Maximum Likelihood Methods in Molecular Phylogenetics

Korbinian Sebastian Strimmer

Dissertation
der Fakultät für Biologie
der Ludwig-Maximilians-Universität München

Oktober 1997

1. Gutachter: Priv. Doz. Dr. Arndt von Haeseler
2. Gutachter: Prof. Dr. Rainer Uhl

Tag der mündlichen Prüfung: 24. November 1997

Platzhalter für Verlagstitelblatt

Platzhalter für Impressumsseite

Meinen Eltern

Die vorliegende Doktorarbeit wurde unter der Betreuung von Priv. Doz. Dr. Arndt von Haeseler am Lehrstuhl von Prof. Dr. Svante Pääbo angefertigt.

All denjenigen, die mir dabei in den letzten drei Jahren mit Rat und Tat zur Seite gestanden sind, möchte ich sehr herzlich danken.

Ganz besonders ich möchte mich bei Priv. Doz. Dr. Arndt von Haeseler für die engagierte Betreuung dieser Arbeit bedanken. Es war stets spannend, im Dialog mit ihm neue Ideen zu entwickeln und zu diskutieren. Durch seine Förderung war es mir außerdem möglich, einen zweimonatigen Forschungsaufenthalt an der Universität Cambridge in Großbritannien durchzuführen.

Mein weiterer Dank gilt Prof. Dr. Svante Pääbo und den Mitarbeitern seiner multinationalen Arbeitgruppe für zahlreiche interessante und aufschlußreiche Diskussionen und Seminare sowie für das angenehme Arbeitsklima.

Ich danke außerdem Dr. Nick Goldman für seine Gastfreundschaft, die er mir in Cambridge gewährt hat, sowie für die fruchtbare Zusammenarbeit.

Für die Finanzierung dieser Arbeit danke ich der Deutschen Forschungsgemeinschaft.

Ein Dank auch an alle gegenwärtigen und ehemaligen Mitgliedern der Arbeitsgruppe, mit denen problemlos eine Flasche Wein oder ein paar Bier beseitigt werden konnten.

Ein Teil dieser Arbeit basiert auf den folgenden Veröffentlichungen:

Korbinian Strimmer und Arndt von Haeseler. 1996. Quartet-puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**:964–969.

Korbinian Strimmer und Arndt von Haeseler. 1997. Likelihood-mapping: A simple method to visualize phylogenetic content of a sequence alignment. *Proc. Natl. Acad. Sci. USA* **94**:6815–6819.

Contents

Chapter 1

Overview

1.1 Introduction

Molecular phylogenetics deals with inferring phylogenetic relationships from molecular sequence data. Of the many techniques described (Swofford *et al.*, 1996) maximum-likelihood methods are special due to their conceptual simplicity and their well-defined statistical basis. In principle a maximum-likelihood analysis consists of three parts. First a model of evolutionary change for nucleotides or amino acids is specified. Then, based on this model, different hypotheses about the evolutionary history are evaluated in terms of the probability that the hypothesized history would give rise to the observed data. Finally, the hypothesis is selected that shows the highest probability. Maximum-likelihood often yields estimates with a lower variance than other methods, and it is frequently the estimation method least affected by sampling error (Swofford *et al.*, 1996). In addition, maximum-likelihood also seems to be quite robust against violations of the assumptions used in the underlying model (Huelsenbeck, 1995). This is part of the power of the approach.

However, maximum-likelihood methods have drawbacks as well. Most importantly, they often are computationally very expensive. This can mostly be attributed to two factors. First, in order to find the optimal solution usually a large number of alternative hypotheses have to be evaluated. Consider, e.g., the number of different unrooted binary trees

$$B(N) = \prod_{i=3}^{N}(2i-5). \tag{1.1}$$

for N sequences (Felsenstein, 1978). As $B(N)$ grows exponentially with N it is virtu-

ally impossible for any computer to compare all trees even if the number of sequences is only moderately large. Second, the complexity of computing the probability for a specific hypothesis can be prohibiting as well. The advent of powerful computers has only partially resolved this situation.

1.2 Subject of this thesis

In this thesis we introduce heuristic methods for use in molecular phylogeny that enable the application of maximum-likelihood even for large data sets. First we provide in Chapter 2 an introduction to models of sequence evolution and to maximum-likelihood. Then we study in Chapter 3 the problem to obtain maximum-likelihood estimates for the parameters of a model of sequence evolution. We describe a number of useful simplifications to speed up the estimation of evolutionary parameters for given data set. In Chapter 4 we focus on quartet-puzzling, a heuristic tree search based on maximum-likelihood tree reconstruction for all sets of four sequences that can be formed for a given data set. The overall tree is then recovered by finding a tree topology reconciling all quartet topologies. This method allows the reconstruction of trees for a large number of sequences. In addition, quartet-puzzling computes estimates of support for all internal branches. In Chapter 5 we present likelihood-mapping, an approach for assessing and visualizing the phylogenetic content of a sequence alignment. This method is based on the evaluation of quartets of sequences as well. It allows to quickly determine the phylogenetic signal present in a given data set. Moreover, likelihood-mapping can also be applied to investigate whether a hypothesized grouping of sequences is supported by the data. In contrast to bootstrap techniques this method can be used even for very large data sets. Likelihood-mapping can be viewed as complementary approach to the so-called statistical geometry in sequence space. Finally, we describe in Appendix A the PUZZLE software, which implements all methods introduced here.

Chapter 2

Maximum Likelihood

In this chapter we give an introduction to models of sequence evolution and to maximum-likelihood. We model the substitution of nucleotides and amino acids by a homogeneous stationary stochastic process and assign relative rates to each sequence position using discrete probability distributions. Then we show how to calculate the likelihood of a data set for a given evolutionary history and explain the principle of maximum-likelihood.

2.1 Introduction

Models of nucleotide or amino acid evolution play an essential role in the analysis of molecular sequence data (Swofford *et al.*, 1996). They are a tool to reduce the enormous complexity of the biological mutation process to a comparatively simple pattern that can be described by a small number of parameters. The models of sequence evolution considered here consist of two parts. First, they specify a modus of substitution for nucleotides or amino acids at a given site. Second, they give a prescription how the rate of substitutions is distributed over different positions in a sequence. This is called a model of rate heterogeneity. Examples for models describing the substitution process of nucleotides are the JC model (Jukes and Cantor, 1969) or the HKY model (Hasegawa *et al.*, 1985). A common model of rate variation is to introduce invariable sites (Fitch and Margoliash, 1967; Hasegawa *et al.*, 1985; Churchill *et al.*, 1992) and to assume a Γ-distribution for the rates of the variable positions (Wakeley, 1993; Yang, 1993; Gu *et al.*, 1995).

In the following, we consider a set of aligned sequences that are typically derived

```
Frog    CAGTGATAAACATTGAAC-ATGAGCGAAGCTCGAT
Bird    CAGTAATTAACCTTAAGCAATAAGTGTAACTTGAC
Human   CAGTGATTAACCTTTAGCAATAAACGAAGTTTAAC
Seal    CAGTAATAAAAATTAAGCTATGAACGAAGTTTGAC
Cow     CAGTGACAAAAATTAAGCCATAAACGAAGTTTGAC
Whale   CAGTGATAAAAATTAAGCTATAA-CGAAGTTCGAC
Mouse   CAGTGATAAATATTAAGCAATAAACGAAGTTTGAC
Rat     CAGTGATAAATATTAAGCAATGAACGAAGTTTGAC
```

Figure 2.1: *Example of a nucleotide alignment. The letters A, C, G, and T code for the four nucleotides, dashes represent gaps (insertions or deletions).*

from different species (Figure 2.1). In an alignment sequences positions that trace back to a common ancestor are lined up in the same column. This is done by introducing insertions or deletions (gaps) in one ore more of the sequences. Thus, an alignment identifies homologous positions in a set of DNA sequences. The problem to infer the correct alignment will not be discussed here (Thompson *et al.*, 1994), we assume that an alignment is given. Each column s in the alignment defines a so-called site pattern D_s. Usually site patterns with gaps are excluded from the analysis to remove insertion or deletion events from the data. A site is said to be constant if the site pattern contains only one sort of nucleotide or amino acid. Computation of the probabilities to observe a specific site pattern D_s for a given evolutionary history T is the basic idea behind the maximum-likelihood framework. This, however, requires that a specific model of sequence evolution is specified.

2.2 Substitution process

A DNA sequence is a collection of the four nucleotides adenine (A), cytosine (C), guanine (G), and thymine (T). Therefore there exist $n = 4$ different states for a sequence position. For amino acid sequences the number of possible states is $n = 20$ corresponding to the number of amino acids that are used in protein synthesis. It is commonly assumed that the substitution of nucleotides or amino acids is a stationary stochastic process (Swofford *et al.*, 1996). This implies that nucleotide or amino acid frequencies π_i do not change over time and from sequence to sequence in a data set. The substitution process is described by the transition probability matrix $\mathbf{P}(k)$. It con-

sists of the probabilities $P_{ij}(k)$ to get from state i to state j after k substitutions at a site. Note, k can also take on fractional values. In a tree k is also called branch length. As a probability $P_{ij}(k)$ satisfies

$$\sum_{j=1}^{n} P_{ij}(k) = 1 \tag{2.1}$$

with initial values

$$P_{ij}(0) = \begin{cases} 1 & \text{for } i = j \\ 0 & \text{for } i \neq j\,. \end{cases} \tag{2.2}$$

The total probability h that a substitution occurred at a site is

$$h = 1 - \sum_{i=1}^{n} \pi_i P_{ii}(k). \tag{2.3}$$

The number of positions where two DNA sequences are different divided by the length l of the alignment, i.e. the so-called observed distance, is an estimate of h. Solving Equation 2.3 for k allows to infer the so-called expected distance k, i.e. the number of substitutions that actually occurred. Usually, the substitution process described by $\mathbf{P}(k)$ is also assumed to be reversible, i.e.

$$\pi_i P_{ij}(k) = \pi_j P_{ji}(k). \tag{2.4}$$

This assumption, also known as detailed balance, ensures local substitution equilibrium. Moreover, it turns out that it simplifies many calculations. If k is small it is possible to linearly approximate the transition probability matrix $\mathbf{P}(k)$ by

$$\mathbf{P}(k) \approx \mathbf{P}(0) + k\mathbf{R}. \tag{2.5}$$

$\mathbf{R}$ is called rate matrix (Tavaŕe, 1986). In order not to violate Equation 2.1 it satisfies

$$\sum_{j=1}^{n} R_{ij} = 0. \tag{2.6}$$

As the expected and the observed number of substitutions are identical for small k it follows from Equations 2.3 and 2.5 that $\mathbf{R}$ also obeys

$$\sum_{i=1}^{n} \pi_i R_{ii} = -1. \tag{2.7}$$

For a reversible process the rate matrix $\mathbf{R}$ can be decomposed into so-called rate parameters Q_{ij} and frequencies π_i (Yang, 1994a)

$$R_{ij} = \begin{cases} Q_{ij}\pi_j & \text{for } i \neq j \\ -\sum_{m=1}^{n} Q_{im}\pi_m & \text{for } i = j\,. \end{cases} \tag{2.8}$$

The matrix $\mathbf{Q} = (Q_{ij})$ is symmetric, $Q_{ij} = Q_{ji}$, and has vanishing diagonal entries, $Q_{ii} = 0$. As a consequence of Equation 2.7 $\mathbf{Q}$ is constrained by

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \pi_i Q_{ij} \pi_j = 1. \tag{2.9}$$

If we further assume that the substitution of nucleotides or amino acids is a homogeneous Markov process then $\mathbf{R}$ determines $\mathbf{P}(k)$ for all k by

$$\mathbf{P}(k) = \mathbf{P}(0) + k\mathbf{R} + \frac{(k\mathbf{R})^2}{2!} + \frac{(k\mathbf{R})^3}{3!} + \ldots = \exp(\mathbf{R}k). \tag{2.10}$$

This matrix exponential can be computed for reversible $\mathbf{R}$ by spectral decomposition

$$P_{ij}(k) = \sum_{m=1}^{n} \exp(\lambda_m k) U_{mi} U_{jm}^{-1} \tag{2.11}$$

where the λ_i are the eigenvalues of $\mathbf{R}$, $\mathbf{U} = (U_{ij})$ is the matrix with corresponding eigenvectors (note indices!), and $\mathbf{U}^{-1}$ is the inverse of $\mathbf{U}$.

2.3 Models of substitution

Choosing a model of nucleotide or amino acid substitution amounts to specifying explicit values for $\mathbf{Q}$ and π_i. In the most general case there are $n - 1$ independent frequency parameters π_i because of $\sum_{i=1}^{n} \pi_i = 1$ and $n(n-1)/2 - 1$ independent rate parameters because of $Q_{ij} = Q_{ji}$, $Q_{ii} = 0$, and Equation 2.9.

For nucleotides ($n = 4$) a number of symmetries of the substitution process can be assumed that help to further reduce the number of independent parameters. Substitutions of nucleotides can be divided into two groups, so-called transitions (Ts) and transversions (Tv). Substitutions between pyrimidines (C $\rightleftharpoons$ T) are pyrimidine transitions (Ts_Y) and substitutions between purines (A $\rightleftharpoons$ G) are purine transitions (Ts_R). All other substitutions where a purine is exchanged by a pyrimidine or vice versa (A $\rightleftharpoons$ C, A $\rightleftharpoons$ T, C $\rightleftharpoons$ G, G $\rightleftharpoons$ T) are called transversions. Each of the six possibilities for a substitution are represented by an entry in $\mathbf{Q}$.

In order to simplify calculations the four rate parameters describing transversions are assumed to be identical. If we moreover assume that the two possible transitions have the same rate parameter and that the ratio of the Ts rate parameter to the Tv rate parameter is $2\,t$ then the matrix $\mathbf{Q}$ reduces to

$$Q_{ij}^{*} = \begin{cases} 2t & \text{for Ts} \\ 1 & \text{Tv}\,. \end{cases} \tag{2.12}$$

This is the so-called HKY model (Hasegawa *et al.*, 1985) with parameter t. Note, the choice of 1 for the Tv rate parameter is arbitrary as only relative values are important. The star indicates that the matrix still has to be rescaled to conform to Equation 2.9. If the base frequencies are uniform ($\pi_i = 1/4$) the HKY model degenerates to the Km model (Kimura, 1980). For $t = 1/2$ the HKY model is called F81 model (Felsenstein, 1981), and the Km model further reduces to the JC model (Jukes and Cantor, 1969).

However, if the assumption of identical Ts rate parameters is relaxed by introducing the parameter γ, the ratio of the Ts_Y rate parameter to the Ts_R rate parameter, then the TN model (Tamura and Nei, 1993) is obtained with

$$Q^*_{ij} = \begin{cases} 2t(\frac{2\gamma}{\gamma+1}) & \text{for} \quad \text{Ts}_\text{Y} \\ 2t(\frac{2}{\gamma+1}) & \quad\quad\ \ \text{Ts}_\text{R} \\ 1 & \quad\quad\ \ \text{Tv}\,. \end{cases} \tag{2.13}$$

This is the most general model of nucleotide substitution discussed here. For $\gamma = 1$ the simpler HKY model is recovered.

Typical characteristics of a model of nucleotide substitution are the expected Ts/Tv ratio and the expected Ts_Y/Ts_R ratio. For any model the fraction, $\#$, of expected Ts_Y, Ts_R, and Tv among all substitutions are

$$\#\text{Ts}_\text{Y} = 2\pi_\text{C} Q_\text{CT} \pi_\text{T}, \tag{2.14}$$

$$\#\text{Ts}_\text{R} = 2\pi_\text{A} Q_\text{AG} \pi_\text{G}, \tag{2.15}$$

$$\#\text{Tv} = 2(\pi_\text{A} Q_\text{AC} \pi_\text{C} + \pi_\text{A} Q_\text{AT} \pi_\text{T} + \pi_\text{C} Q_\text{CG} \pi_\text{G} + \pi_\text{C} Q_\text{CT} \pi_\text{T}). \tag{2.16}$$

The expected Ts/Tv ratio and the expected Ts_Y/Ts_R ratio is then readily obtained:

$$\text{expected Ts/Tv ratio} = \frac{\#\text{Ts}_\text{Y} + \#\text{Ts}_\text{R}}{\#\text{Tv}}\,, \tag{2.17}$$

$$\text{expected } \text{Ts}_\text{Y}/\text{Ts}_\text{R} \text{ ratio} = \frac{\#\text{Ts}_\text{Y}}{\#\text{Ts}_\text{R}}\,. \tag{2.18}$$

It is easy to verify that for the Km model the parameter t is numerically identical to the expected Ts/Tv ratio. The JC model predicts an expected Ts/Tv ratio of $1/2$. Note, both the expected Ts/Tv ratio and the expected Ts_Y/Ts_R ratio cannot directly be observed in a data set. This is because it is not possible to trace multiple substitutions at site. The same reasoning also explains why only the difference h and not the expected number of substitutions k is observable (cf. Equation 2.3).

