universität wien

DISSERTATION

Titel der Dissertation

Goodness of fit and robustness of phylogenetic methods in the light of intermittent evolution

Verfasserin

B.Sc. Minh Anh Thi Nguyen

angestrebter akademischer Grad

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien, 2011

Studienkennzahl lt. Studienblatt:	A 091 490
Dissertationsgebiet lt. Studienblatt:	Molekulare Biologie
Betreuer:	Univ.-Prof.Dr. Arndt von Haeseler

Acknowledgments

First of all, I sincerely thank my supervisor, Arndt von Haeseler, for his continuous support and guidance. Thank you, Arndt, for your encouragement and for all the stimulating discussions and excellent suggestions which keep me motivated. I wish to thank the co-authors of my papers Steffen Kleare, in particular for his inspiring aid in statistics and mathematics, and Tanja Gesell, in particular for her fruitful discussions about studies of the performance of phylogenetic methods. This thesis is written using the personal pronoun 'we' in the scientific sense; it also reflects these vital interactions.

My sincere thanks to Bui Quang Minh for his constant scientific assistance including many detailed explanations, reading our manuscripts and my thesis as well as for other help especially at the beginning of my time in Vienna. I greatly acknowledge Mareike Fischer for helpful comments on our manuscripts previously and now this thesis. I would also like to thank Dirk Metzler and Simon Whelan for accepting to referee my thesis.

I take this chance to thank Le Sy Vinh for introducing me to Arndt; to thank Dinh Quang Huy and my brother Nguyen Dinh Tu for the help with some programming details; to appreciate the great hospitality from and valuable talks with Le Sy Quang and his wife Phan Hang during my Cambridge visit in 2009; and to acknowledge Lars Jermiin for kindly providing two book chapters in my urgent need.

I deeply thank all (former) members of the Center for Integrative Bioinformatics Vienna (CIBIV) for maintaining not only a scientifically open atmosphere where I can ask numerous questions without any hesitation but also a friendly and nutritious environment. Particularly, I wish to thank Ingo Ebersberger and Heiko Schmidt for gradually enriching my negligible knowledge of Bioinformatics at the beginning of my Ph.D. I also want to thank the used-to-be regular dart players for the delightful moments at work, but more importantly these dart games helped me to get along with the group faster. I acknowledge the technical support from Wolfgang Fischl and the help with diverse little things from Tina Köstler. In addition, I appreciate the constantly fast responses from the CIBIV administrative assistants. I have been really enjoying the time with you all.

A warm ‘thank you’ to my many Vietnamese friends and my former flatmates Orsolya, Harald, and Juliia who have made Vienna my ‘new home’ with an unforgettable cheerful time. Thanks also come to *anh* Trung, *chi* Hong for their friendship and for interesting chats about various aspects of life.

I am grateful to my grand mother, my parents, my parents in law, my uncles and aunts *cau* Duc *mo* Sa, *di* Dung *chu* Yen, and my ‘big’ family who have been encouraging and supporting me throughout every moment. Without them this thesis would haven’t been possible.

This thesis is dedicated to my husband, Dinh Viet Cuong, who always inspires me and shares my every up and down. Being with you, I simply find everywhere a ‘full house’.

Abstract

Charles Darwin's theory of 'The Origin of Species' (1859) states that species have evolved from common ancestors. Reconstructing so-called *phylogenetic trees* to elucidate the evolutionary relationships among species has since then become one of the main objectives in biology. In recent years, more and more phylogenetic studies have been published thanks to the advent of massive sequence data and to the development of efficient software packages. However, before drawing biological implications from the inferred evolutionary relationships, several issues should be taken into account. This thesis investigates two interesting issues in more detail:

First, how can one know that the model used describes the data adequately? We present MISFITS, a novel approach to evaluate the goodness of fit between a phylogenetic model and an alignment, which at the same time pinpoints to alignment site patterns that do not fit. MISFITS introduces a minimum number of *extra substitutions* on the inferred tree to provide a biologically motivated justification for the deviation between the observed site pattern frequency and the corresponding expectation. The extra substitutions plus the evolutionary model then fully explain the alignment. Moreover, the significance of the required number of extra substitutions can be determined by conducting a parametric bootstrap analysis. Therefore, MISFITS rejects inadequate models in terms of fit to the data. We demonstrate MISFITS on several examples and present a survey of the goodness of fit of the best-fit models (suggested by model selection) to thousands of alignments in the PANDIT database.

Second, insights into the performance of tree inference methods are essential because they may help to avoid wrong conclusions from the inferred phylogenies due to reconstruction artefacts such as long branch attraction. Among the criteria to evaluate the performance of a phylogenetic method, robustness to model violation is of particular practical importance as complete *a priori* knowledge of evolutionary processes is typ-

ically unavailable. We first develop ImOSM, a convenient tool to imbed *intermittent evolution* as model violation into an alignment. Intermittent evolution refers to extra substitutions occurring randomly on branches of a tree and thus changing alignment site patterns. We then study the robustness of widely used phylogenetic methods: maximum likelihood (ML), maximum parsimony (MP) and a distance-based method (BIONJ) to various scenarios of model violation. We show that violation of rates across sites (RaS) heterogeneity, and simultaneous violation of RaS and the transition transversion ratio along two nonadjacent external branches hinder all methods recovery of the true topology for a four-taxon tree. For an eight-taxon balanced tree these violations cause each of the three methods to infer a different topology: both ML and MP fail whilst BIONJ reconstructs the true tree. Furthermore, we report that several tests including the MISFITS test have enough power to detect such model violations. Thus, for analyses of real data, such reconstruction results require further investigation and these tests are recommended at the first glance.

Parts of this thesis have been published or submitted:

1. M. A. T. Nguyen, S. Klaere, and A. von Haeseler (2011) MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. *Mol. Biol. Evol.*, 28(1): 143-152.

2. M. A. T. Nguyen, T. Gesell and A. von Haeseler. ImOSM: Intermittent evolution and robustness of phylogenetic methods. Submitted to *Mol. Biol. Evol.*

Tools developed from the thesis are available at:

1. MISFITS: http://www.cibiv.at/software/misfits

2. ImOSM: http://www.cibiv.at/software/imosm

The thesis is organized as follows:

Chapter 1: We first present an introduction to phylogeny reconstruction. This consists of descriptions of phylogenetic trees, sequence alignments, models of sequence evolution and tree inference methods. We then briefly describe model test techniques to select the best model for a given data set. Subsequently, several notes that should be taken into the interpretation of the inferred trees are discussed.

Chapter 2: We introduce the term "intermittent evolution". Then, we demonstrate the construction of the one step mutation (OSM) matrix for nucleotide characters to model intermittent evolution. The OSM matrix is the fundamental concept for the methods developed in this thesis.

Chapter 3: We outline several methods for testing model fit in phylogeny inference. We then present and illustrate the MISFITS approach.

Chapter 4: We give a brief review of the performance of phylogenetic methods putting emphasis on studies of the robustness to model violation. Subsequently, we present the ImOSM tool to introduce model violation to the data and report our observation about the robustness of ML, MP and BIONJ.

Chapter 5: We summarize the content of the thesis and discuss an outlook.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Nothing in biology makes sense except in the light of evolution.
Theodosius Dobzhansky

1.1 Introduction to phylogeny reconstruction

Charles Darwin demonstrated in his work 'On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life' (Darwin, 1859) that species have evolved from common ancestors. The evolutionary relationships among species can be depicted by an evolutionary tree, also called *a phylogenetic tree* or *a phylogeny*. In the following, we briefly introduce phylogenetic trees as well as data, mathematical models, and methods used for their reconstruction.

1.1.1 Phylogenetic trees

Phylogenies or phylogenetic trees (illustrated in Figures 1.1a-b) are leaf-labelled trees, where the leaves represent contemporary *taxa* and the internal nodes are hypothetical *ancestors* (see e.g. Vandamme, 2009). An internal branch connects two internal nodes while an external branch connects a taxon with an internal node. Taxa that are connected through a single internal node are *adjacent* taxa. For example, taxa `A` and `B` in Figures 1.1a-b are adjacent but taxa `A` and `C` are nonadjacent. Two adjacent taxa (and the internal node connecting them) form a so-called *cherry*. The branching pattern of

the nodes defines the *topology* of the tree. Branch lengths, if given, usually indicate the number of substitutions per site (see the next sections) between the nodes.

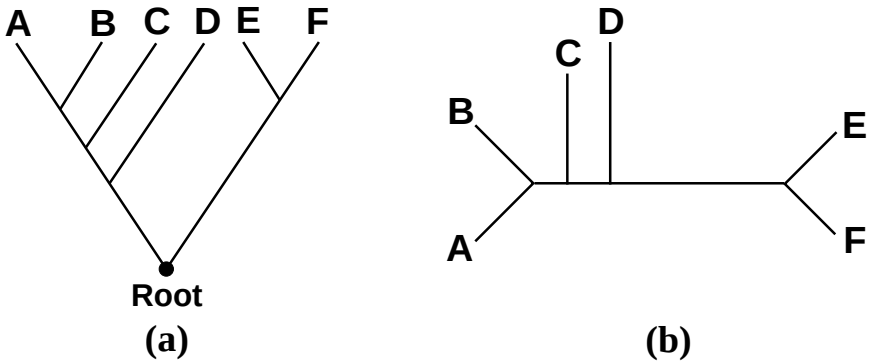


Figure 1.1: A rooted (a) and an unrooted (b) phylogenetic tree. `A, B, C, D, E,` and `F` are taxa or external nodes. The internal nodes (not labelled) are hypothetical ancestors of the taxa. Both trees have the same topology but the rooted tree has a root node which indicates the start of the evolutionary process toward the leaves. In the unrooted tree, the direction of evolution is unknown.

Phylogenies can be *rooted* (Figure 1.1a) or *unrooted* (Figure 1.1b). A rooted phylogeny shows the direction of the evolutionary process, whereas an unrooted tree only displays the relationships of the taxa. Phylogenies are in general *bifurcating* trees: all external nodes have a degree of one, all internal nodes have a degree of three except the "root" node in a rooted tree having a degree of two. A *multifurcating* or *unresolved* phylogeny contains node(s) of degree larger than three.

For n taxa, there are $(2n-5)!! = 1\times3\times\cdots\times(2n-5)$ distinct unrooted bifurcating trees (Felsenstein, 2004, Chapter 3). To compare two trees, the Robinson and Foulds (RF) distance (Robinson and Foulds, 1981) is usually employed. The RF distance between two trees is the number of *bipartitions* present in one of the two trees but not the other. A *bipartition* of a tree is defined by two disjoint subsets of the taxa which are separated by a branch. Two trees are topologically identical if the RF distance between them is zero.

1.1.2 Sequence alignment

As species have evolved from common ancestors, they share common characters including, e.g., morphological characters (phenotype) and genetic characters (genotype).

Characters that have descended, usually with divergence, from a common ancestral character are called **homologous** characters (Fitch, 2000). Phylogenetic trees are inferred based on these homologous characters. Nowadays, genetic data are prevalent in phylogeny reconstruction thanks to the rapidly increasing amount of DNA, RNA and protein sequences in public databases such as GenBank (Benson *et al.*, 2011) and UniProt (UniProt Consortium, 2010).

In DNA sequences, nucleotides exist in four character states: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). They are classified into either purine (A and G) or pyrimidine (C and T). In RNA sequences, Thymine is substituted by Uracil (U). Protein sequences are sequences of amino acids which are produced from nucleotide sequences through the well-known protein synthesis process:

$$\text{DNA} \overset{Transcription}{\longrightarrow} \text{mRNA} \overset{Translation}{\longrightarrow} \text{Protein.}$$

Three consecutive nucleotides in a protein-coding DNA sequence form a triplet called a codon. Every codon either encodes a single amino acid or signals the end of the above process (stop-codons). According to the standard genetic code (e.g. Vandamme, 2009), among the 64 possible codons 61 encode amino acids whereas the remaining three are stop-codons. As multiple codons may encode the same amino acid, there are in total twenty different amino acids as listed in Table 1.1. Thereby, in protein sequences an amino acid is available in any of these twenty character states.

Two homologous nucleotide sequences can be different due to **mutations**, e.g, errors during DNA replication, during DNA repair or due to environmental factors. Mutations are categorizied into (see e.g. Li, 1997, pp 23-30):

Substitutions: replacement of one nucleotide by another. Nucleotide substitutions are classified into **transitions** between the purines (A and G) or between the pyrimidines (C and T) and **transversions** between a purine and a pyrimidine.

Deletions: deletion of one or several nucleotides from the sequence.

Insertions: insertion of one or several nucleotides into the sequence.

Recombination: combination of different parts of a sequence(s) into one sequence fragment.

Inversion: rotation by 180° of a double-stranded DNA segment.

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

Table 1.1: The twenty amino acids and their abbreviations.

In order to reconstruct phylogenetic trees from nucleotide or amino acid sequences, we first align the sequences into a so-called *multiple sequence alignment* such that homologous characters form a column, also called an alignment site (see e.g. Higgins and Lemey, 2009). An alignment of two sequences is called a *pairwise sequence alignment*. Table 1.2 shows a multiple sequence alignment of five DNA fragments from human, chimpanzee, gorilla, rhesus, and mouse. Alignment sites 1-3 contain substitutions. Gap characters (-) introduced to columns 4-5 indicate insertions/deletions. The remaining columns, from 6 to 12, are constant sites i.e. the nucleotides in all sequences are identical in each column.

	1	2	3	4	5	6	7	8	9	10	11	12
Human	A	A	C	C	T	T	T	C	C	A	G	G
Chimpanzee	G	A	C	-	T	T	T	C	C	A	G	G
Gorilla	C	C	G	C	-	T	T	C	C	A	G	G
Rhesus	T	C	T	-	-	T	T	C	C	A	G	G
Mouse	T	G	G	-	T	T	T	C	C	A	G	G

Table 1.2: Example of a multiple sequence alignment. Columns 1-3 contain mismatches which indicate substitutions. Gap characters '-' (insertions/deletions) are introduced to columns 4-5. The remaining columns are all constant sites.

In this thesis, a *pattern* or a *site pattern* refers to any possible combination of the character states in the aligned sequences at one alignment column. An alignment of n nucleotide sequences shows at most 4^n different site patterns. A pattern may occur many times in the alignment while some patterns might not be present at all.

