.56
22 tRNAs		5.6 ± 0.51

The rate was estimated from the average changes of gorilla-chimpanzees and gorilla-human pairs, assuming the divergence of gorilla at 7.7 million years ago.

Substitution rates of tRNA and rRNA genes

As mentioned before, the orangutan tRNAs have many anomalous features (Horai et al. 1992). I therefore used the average tRNA differences between gorilla and common chimpanzee, and gorilla and human to calibrate the substitution rate. Assuming the divergence time of gorilla at 7.7 mya, the average transition and transversion rate over the 22 tRNA genes are estimated at 5.2 x 10^{-9} and 0.5 x 10^{-9} per site per year, respectively. The ratio of the transition to transversion rate is approximately 11. The overall rate (the sum of the transition and transversion rates) becomes 5.6 x 10^{-9} per site per year. This rate is about two thirds of the previous estimate (8.5 x 10^{-9} per site per year) for $tRNA^{His}$, $tRNA^{Ser(AGY)}$, and $tRNA^{Leu(CUN)}$ (Brown et al. 1982). Nevertheless, it is true that tRNAs in mtDNA have evolved much faster than those in nuclear DNA, partly because of higher mutation rates and less selective constraints against mitochondrial tRNAs.

The substitution rates for *12S rRNA* and *16S rRNA* genes are estimated in a similar way as the tRNA coding regions, assuming the divergence time of gorilla at 7.7 mya. The transition and transversion rate for *12S rRNA* are 3.6 x 10^{-9} and 0.5 x 10^{-9} per site per year, respectively. The rates for *16S rRNA* are a little higher, and are 6.0 x 10^{-9} and 1.0 x 10^{-9} per site per year, for transitions and transversions, respectively. The ratios of the transition to transversion rate for the *12S rRNA* and *16S rRNA* are 6.9 and 6.2, respectively. The overall rates for *12S rRNA* and *16S rRNA* become 4.1 x 10^{-9} and 6.9 x 10^{-9} per site per year, respectively

Substitution rates in the noncoding region

In Chapter four, I separated the noncoding region in two parts. One was the major noncoding region that is usually called D-loop. Although a noncoding region, D-loop contains the main regulatory elements concerning replication and transcription, which are expected to be not so much variant among closely related species. However, the observation of the pairwise differences in the shared sites for five species suggested an extreme amount of multiple hit substitutions (Figure 5.4). The rate in this region is, therefore, difficult to estimate even between the two chimpanzees.

Another group of the noncoding region is an assembly of the remaining noncoding regions, which are scattered between the coding regions. Excluding the extremely conserved region (O_L; 32 bp), the substitution rate in such noncoding region was estimated (50 sites).

The base composition in the noncoding portions of the L-strand mtDNA is biased in much the same way as in the third codon positions (Table 7.1). Moreover, the ratio of the observed transversion to transition differences is 1/12=0.08 between gorilla and chimpanzees and 1/15=0.067 between gorilla and human. Thus, in terms of the actual number of substitutions the ratio becomes closer to 1/17=0.06. Using the actual base contents in the noncoding region and the relative transition (α=1) and transversion rate (β=1/17), I carried out a simulation study to make corrections for multiple hit substitutions (Figure 7.2). The estimated substitution rate is $2.68 \pm 0.59 \times 10^{-8}$ per site per year. Although the sampling error is large due to the short length of DNA sequence, the rate is very similar to the average rate of synonymous substitutions. This supports our contention that the mutation rate in mtDNA is uniform over the genome.

Conclusions

In respect of the rates of synonymous and nonsynonymous substitutions, Miyata et al. (1980) examined various eucaryotic genes and concluded that while the rate of nonsynonymous substitutions differ among genes, the rate of synonymous substitution is more or less the same for all genes. However, subsequent analyses suggested that the rate not only at synonymous sites but also in noncoding regions varies considerably in different parts of the genome (Koop et al. 1986, 1989; Li and Tanimura 1987; Li et al. 1987; Maeda et al. 1988; Kawamura et al. 1991). Also suggested was a positive correlation between synonymous and nonsynonymous substitution rates, although the correlation coefficient is only 0.51 (Graur 1985). It was then claimed that variation in the silent evolutionary rate can be accounted for by the differences in GC contents and that the highest rate should occur when the GC content is around 50% (Wolfe et al. 1989; Bulmer et al. 1991). However, the estimates of silent substitution rates are greatly influenced by base compositions (Figure 7.2), and this is particularly so when observed nucleotide differences are close to saturation levels. My conclusion is that silent substitution rates must be examined more carefully so as to avoid artifacts owing to failures of constructing a real.stic model of nucleotide substitutions.

Taking into account the transition bias, base composition bias, and different levels of saturation in TC and AG transitions, I have estimated the rate of substitution. Most significantly, the silent rates — the rate of synonymous substitutions and substitutions in the noncoding region — are very similar, which is about 2.37 x 10^{-8} per site per year. The ratio of transitions to transversions in the silent changes are also similar at 17. The nonsynonymous rate and the substitutions in the tRNA and rRNA genes are similar, and are one order lower than that of the silent rate. This indicate a relaxed mode of evolution of the tRNA and rRNA genes in mtDNA compared to those of nuclear DNA. The above observations suggest that mutations themselves occur more or less with the same rate and bias, except for the hypervariable region localized in the D-loop (Horai and Hayasaka 1990; Kocher and Wilson 1991).

CHAPTER NINE

CONCLUSION AND PROSPECTS

One major characteristic of the vertebrate mtDNA is that it possesses transition and base composition biases throughout the genome. Although not directly proven, biased mutation pressure seems to be a cause for these biases. Such mutation pressure, by affecting the mode of nucleotide substitutions, influences the amino acid usages of protein genes. Further comparative analyses using a variety of mtDNAs are necessary to reveal the mechanism of the biased mutation pressure. Another characteristic is frequent occurrences of insertions and deletions, which are largely responsible for the evolution in the noncoding regions. Comparison of various mtDNAs may lead to precise identification of sequences involved in insertion and deletion events, thereby providing substrates for proteins that catalyze these events. Also, with more information about the enzymology of mtDNA replication and transcription, we will be able to understand the evolution and molecular biology of mtDNA more deeply.

Mitochondrial proteins construct only a small part of the subunits in the respiratory complexes and the majority of the proteins required for the mitochondrial function are encoded in nuclear DNA. Thus, mitochondrial gene products, whether protein or RNA, must interact with those imported from cytoplasms. The way of assembly and the cooperative function of mitochondrial and nuclear subunits in each respiratory complexes need to be understood. The evolution of nuclear coded counterparts also needs to be examined. It is still far from clear how the mitochondrial and nuclear genomes have achieved their coordination in the evolutionary process.

The high rate of synonymous transitions and extreme biases in the base compositions in mtDNA raise serious problems in inferring the actual number of nucleotide substitutions. In comparison among hominoid mtDNAs, synonymous differences are saturated even between the human and chimpanzee pair (diverged around five mya). Too extensive multiple hit substitutions make it impossible to infer the actual number of substitutions from the observed number of nucleotide differences. At present, it seems inevitable to select appropriate nucleotide sites that have experienced theoretically tractable numbers of substitutions. For this purpose, tRNA genes and 12S rRNA genes may be appropriate. With additional data from other primates, the effective range of the mtDNA clock can be examined.

In the present study, the hominoid mtDNA phylogeny is determined by using only unsaturated nucleotide differences. The analysis strengthened the pattern and dating in hominoid diversification inferred from the previous analysis of 4.9 kb region in six hominoid species (among African apes, gorilla diverged first about 7.7 million years ago and then chimpanzee and human became distinct about 4.7 million years ago).

Taking in account the transition bias, base composition bias, different levels of substitutions, and assuming that gorilla diverged 7.7 mya, I have estimated the rate of various protein coding genes, tRNA genes, rRNA genes and noncoding regions. Most significant is the finding that the silent substitution rate is rather uniform over the primate mitochondrial genome and that the ratio of transversions to transitions is independent of regions. These strongly suggest that mutations themselves occur more or less with the same rate and bias, except for the hypervariable region localized in the D-loop (Horai and Hayasaka 1990; Kocher and Wilson 1991), and that the mutation rate in the mtDNA is about 20 times higher and much more biased than in the nuclear genome. The D-loop region showed species-specific divergences, and it is difficult to estimate the rate of substitutions in the same way as other coding regions. Since the D-loop region has been used as a powerful tool to address evolutionary problems within species, a further study should be conducted so as to give a reliable estimate of mutation rate in the D-loop region. The rate will be very useful for studying the evolution of modern humans during the last 200,000 years.

LITERATURES CITED

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1989) In: Molecular biology of the cell (second edition) Garland Publishing, Inc. New York & London

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457-465

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) Complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683-717

Andrews P (1986) Fossil evidence on human origins and dispersal. Cold Spring Harbor Symp Quant Biol 51:419-428

Andrews P, Cronin JE (1982) The relationships of *Sivapithecus* and *Ramapithecus* and the evolution of orangutan. Nature 297:541-546

Aquadro CF, Greenberg BD (1983) Human mitochondrial DNA variation and evolution: Analysis of nucleotide sequences from seven individuals. Genetics 103:287-312.

Asakawa S, Kumazawa Y, Araki T, Himeno H, Miura K, Watanabe K (1991) Strand specific nucleotide composition bias in Echinoderm and vertebrate mitochondrial genomes. J Mol Evol 32:511-520

Attardi G (1985) Animal mitochondrial DNA: An extreme example of genetic economy. In: Genome evolution in prokaryotes and eucaryotes, DC Reanney DC, Chambon P (eds) International Review of Cytology, vol 93, Academic Press, New York

Azad AA (1979) Intermolecular base-paired interaction between complementary sequences present near the 3' ends of 5S rRNA and 18S (16S) rRNA might be involved in the reversible association of ribosomal subunits. Nucleic Aci Res 7:1913-1929

Babcock GT, Wikstrom M (1992) Oxygen activation and the conservation of energy in cell respiration. Nature 356:301-309

Bentzen P, Leggett WC, Brown GG (1988) Length and restriction site heteroplasmy in the mitochondrial DNA of American shad (*Alosa sapidissima*). Genetics 118:509-518

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA (1981) Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167-180

Bogenhagen DF, Applegate EF, Yoza BK (1984) Identification of a promoter for transcription of the heavy strand of human mtDNA: *in vitro* transcription and deletion mutagenesis. Cell 36:1105-1113

Borst P, Flavell RA (1976) Properties of mitochondrial DNAs. In: Handbook of biochemistry and molecular biology 3rd ed., nucleic acids vol II, Fasman GD (ed) CRC Press, Cleveland Ohio, pp 363-374

Brown GG, Gadaleta G, Pepe G,Saccone C, Sbisa E (1986) Structural conservation and variation in the D-loop-containing region of vertebrate mitochondrial DNA. J Mol Biol 192:503-511

Brown GG, Simpson MV (1982) Novel features of animal mtDNA evolution as shown by sequences of rat cytochrome oxidase subunit II genes. Proc Natl Acad Sci USA 79:3246-3250

Brown WM (1981) Mechanisms of evolution in animal mitochondrial DNA, Ann N Y Acad Sci 361:119-134

Brown WM (1985) The mitochondrial genome of animals. In: Molecular evolutionary genetics. MacIntyre (ed) New York, London, Plenuim. pp 95-130

Brown WM, George M Jr, Wilson AC Rapid evolution of animal mitochondrial DNA (1979) Proc Natl Acad Sci USA 76:1967-1971

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225-239

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc Natl Acad Sci USA 88:5974-5978

Caccone A, Powell JR (1989) DNA divergence among hominoids. Evolution 43:925-942

Cann RL, Brown WM, Wilson AC (1984) Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. Genetics 106:479-499

Cantatore P, Gadaleta MN, Roberti M, Saccone C, Wilson AC (1987) Duplication and remoulding of tRNA genes during the evolutionary rearrangement of mitochondrial genomes. Nature 329:853-855

Cantatore P, Roberti M, Rainaldi G, Gadaleta MN, Saccone C (1989) The complete nucleotide sequence, gene organization, and genetic code for the mitochondrial genome of *Paracentrotus lividus*. J Biol Chem 264:10965-10975

Cantatore P, Saccone C (1987) Organization,structure, and evolution of mammalian mitochondrial genes. Int Rev Cytol 108:149-208

Capaldi RA (1990) Structure and function of cytochrome *c* oxidase. Ann Rev Biochem 59:569-596

Cavalier-Smith T (1987) The origin of eukaryotic and archaebacterial cells. Ann N Y Acad Sci 503:17-54

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis models and estimation procedures. Am J Hum Genet 19:233-257

Chang DD, Clayton DA (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. Cell 36:635-643

Chang DD, Clayton DA (1985) Priming of human mitochondrial DNA replication occurs at the light strand promoter. Proc Natl Acad Sci USA 82:351-355