When the substitution of amino acids ($n = 20$) is modelled a simplification of $\mathbf{Q}$ comparable to nucleotides is not available. Instead, a number of different numerically fixed $\mathbf{Q}$ derived from empirical substitution matrices are used. For amino acids encoded on nuclear DNA, e.g., the Dayhoff model (Dayhoff *et al.*, 1978) or the JTT model (Jones *et al.*, 1992) have been suggested. For mitochondrial proteins the mtREV model has been proposed (Adachi and Hasegawa, 1996b).

2.4 Models of rate heterogeneity

It is well-known that the number of substitutions at a site is strongly dependent on its position along the DNA sequence (Li and Graur, 1991). One obvious reason is a functional constraint on a specific site. Then substitutions are unlikely. On the other hand, positions with few functional constraints can easier accumulate substitutions. There exist different approaches to account in a model of sequence evolution for the site-dependent variation of the number of substitutions.

The simplest model is to assume that a certain fraction θ of the sequence sites is invariable. Sites s where substitutions are impossible are a assigned a relative rate $r_s = 0$ whereas for variable positions the relative rate is $r_s = 1$. This model is called the two-rate model (Fitch and Margoliash, 1967; Hasegawa *et al.*, 1985; Churchill *et al.*, 1992). Note, though all invariable sites are constant, not all constant sites in an alignment are necessarily invariable.

The most commonly used Γ-model (Uzzel and Corbin, 1971; Wakeley, 1993) utilizes a Γ-distribution with expectation 1.0 and variance $1/\alpha$

$$g(r) = \frac{\alpha^{\alpha} r^{\alpha-1}}{\exp(\alpha r)\,\Gamma(\alpha)} \tag{2.19}$$

to assign relative rates to the sequence positions. By varying the shape parameter α two different scenarios are taken into account (Figure 2.2). For weak rate heterogeneity over sites ($\alpha > 1$) the distribution is bell-shaped. The relative rates r_s drawn from this distribution are all approximately 1.0. For strong rate heterogeneity ($\alpha < 1$) the Γ-distribution is L-shaped. This indicates that there are positions in the data sets that have very large relative rates r_s whereas many other sites are almost invariable ($r_s \approx 0$).

Calculations involving the continuous Γ-distribution are very tedious. Therefore an approximating discrete distribution is used in practice (Yang, 1994b). In other words, the relative rates r_s for each site are drawn from a set of c different rates $q_1, q_2, \ldots, q_c$.

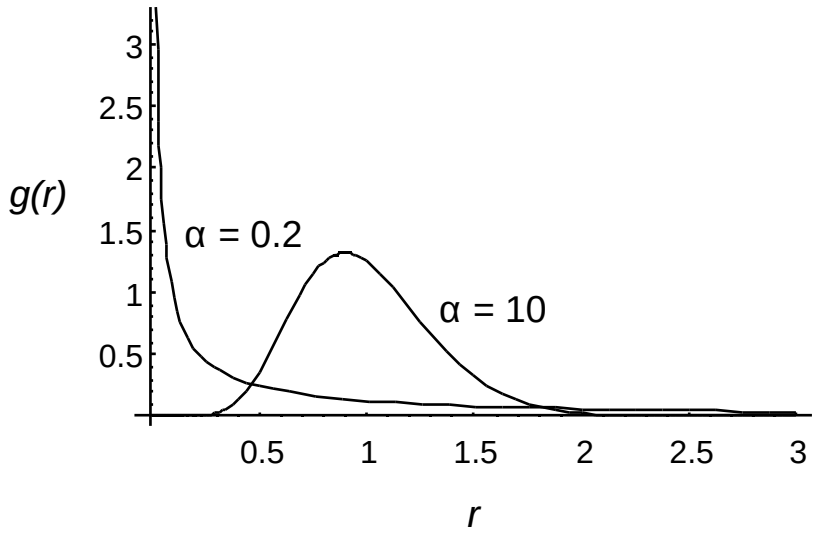

Figure 2.2: *Bell-shaped ($\alpha = 10$) and L-shaped ($\alpha = 0.2$) Γ-distribution.*

Usually, the number of rate categories is considered as fixed in advance and not treated as additional model parameter. To obtain the discrete Γ-distribution first the cumulative density function for $g(r)$ (Equation 2.19)

$$\text{cdf}(x) = \int_0^x g(r)\, dr = 1 - \frac{\Gamma(\alpha, \alpha x)}{\Gamma(\alpha)} \tag{2.20}$$

is computed. $\Gamma(a, b)$ is the incomplete Γ-function with $\Gamma(a, 0) = \Gamma(a)$. If c equally probable rate categories are admitted the rate q_i corresponding to each rate category $i \in \{1, 2, \ldots, c\}$ is obtained using the inverse of the cumulative density function

$$q_i = \text{cdf}^{-1}(\frac{2i-1}{2c}). \tag{2.21}$$

This general approach also works for other continuous probability distributions as well. Note that the mean of a discrete distribution generally deviates slightly from the mean of the corresponding continuous distribution. Therefore the rates q_i are rescaled in order to obtain an average of one. Yang (1994b) showed that the discrete Γ-distribution can provide a good approximation with as few as four rate categories.

A third mixed model simply combines invariable sites with Γ-distributed rates for variable positions (Gu *et al.*, 1995).

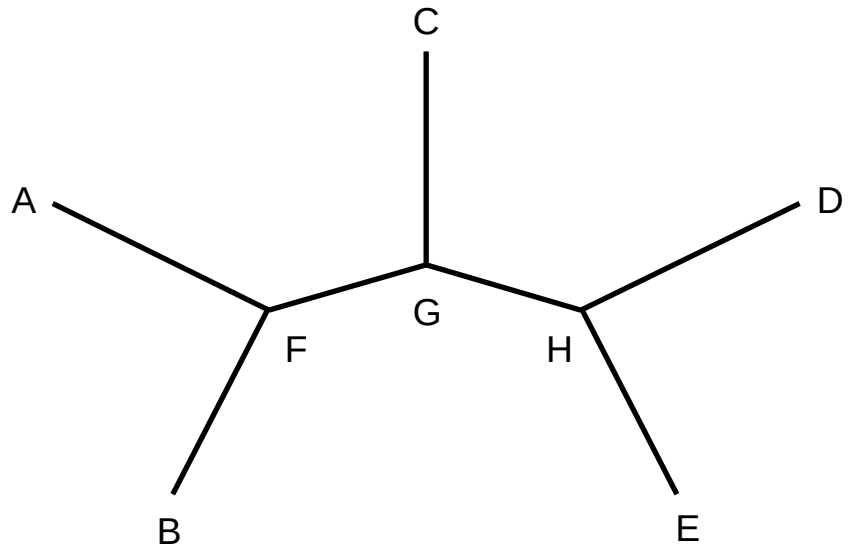

Figure 2.3: *An evolutionary tree of five sequences. The known sequences are at the terminal nodes (A, B, C, D, E) whereas the internal nodes (F, G, H) represent unknown ancestral sequences.*

2.5 Likelihood function

As soon as a model of sequence evolution $M = (t, \gamma, \pi_i, \theta, \alpha)$ is selected it is straightforward to compute the probability to observe a data set D if the evolutionary history T = (branching pattern, branch lengths) underlying the sequences is known. This is done using a so-called likelihood function L, which plays a central role in all applications of maximum-likelihood.

The basic idea is to compute the probability Prob$(D_s|T, M, r_s)$ that the site pattern D_s with relative rate r_s is the result of the evolutionary process M along T. To illustrate the calculation we use a tree connecting five sequences (Figure 2.3). In the tree the site pattern $D_s = (x_{\rm A}, x_{\rm B}, x_{\rm C}, x_{\rm D}, x_{\rm E})$ is observed at the terminal nodes labeled by the sequences A, B, C, D, E. The ancestral states $(x_{\rm F}, x_{\rm G}, x_{\rm H})$ at the internal nodes labeled F, G, H are unknown. In order to compute Prob$(D_s|T, M, r_s)$ we have to choose a hypothetical root node at any convenient location in the tree. As the model of substitution assumed is reversible the choice of the root does not influence the likelihood (Felsenstein, 1981). In the tree illustrated in Figure 2.3 we select node G as root. First the prior probability $\pi_{x_{\rm G}}$ of the state $x_{\rm G}$ at the root node G is determined. Then proceeding from the root to the external nodes of the tree the transition probabilities for

each branch are calculated. The product

$$\begin{aligned} \text{Prob}(D_s|T,M,r_s,x_{\text{F}},x_{\text{G}},x_{\text{H}}) &= \pi_{x_{\text{G}}} P_{x_{\text{G}}x_{\text{C}}}(r_s\, k_{\text{GC}}) \times \\ & P_{x_{\text{G}}x_{\text{F}}}(r_s\, k_{\text{GF}}) P_{x_{\text{F}}x_{\text{A}}}(r_s\, k_{\text{FA}}) \times \\ & P_{x_{\text{F}}x_{\text{B}}}(r_s\, k_{\text{FB}}) P_{x_{\text{G}}x_{\text{H}}}(r_s\, k_{\text{GH}}) \times \\ & P_{x_{\text{H}}x_{\text{D}}}(r_s\, k_{\text{HD}}) P_{x_{\text{H}}x_{\text{E}}}(r_s\, k_{\text{HE}}) \end{aligned} \tag{2.22}$$

is the likelihood to observe D_s given the states $x_{\text{F}}, x_{\text{G}}, x_{\text{H}}$ at the internal nodes. In this formula k_{XY} denotes the expected number of substitutions (branch lengths) that occur between nodes X and Y. By summing over all n^3 possible configurations for the unknown ancestral states $x_{\text{F}}, x_{\text{G}}, x_{\text{H}}$ we obtain

$$\text{Prob}(D_s|T,M,r_s) = \sum_{m=1}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \text{Prob}(D_s|T,M,r_s,i,j,m)\,. \tag{2.23}$$

A lot of computational effort is saved by rearranging Equation 2.23 to a sum with only n terms:

$$\begin{aligned} \text{Prob}(D_s|T,M,r_s) &= \sum_{m=1}^{n} \pi_m P_{mx_{\text{C}}}(r_s\, k_{\text{GC}}) \times \\ & \left(\sum_{i=1}^{n} P_{mi}(r_s\, k_{\text{GF}}) P_{ix_{\text{A}}}(r_s\, k_{\text{FA}}) P_{ix_{\text{B}}}(r_s\, k_{\text{FB}})\right) \times \\ & \left(\sum_{j=1}^{n} P_{mj}(r_s\, k_{\text{GH}}) P_{jx_{\text{D}}}(r_s\, k_{\text{HD}}) P_{jx_{\text{E}}}(r_s\, k_{\text{HE}})\right)\,. \end{aligned} \tag{2.24}$$

Similar formulas are available to compute $\text{Prob}(D_s|T,M,r_s)$ for other tree topologies. In the special case of an invariable site ($r_s = 0$) the formula Equation 2.24 degenerates to $\text{Prob}(D_s|T,M,r_s) = \pi_{x_{\text{A}}}$ where x_{A} is the state observed in all external and internal nodes. The likelihood for a pair of sequences A and B is $\text{Prob}(D_s|x_{\text{A}},x_{\text{B}},k_{\text{AB}},M,r_s) = \pi_{x_{\text{A}}} P_{x_{\text{A}}x_{\text{B}}}(r_s\, k_{\text{AB}})$ with A as arbitrary root.

The likelihood $L = \text{Prob}(D|T,M)$ for a data set D given T and M is the product of the probabilities $\text{Prob}(D_s|T,M,r_s)$ for each site. Note, this implicitly assumes that the rates of the sites are independent from each other. As the assignments of relative rates to sequence positions are usually unknown L is computed as positionwise average of the c rate categories

$$L = \text{Prob}(D|T,M) = \prod_{s=1}^{l} \left(\sum_{i=1}^{c} p_i\, \text{Prob}(D_s|T,q_i)\right) \tag{2.25}$$

where l is the length of the sequence alignment. The rates $q_1, q_2, \ldots, q_c$ and their prior probabilities $p_1, p_2, \ldots, p_c$ depend on the model of rate heterogeneity. For uniform rates over sites there is only one possible rate, therefore $c = 1$, $q_1 = 1, p_1 = 1$. For the two-rate model taking into account variable and invariable sites the rates are $q_1 = 0, q_2 = 1$ and the corresponding prior probabilities are $p_1 = \theta, p_2 = 1 - \theta$. For a discrete Γ-model with c categories rates q_i are determined by Equation 2.21 with $p_i = 1/c$. In a mixed model with in total c categories the rate q_1 equals zero to allow for invariable sites, the other rates $q_2, q_2, \ldots, q_c$ are given by a discrete Γ-distribution with $c - 1$ categories. The corresponding prior probabilities are $p_1 = \theta$ and $p_2 = p_3 = \ldots = p_c = (1 - \theta)/(c - 1)$.

2.6 Principle of maximum-likelihood

Maximum-likelihood aims to maximize L for a given data set with respect to T and M. This task consists of two different optimization problems.

For a fixed tree topology the branch lengths and the model parameters $M = (t, \gamma, \pi_i, \theta, \alpha)$ have to be varied simultaneously in order to maximize L. This is a difficult numerical problem that is typically approached by iteratively optimizing each branch and each model parameter separately using Brent's or Newton-Raphson's method (e.g., Olsen *et al.*, 1994; Felsenstein and Churchill, 1996). A short cut to this time-consuming procedure is, e.g., to use approximate maximum-likelihood branch lengths as described in Chapter 3.

The second difficulty is to find the most likely branching pattern. In order to solve this combinatorial problem it is necessary to evaluate all different tree topologies that are possible with the given number of sequences. It is clear that an exhaustive tree search of this kind is impossible even for moderately sized data sets as the number of different trees grows exponentially with the number of sequences (Equation 1.1). Therefore a number of heuristic tree search algorithms have been devised to exclude unlikely tree topologies from maximum-likelihood evaluation (Swofford *et al.*, 1996). One recent example is the quartet-puzzling method that is presented in Chapter 4.

Chapter 3

Parameter Estimation

We discuss the problem to obtain maximum-likelihood estimates for the parameters of a model of sequence evolution. Then we describe a number of simple strategies to speed up the estimation of evolutionary parameters for given data set. The performance of the methods is examined by a simulation study. As example we estimate the rate heterogeneity of an amino acid alignment consisting of 19 complete vertebrate mtDNA sequences.

3.1 Introduction

Choosing an optimal model of sequence evolution for a given data set is important, e.g., for the computation of branch lengths in a tree (Sullivan *et al.*, 1996). In the most general model discussed in Chapter 2 it requires that optimal values for the parameters $M = (t, \gamma, \pi_i, \theta, \alpha)$ are selected. In principle, they can be determined as maximum-likelihood estimates $\hat{M}$ simultaneously with the estimation of a tree $\hat{T}$ = (branching pattern, branch lengths). However, this is computationally not practicable except for small data sets (Gu *et al.*, 1995; Swofford *et al.*, 1996).

In this chapter we present a number of approximative schemes to speed up the maximum-likelihood estimation of M. As first simplification (P1) we consider the computation of approximate maximum-likelihood branch lengths. Next we discuss a procedure (P2) based on iterative reconstruction of a tree topology using neighbor-joining. Then we introduce quartet-sampling (QS), which focuses on randomly selected sets of four sequences of the data set. In a simulation study using artificially generated nucleotide data following clock-like and non-clock-like evolution the suit-

ability of each procedure to infer the parameters M is investigated. Finally, as an example we estimate the rate heterogeneity of an amino acid alignment containing 19 complete vertebrate mtDNA sequences.