1.1.3 Models of sequence evolution

In order to employ statistical inference techniques to reconstruct phylogenetic trees from molecular data (Section 1.1.4) and to estimate the so-called *genetic distance* between two homologous sequences (a fundamental concept in phylogenetics and sequence analysis), one needs a probabilistic description of the substitution process according to which a sequence evolves, i.e., a model of sequence evolution. Before describing such a model, we first define distances between a pair of homologous sequences. Given two homologous sequences $\mathbf{x}$ and $\mathbf{y}$ of the same length, we distinguish two kinds of distances:

The observed distance between $\mathbf{x}$ and $\mathbf{y}$ is the number of positions at which the character states in $\mathbf{x}$ and $\mathbf{y}$ mismatch divided by the sequence length.

The genetic distance or evolutionary distance between $\mathbf{x}$ and $\mathbf{y}$ is the actual number of substitutions per site which have occurred between $\mathbf{x}$ and $\mathbf{y}$ during evolution.

Thereby, the observed distance is actually the normalized Hamming distance and its computation is straightforward given the two sequences. The observed distance is a proper estimation of the genetic distance if the genetic distance is small. However, if many substitutions (per site) have occurred between the two sequences because of, e.g.,

a high substitution rate (see the next sections), then their observed distance usually underestimates their genetic distance (e.g. Strimmer and von Haeseler, 2009). This is due to the fact that (i) *multiple substitutions*, i.e. two or more substitutions happening at the same site such as A to C and then C to T, are counted as one substitution (from A to T), and (ii) *back substitutions* at a site, e.g. A changed to C and then C changed back to A, are not observed at all. Thereby, one needs to employ statistical techniques to estimate the genetic distance between two homologous sequences assuming a model of sequence evolution.

We outline in the next sections the widely used models of sequence evolution including models of character (nucleotide or amino acid) substitution, models of rates across sites heterogeneity, and more sophisticated, recently developed evolutionary models.

1.1.3.1 Models of nucleotide substitution

For nucleotide sequences, the substitution process is commonly modeled as a *time-homogeneous, time-continuous, stationary Markov process* (Tavaré, 1986). The central component of the process is the so-called *instantaneous substitution rate matrix* $\mathrm{Q} = \{\mathrm{q}_{ij}\}$, which defines the rate of substitution from one nucleotide state (i) to another state (j) per time unit. Usually, time-reversibility is presumed, i.e., the probability of changing from i to j over a time t is the same as changing from j to i. For one unit of evolutionary time, the probability of changing from i to j is the product of the instantaneous substitution rate from i to j, q_{ij}, and the probability of i, π_i. Therefore, the time-reversibility condition implies that $\pi_i \mathrm{q}_{ij} = \pi_j \mathrm{q}_{ji}$. To this end, the most general time-reversible model GTR (Tavaré, 1986) is determined by (see also e.g. Felsenstein, 2004, pp. 196-211; Strimmer and von Haeseler, 2009):

$$\mathrm{Q} = \begin{array}{c} \\ A \\ G \\ C \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & G & C & T \end{array} \\ \begin{pmatrix} . & \pi_G a & \pi_C b & \pi_T c \\ \pi_A a & . & \pi_C d & \pi_T e \\ \pi_A b & \pi_G d & . & \pi_T f \\ \pi_A c & \pi_G e & \pi_C f & . \end{pmatrix} \end{array}, \tag{1.1}$$

where π_A, π_G, π_C and π_T are the stationary equilibrium frequencies of nucleotides A, G, C, and T, respectively ($\pi_A + \pi_G + \pi_C + \pi_T = 1$). Parameters a, b, c, d, e, f indicate the substitution rates between specific nucleotides, thereafter referred to as *relative substitution*

rates to distinguish them from the instantaneous substitution rates. The diagonal elements q_{ii} are determined such that the sum of each row equals zero:

$$q_{ii} = -\sum_{j \neq i} q_{ij}. \tag{1.2}$$

Apart from the time-reversibility assumption, the GTR model assumes four conditions: (i) At any given site in the sequence, the rate of change from nucleotide state i to nucleotide state j is independent of the history of i (*Markov property*). (ii) The substitution rates are constant over time (*time-homogeneity*). (iii) Substitutions between nucleotides can occur at any time during evolution (*time-continuity*); and (iv) the base frequencies $\pi = (\pi_A, \pi_G, \pi_C, \pi_T)$ are at equilibrium (*stationarity*).

The Q matrix is scaled such that the *total substitution rate* over all nucleotides at equilibrium is one. This is equivalent to:

$$2(\pi_A\pi_G a + \pi_A\pi_C b + \pi_A\pi_T c + \pi_G\pi_C d + \pi_G\pi_T e + \pi_C\pi_T f) = -\sum \pi_i q_{ii} = 1 \tag{1.3}$$

This implies that one expects to see one substitution per unit of evolutionary time. Thereby, the GTR model consists of 8 free parameters: 3 for the base frequencies, and 5 for the relative substitution rates between specific nucleotides. Restrictions imposed to the GTR model result in nested (less parameter) models such as JC69 (Jukes and Cantor, 1969) and HKY85 (Hasegawa *et al.*, 1985). Table 1.3 presents a list of widely used nucleotide substitution models together with the substitution types distinguished, equal or unequal base frequencies, and the number of free parameters to be estimated.

The Q matrix of the GTR model can be decomposed into two components: the matrix of relative substitution rates (the R matrix) and the diagonal matrix of base frequencies (the π matrix):

$$Q = R\pi = \begin{pmatrix} . & a & b & c \\ a & . & d & e \\ b & d & . & f \\ c & e & f & . \end{pmatrix} \begin{pmatrix} \pi_A & 0 & 0 & 0 \\ 0 & \pi_G & 0 & 0 \\ 0 & 0 & \pi_C & 0 \\ 0 & 0 & 0 & \pi_T \end{pmatrix}, \tag{1.4}$$

where the diagonal entries of R are determined such that Equation 1.2 is satisfied.

Once the instantaneous rate matrix is specified the substitution probability matrix $P(t) = \{p_{ij}(t)\}$, which provides the probabilities of changing from a nucleotide i to

Model[a]	Substitution types distinguished	Base freq.	#Params
JC69	None (a=b=c=d=e=f)	Equal	0
F81	None (a=b=c=d=e=f)	Unequal	3
K2P	Transitions vs. transversions (a=f, b=c=d=e)	Equal	1
HKY85	Transitions vs. transversions (a=f, b=c=d=e)	Unequal	4
K3ST	Transitions and 2 transversions (a=f, b=e, c=d)	Equal	2
K3STuf	Transitions and 2 transversions (a=f, b=e, c=d)	Unequal	5
TN93ef	2 transitions and transversions (a, f, b=c=d=e)	Equal	2
TN93	2 transitions and transversions (a, f, b=c=d=e)	Unequal	5
SYM	All substitutions (a, b, c, d, e, f)	Equal	5
GTR	All substitutions (a, b, c, d, e, f)	Unequal	8

Table 1.3: List of widely used nucleotide substitution models together with the substitution types distinguished, equal or unequal base frequencies and the number of free parameters. [a]JC69 (Jukes and Cantor, 1969), F81 (Felsenstein, 1981b), K2P or K80 (Kimura, 1980), HKY85 (Hasegawa *et al.*, 1985), K3ST or K81 (Kimura, 1981), TN93 (Tamura and Nei, 1993), SYM (Zharkikh, 1994), and GTR (Tavaré, 1986).

another nucleotide j over an evolutionary time t (e.g. a branch length in a tree), can be calculated by (see e.g. Strimmer and von Haeseler, 2009):

$$\mathrm{P}(t) = \exp(\mathrm{Q}t). \tag{1.5}$$

As Q is normalized so that the total substitution rate is one (Equation 1.3), the evolutionary time t is interpreted as the number of substitutions per site. As R is symmetric, the decomposition of Q as in Equation 1.4 enables an efficient way to compute $\exp(\mathrm{Q}t)$ (Felsenstein, 2004, pp. 196-211).

One may employ $\mathrm{P}(t)$ to estimate the genetic distance between two sequences using the maximum likelihood principle (e.g. Strimmer and von Haeseler, 2009, see also Section 1.1.4.2). Given two homologous nucleotide sequences of length ℓ, $\mathbf{x} = x_1 x_2 \cdots x_\ell$ and $\mathbf{y} = y_1 y_2 \cdots y_\ell$, the likelihood function calculates the probability of observing these two sequences under the evolutionary model specified by Q given that d substitutions

per site have taken place between them:

$$\mathbb{L}(d) = \Pr(\mathbf{x},\mathbf{y} \mid \mathrm{Q}, d) = \prod_{k=1}^{\ell} \pi_{x_k} \mathrm{p}_{x_k y_k}(d). \tag{1.6}$$

The distance $\hat{d}$ which maximizes $\mathbb{L}(d)$ is called the maximum likelihood estimate of the genetic distance between the two sequences (Strimmer and von Haeseler, 2009).

1.1.3.2 Models of amino acid substitution

The substitution process between amino acids is also assumed to be a time-homogeneous, time-continuos, time-reversible, stationary Markov process. However, as twenty possible amino acid states require estimations of too many substitution model parameters, these parameters are usually derived from empirical studies from large amounts of data. Such empirical amino acid substitution models include, e.g., PAM (Dayhoff *et al.*, 1978), JTT (Jones *et al.*, 1992), mtREV (Adachi and Hasegawa, 1996), WAG (Whelan and Goldman, 2001) and LG (Le and Gascuel, 2008).

1.1.3.3 Models of rate heterogeneity

So far we have assumed that the instantaneous substitution rates between nucleotide or amino acid characters are the same for every site in the alignment, i.e. the homogeneous rate model. However, it has been shown that substitution rates may vary across sites in an alignment (see e.g. Li, 1997, pp. 74-78 for a brief review and references therein). This phenomenon is termed *rates across sites (RaS) heterogeneity*. For example, in protein-coding nucleotide sequences, rates of substitution at the third codon positions are typically much larger than those at the first and the second codon positions (see e.g. Rodríguez-Trelles *et al.*, 2006; Bofkin and Goldman, 2007).

Rates across sites heterogeneity has been modeled by, e.g., assuming a proportion of invariable sites (zero rate of change) in the alignment (e.g. Churchill, 1992), a gamma distribution (Uzzell and Corbin, 1971; Yang, 1994b), or a combination of invariable sites and gamma distribution (Gu *et al.*, 1995). More complex models employ site-specific rates, i.e., every site has its own rate (Meyer and von Haeseler, 2003). It should be noted, that in order to make the branch length the number of substitutions per site, the mean of the rates of substitution over all sites must be one, e.g., the Γ-distribution used

to model RaS heterogeneity must have a mean of one. Assume that every alignment site a_i possesses an evolutionary rate r_i with $\frac{1}{\ell}\sum_{i=1}^{\ell} r_i \approx 1$, then the probability of site a_i given a branch length t is computed based on the probability matrix $\mathrm{P}(r_i t) = \exp(r_i \mathrm{Q} t)$. This implies that the "actual" or "average" time of evolution t (the average number of substitutions per site) is scaled by r_i for site a_i.

In general, an evolutionary model involves a substitution model between the characters plus a model of RaS heterogeneity. For instance, 'JC69+I' indicates the JC69 model of nucleotide substitution together with the invariable site model of RaS heterogeneity; '+Γ' specifies a Γ-distribution for RaS heterogeneity; '+I+Γ' for both invariable sites and a Γ-distribution; or simply JC69 for homogeneous rates across sites. The total number of parameters to be estimated for an evolutionary model is the sum of the number of free parameters from the substitution model (c.f. Table 1.3 for nucleotide substitution models) and the number of parameters to model rates across sites, which is zero for homogeneous rates; one for the proportion of invariable sites (I); one for the Γ-shape parameter α; or the number of site patterns in the alignment minus 1 for site-specific rates under the condition that alignment sites showing the same pattern have the same evolutionary rate.

1.1.3.4 More complex models of sequence evolution

We have outlined the widely used models of sequence evolution which comprise an instantaneous rate matrix $\mathrm{Q} = \mathrm{R}\pi$ and a model of rates across sites (homogeneity or heterogeneity). While inferring the phylogeny using these models, we assume a single, constant Q matrix along the tree and across the sites. In the following, we briefly discuss a number of more complex, more "realistic" models which have been proposed recently.

Compositional heterogeneity commonly refers to the differences in the composition of nucleotide or amino acid bases, i.e. different π frequencies, among DNA or protein sequences (see e.g. Jermiin *et al.*, 2009). Models which accommodate compositional heterogeneity assign to each branch a specific π matrix (e.g. Yang and Roberts, 1995), or a specific GC content (e.g. Galtier and Gouy, 1998); and allow for more than one π matrix along the tree, i.e. several branches might have the same π matrix (e.g. Foster, 2004; Blanquart and Lartillot, 2006; Gowri-Shankar and Rattray, 2007). Lartillot and Philippe (2004) developed a model of amino acid

sequence evolution that permits the equilibrium distribution of base frequencies to differ across sites. More recently, Blanquart and Lartillot (2008) incorporated both branch- and site-heterogeneous composition of base frequencies. Tools like SeqVis (Jermiin *et al.*, 2009) detect compositional heterogeneity among aligned nucleotide sequences.

Relative substitution rate heterogeneity. We refer this class of models to those that allow for different R matrices across branches or across sites. These models are mostly applied to nucleotide data only as for amino acid data, the relative substitution rates are usually taken from the empirical model and fixed while inferring the tree. For instance, Huelsenbeck and Nielsen (1999) sampled the transition/transversion ratio for each site of the alignment from a gamma distribution.

Pagel and Meade (2004) developed a mixture model to allow for more than one Q matrix, each has a particular probability or weight at each alignment site. Both constituents, the π and R matrices, may be different among these Q matrices. The evolution of a site is mainly characterized by the Q matrices that have the largest probabilities at this site. As each Q matrix weights differently across sites, this model incorporates both compositional heterogeneity and relative substitution rate heterogeneity across sites.

Changing the π or the R matrices leads to different Q matrices. Therefore, models allowing compositional heterogeneity and/or relative substitution rate heterogeneity relax the assumption of a constant Q matrix along the tree or across alignment sites.

Heterotachy refers to a phenomenon that the substitution rate at a site changes through time (Philippe and Lopez, 2001; Lopez *et al.*, 2002). According to this definition covarion models, first introduced by Fitch and Markowitz (1970); Tuffley and Steel (1998), reflect heterotachy. These models employ, in addition to the substitution process, a switching process between two classes "on" and "off". If a site is in the "on" state, the corresponding (nucleotide/amino acid) character can change to another character state, otherwise it cannot change. The switching between "on" and "off" states of a site during the course of evolution has a rate, called the switching rate. Other covarion-like models allow a proportion of alignment sites to switch among different substitution rates over time (Galtier, 2001). Wang *et al.* (2007) generalized these models so that a site can switch between a variable

state and an invariable state, and if it is in the variable state, it can switch among multiple substitution rates.

