MOLPHY Version 2.3

Programs for Molecular Phylogenetics Based on Maximum Likelihood

Jun Adachi and Masami Hasegawa

The Institute of Statistical Mathematics
4-6-7 Minami-Azabu, Minato-ku, Tokyo 106
E-mail adachi@ism.ac.jp and hasegawa@ism.ac.jp

Contents

1 Introduction **5**

2 Modeling Molecular Evolution **7**

2.1 Modeling Nucleotide Substitutions 8

2.1.1 Markov Models of Nucleotide Substitutions 8

Transition Probability Matrices 8

Poisson Model 9

Proportional Model 10

Hasegawa, Kishino and Yano's (1985) Model 11

Tamura and Nei's (1993) Model 11

General Reversible Markov Model 11

2.1.2 ML Estimate of the Transition Probability Matrix for the REV Model 12

2.1.3 Transition Probability Matrix for the REV Model of Four-Fold Degenerate Sites of Mitochondria 13

2.1.4 Discussion 15

2.2 Modeling Amino Acid Substitution 16

2.2.1 Dayhoff Model 16

2.2.2 Jones, Taylor and Thornton's (1992) Model 19

2.2.3 General Reversible Markov Model for Mitochondrial Proteins 22

Mitochondrial DNA Sequence Data 23

Transition Probability Matrix of the mtREV Model 23

2.2.4 Discussion 30

3 Maximum Likelihood Inference of Molecular Phylogeny **31**

3.1 Evolutionary Tree Reconstruction 32

3.1.1 Phylogenetic Trees 32

3.1.2 Rooted and Unrooted Trees 32

3.2 Algorithm for ML Inference of Molecular Phylogeny 33

3.2.1 Computing the Likelihood of the Data Given a Tree 33

3.2.2 Evaluating Likelihood along a Tree . . . 36
Data Structure of a Tree . . . 36
Partial Likelihood of a Subtree . . . 37
3.2.3 Maximum Likelihood Estimation of Branch Length . . . 38
Internal Branch Length . . . 40
External Branch Length . . . 41
3.2.4 Estimation of Distances by the ML Method . . . 41
Initial Distance Matrix . . . 41
Distance Matrix Estimated by the ML Method . . . 42
3.2.5 Estimation of Initial Branch Lengths . . . 42
Initial Branch Lengths Estimated by the Least Squares Method . . . 42
3.3 Fast Computation of ML for Inferring Evolutionary Trees . . . 45
3.4 Topology Search Strategy for ML Phylogeny . . . 47
3.4.1 Topological Data Structure . . . 47
3.4.2 Automatic Topology Search by Star Decomposition . . . 48
3.4.3 Topology Search by Local Rearrangements . . . 49
3.4.4 Example of Application of the Local Rearrangements . . . 50
3.5 Approximate Likelihood Method for Exhaustive Search . . . 54

4 MOLPHY: A Computer Program Package for Molecular Phylogenetics 57

4.1 Overview of the Input and Output Formats . . . 58
4.1.1 Input Format . . . 58
4.1.2 Basic Statistics of the Data . . . 60
4.1.3 ProtML . . . 61
4.1.4 Nucleotide Sequences . . . 65
4.1.5 NucML . . . 67
4.1.6 TotalML . . . 71
4.2 ProtML: Maximum Likelihood Inference of Protein Phylogeny . . . 72
4.2.1 Options . . . 72
4.2.2 Format of Input Sequences File . . . 73
MOLPHY Format . . . 73
SEQUENTIAL Format . . . 74
COMMON Format . . . 74
INTERLEAVED Format . . . 74
Format of USER TREES File . . . 75
Format of a CONSTRAINT TREE File . . . 76

4.3 NucML: Maximum Likelihood Inference of Nucleic Acid Phylogeny 77
4.3.1 Options 77
4.4 ProtST: Basic Statistics of Protein Sequences 78
4.4.1 Options 78
4.4.2 Output Format 78
4.5 NucST: Basic Statistics of Nucleic Acid Sequences 79
4.5.1 Options 79
4.5.2 Output Format 79
4.6 NJdist: Neighbor Joining Phylogeny from Distance Matrix 80
4.6.1 Options 80
4.6.2 Input Format 80
4.6.3 Output Format 80
4.7 Utilities (Sequence Manipulations) with Perl 81

5 Applications to Biological Problems **82**

5.1 Cytochrome b 82
5.1.1 Sequence Data 82
5.1.2 ProtML Tree of 183 OTUs Obtained by Repeated Local Rearrangements 95
5.1.3 Phylogeny of Cetacea 95
5.1.4 Phylogeny of Artiodactyla 107
5.1.5 Phylogeny of Rodentia 107
5.1.6 Phylogeny of Microchiroptera 108
5.1.7 Phylogeny of Carnivora 108
5.1.8 Phylogeny of Other Mammals 108
5.1.9 Phylogeny of Aves 109
5.1.10 Phylogeny of Galliformes 110
5.1.11 Phylogeny of Fishes 111
5.2 Lysozyme — A Case of Convergent Evolution 115
5.3 Cichlid Fishes in East Africa 122
5.4 Total Evaluation of ML Analyses of Multiple Genes 130

Chapter 1

Introduction

Phylogenetic knowledge is indispensable in evolutionary biology, and molecular phylogenetics has become an important tool in inferring phylogenetic relationships among organisms. Many methods for inferring phylogenetic trees from DNA and protein sequence data have been developed (for review see Felsenstein, 1982[65], 1988[68], 1993[69]; Nei 1987[195]; Miyamoto and Cracraft 1991[185]; Hillis et al. 1996[116]; Swofford et al. 1996[240]). Among these methods, the maximum likelihood (ML) method (Felsenstein 1981[64]) is based on an explict model for the substitution process of nucleotides or amino acids, and, therefore, we can improve the method by improving the model so that it better approximates the real process. The method has a sound statistical ground (e.g., Felsenstein 1983[66]; Ritland and Clegg 1987[214]; Goldman 1990[81]; Reeves 1992[212]; Yang 1994[271]), and has proved to be powerful in recovering correct tree topologies by computer simulation studies (e.g., Hasegawa and Yano 1984[101]; Hasegawa et al. 1991[99]; Hasegawa and Fujiwara 1993[92]; Kuhner and Felsenstein 1994[159]; Gaut and Lewis 1995[76]; Huelsenbeck 1995[122]; Yang 1995[272], 1996[273]). The ML methods for molecular phylogenetic inference were reviewed recently by Hasegawa and Kishino (1996[98]) and by Swofford et al. (1996[240]).

MOLPHY is a free package of programs for molecular phylogenetics based on the ML method. In this monograph, we present the details of the methods implemented in MOLPHY (ver. 2.3), models used in the programs, user's guide for the programs, and several examples of the applications to biological problems[1].

Felsenstein (1981[64]) introduced the ML framework to phylogenetic inference based on nucleotide sequence data, and then implemented it in the program package PHYLIP (program DNAML; Felsenstein 1993[69]). Kishino et al. (1990[148]) developed a ML method for phylogenetic inference based on amino acid sequence data, and then applied it to several biological problems (Kishino et al. 1990[148]; Mukohata et al. 1990[189]; Hasegawa et al. 1990[95]; Iwabe et al. 1991[128]; Miyata et al. 1991[187]). Later, we implemented this method in the MOLPHY package; the program is called ProtML (Adachi and Hasegawa 1992[4]). ProtML proved of great use in inferring evolutionary trees even in situations where

[1] A large part of this work is from the Ph.D. thesis of J. Adachi (1995).

the parsimony method fails (e.g., Hasegawa and Fujiwara 1993[92]), and has now been applied to many phylogenetic problems (Hasegawa et al. 1992[91], 1993[94], 1996[90]; Adachi and Hasegawa, 1992[3], 1995[7], 1995[5], 1995[6], 1996[9]; Adachi et al. 1993[2]; Hasegawa and Adachi 1996[89]; Hashimoto and Hasegawa 1996[105]; Hashimoto et al. 1992[104], 1993[109], 1994[108], 1995[106], 1995[107]; Kojima et al. 1993[153]; Yokobori et al. 1994[274]; Shirakura et al. 1994[226]; Cao et al. 1994[42], 1994[41], 1994[40]; Marsh et al. 1994[179]; Klenk and Zillig 1994[150]; Länge et al. 1994[166]; Nikoh et al. 1994[197]; Kuma and Miyata 1994[160]; Kuma et al. 1995[161]; Golding and Gupta 1995[80]; Clark and Roger 1995[49]; Philippe and Adoutte 1995[207]; Shimada et al. 1995[225]; Ueda and Yoshinaga 1995[253]; Russo et al. 1996[217]; Graur et al. 1996[84]; Janke et al. 1996[130]; D'Erchia et al. 1996[56]; Caspers et al. 1996[43]; Kamaishi et al. 1996[136]; Nakamura et al. 1996[192], 1996[193]; Yamamoto et al. 1996[266]; Keeling and Doolittle 1996[139]; Baldauf et al. 1996[32]; Lawson et al. 1996[165]; Horner et al. 1996[120]; Philippe and Laurent 1996[209]; Orti and Meyer 1996[201]; Zardoya and Meyer 1996[278]; Milinkovitch et al. 1996[183])

In version 2 of MOLPHY, the program NucML for analyzing nucleotide sequences was added, and it has been used in Adachi and Hasegawa (1995[8], 1996[11]), Hasegawa and Adachi (1996[89]), Chow and Kishino (1995[48]), Orti et al. (1996[202]), Zardoya and Meyer (1996[276], 1996[277]) and Aoshima et al. (1996[17]).

Chapter 2

Modeling Molecular Evolution

A basic process in the evolution of DNA and protein sequences is the substitution of nucleotides or 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 based on stochastic models, and several of the models 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[144], 1981[145]; Takahata and Kimura 1981[241]; Gojobori et al. 1982[79], 1982[78]; Hasegawa et al. 1985[100]; Tavaré 1986[245]; Barry and Hartigan 1987[33]; Rodríguez et al. 1990[215]; Saccone et al. 1990[219]; Tamura and Nei 1993[243]; Steel et al. 1993[233]; Yang 1994[270]; Kelly 1994[140]; Adachi and Hasegawa 1996[11]) and for amino acid substitutions (Dayhoff et al. 1978[54]; Kishino et al. 1990[148]; Altschul 1991[14]; Jones et al. 1992[134]; Reeves 1992[212]; Henikoff and Henikoff 1992[113]; Gonnet et al. 1992[83]; Adachi and Hasegawa 1996[10]).

2.1 Modeling 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[118]) were examined in detail by three alternative Markov models (Adachi and Hasegawa 1995[8], 1996[11]); (1) Hasegawa, Kishino and Yano's (1985[100]) model, (2) Tamura and Nei's (1993[243]) model, and (3) the general reversible Markov model (Tavaré 1986[245]; Barry and Hartigan 1987[34], 1987[33]; Zharkikh 1994[279]; Yang 1994[270]; Adachi and Hasegawa 1995[8]). These sites are expected to be relatively free from constraint compared with other sites, and therefore their pattern of substitution should reflect that of mutation. It turned out that, among these 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 branch 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[38]; Miyata et al. 1982[186]; Hasegawa et al. 1984[103]) and hence the multiple-hit effect is great in a comparison between distantly related species, closely related species should be compared in order to accurately estimate the pattern of synonymous nucleotide substitutions of mtDNA. Horai et al. (1992[118]) 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 Matrices

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[148]) 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[54]). 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 λ_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[135]) model. Let δ 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[54]) by the following formula;

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

where π_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[64]), P_{ij} is proportional to the frequency of nucleotide j, π_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 is interpreted to be due to higher substitution probability to that nucleotide. 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[38]). Kimura (1980[144]) 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[100]) combined the Kimura model with the proportional model of Felsenstein, and this is conveniently labelled the HKY85 model. Actually, this model was first suggested in Hasegawa, Yano and Kishino (1984[102]), but since the name of HKY85 is being used widely, we will use this. 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 α and β are relative substitution rates of transition and transversion, respectively. If we fix $\beta = 1$, then α 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. Note that here $\boldsymbol{R}$ is not the overall rate matrix (e.g., as given in Swofford et al. 1996[240]), but rather this matrix with the effect of the base frequencies removed (hence relative, and not absolute rates of substitution).

Tamura and Nei's (1993) Model

Tamura and Nei (1993[243]) proposed a slightly more general model, which we call the TN93 model, than the HKY85 model. It 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 α_{Y} is the relative substitution rate between pyrimidines, α_{R} is that between purines, and β is the relative transversion rate. Given $\beta = 1$, α_{Y} and α_{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[242]) 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[118]) 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. The most general reversible model is described by Tavaré (1986[245]) and Barry and

Hartigan (1987[34], 1987[33]). Subsequently, Yang (1994[270]) estimated 4×4 transition matrices of the most general reversible Markov model (REV model) with ML. He did this for primate $\psi\eta$-globin pseudogenes and for primate mtDNA sequences including all codon positions as well as tRNAs (see also Adachi and Hasegawa 1995[8]). 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[219]) and Rodríguez et al. (1990[215]) also proposed the general reversible model. Saccone et al. (1990[219]), Tavaré (1986[245]), and Tamura (1994[242]) 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. It is desirable to estimate a single transition probability matrix from a tree, and Yang (1994[270]) 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, we will give the details of the method in this monograph, and we will further estimate the transition probability matrices of the REV model for the four-fold degenerate sites of mtDNA. We have applied this method in Adachi and Hasegawa (1995[8], 1996[11]).

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 branch, $t_1, \ldots, t_m$ (m: number of branches in the tree) by the ML method;

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

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

Our procedure to achieve the likelihood maximization is: (1) set the initial value of $\boldsymbol{R}$ by assuming the Proportional model and that of $\boldsymbol{t}$ as the ML estimate under the model. (2) Iterate the likelihood estimations of $\boldsymbol{R}$ by the Brent method and of $\boldsymbol{t}$ by the Newton-Raphson method alternately (described later in subsection3.2.3. On the iteration when the differences of all parameters between the preceding two steps are less than ϵ, a given constant, 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
    $\boldsymbol{t}^{\text{old}} \leftarrow$ the 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 procedure for the ML estimation of $\boldsymbol{R}$ under given $\boldsymbol{X}$ and $\boldsymbol{t}$, whose pseudocode is given by:

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
            ▷ maximum likelihood estimate by the Brent method
            maximize $l(R_{ij}|\boldsymbol{X}, \boldsymbol{t}, \boldsymbol{R}^*_{ij})$ ▷ $\boldsymbol{R}^*_{ij}$ is $\boldsymbol{R}$ without R_{ij}
    return $\boldsymbol{R}$
end.

MLE-Branch-Length ($\boldsymbol{X}$, $\boldsymbol{R}$, $\boldsymbol{t}^{\text{old}}$) is 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 (MOLPHY) for inferring a ML tree from nucleotide sequences.

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

2.1.3 Transition Probability Matrix for the REV Model of Four-Fold Degenerate Sites of Mitochondria

The following protein-encoding regions from Anderson et al. (1981[15]) and Horai et al. (1992[118], 1993[119]) were used. ND1 (4123–4260 using 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.

We estimated the relative substitution rate $\boldsymbol{R}$ of the REV model from the 611 sites data by the ML method 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 the unit time interval was obtained as shown in Table 2.2 (Adachi and Hasegawa 1995[8]).

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	
$\boldsymbol{\pi}$	0.167	0.421	0.366	0.046

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

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

↗	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[8]).

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[244]) 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[12], 1974[13]) defined by AIC = –2 × (log-likelihood) + 2 × (number of adjustable 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.

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. Adachi and Hasegawa (1996[11]) estimated the transition probability matrix for the REV model of the four-fold degenerate sites by using the complete mitochondrial DNA from human, common chimpanzee, bonobo, gorilla, and orangutan (Horai et al. 1995[117]), and obtained essentially the same result presented in this section.

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 ML, further complication of the model may not be 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, tRNAs, and rRNAs complications due to unequal evolutionary rate across sites and other factors become necessary as discussed by Yang (1994[270]). 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. Note that there are suggestions of a non-homogeneous, and therefore potentially non-stationary model for these same data in the work of Adachi and Hasegawa (1996[11]) and Waddell and Steel (1996[258]). So we should be cautious about this. This problem may become serious when we compare different mammalian orders (Cao et al. 1994[40]).

The different nucleotide frequencies between species is often a serious problem in inferring trees (e.g., Hasegawa and Hashimoto 1993[93]; Weisburg et al. 1989[260]). Where genomes have acquired similar nucleotide frequencies 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[164]), Lockhart et al. (1994[173]), and Galtier and Gouy (1995[74]) in the framework of distance methods, but it remains to be studied in the framework of the ML method.

2.2 Modeling 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 so it seems preferable to model the amino acid substitution.

Since selective constraints are 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 amassed data sets which consist mainly of nuclear-encoded proteins (Dayhoff et al. 1978[54]; Jones et al. 1992[134]).

For amino acid substitutions, $\boldsymbol{R}$ (the relative substitution rates) is related to the accepted mutation matrix $\boldsymbol{A}$ in Fig. 80 of Dayhoff et al. (1978[54]) by the following formula;

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

where π_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[54])).

Table 2.3: Relative substitution rate matrix, $\mathbf{R}$, of the Dayhoff model.

	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 2.4: Transition probability matrix, **M**, for the Dayhoff model.

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

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid i being replaced by the amino acid j durin substitution per 100 amino acids (1PAM) for the Dayhoff model, and average amino acid frequencies π o

Table 2.5: Transition probability matrix for the Dayhoff-F model of mtDNA-encoded p

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

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid i being replaced by the amino acid j durin substitution per 100 amino acids (1PAM) for the Dayhoff-F model, and average amino acid frequencies proteins used in the mtREV22 model (Adachi and Hasegawa 1996[10]).

Table 2.3 gives the relative substitution rate matrix $\boldsymbol{R}$ of the Dayhoff model, and Table 2.4 shows the transition probability matrix $\boldsymbol{M}$ for the model. The transition probability matrix for the Dayhoff-F model with average amino acid frequencies of the mtDNA-encoded proteins is also given in Table2.5.

2.2.2 Jones, Taylor and Thornton's (1992) Model

Table 2.6 gives the relative substitution rate matrix $\boldsymbol{R}$ of Jones, Taylor and Thornton (1992[134])(the JTT model). and Table 2.7 shows the transition probability matrix $\boldsymbol{M}$ for the model. Table 2.8 gives transition probability matrix of Jones, Taylor and Thornton's (1992[134]) model of nuclear-encoded proteins adjusted with the amino acid frequencies of the mtDNA-encoded proteins as the equilibrium frequencies (JTT-F model; Cao et al. 1994[41]).

Table 2.6: 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 2.7: Transition probability matrix for the JTT model.

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

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid i being replaced by the amino acid j durin substitution per 100 amino acids (1PAM) for the JTT model, and average amino acid frequencies π of th et al. (1992[134]).

Table 2.8: Transition probability matrix for the JTT-F model of mtDNA-encoded pro

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

Transition probability matrix $\boldsymbol{M}$ ($\times 10^5$) of the amino acid i being replaced by the amino acid j durin substitution per 100 amino acids (1PAM) for the JTT-F model, and average amino acid frequencies π proteins used in the mtREV22 model (Adachi and Hasegawa 1996[10]).

2.2.3 General Reversible Markov Model for Mitochondrial Proteins

The transition matrices of Dayhoff et al. (1978[54]) and Jones et al. (1992[134]) 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[50]; Perna and Kocher 1995[205]).

Collins et al. (1994[50]) pointed out that, in the presence of compositional bias, the transition probability matrix estimated by the parsimony method 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[126]), they demonstrated this trend for both of the data sets. It is clear that this is due to the bias of the parsimony method in inferring the ancestral state when the compositional bias exists. Perna and Kocher (1995[205]) also demonstrated the same characteristic of the parsimony method. Furthermore, since the parsimony method has no time structure (Goldman 1990[81]), it is desirable to estimate the matrix by using the ML method (Yang 1994[270]).

Naylor et al. (1995[194]) have pointed out that, since the bias for T and C at second codon position is directly correlated with the 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 perhaps the most reliable for inferring evolutionary histories of distantly related species, may actually carry less phylogenetic information than the faster 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. However, since the mitochondrial code is different to the universal code, and since most of the mtDNA-encoded proteins are membranous, the transition probability matrix of the mtDNA-encoded proteins might be quite distinct from that estimated from nuclear-encoded proteins. Thus, it seemed desirable to model the amino acid substitution of mtDNA-encoded proteins, and therefore Adachi and Hasegawa (1996[10]) 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 extension to amino acid of the general reversible Markov model of nucleotide substitution proposed by Tavaré (1986[245]), Barry and Hartigan (1987[34], 1987[33]) and Yang (1994[270]). Adachi and Hasegawa (1996[10]) estimated the $\boldsymbol{R}$ matrix by the ML method from the complete sequence data of mtDNA of 20 vertebrate species (including 3 sequences from human and hence 22 sequences in total; mtREV22 model).[1] In ProtML ver. 2.3, a revised matrix estimated with the

[1] In Fig. 1 of Adachi and Hasegawa (1996[10]), *Ornithorhynchus anatinus* (platypus) was included by mistake. The transition probability matrix presented in that paper was estimated without the platypus sequence.

two additional species (hedgehog and platypus) is used, and it is called mtREV24 model (the number 24 refers to the number of sequences used in estimating the matrix). 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.

Mitochondrial DNA Sequence Data

The matrix was estimated through ML method by using the 24 complete mtDNA sequences of vertebrates listed in Table 2.9. 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[96]). Positions with gaps and regions where the alignment was ambiguous were excluded as in Adachi and Hasegawa (1996[10]). The total number of deduced amino acid sites was 3360.

Table 2.9: List of data used in estimating the mtREV24 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[23]	X61145
Balmu	*Balaenoptera musculus*	blue whale	Árnason and Gullberg 1993[19]	X72204
Phovi	*Phoca vitulina*	harbor seal	Árnason and Johnsson 1992[24]	X63726
Halgr	*Halichoerus grypus*	grey seal	Árnason et al. 1993[22]	X72004
Equca	*Equus caballus*	horse	Xu and Árnason 1994[265]	X79547
Anderson	*Homo sapiens*	European	Anderson et al. 1981[15]	J01415*
DCM1	*Homo sapiens*	Japanese	Ozawa et al. 1991[203]	
SB17F	*Homo sapiens*	African	Horai et al. 1995[117]	D38112
Pantr	*Pan troglodytes*	chimpanzee	Horai et al. 1995[117]	D38113
Panpa	*Pan paniscus*	bonobo	Horai et al. 1995[117]	D38116
Gorgo	*Gorilla gorilla*	gorilla	Horai et al. 1995[117]	D38114
Ponpy	*Pongo pygmaeus*	orangutan	Horai et al. 1995[117]	D38115
Musmu	*Mus musculus*	mouse	Bibb et al. 1981[35]	V00711
Ratno	*Rattus norvegicus*	rat	Gadaleta et al. 1989[73]	X14848
Erieu	*Erinaceus europeus*	hedgehog	Krettek et al. 1995[157]	X88898
Didvi	*Didelphis virginiana*	opossum	Janke et al. 1994[129]	Z29573
Ornan	*Ornithorhynchus anatinus*	platypus	Janke et al. 1996[130]	X83427
Galga	*Gallus gallus*	chicken	Desjardins and Morais 1990[57]	X52392
Xenla	*Xenopus laevis*	clawed frog	Roe et al. 1985[216]	X02890
Cypca	*Cyprinus carpio*	carp	Chang et al. 1994[45]	X61010
Crola	*Crossostoma lacustre*	loach	Tzeng et al. 1992[252]	M91245
Oncmy	*Oncorhynchus mykiss*	trout	Zardoya et al. 1995[275])	L29771
Petma	*Petromyzon marinus*	sea lamprey	Lee and Kocher 1995[168]	U11880

*: revised according to Horai et al. (1995[117]).

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 branch, $t_1, \ldots, t_m$

(m: number of branches in the tree) by the same procedure as that presented in subsection2.1.2;

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

where l is the 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 ϵ, we stop the procedure.

Fig. 2.1 shows the unrooted tree (Cao et al. 1994[41]; Janke et al. 1994[129], 1996[130]; Horai et al. 1995[117]), among species from which complete mtDNA sequences are available, assumed in the estimation of the transition probability matrix. The placement of lamprey in this figure is not from the ML tree, but from the 2nd highest likelihood tree (((Birds, Mammals), (Xenopus, Fishes), Lamprey) as shown in Fig. 2.2 is the ML tree). Since the difference of log-likelihood of this tree from that of the ML tree is minor (12.8 $\pm$ 16.2 where $\pm$ is 1SE estimated by the formula in Kishino and Hasegawa 1989[147]), we used this biologically more reasonable tree. Since the branching orders among Carnivora, Perissodactyla and the Cetacea/Artiodactyla clade, and among hedgehog, Rodentia and the other placentals cannot be resolved by the mtDNA data, they were left as trifurcations.[2] The estimated transition probability matrix is not sensitive to the choice of the tree (Yang 1994[270]; Adachi 1995[1]; Adachi and Hasegawa 1996[10]). The log-likelihood of this tree for the mtREV24 model is -52278.9, while that for the JTT-F model is -53205.7, showing much improved fitting of the mtREV24 model to the mtDNA-encoded protein data.

The tree in Fig. 2.1 might be unexpected with respect to the relationship among monotremes, marsupials and placentals. The traditional taxonomy conceives that, because of the primitive characters of monotremes such as egg-laying, monotremes represent the earliest offshoot among the extant mammalian lineages. By sequencing the complete mitochondrial genome of the platypus and by analyzing protein-encoding genes, however, Janke et al. (1996[130]) suggested the marsupial/monotreme clade excluding placentals. Our analysis also supports their hypothesis (Table 2.10; Adachi and Hasegawa 1995[6]). While another unexpected clade of placental/monotreme cannot be excluded, the traditional tree with the placental/marsupial clade is very unlikely by any of the models (Table 2.10). Although Janke et al.'s hypothesis of the marsupial/monotreme clade might seem to contradict morphological evidence, some morphologists have already suggested it (Gregory 1947[87]; Kühne 1973[137], 1975[138]), and the existing molecular data does not support the traditional tree (Retief et al. 1994[213]; Gemmell and Westerman 1994[77]). Therefore, we will adopt Janke et al.'s hypothesis in estimating the transition probability matrix for the mtREV24 model. It must be noted again, however, that the estimated transition probability matrix is apparently not sensitive to the choice of the tree.

[2]The recent data from guinea-pig (*Cavia procellus*), rabbit (*Oryctolagus cuniculus*) (D'Erchia et al. 1996[56]) and cat (database accession number: U20753) help to resolve these trifurcations (unpublished); existence of the Perissodactyla/Carnivora clade and the sister-group relationship of the hedgehog with a clade formed by Rodentia and the other placentals.

Table 2.10: ProtML analyses of mtDNA-encoded proteins on the relationship among monotremes, marsupials and placentals using several alternative models for amino acid substitution.

Model	Placental/Marsupial		Marsupial/Monotreme		Placental/Monotreme	
Poisson	-38.2 ± 22.5	(.0128)	$< -61364.3 >$	(.5815)	-5.9 ± 25.9	(.4057)
Proportional	-26.7 ± 19.6	(.0318)	$< -58112.7 >$	(.5690)	-5.1 ± 22.0	(.3992)
Dayhoff	-35.4 ± 17.9	(.0056)	$< -56401.0 >$	(.6662)	-9.5 ± 21.1	(.3282)
Dayhoff-F	-31.5 ± 16.6	(.0138)	$< -53690.3 >$	(.8401)	-19.0 ± 18.3	(.1461)
JTT	-31.2 ± 17.1	(.0081)	$< -55038.5 >$	(.6534)	-8.2 ± 20.0	(.3385)
JTT-F	-28.0 ± 15.9	(.0169)	$< -53205.7 >$	(.7686)	-13.8 ± 17.8	(.2145)
mtREV24	-26.7 ± 14.4	(.0117)	$< -52278.9 >$	(.7763)	-12.8 ± 16.3	(.2120)

The log-likelihood of the ML tree is given in $< \cdots >$, and the differences in log-likelihood of alternative trees from that of the ML tree are shown with their SE (following $\pm$) which were estimated by Kishino and Hasegawa's (1989[147]) formula. The bootstrap probabilities given in parentheses were estimated by the RELL method (Kishino et al. 1990[148]; Hasegawa and Kishino 1994[97]) with 10^4 replications.

Table 2.11 is the relative substitution rate matrix $\boldsymbol{R}$ of the mtREV24 model, and Table 2.12 gives the estimated transition probability matrix for the mtREV24 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 (Adachi and Hasegawa 1996[10]). This is probably 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 $\leftrightarrow$ Lys substitutions are much less frequent in vertebrate mitochondria. This observation demonstrates the importance of the mutation-driven neutral evolution (Kimura 1968[143], 1983[146]) under the constraint of the code.

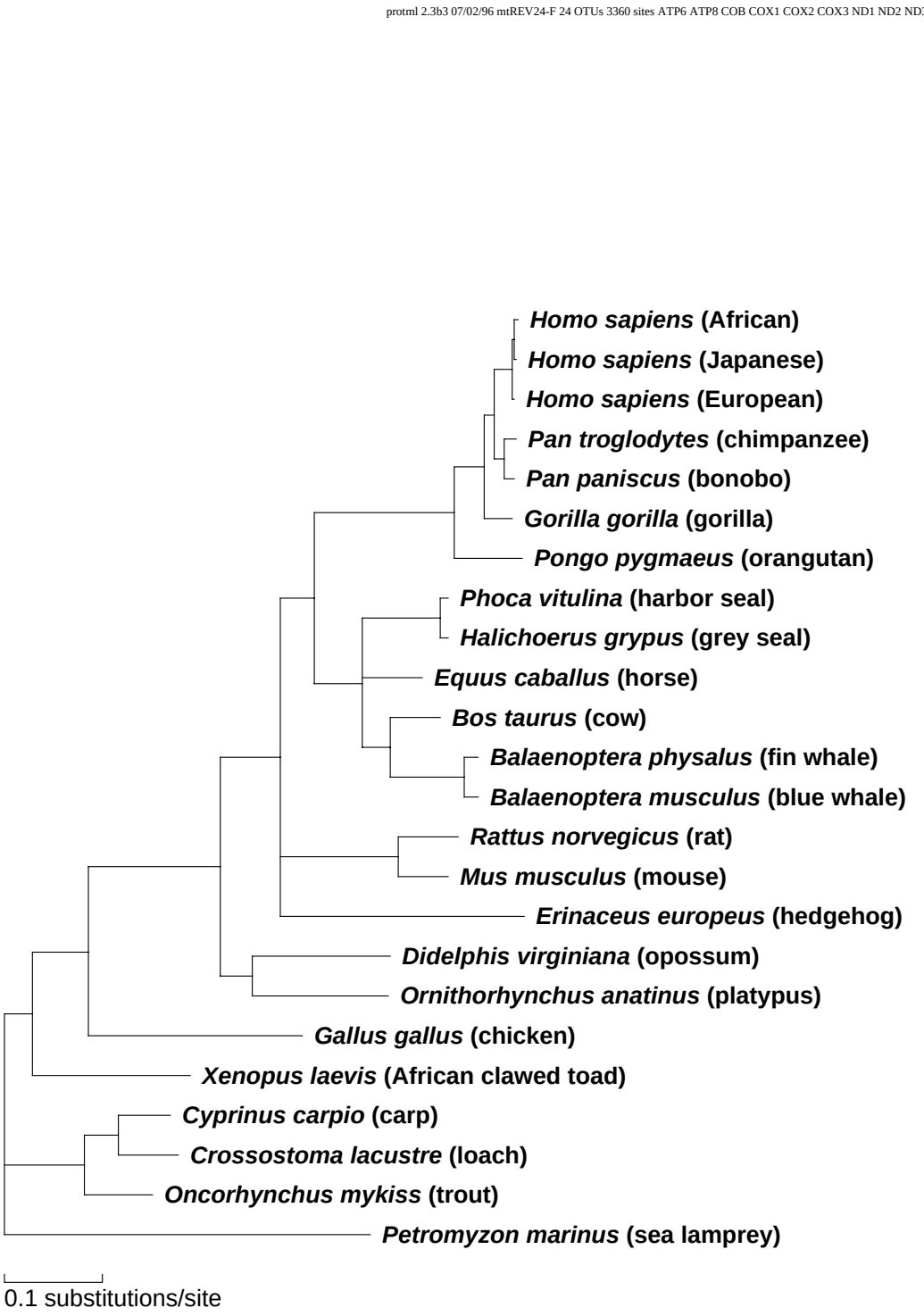


Figure 2.1: The tree used in estimating the transition probability matrix of the mtREV24 model.

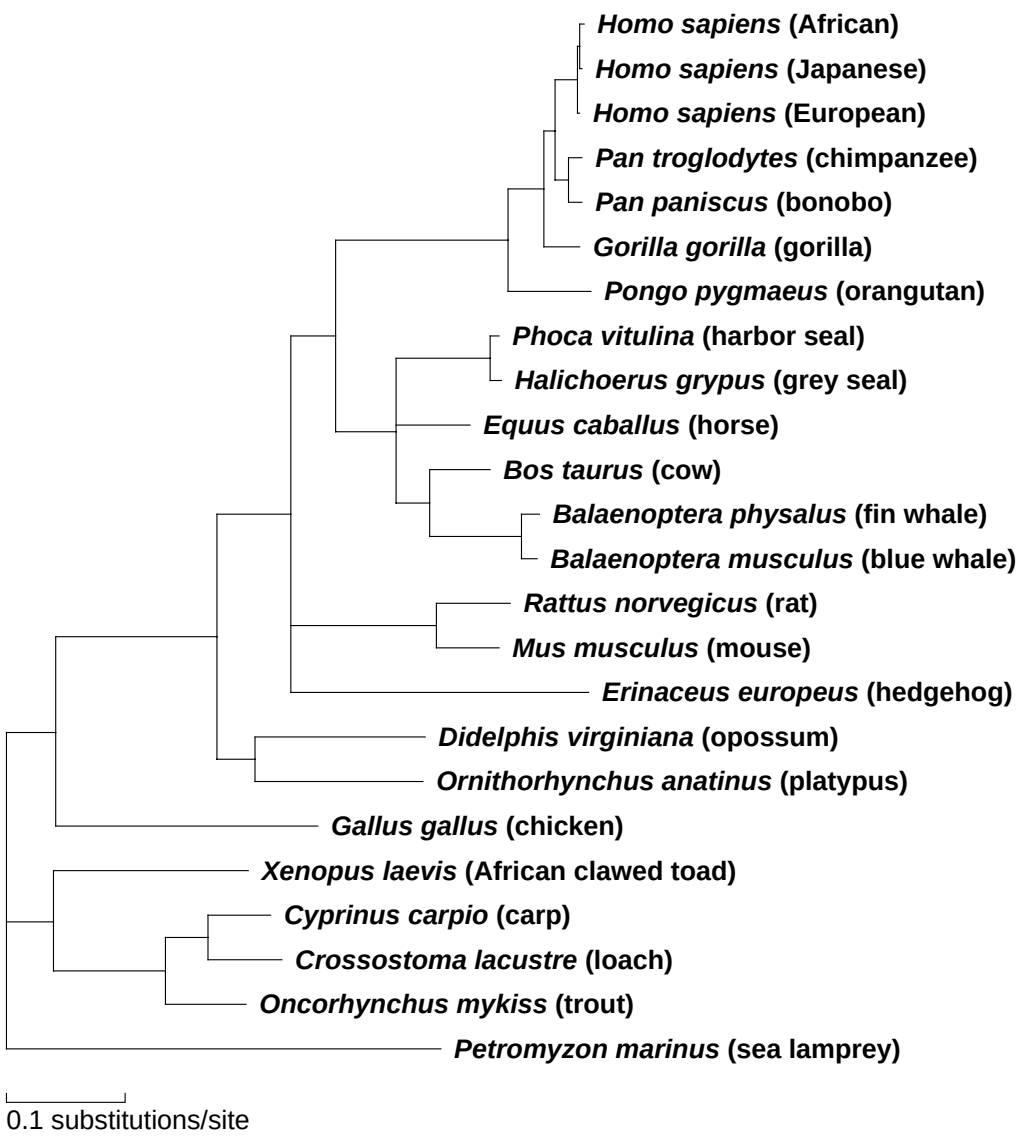


Figure 2.2: The ML tree of mtDNA-encoded proteins.

Table 2.11: Relative substitution rate matrix of mtREV24 model.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala		122	142	93	315	10	51	635	73	508	134	44	747	34	286	2041	2530	10	34	1027
Arg	122		70	10	544	1163	10	121	870	10	82	744	10	25	124	32	11	116	10	40
Asn	142	70		4181	310	913	332	281	2611	143	80	3204	344	80	386	2602	1255	56	1007	10
Asp	93	10	4181		10	291	3071	299	600	23	10	12	10	26	71	363	147	105	112	10
Cys	315	544	310	10		396	10	162	745	330	135	10	33	373	165	1458	947	177	1341	10
Gln	10	1163	913	291	396		1650	36	3065	44	209	2450	249	101	723	285	500	10	204	100
Glu	51	10	332	3071	10	1650		149	259	17	10	1652	10	14	68	288	78	10	69	111
Gly	635	121	281	299	162	36	149		10	31	13	120	10	10	10	663	59	57	17	13
His	73	870	2611	600	745	3065	259	10		65	60	672	63	253	321	408	236	37	3527	10
Ile	508	10	143	23	330	44	17	31	65		1732	103	2726	446	109	251	1939	10	132	6437
Leu	134	82	80	10	135	209	10	13	60	1732		78	2829	1137	211	387	665	171	232	482
Lys	44	744	3204	12	10	2450	1652	120	672	103	78		481	34	264	557	718	126	269	10
Met	747	10	344	10	33	249	10	10	63	2726	2829	481		478	99	585	2780	114	210	2040
Phe	34	25	80	26	373	101	14	10	253	446	1137	34	478		91	338	178	41	2450	33
Pro	286	124	386	71	165	723	68	10	321	109	211	264	99	91		894	675	22	85	43
Ser	2041	32	2602	363	1458	285	288	663	408	251	387	557	585	338	894		3143	203	342	10
Thr	2530	11	1255	147	947	500	78	59	236	1939	665	718	2780	178	675	3143		53	204	1077
Trp	10	116	56	105	177	10	10	57	37	10	171	126	114	41	22	203	53		138	28
Tyr	34	10	1007	112	1341	204	69	17	3527	132	232	269	210	2450	85	342	204	138		10
Val	1027	40	10	10	10	100	111	13	10	6437	482	10	2040	33	43	10	1077	28	10	

Table 2.12: Transition probability matrix for the mtREV24 model.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile
Ala	9904144	379	904	289	309	41	202	5812	335	7301
Arg	1435	9979648	444	31	533	4750	39	1109	3978	144
Asn	1668	216	9887772	12977	304	3731	1301	2566	11944	2050
Asp	1094	31	26638	9944387	10	1188	12042	2733	2744	328
Cys	3710	1688	1976	31	9933516	1617	39	1479	3406	4747
Gln	118	3610	5820	903	388	9930974	6471	325	14021	631
Glu	605	31	2114	9533	10	6740	9965182	1362	1183	250
Gly	7473	376	1787	927	158	145	584	9977898	46	452
His	860	2699	16637	1862	730	12519	1014	91	9925195	928
Ile	5973	31	909	71	324	179	68	288	295	9838322
Leu	1576	254	508	31	132	853	39	116	277	24900
Lys	518	2310	20411	38	10	10008	6477	1094	3074	1481
Met	8783	31	2193	31	32	1018	39	91	288	39192
Phe	394	77	510	81	365	411	55	91	1159	6406
Pro	3362	386	2458	219	161	2951	265	91	1468	1561
Ser	24011	99	16578	1128	1429	1163	1129	6063	1865	3609
Thr	29760	34	7996	458	928	2041	306	538	1078	27877
Trp	118	359	358	324	173	41	39	526	171	144
Tyr	401	31	6417	347	1314	834	271	154	16134	1892
Val	12076	125	64	31	10	408	436	122	46	92532
	0.072	0.019	0.039	0.019	0.006	0.025	0.024	0.056	0.028	0.088

	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	3678	165	6587	334	2521	24011	35547	47	184	7212
Arg	2250	2796	88	246	1097	374	154	547	54	282
Asn	2189	12037	3037	797	3404	30606	17633	266	5430	70
Asp	274	46	88	261	624	4273	2071	495	602	70
Cys	3705	38	287	3714	1452	17151	13308	838	7229	70
Gln	5735	9207	2199	1002	6374	3350	7019	47	1101	702
Glu	274	6207	88	140	596	3387	1096	47	372	782
Gly	349	449	88	100	88	7796	826	272	91	94
His	1660	2525	556	2526	2831	4795	3312	177	19015	70
Ile	47537	387	24050	4441	958	2953	27243	47	710	45214
Leu	9913513	294	24957	11332	1862	4557	9346	809	1253	3389
Lys	2149	9926774	4243	338	2326	6549	10080	598	1452	70
Met	77645	1807	9801271	4763	875	6882	39055	541	1134	14328
Phe	31210	127	4217	9933969	804	3980	2503	196	13210	235
Pro	5793	991	875	908	9957642	10518	9481	105	460	304
Ser	10633	2092	5161	3372	7888	9866745	44160	962	1842	70
Thr	18258	2696	24523	1775	5953	36971	9829898	249	1099	7562
Trp	4685	475	1008	411	196	2388	738	9986903	745	199
Tyr	6377	1012	1855	24419	753	4019	2864	654	9930182	70
Val	13242	38	17994	333	382	118	15125	134	54	9846733
	0.167	0.023	0.054	0.061	0.054	0.072	0.086	0.029	0.033	0.043

Transition probability matrix $\boldsymbol{M}$ ($\times 10^7$) 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 mtREV24 model, and average amino acid frequencies π of the mtDNA-encoded proteins.

2.2.4 Discussion

Previously, the JTT model for nuclear-encoded proteins was used even in the ML analyses of mtDNA-encoded proteins (Cao et al. 1994[41]; Adachi and Hasegawa 1995[7]), mainly because no appropriate model for mtDNA-encoded proteins was available. The conclusions of these phylogenetic analyses hold when the mtREV model 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[72]; Hasegawa and Fujiwara 1993[92]). 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 reasonable stochastic model (such as shown in Table 2.12) is obtained, the ML method would be the preferred method of inferring trees from mtDNA-encoded protein sequences (Felsenstein 1981[64]; Kishino et al. 1990[148]; Edwards 1995[59]). 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 can adjust the equilibrium frequencies of the model to the actual frequencies of the protein under study (F-option). This should also ensure some robustness.

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[223]; Goldman and Yang 1994[82]; Muse and Gaut 1994[190]) 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; Adachi and Hasegawa 1996[11]), and so 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 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[251]). The parsimony method is used widely (Stewart 1993[234]), 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[229]; Swofford et al. 1996[240]).

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[196]) was the first to use the maximum likelihood method to estimate evolutionary trees from DNA sequences based on a stochastic model, and Felsenstein (1981[64]) developed a practical method, from which the maximum likelihood methods used widely at present stem (Kishino et al. 1990[148]; Adachi and Hasegawa 1992[4]; Yang 1993[269]; Felsenstein 1993[69]; Olsen et al. 1994[200]; Swofford et al. 1996[240]).

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 are 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 has only two immediate descendant lineages, but multifurcating if it has 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.

3.2 Algorithm for ML Inference of Molecular Phylogeny

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

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

Let us write the whole data set as matrix $\boldsymbol{X}$, 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 of, and identically with, all others. We further assume that, after speciation, the two separated lineages evolve independently, and that the same stochastic process of substitution applies in all lineages, although the rate parameter of the process might differ among different lineages (i.e., branch lengths can be different).

3.2.1 Computing the Likelihood of the Data Given 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[64]).

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

$$L = \text{Prob}(\boldsymbol{X}|T, \boldsymbol{\theta}) \tag{3.1}$$

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

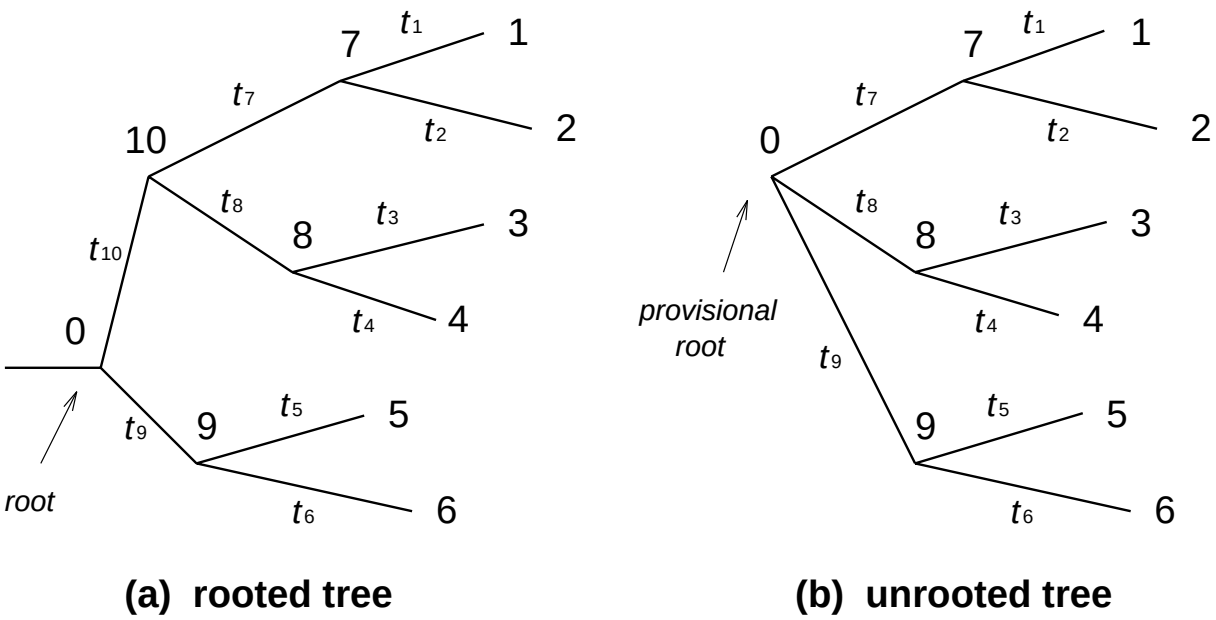


Figure 3.1: The rooted tree and unrooted tree used in the following discussion.

It is possible to write a general expression for the likelihood of a tree, but it will be more illustrative 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. Further, let us assume the data consist of just 1 site. 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 data on that tree would be the product of the probabilities of change in each branch, times the prior probability π_{x_0} of state x_0, so that it would be

$$\begin{aligned} f(\boldsymbol{x}) = \quad & \pi_{x_0} \quad P_{x_0 x_{10}}(t_{10}) P_{x_{10} x_7}(t_7) P_{x_7 x_1}(t_1) P_{x_7 x_2}(t_2) P_{x_{10} x_8}(t_8) P_{x_8 x_3}(t_3) P_{x_8 x_4}(t_4) \\ & \times \quad P_{x_0 x_9}(t_9) P_{x_9 x_5}(t_5) P_{x_9 x_6}(t_6) \end{aligned} \tag{3.2}$$

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

The vector $\boldsymbol{\pi}$ is the prior probability of finding each state at node 0 on the tree. Since we are assuming an evolutionarily steady state, i.e., base composition (amino acid frequency) is not expected to alter, then $\boldsymbol{\pi}$ reflects the overall base composition (amino acid frequency) in the group under study. One of the convenient properties of some Markov process models (like that given in Chapter 2) of base (amino acid) substitution is known as "reversibility" (Felsenstein 1981[64]). This means that the result of base (amino acid) substitution will look the same irrespective of whether going forward or backward in time. Reversibility requires that for all i, j and t

$$\pi_i P_{ij}(t) = \pi_j P_{ji}(t) \tag{3.3}$$

which is easily proven using Eq. 2.10.

Since reversibility and the "pulley principle" (Felsenstein 1981[64]), 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 data given tree topology $T = \big((1,2),(3,4),(5,6)\big)$ in Fig. 3.1b would be

$$\begin{aligned} f(\boldsymbol{x}) = \quad & \pi_{x_0} \quad P_{x_0x_7}(t_7)P_{x_7x_1}(t_1)P_{x_7x_2}(t_2) \\ & \times \quad P_{x_0x_8}(t_8)P_{x_8x_3}(t_3)P_{x_8x_4}(t_4) \\ & \times \quad P_{x_0x_9}(t_9)P_{x_9x_5}(t_5)P_{x_9x_6}(t_6) \end{aligned} \tag{3.4}$$

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 internal nodes on the tree in Fig. 3.2. The probability of realizing $\boldsymbol{x} = (x_1, x_2, \cdots, x_6)^T$ at a site in species $1, 2, \ldots, 6$ respectively, is given by

$$\begin{aligned} f(\boldsymbol{x}) = \sum_{i=1}^{m} \quad & \pi_i \quad \left(\sum_{j=1}^{m} P_{ij}(t_7)P_{jx_1}(t_1)P_{jx_2}(t_2)\right) \\ & \times \quad \left(\sum_{k=1}^{m} P_{ik}(t_8)P_{kx_3}(t_3)P_{kx_4}(t_4)\right) \\ & \times \quad \left(\sum_{l=1}^{m} P_{il}(t_9)P_{lx_5}(t_5)P_{lx_6}(t_6)\right) \end{aligned} \tag{3.5}$$

where m is 4 for bases and 20 for amino acids.

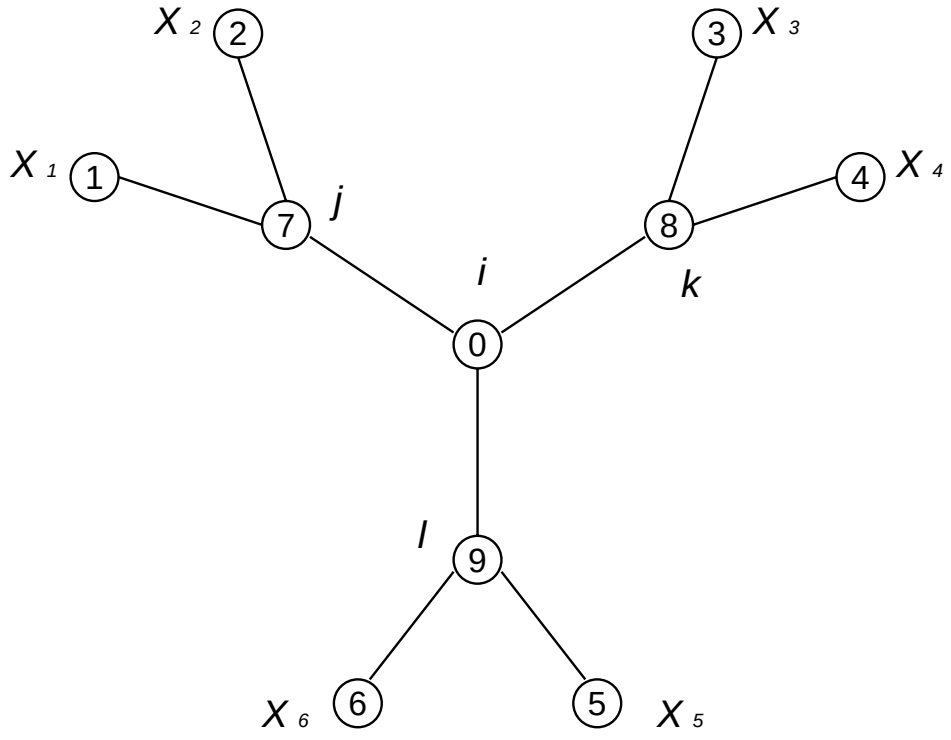


Figure 3.2: The unrooted tree (from Fig. 3.1) used in the discussion of computing the likelihood.

The log-likelihood of the data given this tree is

$$l(\boldsymbol{\theta}|\boldsymbol{X}, T) = \sum_{h=1}^{n} \log f(\boldsymbol{X}_h|T, \boldsymbol{\theta}) \tag{3.6}$$

where

$$\boldsymbol{\theta} = (t_1, t_2, \cdots, t_9)^T . \tag{3.7}$$

The log-likelihood of the data is rewritten as

$$\begin{aligned} l(\boldsymbol{\theta}|\boldsymbol{X}, T) = \sum_{h=1}^{n} \log \Bigg\{ & \sum_{i=1}^{m} \pi_i \left(\sum_{j=1}^{m} P_{ij}(t_7) P_{jX_{1h}}(t_1) P_{jX_{2h}}(t_2) \right) \\ & \times \left(\sum_{k=1}^{m} P_{ik}(t_8) P_{kX_{3h}}(t_3) P_{kX_{4h}}(t_4) \right) \\ & \times \left(\sum_{l=1}^{m} P_{il}(t_9) P_{lX_{5h}}(t_5) P_{lX_{6h}}(t_6) \right) \Bigg\} . \end{aligned} \tag{3.8}$$

(Note: while likelihood refers to the probability of the data given the tree, we will sometimes be more slack and call this quantity the likelihhod of the tree).

3.2.2 Evaluating Likelihood along a 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 in MOLPHY, the estimation of each branch length is iterated separately, by using the Newton-Raphson method (Kishino et al. 1990[148]) 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 is described in Felsenstein (1973[61], 1981[64]).

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 the outer current subnode on the tree, given that the current subnode is known to have state i for a site h under consideration. A partial likelihood is a set of conditional likelihoods for a 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.

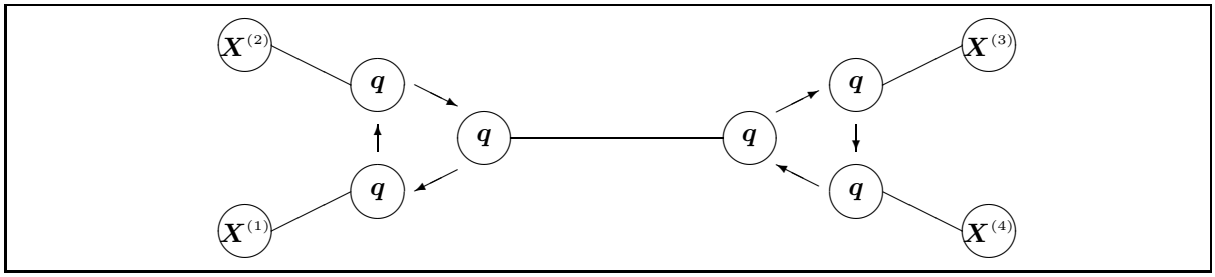


Figure 3.3: Data structure of a tree.

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 the 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 no need for the full matrix $\boldsymbol{q}$ for an 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 the 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:

$$q_i = \begin{cases} \sum_{j=1}^{h} P_{ij}(t) Q_j, & \text{if internal branch} \\ P_{ix}(t), & \text{if external branch} \end{cases} \tag{3.9}$$

where Q_j is product of under partial likelihoods.