3.2 Simplifying procedures

In order to lessen the computational effort involved in the inference of maximum-likelihood estimates $\hat{M} = (\hat{t}, \hat{\gamma}, \hat{\pi}_i, \hat{\theta}, \hat{\alpha})$ simultaneously with $\hat{T}$ = (branching pattern, branch lengths) a number of simplifying procedures are conceivable. The principal aim is to reduce the number of independent parameters of the likelihood function L. Consider, e.g., the stationary frequencies π_i. They can be kept fixed in L as they are equally well estimated from the observed average nucleotide or amino acid composition of the data set.

A prohibiting factor in the inference is the optimization of the branch lengths. As for a completely resolved tree with N sequences there are $2N - 3$ different branches it can be quite time-consuming (Felsenstein, 1981; Olsen *et al.*, 1994). However, we focus on determining maximum-likelihood estimates for the model parameters M and not for branch lengths. Then the tedious optimization of branch lengths can be avoided by employing so-called approximate maximum-likelihood branch lengths. This approximation has first been used by Adachi and Hasegawa (1996a) to single out unlikely tree topologies in an exhaustive tree search. The following procedure P1 to find maximum-likelihood estimates of M for a fixed tree topology takes advantage of this short cut as well.

Procedure P1:

1. Choose start values for M and for the length of each branch (initial step).

2. Obtain estimates $\hat{M}$ by maximizing the likelihood function L.

3. Compute pairwise maximum-likelihood distances based on $\hat{M}$ and obtain approximative maximum-likelihood branch lengths as least-squares fit of the distances to the tree, following Adachi and Hasegawa (1996a).

4. Repeat steps 2 and 3 until $\hat{M}$ does not change any more.

It is much quicker to determine approximate maximum-likelihood branch lengths (step 3) than to find true maximum-likelihood estimates by maximizing the likelihood function L with respect to the branch lengths of the tree. It can be shown that the likelihood value computed on the basis of the non-optimized branch lengths nevertheless is a very good approximation to the maximum-likelihood (Adachi and Hasegawa, 1996a).

Another way to speed up maximum-likelihood estimation of M is to determine the tree topology by a non-maximum-likelihood method. This assumes that estimates of model parameters are not severely biased if a slightly incorrect tree topology is used. There are numerous algorithms to reconstruct phylogenetic trees from molecular sequence data (Swofford *et al.*, 1996). Probably the best non-maximum-likelihood method is neighbor-joining (Saitou and Nei, 1987). It is very fast and performs well even for large data sets (Strimmer and von Haeseler, 1996). Therefore we suggest the following iterative procedure P2 to estimate M without prior knowledge of a tree topology.

Procedure P2:

1. Choose start values for M, compute pairwise maximum-likelihood distances based on M, and reconstruct a neighbor-joining tree (initial step).
2. Find maximum-likelihood estimates $\hat{M}$ using this tree topology.
3. Based on new estimates $\hat{M}$ calculate an improved maximum-likelihood distance matrix and reconstruct a new tree topology.
4. Repeat steps 2 and 3 until $\hat{M}$ does not change any more.

In this way maximum-likelihood optimization has to be performed only for a comparatively small number of tree topologies. Further speed up is gained by using this procedure in conjunction with P1. This is also advantageous because the matrix of pairwise maximum-likelihood distances needs only to be computed once.

A third strategy is to break up the data set into a number of smaller parts to enable a maximum-likelihood estimate of the tree topology. An approximative maximum-likelihood estimate of M for the whole data set is obtained as average over the estimates for the subsets. The following procedure, quartet-sampling (QS), is based on selecting random sets of four sequences (quartets) from the data set:

Procedure QS:

1. Choose a random quartet and find maximum-likelihood estimates $\hat{M}_1$ and $\hat{T}$ (initial step). Note that only three topologies T_1, T_3, T_3 have to be evaluated (Figure 4.1).

2. Select the next random set $z = 2, 3, 4, \ldots$ of four sequences and find maximum-likelihood estimates $\hat{M}_z$ and $\hat{T}$.

3. Compute the average $\bar{M} = (\sum_{i=1}^{z} \hat{M}_i)/z$ of all quartet estimates $\hat{M}_i$.

4. Repeat steps 2 and 3 until $\bar{M}$ does not change any more.

To reach convergence of $\bar{M}$ usually only a small number of quartets have to be investigated. Quartet-sampling can also be combined with procedure P1. Note, however, that quartet-sampling cannot be used to infer parameters of rate heterogeneity as the overall tree structure is important in this case (Sullivan *et al.*, 1996).

3.3 Simulation study

To elucidate the performance of procedures P1 and P2 and of quartet-sampling we conducted a simulation study using a clock-like (H_1) and for a non-clock-like (H_2) tree with branch lengths $a = 0.02$ and $b = 0.19$ (Figure 3.1). We generated 100 artificial nucleotide data sets according to a TN model (Tamura and Nei, 1993) with $t = 15$, $\gamma = 3$, and uniform nucleotide frequencies $\pi_i = 1/4$. A fraction $\theta = 1/6$ of the 1200 sites was assumed to be invariable. For the remaining 1000 sites rate heterogeneity was modelled by a discrete Γ-distribution with four categories and $\alpha = 1$. Therefore the total rate heterogeneity (Gu *et al.*, 1995)

$$\rho = \frac{1 + \theta\alpha}{1 + \alpha} \tag{3.1}$$

was $\rho = 7/12 \approx 0.58$ for the data sets generated.

First we examined how reliable parameters of the substitution process are inferred if rate homogeneity is assumed in the estimation. We tested whether the results depend on an overall tree by comparing procedure P2 with quartet-sampling. Both methods were combined with procedure P1 to avoid optimization of branch lengths. Second we investigated methods to infer rate heterogeneity. Because it is well-known that the

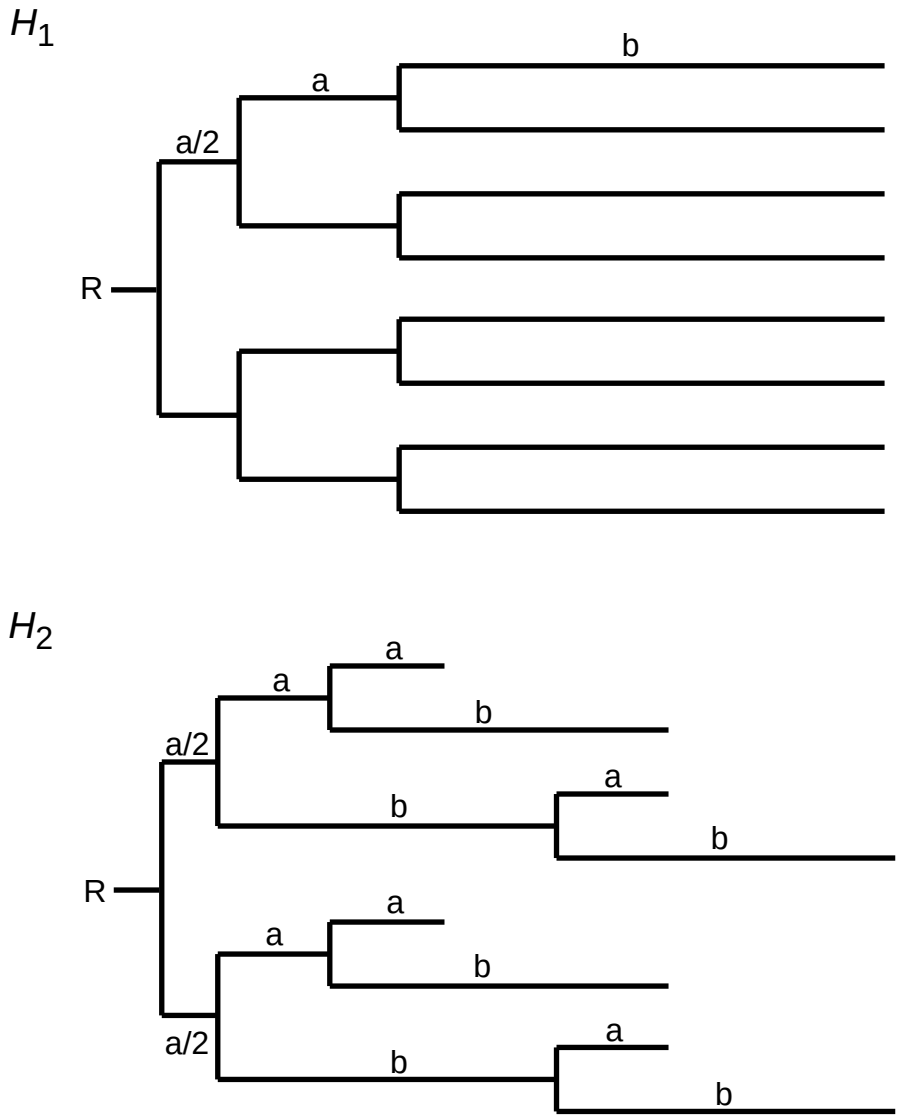

Figure 3.1: *Evolutionary histories H_1 and H_2 with branch lengths a and b. In H_1 a molecular clock is assumed, i.e. all sequences at the tips of the tree have the same distance $3/2a + b$ to the root R. H_2 is an example of non-clock-like evolution where most of the sequences differ in their distance to the root.*

overall tree structure is necessary for determining the amount of rate variation (Sullivan *et al.*, 1996) quartet-sampling was not tested. Instead, we applied the procedure P2 either with computation of exact maximum-likelihood branch lengths or in combination with procedure P1.

For each investigated parameter a 95% confidence interval was determined. In the likelihood framework the so-called observed information $I(x)$ with respect to a parameter x of the likelihood function L is defined by

$$I(x) = -\frac{\partial^2}{\partial x^2} \log L \,. \tag{3.2}$$

An estimate of the standard error σ of the maximum-likelihood estimate x_{ml} is

$$\hat{\sigma} = \frac{1}{\sqrt{I(x_{\text{ml}})}} \tag{3.3}$$

(Lindgren, 1976). An approximative 95% confidence interval for x_{ml} is obtained by

$$x_{\text{ml}} \pm 1.96\,\hat{\sigma} \,. \tag{3.4}$$

If quartet-sampling is used the estimate of the standard error of $\bar{M}$ is

$$\hat{\sigma} = \sqrt{\frac{1}{z(z-1)} \sum_{i=1}^{z} (\hat{M}_i - \bar{M}).} \tag{3.5}$$

Table 3.1 summarizes the estimation results for the simulated data sets. On the whole there was no gross difference in performance of the methods with respect to clock-like (H_1) and non-clock-like (H_2) evolution. When rate heterogeneity was neglected the estimates of t and γ were consistently smaller than the true values, as already found elsewhere (Wakeley, 1996). However, the true parameter values were included in the 95% confidence interval in the non-clock-like case due to slightly larger standard errors compared to clock-like evolution. Both quartet-sampling and procedure P2 resulted in similar estimates $\hat{t}$ and $\hat{\gamma}$. Therefore, the inference of substitution process parameters did not depend on the overall tree structure of the data set.

When rate heterogeneity was taken into account parameters of the substitution process were correctly inferred. It did not matter whether approximate (procedure P1) or exact branch lengths were computed. The total rate heterogeneity ρ was inferred with good accuracy as well though it was slightly overestimated when procedure P1 was used. Estimates for the rate heterogeneity parameters α and θ were much more

Table 3.1: *95% confidence intervals for parameter estimates from simulated data.*

Tree	Method	$\hat{t}$	$\hat{\gamma}$	$\hat{\theta}$	$\hat{\alpha}$	$\hat{\rho}$
H_1	I, P1, QS	9.46–14.20	2.09–2.79	—	—	—
	I, P1, P2	9.56–14.46	2.07–2.77	—	—	—
	II, P1, P2	12.13–18.87	2.38–3.72	0.00–0.04	0.46–0.66	0.56–0.72
	II, P2	11.61–18.11	2.29–3.51	0.18–0.30	0.89–1.75	0.43–0.71
H_2	I, P1, QS	9.28–16.68	2.24–3.18	—	—	—
	I, P1, P2	9.20–16.84	2.29–3.15	—	—	—
	II, P1, P2	12.15–18.03	2.42–3.64	0.00–0.03	0.64–0.66	0.58–0.70
	II, P2	11.45–18.85	2.38–3.68	0.20–0.32	0.89–2.11	0.40–0.72
True values		15.00	3.00	0.17	1.00	0.58

(Abbreviations) I: TN model assuming rate homogeneity, II: TN model with mixed model of rate heterogeneity (discrete Γ-distribution with four categories and invariable sites), P1: approximative maximum-likelihood branch lengths, P2: neighbor-joining procedure, QS:quartet-sampling. The boundaries of the confidence intervals are averaged over 100 data sets.

difficult to obtain. For instance, both for clock-like and for non-clock-like evolution it was not possible at all to infer the fraction of invariable sites when approximate maximum-likelihood branch lengths were used. However, the presence of invariable sites lead to a correspondingly smaller value of $\hat{\alpha}$ so that ρ was still inferred correctly. Unfortunately, an simultaneous estimate of θ and α could only be obtained when the time-saving procedure P1 was not applied.

3.4 Rate heterogeneity of an amino acid alignment

To compare the mixed model of rate heterogeneity used in the simulations to a simple Γ-model or a two-rate model we examined an amino acid alignment of 19 complete mitochondrial sequences. The data set comprised 18 mammals and the frog. More precisely, the species involved were *Xenopus laevis* (frog), *Ornithorhynchus anatinus* (platypus), *Didelphis virginiana* (opossum), *Mus musculus* (mouse), *Rattus norvegicus* (rat), *Hylobates lar* (gibbon), *Pongo pygmaeus* (orang), *Gorilla gorilla* (gorilla), *Pan troglodytes* (chimpanzee), *Pan paniscus* (bonobo), *Homo sapiens* (human), *Bos taurus* (cow), *Balaenoptera musculus* (blue whale), *Balaenoptera physalis* (finback

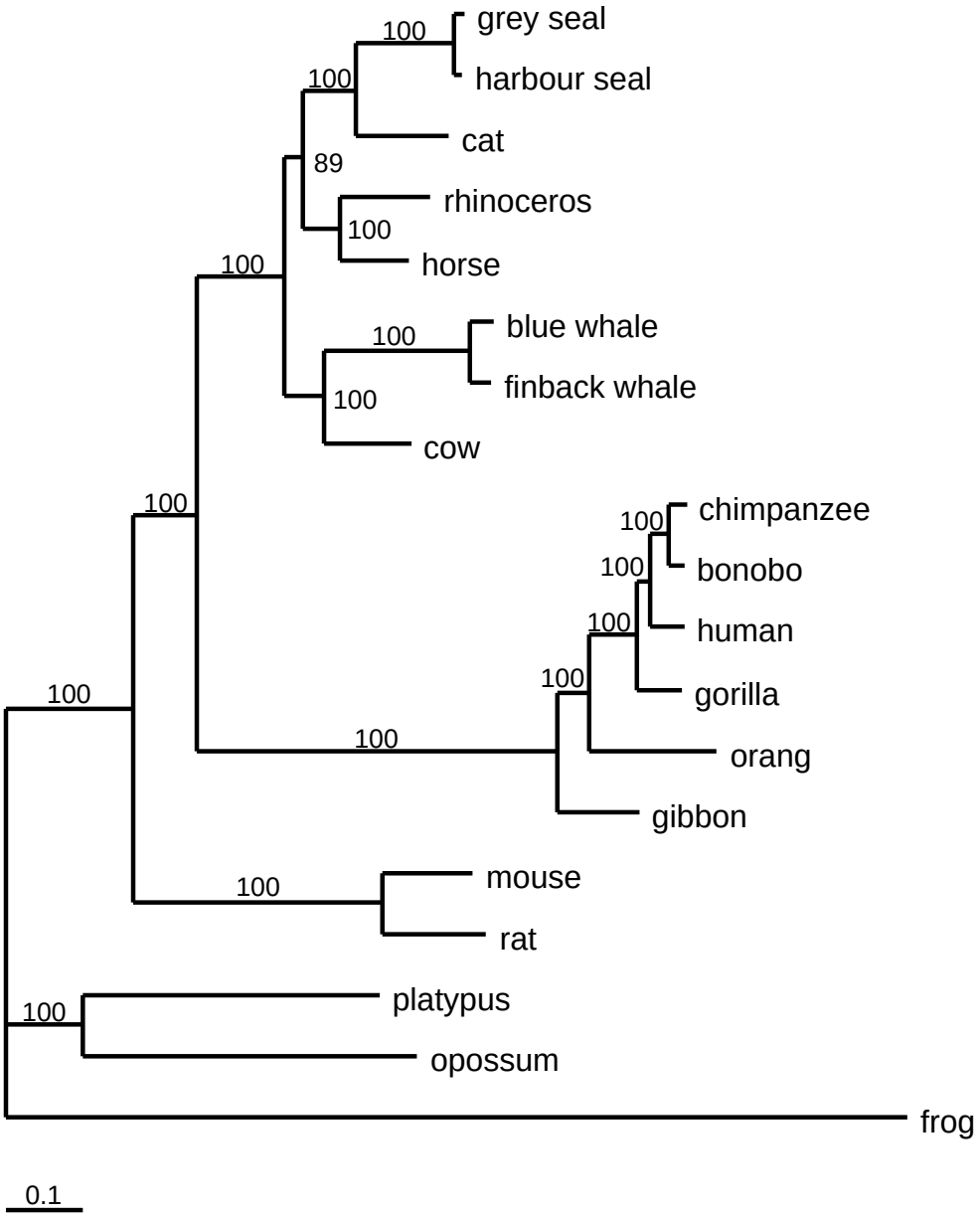

Figure 3.2: *Phylogenetic relationship of the 19 complete mtDNA sequences contained in the amino acid alignment. The tree was reconstructed by quartet-puzzling (Chapter 4) using rate heterogeneity parameters* $\theta = 0.27$ *and* $\alpha = 0.64$*. Maximum-likelihood branch lengths are proportional to substitutions of amino acids per site.*