Mixed branch length models which allow site-specific branch length sets (Tuffley and Steel, 1997; Kolaczkowski and Thornton, 2004; Spencer *et al.*, 2005) are also considered to imply heterotachy though the implication is not so straightforward. The earliest mixed branch length model was described by Tuffley and Steel (1997), the so-called "no-common-mechanism" model, where every site has its own set of branch lengths on the same tree topology. Maximum likelihood with no-common-mechanism and maximum parsimony are shown to be equivalent (Tuffley and Steel (1997), see also Yang (2006), pp. 198-201 and Fischer and Thatte (2010) for more discussion). Other mixed branch length models (Kolaczkowski and Thornton, 2004; Spencer *et al.*, 2005), which were implemented in Kolaczkowski and Thornton (2008), and the one from Pagel and Meade (2008) partition the alignment into different regions where each region evolves under a specific set of branch lengths. In a simple setting, for instance, half of the alignment sites evolved under one set of branch lengths whilst the other half evolved under another set (Kolaczkowski and Thornton, 2004). Mixed branch length models need to be distinguished from RaS heterogeneity models (Figure 1.2a-b). In RaS heterogeneity models (Section 1.1.3.3) a "slow-evolving" site a_i evolves more slowly than a "fast-evolving" site a_j across the whole tree (Figure 1.2a). This implies that the tree which gives rise to a_j is "scaled" by a factor larger than one with respect to the tree (same topology) giving rise to a_i. In mixed branch length models, a site a_i evolves more slowly than another site a_j on some branches, e.g. on the external branches leading to C and to D in Figure 1.2b, but on some other branches (leading to A and to B) a_i evolves faster than a_j. Therefore, one cannot "scale" the tree giving rise to a_i to obtain the tree along which a_j evolves. In comparison with site a_j the evolutionary rate of site a_i shifts, during the course of evolution, from faster on the branches leading to A and to B to slower on the branches leading to C and to D. This might be the reason why mixed branch length models are commonly considered a kind of heterotachy.

More recently, Whelan (2008) employed the so-called "temporal hidden Markov model" to accommodate many kinds of heterogeneity for nucleotide data: compositional heterogeneity, relative substitution rate heterogeneity (reflected in the transition/transversion

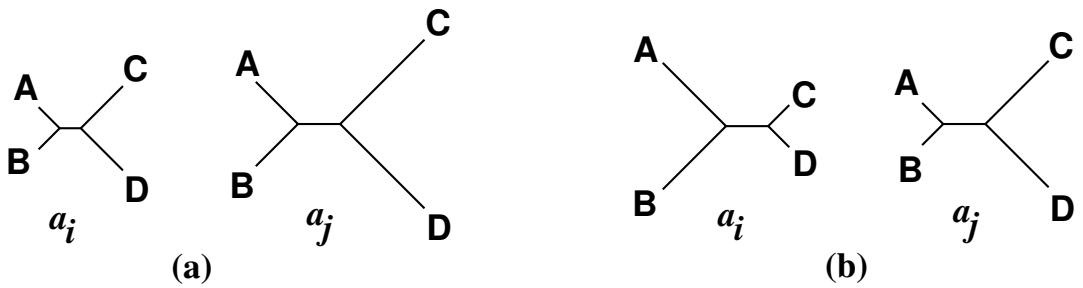


Figure 1.2: Examples of (a) a RaS heterogeneity model and (b) a mixed branch length model.

ratio variation), and substitution rate heterogeneity across both branches on the tree and sites in the alignment. A similar technique was implemented to detect heterotachy in protein sequences (Whelan *et al.*, 2011).

This is an incomplete list of the growing number of sophisticated evolutionary models (codon models, e.g., are not mentioned). Clearly, these complex models cast a substantially large number of parameters to be estimated for likelihood-based phylogeny inference (Section 1.1.4.2). For distance-based tree reconstruction (Section 1.1.4.3), distance adjustments to account for such complex scenarios of sequence evolution exist. For instance, the LogDet (Lockhart *et al.*, 1994) distance accounts for compositional heterogeneity, and the modified Tamura-Nei distance (Tamura and Kumar, 2002) incorporates both compositional and relative rate heterogeneity across branches. Wu and Susko (2009) accommodate general heterotachy by estimating for each pair of sequences an individual Γ-shape parameter α and then computing the pairwise genetic distance accordingly.

1.1.4 Methods for phylogeny reconstruction

Given a sequence alignment $\mathcal{A}$, the task of phylogeny reconstruction is to find a tree (with branch lengths) that best describes the observed sequence data. In order to find a *best* tree, a criterion to score trees is required. There are different optimality criteria that give rise to the following methods:

1.1.4.1 Maximum parsimony

Maximum parsimony (MP) searches for a tree that requires the minimum number of substitutions to explain the variation of the sequences in the alignment (Edwards and Cavalli-Sforza, 1963; Camin and Sokal, 1965; Fitch, 1971). Given an alignment site a_i and a tree, the parsimonious principle assigns sets of possible character states to all internal nodes of the tree to attain a_i at the leaves with the minimum number of substitutions along the branches. The number of substitutions needed is called *parsimony length (PL)*. The *total parsimony length* is the sum of the parsimony lengths over all alignment sites. Tree reconstruction with MP selects the trees with the minimum total parsimony length.

Fitch's algorithm (Fitch, 1971) is often used to compute PL and to derive the possible ancestral character states on a rooted tree for a given site pattern. The algorithm consists of two phases. The first phase simultaneously computes PL and determines the *preliminary* sets of possible character states at the internal nodes. We initially set PL to zero and traverse on the tree from the leaves toward the root. The character set at each leave is always the set containing the corresponding character in the pattern. The preliminary set at an internal node is then either the intersection of the two preliminary sets at its intermediate descendants if their intersection is not empty or their union, otherwise. In case of a union, we increase PL by 1. Figure 1.3a presents an illustration of this phase, where the parsimony length of the pattern $ACGGC$ on the rooted five-taxon tree shown is 2. The second phase of the algorithm introduces several rules to derive the *final* sets of possible character states at the internal nodes from the preliminary sets constructed in the first phase (see Fitch, 1971, for more detail). Applying those rules to our example results in the assignment shown in Figure 1.3b: the character set at every internal node contains exactly one character state, thus yielding a unique assignment of substitutions to the branches.

The Fitch algorithm described above does not distinguish different types of substitutions while other variants (e.g. Sankoff, 1975) weight substitutions differently. In addition, because more than one character state might be possible at the internal nodes of the MP tree (see e.g. Figure 1b in Fitch (1971)), one can assign substitutions on the branches according to (i) *accelerated transformation* ACCTRAN (substitutions are introduced as close to the arbitrary root as possible), or (ii) *delayed transformation* DELTRAN which tries to delay substitutions until the external branches if possible (Swofford and Maddison (1987), see also Farris (1970)).

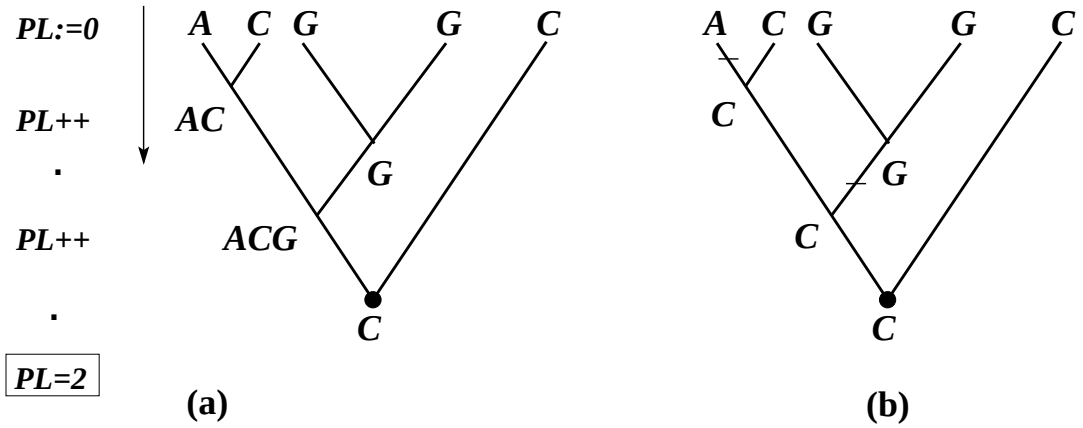


Figure 1.3: An illustration of the Fitch algorithm (Fitch, 1971). (a) illustrates the first phase of the algorithm which simultaneously computes the parsimony length PL of the pattern $ACGGC$ on the tree and determines the preliminary sets of possible ancestral character states. (b) shows the final sets of possible character states at the internal nodes obtained by applying the second phase of the algorithm.

1.1.4.2 Maximum likelihood

Maximum likelihood (ML) searches for a tree that maximizes the conditional probability of the alignment given the tree and a model of sequence evolution $\mathcal{M}$ (Felsenstein, 1981a). The tree with branch lengths $\mathcal{T}_{br}$ and the evolutional model $\mathcal{M}$ with model parameters θ compose a hypothesis H of the evolutionary process that gave rise to the observed data $\mathcal{A}$. The *likelihood* of H is (proportional to) the conditional probability of $\mathcal{A}$ given $\mathcal{M}, \theta,$ and $\mathcal{T}_{br}$. This conditional probability is computed as the product of the conditional probabilities of all sites in the alignment under the assumption that alignment sites evolve independently. Thus:

$$\mathbb{L}(\mathrm{H}) = \Pr(\mathcal{A} \mid \mathrm{H}) = \prod_{i=1}^{\ell} \Pr(a_i \mid \mathcal{M}, \theta, \mathcal{T}_{br}) \tag{1.7}$$

or the log likelihood:

$$\mathrm{L}(\mathrm{H}) = \sum_{i=1}^{\ell} \ln\left(\Pr(a_i \mid \mathcal{M}, \theta, \mathcal{T}_{br})\right), \tag{1.8}$$

where ℓ is the alignment length.

Recall from the last chapter (Section 1.1.3), that the model $\mathcal{M}$ and its parameters θ specify an instantaneous rate matrix Q and thus we can compute the substitution probability matrix $P(t)$ for every branch of a tree according to Equation 1.5 (for models with, e.g., compositional/relative rate heterogeneity across branches, different branches may have different Q matrices). It should be noted that the likelihood function used to estimate the genetic distance between two sequences (Equation 1.6) is a simple case of Equation 1.7 where the tree $\mathcal{T}_{br}$ has only one branch of length d which connects the two sequences (taxa). Having $P(t)$ computed for all branches, it is then straightforward to calculate the conditional probability of each site pattern given the tree and the model using Felsenstein's (1981a) pruning algorithm. Subsequently, the likelihood function in Equation 1.7 is determined. While maximizing the likelihood function, we estimate the model parameters and the branch lengths. Tree reconstruction with ML selects the trees with the maximum likelihood.

1.1.4.3 Distance-based methods

Distance-based methods use a given distance matrix consisting of pairwise distances between every pair of taxa for tree reconstruction. Typically, we use the genetic distance (Section 1.1.3) estimated for every pair of sequences in the alignment under the maximum likelihood principle (see Equation 1.6). Distance-based methods can be classified into three classes:

Least square (LS) method first defines the *predicted* pairwise distance between two taxa on a tree as the sum of the branch lengths along the path connecting these two taxa. Given a distance matrix and a tree, the LS principle estimates the branch lengths of the tree by minimizing either the sum (Cavalli-Sforza and Edwards, 1967) or the weighted sum (Fitch and Margoliash, 1967), denoted as S_δ, of the squared differences between the given and predicted pairwise distances. Tree search under the *LS criterion* selects the trees with the minimum S_δ.

Minimum evolution (ME) method (Rzhetsky and Nei, 1993) uses the tree length (sum of all branch lengths in the tree) as the criterion to compare trees. For a given tree, the least square principle is commonly employed to estimate the branch lengths. Tree search under the *ME criterion* selects the trees with the minimum tree length.

Clustering methods directly compute the tree and its branch lengths by grouping the

subtrees iteratively (a taxon is a trivial subtree). The subtrees are grouped at each step according to the method's principle (e.g. Sokal and Sneath, 1963; Saitou and Nei, 1987; Gascuel, 1997). For example, the unweighted pair-group method with arithmetic averages (UPGMA) (Sokal and Sneath, 1963) agglomerates, at each step, two rooted subtrees with the smallest distance into a new rooted subtree. The distances between the new subtree and the other subtrees are computed; then, the distance matrix is updated with the dimension reduced by one. The process continues until all taxa are grouped into a single tree.

Clustering methods do not necessitate any objective function being 'globally' optimized. Thereby, they do not have to employ any tree search, thus demanding much less computational resources than methods requiring tree search such as LS and ME. Nonetheless, the neighbor joining (NJ) method (Saitou and Nei, 1987), a clustering method, approximately reconstructs the minimum evolution phylogeny (see Saitou and Nei, 1987; Gascuel, 1997). NJ and one of its variants, BIONJ (Gascuel, 1997), are widely used distance-based methods.

1.1.4.4 Bayesian methods

Bayesian methods (see Huelsenbeck *et al.*, 2001, for a review) do not attempt to search for a single best tree but sample a set of plausible trees. They employ the concept of likelihood (therefore, also assuming a model of sequence evolution) to compute the posterior probability of a tree given the observed data (alignment) using Bayes's theorem (Huelsenbeck *et al.*, 2001; Ronquist *et al.*, 2009). In the end, a list of trees that have the largest posterior probabilities are reported. A consensus tree computed on these trees or the *maximum a posteriori* tree can be viewed as the final phylogeny result of the analysis.

Remarks: Because the number of trees grows exponentially concerning the number of taxa, searching for the global optimum or the best tree is impractical for large numbers of taxa. Heuristic search strategies need to be employed. Many algorithms have been developed resulting in a large number of software programs for phylogeny reconstruction such as PHYLIP (Felsenstein, 1993), PAUP* (Swofford, 2002), PhyML (Guindon and Gascuel, 2003), IQPNNI (Vinh and von Haeseler, 2004; Minh *et al.*, 2005), MEGA4 (Kumar *et al.*, 2008), RAxML (Stamatakis, 2006), and MrBayes (Huelsenbeck and Ronquist,

2001) (see Lemey *et al.*, 2009, for a collection).

As the objective functions to search for best trees are different among the methods, they may reconstruct different trees. It is, therefore, exciting to explore the performance with respect to reconstruction accuracy of the various approaches.

