# Modeling of Molecular Evolution and Maximum Likelihood Inference of Molecular Phylogeny

Jun Adachi

DOCTOR OF PHILOSOPHY

Department of Statistical Science
School of Mathematical and Physical Science
The Graduate University for Advanced Studies

1994

(School Year)

# Abstract

This dissertation addresses the statistical methods for inferring molecular phylogeny from sequence data. Chapter 1 presents the outline of the molecular evolution and moleculer phylogeny.

In chapter 2, I studied Markov models of nucleotide substitutions of DNA and of amino acid substitutions of proteins during the course of evolution. Such modelings which approximate the real fundamental process of molecular evolution are prerequisite in inferring evolutionary trees from molecular sequence data. I developed a novel procedure for estimating the transition probability matrix of the general reversible Markov model (REV model) from a tree by using the maximum likelihood (ML), and estimated the nucleotide transition probability matrix from the four-fold degenerate sites of mitochondrial DNA. By using the same procedure, I also estimated the amino acid transition probability matrix of the REV model from protein sequences encoded by mtDNA.

In chapter 3, I developed a ML method for inferring evolutionary trees by using DNA and protein sequences. I developed a fast algorithm for estimating all the parameters (branch lengths) by the ML under a given tree topology. The new algorithm is several times faster than Felsenstein's method. I further developed methods for tree topology search, the star decomposition and the local rearrangement methods, which might be useful in applying the ML to many-OTUs problems. Furthermore, I found a good criterion correlative to the ML of a tree. The criterion is called "approximate likelihood". The approximate likelihood is a probability of a tree with parameters estimated by the least squares method from a ML distance matrix. A calculation of approximate likelihood of a tree is several hundred times faster than that of ML of the same tree topology. By using the approximate likelihood, the exhaustive search of tree topologies can be applicable for many-OTUs problems. We can exhaustively search for a ML tree in a case of about ten OTUs (among about 2 million trees). If constraint on tree topologies is given, we can analyze more OTUs.

In chapter 4, the program package "MOLPHY", which I wrote for molecular phylogenetics, is briefly described. Maximum likelihood programs, ProtML for protein sequences and NucML for nucleotide sequences, are the main programs in this package.

Chapter 5 presents several applications of the ML methods for molecular phylogeny. The internal branch lengths estimated by the distance methods such as neighbor-joining were shown to be biased to be short when the evolutionary rate differs among sites. The variable-invariable model for the site-heterogeneity fits the amino acid sequence data encoded by the mitochondrial DNA from Hominoidea remarkably well. By assuming the orangutan separation to be 13 or 16 Myr old, a ML analysis estimated a young date of 3.6 $\pm$ 0.6 or 4.4 $\pm$ 0.7 Myr ($\pm$: 1SE) for the human/chimpanzee separation, and these estimates turned out to be robust against differences in the assumed model for amino acid substitutions. Although some uncertainties still exist in our estimates, this analysis suggests that humans separated from chimpanzees some 4–5 Myr ago.

From phylogenetic analyses of the 12S and 16S mitochondrial ribosomal DNA and of myoglobin amino acid sequences, Milinkovitch et al. (1993[212]) proposed the hypothesis that one group of toothed whales (Odontoceti), the sperm whales (Physeteridae), is more closely related to the baleen whales (Mysticeti) than to other alleged odontocetes such as dolphins. This hypothesis is in conflict with the traditional view that the odontocetes form a monophyletic clade. From an analysis of the cytochrome *b* gene, Árnason and Gullberg (1994[22]) recently challenged Milinkovitch et al.'s hypothesis as well as the traditional tree, claiming that the mysticetes are closer to the dolphins rather than to the sperm whales. They used the cow as the only outgroup and the giant sperm whale as the only representative of Physeteridae, but the estimated tree may depend on the sampled species. By including many alternative artiodactyl outgroups in their cytochrome *b* dataset, I showed that Árnason and Gullberg's conclusion is shaky, and that the overall evidence favours Milinkovitch et al.'s hypothesis.

I thus demonstrated the importance of species sampling in molecular phylogenetics, and showed that a conclusion drawn from a limited number of species might be unstable. This problem was also examined with the cytochrome *b* data for the phylogenetic relationship among Ruminantia, Suiformes, Cetacea, and other outgroup species of mammals, and the importance of species sampling was again demonstrated even more clearly.

Furthermore, the ProtML program was applied to the 183-OTUs problem of cytochrome *b* data in order to elucidate several phylogenetic problems of mammals and birds, and also to cytochrome oxidase subunit II data of mammals. The NucML was applied to ribosomal RNA data in order to reconfirm recently proposed hypothesis on the phylogenetic origin of myxozoan protists.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Molecular Evolution and Molecular Phylogeny

Molecular evolution encompasses two areas of study: (1) the evolution of macromolecules and (2) the reconstruction of the evolutionary history of organisms. By "evolution of macromolecules" we refer to the rates and patterns of change occurring in the genetic material (e.g., DNA sequences) and its products (e.g., proteins) during evolutionary time and to the mechanisms responsible for such changes. The second area, also known as "molecular phylogenetics," deals with the evolutionary history of organisms as inferred from molecular data.

It might appear that the two areas of study constitute independent fields of inquiry, since the object of the first is to elucidate the causes and effects of evolutionary changes in molecules, while the second uses molecules as a tool to reconstructing the history of organisms. In practice, however, the two disciplines are intimately interrelated, and progress in one area facilitates studies in the other. For instance, phylogenetic knowledge is essential for determining the order of changes in the molecular characters under study. And conversely, knowledge of the pattern and rate of change of a given molecule is crucial in reconstruct the evolutionary history of a group of organisms.

Traditionally, a third area of study, prebiotic evolution or the "origin of life," is also included within the framework of molecular evolution. This subject, however, involves a great deal of speculation and is less amenable to quantitative treatments. Moreover, the rules that govern the process of information transfer in prebiotic systems (i.e., systems devoid of replicable genes) are not known at the present time. Therefore, this thesis will not deal with the origin of life.

The study of molecular evolution has its roots in three disparate disciplines: molecular biology, (population) genetics and statistics. Statistics and population genetics provides the theoretical foundation for the study of evolutionary processes, while molecular biology provides the empirical data.

## 1.1 DNA Sequences

The hereditary information of all living organisms, with the exception of some viruses, is carried by deoxyribonucleic acid (DNA) molecules. DNA usually consists of two complementary chains twisted

around each other to a right-handed helix. Each chain is a linear polynucleotide consisting of four nucleotides. There are two pyrimidines: thymine (T) and cytosine(C), and two purines: adenine (A) and guanine (G). The two chains are joined together by hydrogen bonds between pairs of nucleotides. Adenine pairs with thymine by means of two hydrogen bonds, also referred to as the weak bond, and guanine pairs with cytosine by means of three hydrogen bonds, the strong bond.

Ribonucleic acid (RNA) is found as either a double- or single-stranded molecule. RNA differs from DNA by having ribose, instead of deoxyribose, as its backbone sugar moiety, and by using the nucleotide uracil (U) instead of thymine. Adenine, cytosine, guanine and thymine/uracil are referred to as the standard nucleotides. Some functional RNA molecules, most notably tRNA, contain nonstandard nucleotides, i.e., chemical modifications of standard nucleotides that have been introduced into the RNA after its transcription.

Table 1.1: The four nitrogenous bases and their type.

| Type | Base |
|---|---|
| Pyrimidine | Thymine ( Uracil ) |
| | Cytosine |
| Purine | Adenine |
| | Guanine |

The ultimate step in obtaining genetic data is the determination of the sequence of bases in the DNA molecule. There are four kinds of base T, C, A and G as shown in Table 1.1 (U used instead of T in RNA molecules), and DNA sequence consists of coding as well as noncoding regions. Such data have been collected with different objectives in mind. Constructions of phylogenies for a set of species usually use a single representative sequence from each of the species, while questions concerning variation within species require several sequences from a species.

Table 1.2: Standard base codes.

| Code | Interpretation | Nucleotide group |
|---|---|---|
| A | Adenine | |
| T | Thymine | |
| G | Guanine | |
| C | Cytosine | |
| R | Purines (large) | A or G |
| Y | Pyrimidines (small) | C or T |
| M | Amino (positive charge) | A or C |
| K | Ketone (negative charge) | G or T |
| W | Weak interaction | A or T |
| S | Strong interaction | C or G |
| H | Not G | A or C or T |
| B | Not A | C or G or T |
| V | Not T | A or C or G |
| D | Not C | A or G or T |
| N | Any | A or C or G or T |

## 1.2 Protein-coding Genes

Traditionally, a gene was defined as a segment of DNA that codes for a polypeptide chain or specifies a functional RNA molecule. Recent molecular studies, however, have radically altered our perception of gene, and we shall adopt a somewhat definition. Accordingly, a gene is a sequence of genomic DNA or RNA that is essential for a specific function. Performing the function may not require the gene to be translated or even transcribed.

At present, three type of genes are recognized: (1) protein-coding genes, which are transcribed into RNA and subsequently translated into proteins, (2) RNA-specifying genes, which are only transcribed, and (3) regulatory gene. According to a narrow definition, the third category includes only untranscribed sequences. Transcribed genes for regulation and RNA-specifying genes are also referred to as structural genes. Note that some authors restrict the definition of structural genes to include only protein-coding genes.

A standard eukaryotic protein-coding gene consists of transcribed and nontranscribed parts. The transcription of protein-coding genes starts at the transcription-initiation site (the cap site in the RNA transcript), and ends at the termination site, which may or may not be identical with the polyadenylation or poly(A)-addition site of the mature messenger RNA (mRNA) molecule. The transcribed RNA, also referred to as premessenger RNA (pre-mRNA), contains 5' and 3' untranslated regions, exons, and introns. Introns, or intervening sequences, are those transcribed sequences that are excised during the processing of the pre-mRNA molecule. All genomic sequences that remain in the mature mRNA following splicing are referred to as exons. Exons or parts of exons that are translated are referred to as protein-coding exons or coding regions.

## 1.3 Protein Sequence Data

Electrophoresis can detect change differences among different forms of a protein, but a more complete picture of proteins was provided when it became possible to determine the sequence of amino acids constituting the protein. The variation revealed by this technology allowed more detailed studies of the evolutionary relationships between different species. The protein sequences employ a standard one-letter code for amino acids, as shown in Table 1.3. Protein sequences are now collected into databases so that they are widely accessible.

## 1.4 Genetic Code

The systhesis of proteins involves a process of decoding, whereby the genetic information carried by an mRNA molecule is translated into amino acid through the use of transfer RNA (tRNA) mediator. A list of the 20 primary amino acids and their abbreviations is given in Table 1.3. Translation starts at the translation-initiation site and proceeds to a stop signal. Translation involves the sequential recognition of

Table 1.3: One-letter and three-letter codes for the 20 amino acids.

| One-letter | Three-letter | Amino Acid |
|---|---|---|
| A | Ala | Alanine |
| R | Arg | Arginine |
| N | Asn | Asparagine |
| D | Asp | Aspartic acid |
| C | Cys | Cysteine |
| Q | Gln | Glutamine |
| E | Glu | Glutamic acid |
| G | Gly | Glycine |
| H | His | Histidine |
| I | Ile | Isoleucine |
| L | Leu | Leucine |
| K | Lys | Lysine |
| M | Met | Methionine |
| F | Phe | Phenylalanine |
| P | Pro | Proline |
| S | Ser | Serine |
| T | Thr | Threonine |
| W | Trp | Tryptophan |
| Y | Tyr | Tyrosine |
| V | Val | Valine |
| B | Asx | Aspartic acid or Asparagine |
| Z | Glx | Glutamine or Glutamic acid |
| X | Xaa | Any amino acid |

adjacent nonoverlapping triplets of nucleotides, called codons. The phase at which a sequence is translated is determined by the initiation codon and is referred to as the reading frame. In the translational machinery at the interphase between the ribosome and the mRNA molecule, each codon is translated into a specific amino acid, which is subsequently added to the elongating polypeptide. The correspondence between the codons and the amino acid is determined by a set of rules called the genetic code. With a few exceptions (see later), the genetic code for nuclear protein-coding genes is "universal," i.e., the translation of almost all eukaryotic nuclear genes and prokaryotic genes is determined by the same set of rules.

The universal genetic code is given in Table 1.4.

## 1.5 Mutation

DNA sequences are normally copied exactly during the process of chromosome replication. Rarely, however, errors occur that give rise to sequence. These errors are call mutations. Mutations can occur in either somatic or germ-line cells. Since somatic mutations are not inherited, we can disregard them in an evolutionary context, and throughout this thesis the term "mutation" will denote mutations in germ-line cell.

Mutations may be classified by the length of the DNA sequence affected by the mutational event. For instance, mutations may affect a single nucleotide (point mutations) or several adjacent nucleotides. We

Table 1.4: The universal genetic code.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TTT | Phe | F | TCT | Ser | S | TAT | Tyr | Y | TGT | Cys | C |
| TTC | Phe | F | TCC | Ser | S | TAC | Tyr | Y | TGC | Cys | C |
| TTA | Leu | L | TCA | Ser | S | TAA | Stop | * | TGA | Stop | * |
| TTG | Leu | L | TCG | Ser | S | TAG | Stop | * | TGG | Trp | W |
| CTT | Leu | L | CCT | Pro | P | CAT | His | H | CGT | Arg | R |
| CTC | Leu | L | CCC | Pro | P | CAC | His | H | CGC | Arg | R |
| CTA | Leu | L | CCA | Pro | P | CAA | Gln | Q | CGA | Arg | R |
| CTG | Leu | L | CCG | Pro | P | CAG | Gln | Q | CGG | Arg | R |
| ATT | Ile | I | ACT | Thr | T | AAT | Asn | N | AGT | Ser | S |
| ATC | Ile | I | ACC | Thr | T | AAC | Asn | N | AGC | Ser | S |
| ATA | Ile | I | ACA | Thr | T | AAA | Lys | K | AGA | Arg | R |
| ATG | Met | M | ACG | Thr | T | AAG | Lys | K | AGG | Arg | R |
| GTT | Val | V | GCT | Ala | A | GAT | Asp | D | GGT | Gly | G |
| GTC | Val | V | GCC | Ala | A | GAC | Asp | D | GGC | Gly | G |
| GTA | Val | V | GCA | Ala | A | GAA | Glu | E | GGA | Gly | G |
| GTG | Val | V | GCG | Ala | A | GAG | Glu | E | GGG | Gly | G |

Table 1.5: The mitochondrial genetic code (vertebrates).

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TTT | Phe | F | TCT | Ser | S | TAT | Tyr | Y | TGT | Cys | C |
| TTC | Phe | F | TCC | Ser | S | TAC | Tyr | Y | TGC | Cys | C |
| TTA | Leu | L | TCA | Ser | S | TAA | Stop | * | TGA | **Trp** | **W** |
| TTG | Leu | L | TCG | Ser | S | TAG | Stop | * | TGG | Trp | W |
| CTT | Leu | L | CCT | Pro | P | CAT | His | H | CGT | Arg | R |
| CTC | Leu | L | CCC | Pro | P | CAC | His | H | CGC | Arg | R |
| CTA | Leu | L | CCA | Pro | P | CAA | Gln | Q | CGA | Arg | R |
| CTG | Leu | L | CCG | Pro | P | CAG | Gln | Q | CGG | Arg | R |
| ATT | Ile | I | ACT | Thr | T | AAT | Asn | N | AGT | Ser | S |
| ATC | Ile | I | ACC | Thr | T | AAC | Asn | N | AGC | Ser | S |
| ATA | **Met** | **M** | ACA | Thr | T | AAA | Lys | K | AGA | **Stop** | * |
| ATG | Met | M | ACG | Thr | T | AAG | Lys | K | AGG | **Stop** | * |
| GTT | Val | V | GCT | Ala | A | GAT | Asp | D | GGT | Gly | G |
| GTC | Val | V | GCC | Ala | A | GAC | Asp | D | GGC | Gly | G |
| GTA | Val | V | GCA | Ala | A | GAA | Glu | E | GGA | Gly | G |
| GTG | Val | V | GCG | Ala | A | GAG | Glu | E | GGG | Gly | G |

may also classify mutations by the type of change caused by the mutational event into (1) substitutions, the replacement of one nucleotide by another, (2) deletions, the removal of one or more nucleotides from the DNA, (3) insertions, the addition of one or more nucleotides to the sequence, and (4) inversions, the reversal of polarity of a sequence involving two or more nucleotides.

## 1.6 Nucleotide Substitutions

Nucleotide substitutions are divided into transitions and transversions. Transitions are substitutions between T and C (pyrimidines) or between A and G (purines). Transversions are substitutions between a pyrimidines and a purines.

Nucleotide substitutions occurring in protein-coding regions can also be characterized by their effect on the product of translation, the protein. A substitution is called to be synonymous or silent if it causes no amino acid change. Otherwise, it is nonsynonymous. Nonsynonymous (or amino-acid-altering) mutations are further classified into missense and nonsense mutations. A missense mutation changes the affected codon into a codon that specifies a different amino acid from the one previously encoded. A nonsense mutation changes a codon into one of the termination codons, thus prematurely ending the translation process and ultimately resulting in the production of a truncated protein.

# Chapter 2

# Modeling of Molecular Evolution

A basic process in the evolution of DNA and protein sequences is the change in nucleotides and amino acids with time. This process deserves a detailed consideration since changes in nucleotide and amino acid sequences are used in molecular evolutionary studies both for estimating the rate of evolution and for inferring the evolutionary history of organisms. However, as the processes of nucleotide and amino acid substitutions are usually extremely slow, they cannot be observed within a researcher's life. Therefore, to detect evolutionary changes in DNA and protein sequences, we resort to comparative methods whereby a given sequence is compared with other sequences with which it shared a common ancestry in the evolutionary past. Such comparisons require statistical methods, several of which will be discussed in this chapter.

To study the dynamics of nucleotide and amino acid substitutions, we must make several assumptions regarding the probability of substitution of one nucleotide or amino acid by another. Numerous such mathematical schemes have been proposed in the literature for nucleotide substitutions (Kimura 1980[161], 1981[162]; Takahata and Kimura[286]; Gojobori et al. 1982[90], 1982[89]; Hasegawa et al. 1985[122]; Barry and Hartigan 1987[38]; Rodríigues et al. 1990[247]; Saccone et al. 1990[251]; Tamura and Nei 1993[288]; Yang 1994[317]; Kelly 1994[157]) and for amino acid substitutions (Dayhoff et al. 1978[62]; Kishino et al. 1990[166]; Altschul 1991[13]; Jones et al. 1992[154]; Henikoff and Henikoff 1992[138]; Gonnet et al. 1992[96]).

## 2.1 Modeling of Nucleotide Substitutions

Nucleotide substitutions of the four-fold degenerate sites of mitochondrial DNA (mtDNA) from human (Anderson et al. 1981[15]) , common chimpanzee, bonobo, gorilla, orangutan, and siamang (Horai et al. 1992[141]) were examined in detail by three alternative Markov models; (1) Hasegawa, Kishino and Yano's (1985[122]) model, (2) Tamura and Nei's (1993[288]) model, and (3) the general reversible Markov model (Yang 1994[317]). These sites are expected to be relatively free from constraint compared with other sites, and therefore their pattern of evolution should reflect that of mutation. It turned out that, among the alternative models, the general reversible Markov model best approximates the nucleotide substitutions of the four-fold degenerate sites, while the ML estimates of the numbers of nucleotide substitutions along each branches do not differ significantly among the three models.

### 2.1.1 Markov Models of Nucleotide Substitutions

Nucleotide substitutions of the third positions of four-fold degenerate codon families are always synonymous, and are expected to be relatively free from constraint, and therefore their tempo and mode in evolution should reflect those of mutation. Since the evolutionary rate of animal mtDNA is much higher than that of nuclear DNA (Brown et al. 1982[43]; Miyata et al. 1982[217]; Hasegawa et al. 1984[128]) and hence the multiple-hit effect is great in a comparison between distantly related species, closely related species must be compared in order to examine the pattern of synonymous nucleotide substitutions of mtDNA. Horai et al. (1992[141]) determined 4.8kbp of mtDNA sequences from common chimpanzee (*Pan troglodytes*), pygmy chimpanzee (bonobo; *Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and siamang (*Hylobates syndactylus*). From this data, together with the corresponding sequence from human (*Homo sapiens*) (Anderson et al. 1981[15]), they established that the closest relatives of the human are the two chimpanzees rather than the gorilla. These data from closely related primate species provide us with an opportunity to examine in detail the pattern of synonymous nucleotide substitution of animal mtDNA.

#### Transition Probability Matrix

We assume that each site evolves independently on the other sites according to a reversible Markov process. A probability of a nucleotide $i$ (T, C, A, or G; numbering in this order) being replaced by a nucleotide $j$ in an infinitesimally short time interval, $dt$, is represented by $P_{ij}(dt)$. We would like to derive a transition probability matrix for a finite time $t$,

$$\boldsymbol{P}(t)$$

where

$$\sum_{j=1}^{4} P_{ij}(t) = 1 \qquad (i = 1, \cdots, 4)$$

A time interval during which one nucleotide substitution occurs per 100 sites is taken as a unit of time, and we consider a transition probability matrix $\boldsymbol{M}$ for a unit time interval;

$$\boldsymbol{P}(1) = \boldsymbol{M}$$

Kishino et al. (1990[166]) presented a method for deriving a transition probability matrix $\boldsymbol{P}(t)$ of amino acids from $\boldsymbol{M}$ compiled empirically by Dayhoff et al. (1978[62]). We can extend the method to nucleotide substitutions as described below.

If the unit time interval is sufficiently short, the transition probability matrix $\boldsymbol{P}(t)$ for time interval $t$ is given by

$$\boldsymbol{P}(t) = \exp(t\boldsymbol{W}) \tag{2.1}$$

where $\boldsymbol{W}$ is a function of eigen-values $\lambda_i$ and eigen-vectors $\boldsymbol{u}_i$ of $\boldsymbol{M}$, and is represented by

$$\boldsymbol{W} = \boldsymbol{U} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_4 \end{pmatrix} \boldsymbol{U}^{-1} \tag{2.2}$$

and

$$\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_4) \tag{2.3}$$

Therefore,

$$P_{ij}(t) = \sum_{k=1}^{4} \left( U_{ik} U_{kj}^{-1} \exp(t\lambda_k) \right) \tag{2.4}$$

Thus, if the transition probability matrix $\boldsymbol{M}$ for a unit time is given, the matrix for time $t$ can be calculated.

### Poisson Model

The simplest model for nucleotide substitution is the Poisson model, in which a nucleotide is replaced by any other nucleotides with an equal probability. This model for nucleotide substitution is sometimes called the Jukes-Cantor (1969[156]) model. Let $\delta$ be the number of nucleotide substitutions per site per unit time interval, and we take $\delta = 0.01$. The transition probability for a unit time of the Poisson model is,

$$\boldsymbol{M} = \begin{pmatrix} 1-\delta & \delta/3 & \delta/3 & \delta/3 \\ \delta/3 & 1-\delta & \delta/3 & \delta/3 \\ \delta/3 & \delta/3 & 1-\delta & \delta/3 \\ \delta/3 & \delta/3 & \delta/3 & 1-\delta \end{pmatrix} \tag{2.5}$$

Although the representation of $\boldsymbol{M}$ is thus simple for the Poisson model, it becomes complicated for models in which the transition and transversion rates are distinguished, or in which nucleotide frequencies are unequal. In order to derive $\boldsymbol{M}$ in these models, we define the relative substitution rate $\boldsymbol{R}$ as follows;

$$\begin{aligned} R_{ii} &= 0 \qquad (i = 1, \ldots, 4) \\ R_{ij} &= R_{ji} \geq 0 \quad (i, j = 1, \ldots, 4) \end{aligned}$$

For amino acid substitutions, $\boldsymbol{R}$ is related to the accepted mutation matrix $\boldsymbol{A}$ in Fig. 80 of Dayhoff et al. (1978[62]) by the following formula;

$$R_{ij} = A_{ij}/(20^2 \pi_i^A \pi_j^A), \tag{2.6}$$

where $\pi_i^A$ is the frequency of amino acid $i$ in the data set used in constructing $\boldsymbol{A}$ (given in Table 22 of Dayhoff et al.). The matrix $\boldsymbol{R}$ represents relative frequency of substitutions, and its absolute value has no special meaning. Differing from the transition probability matrix $\boldsymbol{M}$, a summation of a row of $\boldsymbol{R}$ need not be 1. Because of this freedom from the constraint, we can construct the matrix easily.

The relative substitution frequency for the Poisson model is

$$\boldsymbol{R} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} 0 & \alpha & \alpha & \alpha \\ \alpha & 0 & \alpha & \alpha \\ \alpha & \alpha & 0 & \alpha \\ \alpha & \alpha & \alpha & 0 \end{pmatrix} \end{array} \tag{2.7}$$

Usually we take $\alpha = 1$.

From $\boldsymbol{R}$, we can derive $\boldsymbol{M}$ as follows;

$$M_{ij} = \begin{cases} 4\delta R_{ij}/s & (i \neq j) \\ 1 - 4\delta \sum_{k=1}^{4} R_{ik}/s & (i = j) \end{cases} \tag{2.8}$$

where

$$s = \sum_{i=1}^{4} \sum_{j=1}^{4} R_{ij} \tag{2.9}$$

### Proportional Model

In the proportional model which was proposed by Felsenstein (1981[76]), $P_{ij}$ is proportional to the frequency of nucleotide $j$, $\pi_j$ (where $\sum_{j=1}^{4} \pi_j = 1$), and the relative substitution rate is identical with that of the Poisson model (Eq. 2.7). If the nucleotide frequency of the data under analysis is taken as $\boldsymbol{\pi}$, this means that the frequency of the data is at the stationary state of the Markov process. A higher abundance of a particular nucleotide than others is interpreted to be due to higher substitution probability to the nucleotide than to the others. Since the nucleotide composition is highly biased in mtDNA, the introduction of the parameter $\boldsymbol{\pi}$ is important in analyzing mtDNA sequences. The transition probability matrix $\boldsymbol{M}$ for the proportional model is given by

$$M_{ij} = \begin{cases} \delta \pi_j R_{ij}/s & (i \neq j) \\ 1 - \delta \sum_{k=1}^{4} (\pi_k R_{ik})/s & (i = j) \end{cases} \tag{2.10}$$

where

$$s = \sum_{i=1}^{4} \left( \pi_i \sum_{j=1}^{4} (\pi_j R_{ij}) \right). \tag{2.11}$$

By using this transformation, we can easily construct a model dependent on $\boldsymbol{\pi}$.

### Hasegawa, Kishino and Yano's (1985) Model

It is known that transition predominates over transversion particularly in the evolution of animal mtDNA (Brown et al. 1982[43]). Kimura (1980[161]) extended the Poisson model so as to take account of the difference between transition and transversion, but he did not take account of the biased nucleotide composition. Hasegawa, Kishino and Yano (1985[122]) combined the Kimura model with the proportional model of Felsenstein, and we call this the HKY85 model. The relative substitution rate matrix for the HKY85 model is,

$$\boldsymbol{R} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} 0 & \alpha & \beta & \beta \\ \alpha & 0 & \beta & \beta \\ \beta & \beta & 0 & \alpha \\ \beta & \beta & \alpha & 0 \end{pmatrix} \end{array} \tag{2.12}$$

where $\alpha$ and $\beta$ are relative substitution rates of transition and transversion, respectively. We can take $\beta = 1$, and then $\alpha$ represents the transition/transversion ratio. By using the transformation of Eq. 2.10, we can obtain the transition probability matrix $\boldsymbol{M}$ of the HKY85 model for a unit time interval.

### Tamura and Nei's (1993) Model

Tamura and Nei (1993[288]) proposed a more general model, which we call the TN93 model, than the HKY85 model. The model allows different transition rates for purines and pyrimidines. The relative substitution rate for the TN93 model is

$$\boldsymbol{R} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} 0 & \alpha_{\mathrm{Y}} & \beta & \beta \\ \alpha_{\mathrm{Y}} & 0 & \beta & \beta \\ \beta & \beta & 0 & \alpha_{\mathrm{R}} \\ \beta & \beta & \alpha_{\mathrm{R}} & 0 \end{pmatrix} \end{array} \tag{2.13}$$

where $\alpha_{\mathrm{Y}}$ is the relative substitution rate between pyrimidines, $\alpha_{\mathrm{R}}$ is that between purines, and $\beta$ is the relative transversion rate. Given $\beta = 1$, $\alpha_{\mathrm{Y}}$ and $\alpha_{\mathrm{R}}$ represent the transition frequencies between pyrimidines and purines relative to the transversion frequency. By using the transformation of Eq. 2.10, we can obtain the transition probability matrix $\boldsymbol{M}$ of the TN93 model for a unit time interval.

Tamura (1994[287]) showed that the TN93 model is superior to the HKY85 model in approximating the four-fold degenerate sites, as well as all the third codon positions in Horai et al.'s (1992[141]) data of 4.8kbp mtDNA sequences from Hominoidea.

### General Reversible Markov Model

By increasing the number of parameters in $\boldsymbol{R}$, we can construct various Markov models for nucleotide substitutions. Yang (1994[317]) estimated 4 × 4 transition matrices of the most general reversible Markov model (REV model) for primate $\psi\eta$-globin pseudogenes and for primate mtDNA sequences including all

codon positions as well as tRNAs. The relative substitution rate of the REV model is

$$\boldsymbol{R} = \begin{array}{c} \\ \mathrm{T} \\ \mathrm{C} \\ \mathrm{A} \\ \mathrm{G} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{T} & \mathrm{C} & \mathrm{A} & \mathrm{G} \end{array} \\ \begin{pmatrix} 0 & \alpha_{\mathrm{Y}} & \beta_{\mathrm{W}} & \beta_{\mathrm{K}} \\ \alpha_{\mathrm{Y}} & 0 & \beta_{\mathrm{M}} & \beta_{\mathrm{S}} \\ \beta_{\mathrm{W}} & \beta_{\mathrm{M}} & 0 & \alpha_{\mathrm{R}} \\ \beta_{\mathrm{K}} & \beta_{\mathrm{S}} & \alpha_{\mathrm{R}} & 0 \end{pmatrix} \end{array} \tag{2.14}$$

By using the transformation of Eq. 2.10, we can obtain the transition probability matrix $\boldsymbol{M}$ of the REV model for a unit time interval.

Saccone et al. (1990[251]) also proposed a similar reversible model. Saccone et al. (1990[251]) and Tamura (1994[287]) estimated transition matrices for their respective models from pairwise comparisons of sequences, and hence the matrix differs between different species-pairs of the same gene. They did not propose any method to synthesize the estimated matices from several comparisons when one deals with more than three species. It would be desirable to estimate a single transition probability matrix from a tree, and Yang (1994[317]) first gave the ML method for estimating the transition probability matrix from a tree with more than three species. However, the details of the procedure were not given in his paper. Therefore, I will give the details of the method in this thesis, and I will further estimate the transition probablity matrices of the REV model for the four-fold degenerate sites of mtDNA.

All the models described in this chapter were implemented in the NucML program of MOLPHY (Adachi and Hasegawa 1995[5]).

### 2.1.2 ML Estimate of the Transition Probability Matrix for the REV Model

Provided the tree topology which generated the nucleotide sequence data $\boldsymbol{X}$ is known, we estimate the relative substitution rate $\boldsymbol{R}$ and numbers of nucleotide substitutions along each branches, $t_1, \ldots, t_m$ ($m$: number of branches in the tree) by the ML;

$$\text{maximize} \quad l(\boldsymbol{R}, \boldsymbol{t}|\boldsymbol{X}) \tag{2.15}$$

where $l$ is a likelihood function and $\boldsymbol{t} = [t_1, t_2, \ldots, t_m]^T$.

At first we give the initial value of $\boldsymbol{R}$ by assuming the Proportional model and that of $\boldsymbol{t}$ as the ML estimate under the model. Then we iterate ML estimations of $\boldsymbol{R}$ by the Brent method and of $\boldsymbol{t}$ by the Newton-Raphson method alternately. At a step of iteration when the differences of all parameters between the preceding two steps are less than $\epsilon$, we stop the procedure. The procedure of the ML estimation of $\boldsymbol{R}$ and $\boldsymbol{t}$ is shown below by pseudocode with the following conventions; the looping constructs "for" and "repeat - until" have the same meanings as in Pascal, "$\triangleright$" indicates that the remainder of the line is a comment, and the form "$i \leftarrow j$" assigns the value of expression $j$ to a variable $i$.

**Maximum-Likelihood-Procedure** ( $\boldsymbol{X}$ )
begin
    $\boldsymbol{R} \leftarrow$ Proportional Model
    $\boldsymbol{t}^{\text{old}} \leftarrow$ a least squares estimate from distance matrix
    $\boldsymbol{t} \leftarrow$ **MLE-Branch-Length** ( $\boldsymbol{X}$, $\boldsymbol{R}$, $\boldsymbol{t}^{\text{old}}$ )
    repeat
        $\boldsymbol{R}^{\text{old}} \leftarrow \boldsymbol{R}$
        $\boldsymbol{R} \leftarrow$ **MLE-Relative-Substitution-Rate** ( $\boldsymbol{X}$, $\boldsymbol{t}$, $\boldsymbol{R}^{\text{old}}$ )
        $\boldsymbol{t}^{\text{old}} \leftarrow \boldsymbol{t}$
        $\boldsymbol{t} \leftarrow$ **MLE-Branch-Length** ( $\boldsymbol{X}$, $\boldsymbol{R}$, $\boldsymbol{t}^{\text{old}}$ )
    until $|\boldsymbol{R} - \boldsymbol{R}^{\text{old}}| < \epsilon$ and $|\boldsymbol{t} - \boldsymbol{t}^{\text{old}}| < \epsilon$
    return $\boldsymbol{R}$ and $\boldsymbol{t}$
end.

MLE-Relative-Substitution-Rate ( $\boldsymbol{X}$, $\boldsymbol{t}$, $\boldsymbol{R}^{\text{old}}$ ) is the pseudocode of the procedure for the ML estimation of $\boldsymbol{R}$ under given $\boldsymbol{X}$ and $\boldsymbol{t}$.

**MLE-Relative-Substitution-Rate** ( $\boldsymbol{X}$, $\boldsymbol{t}$, $\boldsymbol{R}^{\text{old}}$ )
begin
    $\boldsymbol{R} \leftarrow \boldsymbol{R}^{\text{old}}$
    for $i \leftarrow 1$ to 3
        for $j \leftarrow i+1$ to 4
            $\triangleright$ maximum likelihood estimate by Brent method
            maximize $l(R_{ij}|\boldsymbol{X}, \boldsymbol{t}, \boldsymbol{R}^*_{ij})$ $\triangleright$ $\boldsymbol{R}^*_{ij}$ is excluding $R_{ij}$
    return $\boldsymbol{R}$
end.

MLE-Branch-Length ( $\boldsymbol{X}$, $\boldsymbol{R}$, $\boldsymbol{t}^{\text{old}}$ ) is the pseudocode of the procedure for the ML estimation of $\boldsymbol{t}$ under given $\boldsymbol{X}$ and $\boldsymbol{R}$. The Newton-Raphson method is used for optimizing $\boldsymbol{t}$. We have used the same procedure in the NucML program for inferring a ML tree from nucleotide sequences (Adachi and Hasegawa 1995[5]).

**MLE-Branch-Length** ( $\boldsymbol{X}$, $\boldsymbol{R}$, $\boldsymbol{t}^{\text{old}}$ )
begin
    $\boldsymbol{t} \leftarrow \boldsymbol{t}^{\text{old}}$
    $\triangleright$ maximum likelihood estimate by Newton-Raphson method
    maximizes $l(\boldsymbol{t}|\boldsymbol{X}, \boldsymbol{R})$
    return $\boldsymbol{t}$
end.

### 2.1.3 Fitting of Models to the Four-Fold Degenerate Sites Data

Following protein-encoding regions in Anderson et al. (1981[15]) and Horai et al. (1992[141], 1993[142]) were used. ND1 (4123–4260 in the numbering of Anderson et al.), ND2 (4470–5510), COI (5904–7442), COII (7586–8266), ATPase 8 (8366–8524), ATPase 6 (8575–9024, overlapping region with ATPase8, 8525–8574, was excluded). The total number of deduced codons is 1344, and among these, the number of codons remaining four-fold degenerate during evolution is 611.

I estimated the relative substitution rate $\boldsymbol{R}$ of the REV model from the 611 sites data by the ML based on the tree of the six hominoid species, ((((chimp, bonobo), human), gorilla), orang, siamang), and it is given in Table 2.1. By using the transformation of Eq. 2.10, the transition probability matrix $\boldsymbol{M}$ of the REV model for a unit time interval was obtained as shown in Table 2.2 (Adachi and Hasegawa 1995[7]). The ML tree estimated by this model is represented in Fig. 2.1 with the branch lengths estimated by the HKY85 and TN93 models as well.

Table 2.1: Relative substitution rate matrix of the REV model for the four-fold degenerate sites.

| | T | C | A | G |
|---|---|---|---|---|
| T | | 25.0493 | 2.9367 | 6.3492 |
| C | 25.0493 | | 0.8445 | 1.0967 |
| A | 2.9367 | 0.8445 | | 63.7237 |
| G | 6.3492 | 1.0967 | 63.7237 | |
| $\pi$ | 0.167 | 0.421 | 0.366 | 0.046 |

The relative substitution rate matrix $\boldsymbol{R}$ of the REV model estimated by ML from the four-fold degenerate sites of mtDNA (611 sites). $\pi$ refers to nucleotide frequency.

Table 2.2: Transition probability matrix of the REV model for the four-fold degenerate sites.

| $\nearrow$ | T | C | A | G |
|---|---|---|---|---|
| T | 0.98148 | 0.01640 | 0.00167 | 0.00046 |
| C | 0.00648 | 0.99296 | 0.00048 | 0.00008 |
| A | 0.00076 | 0.00055 | 0.99410 | 0.00459 |
| G | 0.00164 | 0.00072 | 0.03618 | 0.96146 |

The transition probability matrix $\boldsymbol{M}$ of the REV model for a unit time interval (one substitution per 100 sites) estimated by ML from the four-fold degenerate sites of mtDNA (611 sites). From Adachi and Hasegawa (1995[7]).

Table 2.2 shows that the occurrence of nucleotide substitutions at the four-fold degenerate sites is distinctly asymmetric between the two strands of mtDNA. G→A and T→C transitions are 0.03618/0.00648 = 5.6 and 0.01640/0.00459 = 3.6 times more frequent on the L-strand (as represented in the table) than on the H-strand, respectively. This nucleotide substitution bias is roughly consistent with Tanaka and Ozawa's (1994[289]) estimates from the four-fold degenerate sites of the entire mitochondrial genomes of 43 human individuals; that is, G→A and T→C transitions are 9 and 1.8 times more frequent on the L-strand than on the H-strand.

Among the alternative models, we can select the best model by minimizing the Akaike Information Criterion (Akaike 1973[11], 1974[12]) defined by AIC = −2 × (log-likelihood) + 2 × (number of parameters). The REV, TN93 and HKY85 models gave AIC of 5284.4, 5296.6 and 5323.6, and the REV model turned out to be the best among these models in approximating the evolution of the four-fold degenerate sites.

For the alignment of 6 OTUs, $4^6 = 4096$ configurations of nucleotide sites are possible, and probabilities of respective configurations were calculated under the respective models with the branch lengths given in Fig. 2.1. Grouping these configurations into 8 categories of 0-change, 1-TC-transition (configurations which could arise from one transition between T and C), 1-AG-transition, 1-GT-transversion, 1-GC-transversion, 1-AT-transversion, 1-AC-transversion, and ≥2-changes (configurations which could not arise from less than two changes), a $\chi^2$ test for the REV model gave $P$ value of as high as 0.75 (Table 2.3), indicating that the transition probability matrix of Table 2.2 well approximates the evolution of the four-fold degenerate sites and that the site-heterogeneity is not as important as in the case of amino acid sequences studied by Adachi and Hasegawa (1995[4]). Although Kondo et al. (1993[172]) pointed out that the nucleotide frequency of third codon positions differs from genes to genes and hence further complication of the model might become necessary in the future, the transition probabability matrix of Table 2.2 turned out to represent a reasonablly good model in approximating the evolution of third positions of four-fold degenerate codons. Both of $\chi^2$ tests for the TN93 and HKY85 models gave low $P$ value of 0.03 (Table 2.3). Discrepancies of these models with the data are mainly due to more frequent AT-transversions and less frequent AC-transversions than expected.

Table 2.3: Distribution of configurations of the four-fold degenerate sites.

| | | REV model | | TN93 model | | HKY85 model | |
|---|---|---|---|---|---|---|---|
| Configuration | Obs | Exp | $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ | Exp | $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ | Exp | $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ |
| 0-change | 200 | 188.6 | 0.689 | 181.8 | 1.820 | 184.6 | 1.293 |
| 1-TC-transition | 128 | 132.9 | 0.181 | 134.6 | 0.326 | 133.3 | 0.210 |
| 1-AG-transition | 57 | 58.3 | 0.028 | 59.0 | 0.069 | 50.9 | 0.743 |
| 1-GT-transversion | 0 | 0.4 | 0.436 | 0.3 | 0.285 | 0.3 | 0.329 |
| 1-GC-transversion | 1 | 2.9 | 1.232 | 2.8 | 1.173 | 3.0 | 1.299 |
| 1-AT-transversion | 23 | 20.9 | 0.201 | 14.1 | 5.676 | 14.4 | 5.113 |
| 1-AC-transversion | 37 | 45.0 | 1.407 | 55.3 | 6.061 | 56.2 | 6.558 |
| ≥2-changes | 165 | 162.0 | 0.056 | 163.1 | 0.023 | 168.4 | 0.069 |
| total | 611 | 611.0 | $\chi^2 = 4.23$ | 611.0 | $\chi^2 = 15.43$ | 611.0 | $\chi^2 = 15.61$ |
| | | | d.f. = 7 | | d.f. = 7 | | d.f. = 7 |
| | | | $P = 0.75$ | | $P = 0.03$ | | $P = 0.03$ |

Distribution of configurations of the four-fold degenerate nucleotide sites (611 sites) for the REV, TN93 and HKY85 models (ML estimates). The ML estimates of parameters are as follows; $\alpha/\beta = 20.29$ for the HKY85 model, and $(\alpha_Y + \alpha_R)/(2\beta) = 28.69$ and $\alpha_Y/\alpha_R = 0.40$ for the TN93 model.

It is apparent that the transition rate between purines is higher than that between pyrimidines by

about 2 times, and in terms of AIC the TN93 model better approximates the 611 sites data than the HKY85 model does. As for the branch lengths, however, the estimates from the three models do not differ significantly (Fig. 2.1), and therefore the estimates of the evolutionary rate and the branching dates would be robust to some extent to the choice among these models.

### 2.1.4 Discussion

Since the REV model fits to the four-fold degenerate sites data remarkably well when the parameters of the model are estimated by the ML, further complication of the model seems not necessary in approximating the evolution of these sites. Provided these sites are free from constraint, the transition probability matrix shown in Table 2.2 should represent the pattern of mutation in mtDNA.

However, when we deal with the data that include all the codon positions and tRNAs as Yang (1994[317]) did in analyzing mtDNA data, complications due to unequal evolutionary rate across sites and to other factors become necessary as discussed by Yang. Furthermore, even when we deal with the four-fold degenerate sites only, if the nucleotide frequency differs significantly between species, the assumption of stationarity does not hold, and then the REV model may no longer be a good approximation. This factor may become serious when we compare among different mammalian orders (Cao et al. 1994[48]).

The different nucleotide frequency between species is often a serious problem in inferring trees (Hasegawa and Hashimoto 1993[111]). Where genomes have acquired similar nucleotide frequency independently in different lineages, a wrong tree grouping together sequences with similar nucleotide frequency might be obtained. Methods to partially overcome this difficulty have been proposed by Lake (1994[182]) and Lockhart et al. (1994[197]) in the framework of distance methods, but it remains to be studied in the framework of the ML method.

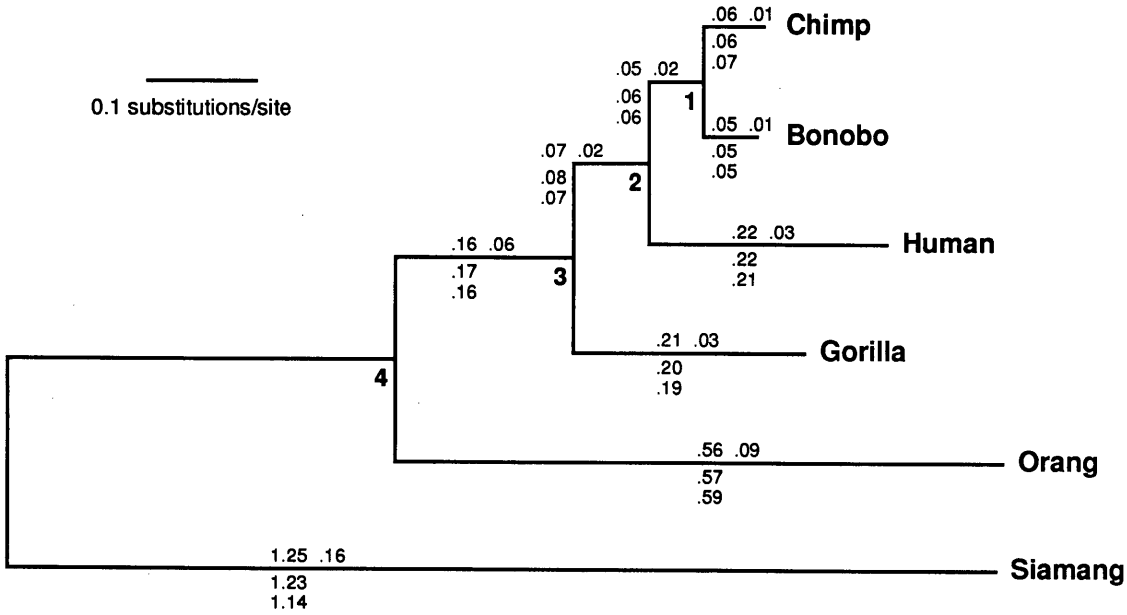



Figure 2.1: The ML tree of the four-fold degenerate sites.

The ML tree of the four-fold degenerate sites (611 sites) based on transition probability matrix of the REV model given in Table 2.2. The horizontal length of each branch is proportional to the estimated number of substitutions. The root of this tree is arbitrarily placed within 4–siamang branch. Branch length estimated by the REV model with its SE is given above a branch, and length estimated by the TN93 (Tamura and Nei 1993[288]) and HKY85 (Hasegawa et al. 1984[127], 1985[122]) models are given below a branch in this order. The NucML program in MOLPHY ver. 2.3 (Adachi and Hasegawa 1995[5]) for the ML inference of DNA or RNA phylogeny was applied.

## 2.2 Modeling of Amino Acid Substitution

### 2.2.1 Dayhoff Model

Any method for inferring molecular phylogeny assumes explicitly or implicitly a model for the fundamental process of evolution, that is, nucleotide or amino acid substitution. Clearly, the assumed model should be as realistic as possible. Dependence among neighbouring nucleotides in a codon complicates the problem in modeling the nucleotide substitution in protein-encoding genes, and it seems preferable to model the amino acid substitution.

Since the selective constraint is more likely to be operating at the codon level rather than at the individual nucleotide level, it would be more realistic to construct a model for amino acid (rather than for nucleotide) substitutions to perform phylogenetic analyses of protein-encoding genes. The transition matrices of amino acid substitutions have previously been estimated by the parsimony method for the data sets which consist mainly of nuclear-encoded proteins (Dayhoff et al. 1978[62]; Jones et al. 1992[154]).

For amino acid substitutions, $\boldsymbol{R}$ is related to the accepted mutation matrix $\boldsymbol{A}$ in Fig. 80 of Dayhoff et al. (1978[62]) by the following formula;

$$R_{ij} = A_{ij}/(20^2 \pi_i^A \pi_j^A), \tag{2.16}$$

where $\pi_i^A$ is the frequency of amino acid $i$ in the data set used in constructing $\boldsymbol{A}$ (given in Table 22 of Dayhoff et al. (1978[62])).

Below is the relative substitution rate matrix $\boldsymbol{R}$ of Dayhoff et al.

Table 2.4: Relative substitution rate matrix of Dayhoff.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | | 30 | 109 | 154 | 33 | 93 | 266 | 579 | 21 | 66 | 95 | 57 | 29 | 20 | 345 | 772 | 590 | 0 | 20 | 365 |
| Arg | 30 | | 17 | 1 | 10 | 120 | 1 | 10 | 103 | 30 | 17 | 477 | 17 | 7 | 67 | 137 | 20 | 27 | 3 | 20 |
| Asn | 109 | 17 | | 532 | 1 | 50 | 94 | 156 | 226 | 36 | 37 | 322 | 1 | 7 | 27 | 432 | 169 | 3 | 36 | 13 |
| Asp | 154 | 1 | 532 | | 0 | 76 | 831 | 162 | 43 | 13 | 1 | 85 | 1 | 0 | 10 | 98 | 57 | 0 | 1 | 17 |
| Cys | 33 | 10 | 1 | 0 | | 0 | 0 | 10 | 10 | 17 | 1 | 0 | 1 | 1 | 10 | 117 | 10 | 1 | 30 | 33 |
| Gln | 93 | 120 | 50 | 76 | 0 | | 422 | 30 | 243 | 8 | 75 | 147 | 20 | 0 | 93 | 47 | 37 | 0 | 1 | 27 |
| Glu | 266 | 1 | 94 | 831 | 0 | 422 | | 112 | 23 | 35 | 15 | 104 | 7 | 0 | 40 | 86 | 31 | 0 | 10 | 37 |
| Gly | 579 | 10 | 156 | 162 | 10 | 30 | 112 | | 10 | 1 | 17 | 60 | 7 | 17 | 49 | 450 | 50 | 1 | 0 | 97 |
| His | 21 | 103 | 226 | 43 | 10 | 243 | 23 | 10 | | 3 | 40 | 23 | 1 | 20 | 50 | 26 | 14 | 3 | 40 | 30 |
| Ile | 66 | 30 | 36 | 13 | 17 | 8 | 35 | 1 | 3 | | 253 | 43 | 57 | 90 | 7 | 20 | 129 | 0 | 13 | 661 |
| Leu | 95 | 17 | 37 | 1 | 1 | 75 | 15 | 17 | 40 | 253 | | 39 | 207 | 167 | 43 | 32 | 52 | 13 | 23 | 303 |
| Lys | 57 | 477 | 322 | 85 | 0 | 147 | 104 | 60 | 23 | 43 | 39 | | 90 | 0 | 43 | 168 | 200 | 0 | 10 | 17 |
| Met | 29 | 17 | 1 | 1 | 1 | 20 | 7 | 7 | 1 | 57 | 207 | 90 | | 17 | 4 | 20 | 28 | 0 | 0 | 77 |
| Phe | 20 | 7 | 7 | 0 | 1 | 0 | 0 | 17 | 20 | 90 | 167 | 0 | 17 | | 7 | 40 | 10 | 10 | 260 | 10 |
| Pro | 345 | 67 | 27 | 10 | 10 | 93 | 40 | 49 | 50 | 7 | 43 | 43 | 4 | 7 | | 269 | 73 | 0 | 1 | 50 |
| Ser | 772 | 137 | 432 | 98 | 117 | 47 | 86 | 450 | 26 | 20 | 32 | 168 | 20 | 40 | 269 | | 696 | 17 | 22 | 43 |
| Thr | 590 | 20 | 169 | 57 | 10 | 37 | 31 | 50 | 14 | 129 | 52 | 200 | 28 | 10 | 73 | 696 | | 0 | 23 | 186 |
| Trp | 0 | 27 | 3 | 0 | 1 | 0 | 0 | 1 | 3 | 0 | 13 | 0 | 0 | 10 | 0 | 17 | 0 | | 6 | 1 |
| Tyr | 20 | 3 | 36 | 1 | 30 | 1 | 10 | 0 | 40 | 13 | 23 | 10 | 0 | 260 | 1 | 22 | 23 | 6 | | 17 |
| Val | 365 | 20 | 13 | 17 | 33 | 27 | 37 | 97 | 30 | 661 | 303 | 17 | 77 | 10 | 50 | 43 | 186 | 1 | 17 | |

Table 2.5: Transition probability matrix for the Dayhoff model.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 98669 | 11 | 40 | 56 | 12 | 34 | 97 | 211 | 8 | 24 | 35 | 21 | 11 | 7 | 126 | 282 | 216 | 0 | 7 | 133 |
| Arg | 23 | 99137 | 13 | 1 | 8 | 93 | 1 | 8 | 80 | 23 | 13 | 370 | 13 | 5 | 52 | 106 | 16 | 21 | 2 | 16 |
| Asn | 87 | 14 | 98198 | 423 | 1 | 40 | 75 | 124 | 180 | 29 | 29 | 256 | 1 | 6 | 21 | 343 | 134 | 2 | 29 | 10 |
| Asp | 104 | 1 | 360 | 98592 | 0 | 51 | 562 | 110 | 29 | 9 | 1 | 57 | 0 | 0 | 7 | 66 | 39 | 0 | 1 | 11 |
| Cys | 32 | 10 | 1 | 0 | 99725 | 0 | 0 | 10 | 10 | 16 | 1 | 0 | 1 | 1 | 10 | 113 | 10 | 1 | 29 | 32 |
| Gln | 78 | 100 | 42 | 64 | 0 | 98754 | 353 | 25 | 203 | 7 | 63 | 123 | 17 | 0 | 78 | 39 | 31 | 0 | 1 | 23 |
| Glu | 169 | 1 | 60 | 528 | 0 | 268 | 98656 | 71 | 15 | 22 | 10 | 66 | 4 | 0 | 25 | 55 | 20 | 0 | 6 | 24 |
| Gly | 207 | 4 | 56 | 58 | 4 | 11 | 40 | 99351 | 4 | 0 | 6 | 21 | 2 | 6 | 17 | 161 | 18 | 0 | 0 | 35 |
| His | 20 | 96 | 211 | 40 | 9 | 227 | 21 | 9 | 99132 | 3 | 37 | 21 | 1 | 19 | 47 | 24 | 13 | 3 | 37 | 28 |
| Ile | 57 | 26 | 31 | 11 | 15 | 7 | 30 | 1 | 3 | 98727 | 217 | 37 | 49 | 77 | 6 | 17 | 111 | 0 | 11 | 568 |
| Leu | 36 | 6 | 14 | 0 | 0 | 28 | 6 | 6 | 15 | 95 | 99465 | 15 | 77 | 62 | 16 | 12 | 19 | 5 | 9 | 113 |
| Lys | 23 | 189 | 128 | 34 | 0 | 58 | 41 | 24 | 9 | 17 | 15 | 99251 | 36 | 0 | 17 | 67 | 79 | 0 | 4 | 7 |
| Met | 61 | 36 | 2 | 2 | 1 | 42 | 15 | 15 | 1 | 121 | 439 | 191 | 98764 | 36 | 8 | 42 | 59 | 0 | 1 | 163 |
| Phe | 16 | 6 | 6 | 0 | 1 | 0 | 0 | 14 | 16 | 71 | 133 | 0 | 14 | 99457 | 6 | 32 | 8 | 8 | 207 | 8 |
| Pro | 215 | 42 | 17 | 6 | 6 | 58 | 25 | 31 | 31 | 4 | 27 | 27 | 2 | 4 | 99260 | 168 | 45 | 0 | 0 | 31 |
| Ser | 350 | 62 | 196 | 44 | 53 | 21 | 39 | 204 | 12 | 9 | 15 | 76 | 9 | 18 | 122 | 98415 | 316 | 8 | 10 | 20 |
| Thr | 323 | 11 | 93 | 31 | 5 | 20 | 17 | 27 | 8 | 71 | 28 | 110 | 15 | 5 | 40 | 381 | 98699 | 0 | 13 | 102 |
| Trp | 1 | 86 | 10 | 0 | 3 | 1 | 1 | 3 | 10 | 1 | 41 | 1 | 1 | 32 | 1 | 54 | 1 | 99733 | 19 | 2 |
| Tyr | 21 | 3 | 38 | 1 | 32 | 1 | 11 | 0 | 42 | 14 | 24 | 11 | 0 | 275 | 1 | 23 | 24 | 6 | 99453 | 18 |
| Val | 178 | 10 | 6 | 8 | 16 | 13 | 18 | 47 | 15 | 323 | 148 | 8 | 38 | 5 | 24 | 21 | 91 | 0 | 8 | 99020 |
| $\pi$ | .087 | .041 | .040 | .047 | .033 | .038 | .050 | .089 | .034 | .037 | .085 | .080 | .015 | .040 | .051 | .070 | .058 | .010 | .030 | .065 |

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid $i$ being replaced by the amino acid $j$ during a time interval of one substitution per 100 amino acids (1PAM) for the Dayhoff model, and average amino acid frequencies $\pi$ of Dayhoff.

Table 2.6: Transition probability matrix for the Dayhoff-F model of mtDNA-encoded proteins.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 98759 | 5 | 37 | 22 | 2 | 21 | 45 | 128 | 6 | 54 | 65 | 6 | 36 | 11 | 128 | 274 | 309 | 0 | 8 | 85 |
| Arg | 18 | 99362 | 12 | 0 | 1 | 58 | 0 | 5 | 63 | 52 | 25 | 102 | 44 | 8 | 53 | 103 | 22 | 58 | 2 | 10 |
| Asn | 69 | 6 | 98697 | 162 | 0 | 25 | 34 | 75 | 141 | 64 | 55 | 71 | 3 | 8 | 22 | 333 | 193 | 7 | 30 | 7 |
| Asp | 82 | 0 | 336 | 99028 | 0 | 32 | 259 | 66 | 23 | 20 | 1 | 16 | 2 | 0 | 7 | 64 | 55 | 0 | 1 | 7 |
| Cys | 25 | 4 | 1 | 0 | 99728 | 0 | 0 | 6 | 8 | 37 | 2 | 0 | 2 | 1 | 10 | 109 | 14 | 3 | 30 | 20 |
| Gln | 62 | 44 | 39 | 24 | 0 | 99093 | 163 | 15 | 160 | 15 | 118 | 34 | 56 | 0 | 79 | 38 | 44 | 1 | 1 | 14 |
| Glu | 134 | 0 | 56 | 202 | 0 | 168 | 99156 | 43 | 12 | 50 | 18 | 18 | 15 | 0 | 26 | 53 | 28 | 1 | 7 | 15 |
| Gly | 164 | 2 | 52 | 22 | 1 | 7 | 18 | 99474 | 3 | 1 | 11 | 6 | 8 | 9 | 18 | 156 | 26 | 1 | 0 | 22 |
| His | 16 | 42 | 197 | 15 | 2 | 142 | 10 | 6 | 99305 | 6 | 70 | 6 | 2 | 27 | 48 | 24 | 19 | 8 | 39 | 18 |
| Ile | 45 | 11 | 29 | 4 | 3 | 4 | 14 | 0 | 2 | 98638 | 407 | 10 | 165 | 111 | 6 | 17 | 159 | 1 | 12 | 362 |
| Leu | 28 | 3 | 13 | 0 | 0 | 18 | 3 | 4 | 12 | 211 | 99205 | 4 | 260 | 90 | 16 | 12 | 28 | 14 | 9 | 72 |
| Lys | 18 | 82 | 119 | 13 | 0 | 36 | 19 | 14 | 7 | 38 | 29 | 99298 | 120 | 0 | 17 | 65 | 114 | 0 | 4 | 4 |
| Met | 49 | 16 | 2 | 1 | 0 | 26 | 7 | 9 | 1 | 269 | 822 | 53 | 98453 | 52 | 9 | 41 | 85 | 1 | 1 | 104 |
| Phe | 13 | 2 | 5 | 0 | 0 | 0 | 0 | 8 | 13 | 159 | 249 | 0 | 45 | 99214 | 6 | 31 | 11 | 22 | 216 | 5 |
| Pro | 170 | 18 | 16 | 2 | 1 | 36 | 11 | 18 | 25 | 10 | 50 | 7 | 8 | 6 | 99371 | 163 | 65 | 0 | 0 | 20 |
| Ser | 277 | 27 | 183 | 17 | 10 | 13 | 18 | 123 | 9 | 20 | 27 | 21 | 31 | 26 | 124 | 98574 | 453 | 21 | 10 | 12 |
| Thr | 256 | 5 | 86 | 12 | 1 | 13 | 8 | 17 | 6 | 158 | 53 | 30 | 52 | 8 | 41 | 371 | 98806 | 0 | 13 | 65 |
| Trp | 1 | 37 | 9 | 0 | 1 | 1 | 0 | 2 | 8 | 2 | 77 | 0 | 2 | 46 | 1 | 52 | 1 | 99739 | 20 | 1 |
| Tyr | 17 | 1 | 36 | 0 | 6 | 1 | 5 | 0 | 33 | 31 | 46 | 3 | 2 | 396 | 1 | 23 | 35 | 18 | 99337 | 11 |
| Val | 141 | 4 | 6 | 3 | 3 | 8 | 8 | 29 | 12 | 721 | 278 | 2 | 127 | 7 | 25 | 20 | 130 | 1 | 9 | 98466 |
| $\pi$ | .072 | .019 | .039 | .019 | .006 | .025 | .024 | .056 | .028 | .087 | .168 | .023 | .053 | .060 | .055 | .072 | .088 | .029 | .033 | .044 |

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid $i$ being replaced by the amino acid $j$ during a time interval of one substitution per 100 amino acids (1PAM) for the Dayhoff-F model, and average amino acid frequencies $\pi$ of the mtDNA-encoded proteins.

### 2.2.2 JTT Model

Below is the relative substitution rate matrix $\boldsymbol{R}$ of Jones et al. (1992[154]).

Table 2.7: Relative substitution rate matrix of JTT.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | | 247 | 216 | 386 | 106 | 208 | 600 | 1183 | 46 | 173 | 257 | 200 | 100 | 51 | 901 | 2413 | 2440 | 11 | 41 | 1766 |
| Arg | 247 | | 116 | 48 | 125 | 750 | 119 | 614 | 446 | 76 | 205 | 2348 | 61 | 16 | 217 | 413 | 230 | 109 | 46 | 69 |
| Asn | 216 | 116 | | 1433 | 32 | 159 | 180 | 291 | 466 | 130 | 63 | 758 | 39 | 15 | 31 | 1738 | 693 | 2 | 114 | 55 |
| Asp | 386 | 48 | 1433 | | 13 | 130 | 2914 | 577 | 144 | 37 | 34 | 102 | 27 | 8 | 39 | 244 | 151 | 5 | 89 | 127 |
| Cys | 106 | 125 | 32 | 13 | | 9 | 8 | 98 | 40 | 19 | 36 | 7 | 23 | 66 | 15 | 353 | 66 | 38 | 164 | 99 |
| Gln | 208 | 750 | 159 | 130 | 9 | | 1027 | 84 | 635 | 20 | 314 | 858 | 52 | 9 | 395 | 182 | 149 | 12 | 40 | 58 |
| Glu | 600 | 119 | 180 | 2914 | 8 | 1027 | | 610 | 41 | 43 | 65 | 754 | 30 | 13 | 71 | 156 | 142 | 12 | 15 | 226 |
| Gly | 1183 | 614 | 291 | 577 | 98 | 84 | 610 | | 41 | 25 | 56 | 142 | 27 | 18 | 93 | 1131 | 164 | 69 | 15 | 276 |
| His | 46 | 446 | 466 | 144 | 40 | 635 | 41 | 41 | | 26 | 134 | 85 | 21 | 50 | 157 | 138 | 76 | 5 | 514 | 22 |
| Ile | 173 | 76 | 130 | 37 | 19 | 20 | 43 | 25 | 26 | | 1324 | 75 | 704 | 196 | 31 | 172 | 930 | 12 | 61 | 3938 |
| Leu | 257 | 205 | 63 | 34 | 36 | 314 | 65 | 56 | 134 | 1324 | | 94 | 974 | 1093 | 578 | 436 | 172 | 82 | 84 | 1261 |
| Lys | 200 | 2348 | 758 | 102 | 7 | 858 | 754 | 142 | 85 | 75 | 94 | | 103 | 7 | 77 | 228 | 398 | 9 | 20 | 58 |
| Met | 100 | 61 | 39 | 27 | 23 | 52 | 30 | 27 | 21 | 704 | 974 | 103 | | 49 | 23 | 54 | 343 | 8 | 17 | 559 |
| Phe | 51 | 16 | 15 | 8 | 66 | 9 | 13 | 18 | 50 | 196 | 1093 | 7 | 49 | | 36 | 309 | 39 | 37 | 850 | 189 |
| Pro | 901 | 217 | 31 | 39 | 15 | 395 | 71 | 93 | 157 | 31 | 578 | 77 | 23 | 36 | | 1138 | 412 | 6 | 22 | 84 |
| Ser | 2413 | 413 | 1738 | 244 | 353 | 182 | 156 | 1131 | 138 | 172 | 436 | 228 | 54 | 309 | 1138 | | 2258 | 36 | 164 | 219 |
| Thr | 2440 | 230 | 693 | 151 | 66 | 149 | 142 | 164 | 76 | 930 | 172 | 398 | 343 | 39 | 412 | 2258 | | 8 | 45 | 526 |
| Trp | 11 | 109 | 2 | 5 | 38 | 12 | 12 | 69 | 5 | 12 | 82 | 9 | 8 | 37 | 6 | 36 | 8 | | 41 | 27 |
| Tyr | 41 | 46 | 114 | 89 | 164 | 40 | 15 | 15 | 514 | 61 | 84 | 20 | 17 | 850 | 22 | 164 | 45 | 41 | | 42 |
| Val | 1766 | 69 | 55 | 127 | 99 | 58 | 226 | 276 | 22 | 3938 | 1261 | 58 | 559 | 189 | 84 | 219 | 526 | 27 | 42 | |

Table 2.8: Transition probability matrix for the JTT model.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 98755 | 27 | 24 | 42 | 12 | 23 | 66 | 130 | 5 | 19 | 28 | 22 | 11 | 6 | 99 | 265 | 268 | 1 | 4 | 194 |
| Arg | 41 | 98964 | 19 | 8 | 21 | 124 | 20 | 102 | 74 | 13 | 34 | 389 | 10 | 3 | 36 | 68 | 38 | 18 | 8 | 11 |
| Asn | 42 | 23 | 98717 | 282 | 6 | 31 | 35 | 57 | 92 | 26 | 12 | 149 | 8 | 3 | 6 | 341 | 136 | 0 | 22 | 11 |
| Asp | 63 | 8 | 233 | 98943 | 2 | 21 | 473 | 94 | 23 | 6 | 6 | 17 | 4 | 1 | 6 | 40 | 25 | 1 | 14 | 21 |
| Cys | 45 | 53 | 14 | 5 | 99444 | 4 | 3 | 41 | 17 | 8 | 15 | 3 | 10 | 28 | 6 | 149 | 28 | 16 | 69 | 42 |
| Gln | 43 | 155 | 33 | 27 | 2 | 98951 | 212 | 17 | 131 | 4 | 65 | 177 | 11 | 2 | 81 | 37 | 31 | 2 | 8 | 12 |
| Glu | 82 | 16 | 25 | 397 | 1 | 140 | 99043 | 83 | 6 | 6 | 9 | 103 | 4 | 2 | 10 | 21 | 19 | 2 | 2 | 31 |
| Gly | 135 | 70 | 33 | 66 | 11 | 10 | 70 | 99371 | 5 | 3 | 6 | 16 | 3 | 2 | 11 | 129 | 19 | 8 | 2 | 32 |
| His | 17 | 164 | 171 | 53 | 15 | 233 | 15 | 15 | 98866 | 10 | 49 | 31 | 8 | 18 | 58 | 51 | 28 | 2 | 189 | 8 |
| Ile | 28 | 12 | 21 | 6 | 3 | 3 | 7 | 4 | 4 | 98702 | 215 | 12 | 114 | 32 | 5 | 28 | 151 | 2 | 10 | 640 |
| Leu | 24 | 19 | 6 | 3 | 3 | 29 | 6 | 5 | 12 | 123 | 99326 | 9 | 90 | 101 | 54 | 40 | 16 | 8 | 8 | 117 |
| Lys | 29 | 336 | 109 | 15 | 1 | 123 | 108 | 20 | 12 | 11 | 13 | 99095 | 15 | 1 | 11 | 33 | 57 | 1 | 3 | 8 |
| Met | 35 | 21 | 14 | 10 | 8 | 18 | 11 | 10 | 7 | 248 | 343 | 36 | 98869 | 17 | 8 | 19 | 121 | 3 | 6 | 197 |
| Phe | 11 | 3 | 3 | 2 | 14 | 2 | 3 | 4 | 11 | 41 | 231 | 1 | 10 | 99356 | 8 | 65 | 8 | 8 | 180 | 40 |
| Pro | 149 | 36 | 5 | 6 | 2 | 65 | 12 | 15 | 26 | 5 | 96 | 13 | 4 | 6 | 99283 | 188 | 68 | 1 | 4 | 14 |
| Ser | 295 | 51 | 213 | 30 | 43 | 22 | 19 | 138 | 17 | 21 | 53 | 28 | 7 | 38 | 139 | 98558 | 276 | 4 | 20 | 27 |
| Thr | 349 | 33 | 99 | 22 | 9 | 21 | 20 | 23 | 11 | 133 | 25 | 57 | 49 | 6 | 59 | 323 | 98677 | 1 | 6 | 75 |
| Trp | 7 | 66 | 1 | 3 | 23 | 7 | 7 | 42 | 3 | 7 | 49 | 5 | 5 | 22 | 4 | 22 | 5 | 99681 | 25 | 16 |
| Tyr | 11 | 12 | 30 | 23 | 43 | 11 | 4 | 4 | 136 | 16 | 22 | 5 | 4 | 224 | 6 | 43 | 12 | 11 | 99371 | 11 |
| Val | 226 | 9 | 7 | 16 | 13 | 7 | 29 | 35 | 3 | 504 | 161 | 7 | 72 | 24 | 11 | 28 | 67 | 3 | 5 | 98771 |
| $\pi$ | .077 | .051 | .043 | .052 | .020 | .041 | .062 | .074 | .023 | .052 | .091 | .059 | .024 | .040 | .051 | .069 | .059 | .014 | .032 | .066 |

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid $i$ being replaced by the amino acid $j$ during a time interval of one substitution per 100 amino acids (1PAM) for the JTT model, and average amino acid frequencies $\pi$ of the proteins used by Jones et al. (1992[154]).

Table 2.9: Transition probability matrix for the JTT-F model of mtDNA-encoded proteins.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 98826 | 9 | 21 | 15 | 4 | 13 | 25 | 94 | 6 | 30 | 49 | 8 | 23 | 8 | 101 | 261 | 378 | 2 | 4 | 122 |
| Arg | 37 | 99239 | 17 | 3 | 6 | 72 | 7 | 74 | 86 | 20 | 60 | 146 | 21 | 4 | 37 | 68 | 54 | 36 | 7 | 7 |
| Asn | 38 | 8 | 98971 | 97 | 2 | 18 | 13 | 42 | 107 | 41 | 22 | 56 | 16 | 4 | 6 | 337 | 192 | 1 | 22 | 7 |
| Asp | 56 | 3 | 203 | 99308 | 1 | 12 | 176 | 68 | 27 | 10 | 10 | 6 | 9 | 2 | 6 | 39 | 35 | 2 | 14 | 13 |
| Cys | 40 | 19 | 12 | 2 | 99454 | 2 | 1 | 30 | 20 | 13 | 27 | 1 | 20 | 40 | 6 | 147 | 39 | 32 | 68 | 26 |
| Gln | 38 | 54 | 29 | 9 | 1 | 99230 | 79 | 13 | 152 | 7 | 113 | 66 | 23 | 3 | 83 | 37 | 43 | 5 | 8 | 8 |
| Glu | 73 | 6 | 21 | 137 | 0 | 81 | 99457 | 60 | 7 | 9 | 16 | 39 | 9 | 3 | 10 | 21 | 27 | 3 | 2 | 19 |
| Gly | 121 | 25 | 29 | 23 | 3 | 6 | 26 | 99529 | 5 | 5 | 11 | 6 | 6 | 3 | 11 | 127 | 26 | 16 | 2 | 20 |
| His | 15 | 57 | 149 | 18 | 4 | 135 | 6 | 11 | 99106 | 15 | 86 | 12 | 16 | 26 | 59 | 50 | 39 | 4 | 185 | 5 |
| Ile | 25 | 4 | 18 | 2 | 1 | 2 | 3 | 3 | 5 | 98606 | 377 | 5 | 241 | 46 | 5 | 28 | 213 | 4 | 10 | 402 |
| Leu | 21 | 7 | 5 | 1 | 1 | 17 | 2 | 4 | 14 | 195 | 99179 | 3 | 190 | 146 | 55 | 40 | 23 | 15 | 8 | 74 |
| Lys | 26 | 118 | 94 | 5 | 0 | 71 | 40 | 15 | 14 | 17 | 24 | 99409 | 31 | 1 | 11 | 32 | 81 | 3 | 3 | 5 |
| Met | 32 | 8 | 12 | 3 | 2 | 11 | 4 | 7 | 9 | 394 | 601 | 14 | 98547 | 25 | 8 | 19 | 171 | 6 | 6 | 124 |
| Phe | 10 | 1 | 3 | 1 | 4 | 1 | 1 | 3 | 12 | 66 | 405 | 1 | 22 | 99171 | 8 | 64 | 12 | 16 | 176 | 25 |
| Pro | 134 | 13 | 4 | 2 | 1 | 38 | 4 | 11 | 30 | 8 | 168 | 5 | 8 | 9 | 99268 | 186 | 96 | 2 | 4 | 9 |
| Ser | 265 | 18 | 185 | 10 | 13 | 13 | 7 | 101 | 20 | 33 | 94 | 10 | 14 | 54 | 142 | 98585 | 391 | 9 | 20 | 17 |
| Thr | 313 | 12 | 86 | 7 | 3 | 12 | 8 | 17 | 13 | 212 | 43 | 21 | 103 | 8 | 60 | 319 | 98706 | 2 | 6 | 47 |
| Trp | 6 | 23 | 1 | 1 | 7 | 4 | 3 | 30 | 4 | 12 | 87 | 2 | 10 | 32 | 4 | 21 | 7 | 99712 | 24 | 10 |
| Tyr | 10 | 4 | 26 | 8 | 13 | 6 | 1 | 3 | 158 | 26 | 39 | 2 | 9 | 323 | 6 | 43 | 17 | 22 | 99277 | 7 |
| Val | 202 | 3 | 6 | 6 | 4 | 4 | 11 | 26 | 3 | 801 | 283 | 3 | 151 | 35 | 11 | 28 | 95 | 7 | 5 | 98316 |
| $\pi$ | .072 | .019 | .039 | .019 | .006 | .025 | .024 | .056 | .028 | .087 | .168 | .023 | .053 | .060 | .055 | .072 | .088 | .029 | .033 | .044 |

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid $i$ being replaced by the amino acid $j$ during a time interval of one substitution per 100 amino acids (1PAM) for the JTT-F model, and average amino acid frequencies $\pi$ of the mtDNA-encoded proteins.

### 2.2.3 General Reversible Markov Model for Mitochondrial Proteins

The transition matrices of Dayhoff et al. (1978[62]) and Jones et al. (1992[154]) were estimated by the parsimony method for the data sets which consist mainly of nuclear-encoded proteins. However, the parsimony method sometimes gives a biased estimate of the transition probability matrix (Collins et al. 1994[57]; Perna and Kocher 1995[235]).

Collins et al. (1994[57]) pointed out that, in the presence of compositional bias, the transition probability matrix estimated by the parsimony might be systematically distorted. From the method, common-to-rare state changes tend to predominate over rare-to-common changes, and therefore in the common ancestral node the estimated compositional bias tends to be more extreme than those of the contemporary species. By using the cytochrome *b* gene sequences from the gastropods (their original data) and from the pecoran ruminants (Irwin et al. 1991[146]), they demonstrated this trend for both of the data sets. It is clear that this is due to the bias of the parsimony in inferring the ancestral state when the compositional bias exists. Perna and Kocher (1995[235]) also demonstrated the same characteristic of the parsimony. Furthermore, the parsimony method has no time structure (Goldman 1990[92]), and therefore it is desirable to estimate the matrix by using the ML method (Yang 1994[317]).

Recently, Naylor et al. (1995[223]) have pointed out that, since the bias for T and C at second codon position is directly correlated with hydrophobicity of an encoded amino acid and since mtDNA-encoded proteins contain a high proportion of hydrophobic amino acids, the second codon positions of mtDNA, hitherto regarded as herhaps the most reliable for inferring evolutionary histories of distantly related species, may actually carry less phylogenetic information than the more fast-evolving first positions whose compositional bias is less skewed. Thus, it seems difficult to take fully into account different constraints operating on different codon positions when the analysis is carried out at the nucleotide sequence level.

Recently, mtDNA sequences encoding proteins have been widely used for inferring the phylogenetic relationships among species (Thomas and Beckenbach 1989[292]; Irwin et al. 1991[146]; Ruvolo et al. 1991[249]; Edwards et al. 1991[72]; Normark et al. 1991[227]; Horai et al. 1992[141]; Garza and Woodruf 1992[88]; Richman and Price 1992[245]; Liu and Beckenbach 1992[195]; Pashley and Ke 1992[234]; DeWalt et al. 1993[66]; Thomas and Martin 1993[293]; Ma et al. 1993[199]; Hedges et al. 1993[135]; Kornegay et al. 1993[173]; Kusmierski et al. 1993[181]; Martin 1993[203]; Block et al. 1993[40]; Avise 1994[28]; Avise et al. 1994[29]; Krajewski and Fetzner 1994[174]; Lanyon 1994[183]; Janke et al. 1994[150]; Cao et al. 1994[49], 1994[48]; Miyamoto et al. 1994[215]; Yokobori et al. 1994[318]; Meyer 1994[209]; Meyer et al. 1994[210]; Hafner et al. 1994[103]; Stern 1994[277]; Arevalo et al. 1994[19]; Weller et al. 1994[306]; Irwin and Árnason 1994[145]; Adkins and Honeycutt 1994[10]; Milinkovitch et al. 1994[211]; Árnason and Gullberg 1994[22]; Árnason et al. 1995[20]; Lento et al. 1995[190]). However, since the mitochondrial code is different from the universal code and since most of the mtDNA-encoded proteins are membranous, the transition probability matrix of the mtDNA-encoded proteins might be different

from that estimated from nuclear-encoded proteins. Thus, it seems desirable to model the amino acid substitution of mtDNA-encoded proteins, and therefore I estimated the 20 × 20 transition probability matrix of the general reversible Markov model (the REV model) for mtDNA-encoded proteins (the mtREV model) by the ML method. This model is an extention to amino acid of the general reversible Markov model of nucleotide substitution proposed by Yang (1994[317]). The matrix was estimated by the ML method from the complete sequence data of mtDNA from 20 vertebrate species. This matrix represents the substitution pattern of the mtDNA-encoded proteins, and shows some differences from the matrix estimated from the nuclear-encoded proteins. The use of this matrix would be recommended in inferring trees from mtDNA-encoded protein sequences by the ML method (Adachi and Hasegawa 1995[5]).

### 2.2.4 Mitochondrial DNA Sequence Data

The matrix was estimated through ML by using the complete mtDNA sequences from the 20 vertebrate species (3 individuals from human) listed in Table 2.10. Only the 12 proteins encoded in the same strand of mtDNA were used and NADH dehydrogenase subunit 6 was omitted, because it is coded on the complementary strand and thus has different nucleotide and accordingly different amino acid compositions (Hasegawa and Kishino 1989[118]). Positions with gaps and regions where the alignment was ambiguous were excluded.

Table 2.10: List of data used in estimating the mtREV matrix.

| Abbrev. | species name | | reference | database |
|---|---|---|---|---|
| Bosta | *Bos taurus* | cow | Anderson et al. 1982[16] | V00654 |
| Balph | *Balaenoptera physalus* | fin whale | Árnason et al. 1991[24] | X61145 |
| Balmu | *Balaenoptera musculus* | blue whale | Árnason and Gullberg 1993[21] | X72204 |
| Phovi | *Phoca vitulina* | harbor seal | Árnason and Johnsson 1992[25] | X63726 |
| Halgr | *Halichoerus grypus* | grey seal | Árnason et al. 1993[23] | X72004 |
| Equca | *Equus caballus* | horse | Xu and Árnason 1994[313] | X79547 |
| Musmu | *Mus musculus* | mouse | Bibb et al. 1981[39] | P00158 |
| Ratno | *Rattus norvegicus* | rat | Gadaleta et al. 1989[87] | P00159 |
| Anderson | *Homo sapiens* | European | Anderson et al. 1981[15] | J01415* |
| DCM1 | *Homo sapiens* | Japanese | Ozawa et al. 1991[232] | |
| SB17F | *Homo sapiens* | African | Horai et al. 1995[140] | D38112 |
| Pantr | *Pan troglodytes* | chimpanzee | Horai et al. 1995[140] | D38113 |
| Panpa | *Pan paniscus* | bonobo | Horai et al. 1995[140] | D38116 |
| Gorgo | *Gorilla gorilla* | gorilla | Horai et al. 1995[140] | D38114 |
| Ponpy | *Pongo pygmaeus* | orangutan | Horai et al. 1995[140] | D38115 |
| Didvi | *Didelphis virginiana* | opossum | Janke et al. 1994[150] | Z29573 |
| Galga | *Gallus gallus* | chicken | Desjardins and Morais 1990[65] | P18946 |
| Xenla | *Xenopus laevis* | clawed frog | Roe et al. 1985[248] | X02890 |
| Cypca | *Cyprinus carpio* | carp | Chang et al. 1994[54] | X61010 |
| Crola | *Crossostoma lacustre* | loach | Tzeng et al. 1992[297] | M91245 |
| Oncmy | *Oncorhynchus mykiss* | trout | Zardaya et al. 1995[319]) | L29771 |
| Petma | *Petromyzon marinus* | sea lamprey | Lee and Kocher 1995[187] | U11880 |

*: revised according to Horai et al. (1995[140]).