Table 3.2: *95% confidence intervals for rate variation parameters of mtDNA sequences.*

Method	$\hat{\theta}$	$\hat{\alpha}$	$\hat{\rho}$	$\log L$
I, P1, P2	0.34–0.36	—	0.34–0.36	-46756.86
II, P1, P2	—	0.27–0.31	0.77–0.79	-45504.67
III, P1, P2	0.00–0.00	0.27–0.31	0.77–0.79	-45499.10
I, P2	0.38–0.42	—	0.38–0.42	-46747.68
II, P2	—	0.31–0.35	0.74–0.76	-45496.39
III, P2	0.25–0.29	0.58–0.70	0.68–0.76	-45468.69

(Abbreviations) I: mtREV model with two-rate model of rate heterogeneity, II: mtREV model assuming discrete Γ-distribution with four categories, III: mtREV model with mixed model of rate heterogeneity (discrete Γ-distribution with four categories and invariable sites), P1: approximative maximum-likelihood branch lengths, P2: neighbor-joining procedure.

whale), *Felis catus* (cat), *Halichoerus grypus* (grey seal), *Phoca vitulina* (harbor seal), *Rhinocerus unicornis* (rhinoceros), and *Equus caballus* (horse). The phylogenetic relationship as inferred by quartet-puzzling (Chapter 4) is shown in Figure 3.2. The frog was used as outgroup to root the tree of mammals. The 13 protein coding genes encoded on the mtDNA sequences (ATP6, ATP8, CO1, CO2, CO3, CYT b, ND1, ND2, ND3, ND4, ND4L, ND5, ND6) were separately aligned with CLUSTAL W (Thompson *et al.*, 1994) and subsequently concatenated. After removing positions with gaps an amino acid alignment of 3,735 sites containing 19 sequences remained. A χ^2-test showed that the composition of amino acids was the same for all sequences. For the substitution process the mtREV model was selected as is was specifically developed to describe the evolution of amino acids encoded on mtDNA (Adachi and Hasegawa, 1996b). For rate heterogeneity three models were considered, the two-rate model taking into account invariable sites (I), a discrete Γ-model with four categories (II), and a mixed model (III) combining models I and II. The parameters corresponding to each model of rate heterogeneity were estimated using the neighbor-joining procedure P2, either with computation of exact maximum-likelihood branch lengths or applying procedure P1. The results of the various methods are shown in Table 3.2.

The best fit to the data according the maximum-likelihood value was obtained when $\hat{\theta}$ and $\hat{\alpha}$ were simultaneously determined without applying P1 (last line in Table 3.2). It's noteworthy that the fit did not improve when more than four categories were al-

lowed in the discrete Γ-distribution. The confidence interval for the total rate heterogeneity ρ was similar in all estimations for models II and III, regardless whether exact or approximate maximum-likelihood branch lengths were computed. In contrast, the two-rate model produced much smaller values for $\hat{\rho}$. However, the corresponding likelihoods (I) indicated that this model is inferior to the others. It was also not possible to obtain a simultaneous estimate of θ and α by applying procedure P1. This was already observed earlier in the simulation study. Very interestingly, the likelihood values for models II and III were all very similar despite their drastically different estimates $\hat{\alpha}$. This reveals a general property of the models of rate heterogeneity. When invariable sites are admitted (III) the rate heterogeneity induced by the Γ-distribution is smaller and the parameter α is correspondingly larger. On the other hand, if invariable sites are not considered (II) the Γ-distribution takes care for them by a smaller α.

3.5 Conclusion

We have presented simple strategies to speed up maximum-likelihood estimation of the parameters of models of substitution and rate heterogeneity. In order to evaluate their performance we have conducted a computer simulation and examined an alignment of 19 amino acid sequences from mitochondrial DNA If only t and γ need to be estimated the quartet-sampling procedure in combination with computation of approximate maximum-likelihood branch lengths (procedure P1) is sufficient. However, to infer parameters of rate heterogeneity an overall tree topology, most conveniently determined by a non-maximum-likelihood method, is necessary. We suggest to use neighbor-joining as applied in procedure P2. As model of rate heterogeneity we recommend the Γ-model. It produces results comparable to the more elaborate mixed model where the fraction of invariable sites is determined as well. In contrast to the mixed model it can also be used in conjunction with procedure P1. The simple two-rate model is not recommended as its fit to the data is insufficient and as it underestimates the total amount of rate heterogeneity.

Chapter 4

Quartet Puzzling

In this chapter we introduce a novel and versatile method, the so-called quartet-puzzling, to reconstruct the topology of a phylogenetic tree based on DNA or amino acid sequence data. This method applies maximum-likelihood tree reconstruction to all possible quartets that can be formed from N sequences. The quartet trees serve as starting points to reconstruct a set of optimal N-sequence trees. The majority-rule consensus of these trees defines the quartet-puzzling tree and shows groupings that are well supported. Computer simulations show that the performance of quartet-puzzling to reconstruct the true tree is always equal to or better than that of neighbor-joining. For some cases with high bias in the expected Ts/Tv ratio quartet-puzzling outperforms neighbor-joining by a factor of more than 10. The application of quartet-puzzling to mitochondrial RNA and $tRNA^{Val}$ sequences from amniotes demonstrates the power of the approach. Using quartet-puzzling we confirm the African origin of human mitochondrial D-loop sequences.

4.1 Introduction

In recent years the maximum-likelihood method for reconstructing phylogenetic relationships (Felsenstein, 1981) has become more popular due to the arrival of powerful computers. The main advantage of a maximum-likelihood approach is the application of a well defined model of sequence evolution to a given data set (Felsenstein, 1988). Although the application of the maximum-likelihood method to biological data is now widespread, its computational complexity prevents computation for a large number of sequences. Generally, only slow programs for analyzing nucleotide

or amino acid sequences are available (Felsenstein, 1993; Yang, 1997) although it is possible to speed up calculations by parallelizing the algorithm or using approximative techniques (Olsen *et al.*, 1994; Adachi and Hasegawa, 1996a). Still, large trees can only be analyzed on massively parallel systems or by constraining the tree topology.

The principal goal of a maximum-likelihood analysis is the determination of a tree and corresponding branch lengths that have the greatest likelihood of generating the data. This task can be split into two parts: determining a tree topology and subsequently assigning branch lengths to the topology to obtain a maximum-likelihood estimate. Because the number of possible tree topologies grows exponentially with the number of sequences, all tree reconstruction methods that optimize an objective function have to rely on heuristic searches to find the best topology. Moreover, the optimization of branch lengths for a given topology is a tedious procedure for maximum-likelihood-based tree reconstruction methods and consumes a lot of computing time (Olsen *et al.*, 1994). While maximum-likelihood procedures are generally slow for the general case of N sequences, the determination of the maximum-likelihood tree based on DNA or amino acid sequences poses no problem for four sequences. On the other hand, methods abound that try to reconstruct a tree topology considering only the branching pattern of the $\binom{N}{4}$ different quartet trees that can be constructed from N sequences (Sattath and Tversky, 1977; Fitch, 1981; Dress *et al.*, 1986; Bandelt and Dress, 1986). It has been shown (Schöniger and von Haeseler, 1993) that these distance-based methods exhibit performance similar to neighbor-joining (Saitou and Nei, 1987) while generally being much slower.

In this chapter we describe a new method, quartet-puzzling, for reconstructing phylogenetic relationships. This method reconstructs the maximum-likelihood tree for each of the $\binom{N}{4}$ possible quartets. In a so-called puzzling step the resulting quartet trees are then combined to an overall tree. During the puzzling step sequences are added sequentially in random order to an already existing subtree. The position of a new sequence is determined by a voting procedure, considering all quartets. Finally, an intermediate tree relating N sequences is obtained. In general, there is no N-sequence tree that fits all the $\binom{N}{4}$ different quartet trees. Therefore the puzzling step is repeated several times thereby elucidating the landscape of possible optimal trees. The quartet-puzzling tree is obtained as a majority-rule consensus (Margush and McMorris, 1981) of all trees that result from multiple runs of the puzzling step. Depending on the phylogenetic information contained in the data, this tree may be binary or multifurcating.

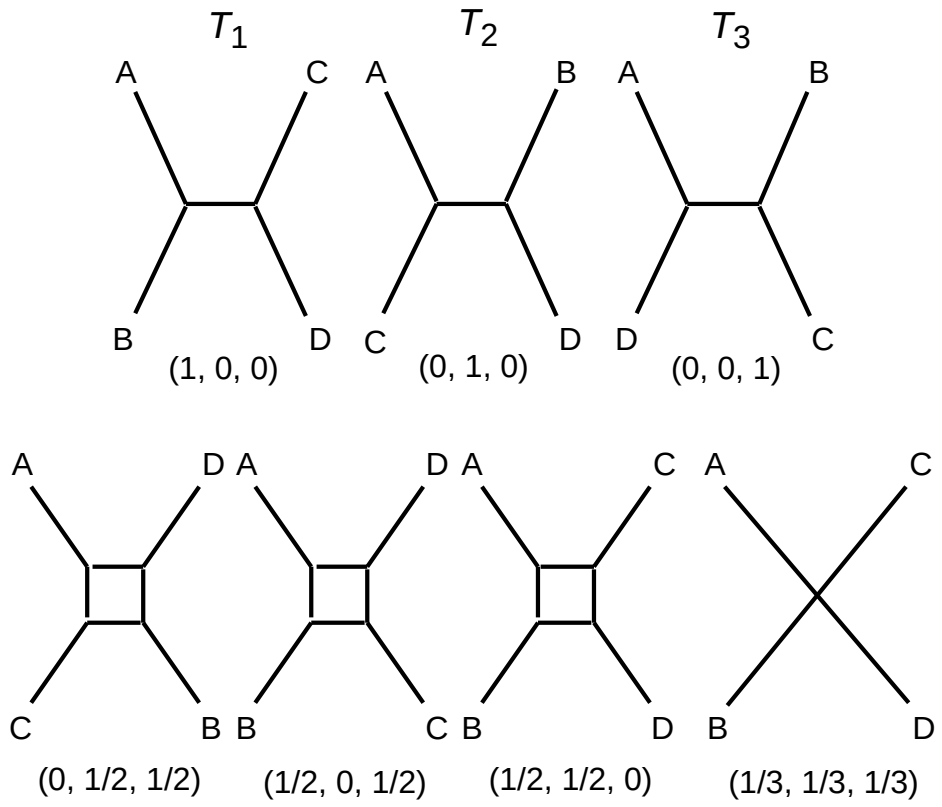

Figure 4.1: *The three unrooted trees T_i and the four graphs for a quartet of sequences (A, B, C, D), and the corresponding discrete weights* (w_1, w_2, w_3).

In addition to the tree topology the quartet-puzzling tree also shows reliability values for each internal branch.

4.2 The quartet-puzzling algorithm

Quartet-puzzling essentially is a three-step procedure, first reconstructing all possible quartet maximum-likelihood trees (maximum-likelihood step), then repeatedly combining the quartet trees to an overall tree (puzzling step), and finally computing the majority-rule consensus of all intermediate trees giving the quartet-puzzling tree (consensus step).

The first step in the quartet-puzzling analysis is the reconstruction of the branching pattern of all possible $\binom{N}{4}$ quartets with maximum-likelihood. For each quartet (A, B, C, D) three different fully-bifurcating tree topologies T_1, T_2, T_3 (Figure 4.1) exist with

corresponding maximum-likelihood values L_1, L_2, L_3. Note that $L_1 + L_2 + L_3 \ll 1$. Evaluation via Bayes' theorem of the three tree topologies given uniform prior information leads to posterior probabilities

$$p_i = \frac{L_i}{L_1 + L_2 + L_3} \tag{4.1}$$

for each quartet T_i (Lindgren, 1976; Kishino and Hasegawa, 1989), with $p_1 + p_2 + p_3 = 1$. Usually, one of the maximum-likelihood values L_i is much larger than the other two. Then $p_{\text{max}} \approx 1, p_{\text{other}} \approx 0$ and the optimal topology T_{max} is stored for the puzzling step.

However, if sequences are short or very closely related and if therefore not all quartet trees can be confidently resolved, the Bayesian posteriors p_i may deviate substantially from this simple picture (Strimmer *et al.*, 1997). In this case the probabilities p_i help us to correctly choose *more than one* T_i from the quartet trees. It is well-known that for four sequences there are not only three different unrooted binary trees but also three partially resolved quartet trees as well as one completely unresolved tree (Eigen *et al.*, 1988). To each corresponds a set of discrete weights w_i (Figure 4.1), according to which bifurcating trees may be obtained by resolving the partially resolved networks. When quartets are examined in the maximum-likelihood step of the quartet-puzzling algorithm we choose among these seven permitted sets of weights by selecting that which minimizes the least-squares distance

$$d = \sum_{i=1}^{3} (p_i - w_i)^2. \tag{4.2}$$

Thus, we approximate the Bayesian probabilities p_i by one of the seven sets of discrete weights w_i. In the end, all topologies T_i that correspond to a non-zero discrete weight are stored for the puzzling step. Note that the special case of one large maximum-likelihood value with $p_{\text{max}} \approx 1, p_{\text{other}} \approx 0$ is also correctly treated with this procedure.

If there is more than one best topology stored, the branching pattern of the quartet (A, B, C, D) is not uniquely defined. In this case we randomly choose among the available topologies each time the branching pattern of (A, B, C, D) is needed. Thus, maximum-likelihood tree reconstruction induces a neighbor relation $||_{\text{ml}}$ between any four sequences A, B, C, D (Bandelt and Dress, 1986). The neighbor relation AB $||_{\text{ml}}$ CD implies that sequences A and B and sequences C and D are neighbors with respect to each other. Note that in the corresponding tree T_1 (Figure 4.1) the paths connecting the sequences A and B and the sequences C and D are disjoint.