1.2 Model selection

In Section 1.1.3 we have discussed a wide range of models of sequence evolution in terms of model complexity. Model based phylogeny inference depends on the model used. Studies have shown that oversimplified models may result in inconsistency of tree reconstruction methods (e.g. Sullivan and Swofford, 2001). These models are inadequate to describe the data. On the other hand, more complex models require more computational resources and may lead to increased uncertainty in parameter estimates as the same amount of data are used for the estimation. For example, Buckley *et al.* (2001) showed that both the GTR+I+Γ and GTR+Γ models yielded a more accurate estimation of branch lengths than a more parameter rich model, GTR plus site-specific rates with 10 rate categories (i.e. 9 free parameters to model rates across sites heterogeneity compared with 2 in the GTR+I+Γ model or 1 in the GTR+Γ model). In some extreme cases, overparameterized models may lead to nonidentifiable parameters (e.g. Rannala, 2002). Therefore, to select the model that best explains the data with a minimum number of free parameters, the so-called *best-fit* model, is necessary.

A typical way to test whether a more complex model significantly fits the data better than a simpler model is the **likelihood ratio test, LRT** (Navidi *et al.*, 1991; Goldman, 1993b; Huelsenbeck and Rannala, 1997; Frati *et al.*, 1997; Sullivan *et al.*, 1997). The LRT statistic is twice the difference between the two maximum log likelihoods under the compared models:

$$\mathrm{LRT} = 2(\mathrm{L}_1 - \mathrm{L}_0), \tag{1.9}$$

where L_1 is the maximum log likelihood under the more complex model (alternative model) and L_0 is the maximum log likelihood under the simpler model (null model). When the compared models are nested (the simpler model is a special case of the more complex model) and the simpler model is correct, this test statistic asymptotically follows a χ^2 distribution with the number of degrees of freedom equal to the difference in the

number of free parameters between the two models (Navidi *et al.*, 1991; Goldman, 1993b; Whelan *et al.*, 2001, and references therein). When the realized LRT value is significantly large (i.e. the P-value is smaller than 0.05 or 0.01), the null model is rejected. On the contrary, a small LRT value indicates that the alternative model does not describe the data significantly better than the null model and we cannot reject the null model.

Whelan and Goldman (1999) showed cases in which the χ^2 distribution is an inappropriate approximation of the null distribution of the LRT statistic, e.g. when the compared models differ by the inclusion of the Γ-shape parameter to model rates across sites heterogeneity in one model (see Goldman and Whelan, 2000, for explanations). In such a case, the null distribution of the test statistic can be achieved via simulation. First, data are simulated under the stochastic process specified by the null model using the ML estimated parameters based on the original data. This is called parametric bootstrapping and the simulated data are called parametric bootstrap replicates (Felsenstein, 2004, pp. 357-358). The simulated data are then analysed under both models and the test statistic is calculated accordingly. Values of the test statistic from many replicates form the null hypothesis distribution to which the realized test statistic is compared (Goldman, 1993b; Whelan *et al.*, 2001).

To choose the best-fit model from a set of nested models (e.g. the GTR "family" shown in Table 1.3) **hierarchical likelihood ratio tests (hLRTs)** are employed (e.g. Huelsenbeck and Crandall, 1997; Huelsenbeck and Rannala, 1997; Posada and Crandall, 1998). One starts, for example, with the simplest model (JC69), then gradually includes more model parameters until the likelihood does not increase significantly. The last model on the trace is the best-fit model. A detailed diagram of the hLRTs to select the best-fit model is depicted in, e.g., Posada and Crandall (1998) and Posada (2009).

For non-nested models, hLRTs are not applicable. As a remedy, one may use information-based criteria like the **Akaike information criterion AIC** (Akaike, 1974) and the **Bayesian information criterion BIC** (Schwarz, 1978). AIC and BIC assess fit via the maximum log likelihood or the logarithm of the maximum *a posteriori* probability plus penalty for overparameterization. Available software packages such as Model-Test (Posada and Crandall, 1998; Posada, 2008) allow to select the best-fit model from a collection of evolutionary models for a given alignment. It should be noted that the above tests compare the fit (as reflected by the likelihood) of two models to the data but do not provide any statement about how well the favored model explains the data.

The best-fit model selected via model selection is not necessary a model that describes the data adequately. Therefore, it is essential to test the adequacy between model and data even if the best-fit model is used for the inference.

1.3 Notes before drawing conclusions based on the inferred trees

We have introduced essential materials for reconstructing phylogenies from molecular data. Table 1.4 depicts a workflow of phylogeny inference recruiting these materials.

Input:	A set of homologous sequences.
Step 1:	Align the sequences into a multiple sequence alignment.
Step 2:	Determine which reconstruction method(s) to use.
Step 3:	Select the best-fit model from a collection of evolutionary models if the methods assume a model of sequence evolution (MP does not explicitly assume an evolutionary model).
Step 4:	Reconstruct the tree using the methods of choice (and the evolutionary models).
Output:	The inferred tree(s).

Table 1.4: A pipeline for inferring phylogeny from molecular data.

Before drawing conclusions about the evolutionary relationships among the taxa with regards to the inferred trees, several issues should be taken into account:

Quality of the alignment: As sequence alignments are the input for phylogeny reconstruction, their quality, i.e., the accuracy of aligning homologous residues (see Edgar and Batzoglou (2006) for a review) influences the phylogeny result (see e.g. Landan and Graur, 2007). Recently, a few approaches to evaluate the reliability of an alignment have been proposed such as the HoT method (Landan and Graur, 2007) and the GUIDANCE method (Penn *et al.*, 2010). Nevertheless, the alignment is typically taken as given in phylogeny inference except several attempts to

simultaneously align the sequences and reconstruct the phylogeny (e.g. Fleissner *et al.*, 2005; Metzler and Fleissner, 2009; Liu *et al.*, 2009; Löytynoja and Goldman, 2009).

Data support for the branches: Heuristic searches for a best tree employ some rearrangements on trees such as nearest neighbor interchange (NNI) and sub-tree pruning and regrafting (SPR) to explore the tree space. This implies that when the data contain little phylogenetic signal, such an operation (e.g. an NNI move) cannot detect any difference between the two topologies under consideration. Consequently, some branches (internal nodes) of the inferred tree may be poorly supported by the data or even unresolved. The bootstrap technique (Efron, 1979), first introduced to phylogenetics by Felsenstein (1985), is widely used to estimate the reliability of the branches. The inferred tree is then reported together with the so-called *bootstrap support* for its internal branches. The bootstrap support for a branch is the proportion of trees estimated from the bootstrap replicates (obtained by sampling the alignment columns with replacement) that contain this branch. The larger the bootstrap support for a branch is, the better do the data agree on this branch.

Goodness of fit between a model and data shows whether or not the data are adequately explained by the model. Model adequacy has been tested by comparing the maximum likelihood yielded by the examined model with the unconstrained likelihood computed directly from the data (Navidi *et al.*, 1991; Goldman, 1993b; Bollback, 2002). We briefly summarize several tests of model adequacy available and present a novel method for evaluating the goodness of fit in Chapter 3.

Insights into the performance of the method may help avoid wrong interpretation due to reconstruction artefacts such as long branch attraction (Felsenstein, 1978). Performance of phylogeny reconstruction methods can be evaluated under several criteria such as consistency (the ability to estimate the correct tree with sufficient data), efficiency (the ability to quickly converge on the correct phylogeny), and robustness (the ability to infer the correct tree in the presence of model violation). Among these, robustness to model violation may be the most practically important as complete *a priori* knowledge of the evolutionary processes is typically not available. Thus, studies to gain more insights into the robustness of phylogenetic methods against different model violations are encouraged. We outline some main outcomes from previous evaluations of the performance in phylogenetics and re-

port new observations about the robustness of different phylogenetic methods in Chapter 4.

Chapter 2

Intermittent evolution: A new view of sequence evolution

The progress of science requires more than new data; it needs novel frameworks and contexts.
Stephen Jay Gould

2.1 Intermittent evolution

Given a tree and an alignment that evolved along the tree, we define *intermittent evolution* as extra substitution(s) occurring randomly on branch(es) of the tree and thus changing site pattern(s) in the alignment. The introduction of intermittent evolution is motivated by the following question: *How does the alignment change, if an additional substitution on an arbitrary branch of the tree took place (Klaere et al., 2008)?* Such an extra substitution implies a stochastic effect that acts somewhere on the tree and disturbs the observed signal (the alignment).

Klaere *et al.* (2008) showed how to model the impact of a single additional substitution on a multiple sequence alignment for binary character states. To this end, they constructed a so-called *one step mutation (OSM) matrix*, a doubly stochastic matrix. For a tree of n taxa the OSM matrix is a $2^n \times 2^n$-dimension permutation matrix where every non-zero entry describes how the extra substitution changes a site pattern into

another site pattern. In the following, we extend the concept of the one step mutation to nucleotide characters by employing the Kimura three parameter (K3ST) model (Kimura, 1981). We then demonstrate the construction of the OSM matrix for a given tree $\mathcal{T}$.

2.2 Modelling intermittent evolution of nucleotide sequences

2.2.1 Impact of an extra substitution on an alignment site pattern

The K3ST model distinguishes three types (classes) of substitutions as summarized in the following permutation matrices:

$$s_1 = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix} \end{array}, \quad s_2 = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{array}, \quad s_3 = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \end{array}.$$

s_1 describes the transitions within purines (A, G) and pyrimidines (C, T). s_2 represents the transversions within the nucleotide pairs (A, C) and (G, T), and s_3 the remaining transversions within the nucleotide pairs (A, T) and (C, G). Using this model, we now study the effect of an extra substitution on a certain branch of the tree. Consider the rooted two-taxon tree in Figure 2.1a. Assume that the mutation history of the nucleotide on this tree is known: The root state is C and a substitution s_2 occurs on the branch leading to taxon 1. Therefore, we observe the nucleotide A in taxon 1. How will the observed nucleotide change if we introduce an extra substitution, e.g. an s_1 substitution, on the branch leading to taxon 1? We can introduce s_1 either "before" or "after" s_2 occurs. If the extra substitution s_1 occurs before s_2, it changes the root nucleotide C into T. Then T is changed into G by s_2; hence G would be observed in taxon 1 instead of A (Figure 2.1b). If the extra substitution occurs after s_2, it changes the observed nucleotide A also into G (Figure 2.1c). Thus, independent of the order of substitutions, the outcome is always a G in taxon 1. Hence, placing an extra substitution on a branch of the tree results in a unique outcome independent of the unknown substitution history of the observed nucleotide. This is essentially due to the fact that the substitution matrices

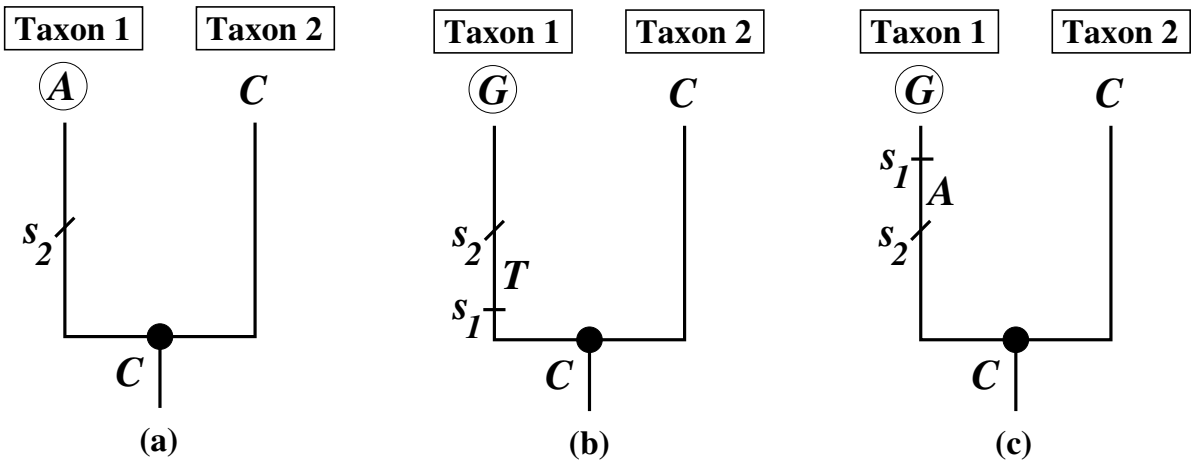


Figure 2.1: Placing an extra substitution on a branch. (a) shows a rooted two-taxon tree, where the mutation history of the nucleotide position is known: a substitution s_2 occurred on the branch leading to taxon 1. An extra substitution s_1 was introduced (b) "before" and (c) "after" the substitution s_2 occurred. Wherever the extra substitution s_1 was placed, the nucleotide observed in taxon 1 is the same.

s_1, s_2, s_3 and the identity matrix s_0 form a commutative group (Klein four-group, see e.g. Humphreys (1996), pp. 26-28) with respect to matrix multiplication. We also note that the K3ST model is the most general model for nucleotide substitution that allows for the formation of such a group.

2.2.2 Formation of the one step mutation (OSM) matrix

The algebraic structure of the K3ST model allows for an analytical way to construct the OSM matrix for a given tree. Figure 2.2 illustrates the connection between the K3ST model and the OSM matrix. For the left branch e_1 of the two-taxon tree (Figure 2.2a), a transition s_1 of the K3ST model (Figure 2.2b) produces a unique 16×16-dimensional (permutation) matrix σ_{s_1,e_1} (Figure 2.2c). The matrix is symmetric, and each row and each column has exactly one non-zero entry (red-horizontal-stripe cell) which describes how a transition on this branch changes a pattern (row) into a new pattern (column). We note that σ_{s_1,e_1} is nothing else but the Kronecker product (see e.g., Horn and Johnson, 1991, pp. 242-243) of the permutation matrix s_1 for the left branch and the identity

matrix s_0 for the right branch:

$$\sigma_{s_1,e_1} = s_1 \otimes s_0.$$

Similarly, for the right branch e_2 the substitution class s_1 generates a permutation matrix σ_{s_1,e_2} shown in Figure 2.2d and we have:

$$\sigma_{s_1,e_2} = s_0 \otimes s_1.$$

For the internal branch e_{12} the substitution class s_1 generates a permutation matrix $\sigma_{s_1,e_{12}}$ shown in Figure 2.2e and

$$\sigma_{s_1,e_{12}} = s_1 \otimes s_1.$$

On the other hand, $\sigma_{s_1,e_1} \cdot \sigma_{s_1,e_2} = (s_1 \otimes s_0) \cdot (s_0 \otimes s_1) = (s_1 \cdot s_0) \otimes (s_0 \cdot s_1) = s_1 \otimes s_1$ because $s_1 \cdot s_0 = s_0 \cdot s_1 = s_1$, where $\cdot$ denotes matrix multiplication. From this we derive:

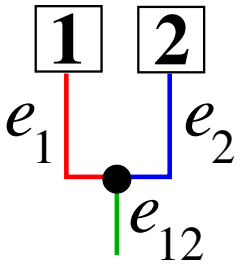
$$\sigma_{s_1,e_{12}} = \sigma_{s_1,e_1} \cdot \sigma_{s_1,e_2}.$$

In other words, for an internal branch the permutation matrix produced by a substitution class $s_i \in \{s_1, s_2, s_3\}$ is the product of the permutation matrices generated by the same substitution class for the descendant branches.