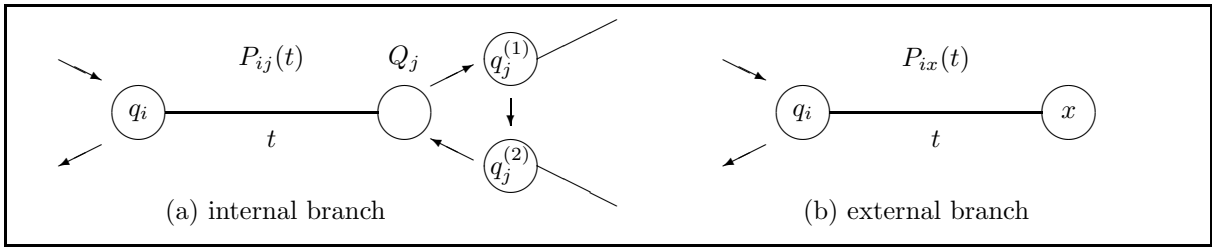


Figure 3.4: Partial likelihood.

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

$$Q_i = \prod_{j=1}^{b} q_i^{(j)} \tag{3.10}$$

where b is a number of branchings.

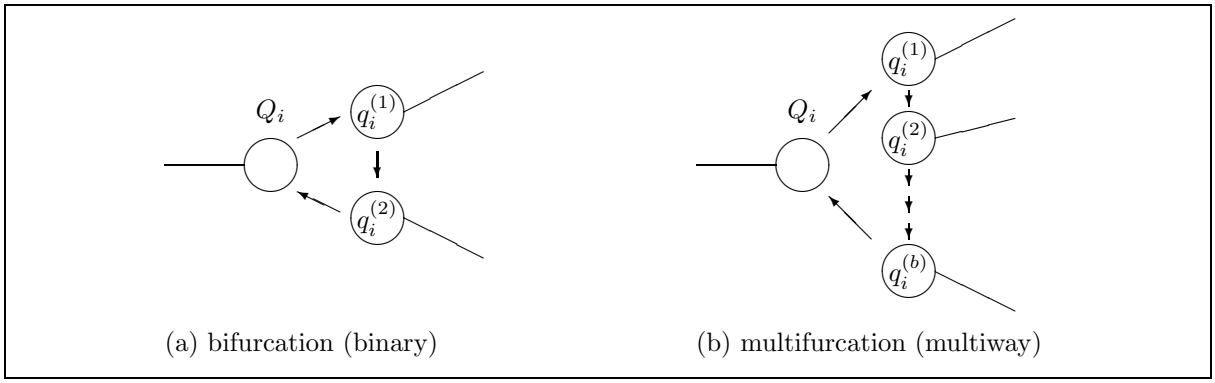


Figure 3.5: Product of partial likelihood.

This process proceeds down the tree towards the root. In an unrooted tree (i.e., reversible model), the root may be placed anywhere. The values of $\boldsymbol{q}$ at the root are then combined in a weighted average

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} \pi_i Q_i^{(\text{ans})} \sum_{j=1}^{m} P_{ij}(t) Q_j^{(\text{des})} = \sum_{i=1}^{m} \pi_i \prod_{j=0}^{b} q_i^{(j)} \tag{3.11}$$

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

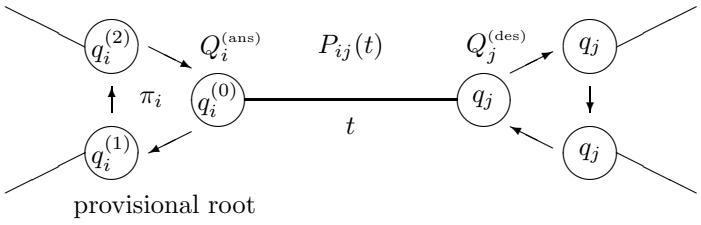


Figure 3.6: Computing the likelihood of a tree.

3.2.3 Maximum Likelihood Estimation of Branch Length

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

$$\text{maximize} \quad \log L(\boldsymbol{\theta}|\boldsymbol{X}, T) \qquad \text{for } \boldsymbol{\theta} \in \boldsymbol{\Theta} \tag{3.12}$$

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

$$\left[\frac{\partial \log L}{\partial \theta_j}\right]_{\widehat{\boldsymbol{\theta}}}^{T} = 0, \tag{3.13}$$

$$\left[\frac{\partial^2 \log L}{\partial \theta_j \partial \theta_h}\right]_{\widehat{\boldsymbol{\theta}}} \quad \text{is} \quad \text{negative definite} \tag{3.14}$$

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

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

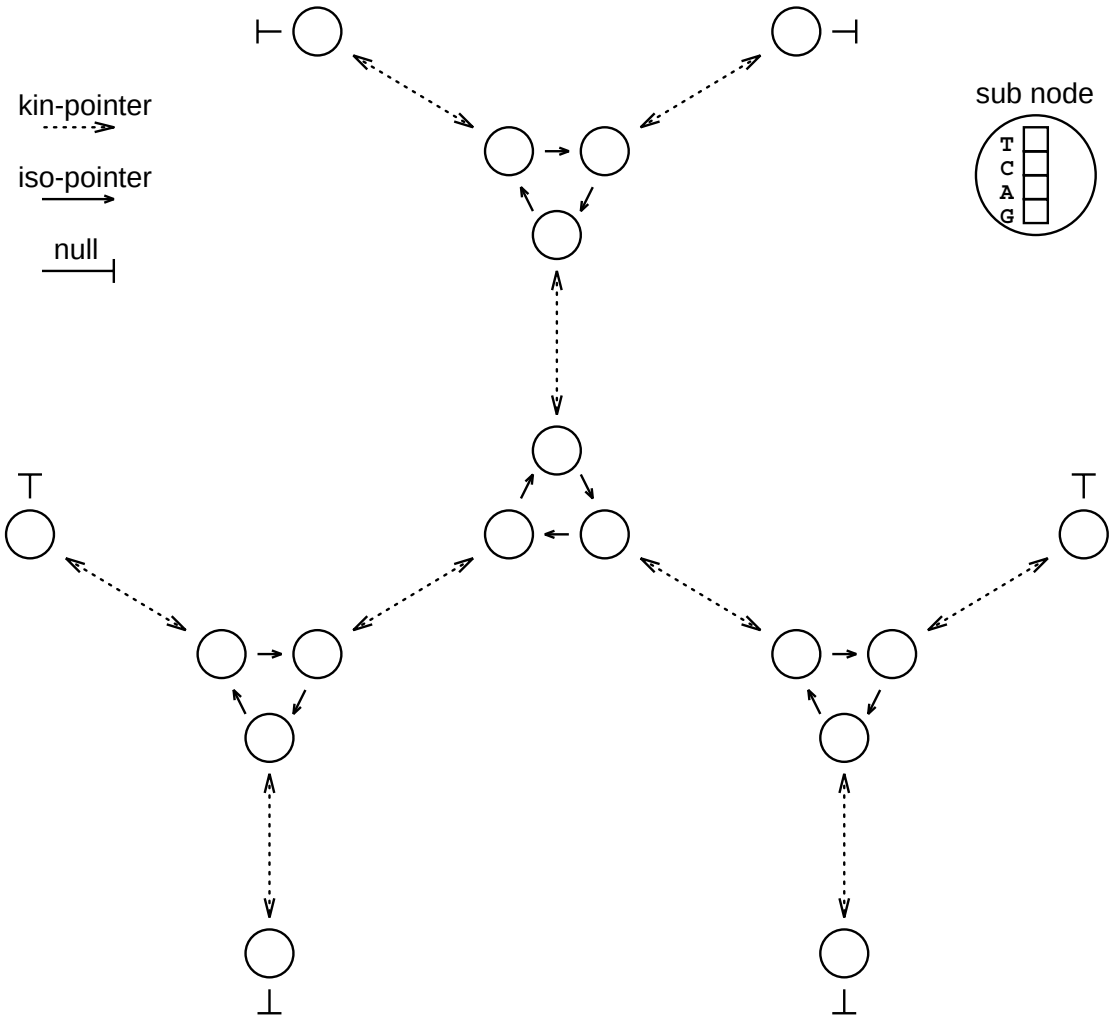


Figure 3.7: Data structure of a tree topology.

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[61], 1981[64]).

We now consider how 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

$$f(x) = 0 \tag{3.15}$$

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 evaluate both the function $f(x)$, its first derivative $f'(x)$, and its second derivative $f''(x)$, at an arbitrary point 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

$$x_{i+1} = x_i - f(x_i) \Big/ \left(\frac{d}{dx} f(x_i)\right). \tag{3.16}$$

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

$$\text{maximize} \quad l(t). \tag{3.17}$$

The problem is to find the maximum point of the 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 an arbitrary point t. The formula is

$$t_{i+1} = t_i - \left(\frac{d}{dt} l(t_i)\right) \Big/ \left(\frac{d^2}{dt^2} l(t_i)\right). \tag{3.18}$$

We can obtain the maximum likelihood estimate of t through the Newton-Raphson method, in which calculations of l, ∇l and $\nabla\nabla^T l$ are necessary (Kishino et al. 1990[148]) and we have

$$P_{ij}(t) = \sum_{k=1}^{m} \left(U_{ik} U_{kj}^{-1} \exp(t\lambda_k)\right) \tag{3.19}$$

$$\frac{d}{dt} P_{ij}(t) = \sum_{k=1}^{m} \left(U_{ik} U_{kj}^{-1} \lambda_k \exp(t\lambda_k)\right) \tag{3.20}$$

$$\frac{d^2}{dt^2} P_{ij}(t) = \sum_{k=1}^{m} \left(U_{ik} U_{kj}^{-1} \lambda_k^2 \exp(t\lambda_k)\right) \tag{3.21}$$

where U_{ij} is an entry in the eigenvectors of P_{ij}.

Internal Branch Length

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

$$l(t_k) = \sum_{h=1}^{n} \log\left(\sum_{i=1}^{m} \pi_i Q_{hi}^{(\text{ans})} \sum_{j=1}^{m} P_{ij}(t_k) Q_{hj}^{(\text{des})}\right). \tag{3.22}$$

From Eqs. 3.20 and 3.21 we can compute the first derivative and the second derivative of the log-likelihood function with respect to the k-th internal branch length

$$\frac{d}{dt} l(t_k) = \sum_{h=1}^{n} \log\left(\sum_{i=1}^{m} \pi_i Q_{hi}^{(\text{ans})} \sum_{j=1}^{m} \frac{d}{dt} P_{ij}(t_k) Q_{hj}^{(\text{des})}\right) \tag{3.23}$$

$$\frac{d^2}{dt^2}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\mathrm{ans})}\sum_{j=1}^{m}\frac{d^2}{dt^2}P_{ij}(t_k)Q_{hj}^{(\mathrm{des})}\right). \tag{3.24}$$

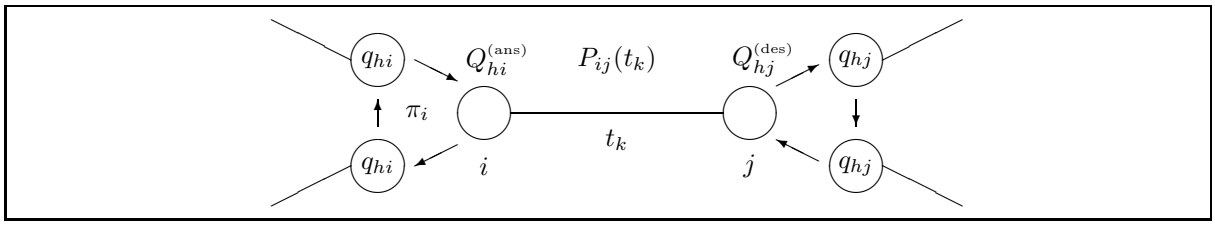


Figure 3.8: MLE of an internal branch length by Newton-Raphson method.

External Branch Length

Similarly, the log-likelihood of the tree at the k-th external branch is rewitten as

$$l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\mathrm{ans})}P_{iX_{kh}}(t_k)\right). \tag{3.25}$$

From Eqs. 3.20 and 3.21 we can compute the first derivative and the second derivative of the log-likelihood function with respect to the k-th external branch length

$$\frac{d}{dt}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\mathrm{ans})}\frac{d}{dt}P_{iX_{kh}}(t_k)\right) \tag{3.26}$$

$$\frac{d^2}{dt^2}l(t_k) = \sum_{h=1}^{n}\log\left(\sum_{i=1}^{m}\pi_i Q_{hi}^{(\mathrm{ans})}\frac{d^2}{dt^2}P_{iX_{kh}}(t_k)\right). \tag{3.27}$$

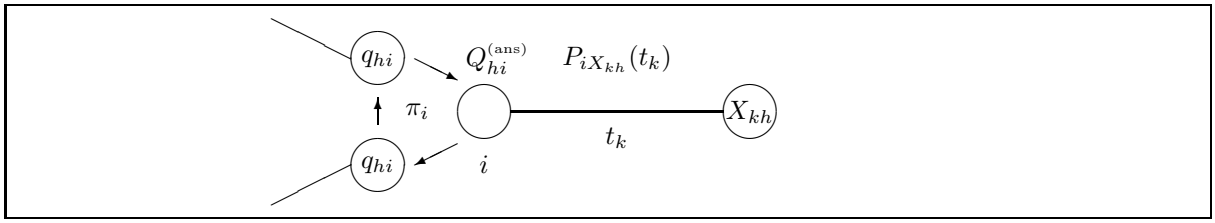


Figure 3.9: MLE of an external branch length by Newton-Raphson method.

Using a new method that will be described in Section 3.3, 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.

3.2.4 Estimation of Distances by the ML Method

Initial Distance Matrix

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

$$D_{ij}^{(\mathrm{init})} = -\frac{m-1}{m}\log\left(1 - \frac{mD_{ij}^{(\mathrm{diff})}}{n(m-1)}\right) \tag{3.28}$$

where n is the length of the sequence, m is the number of states ($m = 4$ for bases and $m = 20$ for amino acids), and $D_{ij}^{(\mathrm{diff})}$ is the number of differences between i-th and j-th sequences (e.g., see Kishino et al. 1990[148]; Felsenstein 1993[69]; Swofford et al. 1996[240]). This estimate is used as an initial distance provided for the ML analysis.

Distance Matrix Estimated by the ML Method

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, i.e., Eq. 3.20 and 3.21. This optimization can be done by a direct search.

The initial value of D_{ij}, denoted by $D_{ij}^{(\mathrm{init})}$, is calculated assuming the Poisson process. Then reestimate D_{ij} by the Newton-Raphson method to maximize

$$l(D_{ij}|\boldsymbol{X}^{(i)}, \boldsymbol{X}^{(j)}) = \sum_{h=1}^{n} \log\left(P_{X_{ih}X_{jh}}(D_{ij})\right) \tag{3.29}$$

where D_{ij} is the number of substitutions per site between i-th and j-th sequences (see also Felsenstein 1993[69], PHYLIP 3.5 documentation).

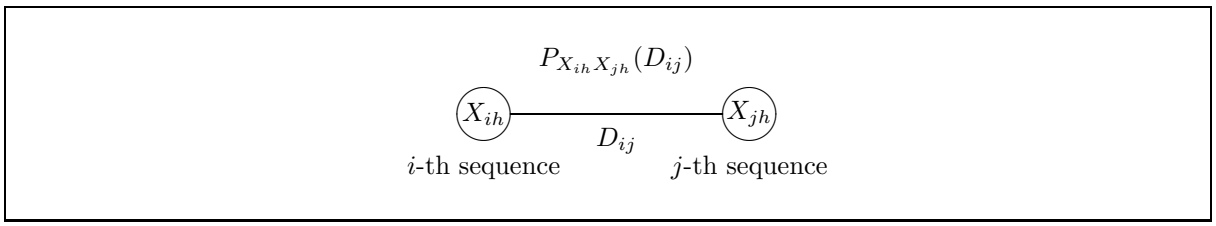


Figure 3.10: MLE of a distance by Newton-Raphson method.

3.2.5 Estimation of Initial Branch Lengths

Initial Branch Lengths Estimated by the Least Squares Method

We have the observed corrected distances in an $(n \times 1)$ vector $\boldsymbol{D}$ where $n = N(N-1)/2$ (where N is number of OTUs) and an $(n \times k)$ incidence 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. Least squares assumes $\boldsymbol{D}$ is generated as

$$\boldsymbol{D} = \boldsymbol{A}\boldsymbol{t} + \boldsymbol{\epsilon} \tag{3.30}$$

where $\boldsymbol{t}$ is a $(k \times 1)$ vector of unknown coefficients, and $\boldsymbol{\epsilon}$ is an $(n \times 1)$ vector of independent normal variates with zero mean and unknown variance σ^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 $\widehat{\boldsymbol{t}}$ by minimizing

$$\min\{\boldsymbol{S}(\boldsymbol{t})\} = \min\{(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{t})^T(\boldsymbol{D} - \boldsymbol{A}\boldsymbol{t})\} \tag{3.31}$$

(Chakraborty 1977[44]).

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

$$\widehat{\boldsymbol{t}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{D} \tag{3.32}$$

with (asymptotic) covariance matrix

$$V\widehat{\boldsymbol{t}} = \sigma^2(\boldsymbol{A}^T\boldsymbol{A})^{-1}. \tag{3.33}$$

where

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

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

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

$$\widehat{\boldsymbol{t}} = \begin{bmatrix} D_{12}/2 + (D_{13} + D_{14} + D_{15} + D_{16})/8 - (D_{23} + D_{24} + D_{25} + D_{26})/8 \\ D_{12}/2 + (D_{23} + D_{24} + D_{25} + D_{26})/8 - (D_{13} + D_{14} + D_{15} + D_{16})/8 \\ D_{34}/2 + (D_{13} + D_{23} + D_{35} + D_{36})/8 - (D_{14} + D_{24} + D_{45} + D_{46})/8 \\ D_{34}/2 + (D_{14} + D_{24} + D_{45} + D_{46})/8 - (D_{13} + D_{23} + D_{35} + D_{36})/8 \\ D_{56}/2 + (D_{15} + D_{25} + D_{35} + D_{45})/8 - (D_{16} + D_{26} + D_{36} + D_{46})/8 \\ D_{56}/2 + (D_{16} + D_{26} + D_{36} + D_{46})/8 - (D_{15} + D_{25} + D_{35} + D_{45})/8 \\ (D_{13} + D_{14} + D_{15} + D_{16} + D_{23} + D_{24} + D_{25} + D_{26})/8 - D_{12}/2 - (D_{35} + D_{36} + D_{45} + D_{46})/8 \\ (D_{13} + D_{14} + D_{23} + D_{24} + D_{35} + D_{36} + D_{45} + D_{46})/8 - D_{34}/2 - (D_{15} + D_{16} + D_{25} + D_{26})/8 \\ (D_{15} + D_{16} + D_{25} + D_{26} + D_{35} + D_{36} + D_{45} + D_{46})/8 - D_{56}/2 - (D_{13} + D_{14} + D_{23} + D_{24})/8 \end{bmatrix} \tag{3.34}$$

3.3 Fast Computation of ML for Inferring Evolutionary Trees

The fast computation algorithm used in MOLPHY is shown in Fig. 3.11.

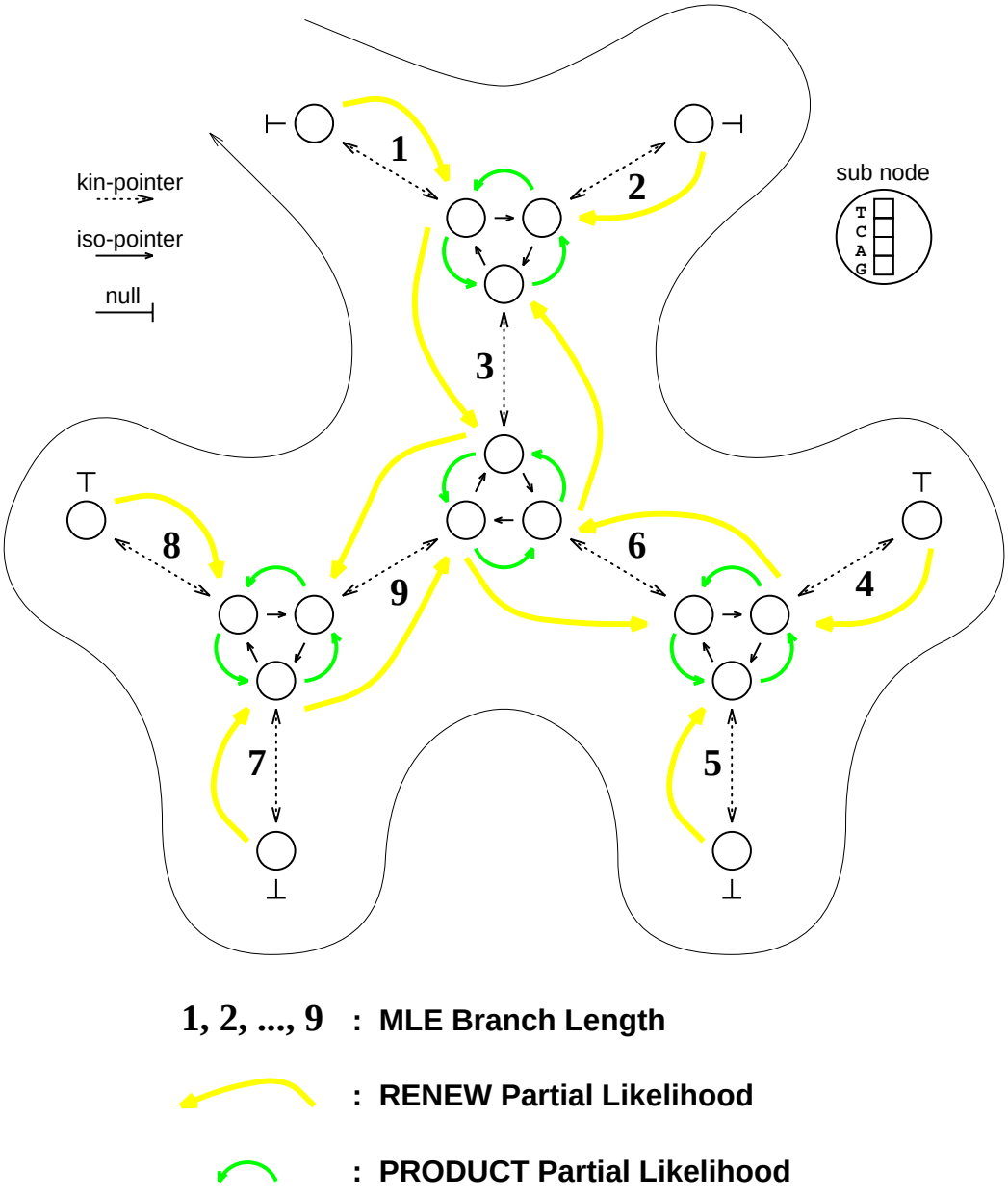


Figure 3.11: Fast computation algorithm.

```
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 3.1: 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

3.4 Topology Search Strategy for ML Phylogeny

3.4.1 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 recalculate likelihood through iteration of a loop multiplied by the times of the number of nodes in revising the estimate of each branch length, but need only revise the partial likelihoods of the two nodes of each branch.

We extend this data structure so that a multifurcating tree can also be represented. Since branches are connected dynamically by pointers, the data structure can easily be revised when a different tree topology is adopted, and 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.

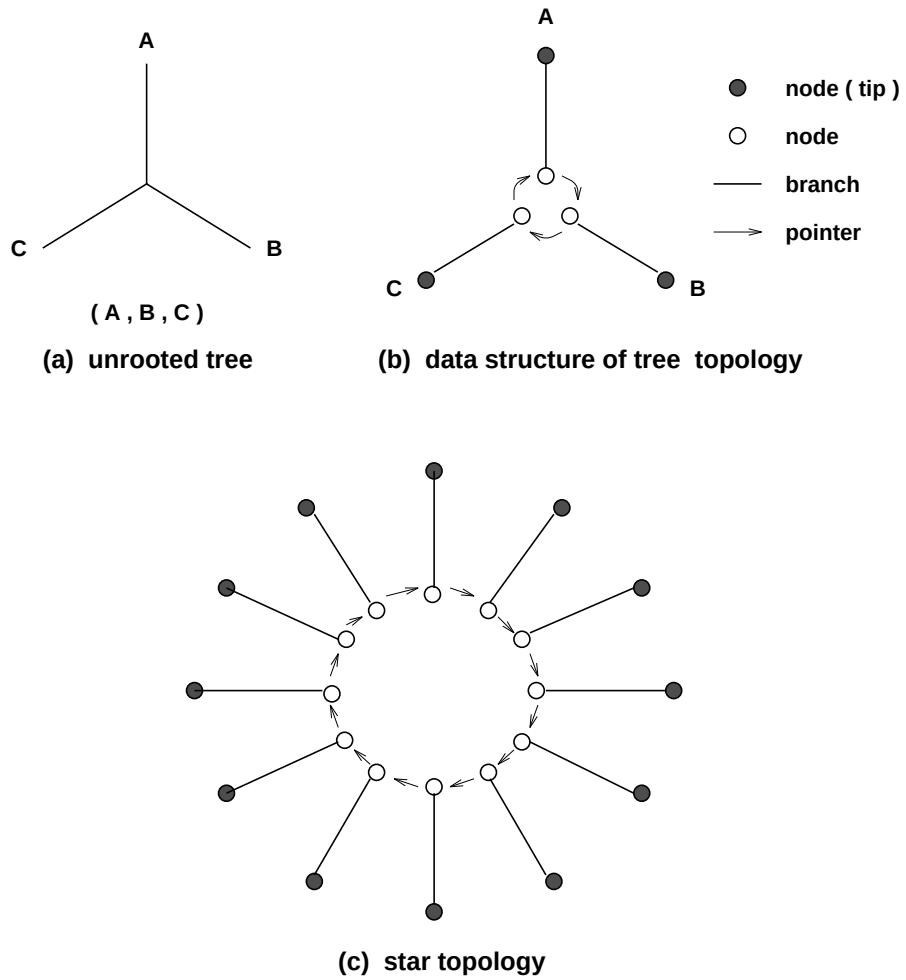


Figure 3.12: Topological data structure.

3.4.2 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[63]).

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 with only these three. Then, the fourth species is taken, and where it should be placed in the tree is evaluated. All possibilities (bifurcating trees) when adding the fourth species are examined. The best one by the likelihood criterion is chosen as the basis for further operations. Then, the fifth species is added, and again the best placement 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[69]). The tree resulting from this procedure depends on the order of the input species. Hence, Felsenstein recommends performing a number of runs with different orderings of the input species.

An alternative strategy which we employ in the automatic and semi-automatic search options of ProtML is called "star decomposition"(Adachi and Hasegawa 1992[4]; Saitou 1990[220]). This is similar to the procedure employed by the neighbor-joining (NJ) algorithm for a distance matrix (Saitou and Nei 1987[221]; for a worked example see Swofford et al. 1996[240]). This procedure starts with a star-like tree. After decomposing (joining branches) in the star-like tree step by step, we obtain a bifurcating tree if all multifurcations can be resolved. Since the information from all of the species under analysis is used from the beginning, the inference of the tree topology may hopefully be stable by this procedure.

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 the AIC measure (Akaike 1973[12], 1974[13]), the program decides whether the multifurcation should be further resolved or not. This criterion works nicely when the substitution model assumed in the phylogenetic analysis represents the real process which has generated the data. However, when there exists a discrepancy between the assumed model and the real process as is always the case in analyzing real data, this criterion tends to prefer a more resolved bifurcating tree to a multifurcating tree (Hasegawa, unpublished). In this situation, Kishino and Hasegawa's (1989[147]) test among the alternatively bifurcating trees might help to decide whether the multifurcation should be further resolved.

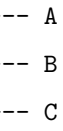
Although the star decomposition algorithm seems efficient in finding the ML tree for problems in which the number of OTUs is about 10 (e.g., Russo et al. 1996[217]), it is not very efficient with many OTUs. The final tree by the star decomposition is uniquely defined, and when erroneous relationships occur in early stages of the procedure, they cannot be corrected in later stages. The local rearrangement method described in the next subsection might be more useful in a wider range of problems. By using many alternative starting trees, the method can produce many candidate trees which can be compared

with the likelihood criterion.

3.4.3 Topology Search by Local Rearrangements

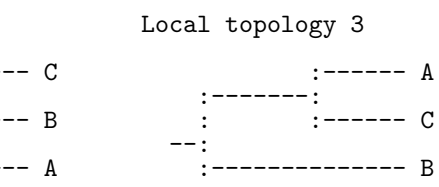
Once an approximate tree topology is obtained by star decomposition as mentioned in the preceding subsection, using either a distance matrix or the parsimony method, the search for better tree topologies by the likelihood criterion can be conducted through local rearrangement which is similar to the method used in the DNAML program of PHYLIP (Felsenstein 1993[69]) and will be described below. These rearrangements are commonly called nearest-neighbor interchanges (abbreviated NNI; e.g., see Swofford et al. 1996[240]).

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



where A, B, C, and the outgroup are subtrees.

A local rearrangement considers the two alternative trees;



and the program also estimates approximate bootstrap probabilities (Felsenstein 1985[67]) among these three trees by the RELL method (Kishino et al. 1990[148]; Hasegawa and Kishino 1994[97]). Since the branching orders within the subtrees, A, B, C and outgroup, are fixed, these are not real bootstrap probabilities, and we 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 that particular internal branch when the other parts of the tree are correct.

If it turned out that another local tree topology has higher likelihood than Local topology 1 and hence higher LBP, then a rearrangement is carried out. This procedure is repeated until all the internal branches are traversed. Since a rearrangement around a branch may make the previously established branches not optimal, the local rearrangements do not end until the program traverses the entire tree without finding any further improvement of 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 are uncertain, then there are 15, 105, or 945 alternative topologies rearranging these branches, and

we can consider them all looking for a better tree topology. By using this modified 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 has the highest likelihood, and it may still depend on the initial tree. For this reason, use of several alternative initial trees is recommended, and a tree with the highest likelihood from several runs should be chosen. For example, NJ analyses with bootstrap resampling might be useful in order to generate alternative initial trees.

Recently, Strimmer and von Haeseler (1996[236]) devised a new method of topology search for the ML tree, which is called "quartet puzzling". Since quartet puzzling does not always find the highest likelihood tree, it too might benefit from local rearrangements.

3.4.4 Example of Application of the Local Rearrangements

Here we give an example of the application of the local rearrangement method described in the preceding subsection. We will apply this method to the amino acid sequences of elongation factor 1α (EF-1α), as used in Hashimoto et al. (1995[106]) and listed in Table 3.2

Table 3.2: List of EF-1α data.

Abbrev.	species name	reference	database
Metazoa			
Homsa	*Homo sapiens*	Uetsuki et al. 1989[254])	X03558
Xenla	*Xenopus laevis*	Krieg et al. 1989[158]	X52975
Drome	*Drosophila melanogaster*	Hoveman and Richer 1988[121]	X06869
Artsa	*Artemia salina*	van Hemert et al. 1984[255]	X03349
Fungi			
Sacce	*Saccharomyces cerevisiae*	Nagashima et al. 1986[191]	X00779
Canal	*Candida albicans*	Sundstrom et al. 1990[238]	M29934
Mucra	*Mucor racemosus*	Linz et al. 1986[172]	J02605
Absgl	*Absidia glauca*	Burmester (unpubl.)	X54730
Plantae			
Arath	*Arabidopsis thaliana*	Liboz et al. 1989[171]	X16430
Lyces	*Lycopersicon esculentum*	Pokalsky et al. 1989[210]	X53043
Protista			
Dicdi	*Dictyostelium discoideum*	Yang et al. 1990[268]	X55972
Euggr	*Euglena gracilis*	Montandon and Stutz 1990[188]	X16890
Trycr	*Trypanosoma cruzi*	Hashimoto et al. 1995[106]	D29834
Tetpy	*Tetrahymena pyriformis*	Kurasawa et al. 1992[163]	D11083
Plafa	*Plasmodium falciparum*	Williamson (unpubl.)	X60488
Enthi	*Entamoeba histolytica*	De Meester et al. 1991[55]	M34256
Giala	*Giardia lamblia*	Hashimoto et al. 1994[108]	D14342
Archaebacteria			
Sulac	*Sulfolobus acidocaldarius*	Auer et al. 1990[26]	X52382
Metva	*Methanococcus vannielii*	Lechner and Böck 1987[167]	X05698
Halma	*Halobacterium marismortui*	Baldacci et al. 1990[30]	X16677

Fig. 3.13 shows the NJ tree of EF-1α in which the branch lengths and LBPs were estimated by ProtML. The distance matrix provided for the NJ analysis was estimated with 2-OTUs trees by ProtML using the JTT-F model. In this tree, animals do not form a monophyletic clade; i.e., fungi cluster with *H. sapiens*/*X. laevis*, and another group of animals, *D. melanogaster*/*A. salina*, is an outgroup to them. However, the LBP for the fungi/*H. sapiens*/*X. laevis* clustering is only 32% by the ProtML analysis, so

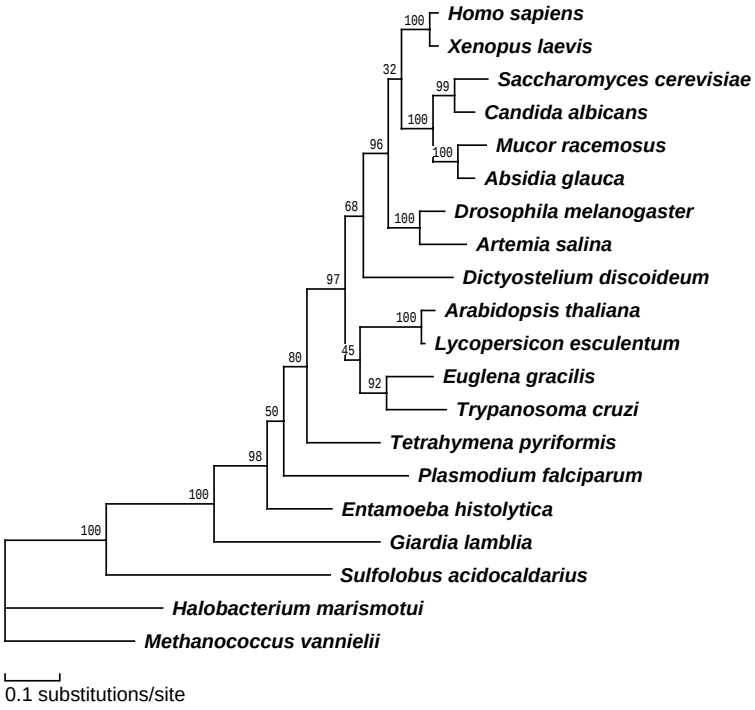


Figure 3.13: NJ tree of EF-1α in which the branch lengths and LBPs were estimated by the ProtML (JTT-F model).

this odd tree might be improved by an ML search.

The process of the local rearrangements applied to the EF-1α data starting with the NJ tree (Fig. 3.13) is shown below.

```
protml 2.3b3 (07/03/96) JTT-F 20 OTUs 382 sites. EF-1a
#1
                                        :-1 Homsa
                                     :--21 100
                                     :   :-2 Xenla
                                  **25 29 71 21&26
                                  :  :       :--5 Sacce
                                  :  :   :--22 98
                                  :  :   :   :--6 Canal
                                  :  :--24 100
                                  :      :   :--7 Mucra
                                  :      :--23 99
                                  :          :-8 Absgl
                               :--27 94
                               :   :   :--3 Drome
                               :   :--26 99
                               :       :---4 Artsa
                            :--28 69
                            :   :-----11 Dicdi
                         :---32 97
                         :    :        :-9 Arath
                         :    :  :----29 100
                         :    :  :     :-10 Lyces
                         :    **31 49 51 28&30
                         :       :   :---12 Euggr
                         :       :--30 91
                         :           :----13 Trycr
                      :--33 81
                      :   :-----14 Tetpy
                   :-34 53
                   :  :-------15 Plafa
              :---35 98
              :    :----16 Enthi
:-----------36 100
:           :----------17 Giala
:
:----------------18 Sulac
:
:  :----------19 Halma
:-37 100
   :---------20 Metva
```

LBP (in %) is given to the right of each internal branch (or node) number. When the local branching order is not optimum, the branch is represented by asterisks. In this example, two branches are indicated by asterisks. For the branch 25, it notes

```
 **25  29  71  21&26
```

This means that branch 25 has 29% LBP[1], but if node 21 and node 26 are linked, LBP becomes 71%. Furthermore, for branch 46, it notes

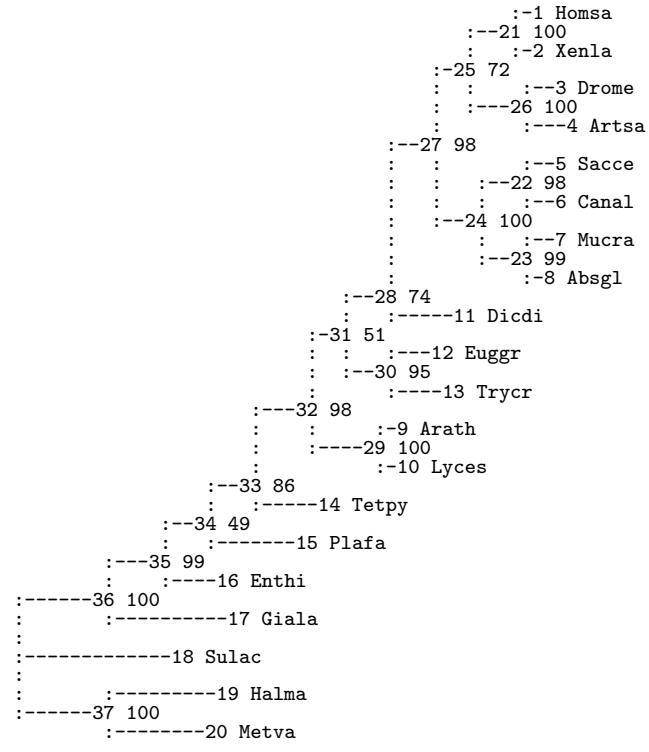
```
 **31  49  51  28&30
```

Rearrangements are done:

```
% 25  21<->26   ln L:   -7110.941 +     4.6392438238
% 31  28<->30   ln L:   -7110.941 +     0.2093060069
```

These numbers mean that, by linking node 21 with node 26, the log-likelihood of the preceding tree (-7110.941) is improved by 4.64, and by linking node 28 with node 30, log-likelihood is improved by 0.21. The final tree, which cannot be improved by local rearrangement, is as follows;

```
((((((((((Homsa,Xenla),(Drome,Artsa)),((Sacce,Canal),(Mucra,Absgl))),Dicdi),
(Euggr,Trycr)),(Arath,Lyces)),Tetpy),Plafa),Enthi),Giala),Sulac,(Halma,Metva));
```



No.1	ext. branch	S.E.	int. branch	S.E.	LBP	2nd	pair
Homsa	1 1.57	0.70	21 4.69	1.29	1.0	0.0	1&26
Xenla	2 1.40	0.66	22 4.05	1.20	0.985	0.011	23&6
Drome	3 4.52	1.21	23 4.58	1.25	0.993	0.004	22&8
Artsa	4 8.29	1.60	24 4.66	1.29	0.996	0.003	22&25
Sacce	5 6.05	1.38	25 2.36	0.98	0.723	0.251	21&24
Canal	6 3.54	1.09	26 7.09	1.52	1.0	0.0	21&4
Mucra	7 5.02	1.25	27 4.60	1.34	0.981	0.018	25&11
Absgl	8 3.27	1.03	28 3.49	1.27	0.739	0.253	27&30
Arath	9 2.41	0.84	29 11.45	2.00	1.0	0.0	31&10
Lyces	10 0.54	0.50	30 4.49	1.36	0.954	0.031	28&13
Dicdi	11 16.14	2.33	31 2.57	1.16	0.514	0.378	29&30
Euggr	12 9.77	1.81	32 7.26	1.68	0.982	0.018	14&31

[1]In Fig. 3.13, the LBP for this branch is 32%, not 29% as shown here. This difference is due to the LBPs in Fig. 3.13 being estimated by the RELL method (Kishino et al. 1990[148]) with 10^4 replications, those in the latter were estimated with 10^3 replications.

```
Trycr       13   9.84  1.81   33   3.66  1.41  0.856  0.096  32&15
Tetpy       14  13.22  2.17   34   3.31  1.41  0.493  0.450  33&16
Plafa       15  22.93  2.89   35   9.51  2.39  0.986  0.011  34&17
Enthi       16  11.61  2.12   36  19.45  3.52  1.0    0.0    18&17
Giala       17  30.79  3.68   37  18.70  3.40  0.996  0.004  36&20
Sulac       18  40.74  4.51  TBL :     360.10  iter: 1
Halma       19  28.97  3.74  ln L:   -7105.02 +- 272.50
Metva       20  23.57  3.42  AIC :   14322.04
```

"Branch" (branch length) refers to the estimated number of substitutions per 100 sites, and S.E. is the standard error of this number.

Fig. 3.14 is the printout of the EPS file of the final tree. The log-likelihood of the NJ tree is −7110.9, while that of the resultant ProtML tree is −7105.0, showing an improvement of log-likelihood by 5.9.

Although, in the NJ tree, the fungi clade ((Sacce, Canal), (Mucra, Absgl)) intrudes into metazoa, linking with vertebrates (Homsa, Xenla), leaving arthropoda (Drome, Artsa) as an outgroup, in the ProtML tree obtained by the local rearrangement, metazoa is monophyletic and is a sister group to fungi (Hasegawa et al. 1993[94]; Baldauf and Palmer 1993[31]; Wainright et al. 1993[259]; Nikoh et al. 1994[197]). The ProtML tree is biologically more reasonable than the NJ tree in this respect.

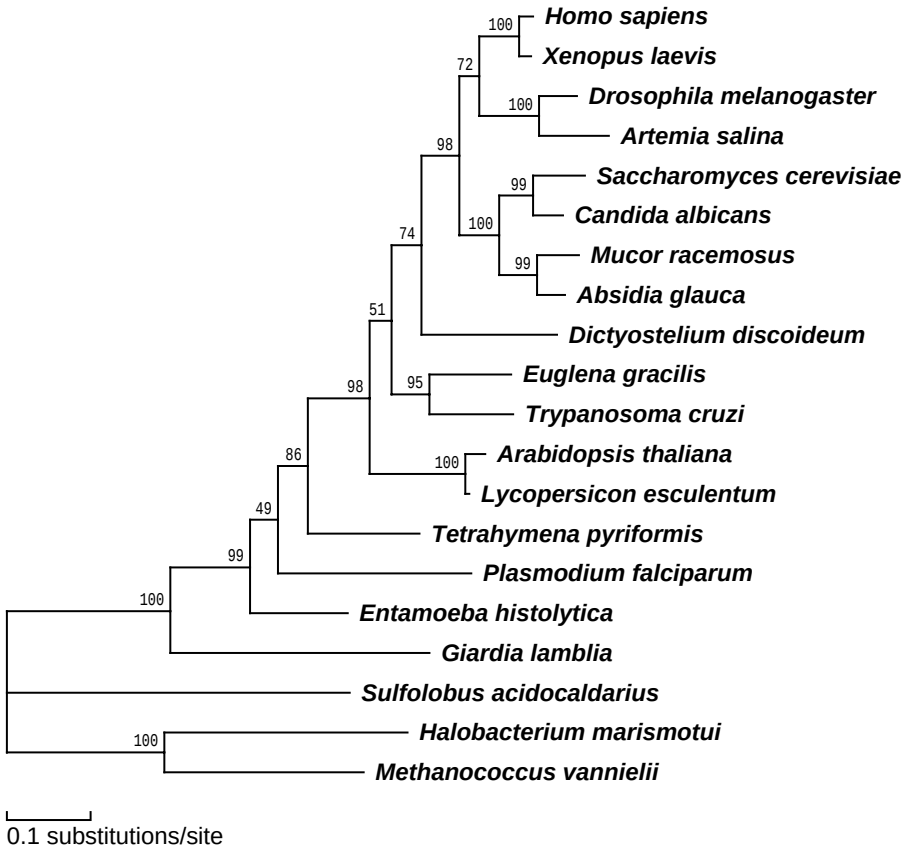


Figure 3.14: ProtML tree of EF-1α obtained by the local rearrangement (JTT-F model).

3.5 Approximate Likelihood Method for Exhaustive Search

Several 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 might be limited (e.g., Nei 1987[195]; Hillis et al. 1994[115]). While it is true that the ML method is computationally intensive and that, at present, there exist several limitations in applying the method to real problems, computational ability is rapidly improving. Furthermore, several methods to reduce the computational burden of ML analyses are being invented. One is the approximate likelihood method presented below (Adachi 1995[1]; and also considered in Waddell 1995[257], called non-iterated likelihood).

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 very bad and unpromising. If we can quickly eliminate these trees by an approximate method, the ML criterion can be applied to many species. 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 for the Newton-Raphson method are given by the ordinary least squares method. It appears that there is a remarkablly good correlation between the likelihood calculated from the intitial values, which is called the approximate likelihood (AL) (or non-iterated likelihood in Waddell 1995[257]), and the optimized likelihood. Therefore, we can exclude unpromising trees by using the AL which can be calculated rather quickly.

The approximate log-likelihood of a tree is

$$l(\widehat{\boldsymbol{t}}|\boldsymbol{X},T)=\sum_{h=1}^{n}\log f(\boldsymbol{X}_h|T,\widehat{\boldsymbol{t}}) \tag{3.35}$$

where

$$\widehat{\boldsymbol{t}}=\left(\widehat{t}_1,\widehat{t}_2,\cdots,\widehat{t}_9\right)^T. \tag{3.36}$$

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

$$\widehat{\boldsymbol{t}}=(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{D}. \tag{3.37}$$

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 may still be attainable. Thus we can exclude unpromising trees by the AL method, and can select the best, say 1000 or 2000, trees (by the AL criterion) that are provided for the full ML analysis.

Fig. 3.15 gives an example of the relationship between the approximate likelihood and the optimized likelihood, here for the possible 945 trees of EF-1α sequences from 7 species chosen from the list in

Table 3.2; *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 method.

Fig. 3.16 gives the relationship between the AL and the likelihood estimated by the ML for the EF-1α data from 5 species chosen from the list in Table 3.2 plus additional two archaebacterial species; *Homo sapiens*, *Entamoeba histolytica*, *Sulfolobus acidocaldarius*, *Methanococcus vannielii*, *Halobacterium marismortui*, *Thermococcus celer* (Auer et al. 1990[25]), and *Thermoplasma acidophilum* (Tesch and Klink 1990[247]). This data set contains more diverse species (including both eukaryotes and archaebacteria) than the preceding one: the correlation between 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.

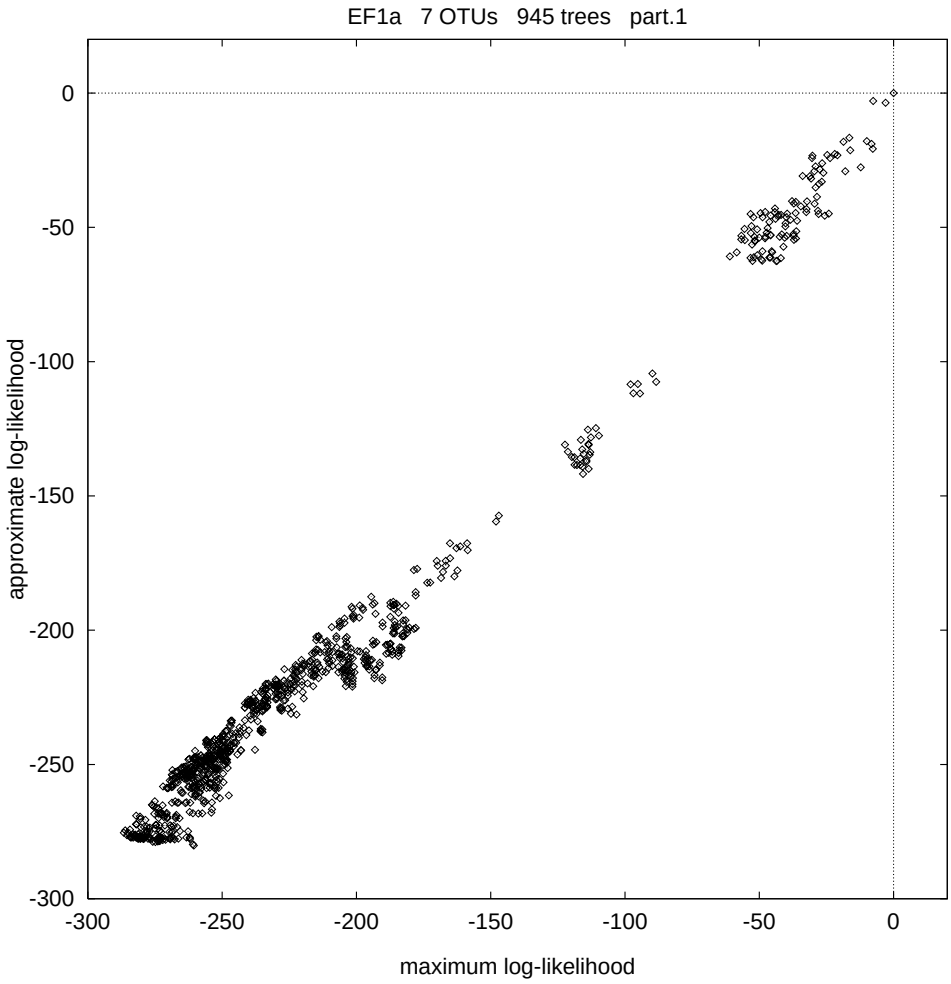


Figure 3.15: Maximum likelihood vs. Approximate likelihood. Only log-likelihood differences from the highest likelihood tree are shown.

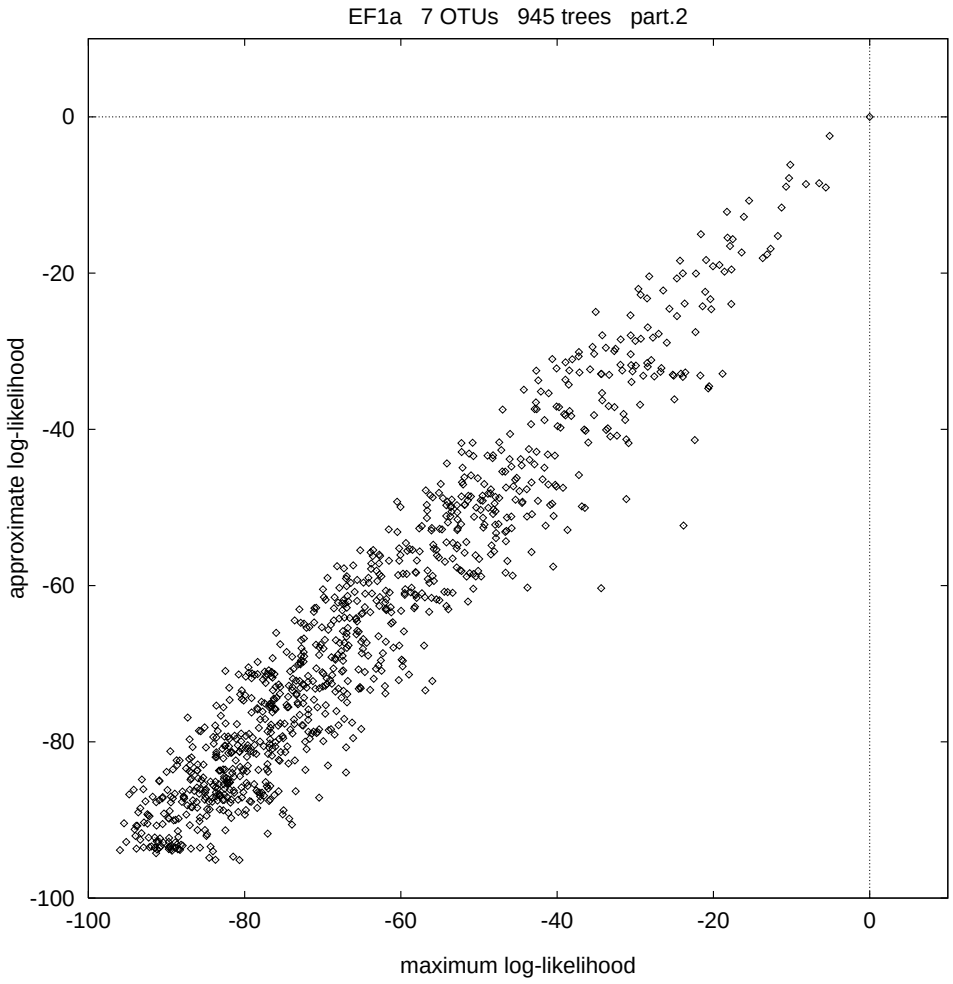


Figure 3.16: Maximum likelihood vs. Approximate likelihood. Only log-likelihood differences from the highest likelihood tree are shown.

Chapter 4

MOLPHY: A Computer Program Package for Molecular Phylogenetics

Readme

```
        This is the MOLPHY (ProtML) distribution,  version 2.3.
        Copyright (c) 1992-1996, 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 specified codon sites
  molinfo:  get varied 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).
However, MOLPHY has NOT been tested on VAX, IBM-PC, and Macintosh.

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

Next are the users manuals for MOLPHY.
```

Installation

To instal MOLPHY, UNIX users should be able to type "make" in molphy-2.3/src directory. (Edit the molphy-2.3/src/Makefile if you need to customize it)

```
% cat molphy-2.3.tar.Z | uncompress | tar xvf -
% cd molphy-2.3/src
% make
% make install
```

To 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 Overview of the Input and Output Formats

This test data is a subset of the protein-encoding genes of mitochondrial DNA from primates (listed in Horai et al., 1992[118]).

4.1.1 Input Format

A standard input file for MOLPHY is as follows (this test file is named "pri5.nuc").