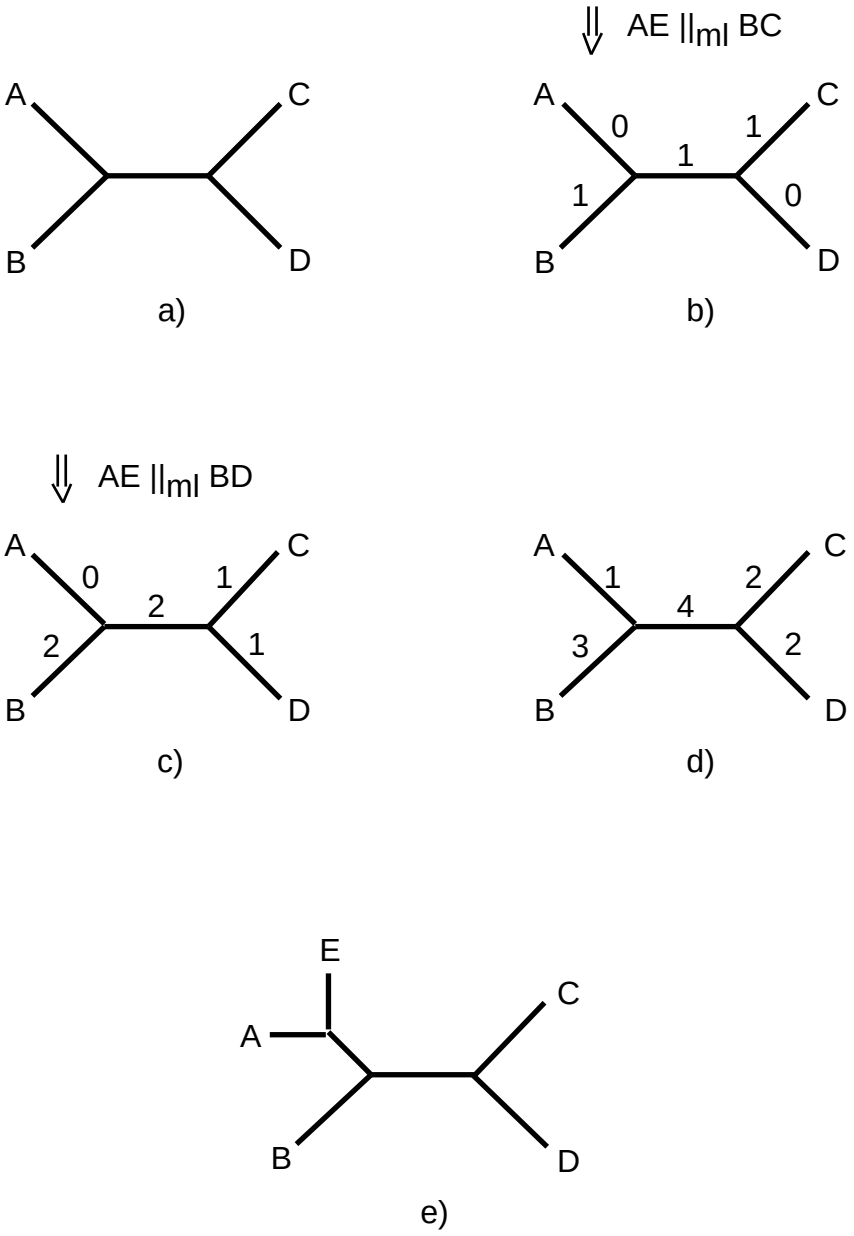

Figure 4.2: *Addition of sequence E to the already existing four-sequence tree (a). The neighbor relations are given by AE $||_{ml}$ BC, AE $||_{ml}$ BD, AC $||_{ml}$ DE, and BD $||_{ml}$ CE. The relation AE $||_{ml}$ BC implies that the branches connecting B and C each get a penalty of 1 (b). (c) shows the penalty of the branches if AE $||_{ml}$ BD is evaluated. If all four quartets are analyzed, the branch leading to sequence A shows the lowest penalty (d). Hence, E is inserted at this branch (e).*

Next, in the puzzling step, we aim to combine the quartet trees to an overall N-sequence tree. Generally the neighbor relation $||_{\mathrm{ml}}$ on the set of all N sequences is not tree-like (Bandelt and Dress, 1986), therefore it is necessary to apply approximation methods to obtain an overall tree topology (Sattath and Tversky, 1977; Fitch, 1981; Dress *et al.*, 1986; Bandelt and Dress, 1986). We suggest the following simple algorithm. First, the input order of the N sequences is randomized; let us assume that the order is A, B, C, D, E, ... The branching pattern of the quartet (A, B, C, D) is now used as a seed for the overall N-sequence tree. Then sequence E is added to the subtree according to the following voting procedure: The neighbor relation $||_{\mathrm{ml}}$ induces for every quartet ($i, j, k,$ E) a clustering i, j versus $k,$ E, for example. It is obvious that sequence E should not be placed on a branch that lies on the path connecting i and j in the subtree. The edges where E should not be placed in the subtree are marked for every quartet ($i, j, k,$ E). Thus, every branch in the subtree is assigned a penalty. If all different quartets containing sequence E and three sequences of the subtree are evaluated, sequence E is inserted at that branch in the tree that shows the lowest penalty. If the minimal penalty is attained for more than one edge, the sequence is inserted randomly at one of the equally good branches. Figure 4.2 illustrates the procedure for five sequences. The addition of a single sequence is repeated until an overall tree of N sequences is obtained. The randomized sequential insertion of sequences may not always lead to the same tree topology for different runs of the puzzling step. Therefore, step two is repeated as often as possible, thereby elucidating the landscape of all possible optimal trees. Generally, the more sequences involved the more runs of the puzzling step are advised.

In the third step of the quartet-puzzling algorithm a majority-rule consensus (Margush and McMorris, 1981) is computed from the intermediate trees resulting from the puzzling steps. We call this consensus tree the quartet-puzzling tree. Depending on the phylogenetic information contained in the data the quartet-puzzling tree is either completely resolved or shows multifurcations. In addition to the tree topology the quartet-puzzling tree also provides information about the number of times a particular grouping occurred in the intermediate trees. If the resolution of phylogenetic relationships between a subset of sequences is unclear, the consensus tree will indicate it by displaying small reliability values for the corresponding internal branches. The repeated randomization of the input order of the sequences and subsequent computation of an intermediate tree results in a collection of locally optimal trees that are generated

independently of each other. In contrast, a collection of trees derived by procedures like branch-swapping from one starting tree produces non-independent trees (Penny *et al.*, 1995). Thus, given the independence, the consensus tree gives a summary of all groups that occur in the majority of the intermediate trees.

The reliability values, i.e. the number of times the group is reconstructed during the puzzling steps, allow a simple interpretation of the phylogenetic information present in the data. Every intermediate tree represents a solution from the set of optimal trees. If we were able to compute all optimal trees, then all clusters that appear in more than 50% of the optimal trees fit into an overall tree (Margush and McMorris, 1981). This not necessarily bifurcating tree represents the total phylogenetic information. However, due to the limited number of puzzling steps, only a subset of all optimal trees are found. Therefore it is advisable to trust only reliability values that are well above 50%. Note that the suggested reliability measure should not be confused with the usual bootstrap values. Whereas reliability values are an intrinsic result of the quartet-puzzling algorithm, bootstrapping is an external procedure that can be applied to any tree building method. Quite remarkably however, it seems that both measures are highly correlated.

Quartet-puzzling therefore is a simple method to get a phylogenetic tree and simultaneously an impression how well the data are suited for a phylogenetic reconstruction.

4.3 Efficiency of quartet-puzzling

It is easy to prove that quartet-puzzling reconstructs the underlying tree if the neighbor relation $||_{\mathrm{ml}}$ is tree-like (Bandelt and Dress, 1986). However, real data hardly ever are tree-like. To study the efficiency of our approach we employed a computer simulation. We compared the efficiency of quartet-puzzling with the performance of neighbor-joining and maximum-likelihood. The simulation settings are analogous to that employed in (Schöniger and von Haeseler, 1993). Maximum-likelihood was used as implemented in the PHYLIP (Felsenstein, 1993) program DNAML version 3.5, quartet-puzzling as implemented in version 2.5 of the PUZZLE program. The results for the performance of neighbor-joining are adopted from Schöniger and von Haeseler (1993). The two investigated evolutionary histories H_1 and H_2 are displayed in Figure 3.1. For each of the two evolutionary scenarios a variety of branch lengths a and b were assumed. Sequences were evolved according to the JC model (Jukes and Cantor, 1969) and the Km model (Kimura, 1980). The expected Ts/Tv ratio t

Table 4.1: *Percentage of correctly reconstructed trees assuming clock-like evolution according to tree* H_1.

Sequence evolution:		JC ($t = 1/2$)			Km ($t = 4$)		
l	a/b	NJ	QP	ML	NJ	QP	ML
500	0.01/0.07	70.5	79.6	87	56.6	69.8	70
	0.02/0.19	52.0	69.7	63	23.1	63.3	48
	0.03/0.42	8.2	28.5	9	1.4	33.4	15
1,000	0.01/0.07	94.7	93.5	96	87.3	89.2	93
	0.02/0.19	86.8	91.5	85	59.3	85.0	85
	0.03/0.42	38.3	52.9	34	10.8	57.2	38

(Abbreviations) NJ: neighbor-joining, QP: quartet-puzzling, ML: maximum-likelihood, l: sequence length, a and b: branch lengths, t: expected Ts/Tv ratio. Estimates of efficiencies are based on 1,000 simulations (NJ, QP) and 100 simulations (ML).

equals $t = 1/2$ in the Jukes-Cantor and $t = 4$ in the Kimura case. Simulations were carried out with sequences of lengths 500 and 1,000. For each setting 1,000 simulations were done. For DNAML, however, only 100 simulations were possible due to the large computational costs involved. All programs were run with their defaults except for the transition-transversion ratio parameter that was set both in DNAML and PUZZLE equal to $1/2$ or 4, according to the mode of assumed sequence evolution. Quartet-puzzling was performed with 1,000 puzzling steps and using only approximate maximum-likelihood branch lengths for the quartet trees. Neighbor-joining results are displayed using Kimura corrected distances (Kimura, 1980).

The results are shown in Tables 4.1 and 4.2. It is obvious that maximum-likelihood generally outperforms neighbor-joining and that its efficiency is similar to or better than quartet-puzzling. Unfortunately, the computational costs of DNAML are prohibitively high when the number of sequences is large. The performance of neighbor-joining and quartet-puzzling is different depending on the choice of parameters. As expected, an increase in sequence length leads to a better performance of each method. If sequences evolved according to a Jukes-Cantor model, both methods show a more or less identical efficiency. Quartet-puzzling is slightly superior if the clock assumption is violated and if branch lengths are large. If sequences evolved under a Kimura model evolution with an expected Ts/Tv ratio of $t = 4$, the quartet-puzzling method

Table 4.2: *Percentage of correctly reconstructed trees for non-clock-like evolution assuming tree H_2.*

Sequence evolution:		JC ($t = 1/2$)			Km ($t = 4$)		
l	a/b	NJ	QP	ML	NJ	QP	ML
500	0.01/0.07	79.7	86.0	91	71.7	80.5	94
	0.02/0.19	64.8	84.9	93	38.6	77.4	92
	0.03/0.42	18.1	47.2	72	3.5	52.1	73
1,000	0.01/0.07	96.1	97.3	99	91.8	95.4	98
	0.02/0.19	91.3	96.2	99	67.6	91.5	99
	0.03/0.42	37.9	69.6	92	7.7	73.6	96

Abbreviations are the same as in Table 4.1.

outperforms neighbor-joining, irrespective whether the tree follows a molecular clock (H_1) or not (H_2). For a high rate of substitutions the efficiency of quartet-puzzling is more than 10 times better than that of neighbor-joining.

4.4 Phylogeny of amniotes

We have reanalyzed the concatenated sequences of amniote mitochondrial 12S rRNA, 16S rRNA, and $tRNA^{Val}$ genes (Hedges, 1994) with the quartet-puzzling method. The data set comprises 15 species, among them six placental mammals, four reptiles, one bird, one frog and three lungfish sequences. More specifically, the species involved are *Neoceratodus forsteri* (lungfish, Australia), *Lepidosiren paradoxa* (lungfish, South America), *Protopterus sp.* (lungfish, Africa), *Xenopus laevis* (frog), *Trachemys scripta* (turtle), *Sphenodon punctatus* (sphenodontid), *Sceloporus undulatus* (lizard), *Alligator mississippiensis* (crocodilian), *Gallus gallus* (bird), *Homo sapiens* (human), *Phoca vitulina* (seal), *Bos taurus* (cow), *Balaenoptera physalis* (whale), *Mus musculus* (mouse), and *Rattus norvegicus* (rat). In addition the corresponding sequences from *Didelphis virginiana* (opossum) (Janke *et al.*, 1994) and *Ornithorhynchus anatinus* (platypus) (Janke *et al.*, 1996) were included in the analysis. The 17 sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994) resulting in an alignment of length 2,903. After removing ambiguous alignment positions as well as gaps 1,998 sites remained for further analysis. In the PUZZLE program the HKY model of sequence evolu-

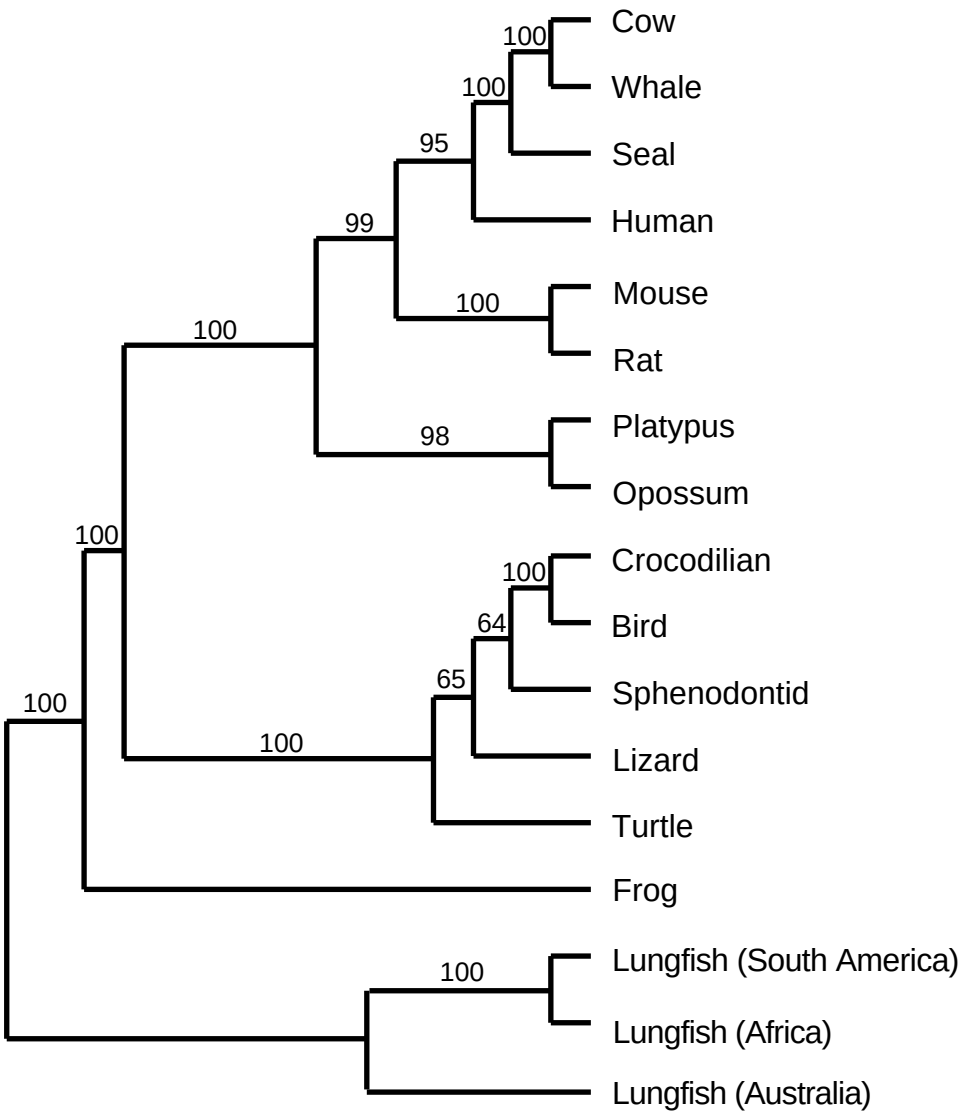

Figure 4.3: *Quartet-puzzling tree topology based on 1,000 puzzling steps. The reliability value of each internal branch indicates in percent how often the corresponding cluster was found among the 1,000 intermediate trees. The lungfishes are used as outgroup to root the tree of the amniotes.*

tion (Hasegawa *et al.*, 1985) was selected and uniform rates over sites were assumed. The transition-transversion ratio parameter $t = 1.28$ was estimated from the data by maximum-likelihood.