Overlapping regions between ATPase subunits 6 and 8, and between NADH dehydrogenase subunits 4 and 4L were also excluded. The following protein-encoding regions were used in this work: ND1 (3322–4050, 4054–4251 in the numbering of Anderson et al. (1981[15])), ND2 (4473–5180, 5184–5423, 5430–5447, 5451–5456, 5460–5471, 5475–5483), COI (5907–6350, 6354–7421), COII (7589–7735, 7739–8245), ATPase8 (8369–8446, 8474–8497, 8501–8503, 8507–8524), ATPase6 (8575–8607, 8644–8703, 8707–8880, 8884–8985, 8989–9030, 9040–9081, 9088–9204), COIII (9210–9272, 9276–9914, 9918–9920, 9924–9989), ND3 (10092–10109, 10116–10154, 10164–10400), ND4L (10476–10496, 10503–10646, 10659–10757), ND4 (10769–11035, 11039–11677, 11690–12007, 12011–12127), ND5 (12355–12372, 12469–12933, 12973–13299, 13303–13680, 13684–13827, 13900–13992, 13996–14028, 14074–14109), and Cyt-b (14750–15598, 15602–15880). The total number of deduced amino acid sites was 3357.

### 2.2.5 Transition Probability Matrix of the mtREV Model

Provided the tree topology which generated the amino acid sequence data $\boldsymbol{X}$ is known, we can estimate the relative substitution rate $\boldsymbol{R}$ and numbers of nucleotide substitutions along each branches, $t_1, \ldots, t_m$ ($m$: number of branches in the tree) by the ML;

$$\text{maximize} \quad l(\boldsymbol{R}, \boldsymbol{t}|\boldsymbol{X}) \tag{2.17}$$

where $l$ is a likelihood function and $\boldsymbol{t} = [t_1, t_2, \ldots, t_m]^T$.

At first we give the initial value of $\boldsymbol{R}$ by assuming the proportional model and that of $\boldsymbol{t}$ as the ML estimate under the model. Then we iterate ML estimations of $\boldsymbol{R}$ by the Brent method and of $\boldsymbol{t}$ by the Newton-Raphson method alternately. At a step of iteration when the differences of all parameters between the preceding two steps are less than $\epsilon$, we stop the procedure.

Fig. 2.2 shows the unrooted tree (Cao et al. 1994[49]; Horai et al. 1995[140]), among species from which complete mtDNA sequences are available, assumed in the estimation of the transition probability matrix. The placing of lamprey as in this figure is not the ML tree but the 2nd highest likelihood tree, and ((Birds, Mammals), (Xenopus, Fishes), Lamprey) shown in Fig. 2.3 is the ML tree. However, since the difference of log-likelihood of this tree from that of the ML tree is minor (9.6 ± 15.6 where ± is 1SE estimated by the formula in Kishino and Hasegawa (1989[164]), we used this biological tree. Since the branching order among Carnivora, Perissodactyla, and the Cetacea/Artiodactyla clade cannot be resolved by the mtDNA data, it was left as a trifurcation. The estimated transition probability matrix is not sensitive to the assumed tree (shown below, and Yang (1994[317])). The log-likelihood of this tree for the mtREV model is −46240, while that for the JTT-F model is −47039, showing much improved fitting of the mtREV model to the mtDNA-encoded protein data.

Table 2.11 is the relative substitution rate matrix $\boldsymbol{R}$ of the mtREV model, and Table 2.12 gives the estimated transition probability matrix for the mtREV model. Table 2.13 shows the difference of the transition probability of the mtREV model estimated from the highest likelihood tree (Fig. 2.3) from

that given in Table 2.12. The differences are minor, suggesting that the estimatd transition probability matrix is robust to the violation of the assumed tree. Table 2.9 gives transition probability matrix of Jones, Taylor and Thornton's (1992[154]) model of nuclear-encoded proteins adjusted with the amino acid frequencies of the mtDNA-encoded proteins (given in the last row of Table 2.12 as the equilibrium frequencies (JTT-F model; Cao et al. 1994[49]; Adachi and Hasegawa 1995[5]).

Table 2.11: Relative substitution rate matrix of mtREV model.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | | 71 | 176 | 74 | 368 | 22 | 73 | 862 | 108 | 731 | 160 | 1 | 903 | 60 | 364 | 2569 | 3289 | 1 | 43 | 1173 |
| Arg | 71 | | 72 | 1 | 788 | 1542 | 1 | 176 | 1024 | 1 | 103 | 955 | 1 | 37 | 197 | 50 | 21 | 167 | 1 | 59 |
| Asn | 176 | 72 | | 5374 | 262 | 950 | 538 | 339 | 3057 | 189 | 136 | 3325 | 241 | 50 | 539 | 3458 | 1391 | 71 | 1013 | 78 |
| Asp | 74 | 1 | 5374 | | 1 | 414 | 4204 | 431 | 915 | 49 | 5 | 82 | 1 | 46 | 62 | 449 | 204 | 63 | 72 | 1 |
| Cys | 368 | 788 | 262 | 1 | | 263 | 1 | 237 | 1094 | 371 | 272 | 1 | 1 | 497 | 111 | 2114 | 1211 | 258 | 1722 | 1 |
| Gln | 22 | 1542 | 950 | 414 | 263 | | 2270 | 52 | 3992 | 87 | 258 | 3179 | 362 | 206 | 942 | 458 | 675 | 1 | 248 | 82 |
| Glu | 73 | 1 | 538 | 4204 | 1 | 2270 | | 142 | 328 | 1 | 1 | 1914 | 1 | 1 | 51 | 414 | 100 | 1 | 137 | 161 |
| Gly | 862 | 176 | 339 | 431 | 237 | 52 | 142 | | 1 | 51 | 8 | 102 | 9 | 1 | 1 | 853 | 61 | 56 | 1 | 20 |
| His | 108 | 1024 | 3057 | 915 | 1094 | 3992 | 328 | 1 | | 115 | 76 | 484 | 1 | 265 | 289 | 426 | 381 | 52 | 4549 | 1 |
| Ile | 731 | 1 | 189 | 49 | 371 | 87 | 1 | 51 | 115 | | 2102 | 66 | 3085 | 501 | 84 | 247 | 2466 | 1 | 199 | 8009 |
| Leu | 160 | 103 | 136 | 5 | 272 | 258 | 1 | 8 | 76 | 2102 | | 56 | 3488 | 1481 | 281 | 522 | 778 | 218 | 282 | 579 |
| Lys | 1 | 955 | 3325 | 82 | 1 | 3179 | 1914 | 102 | 484 | 66 | 56 | | 492 | 63 | 310 | 638 | 885 | 203 | 339 | 11 |
| Met | 903 | 1 | 241 | 1 | 1 | 362 | 1 | 9 | 1 | 3085 | 3488 | 492 | | 532 | 106 | 777 | 3525 | 166 | 203 | 2687 |
| Phe | 60 | 37 | 50 | 46 | 497 | 206 | 1 | 1 | 265 | 501 | 1481 | 63 | 532 | | 117 | 446 | 195 | 63 | 2847 | 41 |
| Pro | 364 | 197 | 539 | 62 | 111 | 942 | 51 | 1 | 289 | 84 | 281 | 310 | 106 | 117 | | 1056 | 835 | 35 | 101 | 59 |
| Ser | 2569 | 50 | 3458 | 449 | 2114 | 458 | 414 | 853 | 426 | 247 | 522 | 638 | 777 | 446 | 1056 | | 3909 | 220 | 362 | 1 |
| Thr | 3289 | 21 | 1391 | 204 | 1211 | 675 | 100 | 61 | 381 | 2466 | 778 | 885 | 3525 | 195 | 835 | 3909 | | 106 | 190 | 1384 |
| Trp | 1 | 167 | 71 | 63 | 258 | 1 | 1 | 56 | 52 | 1 | 218 | 203 | 166 | 63 | 35 | 220 | 106 | | 167 | 42 |
| Tyr | 43 | 1 | 1013 | 72 | 1722 | 248 | 137 | 1 | 4549 | 199 | 282 | 339 | 203 | 2847 | 101 | 362 | 190 | 167 | | 37 |
| Val | 1173 | 59 | 78 | 1 | 1 | 82 | 161 | 20 | 1 | 8009 | 579 | 11 | 2687 | 41 | 59 | 1 | 1384 | 42 | 37 | |

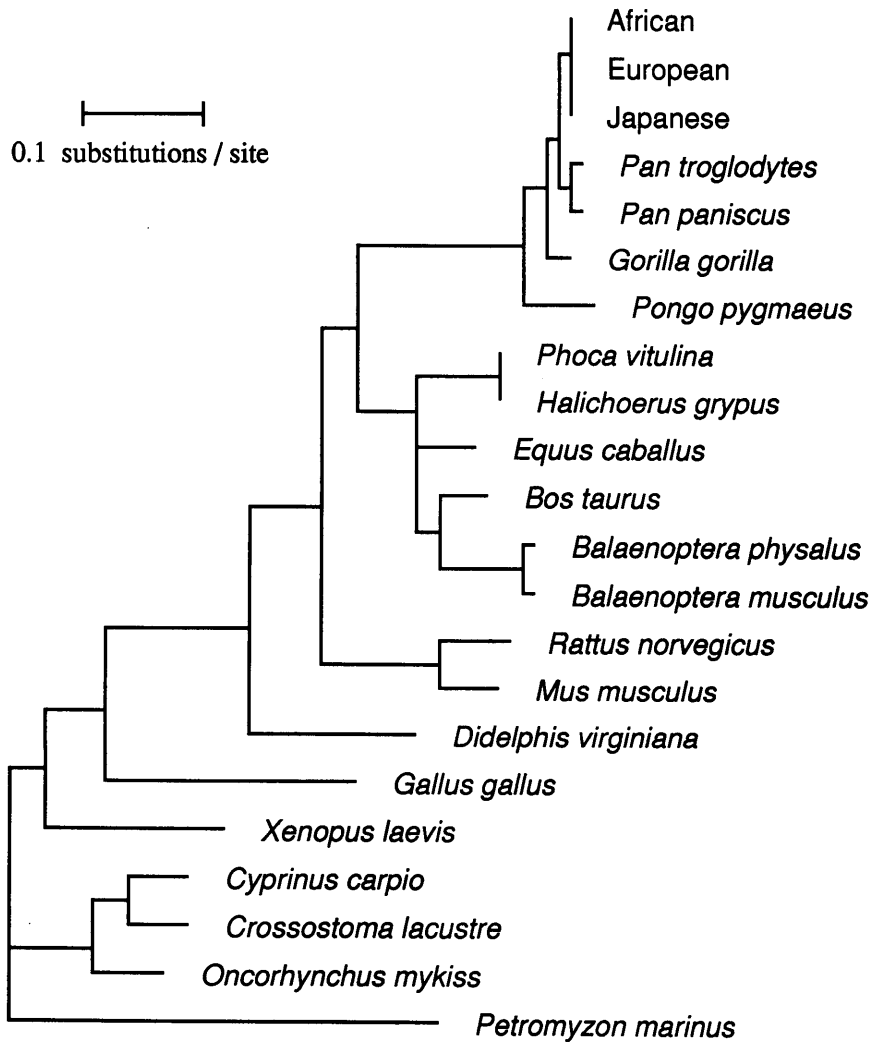



Figure 2.2: The tree used in estimating the transition probability matrix of the mtREV model.

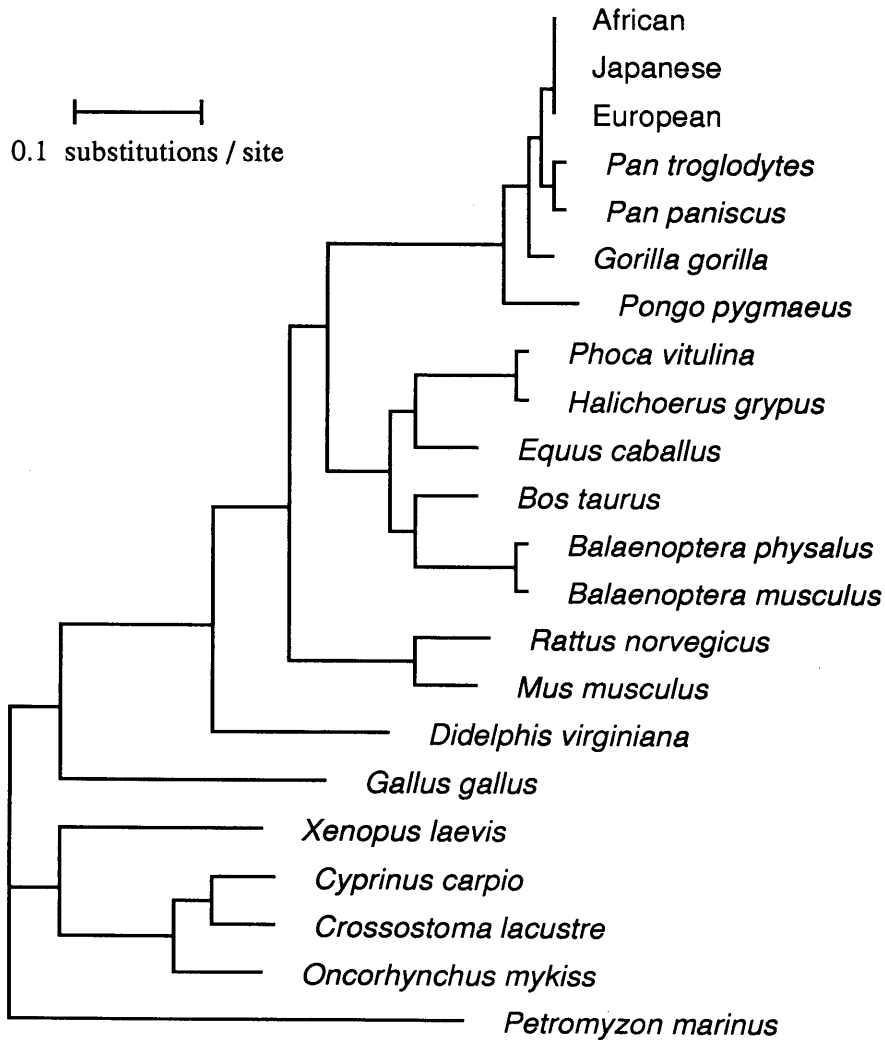



Figure 2.3: The ML tree of mtDNA-encoded proteins.

Table 2.12: Transition probability matrix for the mtREV model.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 99007 | 2 | 9 | 2 | 3 | 1 | 2 | 64 | 4 | 84 | 35 | 0 | 63 | 5 | 26 | 243 | 380 | 0 | 2 | 67 |
| Arg | 7 | 99794 | 4 | 0 | 7 | 51 | 0 | 13 | 38 | 0 | 23 | 29 | 0 | 3 | 14 | 5 | 2 | 6 | 0 | 3 |
| Asn | 17 | 2 | 98905 | 134 | 2 | 31 | 17 | 25 | 114 | 22 | 30 | 102 | 17 | 4 | 39 | 327 | 161 | 3 | 44 | 4 |
| Asp | 7 | 0 | 279 | 99410 | 0 | 14 | 135 | 32 | 34 | 6 | 1 | 3 | 0 | 4 | 4 | 42 | 24 | 2 | 3 | 0 |
| Cys | 35 | 20 | 14 | 0 | 99290 | 9 | 0 | 18 | 41 | 42 | 60 | 0 | 0 | 40 | 8 | 200 | 140 | 10 | 75 | 0 |
| Gln | 2 | 38 | 49 | 10 | 2 | 99261 | 73 | 4 | 148 | 10 | 57 | 98 | 25 | 16 | 68 | 43 | 78 | 0 | 11 | 5 |
| Glu | 7 | 0 | 28 | 105 | 0 | 75 | 99634 | 11 | 12 | 0 | 0 | 59 | 0 | 0 | 4 | 39 | 12 | 0 | 6 | 9 |
| Gly | 82 | 4 | 18 | 11 | 2 | 2 | 5 | 99774 | 0 | 6 | 2 | 3 | 1 | 0 | 0 | 81 | 7 | 2 | 0 | 1 |
| His | 10 | 25 | 159 | 23 | 9 | 132 | 10 | 0 | 99260 | 13 | 17 | 15 | 0 | 21 | 21 | 40 | 44 | 2 | 198 | 0 |
| Ile | 70 | 0 | 10 | 1 | 3 | 3 | 0 | 4 | 4 | 98398 | 465 | 2 | 216 | 40 | 6 | 23 | 285 | 0 | 9 | 461 |
| Leu | 15 | 3 | 7 | 0 | 2 | 9 | 0 | 1 | 3 | 241 | 99142 | 2 | 244 | 118 | 20 | 49 | 90 | 8 | 12 | 33 |
| Lys | 0 | 24 | 173 | 2 | 0 | 105 | 61 | 8 | 18 | 8 | 12 | 99342 | 34 | 5 | 22 | 60 | 102 | 8 | 15 | 1 |
| Met | 86 | 0 | 13 | 0 | 0 | 12 | 0 | 1 | 0 | 354 | 772 | 15 | 98047 | 42 | 8 | 73 | 408 | 6 | 9 | 155 |
| Phe | 6 | 1 | 3 | 1 | 4 | 7 | 0 | 0 | 10 | 57 | 328 | 2 | 37 | 99343 | 8 | 42 | 23 | 2 | 124 | 2 |
| Pro | 35 | 5 | 28 | 2 | 1 | 31 | 2 | 0 | 11 | 10 | 62 | 10 | 7 | 9 | 99583 | 100 | 96 | 1 | 4 | 3 |
| Ser | 246 | 1 | 179 | 11 | 18 | 15 | 13 | 64 | 16 | 28 | 116 | 20 | 54 | 36 | 76 | 98631 | 452 | 8 | 16 | 0 |
| Thr | 315 | 1 | 72 | 5 | 10 | 22 | 3 | 5 | 14 | 283 | 172 | 27 | 247 | 16 | 60 | 369 | 98287 | 4 | 8 | 80 |
| Trp | 0 | 4 | 4 | 2 | 2 | 0 | 0 | 4 | 2 | 0 | 48 | 6 | 12 | 5 | 3 | 21 | 12 | 99866 | 7 | 2 |
| Tyr | 4 | 0 | 53 | 2 | 14 | 8 | 4 | 0 | 169 | 23 | 62 | 10 | 14 | 227 | 7 | 34 | 22 | 6 | 99336 | 2 |
| Val | 112 | 1 | 4 | 0 | 0 | 3 | 5 | 2 | 0 | 918 | 128 | 0 | 188 | 3 | 4 | 0 | 160 | 2 | 2 | 98467 |
| $\pi$ | .072 | .019 | .039 | .019 | .006 | .025 | .024 | .056 | .028 | .087 | .168 | .023 | .053 | .060 | .055 | .072 | .088 | .029 | .033 | .044 |

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid $i$ being replaced by the amino acid $j$ during a time interval of one substitution per 100 amino acids (1PAM) for the mtREV model, and average amino acid frequencies $\pi$ of the mtDNA-encoded proteins.

Table 2.14 is the difference of the transition probability matrix of the mtREV model from that of the JTT-F model. One of the most remarkable chracteristics of the transition probability matrix for the mtREV model is that the transitions between Arg and Lys are very rare compared to those observed in nuclear-encoded proteins. Transition probability of Arg ↔ Lys for 1PAM in the mtREV model is lower by 0.20 fold than that in the JTT-F model. This might be due to the difference between universal and mitochondrial codes. In the universal code, Lys can be substituted by Arg with a one-step change, while in the vertebrate mitochondrial code it requires a two-step change. Therefore, although Arg and Lys are chemically similar (both are basic amino acids) and hence are frequently substituted with each other in nuclear-encoded proteins, Arg ↔ Lys substitutions are much less frequent in vertebrate mitochondria. This probably explains why Arg is the second most conservative amino acid in the mtREV model, while it is only the 9th most conservative in the JTT-F model. These observations demonstrate the importance of the mutation-driven neutral evolution (Kimura 1968[160], 1983[163]) under the constraint of the code.

The substitutions between chemically similar amino acids with a one-step nucleotide change, such as Val ↔ Ile, Ala ↔ Thr, Met ↔ Leu, Ile ↔ Leu, Met ↔ Ile, Ser ↔ Thr and Phe ↔ Leu, are very frequent both in the mtREV and the JTT-F models. In agreement with the neutral theory (Kimura 1968[160], 1983[163]), this suggests that most of the amino acid substitutions in evolution are conservative rather than progressive (McLachlan 1971[208]; Grantham 1974[98]; Hasegawa and Yano 1975[124]). Met ↔ Thr substitutions are more frequent in the mtREV model than in the JTT-F model by 2.4 fold. Again, this might be due to peculiarities of the mitochondrial code, in which there are two codons for Met, while only one in the universal code.

The transition probability of Pro (codons: CCX) ↔ Ala (GCX), in which transversion in a codon is needed, is lower in the mtREV model than in the JTT-F model by 0.26 fold. Increased nucleotide transition rate of mtDNA relative to transversion rate (Brown et al. 1982[43]) might be responsible to this difference. Lower rates of Val (GUX) ↔ Leu (CUX, UUR) and Tyr (UAY) ↔ Phe (UUY) and higher rates of Val (GUX) ↔ Ile (AUY) and Thr (ACX) ↔ Ile (AUY) (in spite of the decreased number of codons for Ile in mitochondria) in the mtREV model than in the JTT-F model might also be due to the difference of transition/transversion mutation ratio between mtDNA and nuclear DNA. However, not all the differences between the mtREV and JTT-F model conform to this expectation. For example, transition probabilities of Pro (CCX) ↔ Leu (CUX, UUR), Pro (CCX) ↔ Ser (UCX, AGY), Val (GUX) ↔ Ala (GCX), and Phe (UUY) ↔ Leu (CUX, UUR), which are achieved by a transition, are lower in the mtREV model than in the JTT-F model by 0.37, 0.54, 0.55, and 0.81 fold, respectively, and the probability of Lys (AAR) ↔ Asn (AAY), which requires a transversion, is higher by 1.84. These differences are not interpretable.

Cys is the 4th most conservative amino acid in the JTT-F model, while it is only the 10th in the mtREV model. This might be due that, since most of the mtDNA-encoded proteins are membranous, cysteines in the mtDNA-encoded proteins are not involved in disulfide bonds so often as in the nuclear-

Table 2.13: Dependence of the estimated transition probability matrix on assumed trees

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 6 | 0 | 0 | 0 | 1 | 1 | 0 | -1 | -1 | 0 | -1 | 0 | 0 | -1 | 0 | 6 | -9 | 0 | 0 | 0 |
| Arg | 0 | 4 | 0 | 0 | 0 | -2 | 0 | 0 | 0 | 0 | 0 | -3 | 0 | 0 | 0 | -1 | 0 | 1 | 0 | 0 |
| Asn | -1 | 0 | 0 | -1 | 1 | 1 | 0 | -1 | 5 | -1 | 1 | 0 | 1 | -2 | -1 | -5 | 2 | 0 | 1 | 1 |
| Asp | 1 | 0 | -2 | 2 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | -3 | 0 | 0 | 1 | 2 | -2 | 1 | 0 | 0 |
| Cys | 5 | 0 | 2 | 0 | -20 | 0 | 0 | 0 | 0 | 11 | 6 | 0 | 0 | 6 | 0 | -1 | -10 | 0 | 0 | 0 |
| Gln | 4 | -1 | 1 | 1 | 0 | -2 | -1 | 0 | 3 | -4 | 5 | -1 | 3 | 0 | -1 | -3 | 1 | 0 | 0 | -2 |
| Glu | 0 | 0 | 0 | 0 | 0 | -1 | 0 | -1 | -1 | 0 | 0 | 2 | 0 | 0 | 0 | 1 | -4 | 0 | -1 | 2 |
| Gly | -1 | 0 | -2 | 0 | 0 | 0 | -1 | 0 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 |
| His | -3 | 1 | 7 | 0 | 0 | 2 | 0 | 0 | 4 | 1 | -2 | -3 | 0 | 0 | 1 | -1 | 1 | 0 | -5 | 0 |
| Ile | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 1 | 0 | -1 | 4 | 0 | 2 | -2 | 0 | 1 | -3 | 0 | 1 | -3 |
| Leu | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | -6 | 0 | 2 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| Lys | 0 | -3 | 0 | -2 | 0 | -1 | 3 | 0 | -4 | -1 | 1 | 1 | -2 | -2 | 3 | 0 | 3 | 0 | 5 | 0 |
| Met | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 3 | -1 | -8 | 4 | -1 | 0 | 1 | 0 | -1 | -1 |
| Phe | -1 | 0 | -1 | 0 | 1 | 0 | 0 | 0 | 0 | -2 | 1 | -1 | 4 | -1 | 1 | 1 | -1 | 1 | 1 | 0 |
| Pro | -1 | 0 | -1 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 2 | 1 | 0 | 1 | 3 | -2 | 0 | 0 | 0 | -1 |
| Ser | 6 | 0 | -2 | 1 | 0 | -1 | 1 | 1 | -1 | 1 | -1 | 0 | 0 | 0 | -1 | 0 | -2 | 0 | -1 | 0 |
| Thr | -8 | 0 | 1 | 0 | -1 | 0 | -1 | -1 | 0 | -3 | 2 | 1 | 1 | -1 | 0 | -2 | 9 | 0 | 1 | 2 |
| Trp | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | -2 | 0 | 0 | -2 | 0 | 2 | -1 | 1 |
| Tyr | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -4 | 2 | 5 | 4 | 0 | 1 | -1 | -2 | 1 | -1 | -4 | 0 |
| Val | -1 | 0 | 0 | 0 | 0 | -1 | 1 | -1 | 0 | -6 | 1 | 1 | -1 | 0 | -2 | 0 | 4 | 0 | -1 | 5 |

Difference between the transition probability matrix of the mtREV model estimated from Fig. 2.2 (Table 2.12) for 1PAM and that of ML tree in Fig. 2.3 ($\times 10^5$).

Table 2.14: Difference between the mtREV and JTT-F matrices.

| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 181 | -7 | -12 | -13 | -1 | -12 | -23 | -30 | -2 | 54 | -14 | -8 | 40 | -3 | -75 | -18 | 2 | -2 | -2 | -55 |
| Arg | -30 | 555 | -13 | -3 | 1 | -21 | -7 | -61 | -48 | -20 | -37 | -117 | -21 | -1 | -23 | -63 | -52 | -30 | -7 | -4 |
| Asn | -21 | -6 | -66 | 37 | 0 | 13 | 4 | -17 | 7 | -19 | 8 | 46 | 1 | 0 | 33 | -10 | -31 | 2 | 22 | -3 |
| Asp | -49 | -3 | 76 | 102 | -1 | 2 | -41 | -36 | 7 | -4 | -9 | -3 | -9 | 2 | -2 | 3 | -11 | 0 | -11 | -13 |
| Cys | -5 | 1 | 2 | -2 | -164 | 7 | -1 | -12 | 21 | 29 | 33 | -1 | -20 | 0 | 2 | 53 | 101 | -22 | 7 | -26 |
| Gln | -36 | -16 | 20 | 1 | 1 | 31 | -6 | -9 | -4 | 3 | -56 | 32 | 2 | 13 | -15 | 6 | 35 | -5 | 3 | -3 |
| Glu | -66 | -6 | 7 | -32 | 0 | -6 | 177 | -49 | 5 | -9 | -16 | 20 | -9 | -3 | -6 | 18 | -15 | -3 | 4 | -10 |
| Gly | -39 | -21 | -11 | -12 | -1 | -4 | -21 | 245 | -5 | 1 | -9 | -3 | -5 | -3 | -11 | -46 | -19 | -14 | -2 | -19 |
| His | -5 | -32 | 10 | 5 | 5 | -3 | 4 | -11 | 154 | -2 | -69 | 3 | -16 | -5 | -38 | -10 | 5 | -2 | 13 | -5 |
| Ile | 45 | -4 | -8 | -1 | 2 | 1 | -3 | 1 | -1 | -208 | 88 | -3 | -25 | -6 | 1 | -5 | 72 | -4 | -1 | 59 |
| Leu | -6 | -4 | 2 | -1 | 1 | -8 | -2 | -3 | -11 | 46 | -37 | -1 | 54 | -28 | -35 | 9 | 67 | -7 | 4 | -41 |
| Lys | -26 | -94 | 79 | -3 | 0 | 34 | 21 | -7 | 4 | -9 | -12 | -67 | 3 | 4 | 11 | 28 | 21 | 5 | 12 | -4 |
| Met | 54 | -8 | 1 | -3 | -2 | 1 | -4 | -6 | -9 | -40 | 171 | 1 | -500 | 17 | 0 | 54 | 237 | 0 | 3 | 31 |
| Phe | -4 | 0 | 0 | 0 | 0 | 6 | -1 | -3 | -2 | -9 | -77 | 1 | 15 | 172 | 0 | -22 | 11 | -14 | -52 | -23 |
| Pro | -99 | -8 | 24 | 0 | 0 | -7 | -2 | -11 | -19 | 2 | -106 | 5 | -1 | 0 | 315 | -86 | 0 | -1 | 0 | -6 |
| Ser | -19 | -17 | -6 | 1 | 5 | 2 | 6 | -37 | -4 | -5 | 22 | 10 | 40 | -18 | -66 | 46 | 61 | -1 | -4 | -17 |
| Thr | 2 | -11 | -14 | -2 | 7 | 10 | -5 | -12 | 1 | 71 | 129 | 6 | 144 | 8 | 0 | 50 | -419 | 2 | 2 | 33 |
| Trp | -6 | -19 | 3 | 1 | -5 | -4 | -3 | -26 | -2 | -12 | -39 | 4 | 2 | -27 | -1 | 0 | 5 | 154 | -17 | -8 |
| Tyr | -6 | -4 | 27 | -6 | 1 | 2 | 3 | -3 | 11 | -3 | 23 | 8 | 5 | -96 | 1 | -9 | 5 | -16 | 59 | -5 |
| Val | -90 | -2 | -2 | -6 | -4 | -1 | -6 | -24 | -3 | 117 | -155 | -3 | 37 | -32 | -7 | -28 | 65 | -5 | -3 | 151 |

Difference of the transition probability matrix of the mtREV (Table 2.12) model for 1PAM from that of the JTT-F (Table 2.9) model ($\times 10^5$).

encoded proteins in which globular proteins occupy a larger portion. All the differences of the transition probability matrix between the mtREV and the JTT-F models are not necessarily interpretable in the straightforward ways. Some of the differences might be due to the biased estimate of the JTT-F matrix by the parsimony method (Collins et al. 1994[57]; Perna and Kocher 1995[235]; Goldman 1990[92]; Yang 1994[317]), and some of the others might be due to the small sample size of the data in estimating the mtREV matrix.

Table 2.15: Comparison of amino acid frequencies between mitochondrial and nuclear-encoded proteins.

| | mt-code | mt-proteins (H-strand encoded) | nuclear | mt/nuc | ND6 (L-strand) |
|---|---|---|---|---|---|
| Trp | UGR | .029 | .014 | 2.07 | .023 |
| Tyr | UAY | .033 | .032 | 1.03 | .061 |
| Phe | UUY | .060 | .040 | 1.50 | .062 |
| Leu | UUR,CUX | .168 | .091 | 1.85 | .141 |
| Ile | AUY | .087 | .053 | 1.64 | .081 |
| Met | AUR | .053 | .024 | 2.21 | .039 |
| Val | GUX | .044 | .066 | 0.67 | .146 |
| Ala | GCX | .072 | .077 | 0.94 | .066 |
| Pro | CCX | .055 | .051 | 1.08 | .033 |
| Gly | GGX | .056 | .074 | 0.76 | .153 |
| Thr | ACX | .088 | .059 | 1.49 | .039 |
| Ser | UCX,AGY | .072 | .069 | 1.04 | .079 |
| Asn | AAY | .039 | .043 | 0.91 | .006 |
| Asp | GAY | .019 | .052 | 0.37 | .003 |
| Gln | CAR | .025 | .041 | 0.61 | .004 |
| Glu | GAR | .024 | .062 | 0.39 | .022 |
| His | CAY | .028 | .023 | 1.22 | .000 |
| Lys | AAR | .023 | .059 | 0.39 | .009 |
| Arg | CGX | .019 | .051 | 0.37 | .011 |
| Cys | UGY | .006 | .020 | 0.30 | .021 |

Average amino acid frequencies of the mtDNA-encoded proteins (mtREV model) and of the nuclear-encoded proteins (JTT model)

Table 2.15 gives amino acid frequencies of the mtDNA-encoded proteins used in the estimation of the mtREV matrix (12 proteins) and of the proteins used in the estimation of the JTT matrix which consist mainly of nuclear-encoded ones. Cys is scarce in the mtDNA-encoded proteins probably because this amino acid is not involved in disulfide bonds so often as in the nuclear-encoded proteins as mentioned before. The mtDNA-encoded proteins are mostly membranous and probably for this reason hydrophobic amino acids, such as Met, Trp, Leu, Ile and Phe, are more abundant, and hydrophilic amino acids, such as Arg, Lys, Glu, Asp and Gln are more scarce than in the nuclear-encoded proteins. Of course, the abundant Met and Trp in the mtDNA-encoded proteins might also be due to their having two codons in mitochondria, while only one in the universal code. However, in disagreement with the above expectation, the frequencies of hydrophobic amino acids, such as Val (codon: GUX) and Gly (GGX), are less in the mtDNA-encoded proteins than in the nuclear-encoded proteins. This might be due to that the codons of these amino acids contain G which is scarce in the L-strand of mtDNA (the 12 proteins used in this

analysis are encoded by the H-strand, and the mRNAs are complementary to the H-strand). In agreement with this consideration, Val and Gly are more abundant by about 3 fold in ND6 which is encoded by the L-strand (G is abundant in its mRNA) than in the 12 mtDNA-encoded proteins. This suggests that, amino acid frequencies of the mtDNA-encoded proteins are governed not only by the structural-functional requirements of the individual proteins but also by the bias and skewness of mtDNA caused by its asymmetric replication pattern (Thomas and Wilson 1991[294]; Kondo 1992[171]; Tanaka and Ozawa 1994[289]).

### 2.2.6 Discussion

Previously, the JTT model for nuclear-encoded proteins was used even in the ML analyses of mtDNA-encoded proteins (Cao et al. 1994[49]; Adachi and Hasegawa 1995[5]), mainly because no appropriate model for mtDNA-encoded proteins was available. The conclusions of these phylogenetic analyses hold when the mtREV model presented in this paper is used. This suggests that the ML method is robust to some extent against the violation of the assumed model (Fukami-Kobayashi and Tateno 1991[86]; Hasegawa and Fujiwara 1993[110]). Nevertheless, phylogenetic conclusions derived from a realistic model should be more reliable than that from a less realistic one, and therefore we must continue to improve the model. Once a probability model as shown in Table 1, which is realistic to some extent, is obtained, the ML method would be the preferred method in inferring trees from mtDNA-encoded protein sequences (Felsenstein 1981[76]; Kishino et al. 1990[166]; Edwards 1995[70]). Although the amino acid frequencies of the individual protein under analysis might be different from the average frequencies of the 12 proteins used in estimating the transition probability matrix, the ProtML program of our package MOLPHY (Adachi and Hasegawa 1995[5]) can adjust the equilibrium frequencies of the model to the actual frequencies of the protein under study (F-option).

If we are to analyze closely related sequences, synonymous substitutions provide us with important information, and therefore a codon-based model of nucleotide substitution (Schöniger et al. 1990[257]; Goldman and Yang 1994[94]; Muse and Gaut[221]) might be preferable to the amino acid substitution model. However, in constructing the model of nucleotide substitution, it must be noted that the nucleotide frequencies of the 3rd codon positions are significantly different even between closely related species in Hominoidea (T is significantly more scarce and C is more abundant in orangutan than in gorilla), and that the reversible Markov model no longer holds for these sites. One of the advantages of the ML method over the other existing methods in molecular phylogenetics is that, as is demonstrated in this work, we can incorporate complexity in the pattern of substitution and can improve the model as the relevant data accumulate, because the method is based on an explicit model (Thorne et al. 1992[296]). The parsimony method is used widely (Stewart 1993[278]), but it is not based on the explicit model, and therefore it suffers limitations in taking account directly of the complex pattern of the actual process of evolution (Sidow 1994[266]).

# Chapter 3

# Maximum Likelihood Inference of Molecular Phylogeny

Molecular phylogenetics studies evolutionary relationships among organisms by using molecular data. It is one of the areas of molecular evolution that have generated much interest in the last decade, mainly because in many cases phylogenetic relationships are difficult to assess in other ways. The purpose of this chapter is to explain how to infer a phylogenetic tree from molecular data by the maximum likelihood method. Neyman (1971[225]) was the first to use the maximum likelihood method to estimate evolutionary trees from DNA sequences based on a probabilistic model, and Felsenstein (1981[76]) developed a practical method, from which the maximum likelihood methods used widely at present stem (Kishino et al. 1990[166]; Adachi and Hasegawa 1992[3], 1995[5]; Yang 1993[316]; Felsenstein 1993[82]; Olsen et al. 1994[231]).

## 3.1 Evolutionary Tree Reconstruction

### 3.1.1 Phylogenetic Trees

All life forms on the earth share a common origin, and their ancestries can be traced back to one organism that lived approximately 4 billion years ago. Consequently, all animals, fungi, plants, protista, and bacteria are related by descent to each other. Closely related organisms descended from a more recent common ancestor than are distantly related ones. The objectives of phylogenetic studies are (1) to reconstruct the correct genealogical ties between organisms and (2) to estimate the time of divergence between organisms since they last shared a common ancestor.

In phylogenetic studies, the evolutionary relationships among a group of organisms are illustrated by means of a phylogenetic tree. A phylogenetic tree is a graph composed of nodes and branches, in which only one branch connects any two adjacent nodes. The nodes represent the taxonomic units, and the branches define the relationships among the units in terms of descent and ancestry. The branching pattern of a tree is called the topology. The branch length usually represents the number of changes per site that have occurred in that branch. The taxonomic units represented by the nodes can be species, populations, individuals, or genes.

When dealing with phylogenetic trees, we distinguish between external nodes and internal nodes. Terminal nodes are external, whereas all others are internal. External nodes represent the extant taxonomic units under comparison (if we are to deal with ancient DNA from extinct organisms, external nodes may not represent extant taxonomic units, but in any case data are given to external nodes), and are referred to as operational taxonomic units (OTUs). Internal nodes represent ancestral units, and we can only infer the states of the internal nodes.

A node is bifurcating if it have only two immediate descendant lineages, but multifurcating if it have more than two immediate descendant lineages.

### 3.1.2 Rooted and Unrooted Trees

Phylogenetic trees can be either rooted or unrooted. In a rooted tree there exists a particular node, called the root, from which a unique path leads to any other nodes. The direction of each path corresponds to the evolutionary time, and the root is the common ancestor of all the OTUs under study. An unrooted tree is a tree that only specifies the relationships among the OTUs with no time direction.

### 3.1.3 Number of Possible Trees

For three OTUs, there are three different possible rooted trees but only one unrooted tree. The number of bifurcating rooted trees ($N_R$) for $n$ OTUs is given by

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{3.1}$$

for $n \geq 2$ (Felsenstein 1978[75]). The number of bifurcating unrooted trees ($N_U$) for $n \geq 3$ is

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!}. \tag{3.2}$$

Note that the number of possible unrooted trees for $n$ OTUs is equal to the number of possible rooted trees for $n-1$ OTUs. The numbers of possible unrooted trees for up to $n$ OTUs are given in Table 3.1. We see that $N_R$ and $N_U$ increase very rapidly as $n$ increases, and for 10 OTUs there are already 2 million bifurcating unrooted trees. Since only one of these trees represents correctly the true evolutionary relationships among the OTUs, it becomes a hard job to find the true phylogenetic tree when $n$ is large.

Table 3.1: Possible numbers of unrooted trees.

| Number of OTUs | Number of unrooted trees | |
|---|---|---|
| 3 | 1 | 1 |
| 4 | $1 \times 3$ | 3 |
| 5 | $1 \cdot 3 \times 5$ | 15 |
| 6 | $1 \cdot 3 \cdot 5 \times 7$ | 105 |
| 7 | $1 \cdot 3 \cdot 5 \cdot 7 \times 9$ | 945 |
| 8 | $1 \cdot 3 \cdot 5 \cdot 7 \cdot 9 \times 11$ | 10395 |
| 9 | $1 \cdot 3 \cdot 5 \cdot 7 \cdot 9 \cdot 11 \times 13$ | 135135 |
| 10 | $1 \cdot 3 \cdot 5 \cdot 7 \cdot 9 \cdot 11 \cdot 13 \times 15$ | 2027025 |
| 11 | $\cdots \times 17$ | 34459425 |
| 12 | $\cdots \cdot 17 \times 19$ | 654729075 |
| 13 | $\cdots \cdot 17 \cdot 19 \times 21$ | 14 billion |
| 14 | $\cdots \cdot 17 \cdot 19 \cdot 21 \times 23$ | 316 billion |
| 15 | $\cdots \cdot 17 \cdot 19 \cdot 21 \cdot 23 \times 25$ | 7906 billion |
| $\vdots$ | $\cdots$ | $\vdots$ |
| $n$ | $\cdots\cdots \times (2n-5)$ | $\frac{(2n-5)!}{2^{n-3}(n-3)!}$ |

### 3.1.4 True vs. Inferred Trees

The sequence of speciation events that has led to the formation of any group of OTUs must be historically unique. Thus, only one of all the possible trees that can be built with a given number of OTUs should represent the true evolutionary history. Such a phylogenetic tree is called a true tree. A tree that is obtained by using a certain set of data with a certain method of tree reconstruction is called an inferred tree. An inferred tree may or may not be identical with the true tree.

## 3.2 Traditional Methods

Taxonomy is concerned with grouping organisms into a manageable number of groups whose characters are mainly shared. This enables the classification of species. For evolutionary studies, the classification also allows the construction of phylogenies. These activities may shed light on the question of whether the observed patterns of species suggest anything about the nature of evolutionary forces.

Three of the principal methods for ordering species in a phylogenetic history will be mentioned here. Distance matrix methods are based on a set of distances calculated for each pairs of species. The computations are generally quite straightforward, and the quality of the resulting tree depends on the quality of the distance measure. Whereas distances are generally based on some statistical models, the maximum parsimony method does not have an explicit model. It proceeds by seeking to minimize the number of changes among species over the tree. Dependence on statistical models is a feature of the maximum likelihood method. Although this third class of tree reconstruction is computationally demanding, it can provide a basis for statistical inference.

### 3.2.1 Maximum Parsimony Method

The parsimony method takes explicit notice of character values observed for each species, rather than working with the distances between sequences that summarize differences between character values. The approach was introduced for gene frequency data by Edward and Cavalli-Sforza (1963[71]) under the name "Principle of Minimum Evolution". If sequences are available for a set of species, then the most parsimonious topology linking them is sought.

In this method, the sequences of ancestral species are inferred from those of the extant species, considering a particular tree topology, and the minimum number of evolutionary changes that are required to explain all observed differences among the sequences is computed. The number is obtained for all possible topologies, and the topology which shows the smallest number of evolutionary changes is chosen as the final tree. This method is used mainly for finding the topology of a tree, but branch lengths can be estimated under certain assumptions (Fitch 1971[83]).

Parsimony methods have been criticized by Felsenstein (1983[77], 1983[78], 1984[79]) on the grounds that they are not based on statistical principles. Felsenstein points out that parsimony methods, in trying to minimize the number of evolutionary events, implicitly assume that such events are improbable. If the amount of change over the evolutionary time being considered is small, the parsimony methods will be well-justified, but for cases in which there are a large amount of change and a large amount of rate heterogeneity among lineages, the parsimony method can be positively misleading (Felsenstein 1978[74]).

### 3.2.2 Distance Method

Phylogenetic trees may be based on distance matrices, that is, genetic distances between all pairs of OTUs.

Using these distances to group OTUs in a phylogenetic context may employ clustering, and possible approaches to clustering were given by Sneath and Sokal (1973[271]). Clusters are characterized by having more OTUs per unit of, say, gene frequency space than do other areas, and the process of clustering consists of identifying these areas of higher density. Several methods of clustering can be used. When distances are used to construct additive trees, as with the Fitch-Margoliash (1967[84]) algorithm described below, the process is said to use a pairwise method rather than clustering.

### UPGMA

There is a class of strategies used for finding clusters, called sequential, agglomerative, hierarchical, and nonoverlapping by Sneath and Sokal (1973[271]). The most widely used among them is the UPGMA (unweighted pair-group method using an arithmetic average). It defines the intercluster distance as the average of all the pairwise distances for members of two clusters.

The UPGMA method has been widely used for matrices of distances. It must be emphasized that assumption of equal reates is crucial for the UPGMA method to be appropriate. This method is not satisfactory when the rates are unequal in different lineage. When rates are equal, a molecular clock (Zuckerkandl and Pauling 1965[321]; Sarich and Wilson 1967[254]; Wilson et al. 1977[308]; Kimura 1983[163]) is said to be operating.

### Neighbor Joining Method

When the molecular clock does not hold, the UPGMA method is not applicable. But there are several distance methods that perform better than the UPGMA in such a situation (Fitch and Margoliash 1967[84]; Sattath and Tversky 1977[256]; Saitou and Nei 1987[253]; and for review, Nei 1987[224]; Felsenstein 1988[81]; Swofford and Olsen 1990[282]; Li and Graur 1991[192]). Among these method, the neighbor joining (NJ) method of Saitou and Nei (1987[253]) is most widely used in the recent years.

The principle of the NJ method is to find pairs of OTUs (neighboring OTUs) that minimize the total branch length at each stage of clustering of OTUs starting with a star-like tree. This method is very quick, and its performance is generally good even when the molecular clock does not hold as far as the distances are accurately corrected for multiple substitutions (Saitou and Imanishi 1989[252]; Hasegawa et al. 1991[121]; DeBry 1992[64]; Hasegawa and Fujiwara 1993[110]).

## 3.3 Algorithm for ML Inference of Molecular Phylogeny

The aligned molecular (bases or amino acids) sequence data of length $n$ (sites) from $N$ species can be represented as follow:

$$\boldsymbol{X} = (\underbrace{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_h, \cdots, \boldsymbol{X}_n}_{\text{number of sites}}) = \begin{pmatrix} \boldsymbol{X}^{(1)} \\ \boldsymbol{X}^{(2)} \\ \vdots \\ \boldsymbol{X}^{(s)} \\ \vdots \\ \boldsymbol{X}^{(N)} \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1h} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2h} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ X_{s1} & X_{s2} & \cdots & X_{sh} & \cdots & X_{sn} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \vdots \\ X_{N1} & X_{N2} & \cdots & X_{Nh} & \cdots & X_{Nn} \end{pmatrix} \begin{matrix} \\ : \text{Species 1} \\ : \text{Species 2} \\ \\ : \text{Species s} \\ \\ : \text{Species N} \end{matrix}$$

Let us write the whole data set as $\boldsymbol{X}$, the value of the $h$-th site $(X_{1h}, X_{2h}, \cdots, X_{Nh})^T$ as $\boldsymbol{X}_h$ and the value of the $s$-th species $(X_{s1}, X_{s2}, \cdots, X_{sn})$ as $\boldsymbol{X}^{(s)}$. We assume that each site evolves independently and identically with others. We further assume that, after speciation, the two separated lineages evolve indepently, and that the same stochastic process of substitution applies in all lineages, although the rate parameter of the process might differ among different lineages.

### 3.3.1 Computing Likelihood of a Tree

Given that we are willing to assume independence of evolution at different sites, it turns out that the probability of a given set of the data arising on a given tree can be computed site by site, and the product of the probabilities can be taken across sites at the final stage of the computation (Felsenstein 1981[76]).

We may write the likelihood for a given tree topology $T$ and sequence data $\boldsymbol{X}$ as

$$L = \text{Prob}(\boldsymbol{X}|T, \boldsymbol{\theta}) \tag{3.3}$$

where $\boldsymbol{\theta}$ is vector of parameters.

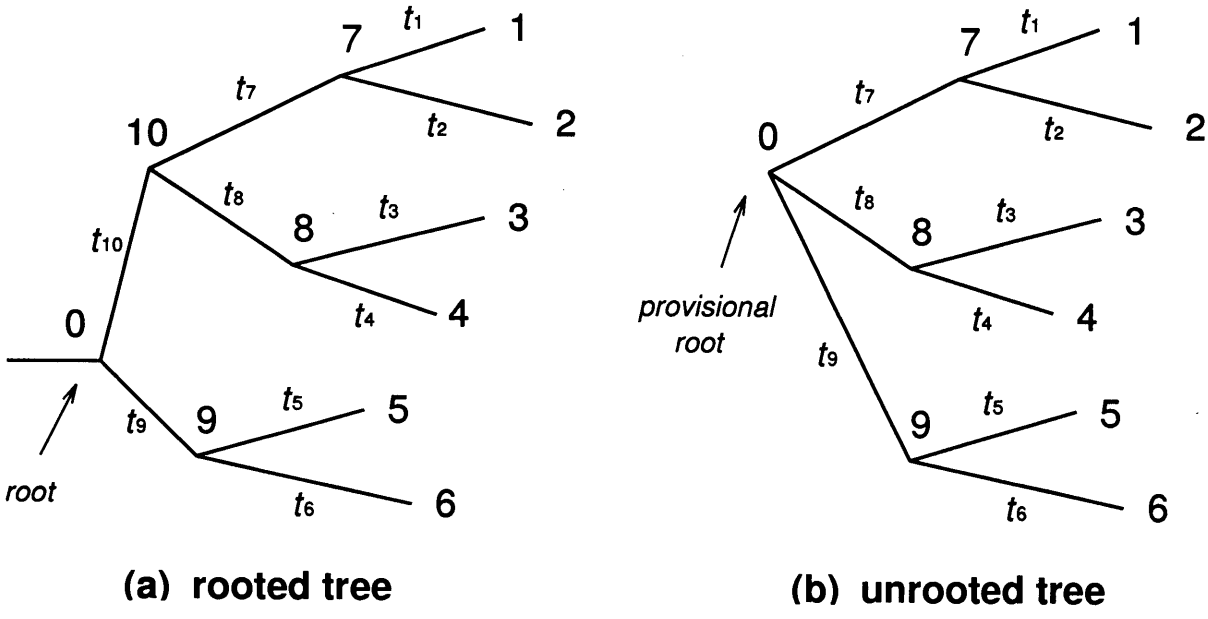



Figure 3.1: The rooted tree and unrooted tree used in the discussion.

It is possible to write a general expression for the likelihood of a tree, but it will be more useful to present the expression for a particular case, the tree topology $T = \big(((1,2),(3,4)),(5,6)\big)$ as in Fig. 3.1a, since the general pattern will become clear from that expression. The lengths of the branchs of the tree are given by the quantities $t_i$. If we know the states (bases or amino acids) at a particular site at nodes 7, 8, 9 and 10 on this tree, and let these be $x_7$, $x_8$, $x_9$ and $x_{10}$, the likelihood of the tree would be the product of the probabilities of change in each branch, times the prior probability $\pi_{x_0}$ of state $x_0$, so that it would be

$$\begin{aligned} f(\boldsymbol{x}) = \quad & \pi_{x_0} \quad P_{x_0 x_{10}}(t_{10}) P_{x_{10} x_7}(t_7) P_{x_7 x_1}(t_1) P_{x_7 x_2}(t_2) P_{x_{10} x_8}(t_8) P_{x_8 x_3}(t_3) P_{x_8 x_4}(t_4) \\ & \times \quad P_{x_0 x_9}(t_9) P_{x_9 x_5}(t_5) P_{x_9 x_6}(t_6) \end{aligned} \tag{3.4}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_6)^T$ is a vector of a sequence data of length 1 and $x_i$ at the internal node is the state at the internal node $i$ in the tree.

The $\pi$ must be the prior probabilities of finding each of the state at node 0 on the tree. Since we are assuming an evolutionarily steady state in base composition (amino acid frequency), they reflect the overall base composition (amino acid frequency) in the group under study. One of the convenient properties of the Markov process model (given in Chapter 2) of base (amino acid) substitution is known as "reversibility" (Felsenstein 1981[76]). This means that the process of base (amino acid) substitution will look the same irrespective of whether forward or backward in time. Reversibility requires that for all $i$, $j$ and $t$

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \tag{3.5}$$

which is easily proven using Eq. 2.10.

From the reversibility and the "pulley principle" (Felsenstein 1981[76]) , the tree in Fig. 3.1b cannot be distinguished from the tree in Fig. 3.1a, for the same $t_i$. The quantity $t_9$ in Fig. 3.1b is equal to $(t_9 + t_{10})$ in Fig. 3.1a. The likelihood of the tree topology $T = \big((1,2),(3,4),(5,6)\big)$ in Fig. 3.1b would be

$$\begin{aligned} f(\boldsymbol{x}) = \quad & \pi_{x_0} \quad P_{x_0 x_7}(t_7) P_{x_7 x_1}(t_1) P_{x_7 x_2}(t_2) \\ & \times \quad P_{x_0 x_8}(t_8) P_{x_8 x_3}(t_3) P_{x_8 x_4}(t_4) \\ & \times \quad P_{x_0 x_9}(t_9) P_{x_9 x_5}(t_5) P_{x_9 x_6}(t_6) \end{aligned} \tag{3.6}$$

where the node 0 is a provisional root of the tree.

In practice we do not know $x_7$, $x_8$ and $x_9$, so the likelihood should be the sum over all possible assignments of bases (amino acids) to those nodes on the tree in Fig. 3.2. A probability of occupying $\boldsymbol{x} = (x_1, x_2, \cdots, x_6)^T$ at a site in species $1, 2, \ldots, 6$ respectively, is given by

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \pi_i \left( \sum_{j=1}^{m} P_{ij}(t_7) P_{jx_1}(t_1) P_{jx_2}(t_2) \right) \times \left( \sum_{k=1}^{m} P_{ik}(t_8) P_{kx_3}(t_3) P_{kx_4}(t_4) \right) \times \left( \sum_{l=1}^{m} P_{il}(t_9) P_{lx_5}(t_5) P_{lx_6}(t_6) \right). \tag{3.7}$$

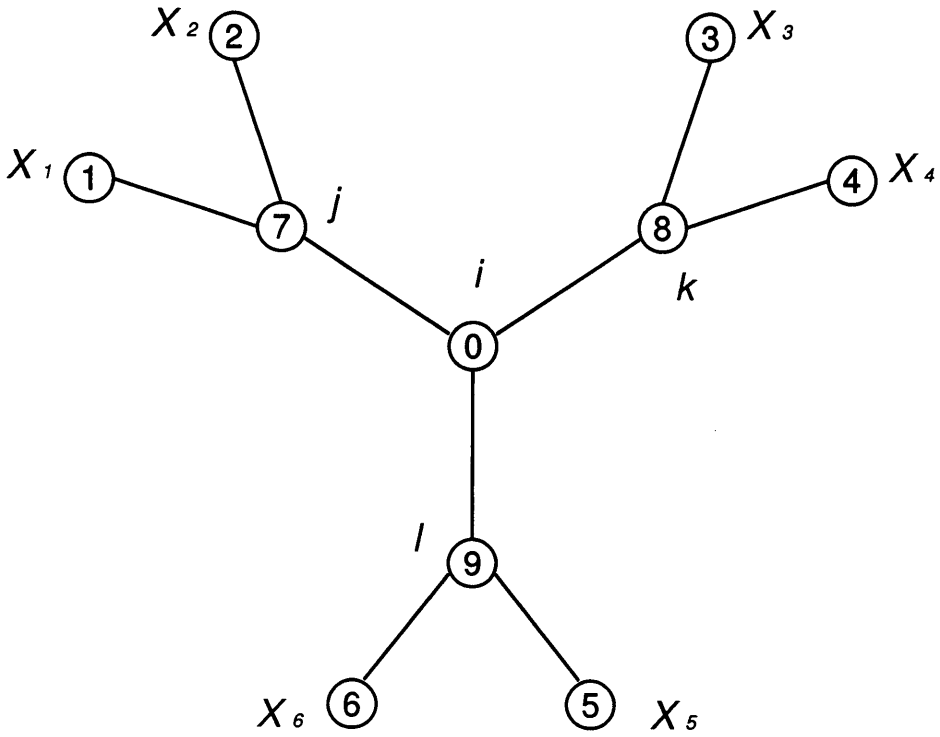



Figure 3.2: The unrooted tree used in the discussion of computing the likelihood.

The log-likelihood of a tree is

$$l(\boldsymbol{\theta}|\boldsymbol{X}, T) = \sum_{h=1}^{n} \log f(\boldsymbol{X}_h|T, \boldsymbol{\theta}) \tag{3.8}$$

where

$$\boldsymbol{\theta} = (t_1, t_2, \cdots, t_9)^T. \tag{3.9}$$

The log-likelihood of the tree is rewritten as

$$l(\boldsymbol{\theta}|\boldsymbol{X}, T) = \sum_{h=1}^{n} \log \left\{ \sum_{i=1}^{m} \pi_i \left( \sum_{j=1}^{m} P_{ij}(t_7) P_{jX_{1h}}(t_1) P_{jX_{2h}}(t_2) \right) \times \left( \sum_{k=1}^{m} P_{ik}(t_8) P_{kX_{3h}}(t_3) P_{kX_{4h}}(t_4) \right) \times \left( \sum_{l=1}^{m} P_{il}(t_9) P_{lX_{5h}}(t_5) P_{lX_{6h}}(t_6) \right) \right\}. \tag{3.10}$$

### 3.3.2 Evaluating Likelihood along the Tree

Given that we can evaluate the likelihood of any given tree topology $T$ for any given parameter value $\boldsymbol{\theta}$, we still have to solve the problem of maximizing the likelihood over all $T$ and all $\boldsymbol{\theta}$.

For a given tree topology, estimation of each branch length is iterated separately, by using the Newton-Raphson method (Kishino et al. 1990[166]) and by repeatedly evaluating the likelihood. This does not require re-evaluation of likelihood throughout the tree each time, because the "pruning" algorithm can be used. This algorithm has been described by Felsenstein (1973[73], 1981[76]).

#### Data Structure of a Tree

We can restate this process in terms of partial likelihood: Let us define $q_{hi}$ as the likelihood based on the descendant data at outer current subnode on the tree, given that current subnode is known to have state $i$ for a site $h$ under consideration. A partial Likelihood is a set of conditional Likelihood of subtree. The partial Likelihood $\boldsymbol{q}$ of length $n$ (sites) for $m$ states can be represented as follow:

$$\boldsymbol{q} = \begin{pmatrix} \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \vdots \\ \boldsymbol{q}_h \\ \vdots \\ \boldsymbol{q}_n \end{pmatrix} = \begin{pmatrix} q_{11} & q_{12} & \cdots & q_{1m} \\ q_{21} & q_{22} & \cdots & q_{2m} \\ \vdots & \vdots & \cdots & \vdots \\ q_{h1} & q_{h2} & \cdots & q_{hm} \\ \vdots & \vdots & \cdots & \vdots \\ q_{n1} & q_{n2} & \cdots & q_{nm} \end{pmatrix}.$$

Let us write the value of the $h$-th site $(q_{h1}, q_{h2}, \cdots, q_{hm})$ as $\boldsymbol{q}_h$. Partial likelihood can be defined at each subnode in an internal node.

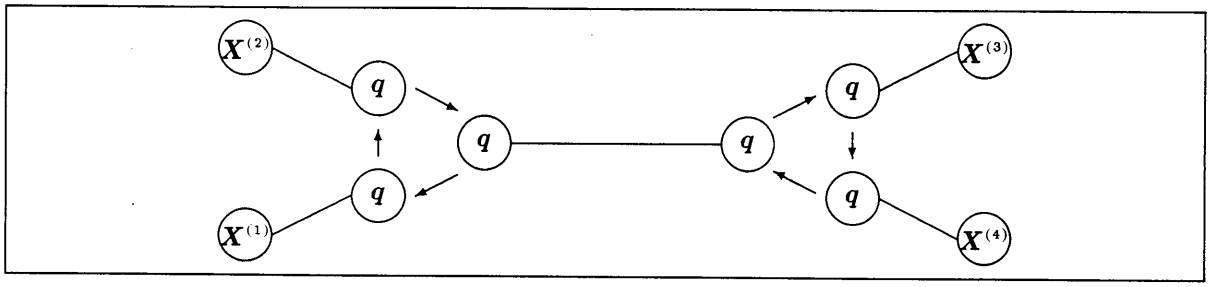



Figure 3.3: Data structure of a tree.

#### Partial Likelihood of a Subtree

Let us define partial Likelihood $q_{hi}$ as the likelihood of the subtree for all data for site $h$ at or above current subnode on the tree, given that site $h$ in current subnode is in state $i$. We can easily determine this for the inner subnode of an external branch in the tree. If, for example, the inner subnode of an external branch shows an $x$ in a site, it follows immediately by its definition that $q_i = P_{ix}(t)$. There is not the matrix $\boldsymbol{q}$ for the external node (outer node of an external branch). We can work down the tree computing $\boldsymbol{q}$ at each site for each subnode of the tree, by making use of the recursion for current subnode whose immediate descendants, subnode 1 and subnode 2, have $q_i$ values that have been previously computed,

and has branch length $t$ leading to them:

$$q_i = \begin{cases} \sum_{j=1}^{h} P_{ij}(t)Q_j, & \text{if internal branch} \\ P_{ix}(t), & \text{if external branch} \end{cases} \tag{3.11}$$

where $Q_j$ is product of under partial likelihoods.

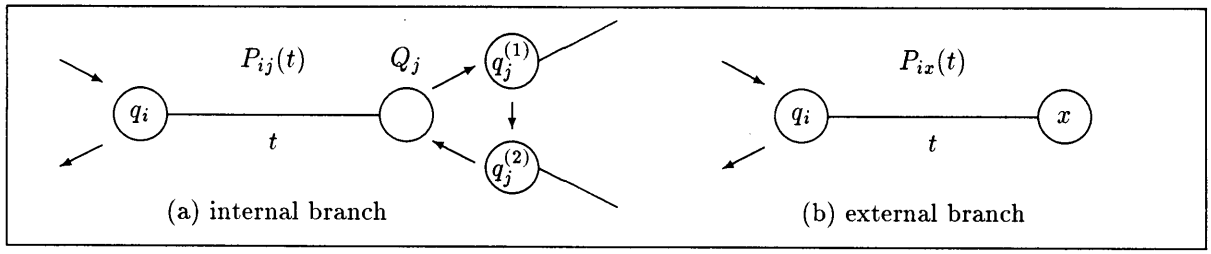



Figure 3.4: Partial likelihood.

Suppose that we define a product of partial Likelihood $Q_i$ as product of each likelihood of the subtree for all data for site $h$ at or above current node on the tree, given that site $h$ in current subnode is in state $i$. We can compute $\boldsymbol{Q}$ at each site for each subnode of internal branch in the tree, by making use of the recursion for current subnode whose immediate descendants, subnode $1, 2, \ldots, b$, have $Q_i$ values that have been previously computed, leading to them:

$$Q_i = \prod_{j=1}^{b} q_i^{(j)} \tag{3.12}$$

where $b$ is a number of branchings.

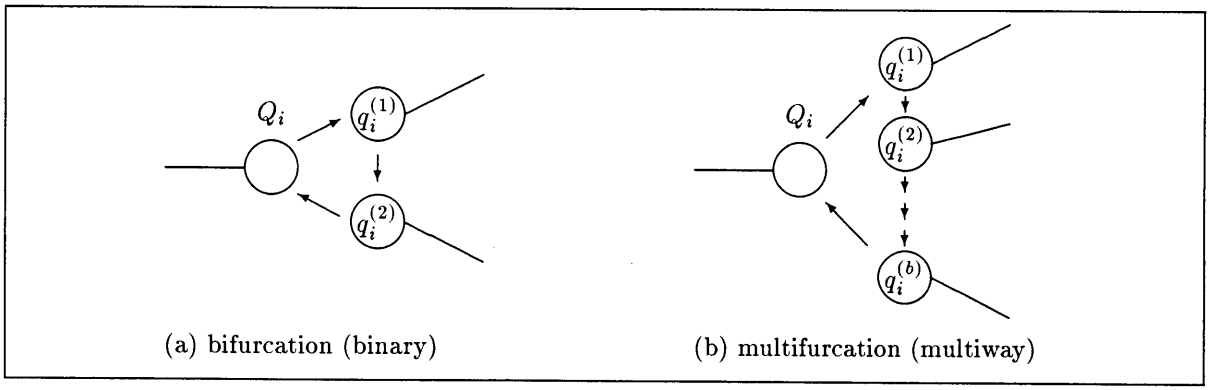



Figure 3.5: Product of partial likelihood.

This process proceeds down the tree towards the root. In an unrooted tree the root may be taken anywhere. The values of $\boldsymbol{q}$ at the root are then combined in a weighted average

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \pi_i Q_i^{(\text{ans})} \sum_{j=1}^{m} P_{ij}(t) Q_j^{(\text{des})} = \sum_{i=1}^{m} \pi_i \prod_{j=0}^{b} q_i^{(j)} \tag{3.13}$$

which computes the likelihood at that site for the whole tree, unconditioned on knowing the state at that node.

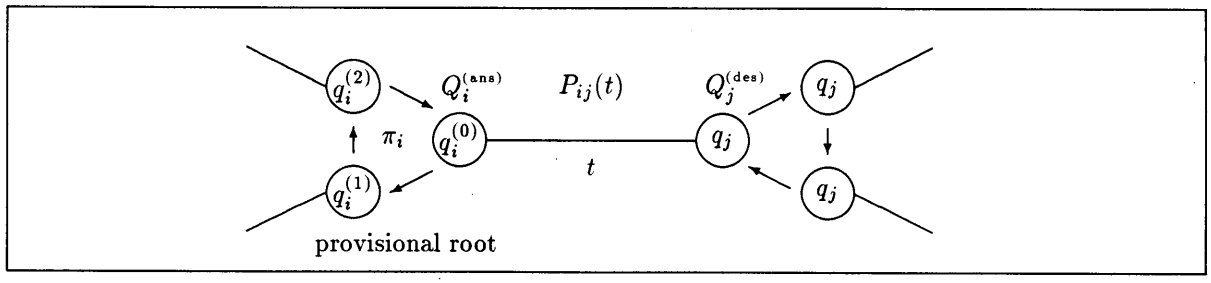



Figure 3.6: Computing the likelihood of a tree.

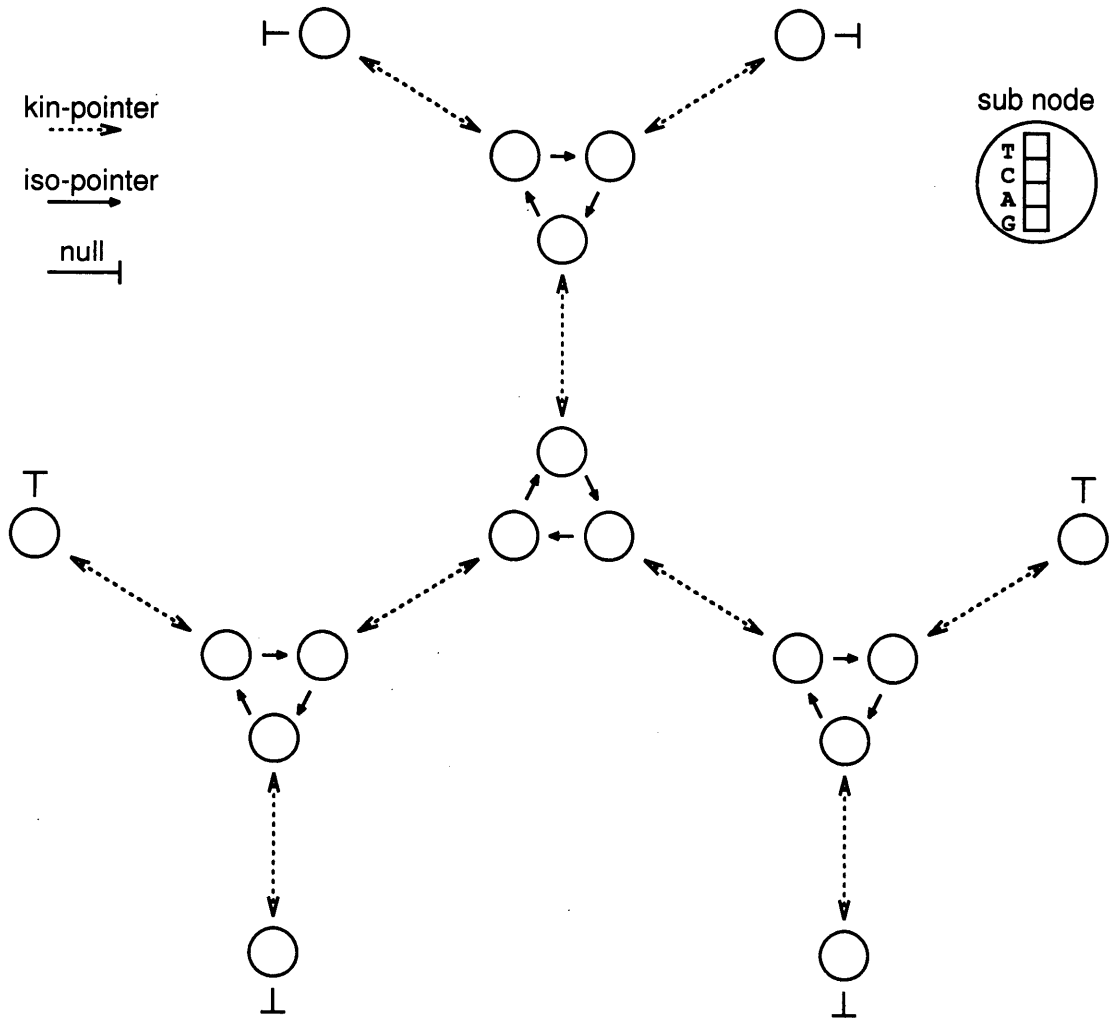



Figure 3.7: Data structure of a tree topology.

### 3.3.3 Maximum Likelihood Estimation of Branch Length

The Maximum Likelihood Estimate (MLE) $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ is the solution of

$$\text{maximize} \quad \log L(\boldsymbol{\theta}|\boldsymbol{X},T) \qquad \text{for } \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{3.14}$$

$\hat{\boldsymbol{\theta}}$ of course satisfies the standard conditions

$$\left[\frac{\partial \log L}{\partial \theta_j}\right]^T_{\hat{\boldsymbol{\theta}}} = 0, \tag{3.15}$$

$$\left[\frac{\partial^2 \log L}{\partial \theta_j, \partial \theta_h}\right]_{\hat{\boldsymbol{\theta}}} \text{ is negative definite} \tag{3.16}$$

provided there is a unique solution at an inner point of $\Theta$. By $\boldsymbol{\theta}$ we denote a vector of unknown parameters located somewhere in the given parameter space $\Theta$.

The preceding process allows us to compute likelihoods for the nodes at both ends of any given branch, by simply assuming the root to be in that branch and "pruning" the likelihoods from the external node down until they arrive at the nodes at the two ends of the branch. We can then use these to find the length of that branch that optimizes the likelihood (Felsenstein 1973[73], 1981[76]).

We now consider to solve an equation numerically. While most equations are born with both a right-hand side and a left-hand side, one traditionally moves all terms to the left, leaving

$$f(x) = 0 \tag{3.17}$$

whose solution is desired. When there is only one independent variable, the problem is one-dimensional, namely to find the root of a function. The Newton-Raphson method requires us to evalute both the function $f(x)$ and its derivative $f'(x)$ at arbitrary points $x$. The formula consists geometrically of extending the tangent line at a current point $x_i$ until it crosses zero, then setting the next guess $x_{i+1}$ to the abscissa of that zero-crossing. The formula is

$$x_{i+1} = x_i - f(x_i) \Big/ \left(\frac{d}{dx} f(x_i)\right). \tag{3.18}$$

Similarly, the MLE $\hat{t}$ of $t$ is the solution of

$$\text{maximize} \quad l(t). \tag{3.19}$$

The problem is to find the maximum point of a function. The Newton-Raphson method requires us to evaluate the function $l(t)$, the first derivative $l'(t)$ and the second derivatives $l''(t)$ at arbitrary points $t$. The formula is

$$t_{i+1} = t_i - \left(\frac{d}{dt} l(t_i)\right) \Big/ \left(\frac{d^2}{dt^2} l(t_i)\right). \tag{3.20}$$

We can obtain the maximum likelihood estimate of $t$ through the Newton-Raphson method, in which calculations of $l$, $\nabla l$ and $\nabla \nabla^T l$ are necessary (Kishino et al. 1990[166]) and we have

$$P_{ij}(t) = \sum_{k=1}^{m} \left(U_{ik} U^{-1}_{kj} \exp(t\lambda_k)\right) \tag{3.21}$$

$$\frac{d}{dt}P_{ij}(t) = \sum_{k=1}^{m}\left(U_{ik}U_{kj}^{-1}\lambda_k \exp(t\lambda_k)\right) \tag{3.22}$$

$$\frac{d^2}{dt^2}P_{ij}(t) = \sum_{k=1}^{m}\left(U_{ik}U_{kj}^{-1}\lambda_k^2 \exp(t\lambda_k)\right). \tag{3.23}$$

**Internal Branch Length**

The log-likelihood of the tree at the $k$-th internal branch is rewitten as

$$l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\text{ans})}\sum_{j=1}^{m}P_{ij}(t_k)Q_{hj}^{(\text{des})}\right). \tag{3.24}$$

From Eqs. 3.22 and 3.23 we can compute the first derivative and the second derivative of the log-likelihood function with respect to the $k$-th internal branch length

$$\frac{d}{dt}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\text{ans})}\sum_{j=1}^{m}\frac{d}{dt}P_{ij}(t_k)Q_{hj}^{(\text{des})}\right) \tag{3.25}$$

$$\frac{d^2}{dt^2}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\text{ans})}\sum_{j=1}^{m}\frac{d^2}{dt^2}P_{ij}(t_k)Q_{hj}^{(\text{des})}\right). \tag{3.26}$$

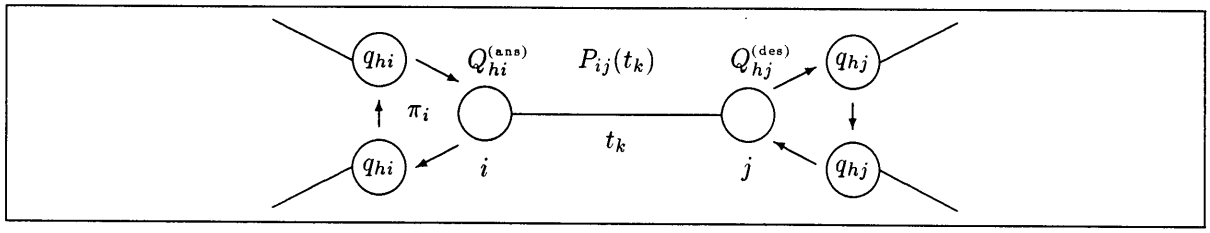



Figure 3.8: MLE internal branch Length by Newton-Raphson method.

**External Branch Length**

The log-likelihood of the tree at the $k$-th external branch is rewitten as

$$l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\text{ans})}P_{iX_{kh}}(t_k)\right). \tag{3.27}$$

From Eqs. 3.22 and 3.23 we can compute the first derivative and the second derivative of the log-likelihood function with respect to the $k$-th external branch length

$$\frac{d}{dt}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\text{ans})}\frac{d}{dt}P_{iX_{kh}}(t_k)\right) \tag{3.28}$$

$$\frac{d^2}{dt^2}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\text{ans})}\frac{d^2}{dt^2}P_{iX_{kh}}(t_k)\right). \tag{3.29}$$

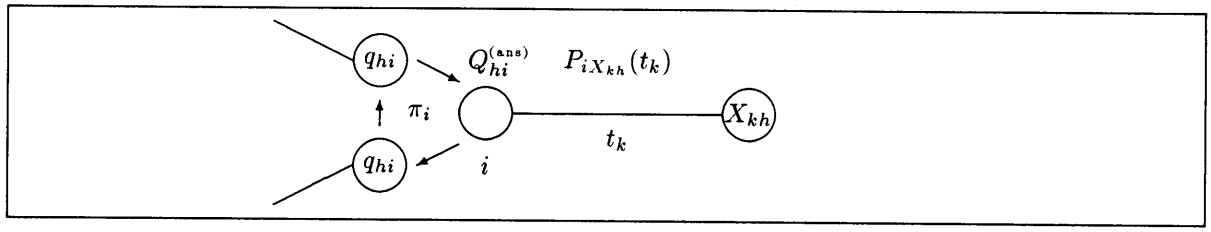



Figure 3.9: MLE external branch Length by Newton-Raphson method.

Using a new method that will be described in Section 3.4, we can recursively compute the quantities $l^{(k)}$ from the $(k = 1)$ -st branch up to the $(k = 2N - 3)$ -th branch. Traversing through the tree, branch lengths are successively optimized until an adequate number of traversals has occurred.

### 3.3.4 Estimation of Distance Matrix by ML

#### Initial Distance Matrix by Poisson Process

If transition probabilities are equal among different pairs of bases (amino acids), the number of substitutions per site between the $i$-th and $j$-th sequences is estimated by

$$D_{ij}^{(\text{init})} = -\frac{m-1}{m}\log\left(1 - \frac{mD_{ij}^{(\text{diff})}}{n(m-1)}\right) \tag{3.30}$$

where $n$ is the length of the sequence, $m$ is the number of base (amino acid) states and $D_{ij}^{(\text{diff})}$ is the number of differences between $i$-th and $j$-th sequences. This estimate is used as an initial distance provided for the ML analysis.

#### Distance Matrix by ML

The maximum likelihood estimate of $D$ is obtained through the Newton-Raphson method, in which calculations of $dl/dt$ and $d^2l/dt^2$ are necesary and we have Eq. 3.22 and 3.23. This can be done by a direct search.

The initial value of $D_{ij}$ denoted by $D_{ij}^{(\text{init})}$, is calculated under the Poisson process. Estimate $D_{ij}$ as a renewal by the Newton-Raphson method that maximizes

$$l(D_{ij}|\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)}) = \sum_{h=1}^{n} \log\left(P_{X_{ih}X_{jh}}(D_{ij})\right) \tag{3.31}$$

where $D_{ij}$ is the number of substitutions per site between $i$-th and $j$-th sequences.

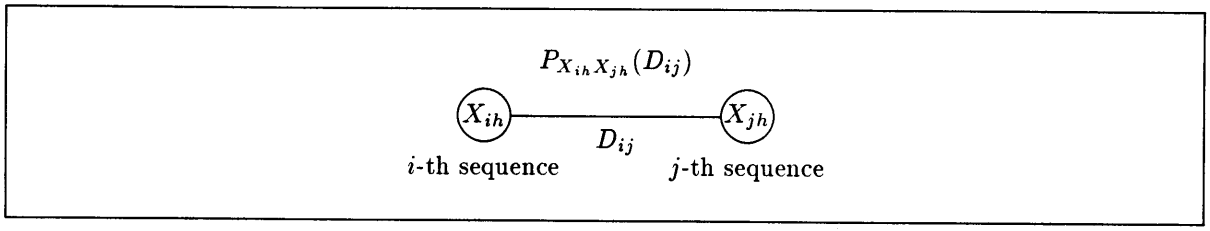



Figure 3.10: MLE distance by Newton-Raphson method.

### 3.3.5 Estimation of Initial Branch Lengths

#### Initial Branch Lengths by Least Squares

We have observed values of an $(n \times 1)$ vector $\boldsymbol{D}$ where $n = N(N-1)/2$ and $N$ is number of OTUs and of an $(n \times k)$ matrix $\boldsymbol{A}$ of full column rank $k$. If the tree is a bifurcating tree then $k = 2N - 3$. $\boldsymbol{A}$ is called a tree topology matrix. We regard $\boldsymbol{D}$ as the realization of a random vector $\boldsymbol{\epsilon}$ that is generated by

$$\boldsymbol{D} = \boldsymbol{A}\boldsymbol{t} + \boldsymbol{\epsilon} \tag{3.32}$$

with $\boldsymbol{t}$ a $(k \times 1)$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ an $(n \times 1)$ vector of independent Normal variates with zero mean and unknown variance $\sigma^2$. For the tree in Fig. 3.1b,

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{t} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \\ t_4 \\ t_5 \\ t_6 \\ t_7 \\ t_8 \\ t_9 \end{bmatrix} \quad \text{and} \quad \boldsymbol{D} = \begin{bmatrix} D_{12} \\ D_{13} \\ D_{14} \\ D_{15} \\ D_{16} \\ D_{23} \\ D_{24} \\ D_{25} \\ D_{26} \\ D_{34} \\ D_{35} \\ D_{36} \\ D_{45} \\ D_{46} \\ D_{56} \end{bmatrix}.$$

We find the least squares estimate $\hat{\boldsymbol{t}}$ by minimizing

$$\min\{\boldsymbol{S}(\boldsymbol{t})\} = \min\{(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{t})^T(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{t})\} \tag{3.33}$$

(Chakraborty 1977[53]).