```
5 357 mtDNA Primates
Chimp Pan troglodytes
CTAATAATCTTAACCTGAATAGGGATATGGTGGCCCCTCATATGAATCATGACCGTCTGA
TATATGGGAATAATATGAAATATGGTAATTTGAGACCAAGCCATCATGATTATGCGTGTC
GTAATGGTCCTAGTAGAGGCAAACTGACCCTCTATTATCTGCACTAGTTCAGTCGTCATA
GTCTTTTCATGAACCATAGACGTTGTTGCTACAATAACTGCCGTATGACCCATAACCCCC
ATAACAGTCACCATATCAAATTACCTACCCTCACCCATAAAAATAAACTACAATAAACCA
GTACTAATCTTCCCTGTCCATCTCACCCAATCAATAACTATAAGCACTATAGTATCC
Human Homo sapiens
CTAATAATCTTAGCCTGAATAGGAATATGATGACCTCTCATATGAGTCATAACCGTCTGA
TACATGGGGATAATATGAAATATGGTGATTTGAGACCATACTATCATAATCATGCGTATC
GTAATGGTCCTAGTAGAAATAAACTGACCCTCTATCACCTGCACTAACTCAGTCGTCATA
GTCCTTTCATGGACCATAGACATTATTGCTACAATGACCACCGTATGGCCCATAACCCCC
ATAACAATCACCATAACAAACTACCTACCCTCACCCATAAAAATAAATTATAACAAACCA
GTACTGATCTTTCCTATCTATCTCACCAAATCAATGACCATAAACACTATAGTATCC
Goril Gorilla gorilla
CTAATAGTTCTAACCTGAATAGGGATATGGTGACCCTTCATATGGATCATAACCGTCTGA
TATATAGGAATAATATGAAATACCATGATTTGAGATCACGCCATCATAATTATACATATC
GTGATAGTCCTAATCGAAACAAATTGATCTTCTATCATCTGCAACAACTCAATCGTCATG
ATCTTCTCATGAACCATAGACGTTGTCGCTACAATGGCCACCGTATGGCCCATAGCCCCA
ATAACAATTACCGTTACAAATTACCTACCCTTAACTATAAAAATAAACTTCTGTAAACCA
GTATTAATTCTTCCTATCTATCTCGCCCAATCAATAACTATAAACGCCATGATATGA
Orang Pongo pygmaeus
CTATCCATCCCAGCCTGGATGGGGATATGATGACTCTTCACATGAATTATATCCATCTGA
CACATAGGAGTCATATGAAACACTATCATCTGGAACCACATCACCATAGTCATACGCATT
GCAATAGTCCCAATTCAAACAAGCTGGCCCCCCGTCATCTGCACTAACTCAATTATTTTA
ATCTTCTCATGGACCATGGACGTCGTTACCTCAATGGCTACCACATGGCTCGTCACTCCA
ACAGCAATCACCCTATCACACCTCCCAACCCCATTTACCAAAACACCCCACGCCAAACTA
ATTCTAGTCTTTCCCGTCCATTTCACCCGACTAATAATCACCAACACTATAACATCC
Siama Hylobates syndactylus
TTCCCTGCCCCAGCCTGGATAGGAATGTGATGGCCTTTCATATGAGTAATATCCGTCTGG
CACATAGGAATAATGTGGGACACCGTAGTCTGAGATCACGCCATTATAGTAATACGTATC
GTGATAATCCTAATCCAGACTAACTGGCCCCCTATCTCTAGCACTAATACGGTCGTTTTA
ATCTTTGCATGAGCCATAGAAATTGTCACTTCCATAACCACCGTGTGACCTATCACATCA
ATAACACTCATAACAATGTACTACCCAGCCTCCCTCATAAACATTCCCCACAACAACCAC
GTACCAATTTTTTCCATTTACCTCACCCAATTAATAACACTAAACACTATAATTTCT
```

This kind of format is called "MOLPHY format". The MOLPHY format is a standard input format used in analyzing sequence data by MOLPHY, and is an ASCII text file. Note, this format is very similar to PHYLIP version 3.4 format. The first line of the file contains the number of OTUs (number of sequences; 5) and sequence length (number of characters; 357) in this order and separated by blanks. There may then follow the title of the data and/or comments. In our test data, specification of the DNA type (mtDNA) and classification of organisms (primates) are given. These comments are shown in the 1st line of the output. The title and comments can be omitted.

The information for each OTU follows, starting on a new line with an abbreviation of the OTU name. Since the abbreviation is used in representing tree topologies, it must be unique in the input file. Scientific name of the organism may follow the abbreviation separated by a blank. The abbreviation should not contain blanks, and hence characters after a blank are regarded as representing a scientific name. For the OTU with a scientific name, the scientific name (in italic) is used instead of the abbreviation in the presentation of the phylogenetic tree by an EPS file (njdist.eps, protml.eps, and nucml.eps). The common name or supplementary information in parentheses may follow the scientific name, and it is printed in roman type within the phylogenetic tree (e.g., Figs. 2.1, 2.2, 5.13, 5.14, 5.20, and 5.21).

Sequence data may start from the next line after the name and comments. The sequence data can be given in free format, and given that the number of characters is as indicated in the 1st line of the file, any representation is allowed; the data can have internal blanks in the sequence. Therefore, a blank should not be used as a symbol for deletion. The standard format we prefer does not contain blanks, and each line (except the last line for each OTU) contains 60 characters.

The standard input data for MOLPHY is in "sequential" format, with all of the data for the first OTU, then all of the characters for the next OTU, and so on. The "interleaved" format (sequences put in aligned form; Felsenstein 1993[69]) can be converted into a MOLPHY format (sequential file) by using a supplied utility "inl2mol".

To repeat, the MOLPHY format is as follows;
'Number of OTUs' 'Number of characters' 'comments' 'Abbreviation of OTU1' 'scientific name for OTU1 (English name)' 'Sequence 1' 'Abbreviation of OTU2' 'scientific name for OTU2 (English name)' 'Sequence 2' .....

Either comments, the scientific name, or the common name may be omitted.

Our test data represents a protein-encoding gene. When a protein gene is analyzed, the sequence should start from a 1st codon position, and ends at a 3rd codon position, and hence the number of nucleotide sites should be a multiple of 3.

There are two alternative ways to analyze this data. One is to translate this data into protein sequences, and then to analyze it by using ProtML. Another is to make three files of each codon position, and analyze them by using the NucML. In the case when the data is analyzed in the nucleotide sequence level, the rate and transition/transversion ratio differ drastically among the different codon positions, so

it is recommended to analyze the three positions separately (e.g., Hasegawa and Adachi, 1996[89]).

To translate the nucleotide sequences into the amino acid sequences, the supplied utility "nuc2ptn" is used. The default is the universal code, and in the case of vertebrate mitochondrial code, -m (mitochondria) option should be used, i.e.,

```
nuc2ptn -m pri5.nuc > pri5.ptn
```

where pri5.ptn is the output file. Extension "nuc" is used for nucleotide sequences, and "ptn" for protein sequences. The output file of amino acid sequences, pri5.ptn, is as follows;

```
5 119 mtDNA Primates
Chimp Pan troglodytes
LMILTWMGMWWPLMWIMTVWYMGMMWNMVIWDQAIMIMRVVMVLVEANWPSIICTSSVVM
VFSWTMDVVATMTAVWPMTPMTVTMSNYLPSPMKMNYNKPVLIFPVHLTQSMTMSTMVS
Human Homo sapiens
LMILAWMGMWWPLMWVMTVWYMGMMWNMVIWDHTIMIMRIVMVLVEMNWPSITCTNSVVM
VLSWTMDIIATMTTVWPMTPMTITMTNYLPSPMKMNYNKPVLIFPIYLTKSMTMNTMVS
Goril Gorilla gorilla
LMVLTWMGMWWPFMWIMTVWYMGMMWNTMIWDHAIMIMHIVMVLIETNWSSIICNNSIVM
IFSWTMDVVATMATVWPMAPMTITVTNYLPLTMKMNFCKPVLILPIYLAQSMTMNAMMW
Orang Pongo pygmaeus
LSIPAWMGMWWLFTWIMSIWHMGVMWNTIIWNHITMVMRIAMVPIQTSWPPVICTNSIIL
IFSWTMDVVTSMATTWLVTPTAITLSHLPTPFTKTPHAKLILVFPVHFTRLMITNTMTS
Siama Hylobates syndactylus
FPAPAWMGMWWPFMWVMSVWHMGMMWDTVVWDHAIMVMRIVMILIQTNWPPISSTNTVVL
IFAWAMEIVTSMTTVWPITSMTLMTMYYPASLMNIPHNNHVPIFSIYLTQLMTLNTMIS
```

4.1.2 Basic Statistics of the Data

Before the data is provided to a phylogenetic analysis, the properties of the data should be examined. For this purpose, the ProtST program, which calculates basic statistics of amino acid sequence data, can be used;

```
protst pri5.ptn > pri5.pst
```

Numbers of amino acid differences, amino acid frequencies, and distances of amino acid composition ("bias") between all pairs of sequences are calculated. The amino acid composition distance between species i and j is measured 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 in species i (Cao et al. 1994[40]; for a euclidian version, see Lockhart et al. 1994[173]). Only sites that have not experienced ins/del are taken into account. These statistics can be used to check whether amino acid composition varies among species due to different base compositions of the genomes (e.g., Hasegawa and Hashimoto 1993[93]; Hashimoto et al. 1994[108], 1995[107]). The output file, pri5.pst, appears as follows;

```
protst 1.2.1  Jun 25 1996    5 OTUs 119 sites  mtDNA Primates

Diff          1   2   3   4   5
            Chi Hum Gor Ora Sia
1   Chimp   Chi  18  32  63  57
2   Human    18 Hum  31  64  51
3   Goril    32  31 Gor  62  58
4   Orang    63  64  62 Ora  61
5   Siama    57  51  58  61 Sia
```

```
            A Ala  R Arg  N Asn  D Asp  C Cys  Q Gln  E Glu  G Gly  H His  I Ile
1   Chimp   0.034  0.008  0.042  0.017  0.008  0.017  0.008  0.017  0.008  0.067
2   Human   0.017  0.008  0.059  0.017  0.008  0.0    0.008  0.017  0.008  0.092
3   Goril   0.050  0.0    0.059  0.017  0.017  0.008  0.008  0.017  0.017  0.109
4   Orang   0.042  0.017  0.034  0.008  0.008  0.008  0.0    0.017  0.042  0.126
5   Siama   0.050  0.008  0.050  0.017  0.0    0.017  0.008  0.017  0.034  0.101
    mean    0.039  0.008  0.049  0.015  0.008  0.010  0.007  0.017  0.022  0.099

            L Leu  K Lys  M Met  F Phe  P Pro  S Ser  T Thr  W Trp  Y Tyr  V Val
1   Chimp   0.059  0.017  0.193  0.017  0.067  0.076  0.101  0.084  0.025  0.134
2   Human   0.067  0.025  0.202  0.008  0.067  0.050  0.126  0.084  0.034  0.101
3   Goril   0.067  0.017  0.193  0.025  0.050  0.042  0.101  0.092  0.025  0.084
4   Orang   0.076  0.017  0.101  0.042  0.076  0.067  0.160  0.084  0.0    0.076
5   Siama   0.059  0.0    0.151  0.034  0.076  0.067  0.109  0.084  0.025  0.092
    mean    0.066  0.015  0.168  0.025  0.067  0.061  0.119  0.086  0.022  0.097

Bias  x1e3   1   2   3   4   5
            Chi Hum Gor Ora Sia
1   Chimp   Chi 101 118 218 118
2   Human   101 Hum 101 210 134
3   Goril   118 101 Gor 193 109
4   Orang   218 210 193 Ora 143
5   Siama   118 134 109 143 Sia
```

In the bias table, it appears that orangutan shows the highest average bias with respect to all other species now considered. Siamang must be the outgroup to all the others including orangutan, but since the evolutionary rate has been higher in the orangutan lineage than in the others (Adachi and Hasegawa 1995[5]), the number of amino acid differences of orangutan from the African apes/human exceed that of siamang. The composition distance relevant to orangutan is further exaggerated, indicating that orangutan has different base composition in mtDNA (Adachi and Hasegawa 1996[11]) and that difference of base composition affects amino acid composition of proteins (Sueoka 1961[237]; Crozier and Crozier 1993[52]; for counter-example, see e.g., Hashimoto et al. 1994[108], 1995[107])

When the -a (alignment) option is used with ProtST by entering

```
protst -a pri5.ptn
```

then, the following representation of the aligned sequences is given.

```
protst 1.2.1  Jun 25 1996    5 OTUs 119 sites  mtDNA Primates

CONSENSUS  LMILAWMGMW WPFMWIMTVW YMGMMWNTVI WDHAIMIMRI VMVLIETNWP SIICTNSVVM
Chimp      ....T..... ..L..... .......M.. ..Q......V ....V.A... .....S....
Human      .......... ..L..V.... .......M.. ...T..... ....V.M... ..T.......
Goril      ..V.T..... .......... ........M. ........H. .........S ....N..I..
Orang      .S.P...... .L.T...SI. H..V....I. .N.IT.V... A..P.Q.S.. PV.....IIL
Siama      FPAP...... .....V.S.. H.....D..V ......V... ..I..Q.... P.SS..T..L
                  10         20         30         40         50         60
CONSENSUS  IFSWTMDVVA TMTTVWPMTP MTIT..NYLP S.MKMN.NKP VLIFPIYLTQ SMTMNTM.S
Chimp      V......... ...A...... ..V.MS.... .P....Y... .....VH... ....S..V.
Human      VL.....II. .......... ....MT.... .P....Y... .........K .......V.
Goril      .......... ..A.....A. ....VT.... LT....FC.. ...L....A. .....A.MW
Orang      .........T S.A.T.LV.. TA..LSHLPT PFT.TPHA.L I.V..VHF.R L.IT...T.
Siama      ..A.A.EI.T S......I.S ..LMTMY.PA .L.NIPH.NH .P..S..... L..L...I.
                  70         80         90        100        110
```

We can thus see the alignment of the data at hand. It must be noted that MOLPHY does not contain any alignment program, and the input file of MOLPHY format should be an aligned one.

4.1.3 ProtML

Let us now consider phylogenetic inference using this amino acid sequence data. Firstly, a simple method of NJ can be applied. Since NJ is a distance method, a distance matrix must be estimated. We can

estimate the distance matrix using pairwise ML by entering;

```
protml -mfD pri5.ptn > pri5.dis
```

The -m option designates the amino acid substitution model for proteins encoded by vertebrate mitochondrial DNA (the mtREV model; section2.2, and Adachi and Hasegawa, 1996[10]); the f-option designates that the amino acid transition matrix is adjusted so that the equilibrium frequencies are the data frequencies; the D-option designates estimate a distance matrix. The distance matrix estimated by pairwise ML is stored in the file "pri5.dis" as follows;

```
5 119 sites mtREV24-F mtDNA Primates
Chimp Pan troglodytes
 0.000000000000 0.164223391360 0.324971183173 0.902582687656 0.776294148912
Human Homo sapiens
 0.164223391360 0.000000000000 0.311311879611 0.896886489077 0.629266051712
Goril Gorilla gorilla
 0.324971183173 0.311311879611 0.000000000000 0.931866113135 0.850510393531
Orang Pongo pygmaeus
 0.902582687656 0.896886489077 0.931866113135 0.000000000000 0.898716655371
Siama Hylobates syndactylus
 0.776294148912 0.629266051712 0.850510393531 0.898716655371 0.000000000000
```

The extension "dis" means a distance matrix. From this distance matrix, an NJ tree can be estimated with the NJdist program by entering;

```
njdist -tpri5 pri5.dis > pri5.nj
```

The result is stored in the file named pri5.nj. The t-option designates store the estimated tree in the file pri5.tpl. The extension "tpl", which means a tree topology file, is automatically attached. Without this t-option, the estimated topology is automatically stored in "njdist.tpl" file. The pri5.nj file contains:

```
njdist 1.2.5 (06/24/96) 5 OTUs 119 sites mtREV24-F mtDNA Primates

            :---1 Chimp
         :--7
         :  :--2 Human
:--------6
:        :------3 Goril
:
:----------------4 Orang
:
:-----------5 Siama
```

On the other hand, the topology file "pri5.tpl" looks like this,

```
1 njdist 1.2.5 (06/24/96) 5 OTUs 119 sites mtREV24-F mtDNA Primates
(((Chimp,Human),Goril),Orang,Siama);
```

The tree is unrooted and in standard parenthetical notation. When NJdist is carried out, a figure of the phylogenetic tree is automatically stored in "njdist.eps" file, which is an EPS (Encapsuled PostScript) file. By using this file, the figure can be printed out with a PostScript printer to give Fig. 4.1.

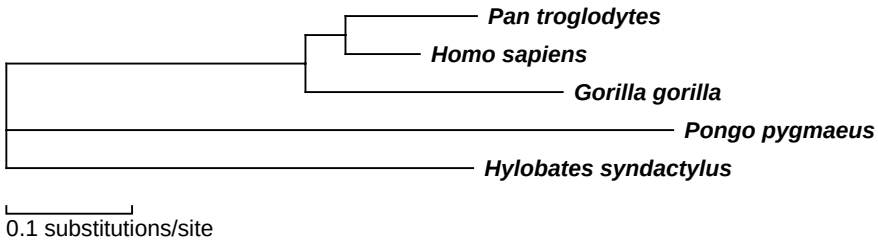


Figure 4.1: A printout of the njdist.eps file.

Every time NJdist is run, njdist.eps is overwritten, and hence you have to change the file name of njdist.eps if you want to keep the earlier results.

Next, we will estimate the phylogenetic tree using maximum likelihood. In looking for the ML tree topology, let us first use the "Local Rearrangement Search" option (subsection 3.4.3) starting from the NJ tree topology. Enter

```
protml -mfR pri5.ptn pri5.tpl > pri5.mlr
```

The R-option designates 'local rearrangement search of tree topology'. In this case, the mtREV model with the amino acid frequencies of the data is used. The amino acid sequence file pri5.ptn and the initial topology file pri5.tpl are designated after the option: the result is stored in the file named pri5.mlr which is shown below;

```
protml 2.3b3 (06/24/96) mtREV24-F 5 OTUs 119 sites. mtDNA Primates
#1
              :---1 Chimp
          :---6 77
          :   :--2 Human
:--------7 99
:         :------3 Goril
:
:---------------4 Orang
:
:----------5 Siama

(((Chimp,Human),Goril),Orang,Siama);

              :---1 Chimp
          :---6 71
          :   :--2 Human
:--------7 100
:         :------3 Goril
:
:---------------4 Orang
:
:----------5 Siama

No.1       ext. branch S.E.  int. branch S.E.   LBP    2nd   pair
Chimp        1   9.91  3.25    6   3.74  2.78  0.712  0.186   1&3
Human        2   6.92  2.76    7  23.55  6.66  0.996  0.004   4&3
Goril        3  19.29  4.90  TBL :     143.35  iter: 1
Orang        4  47.86  9.56  ln L:    -868.79 +- 32.37
Siama        5  32.08  7.56  AIC :    1789.57
```

The first tree is the starting tree. The last tree is the optimal tree obtained after the local rearrangement search, and is the highest likelihood tree found by this limited search. In this case, the tree obtained

by the local rearrangement search coincides with the starting NJ tree. For each internal branch, a local bootstrap probability (LBP; in %) (see page 49) estimated by the RELL method with 10^3 replications is shown after the node number. Branch length refers to the estimated number of substitutions per 100 sites, while S.E. is estimated in the same way as Felsenstein (1993[69]).

The ProtML generates the file protml.eps which stores the phylogenetic tree as an EPS file (Fig. 4.2).

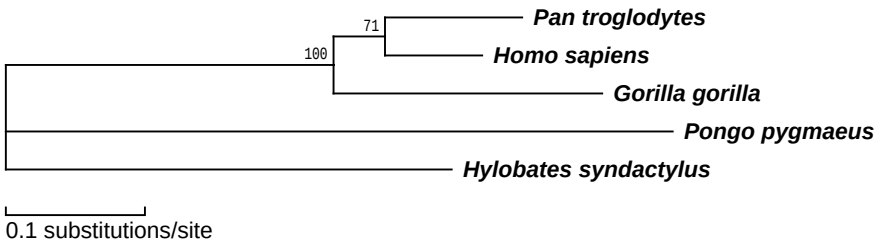


Figure 4.2: The printout of the protml.eps file.

In this figure, the horizontal length of each branch is proportional to the number of amino acid substitutions estimated by ProtML. It must be noted again that protml.eps is overwritten each time the ProtML is carried out.

Sometimes, we may be interested in comparing several competing hypotheses of phylogeny, which can be done by using the user's tree option as follows. First, a tree topology file, which contains candidate tree topologies, must be prepared. We will call this file pri5_user.tpl and for the five taxa it looks like:

```
  3
(((Chimp,Human),Goril),Orang,Siama);
(((Human,Goril),Chimp),Orang,Siama);
(((Chimp,Goril),Human),Orang,Siama);
```

If you then enter

```
protml -mf pri5.ptn pri5_user.tpl > pri5.ml
```

the output "pri5.ml" looks like:

```
protml 2.3b3 (07/12/96) mtREV24-F 5 OTUs 119 sites. mtDNA Primates
#1
              :---1 Chimp
          :---6
          :   :--2 Human
:--------7
:         :------3 Goril
:
:--------------4 Orang
:
:----------5 Siama

No.1       ext. branch S.E.  int. branch S.E.
Chimp        1   9.91  3.25    6   3.74  2.78
Human        2   6.92  2.76    7  23.55  6.66
Goril        3  19.29  4.90  TBL :     143.35  iter: 5
Orang        4  47.86  9.56  ln L:    -868.79 +- 32.37
Siama        5  32.08  7.56  AIC :    1789.57
```

```
#2
             :---2 Human
          :--6
          :  :-------3 Goril
:---------7
:         :---1 Chimp
:
:---------------4 Orang
:
:----------5 Siama

No.2       ext. branch S.E.  int. branch S.E.
Chimp        1  10.32  3.34    6  lower limit
Human        2   7.29  2.86    7  26.39  7.00
Goril        3  22.66  5.13  TBL :     147.15  iter: 4
Orang        4  49.01  9.75  ln L:    -871.06 +- 32.52
Siama        5  31.49  7.53  AIC :    1794.11  lower limit: 0.001
#3
             :---1 Chimp
          :--6
          :  :-------3 Goril
:---------7
:         :--2 Human
:
:---------------4 Orang
:
:---------5 Siama

No.3       ext. branch S.E.  int. branch S.E.
Chimp        1   9.91  3.30    6   0.68  1.69
Human        2   6.94  2.83    7  26.38  6.99
Goril        3  22.33  5.10  TBL :     146.88  iter: 6
Orang        4  49.41  9.81  ln L:    -870.97 +- 32.54
Siama        5  31.24  7.51  AIC :    1793.94

protml 2.3b3 mtREV24-F 3 trees 5 OTUs 119 sites. mtDNA Primates

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-------------------------------------------------------------------
1        -868.8     0.0 <-best   26    1789.6    0.0    ME   0.7172
2        -871.1    -2.3    2.9   26    1794.1    4.5    3.8  0.1038
3        -871.0    -2.2    3.0   26    1793.9    4.4    3.5  0.1790
```

Bootstrap probabilities (BP) among the candidate trees are estimated by the RELL method with 10^4 replications. TBL refers to 'total branch length', and the term ME means the tree with the least sum of edge length after the optimization by ML (in this case). This criterion may be useful at indicating the optimal tree (e.g., see also Waddell 1995[257], p. 314; Rzhetsky and Nei 1993[218]). If this tree is different to the ML tree, it is worth noting.

4.1.4 Nucleotide Sequences

Next, we will show an analysis at the nucleotide sequence level. From the nucleotide sequence file "pri5.nuc", by using our utility 'molcodon', we generate three files for the three different codon positions.

```
molcodon -1 pri5.nuc > pri5f.nuc
molcodon -2 pri5.nuc > pri5s.nuc
molcodon -3 pri5.nuc > pri5t.nuc
```

The options 1, 2, and 3, respectively, choose the 1st, 2nd, and 3rd codon positions. The f, s, and t of the output files refer to first, second, and third positions. By using NucST, we will examine the 2nd and 3rd positions, which show a sharp contrast. Enter

```
nucst pri5s.nuc > pri5s.nst
```

Then, the output file "pri5s.nst" appears as follows;

```
nucst 1.2.1  Jun 25 1996    5 OTUs 119 sites  mtDNA Primates

    Ts       1   2   3   4   5
Tv         Chi Hum Gor Ora Sia
1   Chimp  Chi   4   6  25  14
2   Human    0 Hum   6  25  12
3   Goril    3   3 Gor  24  14
4   Orang    3   3   6 Ora  22
5   Siama    2   2   5   3 Sia


               T      C      A      G     A+T    G+C    Bias   Skew
1   Chimp   0.471  0.261  0.134  0.134  0.605  0.395  0.101  0.462
2   Human   0.471  0.261  0.151  0.118  0.622  0.378  0.101  0.462
3   Goril   0.479  0.244  0.151  0.126  0.630  0.370  0.103  0.445
4   Orang   0.420  0.336  0.109  0.134  0.529  0.471  0.093  0.513
5   Siama   0.437  0.294  0.151  0.118  0.588  0.412  0.086  0.462
    mean    0.455  0.279  0.139  0.126  0.595  0.405  0.094  0.469

Bias  x1e3   1   2   3   4   5
           Chi Hum Gor Ora Sia
1   Chimp  Chi  17  25  76  50
2   Human   17 Hum  17  92  34
3   Goril   25  17 Gor 101  50
4   Orang   76  92 101 Ora  59
5   Siama   50  34  50  59 Sia
```

In the file "pri5s.nst", numbers of transition (Ts) and transversion (Tv) differences are given first in the upper-right half and in the lower-left half of a matrix, then nucleotide frequencies and distance of nucleotide composition ("bias" defined by Eq. 4.1 where f_{ik} is the frequency of the k-th nucleotide of OTU i) follow in this order.

In order to get a list of the alignment, the a-option of NucSt can be used by entering

```
nucst -a pri5s.nuc > pri5s.ali
```

Then, the alignment is given in pri5.ali as follows;

```
nucst 1.2.1  Jun 25 1996    5 OTUs 119 sites  mtDNA Primates

CONSENSUS  TTTTCGTGTG GCTTGTTCTG ATGTTGACTT GAACTTTTGT TTTTTACAGC CTTGCACTTT
Chimp      .......... .......... .......T.. .......... .......... .....G....
Human      .......... .......... .......T.. .......... ......T... ..C.......
Goril      .......... .......... .......... .........A. .......... ....A.....
Orang      .C.C...... .T.C...... .......... ...TC..... C..C...G.. ..........
Siama      .CCC...... .......... .......... .......... .......... ..C.......
                    10         20         30         40         50         60
CONSENSUS  TTCGCTATTC CTCCTGCTCC TCTCTCAATC CCTATAAAAC TTTTCTATCA CTCTACTTC
Chimp      .......... .......... .......... .......... .......... ....G....
Human      .......... .......... .......... .......... .......... .........
Goril      .......... .......... .......... T.....TG.. .......... .........G
Orang      .......... ....C.T... C......TC. .TC.CC.C.T .........G T.TC...C.
Siama      .......... .......... ...TCT..C. .T...C...A .C........ T........
                    70         80         90        100        110
```

Notice that the number of nucleotide substitutions is small in the 2nd codon positions, because a substitution in the 2nd position causes an amino acid substitution which tends to have a deleterious effect.

On the other hand, the numbers of substitutions and alignment at the 3rd positions are as follows;

```
nucst 1.2.1  Jun 25 1996    5 OTUs 119 sites  mtDNA Primates

    Ts       1   2   3   4   5
Tv         Chi Hum Gor Ora Sia
1   Chimp  Chi  26  29  38  37
2   Human    1 Hum  34  32  40
3   Goril    6   5 Gor  40  41
4   Orang   11  10   9 Ora  40
5   Siama   21  20  19  20 Sia

              T      C      A      G     A+T    G+C    Bias   Skew
1   Chimp   0.168  0.328  0.420  0.084  0.588  0.412  0.092  0.496
2   Human   0.151  0.353  0.403  0.092  0.555  0.445  0.092  0.513
3   Goril   0.185  0.328  0.412  0.076  0.597  0.403  0.089  0.479
4   Orang   0.126  0.445  0.353  0.076  0.479  0.521  0.126  0.597
5   Siama   0.185  0.370  0.345  0.101  0.529  0.471  0.066  0.429
    mean    0.163  0.365  0.387  0.086  0.550  0.450  0.088  0.503


Bias  x1e3   1   2   3   4   5
           Chi Hum Gor Ora Sia
1   Chimp  Chi  34  17 118  76
2   Human   34 Hum  42  92  59
3   Goril   17  42 Gor 118  67
4   Orang  118  92 118 Ora  84
5   Siama   76  59  67  84 Sia


nucst 1.2.1  Jun 25 1996    5 OTUs 119 sites  mtDNA Primates

CONSENSUS  AACACAAGAA ACCAACACCA CAAAAAT..T ACCCCA.ATC AACA.AACAC TCCCTCACCA
Chimp      .........G G.....G... TG.....GA. ..A..GTG.. .G..AG.... .T...T....
Human      .......A.. .T........ .GG....GG. ..TT..CG.. .G..A..... ..........
Goril      ..T......G ....G..... T......CG. .T....T... G...C..T.T ....C....G
Orang      .C...GG... .....T.... ...C..CTCC G.....C.CT ....T...G. C......TT.
Siama      CT...G.AG. GT...A...G ....GGCCAC .T..T.A... G...CGT.G. ..T..TG.T.
                  10         20         30         40         50         60
CONSENSUS  CTAACACTTT AGCCAGCACA AACCAACCAC ACAAACCCAA AACTTCTCCA AA.ACTAAC
Chimp      .......... .AT..A...C ......T... .......T.. ...C...... ..T......
Human      ...G...... .........C .......... .....TT... .G........ .GC......
Goril      .C......C. .......... ..T.T.T... .T.....T.. ..T....... ..T..CG.A
Orang      .C.G.G.C.C ..T....CT. .......... .TC....... T...C..... ..CC.....
Siama      ......A.C. CA..GATCA. ...A.G.... C..CT...CC ..T.CTC... ..A....TT
                  70         80         90        100        110
```

From these results, we can see that the 3rd positions are highly variable compared to the other positions, because many of the substitutions in the 3rd positions are synonymous (does not change amino acid).

4.1.5 NucML

Since the three positions in a codon evolve in different rates, it is recommended to analyze the data by taking account of this (e.g., Hasegawa and Adachi 1996[89]). In order to do this, enter

```
protml -topt -l pri5f pri5f.nuc pri5_user.tpl > pri5f.ml
protml -topt -l pri5s pri5s.nuc pri5_user.tpl > pri5s.ml
protml -topt -l pri5t pri5t.nuc pri5_user.tpl > pri5t.ml
```

where "-topt" means estimate the transition/transversion ratio (α/β in Eq. 2.12 of the HKY85 model) by maximizing the likelihood, and "-l pri5*" means that the estimated log-likelihoods of each site are stored in the "pri5*.lls" file which can be used in evaluating the total evidence of different codon positions and/or of different genes. In this example, we estimated optimal transition/transversion ratio for each tree topology. However, when the number of tree topologies is large, this causes a large computational burden. Since the optimal transition/transversion ratio does not appear to depend strongly on the tree

topology, if the optimal ratio is estimated once for a tree topology such as an NJ tree, this ratio might be used in comparing different tree topologies by using "-t estimated ratio" instead of "-topt". However, the ratio should be estimated separately for different codon positions.

The output file "pri5f.ml" appears as follows;

```
nucml 2.3b3 (07/12/96) A/B:opt F 5 OTUs 119 sites. mtDNA Primates
#1
Alpha/Beta: 10.377
                 :----1 Chimp
              :---6
              :   :----2 Human
:-------------7
:             :--------3 Goril
:
:---------------4 Orang
:
:-----------5 Siama

No.1       ext. branch S.E.  int. branch S.E.
Chimp        1   6.69  2.94    6   3.55  3.66
Human        2   6.55  2.96    7  23.49  7.86
Goril        3  16.28  5.02  TBL :     109.46  iter: 9
Orang        4  31.15  8.43  ln L:    -459.53 +- 20.68
Siama        5  21.75  7.32  AIC :     941.06
#2
Alpha/Beta: 10.865
                   :----2 Human
                :--6
                :  :----------3 Goril
:---------------7
:               :----1 Chimp
:
:---------------4 Orang
:
:-----------5 Siama

No.2       ext. branch S.E.  int. branch S.E.
Chimp        1   6.82  2.96    6  lower limit
Human        2   6.57  2.97    7  26.53  8.27
Goril        3  19.86  5.20  TBL :     113.66  iter: 5
Orang        4  31.87  8.68  ln L:    -459.87 +- 20.58
Siama        5  22.02  7.51  AIC :     941.74  lower limit: 0.001
#3
Alpha/Beta: 10.865
                   :----1 Chimp
                :--6
                :  :----------3 Goril
:---------------7
:               :----2 Human
:
:---------------4 Orang
:
:-----------5 Siama

No.3       ext. branch S.E.  int. branch S.E.
Chimp        1   6.82  2.96    6  lower limit
Human        2   6.57  2.97    7  26.52  8.27
Goril        3  19.86  5.20  TBL :     113.66  iter: 5
Orang        4  31.87  8.68  ln L:    -459.87 +- 20.58
Siama        5  22.02  7.51  AIC :     941.73  lower limit: 0.001

nucml 2.3b3 A/B:opt F 3 trees 5 OTUs 119 sites. mtDNA Primates

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-------------------------------------------------------------------
1        -459.5     0.0 <-best   11     941.1    0.0    ME   0.6320
2        -459.9    -0.3    0.9   11     941.7    0.7    4.2  0.1297
3        -459.9    -0.3    0.9   11     941.7    0.7    4.2  0.2383
```

where 'iter' indicates the times the program traversed the entire tree in estimating the branch lengths by the Newton-Raphson method.

The output file for the 2nd codon positions, "pri5s.ml", appears as follows;

```
nucml 2.3b3 (07/12/96) A/B:opt F 5 OTUs 119 sites. mtDNA Primates
#1
Alpha/Beta: 8.128
           :--1 Chimp
        :--6
        :  :--2 Human
:-------7
:       :------3 Goril
:
:------------------4 Orang
:
:-----5 Siama

No.1       ext. branch S.E.  int. branch S.E.
Chimp        1   2.05  1.43    6   0.50  0.96
Human        2   1.48  1.24    7   6.72  3.15
Goril        3   6.23  2.57  TBL :      43.47  iter: 5
Orang        4  20.60  5.33  ln L:    -320.31 +- 20.33
Siama        5   5.89  2.94  AIC :     662.62
#2
Alpha/Beta: 8.277
           :--2 Human
         :--6
         :  :------3 Goril
:--------7
:        :--1 Chimp
:
:------------------4 Orang
:
:-----5 Siama

No.2       ext. branch S.E.  int. branch S.E.
Chimp        1   2.09  1.45    6  lower limit
Human        2   1.51  1.26    7   7.36  3.23
Goril        3   6.72  2.63  TBL :      44.17  iter: 5
Orang        4  20.72  5.37  ln L:    -320.49 +- 20.29
Siama        5   5.79  2.96  AIC :     662.98  lower limit: 0.001
#3
Alpha/Beta: 8.232
           :--1 Chimp
         :--6
         :  :------3 Goril
:--------7
:        :-2 Human
:
:------------------4 Orang
:
:-----5 Siama

No.3       ext. branch S.E.  int. branch S.E.
Chimp        1   1.93  1.38    6   0.51  0.94
Human        2   1.12  1.12    7   7.43  3.22
Goril        3   6.60  2.59  TBL :      44.06  iter: 5
Orang        4  21.14  5.42  ln L:    -320.32 +- 20.27
Siama        5   5.34  2.86  AIC :     662.65

nucml 2.3b3 A/B:opt F 3 trees 5 OTUs 119 sites. mtDNA Primates

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-------------------------------------------------------------------
1        -320.3     0.0 <-best   11     662.6    0.0    ME   0.4099
2        -320.5    -0.2    0.6   11     663.0    0.4    0.7  0.1809
3        -320.3    -0.0    0.9   11     662.6    0.0    0.6  0.4092
```

The output file for the 3rd codon positions, "pri5t.ml", looks as follows;

```
nucml 2.3b3 (07/12/96) A/B:opt F 5 OTUs 119 sites. mtDNA Primates
#1
Alpha/Beta: 37.587
                :-----1 Chimp
          :---------6
          :         :------2 Human
:---------7
:         :------3 Goril
:
:------------------------4 Orang
:
:++++++++++++++++++++++++++-5 Siama

No.1       ext. branch S.E.  int. branch S.E.
Chimp        1  15.63  6.90    6  24.69 11.77
Human        2  17.02  7.02    7  19.76 21.35
Goril        3  19.24 10.67  TBL :     406.43  iter: 10
Orang        4  74.97 29.85  ln L:    -503.75 +- 16.38
Siama        5 235.12 65.23  AIC :    1029.50
#2
Alpha/Beta: 29.740
                    :------2 Human
                  :--6
                  :  :-------------3 Goril
:-----------------7
:                 :------1 Chimp
:
:-----------------4 Orang
:
:++++++++++++++++++++++++++-5 Siama

No.2       ext. branch S.E.  int. branch S.E.
Chimp        1  16.38  6.81    6  lower limit
Human        2  16.54  6.80    7  45.84 24.99
Goril        3  40.25 11.87  TBL :     363.81  iter: 9
Orang        4  50.54 25.36  ln L:    -510.11 +- 17.76
Siama        5 194.26 54.32  AIC :    1042.21  lower limit: 0.001
#3
Alpha/Beta: 29.356
                         :-----1 Chimp
                  :------6
                  :      :-------------3 Goril
:-----------------7
:                 :-2 Human
:
:---------------4 Orang
:
:++++++++++++++++++++++++++-5 Siama

No.3       ext. branch S.E.  int. branch S.E.
Chimp        1  14.49  6.56    6  14.69  6.92
Human        2   3.14  6.59    7  46.00 23.66
Goril        3  40.98 12.05  TBL :     360.15  iter: 50 just before convergence
Orang        4  47.91 23.94  ln L:    -509.98 +- 17.80
Siama        5 192.94 53.65  AIC :    1041.97

nucml 2.3b3 A/B:opt F 3 trees 5 OTUs 119 sites. mtDNA Primates

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-------------------------------------------------------------------
1        -503.8     0.0 <-best   11    1029.5    0.0   46.3  0.9267
2        -510.1    -6.4    4.4   11    1042.2   12.7    3.7  0.0201
3        -510.0    -6.2    4.5   11    1042.0   12.5    ME   0.0532
```

In the last table, TBLs (total branch length) may look strange. Although tree-1 is the best tree by the likelihood criterion, the TBL of tree-1 is much larger than that of tree-3. This is because a much larger α/β ratio was assigned to tree-1 than to trees-2 and 3. The likelihood is not sensitive to the α/β ratio with this data set, and therefore the variance of the estimate of this ratio is very large. Indeed, fixing the α/β ratio at 37.59 does not change the result of the ML analysis as shown below:

```
nucml 2.3b3 A/B:37.59 F 3 trees 5 OTUs 119 sites. mtDNA Primates

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-----------------------------------------------------------------
1        -503.8     0.0 <-best   11    1029.5    0.0    ME   0.9271
2        -510.3    -6.5    4.6   11    1042.5   13.0   28.4  0.0165
3        -510.2    -6.5    4.7   11    1042.4   12.9   25.4  0.0564
```

4.1.6 TotalML

From these results, it is clear that the rate and α/β ratio differ very much among the different codon positions, for which log-likelihoods were estimated separately. The likelihood is the probability that one tree yielded the observed data, and we assume that each codon position evolves independently from other sites. Therefore, the total support for a particular tree can be evaluated by simply summing up the estimated log-likelihoods of the three different codon positions for that tree, and the total log-likelihoods for different trees can then be compared (section 5.4). We can evaluate the total evidence of this protein-encoding data with the "TotalML" program by entering;

```
totalml pri5f.lls pri5s.lls pri5t.lls > pri5.total
```

Then, the "pri5.total" files appears as follows;

```
totalml 1.1(07/12/96) 3 data sets, 357 sites. nucml 2.3b3

tree      1       2       3     total
1      459.5   320.3   503.8   1283.6
          ml      ml      ml       ML
2         0.3     0.2     6.4      6.9
          0.9     0.6     4.3      4.5
3         0.3     0.0     6.2      6.6
          0.9     0.9     4.4      4.7
sites     119     119     119      357

tree      1       2       3      total
1      0.6417  0.4158  0.9263   0.9290
2      0.1229  0.1770  0.0214   0.0162
3      0.2354  0.4072  0.0523   0.0548
```

The 1st, 2nd, and 3rd columns refer to the 1st, 2nd, and 3rd codon positions, "ml" refers to the ML tree topology (for which the estimated negative log-likelihood is given), and for the other tree topologies the differences of log-likelihood from the ML tree are given with their SEs immediately below. In the "total" column, the ML tree is indicated by "ML". Furthermore, bootstrap probabilities (BP) estimated by the RELL method are given for each codon position and for the total.

4.2 ProtML: Maximum Likelihood Inference of Protein Phylogeny

ProtML is a C program for inferring evolutionary trees from protein (amino acid) sequences using the ML method (Kishino et al. 1990[148]). It does not impose any constraint on the constancy of evolutionary rate among lineages.

Features in which the ProtML differs from the DNAML of PHYLIP (up to version 3.4) are as follows:

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

2) Likelihood of multifurcating trees can be estimated. When the information contained by the data is not sufficient to solve branching order, it is preferable to be satisfied with a tree containing multifurcations (e.g., Czelusniak et al. 1990[53]). This is because completely resolved bifurcating trees obtained by using insufficient amount of data could be misleading.

3) Novel methods of topology search ("star decomposition" and "local rearrangement") are adoted.

4) An approximate likelihood method can be used to screen topologies.

5) The Newton-Raphson method is adoted in maximizing likelihood.

6) Bootstrap probabilities of candidate trees can be estimated quickly by using the RELL method (Kishino et al. 1990[148]; Hasegawa and Kishino 1994[97]).

4.2.1 Options

The program allows various options as shown below using switches "-x" in the command line.

```
ProtML 2.3 Maximum Likelihood Inference of Protein Phylogeny
Copyright (C) 1992-1996 J. Adachi & M. Hasegawa. All rights reserved.
Usage: protml [switches] sequence_file [topology_file] > [output_file]
sequence_file = MOLPHY_format | Sequential(-S) | Interleaved(-I)
topology_file = users_trees(-u) | constrained_tree(-e)
Model:
                    -mf  mtREV-F       Adachi & Hasegawa (1996)
-j  JTT (default)   -jf  JTT-F         Jones, Taylor & Thornton (1992)
-d  Dayhoff         -df  Dayhoff-F     Dayhoff et al. (1978)
-p  Poisson         -pf  Proportional  Felsenstein (1981)
-r  users RTF       -rf  users RTF-F   (Relative Transition Frequencies)
-f  with data Frequencies
Search strategy or Mode:
-u  Users trees (requires users_tree file)
-e  Exhaustive search (with/without constrained_tree file)
-R  Local rearrangement search (need starting_tree file; may not result in the ML tree))
-s  Star decomposition search (may not result in the ML tree)
-q  Quick add OTUs search (may not result in the ML tree)
-D  maximum likelihood Distance matrix --> NJDIST
Others:
-n  number of retained top ranking trees by Approx.likelihood(default -e:105,-q:50)
-b  no Bootstrap probabilities (when user trees supplied)
-S  Sequential format   -I  Interleaved format
```

This program has six modes of topology search as shown below; i.e., User tree (manual) mode, Exhaustive search mode, Local rearrangement search mode, Star decomposition search mode, Quick add OTUs search mode and maximum likelihood Distance matrix mode (this last one to be combined with NJdist).

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

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

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

3) "**-R**" : Local rearrangement search mode

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

Unless specified, it starts with a star-like tree.

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

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

The program also has another option;

"**-b**" : no bootstrap option

If the no bootstrap option is not specified, approximate bootstrap probabilities of candidate trees are estimated by the resampling of estimated log-likelihood (RELL) method (Kishino et al. 1990[148]; Hasegawa and Kishino, 1994[97]).

4.2.2 Format of Input Sequences File

MOLPHY Format

A 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; SIZE then 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 comments on the data after the two digits numbers, which are separated by blanks.

SEQUENCES

The following lines of the input file 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 then put the amino acid sequence AFTER a NEWLINE in free format. Separation by whitespace(space, tab or newline) is allowed. The amino acids must be specified by the one letter code (IUPAC-IUB Commission on Biochemical Nomenclature 1968[127]).

SEQUENTIAL Format

Felsenstein's PHYLIP "SEQUENTIAL" format is:

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

The information for each species starts with a TEN-CHARACTER species name (which CAN include punctuation marks and blanks). To run such a file, a user must use SEQUENTIAL FILE, the "-S" Switch, as follows;

```
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, "$" represents newline (or 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 the "-I" Switch as follows;

```
protml -I INTERLEAVED FILE
```

Format of USER TREES File

standard USER TREES file format:

```
3 hominoids
(((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 a comment of the trees after the first number, separated by blanks (space or tab).

MACHINE READABLE TREES

The following lines give sets of (user-defined) machine readable trees. 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 may put the next machine readable tree AFTER a NEWLINE in free format, i.e., separations by whitespace (space, tab or newline) are allowed, for example,

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

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

That is, the above two machine readable tree are the same.

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

```
Unrooted tree (ProtML & NJdist)
    variable rate

( subtree1, subtree2, subtree3 );

    :-----subtree1
    :
    :-----subtree2
    :
    :-----subtree3

    ^provisional root
```

```
Rooted tree (not presently supported)
    constant rate

( subtree1, subtree2 );


    :-----subtree1
    :
    :-----subtree2

    ^root
```

Format of a CONSTRAINT TREE File

standard CONSTRAINT TREE file format:

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

A CONSTRAINT TREE file allows a constrained machine readable tree. A pair of PARENTHESES indicates FIX tree structure, but a pair of BRACES indicates COMBINATION tree structure in a monophyletic group. That is, all branching orders consistent with the group in braces may be considered.

To command of ProtML with the "-e" switch, e.g.,

```
protml -e sequence_file constrained_tree
```

generates 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);
```

where the order of tree topologies is according to the order of approximate likelihood (section3.5). If the number of possible trees exceeds 105, only the best 105 trees by the approximate likelihood criterion are retained. If more tree topologies are needed (say 1000), use the following command;

```
protml -e -n 1000 sequence_file constrained_tree > tree.tpl
```

Then, the best 1000 tree topologies by the approximate likelihood criterion are stored in the tree.tpl file, and can be used in the full likelihood analysis.

4.3 NucML: Maximum Likelihood Inference of Nucleic Acid Phylogeny

NucML is a C program for inferring evolutionary trees from nucleotide sequences by using the ML method.

4.3.1 Options

NucML has several options as shown below;

```
NucML 2.3 Maximum Likelihood Inference of Nucleic Acid Phylogeny
Copyright (C) 1992-1996 J. Adachi & M. Hasegawa. All rights reserved.
Usage: nucml [switches] sequence_file [topology_file] > [output_file]
sequence_file = MOLPHY_format | Sequential(-S) | Interleaved(-I)
topology_file = user_trees(-u) | constraint_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  with equal base frequencies
Search strategy or Mode:
-u  User trees (need user_trees file)
-e  Exhaustive search (with/without constraint_tree file)
-R  Local rearrangement search (need starting_tree file; may not be the ML tree)
-s  Star decomposition search (may not give the ML tree)
-q  Quick add OTUs search (may not give the ML tree)
-D  maximum likelihood Distance matrix --> NJdist
Others:
-n num of retained top ranking trees win Approx.likelihood(default -e:105,-q:50)
-b  no Bootstrap probabilities (with User trees)
-S  Sequential format   -I  Interleaved format
```

4.4 ProtST: Basic Statistics of Protein Sequences

4.4.1 Options

ProtST has several options as follows;

```
ProtST 1.2 Basic Statistics of Protein Sequences
Copyright (C) 1993-1996 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)
```

4.4.2 Output Format

An example of the output of ProtST is shown below;

```
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
```

Bias refers to the distance of amino acid composition between OTUs i and j defined by Eq. 4.1 (see subsection 4.1.2).

4.5 NucST: Basic Statistics of Nucleic Acid Sequences

4.5.1 Options

NucST has several options as follows;

```
NucST 1.2 Basic Statistics of Nucleic Acid Sequences
Copyright (C) 1993-1996 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)
```

4.5.2 Output Format

An example of the output of NucST is shown below;

```
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 x10e3   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
```

Distance of nucleotide composition ('Bias' distance) is defined by Eq. 4.1 where f_{ik} is the frequency of the k-th nucleotide of OTU i.

4.6 NJdist: Neighbor Joining Phylogeny from Distance Matrix

4.6.1 Options

NJdist is a program for inferring a tree from a distance matrix by the neighbor-joining method (Saitou and Nei 1987[221]), and has several options as follows;

```
NJdist 1.3 Neighbor Joining Phylogeny from Distance Matrix
Copyright (C) 1993-1996 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      output of branch length
-l      Least squares estimate of branch length
-o num  branch number of Outgroup (rooting the tree)
-t str  output Tree file name
```

4.6.2 Input Format

An input file of the distance matrix (named "njdist.dis") for the NJdist program appears as follows;

```
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
```

4.6.3 Output Format

Enter

```
njdist njdist.dis > njdist.out
```

Then, the output file "njdist.out" appears as follows;

```
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);
```

4.7 Utilities (Sequence Manipulations) with Perl

Several utilities for sequence manipulations are provided with MOLPHY as listed below;

Conversion of a file between MOLPHY format and formats for other softwares including; Clustal (Higgins et al. 1992[114]), MacClade (Maddison and Maddison 1992[177]; Nexus which is same as PAUP, Swofford 1993[239]), MEGA (Kumar et al. 1993[162]), MUST (Philippe[206]), PHYLIP (Felsenstein[69])

```
clus2mol: Clustal format -> MOLPHY
mc2mol:   MacClade format -> MOLPHY
mega2mol: MEGA format -> MOLPHY
must2mol: MUST format -> MOLPHY
int2mol:  Interleaved format -> MOLPHY
mol2int:  MOLPHY format -> Interleaved
phy2mol:  Sequential format -> MOLPHY
mol2phy:  MOLPHY format -> Sequential
```

Format conversion for sequence manipulation

```
mol2inf:  MOLPHY format -> Inf format
inf2mol:  Inf format -> MOLPHY format
mol2seq:  MOLPHY format -> Seq format
seq2mol:  Seq format -> MOLPHY format
ali2mol:  Ali format -> MOLPHY format
```

Triming of MOLPHY format

```
mol2mol:  MOLPHY format -> standard MOLPHY format (MOLPHY format beautifer)
nuc2NUC:  small letters for nucleotides -> capitals
```

Manipulation of MOLPHY format

```
degene4:  sampling of four-fold degenerate sites
infocode: sampling of codons which have experienced substitution
molcodon: sampling of specified codon positions
molcons:  consensus sequence with decision by majority
molinfo:  sampling of sites which have experienced substitution
mollist:  get identifiers list
molrev:   get complementary sequence of nucleotides
nuc2code: punctuate nucleotide sequence by a blank between codons
nuc2ptn:  translate nucleotide sequences into amino acid sequences
rmid3:    remove codons which contain ins/del sites
rminsdel: remove ins/del sites
molcat:   concatenate sequences of different genes in different files of the same
          set of OTUs
molcut:   extract specified partial sequences
molmerge: merge sequences of different OTUs in different files but for the same gene
molsplit: split sequence data into different files for each OTU
```

Extract sequence data from database

```
egetcds:  extract cds (coding) region from EMBL file
ggetcds:  extract cds (coding) region from Genbank file
```

Chapter 5

Applications to Biological Problems

5.1 Cytochrome b

Cytochrome *b* is one of the most widely used molecular markers in phylogenetic studies of animals. In this section, we will study several phylogenetic problems for vertebrates using this molecule.

5.1.1 Sequence Data

Sequence data used in the phylogenetic analyses are listed below, where the classification is based on traditional taxonomy (Corbet and Hill 1991[51]; Yamashina 1986[267]).