A total of 2,380 four-sequence maximum-likelihood trees were reconstructed in the first step of the quartet-puzzling algorithm. Among all quartet trees there was only one quartet (Sphenodontid-Lizard-Bird-Human) with a completely unresolved branching structure. In this case the corresponding posterior probabilities take on values $p_i \approx 1/3$. A high percentage of quartets of this kind usually indicates that the data set is not very well suited for a phylogenetic analysis. If more than 10–15% completely unresolved quartets are present the quartet-puzzling tree is in general not completely resolved. For the amniote data the number of completely unresolved quartets is negligible and hence the sequences are suitable for a phylogenetic analysis. The notion of unresolved quartets is also important for likelihood-mapping (Chapter 5).

The resulting quartet-puzzling tree, after performing 1,000 puzzling steps, is shown in Figure 4.3. The tree topology coincides more or less with the already published tree (Hedges, 1994). Our analysis supports the view that crocodilians are the closest living relatives of birds. In 100% of trees underlying the quartet-puzzling tree the bird-crocodilian clade is found, indicating the clear separation from the remaining sequences in the tree. This high support from our analysis is matched by a high bootstrap support (Hedges, 1994). Incidentally, the alternative clade placental mammals-bird was never detected in any of the 1,000 intermediate trees. The positions of sphenodontid, lizard and turtle are less clear from the analysis. Though bird, crocodilian, and sphenodontid seem to form a monophyletic group the corresponding reliability value of 64% is quite low. Similarly, the phylogenetic relationship among this group, lizard, and turtle cannot be resolved because of a relatively low reliability of 65% for the corresponding internal branch. Contrary to the tree published in (Hedges, 1994), our branching pattern suggests that within the radiation of placental mammals the rodents branch off first and the humans are a sister group of the Ferungulata, a result in perfect agreement with other studies (Janke *et al.*, 1994). Our results also support the sister group relationship of marsupials and monotremes (Janke *et al.*, 1996). Thus, quartet-puzzling analysis confirms the close relationship of birds and crocodilian and proposes a branching pattern of placental mammals which coincides with other analyses (Janke *et al.*, 1994; Janke *et al.*, 1996).

4.5 African origin of human mtDNA

Quartet-puzzling allows the analysis of data sets with a large number of sequences. To investigate the phylogenetic history of human mitochondrial D-loop sequences we reconstructed a quartet-puzzling tree from 45 human sequences and one chimpanzee sequence (Figure 4.4). The latter was used to root the tree of the human sequences. The data set consisted of 15 American, 15 Asian, and 15 European individuals for which both the hypervariable region I and II are known. For inclusion in the alignment of length 691 bp, kindly provided by Oliva Handt, sequences that contained as few unknown positions as possible were selected from the data sets of Anderson *et al.* (1981), Vigilant *et al.* (1991), and Mountain *et al.* (1995). HKY assuming rate homogeneity over sites was chosen as model of sequence evolution, the transition-transversion ratio parameter $t = 18.5$ was estimated from the data by quartet-sampling. After computing 100,000 puzzling steps a highly multifurcating but nevertheless very structured tree was obtained (Figure 4.4).

The 30 European and Asian sequences of the data set form a group. The quartet-puzzling reliability value of the corresponding internal branch is 83%. With a reliability of 74% the European, Asian, and central African lineages form a cluster, separated from the sequences of southern Africa. It is interesting to observe this general structure in the quartet-puzzling tree, especially as the phylogenetic information contained in the hypervariable regions I and II of human mtDNA sequences is relatively low (Takahata, 1995). The most basal groups in the tree are sequences from Africa whereas the non-African sequences originate later and show a common ancestor with central African lineages. This configuration confirms the hypothesis of an African origin of human D-loop sequences (Vigilant *et al.*, 1991; Penny *et al.*, 1995).

4.6 Discussion

We have presented the quartet-puzzling method to reconstruct tree topologies from sequence data. This method computes the maximum-likelihood tree for all possible quartets. An intermediate N-sequence tree is computed in the so-called puzzling step. The repeated application of the puzzling step allows an assignment of reliability values to the groupings in the final quartet-puzzling tree, a consensus tree built from all intermediate trees. If groups are found only occasionally in different runs of the puzzling

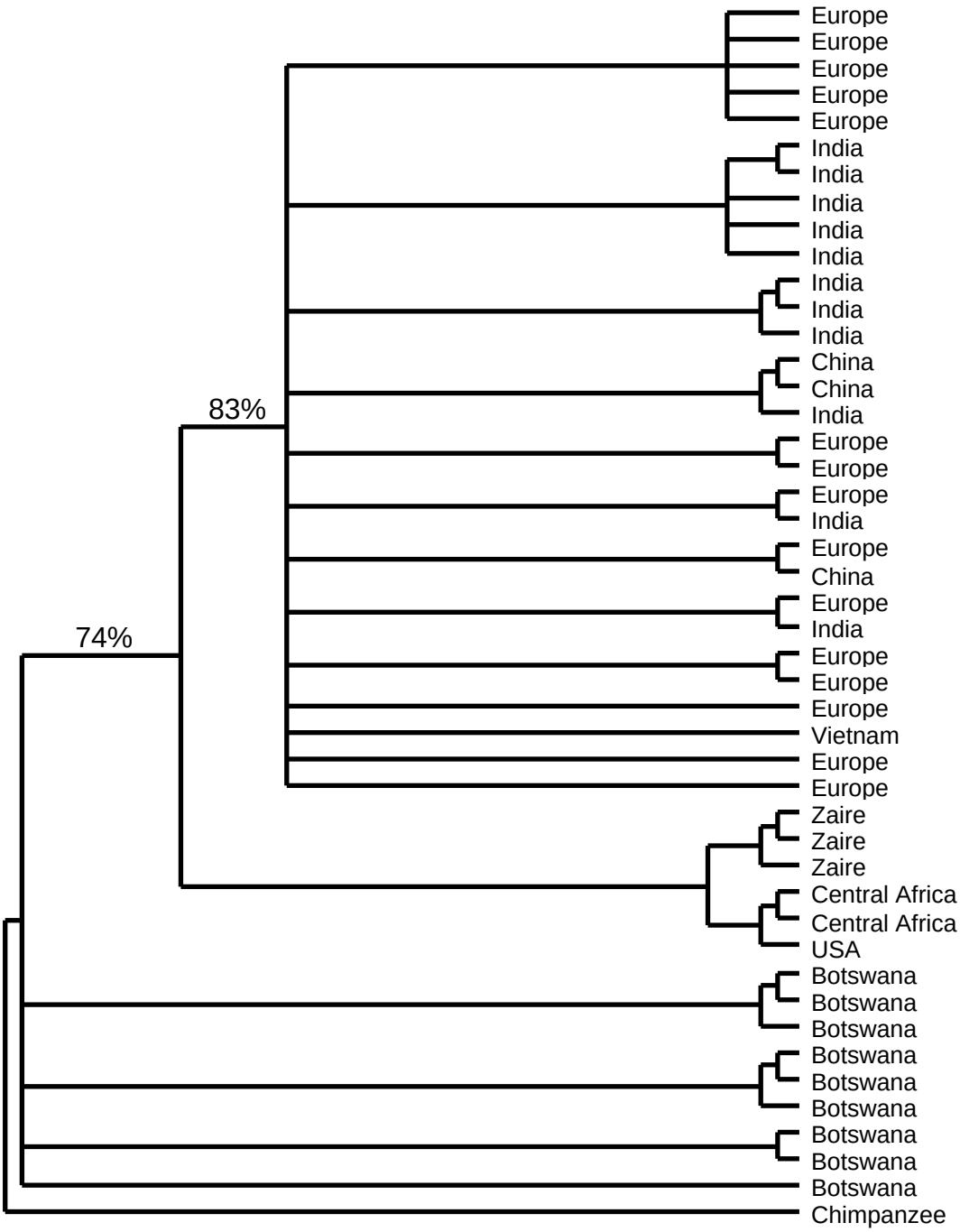

Figure 4.4: *Quartet-puzzling tree from human mtDNA D-loop sequences rooted by a chimpanzee. Sequences are labeled by their geographic origin.*

step, they will obtain a low reliability value. In those situation it is more realistic to assume a multifurcation rather than a bifurcation.

Moreover, we have shown that quartet-puzzling either shows performance comparable to or better than neighbor-joining. If sequences evolved according to the Jukes-Cantor model and obeyed a molecular clock both methods have a similar efficiency. In situations where neighbor-joining performs badly, quartet-puzzling has the advantage of not falling into the traps provided by the complex landscape of the tree space. The repeated application of the puzzling step prevents the method from getting trapped in local optima. This “trap avoiding” property stems from the various averaging procedures that are present in quartet-puzzling. Finally, our analysis of amniote and human mtDNA sequence data shows that quartet-puzzling also performs very well for real data sets.

Chapter 5

Likelihood Mapping

We introduce a graphical method, likelihood-mapping, to visualize the phylogenetic content of a set of aligned sequences. The method is based on an analysis of the maximum-likelihood values for the three different completely resolved tree topologies that exist for four sequences. The three maximum-likelihood values are represented as one point inside an equilateral triangle. The triangle is partitioned in different regions. One region represents star-like evolution, three regions represent a well-resolved phylogeny, and three regions reflect the situation where it is difficult to distinguish between two of the three trees. The location of the likelihoods in the triangle defines the mode of sequence evolution. If N sequences are analyzed then the likelihoods for each subset of four sequences are mapped onto the triangle. The resulting distribution of points shows whether the data are suitable for a phylogenetic reconstruction.

5.1 Introduction

The sequence-based study of phylogenetic relationships among different organisms has become routine. Parallel to the increasing amount of sequence information available a variety of methods have been suggested to reconstruct a phylogenetic tree (Swofford *et al.*, 1996) or a phylogenetic network (Bandelt and Dress, 1992; Dopazo *et al.*, 1993; von Haeseler and Churchill, 1993). So far, few approaches have been proposed to elucidate the phylogenetic content in a set of aligned sequence *a priori* (Eigen *et al.*, 1988; Eigen and Winkler-Oswatitsch, 1990). The so-called statistical geometry in sequence space analyses the distribution of numerical invariants for all possible subsets of four sequences. The resulting distributions make it possible to distinguish be-

tween tree-, star-, and net-like geometry of the data. Moreover, based on the averages of the invariants, the method allows to draw a graph illustrating the mode of evolution. While the description of this diagram is straightforward if sequences consist only of purines and pyrimidines, it gets difficult if more complex alphabets (nucleotides, amino acids) are used (Nieselt-Struwe *et al.*, 1996). Statistical geometry in sequence space has been successfully applied to study the evolution of tRNAs (Eigen *et al.*, 1989) or HIV (Eigen and Nieselt-Struwe, 1990).

Here we present an alternative approach, likelihood-mapping, to display phylogenetic information contained in a sequence alignment. The method is applicable to nucleotides sequences, amino acid sequences, or any other alphabet for which a model of sequence evolution (Swofford *et al.*, 1996; Zarkikh, 1994; Schöniger and von Haeseler, 1994) exists. Our approach allows to visualize the tree-likeness of all quartets in a single graph and therefore provides a tool for a quick interpretation of the phylogenetic content. We exemplify the method by applying it to simulated sequences that evolved on a star-tree or on a completely resolved tree. The analysis of two biological data sets (Zischler *et al.*, 1995; Friedrich and Tautz, 1995) concludes the chapter.

5.2 Method

5.2.1 Four sequences

Let us consider a set of four sequences, a so-called quartet. For this quartet the maximum-likelihood values belonging to the three possible fully resolved tree topologies (Figure 4.1) are computed using any model of sequence evolution (Swofford *et al.*, 1996; Zarkikh, 1994; Schöniger and von Haeseler, 1994). Let L_i be the maximum-likelihood of tree T_i where $i = 1, 2, 3$. Then we can compute posterior probabilities p_i for each tree (Equation 4.1).The probabilities (p_1, p_2, p_3) can be viewed as the barycentric coordinates of the point $\mathbf{P}$ belonging to the two-dimensional simplex

$$\mathbf{S}_2 = \{\sum_{i=1}^{3} p_i \mathbf{e}_i \mid p_1 + p_2 + p_3 = 1, p_i \geq 0\}, \tag{5.1}$$

where the $\mathbf{e}_i$ are real valued and independent. They point to the three corners of the simplex. As a special case $\mathbf{S}_2$ can be illustrated as an equilateral triangle. This construction allows an easy geometric interpretation of the p_i values. For a given point $\mathbf{P} \in \mathbf{S}_2$ the p_i are simply the lengths of the perpendiculars from the point $\mathbf{P}$ to the

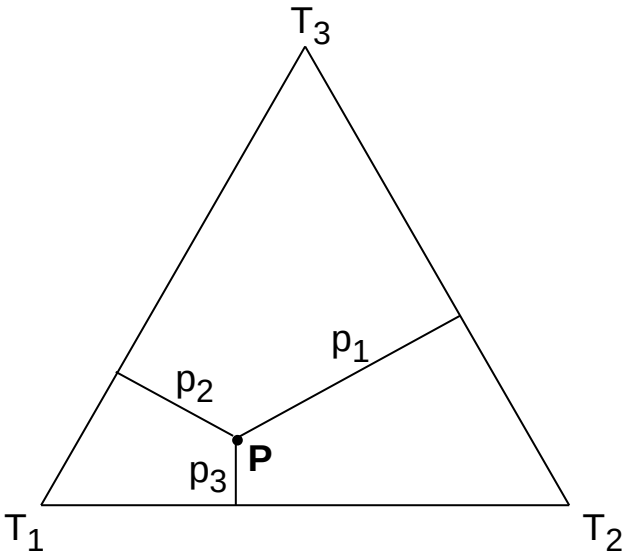

Figure 5.1: *Map of the probability vector* $\mathbf{P} = (p_1, p_2, p_3)$ *onto an equilateral triangle. Barycentric coordinates are used, i.e. the lengths of the perpendiculars from point* $\mathbf{P}$ *to the triangle sides are equal to the probabilities* p_i*. The corners* T_1, T_2, T_3 *represent three quartet topologies with corresponding coordinates* $(1, 0, 0), (0, 1, 0)$*, and* $(0, 0, 1)$*.*

three sides of the triangle (Figure 5.1). In the context of population genetics triangular coordinates of this kind are known as De Finetti diagrams (De Finetti, 1926).

If $\mathbf{P}$ is close to one corner of the triangle, the likelihoods (p_1, p_2, p_3) are clearly favoring one tree over the two others. Thus, every corner of the triangle corresponds to one of the three quartet topologies T_1, T_2, T_3. In a typical maximum-likelihood analysis the tree T_i is chosen with

$$p_i = \max\{p_1, p_2, p_3\}. \tag{5.2}$$

It is easy to compute the corresponding basins of attraction for each tree topology (Figure 5.2 A). The location of a point $\mathbf{P}$ in the simplex gives an immediate impression which tree is preferred.