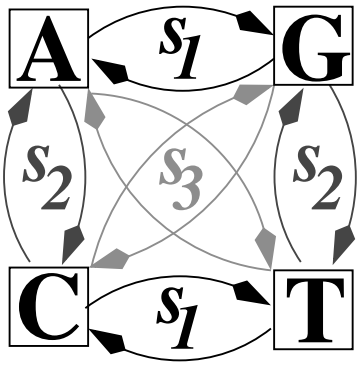
Thus, it is straightforward to construct the permutation matrices for every branch of a tree $\mathcal{T}$. For each of the substitution classes, s_i, we first construct the permutation matrix for every external branch of $\mathcal{T}$ by computing the Kronecker product of the matrix s_i for this branch and the identity matrices s_0 for the other external branches. The order of the terms in the Kronecker product must be the same as the order of the taxa or sequences in the patterns. We then recursively establish the permutation matrices generated by s_i for all internal branches of $\mathcal{T}$ by traversing on the tree from the leaves toward the root. The construction of the OSM matrix $M_{\mathcal{T}}$ for the tree $\mathcal{T}$ is completed by taking into account the relative contribution of each branch in the tree and the probabilities for the three substitution classes for each branch. Thus, we obtain:

$$M_{\mathcal{T}} = \sum_{e \in E} (\alpha_{s_1,e}\sigma_{s_1,e} + \alpha_{s_2,e}\sigma_{s_2,e} + \alpha_{s_3,e}\sigma_{s_3,e})p_e, \tag{2.1}$$

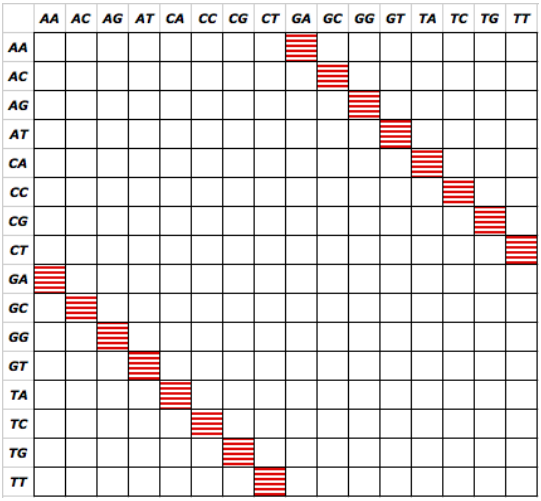
where $\alpha_{s_1,e}, \alpha_{s_2,e}, \alpha_{s_3,e}$ are the probabilities of the three substitution classes for branch e ($\alpha_{s_1,e} + \alpha_{s_2,e} + \alpha_{s_3,e} = 1$), E the set of all branches of $\mathcal{T}$, and p_e the ratio between


1
2
e_1
e_2
e_{12}

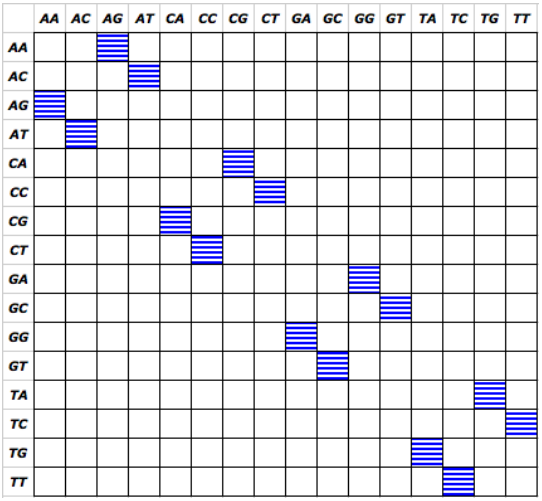
(a)


A
G
C
T
s_1
s_1
s_2
s_2
s_3

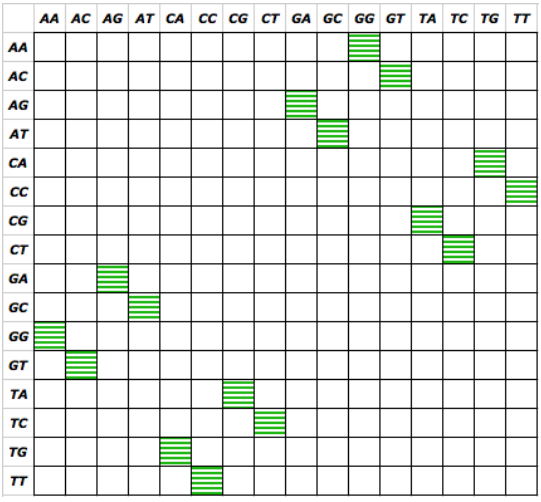
(b)


AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT

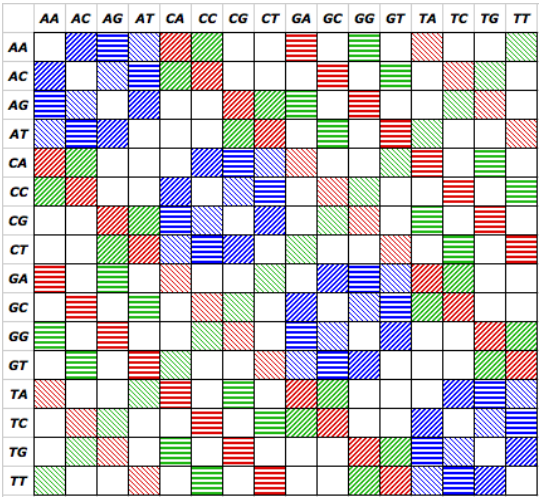
(c)


AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT

(d)


AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT

(e)


AA AC AG AT CA CC CG CT GA GC GG GT TA TC TG TT

(f)

Figure 2.2: (a) A rooted tree with leaves 1 and 2. (b) The K3ST model (Kimura, 1981) distinguishes transitions s_1 (black) and two transversions s_2 and s_3 (gray and light-gray, respectively). A transition s_1 on the left branch e_1 (the red branch) changes a pattern into exactly one new pattern as depicted by the red horizontal stripe cells of the permutation matrix σ_{s_1,e_1} (c). The matrix has 16 rows and 16 columns representing the possible site patterns for the alignment of two nucleotide sequences. The permutation matrices generated by s_1 for the right branch e_2 (blue) and for the branch leading to the "root" e_{12} (green) are displayed in (d) and (e), respectively. The sum of all the permutation matrices generated by all substitution classes for all branches is the unweighted OSM matrix of the tree (including the branch leading to the root in this example) as shown in (f). Horizontal stripe cells ▤ represent the transitions s_1; diagonal stripe ▨ the transversions s_2; and thin reverse diagonal stripe ▧ the transversions s_3. The colors of these cells follow the colors of the branches as in (a) and thus, depicting the branch origin of the substitution.

the length of branch e and the sum of all branch lengths ($p_e \geq 0$ and $\sum_{e \in E} p_e = 1$). p_e is called the relative branch length of e. $M_{\mathcal{T}}$ is the weighted exchangeability matrix for all patterns given that an extra substitution occurs somewhere on the tree $\mathcal{T}$. Figure 2.2f depicts the unweighted OSM matrix (just the sum of all permutation matrices without multiplying by α_e or p_e) for the tree in Figure 2.2a (including the branch leading to the root).

The constructed OSM matrix conveys both the tree and the substitution model to the description of sequence evolution. A positive $M_{\mathcal{T}}(\mathbf{a},\mathbf{b})$ entry indicates that there is a branch e on the tree $\mathcal{T}$ and a substitution class s_i so that a substitution s_i occurring on e changes the pattern $\mathbf{a}$ to the pattern $\mathbf{b}$. In each row and each column of $M_{\mathcal{T}}$, every relative branch length is presented exactly three times, each time together with the probability of one substitution class. Therefore, the sum of each row or each column is always one and the k-th power of $M_{\mathcal{T}}$ produces the probabilities to move from one pattern to another in k substitutions.

Modelling the impact of intermittent evolution on an alignment by the OSM matrix

provides an alternative description of evolutionary processes. Klaere *et al.* (2008) showed how to compute the exact posterior probability of the number of substitutions on a given tree with branch lengths that give rise to an alignment site pattern. They also demonstrated how to derive the parsimony length and the likelihood for a pattern from the OSM matrix. Moreover, the pairwise distance between two sequences can be computed based on the OSM matrix of a tree of two taxa. Thereby, it is possible to use the OSM description of sequence evolution in MP, ML and distance-based tree reconstruction.

In this thesis, it suffices to employ the OSM matrix as in Equation 2.1. We note that the OSM matrix allow us to describe heterogeneous substitution processes along the tree, e.g., the relative substitution rates between nucleotides as reflected by $(\alpha_{s_1,e}, \alpha_{s_2,e}, \alpha_{s_3,e})$ can be different across branches.

Chapter 3

MISFITS: Evaluating the Goodness of Fit between a Phylogenetic Model and an Alignment

Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.
George E. P. Box and Norman R. Draper

3.1 Introduction to goodness of fit tests in phylogeny inference

In recent years, the complexity of models of sequence evolution steadily increased (cf. Swofford *et al.*, 1996; Felsenstein, 2004, see also Section 1.1.3). The general time reversible model allows for the estimation of nucleotide-specific substitution rates (e.g., Yang, 1994a), the assumption of different rates across sites is included (e.g., Yang, 1993, 1994b; Gu *et al.*, 1995) and even heterotachy and change in evolutionary substitution models along a tree can be modeled (e.g., Tuffley and Steel, 1998; Foster, 2004). Moreover, we are able to compute the likelihood of a hypothesis (a model and a tree) by using rapid maximum likelihood tree reconstruction methods (Stamatakis, 2006; Minh *et al.*, 2005; Jobb *et al.*, 2004; Guindon and Gascuel, 2003; Huelsenbeck and Ronquist, 2001). Tools are also available to select the best model from a collection of available

models (Posada and Crandall, 1998; Posada, 2008, see also Section 1.2). Recent surveys (Ripplinger and Sullivan, 2008; Sullivan and Joyce, 2005) indicate that the most complex model is typically selected. The selected model leads to a tree that yields a significantly higher likelihood than trees based on other models do. However, in many instances the selected model fails to explain the data adequately as reflected by large deviations between the observed pattern frequencies and the corresponding expectation. Thus, the next step in a regular phylogenetic analysis would be to evaluate how well the selected model fits the alignment. The easiest such approach is a parametric form of the classical likelihood ratio statistics, where the likelihood under the assumed model is compared to the unconstrained likelihood (Navidi *et al.*, 1991; Goldman, 1993a,b). The unconstrained log likelihood or the multinomial log likelihood sets an upper bound on the log likelihood under any model of sequence evolution and is determined by:

$$\mathrm{L}^{unc} = \ln\left(\prod_{i=1}^{N}\left(\frac{\ell_i}{\ell}\right)^{\ell_i}\right) = \left(\sum_{i=1}^{N}\ell_i \ln(\ell_i)\right) - \ell\,\ln(\ell), \tag{3.1}$$

where N is the number of distinct site patterns in the alignment, ℓ_i the number of alignment sites showing the ith pattern, and ℓ the alignment length ($\sum_{i=1}^{N}\ell_i = \ell$). This equation is applicable to gapless alignments with unambiguous character states. For data with gap characters and ambiguous character states, Waddell (2005) showed a means to compute the unconstrained likelihood.

Goldman (1993b) applied Cox's (1961; 1962) method to test the goodness of fit between a model and an alignment, thereafter this test is referred to as the Goldman-Cox test or shortly the Cox-test. First the ML tree with branch lengths and model parameters are estimated under the examined model. The Goldman-Cox test then computes as the test statistic the observed difference between the unconstrained log likelihood as determined by Equation 3.1 and the maximum log likelihood from the ML inference. A number of parametric bootstrap replicates are generated using the examined model and the ML tree. For each replicate, the difference between the unconstrained log likelihood and the log likelihood under the examined model (both are based on the replicate) is then calculated. This difference represents the expected difference under the null hypothesis of an adequate fit as the tested model was used to generate the data. Thus, the differences obtained from all the replicates form the null distribution to which the observed difference is compared. The P-value is the proportion of the replicates where the difference is larger than or equal to the observed difference. A small P-value (e.g.

less than 5%) indicates that the model cannot explain the data sufficiently, and thus should be rejected. On the other hand, a large P-value implies that the model may describe the data adequately (a schematic workflow to perform this test is provided in Appendix A, Table A.1).

Similarly, Bollback (2002) used the multinomial log likelihood as the test statistic to test model adequacy in a Bayesian framework. First, a Bayesian analysis is conducted under the examined model to provide posterior distribution of tree topologies, branch lengths, and model parameters. Simulated data are generated using the tested model but for each replicate the tree topology, branch lengths, and model parameters are sampled from the corresponding marginal posterior distributions obtained from the Bayesian analysis. The multinomial log likelihoods computed on the simulated data provide the posterior predictive distribution to which the multinomial log likelihood from the original data, i.e. the realized test statistic, is compared. The P-value is interpreted analogously as in the Goldman-Cox test.

The Goldman-Cox test and Bollback's (2002) are two tests of absolute model fit that are accessible to practical analyses. Waddell *et al.* (2009) introduced another method, which has more power, using marginal tests. Nevertheless, all these tests are typically not applied possibly due to the unpleasant outcome that the model and the inferred tree do not explain the alignment very well. However, it has been shown that a careful combination of such tests and then going back to the alignment can help to improve the phylogenetic analysis (Schöniger and von Haeseler, 1999). Unfortunately, this analysis was instance based, and it is not possible to apply it routinely to alignments.

Here we present MISFITS, a novel approach to evaluate the goodness of fit between a phylogenetic model and an alignment. At the same time the method suggests alignment positions that may not fit and a biologically plausible explanation for the deviation.