Chang DD, Hauswirth WW, Clayton DA (1987) A novel endoribonuclease cleaves at a priming site of mouse mitochondrial DNA replication. EMBO J 6:409-417

Chomyn A, Mariottini P, Cleeter M, Ragan F, Matsuno-Yagi A, Hatefi Y, Doolittle R, Attardi G (1985) Six unidentified reading frames of human mitochondrial DNA encode components of the respiratory-chain NADH dehydrogenase. Nature 314: 592-597

Chou PY, Fasman GD (1978) Empirical predictions of protein conformations Ann Rev Biochem 47:251-76

Christianson TW, Clayton DA (1988) A tridecamer DNA sequence supports human mitochondrial RNA 3'-end formation in vitro. Mol Cell Biol 8:4502-4509

Clary DO, Wolstenholm DR (1985) The mitochondrial DNA molecule of *Drosophila yakuba*: Nucleotide sequence, gene organization, and genetic code. J Mol Evol 22:252-271

Clayton DA (1982) Replication of animal mitochondrial DNA. Cell 28:693-705

Clayton DA (1984) Transcription of the mammalian mitochondrial genome. Annu Rev Biochem 53:573-594

Clayton DA (1991) Replication and transcription of vertebrate mitochondrial DNA. Annu Rev Cell Biol 7:453-478

Cox CB, Fimmel AC, Gibson F, Hatch L (1986) Biochim Biophys Acta 849:52-69

Darwin C (1859) In: The origin of species by means of natural selection. John Murray, London

De Giorgi, Lanave C, Musci MD, Saccone C (1991) Mitochondrial DNA in the sea urchin *Arbacia lixula*: evolutionary inferences from nucleotide sequence analysis. Mol Biol Evol 8:515-529

De Salle R, Freeman T, Prager EM, Wilson AC (1987) Tempo and mode of sequence evolution in mitochondrial DNA of Hawaiian *Drosophila*. J Mol Evol 26:157-164

Desjardins P, Morais R (1990) Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates. J Mol Biol 212:599-634

Djian P, Green H (1989) Vectorial expansion of the involucrin gene and the relatedness of the hominoids. Proc Natl Acad Sci USA 86:8447-8451

Eisenberg D, Schwarz M, Komaromy M, Wall R (1984) Analysis of membrane and suface protein sequences with the hydrophobic moment plot. J Mol Biol 179:125-142

Fearnley IM, Walker JE (1992) Conservation of sequences of subunits of mitochondrial complex I and their relation ships with other proteins. Biochim Biophys Acta 1140:105-134

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376

Felsenstein J (1990) PHYLIP manual version 3.3. University Herbarium, University of California, Berkeley

Ferris SD, Wilson AC, and Brown WM (1981) Evolutionary tree for apes and humans based on cleavage maps of mitochondrial DNA. Proc Natl Acad Sci USA 78:2432-2436

Fisher RP, Topper JN, Clayton D A (1987) Promoter selection in human mitochondria involves binding of a transcription factor to orientation-independent upstream regulatory elements. Cell 50:247-258

Fitch WM (1986) The estimate of total nucleotide substitutions from pairwise differences is biased. Phil Trans R Soc Lond B 312:317-324

Foran DR, Hixson JE, Brown WM (1988) Comparisons of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. Nucleic Aci Res 16:5841-5861

Fox TD (1987) Natural variation in the genetic code. Annu Rev Genet 21:67-91

Friedberg EC (1985) In: DNA repair. WH Freeman and Company, New York.

Gadaleta G, Pepe G, De Candia G, Quagliariello C, Sbisá E, Saccone C (1989) The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: Cryptic signals revealed by comparative analysis between vertebrates. J Mol Evol 28:497-516

Gibbons A (1990) Our chimp cousins get that much closer. Science 250:376

Giles RE, Blanc H, Cann HM, Wallace DC (1980) Maternal inheritance of mitochondrial DNA. Proc Natl Acad Sci USA 77:6715-6719

Gonzalez IL, Sylvester JE, Smith TF, Stambolian D, Schmickel RD (1990) Ribosomal RNA gene sequences and hominoid phylogeny. Mol Biol Evol 7:203-219

Goodman M (1963) Serological analysis of the systematics of recent hominoids. Human Biol 35:377-436

Goodman M, Braunitzer G, Stangl A, Schrank B (1983) Evidence on human origin from haemoglobins of African apes. Nature 303:546-548

Goodman M, Koop BF, Czelusniak J, Fitch DHA, Tagle DA, Slightom JL (1989) Molecular phylogeny of the family of apes and humans. Genome 31:316-335

Goodman M, Tagel DA, Fitch DHA, Bailey W Czelusniak J (1990) Primate evolution at the DNA level and a classification of hominoids. J Mol Evol 30:260-266

Goto Y, Nonaka I, Horai S (1990) A mutation in the tRNA$^{Leu(UUR)}$ gene associated with the MELAS subgroup of mitochondrial encephalomyopachies. Nature 348:651-653

Gould SJ (1980) Our natural place. In: Hen's teeth and horse's toes. WW Norton and Company, Inc., New York, pp 241

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53-62

Gray MW (1989) Origin and evolution of mitochondrial DNA. Annu Rev Cell Biol 5:25-50

Greenberg BD, Newbold JE, Sugino A (1983) Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene 21:33-49

Gyllensten U, Wharton D, Josefsson A, Wilson AC (1991) Paternal inheritance of mitochondrial DNA in mice. Nature 352:255-257

Hasegawa M (1990) Phylogeny and molecular evolution in primates. Jpn J Genet 65:243-266

Hasegawa M, Kishino H (1991) DNA sequence analysis and evolution of Hominoidea. In: New aspects of the genetics of molecular evolution. Kimura M, Takahata N (eds) Springer / Verlag, Tokyo, Berlin, pp 303

Hasegawa M, Kishino H, Yano T (1985) Dating of the human - ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160-174

Hasegawa M, Kishino H, Yano T (1987) Man's place in Homineadea as inferred from molecular clocks of DNA. J Mol Evol 26:132-147

Hatafi Y (1985) The mitochondrial electron transport and oxidative phosphorylation system. Annu Rev Biochem 54:1015-1069

Hayasaka K, Gojobori T, Horai S (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5(6):626-644

Higuchi RG, Ochman H (1989) Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Aci Res 17:5865

Hixson JE, Brown WM (1986) A comparison of the small ribosomal RNA genes from the mitochondrial DNA of great apes and humans: sequence, structure, evolution, and phylogenetic implications. Mol Biol Evol 3:1-18

Hixson JE, Clayton DA (1985) Initiation of transcription from each of the two human mitochondrial promoters requires unique nucleotides at the transcriptional start sites. Proc Natl Acad Sci USA 82:2660-2664

Holmquist R, Miyamoto M, Goodman M (1988) Higher-primate phylogeny - Why can't we decide? Mol Biol Evol 5(3):201-216

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46:828-842

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man's place in homineadea revealed by mitochondrial DNA genealogy. J Mol Evol 35:32-43

Howell N (1989) Evolutionary conservation of protein regions in the protein-motive cytochrome *b* and their possible roles in redox catalysis. J Mol Evol 29:157-169

Hutchison CA III, Newbold JE, Potter SS, Edgell MH (1974) Maternal inheritance of mammalian mitochondrial DNA. Nature 251:536-538

Huxley TH (1894) In: Evolution and ethics and other essays, D Appleton and Company, New York

Hyman BC, Beck JL, Weiss KC (1988) Sequence amplification and gene rearrangement in parasitic nematode mitochondrial DNA. Genetics 120:707-712

Ikemura T (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. J Mol Biol 151:389-409

Ikemura T (1982) Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J Mol Biol 158:573-597

Irwin DM, Kocher TM, Wilson AC (1991) Evolution of cytochrome *b* gene in mammals. J Mol Evol 32:128-144

Ishida T, Yamamoto K (1987) Survey of nonhuman primates for antibodies reactive with Estein-Barr virus (EBV) antigens and susceptibility of their lymphocytes for immortalization with EBV. J. Med. Primatol. 16:359-371

Jacobs HT, Elliot D, Math VB, Farquharson A (1988) Nucleotide sequence and gene organization of sea urchin mitochondrial DNA. J Mol Biol 202:185-217

Jacobs HT, Asakawa S, Araki T, Miura K, Smith MJ, Watanabe K (1989) Conserved tRNA gene cluster in starfish mitochondrial DNA. Curr Genet 15:193-206

Jukes TH (1969) Recent advances in studies of evolutionary relationships between proteins and nucleic acids. Space Life Sci. 1:469-490

Jukes TH, Osawa S (1990) The genetic code in mitochondria and chroloplasts. Experientia 46:1117-1126

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Mammalian protein metabolism. Munro HN (ed) New York Academic Press, pp 21-132

Kawamura S, Tanabe H, Watanabe Y, Kurosaki K, Saitou N, Ueda S (1991) Evolutionary rate of immunoglobulin alpha noncoding region is greater in hominoids than in old world monkeys. Mol Biol Evol 8(6):743-752

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624-626

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequence. J Mol Evol 16:111-120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454-458

Kimura M (1983) In: The neutral theory of molecular evolution. Cambridge University Press, Cambridge

King MP, Attardi G (1989) Human cells lacking mtDNA: repopulation with exogeneous mitochondria by complementation. Science 246:500-503

King TC, Low RL (1987) Mapping of control elements in the displaacement loop region of bovine mitochondrial DNA. J Biol Chem 262:6204-6213

Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC (1989) Dynamics of mitochondrial DNA evolution in mammals: amplification and sequencing with conserved primers. Proc Natl Acad Sci USA 86:6196-6200

Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees. In: Evolution of life. Osawa S, Honjo T (eds) Springer/Verlag, Tokyo, pp 391-413

Kohne DE, Chiscon JA, Hoyer BH (1972) Evolution of primate DNA sequences. J Hum Evol 1:627-644

Kondo R, Matsuura ET, Chigusa SI (1992) Further observation of paternal transmission of *Drosophila* mitochondrial DNA by PCR selective amplification method. Genet Res Camb 59:81-84

Kondo R, Satta Y, Matsuura ET, Ishiwa H, Takahata N, and Chigusa SI (1990) Incomplete maternal transmission of mitochondrial DNA in Drosophila. Genetics 126:657-663

Koop BF, Goodman M, Xu P, Chan K, Slightom JL (1986) Primate η-globin DNA sequences and man's place among the great apes. Nature 319:234-238

Koop BF, Tagle DA, Goodman M, Slightom JL (1989) A molecular view of primate phylogeny and important systematic and evolutionary questions. Mol Biol Evol 6(6):580-612

Kruse B, Narashimham N, Attardi G (1989) Termination of transcription in human mitochondria: identification and purification of a DNA binding protein factor that promotes termination. Cell 58:391-397

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105-132

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86-93

Lewin R (1988) In: In the Age of Mankind. Smithsonian Book, Washington D.C.

Li W-H, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93-96

Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330-342

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150-74

Little J W I. Lehman R, and Kaiser AD (1967) An exonuclease induced by bacteriophage λ. I. Preparation of the crystalline enzyme. J Biol Chem 242:672

Maeda N, Wu C-I, Bliska J, Reneke J (1988) Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol 5:1-20

Maly P, Brimacombe R (1983) Refined secondary structure models for the 16S and 23S ribosomal RNA of *Escherichia coli*. Nucleic Aci Res 11:7263-7286

Margulis I (1981) In: Symbiosis in cell evolution. WH Freeman, New York

Mellars P, Stringer C (1989) In: The human revolution: behavioral and biological perspectives on the origin of modern humans. Princeton University Press, New Jersey

Miyamoto MM, Slightom JL, Goodman M (1987) Phylogenetic relations of humans and African apes from DNA sequences in the ψη-globin region. Science 238:369-373

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328-7332

Montoya J, Christianson T, Levens D, Rabinowitz M, Attardi G (1982) Identification of initiation sites heavy-strand and light-stand transcription in human mitochondrial DNA. Proc Natl Acad Sci USA 79:7195-7199

Moritz C, Brown WM (1987) Tandem duplication in animal mitochondrial DNAs: variation in incidence and gene content among lizards. Proc Natl Acad Sci USA. 84:7183-7187

Moritz C, Dowling TE, Brown WM (1987) Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Ann Rev Ecol Syst 18:269-192

Nagley P (1988) Eukaryote membrane genetics: the F_0 sector of mitochondrial ATP synthase.Trends Genet 4:46-52

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol Biol Evol 3:418-426

Nei M, Stephens JC, Saitou N (1985) Methods of computing the standard errors of branching points in an evolutionary tree and their application to molecular data for humans and apes. Mol Biol Evol 2:66-85

Nei M.(1987) In: Molecular evolutionary genetics. Columbia University Press, New York

Ojala D, Merkel C, Gelfand R, Attardi G. (1981) tRNA punctuates the reading of genetic information in human mitochondrial DNA. Nature 290:470-474

Osawa S, Muto A, Jukes TH, Ohama T (1990) Evolutionary changes in the genetic code. Proc R Soc Lond B 241:19-28

Pääbo et al. (1991) Rearrangements of mitochondrial transfer RNA genes in marsupials. J Mol Evol 33:426-430

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568-583

Pilbeam DR (1984) The descent of hominoids and hominids. Sci Am 250:60-69

Potter SS, Newbold JE, Hutchison CA III, Edgell MH (1975) Specific cleavage analysis of mammalian mitochondrial DNA. Proc Natl Acad Sci USA 72:4496-4500

Poyton RO, Duhl DMJ, Clarkson GHD (1992) Protein export from the mitochondrial matrix. Trends Cell Biol 2:369-375

Rao MJK, Argor P (1986) A conformational preference parameter to predict helices in integral membrane proteins. Biochim Biophys Acta 869(2):197-214

Roe BA, Ma D-P, Wilson RK, Wong J F-H (1985) The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. J Bio Chem 260(17):9759-9774

Ruvolo M, Disotell TR, Allard MW, Brown WM, Honeycutt RL (1991) Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. Proc Natl Acad Sci USA 88:1570-1574

Saiki RK, Gelfand DH, Stoffen S, Scharf SH, Higuchi R, Horn GT, Mullis KB, Erlich HA (1988) Primer-directed enzymatic amplification of beta-globin genomic sequences and restriction sites analysis for analysis for sickel cell anemia. Science 239:487-491

Saitou N (1991) Reconstruction of molecular phylogeny of extant hominoids from DNA sequence data. Am J Phys Anthr 84:75-85

Saitou N, Nei M (1986) The number of nucleotides required to determine the branching order of three species with special reference to the human-chimpanzee-gorilla divergence. J Mol Evol 24:189-204.