The standard Ordinary Least Squares (OLS) estimator of $\boldsymbol{t}$ ,

$$\hat{\boldsymbol{t}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{D} \tag{3.34}$$

with (asymptotic) covariance matrix

$$V\hat{\boldsymbol{t}} = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}. \tag{3.35}$$

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 5 & 1 & 1 & 1 & 1 & 1 & 4 & 2 & 2 \\ 1 & 5 & 1 & 1 & 1 & 1 & 4 & 2 & 2 \\ 1 & 1 & 5 & 1 & 1 & 1 & 2 & 4 & 2 \\ 1 & 1 & 1 & 5 & 1 & 1 & 2 & 4 & 2 \\ 1 & 1 & 1 & 1 & 5 & 1 & 2 & 2 & 4 \\ 1 & 1 & 1 & 1 & 1 & 5 & 2 & 2 & 4 \\ 4 & 4 & 2 & 2 & 2 & 2 & 8 & 4 & 4 \\ 2 & 2 & 4 & 4 & 2 & 2 & 4 & 8 & 4 \\ 2 & 2 & 2 & 2 & 4 & 4 & 4 & 4 & 8 \end{bmatrix}$$

$$(\boldsymbol{A}^T\boldsymbol{A})^{-1} = \begin{bmatrix} 3/8 & 1/8 & 0 & 0 & 0 & 0 & -1/4 & 0 & 0 \\ 1/8 & 3/8 & 0 & 0 & 0 & 0 & -1/4 & 0 & 0 \\ 0 & 0 & 3/8 & 1/8 & 0 & 0 & 0 & -1/4 & 0 \\ 0 & 0 & 1/8 & 3/8 & 0 & 0 & 0 & -1/4 & 0 \\ 0 & 0 & 0 & 0 & 3/8 & 1/8 & 0 & 0 & -1/4 \\ 0 & 0 & 0 & 0 & 1/8 & 3/8 & 0 & 0 & -1/4 \\ -1/4 & -1/4 & 0 & 0 & 0 & 0 & \frac{7}{16} & -1/16 & -1/16 \\ 0 & 0 & -1/4 & -1/4 & 0 & 0 & -1/16 & 7/16 & -1/16 \\ 0 & 0 & 0 & 0 & -1/4 & -1/4 & -1/16 & -1/16 & 7/16 \end{bmatrix}$$

$$\boldsymbol{A}^T\boldsymbol{D} = \begin{bmatrix} D_{12}+D_{13}+D_{14}+D_{15}+D_{16} \\ D_{12}+D_{23}+D_{24}+D_{25}+D_{26} \\ D_{13}+D_{23}+D_{34}+D_{35}+D_{36} \\ D_{14}+D_{24}+D_{34}+D_{45}+D_{46} \\ D_{15}+D_{25}+D_{35}+D_{45}+D_{56} \\ D_{16}+D_{26}+D_{36}+D_{46}+D_{56} \\ D_{13}+D_{14}+D_{15}+D_{16}+D_{23}+D_{24}+D_{25}+D_{26} \\ D_{13}+D_{14}+D_{23}+D_{24}+D_{35}+D_{36}+D_{45}+D_{46} \\ D_{15}+D_{16}+D_{25}+D_{26}+D_{35}+D_{36}+D_{45}+D_{46} \end{bmatrix}$$

$$\hat{\boldsymbol{t}} = \begin{bmatrix} D_{12}/2+(D_{13}+D_{14}+D_{15}+D_{16})/8-(D_{23}+D_{24}+D_{25}+D_{26})/8 \\ D_{12}/2+(D_{23}+D_{24}+D_{25}+D_{26})/8-(D_{13}+D_{14}+D_{15}+D_{16})/8 \\ D_{34}/2+(D_{13}+D_{23}+D_{35}+D_{36})/8-(D_{14}+D_{24}+D_{45}+D_{46})/8 \\ D_{34}/2+(D_{14}+D_{24}+D_{45}+D_{46})/8-(D_{13}+D_{23}+D_{35}+D_{36})/8 \\ D_{56}/2+(D_{15}+D_{25}+D_{35}+D_{45})/8-(D_{16}+D_{26}+D_{36}+D_{46})/8 \\ D_{56}/2+(D_{16}+D_{26}+D_{36}+D_{46})/8-(D_{15}+D_{25}+D_{35}+D_{45})/8 \\ (D_{13}+D_{14}+D_{15}+D_{16}+D_{23}+D_{24}+D_{25}+D_{26})/8-D_{12}/2-(D_{35}+D_{36}+D_{45}+D_{46})/8 \\ (D_{13}+D_{14}+D_{23}+D_{24}+D_{35}+D_{36}+D_{45}+D_{46})/8-D_{34}/2-(D_{15}+D_{16}+D_{25}+D_{26})/8 \\ (D_{15}+D_{16}+D_{25}+D_{26}+D_{35}+D_{36}+D_{45}+D_{46})/8-D_{56}/2-(D_{13}+D_{14}+D_{23}+D_{24})/8 \end{bmatrix} \tag{3.36}$$

**Estimation of Branch Lengths by a New Simple Method**

$$t = \frac{1}{lr}\sum_{i\in L}\sum_{j\in R} D_{ij} - \frac{r}{lr(l-1)}\sum_{i\in L}\sum_{j\in L} D_{ij} - \frac{l}{lr(r-1)}\sum_{i\in R}\sum_{j\in R} D_{ij} \tag{3.37}$$

## 3.4 Fast Computation of ML for Inferring Evolutionary Trees

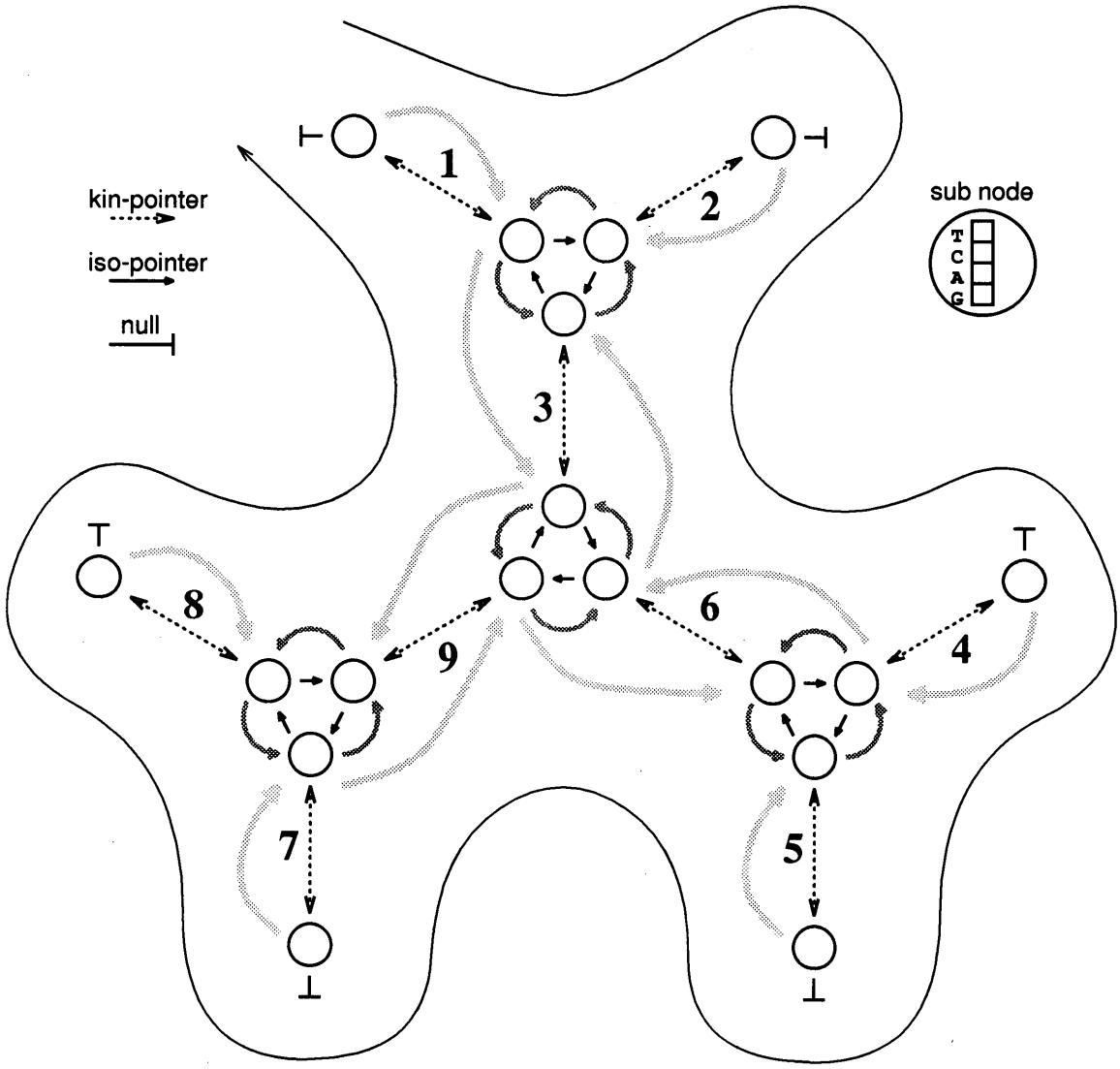



Figure 3.11: Fast computation algorithm.

```
cp = rp = tree->rootp;
do {
    cp = cp->isop->kinp;
    PRODUCT_Partial_Likelihood(cp->kinp->isop);
    if (cp->isop == NULL) { /* external node */
        cp = cp->kinp;
        MLE_Branch_Length(cp);
        RENEW_Partial_Likelihood(cp);
    } else {                /* internal node */
        if (cp->descen)
            RENEW_Partial_Likelihood(cp);
        else
            MLE_Branch_Length(cp);
            RENEW_Partial_Likelihood(cp);
    }
} while (cp != rp);
```

Table 3.2: Constant factors in comparing procedures.

| branch | method | DNAML | Prot/NucML |
|---|---|---|---|
| internal | MLE Branch Length | 1 | 1 |
| branch | RENEW Partial Likelihood | 4 | **2** |
| (N-3) | PRODUCT Partial Likelihood | 2 | 2 |
| external | MLE Branch Length | 1 | 1 |
| branch | RENEW Partial Likelihood | 2 | **1** |
| (N) | PRODUCT Partial Likelihood | 1 | 1 |

## 3.5 Topology Search Strategy for ML Phylogeny

### 3.5.1 Topological Data Structure

As a data structure representing the unrooted tree shown in Fig. 3.12a, Felsenstein considered Fig. 3.12b, where each internal node (excluding external nodes or tips) is decomposed into elements, the number of which coincides with those of branches stemming from the node. The elements are connected circularly through the pointers.

By adopting such data structure, a partial likelihood of a sub-tree stemming from the node can be stored. This means that, when the likelihood of the tree is estimated, we need not calculate likelihood through iteration of a loop by the times of the number of nodes in revising the estimate of each branch length, but need only revise the partial likelihoods of two nodes of each branch.

We extend this data structure so that a multifurcating tree can be represented. Since branches are connected dynamically by pointers, the data structure can easily be revised when different tree topology is adopted, and furthermore not only bifurcating trees but also multifurcating trees can be represented quite easily. The extreme of a multifurcating tree is the star-like tree shown in Fig. 3.12c.

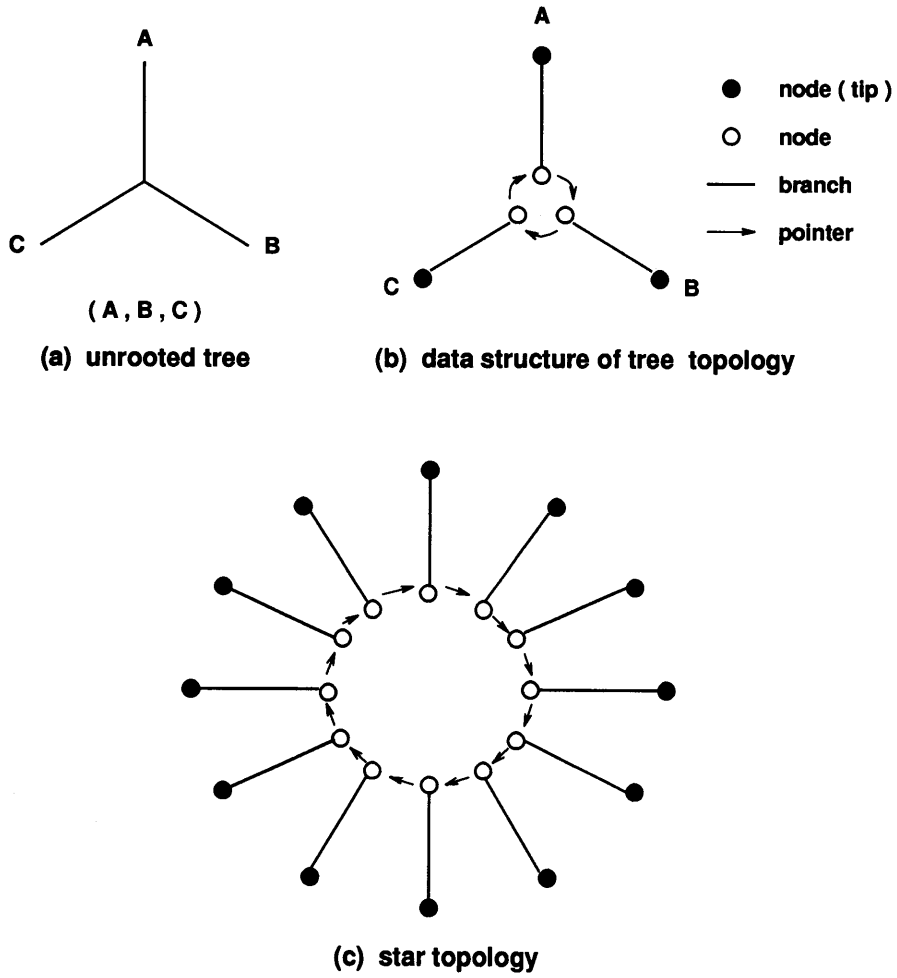



Figure 3.12: Topological data structure.

### 3.5.2 Automatic Topology Search by Star Decomposition

The straightforward approach in inferring a tree would be to evaluate all possible tree topology one after another and pick the one which gives the highest likelihood. This would not be possible for a large number of species, since the number of possible tree topologies is enormous (Felsenstein 1978[75]).

The strategy that Felsenstein's DNAML employs is as follows: the species are taken in the order in which they appear in the input file. The first three are taken and an unrooted tree is constructed containing only these three. Then the fourth species is taken, and it is evaluated where it might be added to the tree. All possibilities (bifurcating trees) of adding the fourth species are examined. The best one in the likelihood criterion is chosen as the basis for further operations. Then the fifth species is added, and again the best of the placing of it is chosen, and so on. At each step, local rearrangements of a tree are examined. This procedure is continued until a bifurcating tree connecting all the species is obtained (Felsenstein 1993[82]).

The resulting tree of this procedure generally depends on the order of the input species. Hence, Felsenstein recommends to perform a number of runs with different orderings of the input species.

The alternative strategy which we employ in the automatic and semi-automatic search options of ProtML is called "star decomposition"(Adachi and Hasegawa 1992[3]). This is similar to the procedure employed in the neighbor-joining method using distance matrix (Saitou and Nei 1987[253]). This starts with a star-like tree (Fig. 3.12c). Decomposing the star-like tree step by step, we finally obtain a bifurcating tree if a multifurcation can be resolved with statistical confidence. Since the information from all of the species under analysis is used from the beginning, the inference of the tree topology is likely to be stable by this procedure.

Let $n$ be the number of species under analysis. At first, a star-like tree containing $n$ species is constructed (Fig. 3.13a, $n = 6$ in this case). Then, a pair of species is separated from others as shown in Fig. 3.13b. Among all possible pairwise combinations of species, a pairing that gives the highest likelihood is chosen (Fig. 3.13c). The resulting tree can be regarded as a star-like tree with $n - 1$ groups (a single species may form a group), if the selected pair is regarded to form a group. This procedure is continued until all multifurcating nodes are resolved into bifurcating ones as shown in Fig. 3.13f.

When the information content of the data is not large enough to discriminate among alternative branching orders, it might be misleading to resolve all the multifurcations into bifurcations. Hence, by using AIC (Akaike 1973[11], 1974[12]), the program decides whether the multifurcation should further be resolved or not.

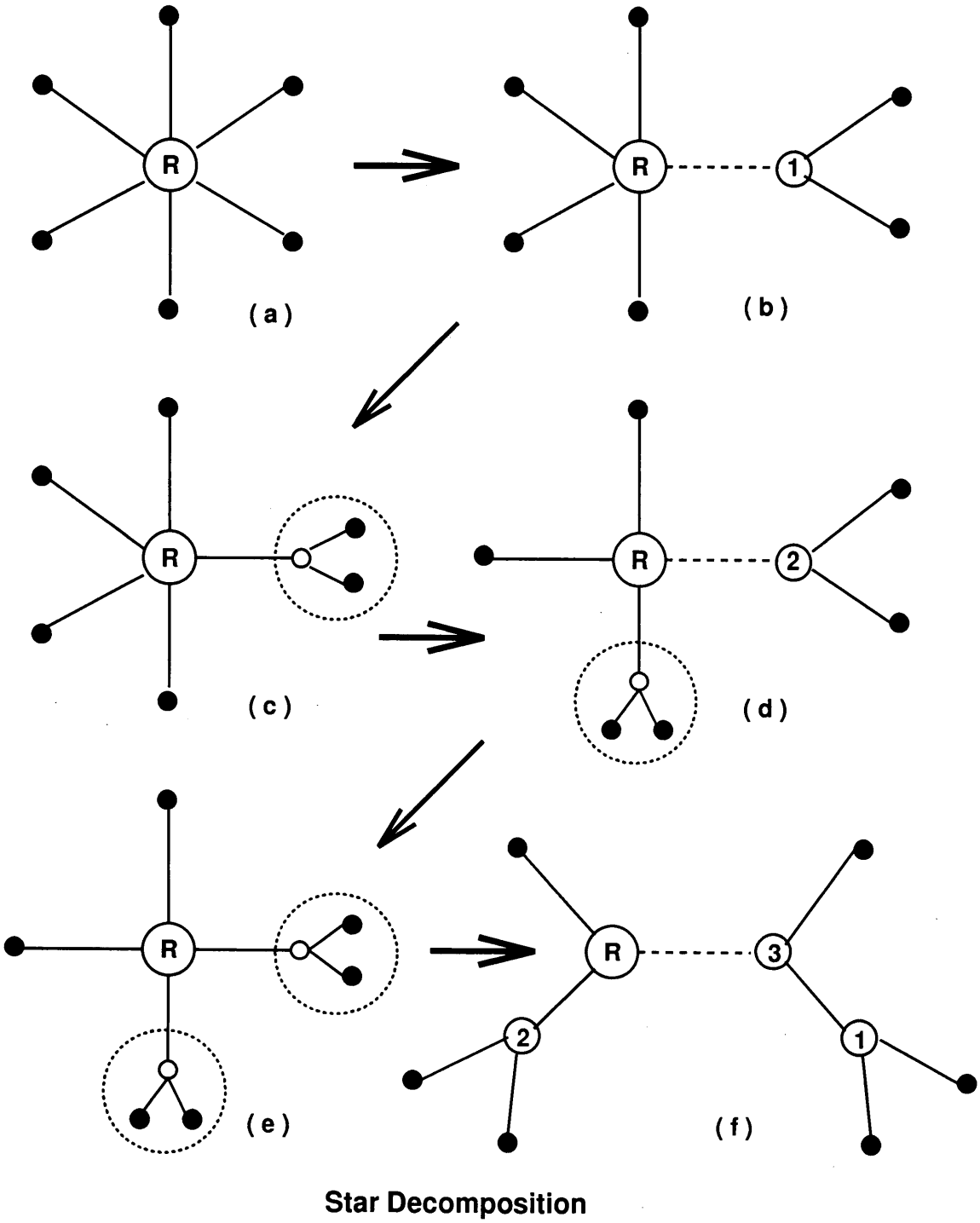
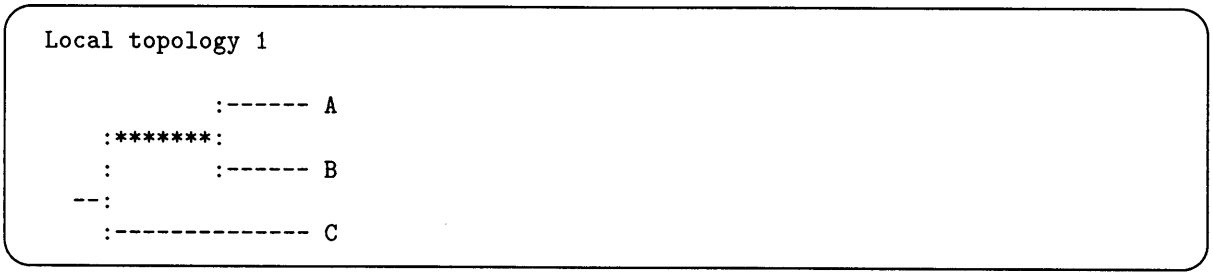



Figure 3.13: Star decomposition.

### 3.5.3 Topology Search by Local Rearrangements

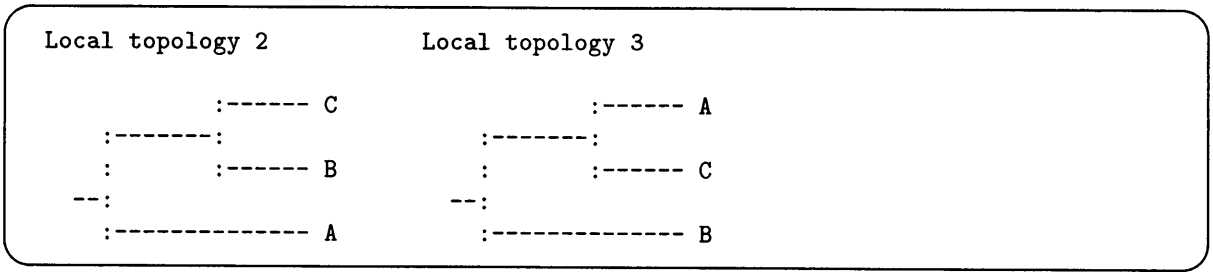
Once an approximate tree topology is obtained by the star decomposition mentioned in the preceding subsection, by either method of distance matrix, or by the parsimony, the search for better tree topologies by the likelihood criterion can be conducted through local rearrangement which is similar to the method used in the DNAML program of PHYLIP (Felsenstein 1993[82]) and will be described below.

Suppose we have obtained an approximate tree topology by either method. Each internal branch of the tree is of the following form;




where A, B, and C are subtrees.

A local rearrangement takes account of the additional two alternative trees;




and estimates bootstrap probabilities among these three trees by the RELL method (Kishino et al. 1990[166]; Hasegawa and Kishino 1994[119]). Since the branching orders within the subtrees, A, B and C, are fixed, these are not real bootstrap probabilities, and I will call them local bootstrap probabilities (LBPs). It must be noted that the LBP might be misleading when the relationships within respective groups (subtrees) attached to the branch are incorrect. LBP can be interpreted as bootstrap probability of the particular internal branch when the other parts of the tree is correct.

If it turned out that another local tree topology than Local topology 1 has higher likelihood and hence higher LBP, then the local rearrangement is carried out for this branch. This procedure is repeated until all the internal branches are traversed. But a rearrangement around a branch may make the previously established branches need to be rearranged, and the phase of local rearrangements does not end until the program can traverse the entire tree, attempting local rearrangements, without finding any that improve the likelihood. Suppose we have obtained a tree for which no local rearrangement can improve the likelihood. When two, three, or four contiguous branches in the tree is ambiguous, then we can examine 15, 105, or 945 alternative topologies relevant to these branches, and we can look for a better tree topology. By using this procedure (extended local rearrangement), we may be able to reduce the possibility of being trapped in a local optimum.

It is not guaranteed that the tree obtained by this procedure is the highest likelihood tree, and the tree depends on the initial tree. For this reason, use of several alternative initial trees is recommended, and a tree with the highest likelihood among several runs should be chosen. For example, NJ analyses with bootstrap resampling might be useful in order to generate alternative intial trees.

### 3.5.4 Example of Application of the Local Rearrangements

I will show an example of application of the local rearrangement method described in the preceding subsection. I will apply this method to the amino acid sequences of elongation factor 1$\alpha$ (EF-1$\alpha$), which were used in Hashimoto et al. (1995[130]) and are listed in Table 3.3

Table 3.3: List of EF-1$\alpha$ data.

| Abbrev. | species name | reference | database |
|---|---|---|---|
| Metazoa | | | |
| Homsa | *Homo sapiens* | Uetsuki'89[299]) | X03558 |
| Xenla | *Xenopus laevis* | Krieg'89[176] | X52975 |
| Drome | *Drosophila melanogaster* | Hoveman'88[143] | X06869 |
| Artsa | *Artemia salina* | van Hemert'84[301] | X03349 |
| Fungi | | | |
| Sacce | *Saccharomyces cerevisiae* | Nagashima'86[222] | X00779 |
| Canal | *Candida albicans* | Sundstrom'90[281] | M29934 |
| Mucra | *Mucor racemosus* | Linz'86[194] | J02605 |
| Absgl | *Absidia glauca* | Burmester (unpubl.) | X54730 |
| Plantae | | | |
| Arath | *Arabidopsis thaliana* | Liboz'89[193] | X16430 |
| Lyces | *Lycopersicon esculentum* | Pokalsky'89[242] | X53043 |
| Protista | | | |
| Dicdi | *Dictyostelium discoideum* | Yang'90[315] | X55972 |
| Euggr | *Euglena gracilis* | Montandon'90[219] | X16890 |
| Trycr | *Trypanosoma cruzi* | Hashimoto'95[130] | D29834 |
| Tetpy | *Tetrahymena pyriformis* | Kurosawa'92[180] | D11083 |
| Plafa | *Plasmodium falciparum* | Williamson (unpubl.) | X60488 |
| Enthi | *Entamoeba histolytica* | De Meester'91[63] | M34256 |
| Giala | *Giardia lamblia* | Hashimoto'94[132] | D14342 |
| Archaebacteria | | | |
| Sulac | *Sulfolobus acidocaldarius* | Auer'90[27] | X52382 |
| Metva | *Methanococcus vannielii* | Lechner'87[185] | X05698 |
| Halma | *Halobacterium marismortui* | Baldacci'90[35] | X16677 |

Fig. 3.14 shows the process of the local rearrangements applied to the EF-1$\alpha$ data. The NJ tree is shown in Fig. 3.14 (a). The distance matrix provided for the NJ analysis was estimated for 2-OTUs trees by the ProtML based on the JTT-F model. In this figure, a LBP (in %) is given following internal branch (or node) number. When the local branching order is not optimum, an asterisk is attached to the LBP value. In this example, two branches are attached by asterisks. For the branch 25, it notes

```
-25  30*  68(21,26)
```

This means that branch 25 has 30% LBP, but if node 21 and node 26 are linked, LBP becomes 68%. Furthermore, for branch 46, it notes

```
-31   46*  47(28,30)
```

Rearrangement is done at first for a branch for which the largest amount of LBP change is obtained as indicated at the bottom of Fig. 3.14(a)

```
#1  25  (21, 26)  ln L: -7111.97 + 4.66
```

This means that, by linking node 21 with node 26, log-likelihood of the preceding tree ($-7111.97$) is improved by 4.66.

Fig. 3.14(b) gives the tree obtained by linking node 21 with node 26 in the preceding tree, and an astrisk is attached on the branches 31 and 34. By linking node 28 with node 30 as indicated in the figure, log-likelihood is improved by 1.28. In this way, we have a final tree shown in Fig. 3.14(c), which gives also branch lengths (with SE), LBPs, and LBPs of second best local arrangements of branches. For the internal branch 31, the following is noted

```
LBP
0.471  0.413 (29,30)
```

This means that the LBP of this branch is 47.1%, and rearrangement of linking node 29 with node 30 is the second best giving LBP of 41.3%.

Fig. 3.14(c) shows the ProtML tree (based on the JTT-F model) which cannot be improved any more by the local rearrangements. The log-likelihood of the NJ tree is $-7111.97$, while that of the resultant ProtML tree is $-7106.02$, showing an improvement of log-likelihood by 5.95.

In the NJ tree, the fungi clade ((Sacce, Canal), (Mucra, Absgl)) intrudes into metazoa, and links with vertebrates (Homsa, Xenla) excluding arthropoda (Drome, Artsa) as an outgroup. This relationship cannot be accepted from biological background. In the ProtML tree, on the otherhand, metazoa is monophyletic and is a sister group to fungi. The ProtML tree is more reasonable than the NJ tree in this respect.

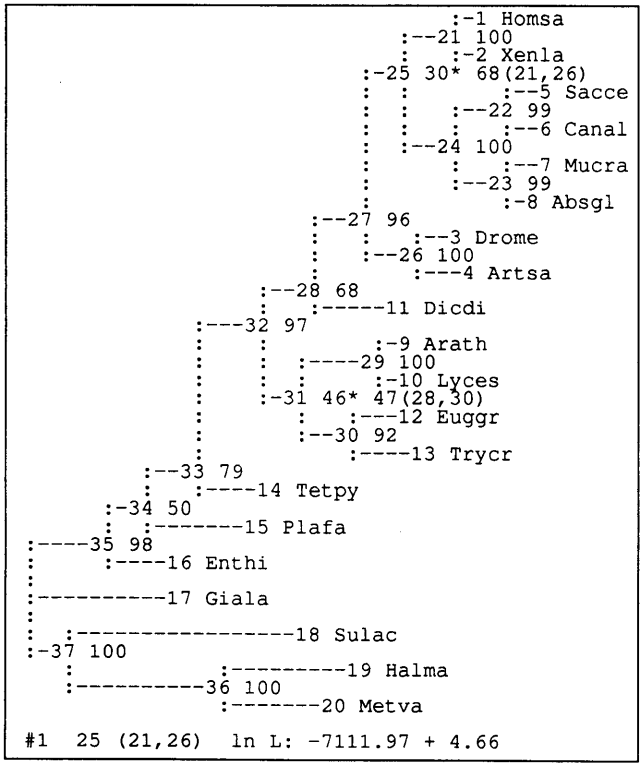



Figure 3.14: (a). Example of application of the local rearrangements, part 1.

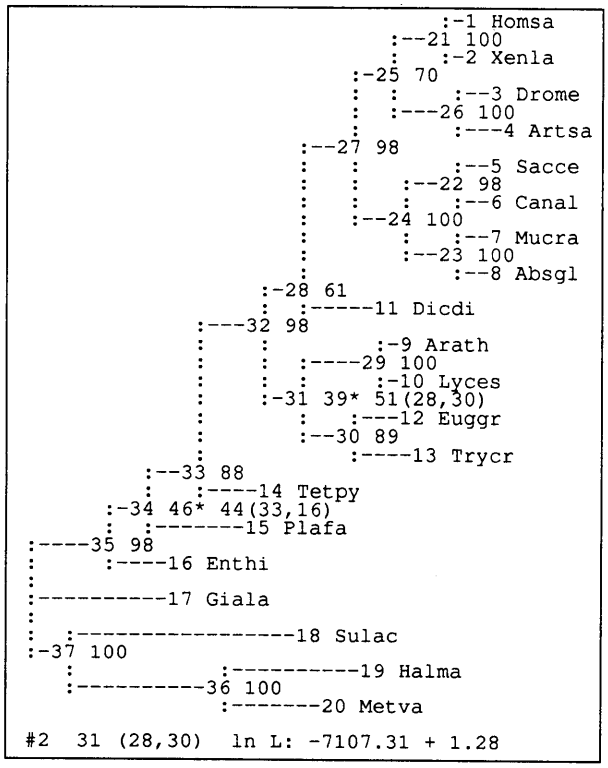
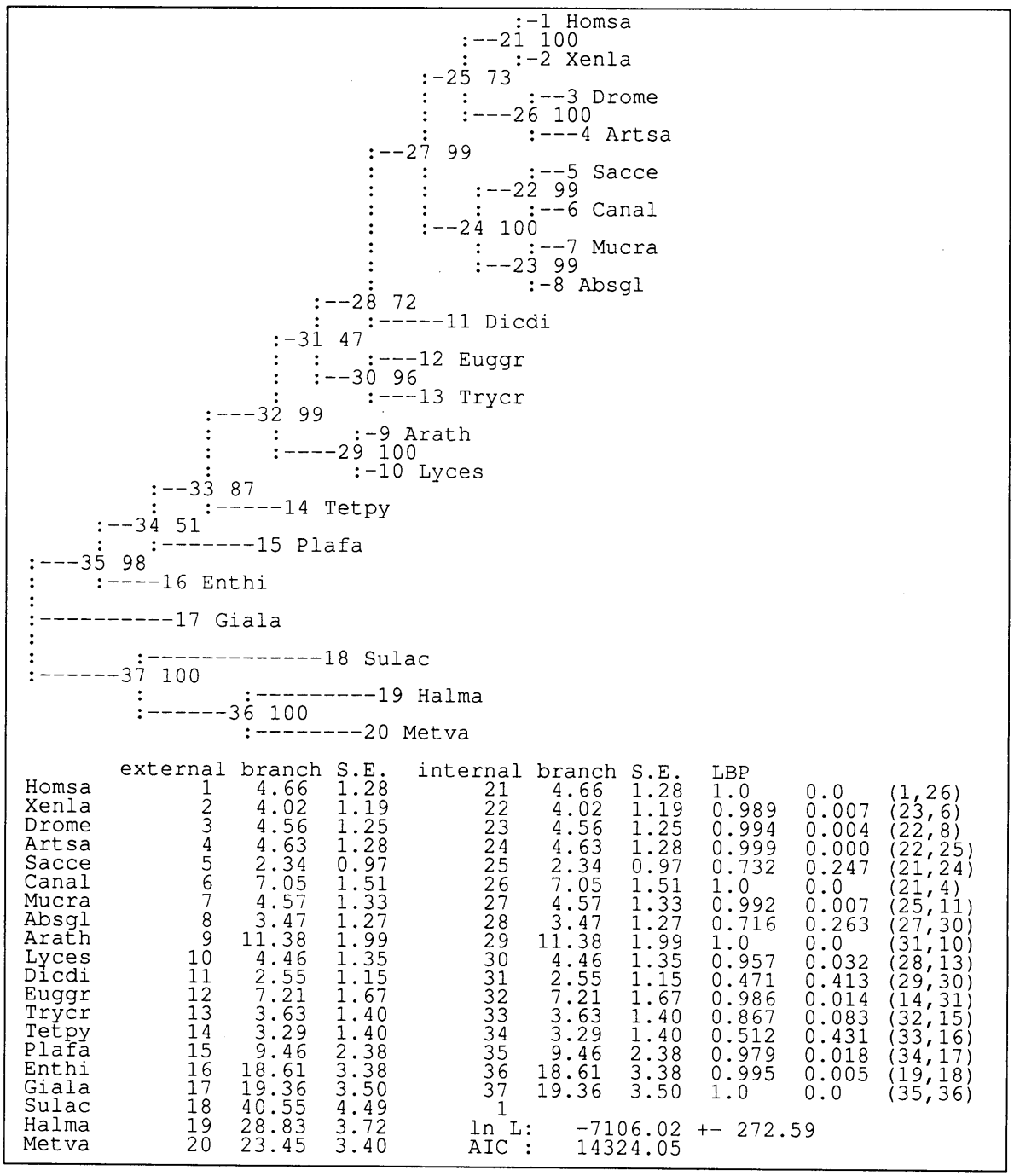



Figure 3.14: (b). Example of application of the local rearrangements, part 2.




| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Homsa | 1 | 4.66 | 1.28 | 21 | 4.66 | 1.28 | 1.0 | 0.0 | (1,26) |
| Xenla | 2 | 4.02 | 1.19 | 22 | 4.02 | 1.19 | 0.989 | 0.007 | (23,6) |
| Drome | 3 | 4.56 | 1.25 | 23 | 4.56 | 1.25 | 0.994 | 0.004 | (22,8) |
| Artsa | 4 | 4.63 | 1.28 | 24 | 4.63 | 1.28 | 0.999 | 0.000 | (22,25) |
| Sacce | 5 | 2.34 | 0.97 | 25 | 2.34 | 0.97 | 0.732 | 0.247 | (21,24) |
| Canal | 6 | 7.05 | 1.51 | 26 | 7.05 | 1.51 | 1.0 | 0.0 | (21,4) |
| Mucra | 7 | 4.57 | 1.33 | 27 | 4.57 | 1.33 | 0.992 | 0.007 | (25,11) |
| Absgl | 8 | 3.47 | 1.27 | 28 | 3.47 | 1.27 | 0.716 | 0.263 | (27,30) |
| Arath | 9 | 11.38 | 1.99 | 29 | 11.38 | 1.99 | 1.0 | 0.0 | (31,10) |
| Lyces | 10 | 4.46 | 1.35 | 30 | 4.46 | 1.35 | 0.957 | 0.032 | (28,13) |
| Dicdi | 11 | 2.55 | 1.15 | 31 | 2.55 | 1.15 | 0.471 | 0.413 | (29,30) |
| Euggr | 12 | 7.21 | 1.67 | 32 | 7.21 | 1.67 | 0.986 | 0.014 | (14,31) |
| Trycr | 13 | 3.63 | 1.40 | 33 | 3.63 | 1.40 | 0.867 | 0.083 | (32,15) |
| Tetpy | 14 | 3.29 | 1.40 | 34 | 3.29 | 1.40 | 0.512 | 0.431 | (33,16) |
| Plafa | 15 | 9.46 | 2.38 | 35 | 9.46 | 2.38 | 0.979 | 0.018 | (34,17) |
| Enthi | 16 | 18.61 | 3.38 | 36 | 18.61 | 3.38 | 0.995 | 0.005 | (19,18) |
| Giala | 17 | 19.36 | 3.50 | 37 | 19.36 | 3.50 | 1.0 | 0.0 | (35,36) |
| Sulac | 18 | 40.55 | 4.49 | 1 | | | | | |
| Halma | 19 | 28.83 | 3.72 | ln L: | -7106.02 | +- 272.59 | | | |
| Metva | 20 | 23.45 | 3.40 | AIC : | 14324.05 | | | | |

Figure 3.14: (c). Example of Application of the Local Rearrangements, part 3.

## 3.6 Approximate Likelihood Method for Exhaustive Search

Many authors wrote that, since the ML method is vastly more computationally intensive than the NJ and MP methods, the usefulness of the ML method in molecular phylogenetics may be limited (e.g., Nei 1987[224]; Hillis et al. 1994[139]). It is true that the ML method is computationally intensive and that, at present, there exists several limitations in applying the method to real problems, but our computational environment is rapidly improving now. Furthermore, several methods to reduce the computational burden of the ML analyses are being invented. One of them is that, once the ML analysis of the data has been done adequately, we can estimate the BPs quite easily by using the RELL method without performing ML estimation for each resampled data set. Another is the approximate likelihood method developed in this thesis.

The most serious problem of the ML method when applied to data from many species is the explosively increasing number of possible tree topologies. However, most of these trees are too bad and too unpromising in the likelihood criterion to be taken as candidates. If we can quickly eliminate these trees by an approximate way, the ML method can be applied to many species cases. In estimating the branch lengths for each tree topologies by the ML, we use the Newton-Raphson method which is time consuming. The intitial values of the Newton-Raphson method are given by the least squares. It is shown that there is a remarkablly good correlation between the likelihood calculated from the intitial values, that is called the approximate likelihood (AL), and the likelihood estimated by the ML. Therefore, we can exclude unpromising trees by using an AL which can be calculated rather quickly.

The approximate log-likelihood of a tree is

$$l(\hat{\boldsymbol{t}}|\boldsymbol{X},T)=\sum_{h=1}^{n}\log f(\boldsymbol{X}_h|T,\hat{\boldsymbol{t}}) \tag{3.38}$$

where

$$\hat{\boldsymbol{t}}=\left(\hat{t}_1,\hat{t}_2,\cdots,\hat{t}_9\right)^T. \tag{3.39}$$

We have observed values of a distance vector $\boldsymbol{D}$ and a tree topology matrix $\boldsymbol{A}$. The $\boldsymbol{t}$ a vector of unknown coefficients is branch lengths. For the tree in Fig. 3.1b, The standard Ordinary Least Squares (OLS) estimator of $\boldsymbol{t}$

$$\hat{\boldsymbol{t}}=(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{D}. \tag{3.40}$$

For example, if we are dealing with 10 species, the number of possible unrooted tree topologies which should be examined are 2,027,025. Although this number may seem terribly large, we can examine all these topologies with the AL method by using a workstation within a reasonable time. Even when we are dealing with more than 10 species, if species can be clustered in advance into 10 or less groups, full topology search among these groups is attainable. Thus we can exclude unpromising trees by the AL method, and can select the best, say 500 or 1000, trees (by the AL criterion) that are provided for the full ML analysis.

Fig. 3.15 gives the relationship between the approximate likelihood and the likelihood estimated by the ML for the possible 945 trees of EF-1$\alpha$ sequences from 7 species chosen from the list in Table 3.3; *Homo sapiens*, *Drosophila melanogaster*, *Candida albicans*, *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Euglena gracilis*, and *Entamoeba histolytica*. These species are all eukaryotes, and it turned out that the AL is a good approximation of the likelihood estimated by the ML.

Fig. 3.16 gives the relationship between the AL and the likelihood estimated by the ML for the EF-1$\alpha$ data from 5 species chosen from the list in Table 3.3 with additional two archaebacterial species; *Homo sapiens*, *Entamoeba histolytica*, *Sulfolobus acidocaldarius*, *Methanococcus vannielii*, *Halobacterium marismortui*, *Thermococcus celer* (Auer et al. 1990[26]), and *Thermoplasma acidophilum* (Tesch and Klink 1990[291]). This data set contains more diversed species (including eukaryotes and archaebacteria) than the preceding one, and the correlation between the AL and ML is not as good as that shown in Fig. 3.15, but still the correlation seems to be good enough for the AL method to be applicable.

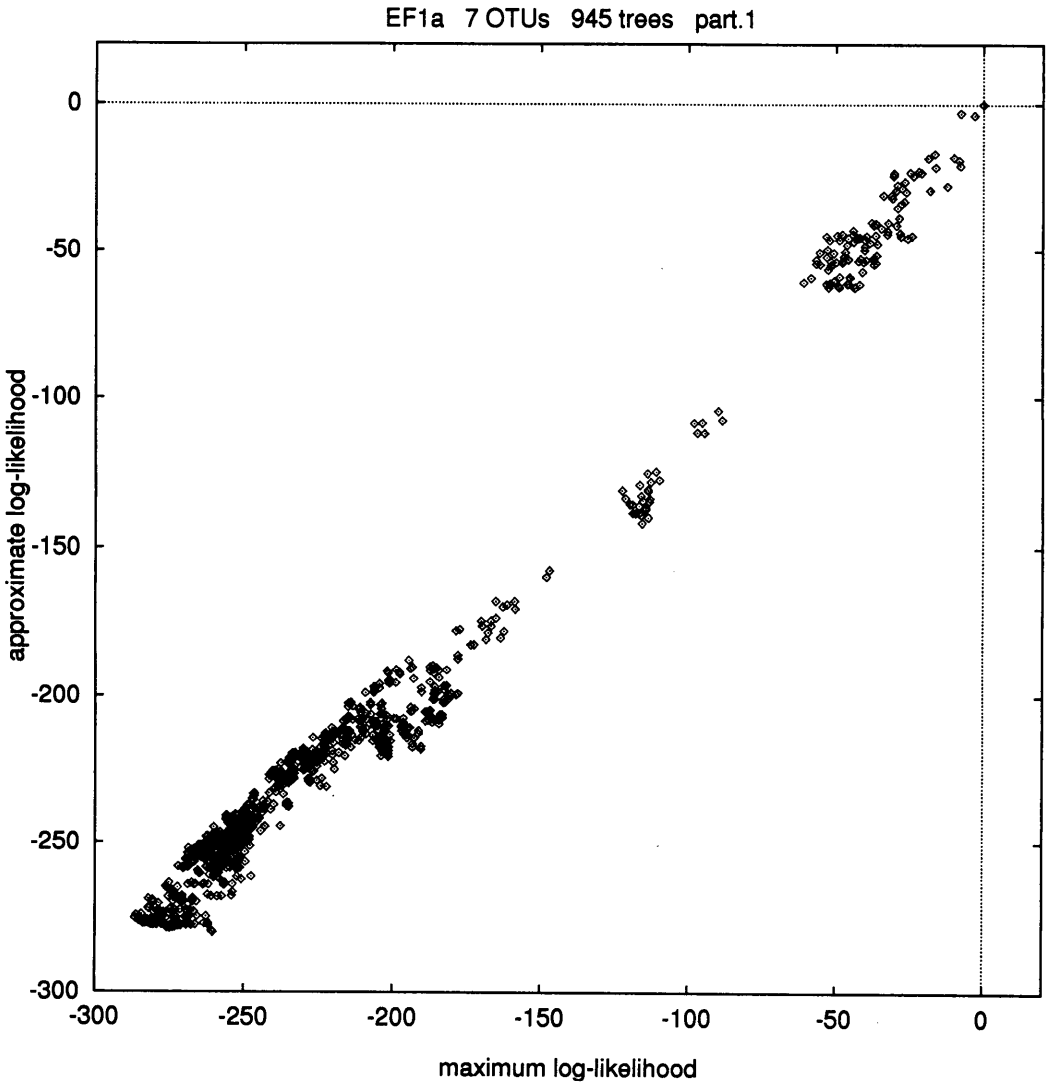



Figure 3.15: Maximum likelihood vs. Approximate likelihood, part 1.

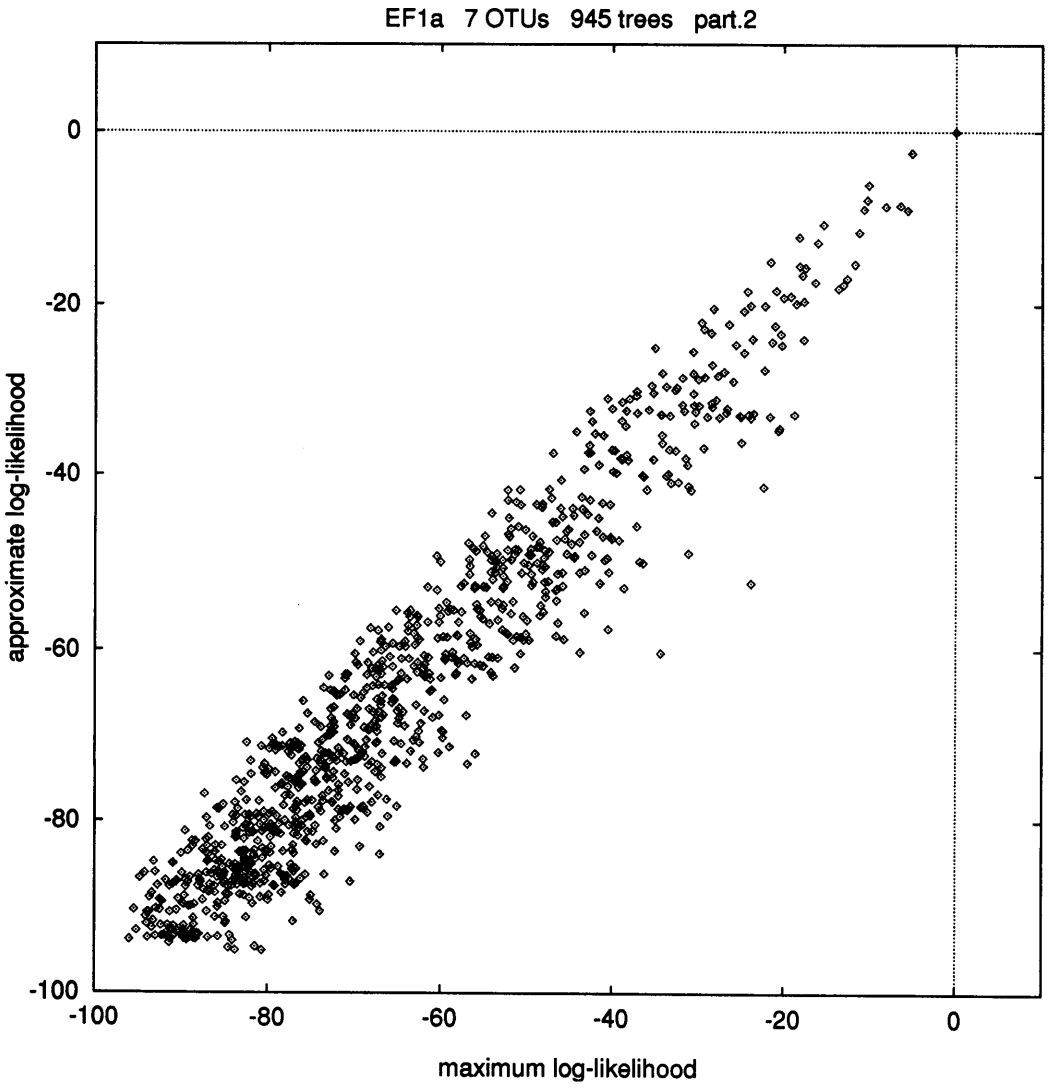



Figure 3.16: Maximum likelihood vs. Approximate likelihood, part 2.

# Chapter 4

# MOLPHY: Computer Programs for Molecular Phylogenetics

**Readme**

```
        This is the MOLPHY (ProtML) distribution,  version 2.3.
        Copyright (c) 1992-1995, Jun Adachi & Masami Hasegawa.
                        All rights reserved.

        MOLPHY is a program package for MOLecular PHYlogenetics.

ProtML is a main program in MOLPHY for inferring evolutionary trees from
PROTein (amino acid) sequences by using the Maximum Likelihood method.

Programs (C language)
  ProtML: Maximum Likelihood Inference of Protein Phylogeny
  NucML:  Maximum Likelihood Inference of Nucleic Acid Phylogeny
  ProtST: Basic Statistics of Protein Sequences
  NucST:  Basic Statistics of Nucleic Acid Sequences
  NJdist: Neighbor Joining Phylogeny from Distance Matrix

Utilities (Perl)
  mollist:  get identifiers list        molrev:   reverse DNA sequences
  molcat:   concatenate sequences       molcut:   get partial sequences
  molmerge: merge sequences             nuc2ptn:  DNA -> Amino acid
  rminsdel: remove INS/DEL sites        molcodon: get 3rd(1st,2nd) codons
  molinfo:  get (non)infomation sites   mol2mol:  MOLPHY format beautifer
  inl2mol:  Interleaved -> MOLPHY       mol2inl:  MOLPHY -> Interleaved
  mol2phy:  MOLPHY -> Sequential        phy2mol:  Sequential -> MOLPHY
  must2mol: MUST -> MOLPHY              etc.

MOLPHY is a free software, and you can use and redistribute it.
The programs are written in a standard subset of C with UNIX-like OS.
The utilities are written in the "Perl" (Ver.4.036) with UNIX-like OS.
MOLPHY has been tested on SUN4's (cc & gcc with SUN-OS 4.1.3) and
HP9000/700 (cc, c89 & gcc with HP-UX 9.05).
But, MOLPHY has NOT been tested on VAX, IBM-PC, and Macintosh.

NETWORK DISTRIBUTION ONLY: The latest version of MOLPHY are always available
by anonymous ftp in sunmh.ism.ac.jp(133.58.12.20): /pub/molphy*.

Followings are users manuals of the ProtML and others contained in the MOLPHY.
```

**Installation**

To build MOLPHY, UNIX users should be able to type "make" in molphy-2.2/src directory. (Edit the molphy-2.2/src/Makefile if you need to customize it)

```
% cat molphy-2.2.tar.Z | uncompress | tar xvf -
% cd molphy-2.2/src
% make
% make install
```

**Test**

```
% cd ..
% njdist.sh > njdist.out
% diff NJDIST.EXA njdist.out
% protml.sh > protml.out
% diff PROTML.EXA protml.out
% nucml.sh > nucml.out
% diff NUCML.EXA nucml.out
```

## 4.1 ProtML: Maximum Likelihood Inference of Protein Phylogeny

ProtML is a C program for inferring evolutionary trees from protein (amino acid) sequences by using maximum likelihood.

A maximum likelihood method for inferring trees from DNA or RNA sequences was developed by Felsenstein (1981[76]). The method does not impose any constraint on the constancy of evolutionary rate among lineages. He wrote a program DNAML implementing the method, and included it in his program package PHYLIP. The program has been used extensively and has proved of great use in phylogenetic studies (Hasegawa and Yano 1984[126]; Hasegawa et al. 1985[122], 1985[115], 1990[120]; Hasegawa and Kishino 1989[117]; Kishino and Hasegawa, 1989[164]; Zillig et al. 1989[320]; Hasegawa 1990[104], 1991[105], 1994[107]; Golenberg et al. 1990[95]; Adkins and Honeycutt 1991[8]; Doebley et al. 1990[68]; Edwards et al. 1991[72]; Les et al. 1991[191]; Ruvolo et al. 1991[249]; Disotell et al. 1992[67]; Lockhart et al. 1992[196]; Cooper et al. 1992[58]). Computer simulations demonstrated that the method is highly efficient in estimating a true tree under various situations such as a violation of rate constancy among lineages (Hasegawa and Yano, 1984[125]; Hasegawa et al. 1991[121]; Kuhner and Felsenstein 1994[177]).

DNAML, however, is confined to DNA or RNA sequence data, and is not applicable to protein sequence data. In phylogenetic studies of deep branchings, such as those among the three major kingdoms of eukaryotes, archaebacteria and eubacteria, and those in the early evolution of eukaryotes, ribosomal RNA sequence data has been used widely (e.g., Woese 1987[309]; Sogin et al. 1989[272]). In spite of many works on the ribosomal RNAs, the universal root of all contemporary organisms on the earth including

eukaryotes, archaebacteria and eubacteria remained uncertain. Miyata and his coworkers demonstrated the usefulness of using amino acid sequence data encoded by duplicated genes (duplicated prior to the divergence among the major kingdoms) in establishing the universal root (Iwabe et al. 1989[148]; Miyata et al. 1991[218]). Furthermore, an evolutionary tree inferred from ribosomal RNA data is sometimes misleading when base composition differs widely among lineages, and a tree inferred from protein sequences is more reliable in such cases (Loomis and Smith, 1990[198]; Hashimoto et al. 1992[129], 1994[132]; Hasegawa et al. 1992[112], 1993[113]; Hasegawa and Hashimoto, 1993[111]).

Because no program was available for inferring a protein tree by the maximum likelihood based on a reasonable model of amino acid substitutions, many authors used DNAML in analyzing protein-encoding DNA sequences. As is well known, third positions of codons evolve more rapidly than other positions, and therefore DNAML was designed so that a user can specify the relative rates of substitutions among several categories of positions. This approach seems to be satisfactory in many cases when one is interested on phylogenetic relationships among closely related species.

Even if the rate difference among positions in a codon is taken into account, however, inclusion of the third positions in the analysis can sometimes be misleading when the pattern of codon usage differs among lineages. Furthermore, the assumption (in DNAML) of independent evolution among three positions of a codon can be a serious defect when we are interested in tracing deep branchings, because a (negative) selection is likely to be operating at the codon level rather than at the individual sites in the codon. Even if nucleotide frequencies of protein-encoding genes differ among lineages, amino acid frequencies may not differ significantly. Although the nucleotide frequency of the genome can affect the amino acid content of proteins through the genetic code (Sueoka 1961[280]; Jukes and Bhushan 1986[155]; D'Onofrio et al. 1991[69]; Crozier and Crozier 1993[60]), the effect is indirect (Adachi and Hasegawa, 1992[2]), and, for conservative proteins used in phylogenetic studies of deep branchings, the amino acid content is almost uninfluenced by the nucleotide frequency (Hashimoto et al. 1992[129], 1994[132]; Hasegawa et al. 1992[112], 1993[113]; Hasegawa and Hashimoto, 1993[111]).

Therefore, if the amino acid substitution process can be represented by an appropriate model, it seems better to handle amino acid sequences rather than to handle nucleotide sequences in estimating orders of deep branchings from data of a protein-encoding gene, and there is an increasing demand for a maximum likelihood program to infer protein phylogenies.

Kishino et al. (1990[166]) developed a maximum likelihood method for inferring protein phylogenies that takes account of unequal transition probabilities among pairs of amino acids by using an empirical transition matrix compiled by Dayhoff et al. (1978[62]), and the model is called the Dayhoff model (Hasegawa et al. 1992[108]). Although the transition probability matrix was constructed from a limited data set (accumulated up to 1978) of proteins encoded in nuclear DNA, the Dayhoff model is not necessarily specific only to those proteins used in constructing the model, but is appropriate to some extent in approximating the amino acid substitutions of wider protein species including mitochondrial

ones (Adachi and Hasegawa 1992[2]; Adachi et al. 1993[1]; Hasegawa et al. 1993[113]).

The original program used in Kishino et al. (1990[166]), Mukohata et al. (1990[220]), Hasegawa et al. (1990[116]), Iwabe et al. (1991[149]), and Miyata et al. (1991[218]) was written by FORTRAN and the number of species in the maximum likelihood analysis was confined to five. In writing the program "ProtML" for public use, I took advantage of another computer language C in representing the tree structure of the data. In this program, there is no limit of the number of species if the computer is big enough.

Since the number of possible tree topologies increases explosively as the number of species increases (Felsenstein 1978[75]), it is a serious problem to find the best tree among the huge number of alternatives. I developed a novel algorithm for searching tree topologies, called "star decomposition" (section 3.5.2), which seems to be effective in finding the best tree.

The parsimony method has been used widely in molecular phylogenetics, but it may be positively misleading when the evolutionary rate differs among lineages (Felsenstein, 1978[74]). ProtML has proved of great use in inferring evolutionary trees even in such situations (Hasegawa and Fujiwara 1993[110]), and has been applied to several phylogenetic problems (Hasegawa et al. 1992[108], 1993[113]; Adachi and Hasegawa, 1992[2]; Adachi et al. 1993[1]; Hashimoto et al. 1992[129], 1993[133], 1994[132], 1995[130], 1995[131]; Kojima et al. 1993[170]; Yokobori et al. 1994[318]; Shirakura et al. 1994[260]; Cao et al. 1994[50], 1994[49], 1994[48]; Adachi and Hasegawa 1995[6], 1995[4]; Marsh et al. 1994[202]; Klenk and Zillig 1994[168]; Länge et al. 1994[184]; Nikoh et al. 1994[226]; Kuma and Miyata 1994[178]; Kuma et al. 1995[179]; Golding and Gupta 1995[91])

The input format of ProtML is quite similar to that of DNAML. Features where ProtML differs from DNAML (up to version 3.4) are as follows:

1) Amino acid sequence data are analyzed based on several alternative models of amino acid substitutions described in section 2.2.

2) Likelihood of multifurcating trees can be estimated. When the information contained by the data is not sufficient enough to solve branching order, it would be preferable to be satisfied with a tree containing multifurcations (Czelusniak et al. 1990[61]). Publication of completely resolved bifurcating trees by using insufficient amount of data would be misleading. For this reason, it would be important to be able to evaluate the likelihood of a multifurcating tree.

3) A novel method of topology search ("star decomposition") is adoted.

4) Newton-Raphson method is adoted in the maximization of likelihood.

5) Bootstrap probabilities of candidate trees can be estimated.

### 4.1.1 Options

The program allows various options by switches "-$x$" on command line.

```
ProtML 2.3 Maximum Likelihood Inference of Protein Phylogeny
Copyright (C) 1992-1995 J. Adachi & M. Hasegawa. All rights reserved.
Usage: protml [switches] sequence_file [topology_file]
sequence_file = MOLPHY_format | Sequential(-S) | Interleaved(-I)
topology_file = users_trees(-u) | constrained_tree(-e)
Model:
-j  JTT (default)    -jf  JTT-F          Jones, Taylor & Thornton(1992)
-d  Dayhoff          -df  Dayhoff-F      Dayhoff et al.(1978)
-p  Poisson          -pf  Proportional
-r  users RTF        -rf  users RTF-F    (Relative Transition Frequencies)
-f  with data Frequencies
Search strategy or Mode:
-u  Users trees (need users_trees file)
-e  Exhaustive search (with/without constrained_tree file)
-s  Star decomposition search (may not be the ML tree)
-q  Quick add OTUs search (may not be the ML tree)
-D  maximum likelihood Distance matrix --> NJDIST
Others:
-n num  retained top ranking trees win Approx.likelihood(default -e:105,-q:50)
-b  no Bootstrap probabilities (Users trees)
-S  Sequential format    -I  Interleaved format
-v  verbose to stderr    -i, -w  output some infomation
```

This program has five mode of topology search; i.e., User tree (manual) mode, Exhaustive search mode, Star decomposition search mode, Quick add OTUs search mode and maximum likelihood Distance matrix mode.

"**-u**" : User tree mode

User tree (manual) mode is similar with the "U" option in Felsenstein's DNAML. This mode calculates likelihood of all user defined topologies. Different from DNAML, this program allows multifurcating trees as user trees.

"**-e**" : Exhaustive search mode

"**-s**" : Star decomposition mode

Unless specified, it is automatic mode that starts with a star-like tree.

"**-q**" : Quick add OTUs search mode

"**-D**" : maximum likelihood Distance matrix mode

"**-b**" : no Bootstrap option

This option can give the approximate bootstrap probabilities of candidate trees by a resampling of estimated log-likelihood (RELL) method (Kishino et al. 1990[166]; Hasegawa and Kishino, 1994[119]).

### 4.1.2 Format of Input Sequences File

#### MOLPHY Format

The standard MOLPHY input sequence data format:

```
4 90
Data1
MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMKPTLLTATSVFIIAFIHAPPVDK
DGHREPVSGSGRVINTWADIINRANLGMEV
Data2
MTTALRQRESANAWEQFCQWIASTENRLYVGWFGVIMKPTLLTATICFIIAFIHAPPVDK
DGHREPVAGSGRVISTWADILNRANLGFEV
Data3
MTTALQRRESASLWQQFCEWVTSTDNRLYVGWFGVLMKPTLLTATICFIVAFIHAPPVDK
DGHREPVAGSGRVINTWADVLNRANLGMEV
Data4
MTTTLQQRSRASVWDRFCEWITSTENRIYIGWFGVLMKPTLLAATACFVIAFIHAPPVDK
DGHREPVAGSGRVIATWADVINRANLGMEV
```

An input file has two parts of data; SIZE and SEQUENCES.

**SIZE**

The first line of the file contains the number of species(OTUs) and the length of amino acid sequences, in free format, separated by blanks(space or tab). A user can write comment of the data after two digits numbers, separated by blanks.

**SEQUENCES**

The following lines give sets of species name and amino acid sequence data. Names are made up of letters and digits; the first character must be a letter. The underscore "_" is regarded as a letter. Upper case and lower case letters are distinct, so "spc_1", "Spc_1" and "SPC_1" are three different names. Name can NOT include blanks. You must put the amino acid sequence AFTER NEWLINE in free format. Separated by whitespace(space, tab or newline) is allowed. The amino acids must be specified by the one letter codes adopted by Table 1.3 (IUPAC-IUB Commission on Biochemical Nomenclature 1968[147]).

### SEQUENTIAL Format

Felsenstein's PHYLIP "SEQUENTIAL" format:

```
   4   90
Data1      MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFII
AFIAAPPVDIDGIREPVSGSGRVINTWADIINRANLGMEV
Data2      MTTALRQRESANAWEQFCQWIASTENRLYVGWFGVIMIPTLLTATICFII
AFIAAPPVDIDGIREPVAGSGRVISTWADILNRANLGFEV
Data3      MTTALQRRESASLWQQFCEWVTSTDNRLYVGWFGVLMIPTLLTATICFIV
AFIAAPPVDIDGIREPVAGSGRVINTWADVLNRANLGMEV
Data4      MTTTLQQRSRASVWDRFCEWITSTENRIYIGWFGVLMIPTLLAATACFVI
AFIAAPPVDIDGIREPVAGSGRVIATWADVINRANLGMEV
```

The information for each species follows, starting with a TEN-CHARACTER species name (which CAN include punctuation marks and blanks). A user must use SEQUENTIAL FILE with "-S" Switch, follow as:

```
protml -S SEQUENTIAL FILE
```

### COMMON Format

MOLPHY and PHYLIP common format:

```
    4   90$
Data1     $
MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFIIAFIAAPPVDI$
DGIREPVSGSGRVINTWADIINRANLGMEV$
Data2     $
MTTALRQRESANAWEQFCQWIASTENRLYVGWFGVIMIPTLLTATICFIIAFIAAPPVDI$
DGIREPVAGSGRVISTWADILNRANLGFEV$
Data3     $
MTTALQRRESASLWQQFCEWVTSTDNRLYVGWFGVLMIPTLLTATICFIVAFIAAPPVDI$
DGIREPVAGSGRVINTWADVLNRANLGMEV$
Data4     $
MTTTLQQRSRASVWDRFCEWITSTENRIYIGWFGVLMIPTLLAATACFVIAFIAAPPVDI$
DGIREPVAGSGRVIATWADVINRANLGMEV$
```

Note, "$" is newline (return) code.

### INTERLEAVED Format

PHYLIP and other packages "INTERLEAVED" format:

```
    4   90
Data1     MTAILERRESESLWGRFCNWITSTENRLYIGWFGVLMIPTLLTATSVFII
Data2     MTTALRQRESANAWEQFCQWIASTENRLYVGWFGVIMIPTLLTATICFII
Data3     MTTALQRRESASLWQQFCEWVTSTDNRLYVGWFGVLMIPTLLTATICFIV
Data4     MTTTLQQRSRASVWDRFCEWITSTENRIYIGWFGVLMIPTLLAATACFVI

AFIAAPPVDIDGIREPVSGSGRVINTWADIINRANLGMEV
AFIAAPPVDIDGIREPVAGSGRVISTWADILNRANLGFEV
AFIAAPPVDIDGIREPVAGSGRVINTWADVLNRANLGMEV
AFIAAPPVDIDGIREPVAGSGRVIATWADVINRANLGMEV
```

A user must use INTERLEAVED FILE with "-I" Switch, follow as:

```
protml -I INTERLEAVED FILE
```

### Format of USERS TREES File

standard USERS TREES file format:

```
3
(((HUMAN,(CHIMP,PYGMY)),GORIL),ORANG,SIAMA);
((HUMAN,((CHIMP,PYGMY),GORIL)),ORANG,SIAMA);
(((HUMAN,GORIL),(CHIMP,PYGMY)),ORANG,SIAMA);
```

An input file has two parts of data; SIZE and MACHINE READABLE TREES.

**SIZE**

The first line of the file contains the number of machine readable trees. A user can write comment of the trees after one digits number, separated by blanks(space or tab).

### MACHINE READABLE TREES

The following lines give sets of (user-defined) machine readable tree. The tree is specified by the nested pairs of parentheses, enclosing names and separated by commas. Semicolon ";" is tree terminator. The pattern of the parentheses represents the tree topology by having each pair of parentheses which encloses all the members of a monophyletic group. A user must put the next machine readable tree AFTER NEWLINE in free format, being allowed to be separated by whitespace(space, tab or newline). for example,

```
(((HUMAN,(CHIMP,PYGMY)),GORIL),ORANG,SIAMA);

(
    (
        (
            HUMAN,
            (
                CHIMP,
                PYGMY
            )
        ),
        GORIL
    ),
    ORANG,
    SIAMA
);
```

the above two machine readable tree are the same.

Note that the machine readable tree is an UNROOTED one, and therefore its base must be multifurcation with a multiplicity of greater than or equal to three.

```
Unrooted tree (ProtML & DISTNJ)
    variable rate
( subtree1, subtree2, subtree3 );
    :-----subtree1
    :
    :-----subtree2
    :
    :-----subtree3
    ^provisional root
```

```
Rooted tree (not allowed)
    constant rate
( subtree1, subtree2 );
    :-----subtree1
    :
    :-----subtree2
    ^root
```

### Format of CONSTRAINED TREE File

standard CONSTRAINED TREE file format:

```
( { HUMAN,CHIMP,PYGMY,GORIL }, ORANG, SIAMA );
```

CONSTRAINED TREE file allows a constrained machine readable tree. A pair of PARENTHESIS indicates FIX tree structure, but a pair of BRACE indicates COMBINATION tree structure in a monophyletic group.

above CONSTRAINED TREE input ProtML with "-e" switch

```
protml -e sequence_file constrained_tree
```

automatic generation of all possible trees.

```
15
(((HUMAN,(CHIMP,PYGMY)),GORIL),ORANG,SIAMA);
((HUMAN,((CHIMP,PYGMY),GORIL)),ORANG,SIAMA);
(((HUMAN,GORIL),(CHIMP,PYGMY)),ORANG,SIAMA);
((((HUMAN,PYGMY),CHIMP),GORIL),ORANG,SIAMA);
((((HUMAN,CHIMP),PYGMY),GORIL),ORANG,SIAMA);
((HUMAN,(CHIMP,(PYGMY,GORIL))),ORANG,SIAMA);
((HUMAN,((CHIMP,GORIL),PYGMY)),ORANG,SIAMA);
((((HUMAN,GORIL),PYGMY),CHIMP),ORANG,SIAMA);
((((HUMAN,CHIMP),GORIL),PYGMY),ORANG,SIAMA);
(((HUMAN,CHIMP),(PYGMY,GORIL)),ORANG,SIAMA);
((((HUMAN,GORIL),CHIMP),PYGMY),ORANG,SIAMA);
(((HUMAN,(PYGMY,GORIL)),CHIMP),ORANG,SIAMA);
(((HUMAN,(CHIMP,GORIL)),PYGMY),ORANG,SIAMA);
(((HUMAN,PYGMY),(CHIMP,GORIL)),ORANG,SIAMA);
((((HUMAN,PYGMY),GORIL),CHIMP),ORANG,SIAMA);
```

### 4.1.3 Output Format

The output usually consists of (1) the name of the program and its version number, (2) the input information printed out, (3) a series of trees, some with associated information indicating how much change there was in each character or on each part of the tree.

The tree grows from left to right and has branches that are approximately proportional in length to the lengths that the program estimates. In some cases when branches are estimated to be very short it makes them a three spaces long so that the topology is clearly shown. Here is what a typical tree looks like:

```
ProtML  JTT-F  6 OTUs 1344 sites. mt5k
#1
            :-1 Chimp
          :--7
          :  :-2 Bonobo
       :--8
       :  :--3 Human
:------9
:      :---4 Gorilla
:
:-------5 Orang
:
:------6 Siamang

No.1   external branch S.E.  internal branch S.E.
Chimp         1   0.72  0.24        7   0.94  0.29
Bonobo        2   0.91  0.27        8   0.97  0.32
Human         3   1.43  0.35        9   3.14  0.54
Gorilla       4   2.58  0.47        4
Orang         5   6.96  0.77      ln L:   -5510.60 +- 103.75
Siamang       6   4.92  0.65      AIC :   11077.21
#2
            :-1 Chimp
          :--7
          :  :-2 Bonobo
       :--8
       :  :----4 Gorilla
:------9
:      :--3 Human
:
:---------5 Orang
:
:------6 Siamang

No.2   external branch S.E.  internal branch S.E.
Chimp         1   0.73  0.24        7   0.88  0.28
Bonobo        2   0.90  0.26        8   0.38  0.19
Human         3   1.25  0.33        9   3.56  0.57
Gorilla       4   3.26  0.51        7
Orang         5   7.01  0.77      ln L:   -5520.16 +- 104.15
Siamang       6   4.92  0.65      AIC :   11096.32

ProtML "JTT-F" 2 trees 6 OTUs 1344 sites. mt5k

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC  Boot P
-------------------------------------------------------------
1       -5510.6     0.0 <-best   28   11077.2    0.0  0.8190  Base
2       -5520.2    -9.6   11.2   28   11096.3   19.1  0.1810 0.196
```

Length refers to the estimated number of substitutions per 100 amino acid sites along the branch leading to the node (or leaf) indicated by the number, and S.E. refers to its standard error estimated by the formula of Kishino and Hasegawa (1989[164]).

## 4.2 NucML: Maximum Likelihood Inference of Nucleic Acid Phylogeny

NucML is a C program for inferring evolutionary trees from nucleotide sequences by using maximum likelihood.

### 4.2.1 Options

```
NucML 2.3 Maximum Likelihood Inference of Nucleic Acid Phylogeny
Copyright (C) 1992-1995 J. Adachi & M. Hasegawa. All rights reserved.
Usage: nucml [switches] sequence_file [topology_file]
sequence_file = MOLPHY_format | Sequential(-S) | Interleaved(-I)
topology_file = users_trees(-u) | constrained_tree(-e)
Model:
-t n1     n1: Alpha/Beta ratio    (default:4.0)  Hasegawa, Kishino & Yano(1985)
-t n1,n2  n2: AlphaY/AlphaR ratio (default:1.0)  Tamura & Nei(1993)
-p  Proportional     -pf  Poisson
-r  users RTF-F      -rf  users RTF     (Relative Transition Frequencies)
-f  withOUT data Frequencies
Search strategy or Mode:
-u  Users trees (need users_trees file)
-e  Exhaustive search (with/without constrained_tree file)
-s  Star decomposition search (may not be the ML tree)
-q  Quick add OTUs search (may not be the ML tree)
-D  maximum likelihood Distance matrix --> NJDIST
Others:
-n num  retained top ranking trees win Approx.likelihood(default -e:105,-q:50)
-b  no Bootstrap probabilities (Users trees)
-S  Sequential format    -I  Interleaved format
-v  verbose to stderr    -i, -w  output some infomation
```

### 4.2.2 Output Format

```
NucML A/B:opt F  6 OTUs 1344 sites. mt5k
#1
Alpha/Beta: 30.569
                          :----1 Chimp
                  :------7
                  :       :--2 Bonobo
          :------8
          :       :----------3 Human
:---------9
:         :----------4 Gorilla
:
:-----------------------5 Orang
:
:-------------------------------------6 Siamang

No.1   external branch S.E.  internal branch S.E.
Chimp         1   6.47  0.92        7   9.42  1.74
Bonobo        2   4.00  0.82        8   9.40  2.19
Human         3  19.82  2.10        9  16.48  3.94
Gorilla       4  20.73  2.58        9
Orang         5  47.49  5.81      ln L:   -5540.01 +- 79.97
Siamang       6  78.35  8.60      AIC :   11106.02
#2
Alpha/Beta: 29.466
                          :----1 Chimp
                  :------7
                  :       :--2 Bonobo
              :--8
              :  :--------------4 Gorilla
:-------------9
:             :----------3 Human
:
:----------------------5 Orang
:
:-----------------------------------6 Siamang

No.2   external branch S.E.  internal branch S.E.
Chimp         1   6.49  0.92        7  10.23  1.78
Bonobo        2   3.96  0.81        8  lower limit
Human         3  19.86  2.12        9  22.91  4.50
Gorilla       4  28.91  2.71        8
Orang         5  46.66  5.83      ln L:   -5559.28 +- 81.23
Siamang       6  75.50  8.30      AIC :   11144.55

nucml "A/B:opt F" 2 trees 6 OTUs 1344 sites. mt5k

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC  Boot P
-------------------------------------------------------------
1       -5540.0     0.0 <-best   13   11106.0    0.0  0.9980  Base
2       -5559.3   -19.3    8.1   13   11144.6   38.5  0.0020 0.009
```

## 4.3 ProtST: Basic Statistics of Protein Sequences

Bias refers to the distance of amino acid composition between OTUs $i$ and $j$ defined by

$$D_{ij} = \sum_k |f_{ik} - f_{jk}|/2, \tag{4.1}$$

where $f_{ik}$ is the frequency of the $k$-th amino acid of OTU $i$ (Cao et al. 1994[48]).

### 4.3.1 Options

```
ProtST 1.2 Basic Statistics of Protein Sequences
Copyright (C) 1993-1995 J. Adachi & M. Hasegawa. All rights reserved.
Usage: protst [switches] sequence_file
Switches:
-a      Alignments viewer
-c num  column size
-S      Sequential input format (PHYLIP)
-I      Interleaved input format (other packages)
```

### 4.3.2 Output Format

```
protst 1.2  6 OTUs 1344 sites  mt5k

Diff          1   2   3   4   5   6
            Chi Bon Hum Gor Ora Sia
1   Chimp   Chi  22  39  61 141 127
2   Bonobo   22 Bon  43  64 136 123
3   Human    39  43 Hum  61 139 116
4   Gorill   61  64  61 Gor 138 121
5   Orang   141 136 139 138 Ora 142
6   Siaman  127 123 116 121 142 Sia

            A Ala  R Arg  N Asn  D Asp  C Cys  Q Gln  E Glu  G Gly  H His  I Ile
1   Chimp   0.065  0.019  0.040  0.020  0.003  0.026  0.022  0.057  0.025  0.085
2   Bonobo  0.062  0.018  0.042  0.020  0.004  0.026  0.022  0.057  0.025  0.083
3   Human   0.065  0.019  0.042  0.020  0.003  0.025  0.022  0.057  0.025  0.086
4   Gorill  0.068  0.018  0.042  0.021  0.004  0.025  0.022  0.057  0.025  0.086
5   Orang   0.070  0.019  0.039  0.022  0.003  0.025  0.022  0.057  0.028  0.092
6   Siaman  0.068  0.019  0.042  0.020  0.002  0.026  0.022  0.057  0.025  0.089
    mean    0.067  0.018  0.041  0.020  0.003  0.026  0.022  0.057  0.025  0.087

            L Leu  K Lys  M Met  F Phe  P Pro  S Ser  T Thr  W Trp  Y Tyr  V Val
1   Chimp   0.152  0.028  0.062  0.055  0.068  0.065  0.094  0.029  0.034  0.050
2   Bonobo  0.150  0.028  0.062  0.057  0.068  0.064  0.098  0.029  0.034  0.051
3   Human   0.153  0.029  0.062  0.055  0.069  0.061  0.095  0.029  0.035  0.048
4   Gorill  0.154  0.028  0.059  0.055  0.067  0.062  0.096  0.030  0.035  0.047
5   Orang   0.154  0.028  0.048  0.058  0.070  0.062  0.096  0.029  0.033  0.046
6   Siaman  0.154  0.027  0.053  0.056  0.068  0.060  0.097  0.029  0.035  0.050
    mean    0.153  0.028  0.058  0.056  0.069  0.062  0.096  0.029  0.034  0.048

Bias x10e3    1   2   3   4   5   6
            Chi Bon Hum Gor Ora Sia
1   Chimp   Chi   8   8  13  26  18
2   Bonobo    8 Bon  13  15  29  19
3   Human     8  13 Hum   9  23  15
4   Gorill   13  15   9 Gor  19  13
5   Orang    26  29  23  19 Ora  18
6   Siaman   18  19  15  13  18 Sia
```

## 4.4 NucST: Basic Statistics of Nucleic Acid Sequences

Bias is defined by Eq. 4.1 where $f_{ik}$ is the frequency of the $k$-th nucleotide of OTU $i$.