3.2 The MISFITS method

In a nutshell, MISFITS does the following: Based on the alignment, the substitution model, and the inferred ML tree we compute the conditional probability given the tree and the model (shortly called the likelihood) of the site patterns in the alignment and the corresponding unconstrained likelihood (Navidi *et al.*, 1991; Goldman, 1993a,b). A

confidence region is then computed to determine a set of over-represented patterns and a set of under-represented patterns with respect to the expected number of occurrence. We then apply a maximum parsimony (MP) approach to determine the minimal number of extra substitutions on the ML tree necessary to convert an alignment column that belongs to an over-represented pattern into a pattern that is under-represented in the alignment. The theoretical basis to compute the minimal number of substitutions utilizes the concept of the one step mutation (OSM) matrix (Klaere *et al.*, 2008). Subsequently, a parametric bootstrap analysis is performed to determine whether the number of extra substitutions is significantly elevated. Moreover, the over-represented patterns are mapped back to the alignment to pinpoint to potentially problematic regions in the alignment and to enable a more thorough analysis.

Table 3.1 presents a schematic workflow of MISFITS. We next describe the steps in more detail.

Step 1:	Count the observed frequency of patterns in the alignment.
Step 2:	Compute pattern likelihood under the model and the inferred tree.
Step 3:	Determine the set of over-represented patterns $\mathcal{D}^+$ and the set of under-represented patterns $\mathcal{D}^-$.
Step 4:	For all pairs of patterns $(a, a'), a \in \mathcal{D}^+$ and $a' \in \mathcal{D}^-$, compute the minimal number of extra substitutions to convert a into a'.
Step 5:	Select a matching which pairs every pattern in $\mathcal{D}^+$ with one pattern in $\mathcal{D}^-$ such that the total number of extra substitutions is minimal.
Step 6:	Map the extra substitutions on the tree.
Step 7:	Determine the significance of the number of extra substitutions computed in Step 5 by parametric bootstrap.

Table 3.1: Schematic workflow of the MISFITS method.

Step 1 and 2: Consider a gap free, multiple nucleic acid alignment of n sequences with length ℓ, a nucleotide substitution model and the inferred ML tree. For n taxa, a total of 4^n site patterns are possible. The sites of the alignment constitute a subset

of these patterns. Given the ML tree and the substitution model (thereafter jointly referred to as tree-model), we compute the expected pattern-likelihood vector (p^{tree}) for the patterns in the alignment using, e.g. TREE-PUZZLE (Schmidt *et al.*, 2002), PHYML (Guindon and Gascuel, 2003), IQPNNI (Vinh and von Haeseler, 2004; Minh *et al.*, 2005). The unconstrained likelihood vector (p^{unc}) of the patterns is simply the number of alignment sites showing the pattern divided by the length of the alignment (Navidi *et al.*, 1991; Goldman, 1993a,b). p^{unc} is actually the observed frequency of the patterns in the alignment. Thus, it will be called observed pattern frequency vector.

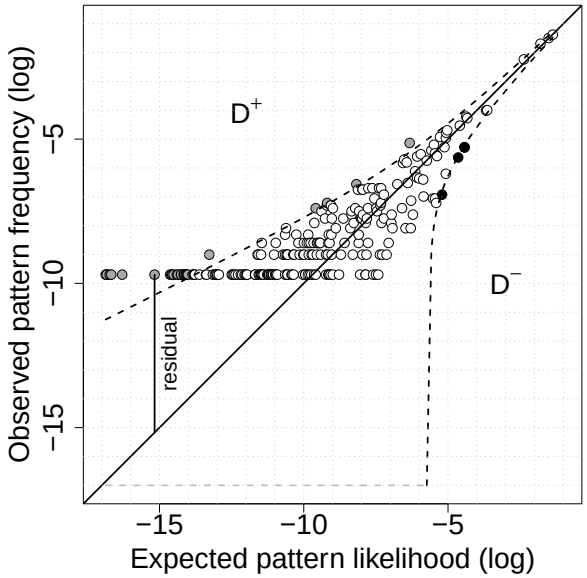


Figure 3.1: Observed pattern frequencies and expected pattern likelihood under the tree-model. Each circle represents a pattern in the alignment by its expected log likelihood under the tree-model (x-axis) and the logarithm of its frequency or unconstrained likelihood (y-axis). The dashed lines indicate the 95% Gold confidence region. The open circles represent patterns within the confidence region; the black-filled circles are under-represented patterns while the gray-filled circles are over-represented patterns.

Step 3: If the tree-model is an adequate description of the data, the difference between the two vectors p^{tree} and p^{unc} should be small. In fact, they are the basis for the Cox-test suggested by Goldman (1993b). Instead of looking at the overall fit, we compare the two vectors position-wise. Figure 3.1 displays a parametric plot of the logarithms of the two likelihood vectors computed on a primate complete mitochondrial genome dataset

under the GTR model. The x-axis displays the logarithm of the entries in p^{tree} and the y-axis the logarithm of the unconstrained likelihood p^{unc}. If the tree-model describes the data adequately, all points will approximately lie on the identity line. However, residuals (deviation from the identity line) are often observed. To account for variability in the data, we compute a simultaneous $\alpha = 95\%$ Gold confidence interval for multivariate proportions (Gold, 1963) for the entries in p^{tree}:

$$\mathrm{CI}(p_{a_i}^{tree}) = p_{a_i}^{tree} \pm \sqrt{\kappa^2 \cdot \frac{p_{a_i}^{tree}(1 - p_{a_i}^{tree})}{\ell}}. \tag{3.2}$$

$\mathrm{CI}(p_{a_i}^{tree})$ is the confidence interval for the likelihood of pattern a_i under the tree-model where $\kappa^2 = \chi_k^2(\alpha/(2\ell))$ is the $1-\alpha/(2\ell)$-quantile of the χ^2 distribution with k degrees of freedom. The number of degrees of freedom in this case is the total number of estimated variables, i.e. the sum of the number of parameters of the evolutionary model (see Section 1.1.3) and the number of branches of the ML tree, i.e $(2n-3)$ for n taxa. The dashed lines in Figure 3.1 show the logarithms of the upper bound and the lower bound of the confidence interval.

We call pattern a_i over-represented if $p_{a_i}^{unc}$ is greater than the upper bound of $\mathrm{CI}(p_{a_i}^{tree})$. If $p_{a_i}^{unc}$ is smaller than the lower bound of $\mathrm{CI}(p_{a_i}^{tree})$, then pattern a_i is under-represented in the alignment. We denote the set containing the over-represented patterns $\mathcal{D}^+$ and the set of the under-represented patterns $\mathcal{D}^-$. $\mathcal{D}^-$ also contains patterns not observed in the alignment, where the pattern likelihood under the tree-model is larger than $1/\ell$. Thus, we would expect to find them at least once in an alignment of length ℓ. These patterns can be easily constructed using the unweighted OSM matrix (Section 2.2.2). Thereby:

$$\mathcal{D}^+ = \{a_i \mid p_{a_i}^{unc} > \text{upper bound of } \mathrm{CI}(p_{a_i}^{tree})\}, \tag{3.3}$$

$$\mathcal{D}^- = \{a_j \mid 0 < p_{a_j}^{unc} < \text{lower bound of } \mathrm{CI}(p_{a_j}^{tree})\} \cup \{\text{unobserved } a_m \mid p_{a_m}^{tree} > 1/\ell\}. \tag{3.4}$$

The over-represented site patterns indicate alignment sites that occur more often than expected under the tree-model. This means the tree-model does not capture these alignment sites adequately. On the other hand, the under-represented patterns are expected to occur more often in the alignment than they actually do. Thus, it appears plausible to compute the minimal number of substitutions that are required to change the over-represented sites in the alignment (site patterns in $\mathcal{D}^+$) into patterns that are more

likely to occur given the ML tree (patterns in $\mathcal{D}^-$). The number of extra substitutions can then be used as a measure to evaluate the goodness of fit of a model to the data: the less the number, the better the fit.

We note that for an over-represented pattern a_i, the number of alignment columns showing this pattern that should be converted into under-represented patterns is determined by $p_{a_i}^{unc} \cdot \ell - \lfloor \text{upper bound of CI}(p_{a_i}^{tree}) \cdot \ell \rceil$, where $\lfloor \cdot \rceil$ denotes the ordinary rounding of a real number (rounding a real number to its closest integer). For an under-represented pattern a_j the total number of alignment columns it can occupy should not exceed $\lfloor \text{upper bound of CI}(p_{a_j}^{tree}) \cdot \ell \rceil$.

Step 4: We now describe how to compute the minimal number of extra substitutions to convert a pattern into another pattern. The mathematical intricacies will be described elsewhere (Klaere S, Nguyen MAT, Fischer M and von Haeseler A, in preparation). For this work it suffices to recapitulate the OSM matrix (Klaere *et al.*, 2008) applied to the K3ST model (Kimura, 1981) (see Section 2.2). The algebraic structure of the K3ST model allows for an efficient way to convert an alignment pattern a into another pattern a' by putting a minimal number of extra substitutions on the tree. In a straightforward approach, one could simply generate all possible patterns from a by putting a number of extra substitutions on branches of the tree until a' is produced. This means one would need to compute the minimal power of the unweighted OSM matrix (excluding α_e and p_e from Equation 2.1) such that the cell corresponding to the two patterns a and a' is non-zero. However, this approach is computationally infeasible for large numbers of taxa. Klaere S, Nguyen MAT, Fischer F and von Haeseler A (in preparation) show that a parsimony algorithm produces the required number of extra substitutions. Here, we discuss an example.

Consider the rooted four-taxon tree in Figure 3.2a and the pattern $GTAA$ at the leaves. Assume that the pattern $GTAA$ is to be converted into $ACAA$. By comparing patterns position-wise, we need a substitution s_1 on the branch leading to taxon 1 to convert G into A at the first position (the first taxon). Similarly, we need a substitution s_1 on the branch leading to taxon 2; no changes are needed for taxa 3 and 4. Thus, a series of substitutions (s_1, s_1, s_0, s_0) on the four external branches of the tree transfers the pattern $GTAA$ into the pattern $ACAA$. Since taxon 1 and taxon 2 form a cluster on the tree and the two substitutions are from the same matrix s_1, they are equivalent to a substitution s_1 on the corresponding internal branch. As shown before, the outcome

is independent of the order of the extra substitutions to other substitutions; therefore, the extra substitution s_1 on the internal branch is enough to switch the pattern $GTAA$ into the pattern $ACAA$.

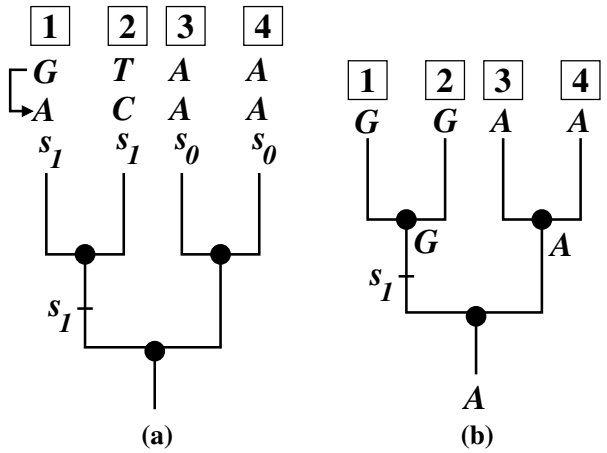


Figure 3.2: Exchanging two patterns on the tree. (a) displays a rooted four-taxon tree with pattern $GTAA$ observed at the leaves. If we want to convert the pattern $GTAA$ into $ACAA$, we may introduce a series of substitutions (s_1, s_1, s_0, s_0) to the 4 external branches. Under the K3ST model and the OSM setting (Klaere *et al.*, 2008), this series is equivalent to an "extra substitution" s_1 on the internal branch leading to taxa 1 and 2 as the two taxa form a cluster on the tree. Therefore, the one extra substitution is enough to switch the observed pattern $GTAA$ into $ACAA$ regardless of the substitution process the pattern $GTAA$ has undergone. On the other hand, the above substitution series converts a constant pattern $AAAA$ into a unique pattern $GGAA$. Hence, converting the pattern $GTAA$ into $ACAA$ is equivalent to evolving the ancestral character A along the tree such that pattern $GGAA$ is obtained at the leaves. Applying the Fitch algorithm (Fitch, 1971) to the latter results in a unique assignment in (b): an s_1 substitution, which changes A into G, occurs on the internal branch leading to taxa 1 and 2. This assignment is identical to the assignment in (a).

On the other hand, we have shown in Section 2.2 that the series of substitutions (s_1, s_1, s_0, s_0) can also act on any other pattern to produce another unique pattern. Applying this series of substitutions on a constant pattern, $AAAA$, leads to the pattern $GGAA$. Therefore, converting the pattern $GTAA$ into the pattern $ACAA$ is equivalent

to converting the constant pattern $AAAA$ into the pattern $GGAA$. Hence, computing the minimal number of substitutions to change the pattern $GTAA$ into the pattern $ACAA$ is equivalent to computing the minimal number of character changes required along the tree to explain the pattern $GGAA$ observed at the leaves given that the root state is A. The latter can be efficiently computed using the first part of Fitch algorithm (Fitch, 1971, see also Section 1.1.4.1) with the extension that if the root character set does not contain A, we increase the number of character changes by 1. The Fitch algorithm also assigns the substitutions on the branches of the tree. In our example, this results in an unique assignment in Figure 3.2b, which agrees with the assignment in Figure 3.2a.

Step 5: After computing the number of substitutions to convert each pattern in $\mathcal{D}^+$ into every pattern in $\mathcal{D}^-$, we determine a matching which pairs every pattern in $\mathcal{D}^+$. This is done by applying the Munkres algorithm for assignment and transportation problems (Munkres, 1957). The minimal number of substitutions, thereafter referred to as the number of extra-substitions and denoted as m, is then considered as the minimal "cost" to fit the tree-model to the observed data.

Step 6: Subsequently, we apply the second part of Fitch algorithm (Fitch, 1971) to assign extra substitutions to the branches of the tree to exchange the paired patterns between $\mathcal{D}^+$ and $\mathcal{D}^-$.

Step 7: Finally, we assess the significance of the number of extra substitutions using parametric bootstrap. We generate a number of alignments (e.g. 1,000 alignments) on the tree under the substitution model with the respective parameter values using a sequence generator program such as SEQ-GEN (Rambaut and Grassly, 1997). We then re-estimate the tree and compute the number of extra substitutions for each simulated alignment. Subsequently, we determine whether the number of extra substitutions computed on the original alignment (m_0) is significantly large according to a given significance level (5%). It should be noted that if m_0 is close the critical value (5% point), one should increase the number of simulated alignments for a more accurate estimation of the P-value.

3.3 Results

3.3.1 Artificial alignments

We used artificial alignments to investigate the positions of over-represented patterns MISFITS recognizes. Two cases were simulated.