	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.)[141]	D34635
	Bosja	*Bos javanicus*	Banteng	Kikkawa (unpubl.)[141]	D34636
	Bubbu1	*Bubalus bubalis*	Asian water buffalo	Kikkawa (unpubl.)[142]	D34637
	Bubbu2	*Bubalus bubalis*	Asian water buffalo	Kikkawa (unpubl.)[142]	D34638
	Budtb	*Budorcas taxicolor bedfordi*	Golden takin	Groves (unpubl.)[88]	U17867
	Budtt	*Budorcas taxicolor taxicolor*	Mishmi takin	Groves (unpubl.)[88]	U17868
	Capcr	*Capricornis crispus*	Japanese serow	Chikuni'94[47]	D32191
	Nemca	*Nemorhaedus caudatus*	Chinese goral	Groves (unpubl.)[88]	U17861
	Ovimo	*Ovibos moschatus moschatus*	Muskox	Groves (unpubl.)[88]	U17862
	Oviar	*Ovis aries*	Domestic sheep	Irwin'91[126]	X56284
	Caphi	*Capra hircus*	Domestic goat	Irwin'91[126]	X56289
	Cerni	*Cervus nippon*	Sika deer	Chikuni'94[47]	D32192
	Odohe	*Odocoileus hemionus*	Black-tailed deer	Irwin'91[126]	X56291
	Damda	*Dama dama*	Fallow deer	Irwin'91[126]	X56290
	Girca	*Giraffa camelopardalis*	Giraffe	Irwin'91[126]	X56287
	Antam	*Antilocapra americana*	Pronghorn	Irwin'91[126]	X56286
	Trana	*Tragulus napu*	Greater Malay chevrotain	Irwin'91[126]	X56288
	Traja	*Tragulus javanicus*	Lesser Malay chevrotain	Chikuni (unpubl.)[46]	D32189
	Camdr1	*Camelus dromedarius*	One-humped camel	Irwin'91[126]	X56281
	Camdr2	*Camelus dromedarius*	One-humped camel	Stanley'94[232]	U06426
	Camba	*Camelus bactrianus*	Two-humped camel	Stanley'94[232]	U06427
	Lamgu	*Lama guanicoe*	Guanaco	Stanley'94[232]	U06428
	Lamgl	*Lama glama*	Llama	Stanley'94[232]	U06429
	Lampa	*Lama pacos*	Alpaca	Stanley'94[232]	U06425
	Vicvi	*Vicugna vicugna*	Vicuna	Stanley'94[232]	U06430
	Hipam	*Hippopotamus amphibius*	Hippopotamus	Irwin'94[125]	U07565
	Tayta	*Tayassu tajacu*	Collared peccary	Irwin'91[126]	X56296

Sussc	*Sus scrofa*	Pig	Irwin'91[126]	X56295
I-2. Cetacea				
Stelo	*Stenella longirostris*	Long-beaked dolphin	Irwin'91[126]	X56293
Steat	*Stenella attenuata*	Narrow-snouted dolphin	Irwin'91[126]	X56294
Phyma	*Physeter macrocephalus*	Sperm whale	Arnason'94[20]	X75589
Balph	*Balaenoptera physalus*	Fin whale	Arnason'91[23]	X61145
Balmu	*Balaenoptera musculus*	Blue whale	Arnason'93[19]	X72204
Balac	*Balaenoptera acutorostrata*	Minke whale	Arnason'94[20]	X75753
Balbon	*Balaenoptera bonaerensis*	Antarctic minke whale	Arnason'94[20]	X75581
Balbor	*Balaenoptera borealis*	Sei whale	Arnason'94[20]	X75582
Baled	*Balaenoptera edeni*	Bryde's whale	Arnason'94[20]	X75583
Megno	*Megaptera novaeangliae*	Humpback whale	Arnason'94[20]	X75584
Escro	*Eschrichtius robustus*	California gray whale	Arnason'94[20]	X75585
Balmy	*Balaena mysticetus*	Bowhead whale	Arnason'94[20]	X75588
Balgl	*Balaena glacialis*	Right whale	Arnason'94[20]	X75587
Capma	*Caperea marginata*	Pygmy right whale	Arnason'94[20]	X75586
I-3. Pinnipedia				
Phovi1	*Phoca vitulina*	Harbor seal	Arnason'92[24]	X63726
Phovi2	*Phoca vitulina*	Harbor seal	Arnason'95[18]	X82306
Phofa	*Phoca fasciata*	Ribbon seal	Arnason'95[18]	X82302
Phola	*Phoca largha*	Spotted seal	Arnason'95[18]	X82305
Phohi	*Phoca hispida*	Ringed seal	Arnason'95[18]	X82304
Phogr	*Phoca groenlandica*	Harp seal	Arnason'95[18]	X82303
Halgr	*Halichoerus grypus*	Grey seal	Arnason'93[22]	X72004
Eriba	*Erignathus barbatus*	Bearded seal	Arnason'95[18]	X82295
Hydle	*Hydrurga leptonyx*	Leopard seal	Arnason'95[18]	X82297
Monsc	*Monachus schauinslandi*	Hawaiian monk seal	Arnason'95[18]	X72209
Cyscr	*Cystophora cristata*	Hooded seal	Arnason'95[18]	X82294
Mirle	*Mirounga leonina*	Southern elephant seal	Arnason'95[18]	X82298
Arcga	*Arctocephalus gazella*	Antarctic fur seal	Arnason'95[18]	X82292
Arcfo	*Arctocephalus forsteri*	New Zealand fur seal	Arnason'95[18]	X82293
Zalca	*Zalophus californianus*	California sea lion	Arnason'95[18]	X82310
Eumju	*Eumetopias jubatus*	Northern sea lion	Arnason'95[18]	X82311
Odoro	*Odobenus rosmarus rosmarus*	Atlantic walrus	Arnason'95[18]	X82299
I-4. Carnivora				
Ursam	*Ursus americanus*	American black bear	Arnason'95[18]	X82307
Ursar	*Ursus arctos*	Brown bear	Arnason'95[18]	X82308
Ursma	*Ursus maritimus*	Polar bear	Arnason'95[18]	X82309
Feldo	*Felis domesticus*	Domestic cat	Arnason'95[18]	X82296
Panle	*Panthera leo*	Lion	Arnason'95[18]	X82300
Panti	*Panthera tigris*	Tiger	Arnason'95[18]	X82301
I-5. Perissodactyla				
Equca	*Equus caballus*	Domestic horse	Xu'94[265]	X79547
Equgr	*Equus grevyi*	Grevy's zebra	Irwin'91[126]	X56282
Dicbi	*Diceros bicornis*	Black rhinoceros	Irwin'91[126]	X56283
I-6. Rodentia				
Musmu	*Mus musculus*	House mouse	Bibb'81[35]	P00158
Ratno	*Rattus norvegicus*	Common rat	Gadaleta'89[73]	P00159
Papbu	*Pappogeomys bulleri*	Buller's pocket gopher	DeWalt'93[58]	L11900
Geobu	*Geomys bursarius juggosicularis*	Plains pocket gopher	DeWalt'93[58]	L11901
Craca	*Cratogeomys castanops castanops*	Yellow-faced pocket gopher	DeWalt'93[58]	L11902
Crafu	*Cratogeomys fumosus*	Smoky pocket gopher	DeWalt'93[58]	L11903
Crago	*Cratogeomys goldmani goldmani*	Goldman's pocket gopher	DeWalt'93[58]	L11904
Cragy	*Cratogeomys gymnurus*	Llano pocket gopher	DeWalt'93[58]	L11905
Crame	*Cratogeomys merriami*	Merriam's pocket gopher	DeWalt'93[58]	L11906
Craru	*Cratogeomys goldmani rubellus*		DeWalt'93[58]	L11907
Crata	*Cratogeomys castanops tamaulipensis*		DeWalt'93[58]	L11908
Craty	*Cratogeomys tylorhinus*	Taylor's pocket gopher	DeWalt'93[58]	L11909

Scini	*Sciurus niger*	Eastern fox squirrel	Wettstein'95[261]	U10180
Sciab	*Sciurus aberti*	Abert squirrel	Wettstein'95[261]	U10163
Speri	*Spermophilus richardsonii*	Richardson's ground squirrel	Thomas'93[248]	S73150
Hysaf	*Hystrix africaeaaustralis*	African porcupine	Ma'93[176]	X70674
Cavpo	*Cavia porcellus*	Guinea pig	Ma'93[176]	
I-7. Lagomorpha				
Orycu	*Oryctolagus cuniculus*	Rabbit	Irwin'94[125]	U07566
I-8. Proboscidea				
Loxaf	*Loxodonta africana*	African elephant	Irwin'91[126]	X56285
I-9. Sirenia				
Dugdu	*Dugong dugong*	Dugong	Irwin'94[125]	U07564
I-10. Primates				
Europ	*Homo sapiens*	European	Anderson'81[15]	J01415
Japan	*Homo sapiens*	Japanese (DCM1)	Ozawa'91[203]	
Afric	*Homo sapiens*	African (SB17F)	Horai'95[117]	D38112
Pantr	*Pan troglodytes*	Chimpanzee	Horai'95[117]	D38113
Panpa	*Pan paniscus*	Bonobo	Horai'95[117]	D38116
Gorgo	*Gorilla gorilla*	Gorilla	Horai'95[117]	D38114
Ponpy	*Pongo pygmaeus*	Orangutan	Horai'95[117]	D38115
I-11. Chiroptera				
Chido	*Chiroderma doriae*		Baker'95[29]	L28937
Chiim	*Chiroderma improvisum*	Guadeloupe white-lined bat	Baker'95[29]	L28938
Chisa	*Chiroderma salvini*	Salvin's white-lined bat	Baker'95[29]	L28939
Chitr	*Chiroderma trinitatum*	Goodwin's bat	Baker'95[29]	L28942
Chivi	*Chiroderma villosum*	Shaggy-haired bat	Baker'95[29]	L28943
Plahe	*Platyrrhinus helleri*	Heller's broad-nosed bat	Baker'95[29]	L28940
Urobi	*Uroderma bilobatum*	Tent-building bat	Baker'95[29]	L28941
I-12. Marsupialia				
Didvi	*Didelphis virginiana*	North American opossum	Janke'94[129]	Z29573
Mondo	*Monodelphis domestica*	South American opossum	Ma'93[176]	X70673
Plama	*Planigale maculata sinualis*	Common planigale	Painter (unpubl.)[204]	U10318
Plain	*Planigale ingrami*	Long-tailed planigale	Painter (unpubl.)[204]	U10319
Plate	*Planigale tenuirostris*	Narrow-nosed planigale	Krajewski'94[156]	U07591
Plagi	*Planigale gilesi*	Paucident planigale	Krajewski'94[156]	U07589
Smimu	*Sminthopsis murina*	Dunnart	Krajewski'94[156]	U07594
II. Class Aves				
II-1. Galliformes				
Galga	*Gallus gallus*	Chicken	Desjardins'90[57]	P18946
Cotco	*Coturnix coturnix*	Japanese quail	Kornegay'93[154]	L08377
Alech	*Alectoris chukar*	Chukar partridge	Kornegay'93[154]	L08378
Pavcr	*Pavo cristatus*	Peafowl	Kornegay'93[154]	L08379
Lopny	*Lophura nycthemera*	Silver pheasant	Kornegay'93[154]	L08380
Melga	*Meleagris gallopavo*	Turkey	Kornegay'93[154]	L08381
Lopga	*Lophortyx gambelii*	Gambel quail	Kornegay'93[154]	L08382
Numme	*Numida meleagris*	Guinea fowl	Kornegay'93[154]	L08383
Ortve	*Ortalis vetula*	Chachalaca	Kornegay'93[154]	L08384
II-2. Anseriformes				
Caimo	*Cairina moschata*	Muscovy duck	Kornegay'93[154]	L08385
II-3. Gruiformes				
Gruru1	*Grus rubicunda*	Brolga	Krajewski'94[155]	U11062
Gruru2	*Grus rubicunda*	Brolga	Leeton'94[169]	U13622
Gruja	*Grus japonensis*	Manchurian crane	Krajewski'94[155]	U11063
Gruan	*Grus antigone*	Sarus crane	Krajewski'94[155]	U11064
Gruvi	*Grus vipio*	White-naped crane	Krajewski'94[155]	U11065
II-4. Psittaciformes				
Calba	*Calyptorhynchus banksii*	Red-tailed black-cockatoo	Leeton'94[169]	U13620
Geooc	*Geopsittacus occidentalis*	Night parrot	Leeton'94[169]	U13621
Melun	*Melopsittacus undulatus*	Budgeriger	Leeton'94[169]	U13623
Pezwa	*Pezoporus wallicus*	Ground parrot	Leeton'94[169]	U13625
Plaix	*Platycercus icterotis xanthogenis*	Western rosella	Leeton'94[169]	U13626
Polan	*Polytelis anthopeplus*	Regent parrot	Leeton'94[169]	U13627

	westralis			
Strha	*Strigops habroptilis*	Kakapo	Leeton'94[169]	U13628
II-5. Piciformes				
Colru	*Colaptes rupicola*	Andean flicker	Edwards'91[60]	X60949
II-6. Passeriformes				
Empmi	*Empidonax minimus*	Least flycatcher	Helm-Bychowski'93[111]	X74251
Scyma	*Scytalopus magellanicus*	Andean tapaculo	Edwards'91[60]	X60945
Thrdo	*Thripophaga dorbignyi*	Creamy-breasted canastero	Edwards'91[60]	X60946
Ampst	*Ampelion stresemanni*	White-cheeked cotinga	Edwards'91[60]	X60947
Pitso	*Pitta sordida*	Hooded pitta	Edwards'91[60]	X60948
Pomte	*Pomatostomus temporalis*	Grey-crowned babbler	Edwards'91[60]	X60936
Pomru	*Pomatostomus ruficeps*	Chestnut-crowned babbler	Edwards'91[60]	X60937
Pomis	*Pomatostomus isidori*	Rufous babbler	Edwards'91[60]	X60938
Ambma	*Amblyornis macgregoriae*	MacGregor's bowerbird	Edwards'91[60]	X60940
Epial	*Epimachus albertisii*	Buff-tailed sicklebill	Edwards'91[60]	X60941
Ptipl	*Ptiloprora plumbea*	Leaden honeyeater	Edwards'91[60]	X60943
Gymti	*Gymnorhina tibicen*	Australian magpie	Edwards'91[60]	X60942
Parin	*Parus inornatus*	Plain titmouse	Edwards'91[60]	X60944
Catgu1	*Catharus guttatus*	Hermit thrush	Edwards'91[60]	X60939
Catgu2	*Catharus guttatus*	Hermit thrush	Helm-Bychowski'93[111]	X74261
Ailme	*Ailuroedus melanotus*	Spotted catbird	Helm-Bychowski'93[111]	X74257
Cyacr	*Cyanocitta cristata*	Blue jay	Helm-Bychowski'93[111]	X74258
Dipma	*Diphyllodes magnificus*	Magnificent bird of paradise	Helm-Bychowski'93[111]	X74255
Epifa	*Epimachus fastuosus*	Black sicklebill	Helm-Bychowski'93[111]	X74253
Lanlu	*Lanius ludovicianus*	Loggerhead shrike	Helm-Bychowski'93[111]	X74259
Manke	*Manucodia keraudrenii*	Trumpet bird	Helm-Bychowski'93[111]	X74252
Ptipa	*Ptiloris paradiseus*	Paradise riflebird	Helm-Bychowski'93[111]	X74254
Ptivi	*Ptilonorhynchus violaceus*	Satin bowerbird	Helm-Bychowski'93[111]	X74256
Virol	*Vireo olivaceus*	Red-eyed vireo	Helm-Bychowski'93[111]	X74260
II-7. Falconiformes				
Tortr	*Torgos tracheliotus*	Lappet-faced vulture	Avise'94[27]	U08934
Neope	*Neophron percnopterus*	Egyptian vulture	Avise'94[27]	U08942
Gypba	*Gypaetus barbatus*	Lammergeier	Avise'94[27]	U08943
Vulgr	*Vultur gryphus*	Andean condor	Avise'94[27]	U08944
Catbu	*Cathartes burrovianus*	Lesser yellow-headed vulture	Avise'94[27]	U08945
Corat	*Coragyps atratus*	Black vulture	Avise'94[27]	U08946
Gymca	*Gymnogyps californianus*	California condor	Avise'94[27]	U08947
II-8. Ciconiiformes				
Scoum	*Scopus umbretta*	Hammerkop	Avise'94[27]	U08936
Balre	*Balaeniceps rex*	Whale-headed stork	Avise'94[27]	U08937
Mycib	*Mycteria ibis*	Yellow-billed stork	Avise'94[27]	U08948
Mycam	*Mycteria americana*	American wood ibis	Avise'94[27]	U08949
Lepcr	*Leptoptilos crumeniferus*	Marabou stork	Avise'94[27]	U08950
Jabmy	*Jabiru mycteria*	Jabiry	Avise'94[27]	U08951
Plaal	*Platalea alba*	African spoonbill	Avise'94[27]	U08941
II-9. Pelecaniformes				
Peler	*Pelecanus erythrorhynchus*	American white pelican	Avise'94[27]	U08938
II-10. Phoenicopteriformes				
Phoru	*Phoenicopterus ruber*	Greater flamingo	Avise'94[27]	U08940
II-11. Cuculiformes				
Cocam	*Coccyzus americanus*	Yellow-billed cuckoo	Avise'94[28]	U09265
Cocer	*Coccyzus erythropthalmus*	Black-billed cuckoo	Avise'94[28]	U09266
Crosu	*Crotophaga sulcirostris*	Groove-billed ani	Avise'94[28]	U09260
Cucpa	*Cuculus pallidus*	Pallid cuckoo	Avise'94[28]	U09262
Piaca	*Piaya cayana*	Squirrel cuckoo	Avise'94[28]	U09263
Phacu	*Phaenicophaeus curvirostris*		Avise'94[28]	U09264
II-12. Opisthocomiformes				
OpihoA	*Opisthocomus hoazin*	Hoatzin	Avise'94[28]	U09257
OpihoB	*Opisthocomus hoazin*	Hoatzin	Avise'94[28]	U09258
OpihoC	*Opisthocomus hoazin*	Hoatzin	Avise'94[28]	U09259
III. Class Amphibia				
Xenla	*Xenopus laevis*	Clawed frog	Roe'85[216]	X02890
IV. Class Osteichthyes (Bony fishes)				

IV-1. Cypriniformes				
Cypca	*Cyprinus carpio*	Carp	Chang'94[45]	X61010
Lytat	*Lythrurus atrapiculus*	Blacktip shiner	Schmidt'95[224]	U17271
Lytar	*Lythrurus ardens*	Rosefin shiner	Schmidt'95[224]	U17268
Lytfu	*Lythrurus fumeus*	Ribbon shiner	Schmidt'95[224]	U17269
Lytli	*Lythrurus lirus*	Mountain shiner	Schmidt'95[224]	U17273
Lytsn	*Lythrurus snelsoni*	Ouchita mountain shiner	Schmidt'95[224]	U17272
Lytum	*Lythrurus umbratilis*	Redfin shiner	Schmidt'95[224]	U17274
Opsem	*Opsopoeodus emilae*	Pugnose minnow	Schmidt'95[224]	U17270
Crola	*Crossostoma lacustre*	Oriental stream loach	Tzeng'92[252]	M91245
IV-2. Salmoniformes				
Oncmy	*Oncorhynchus mykiss*	Rainbow trout	Zardoya'95[275]	L29771
IV-3. Perciformes				
Sarsa	*Sarda sarda*	Atlantic bonito	Cantatore'94[39]	X81562
Thuth	*Thunnus thynnus*	Albacore	Cantatore'94[39]	X81563
Scosc	*Scomber scombrus*	Atlantic mackerel	Cantatore'94[39]	X81564
Oremo	*Oreochromis mossambicus*		Cantatore'94[39]	X81565
Dicla	*Dicentrarchus labrax*	European seabass	Cantatore'94[39]	X81566
Boobo	*Boops boops*		Cantatore'94[39]	X81567
Tratr	*Trachurus trachurus*	Horse mackerel	Cantatore'94[39]	X81568
IV-4. Gadiformes				
Gadmo	*Gadus morhua*	Atlantic cod	Johansen'94[131]	X76365
IV-5. Acipenseriformes				
Acitr	*Acipenser transmontanus*	White sturgeon	Brown'89[37]	X14944
V. Class Chondrichthyes (Cartilaginous fishes)				
V-1. Carcharhiniformes				
Carpl	*Carcharhinus plumbeus*	Sandbar shark	Martin'93[181]	L08032
Carpo	*Carcharhinus porosus*	Smalltail shark	Martin'93[181]	L08033
Prigl	*Prionace glauca*	Blue shark	Martin'93[181]	L08040
Negbr	*Negaprion brevirostris*	Lemon shark	Martin'93[181]	L08039
Sphtive	*Sphyrna tiburo vespertina*	Pacific bonnethead	Martin'93[181]	L08043
Sphtiti	*Sphyrna tiburo tiburo*	Atlantic bonnethead	Martin'93[181]	L08042
Sphle	*Sphyrna lewini*	Scalloped hammerhead	Martin'93[181]	L08041
Galcu	*Galeocerdo cuvier*	Tiger shark	Martin'93[181]	L08034
V-2. Lamniformes				
Carca	*Carcharhodon carcharias*	White shark	Martin'93[181]	L08031
Isuox	*Isurus oxyrhynchus*	Shortfin mako	Martin'93[181]	L08036
Isupa	*Isurus paucus*	Longfin mako	Martin'93[181]	L08037
Lamna	*Lamna nasus*	Porbeagle	Martin'93[181]	L08038
V-3. Heterodontiformes				
Hetfr	*Heterodontus francisci*	Horn shark	Martin'93[181]	L08035
VI. Class Agnatha				
VI-1. Petromyzontiformes				
Petma	*Petromyzon marinus*	Sea lamprey	Lee'95[168]	U11880

The alignment of the cytochrome *b* sequences is shown in Figs. 5.1 (mammals) and 5.2 (other vertebrates).

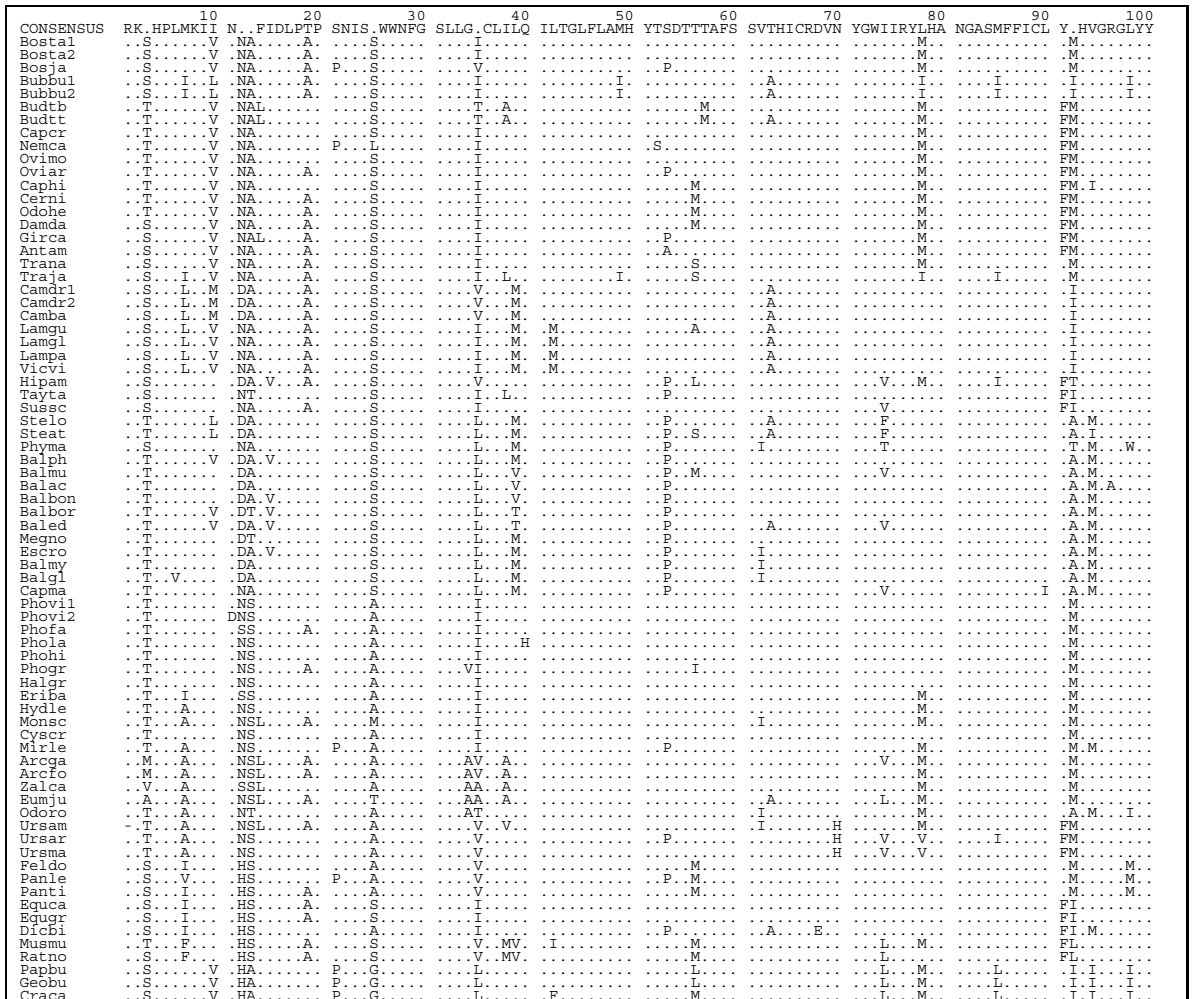


Figure 5.1: (a). The alignment of cytochrome b (mammals), part 1.

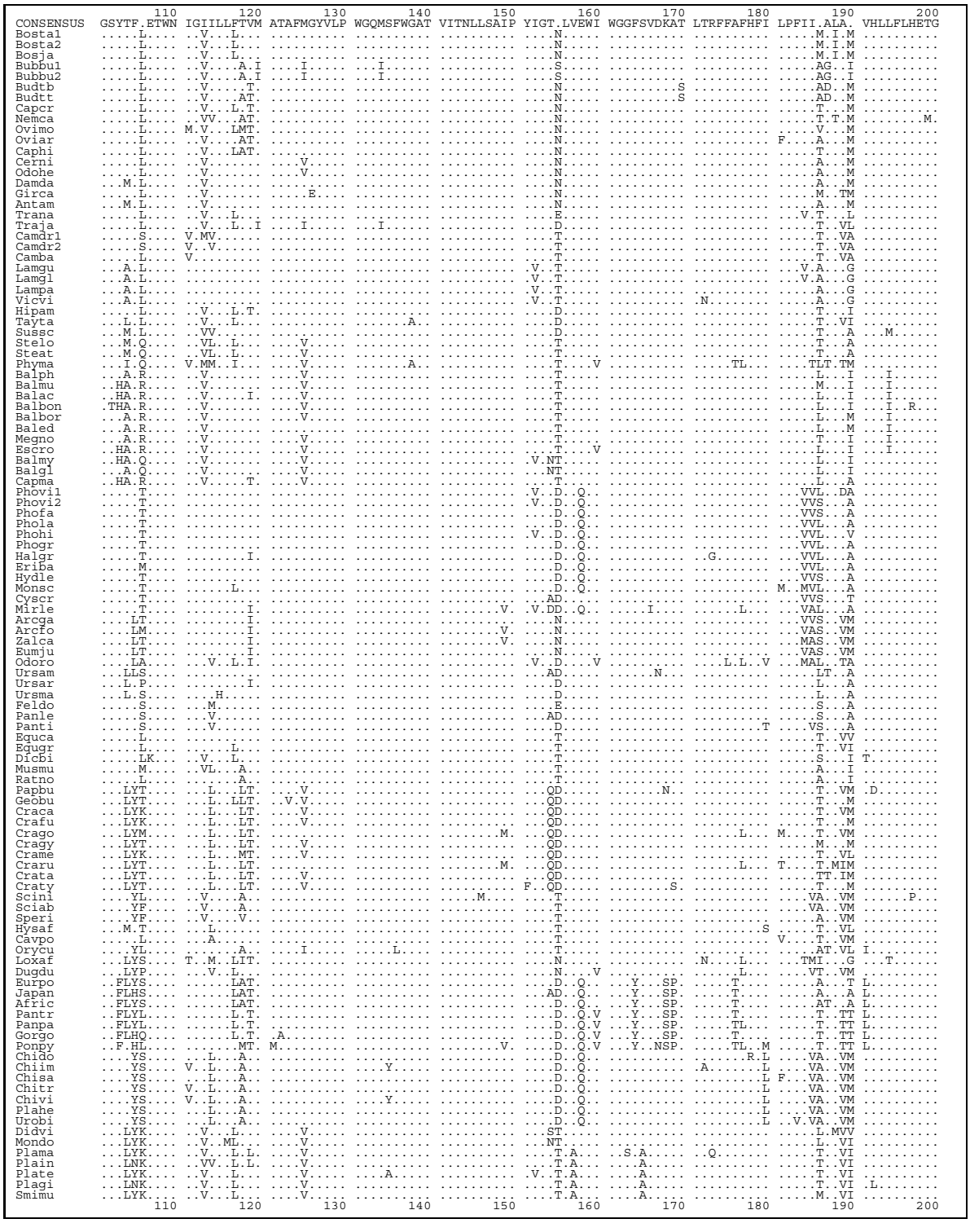


Figure 5.1: (b). The alignment of cytochrome b (mammals), part 2.

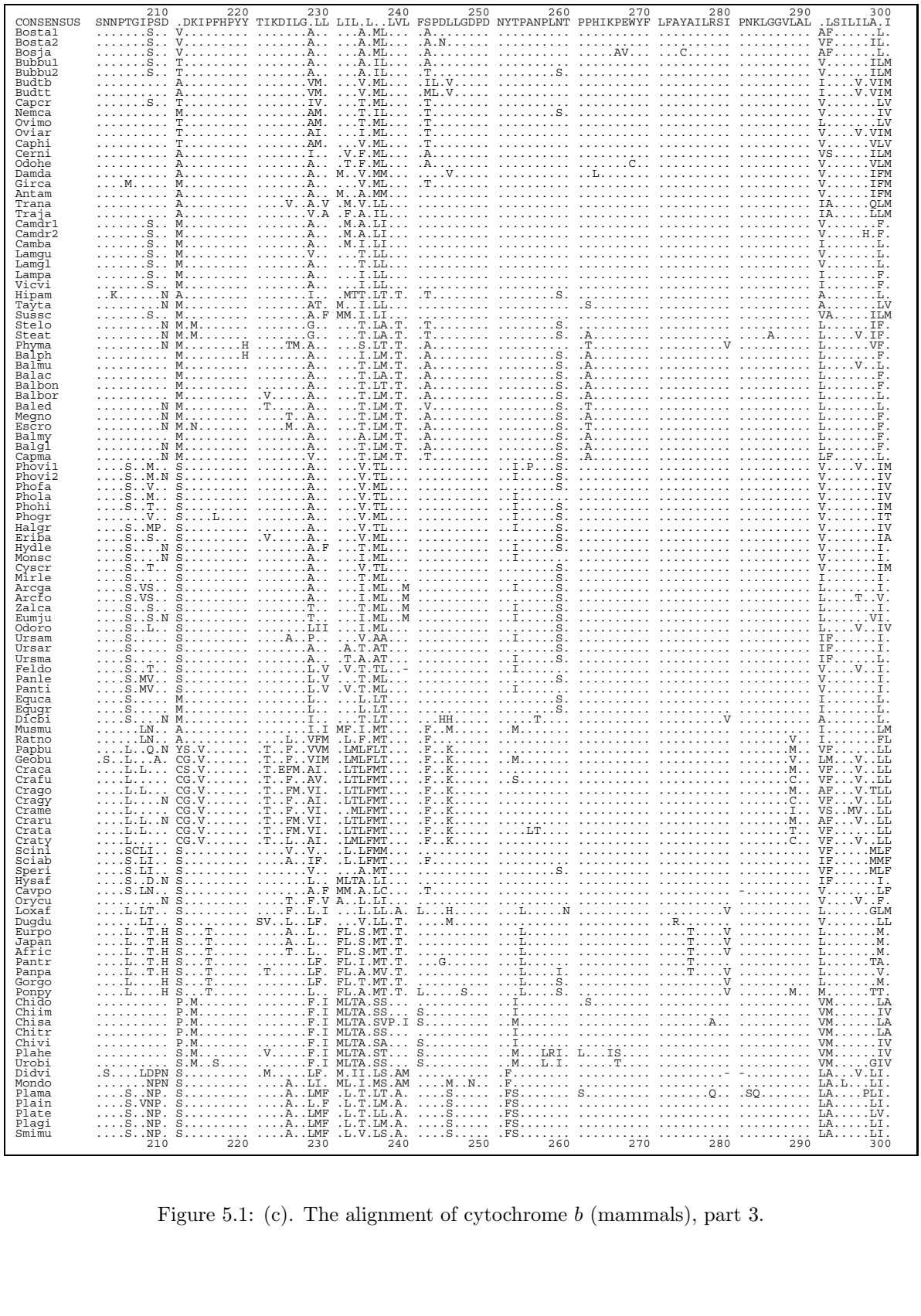


Figure 5.1: (c). The alignment of cytochrome b (mammals), part 3.

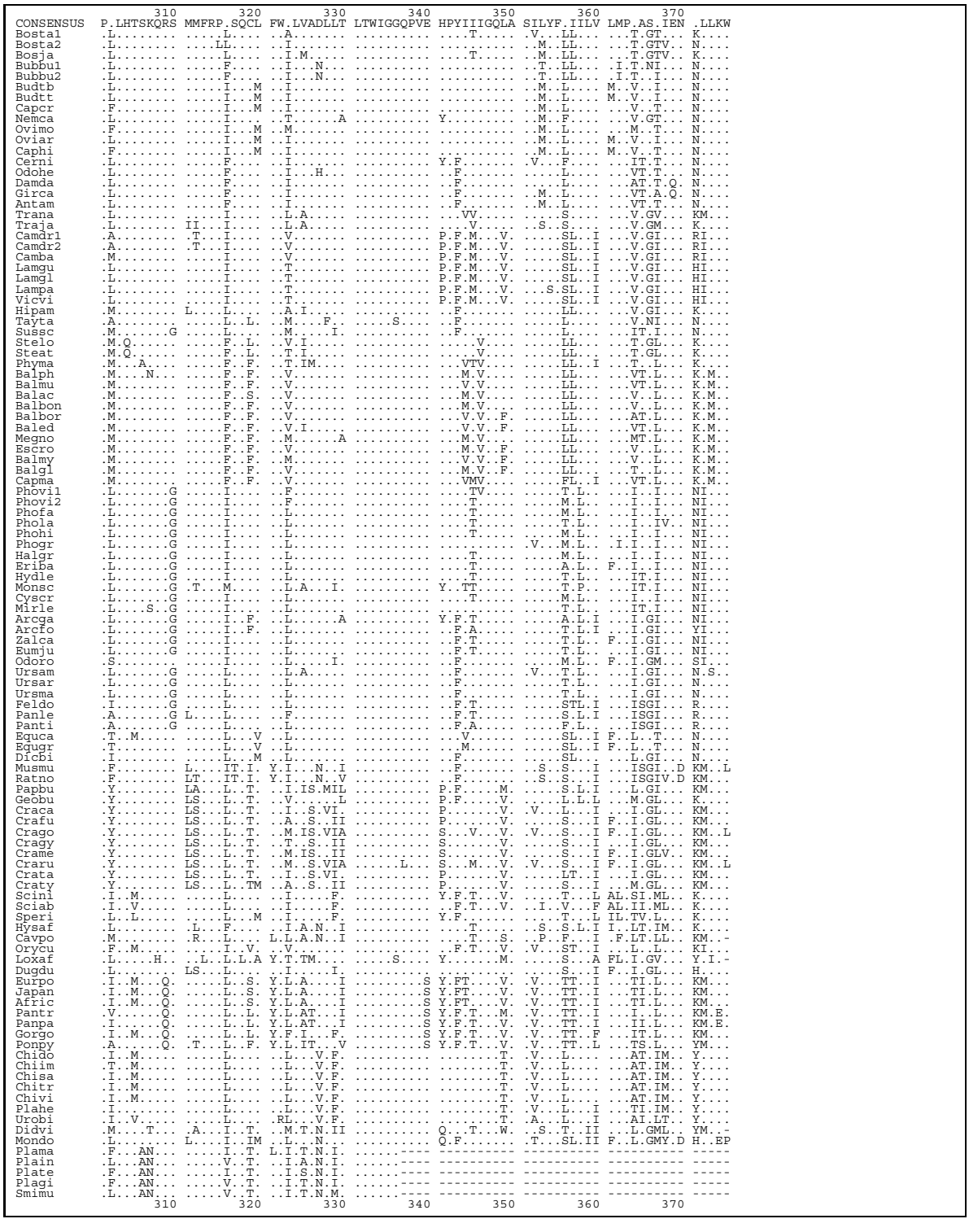


Figure 5.1: (d). The alignment of cytochrome b (mammals), part 4.

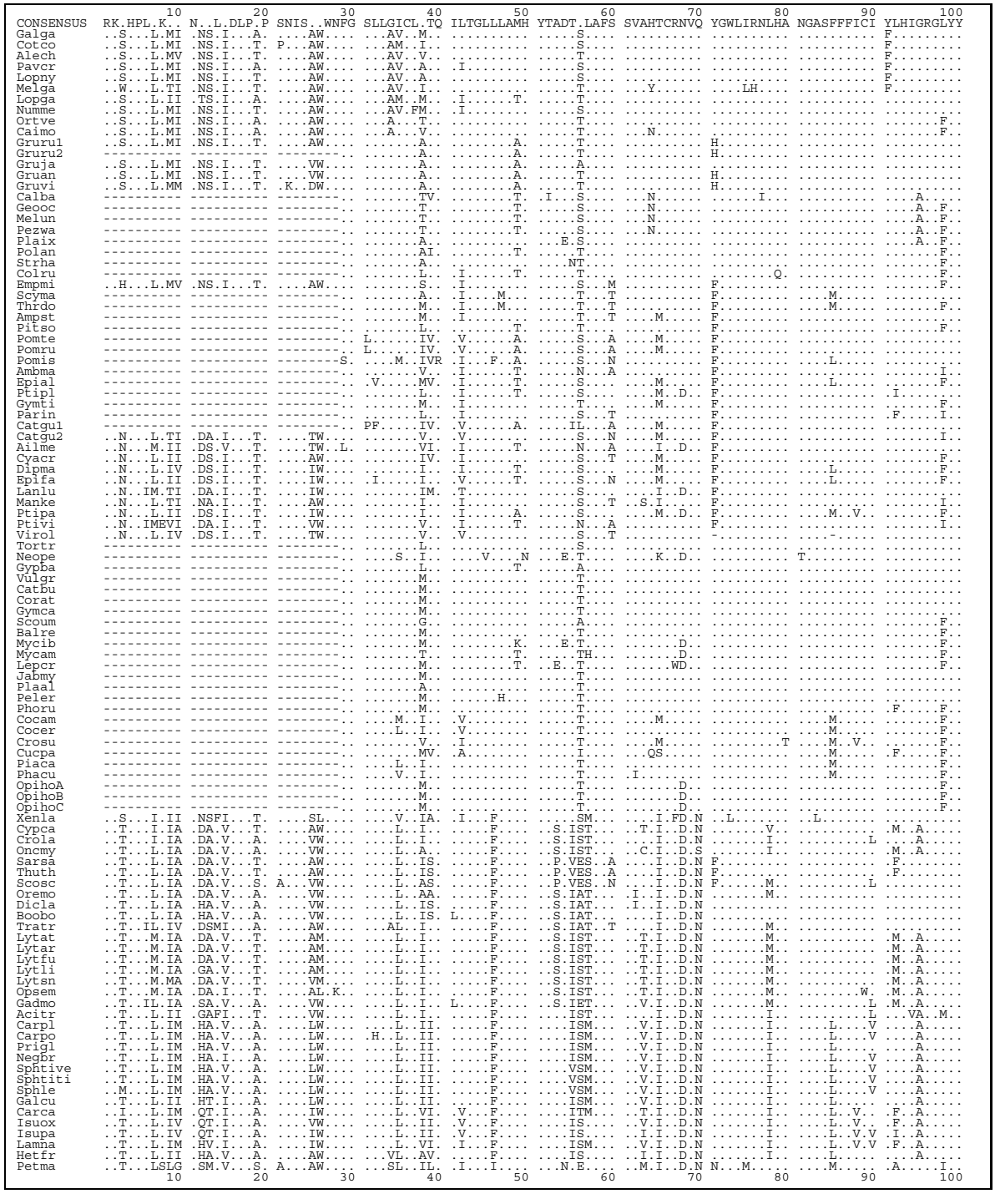


Figure 5.2: (a). The alignment of cytochrome *b* (except mammals), part 1.

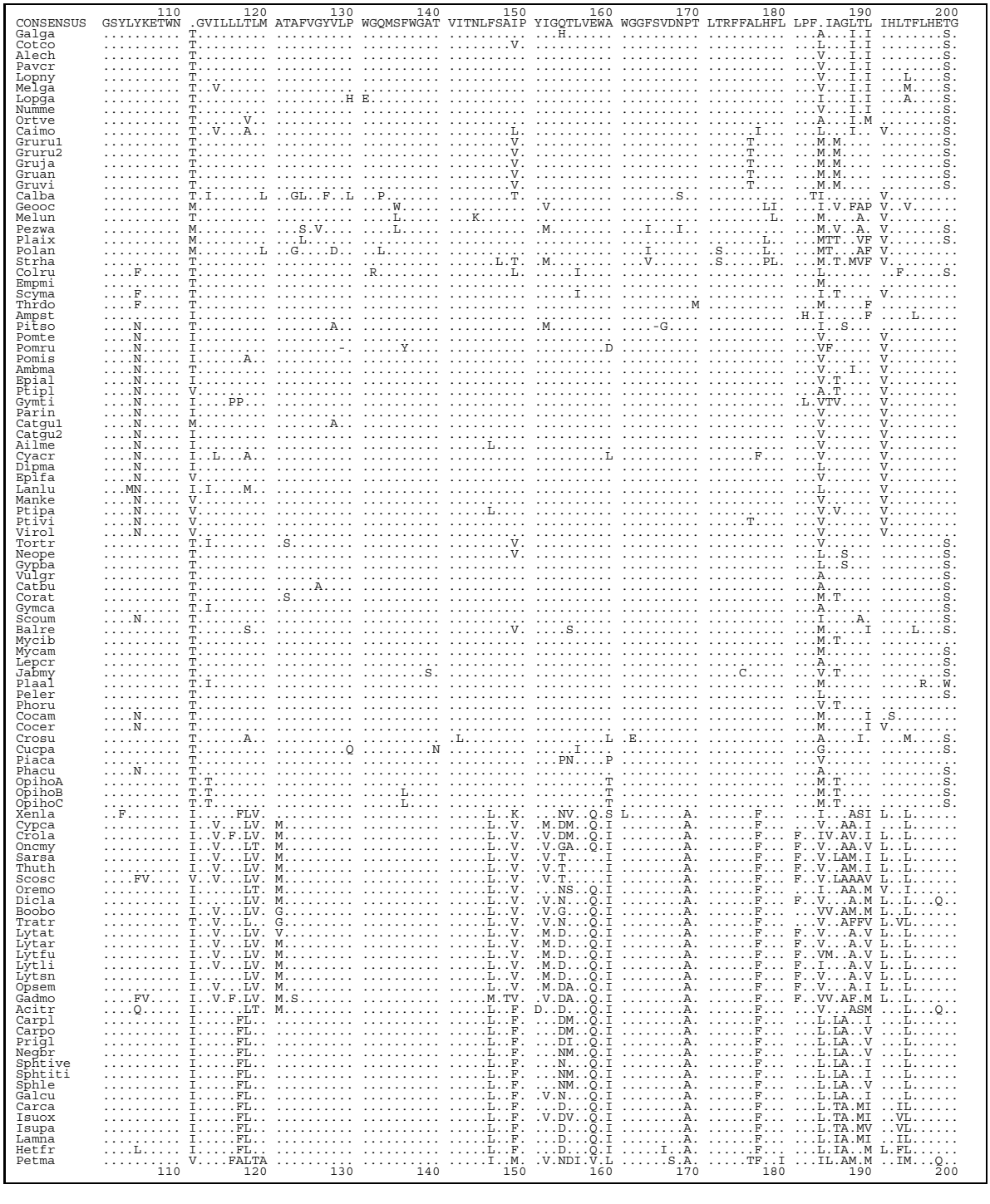


Figure 5.2: (b). The alignment of cytochrome b (except mammals), part 2.

Figure 5.2: (c). The alignment of cytochrome *b* (except mammals), part 3.

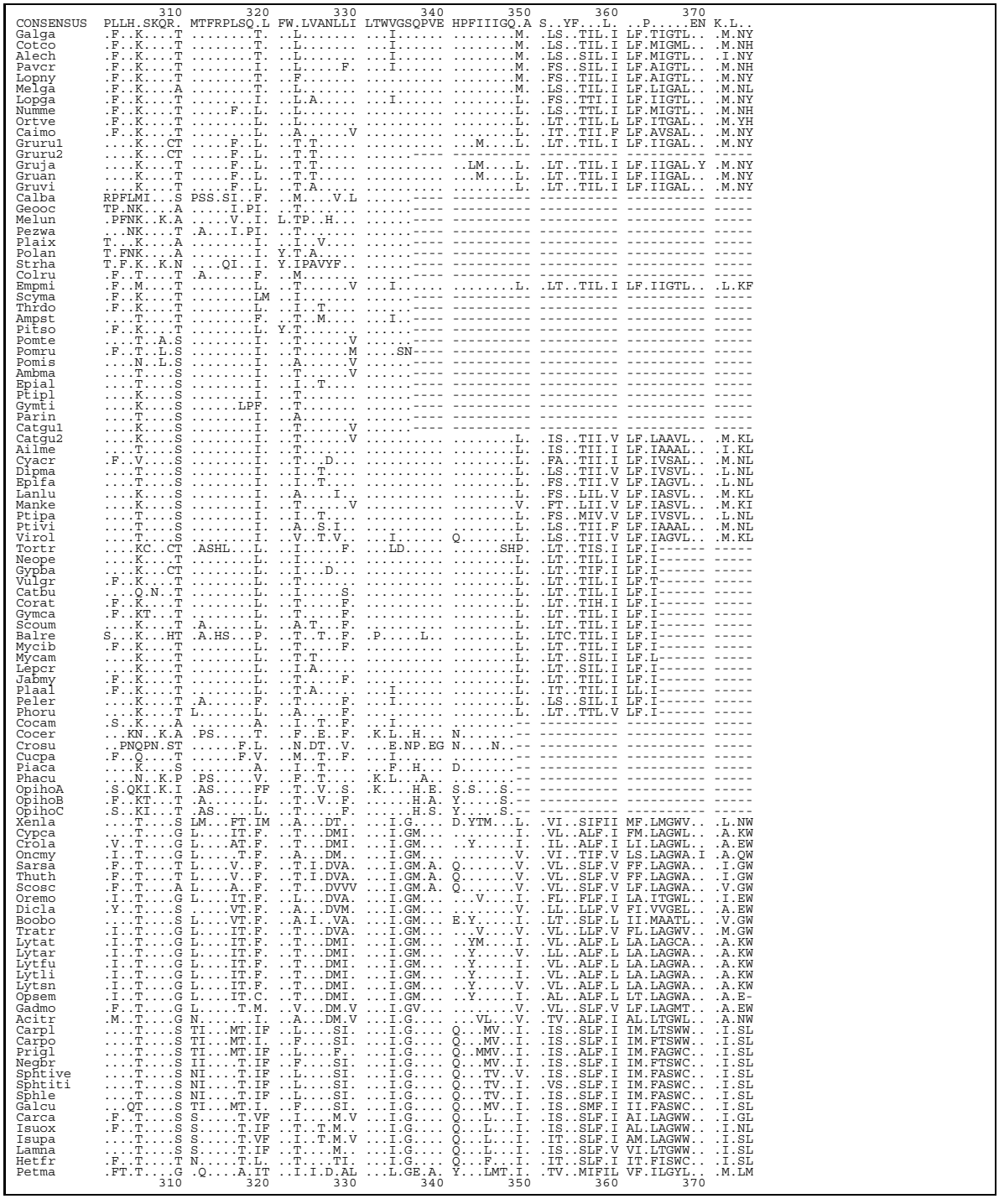


Figure 5.2: (d). The alignment of cytochrome b (except mammals), part 4.

5.1.2 ProtML Tree of 183 OTUs Obtained by Repeated Local Rearrangements

Figs. 5.3 and 5.4 show the NJ tree of cytochrome *b* from 182 OTUs of mammals and birds with a frog as an outgroup (so 183 in total). The distance matrix provided for the NJ analysis was estimated for 2-OTUs trees by ProtML based on the mtREV24-F model. Starting from this tree, a search for better tree topologies by the likelihood criterion was conducted by repeated local (and extended local) rearrangements as described in subsection 3.4.3. Figs. 5.5, 5.6 and 5.7 give the ProtML tree (based on the mtREV24-F model) which cannot be improved by local rearrangements any more. The log-likelihood of the NJ tree is -19177.9, while that of the resultant ProtML tree is -18852.6, showing an improvement of likelihood by 325.3 through the local rearrangement procedure. Since a single gene does not always contain enough information to resolve phylogenetic problems (e.g., Cao et al. 1994[41]), the tree in Fig. 5.5 contains several biologically unreasonable relationships, which might be artifacts. Overall, however, the tree still provides many useful insights on phylogenetics as we will see below.

Note that, since LBP numbers in Fig. 5.5 are estimated by assuming that the relationships within subtrees attached to the relevant branch are correct, they might be misleading when the assumed relationships are not true (see page 49). In that case, even if the LBP is high, the support might be artificial.

5.1.3 Phylogeny of Cetacea

Although the dolphin/sperm whale clade (traditional tree of toothed whale monophyly) is suggested by the NJ tree, the sperm/baleen whales clade with Delphinoids as an outgroup (the Milinkovitch tree; Milinkovitch et al. 1993[184]) is favoured in the ProtML tree with 73% LBP (branch 213; Fig. 5.6a). The second most likely relationship concerning this branching is the traditional tree, and its LBP is 21% (Fig. 5.7a). Therefore, the dolphin/baleen whale clade with sperm whales as an outgroup (the Árnason tree; Árnason and Gullberg 1994[20]) has only 6% LBP, and is least likely from the cytochrome *b* data. Although the support of the Milinkovitch tree is not sufficient to exclude alternative hypotheses in this analysis, increasing the numbers of ingroup species in Delphinoids (Árnason and Gullberg's (1996[21]) data) in the cytochrome *b* analysis helps. Further, the total evidence approach (see section 5.4) using all the relevant molecular data increases the support for the Milinkovitch tree and rejects the traditional and Árnason trees (Hasegawa, Adachi and Milinkovitch, 1996[90]).

Hippopotamus amphibius appears as the most closely related species to Cetacea within Artiodactyla in accord with Irwin and Árnason (1994[125]) and Gatesy et al. (1996[75]), and this relationship is supported with 94% LBP (branch 214). The possible paraphyly of Artiodactyla is most interesting also with respect to the hypothesis of Graur and Higgins (1994[86]) who claim the Ruminantia/Cetacea grouping. More effort should be devoted to resolving this issue with additional sequence data and with improved analyses of the data (Hasegawa and Adachi 1996[89]).

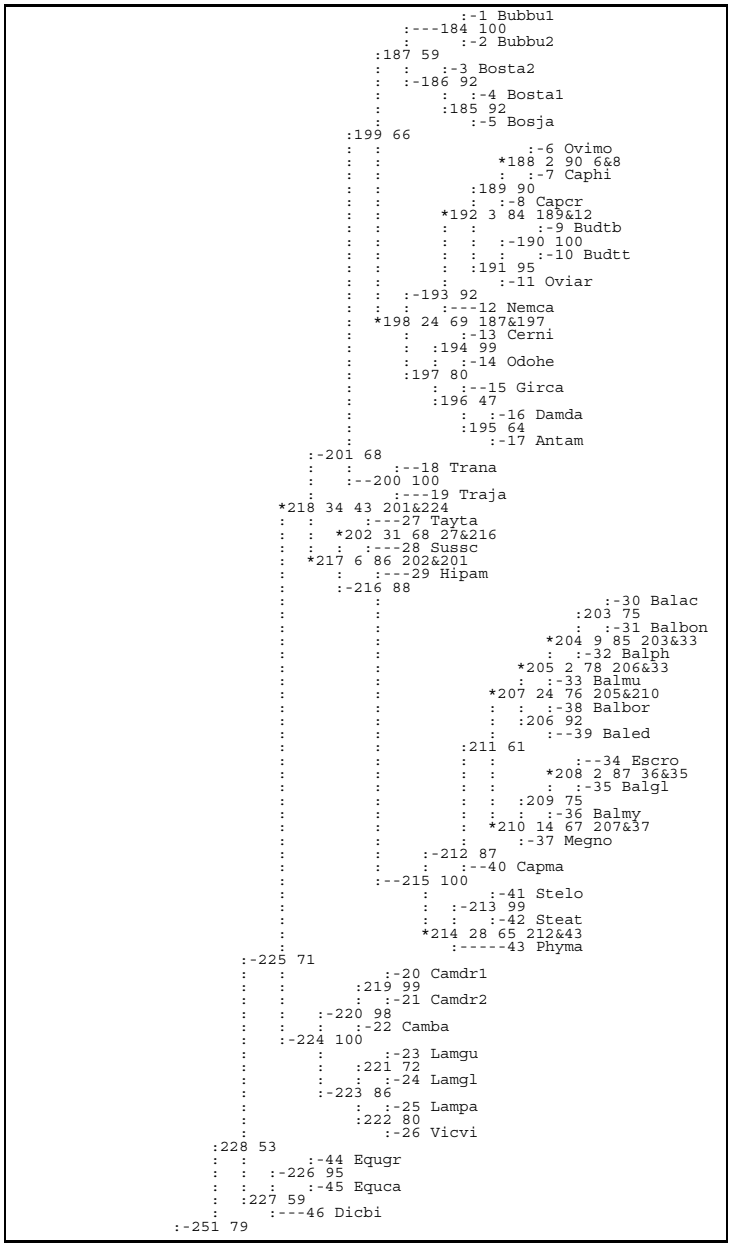


Figure 5.3: (a). The NJ tree of cytochrome b with LBPs estimated by ProtML, part 1.

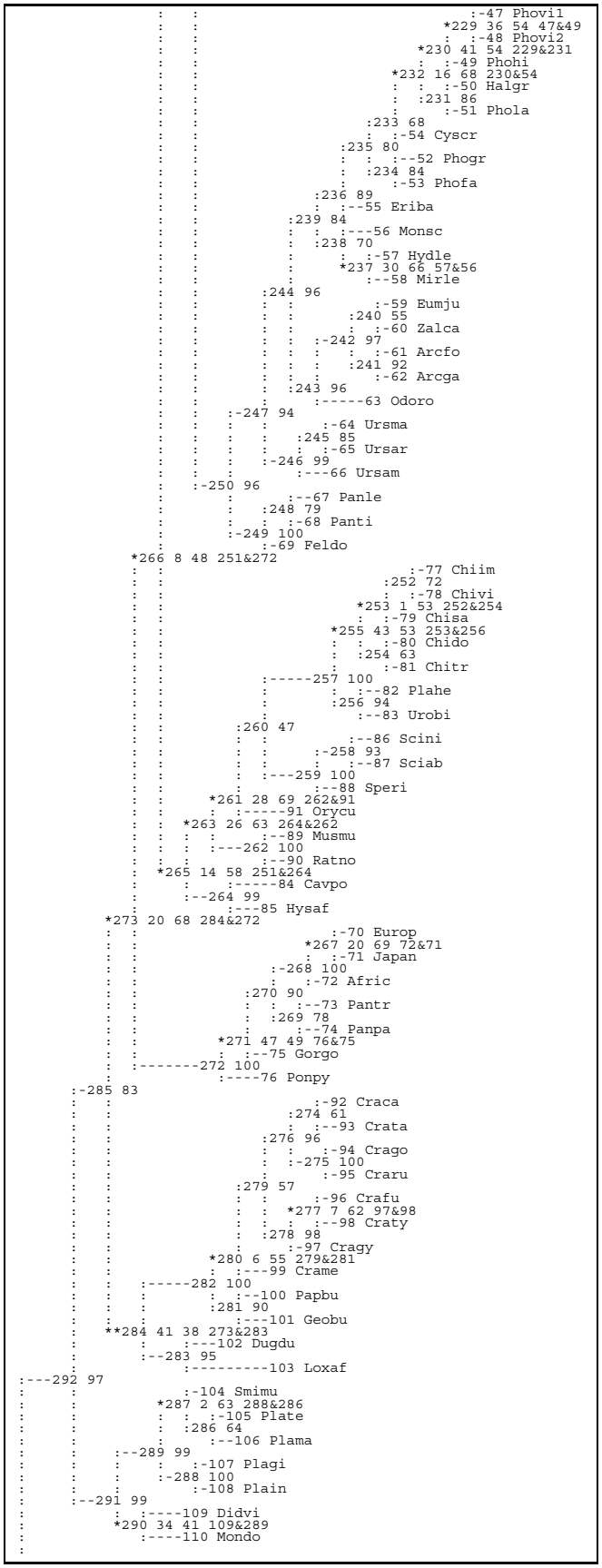


Figure 5.3: (b). The NJ tree of cytochrome *b* with LBPs estimated by ProtML, part 2.