### 4.4.1 Options

```
NucST 1.2 Basic Statistics of Nucleic Acid Sequences
Copyright (C) 1993-1995 J. Adachi & M. Hasegawa. All rights reserved.
Usage: nucst [switches] sequence_file
Switches:
-a      Alignments viewer
-c num  column size
-S      Sequential input format (PHYLIP)
-I      Interleaved input format (other packages)
```

### 4.4.2 Output Format

```
nucst 1.2  6 OTUs 1344 sites  mt5k3

    Ts       1    2    3    4    5    6
Tv          Chi  Bon  Hum  Gor  Ora  Sia
1   Chimp   Chi  114  292  312  356  382
2   Bonob     9  Bon  286  293  363  366
3   Human    15   16  Hum  331  356  398
4   Goril    46   47   45  Gor  365  391
5   Orang    93   92   90   95  Ora  361
6   Siama   121  118  122  129  138  Sia

              T      C      A      G    A+T    G+C    Bias   Skew
1   Chimp  0.184  0.393  0.377  0.046  0.561  0.439  0.110  0.540
2   Bonob  0.190  0.389  0.378  0.043  0.568  0.432  0.110  0.534
3   Human  0.167  0.410  0.365  0.057  0.533  0.467  0.110  0.551
4   Goril  0.193  0.388  0.365  0.054  0.559  0.441  0.099  0.506
5   Orang  0.152  0.432  0.365  0.051  0.517  0.483  0.127  0.594
6   Siama  0.189  0.388  0.376  0.046  0.565  0.435  0.107  0.530
    mean   0.179  0.400  0.371  0.050  0.550  0.450  0.110  0.542

Bias x1e3    1    2    3    4    5    6
            Chi  Bon  Hum  Gor  Ora  Sia
1   Chimp   Chi    7   28   17   44    5
2   Bonob     7  Bon   35   14   51    3
3   Human    28   35  Hum   26   22   33
4   Goril    17   14   26  Gor   44   12
5   Orang    44   51   22   44  Ora   48
6   Siama     5    3   33   12   48  Sia
```

## 4.5 NJdist: Neighbor Joining Phylogeny from Distance Matrix

### 4.5.1 Options

```
NJDist 1.3 Neighbor Joining Phylogeny from Distance Matrix
Copyright (C) 1993-1995 J. Adachi & M. Hasegawa. All rights reserved.
Ref: N. Saitou & M. Nei 1987. Molecular Biology and Evolution 4:406-425
Usage: njdist [switches] distance_matrix_file
Switches:
-w      branch length
-l      Least squares
-S      Sequential input format (PHYLIP)
-O num  branch number of Out group
-T str  output Tree file name
```

### 4.5.2 Input Format

```
6 1344 sites JTT-F mt5k
Chimp
 0.000000000000 0.016309763506 0.029127330244 0.046248695626 0.111674086959
 0.099339573872
Bonobo
 0.016309763506 0.000000000000 0.032187054742 0.048634269105 0.107657113491
 0.096145625286
Human
 0.029127330244 0.032187054742 0.000000000000 0.046322178390 0.110634307362
 0.090756861511
Gorilla
 0.046248695626 0.048634269105 0.046322178390 0.000000000000 0.109596357665
 0.095265576246
Orang
 0.111674086959 0.107657113491 0.110634307362 0.109596357665 0.000000000000
 0.113685178041
Siamang
 0.099339573872 0.096145625286 0.090756861511 0.095265576246 0.113685178041
 0.000000000000
```

### 4.5.3 Output Format

```
njdist 1.3  6 OTUs 1344 sites JTT-F mt5k

           :-1 Chimp
        :--8
        :  :-2 Bonobo
     :--9
     :  :--3 Human
:----7
:    :---4 Gorilla
:
:-------5 Orang
:
:-----6 Siamang

((((Chimp,Bonobo),Human),Gorilla),Orang,Siamang);
```

## 4.6 Utilities (Sequence Manipulations) in Perl

```
mollist:  get identifiers list
molrev:   reverse DNA sequences
molcat:   concatenate sequences
molcut:   get partial sequences
molmerge: merge sequences
nuc2ptn:  DNA -> Amino acid
rminsdel: remove INS/DEL sites
molcodon: get 3rd(1st,2nd) codons
molinfo:  get (non)infomation sites
mol2mol:  MOLPHY format beautifer
inl2mol:  Interleaved -> MOLPHY
mol2inl:  MOLPHY -> Interleaved
mol2phy:  MOLPHY -> Sequential
phy2mol:  Sequential -> MOLPHY
must2mol: MUST -> MOLPHY
```

# Chapter 5

# Applications

## 5.1 Improved Dating of the Human-Chimpanzee Separation in the Mitochondrial DNA Tree: Heterogeneity Among Amino Acid Sites

The internal branch lengths estimated by the distance methods such as neighbor-joining are shown to be biased to be short when the evolutionary rate differs among sites. The variable-invariable model for the site-heterogeneity fits the amino acid sequence data encoded by the mtDNA from Hominoidea remarkably well. By assuming the orangutan separation to be 13 or 16 Myr old, a maximum likelihood analysis estimates a young date of 3.6 ± 0.6 or 4.4 ± 0.7 Myr (±: 1SE) for the human/chimpanzee separation, and these estimates turn out to be robust against differences in the assumed model for amino acid substitutions. Although some uncertainties still exist in our estimates, this analysis suggests that humans separated from chimpanzees some 4–5 Myr ago. The content of this section appears in Adachi and Hasegawa (1995[4]).

### 5.1.1 Problems Inherent in the Previous Estimates of Branching Dates

Although molecular phylogenetics has established that the human/chimpanzee separation is younger than 10 Myr, there is still a wide range of variation in the estimate among the researchers depending on the data and the method they use (Sarich and Wilson 1967[254]; Andrews and Cronin 1982[18]; Sibley and Ahlquist 1984[261], 1987[263]; Hasegawa et al. 1985[122], 1987[123], 1990[120]; Ueda et al. 1989[298]; Kishino and Hasegawa 1990[165]; Gonzalez et al. 1990[97]; Hasegawa 1991[105]; Bailey et al. 1992[31]).

Horai et al. (1992[141]) determined 4.8kbp of mtDNA sequences from common chimpanzee (*Pan troglodytes*), pygmy chimpanzee (bonobo; *Pan paniscus*), gorilla (*Gorilla gorilla*), orangutan (*Pongo pygmaeus*), and siamang (*Hylobates syndactylus*). The sequences cover genes coding for ND2, COI, COII, ATPase 8 and 11 tRNAs, and partially cover genes for ND1 and ATPase 6. Since mtDNA evolves much more rapidly than nuclear DNA (Brown et al. 1982[43]), this data together with the corresponding sequences of human (*Homo sapiens*) (Anderson et al. 1981[15]) should contain more information than the nuclear DNA data published up to date to elucidate the phylogenetic place of humans within Hominoidea.

From these sequences, they established that the closest relatives of the human are the two chimpanzees rather than the gorilla in accord with the preceding works (Sibley and Ahlquist 1984[261], 1987[263]; Hasegawa and Yano 1984[126]; Miyamoto et al. 1987[216]; Kishino and Hasegawa 1989[164]; Caccone and Powell 1989[44]; Sibley et al. 1990[265]; Ruvolo et al. 1991[249]; Hasegawa 1992[106]). By assuming the orangutan separation to be 13 Myr ago, they further estimated the dates of branchings within the African apes/human clade. From the data set that consists of the tRNAs, and first and second codon positions (their DATA1), they estimated the human/chimpanzee separation to be 4.3 and 5.6 Myr ago, respectively, by the maximum likelihood (ML) method for DNA phylogeny (Felsenstein 1981[76]; the DNAML program in Felsenstein's package PHYLIP) and the neighbor-joining (NJ) method (Saitou and Nei 1987[253]). They noted that the ML method gave a shorter divergence times than the NJ, and they attributed the difference to the problematic synonymous changes in *Leu* codons. It is likely that synonymous changes at the first positions of *Leu* codons have substantial effects on the estimation as they thought. But this must be a problem not only in the ML but also in the NJ method, and hence this does not explain the difference of the estimates between the two methods.

We think that the difference of the estimates is due to a defect of the NJ in estimating branch lengths as will be shown later. Because of the problem of *Leu* codons, Horai et al. excluded synonymous transition in the first codon positions. They included synonymous transversions in the third codon positions. They applied the NJ method to this data set (their DATA3; the DNAML program cannot be applied to such a data set). Their estimate of the human/chimpanzee separation was 4.7 $\pm$ 0.5 Myr ($\pm$: 1SE). They attributed a younger estimate of 3.9 $\pm$ 0.7 Myr for this separation by Hasegawa et al. (1990[120]) to the relatively small region compared (896bp of Brown et al. 1982[43]).

Hasegawa et al.'s (1990[120]) estimate was done by classifying sites into two classes; third codon positions and the remainder, and suffers from the problem of the synonymous changes at the first positions of *Leu* codons. Therefore, we admit that Hasegawa et al.'s estimate should be reexamined by an improved method with more abundant data. This does not necessarily mean that Horai et al.'s (1992[141]) estimate is the most reliable from their data.

In Horai et al.'s data set DATA3, they included nonsynonymous differences and synonymous transversion differences in protein-encoding genes, and all differences in tRNA genes. They considered that the differences between species under consideration were small enough to be far from the saturation level, and hence they did not take account of multiple substitutions in a site in their NJ analysis. Since the number of differences between even the most distant pair is only a small fraction of the total number of sites, the multiple-hit correction should be negligibly small by conventional formulas such as of Jukes and Cantor (1969[156]) and of Kimura (1980[161]), and therefore their procedure might seem to be justified at a first glance.

Actually, however, variability differs among sites (even among nonsynonymous sites), and all the sites under consideration are not equally variable (Fitch and Markowitz 1970[85]; Uzzell and Corbin

1971[300]; Hasegawa et al. 1985[122], 1993[109]; Hasegawa and Horai 1991[114]; Kocher and Wilson 1991[169]; Reeves 1992[244]; Sidow et al. 1992[267]; Yang 1993[316]). Although the human/chimpanzee clade has been firmly established for the 4.8kbp data of Horai et al., there are still many sites in DATA3 that support other branchings by the parsimony principle, indicating multiple substitutions in these sites. Such a multiple-hit effect has not been taken into account in their NJ analysis, while it can be taken into account to some extent by the ML analysis as will be shown later. Since the multiple-hit effect is more serious in a longer branch than in a shorter one, their dating of the human/chimpanzee branching could be biased to be older. We attribute this effect to the cause of the difference of the estimates between the NJ and ML methods for DATA1.

Since a more realistic model is available for amino acid substitutions than for nucleotide substitutions in protein-encoding genes (Kishino et al. 1990[166]; Adachi and Hasegawa 1992[3]), I reexamined their data by the ML method at the amino acid sequence level taking account of the heterogeneity of rate among amino acid sites.

**Comparison Between the ML and NJ Methods in Estimating Branch Lengths**

All phylogenetic inferences depend on their underlying models. To have confidence in inferences, it is necessary to have confidence in the models (Goldman 1993[93]). Adachi and Hasegawa (1992[3]) published the ProtML program for the ML inference of protein phylogeny based on the Dayhoff model (Dayhoff et al. 1978[62]), and it has been used widely. Consequently, it turned out that this model is by far more appropriate than the Proportional and Poisson models (Hasegawa et al. 1992[108]) in approximating the evolution of the diverse protein data (Hasegawa et al. 1993[113]; Adachi et al. 1993[1]; Hashimoto et al. 1993[133], 1994[132]). Recently, Jones, Taylor and Thornton (1992[154]) updated the amino acid substitution matrix by using about 40 times more abundant substitution data than those of Dayhoff et al. The new version of ProtML (version 2.2) allows us to use this model (called the JTT model) as well as the Dayhoff, Proportional and Poisson models, and it turned out that the JTT model better approximates the evolution of diverse proteins than the Dayhoff model, except for globins (Cao et al. 1994[50]).

Both the Dayhoff and the JTT models assume the averaged amino acid frequencies of the proteins that were used in estimating the respective substitution matrices as the equilibrium frequencies. However, the amino acid frequencies of the individual protein species under analysis generally differ from those of the average one, and hence it might be better to use the actual amino acid frequencies of the protein under analysis as the equilibrim frequencies. The new version of ProtML (version 2.2) allows us to use this option for the JTT, Dayhoff and Poisson models (the 'F' option; the Proportional model corresponds to the F option of the Poisson model). When it was applied to mtDNA-encoded proteins of tetrapods, it turned out that, among the alternative models, the JTT-F model best approximates the evolution of all the 13 proteins encoded by mtDNA (Cao et al. 1994[49]).

In this work, the JTT-F, JTT, and Poisson models were used. For the purpose of comparison with the ML, the NJ method was also used. The distances estimated by the ProtML for 2-species trees based on the respective models were used in the NJ analysis.

Following protein-encoding regions in Anderson et al. (1981[15]) and Horai et al. (1992[141]) were used in this work: ND1 (4123–4260 in the numbering of Anderson et al.), ND2 (4470–5510), COI (5904–7442), COII (7586–8266), ATPase 8 (8366–8524), ATPase 6 (8575–9024, overlapping region with ATPase8, 8525–8574, was excluded). The total number of deduced codons is 1344.

Fig. 5.1 shows the ML tree estimated from the JTT-F model assuming homogeneity across sites. The left hand side of Table 5.1 gives the branch lengths estimated by the NJ and ML methods based on the JTT-F, JTT, and Poisson models that assume the site-homogeneity. It is apparent that, although the terminal branch lengths do not differ systematically between the NJ and ML methods, the internal branch lengths estimated by the NJ are consistently shorter than those by the ML. This is particularly true for the two most internal branches 4–3 and 3–2, for which the ratios of NJ to ML estimates are nearly 0.7–0.8. This discrepancy between the two methods can be attributed to the fact that the multiple-substitutions are underestimated in the NJ method because it does not take account of the states of the internal nodes.

Table 5.2 gives numbers of differences in the 1344 amino acid sites. The difference between siamang and orangutan is significantly larger than those between siamang and the members of the African apes/human clade. Furthermore, the differences between orangutan and the African apes/human are even larger than those between siamang and the African apes/human. Since the siamang is highly likely to be the outgroup to all the other species used in this analysis (Hayasaka et al. 1988[134]; Hasegawa et al. 1990[120]), these indicate that the evolutionary rate in the orangutan lineage accelerated relative to the African apes and human lineages as suggested by Horai et al. (1992[141]). Except for this violation of the molecular clock, the relative rate tests (Sarich and Wilson 1967[255]; Hasegawa et al. 1987[123]) at the amino acid level do not suggest any rate variation among chimpanzee, bonobo, human, and gorilla, allowing molecular clock analyses of these data.

From the estimates of branch legths, we estimated branching dates by the following procedure similar to Horai et al.'s. A depth of a node (numbered through 1—4 as in Fig. 5.1) from tips was estimated as follows from branch lengths represented as $l_{XY}$ between $X$ and $Y$ (either nodes or tips):

$$d_1 = (l_{1C} + l_{1B})/2, \tag{5.1}$$

$$d_2 = (l_{2H} + l_{21} + d_1)/2, \tag{5.2}$$

$$d_3 = (l_{3G} + l_{32} + d_2)/2, \tag{5.3}$$

$$d_4 = l_{43} + d_3. \tag{5.4}$$

Since the rate in the orangutan lineage is higher than in other lineages, $l_{4O}$ was not used in estimating $d_4$. By assuming 13 Myr for the node-4 (Pilbeam 1988[241]; Andrews 1992[17]; McCrossin and Benefit

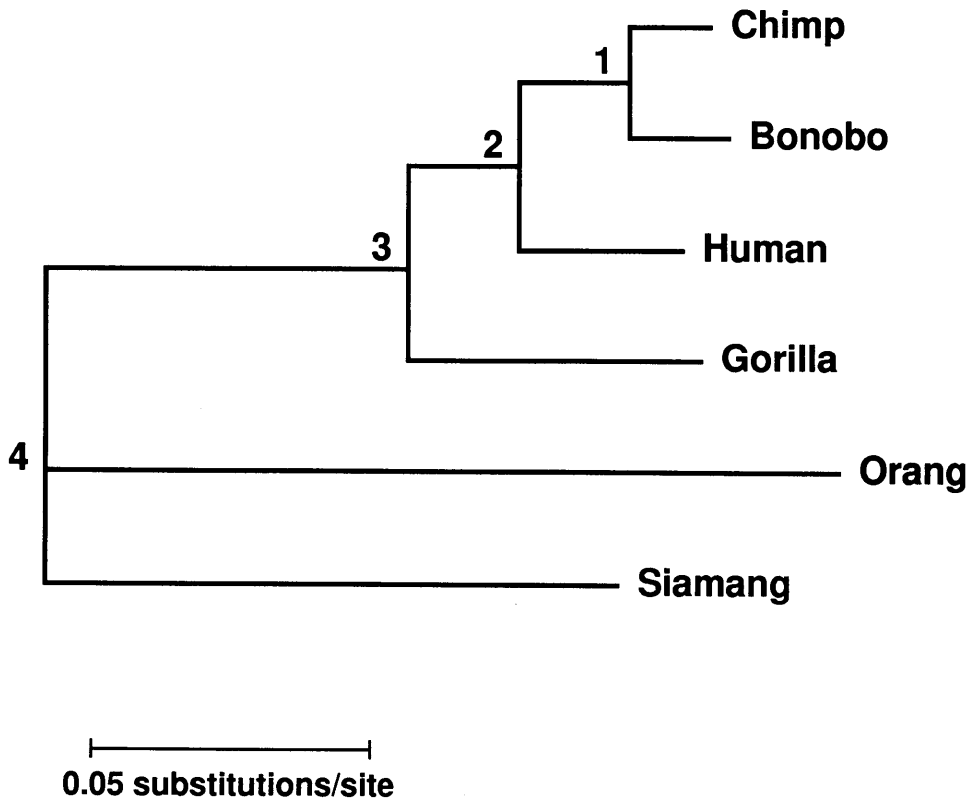



Figure 5.1: The ML tree of the mtDNA.

The ML tree of the mtDNA encoded proteins based on the JTT-F model. The horizontal length of each branch is proportional to the estimated number of substitutions. The root of this tree is located somewhere within the 4–siamang branch. Among several models implemented in the ProtML program (version 2.2), which assume homogeneity among sites, the JTT-F model best approximates the data.

Table 5.1: Branch lengths (numbers of substitutions per 100 amino acids) and branching dates estimated from the amino acid sequences of mtDNA encoded proteins by the NJ and ML methods.

| | Homogeneous | | | | | | | Heterogeneous | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | JTT-F | | JTT | | Poisson | | | Poisson | | |
| | NJ | ML | NJ | ML | NJ | ML | $\frac{NJ}{ML}$ | NJ | ML | $\frac{NJ}{ML}$ |
| Terminal branch | | | | | | | | | | |
| $l_{1C}$ (1–C) | 0.78 | 0.72 | 0.79 | 0.72 | 0.79 | 0.71 | 1.12 | 0.83 | 0.72 ± 0.24 (0.24) | 1.15 |
| $l_{1B}$ (1–B) | 0.85 | 0.91 | 0.86 | 0.93 | 0.85 | 0.93 | 0.91 | 0.85 | 0.94 ± 0.28 (0.28) | 0.91 |
| $l_{2H}$ (2–H) | 1.43 | 1.43 | 1.45 | 1.51 | 1.43 | 1.38 | 1.03 | 1.46 | 1.41 ± 0.37 (0.36) | 1.04 |
| $l_{3G}$ (3–G) | 2.36 | 2.58 | 2.37 | 2.57 | 2.34 | 2.38 | 0.98 | 2.50 | 2.48 ± 0.48 (0.49) | 1.01 |
| $l_{4O}$ (4–O) | 6.41 | 6.96 | 6.40 | 6.91 | 6.26 | 6.82 | 0.92 | 7.70 | 7.75 ± 0.88 (0.87) | 0.99 |
| $l_{4S}$ (4–S) | 4.96 | 4.92 | 4.97 | 5.00 | 4.88 | 4.89 | 1.00 | 5.74 | 5.35 ± 0.73 (0.72) | 1.07 |
| Internal branch | | | | | | | | | | |
| $l_{21}$ (2–1) | 0.82 | 0.94 | 0.83 | 0.93 | 0.84 | 1.00 | 0.84 | 0.92 | 1.02 ± 0.32 (0.31) | 0.91 |
| $l_{32}$ (3–2) | 0.80 | 0.97 | 0.81 | 0.97 | 0.81 | 1.14 | 0.71 | 0.91 | 1.15 ± 0.37 (0.36) | 0.79 |
| $l_{43}$ (4–3) | 2.20 | 3.14 | 2.22 | 3.16 | 2.20 | 3.06 | 0.72 | 2.73 | 3.23 ± 0.59 (0.59) | 0.85 |
| Branching date (Myr) | | | | | | | | | | |
| $t_1$ (C/B) | 2.33 | 1.85 | 2.34 | 1.86 | 2.35 | 1.90 | 1.24 | 2.08 | 1.84 ± 0.43 (0.46) | 1.13 |
| $t_2$ (H/C) | **4.38** | **3.63** | **4.40** | **3.69** | **4.41** | **3.70** | 1.19 | **4.00** | **3.60 ± 0.58** (0.70) | 1.11 |
| $t_3$ (H/G) | 6.70 | 5.86 | 6.71 | 5.85 | 6.70 | 5.92 | 1.13 | 6.22 | 5.83 ± 0.72 (0.99) | 1.07 |
| $t_4$ (H/O) | 13 | 13 | 13 | 13 | 13 | 13 | | 13 | 13 | |
| $\ln L$ | −5510.6 | | −5741.7 | | −6144.9 | | | −5747.7 | | |
| df | 28 | | 9 | | 9 | | | 10 | | |
| AIC | 11077.2 | | 11501.4 | | 12307.8 | | | 11515.4 | | |
| ΔAIC | 0 | | 424.2 | | 1230.6 | | | 438.2 | | |

The homogeneous model assumes that all 1344 amino acid sites are equally variable. The heterogeneous model assumes that some portion of the sites are invariable and the remainings are equally variable. Branch lengths are represented as the averages of all sites irrespective of variable or invariable. ± refers to 1SE estimated by replicating bootstrap resampling (Felsenstein 1985[80]) (1000 replications). The SEs estimated from the curvature of likelihood surface (given by ProtML) are shown in parentheses. Log-likelihood for the heterogeneous Poisson model is given by $\log L = \log L_{var} - (\text{number of invariable sites}) \times \log 20$, where $\log L_{var}$ is the total log-likelihood for the variable sites. df refers to a degree of freedom of the model for the ML method. For the JTT and Poisson models, 9 branch lengths are estimated, for the JTT-F model, 20 amino acid frequencies are additionally estimated under the constraint that the summation is 1 (additional 19 df), and for the heterogeneous model, the fraction of variable sites is estimated (additional 1 df).

Table 5.2: Numbers of amino acid differences in the 1344 sites of mtDNA-encoded proteins from Hominoidea.

| | Orang | Gorilla | Human | Chimp | Bonobo |
|---|---|---|---|---|---|
| Siamang | 142 | 121 | 116 | 127 | 123 |
| Orang | | 138 | 139 | 141 | 136 |
| Gorilla | | | 61 | 61 | 64 |
| Human | | | | 39 | 43 |
| Chimp | | | | | 22 |

1993[207]), dates of the other nodes are estimated by

$$t_i = (d_i/d_4) \times 13 \quad (i = 1, 2, \text{and } 3). \tag{5.5}$$

The human/chimpanzee separation is estimated to be 4.4 and 3.7 (or 3.6 for the JTT-F model) Myr old, respectively, from the NJ and ML methods, when rate homogeneity among sites is assumed. The older estimate by NJ than that by ML is due to the underestimate of the internal branch lengths by NJ. The JTT-F model has turned out to be the best among the alternative models in approximating the data, but the estimated branch lengths and the divergence dates are almost the same among different models as far as the site-homogeneity is assumed, and hence we shall examine the Poisson model further in detail because of its simplicity.

**Heterogeneity Among Sites in the Evolution of Amino Acid Sequences**

The left hand side of Table 5.3 shows a comparison of the observed distribution of configurations of amino acid sites with that expected from the homogeneous Poisson model. The fitting of the model to the data is terribly bad ($\chi^2 = 116.27$ with 10 df) as was pointed out by Reeves (1992[244]) for the mtDNA-encoded proteins of tetrapods. This may be attributed to the fact that not all sites are equally variable, and that some of the sites are invariable due to functional constraints. Therefore, we assume that some portion of the sites are invariable, and that the remaining sites are equally variable (Hasegawa et al. 1985[122]; Hasegawa and Horai 1991[114]). When this heterogeneous Poisson model is applied, the fraction of variable sites turns out to be $372/1344 = 0.277$, and the fitting to the data improves drastically ($\chi^2 = 3.59$ with 9 df) (Table 5.3). Consequently, the AIC of the heterogeneous Poisson model improves over that of the homogeneous Poisson model (Table 5.1). The estimates of branching dates by ML remain almost unchanged by this improvement of the model, while those estimated by NJ become nearer to those by the ML (Table 5.1).

A combination of the heterogeneous model with the JTT-F model should further improve the fit to the data, but we did not take this approach, because of the ambiguity in removing sites with this model. The variable-invariable classification is only an approximation, and the rate variation among sites must be more continuous (Kocher and Wilson 1991[169]; Yang 1993[316]; Tamura and Nei 1993[288]). Nevertheless, it is clear that the ML estimates of the branching dates would remain almost unchanged by these further improvements of the model. It is noteworthy in Table 5.1 that branch lengths estimated by ML are affected only slightly by taking account of the site heterogeneity. Those estimated by NJ are affected more greatly, particularly for the deepest internal branch 4–3. This indicates that, while the multiple-hit effect is taken into account automatically to some extent in ML even under the homogeneity assumption because the method takes account of the states of the internal nodes, it is underestimated by distance methods such as the NJ.

Table 5.3: Distribution of configurations of amino acid sites.

| Configuration | number of changes | homogeneous model Obs | Exp | $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ | heterogeneous model Obs | Exp | $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ |
|---|---|---|---|---|---|---|---|
| (C,B,H,G,O,S) | 0 | 1128 | 1074.4 | 2.67 | 156 | 156.0 | — |
| (C,B,H,G,S)(O) | 1 | 53 | 76.1 | 7.01 | 53 | 50.5 | 0.12 |
| (C,B,H,G,O)(S) | 1 | 39 | 54.2 | 4.26 | 39 | 33.6 | 0.86 |
| (C,B,H,G)(O,S) | 1 | 20 | 33.7 | 5.57 | 20 | 20.0 | 0.00 |
| (C,B,H,O,S)(G) | 1 | 11 | 26.0 | 8.65 | 11 | 14.7 | 0.94 |
| (C,B,G,O,S)(H) | 1 | 5 | 15.0 | 6.67 | 5 | 8.2 | 1.24 |
| (C,B,H)(G,O,S) | 1 | 7 | 12.4 | 2.35 | 7 | 6.8 | 0.01 |
| (C,B)(H,G,O,S) | 1 | 5 | 10.9 | 3.19 | 5 | 5.9 | 0.13 |
| (C,H,G,O,S)(B) | 1 | 6 | 10.1 | 1.66 | 6 | 5.4 | 0.06 |
| (B,H,G,O,S)(C) | 1 | 5 | 7.7 | 0.95 | 5 | 4.1 | 0.19 |
| others | $\geq 2$ | 65 | 23.5 | 73.29 | 65 | 66.7 | 0.04 |
| total | | 1344 | 1344.0 | $\chi^2 = 116.27$ | 372 | 372.0 | $\chi^2 = 3.59$ |
| | | | | df = 10 | | | df = 9 |
| | | | | $P \ll 0.00001$ | | | P = 0.94 |

Distribution of configurations of amino acid sites for the homogeneous and heterogeneous Poisson models (ML estimates). C, B, H, G, O, and S refer to common chimpanzee, bonobo, human, gorilla, orangutan, and siamang. In the specification of a configuration of a site, the amino acids of the species within common parentheses are the same, while those in different parentheses are different. For the heterogeneous model, 0-change sites were deleted one by one until the expected number of the 0-change sites coincides with the observed number for the remainder that were assumed to evolve homogeneously across sites. When 972 sites were deleted from the 1128 sites of 0-change, the coincidence was attained.

For the ML analysis of the heterogeneous Poisson model, SEs of branch lengths and branching dates were estimated by replicating bootstrap resampling (Felsenstein 1985[80]) and from the curvature of likelihood surface (given by ProtML) as well (Table 5.1). The SEs of each branch length are nearly identical between the two methods of estimation, suggesting that the SEs estimated in the ProtML are good approximations. However, since the covariances between different branches are neglected in the estimation from the curvature (ProtML does not estimate covariances), the SEs of the branching dates turned out to be over-estimated.

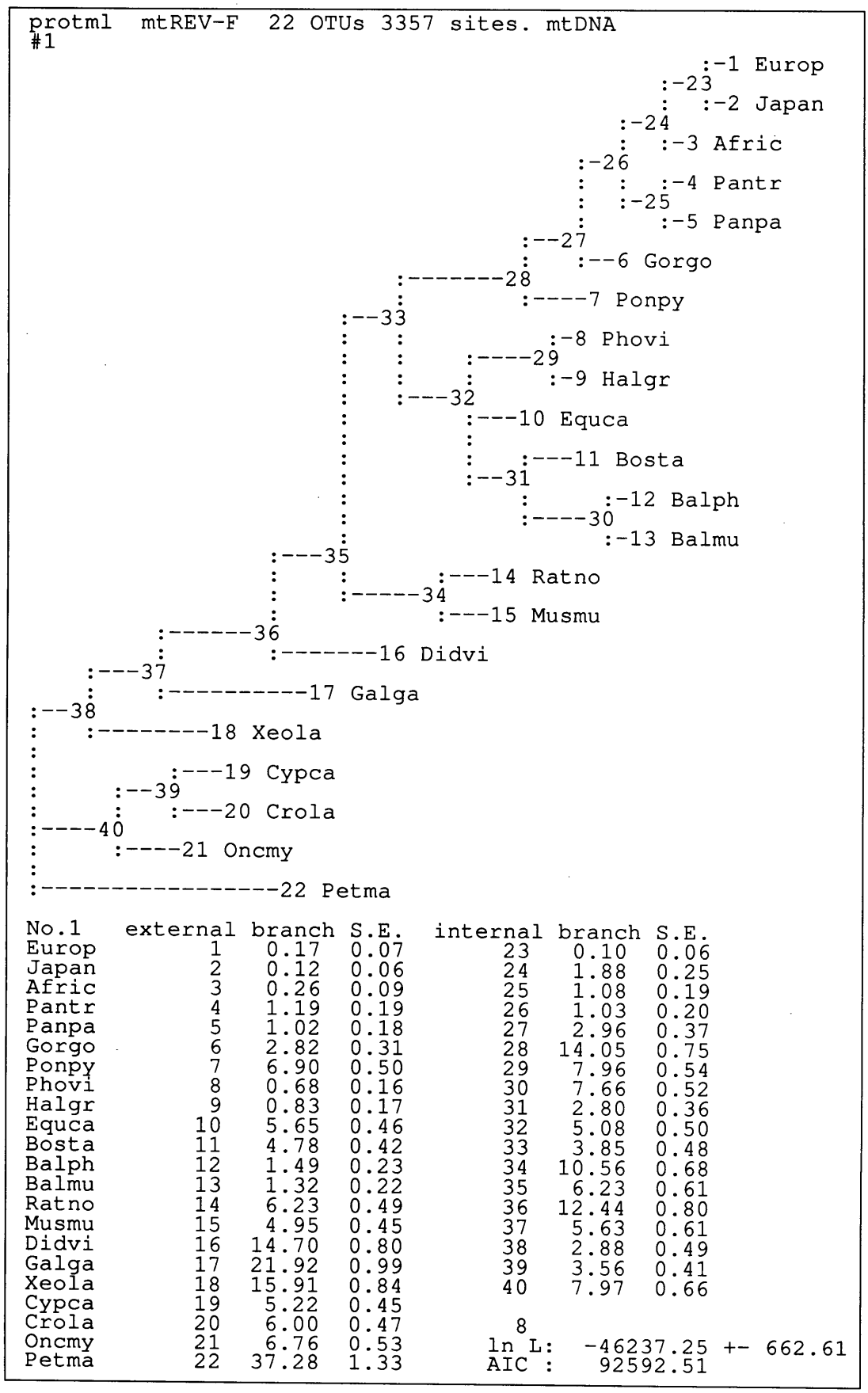
From the ML analysis of the heterogeneous Poisson model, we estimate 1.84 ± 0.43 Myr for the chimpanzee/bonobo separation, 3.60 ± 0.58 Myr for the human/chimpanzee, and 5.83 ± 0.72 Myr for the human/gorilla (Table 5.1). The latter two estimates are in accord with the previous estimates of 3.9 ± 0.7 and 5.1 ± 0.8 Myr from shorter mtDNA sequences (Hasegawa et al. 1990[120]). The remarkable fit of the heterogeneous model to the data and the robustness of the ML estimates of branching dates to changes in model assumptions raises the possibility that the human/chimpanzee separation was more recent than has been generally thought even by molecular evolutionists. However, there are two factors that may cause our estimate to be too young. First, we assumed the orangutan separation to be 13 Myr old. If it was 16 Myr, which is probably the oldest limit (Pilbeam 1988[241]; Andrews 1992[17]; McCrossin and Benefit 1993[207]), the human/chimpanzee separation is estimated to be 4.43 ± 0.71 Myr old. Second, there may have been variation of the evolutionary rate which cannot be detected by the relative rate test. If the rate along the 4–3 branch was as high as that along the orangutan (4–O) branch, the human/chimpanzee separation is estimated to be 4.70 ± 0.99 Myr old. These possibilities cannot be excluded, and therefore some uncertainties exist in our estimates.

### 5.1.2 Date of the Deepest Root of the Human MtDNA Tree

Fig 2.2 is reproduced in Fig. 5.2 with the ML estimates of branch lengths (number of amino acid substitution per 100 sites).

Table 5.4 gives branch lengths estimated from the amino acid sequences (3357 sites; the mtREV model of Table 2.12), from the four-fold degenerate sites (1667 sites; the REV model of Table 5.7), and from total of the third codon positions (3569 sites; the REV model of Table 5.8) in the tree of Hominoidea. The branch lengths leading to the African (SB17F) are estimated to be 0.26 ± 0.09, 0.60 ± 0.19 and 0.52 ± 0.14, respectively, from the amino acids, the four-fold degenerate sites and the total third codon positions. And those leading to chimpanzees (average of *Pan troglodytes* and *P. paniscus*) are estimated to be 1.11 ± 0.18, 5.67 ± 0.83 and 5.60 ± 0.57. The ratios of (24–African)/(25–chimpanzees) are estimated to be 0.11 ± 0.04 and 0.09 ± 0.03, respectively, from the four-fold degenerate sites and total of the third codon positions.

Provided that the two chimpanzee species diverged some 2 Myr ago, these ratios for the four-fold degenerate sites and the total third positions suggest some 200,000 years ago for the time of the deepest




| No.1 | external | branch | S.E. | internal | branch | S.E. |
|---|---|---|---|---|---|---|
| Europ | 1 | 0.17 | 0.07 | 23 | 0.10 | 0.06 |
| Japan | 2 | 0.12 | 0.06 | 24 | 1.88 | 0.25 |
| Afric | 3 | 0.26 | 0.09 | 25 | 1.08 | 0.19 |
| Pantr | 4 | 1.19 | 0.19 | 26 | 1.03 | 0.20 |
| Panpa | 5 | 1.02 | 0.18 | 27 | 2.96 | 0.37 |
| Gorgo | 6 | 2.82 | 0.31 | 28 | 14.05 | 0.75 |
| Ponpy | 7 | 6.90 | 0.50 | 29 | 7.96 | 0.54 |
| Phovi | 8 | 0.68 | 0.16 | 30 | 7.66 | 0.52 |
| Halgr | 9 | 0.83 | 0.17 | 31 | 2.80 | 0.36 |
| Equca | 10 | 5.65 | 0.46 | 32 | 5.08 | 0.50 |
| Bosta | 11 | 4.78 | 0.42 | 33 | 3.85 | 0.48 |
| Balph | 12 | 1.49 | 0.23 | 34 | 10.56 | 0.68 |
| Balmu | 13 | 1.32 | 0.22 | 35 | 6.23 | 0.61 |
| Ratno | 14 | 6.23 | 0.49 | 36 | 12.44 | 0.80 |
| Musmu | 15 | 4.95 | 0.45 | 37 | 5.63 | 0.61 |
| Didvi | 16 | 14.70 | 0.80 | 38 | 2.88 | 0.49 |
| Galga | 17 | 21.92 | 0.99 | 39 | 3.56 | 0.41 |
| Xeola | 18 | 15.91 | 0.84 | 40 | 7.97 | 0.66 |
| Cypca | 19 | 5.22 | 0.45 | | | |
| Crola | 20 | 6.00 | 0.47 | 8 | | |
| Oncmy | 21 | 6.76 | 0.53 | ln L: | -46237.25 +- 662.61 | |
| Petma | 22 | 37.28 | 1.33 | AIC : | 92592.51 | |

Figure 5.2: The tree used in estimating the transition probability matrix of the mtREV model.

Table 5.4: Comparison of branch lengths between the amino acids and synonymous sites of mtDNA.

| | amino acid | 4-fold degenerate | 3rd codon position |
|---|---|---|---|
| 23–Europ | 0.17 ± 0.07 | 0.17 ± 0.10 | 0.22 ± 0.08 |
| 23–Japan | 0.12 ± 0.06 | 0.54 ± 0.18 | 0.39 ± 0.10 |
| 23–Europ/Japan | 0.15 ± 0.06 | 0.36 ± 0.14 | 0.31 ± 0.09 |
| 24–23 | 0.10 ± 0.06 | 0.00 | 0.29 ± 0.13 |
| 24–Afric | 0.26 ± 0.09 | 0.60 ± 0.19 | 0.52 ± 0.14 |
| 26–24 | 1.88 ± 0.25 | 21.36 ± 1.98 | 20.04 ± 1.40 |
| 25–Pantr | 1.19 ± 0.19 | 5.63 ± 0.83 | 5.81 ± 0.58 |
| 25–Panpa | 1.02 ± 0.18 | 5.70 ± 0.84 | 5.39 ± 0.57 |
| 25–Pantr/Panpa | 1.11 ± 0.18 | 5.67 ± 0.83 | 5.60 ± 0.57 |
| 26–25 | 1.08 ± 0.19 | 8.62 ± 1.52 | 10.07 ± 1.17 |
| 27–26 | 1.03 ± 0.20 | 8.85 ± 1.92 | 9.83 ± 1.52 |
| 27–Gorgo | 2.82 ± 0.31 | 21.11 ± 2.31 | 22.07 ± 1.78 |

Comparison of branch lengths between the amino acids and synonymous sites of mtDNA of Hominoidea. Branch lengths (number of amino acid substitution per 100 amino acid sites) estimated from the amino acid sequences (3357 sites; the mtREV model of Table 2.12), the four-fold degenerate sites (1667 sites; the REV model of Table 5.7), and total of the third codon positions (3569 sites; the REV model of Table 5.8) in the tree of Hominoidea. Numbering of the nodes corresponds to that of the preceding tree.

Table 5.5: Numbers of amino acid differences of mtDNA-encoded proteins from Hominoidea.

| | Eur | Afr | Jap | Pan | Pan | Gor | Pon |
|---|---|---|---|---|---|---|---|
| Europ | | 18 | 10 | 131 | 132 | 180 | 359 |
| Afric | 18 | | 16 | 136 | 137 | 183 | 366 |
| Japan | 10 | 16 | | 135 | 136 | 182 | 361 |
| Pantr | 131 | 136 | 135 | | 74 | 186 | 369 |
| Panpa | 132 | 137 | 136 | 74 | | 178 | 361 |
| Gorgo | 180 | 183 | 182 | 186 | 178 | | 371 |
| Ponpy | 359 | 366 | 361 | 369 | 361 | 371 | |

For the data used in estimating the transition probability matrix in Table 2.12 (3357 sites).

root of the human mtDNA tree even if the same rate is assumed both for the humans and chimpanzees. However, since higher transition rate was suggested for the humans than in the chimpanzees, the date of 200,000 years old is likely to be the older estimate. My estimates for this date were 70,000 ± 20,000 and 80,000 ± 20,000 years old from the best models in Tables 5.13 and 5.14 when the 13 Myr was assumed for the orangutan separation. Although the calibration of the clock by using the orangutan separation contains some uncertainty, these estimates strongly suggest that the deepest root of the human mtDNA tree is younger than 200,000 years ago. On the otherhand, the ratio of (24–African)/(25–chimpanzees) is estimated to be 0.23 ± 0.09 from the amino acid sequences. Although marginally, the ratio estimated from the amino acid sequences differs from those of the four-fold degenerate sites and of the total third positions. Consistently with this difference, Horai et al. (1995[140]) estimated 450,000 ± 70,000 years old for the deepest root of the human mtDNA tree from the nonsynonymous sites and the RNA genes, while they gave 143,000 ± 18,000 years date from the synonymous sites. In spite of this apparent discrepancy, Horai et al. did not mention on this difference.

The relative rate test by interspecies comparison of amino acid differences of mtDNA-encoded proteins does not suggest higher rate in humans than in chimpanzees. Table 5.5 gives numbers of amino acid differences in Hominoidea. Numbers of differences of humans from gorilla (180–183) do not differ significantly from those of chimpanzees from gorilla (178–186). Therefore, if the difference of the (24–African)/(25–chimpanzees) ratio between amino acid and synonymous site levels is real, the acceleration of amino acid substitution is likely to have occurred in the human lineage quite recently in the evolutionary time scale long after the human/chimpanzee separation. Takahata (1993[285]) proposed the hypothesis that the relaxation of selective constraint began with the emergence of *Homo sapiens*. The above finding is highly interesting in this respect.

### 5.1.3 Discussion

In spite of the uncertainties discussed above, it seems unlikely from our analysis that the human/chimpanzee separation in the mtDNA tree was much older than 5 Myr, and the most likely date would be 4–5 Myr. Our dating of the human/chimpanzee separation is closely relevant to the dating of the deepest root of the human mtDNA tree, and is in favour of the recent origin hypothesis of modern humans (Cann et al. 1987[46]; Kocher and Wilson 1991[169]; Hasegawa et al. 1993[109]; Ruvolo et al. 1993[250]) rather than the more ancient origin hypothesis (Thorne and Wolpoff 1992[295]; Pesole et al. 1992[236]).

Molecular clock analyses that take account of the rate heterogeneity among lineages (Kishino and Hasegawa 1990[165]; Hasegawa 1991[105]) gave 4.0 ± 1.1 and 4.7 ± 0.8 Myr dates for the human/chimpanzee separation from the ribosomal internal transcribed spacers (ITS1) (Gonzalez et al. 1990[97]) and the immunoglobulin $\epsilon$ pseudogene (Ueda et al. 1989[298]), and 6.3 ± 0.9 and 7.4 ± 0.8 Myr from the intergenic spacer between $\eta$ and $\delta$-globin genes (Maeda et al. 1988) and the $\eta$-globin pseudogene (Miyamoto et al. 1987[216]) for the trifurcation among human, chimpanzee and gorilla (the tricotomy could not be resolved

by these data) with the same reference of 13 Myr for the orangutan separation. Although the estimate from ITS1 is consistent with that from mtDNA, the estimates from the other nuclear genes are older. It should be noted that such gene trees do not necessarily agree with the species tree mainly because of ancestral polymorphism (e.g., Nei 1987[224]). Older coalescence is expected for some nuclear genes.

The expected duration time of polymorphism is proportional to the effective population size under neutrality (Kimura 1983[163]). Since the effective population size of mtDNA is about one-fourth of nuclear genes because of its maternal inheritance and of the haploid nature (Takahata 1985[284]), polymorphism is likely to be maintained for a longer time in the nuclear genes than in the mtDNA. The discrepancy among the dates of human/chimpanzee separation estimated from different genes is thus likely to be due to polymorphism particularly of the $\eta$-globin pseudogene and of the globin spacer in the common ancestral species of human and the African apes (Hasegawa et al. 1987[123]; Hasegawa 1991[105]). If this is the case, it would be reasonable to consider that humans and chimpanzees diverged 4–5 Myr ago as suggested by the mtDNA and ITS1 clocks.

## 5.2 Tempo and Mode of Synonymous Substitution in mtDNA

In section 2.1, Markov models of synonymous substitutions in mtDNA were studied by using Horai et al.'s (1992[141]) 4.8kbp sequence data from Hominoidea. Recently, Horai et al. (1995[140]) published complete sequence data from human, common chimpanzee, bonobo, gorilla, and orangutan. In this section, I reexamine the synonymous substitutions of mtDNA by using this abundant data set. It is shown that the transition rate of the four-fold degenerate sites during evolution, and therefore the transitional mutation rate of mtDNA, are higher in human than in chimpanzees and gorilla probably by about 2 times.

### 5.2.1 Sequence Data

Following Horai et al. (1995[140]), the three human sequences are used; an European (Anderson et al. 1981[15]) designated as Anderson, a Japanese (Ozawa et al. 1991[232]) designated as DCM1, and an African (Horai et al. 1995[140]) designated as SB17F. Anderson is unique in differing by transversions at seven sites where all other human sequences have identical bases. To examine the possibility of inaccuracies in the published Anderson sequence at the seven sites, Horai et al. (1995[140]) determined the nucleotides at the seven sites from an additional six individuals. These six sequences, as well as the 10 Japanese sequences of Ozawa et al. (1991[232]), have no transversions at these sites. Also, four species of nonhuman hominoids have identical bases at four of the seven sites, and all other substitutions are transitions. From these observations, Horai et al. (1995[140]) concluded that the bases at the seven sites of Anderson are incorrectly identified. Thus, following their suggestion, we have used an edited version of the Anderson sequence, obtained by replacing bases at the seven sites with those shared by other humans. Furthermore, Horai et al.'s (1995[140]) sequences from chimpanzee, bonobo, gorilla, and orangutan are used.

Following protein-encoding regions encoded by the mtDNA are used: ND1 (3310–4260 in the numbering of Anderson et al. 1981[15]), ND2 (4473–5510), COI (5907–7442), COII (7589–8266), ATPase 8 (8369–8524), ATPase 6 (8575–9204, overlapping region with ATPase8, 8525–8574, was excluded), COIII (9210–9989), ND3 (10062–10403), ND4L (10473–10757), ND4 (10769–12136), ND5 (12340–14145), and Cyt-b (14750–15886). The total number of deduced codons is 3569. Among these, the number of codons remaining four-fold degenerate during evolution is 1667.

### 5.2.2 Models of Synonymous Nucleotide Substitutions

Table 5.6 gives numbers of transition and transversion differences between species at the four-fold degenerate sites and at the total of the third codon positions.

The transition probability matrices of the REV model were estimated from the 1667 sites data and from the 3569 sites data by the ML method described in section 2.1 based on the tree of the five hominoid species, (((chimp, bonobo), ((Anderson, DCM1), SB17F), gorilla, orang), and they are given in Tables 5.7 and 5.8. Table 5.7 shows that the occurrence of nucleotide substitutions at the four-fold degenerate sites

Table 5.6: Numbers of transition and transversion nucleotide differences among Hominoidea

| | Ander | | DCM1 | | SB17F | | Chimp | | Bonobo | | Gorilla | | Orang | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ander | | | 12 | (21) | 11 | (34) | 329 | (758) | 322 | (741) | 355 | (855) | 357 | (872) |
| DCM1 | 0 | (1) | | | 17 | (41) | 332 | (760) | 327 | (747) | 353 | (851) | 360 | (872) |
| SB17F | 2 | (3) | 2 | (2) | | | 327 | (757) | 320 | (738) | 350 | (848) | 355 | (869) |
| Chimp | 47 | (54) | 47 | (53) | 49 | (55) | | | 142 | (318) | 343 | (818) | 364 | (896) |
| Bonobo | 48 | (53) | 48 | (52) | 50 | (54) | 19 | (23) | | | 328 | (793) | 344 | (873) |
| Gorilla | 92 | (118) | 92 | (117) | 94 | (119) | 81 | (110) | 82 | (111) | | | 368 | (895) |
| Orang | 204 | (263) | 204 | (262) | 206 | (264) | 195 | (257) | 198 | (258) | 200 | (267) | | |

Numbers of transition nucleotide differences (upper right half) and those of transversion differences (lower left half) of four-fold degenerate sites (1667 sites) and total of the third codon positions (3569 sites; in parentheses) among Hominoidea.

Table 5.7: Transition probability matrix of the REV model for the four-fold degenerate sites.

| ↗ | T | C | A | G |
|---|---|---|---|---|
| T | 0.97893 | 0.01895 | 0.00175 | 0.00037 |
| C | 0.00627 | 0.99306 | 0.00054 | 0.00013 |
| A | 0.00066 | 0.00061 | 0.99432 | 0.00441 |
| G | 0.00132 | 0.00144 | 0.04124 | 0.95599 |
| $\pi$ | 0.144 | 0.434 | 0.381 | 0.041 |

The transition probability matrix $\boldsymbol{M}$ of the REV model for a unit time interval (one substitution per 100 sites) estimated by ML from the four-fold degenerate sites (1667 sites).

Table 5.8: Transition probability matrix of the REV model for the total of third codon positions.

| ↗ | T | C | A | G |
|---|---|---|---|---|
| T | 0.97980 | 0.01911 | 0.00086 | 0.00024 |
| C | 0.00707 | 0.99251 | 0.00030 | 0.00012 |
| A | 0.00037 | 0.00035 | 0.99491 | 0.00437 |
| G | 0.00086 | 0.00119 | 0.03702 | 0.96093 |
| $\pi$ | 0.159 | 0.429 | 0.369 | 0.043 |

The transition probability matrix $\boldsymbol{M}$ of the REV model for a unit time interval (one substitution per 100 sites) estimated by ML from the total of third codon positions (3569 sites).

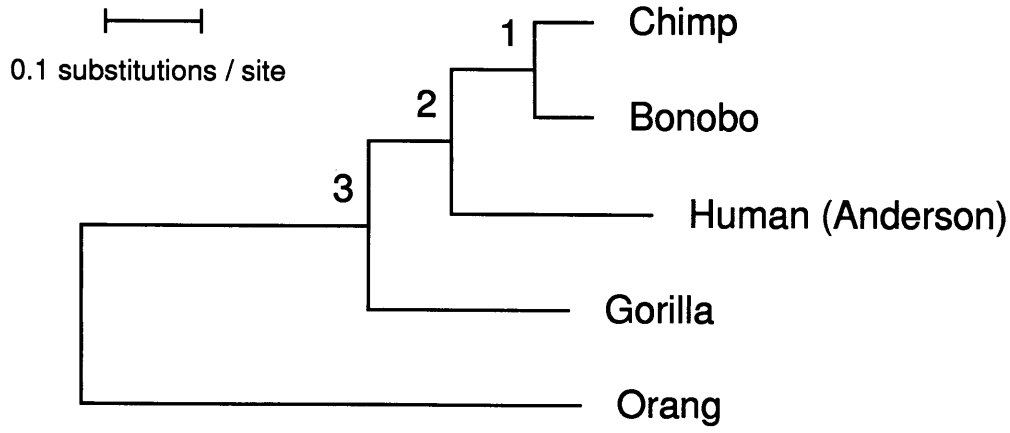



Figure 5.3: The ML tree of the four-fold degenerate sites.

The ML tree of the four-fold degenerate sites (1669 sites) based on the REV model. The horizontal length of each branch is proportional to the estimated number of substitutions. The root of this tree is located somewhere within the 3–Orang branch.

is distinctly asymmetric between the two strands of mtDNA as was discussed in subsection 2.1.3. G→A and T→C transitions are 0.04124/0.00627 = 6.6 and 0.01895/0.00441 = 4.3 times more frequent on the L-strand (as represented in the table) than on the H-strand, respectively. This nucleotide substitution bias is again roughly consistent with Tanaka and Ozawa's (1994[289]) estimates from the four-fold degenerate sites of the entire mitochondrial genomes of 43 human individuals; that is, G→A and T→C transitions are 9 and 1.8 times more frequent on the L-strand than on the H-strand (subsection 2.1.3).

The ML tree estimated by this model for the four-fold degenerate sites data is represented in Fig. 5.3. In this figure, only Anderson was used from human, and number of four-fold degenerate sites is 1669 for this data set. The branch lengths of the tree estimated by the REV, TN93, and HKY85 models for the four-fold degenerate sites are given in Table 5.9. The branch length leading to human is significantly longer than those leading to the African apes.

The REV, TN93 and HKY85 models gave AIC of 11,253.3, 11,287.7, and 11,349.2 for the four-fold degenerate sites, and the REV model turned out to be the best among these models in approximating the evolution of the four-fold degenerate sites consistently with the analysis in section 2.1 with a more limited data set, and although the TN93 model is inferior to the REV model, it is much better than the HKY85 model.

### 5.2.3 Fitting of Models to the Data

For the alignment of 5 OTUs, $4^5 = 1024$ configurations of nucleotide sites are possible, and probabilities of respective configurations were calculated under the respective models with the branch lengths given in Table 5.9. Grouping these configurations into 8 categories of 0-change, 1-TC-transition (configurations

Table 5.9: Branch lengths of the tree of four-fold degenerate sites for the REV, TN93 and HKY85 models.

| | REV | TN93 | HKY85 |
|---|---|---|---|
| including both chimp and bonobo | | | |
| 1–Chimp | 5.71 ± 0.83 | 5.64 ± 0.83 | 5.90 ± 0.83 |
| 1–Bonobo | 5.66 ± 0.83 | 5.75 ± 0.84 | 5.31 ± 0.80 |
| 2–Man | 21.54 ± 1.98 | 21.25 ± 1.97 | 21.32 ± 1.98 |
| 3–Gorilla | 21.10 ± 2.30 | 21.35 ± 2.34 | 20.55 ± 2.36 |
| 3–Orang | 82.61 ± 7.00 | 81.66 ± 7.00 | 86.37 ± 7.64 |
| 2–1 | 8.35 ± 1.51 | 8.52 ± 1.52 | 8.14 ± 1.52 |
| 3–2 | 8.86 ± 1.91 | 8.68 ± 1.91 | 9.45 ± 1.99 |
| 2–1–Chimp | 14.06 ± 1.72 | 14.16 ± 1.73 | 14.04 ± 1.73 |
| 2–1–Bonobo | 14.01 ± 1.72 | 14.27 ± 1.74 | 13.45 ± 1.72 |
| including only chimp | | | |
| 2–Chimp | 13.76 ± 1.76 | | |
| 2–Man | 21.29 ± 2.05 | | |
| 3–Gorilla | 21.16 ± 2.37 | | |
| 3–Orang | 84.86 ± 7.13 | | |
| 3–2 | 9.28 ± 2.00 | | |
| including only bonobo | | | |
| 2–Bonobo | 13.45 ± 1.74 | | |
| 2–Man | 21.41 ± 2.04 | | |
| 3–Gorilla | 20.41 ± 2.29 | | |
| 3–Orang | 82.60 ± 7.00 | | |
| 3–2 | 8.63 ± 1.91 | | |

Branch lengths (numbers of substitutions per 100 sites) of the four-fold degenerate nucleotide sites (1669 sites) for the REV, TN93 and HKY85 models (ML estimates). The ML estimates of parameters are as follows; $\alpha/\beta = 23.6$ for the HKY85 model, and $(\alpha_Y + \alpha_R)/(2\beta) = 31.0$ and $\alpha_Y/\alpha_R = 0.4$ for the TN93 model. ± refers to 1SE.

Table 5.10: Distribution of configurations of four-fold degenerate sites.

| Configuration | Obs. | REV model Expec. | REV model $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ | TN93 model Expec. | TN93 model $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ | HKY85 model Expec. | HKY85 model $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ |
|---|---|---|---|---|---|---|---|
| 0-change | 764 | 756.0 | 0.085 | 739.6 | 0.805 | 735.7 | 1.089 |
| 1-TC-transition | 365 | 366.3 | 0.005 | 372.5 | 0.151 | 389.7 | 1.566 |
| 1-AG-transition | 152 | 153.2 | 0.009 | 156.3 | 0.118 | 127.9 | 4.541 |
| 1-GT-transversion | 1 | 1.4 | 0.114 | 0.8 | 0.050 | 0.9 | 0.011 |
| 1-GC-transversion | 5 | 6.7 | 0.431 | 6.1 | 0.198 | 6.9 | 0.523 |
| 1-AT-transversion | 59 | 45.7 | 3.871 | 29.6 | 29.20 | 29.0 | 21.034 |
| 1-AC-transversion | 104 | 102.1 | 0.035 | 129.7 | 5.092 | 130.1 | 5.236 |
| $\geq$2-changes | 219 | 237.6 | 1.456 | 234.4 | 1.012 | 248.8 | 3.569 |
| total | 1669 | 1669.0 | $\chi^2 = 6.006$ | 1669.0 | $\chi^2 = 36.628$ | 1669.0 | $\chi^2 = 47.569$ |
| | | | df = 7 | | df = 7 | | df = 7 |
| | | | $P = 0.54$ | | $P < 10^{-5}$ | | $P < 10^{-5}$ |

Distribution of configurations of four-fold degenerate sites (1669 sites) for the REV, TN93 and HKY85 models (ML estimates). The ML estimates of parameters are as follows; $\alpha/\beta = 23.6$ for the HKY85 model, and $(\alpha_Y + \alpha_R)/(2\beta) = 31.0$ and $\alpha_Y/\alpha_R = 0.4$ for the TN93 model.

which could arise from one transition between T and C), 1-AG-transition, 1-GT-transversion, 1-GC-transversion, 1-AT-transversion, 1-AC-transversion, and $\geq$2-changes (configurations which could not arise from less than two changes), a $\chi^2$ test for the REV model gave $P$ value of as high as 0.54 (Table 5.10), indicating that the transition probability matrix of Table 5.7 well approximates the evolution of four-fold degenerate sites consistently with my previous analysis on the more limited data set (Section 2.1). Both of $\chi^2$ tests for the TN93 and HKY85 models gave $P$ values lower than $10^{-5}$. Discrepancies of these models with the data are mainly due to more frequent AT-transversions and less frequent AC-transversions than expected.

The same test shown in Table 5.11 of the total third codon positions for the REV model (the transition probability matrix in Table 5.8) also gave a high $P$ value (0.46). However, a detailed study clarifies some discrepancy of the REV model from the data. The most significant discrepancy is found for the configuration CCCCT (in the order of chimpanzee, bonobo, human, gorilla, and orangutan). The observed numbers of sites of the CCCCT configuration are 68 and 147 for the four-fold degenerate sites and the total third codon positions, respectively, while the expected numbers for the REV model are 96.6 and 222.8. These discrepancies would be due to the unequal base composition of orangutan from the other Hominoidea species. The $\pi_T$'s of SB17F, chimpanzee, bonobo, gorilla and orangutan are 0.139, 0.154, 0.153, 0.159 and 0.123, respectively, and the $\pi_C$'s are 0.438, 0.424, 0.424, 0.421 and 0.457 (Table 5.12). Suppose $n_{ij}$ be the number of sites in which gorilla has a base $i$ and orangutan has $j$ in the four-fold degenerate sites (number of sites $n = 1667$), and suppopse $n_{i*} = \Sigma_j n_{ij}$ and $n_{*j} = \Sigma_i n_{ij}$. Then, $n_{T*} - n_{*T} = 60$, suggesting lower T content in the orangutan than in the gorilla. In order to test whether this difference is significant, the variance of this difference is estimated by the following formula

Table 5.11: Distribution of configurations of third codon positions.

| Configuration | Observed | Expected | $\frac{(\text{Obs}-\text{Exp})^2}{\text{Exp}}$ |
|---|---|---|---|
| 0-change | 1668 | 1622.3 | 1.287 |
| 1-TC-transition | 868 | 904.8 | 1.497 |
| 1-AG-transition | 347 | 365.3 | 0.917 |
| 1-GT-transversion | 1 | 2.4 | 0.817 |
| 1-GC-transversion | 11 | 11.7 | 0.042 |
| 1-AT-transversion | 69 | 58.9 | 1.732 |
| 1-AC-transversion | 132 | 125.5 | 0.337 |
| $\geq$2-changes | 473 | 478.1 | 0.054 |
| total | 3569 | 3569.0 | $\chi^2 = 6.682$ |
| | | | df = 7 |
| | | | $P = 0.46$ |

Distribution of configurations of third codon positions (3569 sites) for the REV model (ML estimates).

Table 5.12: Base composition of four-fold degenerate sites of mtDNA (1667 sites).

| | T | C | A | G |
|---|---|---|---|---|
| Anderson | 0.138 | 0.440 | 0.377 | 0.045 |
| DCM1 | 0.140 | 0.439 | 0.377 | 0.044 |
| SB17F | 0.139 | 0.438 | 0.380 | 0.043 |
| Chimp | 0.154 | 0.424 | 0.380 | 0.043 |
| Bonobo | 0.153 | 0.424 | 0.390 | 0.033 |
| Gorilla | 0.159 | 0.421 | 0.378 | 0.043 |
| Orang | 0.123 | 0.457 | 0.386 | 0.035 |

(Hasegawa and Kishino 1989[118]),

$$\text{Var}(n_{T*} - n_{*T}) = n_{T*} + n_{*T} - 2n_{TT} - (n_{T*} - n_{*T})^2/n. \tag{5.6}$$

The SE of $n_{T*} - n_{*T}$ is estimated to be 18.7, indicating that the orangutan has significantly lower T content than the gorilla. A similar analysis for the C content shows that $n_{C*} - n_{*C} = -60$ and its SE is 19.9, suggesting that the oragutan has significantly higher C content than the gorilla. These significant differences of T and C contents of the orangutan from the other Hominoids hold not only for the gorilla but also for chimpanzee and bonobo, and these hold for the total third codon positions (data not shown).

It is apparent that the transition rate between purines is higher than that between pyrimidines by about 2 times, and in terms of AIC the TN93 model better approximates the 1669 sites data than the HKY85 model does. As for the branch lengths, however, the estimates from the three models do not differ significantly (Table 5.9), and therefore the estimates of the evolutionary rate and the branching dates would be robust to the choice among these models. For this reason we shall use the HKY85 model in estimating the evolutionary rate and the branching dates in Hominoidea because of its simplicity.

### 5.2.4 Rate Heterogeneity among Lineages

Fig. 5.3 and Table 5.9 suggest longer branch length, and therefore higher evolutionary rate, in human than in chimpanzee and bonobo. Kishino and Hasegawa (1990[165]) devised a method to estimate simultaneously the evolutionary rate and the branching dates from difference data of nucleotides (such as given in Table 5.6) by the generalized least-squares. The method assumes the HKY85 model and allows rate variation among lineages; that is, $\alpha$ and $\beta$ in Eq. 2.12 can differ among branches, and we can assign different rate parameters to different branches. Among the alternative models for rate variation, we can select the best model by minimizing the AIC, in which a penalty is imposed in introducing too many parameters.

By assuming 13 Myr for the orangutan separation as in section 5.1 (Pilbeam 1988[241]; Andrews 1992[17]; McCrossin et al. 1993[207]), I estimate the evolutionary rates and the branching dates based on several models for rate variation (Table 5.13). At first, the simplest model assuming constant rate is used. For this model, sum of squares of residuals (SSR) is 51.79. SSR follows a $\chi^2$ distribution with the degree of freedom (df) equal to $s(s-1)$ ($s$: number of OTUs) minus the number of free parameters. The $P$ value of this model is 0.03 which is not satisfactorily good. Fig. 5.4 shows generalized least-squares fitting of the relationship between $S/n$ and $V/n$, where $S$ and $V$ refer to the numbers of transition and transversion differences, and $n$ is the number of nucleotides, based on Model 1. The plots of human vs two chimpanzees (node 2) are above the theoretical curves of the HKY85 and REV models. This deviation would likely to be due to change of the pattern of nucleotide substitution in the human lineage relative to the others. It should be noted that, when the transition probability matrix of Table 5.7 holds for all the lineages, all the plots in Fig. 5.4 should be distributed along the theoretical curve even if the absolute rate differs among different lineages. Model 2 (Table 5.13) which allows rate variation in the human lineages improves AIC. The estimate of $\beta_2$ of human in Model 2 does not differ significantly from $\beta_1$ of the other branches, and hence, in Model 3, $\beta$ of human is assumed to be identical to those of the other branches. This reduction of the number of free parameters by one from Model 2 improves AIC.

Introduction of more complicated models does not improve AIC, and it turned out that Model 3 is the best model in approximating the data of four-fold degenerate sites. The $P$ value for Model 3 is as high as 0.28. The estimates of Model 3 are 6.0 $\pm$ 0.7, 3.5 $\pm$ 0.5, and 1.6 $\pm$ 0.2 Myr old ($\pm$: 1SE), respectively, for the separations of gorilla, human/chimpanzee, and chimpanzee/bonobo. These estimates are all consistent with those estimated by the ML from the amino acid sequences (section 5.1; Adachi and Hasegawa 1995[4]).

The estimates of transition rate $v_S = 2(\pi_T\pi_C + \pi_A\pi_G)\alpha$ are 0.0674 $\pm$ 0.0152 and 0.0314 $\pm$ 0.0057 /Myr/site for the human and the other lineages, respectively. The estimate of transversion rate $v_V = 2(\pi_T + \pi_C)(\pi_A + \pi_G)\beta$ is 0.0050 $\pm$ 0.0004 /Myr/site for Model 3. Therefore, the estimates of the total substitution rate, $v = v_S + v_V$, of four-fold degenerate sites is 0.0724 $\pm$ 0.0154 and 0.0363 $\pm$

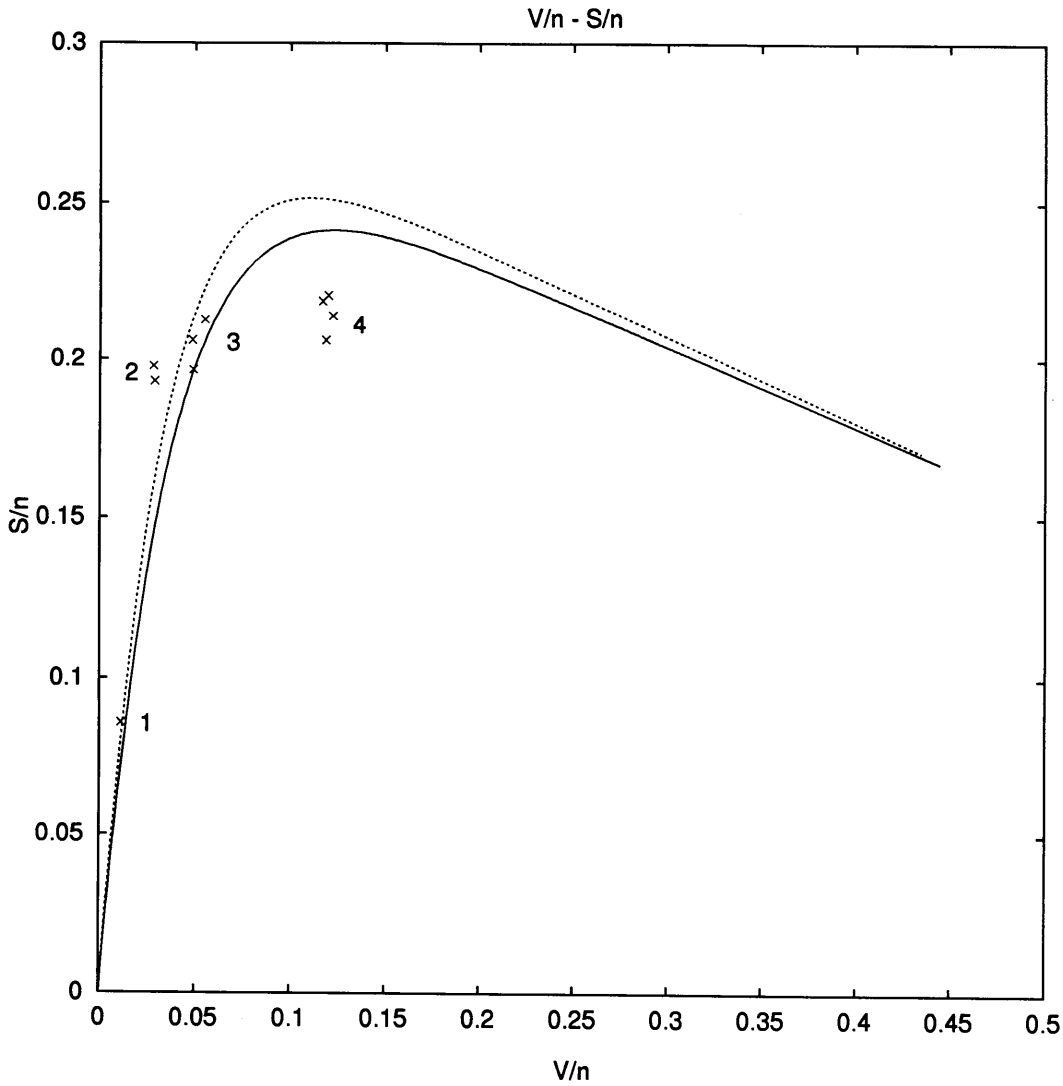



Figure 5.4: Generalized least-squares fitting.

Generalized least-squares fitting of the relationship between $S/n$ and $V/n$ for the four-fold degenerate sites (1669 sites). Theoretical curves are based on the HKY85 model (solid line; Model1 in Table 5.13) and on the REV model (dotted line; Table 5.7). Numberings refer to those of nodes in Fig. 5.3, and node 4 corresponds to the orangutan separation.

Table 5.13: Branching dates and evolutionary rates estimated from the four-fold degenerate sites of mtDNA.

| Model | 1 | 2 | 3 |
|---|---|---|---|
| | constant rate | rate change in human | |
| Rates in branches | | | |
| Human | $\alpha_1, \beta_1$ | $\boldsymbol{\alpha_2, \beta_2}$ | $\boldsymbol{\alpha_2}, \beta_1$ |
| Other branches | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ |
| Rates ($10^{-3}$/Myr) | | | |
| $\widehat{\alpha}_1$ | 265.3 ± 60.7 | 203.4 ± 39.6 | 200.8 ± 36.6 |
| $\widehat{\alpha}_2$ | — | 449.7 ± 120.8 | 431.6 ± 97.3 |
| $\widehat{\beta}_1$ | 10.3 ± 0.9 | 10.1 ± 0.8 | 10.2 ± 0.8 |
| $\widehat{\beta}_2$ | — | 11.1 ± 3.0 | — |
| Branching dates (Myr) | | | |
| $t_1$ (Orang) | 13 | 13 | 13 |
| $\widehat{t}_2$ (Gorilla) | 5.57 ± 0.83 | 5.88 ± 0.71 | 5.95 ± 0.66 |
| $\widehat{t}_3$ **(Human/Chimp)** | **3.90 ± 0.63** | **3.38 ± 0.62** | **3.49 ± 0.51** |
| $\widehat{t}_4$ (Chimp/Bonobo) | 1.30 ± 0.27 | 1.54 ± 0.25 | 1.56 ± 0.24 |
| $\widehat{t}_5$ (SB17F/(Ander,DCM1)) | 0.11 ± 0.03 | 0.07 ± 0.02 | 0.07 ± 0.02 |
| $\widehat{t}_6$ (Ander/DCM1) | 0.08 ± 0.03 | 0.05 ± 0.02 | 0.05 ± 0.02 |
| SSR | 51.79 | 38.24 | 38.36 |
| df | 35 | 33 | 34 |
| $P$ | 0.03 | 0.25 | 0.28 |
| AIC | 163.64 | 154.08 | 152.21 |
| | | | ↑ minimum AIC |

Branching dates and evolutionary rates estimated from the four-fold degenerate sites of mtDNA (1667 sites) by Kishino and Hasegawa's (1990[165]) method. The tree topology and numbering of nodes are shown in Fig. 5.5 (node 1 is the root of the tree). ± is 1SE. SSR refers to the sum of squares of residuals in the generalized least squares of $\mathbf{D} = (\ldots, V_{ij}, \ldots, S_{ij}, \ldots)$, where $V_{ij}$ and $S_{ij}$ are numbers of transversion and transition differences between species $i$ and $j$ given in Table 5.6.

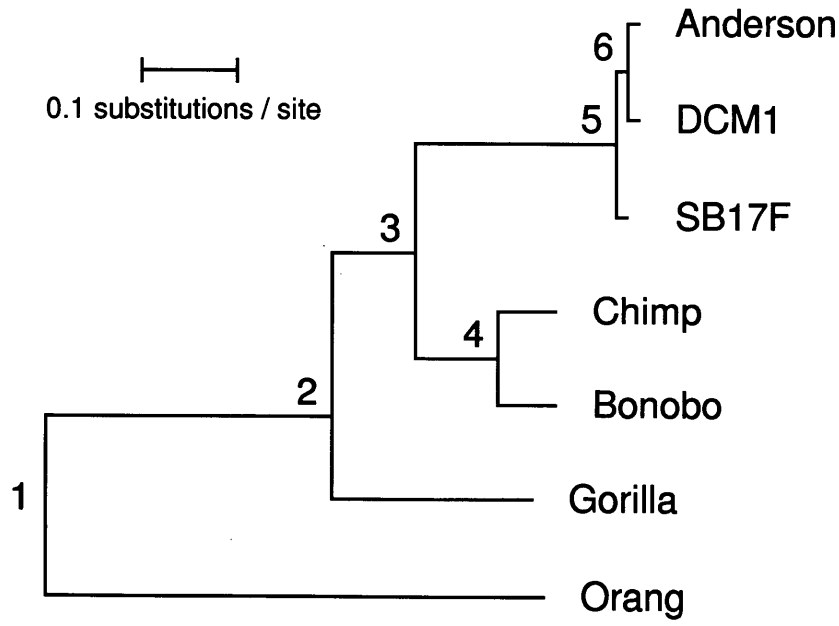



Figure 5.5: The ML tree of the four-fold degenerate sites.

The ML tree of the four-fold degenerate sites (1667 sites) based on the REV model. The horizontal length of each branch is proportional to the estimated number of substitutions. The root (node 1) is arbitrarily placed between node 2 and Orangutan.

0.0059/Myr/site, respectively, for human and the African apes.

The ratio of transition rate parameter $\alpha$ of human to that of others is $2.15 \pm 0.42$, suggesting significant accerelation of transition rate in human. Similar analysis of all third position data (3569 sites) also chooses Model 3 as the best (Table 5.14), and estimate the ratio to be $2.12 \pm 0.31$. However, the fittings of this model to the 3569 sites data ($P = 0.08$) is not as good as to the 1667 sites data ($P = 0.28$), probably because of the heterogeneity of the rate between third positions of two-fold and four-fold degenerate codons.

Another remarkable feature in Fig. 5.4 is that the plots for the orangutan divergence are located significantly under the theoretical curves. By using the 4.8kbp data of Horai et al. (1992[141]) with siamang as an outgroup, Adachi and Hasegawa (1995c[7]) suggested that transversion rate in the orangutan lineages is higher by about 1.5 fold than in the African apes/human clade, while the transition rate in the orangutan does not differ significantly from the others. The deviation of the plots for the orangutan divergence might be due to this putative difference of the substitution pattern in orangutan from the others.

### 5.2.5 Including Siamang as an Outgroup

In order to evaluate how much extent, if any, the evolutionary rate in the orangutan lineage is different from those of other hominoid species, we will analyze Horai et al.'s (1992[141]) 4.8kbp data which include

Table 5.14: Branching dates and evolutionary rates estimated from the total third codon positions of mtDNA.

| Model | 1 | 2 | 3 |
|---|---|---|---|
| | constant rate | rate change in human | |
| Rates in branches | | | |
| Human | $\alpha_1, \beta_1$ | $\boldsymbol{\alpha_2, \beta_2}$ | $\boldsymbol{\alpha_2}, \beta_1$ |
| Other branches | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ |
| Rates ($10^{-3}$/Myr) | | | |
| $\widehat{\alpha}_1$ | 242.0 ± 46.8 | 194.2 ± 23.2 | 198.2 ± 24.2 |
| $\widehat{\alpha}_2$ | — | 389.9 ± 76.8 | 419.5 ± 76.6 |
| $\widehat{\beta}_1$ | 5.8 ± 0.5 | 5.9 ± 0.4 | 5.9 ± 0.4 |
| $\widehat{\beta}_2$ | — | 4.8 ± 1.2 | — |
| Branching dates (Myr) | | | |
| $t_1$ (Orang) | 13 | 13 | 13 |
| $\widehat{t}_2$ (Gorilla) | 6.22 ± 0.84 | 6.50 ± 0.55 | 6.39 ± 0.55 |
| $\widehat{t}_3$ **(Human/Chimp)** | **4.25 ± 0.62** | **3.72 ± 0.50** | **3.52 ± 0.44** |
| $\widehat{t}_4$ (Chimp/Bonobo) | 1.34 ± 0.24 | 1.58 ± 0.18 | 1.55 ± 0.17 |
| $\widehat{t}_5$ (SB17F/(Ander,DCM1)) | 0.13 ± 0.03 | 0.09 ± 0.02 | 0.08 ± 0.02 |
| $\widehat{t}_6$ (Ander/DCM1) | 0.07 ± 0.02 | 0.05 ± 0.01 | 0.04 ± 0.01 |
| SSR | 70.06 | 45.27 | 46.02 |
| df | 35 | 33 | 34 |
| $P$ | 0.0004 | 0.08 | 0.08 |
| AIC | 200.96 | 180.17 | 178.92 |
| | | | ↑ minimum AIC |

Branching dates and evolutionary rates estimated from the total third codon positions of mtDNA (3569 sites) by Kishino and Hasegawa's (1990[165]) method. The tree topology and numbering of nodes are shown in Fig. 5.5 (node 1 is the root of the tree).

siamang as an outgroup. The 1344 codons analyzed in section 5.1 is used. Among these, the number of codons remaining four-fold degenerate during evolution is 611. Table 5.15 gives numbers of transition and transversion differences between species at these 611 sites. The numbers of transition differences between gorilla and the two chimpanzees (112 and 114) are smaller than that between gorilla and human (134), and are even smaller than those of the more closely related pairs of human and chimpanzees (128 and 124) suggesting rate variation among these species.

Table 5.15: Numbers of transition and transversion differences.

| | Siamang | Orang | Gorilla | Human | Chimp | Bonobo |
|---|---|---|---|---|---|---|
| Siamang | | 148 (107) | 152 (98) | 160 (96) | 154 (93) | 158 (88) |
| Orang | 142 | | 151 (65) | 150 (65) | 144 (66) | 144 (65) |
| Gorilla | 121 | 138 | | 134 (32) | 112 (31) | 114 (30) |
| Human | 116 | 139 | 61 | | 128 (13) | 124 (12) |
| Chimp | 127 | 141 | 61 | 39 | | 52 (7) |
| Bonobo | 123 | 136 | 64 | 43 | 22 | |

Numbers of transition nucleotide differences followed by those of transversion differences (in parentheses) of four-fold degenerate sites (611 sites) among Hominoidea (upper right half), and numbers of amino acid differences (1334 sites; lower left half).

At first, I use the constant rate model (Model 1; Table 5.16). For this model, sum of squares of residuals (SSR) is 35.88. SSR follows a $\chi^2$ distribution with the degree of freedom (df) equal to $s(s-1)$ ($s$: number of OTUs) minus the number of free parameters. The $P$ value of this model is 0.06 which is not satisfactorily good. Fig. 5.6 shows generalized least-squares fitting of the relationship between $S/n$ and $V/n$, where $S$ and $V$ refer to the numbers of transition and transversion differences, and $n$ is the number of nucleotides, based on Model 1. The plots of human vs two chimpanzees (node 2) are above the theoretical curve, while the plots of orangutan vs the African apes and humans (node 4) are below the curve. These deviations would likely to be due to change of evolutionary rate in the orangutan and the human lineages relative to the other lineages. Models 2 and 3 which allow rate variation, respectively, in the orangutan and the human lineages improve AIC. Model 4 which allows independent rate change both in the orangutan and the human lineages is inferior to Model 3 by AIC. The estimate of $\alpha_2$ of orangutan and that of $\beta_3$ of human in Model 4 do not differ significantly from $\alpha_1$ and $\beta_1$ of the other branches, and hence, in Model 5, $\alpha$ of orangutan and $\beta$ of human are assumed to be identical to those of the other branches. This reduction of the number of free parameters by two from Model 4 improves AIC.

The model, that assigns $\alpha_2$ and $\beta_2$ to all branches in the African apes/human clade except the human lineage assigned by $\alpha_3$ and $\beta_3$, is much inferior (AIC = 145.73) to Models 4 and 5, and this model estimates the human/chimpanzee separation to be 3.70 $\pm$ 1.20 Myr old. Introduction of more complicated models does not improve AIC, and it turned out that Model 5 is the best model in approximating the data of four-fold degenerate sites. The $P$ value for Model 5 is as high as 0.80. The estimates of Model 5 are 24.4 $\pm$ 4.1, 7.0 $\pm$ 0.9, 3.5 $\pm$ 0.9, and 2.1 $\pm$ 0.4 Myr old, respectively, for the separations of siamang, gorilla,

Table 5.16: Branching dates and evolutionary rates estimated from the four-fold degenerate sites including siamang.

| Model | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| | constant | rate change | | | |
| | rate | in orang | in human | in orang & human | |
| Rates in branches | | | | | |
| 4–Orang | $\alpha_1, \beta_1$ | $\boldsymbol{\alpha_2}, \boldsymbol{\beta_2}$ | $\alpha_1, \beta_1$ | $\boldsymbol{\alpha_2}, \boldsymbol{\beta_2}$ | $\alpha_1, \boldsymbol{\beta_2}$ |
| 2–Human | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ | $\boldsymbol{\alpha_3}, \boldsymbol{\beta_3}$ | $\boldsymbol{\alpha_3}, \boldsymbol{\beta_3}$ | $\boldsymbol{\alpha_3}, \beta_1$ |
| Other branches | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ | $\alpha_1, \beta_1$ |
| Rates ($10^{-3}$/Myr) | | | | | |
| $\widehat{\alpha}_1$ | 172.3 ± 34.3 | 162.7 ± 42.3 | 136.6 ± 21.3 | 116.0 ± 29.2 | 130.2 ± 20.1 |
| $\widehat{\alpha}_2$ | — | 123.9 ± 62.0 | — | 156.5 ± 54.5 | — |
| $\widehat{\alpha}_3$ | — | — | 420.7 ± 141.7 | 365.6 ± 138.8 | 414.3 ± 140.6 |
| $\widehat{\beta}_1$ | 8.6 ± 1.1 | 6.9 ± 1.7 | 9.3 ± 1.0 | 7.4 ± 1.7 | 7.8 ± 1.3 |
| $\widehat{\beta}_2$ | — | 11.9 ± 1.9 | — | 11.9 ± 1.9 | 11.7 ± 1.9 |
| $\widehat{\beta}_3$ | — | — | 6.9 ± 4.0 | 6.1 ± 3.6 | — |
| Branching dates (Myr) | | | | | |
| $\widehat{t}_5$ (Siamang) | 22.93 ± 3.17 | 27.31 ± 6.53 | 21.49 ± 2.68 | 25.70 ± 5.46 | 24.38 ± 4.12 |
| $t_4$ (Orang) | 13 | 13 | 13 | 13 | 13 |
| $\widehat{t}_3$ (Gorilla) | 6.60 ± 1.00 | 7.09 ± 1.71 | 6.48 ± 0.79 | 7.65 ± 1.62 | 6.96 ± 0.94 |
| $\widehat{t}_2$ (Human/Chimp) | 4.73 ± 0.70 | 5.20 ± 1.23 | 3.42 ± 0.82 | 3.96 ± 1.18 | 3.51 ± 0.86 |
| $\widehat{t}_1$ (Chimp/Bonobo) | 1.74 ± 0.36 | 1.88 ± 0.52 | 1.92 ± 0.32 | 2.29 ± 0.58 | 2.07 ± 0.37 |
| SSR | 35.88 | 30.64 | 18.52 | 15.81 | 16.25 |
| df | 24 | 22 | 22 | 20 | 22 |
| $P$ | 0.06 | 0.10 | 0.67 | 0.73 | 0.80 |
| AIC | 148.47 | 147.23 | 135.11 | 136.40 | 132.84 |
| | | | | | ↑ minimum AIC |

Branching dates and evolutionary rates estimated from the four-fold degenerate sites of mtDNA (611 sites). The tree topology and numbering of nodes are shown in Fig. 5.1 (node 5 is the root of the tree).