3.3.1.1 Alignments containing sites with random nucleotides

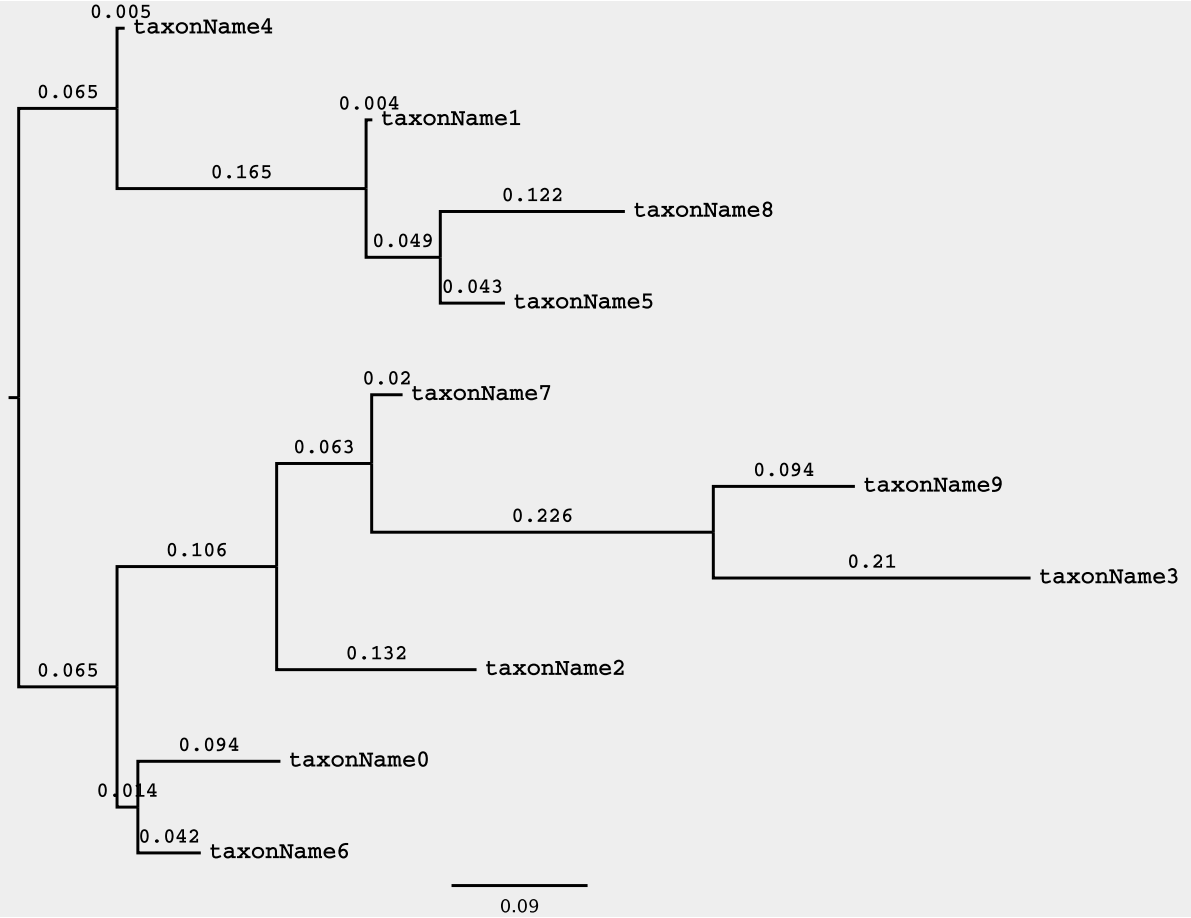


Figure 3.3: The ten-taxon tree used to generate artificial alignments.

The artificial alignments contain 10 sequences of length $\ell = 10,000$. Site patterns in the alignment are partitioned into two regions:

(i) Region of random nucleotides, i.e. the nucleotides in each site are randomly drawn from the uniform distribution. The length of this region, denoted as ℓ_R, varies

from 1% to 50% of the alignment length (1, 10, 20, 30, 40, and 50%).

(ii) Region of site patterns simulated under the JC69 model (i.e. the true model) along the ten-taxon tree shown in Figure 3.3.

For each ℓ_R, we generated 100 alignments and then reconstructed the ML tree using the JC69 model. Subsequently, we investigated the location of the over-represented patterns reported by MISFITS.

Figures 3.4a-c display the results. Firstly, we observed that the estimated number of over-represented sites (the box plots in Figure 3.4a) is a bit larger than the number of sites containing random nucleotides as the line connecting the mean of these boxes is constantly above the identity line. Secondly, MISFITS recognizes from 90% to approximately 100% of the sites containing random nucleotides as over-represented site patterns (Figure 3.4b). Lastly, the percentage of over-represented sites located at Region (i) increases to more than 80% as the length of this region increases to 50% of the alignment length (Figure 3.4c). Thus, MISFITS detects random alignment positions.

3.3.1.2 Alignments containing sites from different models

The artificial alignments are generated along the ten-taxon tree (Figure 3.3) with length $\ell = 10,000$. The alignment is partitioned into two regions, where each region evolves according to a different model.

(i) Region (i) evolves according to the GTR+I+Γ with relative rates (5.0, 10.0, 3.0, 1.0, 15.0, 1.0). The nucleotide frequencies for A, C, G, T are 0.30, 0.25, 0.15, and 0.30, respectively. The proportion of invariant sites is 0.5 and the Γ-shape parameter α is 0.5. The length of this region, denoted as ℓ_G ("guest" model), varies from 1% to 50% of the alignment length (1, 10, 20, 30, 40, and 50%).

(ii) Region of site patterns simulated under the JC69 model, i.e. the "true" model.

For each ℓ_G, we generated 100 alignments and then reconstructed the ML tree under the true model. Subsequently, we investigated the location of the over-represented patterns reported by MISFITS.

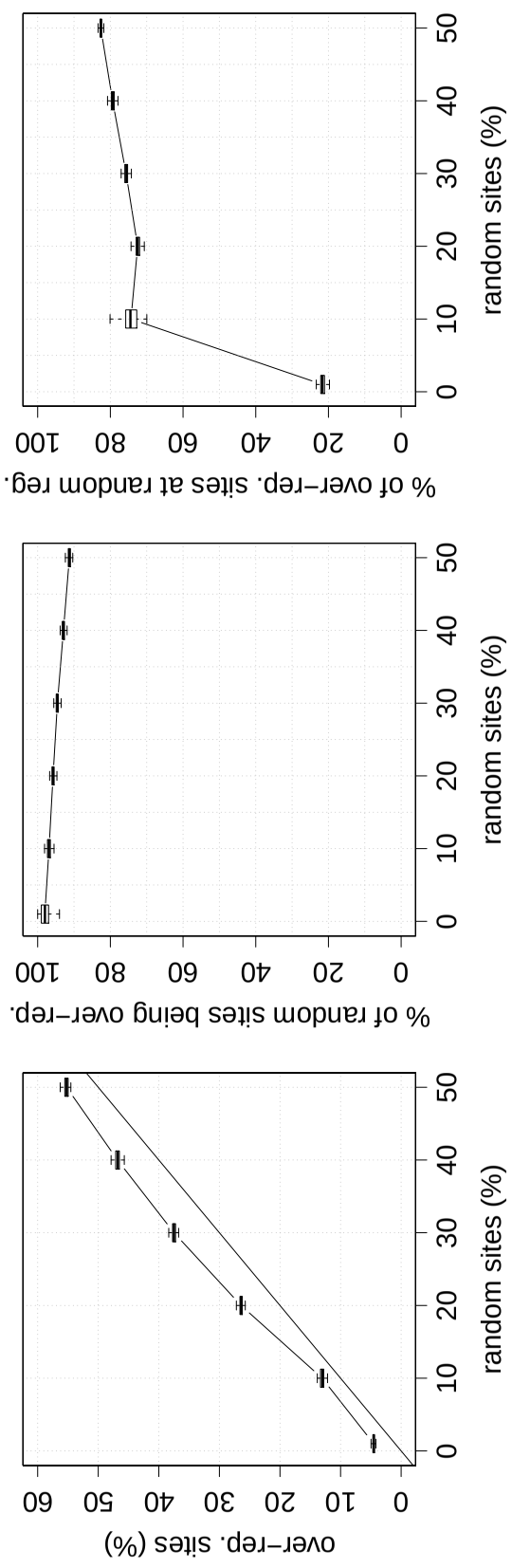
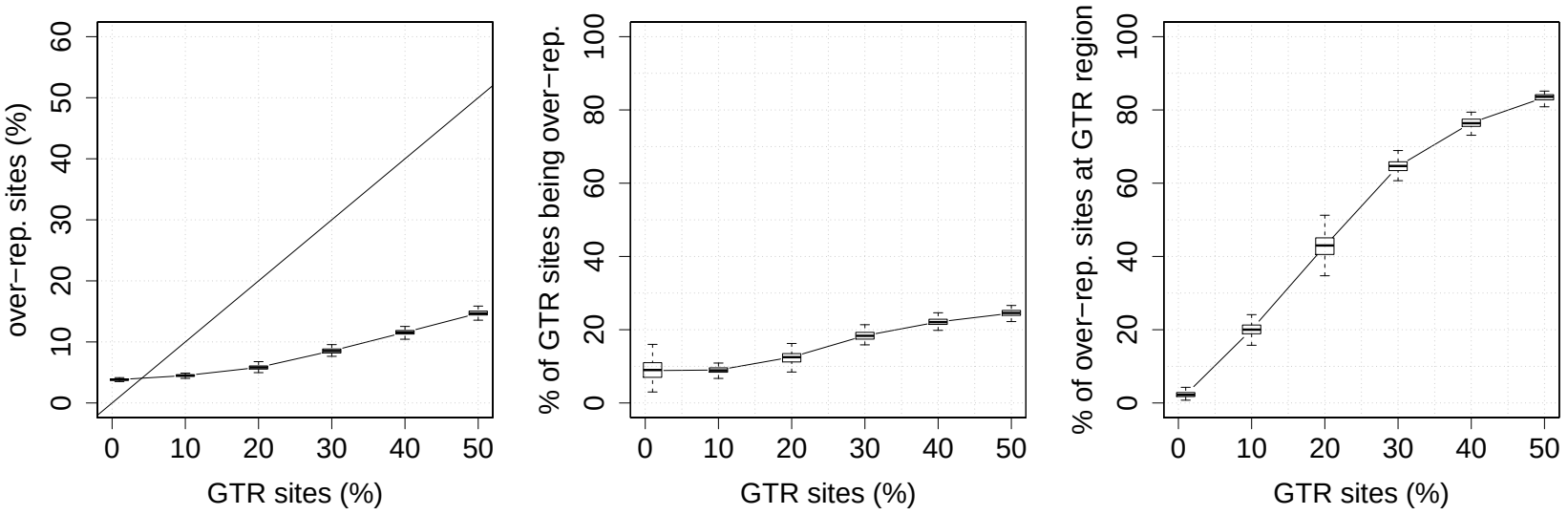


Figure 3.4: MISFITS recognizes sites of random nucleotides in artificial alignments as over-represented patterns. The x-axis shows the length (in percentage of the alignment length) of the region containing alignment sites with random nucleotides, referred to as random sites/region. The remaining sites in the alignments are generated along a ten-taxon tree using the JC69 model. The ML trees are then reconstructed using the JC69 model. The y-axis shows (a) the percentage of the alignment sites which are recognized as over-presented sites, (b) the percentage of the random sites being recognized as over-presented and (c) the percentage of the over-represented sites located at the random region.



(a) (b) (c)

Figure 3.5: MISFITS recognizes alignment sites evolving under a different (guest) model as over-represented patterns. The x-axis shows the length (in percentage of the alignment length) of the region containing alignment sites evolving under the guest model GTR+I+Γ, referred to as GTR sites/region for short. The remaining sites in the alignments are generated according to the JC69 model, i.e. the true model. The ML trees are then reconstructed using the true model. The y-axis shows (a) the percentage of the alignment sites which are recognized as over-represented sites, (b) the percentage of the GTR sites being recognized as over-represented and (c) the percentage of over-represented sites located at the GTR region.

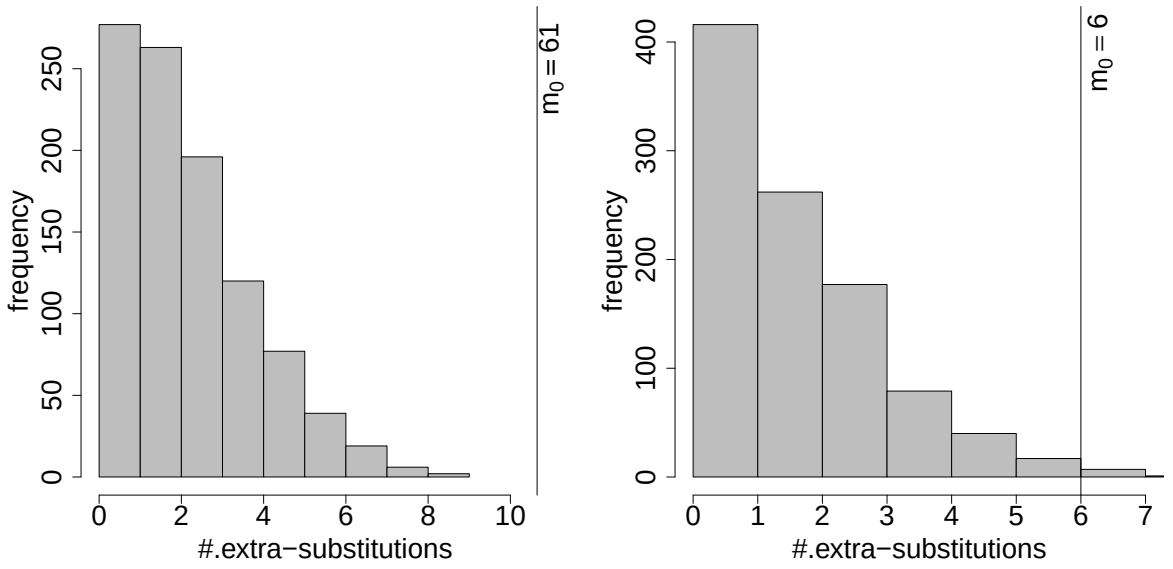
Figures 3.5a-c display the results. Firstly, the inferred number of over-represented sites is much less than (namely, about one third of) the number of sites evolving under the GTR+I+Γ model (Figure 3.5a). Therefore, the percentage of the GTR+I+Γ sites being recognized as over-represented sites is also small. Nevertheless, this percentage increases as the number of GTR+I+Γ sites increases (Figure 3.5b). Moreover, similar to the above study, the percentage of over-represented sites located at the GTR+I+Γ region also increases to more than 80% as the length of this region increases to 50% of the alignment length (Figure 3.5c).