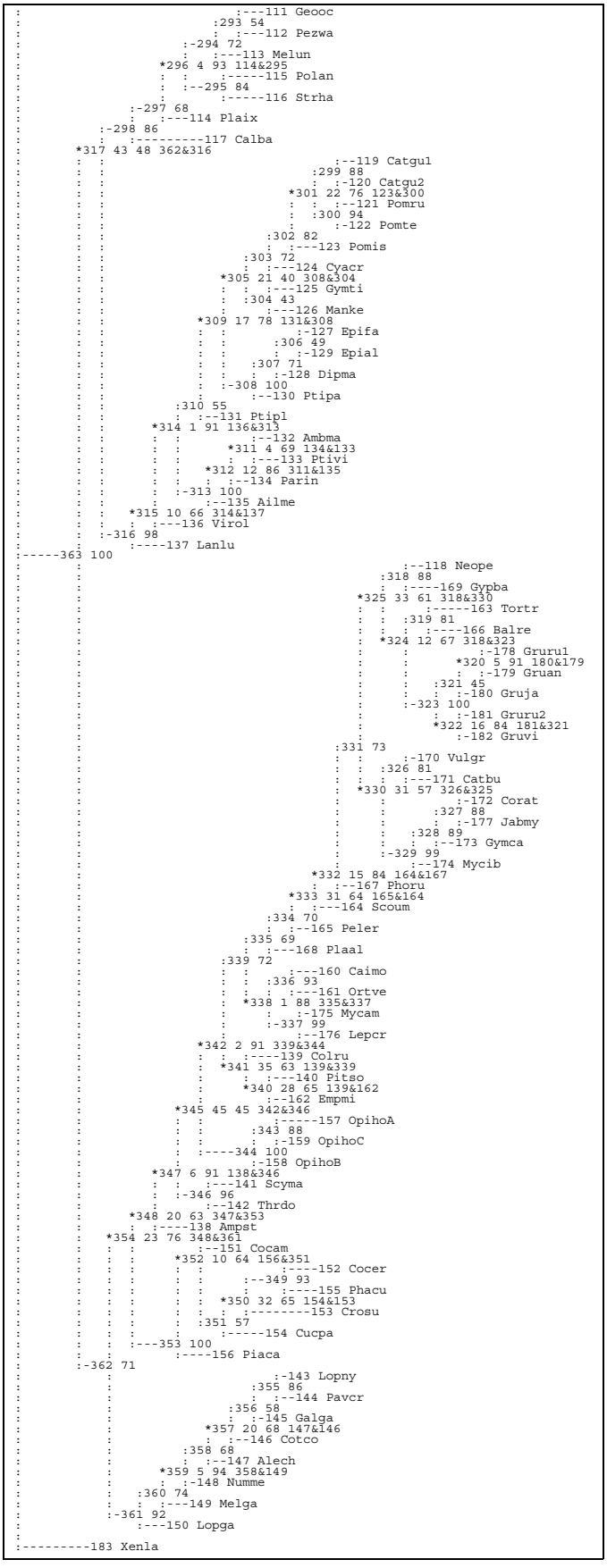


Figure 5.3: (c). The NJ tree of cytochrome b with LBPs estimated by ProtML, part 3.

No.1	ext.	branch	S.E.	int.	branch	S.E.	LBP	2nd	pair
Bubbu1	1	0.18	0.29	184	6.39	1.38	1.0	0.0	1&186
Bubbu2	2	0.62	0.44	185	1.04	0.54	0.915	0.072	3&5
Bosta2	3	0.61	0.43	186	1.95	0.78	0.924	0.070	3&184
Bosta1	4	1.00	0.53	187	1.20	0.65	0.589	0.352	184&198
Bosja	5	1.93	0.73	188	lower limit		0.017*	0.898	6&8
Ovimo	6	1.61	0.66	189	0.78	0.51	0.899	0.065	188&191
Caphi	7	1.29	0.59	190	3.40	0.97	1.0	0.0	11&10
Capcr	8	1.05	0.54	191	1.81	0.74	0.952	0.032	189&11
Budtb	9	0.42	0.36	192	0.23	0.31	0.032*	0.841	189&12
Budtt	10	0.37	0.34	193	2.55	0.92	0.925	0.072	192&197
Oviar	11	1.11	0.56	194	1.66	0.71	0.989	0.009	13&196
Nemca	12	4.18	1.10	195	0.92	0.52	0.635	0.361	16&15
Cerni	13	1.79	0.71	196	0.38	0.38	0.467	0.448	194&195
Odohe	14	1.02	0.54	197	1.16	0.63	0.800	0.167	194&193
Girca	15	2.70	0.87	198	0.48	0.44	0.242*	0.692	187&197
Damda	16	1.70	0.69	199	1.37	0.75	0.658	0.207	200&198
Antam	17	0.42	0.37	200	4.11	1.17	0.995	0.005	199&19
Trana	18	2.03	0.83	201	2.99	1.01	0.682	0.299	199&217
Traja	19	4.96	1.23	202	1.45	0.73	0.309*	0.683	27&216
Camdr1	20	0.26	0.26	203	0.89	0.53	0.746	0.217	32&31
Camdr2	21	0.27	0.27	204	0.28	0.34	0.091*	0.850	203&33
Camba	22	0.34	0.34	205	lower limit		0.015*	0.778	206&33
Lamgu	23	0.53	0.38	206	1.74	0.74	0.915	0.079	38&205
Lamgl	24	0.00	----	207	0.52	0.43	0.241*	0.758	205&210
Lampa	25	0.27	0.27	208	lower limit		0.016*	0.871	36&35
Vicvi	26	0.27	0.27	209	1.30	0.65	0.752	0.181	208&37
Tayta	27	5.15	1.23	210	0.23	0.29	0.144*	0.670	207&37
Sussc	28	3.93	1.09	211	0.71	0.51	0.611	0.313	207&40
Hipam	29	5.24	1.30	212	2.17	0.92	0.872	0.085	211&214
Balac	30	1.26	0.60	213	3.38	1.08	0.986	0.014	43&42
Balbon	31	0.65	0.45	214	1.04	0.72	0.277*	0.653	212&43
Balph	32	1.25	0.61	215	5.06	1.30	0.999	0.001	29&214
Balmu	33	1.93	0.76	216	3.00	1.02	0.885	0.072	29&202
Escro	34	1.95	0.77	217	lower limit		0.059*	0.863	202&201
Balgl	35	0.99	0.60	218	1.09	0.65	0.344*	0.434	201&224
Balmy	36	1.74	0.74	219	1.60	0.66	0.992	0.008	22&21
Megno	37	1.78	0.72	220	2.24	0.85	0.982	0.014	223&22
Balbor	38	0.65	0.44	221	0.35	0.35	0.718	0.250	222&24
Baled	39	2.13	0.76	222	0.72	0.46	0.802	0.181	25&221
Capma	40	2.93	0.93	223	2.11	0.81	0.856	0.138	220&222
Stelo	41	0.08	0.28	224	3.61	1.09	0.997	0.002	220&218
Steat	42	1.54	0.66	225	2.14	0.89	0.707	0.273	227&224
Phyma	43	9.41	1.69	226	3.16	1.01	0.954	0.046	44&46
Equgr	44	0.40	0.39	227	1.66	0.85	0.591	0.400	226&225
Equca	45	0.70	0.48	228	1.51	0.80	0.528	0.223	225&250
Dicbi	46	5.21	1.27	229	0.26	0.27	0.362*	0.535	47&49
Phovi1	47	1.52	0.65	230	0.34	0.33	0.409*	0.544	229&231
Phovi2	48	0.87	0.50	231	lower limit		0.859	0.103	230&51
Phohi	49	0.71	0.45	232	0.02	0.35	0.161*	0.683	230&54
Halgr	50	0.79	0.46	233	0.28	0.28	0.685	0.245	234&54
Phola	51	1.06	0.53	234	0.51	0.38	0.841	0.109	233&53
Phogr	52	2.17	0.77	235	0.86	0.55	0.803	0.180	55&234
Phofa	53	0.79	0.46	236	0.96	0.57	0.886	0.099	235&238
Cyscr	54	1.84	0.70	237	lower limit		0.301*	0.658	57&56
Eriba	55	2.16	0.79	238	0.34	0.33	0.705	0.256	56&236
Monsc	56	4.53	1.12	239	1.16	0.59	0.845	0.146	236&243
Hydle	57	0.75	0.46	240	0.60	0.43	0.553	0.409	59&241
Mirle	58	3.85	1.03	241	1.02	0.57	0.916	0.078	61&240
Eumju	59	1.62	0.68	242	2.41	0.88	0.971	0.029	240&63
Zalca	60	1.70	0.70	243	1.86	0.77	0.957	0.033	242&239
Arcfo	61	1.86	0.71	244	1.52	0.71	0.958	0.025	246&243
Arcga	62	1.34	0.60	245	0.74	0.48	0.848	0.136	64&66
Odoro	63	8.65	1.59	246	1.98	0.79	0.990	0.009	244&66
Ursma	64	0.89	0.51	247	2.46	0.88	0.937	0.062	244&249
Ursar	65	1.61	0.67	248	0.87	0.53	0.791	0.135	69&68
Ursam	66	4.18	1.09	249	2.87	0.99	0.995	0.005	248&247
Panle	67	2.47	0.84	250	2.63	1.01	0.960	0.037	247&228
Panti	68	1.57	0.67	251	3.30	1.08	0.793	0.206	228&265
Feldo	69	1.50	0.69	252	0.94	0.55	0.717	0.270	79&78
Europ	70	0.20	0.26	253	lower limit		0.007*	0.529	252&254
Japan	71	0.59	0.41	254	0.51	0.44	0.633	0.334	253&81
Afric	72	0.45	0.40	255	0.81	0.59	0.434*	0.526	253&256
Pantr	73	2.09	0.76	256	1.80	0.78	0.944	0.050	255&83
Panpa	74	2.17	0.79	257	10.12	1.83	1.0	0.0	259&256
Gorgo	75	2.95	0.97	258	2.98	1.05	0.927	0.051	86&88
Ponpy	76	7.70	1.56	259	5.86	1.41	1.0	0.0	257&88
Chiim	77	0.54	0.38	260	1.11	0.80	0.472	0.349	91&259
Chivi	78	0.26	0.27	261	1.29	0.86	0.275*	0.693	262&91
Chisa	79	1.86	0.76	262	7.60	1.56	1.0	0.0	89&261
Chido	80	0.84	0.48	263	1.58	0.82	0.261*	0.632	264&262
Chitr	81	0.23	0.27	264	4.57	1.27	0.994	0.006	84&263
Plahe	82	2.15	0.79	265	0.44	0.63	0.135*	0.581	251&264
Urobi	83	2.65	0.88	266	0.93	0.79	0.085*	0.480	251&272
Cavpo	84	8.38	1.64	267	0.14	0.27	0.197*	0.689	72&71
Hysaf	85	5.46	1.34	268	3.65	1.03	1.0	0.0	267&269
Scini	86	2.98	0.97	269	0.72	0.51	0.784	0.166	268&74
Sciab	87	3.20	1.00	270	1.91	0.79	0.899	0.069	268&75
Speri	88	2.73	1.02	271	1.39	0.86	0.474*	0.492	76&75
Musmu	89	2.91	0.98	272	15.03	2.24	1.0	0.0	266&76
Ratno	90	3.13	1.01	273	1.25	0.92	0.202*	0.684	284&272
Orycu	91	9.03	1.68	274	0.43	0.39	0.611	0.254	92&275

Figure 5.4: (a). Branch lengths and LBPs of the NJ tree of cytochrome *b* estimated by the ProtML, part 1.

Craca	92	1.30	0.63	275	3.33	0.99	1.0	0.0	274&95
Crata	93	3.48	1.00	276	1.56	0.69	0.956	0.040	274&278
Crago	94	1.24	0.61	277	0.01	0.28	0.074*	0.617	97&98
Craru	95	1.82	0.72	278	1.34	0.64	0.980	0.020	276&97
Crafu	96	1.89	0.74	279	0.29	0.31	0.572	0.332	99&278
Cragy	97	1.16	0.58	280	lower limit		0.058*	0.549	279&281
Craty	98	1.98	0.75	281	1.47	0.68	0.898	0.084	100&280
Crame	99	4.14	1.08	282	11.26	1.95	1.0	0.0	280&283
Papbu	100	3.87	1.06	283	5.15	1.43	0.949	0.040	102&282
Geobu	101	4.51	1.16	284	2.37	1.15	0.406*	0.381	273&283
Dugdu	102	4.65	1.40	285	3.63	1.21	0.831	0.169	291&284
Loxaf	103	16.93	2.44	286	0.51	0.51	0.637	0.334	105&104
Smimu	104	1.29	0.71	287	lower limit		0.023*	0.626	288&286
Plate	105	1.66	0.75	288	3.39	1.06	1.0	0.0	107&287
Plama	106	3.87	1.16	289	4.83	1.37	0.991	0.008	287&290
Plagi	107	0.96	0.58	290	1.59	0.91	0.342*	0.412	109&289
Plain	108	1.64	0.74	291	5.19	1.51	0.991	0.007	289&285
Didvi	109	7.39	1.56	292	6.99	1.68	0.969	0.031	363&291
Mondo	110	6.27	1.44	293	1.54	0.79	0.539	0.252	111&113
Geooc	111	4.66	1.33	294	2.78	1.12	0.716	0.224	293&295
Pezwa	112	4.64	1.32	295	3.99	1.30	0.839	0.160	115&294
Melun	113	4.06	1.25	296	0.46	0.64	0.039*	0.927	114&295
Plaix	114	4.16	1.30	297	2.65	1.16	0.683	0.303	296&117
Polan	115	8.78	1.87	298	2.40	1.07	0.856	0.124	316&117
Strha	116	7.84	1.76	299	1.77	0.81	0.878	0.119	119&300
Calba	117	16.48	2.63	300	1.50	0.80	0.936	0.064	299&122
Neope	118	3.82	1.13	301	0.74	0.62	0.221*	0.755	123&300
Catgu1	119	2.49	0.95	302	1.52	0.78	0.824	0.101	124&123
Catgu2	120	1.82	0.80	303	0.98	0.57	0.722	0.190	302&304
Pomru	121	3.53	1.13	304	0.53	0.52	0.429	0.502	303&126
Pomte	122	0.46	0.48	305	lower limit		0.206*	0.405	308&304
Pomis	123	4.89	1.34	306	0.33	0.41	0.493	0.355	128&129
Cyacr	124	4.80	1.21	307	0.89	0.54	0.714	0.248	130&128
Gymti	125	4.99	1.34	308	2.47	0.87	0.999	0.001	307&305
Manke	126	4.08	1.12	309	0.35	0.43	0.171*	0.784	131&308
Epifa	127	1.53	0.69	310	0.74	0.56	0.547	0.411	313&131
Dipma	128	1.47	0.66	311	lower limit		0.045*	0.694	134&133
Epial	129	1.30	0.68	312	0.70	0.65	0.125*	0.860	311&135
Ptipa	130	3.09	0.95	313	3.43	1.04	0.998	0.002	310&135
Ptipl	131	3.80	1.16	314	lower limit		0.014*	0.914	136&313
Ambma	132	2.20	0.89	315	lower limit		0.103*	0.657	314&137
Ptivi	133	4.02	1.12	316	2.86	1.10	0.975	0.018	298&137
Parin	134	3.46	1.12	317	1.70	1.08	0.429*	0.476	362&316
Ailme	135	3.04	0.94	318	1.74	0.80	0.882	0.113	118&324
Virol	136	4.73	1.20	319	0.76	0.58	0.809	0.158	323&166
Lanlu	137	6.30	1.38	320	lower limit		0.050*	0.908	180&179
Ampst	138	6.64	1.55	321	0.69	0.46	0.446	0.448	320&322
Colru	139	5.92	1.50	322	lower limit		0.156*	0.844	181&321
Pitso	140	5.57	1.46	323	3.08	0.95	1.0	0.0	321&319
Scyma	141	3.94	1.18	324	lower limit		0.125*	0.672	318&323
Thrdo	142	2.27	0.91	325	0.28	0.29	0.329*	0.606	318&330
Lopny	143	1.13	0.57	326	1.27	0.68	0.810	0.172	170&329
Pavcr	144	2.42	0.82	327	0.67	0.47	0.880	0.118	173&177
Galga	145	1.46	0.66	328	1.32	0.68	0.889	0.096	327&174
Cotco	146	2.39	0.82	329	2.82	0.98	0.988	0.011	328&326
Alech	147	2.10	0.77	330	0.43	0.55	0.312*	0.575	326&325
Numme	148	1.59	0.69	331	0.46	0.40	0.726	0.186	325&167
Melga	149	4.77	1.18	332	lower limit		0.151*	0.836	164&167
Lopga	150	5.23	1.25	333	0.26	0.29	0.306*	0.640	165&164
Cocam	151	2.88	1.03	334	0.58	0.41	0.702	0.275	168&165
Cocer	152	6.14	1.54	335	0.35	0.35	0.691	0.206	338&168
Crosu	153	15.23	2.48	336	1.67	0.81	0.932	0.065	337&161
Cucpa	154	8.47	1.77	337	2.51	0.92	0.993	0.006	175&336
Phacu	155	6.40	1.59	338	lower limit		0.013*	0.878	335&337
Piaca	156	5.95	1.47	339	1.00	0.61	0.724	0.265	335&341
OpihoA	157	7.85	1.65	340	0.99	0.69	0.283*	0.649	139&162
OpihoB	158	1.19	0.67	341	1.15	0.80	0.354*	0.633	139&339
OpihoC	159	0.87	0.58	342	lower limit		0.020*	0.914	339&344
Caimo	160	5.66	1.31	343	1.57	0.74	0.879	0.069	158&159
Ortve	161	5.13	1.26	344	8.33	1.74	1.0	0.0	342&158
Empmi	162	3.73	1.13	345	0.74	0.60	0.447*	0.451	342&346
Tortr	163	8.41	1.68	346	2.85	1.03	0.963	0.037	345&142
Scoum	164	4.77	1.26	347	0.22	0.60	0.061*	0.909	138&346
Peler	165	3.39	1.04	348	0.78	0.65	0.195*	0.633	347&353
Balre	166	6.57	1.47	349	4.06	1.26	0.929	0.071	153&155
Phoru	167	3.14	1.00	350	0.85	0.71	0.324*	0.647	154&153
Plaal	168	4.02	1.13	351	1.06	0.74	0.569	0.274	350&151
Gypba	169	6.27	1.45	352	0.18	0.43	0.104*	0.644	156&351
Vulgr	170	0.88	0.59	353	7.00	1.62	1.0	0.0	352&348
Catbu	171	4.57	1.22	354	1.03	0.72	0.229*	0.762	348&361
Corat	172	1.53	0.69	355	1.05	0.55	0.858	0.126	145&144
Gymca	173	3.02	0.98	356	0.55	0.45	0.578	0.352	146&145
Mycib	174	2.79	0.97	357	lower limit		0.195*	0.681	147&146
Mycam	175	1.74	0.79	358	1.19	0.62	0.680	0.310	357&148
Lepcr	176	2.63	0.94	359	0.28	0.33	0.054*	0.945	358&149
Jabmy	177	0.91	0.53	360	0.89	0.69	0.737	0.126	359&150
Gruru1	178	1.65	0.68	361	3.31	1.04	0.920	0.080	360&354
Gruan	179	0.00	----	362	3.64	1.22	0.714	0.284	354&317
Gruja	180	1.36	0.61	363	11.40	2.12	1.0	0.0	317&292
Gruru2	181	0.33	0.33	TBL :	896.02		iter: 1		
Gruvi	182	1.93	0.73	ln L:	-19177.88		+- 998.18		
Xenla	183	17.19	2.54	AIC :	39119.76		lower limit: 0.001		

Figure 5.4: (b). Branch lengths and LBPs of the NJ tree of cytochrome *b* estimated by the ProtML, part 2.

protml 2.3b3 05/30/96 mtREV24-F 183 OTUs 375 sites cyt-b Vertebrate '95.3.31

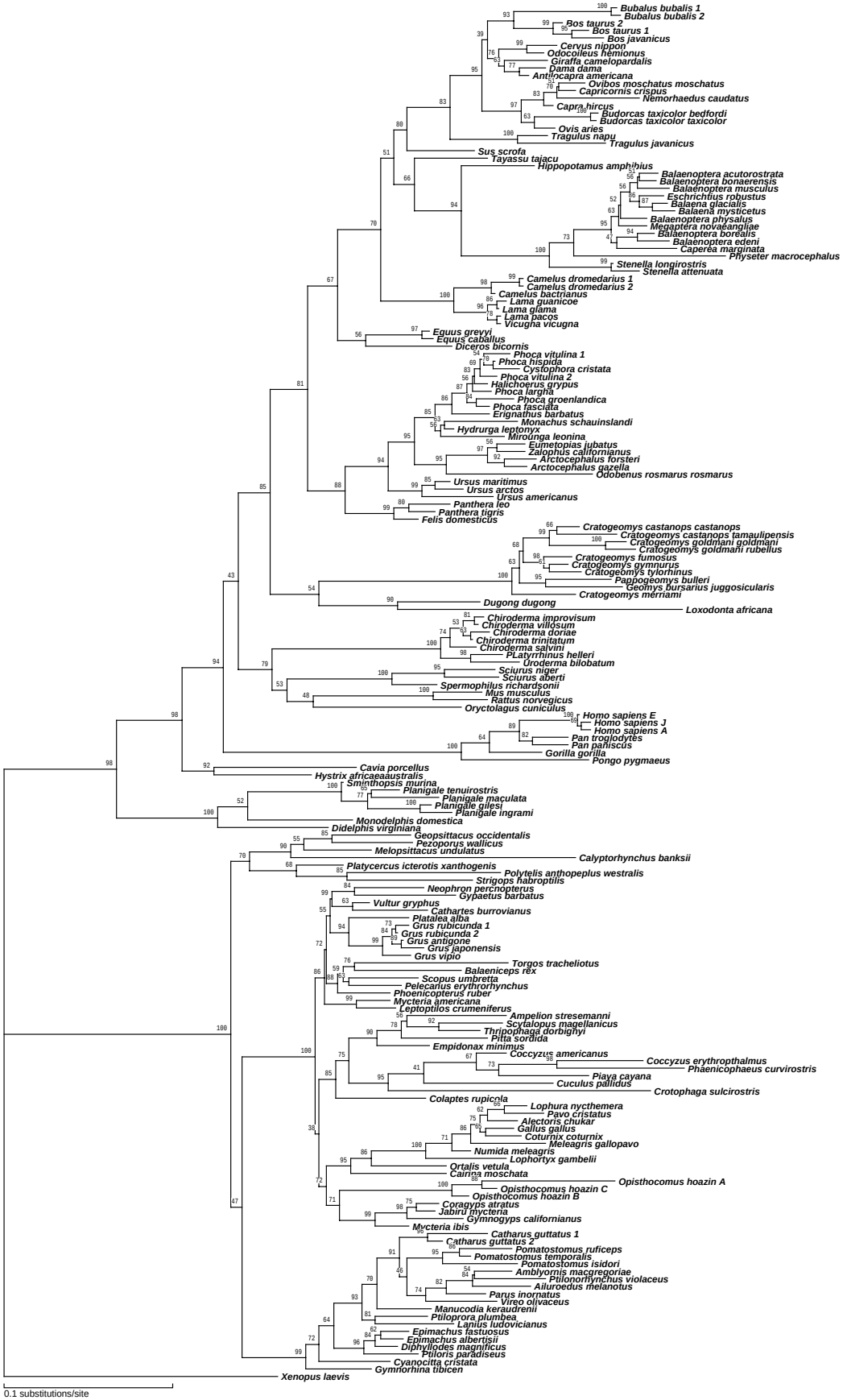


Figure 5.5: ProtML tree of cytochrome b obtained by local rearrangements (mtREV24-F model).

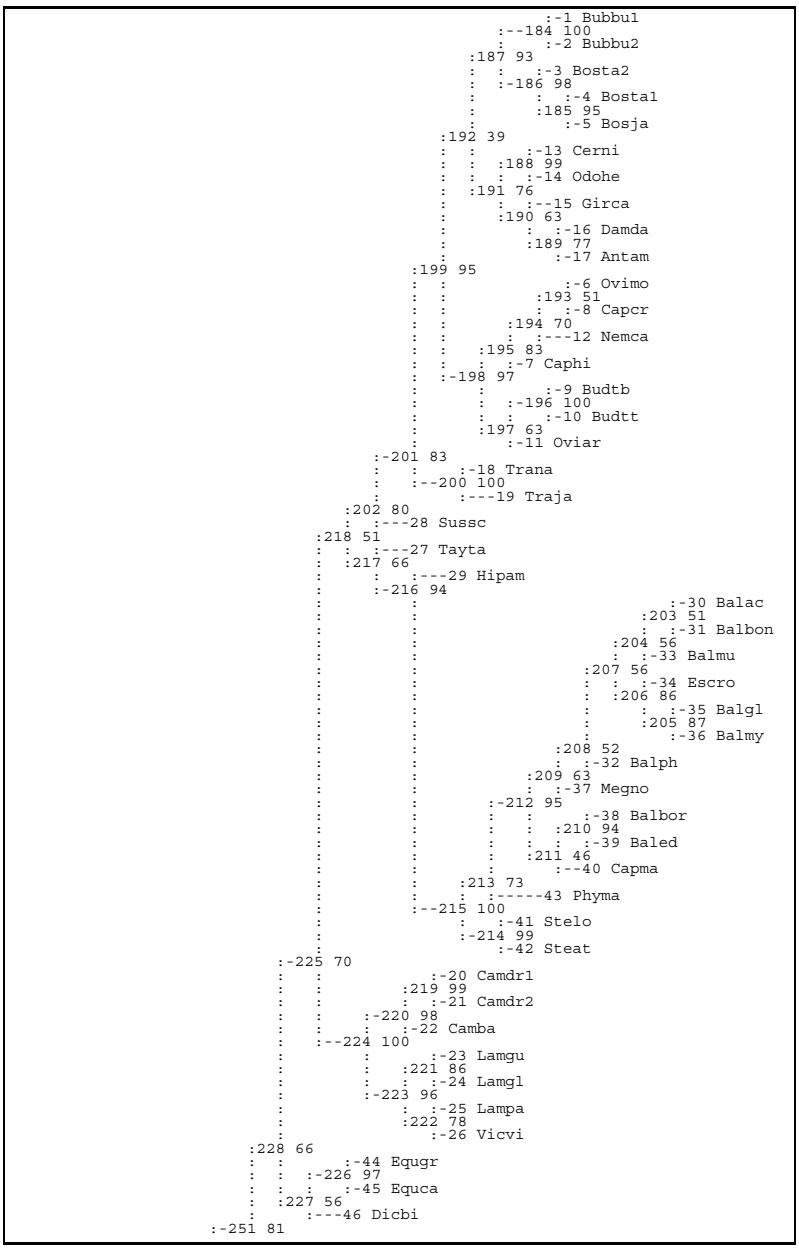


Figure 5.6: (a). The ML tree of cytochrome b, part 1.

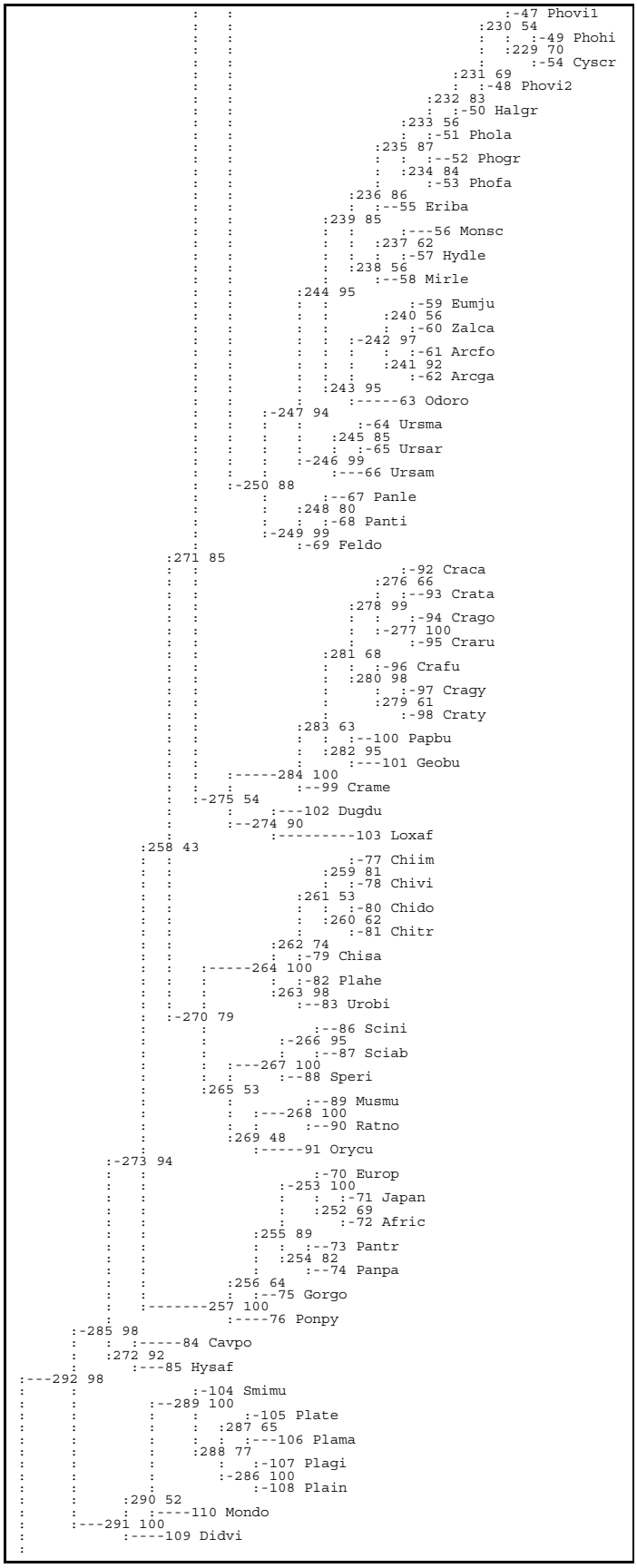


Figure 5.6: (b). The ML tree of cytochrome *b*, part 2.

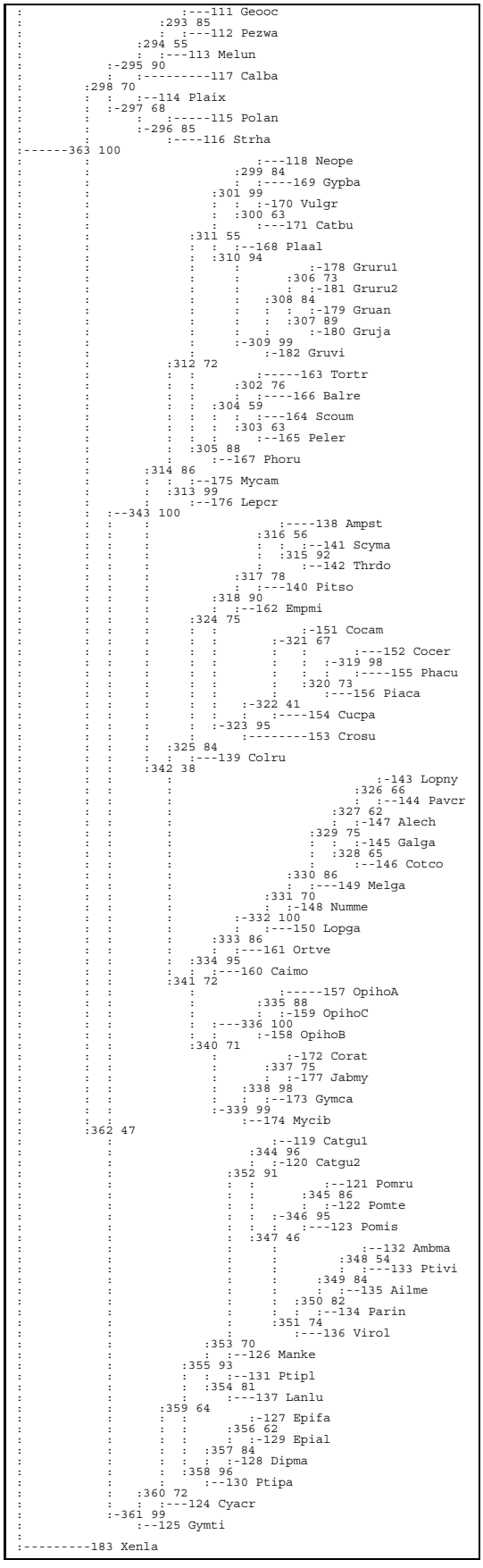


Figure 5.6: (c). The ML tree of cytochrome b, part 3.

	ext.	branch	S.E.	int.	branch	S.E.	LBP	2nd	pair
Bubbu1	1	0.29	0.30	184	5.77	1.30	1.0	0.0	1&186
Bubbu2	2	0.51	0.38	185	1.07	0.54	0.953	0.040	3&5
Bosta2	3	0.57	0.40	186	2.36	0.83	0.985	0.010	3&184
Bosta1	4	1.01	0.53	187	1.53	0.68	0.930	0.040	184&191
Bosja	5	1.92	0.73	188	1.67	0.71	0.989	0.003	13&190
Ovimo	6	1.58	0.65	189	0.93	0.52	0.771	0.212	16&15
Caphi	7	0.53	0.38	190	0.31	0.31	0.628	0.209	188&189
Capcr	8	1.00	0.52	191	0.66	0.46	0.761	0.201	188&187
Budtb	9	0.42	0.36	192	0.31	0.36	0.394	0.330	187&198
Budtt	10	0.36	0.34	193	0.04	0.28	0.510	0.453	6&12
Oviar	11	1.18	0.58	194	0.78	0.46	0.697	0.296	193&7
Nemca	12	4.82	1.14	195	1.28	0.62	0.834	0.160	197&7
Cerni	13	1.79	0.71	196	3.33	0.96	1.0	0.0	11&10
Odohe	14	1.02	0.54	197	0.75	0.48	0.630	0.230	196&195
Girca	15	2.57	0.84	198	2.31	0.83	0.973	0.020	192&197
Damda	16	1.58	0.65	199	1.87	0.83	0.949	0.047	200&198
Antam	17	0.53	0.38	200	4.02	1.16	0.996	0.003	199&19
Trana	18	1.75	0.77	201	2.56	0.91	0.826	0.113	28&200
Traja	19	5.18	1.25	202	0.78	0.55	0.804	0.149	217&28
Camdr1	20	0.26	0.26	203	0.07	0.31	0.506	0.399	30&33
Camdr2	21	0.27	0.27	204	0.45	0.37	0.556	0.384	203&206
Camba	22	0.23	0.27	205	0.81	0.47	0.868	0.120	35&34
Lamgu	23	0.53	0.38	206	0.56	0.40	0.858	0.089	34&204
Lamgl	24	0.00	----	207	0.52	0.38	0.557	0.392	204&32
Lampa	25	0.27	0.27	208	0.09	0.27	0.524	0.305	37&32
Vicvi	26	0.27	0.27	209	0.48	0.39	0.632	0.282	208&211
Tayta	27	4.35	1.14	210	1.26	0.62	0.936	0.060	40&39
Sussc	28	4.02	1.10	211	0.31	0.33	0.465	0.436	209&210
Hipam	29	4.25	1.17	212	2.21	0.87	0.949	0.046	211&43
Balac	30	1.06	0.54	213	1.41	0.71	0.729	0.208	214&43
Balbon	31	1.02	0.53	214	3.60	1.06	0.994	0.006	213&42
Balph	32	1.50	0.65	215	5.15	1.29	1.0	0.0	29&214
Balmu	33	1.86	0.71	216	2.71	0.95	0.940	0.044	27&215
Escro	34	1.40	0.62	217	1.23	0.63	0.659	0.331	27&202
Balgl	35	0.87	0.50	218	0.78	0.58	0.510	0.456	202&224
Balmy	36	1.27	0.60	219	1.64	0.67	0.992	0.008	22&21
Megno	37	1.61	0.66	220	2.22	0.83	0.983	0.012	223&22
Balbor	38	0.99	0.54	221	0.55	0.40	0.861	0.123	222&24
Baled	39	1.83	0.71	222	0.52	0.39	0.777	0.185	25&221
Capma	40	3.48	0.98	223	2.02	0.80	0.955	0.043	220&222
Stelo	41	0.00	----	224	4.26	1.17	0.998	0.002	220&218
Steat	42	1.62	0.66	225	2.49	0.94	0.704	0.237	227&224
Phyma	43	8.89	1.63	226	3.25	1.02	0.966	0.034	44&46
Equgr	44	0.39	0.39	227	1.70	0.85	0.555	0.435	225&226
Equca	45	0.71	0.48	228	1.79	0.88	0.665	0.229	227&250
Dicbi	46	5.08	1.26	229	0.50	0.40	0.697	0.301	49&47
Phovi1	47	1.55	0.65	230	0.26	0.27	0.537	0.363	47&48
Phovi2	48	1.01	0.53	231	0.31	0.31	0.694	0.227	50&48
Phohi	49	0.10	0.26	232	lower limit		0.829	0.155	231&51
Halgr	50	0.79	0.46	233	0.28	0.28	0.563	0.362	232&234
Phola	51	1.06	0.53	234	0.51	0.38	0.841	0.072	52&233
Phogr	52	2.16	0.77	235	0.87	0.55	0.867	0.120	55&234
Phofa	53	0.80	0.46	236	0.95	0.57	0.858	0.109	235&238
Cyscr	54	1.56	0.65	237	0.26	0.27	0.625	0.333	58&57
Eriba	55	2.17	0.79	238	0.29	0.29	0.557	0.361	237&236
Monsc	56	4.32	1.09	239	1.16	0.59	0.846	0.145	236&243
Hydle	57	0.53	0.38	240	0.60	0.43	0.555	0.407	59&241
Mirle	58	3.80	1.02	241	1.02	0.57	0.916	0.079	61&240
Eumju	59	1.62	0.68	242	2.41	0.88	0.972	0.028	240&63
Zalca	60	1.70	0.70	243	1.86	0.77	0.953	0.038	242&239
Arcfo	61	1.86	0.71	244	1.54	0.71	0.954	0.028	246&243
Arcga	62	1.34	0.60	245	0.75	0.48	0.848	0.145	64&66
Odoro	63	8.64	1.59	246	1.96	0.78	0.989	0.009	244&66
Ursma	64	0.88	0.51	247	2.56	0.90	0.937	0.062	244&249
Ursar	65	1.61	0.67	248	0.88	0.53	0.798	0.135	69&68
Ursam	66	4.18	1.09	249	2.82	0.99	0.990	0.010	248&247
Panle	67	2.47	0.84	250	2.22	0.93	0.879	0.117	247&228
Panti	68	1.57	0.67	251	2.25	0.98	0.807	0.112	228&275
Feldo	69	1.48	0.68	252	0.26	0.26	0.688	0.197	71&70
Europ	70	0.00	----	253	3.41	0.99	1.0	0.0	70&254
Japan	71	0.53	0.37	254	0.75	0.53	0.824	0.120	253&74
Afric	72	0.53	0.37	255	1.76	0.76	0.890	0.072	253&75
Pantr	73	2.11	0.76	256	1.62	0.85	0.644	0.333	76&75
Panpa	74	2.09	0.77	257	14.02	2.16	1.0	0.0	258&76
Gorgo	75	3.13	0.99	258	0.85	0.67	0.428	0.316	270&257
Ponpy	76	7.44	1.52	259	0.61	0.43	0.812	0.152	77&260
Chiim	77	0.53	0.38	260	0.46	0.38	0.625	0.318	259&81
Chivi	78	0.26	0.27	261	0.75	0.46	0.527	0.466	79&260
Chisa	79	1.55	0.67	262	0.59	0.47	0.741	0.170	263&261
Chido	80	1.07	0.53	263	1.74	0.73	0.978	0.017	262&83
Chitr	81	0.00	----	264	9.97	1.81	1.0	0.0	262&265
Plahe	82	1.91	0.75	265	0.90	0.73	0.528	0.344	264&267
Urobi	83	2.87	0.91	266	3.12	1.05	0.950	0.036	86&88
Cavpo	84	8.40	1.64	267	6.21	1.45	0.998	0.002	266&269
Hysaf	85	5.68	1.36	268	7.08	1.53	1.0	0.0	91&90
Scini	86	2.99	0.97	269	1.59	0.82	0.475	0.421	267&268
Sciab	87	3.17	0.99	270	1.95	0.86	0.789	0.132	271&264
Speri	88	2.67	0.99	271	1.92	0.86	0.852	0.131	270&251
Musmu	89	2.87	0.98	272	1.89	0.92	0.920	0.061	84&273
Ratno	90	3.19	1.02	273	2.37	0.96	0.935	0.056	257&272
Orycu	91	8.69	1.66	274	4.66	1.39	0.897	0.087	102&284

Figure 5.7: (a). Branch lengths and LBPs of the ML tree of cytochrome *b*, part 1.

Craca	92	1.27	0.62	275	2.86	1.16	0.538	0.461	251&274
Crata	93	3.51	1.00	276	0.49	0.42	0.662	0.226	92&277
Crago	94	1.24	0.61	277	3.26	0.98	0.999	0.000	276&95
Craru	95	1.82	0.72	278	1.57	0.69	0.986	0.012	276&280
Crafu	96	1.67	0.69	279	0.29	0.30	0.614	0.311	97&96
Cragy	97	1.09	0.55	280	1.16	0.58	0.981	0.010	278&279
Craty	98	1.91	0.73	281	0.26	0.28	0.676	0.241	278&282
Crame	99	3.71	1.06	282	1.50	0.69	0.953	0.041	100&281
Papbu	100	3.88	1.06	283	0.41	0.52	0.632	0.267	99&282
Geobu	101	4.51	1.16	284	11.39	1.98	1.0	0.0	274&99
Dugdu	102	4.89	1.44	285	3.81	1.20	0.982	0.016	291&273
Loxaf	103	16.72	2.43	286	3.07	1.00	0.998	0.002	107&287
Smimu	104	0.00	----	287	0.27	0.37	0.651	0.308	105&286
Plate	105	1.66	0.76	288	1.53	0.71	0.767	0.226	104&286
Plama	106	3.94	1.17	289	5.54	1.44	0.995	0.005	104&110
Plagi	107	0.68	0.48	290	1.74	1.00	0.522	0.309	289&109
Plain	108	1.88	0.77	291	5.93	1.51	1.0	0.0	285&290
Didvi	109	6.51	1.46	292	6.57	1.66	0.977	0.023	363&291
Mondo	110	6.19	1.50	293	1.63	0.81	0.851	0.096	111&113
Geooc	111	4.59	1.31	294	0.78	0.66	0.546	0.397	117&113
Pezwa	112	4.70	1.32	295	2.37	1.00	0.903	0.052	294&297
Melun	113	4.01	1.22	296	3.01	1.12	0.849	0.127	114&116
Plaix	114	2.75	1.12	297	2.75	1.09	0.681	0.291	114&295
Polan	115	9.07	1.91	298	1.13	0.89	0.699	0.180	295&362
Strha	116	7.42	1.71	299	1.36	0.71	0.844	0.135	300&169
Calba	117	16.80	2.62	300	1.21	0.62	0.627	0.368	299&171
Neope	118	4.12	1.17	301	lower limit		0.992	0.007	299&310
Catgu1	119	3.54	1.14	302	0.70	0.58	0.757	0.161	163&303
Catgu2	120	0.83	0.56	303	0.31	0.35	0.627	0.257	164&302
Pomru	121	3.12	1.05	304	0.30	0.31	0.592	0.322	167&303
Pomte	122	0.67	0.48	305	0.62	0.45	0.882	0.111	304&311
Pomis	123	4.46	1.27	306	0.27	0.27	0.726	0.147	178&307
Cyacr	124	4.14	1.11	307	0.54	0.38	0.893	0.090	179&306
Gymti	125	3.86	1.18	308	0.52	0.39	0.835	0.153	182&307
Manke	126	3.24	0.99	309	1.97	0.75	0.992	0.008	168&182
Epifa	127	1.64	0.70	310	1.08	0.56	0.944	0.056	168&301
Dipma	128	1.34	0.62	311	0.27	0.27	0.548	0.423	305&310
Epial	129	1.28	0.67	312	lower limit		0.724	0.276	305&313
Ptipa	130	3.21	0.96	313	1.88	0.79	0.991	0.007	175&312
Ptipl	131	3.10	1.05	314	0.57	0.41	0.864	0.135	312&342
Ambma	132	2.24	0.89	315	1.87	0.83	0.924	0.054	138&142
Ptivi	133	4.18	1.11	316	0.33	0.52	0.563	0.364	138&140
Parin	134	2.32	0.92	317	1.48	0.77	0.778	0.219	316&162
Ailme	135	3.40	1.00	318	1.60	0.78	0.900	0.087	317&323
Virol	136	4.18	1.14	319	3.37	1.15	0.977	0.021	156&155
Lanlu	137	4.58	1.19	320	1.32	0.81	0.732	0.237	319&151
Ampst	138	5.86	1.45	321	3.12	1.18	0.670	0.315	154&320
Colru	139	5.25	1.37	322	2.05	1.03	0.412	0.335	153&154
Pitso	140	5.05	1.36	323	2.34	1.04	0.951	0.042	322&318
Scyma	141	3.79	1.17	324	0.72	0.61	0.747	0.140	139&323
Thrdo	142	2.43	0.95	325	1.00	0.68	0.845	0.148	341&139
Lopny	143	1.37	0.64	326	0.95	0.54	0.659	0.335	147&144
Pavcr	144	2.28	0.82	327	0.44	0.41	0.617	0.234	326&328
Galga	145	1.66	0.68	328	0.29	0.30	0.647	0.285	327&146
Cotco	146	2.15	0.78	329	0.53	0.40	0.746	0.195	327&149
Alech	147	1.77	0.72	330	1.13	0.61	0.858	0.122	329&148
Numme	148	1.12	0.60	331	1.52	0.72	0.705	0.290	150&148
Melga	149	4.37	1.11	332	3.22	1.04	0.998	0.002	161&150
Lopga	150	4.70	1.20	333	1.20	0.71	0.860	0.086	160&161
Cocam	151	1.80	0.88	334	1.43	0.71	0.953	0.039	340&160
Cocer	152	5.05	1.38	335	1.78	0.81	0.883	0.069	158&159
Crosu	153	15.45	2.48	336	6.56	1.54	1.0	0.0	339&158
Cucpa	154	7.61	1.70	337	0.61	0.44	0.754	0.211	173&177
Phacu	155	7.48	1.69	338	1.85	0.80	0.983	0.012	337&174
Piaca	156	5.31	1.40	339	2.15	0.86	0.988	0.011	336&174
OpihoA	157	7.85	1.66	340	0.81	0.58	0.706	0.256	336&334
OpihoB	158	0.99	0.67	341	0.46	0.42	0.716	0.176	325&340
OpihoC	159	0.86	0.58	342	0.27	0.28	0.384	0.404	325&314
Caimo	160	5.62	1.29	343	4.27	1.20	1.0	0.0	361&342
Ortve	161	4.46	1.15	344	1.68	0.76	0.964	0.026	119&347
Empmi	162	3.10	1.02	345	0.99	0.64	0.857	0.143	121&123
Tortr	163	9.03	1.75	346	2.09	0.91	0.947	0.041	351&123
Scoum	164	4.09	1.17	347	0.39	0.38	0.455	0.461	344&346
Peler	165	3.12	1.00	348	lower limit		0.541	0.326	135&133
Balre	166	6.29	1.44	349	1.55	0.73	0.843	0.145	348&134
Phoru	167	3.10	0.99	350	1.19	0.71	0.815	0.094	136&134
Plaal	168	3.48	1.06	351	1.15	0.67	0.742	0.226	346&136
Gypba	169	5.98	1.41	352	1.34	0.69	0.910	0.059	126&344
Vulgr	170	0.96	0.56	353	0.88	0.55	0.697	0.277	354&126
Catbu	171	4.38	1.19	354	0.72	0.56	0.808	0.186	353&137
Corat	172	1.31	0.65	355	1.63	0.70	0.926	0.066	358&354
Gymca	173	3.31	1.03	356	0.37	0.44	0.615	0.381	128&129
Mycib	174	1.95	0.81	357	0.65	0.48	0.836	0.089	356&130
Mycam	175	2.00	0.82	358	1.81	0.74	0.956	0.043	355&130
Lepcr	176	2.36	0.89	359	0.90	0.58	0.637	0.314	124&358
Jabmy	177	1.13	0.60	360	0.77	0.63	0.719	0.158	359&125
Gruru1	178	0.82	0.48	361	3.78	1.15	0.991	0.009	343&125
Gruan	179	0.00	----	362	0.63	0.59	0.471	0.362	343&298
Gruja	180	1.36	0.61	363	13.28	2.24	1.0	0.0	298&292
Gruru2	181	0.00	----	TBL :	869.79		iter: 1		
Gruvi	182	1.67	0.69	ln L:	-18852.56		+- 973.53		
Xenla	183	16.12	2.45	AIC :	38469.12		lower limit: 0.001		

Figure 5.7: (b). Branch lengths and LBPs of the ML tree of cytochrome *b*, part 2.

5.1.4 Phylogeny of Artiodactyla

Hippopotamus is traditionally considered to belong to Suiformes, but does 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 (the chevrotains) appear as a sister group to all the other true ruminants (pecora). The monophyly of pecora is supported with 95% LBP (branch 199), and the monophyly of true ruminants with 83% LBP (branch 201).

The possible paraphyly of Bovidae (species 1–12) has been suggested by the previous analyses of cytochrome *b* sequences (Irwin et al. 1991[126]); Irwin and Árnason 1994[125]), and our analysis also favours the paraphyly. However the support is only 39% LBP (branch 192), and the monophyly of Bovidae has 33% LBP (Fig. 5.7a). 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 can be unstable (e.g., Philippe and Douzery 1994[208]; Adachi and Hasegawa 1996[9]).

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 77% LBP (branch 189) consistently with the previous analyses by Irwin et al. (1991[126] and Irwin and Árnason (1994[125]). Further study is needed to prove or disprove this morphologically unexpected relationship.

5.1.5 Phylogeny of Rodentia

The separate origin of Geomyidae (pocket gophers) from the other rodent groups in Figs. 5.5 and 5.6b is in accord with the NJ analysis of a more limited data set of cytochrome *b* by Philippe and Douzery (1994[208]). Geomyidae, which belongs to Sciuromorpha by traditional taxonomy (Nowak 1991[199]), does not cluster with another Sciuromorpha group, Sciuridae (squirrels), not even with Hystricomorpha or Myomorpha in our analysis. Philippe and Douzery attributed this unexpected placement of Geomyidae to a higher rate of molecular evolution in Geomyidae (DeWalt et al. 1993[58]). Some unusual evolution might well have occurred in the cytochrome *b* gene of Geomyidae.

Within Geomyidae, *Cratogeomys* forms a monophyletic clade in the parsimony and Fitch-Margoliash trees (Fitch and Margoliash 1967[70]) of DeWalt et al. (1993[58]) as well as in our NJ tree, while *C. merriami* is an outgroup to all the other pocket gophers including *Pappogeomys* and *Geomys* in the ProtML tree. The relevant LBP is low (63%: branch 283) and the LBP of *Cratogeomys*-monophyly is 10%. Further studies are needed to settle the issue.

Our analysis support a *Cavia*/*Hystrix* clade with 92% LBP (branch 272), consistently with Ma et al. (1993[176]) and with Cao et al. (1994[42]). The close relationship between the South American and the African Hystricomorpha is in accord with the hypothesis that South American rodents originated in Africa (Wyss et al. 1993[264]).

The ProtML analysis of cytochrome *b* by Cao et al. (1994[42]) gave a rodent-monophyly tree with a Myomorpha/Caviomorpha clade. Although Fig. 5.5 gives a tree similar to the rodent-polyphyly hypothesis proposed by Graur et al. (1991[85]), the relevant branches are very poorly supported. Given the abundant database of other sequences relevant to this problem (Cao et al. 1994[42]; Kuma and Miyata 1994[160]; Frye and Hedges 1995[71]; Martignetti and Brosius 1993[180]), Graur et al.'s hypothesis seems unlikely.

5.1.6 Phylogeny of Microchiroptera

The five species of *Chiroderma* form a monophyletic clade in Fig. 5.5, and *Platyrrhinus* is a sister-group to *Uroderma* with 98% LBP (branch 263; Fig. 5.6b).

5.1.7 Phylogeny of Carnivora

Our ProtML tree suggests a *Arctocephalus*/sea lion clade (97% LBP: branch 242) which is a sister-group to *Odobenus* (walrus) (95% LBP: branch 243) in accord with Árnason et al. (1995[18]). Within the northern phocids, *Erignathus barbatus* (bearded seal) is an outgroup to all the others with 86% 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 95% LBP (branch 244). Although some morphologists maintain independent origins for phocids and otariids (e.g., Tedford 1976[246]), our result is consistent with both previous molecular studies (Vrana et al. 1994[256]; Árnason et al. 1995[18]) and recent morphological studies (Wyss 1988[262]; Wyss and Flynn 1993[263]).

The Pinnipedia are a sister-group to *Ursus* with 94% LBP (branch 247) leaving the *Felis*/*Panthera* clade as an outgroup to the other Carnivora (Vrana et al. 1994[256]; Árnason et al. 1995[18]; Lento et al. 1995[170]).

5.1.8 Phylogeny of Other Mammals

The association of *Loxodonta* (elephant) with *Dugong* is supported with 90% LBP (branch 274; Fig. 5.6b) in accord with Irwin and Árnason (1994[125]), Kleinschmidt et al. (1986[149]), Springer and Kirsch (1993[230]), Porter et al. (1996[211]) and Stanhope et al. (1996[231]).

The ProtML tree in Fig. 5.5 places Perissodactyla as a sister-group to the Cetacea/Artiodactyla clade with 66% LBP (branch 228; Fig. 5.6a). However, the LBP is low and this relationship might not be true, because a recent addition of the cat (*Felis catus*) data (database accession number U20753) to the complete mtDNA sequence data set presented in Table 2.9 suggests that Perissodactyla is closer to Carnivora rather than to Cetacea/Artiodactyla.

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 100% LBP (branch 289) in the ProtML tree.

5.1.9 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. 5.5. 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. 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[60]) and by Helm-Bychowski and Cracraft (1993[111]) and with Sibley and Ahlquist (1990[228]). In the NJ tree of Fig. 5.3, 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 95% LBP (branch 334; Fig. 5.6c) in accord with Sibley and Ahlquist's (1990[228]) classification based on DNA-DNA hybridization. The place of *Opisthocomus hoazin* is obscured by this analysis as in Avise et al. (1994[28]).

The most important feature of the Aves part of Fig. 5.5 might be that Falconiformes, Ciconiiformes, Pelicaniformes, and Phoenicopteriformes are intermixed on the tree, consistently to some extent with Sibley and Ahlquist's (1990[228]) 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 (99% LBP: branch 313) in accord with Avise et al. (1994[27]), no other clade in this group is strongly supported, and therefore no resolution of branching order is attainable from just the cytochrome *b* data. Given that the overall features of the ProtML tree of cytochrome *b* are 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[228]) 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, further 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[110]) analysis of mitochondrial ribosomal RNAs, although the number of relevant species they used is less than that of ours.

5.1.10 Phylogeny of Galliformes

The Galliformes part of the tree is mostly consistent with that of Kornegay et al. (1993[154]). The sister-group of *Ortalis vetula* (chachalaca) to all the other Galliformes analyzed in this work is supported with 93% LBP (branch 333; Fig. 5.6c).

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[132]; Jollès and Jollès 1984[133]; Malcolm et al. 1990[178]). 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[124]) 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 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. Our Fig. 5.5 is consistent with Kornegay et al.'s tree, but the outgroup position of the New World quail is only poorly supported (70% LBP: branch 331), and the lysozyme tree has 29% LBP (Fig. 5.7b). Avise et al. (1994[28]) 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.

Placement of the New World quail outside phasianoids, turkey and guinea fowl as suggested by Sibley and Ahlquist (1985[227]) and by Kornegay et al. (1993[154]) implies that coupled amino acid replacements of lysozyme 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 for lysozyme has been demonstrated by Stewart et al. (1987[235]) 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 convincing evidence for such highly interesting evolution. We believe that further studies are needed to clarify this.