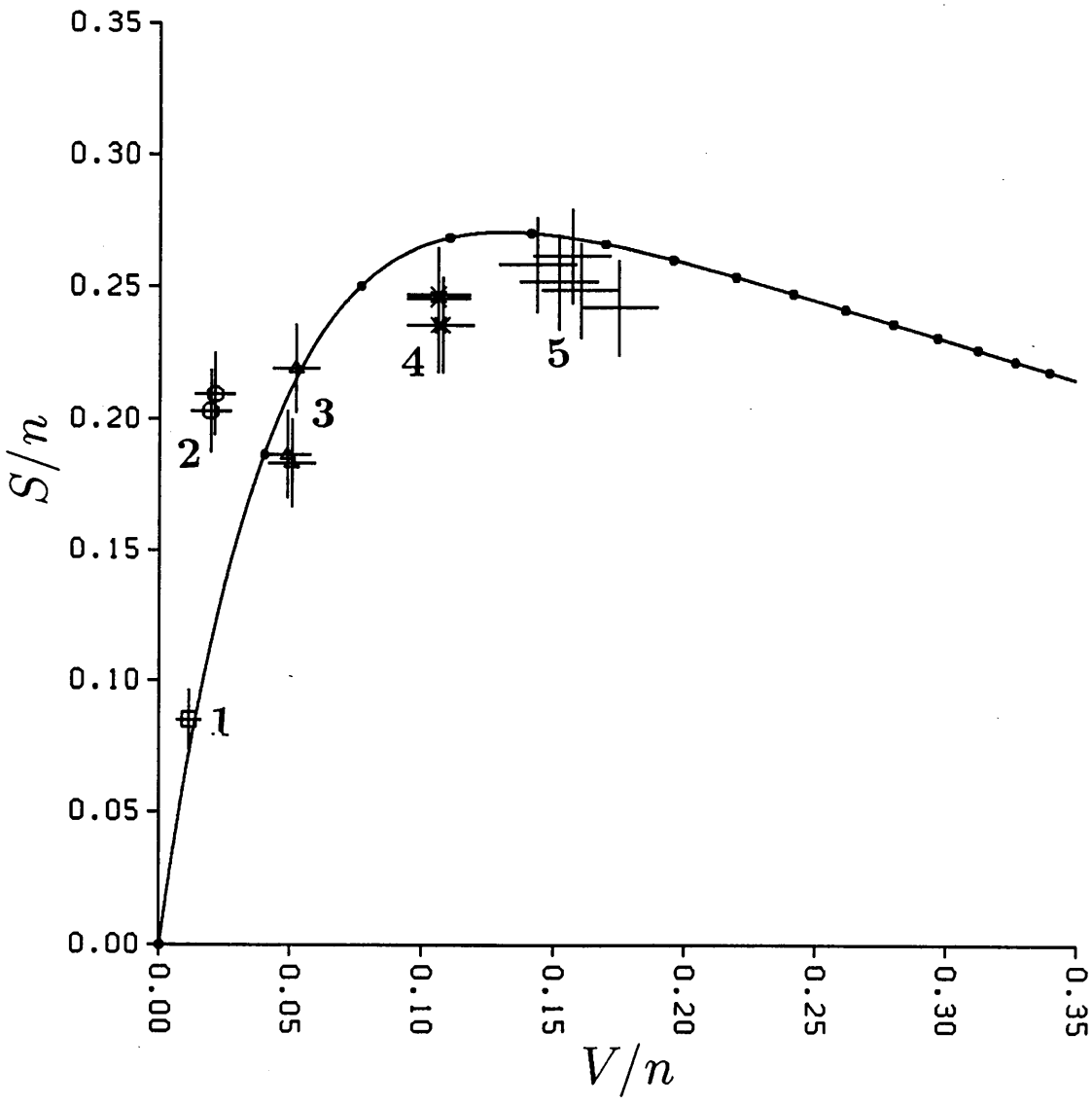



Figure 5.6: The relationship between S/n and V/n.

Generalized least-squares fitting of the relationship between $S/n$ and $V/n$ for the four-fold degenerate sites (611 sites). Theoretical curve is based on the HKY85 model (Model1 in Table 5.16). Numberings refer to those of nodes in Fig. 2.1, and node 5 corresponds to the siamang separation. Vertical and horizontal lines indicate SEs of $S_{ij}/n$ and $V_{ij}/n$, respectively. The interval between neighboring small circles along the curve is 5 Myr.

human/chimpanzee, and chimpanzee/bonobo. These estimates are all consistent with those estimated by the ML from the amino acid sequences (section 5.1; Adachi and Hasegawa 1995[4]) except for the siamang separation that was not estimated.

The estimates of transition rate $v_S = 2(\pi_T\pi_C+\pi_A\pi_G)\alpha$ are $0.072 \pm 0.025$ and $0.023 \pm 0.004$ /Myr/site for the human and the other lineages, respectively. The estimates of transversion rate $v_V = 2(\pi_T + \pi_C)(\pi_A + \pi_G)\beta$ are $0.006 \pm 0.001$ and $0.004 \pm 0.001$ /Myr/site for the orangutan and the other lineages, respectively. Therefore, the estimates of the total substitution rate, $v = v_S + v_V$, of four-fold degenerate sites is $0.076 \pm 0.025$ and $0.026 \pm 0.004$/Myr/site, respectively, for human and the African apes.

The ratio of transition rate parameter $\alpha$ of human to that of others is $3.18 \pm 1.00$, suggesting significant accerelation of transition rate in human. Similar analyses of all third position data (1344 sites) and of first position data of *Leu* codons of *UUR* and *CUR* (*R*: purine) in addition to the 1344 sites data (1449 sites in total) also choose Model 5 as the best, and estimate the ratio to be $2.39 \pm 0.60$ and $2.84 \pm 0.71$, respectively. However, the fittings of this model to the 1344 and 1449 sites data ($P = 0.11$ and 0.13) are not as good as to the 611 sites data ($P = 0.80$), probably because of the heterogeneity of the rate between third positions of two-fold and four-fold degenerate codons as mentioned in sectionratehetero.

The accelerated transition rate of mtDNA in human could theoretically be an artefact of sequencing error. Ozawa et al. (1991[232]) sequenced complete mtDNAs from six Japanese patients with mitochondrial diseases, and we used a sequence labeled FICM which is most divergent among their data from that of Anderson et al.'s (1981[15]) sequence to examine the robustness of our estimate. FICM differs from Anderson et al.'s sequence by 4 transitions and by 1 transversion in the 611 sites. When FICM was used instead of Anderson et al.'s, Model 5 again turned out to be the best model among the alternatives, suggesting that the accelerated transition in human observed before is not an artefact, but is real. The ratio of human $\alpha$ to that of others is estimated to be $3.20 \pm 0.99$ for the 611 sites. Obviously, the transition rate is higher in human than in apes probably by more than 2 times.

In contrast to the obvious accerelation of transition rate in human, there is no indication of higher transversion rate of four-fold degenerate sites in human than in other hominoids. Furthermore, numbers of amino acid differences in the 1344 sites of mtDNA-encoded proteins of gorilla are 61, 61, and 64, respectively, from human, chimpanzee, and bonobo (Table 5.15), indicating no higher amino acid substitution rate in human (Adachi and Hasegawa 1995[4]). On the other hand, it is clear that the transversion rate of four-fold degenerate sites and the amino acid substitution rate are higher in orangutan than in others. The transition rate may also be higher in orangutan than in the African apes, but this is not obvious probably because of saturation. It is noteworthy that the length of $0.078 \pm 0.009$ per site for the 4–orang branch estimated by in section 5.1 from the amino acid sequences is significantly longer than that of $0.054 \pm 0.007$ for 4–siamang, while the estimated length of $0.563 \pm 0.089$ per site for 4–orang from the four-fold degenerate sites is shorter than that of $1.252 \pm 0.158$ for 4–siamang. The ratio of 4–siamang/4–orang for the four-fold degenerate sites is $2.23 \pm 0.45$, and that for the amino acid sequences is $0.69 \pm 0.12$, which

is significantly smaller than the former. If the substitution rate of four-fold degenerate sites represents mutation rate, these observations indicate that mutation rate differs among different lineages and that the extent of constraints operating on proteins also differs among lineages.

### 5.2.6 Discussion

Various hypotheses have been proposed to explain the apparently higher rate of mtDNA evolution in warm-blooded vertebrates than in cold-blooded ones (Thomas and Beckenbach 1989[292]; Adachi et al. 1993[1]; Martin et al. 1992[204]; Martin and Palumbi 1993[205]). It is known that oxygen radicals damage DNA, and oxidative damage is greatest to mtDNA (Richter et al. 1988[246]). Although no single factor cannot explain all variations in rates of mtDNA evolution, species with higher metabolic rates and accordingly with higher content of oxygen radicals are likely to have higher mutation rate of mtDNA (Martin et al. 1992[204]; Martin and Palumbi 1993[205]). The rate difference demonstrated in this work, however, is between the closely related species of human and chimpanzees, with presumably similar metabolic rate, and the mechanism of this difference remains to be studied. In this context, the higher rate of oxygen radical production in rat liver mitochondria than in mouse (Sohal et al. 1990[273]) is interesting. This might explain that rat has a higher evolutionary rate of mtDNA than mouse in spite of larger body size and presumably of lower metabolic rate (Martin and Palumbi 1993[205]).

It is now clear that no universal clock for the evolution of mtDNA can be assumed in phylogenetic analyses, and this underscores the attempt of dating by using the simple clock. The dating is justified only by careful analyses taking account of the possible rate variation among lineages (Kishino and Hasegawa 1990[165]). Even by these analyses, we must take the estimates as approximate, because the model we use is always approximate. Furthermore, there is always ambiguity in calibrating the clock. If the orangutan separation was 16 Myr old which is probably the oldest limit (Pilbeam 1988[241]; Andrews 1992[17]; McCrossin et al. 1993[207]), rather than 13 Myr, the estimate of the human/chimpanzee separation from the 611 sites data becomes 4.3 $\pm$ 1.1 Myr old. Taking account of the analyses of the four-fold degenerate sites and of the amino acid sequences (Adachi and Hasegawa 1995[4], 1995[7]) as well, the overall evidence seems to suggest that the human/chimpanzee separation in the mtDNA tree was some 4–5 Myr old. Although some nuclear genes suggest earlier divergence between human and chimpanzee (Kishino and Haegawa 1990[165]), the discrepancy can be regarded to be due to ancestral polymorphism of the nuclear genes (Hasegawa 1991[105]; Adachi and Hasegawa 1995[4]).

The higher transitional mutation rate of human suggested in this work is closely relevant to the dating of the deepest root of the human mtDNA tree, and is in favour of the recent origin hypothesis of modern humans (Cann et al. 1987[46]; Vigilant et al. 1991[302]; Hasegawa et al. 1993[109]), because, if this suggestion is real, the dating by a constant rate clock with the human/chimpanzee separation as a reference must be an older estimate.

## 5.3 Phylogeny of Whales

### 5.3.1 Dependence of the Inference on Species Sampling

From phylogenetic analyses of the 12S and 16S mitochondrial ribosomal DNA and of myoglobin amino acid sequences, Milinkovitch et al. (1993[212]) proposed the hypothesis that one group of toothed whales (Odontoceti), the sperm whales (Physeteridae), is more closely related to the baleen whales (Mysticeti) than to other alleged odontocetes such as dolphins. This hypothesis is in conflict with the traditional view that the odontocetes form a monophyletic clade (Barnes et al. 1985[37]; Novacek 1993[229]). From an analysis of the cytochrome *b* gene, Árnason and Gullberg (1994[22]) recently challenged Milinkovitch et al.'s hypothesis as well as the traditional tree, claiming that the mysticetes are closer to the dolphins rather than to the sperm whales. They used the cow as the only outgroup and the giant sperm whale as the only representative of Physeteridae, but the estimated tree may depend on the sampled species (Lecointre et al. 1993[186]; Cao et al. 1994[50]). By including many alternative artiodactyl outgroups (Irwin et al. 1991[146]) in their cytochrome *b* dataset, I will show that Árnason and Gullberg's conclusion is shaky, and that the overall evidence favours Milinkovitch et al.'s hypothesis. The content of this subsection appeared in Adachi and Hasegawa (1995[6]).

Fig. 5.7 illustrates the three competing hypotheses on the relationships among mysticetes, sperm whales and dolphins; i.e., (1) Milinkovitch's tree in which sperm whale is closer to mysticetes rather than to dolphins, (2) the traditional tree of Odontoceti monophyly, and (3) Árnason's tree in which dolphins are closer to mysticetes than to sperm whales.

The ProtML program based on the JTT model is used in this analysis. Fig. 5.8 shows that the inferred tree is highly sensitive to the choice of the outgroup species, and although Árnason's tree is favoured when only the cow is used as an outgroup (with 70% bootstrap probability), this does not necessarily hold when other species are used. For example, when goat is used as an outgroup, Árnason's tree is least supported among the alternative trees only with 11% bootstrap probability. It is preferable to cut long branches on the tree (hence, allowing for a better polarization of characters) by using at least two divergent outgroup species. When two species are used as outgroups, Árnason's tree is favoured only in 3 out of 24 cases, while Milinkovitch's tree is favoured in 14 cases and the traditional tree is favoured in 7 cases. Furthermore, in 18 out of 24 cases the bootstap probability of Árnason's tree is the lowest of the three alternatives. Consequently, Árnason and Gullberg's (1994[22]) hypothesis is the least likely among the alternatives as far as their cytochrome *b* data are concerned.

Furthermore, Milinkovitch et al. (1995[213]) showed that transition differences are saturated in Árnason and Gullberg's cytochrome *b* data set, and that, when only transversions are taken into account, Árnason and Gullberg's data support the Milinkovitch tree.

It must be noted that, although Milinkovitch et al.'s (1993[212]) analysis included only two rorquals as baleen-whales representatives, Milinkovitch et al. did not suggest paraphyly of mysticetes as Árnason

and Gullberg (1994[22]) misinterpret, but suggested a sister relationship between sperm whales and all mysticetes. Although more data are obviously needed to rule out the traditional tree of Odontoceti monophyly, the cytochrome *b* data favours Milinkovitch's tree consistently with the 12S, 16S and myoglobin data. Many species included in Árnason and Gullberg's data set are very closely related baleen whales, hence, they are poorly informative for testing Milinkovitch et al.'s hypothesis. This study demonstrates that an argument based on a small data set with respect to the number of relevant species may be unstable (Lecointre et al. 1993[186]; Cao et al. 1994[50]).

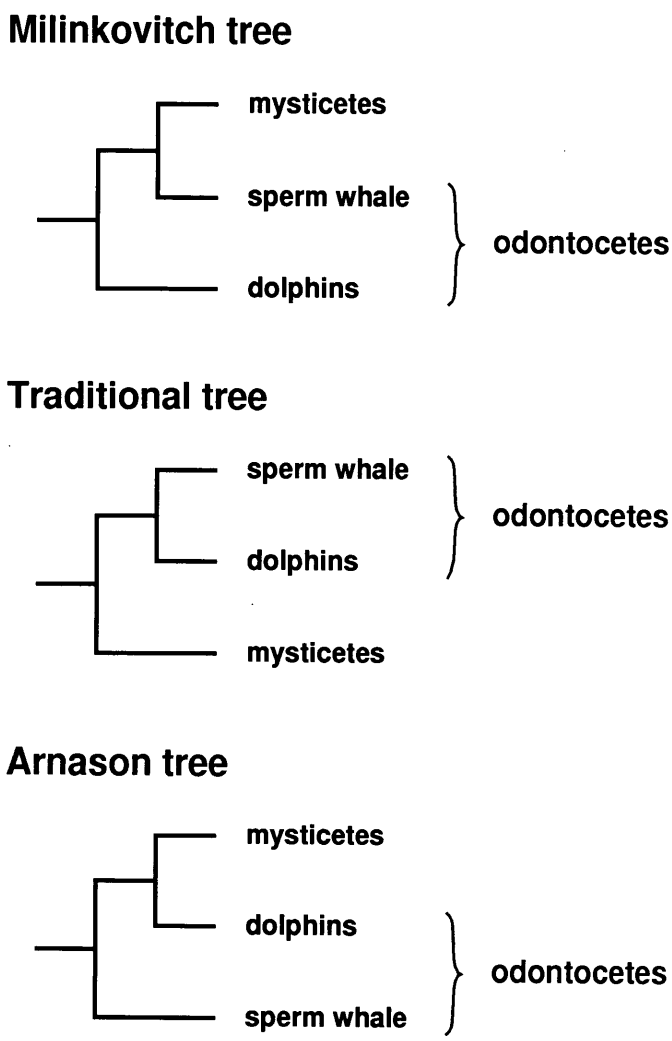



Figure 5.7: Three competing phylogenies of whales.

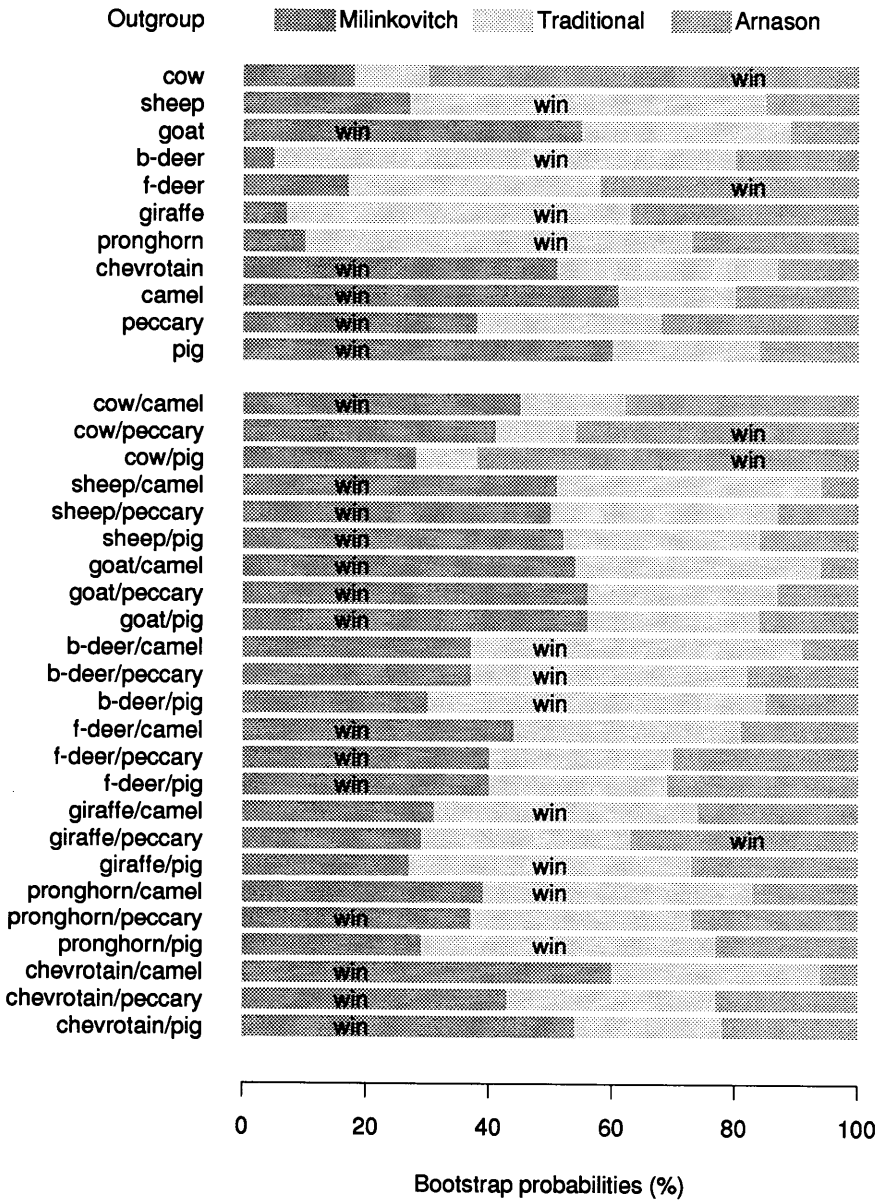



Figure 5.8: Bootstrap probabilities.

Analysis of the cytochrome *b* amino acid sequences (Árnason and Gullberg 1994[22]; Irwin et al. 1991[146]) by the maximum likelihood program ProtML based on the JTT model. Bootstrap probabilities estimated by the RELL method (Kishino et al. 1990[166]; Hasegawa and Kishino 1994[119]) with $10^4$ replications are shown. The same data set from whales as in Árnason and Gullberg (1994[22]) was used. Within the Mysticeti, ((bowhead, right), pygmy right, ((Antarctic minke, N.Atlantic minke), grey, humpback, blue, fin, (sei, Bryde's whale))) was assumed according to the local ProtML analysis, and this is compatible with the Mysticeti part of Fig. 5.7 in Árnason and Gullberg. Firstly, one species was chosen as an outgroup from cow, sheep, goat, black-tailed deer (b-deer), fallow deer (f-deer), giraffe, pronghorn, chevrotain, camel, peccary and pig, and secondly, two divergent species, one from cow, sheep, goat, b-deer, f-deer, giraffe, pronghorn and chevrotain, and the other from camel, peccary and pig, were chosen.

### 5.3.2 Further Study of Cetacean Phylogeny by Partial Cytochrome b Sequences

Milinkovitch et al. (1994[211]) reinforced their hypothesis[212] by using partial cytochrome *b* gene sequences (424bp) from several whale species listed in Table 5.17.

In this subsection, I will reanalyze their data by using the ProtML program. I will analyze 20 whale species with three artiodactyles (cow, camel, and peccary) as an outgroup. At first, by carrying out several analyses of local as well as overall groups and by using conventional taxonomy, I confirmed that the 20 whale species listed in Table 5.17 can be classified into 7 groups as follows;

Group 1: Phocoenidae (*Phocoena phocoena, P. spinipinnis*)

Group 2: Delphinidae

(*Cephalorhynchus eutropia, Delphinus delphis, Lagenorhynchus albirostris, L. obscurus, Lissodelphis peronii, Globicephala melaena*)

Since the branching order within Group 2 cannot be resolved, multifurcation was assumed in this group.

Group 3: *Delphinapterus leucas*

Group 4: Ziphiidae

((*Mesoplodon europaeus, M. peruvianus*), *Ziphius cavirostris)*

Group 5: Physeteridae

((*Physeter catodon* Pcato1, Pcato2), (*Kogia breviceps, K. simus*))

Group 6: Mysticeti

((*Balaenoptera physalus, Megaptera novaeangliae*), *Eschrichtiust robustus, Balaena mysticetus*).

Trifurcation among (*B. physalus, M. novaeangliae*), *E. robustus*, and *B. mysticetus* was assumed, because local ML analysis did not resolve the branching order among them.

Outgroup:

Bovine, camel, and peccary are used as outgroup species, and trifurcation is assumed among them.

There are 945 possible trees that link these 6 groups and the outgroup, and these trees were examined by the ProtML program with the JTT-F model. The highest likelihood tree is given in Fig. 5.9, and it links Mysticeti (Group 6) with Physeteridae (Group 5) excluding other Odontoceti as outgroups. Subtotal of bootstrap probabilities of trees with the Mysticeti-Physeteridae clade excluding other whales is 60%(Table 5.18). The ProtML tree coincides with Milinkovitch et al.'s (1994[211]) tree except linking Ziphiidae with the Mysticeti-Physeteridae clade. Subtotal of bootstrap probabilities of trees with the Mysticeti-Physeteridae- Ziphiidae clade is 64%. Monophyly of Odontoceti is unlikely from this analysis with the bootstrap probability of as low as 4%.

The relationship among Phocoenidae (Group 1), Delphinidae (Group 2), and *Delphinapterus* (Group 3) is uncertain, and left unresolved in Figure 5.9.

In the ML tree of Fig. 5.9, Ziphiidae (the beaked whales) is a sister group to the Physeteridae/Mysticeti clade excluding Delphinoidea (Phocoenidae + Delphinidae + Monodontidae) as an outgroup. In the trees shown in Milinkovitch et al. (1993[212], 1994[211]), Ziphiidae is represented as the sister group to all other whales, although the authors noted that such a relationship is of no statistical significance. Milinkovitch et al. (1994[211]) also noted that, some authors classify the beaked whales as the sister group to sperm whales in the superfamily Physeteroidea (Barnes 1984[36]), and that depending on the method or the variations of analyses, Ziphiidae is positioned at the base of cetaceans, at the base of the Delphinoidea/*Inia* clade, or with the Physeteridae/Mysticeti clade. The grouping of Ziphiidae with the Physeteridae/Mysticeti clade is favoured by myoglobin data (Milinkovitch et al. 1993[212]) and possibly by one morphological trait, the throat grooves, found only in sperm, baleen and beaked whales (Milinkovitch et al. 1994[211]). It might be noteworthy that, in my ProtML analysis, a Mysticeti-Physeteridae-Ziphiidae clade is supported with 64% BP. Although none of the alternative hypotheses on the place of the beaked whales is strongly supported and hence the issue is still open, our ML analysis of the cytochrome *b* data favours their grouping with the sperm and baleen whales.

In Table 5.18, BPs estimated from the best 100 trees by the approximate likelihood criterion (Section 3.6) are also shown, and it turned out that they are good approximations of those estimated by using all the 945 trees.

Table 5.17: List of partial cytochrome b sequences (134 amino acids) of Cetacea

| species name | | classification | database |
|---|---|---|---|
| *Phocoena phocoena* | harbour porpoise | Phocoenidae | U13143 |
| *Phocoena spinipinnis* | Burmeister's porpoise | Phocoenidae | U13144 |
| *Cephalorhynchus eutropia* | black dolphin | Delphinidae | U13128 |
| *Delphinus delphis* | common dolphin | Delphinidae | U13129 |
| *Tursiops truncatus* | bottle-nosed dolphin | Delphinidae | U13145 |
| *Lagenorhynchus albirostris* | white-beaked dolphin | Delphinidae | U13136 |
| *Lagenorhynchus obscurus* | dusky dolphin | Delphinidae | U13137 |
| *Lissodelphis peronii* | southern right whale dolphin | Delphinidae | U13138 |
| *Globicephala melaena* | long-finned pilot whale | Delphinidae | U13132 |
| *Delphinapterus leucas* | beluga | Monodontidae | U13130 |
| *Mesoplodon europaeus* | Gervais' beaked whale | Ziphiidae | U13139 |
| *Mesoplodon peruvianus* | Peruvian beaked whale | Ziphiidae | U13141 |
| *Ziphius cavirostris* | goose-beaked whale | Ziphiidae | U13146 |
| *Physeter catodon* | sperm whale | Physeteridae | U13142 |
| *Kogia breviceps* | pygmy sperm whale | Physeteridae | U13134 |
| *Kogia simus* | dwarf sperm whale | Physeteridae | U13135 |
| *Balaenoptera physalus* | finback whale | Balaenopteridae | U13126 |
| *Megaptera novaeangliae* | humpback whale | Balaenopteridae | U13140 |
| *Eschrichtius robustus* | gray whale | Eschrichtiidae | U13131 |
| *Balaena mysticetus* | bowhead whale | Balaenidae | U13125 |

Milinkovitch et al. (1994[211]).

Table 5.18: Bootstrap probabilities of cetacean clade estimated from the partial cytochrome *b* gene sequences

| clade | among 945 trees | among best 100 trees |
|---|---|---|
| (Mysticeti, Physeteridae, Ziphiidae) clade | 0.6437 | 0.6857 |
| Mysticeti/Physeteridae clade | 0.5952 | 0.6167 |
| Physeteridae/Ziphiidae clade | 0.3065 | 0.2987 |
| Mysticeti/Ziphiidae clade | 0.0342 | 0.0361 |
| Monophyly of Odontoceti | 0.0448 | 0.0465 |

Bootstrap probabilities estimated by the RELL method with $10^4$ replications (the JTT-F model). 'Among best 100 trees' means BP among the best 100 trees in the approximate likelihood criterion.

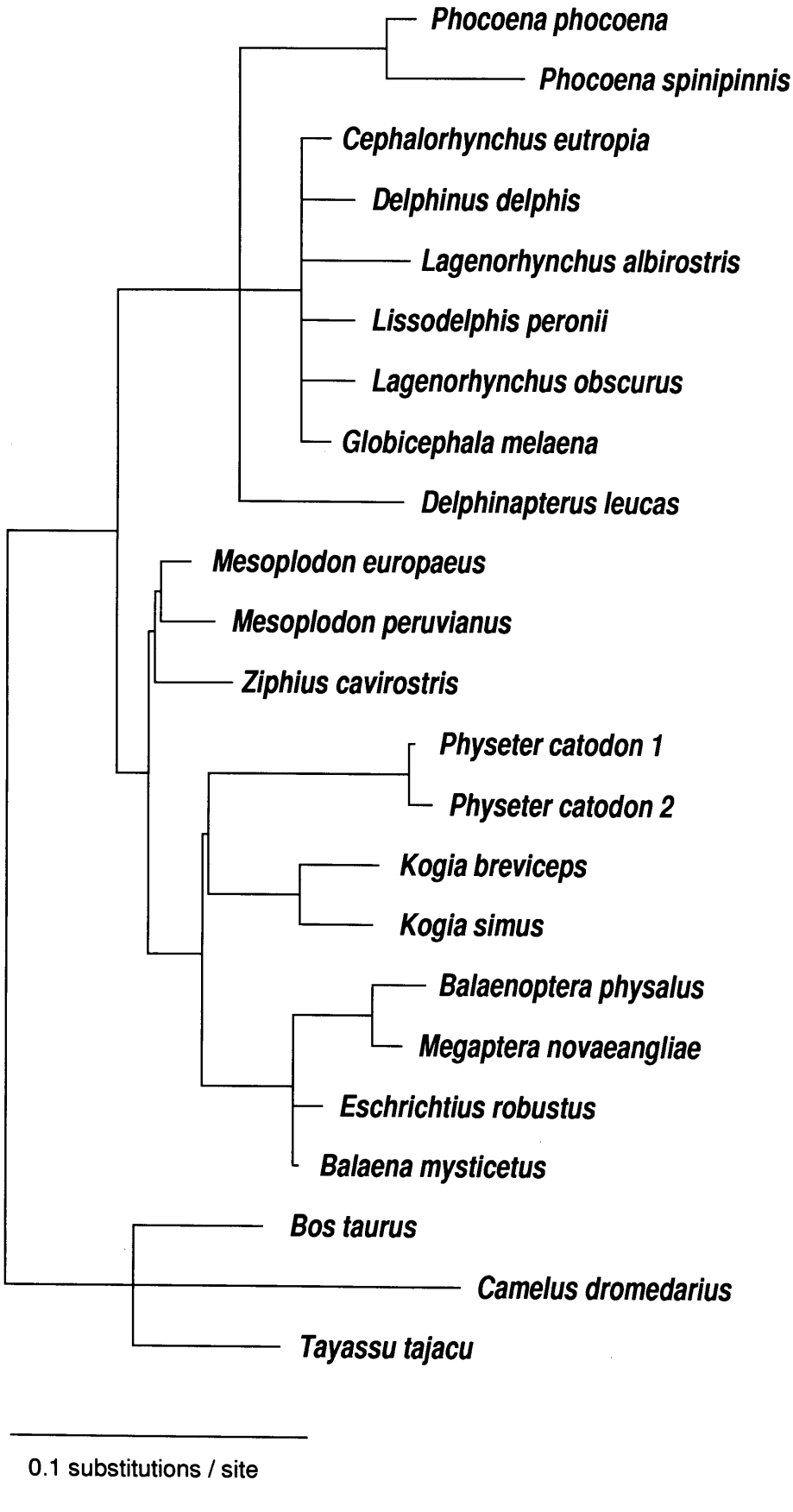



Figure 5.9: The ProtML tree of partial cytochrome $b$ sequences from whales.

## 5.4 Quartet Analyses of Molecular Sequence Data by the ML Method

Using both the maximim parsimony (MP) and neighbor joining (NJ) methods of molecular phylogeny, Philippe and Douzery (1994[240]) demonstrated that it is possible to find a quartet of species which provides a high bootstrap proportion (or bootstrap probability; BP) for each of the three possible trees among Ruminantia, Suiformes, Cetacea, and other outgroup species of mammals. Since their finding poses a serious problem in molecular phylogenetics, it seems to be necessary to examine how much extent the instability of estimated trees of four taxa holds for the maximum likelihood (ML) method, another method of molecular phylogenetics not examined by them. Therefore, by using the ML method, I carry out a similar analysis as that of Philippe and Douzery, and discuss the problem posed by them. I take only the cytochrome *b* (Irwin et al 1991[146]; Ma et al. 1993[199]) encoded by mtDNA as an example among the data used by Philippe and Douzery, and analyze the data at the amino acid sequence level by the ProtML program, and assume the JTT-F model for amino acid substitution. The BP values are estimated by the RELL method (Kishino et al. 1990[166]). This method is a good approximation to the computationally intensive bootstrap method (Felsenstein 1985[80]) in estimating the BP (Hasegawa and Kishino 1994[119]).

As was done by Philippe and Douzery, 1280 combinations of four taxa were generated by taking one species per group (among 8 species in Ruminantia, 2 in Suiformes, 4 in Cetacea, and 20 in outgroup). Among these combinations, and for each of the three possible trees, we selected the highest BP to summarize the results (Table 5.19). When Ruminantia are represented by *Dama dama*, Suiformes by *Sus scrofa*, Cetacea by *Stenella longirostris* 1a, and outgroup by *Diceros bicornis*, the Artiodactyla monophyly (Tree-1) is supported in 99.4% of the replicates. When respective groups are represented by *Odocoileus hemionus*, *Tayassu tajacu*, *Balaenoptera physalus*, and *D. bicornis*, the Ruminantia/Cetacea grouping (Tree-2), which is advocated by Graur and Higgins (1994[100]), is supported in 94.4%. Futhermore, when they are represented by *Tragulus napu*, *T. tajacu*, *Stenella attenuata*, and *Rattus norvegicus* 1, the Suiformes/Cetacea grouping (Tree-3) is supported in 94.7% of the replicates. Thus, for each of the three possible trees, we can find a quartet of species which provides a high BP even by using the ML method.

However, if Tree-1 is the real tree as the traditional taxonomy assumes, our result is reasonable and does not pose such a serious problem in applying the ML method to molecular phylogeny. It is true that, if we happen to sample *O. hemionus*, *T. tajacu*, *B. physalus* and *D. bicornis*, we may tend to accept the putatively erroneous Tree-2 as the real tree. However, rejection of the putatively correct Tree-1 is not significant at the 5% level (log-likelihood difference is less than 2SEs) even for this extreme case. Furthermore, if we sample species randomly, the probability of getting a combination of species which supports the putatively erroneous tree with the 5% significance level would be much smaller than 5%. Actually, frequencies of getting higher BP values than 95% for Tree-2 and Tree-3 are null out of 1280

Table 5.19: The highest BP combinations of four species.

| Ruminantia | Suiformes | Cetacea | outgroup | Tree-1 Rumi./Suif. | Tree-2 Rumi./Ceta. | Tree-3 Suif./Ceta. |
|---|---|---|---|---|---|---|
| *Dama dama* | *Sus scrofa* | *Stenella longirostris* 1a | *Diceros bicornis* | ML (0.9942) | −22.3 ± 9.8 (0.0056) | −23.4 ± 9.4 (0.0002) |
| *Odocoileus hemionus* | *Tayassu tajacu* | *Balaenoptera physalus* | *Diceros bicornis* | −12.5 ± 8.2 (0.0554) | ML (0.9438) | −14.5 ± 7.3 (0.0008) |
| *Tragulus napu* | *Tayassu tajacu* | *Stenella attenuata* | *Rattus norvegicus* 1 | −12.4 ± 8.0 (0.0383) | −13.3 ± 7.8 (0.0152) | ML (0.9465) |

For the 1280 combinations of four species, the maximum BPs of ML trees of cytochrome *b* are selected for each of the three possible tree topologies. Cytochrome *b* was treated through amino acid sequences, and the ML analyses were carried out based on the JTT-F model for amino acid substitution. The highest likelihood tree is indicated as 'ML', and the differences in log-likelihood of alternative trees from that of the ML tree are shown with their SE (following ±) which were estimated by Kishino and Hasegawa's (1989[164]) formula. The BPs given in parentheses were estimated by the RELL method with $10^4$ replications.

cases (Fig. 5.10). Thus, our analysis of the cytochrome *b* data supports Tree-1. Although Philippe and Douzery's finding of the instability of quartet analysis is highly important in molecular phylogenetics, it may not be so serious as they imagine as far as the ML method is used and the confidence level of the inferred tree is estimated adequately.

The dependence of the inferred tree on species sampling has been studied extensively (Lecointre et al. 1993[186]; Cao et al. 1994[50]; Adachi and Hasegawa 1995[6]), and it is clearly dangerous to accept a tree inferred from a single gene. By using complete sequence data of mitochondrial genomes from 6 eutherian species with opossum, chicken and xenopus as outgroups, Janke et al. (1994[150]) and Cao et al. (1994[49]) showed that the relation of (Rodentia, (Primates, (Carnivora, (Artiodactyla, Cetacea)))) is highly likely to be the case. However, analyses of individual genes of mitochondria do not necessarily conform to this conclusion, and some of the genes reject the putatively correct tree with nearly 5% significance. Cao et al. (1994[49]) argued that, if the branching among the orders in question occurred within a short period, such a discrepancy can occur in 5% of the cases. Due to uncertainty about the assumed model underlying the phylogenetic inference, this can occur even more frequently. In this situation, in order to avoid the danger of concluding an erroneous tree, it is important to carry out analyses based on as many different genes as possible, and to synthesize the results (Kishino and Hasegawa 1989[164]; Hasegawa et al. 1992[108]; Cao et al. 1994[50], 1994[49]). The ML and MP methods have merit in allowing the results from different genes to be summed, while it is not attainable by the NJ method.

As Philippe and Douzery showed, it is now clear that an argument based on a quartet analysis of a single gene is very dangerous. Furthermore, even if many ingroup species are included, when only single

species is used as an outgroup as in Árnason and Gullberg's (1994[22]) analysis, the conclusion may be unreliable (subsection 5.3.1; Adachi and Hasegawa 1995[6]).

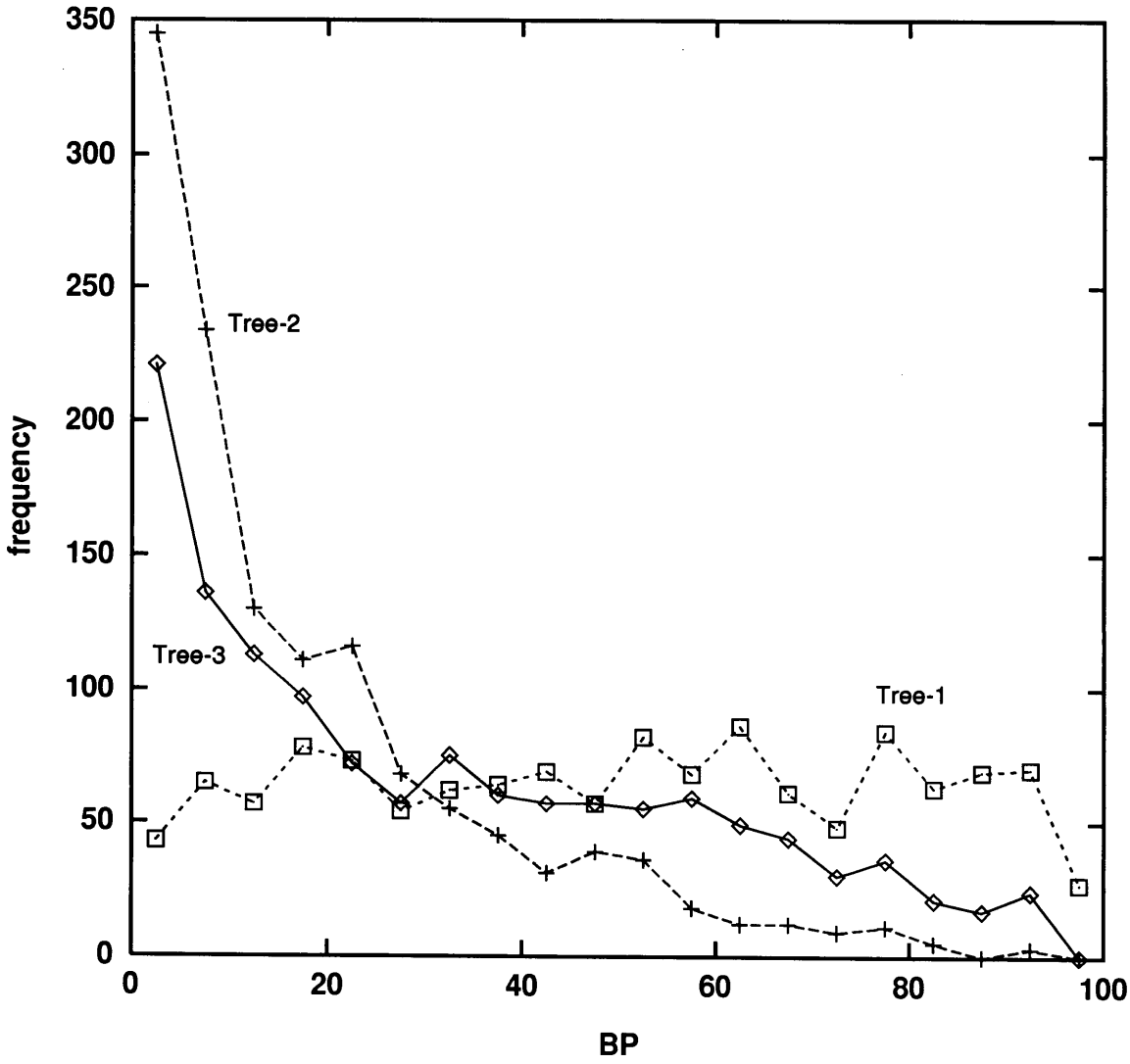



Figure 5.10: Frequencies of BPs for each trees.

Frequencies of BPs for each trees among the 1280 combinations of four species. Plots at $x$ show frequencies of species combinations that give BP values between $x - 5\%$ to $x\%$ for respective trees.

## 5.5 Phylogenetic Analyses by Using MtDNA-encoded Proteins

From the discussions in the preceding section, it is now clear that a phylogenetic conclusion based on a small number of relevant species is often unstable. Therefore, in this section, I will analyze simultaneously as many species as possible by using two mtDNA-encoded proteins; cytochrome *b* and cytochrome oxidase subunit II.

### 5.5.1 Cytochrome b

Cytochrome *b* is one of the most widely used molecular marker in phylogenetic studies. In this subsection, I will study several phylogenetic problems of vertebrates by using this molecule.

#### Sequence Data

Sequence data used in the phylogenetic analyses are listed below, where classification is based on the traditional taxonomy (Corbert and Hill 1991[59]; Yamashina 1986[314]).

| Abbrev. | Species name | Common name | Reference | Database |
|---|---|---|---|---|
| I. Class Mammalia | | | | |
| I-1. Artiodactyla | | | | |
| Bosta1 | *Bos taurus* | Domestic cow | Anderson'82[16] | V00654 |
| Bosta2 | *Bos taurus* | Domestic cow | Kikkawa (unpubl.)[158] | D34635 |
| Bosja | *Bos javanicus* | Banteng | Kikkawa (unpubl.)[158] | D34636 |
| Bubbu1 | *Bubalus bubalis* | Asian water buffalo | Kikkawa (unpubl.)[159] | D34637 |
| Bubbu2 | *Bubalus bubalis* | Asian water buffalo | Kikkawa (unpubl.)[159] | D34638 |
| Budtb | *Budorcas taxicolor bedfordi* | Golden takin | Groves (unpubl.)[102] | U17867 |
| Budtt | *Budorcas taxicolor taxicolor* | Mishmi takin | Groves (unpubl.)[102] | U17868 |
| Capcr | *Capricornis crispus* | Japanese serow | Chikuni'94[56] | D32191 |
| Nemca | *Nemorhaedus caudatus* | Chinese goral | Groves (unpubl.)[102] | U17861 |
| Ovimo | *Ovibos moschatus moschatus* | Muskox | Groves (unpubl.)[102] | U17862 |
| Oviar | *Ovis aries* | Domestic sheep | Irwin'91[146] | X56284 |
| Caphi | *Capra hircus* | Domestic goat | Irwin'91[146] | X56289 |
| Cerni | *Cervus nippon* | Sika deer | Chikuni'94[56] | D32192 |
| Odohe | *Odocoileus hemionus* | Black-tailed deer | Irwin'91[146] | X56291 |
| Damda | *Dama dama* | Fallow deer | Irwin'91[146] | X56290 |
| Girca | *Giraffa camelopardalis* | Giraffe | Irwin'91[146] | X56287 |
| Antam | *Antilocapra americana* | Pronghorn | Irwin'91[146] | X56286 |
| Trana | *Tragulus napu* | Greater Malay chevrotain | Irwin'91[146] | X56288 |
| Traja | *Tragulus javanicus* | Lesser Malay chevrotain | Chikuni (unpubl.)[55] | D32189 |
| Camdr1 | *Camelus dromedarius* | One-humped camel | Irwin'91[146] | X56281 |
| Camdr2 | *Camelus dromedarius* | One-humped camel | Stanley'94[276] | U06426 |
| Camba | *Camelus bactrianus* | Two-humped camel | Stanley'94[276] | U06427 |
| Lamgu | *Lama guanicoe* | Guanaco | Stanley'94[276] | U06428 |
| Lamgl | *Lama glama* | Llama | Stanley'94[276] | U06429 |
| Lampa | *Lama pacos* | Alpaca | Stanley'94[276] | U06425 |
| Vicvi | *Vicugna vicugna* | Vicuna | Stanley'94[276] | U06430 |
| Hipam | *Hippopotamus amphibius* | Hippopotamus | Irwin'94[145] | U07565 |
| Tayta | *Tayassu tajacu* | Collared peccary | Irwin'91[146] | X56296 |
| Sussc | *Sus scrofa* | Pig | Irwin'91[146] | X56295 |
| I-2. Cetacea | | | | |
| Stelo | *Stenella longirostris* | Long-beaked dolphin | Irwin'91[146] | X56293 |

| | | | | |
|---|---|---|---|---|
| Steat | *Stenella attenuata* | Narrow-snouted dolphin | Irwin'91[146] | X56294 |
| Phyma | *Physeter macrocephalus* | Sperm whale | Arnason'94[22] | X75589 |
| Balph | *Balaenoptera physalus* | Fin whale | Arnason'91[24] | X61145 |
| Balmu | *Balaenoptera musculus* | Blue whale | Arnason'93[21] | X72204 |
| Balac | *Balaenoptera acutorostrata* | Minke whale | Arnason'94[22] | X75753 |
| Balbon | *Balaenoptera bonaerensis* | Antarctic minke whale | Arnason'94[22] | X75581 |
| Balbor | *Balaenoptera borealis* | Sei whale | Arnason'94[22] | X75582 |
| Baled | *Balaenoptera edeni* | Bryde's whale | Arnason'94[22] | X75583 |
| Megno | *Megaptera novaeangliae* | Humpback whale | Arnason'94[22] | X75584 |
| Escro | *Eschrichtius robustus* | California gray whale | Arnason'94[22] | X75585 |
| Balmy | *Balaena mysticetus* | Bowhead whale | Arnason'94[22] | X75588 |
| Balgl | *Balaena glacialis* | Right whale | Arnason'94[22] | X75587 |
| Capma | *Caperea marginata* | Pygmy right whale | Arnason'94[22] | X75586 |
| I-3. Pinnipedia | | | | |
| Phovi1 | *Phoca vitulina* | Harbor seal | Arnason'92[25] | X63726 |
| Phovi2 | *Phoca vitulina* | Harbor seal | Arnason'95[20] | X82306 |
| Phofa | *Phoca fasciata* | Ribbon seal | Arnason'95[20] | X82302 |
| Phola | *Phoca largha* | Spotted seal | Arnason'95[20] | X82305 |
| Phohi | *Phoca hispida* | Ringed seal | Arnason'95[20] | X82304 |
| Phogr | *Phoca groenlandica* | Harp seal | Arnason'95[20] | X82303 |
| Halgr | *Halichoerus grypus* | Grey seal | Arnason'93[23] | X72004 |
| Eriba | *Erignathus barbatus* | Bearded seal | Arnason'95[20] | X82295 |
| Hydle | *Hydrurga leptonyx* | Leopard seal | Arnason'95[20] | X82297 |
| Monsc | *Monachus schauinslandi* | Hawaiian monk seal | Arnason'95[20] | X72209 |
| Cyscr | *Cystophora cristata* | Hooded seal | Arnason'95[20] | X82294 |
| Mirle | *Mirounga leonina* | Southern elephant seal | Arnason'95[20] | X82298 |
| Arcga | *Arctocephalus gazella* | Antarctic fur seal | Arnason'95[20] | X82292 |
| Arcfo | *Arctocephalus forsteri* | New Zealand fur seal | Arnason'95[20] | X82293 |
| Zalca | *Zalophus californianus* | California sea lion | Arnason'95[20] | X82310 |
| Eumju | *Eumetopias jubatus* | Northern sea lion | Arnason'95[20] | X82311 |
| Odoro | *Odobenus rosmarus rosmarus* | Atlantic walrus | Arnason'95[20] | X82299 |
| I-4. Carnivora | | | | |
| Ursam | *Ursus americanus* | American black bear | Arnason'95[20] | X82307 |
| Ursar | *Ursus arctos* | Brown bear | Arnason'95[20] | X82308 |
| Ursma | *Ursus maritimus* | Polar bear | Arnason'95[20] | X82309 |
| Feldo | *Felis domesticus* | Domestic cat | Arnason'95[20] | X82296 |
| Panle | *Panthera leo* | Lion | Arnason'95[20] | X82300 |
| Panti | *Panthera tigris* | Tiger | Arnason'95[20] | X82301 |
| I-5. Perissodactyla | | | | |
| Equca | *Equus caballus* | Domestic horse | Xu'94[313] | X79547 |
| Equgr | *Equus grevyi* | Grevy's zebra | Irwin'91[146] | X56282 |
| Dicbi | *Diceros bicornis* | Black rhinoceros | Irwin'91[146] | X56283 |
| I-6. Rodentia | | | | |
| Musmu | *Mus musculus* | House mouse | Bibb'81[39] | P00158 |
| Ratno | *Rattus norvegicus* | Common rat | Gadaleta'89[87] | P00159 |
| Papbu | *Pappogeomys bulleri* | Buller's pocket gopher | DeWalt'93[66] | L11900 |
| Geobu | *Geomys bursarius juggosicularis* | Plains pocket gopher | DeWalt'93[66] | L11901 |
| Craca | *Cratogeomys castanops castanops* | Yellow-faced pocket gopher | DeWalt'93[66] | L11902 |
| Crafu | *Cratogeomys fumosus* | Smoky pocket gopher | DeWalt'93[66] | L11903 |
| Crago | *Cratogeomys goldmani goldmani* | Goldman's pocket gopher | DeWalt'93[66] | L11904 |
| Cragy | *Cratogeomys gymnurus* | Llano pocket gopher | DeWalt'93[66] | L11905 |
| Crame | *Cratogeomys merriami* | Merriam's pocket gopher | DeWalt'93[66] | L11906 |
| Craru | *Cratogeomys goldmani rubellus* | | DeWalt'93[66] | L11907 |
| Crata | *Cratogeomys castanops* | | DeWalt'93[66] | L11908 |

| | | | | |
|---|---|---|---|---|
| | *tamaulipensis* | | | |
| Craty | *Cratogeomys tylorhinus* | Taylor's pocket gopher | DeWalt'93[66] | L11909 |
| Scini | *Sciurus niger* | Eastern fox squirrel | Wettstein'95[307] | U10180 |
| Sciab | *Sciurus aberti* | Abert squirrel | Wettstein'95[307] | U10163 |
| Speri | *Spermophilus richardsonii* | Richardson's ground squirrel | Thomas'93[293] | S73150 |
| Hysaf | *Hystrix africaeaaustralis* | African porcupine | Ma'93[199] | X70674 |
| Cavpo | *Cavia porcellus* | Guinea pig | Ma'93[199] | |
| I-7. Lagomorpha | | | | |
| Orycu | *Oryctolagus cuniculus* | Rabbit | Irwin'94[145] | U07566 |
| I-8. Proboscidea | | | | |
| Loxaf | *Loxodonta africana* | African elephant | Irwin'91[146] | X56285 |
| I-9. Sirenia | | | | |
| Dugdu | *Dugong dugong* | Dugong | Irwin'94[145] | U07564 |
| I-10. Primates | | | | |
| Europ | *Homo sapiens* | European | Anderson'81[15] | J01415 |
| Japan | *Homo sapiens* | Japanese (DCM1) | Ozawa'91[232] | |
| Afric | *Homo sapiens* | African (SB17F) | Horai'95[140] | D38112 |
| Pantr | *Pan troglodytes* | Chimpanzee | Horai'95[140] | D38113 |
| Panpa | *Pan paniscus* | Bonobo | Horai'95[140] | D38116 |
| Gorgo | *Gorilla gorilla* | Gorilla | Horai'95[140] | D38114 |
| Ponpy | *Pongo pygmaeus* | Orangutan | Horai'95[140] | D38115 |
| I-11. Chiroptera | | | | |
| Chido | *Chiroderma doriae* | | Baker'95[34] | L28937 |
| Chiim | *Chiroderma improvisum* | Guadeloupe white-lined bat | Baker'95[34] | L28938 |
| Chisa | *Chiroderma salvini* | Salvin's white-lined bat | Baker'95[34] | L28939 |
| Chitr | *Chiroderma trinitatum* | Goodwin's bat | Baker'95[34] | L28942 |
| Chivi | *Chiroderma villosum* | Shaggy-haired bat | Baker'95[34] | L28943 |
| Plahe | *Platyrrhinus helleri* | Heller's broad-nosed bat | Baker'95[34] | L28940 |
| Urobi | *Uroderma bilobatum* | Tent-building bat | Baker'95[34] | L28941 |
| I-12. Marsupialia | | | | |
| Didvi | *Didelphis virginiana* | North American opossum | Janke'94[150] | Z29573 |
| Mondo | *Monodelphis domestica* | South American opossum | Ma'93[199] | X70673 |
| Plama | *Planigale maculata sinualis* | Common planigale | Painter (unpubl.)[233] | U10318 |
| Plain | *Planigale ingrami* | Long-tailed planigale | Painter (unpubl.)[233] | U10319 |
| Plate | *Planigale tenuirostris* | Narrow-nosed planigale | Krajewski'94[175] | U07591 |
| Plagi | *Planigale gilesi* | Paucident planigale | Krajewski'94[175] | U07589 |
| Smimu | *Sminthopsis murina* | Dunnart | Krajewski'94[175] | U07594 |
| II. Class Aves | | | | |
| II-1. Galliformes | | | | |
| Galga | *Gallus gallus* | Chicken | Desjardins'90[65] | P18946 |
| Cotco | *Coturnix coturnix* | Japanese quail | Kornegay'93[173] | L08377 |
| Alech | *Alectoris chukar* | Chukar partridge | Kornegay'93[173] | L08378 |
| Pavcr | *Pavo cristatus* | Peafowl | Kornegay'93[173] | L08379 |
| Lopny | *Lophura nycthemera* | Silver pheasant | Kornegay'93[173] | L08380 |
| Melga | *Meleagris gallopavo* | Turkey | Kornegay'93[173] | L08381 |
| Lopga | *Lophortyx gambelii* | Gambel quail | Kornegay'93[173] | L08382 |
| Numme | *Numida meleagris* | Guinea fowl | Kornegay'93[173] | L08383 |
| Ortve | *Ortalis vetula* | Chachalaca | Kornegay'93[173] | L08384 |
| II-2. Anseriformes | | | | |
| Caimo | *Cairina moschata* | Muscovy duck | Kornegay'93[173] | L08385 |
| II-3. Gruiformes | | | | |
| Grurul | *Grus rubicunda* | Brolga | Krajewski'94[174] | U11062 |
| Gruru2 | *Grus rubicunda* | Brolga | Leeton'94[188] | U13622 |
| Gruja | *Grus japonensis* | Manchurian crane | Krajewski'94[174] | U11063 |
| Gruan | *Grus antigone* | Sarus crane | Krajewski'94[174] | U11064 |
| Gruvi | *Grus vipio* | White-naped crane | Krajewski'94[174] | U11065 |
| II-4. Psittaciformes | | | | |

| | | | | |
|---|---|---|---|---|
| Calba | *Calyptorhynchus banksii* | Red-tailed black-cockatoo | Leeton'94[188] | U13620 |
| Geooc | *Geopsittacus occidentalis* | Night parrot | Leeton'94[188] | U13621 |
| Melun | *Melopsittacus undulatus* | Budgeriger | Leeton'94[188] | U13623 |
| Pezwa | *Pezoporus wallicus* | Ground parrot | Leeton'94[188] | U13625 |
| Plaix | *Platycercus icterotis xanthogenis* | Western rosella | Leeton'94[188] | U13626 |
| Polan | *Polytelis anthopeplus westralis* | Regent parrot | Leeton'94[188] | U13627 |
| Strha | *Strigops habroptilis* | Kakapo | Leeton'94[188] | U13628 |
| II-5. Piciformes | | | | |
| Colru | *Colaptes rupicola* | Andean flicker | Edwards'91[72] | X60949 |
| II-6. Passeriformes | | | | |
| Empmi | *Empidonax minimus* | Least flycatcher | Helm-Bychowski'93[137] | X74251 |
| Scyma | *Scytalopus magellanicus* | Andean tapaculo | Edwards'91[72] | X60945 |
| Thrdo | *Thripophaga dorbignyi* | Creamy-breasted canastero | Edwards'91[72] | X60946 |
| Ampst | *Ampelion stresemanni* | White-cheeked cotinga | Edwards'91[72] | X60947 |
| Pitso | *Pitta sordida* | Hooded pitta | Edwards'91[72] | X60948 |
| Pomte | *Pomatostomus temporalis* | Grey-crowned babbler | Edwards'91[72] | X60936 |
| Pomru | *Pomatostomus ruficeps* | Chestnut-crowned babbler | Edwards'91[72] | X60937 |
| Pomis | *Pomatostomus isidori* | Rufous babbler | Edwards'91[72] | X60938 |
| Ambma | *Amblyornis macgregoriae* | MacGregor's bowerbird | Edwards'91[72] | X60940 |
| Epial | *Epimachus albertisii* | Buff-tailed sicklebill | Edwards'91[72] | X60941 |
| Ptipl | *Ptiloprora plumbea* | Leaden honeyeater | Edwards'91[72] | X60943 |
| Gymti | *Gymnorhina tibicen* | Australian magpie | Edwards'91[72] | X60942 |
| Parin | *Parus inornatus* | Plain titmouse | Edwards'91[72] | X60944 |
| Catgu1 | *Catharus guttatus* | Hermit thrush | Edwards'91[72] | X60939 |
| Catgu2 | *Catharus guttatus* | Hermit thrush | Helm-Bychowski'93[137] | X74261 |
| Ailme | *Ailuroedus melanotus* | Spotted catbird | Helm-Bychowski'93[137] | X74257 |
| Cyacr | *Cyanocitta cristata* | Blue jay | Helm-Bychowski'93[137] | X74258 |
| Dipma | *Diphyllodes magnificus* | Magnificent bird of paradise | Helm-Bychowski'93[137] | X74255 |
| Epifa | *Epimachus fastuosus* | Black sicklebill | Helm-Bychowski'93[137] | X74253 |
| Lanlu | *Lanius ludovicianus* | Loggerhead shrike | Helm-Bychowski'93[137] | X74259 |
| Manke | *Manucodia keraudrenii* | Trumpet bird | Helm-Bychowski'93[137] | X74252 |
| Ptipa | *Ptiloris paradiseus* | Paradise riflebird | Helm-Bychowski'93[137] | X74254 |
| Ptivi | *Ptilonorhynchus violaceus* | Satin bowerbird | Helm-Bychowski'93[137] | X74256 |
| Virol | *Vireo olivaceus* | Red-eyed vireo | Helm-Bychowski'93[137] | X74260 |
| II-7. Falconiformes | | | | |
| Tortr | *Torgos tracheliotus* | Lappet-faced vulture | Avise'94[29] | U08934 |
| Neope | *Neophron percnopterus* | Egyptian vulture | Avise'94[29] | U08942 |
| Gypba | *Gypaetus barbatus* | Lammergeier | Avise'94[29] | U08943 |
| Vulgr | *Vultur gryphus* | Andean condor | Avise'94[29] | U08944 |
| Catbu | *Cathartes burrovianus* | Lesser yellow-headed vulture | Avise'94[29] | U08945 |
| Corat | *Coragyps atratus* | Black vulture | Avise'94[29] | U08946 |
| Gymca | *Gymnogyps californianus* | California condor | Avise'94[29] | U08947 |
| II-8. Ciconiiformes | | | | |
| Scoum | *Scopus umbretta* | Hammerkop | Avise'94[29] | U08936 |
| Balre | *Balaeniceps rex* | Whale-headed stork | Avise'94[29] | U08937 |
| Mycib | *Mycteria ibis* | Yellow-billed stork | Avise'94[29] | U08948 |
| Mycam | *Mycteria americana* | American wood ibis | Avise'94[29] | U08949 |
| Lepcr | *Leptoptilos crumeniferus* | Marabou stork | Avise'94[29] | U08950 |
| Jabmy | *Jabiru mycteria* | Jabiry | Avise'94[29] | U08951 |
| Plaal | *Platalea alba* | African spoonbill | Avise'94[29] | U08941 |
| II-9. Pelecaniformes | | | | |
| Peler | *Pelecanus erythrorhynchus* | American white pelican | Avise'94[29] | U08938 |
| II-10. Phoenicopteriformes | | | | |
| Phoru | *Phoenicopterus ruber* | Greater flamingo | Avise'94[29] | U08940 |
| II-11. Cuculiformes | | | | |
| Cocam | *Coccyzus americanus* | Yellow-billed cuckoo | Avise'94[30] | U09265 |

| | | | | |
|---|---|---|---|---|
| Cocer | *Coccyzus erythropthalmus* | Black-billed cuckoo | Avise'94[30] | U09266 |
| Crosu | *Crotophaga sulcirostris* | Groove-billed ani | Avise'94[30] | U09260 |
| Cucpa | *Cuculus pallidus* | Pallid cuckoo | Avise'94[30] | U09262 |
| Piaca | *Piaya cayana* | Squirrel cuckoo | Avise'94[30] | U09263 |
| Phacu | *Phaenicophaeus curvirostris* | | Avise'94[30] | U09264 |
| II-12. Opisthocomiformes | | | | |
| OpihoA | *Opisthocomus hoazin* | Hoatzin | Avise'94[30] | U09257 |
| OpihoB | *Opisthocomus hoazin* | Hoatzin | Avise'94[30] | U09258 |
| OpihoC | *Opisthocomus hoazin* | Hoatzin | Avise'94[30] | U09259 |
| III. Class Amphibia | | | | |
| Xenla | *Xenopus laevis* | Clawed frog | Roe'85[248] | X02890 |
| IV. Class Osteichthyes (Bony fishes) | | | | |
| IV-1. Cypriniformes | | | | |
| Cypca | *Cyprinus carpio* | Carp | Chang'94[54] | X61010 |
| Crola | *Crossostoma lacustre* | Oriental stream loach | Tzeng'92[297] | M91245 |
| IV-2. Salmoniformes | | | | |
| Oncmy | *Oncorhynchus mykiss* | Rainbow trout | Zardaya'95[319] | L29771 |
| IV-3. Perciformes | | | | |
| Sarsa | *Sarda sarda* | Atlantic bonito | Cantatore'94[47] | X81562 |
| Thuth | *Thunnus thynnus* | Albacore | Cantatore'94[47] | X81563 |
| Scosc | *Scomber scombrus* | Atlantic mackerel | Cantatore'94[47] | X81564 |
| Oremo | *Oreochromis mossambicus* | | Cantatore'94[47] | X81565 |
| Dicla | *Dicentrarchus labrax* | European seabass | Cantatore'94[47] | X81566 |
| Boobo | *Boops boops* | | Cantatore'94[47] | X81567 |
| Tratr | *Trachurus trachurus* | Horse mackerel | Cantatore'94[47] | X81568 |
| IV-4. Cypriniformes | | | | |
| Lytat | *Lythrurus atrapiculus* | Blacktip shiner | Schmidt'95[258] | U17271 |
| Lytar | *Lythrurus ardens* | Rosefin shiner | Schmidt'95[258] | U17268 |
| Lytfu | *Lythrurus fumeus* | Ribbon shiner | Schmidt'95[258] | U17269 |
| Lytli | *Lythrurus lirus* | Mountain shiner | Schmidt'95[258] | U17273 |
| Lytsn | *Lythrurus snelsoni* | Ouchita mountain shiner | Schmidt'95[258] | U17272 |
| Lytum | *Lythrurus umbratilis* | Redfin shiner | Schmidt'95[258] | U17274 |
| Opsem | *Opsopoeodus emilae* | Pugnose minnow | Schmidt'95[258] | U17270 |
| IV-5. Gadiformes | | | | |
| Gadmo | *Gadus morhua* | Atlantic cod | Johansen'94[151] | X76365 |
| IV-6. Acipenseriformes | | | | |
| Acitr | *Acipenser transmontanus* | White sturgeon | Brown'89[42] | X14944 |
| V. Class Chondrichthyes (Cartilaginous fishes) | | | | |
| V-1. Carcharhiniformes | | | | |
| Carpl | *Carcharhinus plumbeus* | Sandbar shark | Martin'93[206] | L08032 |
| Carpo | *Carcharhinus porosus* | Smalltail shark | Martin'93[206] | L08033 |
| Prigl | *Prionace glauca* | Blue shark | Martin'93[206] | L08040 |
| Negbr | *Negaprion brevirostris* | Lemon shark | Martin'93[206] | L08039 |
| Sphtive | *Sphyrna tiburo vespertina* | Pacific bonnethead | Martin'93[206] | L08043 |
| Sphtiti | *Sphyrna tiburo tiburo* | Atlantic bonnethead | Martin'93[206] | L08042 |
| Sphle | *Sphyrna lewini* | Scalloped hammerhead | Martin'93[206] | L08041 |
| Galcu | *Galeocerdo cuvier* | Tiger shark | Martin'93[206] | L08034 |
| V-2. Lamniformes | | | | |
| Carca | *Carcharhodon carcharias* | White shark | Martin'93[206] | L08031 |
| Isuox | *Isurus oxyrhynchus* | Shortfin mako | Martin'93[206] | L08036 |
| Isupa | *Isurus paucus* | Longfin mako | Martin'93[206] | L08037 |
| Lamna | *Lamna nasus* | Porbeagle | Martin'93[206] | L08038 |
| V-3. Heterodontiformes | | | | |
| Hetfr | *Heterodontus francisci* | Horn shark | Martin'93[206] | L08035 |
| VI. Class Agnatha | | | | |
| VI-1. Petromyzontiformes | | | | |
| Petma | *Petromyzon marinus* | Sea lamprey | Lee'95[187] | U11880 |

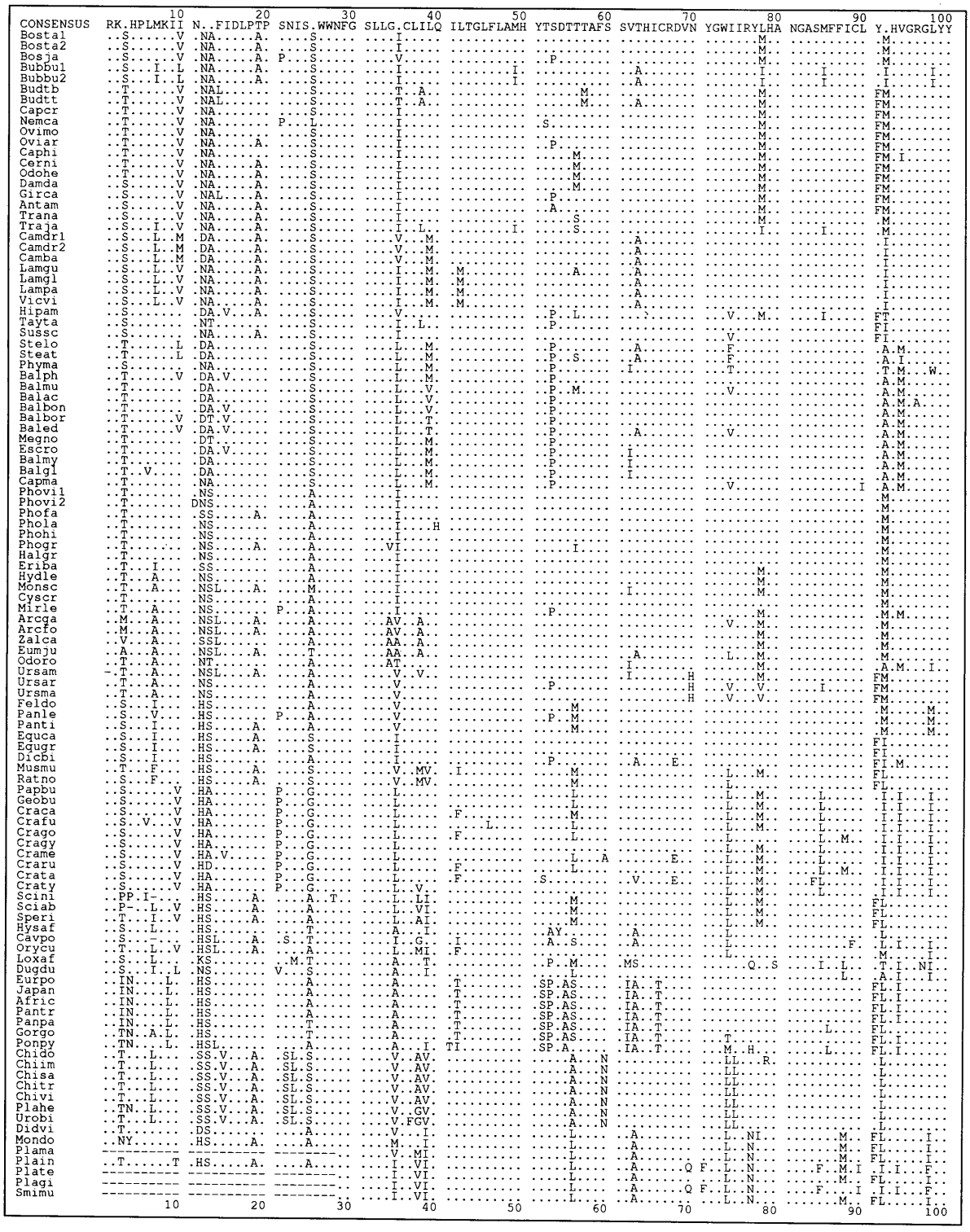


Figure 5.11: (a). The alignment of cytochrome b (mammal), part 1.

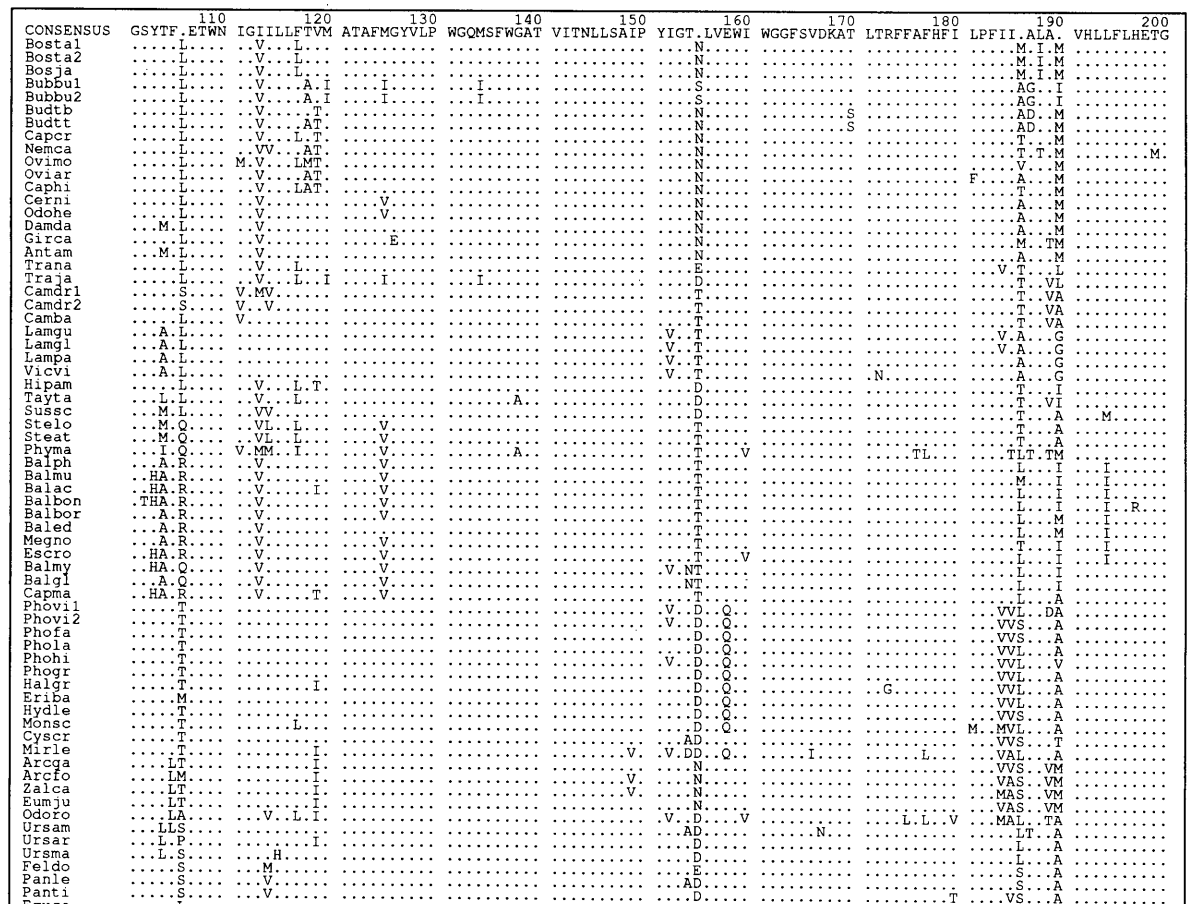



Figure 5.11: (b). The alignment of cytochrome b (mammal), part 2.


Figure 5.11: (c). The alignment of cytochrome b (mammal), part 3.

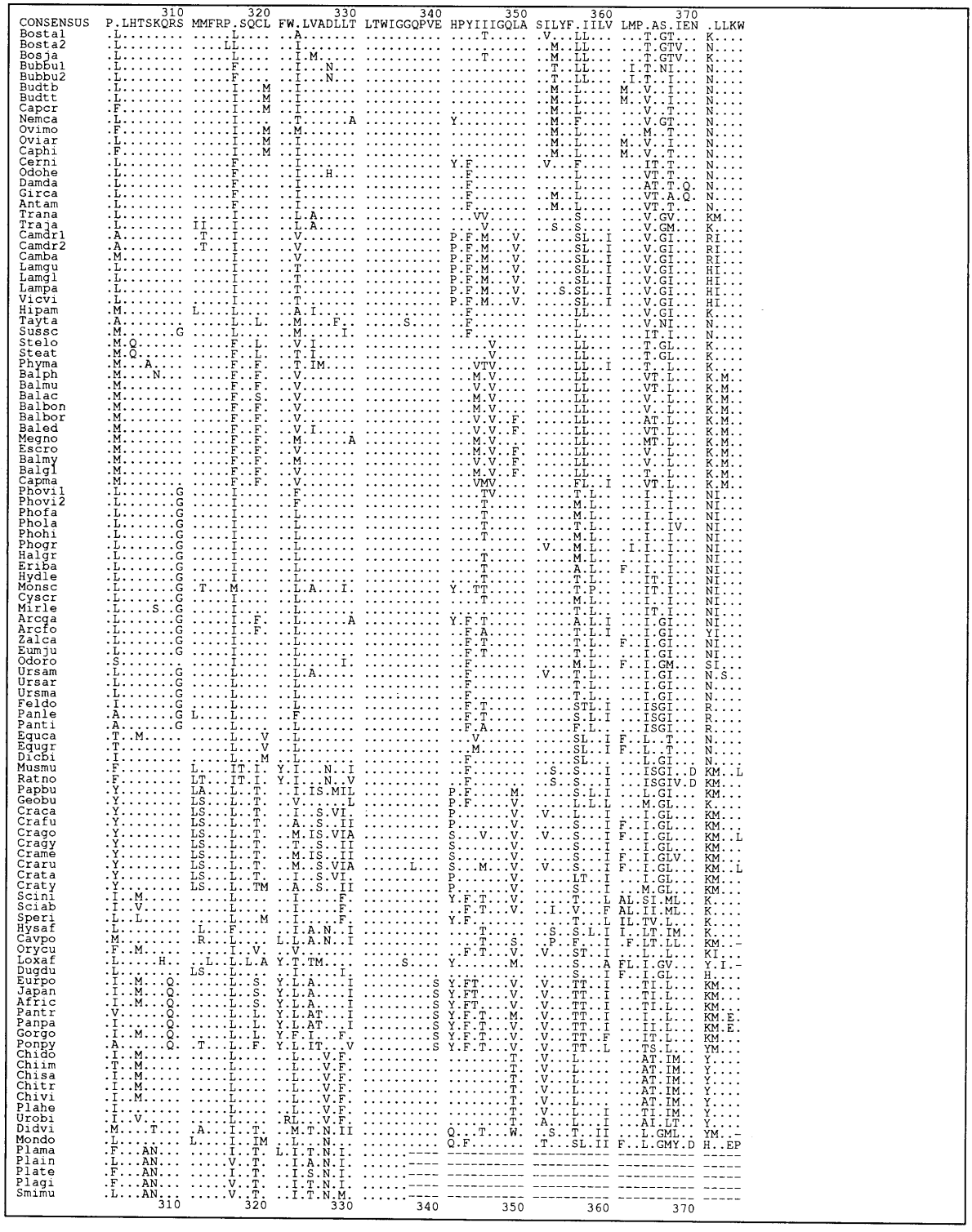



Figure 5.11: (d). The alignment of cytochrome b (mammal), part 4.

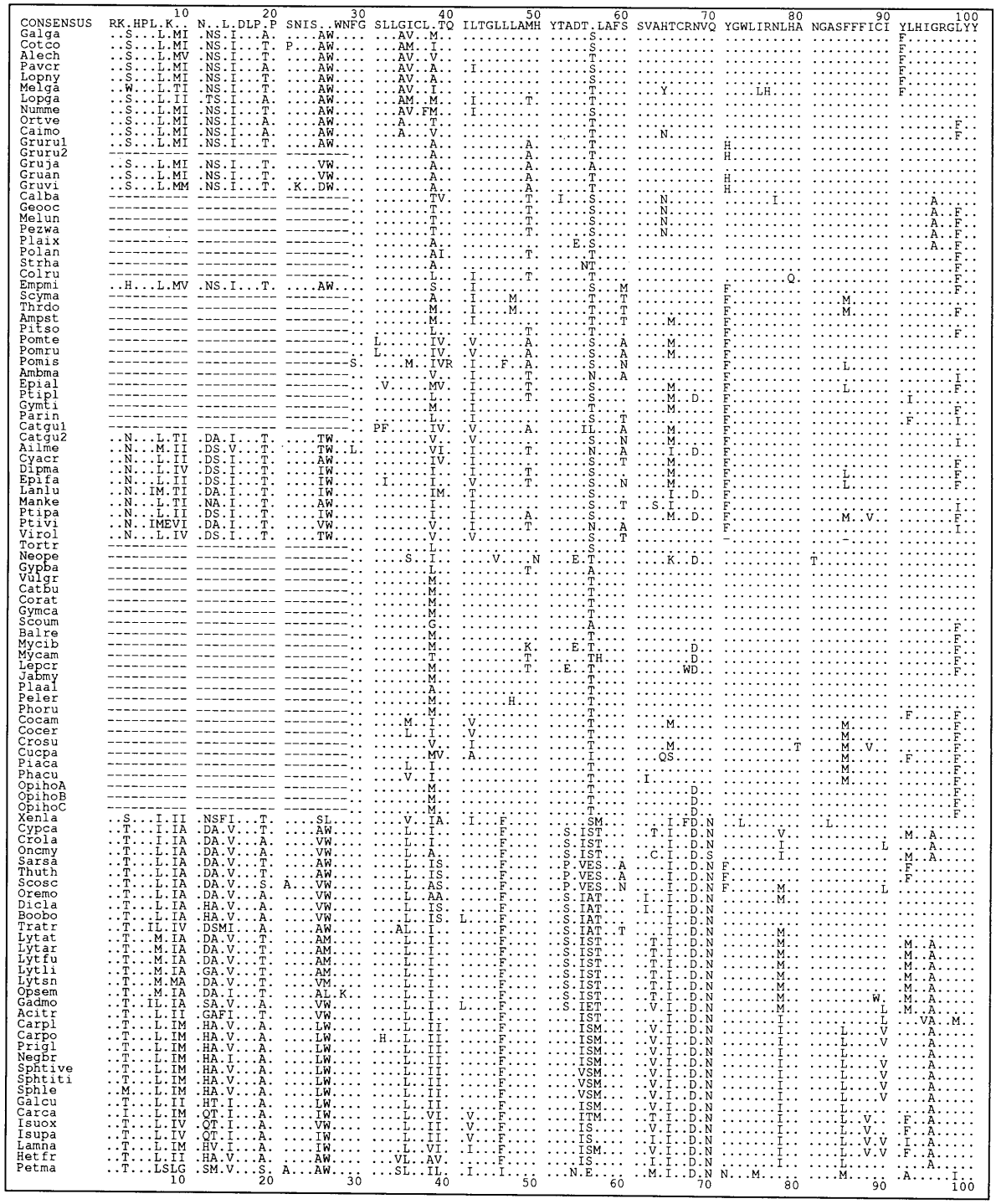


Figure 5.12: (a). The alignment of cytochrome b (except mammal), part 1.