Unfortunately, this picture is too optimistic. For real data it is not always possible to resolve the phylogenetic relationships of four sequences. This is either a consequence of limiting resolution due to short sequences ("noise") or the true evolutionary tree was a star phylogeny. To account for this case, we introduce a region in the triangle $\mathbf{S}_2$ representing the star phylogeny. The center $\mathbf{c}$ of the simplex is the point where all probabilities take on the value $p_i = 1/3$ which means that the three trees are equally

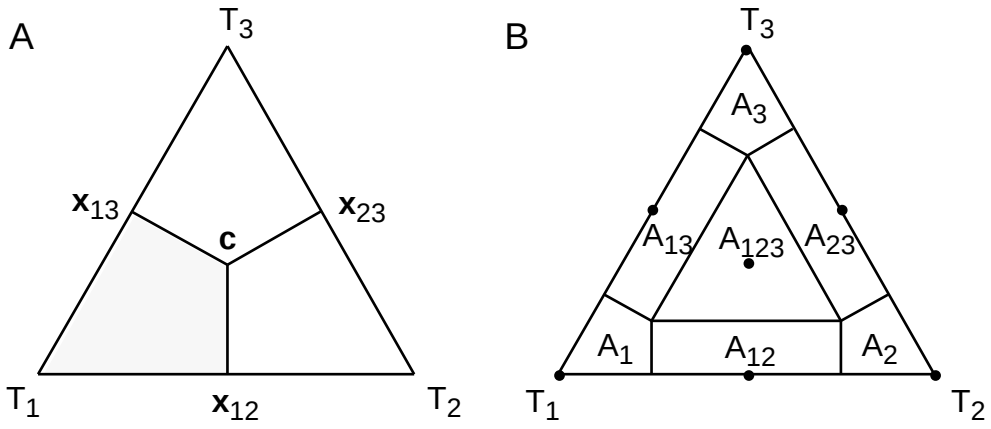

Figure 5.2: *(A) Basins of attraction for the three topologies $T_1, T_2,$ and T_3. The grey area shows the region where the probability for tree T_1 is largest. In the center $\mathbf{c} = (1/3, 1/3, 1/3)$ all trees are equally likely, at the points $\mathbf{x}_{12} = (1/2, 1/2, 0)$, $\mathbf{x}_{13} = (1/2, 0, 1/2)$, and $\mathbf{x}_{23} = (0, 1/2, 1/2)$ two trees have the same likelihood whereas the remaining one has probability 0. (B) shows the seven basins of attraction allowing not only fully resolved trees but also the star phylogeny and three regions where it is not possible to decide between two topologies. The dots indicate the corresponding seven attractors. A_1, A_2, A_3 are the tree-like regions. A_{12}, A_{13}, A_{23} represent the net-like regions and A_{123} is the star-like area.*

likely. Thus, if $\mathbf{P}$ is near the center the phylogenetic relationship cannot be resolved and is better displayed by a star phylogeny. On the other hand it also might be possible that one of the three trees can be excluded but the two others still remain undifferentiated. This is the case, if T_1 and T_2 show probabilities $p_1 = p_2 = 1/2$ and if $p_3 = 0$, for example. Near point $\mathbf{x}_{12}$ (see Figure 5.2 A) the phylogenetic relationship is best displayed by a net-like geometry that excludes tree T_3. Similarly, near points $\mathbf{x}_{13}$ and $\mathbf{x}_{23}$ it is impossible to unambiguously favor one tree. Based on these seven attractors in the triangle (marked with dots in Figure 5.2 B) the corresponding basins of attraction are easily computed. Each point in one of the seven regions has smallest Euclidean distance to its attractor. By A_{123} we denote the region where the star tree is the optimal tree. Its area equals the sum of the areas of A_1, A_2, A_3, the regions where one tree is clearly better then the remaining ones. The regions A_{ij} represent the situation where we can not distinguish trees T_i and T_j. The area of A_{ij} equals the sum of the area of A_i and A_j.

There is yet another way to describe the basins of attraction. If the three-dimensional simplex $\mathbf{S}_3$ is considered where the fourth corner represents the star phylogeny the basins of attraction can be viewed as projections of their corresponding volumes of the tetrahedron $\mathbf{S}_3$ onto the two-dimensional plane.

5.2.2 The general case

For a set of N aligned sequences there are exactly $\binom{N}{4}$ different possible quartets of sequences. To get an overall impression of the phylogenetic signal present in the data we compute the probability-vectors $\mathbf{P}$ for the quartets and draw the corresponding points in the simplex. If only few sequences are analyzed, $\mathbf{P}$ vectors of all $\binom{N}{4}$ quartets are considered, otherwise a random sample of, e.g., 1,000 quartets is sufficient to obtain a comprehensive picture of the phylogenetic quality of the data set. The resulting distribution of points in the triangle $\mathbf{S}_2$ forms a distinct pattern allowing us to predict *a priori* whether an N-sequence tree will show a good resolution. If most of the points $\mathbf{P}$ are found, e.g., in regions $A_{12}, A_{13}, A_{23},$ or in the star-tree region A_{123}, it is clear that the overall tree will be highly multifurcating. That is, evolution was either star-like or not tree-like at all. However, the opposite conclusion is not necessarily true. Even if all quartets are completely resolved, that is almost all $\mathbf{P}$-vectors are in A_1, A_2, A_3, it is possible that the overall N-sequence tree is not completely resolved (Bandelt and Dress, 1986).

5.2.3 Four-cluster likelihood-mapping

Instead of looking at all quartets, the analysis of tree-likeness for four disjoint groups of sequences (clusters) is also possible. Let C_1, C_2, C_3, and C_4 be a set of four clusters with c_1, c_2, c_3, and c_4 sequences. Then, we compute the probability-vectors $\mathbf{P}$ for the $c_1 \cdot c_2 \cdot c_3 \cdot c_4$ possible quartets and plot the corresponding points on the triangle $\mathbf{S}_2$. While the p_i values are randomly assigned to the trees T_1, T_2, T_3, when all quartets are studied, the assignment of p_i to tree T_i is now fixed. Each tree represents one of the three possible phylogenetic relationship among the clusters. As an illustration, think of the A, B, C, D at the T_1, T_2, T_3 (Figure 4.1) as a representative of the clusters C_i. The distribution of the $c_1 \cdot c_2 \cdot c_3 \cdot c_4$ probability vectors over the basins of attractions allows not only to identify the correct phylogenetic relationship of the four clusters but also shows the support for this and alternative groupings. This type of likelihood-mapping

Table 5.1: *Distribution of likelihood vectors* $\mathbf{P}$ *over the basins of attraction as a function of sequence length.*

	star tree			bifurcating tree		
length	$\sum A_i$	$\sum A_{ij}$	A_{123}	$\sum A_i$	$\sum A_{ij}$	A_{123}
50	17.9	4.8	77.3	61.1	6.8	32.1
100	16.2	4.2	79.6	82.0	3.7	14.3
200	11.1	3.6	85.3	91.5	3.3	5.2
500	9.8	3.7	86.5	100.0	0.0	0.0

Occupancies are shown as cumulative percentages for the three resolved regions (A_1, A_2, A_3), the three net-like regions (A_{12}, A_{13}, A_{23}), and the star-like region (A_{123}). Simulation of the data assumed a star phylogeny or a perfectly bifurcating tree.

analysis is a helpful tool to illustrate how well supported an internal branch of a given tree topology is.

5.3 Results

5.3.1 Simulation studies

Figure 5.3 displays the result of a typical likelihood-mapping analysis. A simulated set of 16 DNA-sequences was used to show the distribution of probability vectors $\mathbf{P}$ as a function of sequence length and the evolutionary history.

If evolution was according to a star topology then the probability-vectors are concentrated in the center of the simplex with rays emanating to the corners of the triangle. This picture does not change with increasing sequence length. However, the proportion of quartets found in area A_{123} increases (Table 5.1). If sequence evolution followed a completely resolved tree then the proportion of points $\mathbf{P}$ located inside $A_1 + A_2 + A_3$ increases with longer sequences, as an indication that noise due to sampling artifacts is diminished. Correspondingly, the number of quartets in the remaining regions decreases. For sequences of length 500 base pairs the non tree-like regions of the triangle are empty (Table 5.1). Thus, Figure 5.3 illustrates that likelihood-mapping enables an easy distinction between star-like or tree-like evolution. The influence of sequence length ("noise") on tree-likeness of the data is easily recognized.

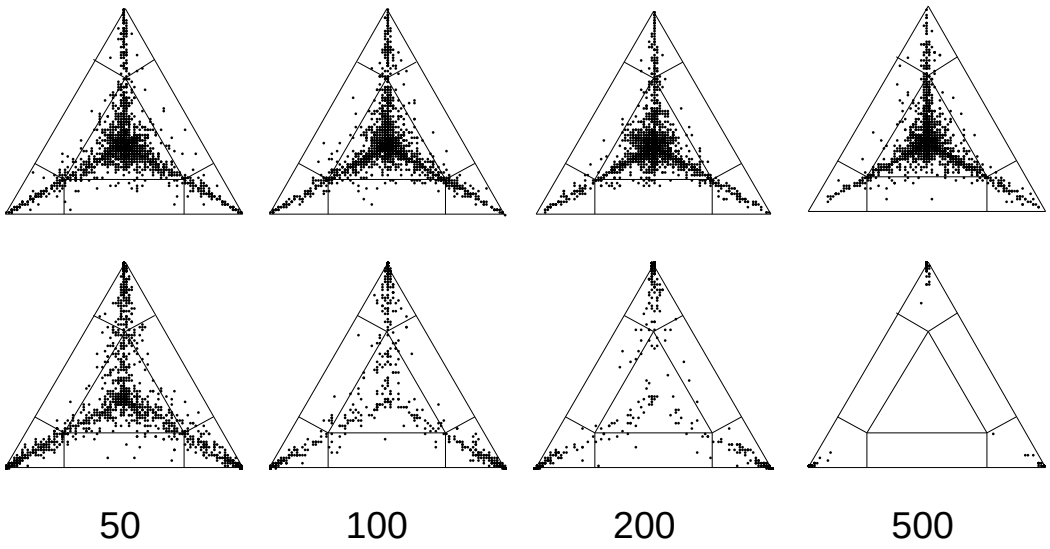

Figure 5.3: *Effect of sequence length (50, 100, 200, and 500 bp) on the distribution of* **P***-vectors for a simulated data set with 16 sequences.* (Upper) *Sequences evolving along a perfect star phylogeny.* (Lower) *Sequences evolving along a completely resolved tree. Sequences evolved according to the Jukes-Cantor model. The number of substitutions per site and per branch was 0.1. Each triangle shows a result of one simulation where all possible 1,820* **P***-vectors were computed. If tree-like data were generated* (Lower) *the number of* **P***-vectors seems to decrease with increasing sequence length. This effect is due to the fact that identical* **P***-vectors fall on top of each other. Longersequences increase the probability that one of the tree favored equals one. That is, most of the 1,820* **P***-vectors superimpose each other in the corners of the triangles (cf. Table 5.1).*

5.3.2 Data analysis

We illustrate the power of likelihood-mapping using two data sets published recently (Zischler *et al.*, 1995; Friedrich and Tautz, 1995). The first set (Zischler *et al.*, 1995) comprises eight partial cytochrome-b sequences (135 bp) and nine putative dinosaur sequences (Woodward *et al.*, 1994). The second alignment (1,850 bp) consists of ribosomal DNA from major arthropod classes (three myriapods, two chelicerates, two crustaceans, three hexapods) and six other sequences (human, *Xenopus, Tubifex, Caenorhabditis*, mouse, and rat). Likelihood-mapping suggests (Figure 5.4) that the Zischler *et al.* (1995) data show a fair amount of star-likeness with 17.5% of all quar-

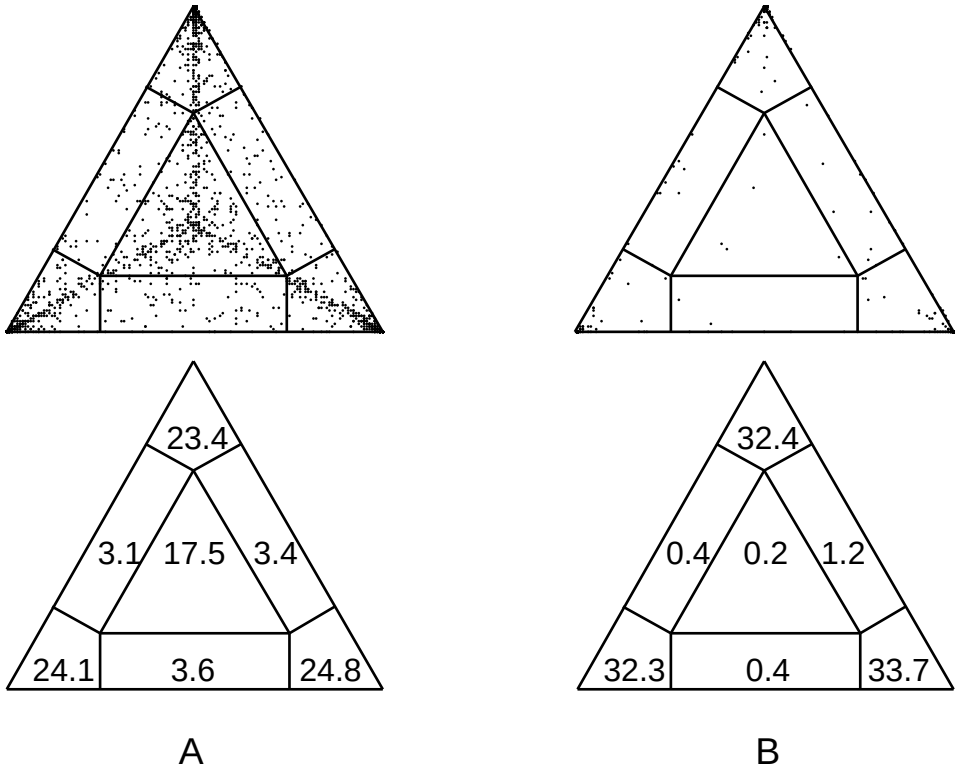

Figure 5.4: *Likelihood-mapping analysis for two biological data sets.* (Upper) *The distribution patterns.* (Lower) *The occupancies (in percent) for the seven areas of attraction. (A) cytochrome-b data (Zischler* et al.*, 1995). (B) Ribosomal DNA of major arthropod groups (Friedrich and Tautz, 1995).*

tet points in region A_{123} in contrast to only 0.2% for the ribosomal DNA. This result is corroborated by a bootstrap analysis (Zischler *et al.*, 1995; Friedrich and Tautz, 1995). Because of the short sequence length the percentage of quartets mapped into regions $A_{12}, A_{13},$ and A_{23} is with 10.1% for the sequences from Zischler *et al.* (1995) very high compared to 1.6% for the rDNA sequences. However, the cytochrome-b data still contain a reasonable amount of tree-likeness as 72.4% of all quartets are placed in the areas $A_1, A_2,$ and A_3. The tree-likeness of the ribosomal DNA is extremely high ($A_1+A_2+A_3 = 98.3\%$). The *a posteriori* analysis based on bootstrap values (Friedrich and Tautz, 1995) shows that all groupings in the tree receive high support.

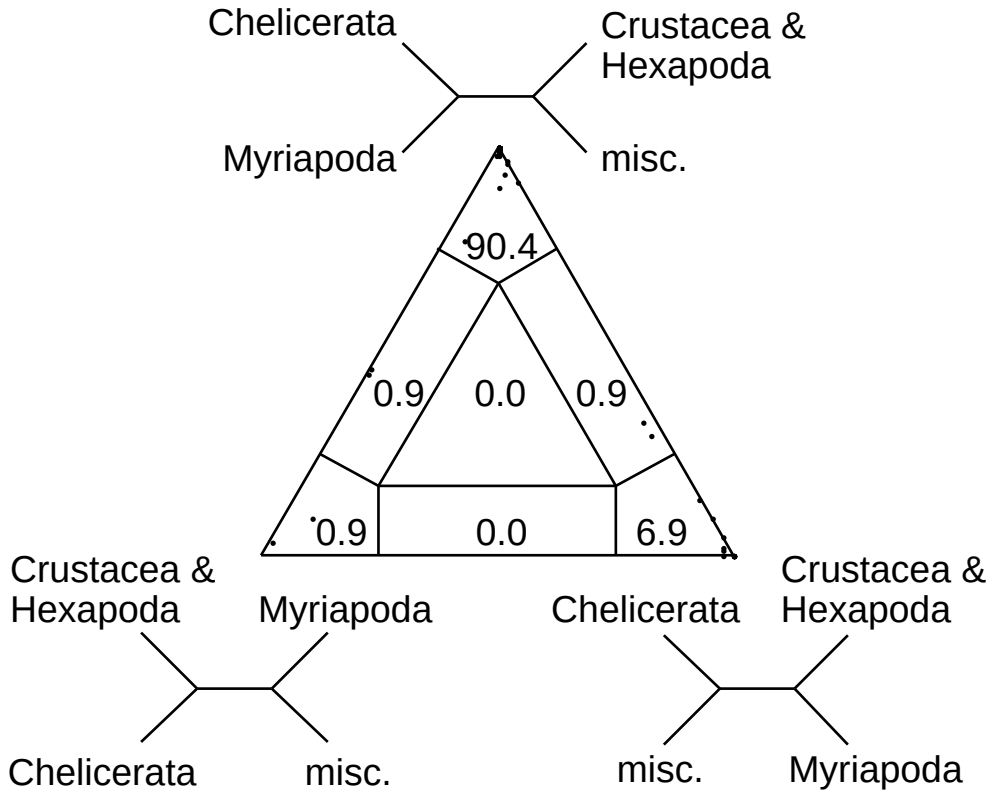

Figure 5.5: *Four-cluster likelihood-mapping of ribosomal DNA (Friedrich and Tautz, 1995). Sequences were split in four disjoint groups, misc. represents the non-arthropod sequences. The corners of the triangle are labeled with the corresponding tree topologies.*

5.3.3 Four-cluster likelihood-mapping

A further application of likelihood-mapping allows testing of an internal edge of a tree as given from any tree reconstruction method. As an example we consider the sister group status of myriapods and chelicerates as suggested by Friedrich and Tautz (1995). Figure 5.5 shows that 90.4% of all quartets between the four corresponding clusters support the branching pattern that groups chelicerates and myriapods versus crustaceans, hexapods and the remaining sequences. We find only very low support (6.9%) for the topology that pairs myriapods with crustaceans plus hexapods rather than with chelicerates or with the rest. Based on likelihood-mapping we can not reject the hypothesis of monophyly of myriapods and chelicerates. However, the outcome of statistical tests as suggested in Rzhetsky *et al.* (1995) remains to be seen.