5.1.11 Phylogeny of Fishes

Fig. 5.8 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 mtREV24-F model. Starting from this tree, the search for better tree topologies by the likelihood criterion was conducted by repeated local rearrangements as described in subsection 3.4.3. Fig. 5.9 gives the ProtML tree (based on the mtREV24-F model) which cannot be improved by local rearrangements. The log-likelihood of the NJ tree is -4687.1, while that of the resultant ProtML tree is -4680.3, and the two trees do not differ 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 92% LBP (branch 49; Fig 5.10). Perciformes is monophyletic with 99% LBP (branch 47). Within Perciformes, a ((*Sarda sarda*, *Thunnus thynnus*), *Scomber scombrus*) clade is supported with 100% LBP (branch 43) in accord with Cantatore et al. (1994[39]).

Within Chondrichthyes, Heterodontiformes is closer to Carcharhiniformes than to Lamniformes with 81% LBP (branch 57), and the outgroup status of Heterodontiformes to all the others has only 13% LBP (Fig. 5.11). These three orders of Chondrichthyes are monophyletic, respectively, in accord with Martin and Palumbi (1993[181])

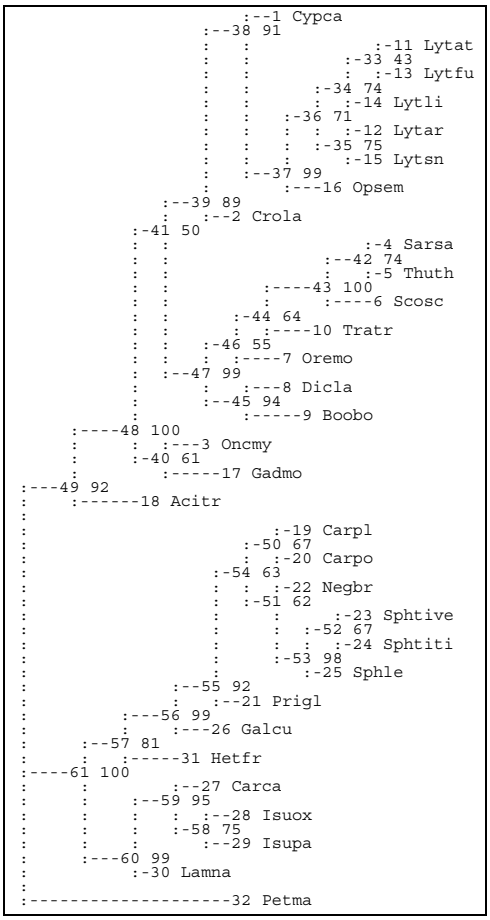


Figure 5.10: The ProtML tree of fish cytochrome b obtained by repeating local rearrangements (mtREV24-F model).

No.1	ext.	branch	S.E.	int.	branch	S.E.	LBP	2nd	pair
Cypca	1	2.57	0.90	33	lower limit		0.426	0.292	11&14
Crola	2	3.44	1.05	34	lower limit		0.743	0.255	35&14
Oncmy	3	4.52	1.20	35	0.27	0.27	0.746	0.127	12&34
Sarsa	4	0.00	----	36	0.60	0.56	0.713	0.273	34&16
Thuth	5	0.55	0.39	37	3.43	1.03	0.993	0.007	36&1
Scosc	6	5.53	1.34	38	1.96	0.81	0.911	0.066	1&2
Oremo	7	6.09	1.37	39	2.14	0.88	0.892	0.103	38&47
Dicla	8	3.66	1.10	40	1.62	0.83	0.613	0.358	41&17
Boobo	9	7.27	1.53	41	0.40	0.45	0.496	0.270	39&40
Tratr	10	6.60	1.48	42	2.60	0.99	0.740	0.255	4&6
Lytat	11	0.80	0.46	43	6.58	1.47	1.0	0.0	42&10
Lytar	12	0.27	0.27	44	1.12	0.76	0.640	0.321	7&10
Lytfu	13	0.27	0.27	45	3.38	1.09	0.941	0.059	46&9
Lytli	14	0.53	0.38	46	0.75	0.58	0.550	0.232	7&45
Lytsn	15	0.80	0.46	47	2.69	0.98	0.990	0.007	39&46
Opsem	16	3.80	1.07	48	5.36	1.45	0.996	0.004	18&40
Gadmo	17	8.69	1.66	49	5.15	1.56	0.922	0.066	48&61
Acitr	18	9.06	1.79	50	0.44	0.40	0.670	0.284	51&20
Carpl	19	1.13	0.57	51	0.55	0.40	0.618	0.354	50&22
Carpo	20	0.89	0.53	52	0.26	0.27	0.669	0.145	25&23
Prigl	21	2.14	0.81	53	1.37	0.64	0.978	0.017	25&22
Negbr	22	1.31	0.63	54	0.62	0.50	0.632	0.296	51&21
Sphtive	23	0.53	0.38	55	2.00	0.86	0.925	0.043	54&26
Sphtiti	24	0.27	0.27	56	3.84	1.18	0.989	0.009	31&26
Sphle	25	0.27	0.27	57	2.08	0.99	0.807	0.134	56&60
Galcu	26	4.03	1.12	58	1.63	0.74	0.749	0.236	27&29
Carca	27	2.47	0.89	59	2.58	0.97	0.952	0.045	27&30
Isuox	28	2.14	0.85	60	4.11	1.25	0.986	0.012	57&30
Isupa	29	2.71	0.93	61	6.67	1.74	0.996	0.004	57&49
Lamna	30	1.77	0.81	TBL :		190.95	iter: 1		
Hetfr	31	7.76	1.59	ln L:		-4680.32	+- 256.64		
Petma	32	34.87	3.76	AIC :		9520.65	lower limit: 0.001		

Figure 5.11: Branch lengths and LBPs of the ProtML tree of fish cytochrome b.

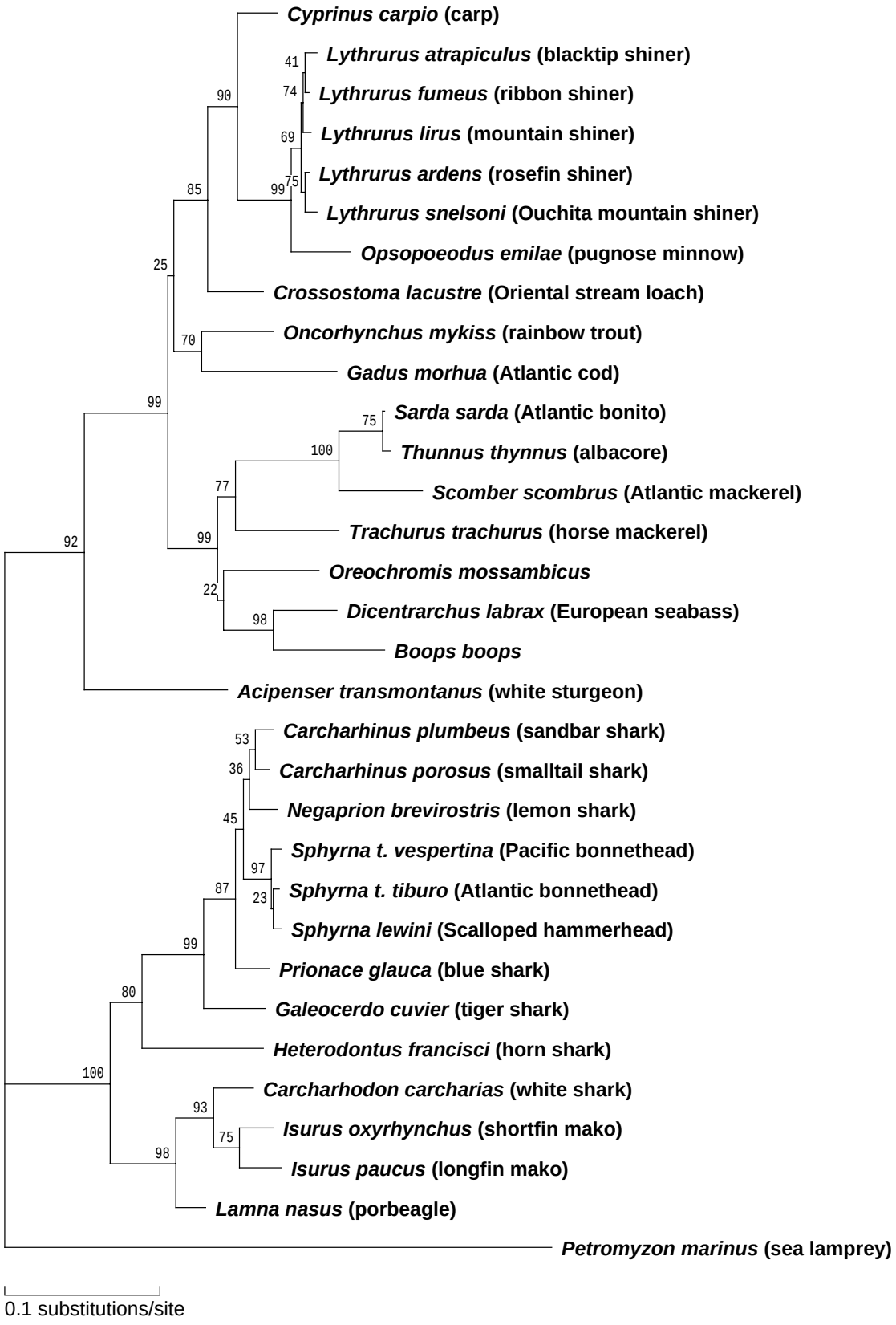


Figure 5.8: The NJ tree of fish cytochrome b in which the branch lengths and LBPs were estimated by the ProtML (mtREV24-F model).

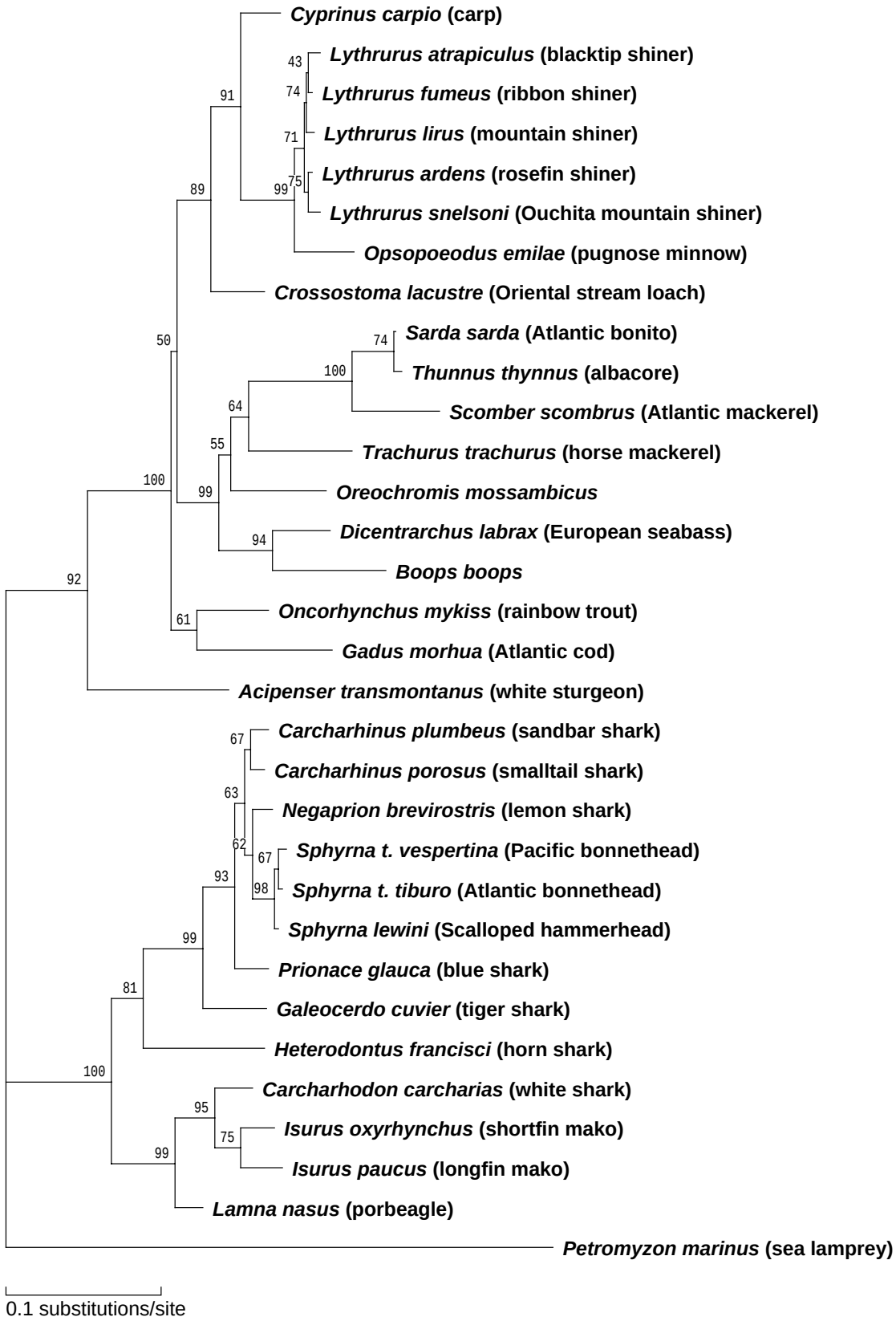


Figure 5.9: The ProtML tree ("protml.eps") of fish cytochrome b obtained by repeating local rearrangements (mtREV24-F model).

5.2 Lysozyme — A Case of Convergent Evolution

Most of the molecular changes during evolution are considered to be selectively neutral (Kimura 1968[143], 1983[146]), but sometimes adaptive evolution does occur (e.g., Stewart et al. 1987[235]; Hughes and Nei 1988[123]). Among the cases of putatively adaptive molecular evolution, the lysozyme protein is interesting to molecular phylogenetics in the sense that the adaptive evolution might mislead phylogeny estimation. A fermentative foregut has evolved independently and convergently in two groups of mammals; i.e., the ruminants in Artiodactyla (for example the cow) and the colobine monkeys in Primates (for example the langur). The appearance of this mode of digestion has been accompanied by the recruitment of lysozyme as a bacteriolytic enzyme in the stomach both in the ruminants and in the colobine monkeys. Stewart et al. (1987[235]) demonstrated that sequence convergence has happened in the amino acid sequence of the stomach lysozymes of the two mammalian lineages, and that such molecular evolution is the basis for these two groups sharing some physicochemical and catalytic properties that adapt their lysozymes for functioning in stomach fluid.

Table 5.1 gives a list of lysozyme sequence data available from the database, and Fig. 5.12 shows its alignment. It is clear that hanuman langur, a species of colobine monkey, independently acquired the same amino acids as present in the stomach lysozymes of ruminants (Bosta2, Caphi1, Caphi2, Oviar2, Axiax); i.e., K, E, D, and N in the 21th, 50th, 75th, and 87th sites.

Figs. 5.14 and 5.15 give the ProtML tree of lysozyme obtained by starting from the NJ tree of Fig. 5.13 and by repeating local rearrangements. In these figures, stomach lysozymes of ruminants (Bosta2, Caphi1, Caphi2, Oviar2, Axiax) and lysozymes of cammel and pig form a monophyletic clade separate from the langur lysozyme (Preen) which is located within the Primate group. On the other hand, when the number of OTUs in the phylogenetic analysis is confined to 6, both the NJ and ProtML analyses give odd results such that the langur clusters with the cow, excluding baboon and human (Figs. 5.17 and 5.18). This is clearly an artifact due to convergent evolution between the ruminants and the langur. When the number of OTUs increases such as in Fig. 5.14, we can get a reasonable tree in spite of the presence of convergent evolution. Indeed, convergent evolution is a serious problem in molecular phylogenetics, and we do not take account of such a possibility in inferring trees using the existing methods of molecular phylogenetics. Therefore, if we encounter an odd tree which drastically contradicts with the traditional view, the possibility of an artifact due to convergent evolution should be considered. Hopefully, when the number of OTUs increases as in Fig. 5.14, we will be safer from such a danger than when we deal with a small number of OTUs.

Table 5.1: List of lysozyme data.

Abbrev.	scientific name	(English name)	database
Bosta2	*Bos taurus*	(bovine 2 rumen)	P04421
Caphi1	*Capra hircus*	(goat 1 rumen)	P37713
Caphi2	*Capra hircus*	(goat 2 rumen)	P37714
Oviar2	*Ovis aries*	(sheep 2 rumen)	P17607
Axiax	*Axis axis*	(axis deer rumen)	P12066
Preen	*Presbytis entellus*	(hanuman langur)	P07232
Cerae	*Cercopithesus aethiops*	(green monkey)	P30200
Macmu	*Macaca mulatta*	(rhesus macaque)	P30201
Papan	*Papio anubis*	(olive baboon)	P00696
Homsa	*Homo sapiens*	(human)	P00695
Camdr	*Camelus dromedarius*	(Arabian camel)	P37712
Bosta1	*Bos taurus*	(bovine 1)	P80189
Oviar1	*Ovis aries*	(sheep 1)	P80190
Sussc1	*Sus scrofa*	(pig 1)	P12067
Sussc2	*Sus scrofa*	(pig 2)	P12068
Sussc3	*Sus scrofa*	(pig 3)	P12069
Equca	*Equus caballus*	(horse)	P11376
Equas	*Equus asinus*	(donkey)	P11375
Orycu	*Oryctolagus cuniculus*	(rabbit)	P16973
Ratno1	*Rattus norvegicus*	(rat 1)	P00697
Ratno2	*Rattus norvegicus*	(rat 2)	Q05820
MusmuM	*Mus musculus*	(mouse M)	P08905
MusmuP	*Mus musculus*	(mouse P)	P17897
Tacac	*Tachyglossus aculeatus*	(echidna)	P37156
Anapl1	*Anas platyrhynchos*	(domestic duck 1)	P00705
Anapl2	*Anas platyrhynchos*	(domestic duck 2)	P00706
Colvi	*Colinus virginianus*	(bobwhite quail)	P00700
Lopca	*Lophortyx californica*	(California quail)	P00699
Numme	*Numida meleagris*	(helmeted guineafowl)	P00704
Galga	*Gallus gallus*	(chicken)	P00698
Chram	*Chrysolophus amherstiae*	(Lady Amherst's pheasant)	P22910
Lople	*Lophura leucomelana*	(kalij pheasant)	P24364
Melga	*Meleagris gallopavo*	(common turkey)	P00703
Pavcr	*Pavo cristatus*	(Indian peafowl)	P19849
Phaco	*Phasianus colchicus*	(ring-necked pheasant)	P00702
Syrre	*Syrmaticus reevesii*	(Reeves' pheasant)	P24533
Ortve	*Ortalis vetula*	(plain chachalaca)	P00707

```
CONSENSUS  KVF.RCELAR .LKRLGLDGY RG.SLANWVC LAK.ESNYNT .ATNYN..D. STDYGIFQIN SRWWCNDGKT
Bosta2     ...E..... T..K..... K.V.....L. .T.W..S... K.....PSSE .......... .K........
Caphi1     ...E..... T..K....D. K.V.....L. .T.W..G... K.....PGSE .......... .KF.......
Caphi2     ...E..... T..E..... K.V.....L. .T.W..S... K.....PGSE .......... .KF.......
Ovlar2     ...E..... T..E..... K.V.....L. .T.W..S... K.....PGSE .......... .K........
Axiax      ...E..... T..E..... K.V.....L. .T.W..S... K.....PGSE .......... .K...D....
Preen      .I.E..... T..K..... K.V..... ...W..G... E.....PG.E .......... ..Y...N...
Cerae      .I.E..... T..... ..I..... ...W..G... Q.....PG.Q .......... .HY...N...
Macmu      .I.E..... T..... ..I..... ...W..... Q.....PG.Q .......... .HY...N...
Papan      .I.E..... T..... ..I..... ...W..D... Q.....PG.Q .......... .HY.......
Homsa      ...E..... T....M... ..I....M. ...W..G... R.....AG.R .......... ..Y.......
Camdr      ..WE..A... K..E..M... ..V.....M. .T.W..D... D.....PSSE .......... ..Y...N...
Bosta1     ...E..... S...F.M.NF ..I.....M. ..RW..... Q.....AG.Q .......... .H........
Oviar1     ...E..... T...F.M..F ..I.....M. ..RW..S... Q.....SG.R .......... .H........
Sussc1     ..YD...F.. I..KS.M... ..V..... ...W..DF.. K.I.R.VGS- -.......... ..Y.......
Sussc2     ..YD...F.. I..KS.M... ..V..... ...W..DF.. K.I.H.VGS- -.......... ..Y.......
Sussc3     ..YD...F.. I..KS.M... ..V..... ...W...F.. K.....PGSQ .......... ..Y.......
Equca      ...SK....H K..AQEM..F G.Y..... M.EY...F.. R.F.GKNANG .S...L..L. NK...K.N.R
Equas      ...SK....H K..AQEM..F G.Y..... M.EY...F.. R.F.GKNANG .Y...L..L. .K...K.N.R
Orycu      .IYE..... T..K..... K.V.....M. ...W..S... R.....PG.K .......... ..Y.......
Ratno1     .IYE..QF.. T...N.MS.. Y.V...D... ..QH..... Q.R...PG.Q .......... ..Y.......
Ratno2     ...KH..... I.RSSA.A.. ..V..E..M. M.QH...FD. E.I...ST.Q .......... ..Y.......
MusmuM     ..YE...F.. T...N.MA.. Y.V...D... ..QH..... R.....RG.Q .......... ..Y.......
MusmuP     ..YN..... I...N.M... ..VK..D... ..QH..... R.....RG.R .......... ..Y.......
Tacac      .ILKKQ..CK N.VAQ.MN.. QHIT.P.... T.FH..S... R...H.T-.G ......L... ..Y..H....
Anapl1     ..YS.....A AM......N. ..Y..G.... A.NY..GF.. Q...R.T-.G ......L... .....DN...
Anapl2     ..YE.....A AM......N. ..Y..G.... A.NY..SF.. Q...R.T-.G .....LE... .....DN...
Colvi      ...G.....A AM..H...N. ..Y..G.... A..F...F.S Q...R.T-.G .....VL... ..........
Lopca      ...G.....A AM..H...N. ..Y..G.... A..F...F.S Q...R.T-.G .....VL... ........R.
Numme      ...G.....A AM..H...N. ..Y..G.... A..F...F.S Q...R.T-.G .....VL... ........R.
Galga      ...G.....A AM..H...N. ..Y..G.... A..F...F.. Q...R.T-.G ......L... ........R.
Chram      ..YG.....A AM......N. ..Y..G.... A..F...F.. H...R.T-.G ......L... ........R.
Lople      ..YG.....A AM......N. ..Y..G.... A..Y...F.. H...R.T-.G ......L... ..........
Melga      ..YG.....A AM......N. ..Y..G.... A..F...F.. H...R.T-.G ......L... ........R.
Pavcr      ..YG.....A AM......N. ..Y..G.... A..F...F.. H...R.T-.G ......L... ........R.
Phaco      ..YG.....A AM..M...N. ..Y..G.... A..F...F.. G...R.T-.G ......L... ........R.
Syrre      ..YG.....A AM......N. ..Y..G.... A..F...F.. H...R.T-.G ......L... ........R.
Ortve      .IYK.....A AM..Y...N. ..Y..G.... A.RY..... Q...R.S-NG ......L... ........R.
                  10         20         30         40         50         60         70

CONSENSUS  PGAVNACHI. CSALL..DIT .AV.CAKRIV SD.QGI.AWV AWR.HC...D VS.YIRGC.L
Bosta2     .N..DG..VS .RE.MEN..A K..A...H.. .E-...T... ..KS..RDH. ..S.VE..T.
Caphi1     .D..DG..VS ..E.MEN..E K..A...H.. .E-...T... ..KS..RDH. ..S.VE..T.
Caphi2     .N..DG..VS ..E.MENN.A K..A...Q.. .E-...T... ..KS..RDH. ..S.VE..T.
Ovlar2     .N..DG..VS ..E.MENN.A K..A...H.. .E-...T... ..KS..RDH. ..S.VE..S.
Axiax      .N..DG..VA ..E.MENN.D K..T...Q.. RE-...T... ..KS..RGH. ..S.VE..T.
Preen      ....D....S .....QNN.A D..A....V. ..P...R... ...N..QNK. ..Q.VK..GV
Cerae      .........S .N...QDN.A D..T....V. R.P...R... ...N..QNR. ..Q.VQ..GV
Macmu      .........S .N...QDN.A D..T....V. ..P...R... ...N..QNR. ..Q.VQ..GV
Papan      .........S .N...QDN.. D..A....V. ..P...R... ...N..QNR. ..Q.VQ..GV
Homsa      ........LS .....QDN.A D..A....V. R.P...R... ...NR.QNR. .RQ.VQ..GV
Camdr      .H...G.G.N .NV..ED... K..Q....V. R.P..VR... ..KN..EGH. .EQ.VE..D.
Bosta1     ........LP .G...QD... Q..A....V. ..P...R... ...S..QNQ. LTS..Q..GV
Oviar1     .........P .....QD... Q..A....V. ..P...R... ...S..QNQ. LTS..Q..GV
Sussc1     .K.......S .KV..DD.LS QDIE....V. R.P...K... ...T..QNK. ..Q.....K.
Sussc2     .K.......S .KV..DD.LS QDIE....V. R.PL.VK... ...A..QNK. ..Q.....K.
Sussc3     .K.......S .KV..DD.LS QDIE....V. R.P...K... ..KA..QNK. ..Q.....K.
Equca      SSS-...N.M ..K..DEN.D DDIS....V. R.PK.MS..K ..VK..KDK. L.E.LAS.N.
Equas      SSS-...N.M ..K..DDN.D DDIS....V. R.PK.MS..K ..VK..KDK. L.E.LAS.N.
Orycu      .R.......P ..D..KD... Q..A....V. ..P...R... ...N..QNQ. LTP.....GV
Ratno1     .R.K...G.P .....QD... Q.IQ....V. R.P...R... ..QR..KNR. L.G...N.GV
Ratno2     .R.....G.P .....QD... Q.IQ....V. R.P...R... ..QR..QNR. L.G...N.GV
MusmuM     .R.....G.N .....QD... A.IQ....V. R.P...R... ...A..QNR. L.Q...N.GV
MusmuP     .RSK...G.N .....QD... A.IQ....V. R.P...R... ...TQ.QNR. L.Q...N.GV
Tacac      ..SK...N.S ..K..DD... DDLK...K.A GEAK.LTP.. ..KSK.RGH. L.KF-K-.--
Anapl1     .RSK...G.P ..V..RS... E..R...... ..GD.MN... ...NR.RGT. ..KW....R.
Anapl2     .R.K...G.P ..V..RS... E..K...... ..GD.MN... ...NR.KGT. ..RW....R.
Colvi      ..SR.L.N.P .....SS... AT.N...K.. ..G-.MN... ...NR.KGT. .QAW....R.
Lopca      ..SR.L.N.P .....SS... AT.N...K.. ..GN.MN... ...NR.KGT. .HAW....R.
Numme      ..SR.L.N.P ....QSS... ATAN...K.. ..GN.MN... ...K..KGT. .RVW.K..R.
Galga      ..SR.L.N.P .....SS... AS.N...K.. ..GN.MN... ...NR.KGT. .QAW....R.
Chram      ..SR.L...P .....SS... AS.N...K.. ..GN.MN... ...NR.KGT. .NAWT...R.
Lople      ..SR.L...P .....SS... AS.N...K.. ..GN.MN... ...NR.KGT. ..VWT...R.
Melga      ..SK.L.N.P .....SS... AS.N...K.A .GGN.MN... ...NR.KGT. .HAW....R.
Pavcr      ..SR.L.N.P .....SS... AS.N...K.. ..RN.MN... ...NR.KGT. .HAW....R.
Phaco      ..SK.L...P .....SS... AS.N...K.. ..GN.MN... ...K..KGT. .NVW....R.
Syrre      ..SR.L...S .....SS... AS.N...K.. ..RN.MN... ...NR.KGT. .NAW....R.
Ortve      ..TK.L...S ....MGA..A PS.R...... ..GD.MN... ...K..KGT. ..TW.KD.K.
                  80         90        100        110        120        130
```

Figure 5.12: The alignment of lysozyme.

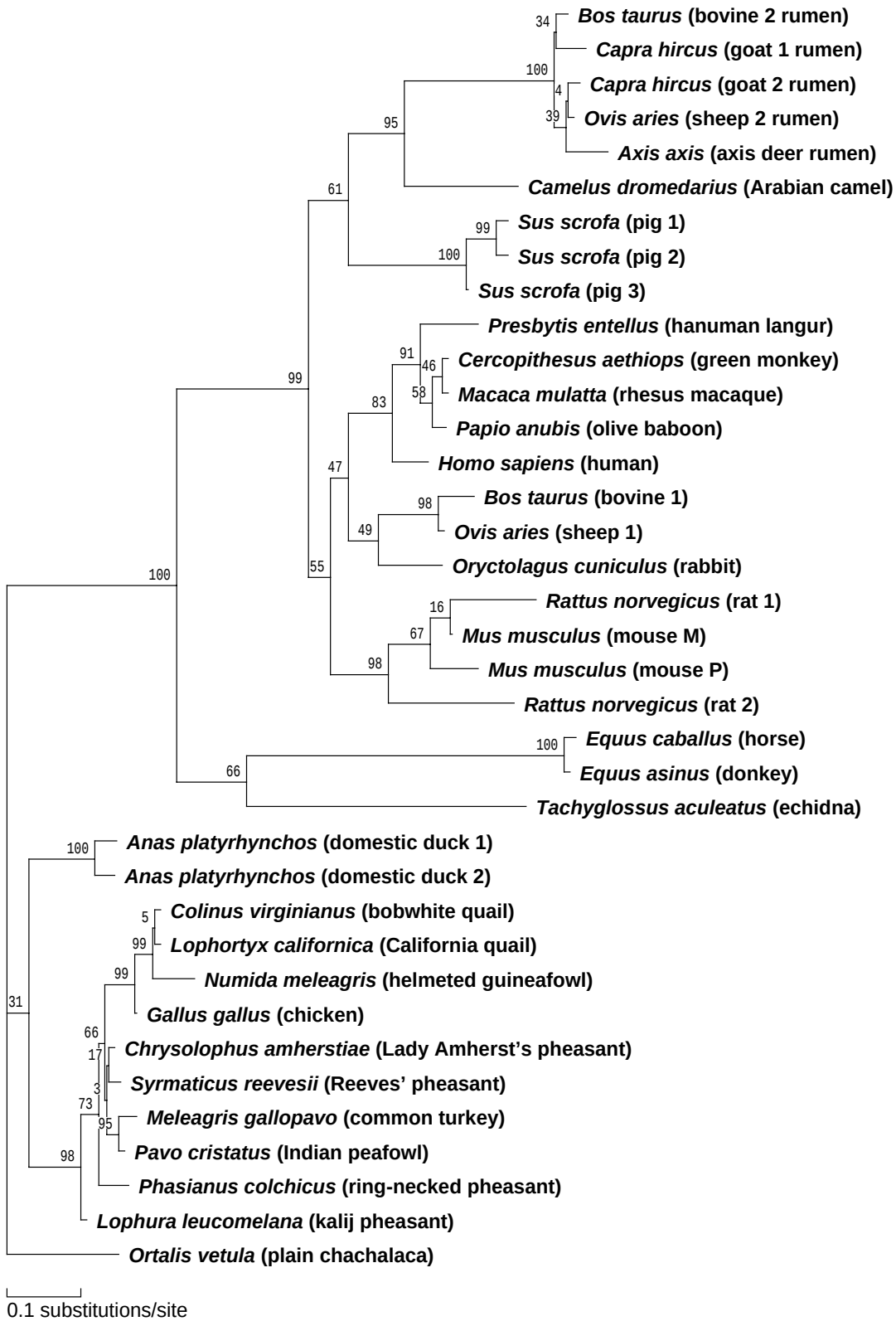


Figure 5.13: NJ tree of lysozyme in which branch lengths and LBPs were estimated by the ProtML (JTT-F model).

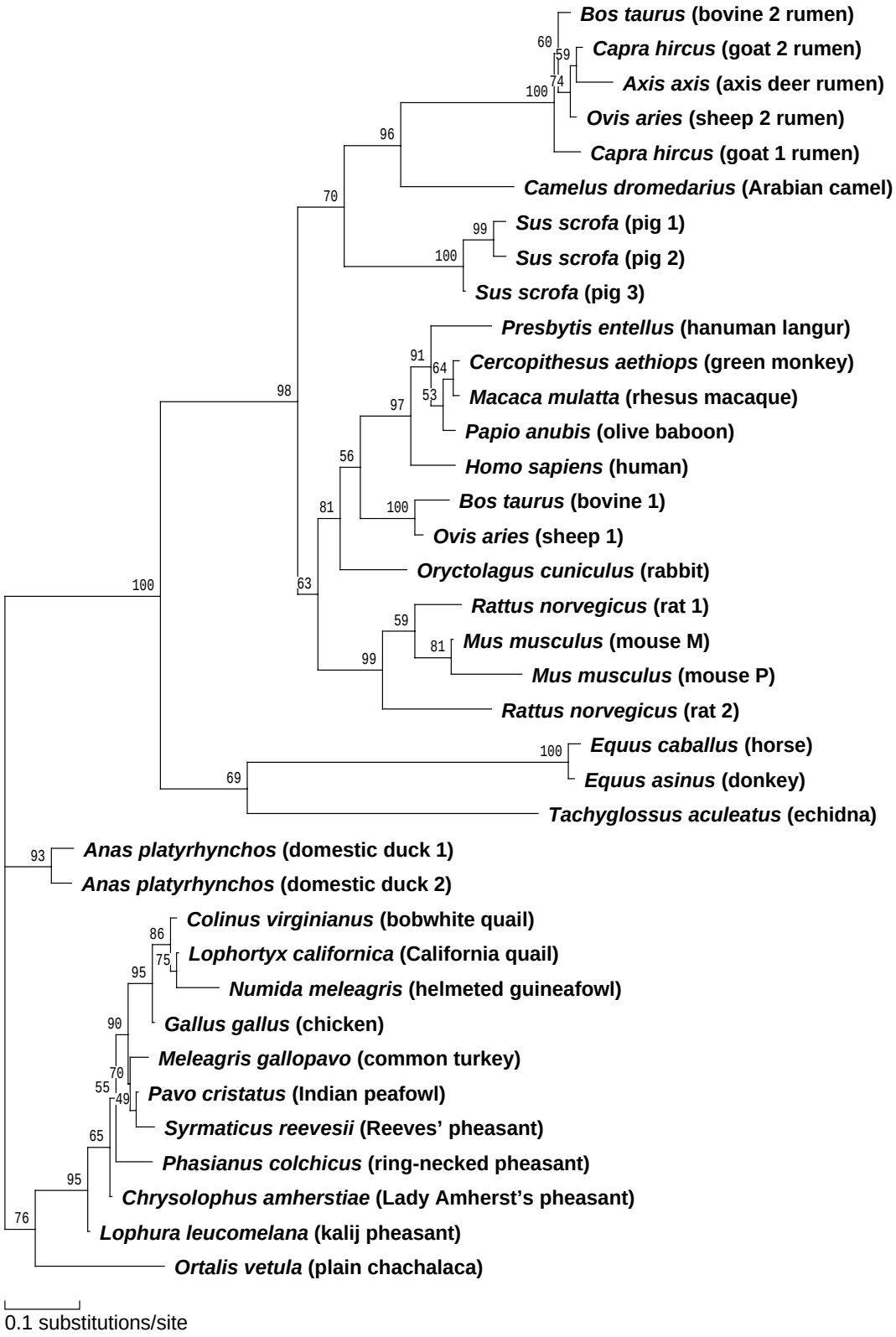


Figure 5.14: ProtML tree of lysozyme obtained by the local rearrangement starting from the NJ tree (JTT-F model).

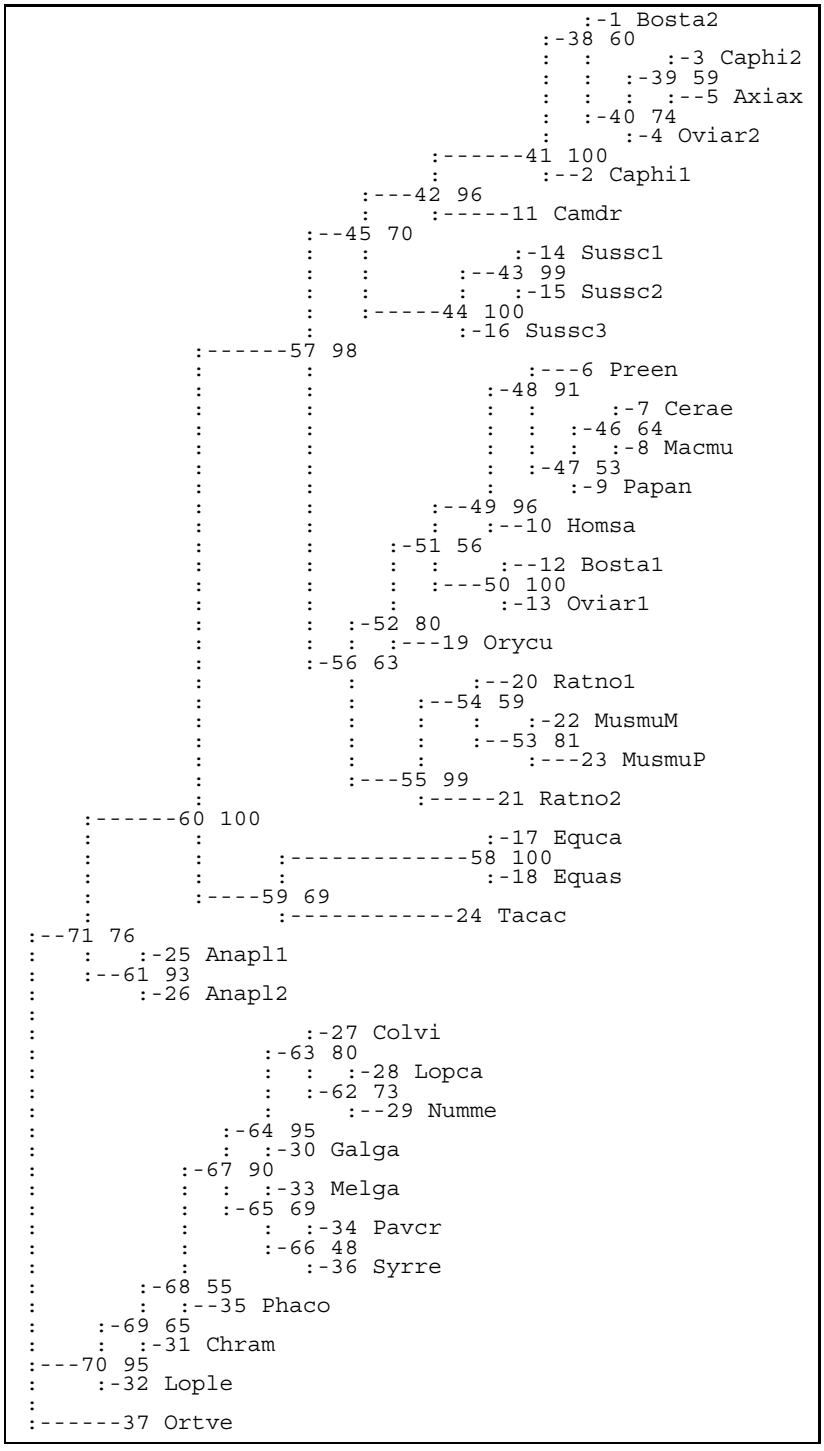


Figure 5.15: The ProtML tree of lysozyme.

No.1	ext.	branch	S.E.	int.	branch	S.E.	LBP	2nd	pair
Bosta2	1	1.55	1.10	38	0.53	0.94	0.599	0.343	1&2
Caphi1	2	3.47	1.73	39	0.76	0.77	0.591	0.388	4&5
Caphi2	3	0.77	0.78	40	1.54	1.09	0.740	0.255	1&4
Oviar2	4	0.77	0.77	41	20.15	4.52	1.0	0.0	11&2
Axiax	5	4.75	1.95	42	7.56	3.08	0.959	0.026	44&11
Preen	6	7.95	2.60	43	4.06	1.84	0.987	0.012	16&15
Cerae	7	0.80	0.80	44	15.74	4.00	0.999	0.001	42&16
Macmu	8	0.77	0.79	45	6.01	2.76	0.698	0.284	42&56
Papan	9	1.72	1.24	46	1.42	1.14	0.644	0.338	9&8
Homsa	10	5.86	2.25	47	1.52	1.18	0.528	0.418	46&6
Camdr	11	14.83	3.94	48	2.65	1.60	0.913	0.080	6&10
Bosta1	12	4.57	1.97	49	6.69	2.49	0.965	0.029	48&50
Oviar1	13	1.04	1.05	50	7.19	2.57	0.996	0.003	49&13
Sussc1	14	1.56	1.14	51	2.58	1.61	0.562	0.424	50&19
Sussc2	15	1.65	1.18	52	2.95	1.95	0.805	0.097	55&51
Sussc3	16	0.00	----	53	4.81	2.12	0.809	0.145	20&22
Equca	17	1.51	1.18	54	4.33	2.06	0.589	0.407	21&20
Equas	18	0.84	0.95	55	8.39	2.97	0.992	0.004	54&52
Orycu	19	8.90	2.87	56	2.74	2.06	0.626	0.234	52&45
Ratno1	20	6.08	2.36	57	17.98	4.88	0.975	0.021	45&59
Ratno2	21	14.42	3.68	58	42.40	7.84	1.0	0.0	17&24
MusmuM	22	0.00	----	59	11.42	4.81	0.691	0.202	57&24
MusmuP	23	9.31	2.83	60	20.46	5.19	0.998	0.002	57&61
Tacac	24	38.33	7.32	61	6.03	2.60	0.929	0.036	60&26
Anapl1	25	3.01	1.72	62	0.65	0.78	0.731	0.231	29&27
Anapl2	26	2.66	1.65	63	2.34	1.35	0.802	0.198	30&27
Colvi	27	0.78	0.78	64	3.11	1.56	0.952	0.048	63&65
Lopca	28	0.13	0.77	65	lower limit		0.694	0.297	66&64
Numme	29	5.46	2.08	66	0.77	0.77	0.485	0.511	33&34
Galga	30	0.00	----	67	1.55	1.10	0.896	0.104	35&64
Chram	31	0.00	----	68	0.77	0.78	0.548	0.452	31&67
Lople	32	0.29	0.84	69	2.82	1.54	0.648	0.348	32&31
Melga	33	2.33	1.35	70	6.92	2.65	0.954	0.036	71&32
Pavcr	34	0.00	----	71	4.09	2.34	0.758	0.200	61&70
Phaco	35	4.78	1.96	TBL :	393.26		iter: 1		
Syrre	36	2.31	1.34	ln L:	-2773.37		+- 164.96		
Ortve	37	17.11	4.03	AIC :	5726.74		lower limit: 0.001		

protml 2.3b3 07/05/96 JTT-F 6 OTUs 130 sites

protdist 2.3b3 07/05/96 6 OTUs 130 sites JTT-F

Figure 5.16: Branch lengths and LBPs of the ProtML tree of lysozyme.

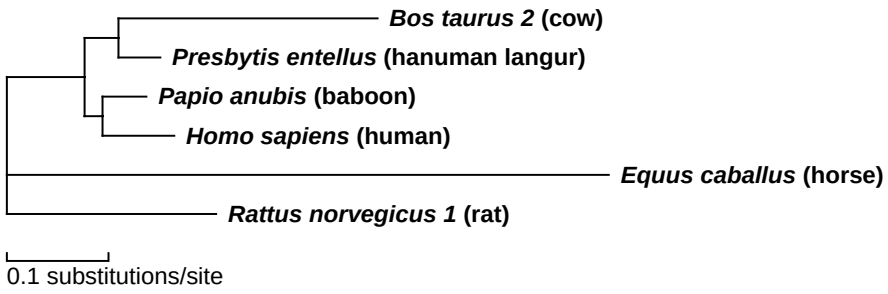


Figure 5.17: NJ tree of 6 lysozyme sequences.

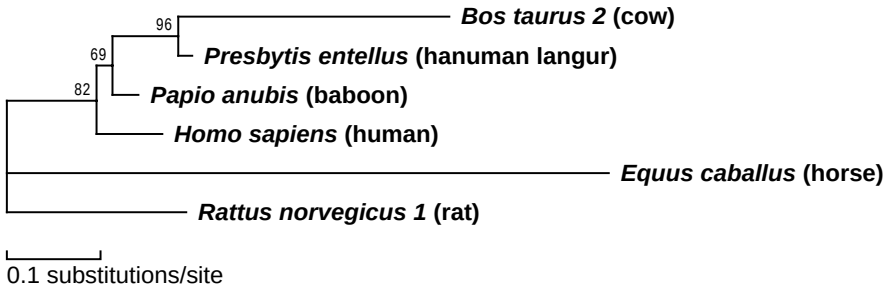


Figure 5.18: ProtML tree of 6 lysozymes obtained by the local rearrangement starting from the NJ tree (JTT-F model).

5.3 Cichlid Fishes in East Africa

The explosive speciation of cichlid fish in the lakes of East Africa has been a focus of much interest among evolutionists. Particularly interesting is that similar highly derived morphologies are found among species in different lakes. These similarities have been variously interpreted either as evidence for migration of ancestral species between the lakes, or of striking convergence of morphology. Molecular phylogenetic studies (Meyer et al. 1990[182]; Kocher et al. 1993[151], 1995[152]) demonstrated that convergent evolution is actually the case.

In this section, we will reanalyze the ND2 data of Kocher et al. (1995[152]) as a further example of application of the NucML program to a real biological problem. The data provided for the analysis (Table 5.2 and alignment in Fig. 5.19) are the 1044 nucleotides of the 31 species of cichlids in Lake Tanganyika and in Lake Malawi both of East Africa. Since we are dealing with relatively closely related species, synonymous substitutions predominate over nonsynonymous ones, and the multiple-hit effect might not be serious. Therefore, we did not distinguish among codon positions in this analysis.

Table 5.2: List of ND2 data of cichlid fish (database accession numbers: U07239–U07270).

Abbrev.	scientific name	(tribe or location)
Pseze	*Pseudotropheus zebra*	(Malawi)
Bucle	*Buccochromis lepturus*	(Malawi)
Chasp	*Champsochromis spilorhynchus*	(Malawi)
Letau	*Lethrinops auritus*	(Malawi)
Rhasp	*Rhamphochromis* sp.	(Malawi)
Lobla	*Lobochilotes labiatus*	(Tropheini)
Petor	*Petrochromis orthognathus*	(Tropheini)
Gnapf	*Gnathochromis pfefferi*	(Limnochromini)
Tromo	*Tropheus moorii*	(Tropheini)
Calma	*Callochromis macrops*	(Ectodini)
Carsc	*Cardiopharynx schoutedeni*	(Ectodini)
Optve	*Opthalmotilapia ventralis*	(Ectodini)
Xenfl	*Xenotilapia flavipinnus*	(Ectodini)
Xensi	*Xenotilapia sima*	(Ectodini)
Chapo	*Chalinochromis popeleni*	(Lamprologini)
Julma	*Julidochromis marlieri*	(Lamprologini)
Telte	*Telmatochromis temporalis*	(Lamprologini)
Neobr	*Neolamprologus brichardi*	(Lamprologini)
Neote	*Neolamprologus tetracanthus*	(Lamprologini)
Lamca	*Lamprologus callipterus*	(Lamprologini)
Lepel	*Lepidiolamprologus elongatus*	(Lamprologini)
Permi1	*Perissodus microlepis 1*	(Perissodini)
Permi2	*Perissodus microlepis 2*	(Perissodini)
Cypfr	*Cyphotilapia frontosa*	(Tropheini)
Tanir	*Tanganicodus irsacae*	(Eretmodini)
Limau	*Limnochromis auritus*	(Limnochromini)
Parbr	*Paracyprichromis brieni*	(Cyprichromini)
Oreni	*Oreochromis niloticus*	(Tilapiini)
Tylpo	*Tylochromis polylepis*	(Tylochromini)
Boumi	*Boulengerochromis microlepis*	(Tilapiini)
Batsp	*Bathybates* sp.	(Bathybatini)
Cicci	*Cichlasoma citrinellum*	(Central America)

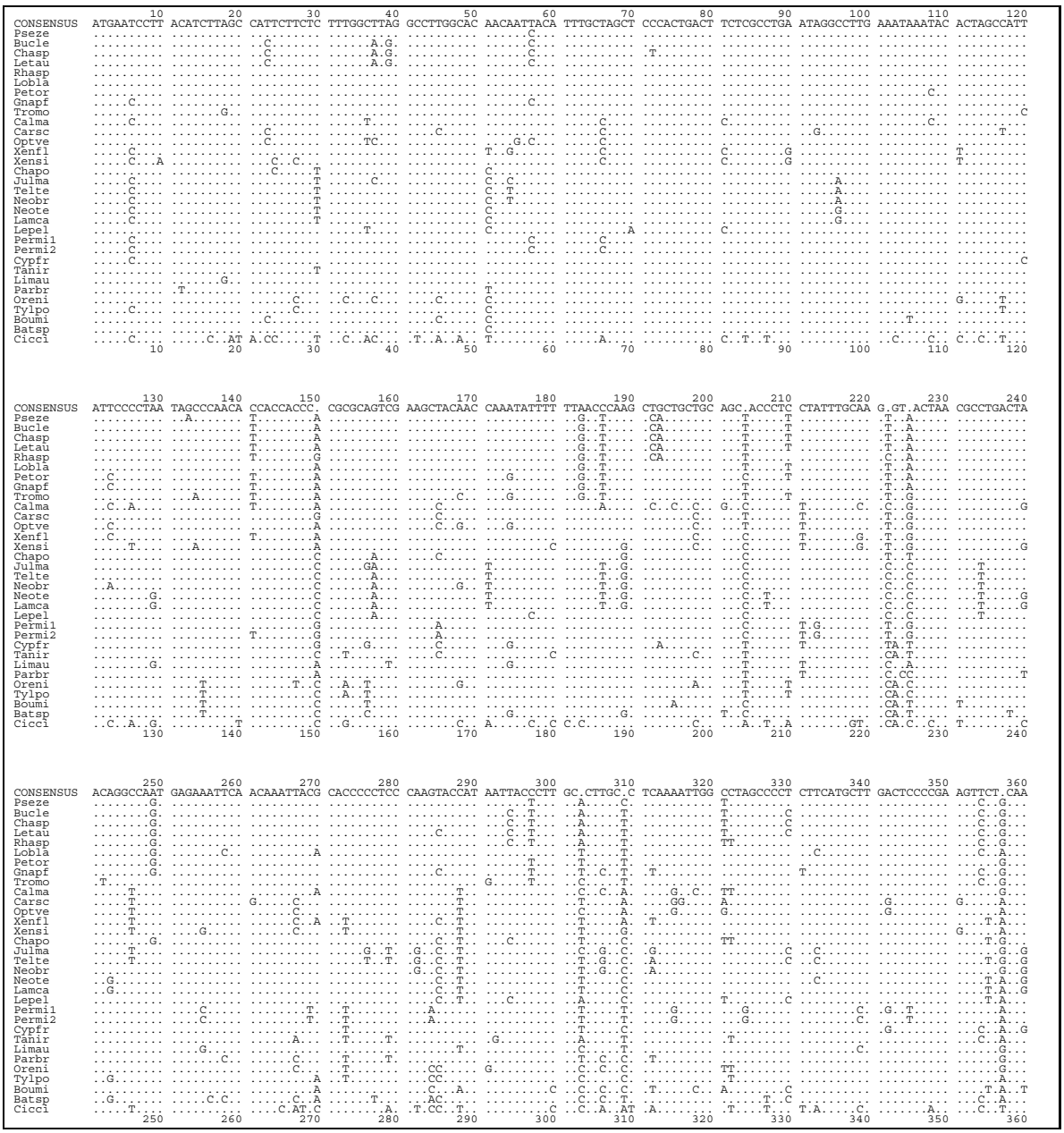


Figure 5.19: (a). The alignment of ND2 of cichlid fishes, part 1.

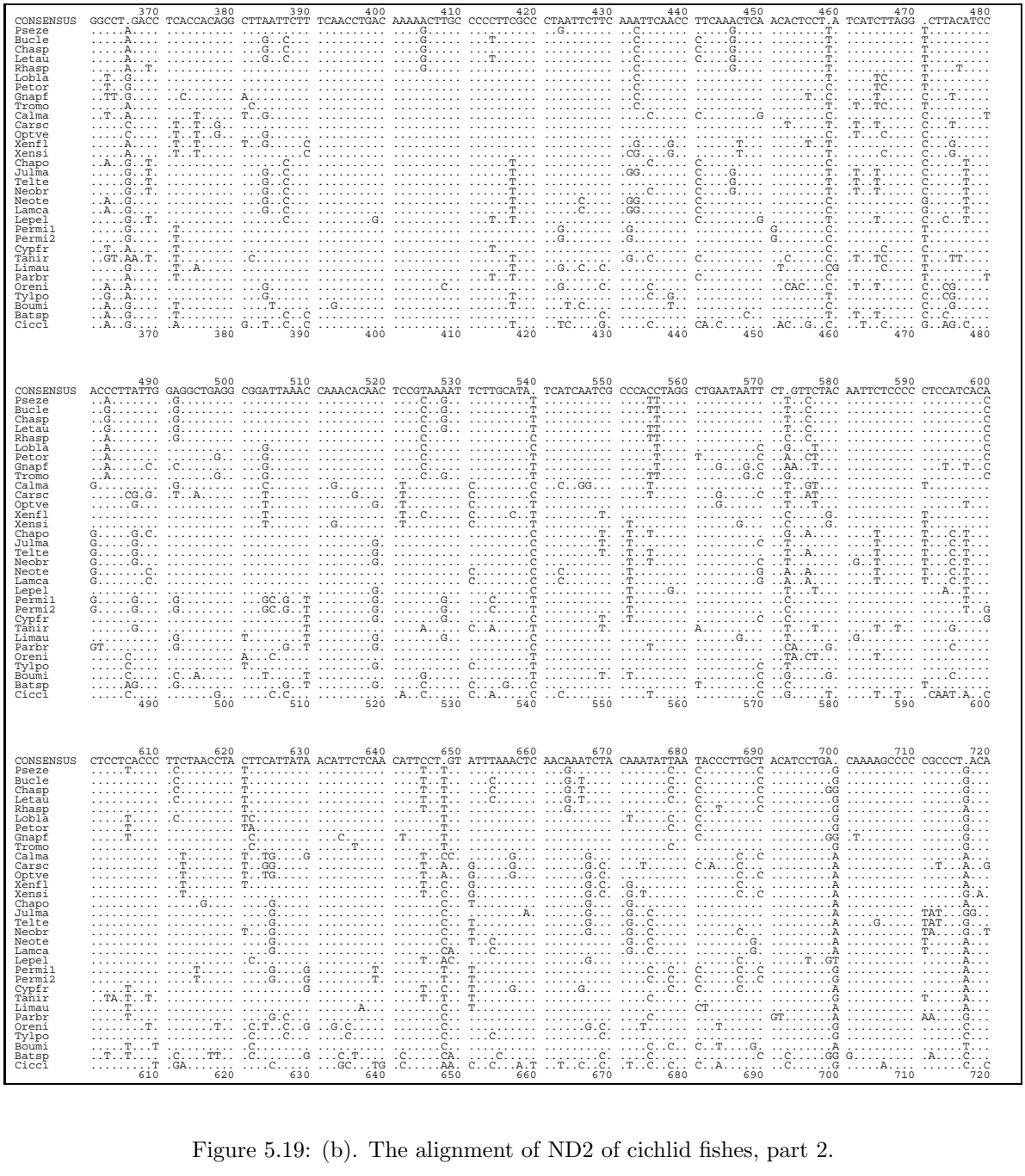


Figure 5.19: (b). The alignment of ND2 of cichlid fishes, part 2.