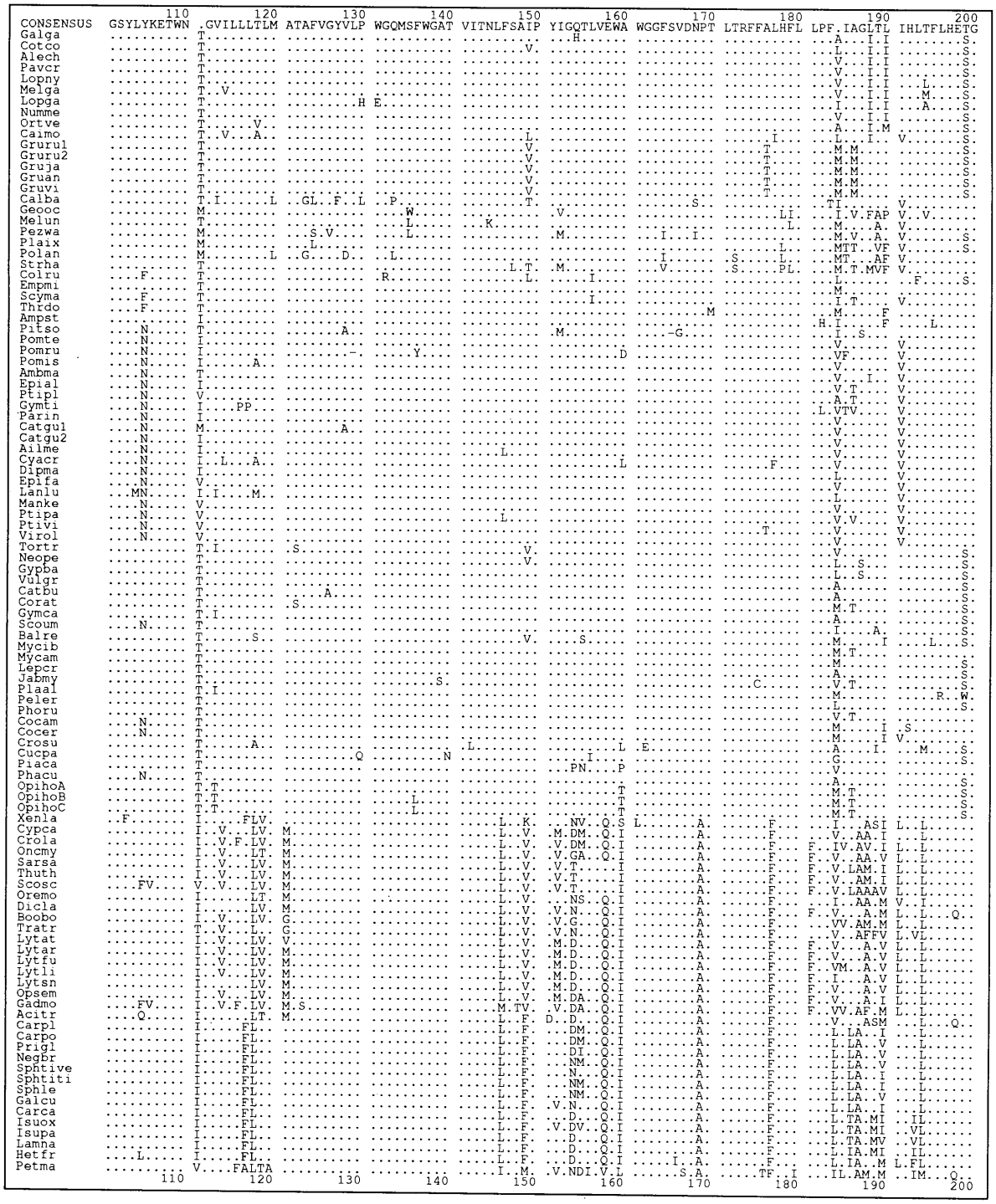


Figure 5.12: (b). The alignment of cytochrome b (except mammal), part 2.

Figure 5.12: (c). The alignment of cytochrome b (except mammal), part 3.

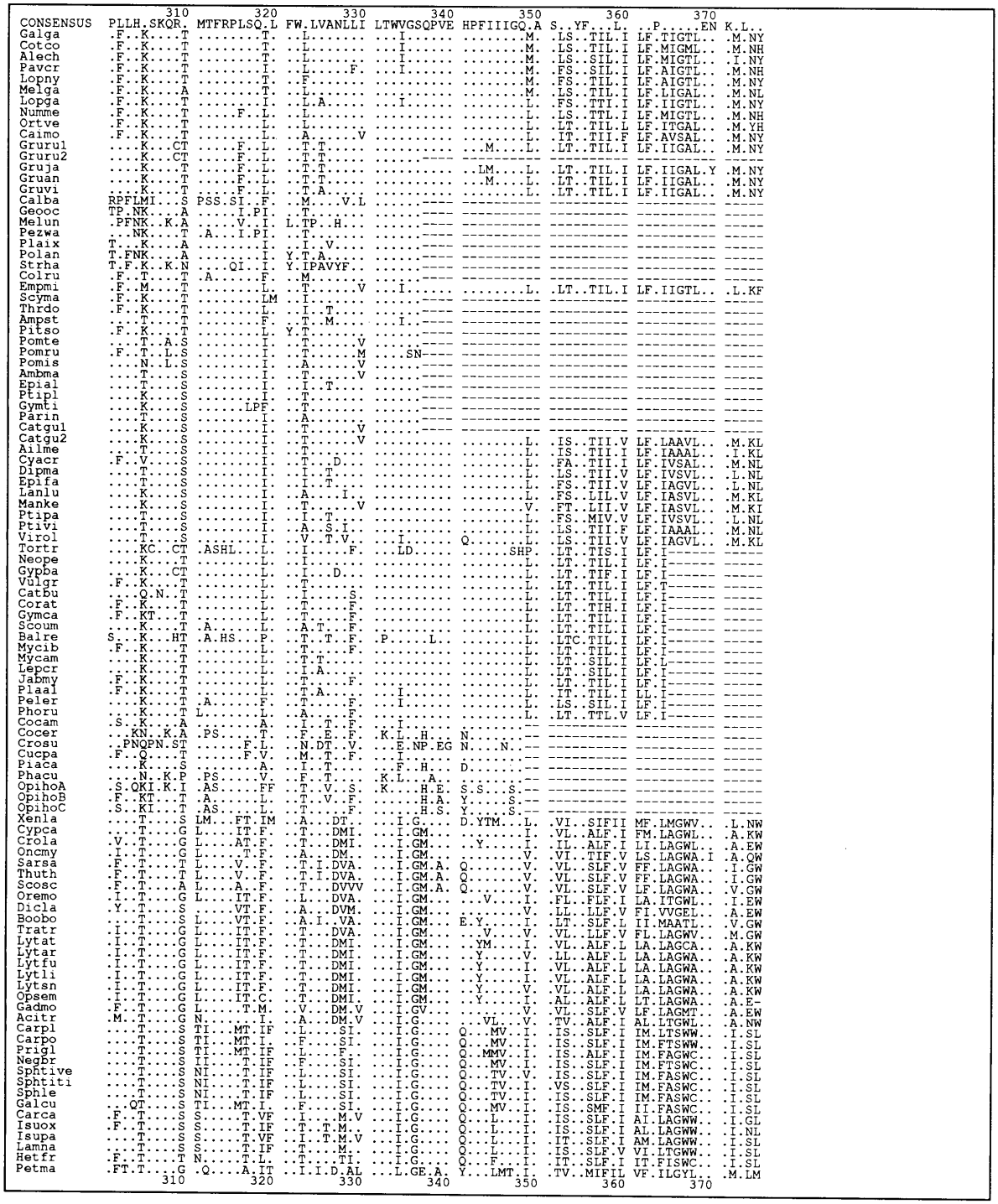



Figure 5.12: (d). The alignment of cytochrome b (except mammal), part 4.

### ProtML Tree of 183 OTUs Obtained by Repeating Local Rearrangements

Fig. 5.13 shows the NJ tree of cytochrome *b* from 182 OTUs of mammals and birds with a frog as an outgroup. The distance matrix provided for the NJ analysis was estimated for 2-OTUs trees by the ProtML based on the JTT-F model. Starting from this tree, the search for better tree topologies by the likelihood criterion was conducted by repeating local (and extended local) rearrangements as described in subsection 3.5.3. Fig. 5.11 gives the ProtML tree (based on the JTT-F model) which cannot be improved any more by the local rearrangements. The log-likelihood of the NJ tree is $-19376.24$, while that of the resultant ProtML tree is $-19044.73$, showing a great improvement of likelihood by 331.51 through the local rearrangement procedure.

### Phylogeny of Cetacea

Although the dolphin/sperm whale clade (monophyly of toothed whales; the traditional tree in section 5.3) is suggested by the NJ tree, the sperm/baleen whales clade (the Milinkovitch tree) is favoured in the ProtML tree, but only with 52% LBP (branch 214). The second most likely relationship concerning this branch is the traditional tree, and its LBP is 40%. Therefore, the Árnason tree has only 8% LBP, and is least likely from the cytochrome *b* data.

*Hippopotamus amphibius* is the most closely related species to Cetacea within Artiodactyla in accord with Irwin and Árnason (1994[145]), and this relationship is supported with 97% LBP (branch 217). The possible paraphyly of Artiodactyla is most interesting also with respect to the recently proposed hypothesis of Graur and Higgins (1994[100]) who claim the Ruminantia/Cetacea grouping, and more effort should be devoted to resolve this issue with additional sequence data. *Hippopotamus* traditionally considered to belong to Suiformes do not group with *Sus* and *Tayassu*. Camelidae, including the Old World and New World species, form a monophyletic group with 100% LBP (branch 224). Tragulidae is a sister group to all the other true ruminants (pecora). The monophyly of pecora is supported with 96% LBP (branch 199), and the monophyly of true ruminants with 90% LBP (branch 201).

### Phylogeny of Artiodactyla

In the NJ tree, Perissodactyla (species 44–46) is within the Cetacea/Artiodactyla group excluding Camelidae (species 20–26) as an outgroup, but the relationship is poorly supported by ProtML (only 3% LBP for branch 221 in Fig. 5.13), and the ProtML tree in Fig. 5.15 places Perissodactyla as a sister-group to the Cetacea/Artiodactyla clade with 97% LBP (branch 225).

The possible paraphyly of Bovidae (species 1–12) has been suggested by the previous analyses of cytochrome *b* sequences (Irwin et al. 1991[146]); Irwin and Árnason 1994[145]), and my analysis also favours the paraphyly. However the support is only 60% LBP (branch 192), and the monophyly of Bovidae has 29% LBP. It might be worth mentioning that, in Irwin and Árnason's parsimony analysis of

amino acid sequences, the paraphyly (sheep and goat are closer to other ruminant families than to cow) is supported with 100% BP. They used only three species from Bovidae, and the conclusion drawn from a limited number of species might be unstable.

The two groups of Cervidae, *Dama* and *Cervus/Odocoileus*, do not form a monophyletic clade, and *Dama* is most closely related to *Antilocapra americana* (pronghorn) with 81% LBP consistently with the previous analyses by Irwin et al. (1991[146] and Irwin and Árnason (1994[145]). Further study is needed to prove or disprove this morphologically unexpected relationship.

### Phylogeny of Rodentia

The separate origin of Geomyidae (pocket gophers) from the other rodent groups in Fig. 5.15 is in accord with the NJ analysis of more limited data set of cytochrome *b* by Philippe and Douzery (1994[240]). Geomyidae, which belongs to Sciuromorpha by the traditional taxonomy (Nowak 1991[230]), does not cluster even with another Sciuromorpha group, Sciuridae (squirrels), not only with Hystricomorpha and Myomorpha by my analysis. Philippe and Douzery attributed this unexpected placement of Geomyidae to a higher rate of evolution in Geomyidae (DeWalt et al. 1993[66]). Some unusual evolution might have occurred in the cytochrome *b* gene of Geomyidae.

Within Geomyidae, *Cratogeomys* form a monophyletic clade in the parsimony and Fitch trees of DeWalt et al. (1993[66]), while *C. merriami* is an outgroup to all the other pocket gophers including *Pappogeomys* and *Geomys* in the ProtML tree. The relevant LBP is not very high (66%: branch 283) and the LBP of *Cratogeomys*-monophyly is 19%. Further studies are needed to settle the issue.

My analysis strongly support a *Cavia/Hystrix* clade with 97% LBP, consistently with Ma et al. (1993[199]) and with Cao et al. (1994[50]). The close relationship between the South American and the African Hystricomorpha is in accord with the hypothesis that the South American ones originated in Africa (Wyss et al. 1993[312]).

The ProtML analysis of cytochrome *b* by Cao et al. (1994[50]) gave a rodent-monophyly tree with a Myomorpha/Caviomorpha clade. Although Fig. 5.15 gives a tree similar to the rodent-polyphyly hypothesis proposed by Graur et al. (1991[99]), the relevant branches are very poorly supported. Given the abundant database relevant to this problem (Cao et al. 1994[50]), Graur et al.'s hypothesis seems unlikely.

### Phylogeny of Microchiroptera

Five species of *Chiroderma* form a monophyletic clade in Fig. 5.15, and *Platyrrhinus* is a sister-group to *Uroderma* with 92% LBP (branch 269).

### Phylogeny of Carnivora

My ProtML tree suggest a *Arctocephalus*/sea lion clade (99% LBP: branch 242) which is a sister-group to *Odobenus* (walrus) (95% LBP: branch 243) in accord with Árnason et al. (1995[20]). Within the northern phocids, *Erignathus barbatus* (bearded seal) is an outgroup to all the others with 91% LBP (branch 236). The genus *Phoca* is highly likely to be paraphyletic, and *Halichoerus* represented by the grey seal and *Cystophora* represented by the hooded seal might be included in the genus.

The monophyly of Pinnipedia is strongly supported with 97% LBP (branch 244). Although some morphologists maintain independent origins for phocids and otariids (e.g., Tedford 1976[290]), my result is consistent with previous molecular studies (Vrana et al. 1994[304]; Árnason et al. 1995[20]) as well as with the recent morphological studies (Wyss 1988[310]; Wyss and Flynn 1993[311]).

The Pinnipedia is a sister-group to *Ursus* with 93% LBP (branch 247) excluding the *Felis/Panthera* clade as an outgroup in Carnivora (Vrana et al. 1994[304]; Árnason et al. 1995[20]; Lento et al. 1995[190]).

### Phylogeny of Other Mammals

The association of *Loxodonta* (elephant) with *Dugong* is supported with 100% LBP in accord with Irwin and Árnason (1994[145]), with Kleinschmidt et al. (1986[167]), and with Springer and Kirsch (1993[274]).

Paraphyly of Perissodactyla suggested by branch 227 (70% LBP) is unexpected, but it seems to be artificial, and the monophyly tree has 28% LBP.

Within subfamily Sminthopsinae of Australian marsupials, although *Planigale* is paraphyletic in the NJ tree, the four *Planigale* species form a monophyletic clade which is a sister-group to *Sminthopsis* with 78% LBP (branch 288) in the ProtML tree.

### Phylogeny of Aves

Many of the Aves orders, such as Gruiformes, Psittaciformes, Cuculiformes, and Galliformes, respectively form monophyletic clades within the ProtML tree of Fig. 5.15. Passeriformes are separated into two monophyletic groups in the tree, that is, Suboscines and Oscines, but the possibility of Passeriformes monophyly cannot be evaluated adequately in the presence of huge number of possible trees. Nevertheless, I think the apparantly separate placing of the two groups of Passeriformes may not be real, because LBP's of the branchs separating the two groups are low. Suboscines include *Scytalopus magellanicus* (Andean tapaculo), *Thripophaga dorbignyi* (creamy-breasted canastero), *Ampelion stresemanni* (white-cheeked cotinga), *Pitta sordida* (hooded pitta), and *Empidonax minimus* (least flycatcher), and Oscines include all the other Passeriformes species analyzed in this thesis. Monophyly of respective groups of Suboscines and Oscines is consistent with the previous analyses of cytochrome *b* by Edwards et al. (1991[72]) and by Helm-Bychowski and Cracraft (1993[137]) and with Sibley and Ahlquist (1990[264]). In the NJ tree of Fig. 5.13, two groups of Suboscines, (*Pitta sordida*, *Empidonax minimus*) and ((*Scytalopus*

*magellanicus*, *Thripophaga dorbignyi*), *Ampelion stresemanni*), are separate, and furthermore the latter is paraphyletic. The ProtML tree seems more reasonable in this respect.

Galliformes are not monophyletic in the NJ tree; *Ortalis vetula* (chachalaca; species 161) clusters with an Anseriformes species, *Cairina moschata* (Muscovy duck), and this group is distantly separate from the other Galliformes (species 143–150). However, it turned out that all the Galliformes birds form a monophyletic clade with Anseriformes as a sister-group in the ProtML tree, which might be more reasonable than the NJ tree in this respect. The association between Anseriformes and Galliformes is supported with 97% LBP (branch 323) in accord with Sibley and Ahlquist's (1990[264]) classification based on DNA-DNA hybridization. The place of *Opisthocomus hoazin* is obscured by this analysis as in Avise et al. (1994[30]).

The most important feature of the Aves part of Fig. 5.15 might be that Falconiformes, Ciconiiformes, Pelicaniformes, and Phoenicopteriformes are intermixed on the tree, consistently to some extent with Sibley and Ahlquist's (1990[264]) classification based on DNA-DNA hybridization. Except that *Mycteria americana* (American wood ibis) and *Leptoptilos crumeniferus* (Marabou stork) are each others closest relatives in the tree (98% LBP: branch 299) in accord with Avise et al. (1994[29]), no other clade in this group is strongly supported, and therefore no resolution of branching order is attainable only from the cytochrome *b* data. Given that the overall feature of the ProtML tree of cytochrome *b* is reasonable, however, the intermixing among Falconiformes, Ciconiiformes, Pelicaniformes, Phoenicopteriformes and Gruiformes might reflect the real evolutionary history of these birds to some extent.

The separation of a (((*Coragyps atratus*, *Jabiru mycteria*), *Gymnogyps californianus*), *Mycteria ibis*) clade from the other members of Falconiformes and Ciconiiformes, and from Pelicaniformes, Phoenicopteriformes and Gruiformes are likely to be an artifact, and these birds form a monophyletic clade in the NJ tree. Based on the DNA-DNA hybridization data, Sibley and Ahlquist (1990[264]) included Falconides (Old World vultures, eagles) and Ciconiides in their suborder Ciconii of order Ciconiiformes, and Pelicanoidea (pelicans and shoebill), Phoenicopteridae (flamingos), Threskiornithoidea (ibises and spoonbills), and Ciconioidea (New World vultures, condors, storkes, jabiru) in infraorder Ciconiides. Gruiformes form a separate order in their classification. In order to clarify the relationships among these birds, futher studies of different genes are needed.

It seems contradictory that *Vultur gryphus* (Andean condor) and *Gymnogyps californianus* (California condor) do not form a clade in the cytochrome *b* tree, while the clade is supported by 99% BP in Hedges and Sibley's (1994[136]) analysis of mitochondrial ribosomal RNAs, although the number of relevant species they used is less than that of mine.

### Phylogeny of Galliformes

The Galliformes part of the tree is mostly consistent with that of Kornegay et al. (1993[173]). The sister-group of *Ortalis vetula* (chachalaca) to all the other Galliformes analyzed in this work is supported

with 93% LBP (branch 322).

The egg-white lysozyme *c* sequences of Galliformes possess a unique pattern of amino acid replacements at three internally clustered residues. These positions are occupied in all characterized galliform bird lysozymes by Thr 40, Ile 55, and Ser 91 (TIS), with the exception of the guinea fowl (Numididae) and the New World quail (Odontophoridae) lysozymes, which have Ser 40, Val 55, and Thr 91 (SVT) at these positions (Jollès et al. 1976[152]; Jollès and Jollès 1984[153]; Malcolm et al. 1990[200]). Therefore, amino acid sequences of these lysozymes suggest that the guinea fowl and the New World quail form a clade excluding Phasianidae and Meleagrididae (turkey) as outgroups. However, this suggestion is not supported by morphological and other molecular evidence, and Ibrahimi et al. (1979[144]) viewed this as an unusual case of coupled amino acid replacements in the lysozyme *c* which occurred independently in the two lineages of Galliformes.

From the analysis of cytochrome *b* genes, Kornegay et al. placed the New World quail *Lophortyx gambelii* outside *Numida meleagris* (Guinea fowl), Phasianidae and Meleagrididae, and claimed the independent occurrences of coupled amino acid replacements in the lysozyme *c* in the two lineages. However, in spite of the presentation of detailed comparison of several phylogenetic hypotheses by the ML method in their Table 4, Kornegay et al. did not show the evaluation of the lysozyme tree with a clade of the guinea fowl and the New World quail. My Fig. 5.15 is consistent with Kornegay et al.'s tree, but the outgroup position of the New World quail is only poorly supported (71% LBP: branch 320), and the lysozyme tree has 29% LBP. Recently, Avise et al. (1994[30]) published the cytochrome *b* sequence from California quail, which is another species of New World quails. The data is a partial sequence (covers 320 amino acids). When this data is additionally used, the grouping of the New World quails with guinea fowl is preferred by the ProtML analysis (Cao, Adachi, and Hasegawa, unpublished). Therefore, the clustering of the New World quail with the guinea fowl cannot be dismissed as a candidate of the true tree at least as far as the cytochrome *b* data is concerned.

Placement of the New World quail outside phasianoids, turkey and guinea fowl as suggested by Sibley and Ahlquist (1985[262]) and by Kornegay et al. (1993[173]) implies that coupled amino acid replacements of lysozyme *c* occurred independently at least in two lineages of Galliformes. If this is actually the case, this represents a remarkable case of convergent or reversal evolution. A case of convergent evolution of lysozyme has been demonstrated by Stewart et al. (1987[279]) for ruminants and leaf-eating monkeys. A similar situation may of course be possible for the galliform birds, but the data presented by Kornegay et al. does not seem to present a convincing evidence for such a highly interesting evolution. I think that further studies are needed to answer the question whether such evolution actually occurred in the lysozyme of Galliformes.

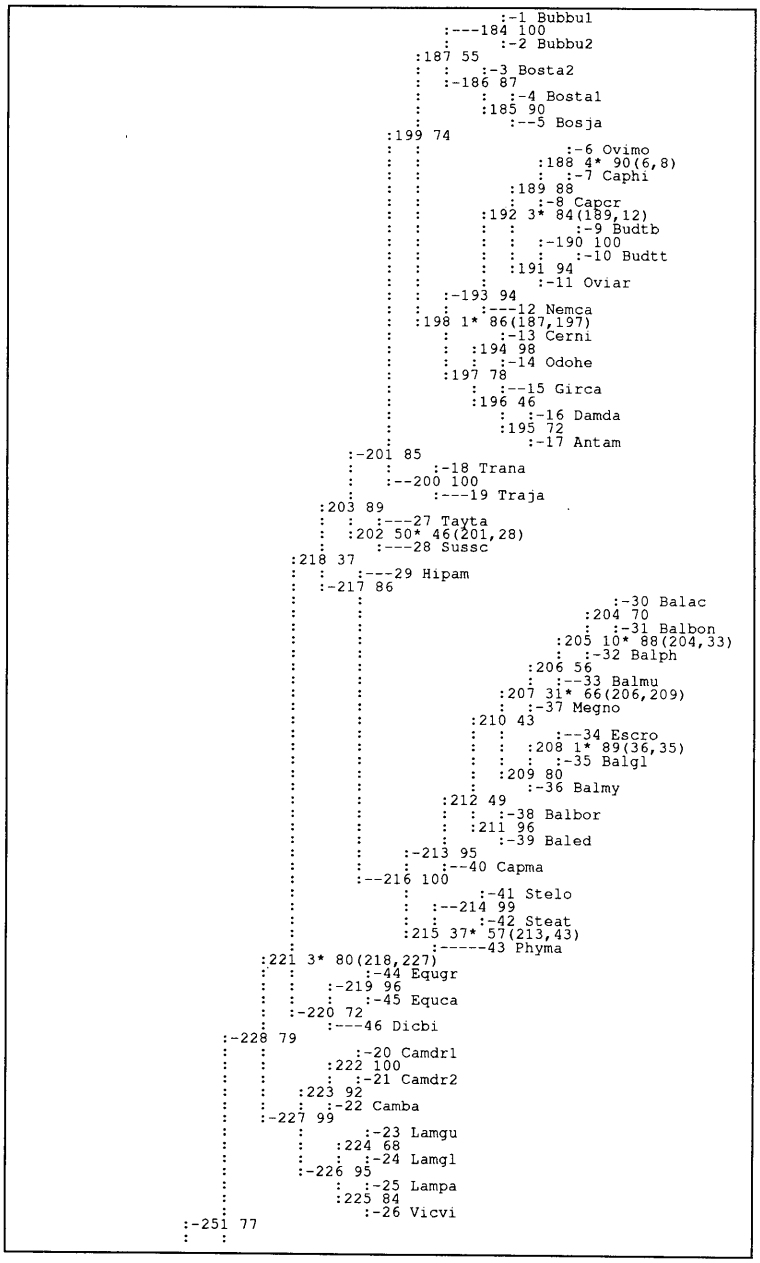



Figure 5.13: (a). The NJ tree of cytochrome b, part 1.

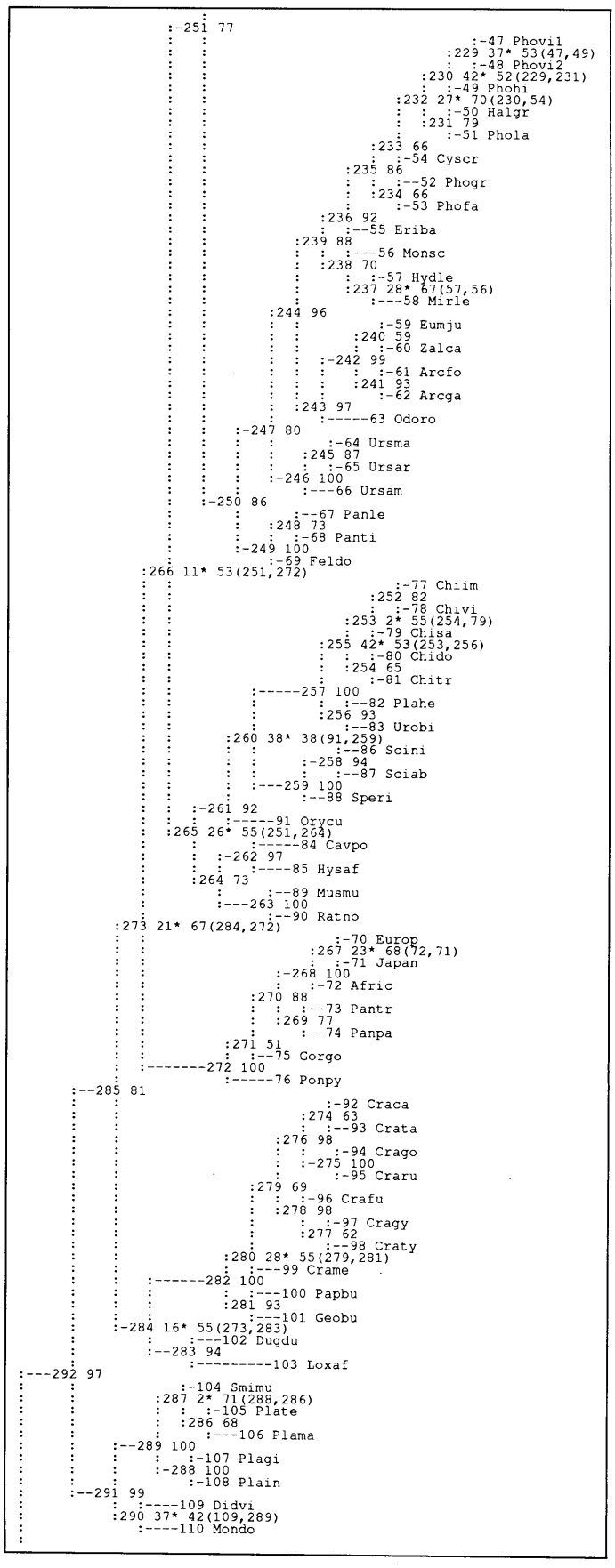



Figure 5.13: (b). The NJ tree of cytochrome b, part 2.

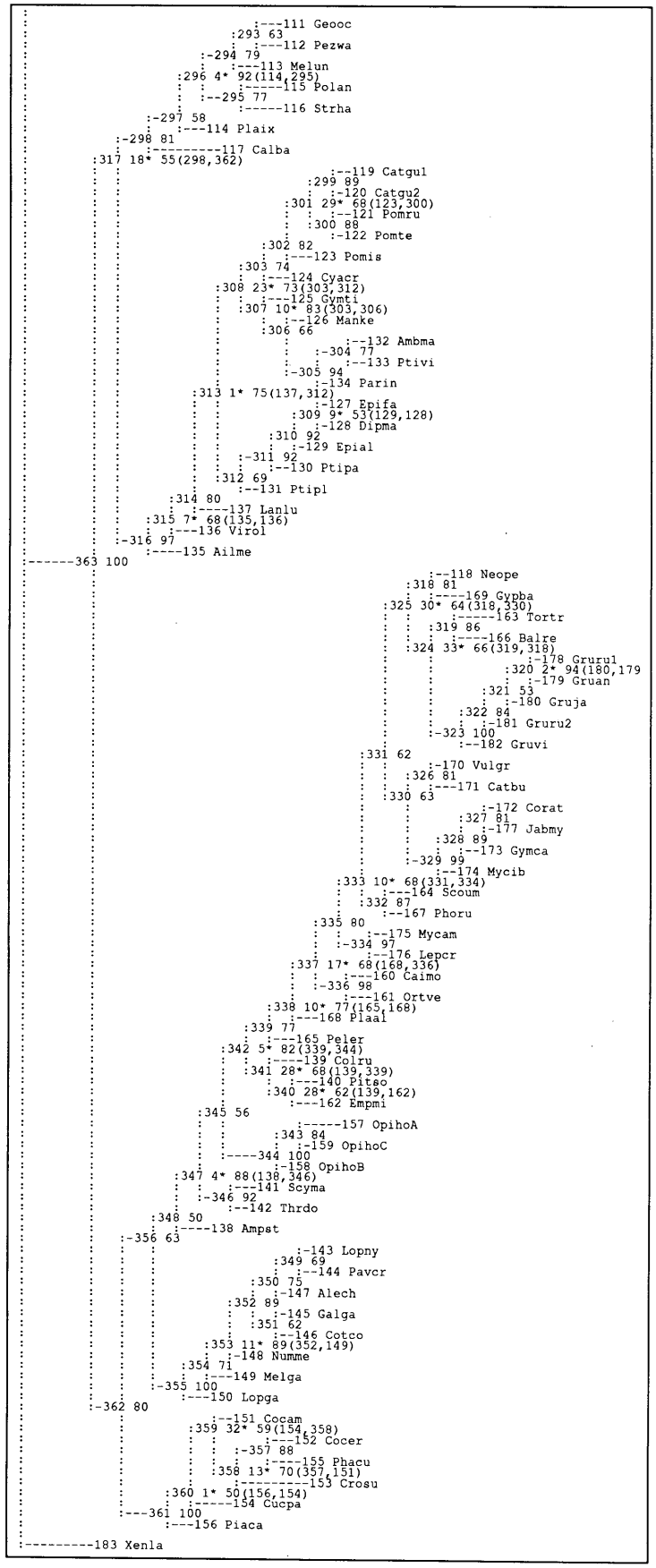



Figure 5.13: (c). The NJ tree of cytochrome b, part 3.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Bubbu1 | 1 | 6.18 | 1.35 | 184 | 6.18 | 1.35 | 1.0 | 0.0 | (1,186) |
| Bubbu2 | 2 | 1.00 | 0.54 | 185 | 1.00 | 0.54 | 0.905 | 0.077 | (3,5) |
| Bosta2 | 3 | 2.15 | 0.80 | 186 | 2.15 | 0.80 | 0.871 | 0.120 | (3,184) |
| Bosta1 | 4 | 1.52 | 0.70 | 187 | 1.52 | 0.70 | 0.547 | 0.399 | (184,198) |
| Bosja | 5 | lower | limit | 188 | lower | limit | 0.036* | 0.898 | (6,8) |
| Ovimo | 6 | 0.83 | 0.53 | 189 | 0.83 | 0.53 | 0.875 | 0.085 | (188,191) |
| Caphi | 7 | 3.39 | 0.97 | 190 | 3.39 | 0.97 | 1.0 | 0.0 | (11,10) |
| Capcr | 8 | 1.74 | 0.74 | 191 | 1.74 | 0.74 | 0.942 | 0.038 | (189,11) |
| Budtb | 9 | 0.22 | 0.29 | 192 | 0.22 | 0.29 | 0.033* | 0.842 | (189,12) |
| Budtt | 10 | 2.61 | 0.93 | 193 | 2.61 | 0.93 | 0.945 | 0.053 | (192,197) |
| Oviar | 11 | 1.62 | 0.69 | 194 | 1.62 | 0.69 | 0.981 | 0.019 | (13,196) |
| Nemca | 12 | 0.92 | 0.52 | 195 | 0.92 | 0.52 | 0.721 | 0.276 | (16,15) |
| Cerni | 13 | 0.31 | 0.32 | 196 | 0.31 | 0.32 | 0.456 | 0.433 | (194,195) |
| Odohe | 14 | 1.15 | 0.64 | 197 | 1.15 | 0.64 | 0.777 | 0.160 | (194,193) |
| Girca | 15 | lower | limit | 198 | lower | limit | 0.013* | 0.857 | (187,197) |
| Damda | 16 | 1.62 | 0.80 | 199 | 1.62 | 0.80 | 0.744 | 0.234 | (187,200) |
| Antam | 17 | 4.33 | 1.18 | 200 | 4.33 | 1.18 | 0.996 | 0.004 | (199,19) |
| Trana | 18 | 2.36 | 0.90 | 201 | 2.36 | 0.90 | 0.846 | 0.152 | (199,202) |
| Traja | 19 | 0.82 | 0.57 | 202 | 0.82 | 0.57 | 0.495* | 0.464 | (201,28) |
| Camdr1 | 20 | 1.45 | 0.75 | 203 | 1.45 | 0.75 | 0.889 | 0.072 | (201,217) |
| Camdr2 | 21 | 0.83 | 0.51 | 204 | 0.83 | 0.51 | 0.702 | 0.263 | (32,31) |
| Camba | 22 | lower | limit | 205 | lower | limit | 0.105* | 0.884 | (204,33) |
| Lamgu | 23 | 0.35 | 0.35 | 206 | 0.35 | 0.35 | 0.557 | 0.375 | (205,37) |
| Lamgl | 24 | lower | limit | 207 | lower | limit | 0.307* | 0.655 | (206,209) |
| Lampa | 25 | lower | limit | 208 | lower | limit | 0.010* | 0.888 | (36,35) |
| Vicvi | 26 | 1.30 | 0.65 | 209 | 1.30 | 0.65 | 0.798 | 0.198 | (208,207) |
| Tayta | 27 | 0.73 | 0.50 | 210 | 0.73 | 0.50 | 0.431 | 0.344 | (207,211) |
| Sussc | 28 | 1.40 | 0.67 | 211 | 1.40 | 0.67 | 0.955 | 0.036 | (38,210) |
| Hipam | 29 | 0.52 | 0.41 | 212 | 0.52 | 0.41 | 0.490 | 0.341 | (210,40) |
| Balac | 30 | 2.10 | 0.90 | 213 | 2.10 | 0.90 | 0.947 | 0.036 | (215,40) |
| Balbon | 31 | 3.79 | 1.12 | 214 | 3.79 | 1.12 | 0.989 | 0.011 | (43,42) |
| Balph | 32 | 1.34 | 0.76 | 215 | 1.34 | 0.76 | 0.374* | 0.567 | (213,43) |
| Balmu | 33 | 4.95 | 1.28 | 216 | 4.95 | 1.28 | 0.997 | 0.002 | (29,215) |
| Escro | 34 | 2.44 | 0.90 | 217 | 2.44 | 0.90 | 0.865 | 0.118 | (29,203) |
| Balgl | 35 | 0.84 | 0.65 | 218 | 0.84 | 0.65 | 0.367 | 0.383 | (203,220) |
| Balmy | 36 | 3.24 | 1.02 | 219 | 3.24 | 1.02 | 0.965 | 0.034 | (44,46) |
| Megno | 37 | 2.24 | 0.94 | 220 | 2.24 | 0.94 | 0.720 | 0.195 | (218,46) |
| Balbor | 38 | 1.02 | 0.71 | 221 | 1.02 | 0.71 | 0.030* | 0.804 | (218,227) |
| Baled | 39 | 1.58 | 0.66 | 222 | 1.58 | 0.66 | 0.995 | 0.005 | (22,21) |
| Capma | 40 | 1.68 | 0.77 | 223 | 1.68 | 0.77 | 0.915 | 0.038 | (222,226) |
| Stelo | 41 | 0.29 | 0.30 | 224 | 0.29 | 0.30 | 0.684 | 0.291 | (225,24) |
| Steat | 42 | 0.77 | 0.47 | 225 | 0.77 | 0.47 | 0.840 | 0.079 | (25,224) |
| Phyma | 43 | 2.62 | 0.90 | 226 | 2.62 | 0.90 | 0.946 | 0.054 | (223,225) |
| Equgr | 44 | 3.68 | 1.11 | 227 | 3.68 | 1.11 | 0.990 | 0.009 | (221,226) |
| Equca | 45 | 2.45 | 0.95 | 228 | 2.45 | 0.95 | 0.794 | 0.154 | (250,227) |
| Dicbi | 46 | 0.26 | 0.26 | 229 | 0.26 | 0.26 | 0.368* | 0.526 | (47,49) |
| Phovi1 | 47 | 0.34 | 0.33 | 230 | 0.34 | 0.33 | 0.425* | 0.524 | (229,231) |
| Phovi2 | 48 | lower | limit | 231 | lower | limit | 0.793 | 0.141 | (230,51) |
| Phohi | 49 | lower | limit | 232 | lower | limit | 0.274* | 0.703 | (230,54) |
| Halgr | 50 | 0.43 | 0.37 | 233 | 0.43 | 0.37 | 0.658 | 0.249 | (234,54) |
| Phola | 51 | 0.36 | 0.34 | 234 | 0.36 | 0.34 | 0.662 | 0.207 | (52,233) |
| Phogr | 52 | 0.86 | 0.52 | 235 | 0.86 | 0.52 | 0.865 | 0.113 | (55,234) |
| Phofa | 53 | 0.93 | 0.55 | 236 | 0.93 | 0.55 | 0.924 | 0.023 | (235,238) |
| Cyscr | 54 | lower | limit | 237 | lower | limit | 0.282* | 0.674 | (57,56) |
| Eriba | 55 | 0.32 | 0.32 | 238 | 0.32 | 0.32 | 0.698 | 0.234 | (56,236) |
| Monsc | 56 | 1.10 | 0.56 | 239 | 1.10 | 0.56 | 0.876 | 0.116 | (236,243) |
| Hydle | 57 | 0.64 | 0.45 | 240 | 0.64 | 0.45 | 0.590 | 0.386 | (59,241) |
| Mirle | 58 | 1.05 | 0.57 | 241 | 1.05 | 0.57 | 0.931 | 0.047 | (61,240) |
| Eumju | 59 | 2.54 | 0.90 | 242 | 2.54 | 0.90 | 0.987 | 0.013 | (240,63) |
| Zalca | 60 | 1.70 | 0.73 | 243 | 1.70 | 0.73 | 0.966 | 0.026 | (242,239) |
| Arcfo | 61 | 1.54 | 0.70 | 244 | 1.54 | 0.70 | 0.960 | 0.028 | (239,246) |
| Arcga | 62 | 0.79 | 0.49 | 245 | 0.79 | 0.49 | 0.873 | 0.120 | (64,66) |
| Odoro | 63 | 2.01 | 0.78 | 246 | 2.01 | 0.78 | 0.995 | 0.005 | (244,66) |
| Ursma | 64 | 2.11 | 0.83 | 247 | 2.11 | 0.83 | 0.803 | 0.197 | (244,249) |
| Ursar | 65 | 0.81 | 0.53 | 248 | 0.81 | 0.53 | 0.730 | 0.213 | (69,68) |
| Ursam | 66 | 3.06 | 1.03 | 249 | 3.06 | 1.03 | 0.995 | 0.004 | (248,247) |
| Panle | 67 | 3.10 | 1.06 | 250 | 3.10 | 1.06 | 0.856 | 0.143 | (247,228) |
| Panti | 68 | 3.23 | 1.07 | 251 | 3.23 | 1.07 | 0.772 | 0.225 | (228,265) |
| Feldo | 69 | 0.91 | 0.53 | 252 | 0.91 | 0.53 | 0.821 | 0.164 | (79,78) |
| Europ | 70 | lower | limit | 253 | lower | limit | 0.015* | 0.545 | (254,79) |
| Japan | 71 | 0.49 | 0.41 | 254 | 0.49 | 0.41 | 0.654 | 0.317 | (253,81) |
| Afric | 72 | 0.71 | 0.54 | 255 | 0.71 | 0.54 | 0.421* | 0.533 | (253,256) |
| Pantr | 73 | 1.84 | 0.77 | 256 | 1.84 | 0.77 | 0.929 | 0.071 | (255,83) |
| Panpa | 74 | 10.44 | 1.83 | 257 | 10.44 | 1.83 | 1.0 | 0.0 | (259,256) |
| Gorgo | 75 | 2.97 | 1.02 | 258 | 2.97 | 1.02 | 0.945 | 0.041 | (86,88) |
| Ponpy | 76 | 5.66 | 1.37 | 259 | 5.66 | 1.37 | 0.998 | 0.002 | (258,257) |
| Chiim | 77 | 0.80 | 0.71 | 260 | 0.80 | 0.71 | 0.382* | 0.377 | (91,259) |
| Chivi | 78 | 2.02 | 0.92 | 261 | 2.02 | 0.92 | 0.917 | 0.040 | (260,264) |
| Chisa | 79 | 2.98 | 1.07 | 262 | 2.98 | 1.07 | 0.972 | 0.020 | (84,263) |
| Chido | 80 | 7.21 | 1.51 | 263 | 7.21 | 1.51 | 1.0 | 0.0 | (89,262) |
| Chitr | 81 | 1.76 | 0.80 | 264 | 1.76 | 0.80 | 0.729 | 0.218 | (261,263) |
| Plahe | 82 | 0.87 | 0.74 | 265 | 0.87 | 0.74 | 0.256* | 0.548 | (251,264) |
| Urobi | 83 | 0.78 | 0.76 | 266 | 0.78 | 0.76 | 0.106* | 0.529 | (251,272) |
| Cavpo | 84 | 0.21 | 0.27 | 267 | 0.21 | 0.27 | 0.227* | 0.685 | (72,71) |
| Hysaf | 85 | 3.70 | 1.03 | 268 | 3.70 | 1.03 | 1.0 | 0.0 | (267,269) |
| Scini | 86 | 0.82 | 0.53 | 269 | 0.82 | 0.53 | 0.773 | 0.173 | (268,74) |
| Sciab | 87 | 1.85 | 0.78 | 270 | 1.85 | 0.78 | 0.885 | 0.094 | (268,75) |
| Speri | 88 | 1.58 | 0.89 | 271 | 1.58 | 0.89 | 0.508 | 0.448 | (76,75) |
| Musmu | 89 | 14.83 | 2.21 | 272 | 14.83 | 2.21 | 1.0 | 0.0 | (266,76) |
| Ratno | 90 | 1.44 | 0.91 | 273 | 1.44 | 0.91 | 0.210* | 0.669 | (284,272) |
| Orycu | 91 | 0.50 | 0.41 | 274 | 0.50 | 0.41 | 0.633 | 0.205 | (92,275) |

Figure 5.14: (a). Branch lengths and LBPs (estimated by ML) of the NJ tree of cytochrome b, part 1.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Orycu | 91 | 0.50 | 0.41 | 274 | 0.50 | 0.41 | 0.633 | 0.205 | (92,275) |
| Craca | 92 | 3.26 | 0.98 | 275 | 3.26 | 0.98 | 1.0 | 0.0 | (274,95) |
| Crata | 93 | 1.55 | 0.67 | 276 | 1.55 | 0.67 | 0.983 | 0.014 | (274,278) |
| Crago | 94 | 0.33 | 0.32 | 277 | 0.33 | 0.32 | 0.620 | 0.270 | (97,96) |
| Craru | 95 | 1.10 | 0.55 | 278 | 1.10 | 0.55 | 0.975 | 0.019 | (276,277) |
| Crafu | 96 | 0.27 | 0.27 | 279 | 0.27 | 0.27 | 0.691 | 0.291 | (99,278) |
| Cragy | 97 | lower | limit | 280 | lower | limit | 0.283* | 0.554 | (279,281) |
| Craty | 98 | 1.50 | 0.69 | 281 | 1.50 | 0.69 | 0.928 | 0.064 | (100,280) |
| Crame | 99 | 11.37 | 1.94 | 282 | 11.37 | 1.94 | 1.0 | 0.0 | (280,283) |
| Papbu | 100 | 5.21 | 1.41 | 283 | 5.21 | 1.41 | 0.939 | 0.050 | (102,282) |
| Geobu | 101 | 1.96 | 1.13 | 284 | 1.96 | 1.13 | 0.160* | 0.551 | (273,283) |
| Dugdu | 102 | 3.77 | 1.25 | 285 | 3.77 | 1.25 | 0.812 | 0.188 | (291,284) |
| Loxaf | 103 | 0.47 | 0.48 | 286 | 0.47 | 0.48 | 0.685 | 0.262 | (105,104) |
| Smimu | 104 | lower | limit | 287 | lower | limit | 0.015* | 0.713 | (288,286) |
| Plate | 105 | 3.38 | 1.06 | 288 | 3.38 | 1.06 | 1.0 | 0.0 | (107,287) |
| Plama | 106 | 4.90 | 1.37 | 289 | 4.90 | 1.37 | 0.995 | 0.005 | (287,290) |
| Plagi | 107 | 1.57 | 0.89 | 290 | 1.57 | 0.89 | 0.366* | 0.417 | (109,289) |
| Plain | 108 | 5.32 | 1.50 | 291 | 5.32 | 1.50 | 0.986 | 0.014 | (289,285) |
| Didvi | 109 | 7.17 | 1.67 | 292 | 7.17 | 1.67 | 0.974 | 0.025 | (363,291) |
| Mondo | 110 | 1.54 | 0.79 | 293 | 1.54 | 0.79 | 0.627 | 0.192 | (113,112) |
| Geooc | 111 | 2.74 | 1.10 | 294 | 2.74 | 1.10 | 0.789 | 0.166 | (293,295) |
| Pezwa | 112 | 3.78 | 1.25 | 295 | 3.78 | 1.25 | 0.774 | 0.225 | (115,294) |
| Melun | 113 | 0.46 | 0.59 | 296 | 0.46 | 0.59 | 0.042* | 0.925 | (114,295) |
| Plaix | 114 | 2.41 | 1.09 | 297 | 2.41 | 1.09 | 0.582 | 0.383 | (296,117) |
| Polan | 115 | 2.58 | 1.08 | 298 | 2.58 | 1.08 | 0.811 | 0.180 | (316,117) |
| Strha | 116 | 1.70 | 0.78 | 299 | 1.70 | 0.78 | 0.891 | 0.107 | (119,300) |
| Calba | 117 | 1.42 | 0.76 | 300 | 1.42 | 0.76 | 0.878 | 0.076 | (299,122) |
| Neope | 118 | 0.69 | 0.59 | 301 | 0.69 | 0.59 | 0.289* | 0.681 | (123,300) |
| Catgu1 | 119 | 1.20 | 0.68 | 302 | 1.20 | 0.68 | 0.816 | 0.103 | (124,123) |
| Catgu2 | 120 | 0.98 | 0.56 | 303 | 0.98 | 0.56 | 0.740 | 0.132 | (307,124) |
| Pomru | 121 | 2.00 | 0.91 | 304 | 2.00 | 0.91 | 0.774 | 0.210 | (134,133) |
| Pomte | 122 | 2.56 | 1.06 | 305 | 2.56 | 1.06 | 0.937 | 0.052 | (304,126) |
| Pomis | 123 | 1.14 | 0.70 | 306 | 1.14 | 0.70 | 0.663 | 0.328 | (126,125) |
| Cyacr | 124 | 0.16 | 0.41 | 307 | 0.16 | 0.41 | 0.095* | 0.831 | (303,306) |
| Gymti | 125 | lower | limit | 308 | lower | limit | 0.233* | 0.729 | (303,312) |
| Manke | 126 | lower | limit | 309 | lower | limit | 0.092* | 0.528 | (129,128) |
| Epifa | 127 | 0.90 | 0.53 | 310 | 0.90 | 0.53 | 0.922 | 0.074 | (309,130) |
| Dipma | 128 | 2.11 | 0.83 | 311 | 2.11 | 0.83 | 0.916 | 0.083 | (131,130) |
| Epial | 129 | 0.86 | 0.66 | 312 | 0.86 | 0.66 | 0.692 | 0.231 | (311,308) |
| Ptipa | 130 | lower | limit | 313 | lower | limit | 0.012* | 0.750 | (137,312) |
| Ptipl | 131 | 1.31 | 0.70 | 314 | 1.31 | 0.70 | 0.799 | 0.171 | (313,136) |
| Ambma | 132 | 0.11 | 0.36 | 315 | 0.11 | 0.36 | 0.074* | 0.685 | (135,136) |
| Ptivi | 133 | 2.78 | 1.04 | 316 | 2.78 | 1.04 | 0.970 | 0.023 | (315,298) |
| Parin | 134 | 0.89 | 0.87 | 317 | 0.89 | 0.87 | 0.185* | 0.545 | (298,362) |
| Ailme | 135 | 1.32 | 0.68 | 318 | 1.32 | 0.68 | 0.813 | 0.186 | (118,324) |
| Virol | 136 | 1.00 | 0.63 | 319 | 1.00 | 0.63 | 0.860 | 0.064 | (323,166) |
| Lanlu | 137 | lower | limit | 320 | lower | limit | 0.020* | 0.944 | (180,179) |
| Ampst | 138 | 0.62 | 0.43 | 321 | 0.62 | 0.43 | 0.528 | 0.327 | (320,181) |
| Colru | 139 | lower | limit | 322 | lower | limit | 0.842 | 0.158 | (182,181) |
| Pitso | 140 | 2.87 | 0.91 | 323 | 2.87 | 0.91 | 1.0 | 0.0 | (319,182) |
| Scyma | 141 | lower | limit | 324 | lower | limit | 0.328* | 0.663 | (319,318) |
| Thrdo | 142 | 0.28 | 0.29 | 325 | 0.28 | 0.29 | 0.296* | 0.637 | (318,330) |
| Lopny | 143 | 1.06 | 0.61 | 326 | 1.06 | 0.61 | 0.807 | 0.161 | (170,329) |
| Pavcr | 144 | 0.65 | 0.46 | 327 | 0.65 | 0.46 | 0.813 | 0.145 | (173,177) |
| Galga | 145 | 1.35 | 0.68 | 328 | 1.35 | 0.68 | 0.891 | 0.088 | (327,174) |
| Cotco | 146 | 3.04 | 1.00 | 329 | 3.04 | 1.00 | 0.992 | 0.008 | (328,326) |
| Alech | 147 | 0.42 | 0.54 | 330 | 0.42 | 0.54 | 0.631 | 0.304 | (326,325) |
| Numme | 148 | 0.49 | 0.40 | 331 | 0.49 | 0.40 | 0.622 | 0.367 | (325,332) |
| Melga | 149 | 0.80 | 0.54 | 332 | 0.80 | 0.54 | 0.866 | 0.078 | (331,167) |
| Lopga | 150 | lower | limit | 333 | lower | limit | 0.096* | 0.685 | (331,334) |
| Cocam | 151 | 2.08 | 0.83 | 334 | 2.08 | 0.83 | 0.971 | 0.024 | (333,176) |
| Cocer | 152 | 0.59 | 0.42 | 335 | 0.59 | 0.42 | 0.803 | 0.129 | (336,334) |
| Crosu | 153 | 2.39 | 0.93 | 336 | 2.39 | 0.93 | 0.985 | 0.009 | (160,335) |
| Cucpa | 154 | lower | limit | 337 | lower | limit | 0.170* | 0.681 | (168,336) |
| Phacu | 155 | lower | limit | 338 | lower | limit | 0.105* | 0.770 | (165,168) |
| Piaca | 156 | 1.10 | 0.63 | 339 | 1.10 | 0.63 | 0.773 | 0.179 | (338,341) |
| OpihoA | 157 | 1.13 | 0.71 | 340 | 1.13 | 0.71 | 0.282* | 0.623 | (139,162) |
| OpihoB | 158 | 0.73 | 0.72 | 341 | 0.73 | 0.72 | 0.282* | 0.675 | (139,339) |
| OpihoC | 159 | 0.12 | 0.32 | 342 | 0.12 | 0.32 | 0.048* | 0.816 | (339,344) |
| Caimo | 160 | 1.34 | 0.69 | 343 | 1.34 | 0.69 | 0.837 | 0.103 | (158,159) |
| Ortve | 161 | 7.90 | 1.67 | 344 | 7.90 | 1.67 | 1.0 | 0.0 | (342,158) |
| Empmi | 162 | 1.14 | 0.74 | 345 | 1.14 | 0.74 | 0.562 | 0.370 | (342,346) |
| Tortr | 163 | 2.73 | 1.01 | 346 | 2.73 | 1.01 | 0.925 | 0.073 | (345,142) |
| Scoum | 164 | 0.26 | 0.60 | 347 | 0.26 | 0.60 | 0.041* | 0.884 | (138,346) |
| Peler | 165 | 1.24 | 0.73 | 348 | 1.24 | 0.73 | 0.499 | 0.306 | (347,355) |
| Balre | 166 | 0.96 | 0.54 | 349 | 0.96 | 0.54 | 0.690 | 0.305 | (147,144) |
| Phoru | 167 | 0.55 | 0.44 | 350 | 0.55 | 0.44 | 0.749 | 0.159 | (351,147) |
| Plaal | 168 | 0.28 | 0.29 | 351 | 0.28 | 0.29 | 0.623 | 0.308 | (350,146) |
| Gypba | 169 | 1.12 | 0.59 | 352 | 1.12 | 0.59 | 0.893 | 0.028 | (350,148) |
| Vulgr | 170 | 0.31 | 0.33 | 353 | 0.31 | 0.33 | 0.106* | 0.889 | (352,149) |
| Catbu | 171 | 0.90 | 0.67 | 354 | 0.90 | 0.67 | 0.707 | 0.244 | (353,150) |
| Corat | 172 | 3.49 | 1.07 | 355 | 3.49 | 1.07 | 0.995 | 0.005 | (354,348) |
| Gymca | 173 | 1.91 | 0.93 | 356 | 1.91 | 0.93 | 0.626 | 0.361 | (348,361) |
| Mycib | 174 | 3.62 | 1.19 | 357 | 3.62 | 1.19 | 0.878 | 0.122 | (153,155) |
| Mycam | 175 | 0.75 | 0.63 | 358 | 0.75 | 0.63 | 0.128* | 0.696 | (357,151) |
| Lepcr | 176 | 0.82 | 0.86 | 359 | 0.82 | 0.86 | 0.317* | 0.594 | (154,358) |
| Jabmy | 177 | 0.33 | 0.50 | 360 | 0.33 | 0.50 | 0.014* | 0.496 | (156,154) |
| Grurul | 178 | 6.29 | 1.56 | 361 | 6.29 | 1.56 | 0.999 | 0.001 | (360,356) |
| Gruan | 179 | 2.47 | 1.08 | 362 | 2.47 | 1.08 | 0.804 | 0.179 | (317,361) |
| Gruja | 180 | 11.46 | 2.08 | 363 | 11.46 | 2.08 | 1.0 | 0.0 | (317,292) |
| Gruru2 | 181 | 0.33 | 0.33 | 1 | | | | | |
| Gruvi | 182 | 1.89 | 0.72 | ln L: | -19376.24 | +- | 1019.84 | | |
| Xenla | 183 | 16.45 | 2.45 | AIC : | 39516.48 | | | | |

Figure 5.14: (b). Branch lengths and LBPs (estimated by ML) of the NJ tree of cytochrome b, part 2.

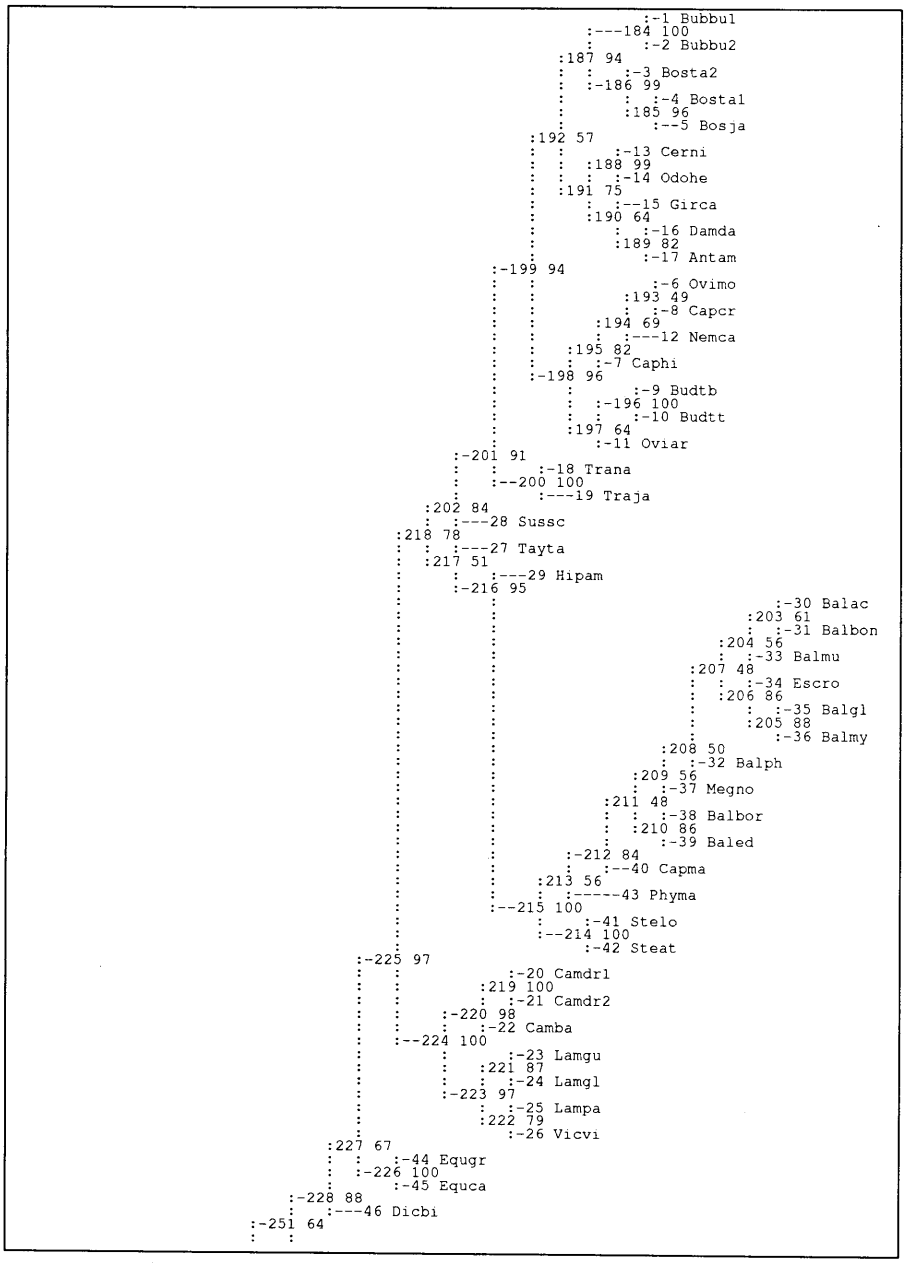



Figure 5.15: (a). The ML tree of cytochrome b, part 1.

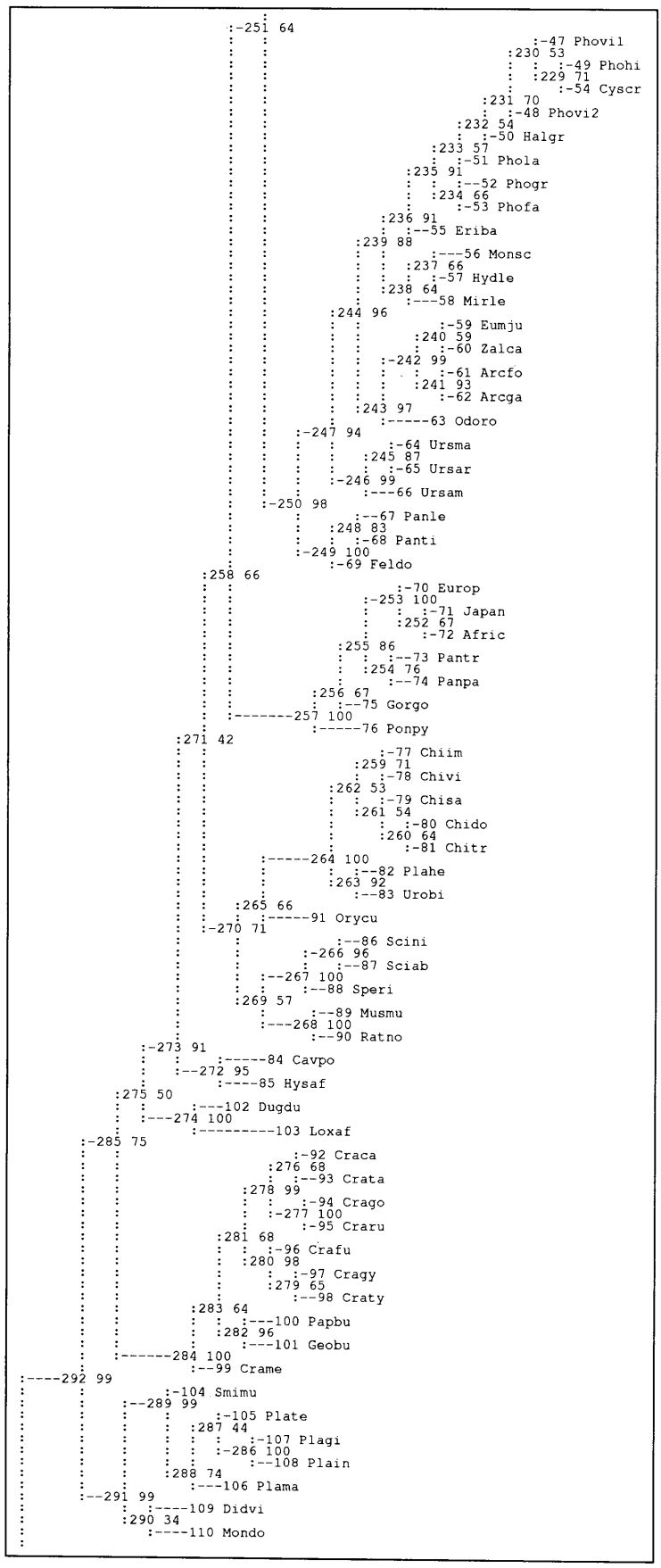



Figure 5.15: (b). The ML tree of cytochrome b, part 2.

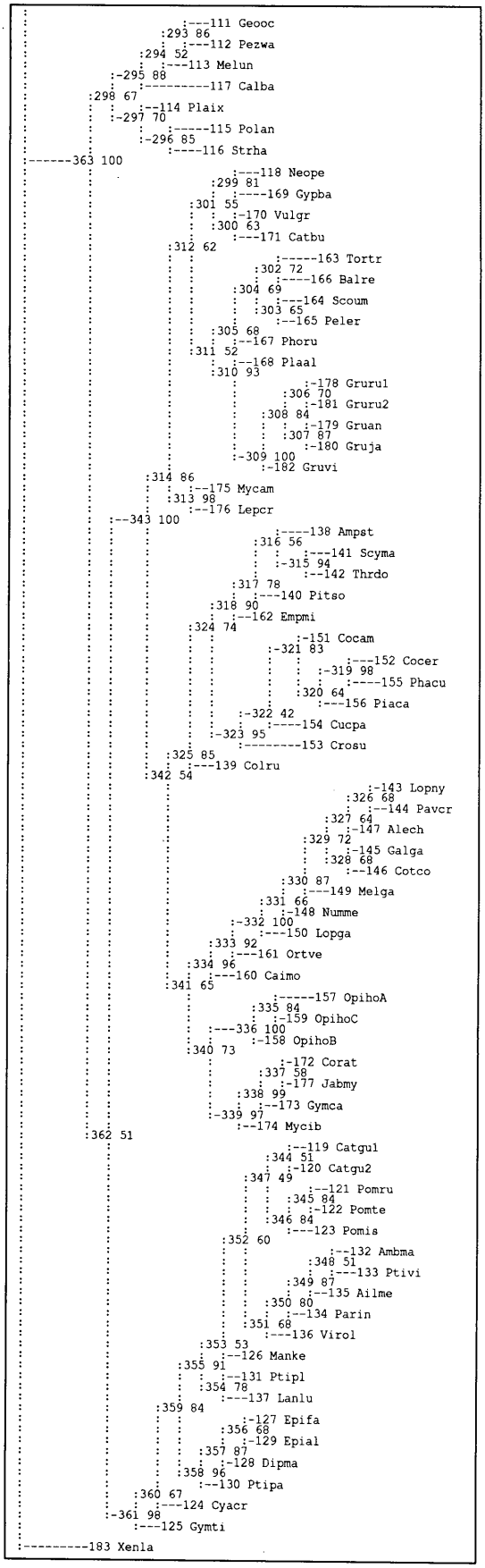



Figure 5.15: (c). The ML tree of cytochrome b, part 3.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Bubbu1 | 1 | 5.71 | 1.27 | 184 | 5.71 | 1.27 | 1.0 | 0.0 | (1,186) |
| Bubbu2 | 2 | 1.08 | 0.54 | 185 | 1.08 | 0.54 | 0.964 | 0.030 | (3,5) |
| Bosta2 | 3 | 2.37 | 0.83 | 186 | 2.37 | 0.83 | 0.988 | 0.007 | (3,184) |
| Bosta1 | 4 | 1.50 | 0.67 | 187 | 1.50 | 0.67 | 0.935 | 0.043 | (191,186) |
| Bosja | 5 | 1.59 | 0.68 | 188 | 1.59 | 0.68 | 0.993 | 0.004 | (13,190) |
| Ovimo | 6 | 0.94 | 0.52 | 189 | 0.94 | 0.52 | 0.819 | 0.166 | (16,15) |
| Caphi | 7 | 0.31 | 0.31 | 190 | 0.31 | 0.31 | 0.639 | 0.337 | (188,189) |
| Capcr | 8 | 0.62 | 0.44 | 191 | 0.62 | 0.44 | 0.746 | 0.216 | (188,187) |
| Budtb | 9 | 0.34 | 0.38 | 192 | 0.34 | 0.38 | 0.572 | 0.301 | (187,198) |
| Budtt | 10 | 0.06 | 0.28 | 193 | 0.06 | 0.28 | 0.490 | 0.478 | (6,12) |
| Oviar | 11 | 0.79 | 0.46 | 194 | 0.79 | 0.46 | 0.693 | 0.302 | (193,7) |
| Nemca | 12 | 1.30 | 0.63 | 195 | 1.30 | 0.63 | 0.820 | 0.175 | (197,7) |
| Cerni | 13 | 3.34 | 0.97 | 196 | 3.34 | 0.97 | 0.999 | 0.001 | (11,10) |
| Odohe | 14 | 0.76 | 0.48 | 197 | 0.76 | 0.48 | 0.640 | 0.202 | (196,195) |
| Girca | 15 | 2.35 | 0.84 | 198 | 2.35 | 0.84 | 0.959 | 0.018 | (195,192) |
| Damda | 16 | 1.87 | 0.83 | 199 | 1.87 | 0.83 | 0.944 | 0.053 | (200,198) |
| Antam | 17 | 3.94 | 1.13 | 200 | 3.94 | 1.13 | 0.999 | 0.001 | (199,19) |
| Trana | 18 | 2.69 | 0.92 | 201 | 2.69 | 0.92 | 0.913 | 0.078 | (28,200) |
| Traja | 19 | 0.88 | 0.60 | 202 | 0.88 | 0.60 | 0.836 | 0.123 | (217,28) |
| Camdr1 | 20 | 0.23 | 0.31 | 203 | 0.23 | 0.31 | 0.611 | 0.287 | (30,33) |
| Camdr2 | 21 | 0.44 | 0.37 | 204 | 0.44 | 0.37 | 0.559 | 0.361 | (203,206) |
| Camba | 22 | 0.82 | 0.47 | 205 | 0.82 | 0.47 | 0.881 | 0.104 | (35,34) |
| Lamgu | 23 | 0.58 | 0.42 | 206 | 0.58 | 0.42 | 0.855 | 0.079 | (204,205) |
| Lamgl | 24 | 0.50 | 0.38 | 207 | 0.50 | 0.38 | 0.485 | 0.467 | (204,32) |
| Lampa | 25 | 0.08 | 0.27 | 208 | 0.08 | 0.27 | 0.502 | 0.386 | (207,37) |
| Vicvi | 26 | 0.46 | 0.41 | 209 | 0.46 | 0.41 | 0.562 | 0.394 | (208,210) |
| Tayta | 27 | 1.53 | 0.68 | 210 | 1.53 | 0.68 | 0.863 | 0.134 | (38,209) |
| Sussc | 28 | 0.41 | 0.40 | 211 | 0.41 | 0.40 | 0.476 | 0.407 | (40,210) |
| Hipam | 29 | 1.96 | 0.82 | 212 | 1.96 | 0.82 | 0.843 | 0.110 | (211,43) |
| Balac | 30 | 1.20 | 0.69 | 213 | 1.20 | 0.69 | 0.556 | 0.373 | (214,43) |
| Balbon | 31 | 3.80 | 1.09 | 214 | 3.80 | 1.09 | 0.999 | 0.001 | (213,42) |
| Balph | 32 | 4.96 | 1.26 | 215 | 4.96 | 1.26 | 1.0 | 0.0 | (29,214) |
| Balmu | 33 | 2.85 | 0.97 | 216 | 2.85 | 0.97 | 0.950 | 0.045 | (27,215) |
| Escro | 34 | 1.08 | 0.59 | 217 | 1.08 | 0.59 | 0.506 | 0.487 | (27,202) |
| Balgl | 35 | 1.28 | 0.68 | 218 | 1.28 | 0.68 | 0.781 | 0.201 | (202,224) |
| Balmy | 36 | 1.64 | 0.67 | 219 | 1.64 | 0.67 | 0.998 | 0.002 | (22,21) |
| Megno | 37 | 2.00 | 0.80 | 220 | 2.00 | 0.80 | 0.975 | 0.012 | (223,22) |
| Balbor | 38 | 0.54 | 0.39 | 221 | 0.54 | 0.39 | 0.870 | 0.120 | (222,24) |
| Baled | 39 | 0.51 | 0.38 | 222 | 0.51 | 0.38 | 0.790 | 0.136 | (25,221) |
| Capma | 40 | 2.23 | 0.84 | 223 | 2.23 | 0.84 | 0.973 | 0.025 | (220,222) |
| Stelo | 41 | 3.79 | 1.09 | 224 | 3.79 | 1.09 | 0.999 | 0.001 | (220,218) |
| Steat | 42 | 2.38 | 0.87 | 225 | 2.38 | 0.87 | 0.973 | 0.020 | (226,224) |
| Phyma | 43 | 3.06 | 0.97 | 226 | 3.06 | 0.97 | 0.998 | 0.002 | (225,45) |
| Equgr | 44 | 0.87 | 0.56 | 227 | 0.87 | 0.56 | 0.670 | 0.306 | (46,226) |
| Equca | 45 | 2.11 | 0.86 | 228 | 2.11 | 0.86 | 0.885 | 0.113 | (250,46) |
| Dicbi | 46 | 0.41 | 0.38 | 229 | 0.41 | 0.38 | 0.713 | 0.251 | (49,47) |
| Phovi1 | 47 | 0.26 | 0.27 | 230 | 0.26 | 0.27 | 0.530 | 0.383 | (47,48) |
| Phovi2 | 48 | 0.30 | 0.30 | 231 | 0.30 | 0.30 | 0.697 | 0.235 | (50,48) |
| Phohi | 49 | lower | limit | 232 | lower | limit | 0.536 | 0.351 | (231,51) |
| Halgr | 50 | 0.43 | 0.37 | 233 | 0.43 | 0.37 | 0.574 | 0.375 | (232,234) |
| Phola | 51 | 0.36 | 0.34 | 234 | 0.36 | 0.34 | 0.664 | 0.254 | (52,233) |
| Phogr | 52 | 0.86 | 0.52 | 235 | 0.86 | 0.52 | 0.914 | 0.083 | (55,234) |
| Phofa | 53 | 0.93 | 0.55 | 236 | 0.93 | 0.55 | 0.913 | 0.039 | (235,238) |
| Cyscr | 54 | 0.26 | 0.27 | 237 | 0.26 | 0.27 | 0.657 | 0.306 | (58,57) |
| Eriba | 55 | 0.27 | 0.28 | 238 | 0.27 | 0.28 | 0.642 | 0.220 | (237,236) |
| Monsc | 56 | 1.09 | 0.56 | 239 | 1.09 | 0.56 | 0.875 | 0.116 | (236,243) |
| Hydle | 57 | 0.64 | 0.45 | 240 | 0.64 | 0.45 | 0.589 | 0.386 | (59,241) |
| Mirle | 58 | 1.05 | 0.57 | 241 | 1.05 | 0.57 | 0.930 | 0.048 | (61,240) |
| Eumju | 59 | 2.54 | 0.90 | 242 | 2.54 | 0.90 | 0.988 | 0.012 | (240,63) |
| Zalca | 60 | 1.70 | 0.73 | 243 | 1.70 | 0.73 | 0.971 | 0.022 | (242,239) |
| Arcfo | 61 | 1.54 | 0.69 | 244 | 1.54 | 0.69 | 0.963 | 0.025 | (239,246) |
| Arcga | 62 | 0.79 | 0.49 | 245 | 0.79 | 0.49 | 0.872 | 0.122 | (64,66) |
| Odoro | 63 | 2.01 | 0.79 | 246 | 2.01 | 0.79 | 0.991 | 0.009 | (244,66) |
| Ursma | 64 | 2.67 | 0.89 | 247 | 2.67 | 0.89 | 0.938 | 0.062 | (244,249) |
| Ursar | 65 | 0.88 | 0.53 | 248 | 0.88 | 0.53 | 0.829 | 0.129 | (69,68) |
| Ursam | 66 | 2.37 | 0.88 | 249 | 2.37 | 0.88 | 0.995 | 0.005 | (248,247) |
| Panle | 67 | 2.49 | 0.93 | 250 | 2.49 | 0.93 | 0.978 | 0.019 | (247,228) |
| Panti | 68 | 2.57 | 1.01 | 251 | 2.57 | 1.01 | 0.643 | 0.353 | (257,250) |
| Feldo | 69 | 0.26 | 0.27 | 252 | 0.26 | 0.27 | 0.670 | 0.223 | (71,70) |
| Europ | 70 | 3.49 | 1.00 | 253 | 3.49 | 1.00 | 1.0 | 0.0 | (70,254) |
| Japan | 71 | 0.83 | 0.53 | 254 | 0.83 | 0.53 | 0.765 | 0.184 | (253,74) |
| Afric | 72 | 1.81 | 0.76 | 255 | 1.81 | 0.76 | 0.863 | 0.102 | (253,75) |
| Pantr | 73 | 1.58 | 0.83 | 256 | 1.58 | 0.83 | 0.673 | 0.305 | (76,75) |
| Panpa | 74 | 13.88 | 2.13 | 257 | 13.88 | 2.13 | 1.0 | 0.0 | (256,251) |
| Gorgo | 75 | 1.50 | 0.85 | 258 | 1.50 | 0.85 | 0.655 | 0.304 | (270,257) |
| Ponpy | 76 | 0.54 | 0.41 | 259 | 0.54 | 0.41 | 0.711 | 0.215 | (77,261) |
| Chiim | 77 | 0.29 | 0.29 | 260 | 0.29 | 0.29 | 0.638 | 0.313 | (80,79) |
| Chivi | 78 | 0.50 | 0.37 | 261 | 0.50 | 0.37 | 0.536 | 0.456 | (259,260) |
| Chisa | 79 | 0.61 | 0.49 | 262 | 0.61 | 0.49 | 0.530 | 0.408 | (259,263) |
| Chido | 80 | 1.62 | 0.72 | 263 | 1.62 | 0.72 | 0.924 | 0.066 | (262,83) |
| Chitr | 81 | 10.93 | 1.85 | 264 | 10.93 | 1.85 | 1.0 | 0.0 | (262,91) |
| Plahe | 82 | 0.81 | 0.81 | 265 | 0.81 | 0.81 | 0.660 | 0.160 | (264,269) |
| Urobi | 83 | 3.04 | 1.01 | 266 | 3.04 | 1.01 | 0.965 | 0.019 | (86,88) |
| Cavpo | 84 | 4.91 | 1.32 | 267 | 4.91 | 1.32 | 1.0 | 0.0 | (268,88) |
| Hysaf | 85 | 7.45 | 1.54 | 268 | 7.45 | 1.54 | 1.0 | 0.0 | (89,267) |
| Scini | 86 | 1.63 | 0.79 | 269 | 1.63 | 0.79 | 0.572 | 0.275 | (267,265) |
| Sciab | 87 | 2.02 | 0.84 | 270 | 2.02 | 0.84 | 0.712 | 0.241 | (265,258) |
| Speri | 88 | 0.80 | 0.62 | 271 | 0.80 | 0.62 | 0.421 | 0.409 | (258,272) |
| Musmu | 89 | 3.75 | 1.16 | 272 | 3.75 | 1.16 | 0.954 | 0.033 | (271,85) |
| Ratno | 90 | 2.36 | 0.99 | 273 | 2.36 | 0.99 | 0.913 | 0.056 | (271,274) |
| Orycu | 91 | 6.13 | 1.50 | 274 | 6.13 | 1.50 | 1.0 | 0.0 | (273,103) |

Figure 5.16: (a). Branch lengths and LBPs of the ML tree of cytochrome b, part 1.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Orycu | 91 | 6.13 | 1.50 | 274 | 6.13 | 1.50 | 1.0 | 0.0 | (273,103) |
| Craca | 92 | 0.98 | 0.64 | 275 | 0.98 | 0.64 | 0.497 | 0.357 | (284,274) |
| Crata | 93 | 0.49 | 0.40 | 276 | 0.49 | 0.40 | 0.683 | 0.234 | (92,277) |
| Crago | 94 | 3.26 | 0.97 | 277 | 3.26 | 0.97 | 1.0 | 0.0 | (276,95) |
| Craru | 95 | 1.55 | 0.67 | 278 | 1.55 | 0.67 | 0.987 | 0.013 | (276,280) |
| Crafu | 96 | 0.30 | 0.30 | 279 | 0.30 | 0.30 | 0.650 | 0.259 | (97,96) |
| Cragy | 97 | 1.10 | 0.55 | 280 | 1.10 | 0.55 | 0.982 | 0.016 | (278,279) |
| Craty | 98 | 0.27 | 0.27 | 281 | 0.27 | 0.27 | 0.682 | 0.176 | (278,282) |
| Crame | 99 | 1.47 | 0.67 | 282 | 1.47 | 0.67 | 0.962 | 0.032 | (100,281) |
| Papbu | 100 | 0.44 | 0.46 | 283 | 0.44 | 0.46 | 0.638 | 0.211 | (281,99) |
| Geobu | 101 | 12.11 | 1.98 | 284 | 12.11 | 1.98 | 1.0 | 0.0 | (283,275) |
| Dugdu | 102 | 2.14 | 0.99 | 285 | 2.14 | 0.99 | 0.750 | 0.244 | (275,291) |
| Loxaf | 103 | 3.39 | 1.06 | 286 | 3.39 | 1.06 | 0.999 | 0.000 | (105,108) |
| Smimu | 104 | 0.12 | 0.33 | 287 | 0.12 | 0.33 | 0.438 | 0.342 | (105,106) |
| Plate | 105 | 0.91 | 0.61 | 288 | 0.91 | 0.61 | 0.744 | 0.197 | (287,104) |
| Plama | 106 | 4.48 | 1.30 | 289 | 4.48 | 1.30 | 0.991 | 0.009 | (104,290) |
| Plagi | 107 | 1.63 | 0.90 | 290 | 1.63 | 0.90 | 0.343 | 0.337 | (289,110) |
| Plain | 108 | 5.28 | 1.46 | 291 | 5.28 | 1.46 | 0.994 | 0.006 | (289,285) |
| Didvi | 109 | 8.03 | 1.75 | 292 | 8.03 | 1.75 | 0.994 | 0.005 | (363,291) |
| Mondo | 110 | 1.56 | 0.78 | 293 | 1.56 | 0.78 | 0.865 | 0.063 | (113,112) |
| Geooc | 111 | 0.85 | 0.69 | 294 | 0.85 | 0.69 | 0.516 | 0.412 | (117,113) |
| Pezwa | 112 | 2.04 | 0.94 | 295 | 2.04 | 0.94 | 0.882 | 0.079 | (294,297) |
| Melun | 113 | 3.16 | 1.11 | 296 | 3.16 | 1.11 | 0.846 | 0.131 | (114,116) |
| Plaix | 114 | 2.66 | 1.06 | 297 | 2.66 | 1.06 | 0.703 | 0.277 | (114,295) |
| Polan | 115 | 1.36 | 0.88 | 298 | 1.36 | 0.88 | 0.674 | 0.235 | (295,362) |
| Strha | 116 | 1.20 | 0.67 | 299 | 1.20 | 0.67 | 0.809 | 0.172 | (300,169) |
| Calba | 117 | 1.15 | 0.62 | 300 | 1.15 | 0.62 | 0.627 | 0.371 | (299,171) |
| Neope | 118 | 0.43 | 0.42 | 301 | 0.43 | 0.42 | 0.545 | 0.408 | (299,311) |
| Catgu1 | 119 | 0.60 | 0.52 | 302 | 0.60 | 0.52 | 0.723 | 0.170 | (163,303) |
| Catgu2 | 120 | 0.29 | 0.34 | 303 | 0.29 | 0.34 | 0.651 | 0.265 | (164,302) |
| Pomru | 121 | 0.30 | 0.30 | 304 | 0.30 | 0.30 | 0.693 | 0.274 | (167,303) |
| Pomte | 122 | 0.31 | 0.32 | 305 | 0.31 | 0.32 | 0.678 | 0.203 | (310,167) |
| Pomis | 123 | 0.27 | 0.27 | 306 | 0.27 | 0.27 | 0.700 | 0.141 | (178,307) |
| Cyacr | 124 | 0.54 | 0.38 | 307 | 0.54 | 0.38 | 0.869 | 0.026 | (179,306) |
| Gymti | 125 | 0.52 | 0.38 | 308 | 0.52 | 0.38 | 0.837 | 0.143 | (182,307) |
| Manke | 126 | 1.95 | 0.75 | 309 | 1.95 | 0.75 | 1.0 | 0.0 | (168,182) |
| Epifa | 127 | 1.07 | 0.55 | 310 | 1.07 | 0.55 | 0.929 | 0.071 | (168,305) |
| Dipma | 128 | 0.27 | 0.27 | 311 | 0.27 | 0.27 | 0.518 | 0.457 | (301,310) |
| Epial | 129 | lower | limit | 312 | lower | limit | 0.617 | 0.381 | (301,313) |
| Ptipa | 130 | 1.81 | 0.77 | 313 | 1.81 | 0.77 | 0.980 | 0.015 | (175,312) |
| Ptipl | 131 | 0.55 | 0.39 | 314 | 0.55 | 0.39 | 0.856 | 0.144 | (312,342) |
| Ambma | 132 | 1.88 | 0.83 | 315 | 1.88 | 0.83 | 0.941 | 0.038 | (138,142) |
| Ptivi | 133 | 0.31 | 0.52 | 316 | 0.31 | 0.52 | 0.559 | 0.353 | (138,140) |
| Parin | 134 | 1.48 | 0.76 | 317 | 1.48 | 0.76 | 0.781 | 0.215 | (316,162) |
| Ailme | 135 | 1.61 | 0.78 | 318 | 1.61 | 0.78 | 0.899 | 0.087 | (317,323) |
| Virol | 136 | 3.10 | 1.08 | 319 | 3.10 | 1.08 | 0.979 | 0.013 | (156,155) |
| Lanlu | 137 | 1.40 | 0.81 | 320 | 1.40 | 0.81 | 0.636 | 0.349 | (319,151) |
| Ampst | 138 | 2.80 | 1.10 | 321 | 2.80 | 1.10 | 0.830 | 0.160 | (154,320) |
| Colru | 139 | 2.14 | 1.00 | 322 | 2.14 | 1.00 | 0.419 | 0.326 | (321,153) |
| Pitso | 140 | 2.32 | 1.01 | 323 | 2.32 | 1.01 | 0.951 | 0.047 | (322,318) |
| Scyma | 141 | 0.73 | 0.61 | 324 | 0.73 | 0.61 | 0.739 | 0.142 | (139,323) |
| Thrdo | 142 | 0.91 | 0.62 | 325 | 0.91 | 0.62 | 0.849 | 0.151 | (341,139) |
| Lopny | 143 | 0.97 | 0.54 | 326 | 0.97 | 0.54 | 0.679 | 0.319 | (147,144) |
| Pavcr | 144 | 0.56 | 0.45 | 327 | 0.56 | 0.45 | 0.635 | 0.247 | (326,328) |
| Galga | 145 | 0.28 | 0.29 | 328 | 0.28 | 0.29 | 0.680 | 0.157 | (145,327) |
| Cotco | 146 | 0.53 | 0.40 | 329 | 0.53 | 0.40 | 0.724 | 0.242 | (327,149) |
| Alech | 147 | 1.13 | 0.61 | 330 | 1.13 | 0.61 | 0.871 | 0.109 | (329,148) |
| Numme | 148 | 1.48 | 0.69 | 331 | 1.48 | 0.69 | 0.664 | 0.330 | (150,148) |
| Melga | 149 | 2.92 | 0.99 | 332 | 2.92 | 0.99 | 0.995 | 0.002 | (331,161) |
| Lopga | 150 | 1.45 | 0.74 | 333 | 1.45 | 0.74 | 0.924 | 0.046 | (160,161) |
| Cocam | 151 | 1.59 | 0.74 | 334 | 1.59 | 0.74 | 0.960 | 0.028 | (340,160) |
| Cocer | 152 | 1.73 | 0.80 | 335 | 1.73 | 0.80 | 0.845 | 0.102 | (158,159) |
| Crosu | 153 | 6.32 | 1.48 | 336 | 6.32 | 1.48 | 1.0 | 0.0 | (339,158) |
| Cucpa | 154 | 0.59 | 0.43 | 337 | 0.59 | 0.43 | 0.580 | 0.401 | (173,177) |
| Phacu | 155 | 1.84 | 0.80 | 338 | 1.84 | 0.80 | 0.992 | 0.008 | (337,174) |
| Piaca | 156 | 2.01 | 0.82 | 339 | 2.01 | 0.82 | 0.969 | 0.028 | (336,174) |
| OpihoA | 157 | 0.82 | 0.57 | 340 | 0.82 | 0.57 | 0.730 | 0.223 | (336,334) |
| OpihoB | 158 | 0.31 | 0.32 | 341 | 0.31 | 0.32 | 0.653 | 0.344 | (325,340) |
| OpihoC | 159 | 0.26 | 0.27 | 342 | 0.26 | 0.27 | 0.544 | 0.349 | (325,314) |
| Caimo | 160 | 4.22 | 1.17 | 343 | 4.22 | 1.17 | 1.0 | 0.0 | (361,342) |
| Ortve | 161 | 1.33 | 0.66 | 344 | 1.33 | 0.66 | 0.512 | 0.488 | (119,346) |
| Empmi | 162 | 1.00 | 0.66 | 345 | 1.00 | 0.66 | 0.836 | 0.164 | (121,123) |
| Tortr | 163 | 1.87 | 0.86 | 346 | 1.87 | 0.86 | 0.836 | 0.150 | (345,344) |
| Scoum | 164 | 1.12 | 0.70 | 347 | 1.12 | 0.70 | 0.493 | 0.438 | (351,346) |
| Peler | 165 | lower | limit | 348 | lower | limit | 0.508 | 0.334 | (132,135) |
| Balre | 166 | 1.48 | 0.69 | 349 | 1.48 | 0.69 | 0.869 | 0.119 | (348,134) |
| Phoru | 167 | 1.03 | 0.68 | 350 | 1.03 | 0.68 | 0.795 | 0.109 | (136,134) |
| Plaal | 168 | 1.27 | 0.66 | 351 | 1.27 | 0.66 | 0.676 | 0.256 | (350,347) |
| Gypba | 169 | 0.94 | 0.61 | 352 | 0.94 | 0.61 | 0.603 | 0.350 | (347,126) |
| Vulgr | 170 | 0.75 | 0.52 | 353 | 0.75 | 0.52 | 0.533 | 0.423 | (354,126) |
| Catbu | 171 | 0.70 | 0.53 | 354 | 0.70 | 0.53 | 0.775 | 0.155 | (353,137) |
| Corat | 172 | 1.58 | 0.68 | 355 | 1.58 | 0.68 | 0.914 | 0.077 | (358,354) |
| Gymca | 173 | 0.44 | 0.48 | 356 | 0.44 | 0.48 | 0.682 | 0.264 | (128,129) |
| Mycib | 174 | 0.61 | 0.45 | 357 | 0.61 | 0.45 | 0.869 | 0.085 | (356,130) |
| Mycam | 175 | 1.77 | 0.72 | 358 | 1.77 | 0.72 | 0.956 | 0.044 | (355,130) |
| Lepcr | 176 | 1.00 | 0.61 | 359 | 1.00 | 0.61 | 0.836 | 0.090 | (124,358) |
| Jabmy | 177 | 0.70 | 0.61 | 360 | 0.70 | 0.61 | 0.673 | 0.204 | (359,125) |
| Grurul | 178 | 3.59 | 1.09 | 361 | 3.59 | 1.09 | 0.980 | 0.020 | (343,125) |
| Gruan | 179 | 0.65 | 0.62 | 362 | 0.65 | 0.62 | 0.514 | 0.240 | (298,361) |
| Gruja | 180 | 12.50 | 2.15 | 363 | 12.50 | 2.15 | 1.0 | 0.0 | (298,292) |
| Gruru2 | 181 | lower | limit | 1 | | | | | |
| Gruvi | 182 | 1.64 | 0.67 | ln L: | -19044.73 | +- 993.45 | | | |
| Xenla | 183 | 15.98 | 2.41 | AIC : | 38853.47 | | | | |

Figure 5.16: (b). Branch lengths and LBPs of the ML tree of cytochrome b, part 2.