Saitou N, Nei M (1987) The neighbour-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406-425

Sanger F, Nicklen S, Coulson AR (1977) DNA sequence with chain terminating inhibitors. Proc Natl Acad Sci USA 74:5463-5467

Saraste M (1990) Structural features of cytochrome oxidase Q Rev Biophys 23:321-366

Sarich VM, Wilson AC (1967) Immunological time scale for hominoid evolution. Science 158:1200-1203

Satta Y, Takahata N (1990) Evolution of Drosophila mitochondrial DNA and the history of the melanogaster subgroup. Proc Natl Acad Sci USA 87:9558-9562

Satta Y, Takahata N, Schönbach C, Gutknecht J, Klein J (1991) Calibrating evolutionary rates at major histocompatibility complex loci. In: Molecular evolution of the majorhistocompatibility complex. Klein J, Klein D (eds) Springer-Verlag, Heidelberg, pp 51-62

Sibley CG, Ahlquist JE (1984) The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. J Mol Evol 20:2-15

Sibley CG, Comstock JA, Ahlquist JE (1990) DNA hybridization evidence of hominoid phylogeny: A reanalysis of the data. J Mol Evol 30:202-236

Silvestri G, Moraes CT, Shanske S, Oh SJ, DiMauro S (1992) A new mtDNA mutation in the tRNA^{Lys} gene associated with myoclonic epilepsy and ragged-red fibers (MERRF). Am J Hum Genet 51:1213-1217

Smith MJ, Banfield DK, Doteval K, Gorski S, Kowbel DJ (1989) Gene arrangements in the sea star mitochondial DNA demonstrates a major inversion event during echinoderm evolution. Gene 76:181-185

Smith MJ, Banfield DK, Doteval K, Gorski S, Kowbel DJ (1990) Nucleotide sequence of nine protein-coding genes and 22 tRNAs in the mitochondial DNA of the sea star *Pisaster ochraceus*. J Mol Evol 31:195-204

Snyder M, Fraser AR, LaRoche J, Gartner-Kepka KE, Zouros E (1987) Atypical mitochondrial DNA from the deep-sea scallop *Placopecten magellanicus*. Proc Natl Acad Sci USA 84:7595-7599

Stringer CB (1990) The emergence of modern humans. Sci Am 263:68-74

Takahata N (1989) Gene genealogy in three related populations: consistency probability between gene and population trees. Genetics 122:957-966

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641-657

Takahata N, Tajima F (1991) Sampling errors in phylogeny. Mol Biol Evol 8:494-502

Thomas WK, Wilson AC (1991) Evolution by base substitution in animal mitochondrial DNA. Int Rev Cytol (in press)

Ueda S, Watanabe Y, Saitou N, Omoto K, Hayashida H, Miyata T, Hisajima H, Honjo T (1989) Nucleotide sequences of immunoglobulin-epsilon pseudogenes in man and apes and their phylogenetic relationships. J Mol Biol 205:85-90

Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African population. Proc Natl Acad Sci USA 86:9350-9354

Walberg MW, Clayton DA (1981) Sequences and properties of the human KB cell and mouse L cell D-loop regions of mitochondrial DNA. Nucleic Aci Res 9:5411-5421

Widger WR, Cramer WA, Herrman RG, Trebst A (1984) Sequence homology and structural similarity between cytochrome *b* of mitochondrial complex III and the chloroplast *b6-f* complex: position of the cytochrome *b* hemes in the membrane. Proc Natl Acad Sci USA 81:674-677.

Wikstrom M, Saraste M, Penttila T (1984) In: The enzymes in biological membranes (vol 4) Martinosi A (ed) Plenum Publishing. pp 11-148

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283-285

Wolstenholm DR, MacFarlane JL, Okimoto R, Clary DO, Wahleithner JA (1987) Bizarre tRNAs inferred from DNA sequences of mitochondrial genomes of nematode worms. Proc Natl Acad Sci USA 84:1324-1328

Yang W, Zhou X (1988) rRNA genes are located far away from the D-loop region in Peking duck mitochondrial DNA. Curr Genet 13:351-355

Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Evolving Genes and Proteins. Bryson Y, Vogel HJ (ed) Academic press, New York, pp 97-166

Árnason Ú, Gullberg A, Widegren B (1991) The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenopera physalus*. J Mol Biol 33:556-568

Árnason Ú, Johnsson (1992) The complete mitochondrial DNA sequences of the Harbor Seal *Phoca vitulina*. J Mol Evol 34:493-505

APPENDIX I

CALCULATIONS

Some traditional models of nucleotide substitutions (e.g., Nei and Gojobori 1986) rely on the assumption that all four nucleotides are equally likely to mutate. However, since this does not hold for the primate mtDNA, we need to use a different method that can reflect biased substitutions. The model I use distinguishes transition (α) and transversion (β) and takes into account different base frequencies (π_j, j = A, T, G, or C) at equilibrium. I use empirical relative values of $\alpha=1$ and $\beta=1/17$. These are close to values suggested by many studies (Brown et al. 1982; Hixson and Brown 1986; Hayasaka et al. 1988; Foran et al. 1988; Horai and Hayasaka 1990). The probability that base i is replaced by base j (P_{ij}) is defined by

$$P_{AG} = \frac{\alpha\pi_G}{\alpha\pi_G+\beta\pi_C+\beta\pi_T}, \quad P_{GA} = \frac{\alpha\pi_A}{\alpha\pi_A+\beta\pi_C+\beta\pi_T},$$

$$P_{CT} = \frac{\alpha\pi_T}{\alpha\pi_T+\beta\pi_A+\beta\pi_G}, \quad P_{TC} = \frac{\alpha\pi_C}{\alpha\pi_C+\beta\pi_A+\beta\pi_G}, \text{ etc.}$$

The number of synonymous sites (s) at the third position of a two-fold degenerate codon is counted as follows: For codons TTC, TAC, CAC, AAC, GAC, TGC, ATC, and AGC, $s=P_{CT}$; for TTT, TAT, CAT, AAT, GAT, TGT, ATT, and AGT, $s=P_{TC}$; for CAA, AAA, GAA, ATA, and TGA, $s=P_{AG}$; and for codons CAG, AAG, GAG, ATG, TGG), $s=P_{GA}$. For leucine codons; $s=P_{TC}+P_{AG}$ for TTA, $P_{TC}+P_{GA}$ for TTG, and $P_{TC}+1$ for CTA and CTG. The third positions in the four-fold degenerate sites are $s=1$. The number of nonsynonymous sites (n) for a given codon is $n=3 - s$.

In the above, I note that under Jukes and Cantor, $\alpha=\beta$ and $\pi_j=1/4$ for all j so that P=1/3 for a twofold-degenerate synonymous site. Under Kimura's two-parameter model, however, $\alpha\neq\beta$ but $\pi_j=1/4$, so $P=\alpha/(\alpha+2\beta)$ for a twofold-degenerate site.

The extents of the differences by TC transitions, AG transitions, and transversions differ from each other and approach different saturation levels. It is therefore necessary to treat them separately. For a given pair of genes from two out of four species, I first compute the total numbers of TC transitions, AG transitions and transversions at the synonymous sites and divide each of them by the total number of synonymous sites ($S=\sum s$) to obtain their per-site differences. The same procedure applies to the remaining five different pairs of species. Based on the difference matrix (d_{ij}, $i<j=1, \ldots, 4$) for each of TC and AG transitions and transversions, I estimate the branch lengths (b_j, $j=1, \ldots, 5$) in Fig. A.1 so as to minimize the total branch lengths (Cavalli-Sforza and Edwards 1967). In the present case of four species, the estimates of b_j can be given by

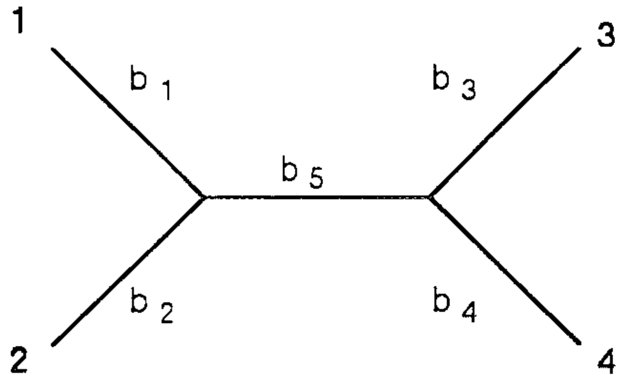
$$\begin{pmatrix} \mathrm{b}_1 \\ \mathrm{b}_2 \\ \mathrm{b}_3 \\ \mathrm{b}_4 \\ \mathrm{b}_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \\ 0 & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} \mathrm{d}_{12} \\ \mathrm{d}_{13} \\ \mathrm{d}_{14} \\ \mathrm{d}_{23} \\ \mathrm{d}_{24} \\ \mathrm{d}_{34} \end{pmatrix}.$$

These b_j values are converted into the actual number of substitutions (x_j) from the respective graph obtained from the simulation study (Figure 7.2). I then take the sum of x_j for each of the TC transitions and AG transitions and transversions to obtain the total number of synonymous substitutions per site (X_j) for branch j. The distance D_{13} between species 1 and 3 in Figure A.1, for example, is defined as $\mathrm{X}_1 + \mathrm{X}_3 + \mathrm{X}_5$. The sampling error of D_{13} is roughly given by $\sqrt{\mathrm{D}_{13}/S}$ (See Takahata and Tajima 1991 for a related argument). Similarly, the total number of synonymous substitutions in the tree, used in Figure 7.1, is estimated as $\sum_{k=1}^{5} \mathrm{X}_k$. Suppose that species 4 in Figure A.1 is gorilla and then denote by T the divergence time from the human and chimpanzee clade. I define the mean pairwise distances between gorilla and the human-chimpanzee clade as

D=(X_1 + X_2 + X_3 + 2X_5)/3 + X_4 and the rate is calculated by $D/(2T)$. The overall synonymous rate is the weighted average of *D*s for all the genes.

Figure A.1 A tree of four species.

Relationship of four species. The b_j stands for the estimated branch length by the ordinary least square method.


1
2
3
4
b 1
b 2
b 3
b 4
b 5

Journal of Molecular Evolution

© Springer-Verlag New York Inc. 1993

APPENDIX II

Evolution of Hominoid Mitochondrial DNA with Special Reference to the Silent Substitution Rate Over the Genome

Rumi Kondo, Satoshi Horai, Yoko Satta,* and Naoyuki Takahata

National Institute of Genetics, The Graduate University for Advanced Studies, 1111 Yata, Mishima 411, Japan

Summary. Focusing on the synonymous substitution rate, we carried out detailed sequence analyses of hominoid mitochondrial (mt) DNAs of ca. 5-kb length. Owing to the outnumbered transitions and strong biases in the base compositions, synonymous substitutions in mtDNA reach rapidly a rather low saturation level. The extent of the compositional biases differs from gene to gene. Such changes in base compositions, even if small, can bring about considerable variation in observed synonymous differences and may result in the region-dependent estimate of the synonymous substitution rate. We demonstrate that such a region dependency is due to a failure to take proper account of heterogeneous compositional biases from gene to gene but that the actual synonymous substitution rate is rather uniform. The synonymous substitution rate thus estimated is $2.37 \pm 0.11 \times 10^{-8}$ per site per year and comparable to the overall rate for the noncoding region. On the other hand, the rate of nonsynonymous substitutions differs considerably from gene to gene, as expected under the neutral theory of molecular evolution. The lowest rate is 0.8×10^{-9} per site per year for *COI* and the highest rate is 4.5×10^{-9} for *ATPase 8*, the degree of functional constraints (measured by the ratio of the nonsynonymous to the synonymous substitution rate) being 0.03 and 0.19, respectively. Transfer RNA (tRNA) genes also show variability in the base contents and thus in the nucleotide differences. The average rate for 11 tRNAs contained in the 5-kb region is 3.9×10^{-9} per site per year. The nucleotide substitutions in the genome suggest that the transition rate is about 17 times faster than the transversion rate.