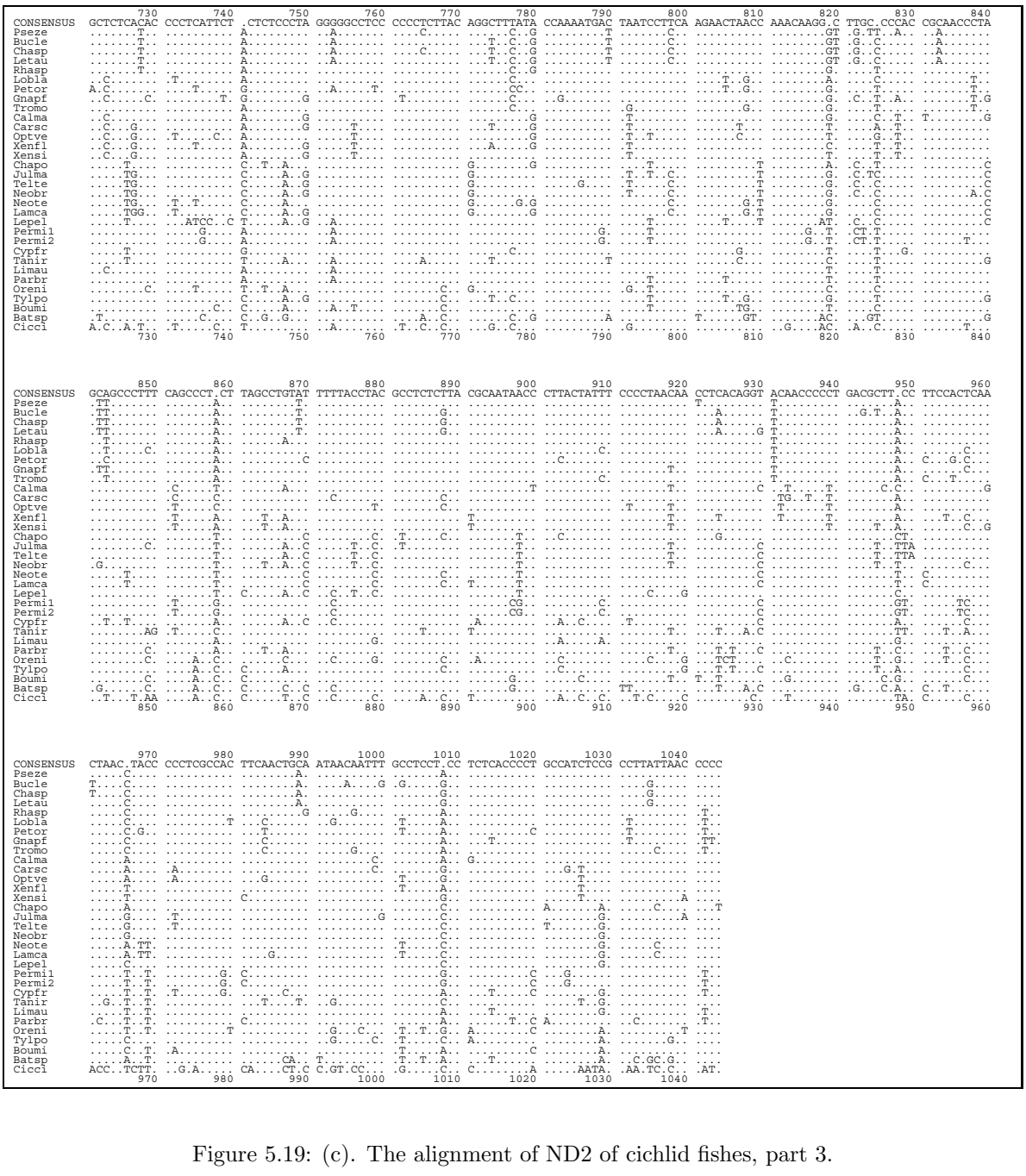


Figure 5.19: (c). The alignment of ND2 of cichlid fishes, part 3.

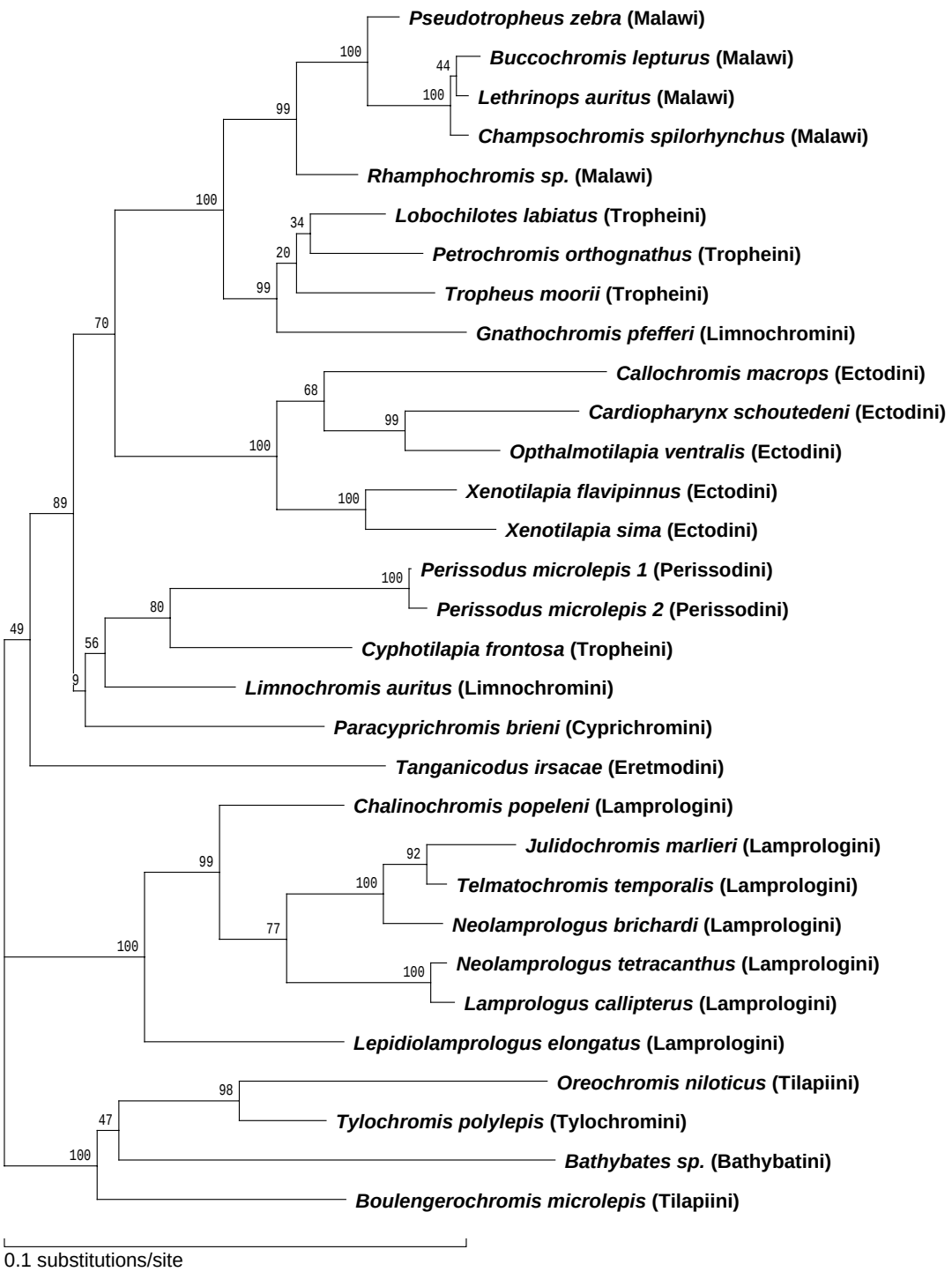


Figure 5.20: NJ tree of ND2 from East African cichlids in which the branch lengths and LBPs were estimated by NucML (HKY85 model; $\alpha/\beta = 6.6$; $\ln L = -7884.4$).

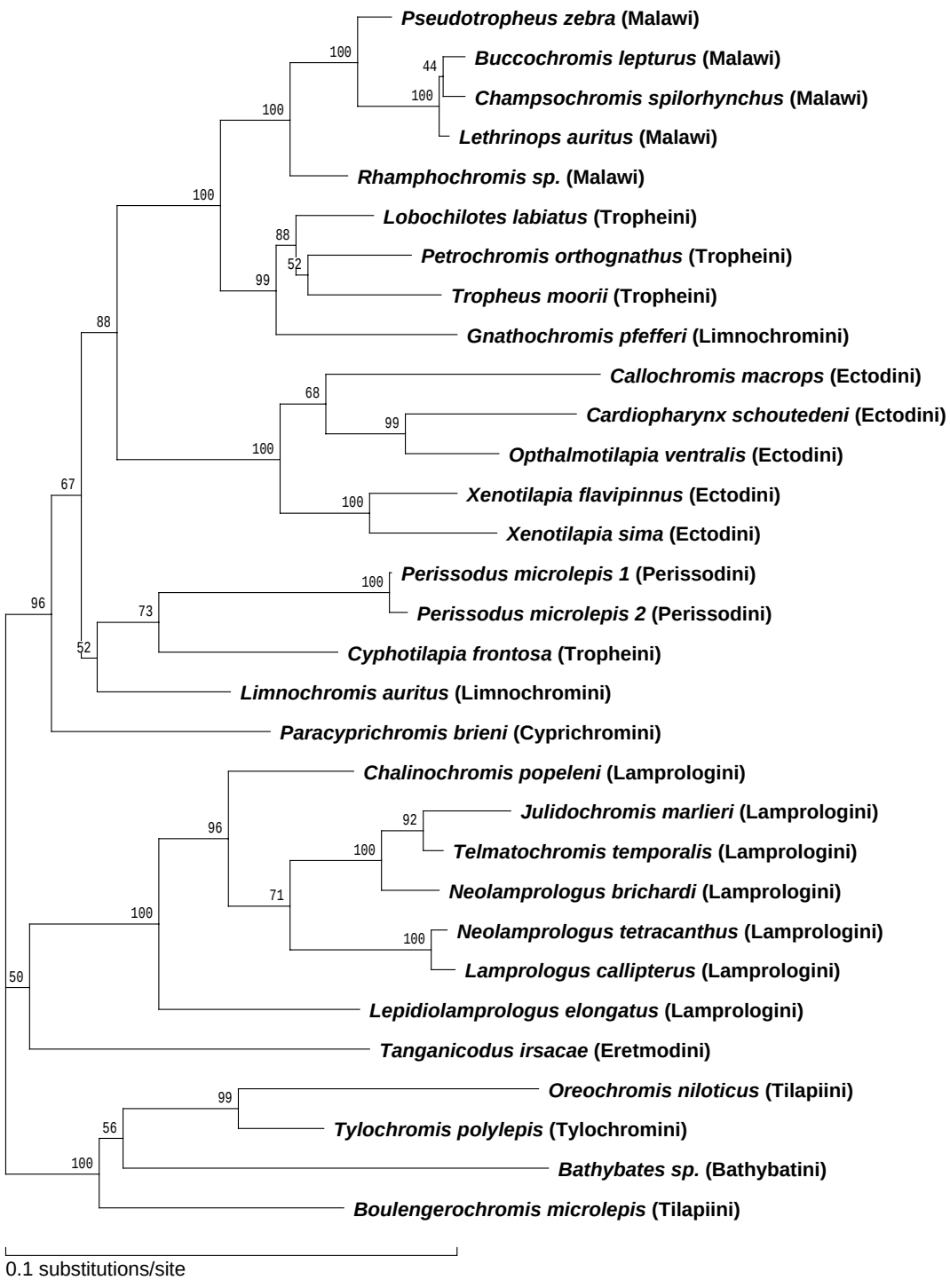


Figure 5.21: NucML tree of ND2 from East African cichlids obtained by replicating the local rearrangements (HKY85 model; $\alpha/\beta = 6.6$; $\ln L = -7879.7$)).

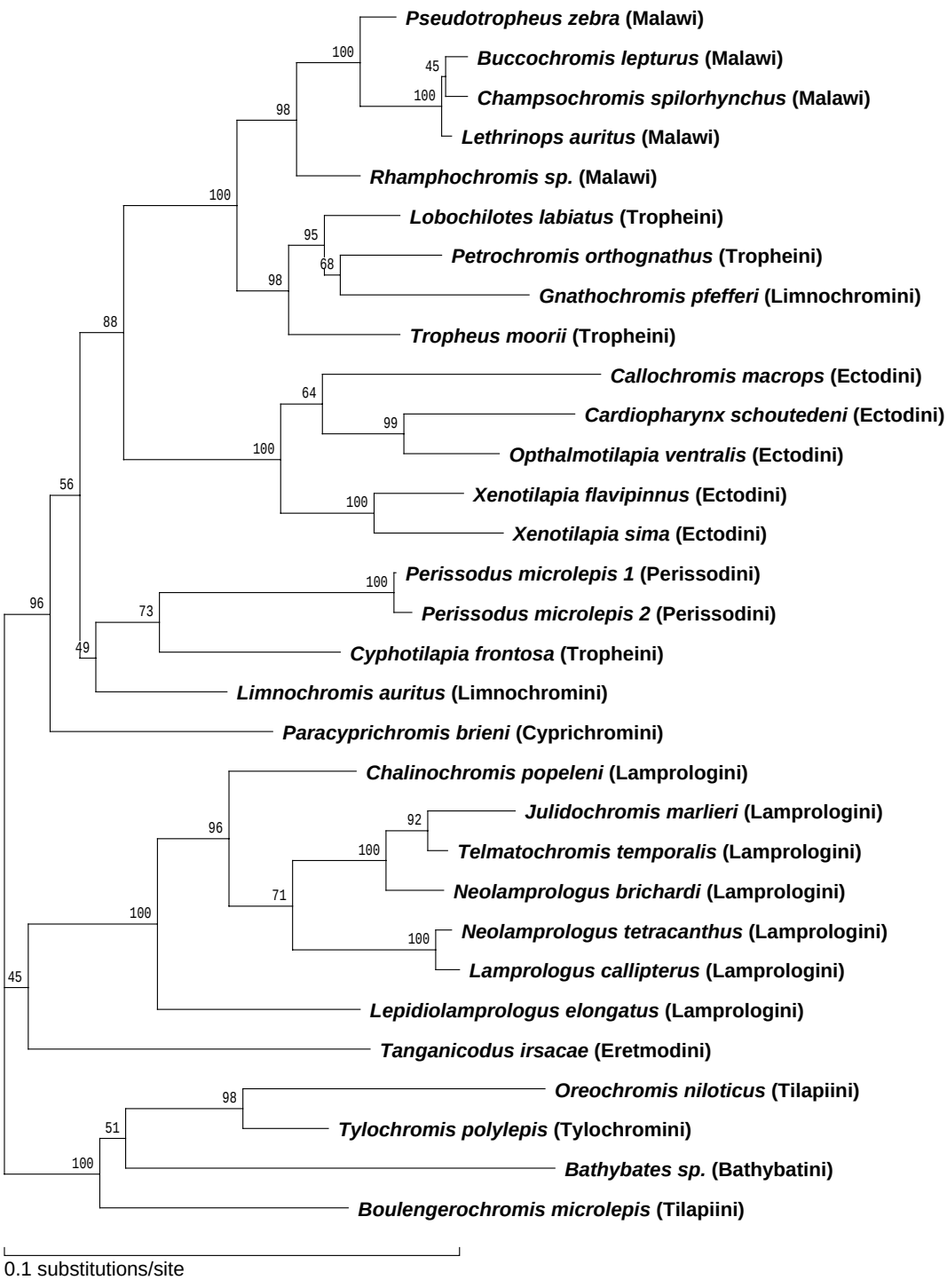


Figure 5.22: NucML tree of ND2 from East African cichlids with log-likelihood higher than that of Fig. 5.22 (HKY85 model; $\alpha/\beta = 6.6$; $\ln L = -7874.0$)).

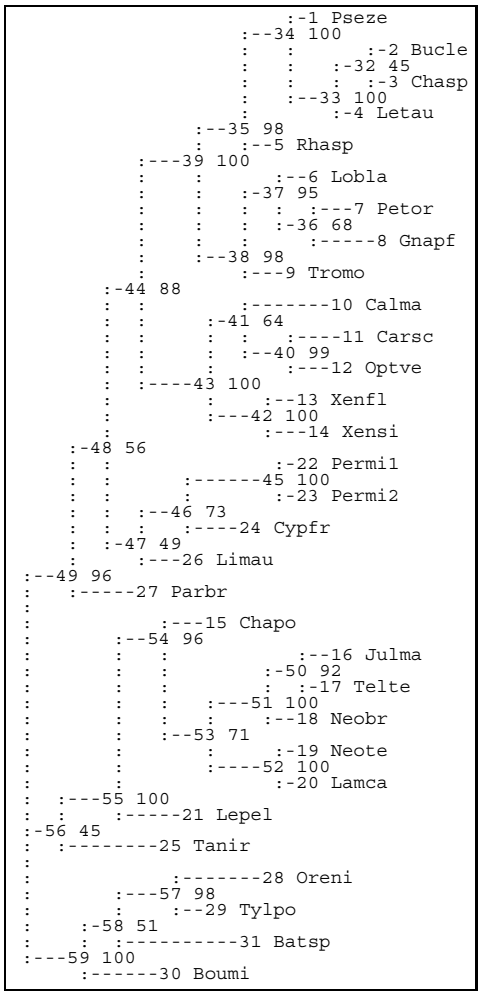


Figure 5.23: The NucML tree of ND2 of cichlid fish (the tree in Fig. 5.22).

No.1	ext.	branch	S.E.	int.	branch	S.E.	LBP	2nd	pair
Pseze	1	0.76	0.29	32	0.09	0.10	0.448	0.435	2&4
Bucle	2	0.49	0.22	33	1.78	0.43	1.0	0.0	32&1
Chasp	3	0.48	0.22	34	1.39	0.39	0.998	0.002	5&33
Letau	4	0.20	0.14	35	1.30	0.40	0.980	0.020	34&38
Rhasp	5	1.38	0.39	36	0.33	0.21	0.676	0.282	7&6
Lobla	6	1.66	0.42	37	0.80	0.32	0.947	0.040	6&9
Petor	7	2.20	0.48	38	1.12	0.37	0.977	0.015	35&9
Gnapf	8	4.15	0.66	39	2.47	0.54	1.0	0.0	35&43
Tromo	9	2.45	0.51	40	1.78	0.49	0.993	0.007	10&12
Calma	10	6.08	0.84	41	0.93	0.37	0.639	0.303	42&40
Carsc	11	3.75	0.65	42	2.04	0.51	0.998	0.002	13&41
Optve	12	2.08	0.50	43	3.43	0.65	1.0	0.0	41&39
Xenfl	13	1.95	0.47	44	0.96	0.39	0.875	0.111	39&47
Xensi	14	2.81	0.55	45	5.15	0.77	1.0	0.0	24&23
Chapo	15	2.79	0.58	46	1.41	0.43	0.731	0.201	26&24
Julma	16	1.94	0.45	47	0.35	0.25	0.486	0.330	46&44
Telte	17	0.43	0.23	48	0.66	0.33	0.565	0.392	44&27
Neobr	18	1.28	0.37	49	1.01	0.37	0.960	0.028	56&27
Neote	19	0.36	0.19	50	0.92	0.31	0.923	0.077	16&18
Lamca	20	0.51	0.23	51	2.04	0.50	1.0	0.0	52&18
Lepel	21	4.43	0.72	52	3.13	0.60	1.0	0.0	51&20
Permi1	22	0.00	----	53	1.38	0.43	0.708	0.290	15&52
Permi2	23	0.38	0.19	54	1.57	0.47	0.964	0.020	21&53
Cypfr	24	3.94	0.68	55	2.82	0.61	0.999	0.001	54&25
Tanir	25	7.49	0.94	56	0.51	0.31	0.453	0.387	49&25
Limau	26	2.90	0.58	57	2.58	0.59	0.984	0.010	31&29
Parbr	27	4.87	0.74	58	0.55	0.39	0.511	0.382	30&31
Oreni	28	6.64	0.88	59	2.11	0.53	1.0	0.0	49&30
Tylpo	29	1.88	0.52	TBL :	129.72	iter: 1			
Boumi	30	5.43	0.80	ln L:	-7874.03	+- 257.37			
Batsp	31	9.40	1.07	AIC :	15874.07				

Figure 5.24: Branch lengths and LBPs of the NucML tree of ND2 (the tree in Fig. 5.22).

In Figs. 5.20 and 5.21, cichlids in Lake Malawi are indicated as "Malawi" in parentheses, and all the others are from Lake Tanganyika. The log-likelihood of the NJ tree is -7884.4, and that of the resultant NucML tree in Fig. 5.21 is -7879.7 (improvement of log-likelihood by 4.7). Although the tree in Fig. 5.21 cannot be improved any more by 1-step local rearrangement, it turned out that the tree of Fig. 5.22 in which *Tropheus moori* and *Gnathochromis pferreri* are transposed has a higher log-likelihood than the tree in Fig. 5.21 by 5.7 ± 9.4. This shows the limitation of 1-step local rearrangements, and more extended rearrangements and/or adoption of alternative initial trees provided to the local rearrangements might be needed in many real problems (e.g., see Swofford 1993[239], PAUP 3.1 manual).

This analysis clearly demonstrates that the 5 Malawi species form a monophyletic clade within the Tanganyika species. In spite of that *Pseudotropheus* and *Rhamphochromis* from Lake Malawi are morphologically very similar, respectively, to *Tropheus* and *Bathybathes* from Lake Tanganyika (Kocher et al. 1993[151]). Furthermore, the cichlids in Lake Malawi are suggested to have derived from an ancestral stock closely related to Tropheini (excluding *Cyphotilapia*) and *Gnathochromis*. These observations are consistent with the previous analyses of Kocher et al. (1993[151], 1995[152]).

5.4 Total Evaluation of ML Analyses of Multiple Genes

Although the analysis of molecular sequence data has become powerful in elucidating the phylogenetic history of organisms, a single gene does not necessarily contain sufficient phylogenetic information to resolve the problem at hand. Therefore, it is necessary to scrutinize as many loci as possible and to evaluate the total evidence. The ML method is particularly suitable for this purpose. Given the model, one can calculate the likelihood as the probability that one tree yielded the observed data, and each gene can reasonably be regarded as evolving independently from other genes. Therefore, the total support for a particular tree can be evaluated by simply summing up the estimated log-likelihoods of individual genes for that tree, and the total log-likelihoods for different trees can then be compared. Importantly, the analyses of tandemly-combined sequences from several genes do not explicitly take into account the differences of tempo and mode of evolution among different genes. On the other hand, if we analyze the different genes separately, we can take into account these differences. We can even evaluate the total evidence combining a ProtML analysis of protein sequences with a NucML analysis of rRNA sequences.

Although insertion/deletion (Thorne et al. 1991[250], 1992[251]; Thorne and Kishino 1992[249]) and gene rearrangements (Sankoff et al. 1992[222]; Boore et al. 1995[36]) are not taken into account in MOLPHY, these data can be analyzed in the framework of the ML, if these events can be represented by adequate models. Thus such data will be able to be included by the total evidence approach, and a preliminary attempt has been done in Kishino et al. (1990[148]). On the other hand, it might be difficult to combine different types of data in the framework of parsimony, because weighting among different types of data must be ambiguous. Therefore, the availability of the total evidence approach might be

one of the most important merits of ML.

We will exemplify how analyses of different genes can be combined in the total evidence approach by using two data sets, hemoglobin α and cytochrome b, among the 10 proteins used in Table 4 of Cao et al. (1994[42]).

From molecular phylogenetic analyses of proteins, Graur et al. (1991[85]) suggested that the order Rodentia may not be monophyletic, and that the guinea pig-like rodents (Caviomorpha) may have a separate evolutionary origin within mammals from that of the rat-like rodents (Myomorpha) and the squirrel-like rodents (Sciuromorpha). They further suggested that the Caviomorpha separated from other rodents before the divergence among Rodentia, Primates and Artiodactyla. Their suggestion contradicts the traditional view of rodent monophyly based mainly on comparative morphology (Luckett and Hartenberger 1985[174], 1993[175]; Novacek 1992[198]).

They used parsimony in estimating the tree, but it is known that the parsimony method is sometimes misleading particularly when the evolutionary rate differs among lineages (Felsenstein 1978[62]) or even if there is a molecular clock (Hendy and Penny 1989[112]). Therefore, Cao et al. (1994[42]) re-examined their data, as well as additional data, with ProtML which is robust against the violation of rate constancy (Hasegawa and Fujiwara 1993[92]). The overall evidence did not support Graur et al.'s hypothesis and supported the traditional view of rodent monophyly. Cao et al.'s analysis suggests that Graur et al.'s conclusion is due to an artifact of the parsimony method caused by rapid molecular evolution in the guinea pig lineage.

The sequence data file and toplogy file for hemoglobin α are shown in Figs. 5.25 and 5.26.

By submitting the command,

```
protml -jf -l hba hba.ptn hba.tpl > hba.ml
```

ProtML analysis is carried out with the JTT-F model and we obtain "hba.ml" which is shown in Fig. 5.27, and "hba.lls" which gives the estimated log-likelihood for each site as shown in Fig. 5.28 and will be used in the total evidence approach later.

The printout of the protml.eps file of the ML tree by this analysis is given in Fig. 5.29.[1]

The sequence data file and toplogy file for cytochrome b are shown in Figs. 5.30 and 5.31. Then, using a command,

```
protml -mf -l cytb cytb.ptn cytb.tpl > cytb.ml
```

ProtML analysis with the mtREV24-F model is carried out for the cytochrome b data, and "cytb.ml" file is obtained as in Fig. 5.32, and estimated log-likelihoods for each site are stored in the "cytb.lls" file (not shown).

[1] In the user's tree option, only the first tree is stored in the protml.eps or nucml.eps file.

```
 12 141 alpha-globin
Oan Ornithorhynchus anatinus (platypus)
MLTDAEKKEVTALWGKAAGHGEEYGAEALERLFQAFPTTKTYFSHFDLSHGSAQIKAHGK
KVADALSTAAGHFDDMDSALSALSDLHAHKLRVDPVNFKLLAHCILVVLARHCPGEFTPS
AHAAMDKFLSKVATVLTSKYR
Tac Tachyglossus aculeatus (Australian echidna)
VLTDAEKKEVTSLWGKASGHAEEYGAEALERLFLSFPTTKTYFSHMDLSKGSAQVKAHGK
RVADALTTAAGHFNDMDSALSALSDLHAHKLRVDPVNFKLLAHCFLVVLARHHPAEFTPS
AHAAMDKFLSRVATVLTSKYR
Dma Didelphis marsupialis (North American opossum)
VLSANDKTNVKGAWSKVGGNSGAYMGEALYRTFLSFPTTKTYFPNYDFSAGSAQIKTQGQ
KIADAVGLAVAHLDDMPTALSSLSDLHAHELKVDPVNFKFLCHNVLVTMAAHLGKDFTPE
IHASMDKFLASVSTVLTSKYR
Mgi Macropus giganteus (eastern gray kangaroo)
VLSAADKGHVKAIWGKVGGHAGEYAAEGLERTFHSFPTTKTYFPHFDLSHGSAQIQAHGK
KIADALGQAVEHIDDLPGTLSKLSDLHAHKLRVDPVNFKLLSHCLLVTFAAHLGDAFTPE
VHASLDKFLAAVSTVLTSKYR
Dvi Dasyurid viverrinus (southeastern quoll)
VLSDADKTHVKAIWGKVGGHAGAYAAEALARTFLSFPTTKTYFPHFDLSPGSAQIQGHGK
KVADALSQAVAHLDDLPGTLSKLSDLHAHKLRVDPVNFKLLSHCLIVTLAAHLSKDLTPE
VHASMDKFFASVATVLTSKYR
Mau Mesocricetus auratus (golden hamster)
VLSAKDKTNISEAWGKIGGHAGEYGAEALERMFFVYPTTKTYFPHFDVSHGSAQVKGHGK
KVADALTNAVGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLANHHPADFTPA
VHASLDKFFASVSTVLTSKYR
Mmu Mus musculus (mouse)
VLSGEDKSNIKAAWGKIGGHGAEYGAEALERMFASFPTTKTYFPHFDVSHGSAQVKGHGK
KVADALASAAGHLDDLPGALSALSDLHAHKLRVDPVNFKLLSHCLLVTLASHHPADFTPA
VHASLDKFLASVSTVLTSKYR
Cpo Cavia porcellus (guinea-pig)
VLSAADKNNVKTTWDKIGGHAAEYVAEGLTRMFTSFPTTKTYFHHIDVSPGSGDIKAHGK
KVADALTTAVGHLDDLPTALSTLSDVHAHKLRVDPVNFKFLNHCLLVTLAAHLGADFTPS
IHASLDKFFASVSTVLTSKYR
Lta Loris tardigradus (slender loris)
VLSPADKTNVKTAWEKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKAHGK
KVADALTTAVSHVDDMPSALSALSDLHAHKLRVDPVNFKLLSHCLLVTLACHHPADFTPA
VHASLDKFLASVSTVLTSKYR
Age Ateles geoffroyi (spider monkey)
VLSPADKSNVKAAWGKVGGHAGDYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHHPADFTPA
VHASLDKFLASVSTVLTSKYR
Cae Cercopithecus aethiops (green monkey)
VLSPADKSNVKAAWGKVGGHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTLAVGHVDDMPHALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
Hsa Homo sapiens (human)
VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGK
KVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPA
VHASLDKFLASVSTVLTSKYR
```

Figure 5.25: Hemoglobin α sequence data ("hba.ptn" file).

```
   3
(((Oan,Tac),((Mgi,Dvi),Dma)),(Lta,(Age,(Cae,Hsa))),(Cpo,(Mmu,Mau)));
(((Oan,Tac),((Mgi,Dvi),Dma)),(Mmu,Mau),(Cpo,(Lta,(Age,(Cae,Hsa)))));
(((Oan,Tac),((Mgi,Dvi),Dma)),Cpo,((Mmu,Mau),(Lta,(Age,(Cae,Hsa)))));
```

Figure 5.26: Tree topology file of hemoglobin α ("hba.tpl" file).

```
protml 2.3b3 (07/05/96) JTT-F 12 OTUs 141 sites. alpha-globin
#1
                  :----1 Oan
   :----------13
   :              :---2 Tac
:-16
:  :             :----4 Mgi
:  :      :---14
:  :      :     :----5 Dvi
:  :----15
:         :-----------3 Dma
:
:  :--9 Lta
:-19
:  :  :-10 Age
:  :-18
:     :  :--11 Cae
:     :-17
:        :-12 Hsa
:
:   :---------8 Cpo
:--21
    :   :---7 Mmu
    :--20
        :---6 Mau

No.1       ext. branch S.E.  int. branch S.E.
Oan          1   7.15  2.55   13  22.71  4.64
Tac          2   5.12  2.25   14   5.47  2.35
Dma          3  23.92  4.72   15   8.05  2.90
Mgi          4   8.79  2.81   16   0.89  1.05
Dvi          5   8.32  2.72   17   1.44  1.03
Mau          6   6.02  2.32   18   2.28  1.56
Mmu          7   6.18  2.36   19   2.22  1.45
Cpo          8  20.19  4.27   20   3.98  2.10
Lta          9   3.24  1.68   21   3.13  1.83
Age         10   1.19  1.03  TBL :     143.96  iter: 8
Cae         11   2.61  1.46  ln L:   -1386.93 +- 88.32
Hsa         12   1.05  0.98  AIC :    2853.86
#2
                  :----1 Oan
   :----------13
   :              :---2 Tac
:-16
:  :             :----4 Mgi
:  :      :---14
:  :      :     :----5 Dvi
:  :----15
:         :-----------3 Dma
:
:    :---7 Mmu
:---17
:    :---6 Mau
:
:  :----------8 Cpo
:-21
   :   :--9 Lta
   :--20
       :   :-10 Age
       :--19
           :  :--11 Cae
           :-18
              :-12 Hsa

No.2       ext. branch S.E.  int. branch S.E.
Oan          1   7.05  2.53   13  22.26  4.56
Tac          2   5.22  2.26   14   5.47  2.37
Dma          3  23.92  4.73   15   8.39  2.94
Mgi          4   8.67  2.79   16   0.95  1.09
Dvi          5   8.46  2.73   17   6.65  2.49
Mau          6   6.09  2.31   18   1.44  1.03
Mmu          7   6.07  2.29   19   2.64  1.61
Cpo          8  22.25  4.45   20   2.75  1.60
Lta          9   2.96  1.62   21  lower limit
Age         10   1.16  1.03  TBL :     146.07  iter: 6
Cae         11   2.56  1.45  ln L:   -1391.74 +- 88.04
Hsa         12   1.10  1.00  AIC :    2863.47  lower limit: 0.001
#3
                  :---1 Oan
   :----------13
   :              :---2 Tac
:-16
:  :             :----4 Mgi
:  :      :---14
:  :      :     :----5 Dvi
:  :----15
:         :-----------3 Dma
:
:----------8 Cpo
:
:       :---7 Mmu
:  :---17
:  :    :---6 Mau
:-21
   :  :--9 Lta
   :-20
      :  :-10 Age
      :-19
         :  :--11 Cae
         :-18
            :-12 Hsa

No.3       ext. branch S.E.  int. branch S.E.
Oan          1   6.79  2.48   13  22.52  4.67
Tac          2   5.49  2.28   14   5.49  2.38
Dma          3  23.95  4.74   15   7.81  2.88
Mgi          4   8.65  2.78   16   0.64  0.99
Dvi          5   8.49  2.73   17   5.49  2.28
Mau          6   5.87  2.26   18   1.44  1.02
Mmu          7   6.26  2.30   19   1.69  1.48
Cpo          8  21.38  4.41   20   2.28  1.46
Lta          9   3.81  1.82   21   2.28  1.86
Age         10   1.29  1.07  TBL :     145.28  iter: 7
Cae         11   2.63  1.47  ln L:   -1390.08 +- 88.11
Hsa         12   1.03  0.98  AIC :    2860.16

protml 2.3b3 JTT-F 3 trees 12 OTUs 141 sites. alpha-globin

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-------------------------------------------------------------------
1       -1386.9     0.0 <-best   40    2853.9     0.0    ME   0.7081
2       -1391.7    -4.8    4.2   40    2863.5     9.6   2.1   0.0246
3       -1390.1    -3.1    5.4   40    2860.2     6.3   1.3   0.2673
```

Figure 5.27: Result of ProtML analysis of hemoglobin α ("hba.ml" file).

```
 3 141 protml 2.3b3 (07/05/96) JTT-F 12 OTUs 141 sites. alpha-globin
 # 1 -1386.9 (((Oan,Tac),((Mgi,Dvi),Dma)),(Lta,(Age,(Cae,Hsa))),(Cpo,(Mmu,Mau)));
  -3.83439689e+00 -3.83439689e+00 -3.83439689e+00 -3.83439689e+00 -3.83439689e+00
  -3.83439689e+00 -7.55063190e+00 -1.21447698e+01 -7.66679333e+00 -2.01027775e+01
  -1.58330310e+01 -7.35795462e+00 -1.37801219e+01 -1.46995969e+01 -9.73655748e+00
  -6.67363395e+00 -2.19008094e+01 -1.21486010e+01 -1.00825120e+01 -2.47746528e+01
  -9.90807926e+00 -1.38959051e+01 -2.72224840e+01 -5.56009199e+00 -5.56009199e+00
  -7.52161061e+00 -6.33670851e+00 -8.94233740e+00 -2.45282331e+01 -3.82532936e+00
  -3.82532936e+00 -3.82532936e+00 -3.82532936e+00 -3.82532936e+00 -3.82532936e+00
  -1.02989902e+01 -1.18570643e+01 -2.76904017e+01 -1.10002557e+01 -3.13304150e+01
  -1.77494586e+01 -1.50635593e+01 -1.72031920e+01 -6.86929417e+00 -5.06656881e+00
  -1.58643555e+01 -2.17411359e+01 -2.04147767e+01 -1.60813759e+01 -2.79972453e+01
  -9.79719378e+00 -3.82254426e+00 -3.82254426e+00 -1.85946657e+01 -1.79654275e+01
  -7.75133350e+00 -7.75133350e+00 -7.96789834e+00 -3.72513038e+00 -3.72513038e+00
  -3.72513038e+00 -3.72513038e+00 -3.72513038e+00 -3.72513038e+00 -2.44533119e+01
  -1.62350153e+01 -1.63018829e+01 -1.88294335e+01 -9.02695767e+00 -3.00318729e+00
  -3.00318729e+00 -3.00318729e+00 -3.00318729e+00 -3.00318729e+00 -3.00318729e+00
  -3.00318729e+00 -6.52537286e+00 -1.54891587e+01 -1.18394431e+01 -9.97194951e+00
  -1.18934033e+01 -1.26596622e+01 -6.35359329e+00 -8.05867825e+00 -1.55825315e+01
  -8.23961881e+00 -7.78867783e+00 -9.72222845e+00 -3.35649761e+00 -3.35649761e+00
  -3.35649761e+00 -3.35649761e+00 -3.35649761e+00 -3.35649761e+00 -3.21889704e+01
  -1.34099107e+01 -1.41978531e+01 -9.79480838e+00 -2.15723860e+01 -1.81911713e+01
  -7.54768438e+00 -3.90516934e+00 -3.90516934e+00 -3.90516934e+00 -3.90516934e+00
  -8.72727161e+00 -2.06501106e+01 -4.50273267e+00 -4.50273267e+00 -4.50273267e+00
  -6.42383099e+00 -1.49125213e+01 -1.15191344e+01 -4.68585646e+00 -4.68585646e+00
  -4.68585646e+00 -4.68585646e+00 -4.68585646e+00 -3.03430947e+01 -2.12939921e+01
  -3.81766391e+01 -1.36269290e+01 -7.19141168e+00 -4.82469238e+00 -4.82469238e+00
  -4.82469238e+00 -4.82469238e+00 -4.82469238e+00 -4.82469238e+00 -5.36541392e+00
  -5.16873735e+00 -5.16873735e+00 -5.16873735e+00 -1.11217140e+01 -8.32696112e+00
  -9.29944198e+00 -3.99331246e+00 -3.99331246e+00 -3.99331246e+00 -3.99331246e+00
  -3.99331246e+00
 # 2 -1391.7 (((Oan,Tac),((Mgi,Dvi),Dma)),(Mmu,Mau),(Cpo,(Lta,(Age,(Cae,Hsa)))));
  -3.85936261e+00 -3.85936261e+00 -3.85936261e+00 -3.85936261e+00 -3.85936261e+00
  -3.85936261e+00 -7.46967141e+00 -1.20776423e+01 -7.68375958e+00 -1.96981971e+01
  -1.57885213e+01 -7.38266416e+00 -1.37309297e+01 -1.46098595e+01 -9.77294891e+00
  -6.72580021e+00 -2.14393877e+01 -1.47301346e+01 -1.01351442e+01 -2.47576015e+01
  -9.94106523e+00 -1.39243295e+01 -2.69633214e+01 -5.58591513e+00 -5.58591513e+00
  -7.54730006e+00 -6.37139549e+00 -8.94352085e+00 -2.43667262e+01 -3.84064634e+00
  -3.84064634e+00 -3.84064634e+00 -3.84064634e+00 -3.84064634e+00 -3.84064634e+00
  -1.03459766e+01 -1.18700112e+01 -2.75715752e+01 -1.08993136e+01 -3.15466244e+01
  -1.77772908e+01 -1.51265389e+01 -1.72502254e+01 -6.90380901e+00 -5.08815134e+00
  -1.58736654e+01 -2.16208411e+01 -2.04224220e+01 -1.61309218e+01 -2.82828922e+01
  -9.76288982e+00 -3.83704992e+00 -3.83704992e+00 -1.84637349e+01 -1.79673869e+01
  -7.76682714e+00 -7.76682714e+00 -7.98327713e+00 -3.74081483e+00 -3.74081483e+00
  -3.74081483e+00 -3.74081483e+00 -3.74081483e+00 -3.74081483e+00 -2.43332678e+01
  -1.62600007e+01 -1.58718906e+01 -1.87748880e+01 -9.02416717e+00 -3.01557582e+00
  -3.01557582e+00 -3.01557582e+00 -3.01557582e+00 -3.01557582e+00 -3.01557582e+00
  -3.01557582e+00 -6.44215555e+00 -1.56656665e+01 -1.18643483e+01 -9.88855245e+00
  -1.44780387e+01 -1.26640229e+01 -6.36581696e+00 -8.05823078e+00 -1.56547341e+01
  -8.25109040e+00 -7.80022708e+00 -9.73037732e+00 -3.36814359e+00 -3.36814359e+00
  -3.36814359e+00 -3.36814359e+00 -3.36814359e+00 -3.36814359e+00 -3.21298219e+01
  -1.50557660e+01 -1.41900865e+01 -9.82855562e+00 -2.14614180e+01 -1.80379027e+01
  -7.54602238e+00 -3.91909597e+00 -3.91909597e+00 -3.91909597e+00 -3.91909597e+00
  -8.72416718e+00 -2.04816446e+01 -4.52159756e+00 -4.52159756e+00 -4.52159756e+00
  -6.46897000e+00 -1.48102124e+01 -1.14543413e+01 -4.71664592e+00 -4.71664592e+00
  -4.71664592e+00 -4.71664592e+00 -4.71664592e+00 -2.97541412e+01 -2.12919863e+01
  -3.78115495e+01 -1.36463815e+01 -7.24330169e+00 -4.85482212e+00 -4.85482212e+00
  -4.85482212e+00 -4.85482212e+00 -4.85482212e+00 -4.85482212e+00 -5.37151057e+00
  -5.18885914e+00 -5.18885914e+00 -5.18885914e+00 -1.11103737e+01 -8.38213153e+00
  -8.92899635e+00 -4.01407733e+00 -4.01407733e+00 -4.01407733e+00 -4.01407733e+00
  -4.01407733e+00
 # 3 -1390.1 (((Oan,Tac),((Mgi,Dvi),Dma)),Cpo,((Mmu,Mau),(Lta,(Age,(Cae,Hsa)))));
  -3.85032777e+00 -3.85032777e+00 -3.85032777e+00 -3.85032777e+00 -3.85032777e+00
  -3.85032777e+00 -7.50449264e+00 -1.21137859e+01 -7.68468932e+00 -1.98769456e+01
  -1.58030286e+01 -7.37271574e+00 -1.37135142e+01 -1.46284922e+01 -9.74643840e+00
  -6.70235985e+00 -2.09679464e+01 -1.49584207e+01 -1.00890431e+01 -2.45511208e+01
  -9.92497451e+00 -1.39898017e+01 -2.68716677e+01 -5.57631813e+00 -5.57631813e+00
  -7.53662537e+00 -6.35849044e+00 -8.89475567e+00 -2.44852371e+01 -3.83501040e+00
  -3.83501040e+00 -3.83501040e+00 -3.83501040e+00 -3.83501040e+00 -3.83501040e+00
  -1.03229334e+01 -1.18636648e+01 -2.74847799e+01 -1.09419415e+01 -3.16921263e+01
  -1.76677146e+01 -1.50959064e+01 -1.72393368e+01 -6.88734683e+00 -5.08014994e+00
  -1.58658697e+01 -2.16596215e+01 -2.05264606e+01 -1.61283286e+01 -2.84695518e+01
  -9.82197811e+00 -3.83165548e+00 -3.83165548e+00 -1.85319087e+01 -1.77499262e+01
  -7.76024806e+00 -7.76024806e+00 -7.97685568e+00 -3.73502868e+00 -3.73502868e+00
  -3.73502868e+00 -3.73502868e+00 -3.73502868e+00 -3.73502868e+00 -2.42836744e+01
  -1.62444311e+01 -1.34942759e+01 -1.86782461e+01 -9.01619892e+00 -3.01100718e+00
  -3.01100718e+00 -3.01100718e+00 -3.01100718e+00 -3.01100718e+00 -3.01100718e+00
  -3.01100718e+00 -6.47704447e+00 -1.57354075e+01 -1.18617826e+01 -9.91892521e+00
  -1.47069036e+01 -1.27177035e+01 -6.36033266e+00 -8.01369259e+00 -1.56433177e+01
  -8.24594011e+00 -7.79492858e+00 -9.72839580e+00 -3.36385183e+00 -3.36385183e+00
  -3.36385183e+00 -3.36385183e+00 -3.36385183e+00 -3.36385183e+00 -3.22159450e+01
  -1.52775105e+01 -1.42832713e+01 -9.85521603e+00 -2.14451810e+01 -1.82020488e+01
  -7.53730910e+00 -3.91393531e+00 -3.91393531e+00 -3.91393531e+00 -3.91393531e+00
  -8.75744022e+00 -2.04504438e+01 -4.51457577e+00 -4.51457577e+00 -4.51457577e+00
  -6.44848089e+00 -1.45886931e+01 -1.14779522e+01 -4.70536484e+00 -4.70536484e+00
  -4.70536484e+00 -4.70536484e+00 -4.70536484e+00 -3.03423428e+01 -2.13663065e+01
  -3.77810638e+01 -1.36691811e+01 -7.22007479e+00 -4.84375985e+00 -4.84375985e+00
  -4.84375985e+00 -4.84375985e+00 -4.84375985e+00 -4.84375985e+00 -5.36923120e+00
  -5.18138888e+00 -5.18138888e+00 -5.18138888e+00 -1.11449229e+01 -8.35682460e+00
  -9.10231830e+00 -4.00644491e+00 -4.00644491e+00 -4.00644491e+00 -4.00644491e+00
  -4.00644491e+00
```

Figure 5.28: Estimated log-likelihood for each site of hemoglobin α ("hba.lls" file).

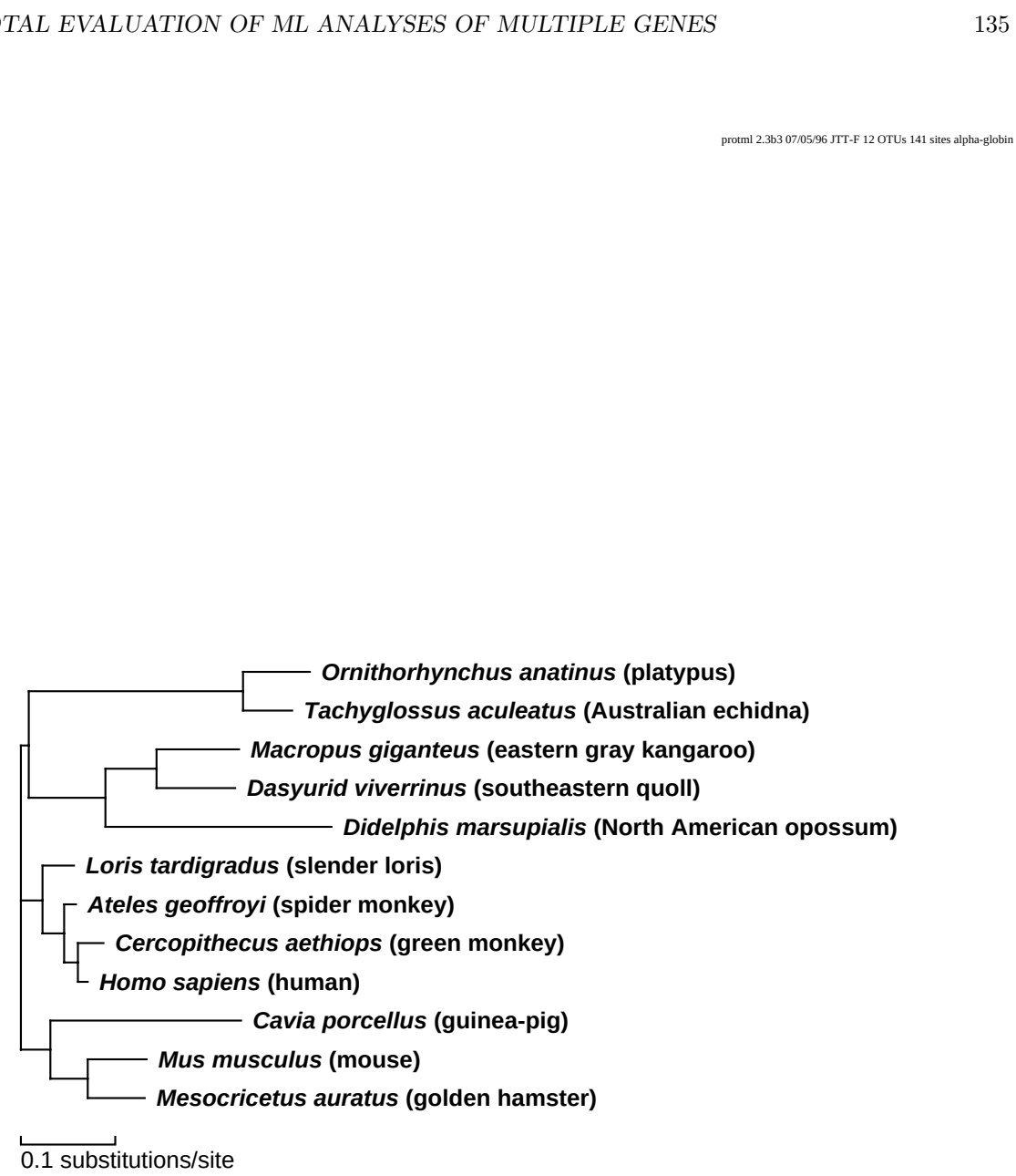


Figure 5.29: ML tree of hemoglobin α (JTT-F model).

```
 10 377 cytochrome b
Cla Crossostoma lacustre (loach)
ASLRKTHPLIKIANDALVDLPAPSNISVWWNFGSLLGLCLITQILTGLFLAMHYTSDIST
AFSSVAHICRDVNYGWLIRNIHANGASFFFICLYLHIARGLYYGSYLYKETWNIGVVLFL
LVMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYVGDMLVQWIWGGFSVDNATLTRFFAF
HFLFPFIVAAVTILHLLFLHETGSNNPAGLNSDADKISFHPYFSYKDLLGFVVMLLGLTT
LALFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSINKLGGVLALLFSILVLM
VVPVLHTSKQRGLTFRPATQFLFWTLVADMIILTWIGGMPVEHPYIIIGQIASILYFALF
LILIPLAGWLENKALEW
Cca Cyprinus carpio (carp)
ASLRKTHPLIKIANDALVDLPTPSNISAWWNFGSLLGLCLITQILTGLFLAMHYTSDIST
AFSSVTHICRDVNYGWLIRNVHANGASFFFICIYMHIARGLYYGSYLYKETWNIGVVLLL
LVMMTAFVGYVLPWGQMSFWGATVITNLLSAVPYMGDMLVQWIWGGFSVDNATLTRFFAF
HFLLPFVIAAATIIHLLFLHETGSNNPIGLNSDADKVSFHPYFSYKDLLGFVIMLLALTL
LALFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSINKLGGVLALLFSILVLM
VVPLLHTSKQRGLTFRPITQFLFWTLVADMIILTWIGGMPVEHPFIIIGQIASVLYFALF
LIFMPLAGWLENKALKW
Xla Xenopus laevis (frog)
ANIRKSHPLIKIINNSFIDLPTPSNISSLWNFGSLLGVCLIAQIITGLFLAMHYTADTSM
AFSSVAHICFDVNYGLLIRNLHANGLSFFFICIYLHIGRGLYYGSFLYKETWNIGVILLF
LVMATAFVGYVLPWGQMSFWGATVITNLLSAKPYIGNVLVQWSLGGFSVDNATLTRFFAF
HFLLPFIIAGASILHLLFLHETGSTNPTGLNSDPDKVPFHPYFSYKDLLGFLIMLTALTL
LAMFSPNLLGDPDNFTPANPLITPPHIKPEWYFLFAYAILRSMNKLGGVLALVLSILILA
LMPLLHTSKQRSLMFRPFTQIMFWALVADTLILTWIGGQPVEDPYTMIGQLASVIYFSIF
IIMFPLMGWVENKLLNW
Gga Gallus gallus (chicken)
ANIRKSHPLLKMINNSLIDLPAPSNISAWWNFGSLLAVCLMTQILTGLLLAMHYTADTSL
AFSSVAHTCRNVQYGWLIRNLHANGASFFFICIFLHIGRGLYYGSYLYKETWNTGVILLL
TLMATAFVGYVLPWGQMSFWGATVITNLFSAIPYIGHTLVEWAWGGFSVDNPTLTRFFAL
HFLLPFAIAGITIIHLTFLHESGSNNPLGISSDSDKIPFHPYYSFKDILGLTLMLTPFLT
LALFSPNLLGDPENFTPANPLVTPPHIKPEWYFLFAYAILRSINKLGGVLALAASVLILF
LIPFLHKSKQRTMTFRPLSQTLFWLLVANLLILTWIGSQPVEHPFIIIGQMASLSYFTIL
LILFPTIGTLENKMLNY
Mdo Monodelphis domestica (grey short-tailed opossum)
TNLRKNYPLMKIINHSFIDLPAPSNISAWWNFGSLLGMCLIIQILTGLFLAMHYTSDTLT
AFSSVAHICRDVNYGWLIRNLHANGASMFFMCLFLHVGRGIYYGSYLYKETWNIGVILML
TVMATAFVGYVLPWGQMSFWGATVITNLLSAIPYIGNTLVEWIWGGFSVDKATLTRFFAF
HFILPFIILALVIVHLLFLHETGSNNPTGINPNSDKIPFHPYYTIKDALGLILMLLILMS
LAMFSPDMLGNPDNFTPANPLNTPPHIKPEWYFLFAYAILRSINKLGGVLALLASLLILL
IIPLLHTSKQRSLMFRPISQIMFWLLVANLLTLTWIGGQPVEQPFIIIGQLASTLYFSLI
IIFMPLAGMYEDHLLEP
Cpo Cavia porcellus (guinea-pig)
THLRKSHPLIKIINHSLIDLPAPSSISTWWNFGSLLGICLGLQIITGLFLAMHYTADTST
AFSSVAHICRDVNYGWLIRYLHANGASMFFIFLYLHIGRGIYYGSYTFLETWNIGIALLF
TVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAF
HFIVPFIITALVMVHLLFLHETGSNNPSGLNSDSDKIPFHPYYTIKDILGALFMMLALLC
LVLFTPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSINKLGGVLALVLSILILA
LFPMLHTSKQRSMRFRPLSQCLLWLLAANLLILTWIGGQPVEHPYITIGQSASIPYFFII
LILFPLTSLLENKMLKW
Haf Hystrix africaeaustralis (African porcupine)
TNIRKSHPLLKIINHSFIDLPTPSNISTWWNFGSLLGACLIIQILTGLFLAMHYTAYTTT
AFSSVAHICRDVNYGWLIRYLHANGASMFFICLYLHVGRGLYYGSYMFTETWNIGILLLF
TVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAF
HFSLPFIITALVLVHLLFLHETGSNNPSGIDSNSDKIPFHPYYTIKDILGLLLMLTALLI
LVLFSPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSINKLGGVLALIFSILILA
IIPLLHTSKQRSMLFRPFSQCLFWILAANLLILTWIGGQPVEHPYITIGQLASISYFSIL
LIIMPLTSIMENKLLKW
Rno Rattus norvegicus (rat)
TNIRKSHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQILTGLFLAMHYTSDTMT
AFSSVTHICRDVNYGWLIRYLHANGASMFFICLFLHVGRGLYYGSYTFLETWNIGIILLF
AVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAF
HFILPFIIAALAIVHLLFLHETGSNNPTGLNSDADKIPFHPYYTIKDLLGVFMLLLFLMT
LVLFFPDLLGDPDNYTPANPLNTPPHIKPEWYFLFAYAILRSINKLGGVVALILSILILA
FLPFLHTSKQRSLTFRPITQILYWILVANLLVLTWIGGQPVEHPFIIIGQLASISYFSII
LILMPISGIVEDKMLKW
Mmu Mus musculus (mouse)
TNMRKTHPLFKIINHSFIDLPAPSNISSWWNFGSLLGVCLMVQIITGLFLAMHYTSDTMT
AFSSVTHICRDVNYGWLIRYMHANGASMFFICLFLHVGRGLYYGSYTFMETWNIGVLLLF
AVMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTTLVEWIWGGFSVDKATLTRFFAF
HFILPFIIAALAIVHLLFLHETGSNNPTGLNSDADKIPFHPYYTIKDILGILIMFLILMT
LVLFFPDMLGDPDNYMPANPLNTPPHIKPEWYFLFAYAILRSINKLGGVLALILSILILA
LMPFLHTSKQRSLMFRPITQILYWILVANLLILTWIGGQPVEHPFIIIGQLASISYFSII
LILMPISGIIEDKMLKL
Hsa Homo sapiens (human)
TPMRKINPLMKLINHSFIDLPTPSNISAWWNFGSLLGACLILQITTGLFLAMHYSPDAST
AFSSIAHITRDVNYGWIIRYLHANGASMFFICLFLHIGRGLYYGSFLYSETWNIGIILLL
ATMATAFMGYVLPWGQMSFWGATVITNLLSAIPYIGTDLVQWIWGGYSVDSPTLTRFFTF
HFILPFIIAALATLHLLFLHETGSNNPLGITSHSDKITFHPYYTIKDALGLLLFLLSLMT
LTLFSPDLLGDPDNYTLANPLNTPPHIKPEWYFLFAYTILRSVNKLGGVLALLLSILILA
MIPILHMSKQQSMMFRPLSQSLYWLLAADLLILTWIGGQPVSYPFTIIGQVASVLYFTTI
LILMPTISLIENKMLKW
```

Figure 5.30: Cytochrome b sequence data ("cytb.ptn" file).

```
    3
((Cla,Cca),Xla,(Gga,(Mdo,(Hsa,((Rno,Mmu),(Cpo,Haf))))));
((Cla,Cca),Xla,(Gga,(Mdo,((Rno,Mmu),((Cpo,Haf),Hsa)))));
((Cla,Cca),Xla,(Gga,(Mdo,((Cpo,Haf),(Hsa,(Rno,Mmu))))));
```

Figure 5.31: Tree topology file of cytochrome b ("cytb.tpl" file).

```
protml 2.3b3 (07/05/96) mtREV24-F 10 OTUs 377 sites. cytochrome b
#1
          :----1 Cla
:---------1i
:          :--2 Cca
:
:---------3 Xla
:
:      :-----------4 Gga
:-----17
       :     :--------5 Mdo
       :----16
             :   :-------------10 Hsa
             :--15
                 :          :---8 Rno
                 :   :-----12
                 :   :      :---9 Mmu
                 :--14
                     :    :-------6 Cpo
                     :---13
                          :----7 Haf

No.1       ext. branch S.E.  int. branch S.E.
Cla          1   5.14  1.32   11  12.07  2.18
Cca          2   2.01  0.96   12   6.46  1.51
Xla          3  12.78  2.21   13   4.22  1.30
Gga          4  15.90  2.43   14   2.30  1.06
Mdo          5  11.06  1.94   15   2.34  1.09
Cpo          6   9.46  1.76   16   5.44  1.56
Haf          7   5.51  1.39   17   6.43  1.71
Rno          8   3.37  1.05  TBL :      125.47  iter: 5
Mmu          9   2.96  0.99  ln L:    -3209.62 +- 121.91
Hsa         10  18.01  2.49  AIC :     6491.24
#2
          :----1 Cla
:---------1i
:          :--2 Cca
:
:---------3 Xla
:
:      :-----------4 Gga
:-----17
       :     :--------5 Mdo
       :----16
             :          :---8 Rno
             :    :-----12
             :    :      :--9 Mmu
             :---15
                  :         :-------6 Cpo
                  :   :---13
                  :   :    :----7 Haf
                  :--14
                      :-------------10 Hsa

No.2       ext. branch S.E.  int. branch S.E.
Cla          1   5.14  1.32   11  12.27  2.20
Cca          2   2.01  0.97   12   6.62  1.53
Xla          3  12.57  2.21   13   3.39  1.24
Gga          4  15.85  2.43   14   1.95  1.00
Mdo          5  11.14  1.96   15   3.24  1.21
Cpo          6   9.52  1.77   16   5.57  1.59
Haf          7   5.43  1.39   17   6.60  1.73
Rno          8   3.47  1.06  TBL :      125.86  iter: 6
Mmu          9   2.86  0.98  ln L:    -3209.72 +- 121.92
Hsa         10  18.22  2.50  AIC :     6491.43
#3
          :----1 Cla
:---------1i
:          :--2 Cca
:
:---------3 Xla
:
:      :-----------4 Gga
:-----17
       :     :--------5 Mdo
       :----16
             :         :-------6 Cpo
             :    :---12
             :    :    :----7 Haf
             :---15
                  :  :-------------10 Hsa
                  :-14
                     :      :---8 Rno
                     :------13
                            :--9 Mmu

No.3       ext. branch S.E.  int. branch S.E.
Cla          1   5.13  1.32   11  12.32  2.21
Cca          2   2.02  0.97   12   4.32  1.32
Xla          3  12.57  2.21   13   6.41  1.51
Gga          4  15.54  2.41   14   1.08  0.76
Mdo          5  11.35  1.96   15   3.30  1.22
Cpo          6   9.46  1.76   16   5.48  1.56
Haf          7   5.52  1.38   17   6.76  1.75
Rno          8   3.43  1.06  TBL :      126.20  iter: 6
Mmu          9   2.90  0.99  ln L:    -3213.16 +- 121.94
Hsa         10  18.59  2.53  AIC :     6498.32

protml 2.3b3 mtREV24-F 3 trees 10 OTUs 377 sites. cytochrome b

Tree     ln L  Diff ln L  S.E. #Para   AIC  Diff AIC   TBL  RELL-BP
-------------------------------------------------------------------
1       -3209.6     0.0 <-best   36    6491.2    0.0    ME   0.4553
2       -3209.7    -0.1    6.6   36    6491.4    0.2    0.4  0.4448
3       -3213.2    -3.5    5.1   36    6498.3    7.1    0.7  0.0999
```

Figure 5.32: Result of ProtML analysis of cytochrome *b* ("cytb.ml" file).