5.4 Discussion

The evaluation of the phylogenetic contents in a data set is of prime importance to avoid false conclusions about evolutionary relationships among organism. Methods abound that evaluate the reliability of a reconstructed tree *a posteriori* (Swofford *et al.*, 1996). Likelihood-mapping can be viewed as a complementary approach to existing methods of *a priori* or *a posteriori* evaluations of tree-likeness. Our method may be helpful when analyzing controversial phylogenies. Similar to statistical geometry in sequence space (Eigen *et al.*, 1988; Eigen and Winkler-Oswatitsch, 1990; Nieselt-Struwe *et al.*, 1996) likelihood-mapping is based on the analysis of quartets, the basic ingredients to reconstruct trees (Bandelt and Dress, 1986). Moreover, the description of seven basins of attraction (Figure 5.2 B) that can be characterized as fully resolved (A_1, A_2, A_3), intermediate between two trees (A_{12}, A_{13}, A_{23}), or star-like (A_{123}) is also of great importance in the quartet-puzzling tree search algorithm (Chapter 4).

Here, we have provided a simple, but versatile, approach to visualize the phylogenetic content of a data set. We have shown that the method has reasonable predictive power. While we have presented only a visual tool to analyze the phylogenetic signal of sequences it is certainly necessary to develop solid statistical tests that provide evidence as to the significance of clusters (Rzhetsky *et al.*, 1995) or to a deviation from tree-likeness. For example, the assumption of equal prior probability for the trees may be debatable. It remains to be seen how approaches like Jeffrey's prior (Lake, 1995) or the inclusion of the variance of likelihood estimates (Hasegawa and Kishino, 1989) will influence the analysis.

Finally, it should kept in mind that the interpretation of the result of a likelihood-mapping analysis strongly depends on sequence length. The alignment of human mitochondrial control-region data (Vigilant *et al.*, 1991) comprises 1,137 positions. 82.5% of the quartets belong to the regions that represent fully resolved trees. Thus, the result suggests that the data are very well suited to reconstruct a well resolved tree. However, we observe 8.3% of all quartets in the star-like region A_{123} of the triangle. This value is too high for a completely resolved phylogeny (see Table 5.1). Therefore, we expect a phylogeny that is well resolved in certain parts of the tree only.

Chapter 6

Summary

In this doctoral thesis a variety of new methods for molecular phylogenetics based on maximum-likelihood were introduced:

Tree reconstruction: Quartet-puzzling, a fast and efficient heuristic tree search procedure for maximum-likelihood was developed. This method does not only determine a tree topology but in addition estimates local support values for internal branches. Moreover, in contrast to many other maximum-likelihood methods it is well-suited to compute trees for large data sets.

Data assessment: Likelihood-mapping, a novel technique to analyze and to visualize the phylogenetic content of a sequence alignment was presented. This method can be viewed as complementary approach to statistical geometry in sequence space. Likelihood-mapping is applicable to large data sets as well.

Intra-tree relationships: New measures for the reliability of internal branches were introduced. Quartet-puzzling support values are a helpful additional tool similar to bootstrap values. Four-cluster likelihood-mapping enables the estimation of the support of a single hypothesized internal branch without reconstructing an overall tree.

Parameter estimation: To speed up maximum-likelihood estimation of the parameters of a model of sequence evolution a number of useful simplification were proposed. The procedures allow the quick and reliable determination of the parameters of the model of the substitution process and of rate heterogeneity.

These methods were implemented in a computer program:

Software: PUZZLE, a user-friendly and platform-independent maximum-likelihood program was developed. It is distributed free of charge over the Internet (see Appendix A). PUZZLE is one of the fastest maximum-likelihood programs for molecular phylogeny currently available.

Data analysis: All computations in this thesis were done using PUZZLE except where indicated. The suitability of PUZZLE to study biological questions was illustrated by studying mitochondrial and ribosomal DNA sequences.

Appendix A

The PUZZLE Software

A PHYLIP (Felsenstein, 1993) compatible program PUZZLE has been developed to analyze nucleotide, two-state, and amino acid sequence data by the maximum-likelihood methods presented in this doctoral thesis.

A.1 Description

PUZZLE is a computer program to reconstruct phylogenetic trees from molecular sequence data by maximum likelihood. It implements a fast tree search algorithm, quartet puzzling, that allows analysis of large data sets and automatically assigns estimations of support to each internal branch. PUZZLE also computes pairwise maximum likelihood distances as well as branch lengths for user specified trees. Branch lengths can be calculated under the clock-assumption. In addition, PUZZLE offers a novel method, likelihood mapping, to investigate the support of a hypothesized internal branch without computing an overall tree and to visualize the phylogenetic content of a sequence alignment. PUZZLE also conducts a number of statistical tests for the data set, e.g., a χ^2-test for homogeneity of base composition over sequences, a likelihood ratio clock test (Felsenstein, 1988), and a test for comparison of different tree topologies (Kishino and Hasegawa, 1989). The models of substitution provided by PUZZLE are TN, HKY, F84, SH for nucleotides, Dayhoff, JTT, mtREV, BLOSUM 62 for amino acids, and F81 for two-state data. Rate heterogeneity is modelled by a discrete Γ-distribution and by allowing invariable sites. The corresponding parameters can be inferred from the data set.

PUZZLE is written in ANSI C and runs on all popular platforms (MacOS, Win-

dows 95/NT, UNIX, VMS). Further details about PUZZLE program can be found in its online manual

`http://www.zi.biologie.uni-muenchen.de/~strimmer/manual.html`.

A.2 Distribution

PUZZLE is distributed over the Internet. It is available free of charge from a number of servers. Currently the official home page of PUZZLE is located at the Institute of Zoology of the University of Munich (Germany)

`http://www.zi.biologie.uni-muenchen.de/~strimmer/puzzle.html`.

In addition, PUZZLE also is distributed by the European Bioinformatics Institute (United Kingdom)

`ftp://ftp.ebi.ac.uk/pub/software/`,

by the Institut Pasteur (France)

`ftp://ftp.pasteur.fr/pub/GenSoft/`,

and by the IUBio archive at the University of Indiana (USA)

`http://iubio.bio.indiana.edu/soft/molbio/evolve/`,
`ftp://iubio.bio.indiana.edu/molbio/evolve/`.

Bibliography

Adachi, J. and Hasegawa, M. 1996a. *MOLPHY: Programs for Molecular Phylogenetics, version 2.3.* Tokyo: Institute of Statistical Mathematics.

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

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., Schreier, P. H., Smith, A. J., Staden, R., and Young, I. G. 1981. Sequence and organization of the human mitochondrial genome. *Nature* **290**:457–465.

Bandelt, H.-J. and Dress, A. 1986. Reconstructing the shape of a tree from observed dissimilarity data. *Adv. Appl. Math.* **7**:309–343.

Bandelt, H.-J. and Dress, A. 1992. A canonical decomposition theory for metrics on a finite set. *Adv. Math.* **92**:47–105.

Churchill, G. A., von Haeseler, A., and Navidi, W. C. 1992. Sample size for a phylogenetic inference. *Mol. Biol. Evol.* **9**:753–769.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. 1978. A model of evolutionary change in proteins. In: *Atlas of Protein Sequences and Structure* (Dayhoff, M. O., ed.) volume 5 pp. 345–352. Natl. Biomed. Res. Found. Silver Springs.

De Finetti, B. 1926. Considerazioni matematiche sull'ereditarietà mendeliana. *Metron* **6**:29–37.

Dopazo, J., Dress, A., and von Haeseler, A. 1993. Split decomposition: A technique to analyze viral evolution. *Proc. Natl. Acad. Sci. USA* **90**:10320–10324.

Dress, A., von Haeseler, A., and Krüger, M. 1986. Reconstructing phylogenetic trees using variants of the four-point condition. *Studien zur Klassifikation* **17**:299–305.

Eigen, M., Lindemann, B. F., Tietze, M., Winkler-Oswatitsch, R., Dress, A., and von Haeseler, A. 1989. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science* **244**:673–679.

Eigen, M. and Nieselt-Struwe, K. 1990. How old is the immunodeficiency virus? *AIDS (suppl. 1)* **4**:S85–S93.

Eigen, M. and Winkler-Oswatitsch, R. 1990. Statistical geometry in sequence space. *Methods in Enzymology* **183**:505–530.

Eigen, M., Winkler-Oswatitsch, R., and Dress, A. 1988. Statistical geometry in sequence space: A method of quantitative comparative sequence analysis. *Proc. Natl. Acad. Sci. USA* **85**:5913–5917.

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

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum-likelihood approach. *J. Mol. Evol.* **17**:368–76.

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

Felsenstein, J. 1993. *PHYLIP: Phylogenetic Inference Package, version 3.5c*. Seattle: Department of Genetics, University of Washington.

Felsenstein, J. and Churchill, G. A. 1996. A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. Evol.* **13**:93–104.

Fitch, W. M. 1981. A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* **18**:30–37.

Fitch, W. M. and Margoliash, E. 1967. A method for estimating the number of invariant amino acid positions in a gene using cytochrome c as a model case. *Biochem. Gen.* **1**:65–71.

Friedrich, M. and Tautz, D. 1995. Ribosomal DNA phylogeny of the major extant arthropod classes and the evolution of myriapods. *Nature* **376**:165–167.

Gu, X., Fu, Y.-X., and Li, W.-H. 1995. Maximum-likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. *Mol. Biol. Evol.* **12**:546–557.

Hasegawa, M. and Kishino, H. 1989. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial mtDNA sequences. *Evolution* **43**:672–677.

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

Hedges, S. B. 1994. Molecular evidence for the origin of birds. *Proc. Natl. Acad. Sci. USA* **91**:2621–2624.

Huelsenbeck, J. P. 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* **44**:17–48.

Janke, A., Feldmaier-Fuchs, G., Gemmell, M., von Haeseler, A., and Pääbo, S. 1996. The complete mitochondrial genome of the platypus (ornithorhynchus anatinus) and the evolution of mammals. *J. Mol. Evol.* **42**:153–159.

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.

Jones, D. T., Taylor, W. R., and Thornton, J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput. Applic. Biosci.* **8**:275–282.

Jukes, T. H. and Cantor, C. R. 1969. Evolution of protein molecules. In: *Mammalian Protein Metabolism* (Munro, H. N., ed.) pp. 21–132. Academic Press New York.

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.

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.

Lake, J. A. 1995. Calculating the probability of multi-taxon evolutionary trees: Bootstrappers gambit. *Proc. Natl. Acad. Sci. USA* **92**:9662–9666.

Li, W.-H. and Graur, D. 1991. *Fundamentals of molecular evolution*. Sunderland, Massachusetts: Sinauer.

Lindgren, B. W. 1976. *Statistical Theory*. New York: Macmillan, 3rd edition.

Margush, T. and McMorris, F. R. 1981. Consensus n-trees. *Bull. Math. Biol.* **43**:239–244.

Mountain, J. L., Hebert, J. M., Bhattacharyya, S., Underhill, P. A., Ottolenghi, C., Gadgil, M., and Cavalli-Sforza, L. L. 1995. Demographic history of India and mtDNA sequence diversity. *Am. J. Hum. Genet.* **56**:979–992.

Nieselt-Struwe, K., Mayer, C. B., and Eigen, M. 1996. Determining the reliability of phylogenies with statistical geometry. Manuscript.

Olsen, G. J., Natsuda, H., Hagstrom, R., and Overbeek, R. 1994. FastDNAML: A tool for construction of phylogenetic trees of DNA sequences using maximum-likelihood. *Comput. Applic. Biosci.* **10**:41–48.

Penny, D., Steel, M., Waddell, P. J., and Hendy, M. D. 1995. Improved analyses of human mtDNA sequences support a recent African origin of Homo sapiens. *Mol. Biol. Evol.* **12**:863–882.

Rzhetsky, A., Kumar, S., and Nei, M. 1995. Four-cluster analysis: A simple method to test phylogenetic hypotheses. *Mol. Biol. Evol.* **12**:163–167.

Saitou, N. and Nei, M. 1987. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**:406–425.

Sattath, S. and Tversky, A. 1977. Additive similarity trees. *Psychometrika* **42**:319–345.

Schöniger, M. and von Haeseler, A. 1993. A simple method to improve the reliability of tree reconstructions. *Mol. Biol. Evol.* **10**:471–483.

Schöniger, M. and von Haeseler, A. 1994. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phyl. Evol.* **3**:240–247.

Strimmer, K., Goldman, N., and von Haeseler, A. 1997. Bayesian probabilities and quartet-puzzling. *Mol. Biol. Evol.* **14**:210–211.

Strimmer, K. and von Haeseler, A. 1996. Accuracy of neigbor joining for n-taxon trees. *Syst. Biol.* **45**:516–523.

Sullivan, J., Holsinger, K., and Simon, C. 1996. The effect of topology on estimates of among-site rate variation. *J. Mol. Evol.* **42**:308–312.

Swofford, D. L., Olsen, G. J., Wadell, P. J., and Hillis, D. M. 1996. Chapter 11: Phylogenetic inference. In: *Systematic Biology* (Hillis, D. M., Moritz, C., and Mable, B. K., eds.). Sinauer Associates Sunderland, Massachusetts.

Takahata, N. 1995. A genetic perspective on the origin and history of humans. *Annu. Rev. Ecol. Syst.* **26**:343–372.

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.

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

Thompson, J. D., Higgins, D. G., and Gibson, T. J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673–4680.

Uzzel, T. and Corbin, K. W. 1971. Fitting discrete probability distributions to evolutionary events. *Science* **172**:1089–1096.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. 1991. African populations and the evolution of human mitochondrial DNA. *Science* **253**:1503–1507.

von Haeseler, A. and Churchill, G. A. 1993. Network models for sequence evolution. *J. Mol. Evol.* **37**:77–85.

Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.* **37**:613–623.

Wakeley, J. 1996. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *TREE* **11**:158–163.

Woodward, S. R., Weynand, N. J., and Bunnell, X. 1994. DNA sequence from Cretaceous period bone fragments. *Science* **266**:1229–1232.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**:1396–1401.

Yang, Z. 1994a. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**:105–111.

Yang, Z. 1994b. Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. *J. Mol. Evol.* **39**:306–314.

Yang, Z. 1997. *Phylogenetic analysis by maximum-likelihood (PAML), version 1.3.* Pennsylvania State University: Institute of Molecular Evolutionary Genetics.

Zarkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* **39**:315–329.

Zischler, H., Höss, M., Handt, O., von Haeseler, A., van der Kuyl, A. C., Goudsmit, J., and Pääbo, S. 1995. Detecting dinosaur DNA. *Science* **268**:1192–1193.