Key words: Hominoid mitochondrial DNA — Nucleotide substitution rate — Transition bias — Base compositions — Functional constraints — Multiple hit corrections — Transfer RNA

There are a number of molecular evolutionary studies on primate mtDNAs. The regions studied include portions of nicotinamide adenine dinucleotide dehydrogenase subunit (*ND*)*4* and *ND5* genes (Brown et al. 1982; Hayasaka et al. 1988; Kocher and Wilson 1991), the complete cytochrome *c* oxydase subunit (*CO*)*II* gene (Ruvolo et al. 1991), three tRNA genes (*His, Ser, Leu*) (Brown et al. 1982; Hayasaka et al. 1988), the 12S ribosomal RNA gene (Hixon and Brown 1986), and the displacement loop (D-loop) (Foran et al. 1988; Greenberg et al. 1983; Vigilant et al. 1989; Horai and Hayasaka 1990; Kocher and Wilson 1991). Although these pieces of DNA sequence information were useful in conceiving rough pictures of the tempo and mode of primate mtDNA evolution and were applicable to the taxonomic study of primates, they apparently do not suffice to support definite conclusions.

Recently, we (Horai et al. 1992) determined the nucleotide sequences of homologous 4,938-bp portions of five hominoid mtDNAs and compared them with the known human counterpart (Anderson et al. 1981). We were mainly concerned with the pattern and dating in hominoid diversification, and used

* *Present address:* Max-Plank Institut für Biologie, Abteilung Immungenetik, Corresstrasse 42, D-7400, Tübingen, Federal Republic of Germany

Offprint requests to: S. Horai

only the unsaturated parts of sequence differences (i.e., nonsynonymous changes, synonymous-transversions, and changes in the tRNA coding regions). Because there was little problem in correcting for multiple hit substitutions and unprecedentedly long regions compared, we believe that the hominoid phylogeny then reconstructed has resolved the trichotomy (the phylogenetic relationships among human, chimpanzee, and gorilla) and that the dating of divergences is fairly accurate. In this paper, we report the results of more detailed examination of the same DNA sequence information in order to reveal evolutionary characteristics of individual genes. Our particular interest is in estimating a reliable synonymous substitution rate.

We base our analyses on the hominoid mtDNA phylogeny obtained by Horai et al. (1992), but focus on the DNA sequences sampled from relatively closely related species for which multiple hit substitutions are not particularly extensive. In doing so, we try to take proper account of the transition bias, the base composition bias, and the functional constraints which are characteristic of primate mtDNA molecules. In the following, the word *differences* is used for the *observed* changes, and the word *substitutions* is used for the *inferred* changes.

DNA Sequence Analyses

An Overview of the Compared Region

In the 4,938-bp aligned region of mtDNA sequences determined for six hominoids (common chimpanzee, pygmy chimpanzee, human, gorilla, orangutan, and siamang), there are six protein-coding genes [*ND2, COI, COII,* adenosine triphosphatase (*ATPase*)*8*, and portions of *ND1* and *ATPase 6*], 11 tRNAs for *Ile, Gln, Met, Trp, Ala, Asn, Cys, Tyr, Ser*(*UCN*), *Asp*, and *Lys* (Anderson et al. 1981; Chomyn et al. 1985), and noncoding regions. The six protein-coding genes are all transcribed from the H-strand, and the L-strand replication origin (O_L) of 32-bp length is included in the sequenced region. The total noncoding region consists of eight short spacer sequences (1–9 bp) and three larger ones (46 bp between *COII* and *tRNA*^{Lys}, 18 bp between *tRNA*^{Tyr} and *COI*, and 32 bp between *tRNA*^{Asp} and *tRNA*^{Cys}).

Among the aligned region, there are 1,438 variable sites (or 29.1%). Of these, single nucleotide differences account for 1,398 of the case. Depending on species pairs, the overall sequence differences range from 3.3% (between the two chimpanzee species) to 16.9% (between orangutan and siamang). The proportion of transitional differences decreases from 95.3% to 75.8% as the divergence time between two species increases. (See also Brown et al. 1982.) The remaining 40 variable sites result from deletions and/or insertions, and they are restricted to the noncoding regions and tRNA loops (in tRNA of *Trp, Asn, Cys, Tyr,* and *Asp*). Most of them are associated with poly-C or -A tracts, caused by single base deletions and/or insertions. Exceptions are a 2-bp deletion in orangutan *tRNA*^{Trp}, 3- and 6-bp insertions in the spacer region between tRNA^{Tyr} and *COI* in gorilla and orangutan, and a 16-bp insertion in the spacer region between *COII* and *tRNA*^{Lys} in orangutan. There are other six single-base deletions and/or insertions in the tRNA-coding regions that are found only in orangutan and siamang. In the following analyses, we treat these deletions and/or insertions separately from nucleotide substitutions, and discard the 58-bp overlapping sites: the 44-bp overlap of *ATPase 8* and *ATPase 6*, 4-bp overlaps in tRNA-coding regions, and other 10 sites which occur in incompletely determined codons in *ND1, ATPase 8,* and *ATPase 6*. The total sites compared for nucleotide substitutions are 4,008 bp in the protein-coding region, 751 bp in the tRNA-coding region, and 82 bp in the noncoding region.

Amino Acid Sequences

In contrast to a rather uniform distribution of nucleotide differences over the entire sequenced region (Fig. 1 in Horai et al. 1992), the amino acid replacement (nonsynonymous) differences are dispersed in clusters (Fig. 1). For instance, a number of amino acid differences are observed in the proximal region of ATPase 8 between orangutan and siamang, whereas only a few occur in COI and COII. Such heterogeneous amino acid differences most likely reflect different degrees of selective constraints against different genes or parts of genes (Kimura 1983). Most conserved are COI and COII and least is ATPase 8. Intermediate are ND1, ND2, and ATPase 6 (Table 1).

Subunits COI and COII contain hemes and coppers which constitute all the four redox centers of the cytochrome oxidase (complex IV). These centers are responsible for the catalytic function of the enzyme (reviewed in Capaldi 1990), so they may be most functionally important among the six proteins under study. In contrast, neither ND1 nor ND2 contains iron sulfur centers of NADH coenzyme Q reductase (complex I). ATP synthase (complex V) has the catalytic part (F_1) and an integral membrane component with proton channel (F_0). ATPase 6 and 8 participate in the F_0 part of the complex V, but their functional roles are poorly understood. How-

ND1

```
                                                    46
HUM     IRTAYPRFRYDQLMHLLWKNFLPLTLALLMWYVSMPITISSIPPQT
CHI                                S    I   T
PYG        T   LC                  S    I   T
GOR                                     I   T
ORA       QT                           HI     T G
SIA        T          Y                 I L  M A T
```

ND2

```
                                                          60
HUM     INPLAQPVIYSTIFAGTLITALSSHWFFTWVGLEMNMLAFIPVLTKKMNPRSTEAAIKYF
CHI            I     LT                           I     S
PYG            I         F  V                           S
GOR            I                    A
ORA            I  L V T             A L                TS
SIA            I             S     LA
                                                         120
HUM     LTQATASMILLMAILFNNMLSGQWTMTNTTNQYSSLMIMMAMAMKLGMAPFHFWVPEVTQ
CHI                    S S
PYG                    S                       T
GOR                    S         T  A         VV
ORA              F     H   F     TA   P       VT L
SIA      V        M    S  L      T  I         LT L
                                                         180
HUM     GTPLTSGLLLLTWQKLAPISIMYQISPSLNVSLLLTLSILSIMAGSWGGLNQTQLRKILA
CHI         M                     S    N
PYG         M                   M S    N
GOR         M             M       S T             L
ORA      V                      MY  VDTNI         LV           H
SIA       T                      F VM  NI   F      V
                                                         240
HUM     YSSITHMGWMMAVLPYNPNMTILNLTIYIILTTTAFLLLNLNSSTTTLLLSRTWNKLTWL
CHI
PYG                                       T
GOR           V                           T      S
ORA               V        I     I   T    T  I D      I
SIA           V   T        I  F      V       A              S
                                                         300
HUM     TPLIPSTLLSLGGLPPLTGFLPKWAIIEEFTKNNSLIIPTIMATITLLNLYFYLRLIYST
CHI                             V                  I
PYG                             V               T  I
GOR                             L         D  T     I
ORA     M   S                         A  DN  A     I S       A    I
SIA     L                       LV   L   GT        IV  I     M
                                            347
HUM     SITLLPMSNNVKMKWQFEHTKPTPFLPTLIALTTLLLPISPFMLMIL
CHI                                   T
PYG                                   T
GOR                     L Y           T              V
ORA                       N   A L    TI  A       LI S P
SIA         F T           NM   LL    TI        A LTFPAP
```

Fig. 1. Amino acid alignment of the six protein genes of six hominoid mtDNAs. The sequence of published human mtDNA (Anderson et al. 1981) is shown in the uppermost line with the single-letter amino acid code. For the other species only the amino acids different from those in the human sequence are shown. The overlapping region of *ATPase 8* and *ATPase 6* is excluded. Abbreviations used: HUM = human (*Homo s. sapiens*); CHI = common chimpanzee (*Pan troglodytes*); PYG = pygmy chimpanzee (*Pan paniscus*); GOR = gorilla (*Gorilla gorilla*); ORA = orangutan (*Pongo pygmaeus*); and SIA = siamang (*Hylobates syndactylus*). Continued on page 000.

COI

```
                                                                   60
HUM      MFADRWLFSTNHKDIGTLYLLFGAWAGVLGTALSLLIRAELGQPGNLLGNDHIYNVIVTA
CHI        T
PYG        T                    T
GOR        T
ORA
SIA

                                                                  120
HUM      HAFVMIFFMVMPIMIGGFGNWLVPLMIGAPDMAFPRMNNMSFWLLPPSLLLLLASAMVEA
CHI
PYG
GOR                                                      F
ORA                  M                                L  F       T
SIA                                                      F
                                                                  180
HUM      GAGTGWTVYPPLAGNYSHPGASVDLTIFSLHLAGVSSILGAINFITTIINMKPPAMTQYQ
CHI                                        I
PYG
GOR                                        I
ORA                                        I                     S
SIA                                                              S
                                                                  240
HUM      TPLFVWSVLITAVLLLLSLPVLAAGITMLLTDRNLNTTFFDPAGGGDPILYQHLFWFFGH
CHI
PYG
GOR
ORA             I
SIA

                                                                  300
HUM      PEVYILILPGFGMISHIVTYYSGKKEPFGYMGMVWAMMSIGFLGFIVWAHHMFTVGMDVD
CHI
PYG
GOR
ORA                         H                 V
SIA                         H

                                                                  360
HUM      TRAYFTSATMIIAIPTGVKVFSWLATLHGSNMKWSAAVLWALGFIFLFTVGGLTGIVLAN
CHI
PYG
GOR                                     T     M
ORA                                     T     I
SIA                                    DT

                                                                  420
HUM      SSLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFIHWFPLFSGYTLDQTYAKIHFTIMFIG
CHI                                                          Q A
PYG                                                          Q A
GOR                                                            A
ORA                                                   N        IT  V
SIA                                       V                    A   V
                                                                  480
HUM      VNLTFFPQHFLGLSGMPRRYSDYPDAYTTWNILSSVGSFISLTAVMLMIFMIWEAFASKR
CHI                                     V
PYG                                     V
GOR                        H
ORA                                         A
SIA

                                       513
HUM      KVLMVEEPSMNLEWLYGCPPPYHTFEEPVYMKS
CHI               A
PYG               A
GOR          I    T                 S
ORA        P I Q  TS                     P
SIA       I  I Q  T                      P
```

Fig. 1. Continued from page 000.