**Phylogeny of Fishes**

Fig. 5.17 shows the NJ tree of cytochrome *b* from 31 OTUs of bony fishes and cartilaginous fishes with a lamprey as an outgroup. The distance matrix provided for the NJ analysis was estimated for 2-OTUs trees by the ProtML based on the JTT-F model. Starting from this tree, the search for better tree topologies by the likelihood criterion was conducted by repeating local rearrangements as described in subsection 3.5.3. Fig. 5.19 gives the ProtML tree (based on the JTT-F model) which cannot be improved any more by the local rearrangements. The log-likelihood of the NJ tree is $-4735.2$, while that of the resultant ProtML tree is $-4723.1$, and the two trees do not differ very much in their topology.

Osteichthyes (bony fishes) and Chondrichthyes (cartilaginous fishes) are clearly separated, and form two monophyletic clades respectively. Within Osteichthyes, Acipenseriformes is a sister group to the others with 100% LBP (branch 48). The order Cypriniformes may be paraphyletic, because *Cyprinus carpio* is closer to Cypriniformes fishes, which form a monophyletic clade with 99% LBP (branch 38), than to *Crossostoma lacustre*, and this relationship is supported wth 91% LBP (branch 39) in the ProtML tree. Perciformes is monophyletic with 100% LBP (branch 47). Within Perciformes, a ((*Sarda sarda*, *Thunnus thynnus*),*Scomber scombrus*) clade is supported with 100% LBP in accord with Cantatore et al. (1994[47]). A (*Oreochromis mossambicus*, *Trachurus trachurus*) clade is sister-group to all the other Perciformes fishes (95% LBP: branch 45). Salmoniformes might be a sister-group to Gadiformes (cod), but the support is not strong enough to draw conclusion (70% LBP: branch 33).

Within Chondrichthyes, Heterodontiformes is closer to Carcharhiniformes than to Lamniformes with 94% LBP (branch 60), and the outgroup status of Heterodontiformes to all the others has only 5% LBP. These three orders of Chondrichthyes are monophyletic, respectively, in accord with Martin and Palumbi (1993[206])

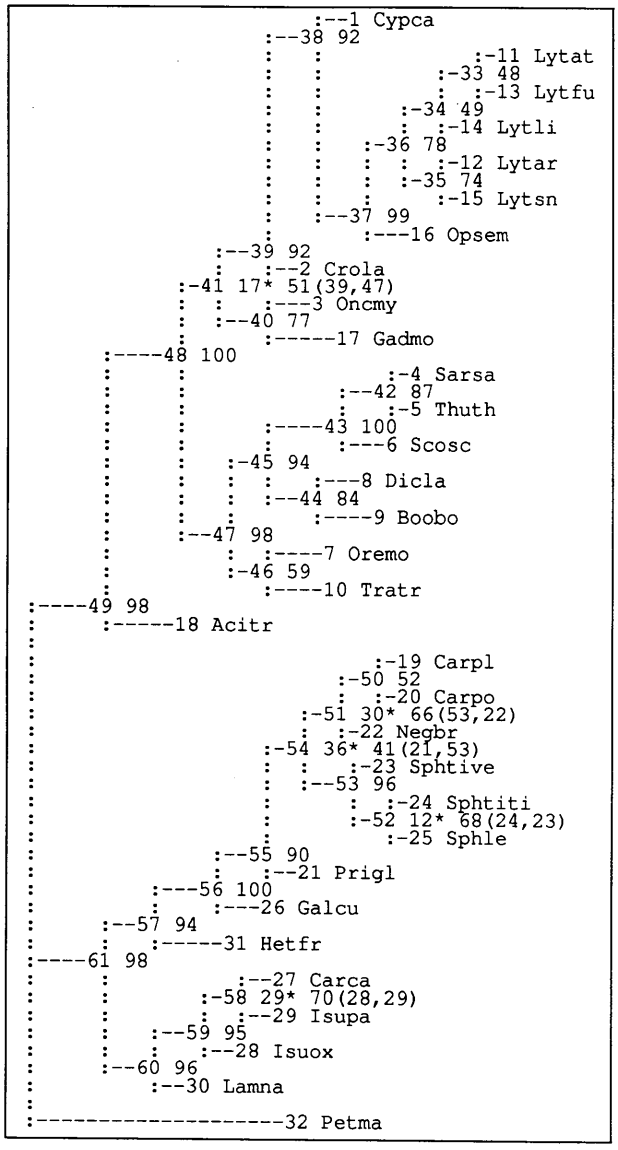



Figure 5.17: The NJ tree of cytochrome b from fishs.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Cypca | 1 | lower | limit | 33 | lower | limit | 0.481 | 0.281 | (11,14) |
| Crola | 2 | lower | limit | 34 | lower | limit | 0.486 | 0.363 | (35,14) |
| Oncmy | 3 | 0.26 | 0.27 | 35 | 0.26 | 0.27 | 0.735 | 0.167 | (12,34) |
| Sarsa | 4 | 0.76 | 0.57 | 36 | 0.76 | 0.57 | 0.781 | 0.186 | (34,16) |
| Thuth | 5 | 3.39 | 1.02 | 37 | 3.39 | 1.02 | 0.993 | 0.007 | (36,1) |
| Scosc | 6 | 2.01 | 0.83 | 38 | 2.01 | 0.83 | 0.918 | 0.061 | (1,2) |
| Oremo | 7 | 2.29 | 0.88 | 39 | 2.29 | 0.88 | 0.915 | 0.076 | (38,40) |
| Dicla | 8 | 1.82 | 0.85 | 40 | 1.82 | 0.85 | 0.772 | 0.225 | (3,39) |
| Boobo | 9 | 0.14 | 0.44 | 41 | 0.14 | 0.44 | 0.167* | 0.511 | (39,47) |
| Tratr | 10 | 3.04 | 1.02 | 42 | 3.04 | 1.02 | 0.866 | 0.133 | (4,6) |
| Lytat | 11 | 5.49 | 1.30 | 43 | 5.49 | 1.30 | 1.0 | 0.0 | (42,44) |
| Lytar | 12 | 1.78 | 0.80 | 44 | 1.78 | 0.80 | 0.841 | 0.152 | (43,9) |
| Lytfu | 13 | 1.67 | 0.76 | 45 | 1.67 | 0.76 | 0.941 | 0.043 | (43,46) |
| Lytli | 14 | 1.09 | 0.69 | 46 | 1.09 | 0.69 | 0.589 | 0.360 | (45,10) |
| Lytsn | 15 | 2.95 | 0.98 | 47 | 2.95 | 0.98 | 0.977 | 0.023 | (45,41) |
| Opsem | 16 | 5.34 | 1.40 | 48 | 5.34 | 1.40 | 0.995 | 0.005 | (41,18) |
| Gadmo | 17 | 5.69 | 1.55 | 49 | 5.69 | 1.55 | 0.983 | 0.015 | (48,61) |
| Acitr | 18 | 0.48 | 0.40 | 50 | 0.48 | 0.40 | 0.516 | 0.461 | (22,20) |
| Carpl | 19 | 0.33 | 0.33 | 51 | 0.33 | 0.33 | 0.300* | 0.658 | (53,22) |
| Carpo | 20 | lower | limit | 52 | lower | limit | 0.125* | 0.681 | (24,23) |
| Prigl | 21 | 1.87 | 0.74 | 53 | 1.87 | 0.74 | 0.964 | 0.035 | (51,52) |
| Negbr | 22 | 0.34 | 0.38 | 54 | 0.34 | 0.38 | 0.364* | 0.409 | (21,53) |
| Sphtive | 23 | 2.39 | 0.92 | 55 | 2.39 | 0.92 | 0.901 | 0.096 | (54,26) |
| Sphtiti | 24 | 3.96 | 1.17 | 56 | 3.96 | 1.17 | 0.996 | 0.003 | (31,26) |
| Sphle | 25 | 2.58 | 1.01 | 57 | 2.58 | 1.01 | 0.944 | 0.040 | (56,60) |
| Galcu | 26 | 0.96 | 0.57 | 58 | 0.96 | 0.57 | 0.291* | 0.704 | (28,29) |
| Carca | 27 | 2.43 | 0.95 | 59 | 2.43 | 0.95 | 0.946 | 0.051 | (58,30) |
| Isuox | 28 | 3.41 | 1.15 | 60 | 3.41 | 1.15 | 0.965 | 0.030 | (57,30) |
| Isupa | 29 | 5.53 | 1.57 | 61 | 5.53 | 1.57 | 0.983 | 0.017 | (57,49) |
| Lamna | 30 | 2.68 | 0.98 | 1 | | | | | |
| Hetfr | 31 | 7.43 | 1.55 | ln L: | -4730.33 | +- 260.38 | | | |
| Petma | 32 | 34.42 | 3.64 | AIC : | 9620.67 | | | | |

Figure 5.18: Branch lengths and LBPs (estimated by ML) of the NJ tree of cytochrome b from fishs.

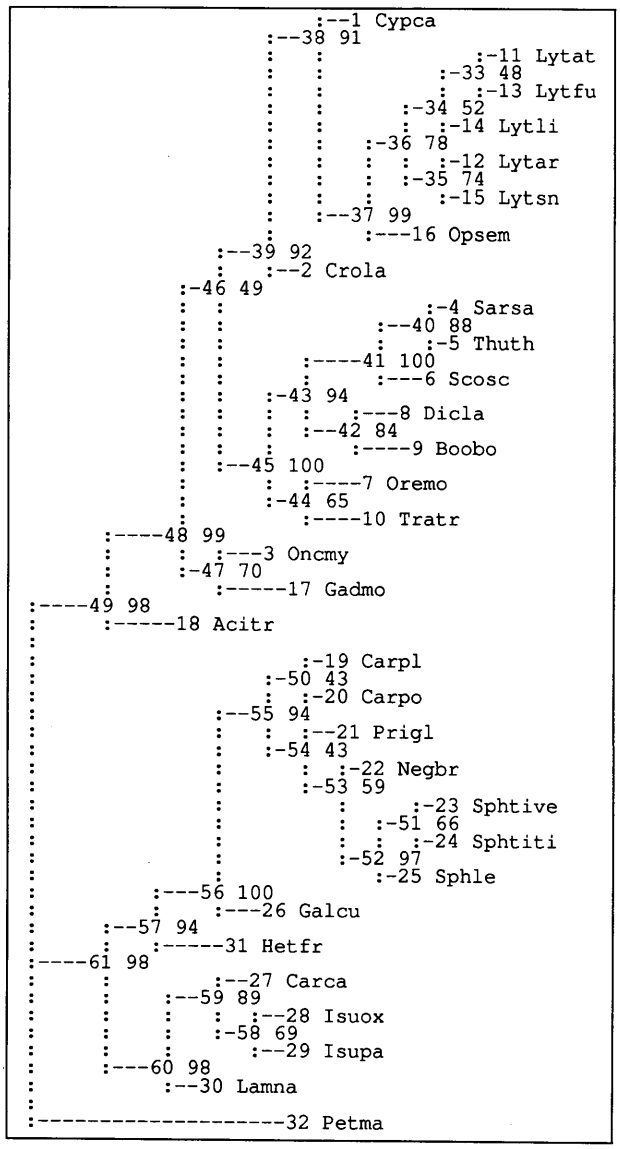



Figure 5.19: The ML tree of cytochrome b from fishs.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Cypca | 1 | lower | limit | 33 | lower | limit | 0.485 | 0.249 | (11,14) |
| Crola | 2 | lower | limit | 34 | lower | limit | 0.517 | 0.279 | (35,14) |
| Oncmy | 3 | 0.26 | 0.27 | 35 | 0.26 | 0.27 | 0.735 | 0.173 | (12,34) |
| Sarsa | 4 | 0.77 | 0.57 | 36 | 0.77 | 0.57 | 0.783 | 0.193 | (34,16) |
| Thuth | 5 | 3.39 | 1.02 | 37 | 3.39 | 1.02 | 0.993 | 0.007 | (36,1) |
| Scosc | 6 | 2.01 | 0.82 | 38 | 2.01 | 0.82 | 0.912 | 0.064 | (1,2) |
| Oremo | 7 | 2.17 | 0.86 | 39 | 2.17 | 0.86 | 0.915 | 0.084 | (38,45) |
| Dicla | 8 | 3.05 | 1.02 | 40 | 3.05 | 1.02 | 0.880 | 0.119 | (4,6) |
| Boobo | 9 | 5.48 | 1.30 | 41 | 5.48 | 1.30 | 1.0 | 0.0 | (40,42) |
| Tratr | 10 | 1.78 | 0.80 | 42 | 1.78 | 0.80 | 0.845 | 0.150 | (41,9) |
| Lytat | 11 | 1.69 | 0.75 | 43 | 1.69 | 0.75 | 0.943 | 0.038 | (41,44) |
| Lytar | 12 | 1.06 | 0.67 | 44 | 1.06 | 0.67 | 0.648 | 0.318 | (43,10) |
| Lytfu | 13 | 2.89 | 0.96 | 45 | 2.89 | 0.96 | 0.997 | 0.003 | (43,39) |
| Lytli | 14 | 0.36 | 0.42 | 46 | 0.36 | 0.42 | 0.490 | 0.321 | (47,45) |
| Lytsn | 15 | 1.72 | 0.82 | 47 | 1.72 | 0.82 | 0.705 | 0.276 | (46,17) |
| Opsem | 16 | 5.34 | 1.41 | 48 | 5.34 | 1.41 | 0.994 | 0.003 | (18,47) |
| Gadmo | 17 | 5.73 | 1.56 | 49 | 5.73 | 1.56 | 0.981 | 0.016 | (48,61) |
| Acitr | 18 | 0.43 | 0.38 | 50 | 0.43 | 0.38 | 0.428 | 0.388 | (19,54) |
| Carpl | 19 | 0.26 | 0.27 | 51 | 0.26 | 0.27 | 0.663 | 0.115 | (23,25) |
| Carpo | 20 | 1.29 | 0.63 | 52 | 1.29 | 0.63 | 0.967 | 0.016 | (22,25) |
| Prigl | 21 | 0.57 | 0.41 | 53 | 0.57 | 0.41 | 0.592 | 0.391 | (21,52) |
| Negbr | 22 | 0.30 | 0.31 | 54 | 0.30 | 0.31 | 0.431 | 0.466 | (50,53) |
| Sphtive | 23 | 2.33 | 0.89 | 55 | 2.33 | 0.89 | 0.945 | 0.046 | (26,54) |
| Sphtiti | 24 | 4.00 | 1.18 | 56 | 4.00 | 1.18 | 0.995 | 0.004 | (31,26) |
| Sphle | 25 | 2.59 | 1.02 | 57 | 2.59 | 1.02 | 0.937 | 0.040 | (56,60) |
| Galcu | 26 | 1.63 | 0.73 | 58 | 1.63 | 0.73 | 0.689 | 0.303 | (27,29) |
| Carca | 27 | 2.54 | 0.96 | 59 | 2.54 | 0.96 | 0.893 | 0.107 | (27,30) |
| Isuox | 28 | 3.70 | 1.20 | 60 | 3.70 | 1.20 | 0.976 | 0.021 | (57,30) |
| Isupa | 29 | 5.66 | 1.59 | 61 | 5.66 | 1.59 | 0.984 | 0.016 | (57,49) |
| Lamna | 30 | 1.90 | 0.85 | 1 | | | | | |
| Hetfr | 31 | 7.49 | 1.55 | ln L: | -4718.13 | +- 259.37 | | | |
| Petma | 32 | 34.42 | 3.64 | AIC : | 9596.27 | | | | |

Figure 5.20: Branch lengths and LBPs of the ML tree of cytochrome b from fishs.

### 5.5.2 Cytochrome Oxidase Subunit II from Mammalia

Sequence data used in the phylogenetic analysis are listed below.

| Abbrev. | Species name | Common name | Reference | Database |
|---|---|---|---|---|
| Homsa | *Homo sapiens* | human | Anderson'81[15] | V00662 |
| Pantr | *Pan troglodytes* | chimpanzee | Horai'95[140] | D38113 |
| Panpa1 | *Pan paniscus* | bonobo | Ruvolo'91[249] | M58009 |
| Panpa2 | *Pan paniscus* | bonobo | Horai'95[140] | D38116 |
| Gorgo1 | *Gorilla gorilla* | gorilla | Ruvolo'91[249] | M58006 |
| Gorgo2 | *Gorilla gorilla* | gorilla | Horai'95[140] | D38114 |
| Ponpy | *Pongo pygmaeus* | orangutan | Horai'95[140] | D38115 |
| Hylsy1 | *Hylobates syndactyllus* | siamang | Ruvolo'91[249] | M58007 |
| Hylsy2 | *Hylobates syndactyllus* | siamang | Horai'93[142] | |
| Macfa | *Macaca fascicularis* | crab-eating macaque | Ruvolo'91[249] | M58008 |
| Cerae | *Cercopithecus aethiops* | green monkey | Ruvolo'91[249] | M58005 |
| Macmu | *Macaca mulatta* | rhesus macaque | Disotell'92[67] | M74005 |
| Papan | *Papio anubis* | anubis baboon | Disotell'92[67] | M74007 |
| Thege | *Theropithecus gelada* | gelada | Disotell'92[67] | M74009 |
| Cerga | *Cercocebus galeritus* | agile mangabey | Disotell'92[67] | M74004 |
| Manle | *Mandrillus leucophaeus* | drill | Disotell'92[67] | M74006 |
| Alopa | *Alouatta palliata* | mantled howler | Adkins'94[10] | L22774 |
| Lagla | *Lagothrix lagothricha* | Humboldt's wooly monkey | Adkins'94[10] | L22779 |
| Tarba | *Tarsius bancanus* | western tarsier | Adkins'94[10] | L22783 |
| Tarsy | *Tarsius syrichta* | Philippine tarsier | Adkins'94[10] | L22784 |
| Eulma | *Eulemur macaco* | black-lemur | Adkins'94[10] | L22777 |
| Hapgr | *Hapalemur griseus* | gray gentle lemur | Adkins'94[10] | L22778 |
| Lemca | *Lemur catta* | ring-tailed lemur | Adkins'94[10] | L22780 |
| Varva | *Varecia variegata* | ruffed lemur | Adkins'94[10] | L22785 |
| Prota | *Propithecus tattersalli* | Tattersal's sifaka | Adkins'94[10] | L22782 |
| Chema | *Cheirogaleus medius* | fat-tailed mouse lemur | Adkins'94[10] | L22775 |
| Dauma | *Daubentonia madagascarensis* | aye-aye | Adkins'94[10] | L22776 |
| Nycco | *Nycticebus coucang* | slow loris | Adkins'94[10] | L22781 |
| Galse | *Galago senegalensis* | lesser bushbaby | Adkins'91[8] | M80905 |
| Cynva | *Cynocephalus variegatus* | Malayan flying lemur | Adkins'91[8] | M80904 |
| Tupgl | *Tupaia glis* | common tree shrew | Adkins'91[8] | M80907 |
| Phyha | *Phyllostomus hastatus* | greater spear-nosed bat | Adkins'91[8] | M80906 |
| Roule | *Rousettus leschenaulti* | Leschenault's rousette | Adkins'91[8] | M80908 |
| Bosta | *Bos taurus* | cow | Anderson'82[16] | V00654 |
| Balph | *Balaenoptera physalus* | fin whale | Arnason'91[24] | X61145 |
| Balmu | *Balaenoptera musculus* | blue whale | Arnason'93[21] | X72204 |
| Phovi | *Phoca vitulina* | harbor seal | Arnason'92[25] | X63726 |
| Halgr | *Halichoerus grypus* | grey seal | Arnason'93[23] | X72004 |
| Musmu | *Mus musculus* | mouse | Bibb'81[39] | V00711 |
| Ratno | *Rattus norvegicus* | rat | Gadaleta'89[87] | X14848 |
| Geoca | *Georychus capensis* | Cape mole-rat | Adkins'93[9] | |
| Dasno | *Dasypus novemcinctus* | nine-banded armadillo | Adkins'91[8] | M80903 |
| Didvi | *Didelphis virginiana* | Virginia opossum | Janke'94[150] | Z29573 |
| Galga | *Gallus gallus* | chicken | Desjardins'90[65] | X52392 |
| Xenla | *Xenopus laevis* | | Roe'85[248] | X02890 |

Figs. 5.22 and 5.24 show, respectively, the NJ tree from the distance matrix estimated by ML (the JTT-F model) and the ProtML tree obtained by replicating local rearrangements starting from the NJ tree. The log-likelihoods of the NJ and ML trees are $-19376.2$ and $-19044.7$, respectively, indicating improvement of log-likelihood by 331.5 through the replication of local rearrangements.

protst 46 OTUs 225 sites cox2

| Diff | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Hom | Pan | Pan | Pan | Gor | Gor | Pon | Hyl | Hyl | Mac | Mac | Pap | The | Cer | Man | Cer | Alo | Lag | Tar | Tar | Eul | Hap | Lem |
| 1 | Homsa | Hom | 5 | 6 | 5 | 6 | 8 | 12 | 13 | 12 | 27 | 27 | 26 | 25 | 24 | 23 | 24 | 67 | 66 | 59 | 57 | 63 | 64 | 64 |
| 2 | Pantr | 5 | Pan | 1 | 0 | 5 | 5 | 9 | 13 | 12 | 28 | 26 | 27 | 26 | 25 | 24 | 25 | 68 | 67 | 60 | 58 | 64 | 67 | 65 |
| 3 | Panpa1 | 6 | 1 | Pan | 1 | 6 | 6 | 10 | 14 | 13 | 29 | 27 | 28 | 27 | 26 | 25 | 26 | 68 | 67 | 61 | 59 | 65 | 67 | 66 |
| 4 | Panpa2 | 5 | 0 | 1 | Pan | 5 | 5 | 9 | 13 | 12 | 28 | 26 | 27 | 26 | 25 | 24 | 25 | 68 | 67 | 60 | 58 | 64 | 67 | 65 |
| 5 | Gorgo1 | 6 | 5 | 6 | 5 | Gor | 2 | 6 | 13 | 12 | 27 | 27 | 24 | 23 | 24 | 23 | 23 | 68 | 69 | 61 | 59 | 65 | 66 | 64 |
| 6 | Gorgo2 | 8 | 5 | 6 | 5 | 2 | Gor | 6 | 14 | 13 | 29 | 29 | 26 | 25 | 26 | 25 | 25 | 69 | 70 | 62 | 60 | 66 | 67 | 65 |
| 7 | Ponpy | 12 | 9 | 10 | 9 | 6 | 6 | Pon | 12 | 11 | 27 | 27 | 26 | 25 | 26 | 27 | 23 | 69 | 70 | 59 | 59 | 63 | 66 | 62 |
| 8 | Hylsy1 | 13 | 13 | 14 | 13 | 13 | 14 | 12 | Hyl | 1 | 25 | 25 | 24 | 23 | 22 | 23 | 20 | 67 | 66 | 57 | 57 | 61 | 64 | 60 |
| 9 | Hylsy2 | 12 | 12 | 13 | 12 | 12 | 13 | 11 | 1 | Hyl | 25 | 25 | 24 | 23 | 22 | 23 | 20 | 67 | 66 | 57 | 57 | 61 | 64 | 60 |
| 10 | Macfa | 27 | 28 | 29 | 28 | 27 | 29 | 27 | 25 | 25 | Mac | 6 | 12 | 14 | 11 | 14 | 13 | 78 | 79 | 70 | 69 | 67 | 71 | 68 |
| 11 | Macmu | 27 | 26 | 27 | 26 | 27 | 29 | 27 | 25 | 25 | 6 | Mac | 12 | 14 | 9 | 12 | 13 | 76 | 77 | 67 | 66 | 66 | 73 | 69 |
| 12 | Papan | 26 | 27 | 28 | 27 | 24 | 26 | 26 | 24 | 24 | 12 | 12 | Pap | 4 | 10 | 11 | 9 | 75 | 76 | 68 | 64 | 62 | 68 | 64 |
| 13 | Thege | 25 | 26 | 27 | 26 | 23 | 25 | 25 | 23 | 23 | 14 | 14 | 4 | The | 10 | 11 | 12 | 76 | 77 | 68 | 64 | 62 | 68 | 64 |
| 14 | Cerga | 24 | 25 | 26 | 25 | 24 | 26 | 26 | 22 | 22 | 11 | 9 | 10 | 10 | Cer | 5 | 10 | 76 | 77 | 68 | 66 | 67 | 72 | 68 |
| 15 | Manle | 23 | 24 | 25 | 24 | 23 | 25 | 27 | 23 | 23 | 14 | 12 | 11 | 11 | 5 | Man | 13 | 76 | 77 | 70 | 66 | 68 | 73 | 69 |
| 16 | Cerae | 24 | 25 | 26 | 25 | 23 | 25 | 23 | 20 | 20 | 13 | 13 | 9 | 12 | 10 | 13 | Cer | 77 | 77 | 66 | 64 | 64 | 70 | 66 |
| 17 | Alopa | 67 | 68 | 68 | 68 | 68 | 69 | 69 | 67 | 67 | 78 | 76 | 75 | 76 | 76 | 76 | 77 | Alo | 6 | 55 | 54 | 64 | 63 | 63 |
| 18 | Lagla | 66 | 67 | 67 | 67 | 69 | 70 | 70 | 66 | 66 | 79 | 77 | 76 | 77 | 77 | 77 | 77 | 6 | Lag | 53 | 53 | 62 | 61 | 61 |
| 19 | Tarba | 59 | 60 | 61 | 60 | 61 | 62 | 59 | 57 | 57 | 70 | 67 | 68 | 68 | 68 | 70 | 66 | 55 | 53 | Tar | 15 | 30 | 31 | 31 |
| 20 | Tarsy | 57 | 58 | 59 | 58 | 59 | 60 | 59 | 57 | 57 | 69 | 66 | 64 | 64 | 66 | 66 | 64 | 54 | 53 | 15 | Tar | 35 | 35 | 36 |
| 21 | Eulma | 63 | 64 | 65 | 64 | 65 | 66 | 63 | 61 | 61 | 67 | 66 | 62 | 62 | 67 | 68 | 64 | 64 | 62 | 30 | 35 | Eul | 11 | 10 |
| 22 | Hapgr | 64 | 67 | 67 | 67 | 66 | 67 | 66 | 64 | 64 | 71 | 73 | 68 | 68 | 72 | 73 | 70 | 63 | 61 | 31 | 35 | 11 | Hap | 8 |
| 23 | Lemca | 64 | 65 | 66 | 65 | 64 | 65 | 62 | 60 | 60 | 68 | 69 | 64 | 64 | 68 | 69 | 66 | 63 | 61 | 31 | 36 | 10 | 8 | Lem |
| 24 | Varva | 59 | 60 | 61 | 60 | 62 | 63 | 60 | 57 | 57 | 64 | 65 | 61 | 61 | 65 | 66 | 63 | 59 | 57 | 32 | 36 | 14 | 14 | 14 |
| 25 | Prota | 65 | 66 | 66 | 66 | 67 | 68 | 67 | 65 | 65 | 74 | 73 | 71 | 71 | 74 | 75 | 72 | 59 | 57 | 26 | 30 | 16 | 8 | 14 |
| 26 | Cheme | 65 | 66 | 67 | 66 | 67 | 68 | 65 | 63 | 63 | 70 | 71 | 67 | 67 | 72 | 73 | 68 | 61 | 59 | 31 | 35 | 13 | 12 | 14 |
| 27 | Dauma | 61 | 64 | 64 | 64 | 63 | 64 | 61 | 59 | 60 | 73 | 72 | 67 | 67 | 70 | 69 | 68 | 64 | 64 | 39 | 38 | 34 | 34 | 34 |
| 28 | Nycco | 67 | 67 | 68 | 67 | 68 | 69 | 67 | 65 | 65 | 73 | 74 | 70 | 70 | 75 | 75 | 71 | 65 | 62 | 37 | 38 | 27 | 27 | 30 |
| 29 | Galse | 64 | 65 | 66 | 65 | 67 | 68 | 66 | 63 | 63 | 73 | 74 | 70 | 70 | 75 | 76 | 71 | 70 | 67 | 42 | 45 | 30 | 31 | 31 |
| 30 | Cynva | 60 | 61 | 61 | 61 | 62 | 63 | 63 | 59 | 59 | 70 | 68 | 64 | 63 | 68 | 68 | 65 | 51 | 51 | 34 | 31 | 44 | 44 | 44 |
| 31 | Tupgl | 61 | 60 | 61 | 60 | 59 | 60 | 59 | 59 | 59 | 69 | 66 | 63 | 63 | 66 | 66 | 65 | 55 | 55 | 20 | 23 | 34 | 35 | 34 |
| 32 | Phyha | 69 | 67 | 68 | 67 | 66 | 67 | 66 | 66 | 66 | 73 | 71 | 68 | 69 | 71 | 71 | 70 | 62 | 63 | 31 | 32 | 44 | 44 | 43 |
| 33 | Roule | 64 | 65 | 66 | 65 | 64 | 65 | 64 | 64 | 64 | 71 | 70 | 66 | 66 | 70 | 70 | 68 | 52 | 54 | 24 | 21 | 37 | 36 | 37 |
| 34 | Bosta | 61 | 62 | 63 | 62 | 61 | 62 | 61 | 61 | 61 | 70 | 68 | 65 | 65 | 68 | 68 | 67 | 53 | 54 | 19 | 18 | 31 | 31 | 31 |
| 35 | Balph | 60 | 61 | 62 | 61 | 64 | 65 | 64 | 62 | 62 | 71 | 69 | 69 | 69 | 69 | 69 | 69 | 55 | 54 | 18 | 18 | 33 | 33 | 33 |
| 36 | Balmu | 60 | 61 | 62 | 61 | 64 | 65 | 64 | 62 | 62 | 71 | 69 | 69 | 69 | 69 | 69 | 69 | 54 | 53 | 17 | 18 | 34 | 34 | 34 |
| 37 | Phovi | 59 | 60 | 61 | 60 | 61 | 62 | 61 | 58 | 59 | 71 | 69 | 66 | 66 | 69 | 69 | 67 | 53 | 52 | 18 | 19 | 33 | 33 | 33 |
| 38 | Halgr | 58 | 59 | 60 | 59 | 60 | 61 | 60 | 57 | 58 | 70 | 68 | 65 | 65 | 68 | 68 | 66 | 53 | 52 | 18 | 19 | 32 | 32 | 32 |
| 39 | Cansi | 60 | 61 | 62 | 61 | 62 | 63 | 62 | 60 | 60 | 71 | 68 | 64 | 64 | 69 | 69 | 67 | 54 | 53 | 20 | 20 | 36 | 36 | 36 |
| 40 | Musmu | 64 | 65 | 66 | 65 | 64 | 65 | 64 | 64 | 64 | 72 | 71 | 66 | 66 | 71 | 71 | 68 | 51 | 51 | 21 | 23 | 35 | 37 | 37 |
| 41 | Ratno | 64 | 65 | 66 | 65 | 64 | 65 | 64 | 64 | 64 | 70 | 69 | 64 | 64 | 69 | 69 | 66 | 53 | 53 | 21 | 23 | 33 | 37 | 37 |
| 42 | Geoca | 66 | 67 | 68 | 67 | 68 | 69 | 66 | 64 | 64 | 74 | 71 | 68 | 68 | 71 | 71 | 69 | 57 | 56 | 27 | 23 | 37 | 39 | 39 |
| 43 | Dasno | 60 | 61 | 60 | 61 | 62 | 63 | 62 | 59 | 59 | 68 | 69 | 65 | 65 | 69 | 69 | 65 | 55 | 53 | 17 | 16 | 29 | 27 | 30 |
| 44 | Didvi | 72 | 73 | 73 | 73 | 72 | 73 | 70 | 69 | 69 | 77 | 76 | 72 | 72 | 76 | 76 | 73 | 58 | 61 | 38 | 35 | 41 | 45 | 46 |
| 45 | Galga | 76 | 79 | 79 | 79 | 76 | 76 | 78 | 79 | 79 | 87 | 85 | 84 | 81 | 82 | 83 | 85 | 79 | 80 | 71 | 70 | 76 | 74 | 74 |
| 46 | Xenla | 71 | 73 | 73 | 73 | 74 | 75 | 72 | 69 | 69 | 75 | 74 | 76 | 76 | 73 | 76 | 76 | 67 | 66 | 50 | 52 | 61 | 60 | 59 |

| Diff | | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Var | Pro | Che | Dau | Nyc | Gal | Cyn | Tup | Phy | Rou | Bos | Bal | Bal | Pho | Hal | Can | Mus | Rat | Geo | Das | Did | Gal | Xen |
| 1 | Homsa | 59 | 65 | 65 | 61 | 67 | 64 | 60 | 61 | 69 | 64 | 61 | 60 | 60 | 59 | 58 | 60 | 64 | 64 | 66 | 60 | 72 | 76 | 71 |
| 2 | Pantr | 60 | 66 | 66 | 64 | 67 | 65 | 61 | 60 | 67 | 65 | 62 | 61 | 61 | 60 | 59 | 61 | 65 | 65 | 67 | 61 | 73 | 79 | 73 |
| 3 | Panpa1 | 61 | 66 | 67 | 64 | 68 | 66 | 61 | 61 | 68 | 66 | 63 | 62 | 62 | 61 | 60 | 62 | 66 | 66 | 68 | 60 | 73 | 79 | 73 |
| 4 | Panpa2 | 60 | 66 | 66 | 64 | 67 | 65 | 61 | 60 | 67 | 65 | 62 | 61 | 61 | 60 | 59 | 61 | 65 | 65 | 67 | 61 | 73 | 79 | 73 |
| 5 | Gorgo1 | 62 | 67 | 67 | 63 | 68 | 67 | 62 | 59 | 66 | 64 | 61 | 64 | 64 | 61 | 60 | 62 | 64 | 64 | 68 | 62 | 72 | 76 | 74 |
| 6 | Gorgo2 | 63 | 68 | 68 | 64 | 69 | 68 | 63 | 60 | 67 | 65 | 62 | 65 | 65 | 62 | 61 | 63 | 65 | 65 | 69 | 63 | 73 | 76 | 75 |
| 7 | Ponpy | 60 | 67 | 65 | 61 | 67 | 66 | 63 | 59 | 66 | 64 | 61 | 64 | 64 | 61 | 60 | 62 | 64 | 64 | 66 | 62 | 70 | 78 | 72 |
| 8 | Hylsy1 | 57 | 65 | 63 | 59 | 65 | 63 | 59 | 59 | 66 | 64 | 61 | 62 | 62 | 58 | 57 | 60 | 64 | 64 | 64 | 59 | 69 | 79 | 69 |
| 9 | Hylsy2 | 57 | 65 | 63 | 60 | 65 | 63 | 59 | 59 | 66 | 64 | 61 | 62 | 62 | 59 | 58 | 60 | 64 | 64 | 64 | 59 | 69 | 79 | 69 |
| 10 | Macfa | 64 | 74 | 70 | 73 | 73 | 73 | 70 | 69 | 73 | 71 | 70 | 71 | 71 | 71 | 70 | 71 | 72 | 70 | 74 | 68 | 77 | 87 | 75 |
| 11 | Macmu | 65 | 73 | 71 | 72 | 74 | 74 | 68 | 66 | 71 | 70 | 68 | 69 | 69 | 69 | 68 | 68 | 71 | 69 | 71 | 69 | 76 | 85 | 74 |
| 12 | Papan | 61 | 71 | 67 | 67 | 70 | 70 | 64 | 63 | 68 | 66 | 65 | 69 | 69 | 66 | 65 | 64 | 66 | 64 | 68 | 65 | 72 | 84 | 76 |
| 13 | Thege | 61 | 71 | 67 | 67 | 70 | 70 | 63 | 63 | 69 | 66 | 65 | 69 | 69 | 66 | 65 | 64 | 66 | 64 | 68 | 65 | 72 | 81 | 76 |
| 14 | Cerga | 65 | 74 | 72 | 70 | 75 | 75 | 68 | 66 | 71 | 70 | 68 | 69 | 69 | 69 | 68 | 69 | 71 | 69 | 71 | 69 | 76 | 82 | 73 |
| 15 | Manle | 66 | 75 | 73 | 69 | 75 | 76 | 68 | 66 | 71 | 70 | 68 | 69 | 69 | 69 | 68 | 69 | 71 | 69 | 71 | 69 | 76 | 83 | 76 |
| 16 | Cerae | 63 | 72 | 68 | 68 | 71 | 71 | 65 | 65 | 70 | 68 | 67 | 69 | 69 | 67 | 66 | 67 | 68 | 66 | 69 | 65 | 73 | 85 | 76 |
| 17 | Alopa | 59 | 59 | 61 | 64 | 65 | 70 | 51 | 55 | 62 | 52 | 53 | 55 | 54 | 53 | 53 | 54 | 51 | 53 | 57 | 55 | 58 | 79 | 67 |
| 18 | Lagla | 57 | 57 | 59 | 64 | 62 | 67 | 51 | 55 | 63 | 54 | 54 | 54 | 53 | 52 | 52 | 53 | 51 | 53 | 56 | 53 | 61 | 80 | 66 |
| 19 | Tarba | 32 | 26 | 31 | 39 | 37 | 42 | 34 | 20 | 31 | 24 | 19 | 18 | 17 | 18 | 18 | 20 | 21 | 21 | 27 | 17 | 38 | 71 | 50 |
| 20 | Tarsy | 36 | 30 | 35 | 38 | 38 | 45 | 31 | 23 | 32 | 21 | 18 | 18 | 18 | 19 | 19 | 20 | 23 | 23 | 23 | 16 | 35 | 70 | 52 |
| 21 | Eulma | 14 | 16 | 13 | 34 | 27 | 30 | 44 | 34 | 44 | 37 | 31 | 33 | 34 | 33 | 32 | 36 | 35 | 33 | 37 | 29 | 41 | 76 | 61 |
| 22 | Hapgr | 14 | 8 | 12 | 34 | 27 | 31 | 44 | 35 | 44 | 36 | 31 | 33 | 34 | 33 | 32 | 36 | 37 | 37 | 39 | 27 | 45 | 74 | 60 |
| 23 | Lemca | 14 | 14 | 14 | 34 | 30 | 31 | 44 | 34 | 43 | 37 | 31 | 33 | 34 | 33 | 32 | 36 | 37 | 37 | 39 | 30 | 46 | 74 | 59 |
| 24 | Varva | Var | 18 | 14 | 33 | 28 | 29 | 43 | 36 | 46 | 40 | 33 | 35 | 36 | 35 | 34 | 37 | 39 | 39 | 41 | 30 | 44 | 74 | 61 |
| 25 | Prota | 18 | Pro | 13 | 36 | 28 | 31 | 41 | 32 | 43 | 33 | 26 | 28 | 29 | 28 | 29 | 32 | 34 | 34 | 36 | 25 | 43 | 70 | 58 |
| 26 | Cheme | 14 | 13 | Che | 35 | 21 | 23 | 44 | 34 | 42 | 38 | 30 | 32 | 33 | 31 | 32 | 35 | 36 | 35 | 38 | 29 | 43 | 75 | 63 |
| 27 | Dauma | 33 | 36 | 35 | Dau | 39 | 43 | 45 | 39 | 45 | 36 | 32 | 37 | 36 | 35 | 34 | 38 | 42 | 42 | 39 | 37 | 46 | 78 | 61 |
| 28 | Nycco | 28 | 28 | 21 | 39 | Nyc | 16 | 49 | 39 | 43 | 42 | 35 | 39 | 40 | 38 | 39 | 40 | 42 | 42 | 42 | 36 | 47 | 80 | 69 |
| 29 | Galse | 29 | 31 | 23 | 43 | 16 | Gal | 53 | 45 | 51 | 49 | 42 | 45 | 46 | 43 | 44 | 46 | 48 | 48 | 49 | 43 | 54 | 76 | 70 |
| 30 | Cynva | 43 | 41 | 44 | 45 | 49 | 53 | Cyn | 33 | 44 | 36 | 37 | 37 | 37 | 33 | 32 | 33 | 35 | 36 | 39 | 29 | 47 | 69 | 62 |
| 31 | Tupgl | 36 | 32 | 34 | 39 | 39 | 45 | 33 | Tup | 29 | 27 | 21 | 22 | 22 | 22 | 22 | 26 | 24 | 24 | 26 | 26 | 37 | 70 | 53 |
| 32 | Phyha | 46 | 43 | 42 | 45 | 43 | 51 | 44 | 29 | Phy | 26 | 25 | 27 | 26 | 28 | 28 | 29 | 33 | 34 | 36 | 32 | 46 | 77 | 60 |
| 33 | Roule | 40 | 33 | 38 | 36 | 42 | 49 | 36 | 27 | 26 | Rou | 17 | 21 | 20 | 22 | 22 | 23 | 23 | 25 | 29 | 24 | 37 | 72 | 56 |
| 34 | Bosta | 33 | 26 | 30 | 32 | 35 | 42 | 37 | 21 | 25 | 17 | Bos | 11 | 9 | 12 | 12 | 14 | 21 | 21 | 29 | 20 | 33 | 73 | 53 |
| 35 | Balph | 35 | 28 | 32 | 37 | 39 | 45 | 37 | 22 | 27 | 21 | 11 | Bal | 3 | 16 | 16 | 18 | 25 | 25 | 31 | 21 | 38 | 72 | 46 |
| 36 | Balmu | 36 | 29 | 33 | 36 | 40 | 46 | 37 | 22 | 26 | 20 | 9 | 3 | Bal | 15 | 15 | 16 | 23 | 23 | 30 | 19 | 37 | 72 | 47 |
| 37 | Phovi | 35 | 28 | 31 | 35 | 38 | 43 | 33 | 22 | 28 | 22 | 12 | 16 | 15 | Pho | 1 | 11 | 25 | 25 | 30 | 21 | 40 | 75 | 56 |
| 38 | Halgr | 34 | 29 | 32 | 34 | 39 | 44 | 32 | 22 | 28 | 22 | 12 | 16 | 15 | 1 | Hal | 10 | 25 | 25 | 30 | 20 | 40 | 76 | 55 |
| 39 | Cansi | 37 | 32 | 35 | 38 | 40 | 46 | 33 | 26 | 29 | 23 | 14 | 18 | 16 | 11 | 10 | Can | 22 | 22 | 31 | 22 | 39 | 75 | 56 |
| 40 | Musmu | 39 | 34 | 36 | 42 | 42 | 48 | 35 | 24 | 33 | 23 | 21 | 25 | 23 | 25 | 25 | 22 | Mus | 3 | 25 | 25 | 30 | 74 | 55 |
| 41 | Ratno | 39 | 34 | 35 | 42 | 42 | 48 | 36 | 24 | 34 | 25 | 21 | 25 | 23 | 25 | 25 | 22 | 3 | Rat | 27 | 25 | 30 | 74 | 57 |
| 42 | Geoca | 41 | 36 | 38 | 39 | 42 | 49 | 39 | 26 | 36 | 29 | 29 | 31 | 30 | 30 | 30 | 31 | 25 | 27 | Geo | 30 | 38 | 75 | 60 |
| 43 | Dasno | 30 | 25 | 29 | 37 | 36 | 43 | 29 | 26 | 32 | 24 | 20 | 21 | 19 | 21 | 20 | 22 | 25 | 25 | 30 | Das | 37 | 71 | 51 |
| 44 | Didvi | 44 | 43 | 43 | 46 | 47 | 54 | 47 | 37 | 46 | 37 | 33 | 38 | 37 | 40 | 40 | 39 | 30 | 30 | 38 | 37 | Did | 76 | 64 |
| 45 | Galga | 74 | 70 | 75 | 78 | 80 | 76 | 69 | 70 | 77 | 72 | 73 | 72 | 72 | 75 | 76 | 75 | 74 | 74 | 75 | 71 | 76 | Gal | 66 |
| 46 | Xenla | 61 | 58 | 63 | 61 | 69 | 70 | 62 | 53 | 60 | 56 | 53 | 46 | 47 | 56 | 55 | 56 | 55 | 57 | 60 | 51 | 64 | 66 | Xen |

Figure 5.21: Number of amino acid diffrences of cytochrome oxidase subunit II.

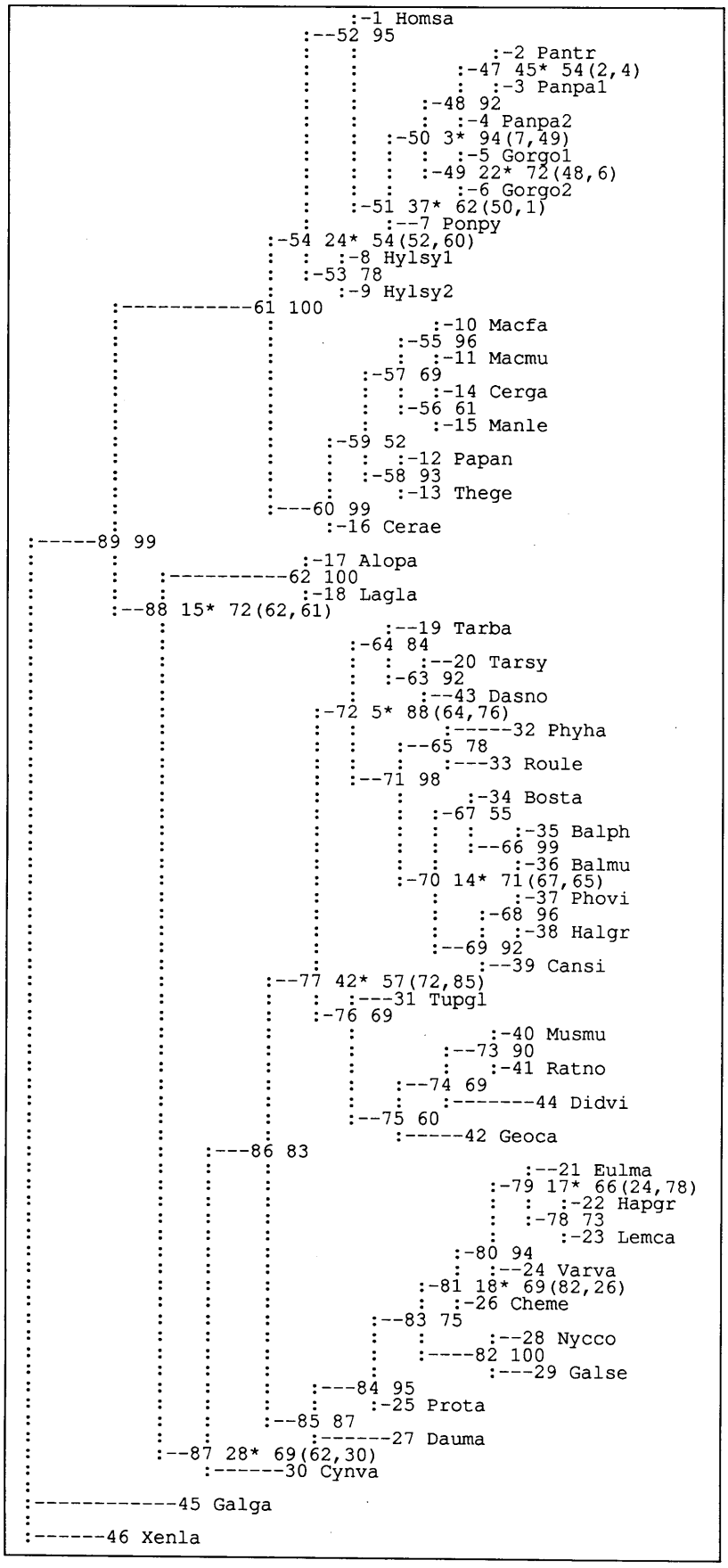



Figure 5.22: The NJ tree of cytochrome oxidase subunit II.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Homsa | 1 | lower | limit | 47 | lower | limit | 0.452* | 0.540 | (2,4) |
| Pantr | 2 | 1.78 | 0.89 | 48 | 1.78 | 0.89 | 0.922 | 0.053 | (47,49) |
| Panpa1 | 3 | lower | limit | 49 | lower | limit | 0.221* | 0.723 | (48,6) |
| Panpa2 | 4 | lower | limit | 50 | lower | limit | 0.028* | 0.940 | (7,49) |
| Gorgo1 | 5 | 1.78 | 0.91 | 51 | 1.78 | 0.91 | 0.372* | 0.625 | (50,1) |
| Gorgo2 | 6 | 2.26 | 1.14 | 52 | 2.26 | 1.14 | 0.946 | 0.044 | (1,53) |
| Ponpy | 7 | 1.81 | 1.04 | 53 | 1.81 | 1.04 | 0.778 | 0.205 | (52,9) |
| Hylsy1 | 8 | 1.46 | 1.03 | 54 | 1.46 | 1.03 | 0.240* | 0.538 | (52,60) |
| Hylsy2 | 9 | 1.90 | 0.95 | 55 | 1.90 | 0.95 | 0.956 | 0.040 | (56,11) |
| Macfa | 10 | 1.10 | 0.76 | 56 | 1.10 | 0.76 | 0.609 | 0.345 | (14,55) |
| Macmu | 11 | 0.64 | 0.64 | 57 | 0.64 | 0.64 | 0.686 | 0.180 | (58,56) |
| Papan | 12 | 1.71 | 0.92 | 58 | 1.71 | 0.92 | 0.929 | 0.057 | (57,13) |
| Thege | 13 | 1.40 | 0.84 | 59 | 1.40 | 0.84 | 0.522 | 0.442 | (57,16) |
| Cerga | 14 | 5.86 | 1.84 | 60 | 5.86 | 1.84 | 0.990 | 0.008 | (54,16) |
| Manle | 15 | 21.30 | 3.71 | 61 | 21.30 | 3.71 | 1.0 | 0.0 | (88,60) |
| Cerae | 16 | 19.76 | 3.48 | 62 | 19.76 | 3.48 | 1.0 | 0.0 | (87,18) |
| Alopa | 17 | 1.89 | 1.01 | 63 | 1.89 | 1.01 | 0.923 | 0.060 | (20,19) |
| Lagla | 18 | 1.12 | 0.78 | 64 | 1.12 | 0.78 | 0.836 | 0.119 | (71,63) |
| Tarba | 19 | 2.04 | 1.14 | 65 | 2.04 | 1.14 | 0.775 | 0.201 | (32,70) |
| Tarsy | 20 | 2.25 | 1.01 | 66 | 2.25 | 1.01 | 0.992 | 0.001 | (34,36) |
| Eulma | 21 | 0.95 | 0.73 | 67 | 0.95 | 0.73 | 0.547 | 0.376 | (34,69) |
| Hapgr | 22 | 1.87 | 0.94 | 68 | 1.87 | 0.94 | 0.956 | 0.044 | (39,38) |
| Lemca | 23 | 2.58 | 1.12 | 69 | 2.58 | 1.12 | 0.920 | 0.077 | (67,39) |
| Varva | 24 | lower | limit | 70 | lower | limit | 0.143* | 0.707 | (67,65) |
| Prota | 25 | 2.40 | 1.08 | 71 | 2.40 | 1.08 | 0.982 | 0.017 | (64,70) |
| Cheme | 26 | lower | limit | 72 | lower | limit | 0.051* | 0.878 | (64,76) |
| Dauma | 27 | 2.70 | 1.34 | 73 | 2.70 | 1.34 | 0.905 | 0.078 | (44,41) |
| Nycco | 28 | 2.34 | 1.32 | 74 | 2.34 | 1.32 | 0.694 | 0.172 | (42,44) |
| Galse | 29 | 2.48 | 1.31 | 75 | 2.48 | 1.31 | 0.604 | 0.361 | (31,42) |
| Cynva | 30 | 0.71 | 0.71 | 76 | 0.71 | 0.71 | 0.688 | 0.239 | (72,75) |
| Tupgl | 31 | 2.07 | 1.13 | 77 | 2.07 | 1.13 | 0.424* | 0.569 | (72,85) |
| Phyha | 32 | 1.01 | 0.74 | 78 | 1.01 | 0.74 | 0.727 | 0.183 | (21,23) |
| Roule | 33 | 0.07 | 0.47 | 79 | 0.07 | 0.47 | 0.169* | 0.661 | (24,78) |
| Bosta | 34 | 1.87 | 0.99 | 80 | 1.87 | 0.99 | 0.938 | 0.052 | (79,26) |
| Balph | 35 | lower | limit | 81 | lower | limit | 0.180* | 0.687 | (82,26) |
| Balmu | 36 | 6.59 | 1.80 | 82 | 6.59 | 1.80 | 1.0 | 0.0 | (81,29) |
| Phovi | 37 | 2.31 | 1.11 | 83 | 2.31 | 1.11 | 0.752 | 0.240 | (81,25) |
| Halgr | 38 | 5.07 | 1.64 | 84 | 5.07 | 1.64 | 0.952 | 0.048 | (83,27) |
| Cansi | 39 | 3.32 | 1.43 | 85 | 3.32 | 1.43 | 0.867 | 0.131 | (84,77) |
| Musmu | 40 | 4.37 | 1.68 | 86 | 4.37 | 1.68 | 0.828 | 0.153 | (77,30) |
| Ratno | 41 | 2.62 | 1.79 | 87 | 2.62 | 1.79 | 0.282* | 0.694 | (62,30) |
| Geoca | 42 | 2.95 | 1.56 | 88 | 2.95 | 1.56 | 0.150* | 0.715 | (62,61) |
| Dasno | 43 | 9.03 | 2.70 | 89 | 9.03 | 2.70 | 0.986 | 0.011 | (61,45) |
| Didvi | 44 | 13.06 | 2.62 | 1 | | | | | |
| Galga | 45 | 23.97 | 3.80 | ln L: | -3930.73 | +- 245.24 | | | |
| Xenla | 46 | 11.94 | 2.78 | AIC : | 8077.47 | | | | |

Figure 5.23: Branch lengths and LBPs (estimated by ML) of the NJ tree of cytochrome oxidase subunit II.

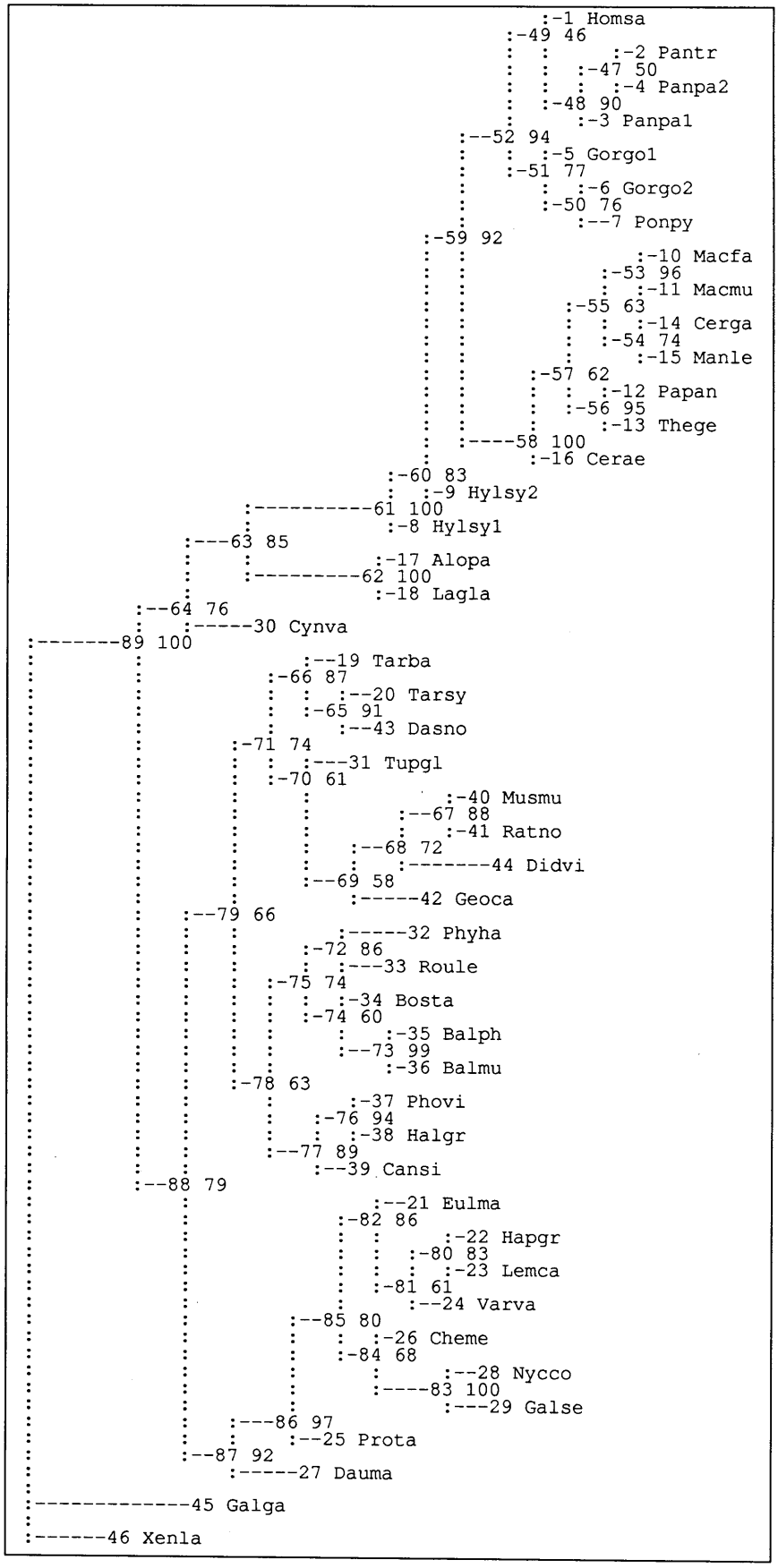



Figure 5.24: The ML tree of cytochrome oxidase subunit II.

| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Homsa | 1 | lower | limit | 47 | lower | limit | 0.496 | 0.504 | (3,4) |
| Pantr | 2 | 0.96 | 0.68 | 48 | 0.96 | 0.68 | 0.899 | 0.064 | (1,3) |
| Panpa1 | 3 | 0.46 | 0.48 | 49 | 0.46 | 0.48 | 0.459 | 0.431 | (51,48) |
| Panpa2 | 4 | 0.44 | 0.45 | 50 | 0.44 | 0.45 | 0.761 | 0.139 | (6,5) |
| Gorgo1 | 5 | 0.94 | 0.69 | 51 | 0.94 | 0.69 | 0.766 | 0.206 | (49,50) |
| Gorgo2 | 6 | 2.66 | 1.14 | 52 | 2.66 | 1.14 | 0.943 | 0.049 | (58,51) |
| Ponpy | 7 | 1.90 | 0.95 | 53 | 1.90 | 0.95 | 0.955 | 0.044 | (54,11) |
| Hylsy1 | 8 | 1.06 | 0.76 | 54 | 1.06 | 0.76 | 0.744 | 0.154 | (14,53) |
| Hylsy2 | 9 | 0.61 | 0.61 | 55 | 0.61 | 0.61 | 0.628 | 0.214 | (56,54) |
| Macfa | 10 | 1.79 | 0.93 | 56 | 1.79 | 0.93 | 0.948 | 0.041 | (55,13) |
| Macmu | 11 | 1.40 | 0.84 | 57 | 1.40 | 0.84 | 0.617 | 0.328 | (55,16) |
| Papan | 12 | 6.81 | 1.84 | 58 | 6.81 | 1.84 | 1.0 | 0.0 | (57,52) |
| Thege | 13 | 1.83 | 0.95 | 59 | 1.83 | 0.95 | 0.916 | 0.074 | (52,9) |
| Cerga | 14 | 0.44 | 0.45 | 60 | 0.44 | 0.45 | 0.831 | 0.169 | (8,9) |
| Manle | 15 | 20.03 | 3.47 | 61 | 20.03 | 3.47 | 1.0 | 0.0 | (60,62) |
| Cerae | 16 | 17.48 | 3.27 | 62 | 17.48 | 3.27 | 1.0 | 0.0 | (17,61) |
| Alopa | 17 | 4.88 | 1.97 | 63 | 4.88 | 1.97 | 0.854 | 0.133 | (30,62) |
| Lagla | 18 | 3.61 | 1.65 | 64 | 3.61 | 1.65 | 0.763 | 0.224 | (88,30) |
| Tarba | 19 | 1.90 | 1.01 | 65 | 1.90 | 1.01 | 0.910 | 0.076 | (20,19) |
| Tarsy | 20 | 0.92 | 0.67 | 66 | 0.92 | 0.67 | 0.868 | 0.130 | (19,70) |
| Eulma | 21 | 2.69 | 1.34 | 67 | 2.69 | 1.34 | 0.884 | 0.084 | (44,41) |
| Hapgr | 22 | 2.38 | 1.32 | 68 | 2.38 | 1.32 | 0.715 | 0.169 | (42,44) |
| Lemca | 23 | 2.42 | 1.29 | 69 | 2.42 | 1.29 | 0.581 | 0.377 | (31,42) |
| Varva | 24 | 0.68 | 0.68 | 70 | 0.68 | 0.68 | 0.613 | 0.242 | (66,69) |
| Prota | 25 | 1.02 | 0.79 | 71 | 1.02 | 0.79 | 0.743 | 0.160 | (78,70) |
| Cheme | 26 | 1.74 | 1.07 | 72 | 1.74 | 1.07 | 0.862 | 0.108 | (74,33) |
| Dauma | 27 | 2.39 | 1.06 | 73 | 2.39 | 1.06 | 0.994 | 0.002 | (34,36) |
| Nycco | 28 | 0.81 | 0.64 | 74 | 0.81 | 0.64 | 0.598 | 0.342 | (34,72) |
| Galse | 29 | 1.00 | 0.82 | 75 | 1.00 | 0.82 | 0.736 | 0.142 | (77,74) |
| Cynva | 30 | 1.51 | 0.87 | 76 | 1.51 | 0.87 | 0.935 | 0.034 | (39,38) |
| Tupgl | 31 | 2.22 | 1.04 | 77 | 2.22 | 1.04 | 0.886 | 0.110 | (75,39) |
| Phyha | 32 | 1.10 | 0.79 | 78 | 1.10 | 0.79 | 0.633 | 0.229 | (75,71) |
| Roule | 33 | 2.27 | 1.20 | 79 | 2.27 | 1.20 | 0.663 | 0.261 | (87,78) |
| Bosta | 34 | 0.88 | 0.65 | 80 | 0.88 | 0.65 | 0.829 | 0.142 | (24,23) |
| Balph | 35 | 0.45 | 0.46 | 81 | 0.45 | 0.46 | 0.608 | 0.231 | (80,21) |
| Balmu | 36 | 1.22 | 0.79 | 82 | 1.22 | 0.79 | 0.864 | 0.098 | (21,84) |
| Phovi | 37 | 6.35 | 1.75 | 83 | 6.35 | 1.75 | 0.999 | 0.001 | (26,29) |
| Halgr | 38 | 0.79 | 0.74 | 84 | 0.79 | 0.74 | 0.683 | 0.206 | (26,82) |
| Cansi | 39 | 2.07 | 1.04 | 85 | 2.07 | 1.04 | 0.801 | 0.172 | (25,84) |
| Musmu | 40 | 5.63 | 1.72 | 86 | 5.63 | 1.72 | 0.971 | 0.028 | (85,27) |
| Ratno | 41 | 2.93 | 1.37 | 87 | 2.93 | 1.37 | 0.920 | 0.074 | (86,79) |
| Geoca | 42 | 2.70 | 1.35 | 88 | 2.70 | 1.35 | 0.788 | 0.148 | (79,64) |
| Dasno | 43 | 12.39 | 2.86 | 89 | 12.39 | 2.86 | 0.998 | 0.002 | (64,45) |
| Didvi | 44 | 13.04 | 2.62 | 1 | | | | | |
| Galga | 45 | 24.46 | 3.83 | ln L: | -3897.47 | +- 241.84 | | | |
| Xenla | 46 | 11.05 | 2.66 | AIC : | 8010.94 | | | | |

Figure 5.25: Branch lengths and LBPs of the ML tree of cytochrome oxidase subunit II.

**Unacceptable Points in the COII Tree**

These two trees contain several characteristics which cannot be accepted from biological ground. First, *Didelphis virginiana* (opossum; Marsupialia) must be the outgroup to all the other mammals used in this analysis (Eutheria), but in the COII tree it is located within Rodentia. Second, although anthropoids (catarrhines and platyrrhines; species 1 – 18) form a monophyletic clade, three major groups in primates, that is, anthropoids, tarsiers (species 19 – 20), and strepsirhines (species 21 – 29) are separated in the trees. Third, *Dasypus novemcinctus* (armadillo) is located within tarsiers, and it is closer to *Tarsius syrichta* than to *Tarsius bancanus*. Fourth, the location of Chiroptera (bats; species 32 – 33) close to the Artiodactyls/Cetacea clade might not be accepted. In these ways, there are several unreasonable features in these trees, and therefore phylogenetic conelusions drawn from this molecule must be taken with care.

The apparent polyphyly of primates might be due to the increased rate of COII in primates relative to other eutherian mammals. The increased amino acid substitution rate of COII in primates is consistent with an increase of the rate of primate cytochrome *c*, suggesting coevolution of these two mutually interacting proteins during electron transport (Brown and Simpson 1982[41]; Cann et al. 1984[45]; Ramharack and Deeley 1987[243]). The ML and NJ methods are known to be robust against the violation of rate constancy among lineages (Hasegawa et al. 1991[121]; DeBry 1992[64]; Hasegawa and Fujiwara 1993[113]; Kuhner and Felsenstein[177]). But if the rate variation is accompanied with the variation of substitution pattern, then the robustness of the method may not hold. Actually for example, numbers of Ala in COII are 12–16 in Anthropoidea, while 4–6 in tarsiers, and the substitution bias might be different in these two lineages. Although it remains to be elucidated whether such a difference is due simply to chance by statistical fluctuation or due to real difference of the fundamental process among different lineages, increase of the number of ingroup species might be helpful to obtain more reliable inference of a tree.

Cao et al. (1994[48]) pointed out that convergent change of base composition is a serious problem in inferring eutherian phylogeny by using mitochondrial ribosomal RNAs. For example, *Didelphis virginiana* (opossum) has the most similar base composition of rRNAs with mouse and rat, while chicken has the most similar composition with human. They also showed that, although indirect, amino acid compositions of mtDNA-encoded proteins are influenced by the biased base composition of the genome, and that convergent amino acid composition can also be a serious problem in analyzing mtDNA-encoded proteins. To minimize the disturbing effect of the composition bias, it might be better to use several ingroup species which are divergent from each other as far as possible, because the compositional bias varies within a mammalian order and the use of divergent species within an order might alleviate an extreme bias of a particular species.

Numbers of amino acid differences in 225 sites of COII between different species are given in Fig. 5.21. Numbers of differences of *Tupaia glis* from Anthropoidea, Tarsiiformes, and Strepsirhini are 55–69, 20–23,

and 32–45, respectively, and those of *Cynocephalus variegatus* (flying lemur) COII from Anthropoidea, Tarsiiformes, and Strepsirhini are 51–70, 31–34, and 41–53. Those of *Phyllostomus hastatus* (microbat) from Anthropoidea, Tarsiiformes, and Strepsirhini are 62–73, 31–32, and 42–51, and those of *Rousettus leschenaulti* (megabat) are 52–71, 21–24, and 33–49. These observations suggest that the increased rate of the COII seems mostly confined to Anthropoidea. Numbers of differences of the COII of *Dasypus novemcinctus* (Edentata), which is considered to be an outgroup to the other Eutherian orders (Novacek 1992[228]), from Anthropoidea, Tarsiiformes, Strepsirhini, and the other eutherian species under analysis are 53–69, 16–17, 25–43, and 19–32. The differences are very small between *D. novemcinctus* and tarsiers probably due to chance under fluctuating evolutionary rate of this protein, and for this reason it turned out that *D. novemcinctus* makes a cluster with tarsiers in my ProtML and NJ trees, which cannot be real. In order to avoid this sort of wrong tree, data from another species of Edentata might be of help. The two bat species, *P. hastatus* and *R. leschenaulti*, are closely related in my tree as will be discussed below, still the numbers of differences from tarsiers are different between the two species (31–32 and 21–24). Therefore, inclusion of the two species would help to provide reliability of the inferred tree than when only one is included.

### Monophyly of Chiroptera

Monophyly of Chiroptera is supported with 85% LBP (branch 65) in accord with Adkins and Honeycutt (1993[9]). From the analysis of neural anatomy in the visual and motor pathways, Pettigrew (1986[237]) contended that the two suborders of Chiroptera (bats), Megachiroptera which are large fruit-eating bats of the Old World and Microchiroptera which are smaller predominantly insectivorous cosmopolitan bats, are not each other's closest relatives, but that the megabats are closer to primates rather than to microbats. He found several features related to the patterns of connection between the retina and midbrain that are shared in primates and megabats. Pettigrew's hypothesis is called the bat-diphyly or "flying primate" hypothesis, and implies that mammals evolved flight twice. Since this hypothesis sharply contradicts the traditional view of bat-monophyly, this problem has been hotly debated by morphologists (Pettigrew 1991[239], 1991[238]; Baker et al. 1991[33]; Simmons et al. 1992[269]). For this type of controversial problems among morphologists, molecular phylogenetic approach is expected to be powerful and several molecular studies have been done, and they consistently support the bat-monophyly hypothesis (Adkins and Honeycutt 1991[8], 1993[9], 1994[10]); Mindell et al. 1991[214]); Ammerman and Hillis 1992[14]; Bailey et al. 1992[32]; Stanhope 1993[275]). My analysis reconfirms this hypothesis.

### Relationships among Cercopithecoidea

My analysis of the COII data reconfirmed Adkins and Honeycutt's (1994[10]) conclusions on the following relationships among Cercopithecoidea. In the ProtML as well as the NJ tree, the tribe Papionini (*Papio*, *Mandrillus*, *Macaca*, *Theropithecus*, and *Cercocebus*) is monophyletic relative to *Cercopithecus* only with

60% LBP. A *Papio/Theropithecus* clade is strongly supported with 95% LBP. A *Mandrillus/Cercocebus* clade is supported with 75% LBP. The relationships among the *Papio/Theropithecus* clade, the *Mandrillus/Cercocebus*, and *Macaca* are obscured from the COII analysis.

### Relationships among Strepsirhines

Consistent with Adkins and Honeycutt's (1994[10]) analysis of the COII data by the parsimony, my analysis suggests sister relationship of the aye-aye (*Daubentonia madagascarensis*) to all other strepsirhine primates (97% LBP for branch 84). The placing of *Daubentonia* has been controversial among morphologists. Schwartz and Tattersall (1985[259]) suggested that *Daubentonia* is sister to the family Indriidae, and Groves (1989[101]), on the otherhand, suggested sister-group relationship of *Daubentonia* to all other strepsirhines. Adkins and Honeycutt's and my analyses favour the latter suggestion, but further independent data might be necessary to have conclusive evidence. Strepsirhini primates of Madagascar (Lemuriformes) most likely form a paraphylytic group, within which Lorisiformes is included.