The printout of the protml.eps file of the ML tree in this analysis is given in Fig. 5.33.

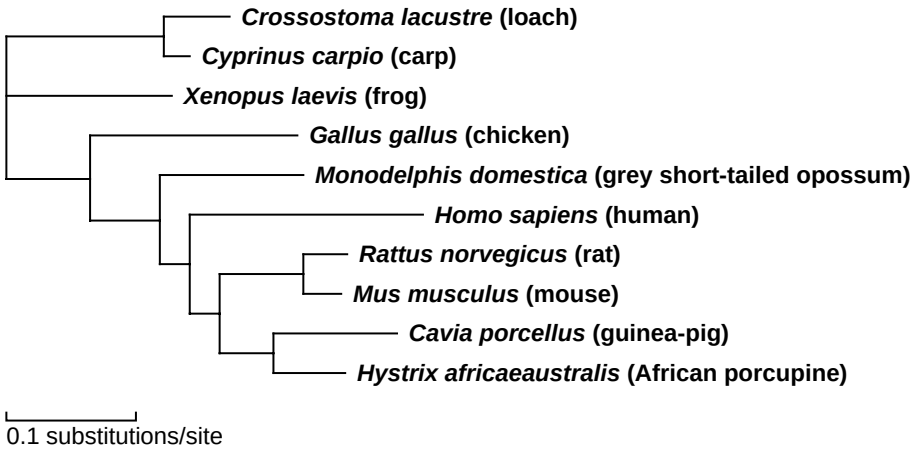


Figure 5.33: ML tree of cytochrome b (mtREV24-F model).

The total evidence of the two analyses of cytochrome b and hemoglobin α is evaluated by using the following command (see page 71);

```
totalml cytb.lls hba.lls > total.ml
```

The "total.ml" file looks like this,

```
totalml 1.1(07/05/96) 2 data sets, 518 sites. protml 2.3b3

tree       1        2     total
1     3209.6   1386.9    4596.6
           ml       ml       ML
2         0.1      4.8      4.9
          6.6      4.2      7.8
3         3.5      3.1      6.7
          5.1      5.4      7.5
sites     377      141      518

tree      1        2      total
1     0.4574   0.7103    0.6690
2     0.4484   0.0215    0.2157
3     0.0942   0.2682    0.1153
```

The 1st and 2nd columns refer to cytochrome b and hemoglobin α. "ml" refers to the ML tree topology (for which the estimated negative log-likelihood is given), and for other tree topologies the differences of log-likelihood from the ML tree are given with their SEs. In the "total" column, the ML tree is indicated by "ML". Furthermore, bootstrap probabilities (BP) estimated by the RELL method are given for each data set and for the total.

Although the two proteins do not have sufficient information to resolve the issue at hand, Graur et al.'s hypothesis (tree-3) is the least likely (with 11.5% BP) in this analysis. In order to resolve the problem, we should increase the number of proteins to analyze, and then we can have a satisfactory resolution in which Myomorpha form a clade with the guinea pig excluding Primates as an outgroup (Cao et al. 1994[42]; Kuma and Miyata 1994[160]). Recently, on the basis of phylogenetic analyses of the complete

mitochondrial genome from the guinea-pig, D'Erchia et al. (1996[56]) concluded that the guinea pig is closer to the Lagomorpha/Primates/Carnivora/Perissodactyla/Artiodactyla/Cetacea clade rather than to Myomorpha (tree-2). However, the support is very marginal by the ProtML analysis (Cao, Okada and Hasegawa, submitted), and their data is too weak to exclude the rodent monophyly hypothesis which is supported by other molecular evidence (e.g., Hasegawa et al. 1992[91]; Martignetti and Brosius 1993[180]; Cao et al. 1994[42]; Kuma and Miyata 1994[160]; Frye and Hedges 1995[71]).

Acknowledgement

We are very grateful to many users for suggestions and complaints. We are particularly grateful to Hirohisa Kishino and Tetsuo Hashimoto for discussions and suggestions, to Ying Cao for collecting the data, and to Korbinian Strimmer and Arndt von Haeseler for pointing out problems in the earlier version of MOLPHY. Thanks are also due to Peter Waddell, Tetsuo Hashimoto, Genshiro Kitagawa, and anonymous reviewers for their many valuable comments on an earlier version of the manuscript. This work was carried out under the Institute of Statistical Mathematics Cooperative Research Program (94-ISM-CRP-A74 and 95-ISM-CRP-A69) and was supported in part by grants from the Ministry of Education, Science, Sports and Culture of Japan. J.A. is a Research Fellow of the Japan Society for the Promotion of Science.

Bibliography

[1] Adachi, J. (1995). *Modeling of Molecular Evolution and Maximum Likelihood Inference of Molecular Phylogeny*. Ph.D. dissertation, The Graduate University for Advanced Studies.

[2] Adachi, J., Cao, Y., and Hasegawa, M. (1993). 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.

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

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

[5] Adachi, J. and Hasegawa, M. (1995). Improved dating of the human-chimpanzee separation in the mitochondrial DNA tree: heterogeneity among amino acid sites. *J. Mol. Evol.*, 40:622–628.

[6] Adachi, J. and Hasegawa, M. (1995). Markov model of amino acid substitution in mitochondrial proteins and maximum likelihood inference of molecular phylogeny. In Nei, M. and Takahata, N., editors, *Current Topics on Molecular Evolution — Proceedings of the U.S.-Japan Workshop*, pages 61–67. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University (USA), and Graduate University for Advanced Studies (Hayama, Japan).

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

[8] Adachi, J. and Hasegawa, M. (1995). Time scale for the mitochondrial DNA tree of human evolution. In Brenner, S. and Hanihara, K., editors, *The Origin and Past of Modern Humans as Viewed from DNA*, pages 46–68. World Scientific Publ., Singapore.

[9] Adachi, J. and Hasegawa, M. (1996). Instability of quartet analyses of molecular sequence data by the maximum likelihood method: The Cetacea/Artiodactyla relationships. *Mol. Phyl. Evol.*, 6:72–76.

[10] Adachi, J. and Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *J. Mol. Evol.*, 42:459–468.

[11] Adachi, J. and Hasegawa, M. (1996). Tempo and mode of synonymous substitutions in mitochondrial DNA of primates. *Mol. Biol. Evol.*, 13:200–208.

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

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

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

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

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

[17] Aoshima, M., Nishibe, Y., Hasegawa, M., Yamagishi, A., and Oshima, T. (1996). Cloning and sequencing of a gene encoding 16S ribosomal RNA from a novel hyperthermophilic archaebacterium NC12. *Gene*, in press.

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

[19] Árnason, Ú. and Gullberg, A. (1993). 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.

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

[21] Árnason, Ú. and Gullberg, A. (1996). Cytochrome *b* nucleotide sequences and the identification of five primary lineages of extant cetaceans. *Mol. Biol. Evol.*, 13:407–417.

[22] Árnason, Ú., Gullberg, A., Johnsson, E., and Ledje, C. (1993). 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.

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

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

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

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

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

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

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

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

[31] Baldauf, S.L. and Palmer, J.D. (1993). Animals and fungi are each other's closest relatives: Congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. USA*, 90:11558–11562.

[32] Baldauf, S.L., Palmer, J.D., and Doolittle, W.F. (1996). The root of the universal tree and the origin of eukaryotes based on elongation factor phylogeny. *Proc. Natl. Acad. Sci. USA*, 93:7749–7754.

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

[34] Barry, D. and Hartigan, J.A. (1987). Statistical analysis of hominoid molecular evolution. *Statist. Sci.*, 2:191–210.

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

[36] Boore, J.L., Collins, T.M., Stanton, D., Daehler, L.L., and Brown, W.M. (1995). Deducing the pattern of arthropod phylogeny from mitochondrial DNA rearrangements. *Nature*, 376:163–165.

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

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

[39] Cantatore, P., Roberti, M., Pesole, G., Ludovico, A., Milella, F., Gadaleta, M.N., and Saccone, C. (1994). 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.

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

[41] Cao, Y., Adachi, J., Janke, A., Pääbo, S., and Hasegawa, M. (1994). 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.

[42] Cao, Y., Adachi, J., Yano, T., and Hasegawa, M. (1994). 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.

[43] Caspers, G.-J., Reinders, G.-J., Leunissen, J.A.M., Wattel, J., and de Jong, W.W. (1996). Protein sequences indicate that turtles branched off from the amniote tree after mammals. *J. Mol. Evol.*, 42:580–586.

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

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

[46] Chikuni, K., Mori, Y., Tabata, T., Saito, M., Monma, M., and Kosugiyama, M. (1995). Molecular phylogeny based on the κ-casein and cytochrome *b* sequences in the mammalian suborder Ruminantia. *J. Mol. Evol.*, 41:859–866.

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

[48] Chow, S. and Kishino, H. (1995). Phylogenetic relationships between tuna species of the genus *Thunnus* (Scombridae: Teleostei): Inconsistent implications from morphology, nuclear and mitochondrial genomes. *J. Mol. Evol.*, 41:741–748.

[49] Clark, C.G. and Roger, A.J. (1995). Direct evidence for secondary loss of mitochondria in *Entamoeba histolytica*. *Proc. Natl. Acad. Sci. USA*, 92:6518–6521.

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

[51] Corbet, G.B. and Hill, J.E. (1991). *A World List of Mammalian Species, Third Edition*. Oxford Univ. Press, Oxford.

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

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

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

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

[56] D'Erchia, A.M., Gissi, C., Pesole, G., Saccone, C., and Árnason, Ú. (1996). The guinea-pig is not a rodent. *Nature*, 381:597–600.

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

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

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

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

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

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

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

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

[65] Felsenstein, J. (1982). Numerical methods for inferring evolutionary trees. *Quart. Rev. Biol.*, 57:379–404.

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

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

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

[69] Felsenstein, J. (1993). *PHYLIP (Phylogeny Inference Package) and manual, version 3.5c.* Dept. Genetics, Univ. of Washington, Seattle.

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

[71] Frye, M.S. and Hedges, S.B. (1995). Monophyly of the order Rodentia inferred from mitochondrial DNA sequences of the genes for 12S rRNA, 16S rRNA, and tRNA-Valine. *Mol. Biol. Evol.*, 12:168–176.

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

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

[74] Galtier, N. and Gouy, M. (1995). Inferring phylogenies from DNA sequences of unequal base compositions. *Proc. Natl. Acad. Sci. USA*, 92:11317–11321.

[75] Gatesy, J., Hayashi, C., Cronin, M.A., and Arctander, P. (1996). Evidence from milk casein genes that cetaceans are close relatives of hippopotamid artiodactyls. *Mol. Biol. Evol.*, 13:954–963.

[76] Gaut, B.S. and Lewis, P.O. (1995). Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.*, 12:152–162.

[77] Gemmell, N.J. and Westerman, M. (1994). Phylogenetic relationships within the class Mammalia: a study using mitochondrial 12S RNA sequences. *J. Mammal. Evol.*, 2:3–23.

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

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

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

[81] Goldman, N. (1990). 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.

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

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

[84] Graur, D., Duret, L., and Gouy, M. (1996). Phylogenetic position of the order Lagomorpha (rabbits, hares and allies). *Nature*, 379:333–335.

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

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

[87] Gregory, W.K. (1947). The monotremes and the palimpsest theory. *Amer. Mus. Nat. Hist. Bull.*, 88:1–52.

[88] Groves, P. and Shields, G.F. Convergent evolution of the Asian takin and Arctic muskox. Unpublished.

[89] Hasegawa, M. and Adachi, J. (1996). Phylogenetic position of cetaceans relative to artiodactyls: Reanalysis of mitochondrial and nuclear sequences. *Mol. Biol. Evol.*, 13:710–717.

[90] Hasegawa, M., Adachi, J., and Milinkovitch, M.C. (1996). Novel phylogeny of whales supported by total molecular evidence. *J. Mol. Evol.*, in press.

[91] Hasegawa, M., Cao, Y., Adachi, J., and Yano, T. (1992). Rodent polyphyly? *Nature*, 355:595–595.

[92] Hasegawa, M. and Fujiwara, M. (1993). Relative efficiencies of the maximum likelihood, maximum parsimony, and neighbor-joining methods for estimating protein phylogeny. *Mol. Phyl. Evol.*, 2:1–5.

[93] Hasegawa, M. and Hashimoto, T. (1993). Ribosomal RNA trees misleading? *Nature*, 361:23–23.

[94] Hasegawa, M., Hashimoto, T., Adachi, J., Iwabe, N., and Miyata, T. (1993). 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.

[95] Hasegawa, M., Iwabe, N., Mukohata, Y., and Miyata, T. (1990). 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.

[96] Hasegawa, M. and Kishino, H. (1989). Heterogeneity of tempo and mode of mitochondrial DNA evolution among mammalian orders. *Jpn. J. Genet.*, 64:243–258.

[97] Hasegawa, M. and Kishino, H. (1994). Accuracies of the simple methods for estimating the bootstrap probability of a maximum likelihood tree. *Mol. Biol. Evol.*, 11:142–145.

[98] Hasegawa, M. and Kishino, H. (1996). *Molecular Phylogenetics (in Japanese)*. Iwanami Publ., Tokyo.

[99] Hasegawa, M., Kishino, H., and Saitou, N. (1991). On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.*, 32:443–445.

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

[101] Hasegawa, M. and Yano, T. (1984). Maximum likelihood method of phylogenetic inference from DNA sequence data. *Bull. Biomet. Soc. Japan*, 5:1–7.

[102] Hasegawa, M., Yano, T., and Kishino, H. (1984). A new molecular clock of mitochondrial DNA and the evolution of hominoids. *Proc. Japan Acad.*, B60:95–98.

[103] Hasegawa, M., Yano, T., and Miyata, T. (1984). Evolutionary implications of error amplification in the self-replicating and protein-synthesizing machinery. *J. Mol. Evol.*, 20:77–85.

[104] Hashimoto, T., Adachi, J., and Hasegawa, M. (1992). Phylogenetic place of *Giardia lamblia*, a protozoan that lacks mitochondria. *Endocytobiosis and Cell Research*, 9:59–69.

[105] Hashimoto, T. and Hasegawa, M. (1996). Origin and early evolution of eukaryotes inferred from the amino acid sequences of translation elongation factors 1α/Tu and 2/G. *Adv. Biophys.*, 32:73–120.

[106] Hashimoto, T., Nakamura, Y., Kamaishi, T., Adachi, J., Nakamura, F., Okamoto, K., and Hasegawa, M. (1995). Phylogenetic place of kinetoplastid protozoa inferred from a protein phylogeny of elongation factor 1α. *Mol. Biochem. Parasitol.*, 70:181–185.

[107] Hashimoto, T., Nakamura, Y., Kamaishi, T., Nakamura, F., Adachi, J., Okamoto, K., and Hasegawa, M. (1995). Phylogenetic place of a mitochondrion-lacking protozoan, *Giardia lamblia*, inferred from amino acid sequences of elongation factor 2. *Mol. Biol. Evol.*, 12:782–793.

[108] Hashimoto, T., Nakamura, Y., Nakamura, F., Shirakura, T., Adachi, J., Goto, N., Okamoto, K., and Hasegawa, M. (1994). 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.

[109] Hashimoto, T., Otaka, E., Adachi, J., Mizuta, K., and Hasegawa, M. (1993). The giant panda is most close to a bear, judged by α- and β-hemoglobin sequences. *J. Mol. Evol.*, 36:282–289.

[110] Hedges, S.B. and Sibley, C.G. (1994). Molecular vs. morphology in avian evolution: The case of the "pelecaniform" birds. *Proc. Natl. Acad. Sci. USA*, 91:9861–9865.

[111] Helm-Bychowski, K. and Cracraft, J. (1993). 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.

[112] Hendy, M.D. and Penny, D. (1989). A framework for the quantitative study of evolutionary trees. *Syst. Zool.*, 38:297–309.

[113] Henikoff, S. and Henikoff, J.G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919.

[114] Higgins, D.G., Bleasby, A.J., and Fuchs, R. (1992). CLUSTAL V: improved software for multiple sequence alignment. *Comput. Appl. Biosci.*, 8:189–191.

[115] Hillis, D.M., Huelsenbeck, J.P., and Swofford, D.L. (1994). Hobglobin of phylogenetics. *Nature*, 369:363–364.

[116] Hillis, D.M., Moritz, C., and Mable, B.K. (1996). *Molecular Systematics, 2nd Edition*. Sinauer Associates, Sunderland, Massachusetts.

[117] Horai, S., Hayasaka, K., Kondo, R., Tsugane, K., and Takahata, N. (1995). The recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA*, 92:532–536.

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

[119] Horai, S., Satta, Y., Hayasaka, K., Kondo, R., Inoue, T., Ishida, T., Hayashi, S., and Takahata, N. (1993). Man's place in Hominoidea revealed by mitochondrial DNA genealogy (Erratum). *J. Mol. Evol.*, 37:89–89.

[120] Horner, D.S., Hirt, R.P., Kilvington, S., Lloyd, D., and Embley, T.M. (1996). Molecular data suggest an early acquisition of the mitochondrion endosymbiont. *Proc. R. Soc. London*, B263:1053–1059.

[121] Hovemann, B. and Richer, S. (1988). Two genes encode related cytoplasmic elongation factors 1-α (EF-1) in *Drosophila melanogaster* with continuous and stage specific expression. *Nucl. Acids. Res.*, 16:3175–3194.

[122] Huelsenbeck, J.P. (1995). The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. *Mol. Biol. Evol.*, 12:843–849.

[123] Hughes, A.L. and Nei, M. (1988). Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature*, 335:167–170.

[124] Ibrahimi, I.M., Prager, E.M., White, T.J., and Wilson, A.C. (1979). Amino acid sequence of California quail lysozyme. Effect of evolutionary substitutions on the antigenic structure of lysozyme. *Biochemistry*, 18:2736–2744.

[125] Irwin, D.M. and Árnason, Ú. (1994). Cytochrome *b* gene of marine mammals: phylogeny and evolution. *J. Mammal. Evol.*, 2:37–55.

[126] Irwin, D.M., Kocher, T.D., and Wilson, A.C. (1991). Evolution of the cytochrome *b* gene of mammals. *J. Mol. Evol.*, 32:128–144.

[127] IUPAC-IUB Commission on Biochemical Nomenclature, . (1968). A one-letter notation for amino acid sequences, tentative rules. *J. Biol. Chem.*, 243:3557–3559.

[128] Iwabe, N., Kuma, K., Kishino, H., Hasegawa, M., and Miyata, T. (1991). Evolution of RNA polymerases and branching patterns of the three major groups of archaebacteria. *J. Mol. Evol.*, 32:70–78.

[129] Janke, A., Feldmaier-Fuchs, G., Thomas, W.K., von Haeseler, A., and Pääbo, S. (1994). The marsupial mitochondrial genome and the evolution of placental mammals. *Genetics*, 137:243–256.

[130] Janke, A., Gemmell, N.J., Feldmaier-Fuchs, G., von Haeseler, A., and Pääbo, S. (1996). The complete mitochondrial genome of a monotreme, the platypus (*Ornithorhynchus anatinus*). *J. Mol. Evol.*, 42:153–159.

[131] Johansen, S. and Johansen, T. (1994). Sequence analysis of 12 structural genes and a novel non-coding region from mitochondrial DNA of Atlantic cod *Gadus morhua*. *Biochim. Biophys. Acta*, 1218:2130–2170.

[132] Jollès, J., Schoentgen, F., Jollès, P., Prager, E.M., and Wilson, A.C. (1976). Amino acid sequence and immunological properties of chachalaca egg white lysozyme. *J. Mol. Evol.*, 8:59–78.

[133] Jollès, P. and Jollès, J. (1984). What's new in lysozyme research? Always a model system, today as yesterday. *Mol. Cell. Biochem.*, 63:165–189.

[134] Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid genaration of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, 8:275–282.

[135] Jukes, T.H. and Cantor, C.R. (1969). Evolution of protein molecules. In Munro, H.N., editor, *Mammalian Protein Metabolism, Vol.III*, pages 21–132. Academic Press, New York.

[136] Kamaishi, T., Hashimoto, T., Nakamura, Y., Nakamura, F., Murata, S., Okada, N., Okamoto, K., Shimizu, M., and Hasegawa, M. (1996). Protein phylogeny of EF-1α suggests microsporidians are extremely ancient eukaryotes. *J. Mol. Evol.*, 42:257–263.

[137] Kühne, W.G. (1973). The systematic position of monotremes reconsidered (Mammalia). *Z. Morph. Tiere*, 75:59–64.

[138] Kühne, W.G. (1975). Marsupium and marsupial bone in mesozoic mammals and in the marsupionta. *Colloque internationl C.N.R.S.*, 218:585–590.

[139] Keeling, P.J. and Doolittle, W.F. (1996). A non-canonical genetic code in an early diverging eukaryotic lineage. *EMBO J.*, 15:2285–2290.

[140] Kelly, C. (1994). A test of Markovian model of DNA evolution. *Biometrics*, 50:653–6641.

[141] Kikkawa, Y., Amano, T., and Suzuki, H. 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.

[142] Kikkawa, Y., Suzuki, H., Yonekawa, H., and Amano, T. Genetic diversity and geographic distribution of asian domestic water buffaloes based on the variations in restriction sites and sequences of mitochondrial DNA. Unpublished.

[143] Kimura, M. (1968). Evolutionary rate at the molecular level. *Nature*, 217:624–626.

[144] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.*, 16:111–120.

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

[146] Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge Univ. Press, Cambridge.

[147] Kishino, H. and Hasegawa, M. (1989). 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.

[148] Kishino, H., Miyata, T., and Hasegawa, M. (1990). Maximum likelihood inference of protein phylogeny, and the origin of chloroplasts. *J. Mol. Evol.*, 31:151–160.

[149] Kleinschmidt, T., Czelusniak, J., Goodman, M., and Braunitzer, G. (1986). Paenungulata: a comparison of the hemoglobin sequences from elephant, hyrax, and manatee. *Mol. Biol. Evol.*, 3:427–435.

[150] Klenk, H.-P. and Zillig, W. (1994). DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J. Mol. Evol.*, 38:420–432.

[151] Kocher, T.D., Conroy, J.A., McKaye, K.R., and Stauffer, J.R. (1993). Similar morphologies of cichlid fish in Lakes Tanganyika and Malawi are due to convergence. *Mol. Phyl. Evol.*, 2:158–165.

[152] Kocher, T.D., Conroy, J.A., McKaye, K.R., Stauffer, J.R., and Lockwood, S.F. (1995). Evolution of NADH dehydrogenase subunit 2 in East African cichlid fish. *Mol. Phyl. Evol.*, 4:420–432.

[153] Kojima, S., Hashimoto, T., Hasegawa, M., Murata, S., Ohta, S., Seki, H., and Okada, N. (1993). Close phylogenetic relationship between Vestimentifera (tube worms) and Annelida revealed by the amino acid sequence of elongation factor-1α. *J. Mol. Evol.*, 37:66–70.

[154] Kornegay, J.R., Kocher, T.D., Williams, L.A., and Wilson, A.C. (1993). Pathways of lysozyme evolution inferred from the sequences of cytochrome *b* in birds. *J. Mol. Evol.*, 37:367–379.

[155] Krajewski, C. and Fetzner, Jr., J.W. (1994). Phylogeney of cranes (Gruiformes: Gruidae) based on cytochrome-*b* DNA sequences. *Auk*, 111:351–365.

[156] Krajewski, C., Painter, J., Buckley, L., and Westerman, M. (1994). Phylogenetic structure of the marsupial family Dasyuridae based on cytochrome *b* DNA sequences. *J. Mammal. Evol.*, 2:25–35.

[157] Krettek, A., Gullberg, A., and Árnason, Ú. (1995). Sequence analysis of the complete mitochondrial DNA molecule of the hedgehog, *Erinaceus europaeus*, and the phylogenetic position of the Lipotyphla. *J. Mol. Evol.*, 41:952–957.

[158] Krieg, P.A., Varnum, S.M., Wormington, W.M., and Melton, D.A. (1989). The mRNA encoding elongation factor 1α (EF-1α) is a major transcript at the midblastula transition in Xenopus. *Dev. Biol.*, 133:93–100.

[159] Kuhner, M.K. and Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol. Biol. Evol.*, 11:459–468.

[160] Kuma, K. and Miyata, T. (1994). Mammalian phylogeny inferred from multiple protein data. *Jpn. J. Genet.*, 69:555–566.

[161] Kuma, K., Nikoh, N., Iwabe, N., and Miyata, T. (1995). Phylogenetic position of *Dictyostelium* inferred from multiple protein data sets. *J. Mol. Evol.*, 41:238–246.

[162] Kumar, S., Tamura, K., and Nei, M. (1993). *MEGA: Molecular Evolutionary Genetics Analysis, ver. 1.01.* Pennsylvania State Univ., University Park.

[163] Kurasawa, Y., Numata, O., Katoh, M., Hirano, H., Chiba, J., and Watanabe, Y. (1992). Identification Tetrahymena 14-nm filament-associated protein as elongation factor 1α. *Exp. Cell Res.*, 203:251–258.

[164] Lake, J.A. (1994). Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. *Proc. Natl. Acad. Sci. USA*, 91:1455–1459.

[165] Lawson, F.S., Charlebois, R.L., and Dillon, J.-A.R. (1996). Phylogenetic analysis of carbamoylphosphate synthetase genes: complex evolutionary history includes an internal duplication within a gene which can root the tree of life. *Mol. Biol. Evol.*, 13:970–977.

[166] Länge, S., Rozario, C., and Müller, M. (1994). Primary structure of the hydrogenosomal adenylate kinase of *Trichomonas vaginalis* and its phylogenetic relationships. *Mol. Biochem. Parasitol.*, 66:297–308.

[167] Lechner, K. and Böck, A. (1987). Cloning and nucleotide sequence of the gene for an archaebacterial protein synthesis elongation factor Tu. *Mol. Gen. Genet.*, 208:523–528.

[168] Lee, W.J. and Kocher, T.D. (1995). Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. *Genetics*, 139:873–887.

[169] Leeton, P.R., Christidis, L., Westerman, M., and Boles, W.E. (1994). Molecular phylogenetic affinities of the night parrot (*Geopsittacus occidentalis*) and the ground parrot (*Pezopotus wallicus*). *Auk*, 111:831–841.

[170] Lento, G.M., Hickson, R.E., Chambers, G.K., and Penny, D. (1995). Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Mol. Biol. Evol.*, 12:28–52.

[171] Liboz, T., Bardet, C., Le Van Thai, A., Axelos, M., and Lescure, B. (1989). The four members of the gene family encoding the *A. thaliana* translation elongation factor. *Plant Mol. Biol.*, 14:107–110.

[172] Linz, J.E., Lira, L.M., and Sypherd, P.S. (1986). The primary structure and the functional domains of an elongation factor-1α from *Mucor racemosus*. *J. Biol. Chem.*, 261:15022–15029.

[173] Lockhart, P.J., Steel, M.A., Hendy, M.D., and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol. Biol. Evol.*, 11:605–612.

[174] Luckett, W.P. and Hartenberger, J.-L. (1985). Evolutionary relationships among rodents: comments and conclusions. In Luckett, W. and Hartenberger, J.-L., editors, *Evolutionary Relationships among Rodents: A Multidisciplinary Analysis*, pages 685–712. Plenum Press, New York.

[175] Luckett, W.P. and Hartenberger, J.-L. (1993). Monophyly or polyphyly of the order Rodentia: possible conflict between morphological and molecular interpretations. *J. Mammal. Evol.*, 1:127–147.

[176] Ma, D.-P., Zharkikh, A., Graur, D., VandeBerg, J.L., and Li, W.-H. (1993). Structure and evolution of opossum, guinea pig, and porcupine cytochrome *b* genes. *J. Mol. Evol.*, 36:327–334.

[177] Maddison, W.P. and Maddison, D.R. (1992). *MacClade, version 3.0*. Sinauer, Sunderland, Massachusetts.

[178] Malcolm, B.A., Wilson, K.P., Matthews, B.W., Kirsch, J.F., and Wilson, A.C. (1990). Ancestral lysozymes reconstructed, neutrality tested, and thermostability linked to hydrocarbon packing. *Nature*, 345:86–89.

[179] Marsh, T.L., Reich, C.I., Whitelock, R.B., and Olsen, G.L. (1994). 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.

[180] Martignetti, J.A. and Brosius, J. (1993). Neural BC1 RNA as an evolutionary marker: Guinea pig remains a rodent. *Proc. Natl. Acad. Sci. USA*, 90:9698–9702.

[181] Martin, A.P. and Palumbi, S.R. (1993). Protein evolution in different cellular environments: cytochrome *b* in sharks and mammals. *Mol. Biol. Evol.*, 10:873–891.

[182] Meyer, A., Kocher, T.D., Basasibwaki, P., and Wilson, A.C. (1990). Monophyletic origin of Lake Victoria cichlid fishes suggested by mitochondrial DNA sequences. *Nature*, 347:550–553.

[183] Milinkovitch, M.C., LeDuc, R.G., Adachi, J., Farnir, F., Georges, M., and Hasegawa, M. (1996). Effects of character weighting and species sampling on phylogeny reconstruction: a case study based on DNA sequence data in cetaceans. *Genetics*, in press.

[184] Milinkovitch, M.C., Orti, G., and Meyer, A. (1993). Revised phylogeny of whales suggested by mitochondrial ribosomal DNA sequences. *Nature*, 361:346–348.

[185] Miyamoto, M.M. and Cracraft, J. (1991). *Phylogenetic Analysis of DNA Sequences*. Oxford Univ. Press, Oxford.

[186] Miyata, T., Hayashida, H., Kikuno, R., Hasegawa, M., Kobayashi, M., and Koike, K. (1982). 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.

[187] Miyata, T., Iwabe, N., Kuma, K., Kawanishi, Y., Hasegawa, M., Kishino, H., Mukohata, Y., Ihara, K., and Osawa, S. (1991). Evolution of archaebacteria: Phylogenetic relationships among archaebacteria, eubacteria, and eukaryotes. In Osawa, S. and Honjo, T., editors, *Evolution of Life: Fossils, Molecules, and Culture*, pages 337–351. Springer-Verlag, Tokyo.

[188] Montandon, P.E. and Stutz, E. (1990). Structure and expression of the *Euglena gracilis* nuclear gene coding for the translation elongation factor EF-1a. *Nucl. Acids. Res.*, 18:75–82.

[189] Mukohata, Y., Ihara, K., Kishino, H., Hasegawa, M., Iwabe, N., and Miyata, T. (1990). Close evolutionary relatedness of archaebacteria with eukaryotes. *Proc. Japan Acad.*, 66B:63–67.

[190] Muse, S.V. and Gaut, B.S. (1994). A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.*, 11:715–724.

[191] Nagashima, K., Kasai, M., Nagata, S., and Kaziro, Y. (1986). Structure of the two genes coding for polypeptide chain elongation factor 1-α (EF-1-α) from *Saccharomyces cerevisiae*. *Gene*, 45:265–273.

[192] Nakamura, Y., Hashimoto, T., Kamaishi, T., Adachi, J., Nakamura, F., Okamoto, K., and Hasegawa, M. (1996). Phylogenetic place of kinetoplastid protozoa inferred from protein phylogenies of elongation factors 1α and 2. *J. Biochem.*, 119:70–79.

[193] Nakamura, Y., Hashimoto, T., Yoshikawa, H., Kamaishi, T., Nakamura, F., Okamoto, K., and Hasegawa, M. (1996). Phylogenetic position of *Blastocystis hominis* that contains cytochrome-free mitochondria, inferred from the protein phylogeny of elongation factor 1α. *Mol. Biochem. Parasitol.*, in press.

[194] Naylor, G.J.P., Collins, T.M., and Brown, W.M. (1995). Hydrophobicity and phylogeny. *Nature*, 373:565–566.

[195] Nei, M. (1987). *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York.

[196] Neyman, J. (1971). Molecular studies of evolution: a source of novel statistical problems. In Gupta, S.S. and Yackel, J., editors, *Statistical Decision Theory and Related Topics*, pages 1–27. Academic Press, New York.

[197] Nikoh, N., Hayase, N., Iwabe, N., Kuma, K., and Miyata, T. (1994). Phylogenetic relationship of the kingdoms Animalia, Plantae, and Fungi inferred from twenty three different protein species. *Mol. Biol. Evol.*, 11:762–768.

[198] Novacek, M.J. (1992). Mammalian phylogeny: shaking the tree. *Nature*, 356:121–125.

[199] Nowak, R.M. (1991). *Walker's Mammals of the World, Fifth Edition*. Johns Hopkins Univ. Press, Baltimore.

[200] Olsen, G.J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *Comput. Appl. Biosci.*, 10:41–48.

[201] Orti, G. and Meyer, A. (1996). Molecular evolution of ependymin and the phylogenetic resolution of early divergences among euteleost fish. *Mol. Biol. Evol.*, 13:556–573.

[202] Orti, G., Petry, P., Porto, J.I.R., Jégu, M., and Meyer, A. (1996). Patterns of nucleotide change in mitochondrial ribosomal RNA genes and the phylogeny of piranhas. *J. Mol. Evol.*, 42:169–182.

[203] Ozawa, T., Tanaka, M., Ino, H., Ohno, K., Sano, T., Wada, Y., Yoneda, M., Tanno, Y., Miyatake, T., Tanaka, T., Itoyama, S., Ikebe, S., Hattori, N., and Mizuno, Y. (1991). Distinct clustering of point mutations in mitochondrial DNA among patients with mitochondrial encephalomyopathies and Parkinson's disease. *Biochem. Biophys. Res. Commun.*, 176:938–946.

[204] Painter, J., Krajewski, C.W., and Westerman, M. Molecular phylogeny for the marsupial genus Planigale (Dasyuridae). Unpublished.

[205] Perna, N.T. and Kocher, T.D. (1995). Unequal base frequencies and the estimation of substitution rates. *Mol. Biol. Evol.*, 12:359–361.

[206] Philippe, H. (1993). MUST: a computer package of management utilities for sequences and trees. *Nucl. Acids. Res.*, 21:5264–5272.

[207] Philippe, H. and Adoutte, A. (1995). How reliable is our current view of eukaryotic phylogeny? In Brugerolle, G. and Mignot, J.-P., editors, *Protistological Actualities (Proceedings of the Second European Congress of Protistology, Clermont-Ferrand, 1995)*, pages 17–33.

[208] Philippe, H. and Douzery, E. (1994). The pitfalls of molecular phylogeny based on four species, as illustrated by the Cetacea/Artiodactyla relationships. *J. Mammal. Evol.*, 2:133–152.

[209] Philippe, H. and Laurent, J. (August 17–24 1996). The clock needs some repair. In *Fifth International Congress of Systematic and Evolutionary Biology*, Budapest, Hungary.

[210] Pokalsky, A.R., Hiatt, W.R., Ridge, N., Rasmussen, R., Houck, C.M., and Shewmaker, C.K. (1989). Structure and expression of elongation factor 1α in tomato. *Nucl. Acids. Res.*, 17:4661–4673.

[211] Porter, C.A., Goodman, M., and Stanhope, M.J. (1996). Evidence on mammalian phylogeny from sequences of exon 28 of the von Willebrand factor gene. *Mol. Phyl. Evol.*, 5:89–101.

[212] Reeves, J.H. (1992). Heterogeneity in the substitution process of amino acid sites of proteins coded for by mitochondrial DNA. *J. Mol. Evol.*, 35:17–31.

[213] Retief, J.D., Winkfein, R.J., and Dixon, G.H. (1994). Evolution of monotremes — the sequences of the protamine P1 genes of platypus and echidna. *Eur. J. Biochem.*, 218:457–461.

[214] Ritland, K. and Clegg, M.T. (1987). Evolutionary analysis of plant DNA sequences. *Am. Nat.*, 130:S74–S100.

[215] Rodríguez, F., Oliver, J.L., Marín, A., and Medina, J.R. (1990). The general stochastic model of nucleotide substitution. *J. Theor. Biol.*, 142:485–501.

[216] Roe, B.A., Ma, D.-P., Wilson, R.K., and Wong, J.F.-H. (1985). The complete nucleotide sequence of the *Xenopus laevis* mitochondrial genome. *J. Biol. Chem.*, 260:9759–9774.

[217] Russo, C.A.M., Takezaki, N., and Nei, M. (1996). Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.*, 13:525–536.

[218] Rzhetsky, A. and Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Mol. Biol. Evol.*, 10:1073–1095.

[219] Saccone, C., Lanave, C., Pesole, G., and Preparata, G. (1990). Influence of base composition on quantitative estimates of gene evolution. *Methods in Enzymology*, 183:570–583.

[220] Saitou, N. (1990). Maximum likelihood methods. *Methods in Enzymology*, 183:584–598.

[221] Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425.

[222] Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., and Cedergren, R. (1992). Gene order comparisons for phylogenetic inference: evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA*, 89:6575–6579.

[223] Schöniger, M., Hofacker, G.L., and Borstnik, B. (1990). Stochastic traits of molecular evolution — acceptance of point mutations in native actin genes. *J. Theor. Biol.*, 143:287–306.

[224] Schmidt, T.R. and Gold, J.R. (1995). Molecular phylogenetics and evolution of the cytochrome *b* gene in the cyprinid genus *Lythrurus* (Actinopterygii: Cypriniformes). *Mol. Phyl. Evol.*, in press.

[225] Shimada, A., Kanai, S., and Maruyama, T. (1995). Partial sequence of ribulose-1,5-bisphosphate carboxylase/oxygenase and the phylogeny of *Prochloron* and *Prochlorococcus* (Prochlorales). *J. Mol. Evol.*, 40:671–677.

[226] Shirakura, T., Hashimoto, T., Nakamura, Y., Kamaishi, T., Cao, Y., Adachi, J., Hasegawa, M., Yamamoto, A., and Goto, N. (1994). Phylogenetic place of a mitochondria-lacking protozoan, *Entamoeba histolytica*, inferred from amino acid sequences of elongation factor 2. *Jpn. J. Genet.*, 69:119–135.

[227] Sibley, C.G. and Ahlquist, J.E. (1985). The relationships of some groups of African birds, based on comparisons of the genetic material, DNA. In Schuchmann, K.-L., editor, *Proceedings of the International Symposium on African Vertebrates: Systematics, Phylogeny and Evolutionary Ecology*, pages 115–161. Zoologisches Forschungsinstitut und Museum Alexander Koenig, Bonn.

[228] Sibley, C.G. and Ahlquist, J.E. (1990). *Phylogeny and Classification of Birds: A Study in Molecular Evolution*. Yale Univ. Press, New Haven.

[229] Sidow, A. (1994). Parsimony or statistics? *Nature*, 367:26–26.

[230] Springer, M.S. and Kirsch, J.A.W. (1993). 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.

[231] Stanhope, M.J., Smith, M.R., Waddell, V.G., Porter, C.A., Shivji, M.S., and Goodman, M. (1996). Mammalian evolution and the interphotoreceptor retinoid binding protein (IRBP) gene: convincing evidence for several superordinal clades. *J. Mol. Evol.*, 43:83–92.

[232] Stanley, H.F., Kadwell, M., and Wheeler, J.C. (1994). Molecular evolution of the Camelidae: a mitochondrial DNA study. *Proc. R. Soc. London*, B256:1–6.

[233] Steel, M.A., Székely, L., Erdös, P.L., and Waddell, P.J. (1993). A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *New Zealand J. Botany*, 31:289–296.

[234] Stewart, C.-B. (1993). The powers and pitfalls of parsimony. *Nature*, 361:603–607.

[235] Stewart, C.-B., Schilling, J.W., and Wilson, A.C. (1987). Adaptive evolution in the stomach lysozymes of foregut fermenters. *Nature*, 330:401–404.

[236] Strimmer, K. and von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 13:964–969.

[237] Sueoka, N. (1961). Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proc. Natl. Acad. Sci. USA*, 47:1141–1149.

[238] Sundstrom, P., Smith, D., and Sypherd, P.S. (1990). Sequence analysis and expression of the two genes for elongation factor 1-α from the dimorphic yeast *Candida albicans*. *J. Bacteriol.*, 172:2036–2045.

[239] Swofford, D.L. (1993). *PAUP: Phylogenetic Analysis Using Parsimony*. Illinois Natural History Survey, Champaign.

[240] Swofford, D.L., Olsen, G.J., Waddell, P.J., and Hillis, D.M. (1996). Phylogenetic inference. In Hillis, D.M., Moritz, C., and Mable, B.K., editors, *Molecular Systematics, 2nd Edition*, pages 407–514. Sinauer Associates, Sunderland, Massachusetts.

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

[242] Tamura, K. (1994). Model selection in the estimation of the number of nucleotide substitutions. *Mol. Biol. Evol.*, 11:154–157.

[243] Tamura, K. and Nei, M. (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.*, 10:512–526.

[244] Tanaka, M. and Ozawa, T. (1994). Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, 22:327–335.

[245] Tavaré, S. (1986). Some probabilistic and statistical problems on the analysis of DNA sequences. *Lec. Math. Life Sci.*, 17:57–86.

[246] Tedford, R.H. (1976). Relationships of pinnipeds to other carnivores (Mammalia). *Syst. Zool.*, 25:363–374.

[247] Tesch, A. and Klink, F. (1990). Cloning and sequencing of the gene coding for the elogation factor 1α from the archaebacterium *Thermoplasma acidophilum*. *FEMS Microbiol. Lett.*, 71:293–298.

[248] Thomas, W.K. and Martin, S.L. (1993). A recent origin of marmots. *Mol. Phyl. Evol.*, 2:330–336.

[249] Thorne, J.L. and Kishino, H. (1992). Freeing phylogenies from artifacts of alignment. *Mol. Biol. Evol.*, 9:1148–1162.

[250] Thorne, J.L., Kishino, H., and Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, 33:114–124.

[251] Thorne, J.L., Kishino, H., and Felsenstein, J. (1992). Inching toward reality: an improved likelihood model of sequence evolution. *J. Mol. Evol.*, 34:3–16.

[252] Tzeng, C.-S., Hui, C.-F., Shen, S.-C., and Huang, P.C. (1992). The complete nucleotide sequence of the *Crossostoma lacustre* mitochondrial genome: conservation and variations among vertebrates. *Nucl. Acids. Res.*, 20:4853–4858.

[253] Ueda, K. and Yoshinaga, K. (1995). Can *rbcL* tell us the phylogeny of green plants? In Nei, M. and Takahata, N., editors, *Current Topics on Molecular Evolution — Proceedings of the U.S.-Japan Workshop*, pages 97–103. Institute of Molecular Evolutionary Genetics, The Pennsylvania State University (USA), and Graduate University for Advanced Studies (Hayama, Japan).

[254] Uetsuki, T., Naito, A., Nagata, S., and Kaziro, Y. (1989). Isolation and characterization of the human chromosomal gene for polypeptide chain elongation factor 1-α. *J. Biol. Chem.*, 264:5791–5798.

[255] van Hemert, F.J., Amons, R., Pluijms, W.J.M., van Ormondt, H., and Möller, W. (1984). The primary structure of elongation factor EF-1α from the brine shrimp *Artemia*. *EMBO J.*, 3:1109–1113.

[256] Vrana, P.B., Milinkovitch, M.C., Powell, J.R., and Wheeler, W.C. (1994). Higher level relationships of arctoid Carnivora based on sequence data and "total evidence". *Mol. Phyl. Evol.*, 3:47–58.

[257] Waddell, P.J. (1995). *Statistical Methods of Phylogenetic Analysis: Including Hadamard Conjugations, LogDet Transforms, and Maximum Likelihood*. Ph.D. dissertation, Massey University.

[258] Waddell, P.J. and Steel, M.A. (1996). Time reversible distances allowing a distribution of rates across sites. *Research Report of Department of Mathematics and Statistics, University of Canterbury, New Zealand*, 143:1–28.

[259] Wainright, P.O., Hinkle, G., Sogin, M.L., and Stickel, S.K. (1993). Monophyletic origins of the metazoa: an evolutionary link with fungi. *Science*, 260:340–342.

[260] Weisburg, W.G., Giovannoni, S.J., and Woese, C.R. (1989). The *Deinococcus-Thermus* phylum and the effect of rRNA composition on phylogenetic tree reconstruction. *System. Appl. Microbiol.*, 11:128–134.

[261] Wettstein, P.J., Strausbauch, M., Lamb, T., States, J., Chakraborty, R., Jin, L., and Riblet, R. (1995). Phylogeny of six *Sciurus aberti* subspecies based on nucleotide sequences of cytochrome *b*. *Mol. Phyl. Evol.*, 4:150–162.

[262] Wyss, A. (1988). Evidence from flipper structure for a single origin of pinnipeds. *Nature*, 334:427–428.

[263] Wyss, A.R. and Flynn, J.J. (1993). A phylogenetic analysis and definition of the carnivora. In Szalay, F.S., Novacek, M.J., and McKenna, M.C., editors, *Mammal Phylogeny — Placentals*, pages 32–52. Springer-Verlag, New York.

[264] Wyss, A.R., Flynn, J.J., Norell, M.A., Swisher III, C.C., Charrier, R., Novacek, M.J., and McKenna, M.C. (1993). South America's earliest rodent and recognition of a new interval of mammalian evolution. *Nature*, 365:434–437.

[265] Xu, X. and Árnason, Ú. (1994). The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region. *Gene*, 148:357–362.

[266] Yamamoto, A., Hashimoto, T., Asaga, E., Hasegawa, M., and Goto, N. (1996). Phylogenetic position of mitochondrion-lacking protozoan, *Trichomonas tenax* based on amino acid sequences of elongation factors 1α and 2. *J. Mol. Evol.*, in press.

[267] Yamashina, Y. (1986). *A World List of Birds with Japanese Names*. Daigakusyorin, Tokyo.

[268] Yang, F., Demma, M., Warren, V., Dharmawardhane, S., and Condeelis, J. (1990). Identification of an actin-binding protein from *Dictyostelium* as elongation factor 1a. *Nature*, 347:494–496.

[269] Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.*, 10:1396–1401.

[270] Yang, Z. (1994). Estimating the pattern of nucleotide substitution. *J. Mol. Evol.*, 39:105–111.

[271] Yang, Z. (1994). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst. Biol.*, 43:329–342.

[272] Yang, Z. (1995). Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. *J. Mol. Evol.*, 40:689–697.

[273] Yang, Z. (1996). Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.*, 42:294–307.

[274] Yokobori, S., Hasegawa, M., Ueda, T., Okada, N., Nishikawa, K., and Watanabe, K. (1994). Relationship among coelacanths, lungfishes and tetrapods; a phylogenetic analysis based on mitochondrial cytochrome oxidase I gene sequences. *J. Mol. Evol.*, 38:602–609.

[275] Zardoya, R., Garrido-Pertierra, A., and Bautista, J.M. (1995). The complete nucleotide sequence of mitochondrial DNA genome of the rainbow trout, *Oncorhynchus mykiss*. *J. Mol. Evol.*, 41:942–951.

[276] Zardoya, R. and Meyer, A. (1996). The complete nucleotide sequence of the mitochondrial genome of the lungfish (*Protopterus dolloi* supports its phylogenetic position as a close relative of land vertebrates. *Genetics*, 142:1249–1263.

[277] Zardoya, R. and Meyer, A. (1996). The origin of tetrapods based on 28S ribosomal RNA sequences. *Proc. Natl. Acad. Sci. USA*, 93:5449–5454.

[278] Zardoya, R. and Meyer, A. (1996). Phylogenetic performance of mitochondrial protein coding genes in resolving relationships among vertebrates. *Mol. Biol. Evol.*, 13:933–942.

[279] Zharkikh, A. (1994). Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, 39:315–329.