COII

```
                                                                 60
HUM    MAHAAQVGLQDATSPIMEELITFHDHALMIIFLICFLVLYALFLTLTTKLTNTNISDAQE
CHI                         I                               S
PYG                         I                               S
GOR                         I                              N
ORA                        VI
SIA                         S            S                    T
                                                                120
HUM    METVWTILPAIILVLIALPSLRILYMTDEVNDPSLTIKSIGHQWYWTYEYTDYGGLIFNS
CHI                                      F
PYG                                      F
GOR       I                         I    F
ORA       I         I           L   I    F
SIA                             L   I    F   A       A
                                                                180
HUM    YMLPPLFLEPGDLRLLDVDNRVVLPIEAPIRMMITSQDVLHSWAVPTLGLKTDAIPGRLN
CHI                                 V   V
PYG                                 V   V
GOR                                 V   V
ORA                                 V   V             T  S
SIA                    E            V             T  S
                                                   227
HUM    QTTFTATRPGVYYGQCSEICGANHSFMPIVLELIPLKIFEMGPVFTL
CHI
PYG
GOR                                                 A
ORA                                                 A
SIA
```

ATPase 8

```
                                                          53
HUM    MPQLNTTVWPTMITPMLLTLFLITQLKMLNTNYHLPPSPKPMKMKNYNKPWEP
CHI          A               V       S
PYG          A    T                  S
GOR                 A             V         L  T     FC
ORA           T L V   T  A        L  SHL P TP  FT T PHA    L
SIA                 I  S       LM   T  MY  P A   L NI PH N   H
```

ATPase 6

```
                                                                 60
HUM    LPAAVLIILFPPLLIPTSKYLINNRLITTQQWLIKLTSKQMMTMHNTKGRTWSLMLVSLI
CHI                  V    H              Q          S
PYG                  V    H              Q
GOR             L                 A      Q           A         MW
ORA        I V       V   HF              R  L    IT            T
SIA         P     S                      Q  L     L            I
                                                                120
HUM    IFIATTNLLGLLPHSFTPTTQLSMNLAMAIPLWAGTVIMGFRSKIKNALAHFLPQGTPTP
CHI       T                               A V    F T
PYG                                         V    F T
GOR                                       A TT     T      L
ORA        S      F Y                    S  A  L F A IS   L
SIA                                         AT L L T  T   L
                                  150
HUM    LIPMLVIIETISLLIQPMALAVRLTANITA
CHI
PYG         I       F
GOR                 F
ORA         I       F   L
SIA         I       F
```

Fig. 1. Continued from page 000.

Table 1. Percent similarity of amino acid sequences among hominoids[a]

Pairs compared	*ND1* (46aa)	*ND2* (347aa)	*COI* (513aa)	*COII* (227aa)	*ATP8* (53aa)	*ATP6* (150aa)
C-P	93.5	97.1	99.6	100.0	96.2	96.7
C-H	93.5	96.3	98.8	97.8	94.3	94.0
P-H	87.0	95.7	98.8	97.8	94.3	94.7
C-G	97.8	93.4	98.3	97.8	83.0	90.7
P-G	91.3	93.7	97.9	97.8	83.0	91.3
H-G	95.7	93.4	98.1	96.5	88.7	92.0
C-O	84.8	83.6	95.3	96.0	62.3	84.0
P-O	82.6	84.2	94.9	96.0	64.2	87.3
H-O	87.0	83.9	95.7	94.7	64.2	84.0
G-O	87.0	83.3	96.1	97.4	62.3	81.3
C-S	82:6	83.9	96.9	94.7	67.9	88.0
P-S	80.4	83.9	96.9	94.7	67.9	91.3
H-S	84.8	84.7	97.5	94.7	69.8	90.0
G-S	84.8	84.4	97.5	94.3	64.2	90.0
O-S	80.4	81.6	96.9	95.2	60.4	86.0

[a] Number of amino acids for each gene is shown in parenthesis. Abbreviations for the species are C (common chimpanzee), P (pygmy chimpanzee), H (human), G (gorilla), O (orangutan), S (siamang)

ever, the hydrophobicity of ATPase 8 (SOAP profile: Kyte and Doolittle 1982) is relatively changeable among hominoids. It is therefore possible that required amino acids at particular sites can be flexible in their biochemical properties and thus the degree of selective constraints is rather low.

It is notable that the termination codons can vary in mtDNA genes (e.g., Gadaleta et al. 1989). In spite of a rather strict constraint on nonsynonymous changes in the *COI* and *COII* genes, some anomalous variation occurs in their termination codons. The *COI* termination codon is AGA in human, chimpanzees, and siamang, but the same codon position is substituted by AAA (Lys) in gorilla and GAG (Asp) in orangutan (Fig. 1 in Horai et al. 1992). Because AAA and GAG are often used in the coding regions (Table 2 in Horai et al. 1992), acquisition of such new termination codons in gorilla and orangutan is unlikely. An alternative explanation is that a termination codon AGG occurring in down-stream is used. This down-stream termination codon would cause a 10-bp overlap between *COI* and *tRNA*$^{Ser(UCN)}$ genes. However, this overlap would not give rise to any serious effect on transcription because these two genes are coded on the opposite strands. Furthermore, amino acid replacements in *COI* are more frequent in the 3′ end than in other regions (Fig. 1), indicating weak functional constraints in this region. Thus elongation of the C-terminal of *COI* may not be functionally defective. Similar instances of *COI* have also been noted in bovine and mouse (Anderson et al. 1982; Bibb et al. 1981). It therefore seems to be a general phenomenon that the 3′ end of *COI* is flexible and the gene has been using different termination codons in different positions. Also, the *COII* termination codon is variable among primates although the position is fixed: TAG in chimpanzees and human, TAA in gorilla, orangutan, and siamang.

Nucleotide Differences in Protein Genes

The orangutan mtDNA is too distantly related from that of chimpanzees and human; about 34% of the third codon positions differ among these species, which is close to the saturation level (Table 1 in Horai et al. 1992). Therefore, including the orangutan mtDNA in the following analysis gives rise to a difficult and commonly recognized problem in estimating accurate nucleotide substitution rates. The gorilla mtDNA also differs substantially from the human and chimpanzee mtDNAs. However, if we exclude it from the analysis, the sampling errors become too large. Actually, the observed differences at the third codon positions are about 27%, which seems not too extensive to make reasonable multiple hit corrections. For these reasons, we use the four DNA sequences sampled from common and pygmy chimpanzees, human, and gorilla. We did not analyze the *ND1* because it represents only a small portion (14.4%) of the entire gene and is subject to large sampling errors.

For a given gene, we first compare the nucleotide differences at each of three codon positions. Since most differences at the first and second codon positions are nonsynonymous and those at the third positions are synonymous, comparison of the differences between these two different codon positions gives some idea about the relative rates of synonymous and nonsynonymous differences. For the five protein coding genes, we depict Fig. 2 to

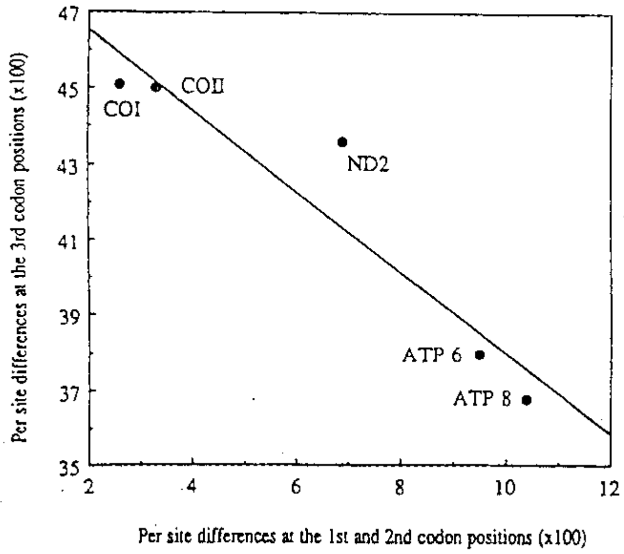


Fig. 2. Negative correlation in the observed number of differences between the first and second codon positions and the third codon positions. The number of differences for each gene was estimated from the total branch length of a tree of common chimpanzee, pygmy chimpanzee, human, and gorilla obtained by the ordinary least squares. (See Appendix). The number of differences at the first and second codon positions and the third codon positions of each gene is divided by the corresponding number of sites, and the 100-fold value (differences per 100 sites) is shown.

show the relationship of the nucleotide differences at the first and second vs those at the third codon positions. It is clear that there is an apparent negative correlation ($r = -0.94$, $t = -4.8$, $P = 0.009$, $df = 3$): The larger the differences at the first and second codon positions, the smaller the differences at the third codon positions. The extents of the nucleotide differences at the first and second codon positions are in agreement with those from the amino acid differences and are most likely related to the different degrees of selective constraints against different genes. The neutral theory (Kimura 1968, 1983) provides a reasonable account for the relationship between selective constraints and functional importance.

Curiously and importantly, however, the nucleotide differences at the third codon positions also differ from gene to gene. Some of those differences are synonymous and the differences may not be entirely attributed to differential selective constraints. Nonetheless, genes such as *COI* and *COII* under stronger functional constraints exhibit relatively large numbers of differences, while *ATPase 8* under least functional constraints exhibits a relatively small number of differences. As a result, the figure shows a negative correlation.

Analysis of codon usages among those genes shows a positive correlation ($r = 0.97$, $t = 6.9$, $P = 0.003$, $df = 3$) between the extent of nonsynonymous differences and the frequencies of A or C at the first and third codon positions (data not shown). Namely, genes exhibiting more nonsynonymous differences use more codons starting and ending with A or C. These codons generally encode hydrophobic amino acids, such as *Ile*, *Met*, and *Thr*, and may possibly be exchangeable without serious functional disruption. In contrast, *COI* and *COII* frequently use amino acids such as *Gly* and *Val* whose codons begin with G. However, these differences in codon usages alone cannot account for the variation in the nucleotide differences at the third codon positions.

However, there is a substantial heterogeneity from gene to gene in terms of the base compositions at the third codon positions. Together with extremely high transition rates, this heterogeneity suggests that the dynamics of the nucleotide substitution process at the third codon positions or at the synonymous sites may be different from gene to gene. It is also important to realize that, because of these biases, the saturation level can differ among genes.

Simulation Study

None of the nucleotide substitution models proposed thus far are particularly good at describing empirical mtDNA data (Fitch 1986). The method of Jukes and Cantor (1969) is inappropriate, because it assumes an equal transition-transversion probability and equal frequency among the four nucleotides. For the same reason, Kimura's two-parameter method (1980) may be unsuited, because it ultimately leads to the equal frequency of the four bases despite the assumed different rates between transitions and transversions. Models that consider both different transition-transversion rates and base compositional biases must be used, and some such are developed by Felsenstein (1981), Kimura (1981), Takahata and Kimura (1981), Lanave et al. (1984), and Hasegawa et al. (1985). Under some of these models satisfying reversibility, biased base compositions in a sequence can be kept constant through the nucleotide substitution process. For simplicity, we used the model of Hasegawa et al. (1985) in which the probability of base i to be replaced by base j is defined by the product of stationary base composition (π_j, j = A, G, C, or T) and the relative transition (α) or transversion rate (β) (For details of the model, see Appendix A.) The parameters α and β affect the initial increase of nucleotide differences and the time required to reach the saturation level (A and B in Fig. 3). On the other hand, the equilibrium frequencies π_j are related to the ultimate saturation level, which is given by $2(\pi_A\pi_G + \pi_C\pi_T)$ for transitions and $2(\pi_A + \pi_G)(\pi_C + \pi_T)$ for transversions (B and C in Fig. 3).