The lemuriform family Cheirogaleidae has been suggested to be more closely related to the lorisiforms, such as *Galago* and *Nycticebus* (loris), than to the other lemuriformes (Szalay and Katz 1973[283]). Although Adkins and Honeycutt's (1994[10]) analysis rejected this relationship and my NJ tree is in accord with their tree, the ProtML tree suggests sister-group relationship of *Cheirogaleus* (dwarf lemurs) with lorisiform primates. Since LBP of this relationship is only 66% in my ProtML tree and the grouping of *Cheirogaleus* with the other lemurs is suggested by the $\gamma$-globin gene sequences (Bailey et al. 1992[31]), further studies are needed to settle the issue.

## 5.6 Phylogenetic Place of Myxozoa

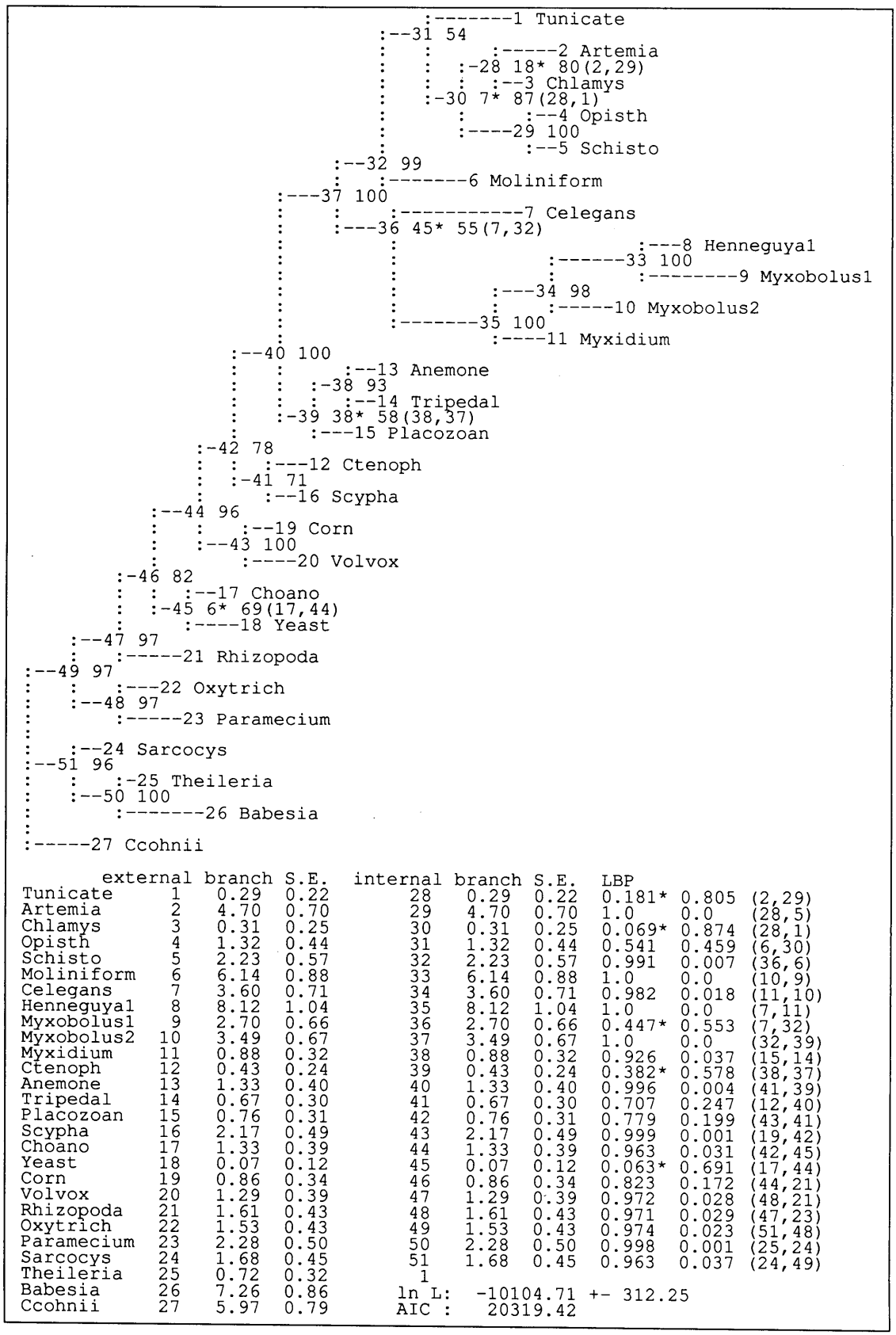
As an example of application of NucML program, I will choose the 18S ribosomal RNA tree, including myxozoan protists, which has been studies by Smothers et al. (1994[270]). Margulis and Schwartz (1988[201]) classified Myxozoa in the phylum Cnidosporidia together with Microsporidia such as *Vairimorpha*, *Nosema*, and *Glugea*. Microsporidia lack mitochondria in spite that they are eukaryotes, and they are considered to have diverged from other eukaryotes in the very early stage of the eukaryotic evolution (Vossbrinck et al. 1987[303]; Cavalier-Smith 1989[51], 1993[52]; Leipe et al. 1993[189]). Unexpectedly, however, Smothers et al. found that myxozoans are likely to be located within the metazoan clade, and that they may be closely related particularly to the bilateral animals in metazoa. They analyzed the 18S ribosomal RNA sequences by the parsimony and NJ methods as well as by the DNAML of Felsenstein. I reanalyzed their data listed in Table 5.20 by using the NucML program.

I used the aligned sequence data provided by Drs. Carol D. von Dohlen and Richard D. Spall, and selected the same sites they used. The HKY85 model was adopted in the NucML analysis. The optimal $\alpha/\beta$ ratio in the NJ tree is 2.70, and this ratio was used for the NucML analysis. Starting from the NJ tree shown in Fig. 5.26, replications of the nearest-neighbour rearrangements produced the NucML tree shown in Fig. 5.27. The NJ and NucML trees do not differ very much, and they are mostly consistent with Smothers et al.'s tree. My analysis reconfirmed their assertion that Myxozoa are located within the metazoan clade, but the precise placing of Myxozoa in metazoa, such as whether Myxozoa are members of the bilateral animal clade or a sister group to them, requires further studies as Smothers et al. admit.

In the NJ and NucML trees, plants are closer to metazoa than to fungi. Although this is consistent with Sidow and Thomas's (1994[268]) analysis of RNA polymerase, the overall molecular evidence accumulated up to now seems to support the closer relationship of metazoa to fungi rather than to plants (Hasegawa et al. 1993[113]; Nikoh et al. 1994[226]). Support of the plants/metazoa clade in the NucML tree of 18S ribosomal RNA is as high as 99% (branch 44), but this result may depend on the alignment and on the species sampling, and Wainright et al.'s (1993[305]) conclusion based also on 18S ribosomal RNA differs in this respect and supports the fugi/metazoa clade.

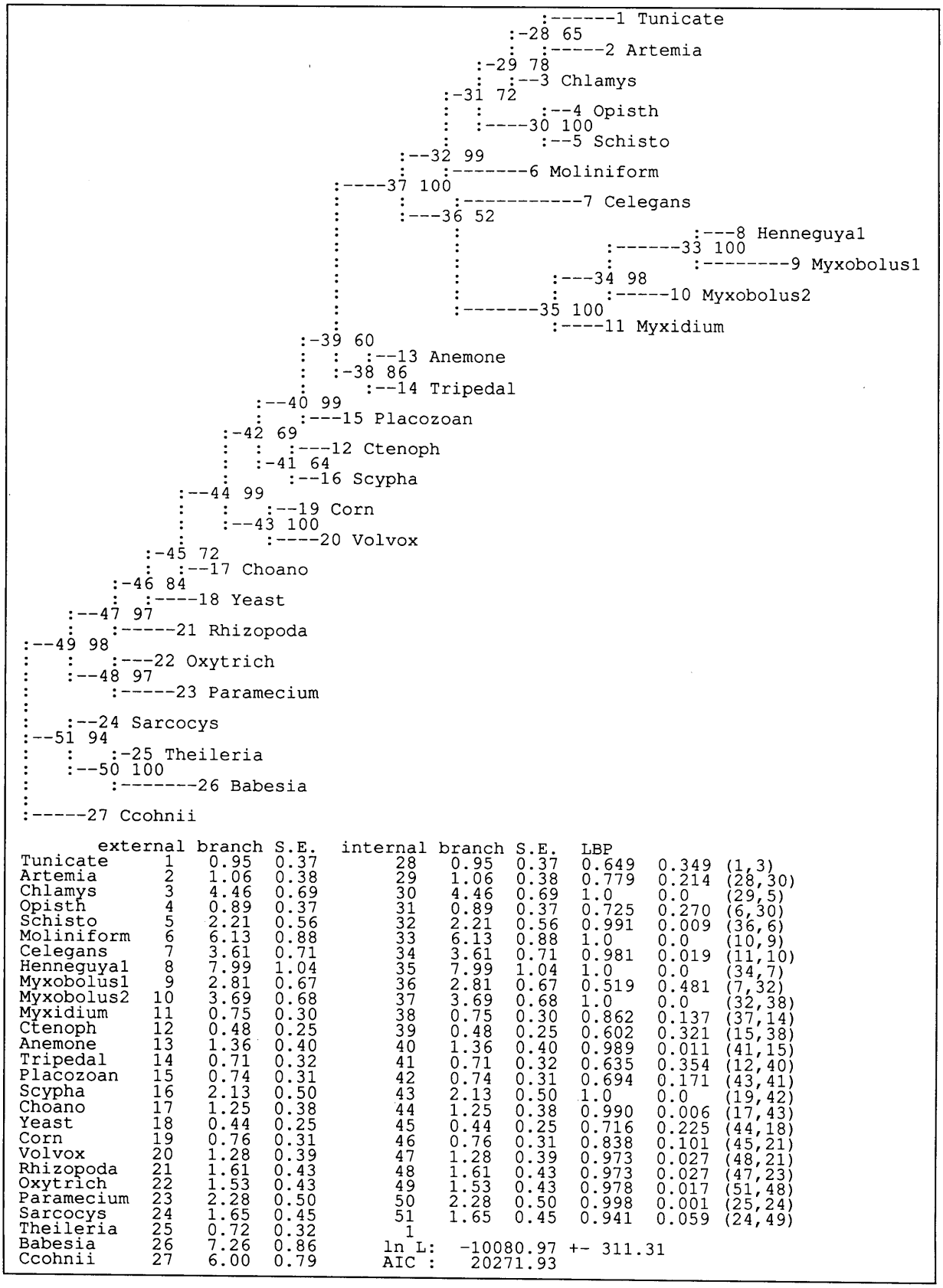
Table 5.20: List of 18S ribosomal RNA sequences used in inferring phylogenetic place of Myxozoa.

| Abbrev. | Species name | Classification |
|---|---|---|
| Tunicate | *Herdmania momus* | Chordata (Bilateria) |
| Artemia | *Artemia salina* | Arthropoda (Bilateria) |
| Chlamys | *Chlamys islandica* | Mollusca (Bilateria) |
| Opisth | *Opisthorchis viverrini* | Platyhelminthes (Bilateria) |
| Schisto | *Schistosoma mansoni* | Platyhelminthes (Bilateria) |
| Moliniform | *Moliniformis moliniformis* | Acanthocephala (Bilateria) |
| Celegans | *Caenorhabditus elegans* | Nematoda (Bilateria) |
| Anemone | *Anemonia sulcata* | Cnidaria |
| Tripedal | *Tripedalia cystophora* | Cnidaria |
| Placozoan | *Trichoplax adhaerens* | Placozoa |
| Ctenoph | *Mnemiopsis leidyi* | Ctenophora |
| Scypha | *Scypha ciliata* | Porifera |
| Yeast | *Saccharomyces cerevisiae* | Fungi |
| Corn | *Zea maya* | Plantae |
| Volvox | *Volvox carteri* | Plantae |
| Henneguya1 | *Henneguya* sp. 1 | Myxozoa |
| Myxobolus1 | *Myxobolus* sp. 1 | Myxozoa |
| Myxobolus2 | *Myxobolus* sp. 2 | Myxozoa |
| Choano | *Diaphanoeca grandid* | Choanozoa |
| Rhizopoda | *Hartmanella vermiformis* | Rhizopoda |
| Oxytrich | *Oxytricha nova* | Ciliophora |
| Paramecium | *Paramecium tetraurelia* | Ciliophora |
| Sarcocys | *Sarcocystis muris* | Apicomplexa |
| Theileria | *Theileria annulata* | Apicomplexa |
| Babesia | *Babesia bovis* | Apicomplexa |
| Ccohnii | *Crypthecodinium cohnii* | Dinozoa |




| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Tunicate | 1 | 0.29 | 0.22 | 28 | 0.29 | 0.22 | 0.181* | 0.805 | (2,29) |
| Artemia | 2 | 4.70 | 0.70 | 29 | 4.70 | 0.70 | 1.0 | 0.0 | (28,5) |
| Chlamys | 3 | 0.31 | 0.25 | 30 | 0.31 | 0.25 | 0.069* | 0.874 | (28,1) |
| Opisth | 4 | 1.32 | 0.44 | 31 | 1.32 | 0.44 | 0.541 | 0.459 | (6,30) |
| Schisto | 5 | 2.23 | 0.57 | 32 | 2.23 | 0.57 | 0.991 | 0.007 | (36,6) |
| Moliniform | 6 | 6.14 | 0.88 | 33 | 6.14 | 0.88 | 1.0 | 0.0 | (10,9) |
| Celegans | 7 | 3.60 | 0.71 | 34 | 3.60 | 0.71 | 0.982 | 0.018 | (11,10) |
| Henneguya1 | 8 | 8.12 | 1.04 | 35 | 8.12 | 1.04 | 1.0 | 0.0 | (7,11) |
| Myxobolus1 | 9 | 2.70 | 0.66 | 36 | 2.70 | 0.66 | 0.447* | 0.553 | (7,32) |
| Myxobolus2 | 10 | 3.49 | 0.67 | 37 | 3.49 | 0.67 | 1.0 | 0.0 | (32,39) |
| Myxidium | 11 | 0.88 | 0.32 | 38 | 0.88 | 0.32 | 0.926 | 0.037 | (15,14) |
| Ctenoph | 12 | 0.43 | 0.24 | 39 | 0.43 | 0.24 | 0.382* | 0.578 | (38,37) |
| Anemone | 13 | 1.33 | 0.40 | 40 | 1.33 | 0.40 | 0.996 | 0.004 | (41,39) |
| Tripedal | 14 | 0.67 | 0.30 | 41 | 0.67 | 0.30 | 0.707 | 0.247 | (12,40) |
| Placozoan | 15 | 0.76 | 0.31 | 42 | 0.76 | 0.31 | 0.779 | 0.199 | (43,41) |
| Scypha | 16 | 2.17 | 0.49 | 43 | 2.17 | 0.49 | 0.999 | 0.001 | (19,42) |
| Choano | 17 | 1.33 | 0.39 | 44 | 1.33 | 0.39 | 0.963 | 0.031 | (42,45) |
| Yeast | 18 | 0.07 | 0.12 | 45 | 0.07 | 0.12 | 0.063* | 0.691 | (17,44) |
| Corn | 19 | 0.86 | 0.34 | 46 | 0.86 | 0.34 | 0.823 | 0.172 | (44,21) |
| Volvox | 20 | 1.29 | 0.39 | 47 | 1.29 | 0.39 | 0.972 | 0.028 | (48,21) |
| Rhizopoda | 21 | 1.61 | 0.43 | 48 | 1.61 | 0.43 | 0.971 | 0.029 | (47,23) |
| Oxytrich | 22 | 1.53 | 0.43 | 49 | 1.53 | 0.43 | 0.974 | 0.023 | (51,48) |
| Paramecium | 23 | 2.28 | 0.50 | 50 | 2.28 | 0.50 | 0.998 | 0.001 | (25,24) |
| Sarcocys | 24 | 1.68 | 0.45 | 51 | 1.68 | 0.45 | 0.963 | 0.037 | (24,49) |
| Theileria | 25 | 0.72 | 0.32 | 1 | | | | | |
| Babesia | 26 | 7.26 | 0.86 | ln L: | -10104.71 | +- 312.25 | | | |
| Ccohnii | 27 | 5.97 | 0.79 | AIC : | 20319.42 | | | | |

Figure 5.26: The NJ tree of 18S ribosomal RNA.




| | external | branch | S.E. | internal | branch | S.E. | LBP | | |
|---|---|---|---|---|---|---|---|---|---|
| Tunicate | 1 | 0.95 | 0.37 | 28 | 0.95 | 0.37 | 0.649 | 0.349 | (1,3) |
| Artemia | 2 | 1.06 | 0.38 | 29 | 1.06 | 0.38 | 0.779 | 0.214 | (28,30) |
| Chlamys | 3 | 4.46 | 0.69 | 30 | 4.46 | 0.69 | 1.0 | 0.0 | (29,5) |
| Opisth | 4 | 0.89 | 0.37 | 31 | 0.89 | 0.37 | 0.725 | 0.270 | (6,30) |
| Schisto | 5 | 2.21 | 0.56 | 32 | 2.21 | 0.56 | 0.991 | 0.009 | (36,6) |
| Moliniform | 6 | 6.13 | 0.88 | 33 | 6.13 | 0.88 | 1.0 | 0.0 | (10,9) |
| Celegans | 7 | 3.61 | 0.71 | 34 | 3.61 | 0.71 | 0.981 | 0.019 | (11,10) |
| Henneguya1 | 8 | 7.99 | 1.04 | 35 | 7.99 | 1.04 | 1.0 | 0.0 | (34,7) |
| Myxobolus1 | 9 | 2.81 | 0.67 | 36 | 2.81 | 0.67 | 0.519 | 0.481 | (7,32) |
| Myxobolus2 | 10 | 3.69 | 0.68 | 37 | 3.69 | 0.68 | 1.0 | 0.0 | (32,38) |
| Myxidium | 11 | 0.75 | 0.30 | 38 | 0.75 | 0.30 | 0.862 | 0.137 | (37,14) |
| Ctenoph | 12 | 0.48 | 0.25 | 39 | 0.48 | 0.25 | 0.602 | 0.321 | (15,38) |
| Anemone | 13 | 1.36 | 0.40 | 40 | 1.36 | 0.40 | 0.989 | 0.011 | (41,15) |
| Tripedal | 14 | 0.71 | 0.32 | 41 | 0.71 | 0.32 | 0.635 | 0.354 | (12,40) |
| Placozoan | 15 | 0.74 | 0.31 | 42 | 0.74 | 0.31 | 0.694 | 0.171 | (43,41) |
| Scypha | 16 | 2.13 | 0.50 | 43 | 2.13 | 0.50 | 1.0 | 0.0 | (19,42) |
| Choano | 17 | 1.25 | 0.38 | 44 | 1.25 | 0.38 | 0.990 | 0.006 | (17,43) |
| Yeast | 18 | 0.44 | 0.25 | 45 | 0.44 | 0.25 | 0.716 | 0.225 | (44,18) |
| Corn | 19 | 0.76 | 0.31 | 46 | 0.76 | 0.31 | 0.838 | 0.101 | (45,21) |
| Volvox | 20 | 1.28 | 0.39 | 47 | 1.28 | 0.39 | 0.973 | 0.027 | (48,21) |
| Rhizopoda | 21 | 1.61 | 0.43 | 48 | 1.61 | 0.43 | 0.973 | 0.027 | (47,23) |
| Oxytrich | 22 | 1.53 | 0.43 | 49 | 1.53 | 0.43 | 0.978 | 0.017 | (51,48) |
| Paramecium | 23 | 2.28 | 0.50 | 50 | 2.28 | 0.50 | 0.998 | 0.001 | (25,24) |
| Sarcocys | 24 | 1.65 | 0.45 | 51 | 1.65 | 0.45 | 0.941 | 0.059 | (24,49) |
| Theileria | 25 | 0.72 | 0.32 | 1 | | | | | |
| Babesia | 26 | 7.26 | 0.86 | ln L: | -10080.97 | +- 311.31 | | | |
| Ccohnii | 27 | 6.00 | 0.79 | AIC : | 20271.93 | | | | |

Figure 5.27: The ML tree of 18S ribosomal RNA.

# Acknowledgement

First I am very grateful to Dr. Masami Hasegawa of The Institute of Statistical Mathematics who was my principal academic adviser during the last three years for introducing me into this field of molecular evolution. I learned a lot from discussions with him. He suggested me the interesting subjects studied in the present thesis. He gave me valuable guidance and helpful discussions.

I would like to Dr. Tetsuo Hashimoto for biological and statistical suggestions, comments and advice. I also thank Dr. Hirohisa Kishino for statistical suggetions. Thanks are also due to my academic advisers Dr. Kunio Tanabe, Dr. Yoshiaki Itoh and Dr. Toshiya Sato.

I would like to thank Dr. David Penny and Dr. Avner Bar-Hen for reading this thesis and for giving very valuable comments, and Dr. Michel Milinkovitch, Dr. Arend Sidow and Dr. Jaxk Reeves for suggestions.

I also thank Ms. Ying Cao and Ms. Miyako Fujiwara for their help with data management and data analysis.

Finally, I thank Haruhiko, Teruko and Makiko for their support and their patience.

# Bibliography


[1] J. Adachi, Y. Cao, and M. Hasegawa. Tempo and mode of mitochondrial DNA evolution in vertebrates at the amino acid sequence level: rapid evolution in warm-blooded vertebrates. *J. Mol. Evol.*, 36:270–281, 1993.

[2] J. Adachi and M. Hasegawa. Amino acid substitution of proteins coded for in mitochondrial DNA during mammalian evolution. *Jpn. J. Genet.*, 67:187–197, 1992.

[3] J. Adachi and M. Hasegawa. *Computer Science Monographs, No. 27. MOLPHY: Programs for Molecular Phylogenetics, I. — PROTML: Maximum Likelihood Inference of Protein Phylogeny.* Institute of Statistical Mathematics, Tokyo, 1992.

[4] J. Adachi and M. Hasegawa. Improved dating of the human-chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J. Mol. Evol.*, 1995.

[5] J. Adachi and M. Hasegawa. *MOLPHY: Programs for Molecular Phylogenetics, ver. 2.3.* Institute of Statistical Mathematics, Tokyo, 1995.

[6] J. Adachi and M. Hasegawa. Phylogeny of whales: dependence of the inference on species sampling. *Mol. Biol. Evol.*, 12:177–179, 1995.

[7] J. Adachi and M. Hasegawa. Time scale for the mitochondrial DNA tree of human evolution. In K. Hanihara, editor, *The Origin and Past of Homo sapiens sapiens as Viewed from DNA — Theoretical Approach.* World Scientific Publ., Singapore, 1995.

[8] R.M. Adkins and R.L. Honeycutt. Molecular phylogeny of the superorder Archonta. *Proc. Natl. Acad. Sci. USA*, 88:10317–10321, 1991.

[9] R.M. Adkins and R.L. Honeycutt. A molecular examination of Archontan and Chiropteran monophyly. In R.D.E. MacPhee, editor, *Primates and Their Relatives in Phylogenetic Perspective*, pages 227–249. Plenum Press, New York, 1993.

[10] R.M. Adkins and R.L. Honeycutt. Evolution of the primate cytochrome *c* oxidase subunit II gene. *J. Mol. Evol.*, 38:215–231, 1994.

[11] H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Second International Symposium on Information Theory*, pages 267–281. Akademiai Kiado, Budapest, 1973.

[12] H. Akaike. A new look at the statistical model identification. *IEEE Trans. Autom. Contr.*, AC-19:716–723, 1974.

[13] S.F. Altschul. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, 219:555–565, 1991.

[14] L.K. Ammerman and D.M. Hillis. A molecular test of bat relationships: monophyly or diphyly? *Syst. Biol.*, 41:222–232, 1992.

[15] S. Anderson, A.T. Bankier, B.G. Barrell, M.H.L. de Bruijn, A.R. Coulson, J. Drouin, I.C. Eperon, D.P. Nierlich, B.A. Roe, F. Sanger, P.H. Schreier, A.L.H. Smith, R. Staden, and I.G. Young. Sequence and organization of the human mitochondrial genome. *Nature*, 290:457–464, 1981.

[16] S. Anderson, M.H.L. de Bruijn, A.R. Coulson, I.C. Eperon, F. Sanger, and I.G. Young. The complete sequence of bovine mitochondrial DNA: conserved features of the mammalian mitochondrial genome. *J. Mol. Biol.*, 156:683–717, 1982.

[17] P. Andrews. Evolution and environment in the Hominoidea. *Nature*, 360:641–646, 1992.

[18] P. Andrews and J.E. Cronin. The relationships of *Sivapithecus* and *Ramapithecus* and the evolution of the orang-utan. *Nature*, 297:541–546, 1982.

[19] E. Arévalo, S.K. Davis, and J.W. Sites, Jr. Mitochondrial DNA sequence divergence and phylogenetic relationships among eight chromosome races of the *Sceloporus grammicus* complex (Phrynosomatidae) in central Mexico. *Syst. Biol.*, 43:387–418, 1994.

[20] Ú. Árnason, K. Bodin, A. Gullberg, C. Ledje, and S. Mouchaty. A molecular view of pinniped relationships with particular emphasis on the true seals. *J. Mol. Evol.*, 40:78–85, 1995.

[21] Ú. Árnason and A. Gullberg. Comparison between the complete mtDNA sequences of the blue and the fin whale, two species that can hybridize in nature. *J. Mol. Evol.*, 37:312–322, 1993.

[22] Ú. Árnason and A. Gullberg. Relationship of baleen whales established by cytochrome b gene sequence comparison. *Nature*, 367:726–728, 1994.

[23] Ú. Árnason, A. Gullberg, E. Johnsson, and C. Ledje. The nucleotide sequence of the mitochondrial DNA molecule of the grey seal, *Halichoerus grypus*, and a comparison with mitochondrial sequences of other true seals. *J. Mol. Evol.*, 37:323–330, 1993.

[24] Ú. Árnason, A. Gullberg, and B. Widegren. The complete nucleotide sequence of the mitochondrial DNA of the fin whale, *Balaenoptera physalus*. *J. Mol. Evol.*, 33:556–568, 1991.

[25] Ú. Árnason and E. Johnsson. The complete mitochondrial DNA sequence of the harbor seal, *Phoca vitulina*. *J. Mol. Evol.*, 34:493–505, 1992.

[26] J. Auer, G. Spicker, and A. Böck. Nucleotide sequence of the gene for the translation elongation factor 1$\alpha$ from the extreme thermophilic archaebacterium *Thermococcus celer*. *Nucl. Acids. Res.*, 18:3989–3989, 1990.

[27] J. Auer, G. Spicker, L. Mayerhofer, G. Pühler, and A. Böck. Organisation and nucleotide sequence of a gene cluster comprising the translation elongation factor 1$\alpha$ from the extreme thermophilic archaebacterium *Sulfolobus acidocaldarius*. *Syst. Appl. Microbiol.*, 14:14–22, 1990.

[28] J.C. Avise. *Molecular Markers, Natural History and Evolution*. Chapman & Hall, New York, 1994.

[29] J.C. Avise, W.S. Nelson, and C.G. Sibley. DNA sequence support for a close phylogenetic relationship between some storks and New World vultures. *Proc. Natl. Acad. Sci. USA*, 91:5173–5177, 1994.

[30] J.C. Avise, W.S. Nelson, and C.G. Sibley. Why one-kilobase sequences from mitochondrial DNA fail to solve the Hoatzin phylogenetic enigma. *Mol. Phyl. Evol.*, 3:175–184, 1994.

[31] W.J. Bailey, K. Hayasaka, C.G. Skinner, S. Kehoe, L.C. Sieu, J.L. Slightom, and M. Goodman. Reexamination of the African hominoid trichotomy with additional sequences from the primate $\beta$-globin gene cluster. *Mol. Phyl. Evol.*, 1:97–135, 1992.

[32] W.J. Bailey, J.L. Slightom, and M. Goodman. Rejection of the "flying primate" hypothesis by phylogenetic evidence from the $\epsilon$-globin gene. *Science*, 256:86–89, 1992.

[33] R.J. Baker, M.J. Novacek, and N.B. Simmons. On the monophyly of bats. *Syst. Zool.*, 40:216–231, 1991.

[34] R.J. Baker, V.A. Taddei, J.L. Hudgeons, and R.A. Den Bussche. Systematic relationships within Chiroderma (Chiroptera: Phyllostomidae) based on cytochrome *b* sequence variation. Unpublished.

[35] G. Baldacci, F. Guinet, J. Tillit, G. Zaccai, and A.M. de Recondo. Functional implications releated to the gene structure of the elongation factor EF-Tu from *Halobacterium marismortui*. *Nucl. Acids. Res.*, 18:507–511, 1990.

[36] L.G. Barnes. In T.W. Broadhead, editor, *Mammals: notes for a short course organized by P.D. Gingerich and C.E. Badgley*, volume 8, pages 139–158. Univ. Tenn. Stud. Geol., 1984.

[37] L.G. Barnes, D.P. Doming, and C.E. Ray. Status of studies on fossil marine mammals. *Mar. Mammal Sci.*, 1:15–53, 1985.

[38] D. Barry and J.A. Hartigan. Asynchronous distance between homologous DNA sequences. *Biometrics*, 43:261–276, 1987.

[39] M.J. Bibb, R.A. Van Etten, C.T. Wright, M.W. Walberg, and D.A. Clayton. Sequence and gene organization of mouse mitochondrial DNA. *Cell*, 26:167–180, 1981.

[40] B.A. Block, J.R. Finnerty, A.F.R. Stewart, and J. Kidd. Evolution of endothermy in fish: mapping physiological traits on a molecular phylogeny. *Science*, 260:210–214, 1993.

[41] G.G. Brown and M.V. Simpson. Novel features of animal mtDNA evolution as shown by sequences of two rat cytochrome oxidase subunit II genes. *Proc. Natl. Acad. Sci. USA*, 79:3246–3250, 1982.

[42] J.R. Brown, T.L. Gilbert, D.J. Kowbel, P.J. O'Hara, N.E. Buroker, A.T. Beckenbach, and M.J. Smith. Nucleotide sequence of the apocytochrome b gene in white sturgeon mitochondrial DNA. *Nucl. Acids. Res.*, 17:4389–4389, 1989.

[43] W.M. Brown, E.M. Prager, A. Wang, and A.C. Wilson. Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J. Mol. Evol.*, 18:225–239, 1982.

[44] A. Caccone and J.R. Powell. DNA divergence among hominoids. *Evolution*, 43:925–942, 1989.

[45] R.L. Cann, W.M. Brown, and A.C. Wilson. Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. *Genetics*, 106:479–499, 1984.

[46] R.L. Cann, M. Stoneking, and A.C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325:31–36, 1987.

[47] P. Cantatore, M. Roberti, G. Pesole, A. Ludovico, F. Milella, M.N. Gadaleta, and C. Saccone. Evolutionary analysis of cytochrome b sequences in some Perciformes: evidence for a slower rate of evolution than in mammals. *J. Mol. Evol.*, 39:589–597, 1994.

[48] Y. Cao, J. Adachi, and M. Hasegawa. Eutherian phylogeny as inferred from mitochondrial DNA sequence data. *Jpn. J. Genet.*, 69:455–472, 1994.

[49] Y. Cao, J. Adachi, A. Janke, S. Pääbo, and M. Hasegawa. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. *J. Mol. Evol.*, 39:519–527, 1994.

[50] Y. Cao, J. Adachi, T. Yano, and M. Hasegawa. Phylogenetic place of guinea pigs: no support of the rodent polyphyly hypothesis from maximum likelihood analyses of multiple protein sequences. *Mol. Biol. Evol.*, 11:593–604, 1994.

[51] T. Cavalier-Smith. Archaebacteria and Archezoa. *Nature*, 339:100–101, 1989.

[52] T. Cavalier-Smith. Kingdom Protozoa and its 18 phyla. *Microbiol. Rev.*, 57:953–994, 1993.

[53] R. Chakraborty. Estimation of time of divergence from phylogenetic studies. *Can. J. Genet. Cytol.*, 19:217–223, 1977.

[54] Y.-s. Chang, F.-l. Huang, and T.-b. Lo. The complete nucleotide sequence and gene organization of carp (*Cyprinus carpio*) mitochondrial genome. *J. Mol. Evol.*, 38:138–155, 1994.

[55] K. Chikuni, Y. Mori, T. Tabata, M. Saito, M . Monma, and M. Kosugiyama. Molecular phylogeny based on the kappa-casein and cytochrome *b* sequences in the Ruminantia. Unpublished.

[56] K. Chikuni, T. Tabata, M. Saito, and M. Monma. Sequencing of mitochondrial cytochrome *b* genes for the identification of meat species. *Anim. Sci. Technol. (Jpn)*, 65:571–579, 1994.

[57] T.M. Collins, P.H. Wimberger, and G.J.P. Naylor. Compositional bias, character-state bias, and character-state reconstruction using parsimony. *Syst. Biol.*, 43:482–496, 1994.

[58] A. Cooper, C. Mourer-Chauviré, G.K. Chambers, A. von Haeseler, A.C. Wilson, and S. Pääbo. Independent origins of New Zealand moas and kiwis. *Proc. Natl. Acad. Sci. USA*, 89:8741–8744, 1992.

[59] G.B. Corbert and J.E. Hill. *A World List of Mammalian Species, Third Edition.* Oxford Univ. Press, Oxford, 1991.

[60] R.H. Crozier and Y.C. Crozier. The mitochondrial genome of the honeybee *Apis mellifera*: complete sequence and genome organization. *Genetics*, 133:97–117, 1993.

[61] J. Czelusniak, M. Goodman, B.F. Koop, D.A. Tagle, J. Shoshani, G. Braunitzer, T.K. Kleinschmidt, W.W. de Jong, and G. Matsuda. Perspectives from amino acid and nucleotide sequences on cladistic relationships among higher taxa of Eutheria. In H.H. Genoways, editor, *Current Mammalogy, Vol. 2*, pages 545–572. Plenum Press, New York, 1990.

[62] M.O. Dayhoff, R.M. Schwartz, and B.C. Orcutt. A model of evolutionary change in proteins. In M.O. Dayhoff, editor, *Atlas of Protein Sequence and Structure, Vol. 5, Suppl. 3*, pages 345–352. National Biomedical Research Foundation, Washington, D.C., 1978.

[63] F. De Meester, R. Bracha, M. Huber, Z. Keren, S. Rozenblatt, and D. Mirelman. Cloning and characterization of an unusual elongation factor-1$\alpha$ cDNA from *Entamoeba histolytica*. *Mol. Biochem. Parasitol.*, 44:23–32, 1991.

[64] R.W. DeBry. The consistency of several phylogeny-inference methods under varying evolutionary rates. *Mol. Biol. Evol.*, 9:537–551, 1992.

[65] P. Desjardins and R. Morais. Sequence and gene organization of the chicken mitochondrial genome: a novel gene order in higher vertebrates. *J. Mol. Biol.*, 212:599–634, 1990.

[66] T.S. DeWalt, P.D. Sudman, M.S. Hafner, and S.K. Davis. Phylogenetic relationships of pocket gophers (*Cratogeomys* and *Pappogeomys*) based on mitochondrial DNA cytochrome *b* sequences. *Mol. Phyl. Evol.*, 2:193–204, 1993.

[67] T.R. Disotell, R.L. Honeycutt, and M. Ruvolo. Mitochondrial DNA phylogeny of the Old-World monkey tribe Papionini. *Mol. Biol. Evol.*, 9:1–13, 1992.

[68] J. Doebley, M. Durbin, E.M. Golenberg, M.T. Clegg, and Ma D.-P. Evolutionary analysis of the large subunit of carboxylase (*rbcL*) nucleotide sequence among the grasses (Gramineae). *Evolution*, 44:1097–1108, 1990.

[69] G. D'Onofrio, D. Mouchiroud, B. Aïssani, C. Gautier, and G. Bernardi. Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *J. Mol. Evol.*, 32:504–510, 1991.

[70] A.W.F. Edwards. Assessing molecular phylogenies. *Science*, 267:253–253, 1995.

[71] A.W.F. Edwards and L.L. Cavalli-Sforza. The reconstruction of evolution. *Ann. Hum. Genet.*, 27:105, 1963.

[72] S.V. Edwards, P. Arctander, and A.C. Wilson. Mitochondrial resolution of a deep branch in the genealogical tree for perching birds. *Proc. Roy. Soc. London*, B243:99–107, 1991.

[73] J. Felsenstein. Maximum-likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Syst. Zool.*, 22:240–249, 1973.

[74] J. Felsenstein. Cases in which parsimony and compatibility methods will be positively misleading. *Syst. Zool.*, 27:401–410, 1978.

[75] J. Felsenstein. The number of evolutionary trees. *Syst. Zool.*, 27:27–33, 1978.

[76] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.*, 17:368–376, 1981.

[77] J. Felsenstein. Methods for inferring phylogenies: a statistical view. In J. Felsenstein, editor, *Numerical Taxonomy*, pages 315–334. Springer-Verlag, Berlin, 1983.

[78] J. Felsenstein. Statistical inference of phylogenies. *J. Roy. Statist. Soc.*, A146:246–272, 1983.

[79] J. Felsenstein. The statistical approach to inferring evolutionary trees and what it tells us about parsimony and compatibility. In T. Duncan and T.F. Stuessy, editors, *Cladistics: Perspectives on the Reconstruction of Evolutionary History*, pages 169–191. Columbia Univ. Press, New York, 1984.

[80] J. Felsenstein. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution*, 39:783–791, 1985.

[81] J. Felsenstein. Phylogenies from molecular sequences: inference and reliability. *Annu. Rev. Genet.*, 22:521–565, 1988.

[82] J. Felsenstein. *PHYLIP, ver. 3.5c.* Univ. of Washington, Seattle, 1993.

[83] W.M. Fitch. Toward defining the course of evolution: minimum change for a specified tree topology. *Syst. Zool.*, 20:406–416, 1971.

[84] W.M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

[85] W.M. Fitch and E. Markowitz. An improved method for determining codon variability in a gene and its application to the rate of fixations of mutations in evolution. *Biochem. Genet.*, 4:579–593, 1970.

[86] K. Fukami-Kobayashi and Y. Tateno. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J. Mol. Evol.*, 32:79–91, 1991.

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

[88] J.C. Garza and D.S. Woodruff. A phylogenetic study of the gibbons (*Hylobates*) using DNA obtained noninvasively from hair. *Mol. Phyl. Evol.*, 1:202–210, 1992.

[89] T. Gojobori, K. Ishii, and M. Nei. Estimation of average number of nucleotide substitutions when the rate of substitution varies with nucleotides. *J. Mol. Evol.*, 18:414–423, 1982.

[90] T. Gojobori, W.-H. Li, and D. Graur. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.*, 18:360–369, 1982.

[91] G.B. Golding and R.S. Gupta. Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol. Biol. Evol.*, 12:1–6, 1995.

[92] N. Goldman. Maximum likelihood inference of phylogenetic trees, with special reference to a Poisson process model of DNA substitution and to parsimony analyses. *Syst. Zool.*, 39:345–361, 1990.

[93] N. Goldman. Statistical tests of models of DNA substitution. *J. Mol. Evol.*, 36:182–198, 1993.

[94] N. Goldman and Z. Yang. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.*, 11:725–736, 1994.

[95] E.M. Golenberg, D.E. Giannasi, M.T. Clegg, C.J. Smiley, M. Durbin, D. Henderson, and G. Zurawski. Chlorolplast DNA sequence from a Miocene Magnolia species. *Nature*, 344:656–658, 1990.

[96] G.H. Gonnet, M.A. Cohen, and S.A. Benner. Exhaustive matching of the entire protein sequence database. *Science*, 256:1443–1445, 1992.

[97] I.L. Gonzalez, J.E. Sylvester, T.F. Smith, D. Stambolian, and R.D. Schmickel. Ribosomal RNA gene sequences and hominoid phylogeny. *Mol. Biol. Evol.*, 7:203–219, 1990.

[98] R. Grantham. Amino acid difference formula to help explain protein evolution. *Science*, 185:862–864, 1974.

[99] D. Graur, W.A. Hide, and W.-H. Li. Is the guinea-pig a rodent? *Nature*, 351:649–652, 1991.

[100] D. Graur and D.G. Higgins. Molecular evidence for the inclusion of Cetaceans within the order Artiodactyla. *Mol. Biol. Evol.*, 11:357–364, 1994.

[101] C.P. Groves. *A Theory of Human and Primate Evolution.* Clarendon Press, Oxford, 1989.

[102] P. Groves and G.F. Shields. Convergent evolution of the Asian takin and Arctic muskox. Unpublished.

[103] M.S. Hafner, P.D. Sudman, F.X. Villablanca, T.A. Spradling, J.W. Demastes, and S.A. Nadler. Disparate rates of molecular evolution in conspeciating hosts and parasites. *Science*, 265:1087–1090, 1994.

[104] M. Hasegawa. Phylogeny and molecular evolution in primates. *Jpn. J. Genet.*, 65:243–265, 1990.

[105] M. Hasegawa. Molecular phylogeny and man's place in Hominoidea. *J. Anthrop. Soc. Nippon*, 99:49–61, 1991.

[106] M. Hasegawa. Evolution of hominoids as inferred from DNA sequences. In T. Nishida, W.C. McGrew, P. Marler, M. Pickford, and F.B.M. de Waal, editors, *Topics in Primatology, Vol. 1. Human Origins*, pages 347–357. Univ. Tokyo Press, Tokyo, 1992.

[107] M. Hasegawa. Inference of evolutionary trees from DNA and protein sequence data. In H. Bozdogan, editor, *The Frontiers of Statistical Modeling: An Informational Approach, Vol. 3, Engineering and Scientific Applications*, pages 241–248. Kluwer Academic Publ., Dordrecht, 1994.

[108] M. Hasegawa, Y. Cao, J. Adachi, and T. Yano. Rodent polyphyly? *Nature*, 355:595–595, 1992.

[109] M. Hasegawa, A. Di Rienzo, T.D. Kocher, and A.C. Wilson. Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.*, 37:347–354, 1993.

[110] M. Hasegawa and M. Fujiwara. Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phyl. Evol.*, 2:1–5, 1993.

[111] M. Hasegawa and T. Hashimoto. Ribosomal RNA trees misleading? *Nature*, 361:23–23, 1993.

[112] M. Hasegawa, T. Hashimoto, and J. Adachi. Origin and evolution of eukaryotes as inferred from protein sequence data. In H. Hartman and K. Matsuno, editors, *The Origin and Evolution of the Cell*, pages 107–130. World Scientific, Singapore, 1992.

[113] M. Hasegawa, T. Hashimoto, J. Adachi, N. Iwabe, and T. Miyata. Early divergences in the evolution of eukaryotes: Ancient divergence of *Entamoeba* that lacks mitochondria revealed by protein sequence data. *J. Mol. Evol.*, 36:380–388, 1993.

[114] M. Hasegawa and S. Horai. Time of the deepest root for polymorphism in human mitochondrial DNA. *J. Mol. Evol.*, 32:37–42, 1991.

[115] M. Hasegawa, Y. Iida, T. Yano, F. Takaiwa, and M. Iwabuchi. Phylogenetic relationships among eukaryotic kingdoms inferred from ribosomal RNA sequences. *J. Mol. Evol.*, 22:32–38, 1985.

[116] M. Hasegawa, N. Iwabe, Y. Mukohata, and T. Miyata. Close evolutionary relatedness of archaebacteria, Methanococcus and Halobacterium, to eukaryotes demonstrated by composite phylogenetic trees of elongation factors EF-Tu and EF-G: eocyte tree is unlikely. *Jpn. J. Genet.*, 65:109–114, 1990.

[117] M. Hasegawa and H. Kishino. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. *Evolution*, 43:672–677, 1989.

[118] M. Hasegawa and H. Kishino. Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. *Jpn. J. Genet.*, 64:243–258, 1989.

[119] M. Hasegawa and H. Kishino. Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. *Mol. Biol. Evol.*, 11:142–145, 1994.

[120] M. Hasegawa, H. Kishino, K. Hayasaka, and S. Horai. Mitochondrial DNA evolution in primates: Transition rate has been extremely low in lemur. *J. Mol. Evol.*, 31:113–121, 1990.

[121] M. Hasegawa, H. Kishino, and N. Saitou. On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.*, 32:443–445, 1991.

[122] M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.*, 22:160–174, 1985.

[123] M. Hasegawa, H. Kishino, and T. Yano. Man's place in Hominoidea as inferred from molecular clocks of DNA. *J. Mol. Evol.*, 26:132–147, 1987.

[124] M. Hasegawa and T. Yano. Classification of amino acids and its implication to the genetic code. *Viva Origino*, 4:11–18, 1975.

[125] M. Hasegawa and T. Yano. Maximum likelihood method of phylogenetic inference from DNA sequence data. *Bull. Biomet. Soc. Japan*, 5:1–7, 1984.

[126] M. Hasegawa and T. Yano. Phylogeny and classification of Hominoidea as inferred from DNA sequence data. *Proc. Japan Acad.*, B60:389–392, 1984.

[127] M. Hasegawa, T. Yano, and H. Kishino. A new molecular clock of mitochondrial DNA and the evolution of hominoids. *Proc. Japan Acad.*, B60:95–98, 1984.

[128] M. Hasegawa, T. Yano, and T. Miyata. Evolutionary implications of error amplification in the self-replicating and protein-synthesizing machinery. *J. Mol. Evol.*, 20:77–85, 1984.

[129] T. Hashimoto, J. Adachi, and M. Hasegawa. Phylogenetic place of *Giardia lamblia*, a protozoan that lacks mitochondria. *Endocytobiosis and Cell Research*, 9:59–69, 1992.

[130] T. Hashimoto, Y. Nakamura, T. Kamaishi, J. Adachi, F. Nakamura, K. Okamoto, and M. Hasegawa. Phylogenetic place of kinetoplastid protozoa inferred from a protein phylogeny of elongation factor 1$\alpha$. *Mol. Biochem. Parasitol.*, in press, 1995.

[131] T. Hashimoto, Y. Nakamura, T. Kamaishi, F. Nakamura, J. Adachi, K. Okamoto, and M. Hasegawa. Phylogenetic place of a mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. *Mol. Biol. Evol.*, in press, 1995.

[132] T. Hashimoto, Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa. Protein phylogeny gives a robust estimation for early divergences of eukaryotes: phylogenetic place of a mitochondria-lacking protozoan, *Giardia lamblia*. *Mol. Biol. Evol.*, 11:65–71, 1994.

[133] T. Hashimoto, E. Otaka, J. Adachi, K. Mizuta, and M. Hasegawa. The giant panda is most close to a bear, judged by $\alpha$- and $\beta$-hemoglobin sequences. *J. Mol. Evol.*, 36:282–289, 1993.

[134] K. Hayasaka, T. Gojobori, and S. Horai. Molecular phylogeny and evolution of primate mitochondrial DNA. *Mol. Biol. Evol.*, 5:626–644, 1988.

[135] S.B. Hedges, C.A. Hass, and L.R. Maxson. Relations of fish and tetrapods. *Nature*, 363:501–502, 1993.

[136] S.B. Hedges and C.G. Sibley. Molecular vs. morphology in avian evolution: The case of the "pelecaniform" birds. *Proc. Natl. Acad. Sci. USA*, 91:9861–9865, 1994.

[137] K. Helm-Bychowski and J. Cracraft. Recovering phylogenetic signal from DNA sequences: Relationships within corvine assemblage (class Aves) as inferred from complete sequences of the mitochondrial DNA cytochrome-*b* gene. *Mol. Biol. Evol.*, 10:1196–1214, 1993.

[138] S. Henikoff and J.G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.

[139] D.M. Hillis, J.P. Huelsenbeck, and D.L. Swofford. Hobglobin of phylogenetics. *Nature*, 369:363–364, 1994.

[140] S. Horai, K. Hayasaka, R. Kondo, K. Tsugane, and N. Takahata. The recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA*, 92:532–536, 1995.

[141] S. Horai, Y. Satta, K. Hayasaka, R. Kondo, T. Inoue, T. Ishida, S. Hayashi, and N. Takahata. Man's place in Hominoidea revealed by mitochondrial DNA genealogy. *J. Mol. Evol.*, 35:32–43, 1992.

[142] S. Horai, Y. Satta, K. Hayasaka, R. Kondo, T. Inoue, T. Ishida, S. Hayashi, and N. Takahata. Man's place in Hominoidea revealed by mitochondrial DNA genealogy (Erratum). *J. Mol. Evol.*, 37:89–89, 1993.

[143] B. Hovemann and S. Richer. Two genes encode related cytoplasmic elongation factors 1-$\alpha$ (EF-1) in *Drosophila melanogaster* with continuous and stage specific expression. *Nucl. Acids. Res.*, 16:3175–3194, 1988.

[144] I.M. Ibrahimi, E.M. Prager, T.J. White, and A.C. Wilson. Amino acid sequence of California quail lysozyme. Effect of evolutionary substitutions on the antigenic structure of lysozyme. *Biochemistry*, 18:2736–2744, 1979.

[145] D.M. Irwin and Ú. Árnason. Cytochrome *b* gene of marine mammals: phylogeny and evolution. *J. Mammal. Evol.*, 2:37–55, 1994.

[146] D.M. Irwin, T.D. Kocher, and A.C. Wilson. Evolution of the cytochrome *b* gene of mammals. *J. Mol. Evol.*, 32:128–144, 1991.

[147] IUPAC-IUB Commission on Biochemical Nomenclature. A one-letter notation for amino acid sequences, tentative rules. *J. Biol. Chem.*, 243:3557–3559, 1968.

[148] N. Iwabe, K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc. Natl. Acad. Sci. USA*, 86:9355–9359, 1989.

[149] N. Iwabe, K. Kuma, H. Kishino, M. Hasegawa, and T. Miyata. Evolution of RNA polymerases and branching patterns of the three major groups of archaebacteria. *J. Mol. Evol.*, 32:70–78, 1991.

[150] A. Janke, G. Feldmaier-Fuchs, W.K. Thomas, A. von Haeseler, and S. Pääbo. The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics*, 137:243–256, 1994.

[151] S. Johansen and T. Johansen. Sequence analysis of twelve structural genes and a novel non-coding region from mitochondrial DNA of Atlantic cod *Gadus morhua*. *Biochim. Biophys. Acta*, 1218:2130–2170, 1994.

[152] J. Jollès, F. Schoentgen, P. Jollès, E.M. Prager, and A.C. Wilson. Amino acid sequence and immunological properties of chachalaca egg white lysozyme. *J. Mol. Evol.*, 8:59–78, 1976.

[153] P. Jollès and J. Jollès. What's new in lysozyme research? always a model system, today as yesterday. *Mol. Cell. Biochem.*, 63:165–189, 1984.

[154] D.T. Jones, W.R. Taylor, and J.M. Thornton. The rapid genaration of mutation data matrices from protein sequences. *Comp. Appl. Biosci.*, 8:275–282, 1992.

[155] T.H. Jukes and V. Bhushan. Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *J. Mol. Evol.*, 24:39–44, 1986.

[156] T.H. Jukes and C.R. Cantor. Evolution of protein molecules. In H.N. Munro, editor, *Mammalian Protein Metabolism, Vol.III*, pages 21–132. Academic Press, New York, 1969.

[157] C. Kelly. A test of Markovian model of DNA evolution. *Biometrics*, 50:653–6641, 1994.

[158] Y. Kikkawa, T. Amano, and H. Suzuki. Analysis of genetic diversity of domestic cattle in east and South-East Asia in terms of variations in restriction sites and sequences of mitochondrial DNA. Unpublished.

[159] Y. Kikkawa, H. Suzuki, H. Yonekawa, and T. Amano. Genetic diversity and geographic distribution of asian domestic water buffaloes based on the variations in restriction sites and sequences of mitochondrial DNA. Unpublished.

[160] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.

[161] M. Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120, 1980.

[162] M. Kimura. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA*, 78:454–458, 1981.

[163] M. Kimura. *The Neutral Theory of Molecular Evolution.* Cambridge Univ. Press, Cambridge, 1983.

[164] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.*, 29:170–179, 1989.

[165] H. Kishino and M. Hasegawa. Converting distance to time: an application to human evolution. *Methods in Enzymology*, 183:550–570, 1990.

[166] H. Kishino, T. Miyata, and M. Hasegawa. Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. *J. Mol. Evol.*, 31:151–160, 1990.

[167] T. Kleinschmidt, J. Czelusniak, M. Goodman, and G. Braunitzer. Paenungulata: a comparison of the hemoglobin sequences from elephant, hyrax, and manatee. *Mol. Biol. Evol.*, 3:427–435, 1986.

[168] H.-P. Klenk and W. Zillig. DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J. Mol. Evol.*, 38:420–432, 1994.

[169] T.D. Kocher and A.C. Wilson. Sequence evolution of mitochondrial DNA in humans and chimpanzees: Control region and a protein-coding region. In S. Osawa and T. Honjo, editors, *Evolution of Life: Fossils, Molecules, and Culture*, pages 391–413. Springer-Verlag, Tokyo, 1991.

[170] S. Kojima, T. Hashimoto, M. Hasegawa, S. Murata, S. Ohta, H. Seki, and N. Okada. Close phylogenetic relationship between Vestimentifera (tube worms) and Annelida revealed by the amino acid sequence of elongation factor-1$\alpha$. *J. Mol. Evol.*, 37:66–70, 1993.

[171] R. Kondo. *Evolution and Phylogeny of Hominoids Inferred from Mitochondrial DNA Sequences.* Ph.D. dissertation, The Graduate University for Advanced Studies, 1992.

[172] R. Kondo, S. Horai, Y. Satta, and N. Takahata. Evolution of hominoid mitochondrial DNA with special reference to the silent substitution rate over the genome. *J. Mol. Evol.*, 36:517–531, 1993.

[173] J.R. Kornegay, T.D. Kocher, L.A. Williams, and A.C. Wilson. Pathways of lysozyme evolution inferred from the sequences of cytochrome *b* in birds. *J. Mol. Evol.*, 37:367–379, 1993.

[174] C. Krajewski and J.W. Fetzner, Jr. Phylogeney of cranes (Gruiformes: Gruidae) based on cytochrome-*b* DNA sequences. *Auk*, 111:351–365, 1994.

[175] C. Krajewski, J. Painter, L. Buckley, and M. Westerman. Phylogenetic structure of the marsupial family Dasyuridae based on cytochrome *b* DNA sequences. *J. Mammal. Evol.*, 2:25–35, 1994.

[176] P.A. Krieg, S.M. Varnum, W.M. Wormington, and D.A Melton. The mRNA encoding elongation factor 1$\alpha$ (EF-1$\alpha$) is a major transcript at the midblastula transition in Xenopus. *Dev. Biol.*, 133:93–100, 1989.

[177] M.K. Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459–468, 1994.

[178] K. Kuma and T. Miyata. Mammalian phylogeny inferred from multiple protein data. *Jpn. J. Genet.*, 69:555–566, 1994.

[179] K. Kuma, N. Nikoh, N. Iwabe, and T. Miyata. Phylogenetic position of *Dictyostelium* inferred from multiple protein data sets. *J. Mol. Evol.*, 1995.

[180] Y. Kurasawa, O. Numata, M. Katoh, H. Hirano, J. Chiba, and Y. Watanabe. Identification Tetrahymena 14-nm filament-associated protein as elongation factor 1$\alpha$. *Exp. Cell Res.*, 203:251–258, 1992.

[181] R. Kusmierski, G. Borgia, R.H. Crozier, and B.H.Y. Chan. Molecular information on bowerbird phylogeny and the evolution of exaggerated male characteristics. *J. Evol. Biol.*, 6:737–752, 1993.

[182] J.A. Lake. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA*, 91:1455–1459, 1994.

[183] S.M. Lanyon. Polyphyly of the blackbird genus *Agelaius* and the importance of assumptions of monophyly in comparative studies. *Evolution*, 48:679–693, 1994.

[184] S. Länge, C. Rozario, and M. Müller. Primary structure of the hydrogenosomal adenylate kinase of *Trichomonas vaginalis* and its phylogenetic relationships. *Mol. Biochem. Parasitol.*, 66:297–308, 1994.

[185] K. Lechner and A. Böck. Cloning and nucleotide sequence of the gene for an archaebacterial protein synthesis elongation factor Tu. *Mol. Gen. Genet.*, 208:523–528, 1987.

[186] G. Lecointre, H. Philippe, H.L.V. Lê, and H. Le Guyader. Species sampling has a major impact on phylogenetic inference. *Mol. Phyl. Evol.*, 2:205–224, 1993.

[187] W.J. Lee and T.D. Kocher. Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. *Genetics*, 139:873–887, 1995.

[188] P.R. Leeton, L. Christidis, M. Westerman, and W.E. Boles. Molecular phylogenetic affinities of the night parrot (*Geopsittacus occidentalis*) and the ground parrot (*Pezopotus wallicus*). *Auk*, 111:831–841, 1994.

[189] D.D. Leipe, J.H. Gunderson, T.A. Nerad, and M.L. Sogin. Small subunit ribosomal RNA$^+$ of *Hexamita inflata* and the quest for the first branch in the eukaryotic tree. *Mol. Biochem. Parasitol.*, 59:41–48, 1993.

[190] G.M. Lento, R.E. Hickson, G.K. Chambers, and D. Penny. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol. Biol. Evol.*, 12:28–52, 1995.

[191] D.H. Les, D.K. Garvin, and C.F. Wimpee. Molecular evolutionary history of ancient aquatic angiosperms. *Proc. Natl. Acad. Sci. USA*, 88:10119–10123, 1991.

[192] W.-H. Li and D. Graur. *Molecular Evolution.* Sinauer Associates, Sunderland, Massachusetts, 1991.

[193] T. Liboz, C. Bardet, A. Le Van Thai, M. Axelos, and B. Lescure. The four members of the gene family encoding the *A. thaliana* translation elongation factor. *Plant Mol. Biol.*, 14:107–110, 1989.

[194] J.E. Linz, L.M. Lira, and P.S. Sypherd. The primary structure and the functional domains of an elongation factor-1$\alpha$ from *Mucor racemosus*. *J. Biol. Chem.*, 261:15022–15029, 1986.

[195] H. Liu and A.T. Beckenbach. Evolution of the mitochondrial cytochrome oxidase II gene among 10 orders of insects. *Mol. Phyl. Evol.*, 1:41–52, 1992.

[196] P.J. Lockhart, C.J. Howe, D.A. Bryant, T.J. Beanland, and A.W.D. Larkum. Substitutional bias confounds inference of cyanelle origins from sequence data. *J. Mol. Evol.*, 34:153–162, 1992.

[197] P.J. Lockhart, M.A. Steel, M.D. Hendy, and D. Penny. Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, 11:605–612, 1994.

[198] W.F. Loomis and D.W. Smith. Molecular phylogeny of *Dictyostelium discoideum* by protein sequence comparison. *Proc. Natl. Acad. Sci. USA*, 87:9093–9097, 1990.

[199] D.-P. Ma, A. Zharkikh, D. Graur, J.L. VandeBerg, and W.-H. Li. Structure and evolution of opossum, guinea pig, and porcupine cytochrome *b* genes. *J. Mol. Evol.*, 36:327–334, 1993.

[200] B.A. Malcolm, K.P. Wilson, B.W. Matthews, J.F. Kirsch, and A.C. Wilson. Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, 345:86–89, 1990.

[201] L. Margulis and K.V. Schwartz. *Five Kingdoms: An Illustrated Guide to the Phyla of Life on Earth. 2nd Edition.* W.H. Freeman, New York, 1988.

[202] T.L. Marsh, C.I. Reich, R.B. Whitelock, and G.L. Olsen. Transcrition factor IID in the Archaea: sequences in the *Thermococcus celer* genome would encode a product closely related to the TATA-binding protein of eukaryotes. *Proc. Natl. Acad. Sci. USA*, 91:4180–4184, 1994.

[203] A. Martin. Hammerhead shark origins. *Nature*, 364:494–494, 1993.

[204] A.P. Martin, G.J.P. Naylor, and S.R. Palumbi. Rates of mitochondrial DNA evolution in sharks are slow compared with mammals. *Nature*, 357:153–155, 1992.

[205] A.P. Martin and S.R. Palumbi. Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. USA*, 90:4087–4091, 1993.

[206] A.P. Martin and S.R. Palumbi. Protein evolution in different cellular environments: cytochrome *b* in sharks and mammals. *Mol. Biol. Evol.*, 10:873–891, 1993.

[207] M.L. McCrossin and B.R. Benefit. Recently recovered *Kenyapithecus* mandible and its implications for great ape and human origins. *Proc. Natl. Acad. Sci. USA*, 90:1962–1966, 1993.

[208] A.D. McLachlan. Tests for comparing related amino-acid sequences. Cytochrome *c* and cytochrome $c_{551}$. *J. Mol. Biol.*, 61:409–424, 1971.

[209] A. Meyer. Molecular phylogenetic studies of fish. In A.R. Beaumont, editor, *Genetics and Evolution of Aquatic Organisms*, pages 219–249. Chapman & Hall, London, 1994.

[210] A. Meyer, J.M. Morrissey, and M. Schartl. Recurrent origin of a sexually selected trait in *Xiphophorus* fishes inferred from a molecular phylogeny. *Nature*, 368:539–542, 1994.

[211] M.C. Milinkovitch, A. Meyer, and J.R. Powell. Phylogeny of all major groups of cetaceans based on DNA sequences from three mitochondrial genes. *Mol. Biol. Evol.*, 11:939–948, 1994.

[212] M.C. Milinkovitch, G. Orti, and A. Meyer. Revised phylogeny of whales suggested by mitochondrial ribosomal DNA sequences. *Nature*, 361:346–348, 1993.

[213] M.C. Milinkovitch, G. Orti, and A. Meyer. Novel phylogeny of whales revisited but not revised. *Mol. Biol. Evol.*, 12:in press, 1995.

[214] D.P. Mindell, C.W. Dick, and R.J. Baker. Phylogenetic relationships among megabats, microbats, and primates. *Proc. Natl. Acad. Sci. USA*, 88:10322–10326, 1991.

[215] M.M. Miyamoto, M.W. Allard, R.M. Adkins, L.L. Janecek, and R.L. Honeycutt. A congruence test of reliability using linked mitochondrial DNA sequences. *Syst. Biol.*, 43:236–249, 1994.

[216] M.M. Miyamoto, J.L. Slightom, and M. Goodman. Phylogenetic relations of humans and African apes from DNA sequences in the $\psi\eta$-globin region. *Science*, 238:369–373, 1987.

[217] T. Miyata, H. Hayashida, R. Kikuno, M. Hasegawa, M. Kobayashi, and K. Koike. Molecular clock of silent substitution: at least six-fold preponderance of silent changes in mitochondrial genes over those in nuclear genes. *J. Mol. Evol.*, 19:28–35, 1982.

[218] T. Miyata, N. Iwabe, K. Kuma, Y. Kawanishi, M. Hasegawa, H. Kishino, Y. Mukohata, K. Ihara, and S. Osawa. Evolution of archaebacteria: Phylogenetic relationships among archaebacteria, eubacteria, and eukaryotes. In S. Osawa and T. Honjo, editors, *Evolution of Life: Fossils, Molecules, and Culture*, pages 337–351. Springer-Verlag, Tokyo, 1991.

[219] P.E. Montandon and E. Stutz. Structure and expression of the *Euglena gracilis* nuclear gene coding for the translation elongation factor EF-1a. *Nucl. Acids. Res.*, 18:75–82, 1990.

[220] Y. Mukohata, K. Ihara, H. Kishino, M. Hasegawa, N. Iwabe, and T. Miyata. Close evolutionary relatedness of archaebacteria with eukaryotes. *Proc. Japan Acad.*, 66B:63–67, 1990.

[221] S.V. Muse and B.S. Gaut. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11:715–724, 1994.

[222] K. Nagashima, M. Kasai, S. Nagata, and Y. Kaziro. Structure of the two genes coding for polypeptide chain elongation factor 1-$\alpha$ (EF-1-$\alpha$) from *Saccharomyces cerevisiae*. *Gene*, 45:265–273, 1986.

[223] G.J.P. Naylor, T.M. Collins, and W.M. Brown. Hydrophobicity and phylogeny. *Nature*, 373:565–566, 1995.

[224] M. Nei. *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York, 1987.

[225] J. Neyman. Molecular studies of evolution: a source of novel statistical problems. In S.S. Gupta and J. Yackel, editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York, 1971.

[226] N. Nikoh, N. Hayase, N. Iwabe, K. Kuma, and T. Miyata. Phylogenetic relationship of the kingdoms Animalia, Plantae, and Fungi inferred from twenty three different protein species. *Mol. Biol. Evol.*, 11:762–768, 1994.

[227] B.B. Normark, A.R. McCune, and R.G. Harrison. Phylogenetic relationships of neopterygian fishes, inferred from mitochondrial DNA sequences. *Mol. Biol. Evol.*, 8:819–834, 1991.

[228] M.J. Novacek. Mammalian phylogeny: shaking the tree. *Nature*, 356:121–125, 1992.

[229] M.J. Novacek. Genes tell a new whale tale. *Nature*, 361:298–299, 1993.

[230] R.M. Nowak. *Walker's Mammals of the World, Fifth Edition*. Johns Hopkins University Press, Baltimore, 1991.

[231] G.J. Olsen, H. Matsuda, R. Hagstrom, and R. Overbeek. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comp. Appl. Biosci.*, 10:41–48, 1994.

[232] T. Ozawa, M. Tanaka, H. Ino, K. Ohno, T. Sano, Y. Wada, M. Yoneda, Y. Tanno, T. Miyatake, T. Tanaka, S. Itoyama, S. Ikebe, N. Hattori, and Y. Mizuno. Distinct clustering of point mutations in mitochondrial DNA among patients with mitochondrial encephalomyopathies and Parkinson's disease. *Biochem. Biophys. Res. Commun.*, 176:938–946, 1991.

[233] J. Painter, C.W. Krajewski, and M. Westerman. Molecular phylogeny for the marsupial genus Planigale (Dasyuridae). Unpublished.

[234] D.P. Pashley and L.D. Ke. Sequence evolution in mitochondrial ribosomal and ND-1 genes in Lepidoptera: implications for phylogenetic analyses. *Mol. Biol. Evol.*, 9:1061–1075, 1992.

[235] N.T. Perna and T.D. Kocher. Unequal base frequencies and the estimation of substitution rates. *Mol. Biol. Evol.*, 12:359–361, 1995.

[236] G. Pesole, E. Ebisá, G. Preparata, and C. Saccone. The evolution of the mitochondrial D-loop region and the origin of modern man. *Mol. Biol. Evol.*, 9:587–598, 1992.

[237] J.D. Pettigrew. Flying primates? Megabats have the advanced pathway from eye to midbrain. *Science*, 231:1304–1306, 1986.

[238] J.D. Pettigrew. A fruitful, wrong hypothesis? Response to Baker, Novacek, and Simmons. *Syst. Zool.*, 40:231–239, 1991.

[239] J.D. Pettigrew. Wings or brain? Convergent evolution in the origins of bats. *Syst. Zool.*, 40:199–216, 1991.

[240] H. Philippe and E. Douzery. The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. *J. Mammal. Evol.*, 2:133–152, 1994.

[241] D. Pilbeam. Human origins and evolution. In A.C. Fabian, editor, *Origins*, pages 89–114. Cambridge Univ. Press, Cambridge, 1988.

[242] A.R. Pokalsky, W.R. Hiatt, N. Ridge, R. Rasmussen, C.M. Houck, and C.K. Shewmaker. Structure and expression of elongation factor $1\alpha$ in tomato. *Nucl. Acids. Res.*, 17:4661–4673, 1989.

[243] R. Ramharack and R.G. Deeley. Structure and evolution of primate cytochrome *c* oxidase subunit II gene. *J. Biol. Chem.*, 262:14014–14021, 1987.

[244] J.H. Reeves. Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J. Mol. Evol.*, 35:17–31, 1992.

[245] A.D. Richman and T. Price. Evolution of ecological differences in the Old World leaf warblers. *Nature*, 355:817–821, 1992.

[246] C. Richter, J.-W. Park, and B.N. Ames. Normal oxidative damage to mitochondrial and nuclear DNA is extensive. *Proc. Natl. Acad. Sci. USA*, 85:6465–6467, 1988.

[247] F. Rodríigues, J.L. Oliver, A. Marín, and J.R. Medina. The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, 142:485–501, 1990.

[248] B.A. Roe, D.-P. Ma, R.K. Wilson, and J.F.-H. Wong. The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. *J. Biol. Chem.*, 260:9759–9774, 1985.

[249] M. Ruvolo, T.R. Disotell, M.W. Allard, W.M. Brown, and R.L. Honeycutt. Resolution of the African hominoid trichotomy by use of a mitochondrial gene sequence. *Proc. Natl. Acad. Sci. USA*, 88:1570–1574, 1991.

[250] M. Ruvolo, S. Zehr, M. von Dornum, D. Pan, B. Chang, and J. Lin. Mitochondrial COII sequences and modern human origins. *Mol. Biol. Evol.*, 10:1115–1135, 1993.

[251] C. Saccone, C. Lanave, G. Pesole, and G. Preparata. Influence of base composition on quantitative estimates of gene evolution. *Methods in Enzymology*, 183:570–583, 1990.

[252] N. Saitou and T. Imanishi. Relative efficiencies of the Fitch-Margolish, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Mol. Biol. Evol.*, 6:514–525, 1989.

[253] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[254] V.M. Sarich and A.C. Wilson. Immunological time scale for hominid evolution. *Science*, 158:1200–1203, 1967.

[255] V.M. Sarich and A.C. Wilson. Rates of albumin evolution in primates. *Proc. Natl. Acad. Sci. USA*, 58:142–148, 1967.

[256] S. Sattath and A. Tversky. Additive similarity trees. *Psychometrika*, 42:319–345, 1977.

[257] M. Schöniger, G.L. Hofacker, and B. Borstnik. Stochastic traits of molecular evolution — acceptance of point mutations in native actin genes. *J. Theor. Biol.*, 143:287–306, 1990.

[258] T.R. Schmidt and J.R. Gold. Molecular phylogenetics and evolution of the cytochrome *b* gene in the cyprinid genus *Lythrurus* (Actinopterygii: Cypriniformes). *Mol. Phyl. Evol.*, in press, 1995.

[259] J.H. Schwartz and I. Tattersall. Evolutionary relationships of living lemurs and lorises (Mammalia, Primates) and their potential affinities with European Eocene Adapidae. *Anthropol. Papers Am. Mus. Nat. Hist.*, 60(1):1–100, 1985.

[260] T. Shirakura, T. Hashimoto, Y. Nakamura, T. Kamaishi, Y. Cao, J. Adachi, M. Hasegawa, A. Yamamoto, and N. Goto. Phylogenetic place of a mitochondria-lacking protozoan, *Entamoeba histolytica*, inferred from amino acid sequences of elongation factor 2. *Jpn. J. Genet.*, 69:119–135, 1994.

[261] C.G. Sibley and J.E. Ahlquist. The phylogeny of the hominoid primates, as indicated by DNA-DNA hybridization. *J. Mol. Evol.*, 20:2–15, 1984.

[262] C.G. Sibley and J.E. Ahlquist. The relationships of some groups of African birds, based on comparisons of the genetic material, DNA. In K.-L. Schuchmann, editor, *Proceedings of the International Symposium on African Vertebrates: Systematics, Phylogeny and Evolutionary Ecology*, pages 115–161. Zoologisches Forschungsinstitut und Museum Alexander Koenig, Bonn, 1985.

[263] C.G. Sibley and J.E. Ahlquist. DNA hybridization evidence of hominoid phylogeny: results from an expanded data set. *J. Mol. Evol.*, 26:99–121, 1987.

[264] C.G. Sibley and J.E. Ahlquist. *Phylogeny and Classification of Birds: A Study in Molecular Evolution.* Yale Univ. Press, New Haven, 1990.

[265] C.G. Sibley, J.A. Comstock, and J.E. Ahlquist. DNA hybridization evidence of hominoid phylogeny: a reanalysis of the data. *J. Mol. Evol.*, 30:202–236, 1990.

[266] A. Sidow. Parsimony or statistics? *Nature*, 367:26–26, 1994.

[267] A. Sidow, T. Nguyen, and T.P. Speed. Estimating the fraction of invariable codons with a capture-recapture method. *J. Mol. Evol.*, 35:253–260, 1992.

[268] A. Sidow and W.K. Thomas. A molecular evolutionary framework for eukaryotic model organisms. *Current Biol.*, 4:596–603, 1994.

[269] N.B. Simmons, M.J. Novacek, and R.J. Baker. Approaches, methods, and the future of the Chiropteran monophyly controversy: a reply to J.D. Pettigrew. *Syst. Zool.*, 40:239–243, 1991.

[270] J.F. Smothers, C.D. von Dohlen, L.H. Smith Jr., and R.D. Spall. Molecular evidence that the myxozoan protists are metazoans. *Science*, 265:1719–1721, 1994.

[271] P.H.A. Sneath and R.R. Sokal. *Numerical Taxonomy: The Principle and Practice of Numerical Classification.* Freeman, San Francisco, 1973.

[272] M.L. Sogin, U. Edman, and H. Elwood. A single kingdom of eukaryotes. In B. Fernholm, K. Bremer, and H. Jörnvall, editors, *The Hierarchy of Life*, pages 133–143. Elsevier Science Publisher, Amsterdam, 1989.

[273] R.S. Sohal, I. Svensson, and U.T. Brunk. Hydrogen peroxide production by liver mitochondria in different species. *Mech. Ageing Dev.*, 53:209–215, 1990.

[274] M.S. Springer and J.A.W. Kirsch. A molecular perspective on the phylogeny of placental mammals based on mitochondrial 12S rDNA sequences, with special reference to the problem of the Paenungulata. *J. Mammal. Evol.*, 1:149–166, 1993.

[275] M.J. Stanhope, W.J. Bailey, J. Czelusniak, M. Goodman, J.-S. Si, J. Nickerson, J.G. Sgouros, G.A.M. Singer, and T.K. Kleinschmidt. A molecular view of primate supraordinal relationships from the analysis of both nucleotide and amino acid sequences. In R.D.E. MacPhee, editor, *Primates and Their Relatives in Phylogenetic Perspective*, pages 251–292. Plenum Press, New York, 1993.

[276] H.F. Stanley, M. Kadwell, and J.C. Wheeler. Molecular evolution of the Camelidae: a mitochondrial DNA study. *Proc. R. Soc. London*, B256:1–6, 1994.

[277] D.L. Stern. A phylogenetic analysis of soldier evolution in the aphid family Hormaphididae. *Proc. R. Soc. London*, B256:203–209, 1994.

[278] C.-B. Stewart. The powers and pitfalls of parsimony. *Nature*, 361:603–607, 1993.

[279] C.-B. Stewart, J.W. Schilling, and A.C. Wilson. Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, 330:401–404, 1987.

[280] N. Sueoka. Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl. Acad. Sci. USA*, 47:1141–1149, 1961.

[281] P. Sundstrom, D. Smith, and P.S. Sypherd. Sequence analysis and expression of the two genes for elongation factor 1-$\alpha$ from the dimorphic yeast *Candida albicans*. *J. Bacteriol.*, 172:2036–2045, 1990.

[282] D.L. Swofford and G.J. Olsen. Phylogeny reconstruction. In D.M. Hillis and C. Moritz, editors, *Molecular Systematics*, pages 411–501. Sinauer Associates, Sunderland, Massachusetts, 1990.

[283] F.S. Szalay and C.C. Katz. Phylogeny of lemurs, galagos and lorises. *Folia Primatol.*, 19:88–103, 1973.

[284] N. Takahata. Population genetics of extranuclear genomes: a model and review. In T. Ohta and K. Aoki, editors, *Population Genetics and Molecular Evolution*, pages 195–212. Japan Sci. Soc. Press, Tokyo, 1985.

[285] N. Takahata. Relaxed natural selection in human populations during the Pleistocene. *Jpn. J. Genet.*, 68:539–547, 1993.

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

[287] K. Tamura. Model selection in the estimation of the number of nucleotide substitutions. *Mol. Biol. Evol.*, 11:154–157, 1994.

[288] K. Tamura and M. Nei. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10:512–526, 1993.

[289] M. Tanaka and T. Ozawa. Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, 22:327–335, 1994.

[290] R.H. Tedford. Relationships of pinnipeds to other carnivores (Mammalia). *Syst. Zool.*, 25:363–374, 1976.

[291] A. Tesch and F. Klink. Cloning and sequencing of the gene coding for the elogation factor 1$\alpha$ from the archaebacterium *Thermoplasma acidophilum*. *FEMS Microbiol. Lett.*, 71:293–298, 1990.

[292] W.K. Thomas and A.T. Beckenbach. Variation in salmonid mitochondrial DNA: evolutionary constraints and mechanisms of substitution. *J. Mol. Evol.*, 29:233–245, 1989.

[293] W.K. Thomas and S.L. Martin. A recent origin of marmots. *Mol. Phyl. Evol.*, 2:330–336, 1993.

[294] W.K. Thomas and A.C. Wilson. Evolution by base substitution in animal mitochondrial DNA. unpublished manuscript, 1991.

[295] A.G. Thorne and M.H. Wolpoff. The multiregional evolution of humans. *Sci. Amer.*, 266(4):76–83, 1992.

[296] J.L. Thorne, H. Kishino, and J. Felsenstein. Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34:3–16, 1992.

[297] C.-S. Tzeng, C.-F. Hui, S.-C. Shen, and P.C. Huang. The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. *Nucl. Acids. Res.*, 20:4853–4858, 1992.

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

[299] T. Uetsuki, A. Naito, S. Nagata, and Y. Kaziro. Isolation and characterization of the human chromosomal gene for polypeptide chain elongation factor 1-$\alpha$. *J. Biol. Chem.*, 264:5791–5798, 1989.

[300] T. Uzzell and K.W. Corbin. Fitting discrete probability distributions to evolutionary events. *Science*, 172:1089–1096, 1971.

[301] F.J. van Hemert, R. Amons, W.J.M. Pluijms, H. van Ormondt, and W. Möller. The primary structure of elongation factor EF-1$\alpha$ from the brine shrimp *Artemia*. *EMBO J.*, 3:1109–1113, 1984.

[302] L. Vigilant, M. Stoneking, H. Harpending, K. Hawkes, and A.C. Wilson. African populations and the evolution of human mitochondrial DNA. *Science*, 253:1503–1507, 1991.

[303] C.R. Vossbrinck, J.V. Maddox, S. Friedman, B.A. Debrunner-Vossbrinck, and C.R. Woese. Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes. *Nature*, 326:411–414, 1987.

[304] P.B. Vrana, M.C. Milinkovitch, J.R. Powell, and W.C. Wheeler. Higher level relationships of arctoid Carnivora based on sequence data and "total evidence". *Mol. Phyl. Evol.*, 3:47–58, 1994.

[305] P.O. Wainright, G. Hinkle, M.L. Sogin, and S.K. Stickel. Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science*, 260:340–342, 1993.

[306] S.J. Weller, D.P. Pashley, J.A. Martin, and J.L. Constable. Phylogeny of noctuoid moths and the utility of combining independent nuclear and mitochondrial genes. *Syst. Biol.*, 43:194–211, 1994.

[307] P.J. Wettstein, M.A. Strausbauch, L. Jin, J. States, R. Chakraborty, T. Lamb, and R. Riblet. Phylogeny of six *Sciurus aberti* subspecies based on nucleotide sequences of cytochrome *b*. Unpublished.

[308] A.C. Wilson, S.S. Carlson, and T.J. White. Biochemical evolution. *Annu. Rev. Biochem.*, 46:573–639, 1977.

[309] C.R. Woese. Bacterial evolution. *Microbiol. Rev.*, 51:221–271, 1987.

[310] A. Wyss. Evidence from flipper structure for a single origin of pinnipeds. *Nature*, 334:427–428, 1988.

[311] A.R. Wyss and J.J. Flynn. A phylogenetic analysis and definition of the carnivora. In F.S. Szalay, M.J. Novacek, and M.C. McKenna, editors, *Mammal Phylogeny — Placentals*, pages 32–52. Springer-Verlag, New York, 1993.

[312] A.R. Wyss, J.J. Flynn, M.A. Norell, C.C. Swisher III, R. Charrier, M.J. Novacek, and M.C. McKenna. South America's earliest rodent and recognition of a new interval of mammalian evolution. *Nature*, 365:434–437, 1993.

[313] X. Xu and Ú. Árnason. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene*, 148:357–362, 1994.

[314] Y. Yamashina. *A World List of Birds with Japanese Names.* Daigakusyorin, Tokyo, 1986.

[315] F. Yang, M. Demma, V. Warren, S. Dharmawardhane, and J. Condeelis. Identification of an actin-binding protein from *Dictyostelium* as elongation factor 1a. *Nature*, 347:494–496, 1990.

[316] Z. Yang. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401, 1993.

[317] Z. Yang. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39:105–111, 1994.

[318] S. Yokobori, M. Hasegawa, T. Ueda, N. Okada, K. Nishikawa, and K. Watanabe. Relationship among coelacanths, lungfishes and tetrapods; a phylogenetic analysis based on mitochondrial cytochrome oxidase I gene sequences. *J. Mol. Evol.*, 38:602–609, 1994.

[319] R. Zardoya, A. Garrido-Pertierra, and J.M. Bautista. The complete nucleotide sequence of mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss. J. Mol. Evol.*, in press, 1995.

[320] W. Zillig, H.-P. Klenk, P. Palm, H. Leffers, G. Pühler, F. Gropp, and R.A. Garrett. Did eukaryotes originate by a fusion event? *Endocytobiosis & Cell Res.*, 6:1–25, 1989.

[321] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H.J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York, 1965.