Since most substitutions at the third codon posi-

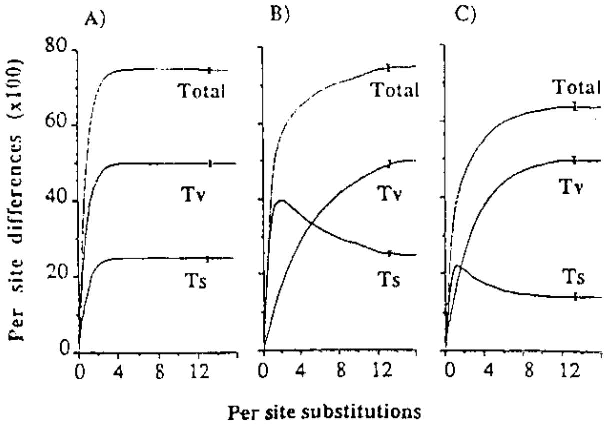


Fig. 3. Simulation study on the effect of π_j, α, and β on the saturation level, in which π_j is the frequency of base j(j = A, T, G, or C) and α and β are relative transition and transversion rates. The substitution matrix for base i to be replaced by base j is defined as follows (Hasegawa et al. 1985):

$$\begin{array}{c|cccc}
i \backslash j & T & C & A & G \\
\hline
T & 1-(\alpha\pi_C+\beta\pi_A+\beta\pi_G) & \alpha\pi_C & \beta\pi_A & \beta\pi_G \\
C & \alpha\pi_T & 1-(\alpha\pi_T+\beta\pi_A+\beta\pi_G) & \beta\pi_A & \beta\pi_G \\
A & \beta\pi_T & \beta\pi_C & 1-(\alpha\pi_G+\beta\pi_T+\beta\pi_C) & \alpha\pi_G \\
G & \beta\pi_T & \beta\pi_C & \alpha\pi_A & 1-(\alpha\pi_A+\beta\pi_T+\beta\pi_C)
\end{array}$$

In Fig. 3, we set the values of parameters as $\alpha = \beta = 1$, $\pi_T = \pi_C = \pi_A = \pi_G = 0.25$ for A), $\alpha = 1$, $\beta = 1/17$, $\pi_T = \pi_C = \pi_A = \pi_G = 0.25$ for B) and $\alpha = 1$, $\beta = 1/17$, $\pi_T = 0.15$, $\pi_C = 0.34$, $\pi_A = 0.47$, $\pi_G = 0.04$ for C). An ancestral sequence of 1 kb is assumed to have a specified base composition in A, B, or C. According to each matrix, we generated a uniform random number U for each nucleotide site. If U is smaller than a specified probability in the matrix, the site is changed. This process is repeated over the entire sequence and stored for the next round of substitutions. The observed number of differences is counted at given numbers of the total substitutions. For each parameter set, we repeated 1,000 times and computed the average and standard deviation. For the transitions (Ts), the transversions (Tv), and the total changes, the observed number of differences were plotted against the actual number of total substitutions. Graphs A and B show that the ultimate saturation level is the same even for different values of α and β, as theoretically expected. Graphs B and C show that the saturation level of the observed differences changes by the base compositions.

tions are presumably neutral, the base compositions and the transition–transversion rates should reflect the actual mutation process in each gene. For a given gene, the base compositions at the third codon positions are rather similar among chimpanzees, human, and gorilla. Importantly, however, there exists a strong bias toward A and C. TG content is extremely low; how low differs between genes (Table 2). To see the effects of such compositional biases on the nucleotide substitutions, we carried out simulations. For each gene, we used the observed base compositions as equilibrium frequency π_j but assumed the same rate, irrespective of genes, for transitions and transversions ($\alpha = 1$, $\beta = 1/17$). Figure 4 shows that even a small change (<5%) in the lower base compositions results in a substantial change in the transition differences. For example, genes with lower T content show much lower levels of the CT transition differences, and genes with lower G content show much lower levels of the AG transition differences. This suggests that variation in the base compositions can be an important factor for determining the synonymous differences. Consequently, synonymous substitutions may be inaccurately estimated if we ignore biased transition rates and base compositions. Below, we examine in more detail this possibility to explain the negative correlation shown in Fig. 2.

Synonymous and Nonsynonymous Substitution Rates

To deal with the synonymous and nonsynonymous substitutions, one must first determine the numbers of such sites per gene. A usual method for counting the number of synonymous sites per gene assumes equal substitution rates among four nucleotides (e.g., Nei and Gojobori 1986), so the number of synonymous sites for a twofold-degenerate site is assigned as one-third. However, this assignment is not valid when transition rates are much faster than transversion rates as in primate mtDNAs. Because there is only a small probability of transversions

Table 2. Base compositions at the third codon positions and noncoding region[a]

	Genes (no. sites)						
	ND1	*ND2*	*COI*	*COII*	*ATPase 8*	*ATPase 6*	Noncoding
	46	347	513	227	53	150	50
A (%)	37	38	36	35	47	39	30
C (%)	43	42	40	40	34	36	39
T (%)	16	16	19	20	15	20	23
G (%)	4	4	5	5	4	5	8
TG (%)	20	20	24	25	19	25	31

[a] Base compositions shown above are the average of common chimpanzee, pygmy chimpanzee, human, and gorilla

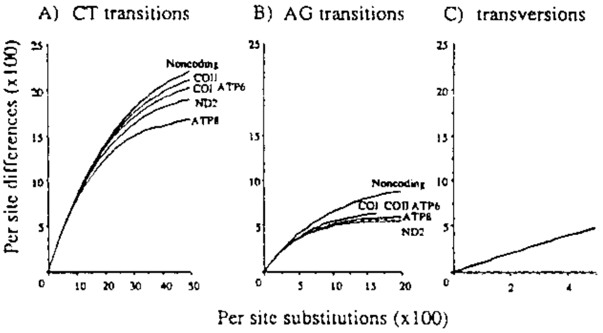


Fig. 4. Simulation study on the relationship between the actual number of substitutions and the observed number of differences. The values of π_j (j = A, T, G, or C) are set as to imitate the base compositions in actual sequences (Table 2) and the relative rates are set as $\alpha = 1$ and $\beta = 1/17$. (See Fig. 3 for details.) Graph **A** is for TC transitions, **B** is for AG transitions, and is **C** for transversions. The actual length of the abscissa in A–C indicates the same length of time. Note that there is a large difference in the saturation level between CT and AG transitions because of lower AG contents.

that result in nonsynonymous changes, the number of synonymous sites for a twofold-degenerate site should be closer to one than to one-third. Clearly, this increases the number of synonymous sites and decreases that of nonsynonymous sites per gene compared to those expected under Jukes and Cantor's model. We claim that the number of synonymous sites must be counted consistently with the model of nucleotide substitutions actually used for multiple-hit corrections. More specifically, Nei and Gojobori's method cannot be used for any substitution model other than Jukes and Cantor's and is certainly inappropriate for mammalian mtDNA.

In order to study quantitatively different ways of counting synonymous sites (nonsynonymous sites = the total number of sites − synonymous site), we compared the correlation of the synonymous and nonsynonymous differences per site. We estimated the number of synonymous sites per gene by Jukes and Cantor's model and Hasegawa et al's. model (Table 3). The number of synonymous sites estimated by Jukes and Cantor's model is approximately 73% with respect to Hasegawa et al.'s model. Accordingly, the nucleotide differences per synonymous site are overestimated by 20% and those per nonsynonymous site are underestimated. Therefore, the relationship between the synonymous and nonsynonymous differences based on Jukes and Cantor's model (indicated by a solid circle) shows a stronger negative correlation than that based on Hasegawa et al's. model (indicated by an open circle).

In general, the number of synonymous sites per twofold-degenerate sites must be determined at least in terms of α, β, and π_j (j = A, T, G, or C). This caution may be applied to many other methods for correcting multiple hit substitutions (e.g., Brown et al. 1982; Li et al. 1985). Unequal rates and base compositions in nucleotide substitutions are essential in calculating the number of synonymous or nonsynonymous sites in mammalian mtDNA.

In order to evaluate the accuracy of correction methods, we applied Kimura's two-parameter model (Kimura 1980) and Hasegawa et al.'s model to the synonymous differences (Fig. 5B). Correction with Kimura's two-parameter model (indicated by a solid circle in Fig. 5B) considerably inflated the inferred number of synonymous substitutions, and the negative correlation becomes even stronger than that in Fig. 5A. On the other hand, the correction based on Hasegawa et al.'s model (see Fig. 4 and Appendix A) gives similar estimates of the synonymous substitutions for all the genes (indicated by an open circle in Fig. 5B). As mentioned, the latter estimates depend on α and β. In the case of *ATPase 8*, which shows the lowest saturation level in Fig. 4A, the estimate of synonymous substitutions is substantially influenced by the ratio of transition to transversion. We chose relative values of α and β so as to erase the negative correlation. They turned out to be $\alpha = 1$ and $\beta = 1/17$, which are very close to the maximum likelihood estimates obtained by the DNAML in PHYLIP (Felsenstein 1990). It is reasonable to expect that the genes on the mtDNA genome have more or less the same mutation rate

Table 3. Observed number of synonymous changes and differences in the noncoding region[a]

	(Genes: no. sites)													
	ND1 44.2		*ND2* 341.7		*COI* 507.0		*COII* 224.5		*ATPase 8* 50.0		*ATPase 6* 154.2		Noncoding region 50.0	
	AG. TC/V	(%)	AG. TC/V	(%)	AG. TC/V	(%)	AG, TC/V	(%)	AG, TC/V	(%)	AG, TC/V	(%)	AG, TC/V	(%)
C-P	1,1/0	(4.5)	7,25/3	(10.2)	25,30/3	(11.4)	5,11/1	(7.6)	0,3/1	(8)	2,8/0	(6.5)	2,0/0	(4)
C-H	2.4/0	(13.6)	26,57/5	(25.8)	38,84/5	(25.1)	14,43/3	(26.7)	2,9/1	(24)	9,24/0	(21.4)	0,7/0	(14)
P-H	1.5/0	(13.6)	25.50/6	(23.7)	34,96/6	(26.8)	13,41/8	(24.9)	2.10/0	(24)	7,20/0	(17.5)	2,7/0	(18)
C-G	6.6.5/1	(30.5)	20,60/14	(27.5)	37,92/15	(28.4)	15,47/7	(30.7)	2,4/2.5	(17)	12,33.5/2	(30.8)	2,10/1	(26)
P-G	5,5.5/1	(26.0)	21,59/15	(27.8)	30.87.5/15.5	(26.2)	12,45/6	(28.1)	2,5/1.5	(17)	12,33.5/2	(30.8)	2,10/1	(26)
H-G	4.9/1	(31.7)	22,74/13	(31.9)	33,106.5/15.5	(30.6)	19,48/8	(33.1)	2,10/1.5	(27)	13,27.5/2	(27.6)	2,13/1	(28)

Observed number of nonsynonymous changes

	(Genes: no. sites and percent)											
	ND1		*ND2*		*COI*		*COII*		*ATPase 8*		*ATPase 6*	
	93.8	(%)	699.3	(%)	1,032.0	(%)	456.5	(%)	109.0	(%)	295.8	(%)
C-P	3	(3.2)	10	(1.4)	2	(0.2)	0	(0)	2	(1.8)	5	(1.7)
C-H	3	(3.2)	13	(1.9)	7	(0.7)	5	(1.1)	3	(2.8)	9	(3.0)
P-H	6	(6.4)	15	(2.2)	7	(0.7)	5	(1.1)	3	(2.8)	8	(2.7)
C-G	1.5	(1.6)	25	(3.6)	10	(1.0)	5	(1.1)	10.5	(9.6)	15.5	(5.2)
P-G	4.5	(4.8)	25	(3.6)	13	(1.3)	5	(1.1)	10.5	(9.6)	14.5	(4.9)
H-G	2	(2.1)	26	(3.7)	12	(1.2)	8	(1.8)	7.5	(6.9)	12.5	(4.2)

[a] The human sequence is taken from (Anderson et al. 1981). The number of synonymous and nonsynonymous sites are the average of the four species, estimated by assuming that the relative rates of transition and transversion are 1 and 1/17, respectively. (See Appendix A.) The observed number of substitutions and its percent in each gene or region (in parenthesis) are shown. For the synonymous changes and the differences in the noncoding region, the numbers of substitutions are shown in the categories of AG (AG transitions), TC (TC transitions), and V (transversions). Abbreviations: C, common chimpanzee; P, pygmy chimpanzee; H, human; G, gorilla

and therefore similar levels of synonymous substitutions. Although the model we used here may still be far from reality, it is clear from Fig. 5 that use of models with simplified or incorrect assumptions leads to inaccurate estimates.

We have concluded that the number of synonymous substitutions per site (distances) is fairly uniform over the genes. This supports the assumption of equal overall synonymous rates for all mitochondrial genes. We take the weighted average of the synonymous substitutions over the genes when compared between gorilla and the remaining three species: common chimpanzee, pygmy chimpanzee, and human (Appendix A). With the divergence time of gorilla at 7.7 million years ago which was estimated by Horai et al. (1992), the synonymous rate for the hominoid mtDNA becomes $2.37 \pm 0.11 \times 10^{-8}$ per site per year. This rate is five to 10 times faster than that of nuclear DNAs (Brown et al. 1982), or even faster (20 times) if the synonymous substitution rate for nuclear genes is 1.2×10^{-9} (Satta et al. 1991).

For the nonsynonymous rate, any multiple hit correction is practically unnecessary among the four species. The rate of nonsynonymous substitutions for each gene is at least five times lower than that of the synonymous change. The nonsynonymous substitution rate per site year ranges from 0.8×10^{-9} for *COI* to 4.5×10^{-9} for *ATPase 8*.

Comparison of the tRNA Coding Region

The location of variable sites and base mispairs in the secondary structures of the 11 tRNAs shows that the number of substitutions differs substantially among various tRNA domains, stems (helical regions), and loops (Fig. 6). Although the nucleotide substitutions are found in all the domains, stems are more conserved than loops and gaps. The most conserved domain is the anticodon loop which contains an anticodon of each tRNA. On the other hand, the most variable domain is the TΨC loop and the second most is the dihydrouridine loop. As noted long ago by many authors (e.g., Jukes 1969; Kimura 1983), the stems accumulate compensatory substitutions which restore Watson-Crick base pairings. Compensatory substitutions are found in the acceptor stems of the orangutan $tRNA^{Lys}$ and $tRNA^{Asp}$ and siamang $tRNA^{Gln}$ and also in the anticodon stem of human $tRNA^{Asn}$ and the TΨC stem of human $tRNA^{Asp}$.

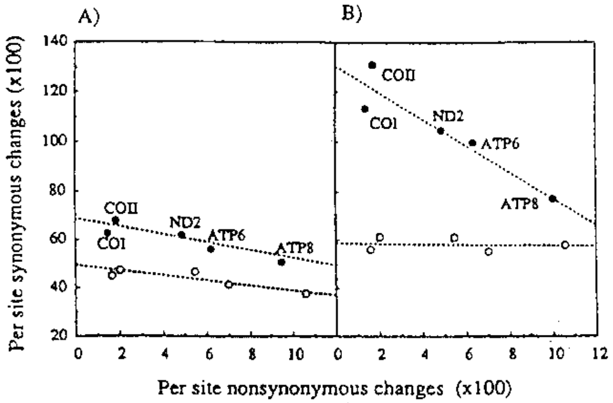


Fig. 5. Observed synonymous and nonsynonymous differences (A) and inferred synonymous and nonsynonymous substitutions (B). The inferred number of synonymous and nonsynonymous sites is calculated by Jukes and Cantor's model, which assumes equal probability for all nucleotide substitutions (indicated by a *solid circle*), and by the model of Hasagawa et al., which considers both biases in base composition and transition rate (indicated by the *open circle*). (See Appendix A). In A, the number of observed synonymous and nonsynonymous differences in each gene is calculated as in Fig. 2 by using the observed number of pairwise differences shown in Table 3. For the *solid circle* in B, the number of synonymous and nonsynonymous substitutions is estimated based on Kimura's two-parameter model (Kimura 1980), and it was converted into the per-site number of changes using the number of sites inferred from the Jukes and Cantor model. This is an example of inconsistent use of different substitution models in computing the number of synonymous sites and correcting multiple hit substitutions. For the *open circle* in B, the number of synonymous and nonsynonymous substitutions is estimated based on Fig. 4 (see Appendix A) and the number of sites inferred from the same Hasegawa et al. model was used.

The proportion of variable sites in the aligned tRNA genes shows a strong heterogeneity in the nucleotide substitutions per site over the 11 tRNAs: The proportion ranges from 6.2% (in $tRNA^{Met}$) to 26.9% (in $tRNA^{Lys}$) (Table 4). The low number of substitutions in $tRNA^{Met}$ is due probably to its important role in the transcription initiation. It is known from $tRNA^{His}$, $tRNA^{Ser(AGY)}$, and $tRNA^{Leu(CUN)}$ that the number of substitutions in tRNA varies inversely with the frequency of the corresponding codons used in the mtDNA (Brown et al. 1982). We examined this rule for the present 10 tRNA genes (excluding $tRNA^{Met}$) and found that there is indeed a negative correlation between codon usages and nucleotide substitutions ($r = -0.70$, $t = -2.8$, $P = 0.012$, $df = 8$). Because mitochondrial tRNA involves in the regulation of transcription, codon usage pattern appears to determine to some extent the rate of tRNA evolution.

We compared base compositions of tRNAs in terms of the AC content in the sense strand (Table 4), excluding all the anticodons. Since GC or AT contents in the stem parts must be 50% for Watson-Crick pairings, deviation of AC contents from 50% indicates that actual compositional biases are much stronger in other parts of tRNAs such as the loops or mispairs in the stems. The AC content is higher in the L-strand than in the H-strand (Borst and Flavell 1976; Brown 1981). This holds true whether tRNAs are coded by the L-strand or by the H-strand. The fact is that the H-strand-coded tRNAs have lower AC contents (34.8–48%) than those L-strand coded (52.5–59.5%), though the reason is unclear.

The observed sequence differences in the 11 tRNA genes do not show any evidence for saturation in the comparison between the six hominoid species (Table 1 in Horai et al. 1992). As mentioned in our previous paper (Horai et al. 1992), however, the orangutan tRNAs have many anomalous features. We therefore used the tRNA sequences for a pair of human and gorilla to calibrate the substitution rate. The average transition and transversion rates over the 11 tRNA genes are 3.7×10^{-9} and 0.2×10^{-9} per site per year, respectively. The ratio of the transversion to transition rate is $0.2/3.7 = 0.054$, which is very close to the value of $\beta/\alpha = 0.06$ used for the protein coding regions. As seen later, this holds true for the noncoding region as well. The overall rate (the sum of the transition and transversion rates) becomes 3.9×10^{-9} per site per year. The overall rate is about one-half the previous estimate (8.5×10^{-9} per site per year) for $tRNA^{His}$, $tRNA^{Ser(AGY)}$, and $tRNA^{Leu(CUN)}$ (Brown et al. 1982). Nevertheless, it is true that tRNAs in mtDNA have evolved much faster than in nuclear DNA, partly because of higher mutation rates and partly because of the less selective constraints against mitochondrial tRNAs.

Noncoding Region

The noncoding region is an assembly of small sequences dispersed between genes which contain 114 bases in total. Among the six hominoids, there are 31 variable sites due to single nucleotide changes (or 27%). Deletions and/or insertions occur at 32 sites, the frequency being as high as 28%. Thus, although single base substitutions in the noncoding region are frequent, deletions and/or insertions are also much more common than in any other regions. An exception is a small region surrounding the O_L where the 11-bp sequences are identical among all the sequences. Also conserved are the tandem T repeats of six to seven nucleotides located in the 12-bp (13 bp in orangutan) proposed loop sequences. (See Fig. 1 in Horai et al. 1992.) We excluded these two extremely conserved regions (32 bp) in estimating the substitution rate in the noncoding region.

The observed number of differences at the remaining 50 noncoding sites is given in Table 3. The

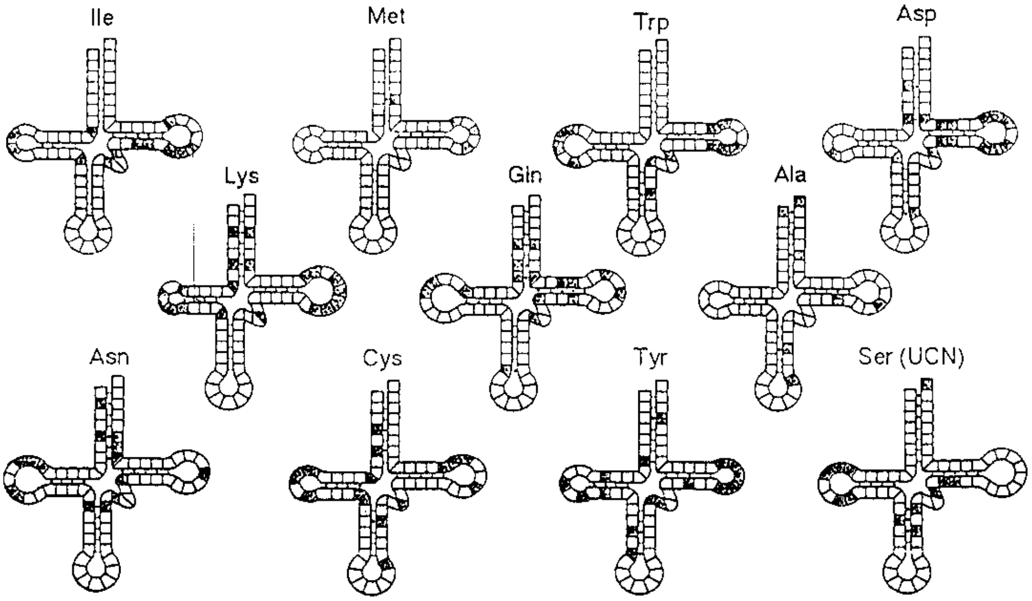


Fig. 6. Secondary structure of 11 mitochondrial tRNAs of hominoids. Each tRNA consists of 14 structurally distinct regions: 5′-acceptor (A) stem–gap–dihydrouridine (D) stem–D loop–D stem–gap–anticodon (AC stem–AC loop–AC stem–variable loop–TΨC stem–TΨC loop–TΨC stem–A stem–3′. Among the hominoid tRNAs, size variation occur only at the D loop, variable loop, and TΨC loop. *Hatched areas* indicate positions at which base substitutions have taken place. *A solid line* between nucleotides (*boxes*) in stem parts indicates a Watson-Crick pair in all the species; a *dot* indicates a non-Watson-Crick pair in some species.

Table 4. Comparison of 11 transfer RNA genes in six hominoid mtDNAs

tRNA	Aligned size[a] (sense strand)	No. variable sites	Proportion variable sites (%)	AC content[b] (%)	tRNA usage[c] (%)
Ile	63 (L)	12	19.0	52.5	8.5
Met	65 (L)	4	6.2	57.1	5.5
Trp	63 (L)	12	19.0	58.1	2.7
Asp	65 (L)	17	26.2	56.3	1.7
Lys	67 (L)	18	26.9	59.5	2.5
Gln	66 (H)	16	24.2	64.7 (35.3)	2.4
Ala	66 (H)	7	10.6	64.4 (35.6)	6.7
Asn	70 (H)	15	21.4	61.8 (38.2)	4.3
Cys	62 (H)	15	24.2	53.7 (46.3)	0.6
Tyr	62 (H)	16	25.8	52.0 (48.0)	3.6
Ser(UCN)	69 (H)	12	17.4	65.2 (34.8)	5.8

[a] The sites for anticodon and overlapping regions were excluded
[b] Calculated as an average AC content in L-strand (H-strand) of the six hominoid mtDNA, excluding the anticodons
[c] Calculated from codon usage in human mtDNA

base composition is biased in much the same way as in the third codon positions (Table 2). Moreover, the ratio of the observed transversion-to-transition differences is $1/12$ = 0.08 between gorilla and chimpanzees and $1/15$ = 0.067 between gorilla and human (Table 3). Thus, in terms of the actual number of substitutions the ratio becomes closer to $1/17$ = 0.06. Using the actual base contents in the noncoding region and the relative transition ($\alpha = 1$) and transversion rate ($\beta = 1/17$), we carried out a simulation study to make corrections for multiple hit substitutions (Fig. 4). The estimated substitution rate is $2.68 \pm 0.59 \times 10^{-8}$ per site per year. Although the sampling error is large due to the short DNA sequence, the rate is very similar to the average rate of synonymous substitutions. This supports our contention that the mutation rate in mtDNA is uniform over the genome.

Discussion

Base composition biases, which are manifest in almost the entire genome, are quite common in animal mtDNAs. The biases in the noncoding region and at the synonymous sites suggest that they are

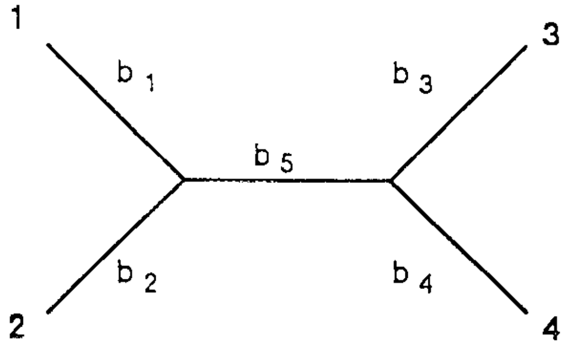


Fig. 7. Relationship of four species. The b_j stands for the estimated branch length by the ordinary least-square method.

not a result of natural selection. A more likely explanation would be preferential codon usages or biased mutation pressure. There is only one tRNA for each codon group in animal mtDNAs, but the codons that match the tRNAs with Watson-Crick pairings are not always preferred (Horai et al. 1992). Rather, codons ending with both A and C are most often used in the four-codon families. In addition, the AC bias in the L-strand is a general feature. Hence, preferential codon usages alone cannot fully explain the base compositional biases.

It is likely that the evolution of mtDNA is influenced by directional mutation pressure, as in low GC gram-positive eubacteria (Osawa et al. 1990). One possible molecular mechanism for directional mutation pressure may be preferential C to T mutations due to cytosine deamination in the single-stranded DNA in replication (Brown and Simpson 1982). When replicating, the H- and L-strands experience the single-stranded state for different periods of time. It was suggested that nucleotide changes tend to decrease C and increase T in the H-strand on one hand and to decrease G and increase A in the L-strand on the other, and that the base bias should be stronger in the regions that have a longer single-stranded state. Were this the case, the extent of base bias would increase according to the gene order of *COI*, *COII*, *ATPase 8*, *ATPase 6*, *ND1*, and *ND2*. This is not what is actually observed, however: There is no such a conspicuous trend and the bias is strongest in *ATPase 8*. Thus, the cytosine deamination alone cannot explain the base composition biases. Whatever the molecular mechanisms are, directional mutation pressure appears to operate in animal mtDNAs.

In respect of the rates of synonymous and nonsynonymous substitutions, Miyata et al. (1980) examined various eukaryotic genes and concluded that while the rate of nonsynonymous substitutions differs among genes, the rate of synonymous substitution is more or less the same for all genes. However, subsequent analyses suggested that the rate not only at synonymous sites but also in noncoding regions varies considerably in different parts of the genome (Koop et al. 1986, 1989; Li and Tanimura 1987; Li et al. 1987; Maeda et al. 1988; Kawamura et al. 1991). Also suggested was a positive correlation between synonymous and nonsynonymous substitution rates, although the correlation coefficient is only 0.51 (Graur 1985). It was then claimed that variation in the silent evolutionary rate can be accounted for by the differences in GC contents and that the highest rate should occur when the GC content is around 50% (Wolfe et al. 1989; Bulmer et al. 1991). However, the estimates of silent substitution rates are greatly influenced by base compositions (Fig. 3), and this is particularly so when observed nucleotide differences are close to saturation levels. Our conclusion is that silent substitution rates must be examined more carefully so as to avoid artifacts owing to failures in constructing a realistic model of nucleotide substitutions.

The high rate of synonymous transitions and extreme biases in the base compositions in mtDNA raise serious problems in inferring the actual number of nucleotide substitutions. Too-extensive multiple hit substitutions make it impossible to infer the actual number from the observed number of nucleotide differences. At present, it seems inevitable that one must select appropriate nucleotide sites that have experienced theoretically tractable numbers of substitutions. In this study, we compared various protein coding genes, tRNA genes, and noncoding regions among closely related species. Most significant is the finding that the silent substitution rate is rather uniform over the primate mitochondrial genome and that the ratio of transversions to transitions is independent of regions. These strongly suggest that mutations themselves occur more or less with the same rate and bias except for the hypervariable region localized in the D-loop (Horai and Hayasaka 1990; Kocher and Wilson 1991). The mutation rate is the mtDNA is about 20 times higher and much more biased than in the nuclear genome.

Acknowledgments. This work was supported by Grants-in-aids for General Scientific Research and for Scientific Research on the Priority Area to S. Horai from the Ministry of Education, Science and Culture, Japan.

References

Anderson S, Bankier AT, Barrell BG, de Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. Nature 290:457–465

Anderson S, de Bruijn MHL, Coulson AR, Eperon IC, Sanger F, Young IG (1982) Complete sequence of bovine mitochondrial DNA: Conserved features of the mammalian mitochondrial genome. J Mol Biol 156:683–717

Bibb MJ, Van Etten RA, Wright CT, Walberg MW, Clayton DA (1981) Sequence and gene organization of mouse mitochondrial DNA. Cell 26:167–180

Borst P, Flavell RA (1976) Properties of mitochondrial DNAs. In: Fasman GD (ed) Handbook of biochemistry and molecular biology 3rd ed, nucleic acids vol II, CRC Press, Cleveland, Ohio, pp 363–374

Brown GG, Simpson MV (1982) Novel features of animal mtDNA evolution as shown by sequences of rat cytochrome oxidase subunit II genes. Proc Natl Acad Sci USA 79:3246–3250

Brown WM (1981) Mechanisms of evolution in animal mitochondrial DNA, Ann NY Acad Sci 361:119–134

Brown WM, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225–239

Bulmer M, Wolfe KH, Sharp PM (1991) Synonymous nucleotide substitution rates in mammalian genes: implications for the molecular clock and the relationship of mammalian orders. Proc Natl Acad Sci USA 88:5974–5978

Capaldi RA (1990) Structure and function of cytochrome *c* oxidase. Annu Rev Biochem 59:569–596

Cavalli-Sforza LL, Edwards AWF (1967) Phylogenetic analysis models and estimation procedures. Am J Hum Genet 19:233–257

Chomyn A, Mariottini P, Cleeter M, Ragan F, Matsuno-Yagi A, Hatefi Y, Doolittle R, Attardi G (1985) Six unidentified reading frames of human mitochondrial DNA encode components of the respiratory-chain NADH dehydrogenase. Nature 314: 592–597

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368–376

Felsenstein J (1990) PHYLIP manual version 3.3. University Herbarium, University of California, Berkeley

Fitch WM (1986) The estimate of total nucleotide substitutions from pairwise differences is biased. Philos Trans R Soc Lond B 312:317–324

Foran DR, Hixson JE, Brown WM (1988) Comparisons of ape and human sequences that regulate mitochondrial DNA transcription and D-loop DNA synthesis. Nucleic Aci Res 16: 5841–5861

Gadaleta G, Pepe G, De Candia G, Quagliariello C, Sbisá E, Saccone C (1989) The complete nucleotide sequence of the *Rattus norvegicus* mitochondrial genome: cryptic signals revealed by comparative analysis between vertebrates. J Mol Evol 28:497–516

Graur D (1985) Amino acid composition and the evolutionary rates of protein-coding genes. J Mol Evol 22:53–62

Greenberg BD, Newbold JE, Sugino A (1983) Intraspecific nucleotide sequence variability surrounding the origin of replication in human mitochondrial DNA. Gene 21:33–49

Hasegawa M, Kishino H, Yano T (1985) Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hayasaka K, Gojobori T, Horai S (1988) Molecular phylogeny and evolution of primate mitochondrial DNA. Mol Biol Evol 5(6):626–644

Hixson JE, Brown WM (1986) A comparison of the small ribosomal RNA genes from the mitochondrial DNA of great apes and humans: sequence, structure, evolution, and phylogenetic implications. Mol Biol Evol 3:1–18

Horai S, Hayasaka K (1990) Intraspecific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. Am J Hum Genet 46:828–842

Horai S, Satta Y, Hayasaka K, Kondo R, Inoue T, Ishida T, Hayashi S, Takahata N (1992) Man's place in homineadea revealed by mitochondrial DNA genealogy. J Mol Evol 35: 32–43

Jukes TH (1969) Recent advances in studies of evolutionary relationships between proteins and nucleic acids. Space Life Sci 1:469–490

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. New York Academic Press, pp 21–132

Kawamura S, Tanabe H, Watanabe Y, Kurosaki K, Saitou N, Ueda S (1991) Evolutionary rate of immunoglobulin alpha noncoding region is greater in hominoids than in old world monkeys. Mol Biol Evol 8(6):743–752

Kimura M (1968) Evolutionary rate at the molecular level. Nature 217:624–626

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequence. J Mol Evol 16:111–120

Kimura M (1981) Estimation of evolutionary distances between homologous nucleotide sequences. Proc Natl Acad Sci USA 78:454–458

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in humans and chimpanzees. In: Osawa S, Honjo T (eds) Evolution of life. Springer/Verlag, Tokyo, pp 391–413

Koop BF, Goodman M, Xu P, Chan K, Slightom JL (1986) Primate η-globin DNA sequences and man's place among the great apes. Nature 319:234–238

Koop BF, Tagle DA, Goodman M, Slightom JL (1989) A molecular view of primate phylogeny and important systematic and evolutionary questions. Mol Biol Evol 6(6):580–612

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Lanave C, Preparata G, Saccone C, Serio G (1984) A new method for calculating evolutionary substitution rates. J Mol Evol 20:86–93

Li W-H, Tanimura M (1987) The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–96

Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 25:330–342

Li W-H, Wu C-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Maeda N, Wu C-I, Bliska J, Reneke J (1988) Molecular evolution of intergenic DNA in higher primates: pattern of DNA changes, molecular clock, and evolution of repetitive sequences. Mol Biol Evol 5:1–20

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions. Mol Biol Evol 3:418–426

Osawa S, Muto A, Jukes TH, Ohama T (1990) Evolutionary changes in the genetic code. Proc R Soc Lond B 241:19–28

Ruvolo M, Disotell TR, Allard MW, Brown WM, Honeycutt RL (1991) Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. Proc Natl Acad Sci USA 88:1570–1574

Satta Y, Takahata N, Schönbach C, Gutknecht J, Klein J (1991) Calibrating evolutionary rates at major histocompatibility complex loci. In: Klein J, Klein D (eds) Molecular evolution of the major histocompatibility complex. Springer-Verlag, Heidelberg, pp 51–62

Takahata N, Kimura M (1981) A model of evolutionary base substitutions and its application with special reference to rapid change of pseudogenes. Genetics 98:641–657

Takahata N, Tajima F (1991) Sampling errors in phylogeny. Mol Biol Evol 8:494–502

Vigilant L, Pennington R, Harpending H, Kocher TD, Wilson AC (1989) Mitochondrial DNA sequences in single hairs from a southern African populations. Proc Natl Acad Sci USA 86:9350–9354

Wolfe KH, Sharp PM, Li W-H (1989) Mutation rates differ among regions of the mammalian genome. Nature 337:283–285

Received June 15, 1992/Revised and accepted November 24, 1992

Appendix A

Some traditional models of nucleotide substitutions (e.g., Nei and Gojobori 1986) rely on the assumption that all four nucleotides are equally likely to mutate. However, since this does not hold for the primate mtDNA, we need to use a different method that can reflect biased substitutions. The model we use distinguishes transition (α) and transversion (β) and takes into account different base frequencies (π_j, j = A, T, G, or C) at equilibrium. We use empirical relative values of $\alpha = 1$ and $\beta = 1/17$. These are close to values suggested by many studies (Brown et al. 1982; Hixon and Brown 1986; Hayasaka et al. 1988; Foran et al. 1988; Horai and Hayasaka 1990). The probability that base i is replaced by base j (P_{ij}) is defined by

$$P_{AG} = \frac{\alpha\pi_G}{\alpha\pi_G + \beta\pi_C + \beta\pi_T} \quad P_{GA} = \frac{\alpha\pi_A}{\alpha\pi_A + \beta\pi_C + \beta\pi_T}$$

$$P_{CT} = \frac{\alpha\pi_T}{\alpha\pi_T + \beta\pi_A + \beta\pi_G} \quad P_{TC} = \frac{\alpha\pi_C}{\alpha\pi_C + \beta\pi_A + \beta\pi_G}$$

etc.

The number of synonymous sites (s) at the third position of a twofold-degenerate codon is counted as follows: For codons TTC, TAC, CAC, AAC, GAC, TGC, ATC, and AGC, $s = P_{CT}$; for TTT, TAT, CAT, AAT, GAT, TGT, ATT, and AGT, $s = P_{TC}$; for CAA, AAA, GAA, ATA, and TGA, $s = P_{AG}$; and for codons CAG, AAG, GAG, ATG, TGG), $s = P_{GA}$. For leucine codons; $s = P_{TC} + P_{AG}$ for TTA, $P_{TC} + P_{GA}$ for TTG, and $P_{TC} + 1$ for CTA and CTG. The third positions in the fourfold-degenerate sites are $s = 1$. The number of nonsynonymous sites (n) for a given codon is $n = 3 - s$.

In the above, we note that under Jukes and Cantor, $\alpha = \beta$ and $\pi_j = 1/4$ for all j, so $P = 1/3$ for a twofold-degenerate synonymous site. Under Kimura's two-parameter model, however, $\alpha \neq \beta$ but $\pi_j = 1/4$, so $P = \alpha/(\alpha + 2\beta)$ for a twofold-degenerate site.

The extents of the differences by TC transitions, AG transitions, and transversions differ from each other and approach different saturation levels. It is therefore necessary to treat them separately. For a given pair of genes from two out of four species, we first compute the total numbers of TC transitions, AG transitions, and transversions at the synonymous sites (Table 3) and divide each of them by the total number of synonymous sites ($S = \Sigma s$) to obtain their per-site differences. The same procedure applies to the remaining five different pairs of species. Based on the difference matrix (d_{ij}, $i < j = 1, \ldots, 4$) for each of TC and AG transitions and transversions, we estimate the branch lengths (b_j, $j = 1, \ldots, 5$) in Fig. 7 so as to minimize the total branch lengths (Cavalli-Sforza and Edwards 1967). In the present case of four species, the estimates of b_j can be given by

$$\begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{2} \\ 0 & -\frac{1}{4} & \frac{1}{4} & -\frac{1}{4} & \frac{1}{4} & \frac{1}{2} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} d_{12} \\ d_{13} \\ d_{14} \\ d_{23} \\ d_{24} \\ d_{34} \end{pmatrix}$$

These b_j values are converted into the actual number of substitutions (x_j) from the respective graph obtained from the simulation study (Figs. 3, 4). We then take the sum of x_j for each of the TC transitions and AG transitions and transversions to obtain the total number of synonymous substitutions per site (X_j) for branch j. The distance D_{13} between species 1 and 3 in Fig. 7, for example, is defined as $X_1 + X_3 + X_5$. The sampling error of D_{13} is roughly given by $\sqrt{D_{13}/S}$. (See Takahata and Tajima 1991 for a related argument.) Similarly, the total number of synonymous substitutions in the tree, used in Fig. 5, is estimated as $\Sigma_{k=1}^{5} X_k$. Suppose that species 4 in Fig. 7 is gorilla and then denote by T the divergence time from the human and chimpanzee clade. We define the mean pairwise distances between gorilla and the human-chimpanzee clade as $D = (X_1 + X_2 + X_3 + 2X_5)/3 + X_4$ and the rate is calculated by $D/(2T)$. The overall synonymous rate is the weighted average of Ds for all the genes.