The two studies demonstrate that MISFITS detects alignment positions that may not fit to the model and the tree, especially when some alignment sites consist of random nucleotides.

3.3.2 Primate mitochondrion, complete genome

The data set under consideration contains the complete mitochondrial DNA from five primates: chimpanzee, bonobo, human, gorilla and orangutan (Horai *et al.*, 1995). The alignment, after removing sites containing gaps, is 16,271 bp long and is composed of 241 distinct patterns. As discussed earlier, Figure 3.1 shows the logarithm of the observed pattern frequencies and the expected pattern likelihood under the GTR model. We counted 207 patterns within the confidence region (open circles), 30 over-represented patterns (gray-filled circles) and four under-represented ones (black-filled circles). Using the OSM matrix (Klaere *et al.*, 2008), we generated 94 patterns one substitution away from the constant patterns, twelve of which are not observed in the alignment but are all expected to occur at least once. The average likelihood of these 12 patterns is $1.09 \cdot 10^{-4}$ while the average likelihood of the 25 over-represented patterns, each occurring once in the alignment, is $5.28 \cdot 10^{-7}$. Thus, the unobserved one-substitution patterns are on average 207.5 times more likely to occur in the alignment. The inferred ML tree is rooted at the external branch leading to the orangutan and the resulting number of extra substitutions was 61. This was excessively high compared to the simulated null hypothesis distribution (Figure 3.6a).

We then included invariant sites (I) and Γ−rate heterogeneity into the GTR model and also examined a simpler model, JC69+I+Γ. Under JC69+I+Γ, the number of extra substitutions on the original alignment (m_0) was 2,168 and it was way out of the range of



(a) (b)

Figure 3.6: Primate complete mitochondrial genome. Histogram of the number of extra substitutions computed on 1,000 generated alignments under (a) GTR and (b) GTR+I+Γ models. The attained value (m_0) was significantly high under both models.

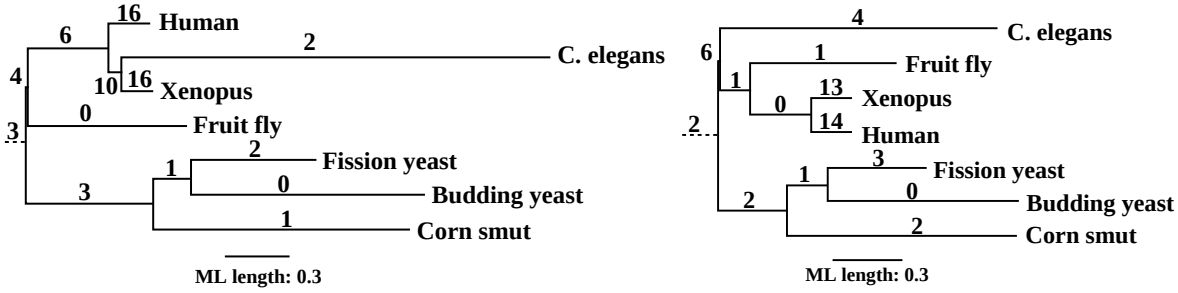
the simulated null hypothesis distribution (data not shown). Remarkably, m_0 estimated under GTR+I+Γ, though very low ($m_0 = 6$), was still significantly high (P-value = 0.002 based on 1,000 simulated alignments, Figure 3.6b). This demonstrates the power of our approach in terms of rejecting models that do not really fit the data.

This study involves a simple model, JC69+I+Γ, and the more complex ones, GTR and GTR+I+Γ. Nevertheless, these models are rejected by the Goldman-Cox test (data not shown) as well as under our approach. Thus, there might be factors in the process of evolution that even the complicated model GTR+I+Γ was unable to cover. A closer look at the data revealed four over-represented site patterns. Two of these are located at genes *ND1* and *ND5*, both at the third codon position: position 747 in the human *ND1* gene and position 981 in the human *ND5* gene (positions 4,053 and 13,317, respectively, in the human mitochondrial genome). The other two are located at the D-loop at positions 151 and 16,293 in the human mitochondrial genome. Moreover, it should be noted that the test of homogeneity of the substitution process on a phylogeny advocated by Weiss and von Haeseler (2003) also rejected GTR and GTR+Γ on this data set. It

implied that there may be heterogeneous substitution processes within the phylogeny describing the data.

3.3.3 Fungi, metazoa CDC45-like region

The protein coding DNA alignment PF02724 was taken from the PANDIT database (Whelan *et al.*, 2006). It encodes the *CDC45*-like protein. The sequences are homologs, as studied by Saha *et al.* (1998), from 7 fungi, metazoa species: *Ustilago maydis* (Corn smut), *Saccharomyces cerevisiae* (Budding yeast), *Schizosaccharomyces pombe* (Fission yeast), *Caenorhabditis elegans* (C. elegans), *Drosophila melanogaster* (Fruit fly), *Xenopus laevis* (Xenopus) and *Homo sapiens* (Human). After removing sites containing gaps, 1,503 sites remain.



(a) (b)

Figure 3.7: Fungi, metazoa *CDC45*-like region. Figures (a) and (b) present the maximum likelihood trees reconstructed under GTR+I+Γ and JC69+I+Γ, respectively. Branch lengths are scaled according to the ML estimation. The number above each branch is the number of extra substitutions assigned by MISFITS.

Model testing under AIC suggested the GTR+I+Γ model. However, the inferred tree failed to recover the generally accepted taxonomic groupings (Figure 3.7a). The internal branch leading to one of the inappropriate groupings (Xenopus, C. elegans) is weakly supported by 31%. Remarkably, the tree inferred by a simpler model, JC69+I+Γ, was congruent with the generally accepted phylogeny (Figure 3.7b). Moreover, this tree needs 15 extra substitutions less than the tree in Figure 3.7a. Figures 3.7a-b also display the assignment of the extra substitutions on the trees using accelerated transformation, ACCTRAN (Farris, 1970; Swofford and Maddison, 1987). The number above each branch

shows the number of extra substitutions. Branch lengths are scaled according to the number of substitutions per site under the ML estimation. The root was placed on the branch separating the fungal species from the metazoa.

		From 1,000 bootstraps			
Branch leads to	mb_0	min	max	mean	P-value
Budding yeast (BY)	0	0	6	0.529	1.000
Fission yeast (FY)	3	0	7	0.778	0.076
(BY, FY)	1	0	7	0.546	0.366
Corn smut (S)	2	0	5	0.393	0.079
(BY, FY, S)	2	0	5	0.450	0.090
Human (H)	14	0	46	5.812	0.109
Xenopus (X)	13	0	33	5.912	0.132
(H, X)	0	0	8	0.587	1.000
Fruit fly (F)	1	0	11	1.204	0.560
(H, X, F)	1	0	10	1.181	0.599
C. elegans (E)	4	0	12	1.815	0.153
(H, X, F, E)	6	0	16	1.818	0.065
Root	2	0	9	1.458	0.358

Table 3.2: Number of extra substitutions assigned to the branches of the tree inferred by JC69+I+Γ for the alignment of *CDC45*-like region (PF02724). mb_0 is the number of extra substitutions assigned to each branch of the tree, computed on the original alignment. The P-value is the proportion of the number of parametric bootstrapped alignments where the number of extra substitutions assigned to a certain branch was greater or equal to that computed on the original alignment.

Notably, we observed the tendency to place extra substitutions on short branches, for instance, on the two external branches leading to Human and Xenopus. A reason may be that substitutions on short branches are rarely captured by the ML model. They are then accounted for by our approach as extra substitutions. We therefore studied the significance of the number of extra substitutions assigned to the branches of the tree under JC69+I+Γ. We generated 1,000 alignments, used them to re-estimate the

branch lengths and then computed the number of extra substitutions on the branches of this tree. Table 3.2 displays the results. The number of assigned extra substitutions on all branches of the tree, including the two external branches leading to Human and Xenopus, are not significantly high (significance level $\alpha = 0.05$).

Highlighting the over-represented positions in the alignment, we observed most of them at the third codon position: 88.6% under GTR+I+Γ and 87.5% under JC69+I+Γ. This is congruent with the fastest evolutionary rate of the nucleotides at the third codon position (Swofford *et al.*, 1996; Rodríguez-Trelles *et al.*, 2006; Bofkin and Goldman, 2007).

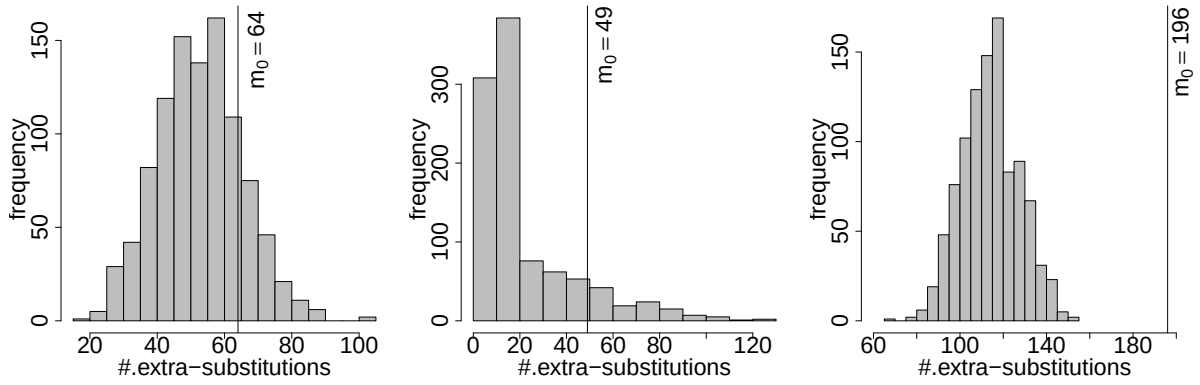


Figure 3.8: Fungi, metazoa *CDC45*-like region. Histogram of the number of extra substitutions computed on 1,000 generated alignments under (a) GTR+I+Γ, (b) JC69+I+Γ and (c) GTR models. The attained value (m_0) falls in the null hypothesis distribution (not significant) under GTR+I+Γ and JC69+I+Γ. It is significantly high under GTR though (c).

Finally, we studied the significance of the number of extra substitutions on the trees under the above two models and GTR (1,000 alignments were generated for each model). Under JC69+I+Γ and GTR+I+Γ, the number of extra substitutions ($m_0 = 49$ and 64, respectively) fell in the corresponding simulated null hypothesis distribution (no significance). However, $m_0 = 196$ under GTR was way too high (see Figure 3.8). It implies that models without rate heterogeneity across sites would be inadequate for this data set.

Most importantly, this alignment demonstrates a case in which a simpler model,

JC69+I+Γ, performed better than a more complex one, GTR+I+Γ, with regard to the inferred trees (c.f. Sullivan and Swofford, 2001) and to the number of extra substitutions. Thus, MISFITS is capable of indicating such a situation.

3.3.4 Study on a large range of data

We applied MISFITS to study a wide range of multiple alignments of protein-coding DNA sequences from the PANDIT database, release 17 (Whelan *et al.*, 2006). The PANDIT database contains 7,738 alignments in total. Alignments with less than four sequences (1,247 alignments) were discarded from our analysis as the tree space (only one unrooted topology) and the pattern space (not more than 64) are too small for a typical phylogenetic analysis. Alignments with more than 100 sequences (320 alignments) were also discarded because the gapless alignment lengths are too short: the average of gapless alignment sites per taxon (alignment length divided by number of sequences) is 1.23. Alignments with short sequence length and large number of taxa may lead to a bias in phylogeny inference (Revell *et al.*, 2005). Thus, the study involved 6,171 alignments containing from 4 to 100 sequences with gapless alignment length ranges from 15 to 6,288 bp.

First, we used jModelTest (Posada, 2008) to select the best model for each alignment using the AIC criterion (Akaike, 1974). Under the selected model, the ML tree and pattern likelihood were computed by using the PHYML package (Guindon and Gascuel, 2003) and the TREE-PUZZLE program (Schmidt *et al.*, 2002), respectively. We observed that the GTR models with and without rate heterogeneity across sites (GTR, GTR+I, GTR+Γ (4 rate categories), GTR+I+Γ) were mostly selected (70.65%). Furthermore, models with one rate across sites were rarely selected (only 1.54%, see Table 3.3).

Subsequently, we studied the goodness of fit of the selected models to the alignments. For 777 alignments (12.59%) the observed frequencies of all patterns are within the confidence region. The number of sequences in these alignments ranges between four and eight. The number of extra substitutions needed for these 777 alignments is 0. For alignments with more than eight sequences, $\mathcal{D}^+$ or $\mathcal{D}^-$ are never empty.

We observed over-represented patterns in the remaining 5,394 alignments. There were 98 alignments in which all patterns are over-represented. Thus, not a single pattern fell into the confidence region. This is attributed to the fact that they contain only singleton

	Rate				
Model[a]	One rate	I	Γ	I + Γ	$\sum_{Model}$
JC	0.05	0.02	0.02	0.00	0.08
F81	0.15	0.16	0.18	0.10	0.58
K80	0.03	0.29	0.58	0.21	1.12
HKY	0.29	1.39	3.34	2.85	7.88
K3ST	0.02	0.13	0.28	0.13	0.55
K3STuf	0.16	0.92	1.70	1.70	4.49
TN93ef	0.03	0.15	0.13	0.19	0.50
TN93	0.19	0.79	1.59	1.81	4.39
SYM	0.13	0.39	3.21	6.03	9.76
GTR	0.49	3.68	26.06	40.43	70.65
$\sum_{Rate}$	1.54	7.92	37.09	53.45	100.00

Table 3.3: Percentages (%) of the selected models for 6,171 alignments in the PANDIT database. [a] refer to Table 1.3 for a detailed description of the models.

patterns (occurring only once in the alignment). Phylogenies based on such alignments with tremendously diverse patterns are probably arbitrary. Therefore, we discarded these alignments from the next steps.

The next step of MISFITS thus comprised 5,296 alignments. However, for 1,028 alignments (19,41%), there were not enough unobserved patterns having a likelihood greater than $1/\ell$, i.e. not enough under-represented patterns, to exchange for all the over-represented patterns. These alignments were also discarded.

Thus, 4,268 alignments went into the final analysis. The percentages of the models being selected for these alignments were similar to those in Table 3.3 (see Table A.2). Thus, the removal of the above alignments did not change the model selection substantially. On average, MISFITS introduced 13.73 extra substitutions per 100 characters (number of extra substitutions per site divided by the number of sequences in the alignment times 100). Figure 3.9 shows the histogram of the number of alignments against the number of extra substitutions per 100 characters.

Based on the parametric bootstrap analysis consisting of 100 simulations for each of

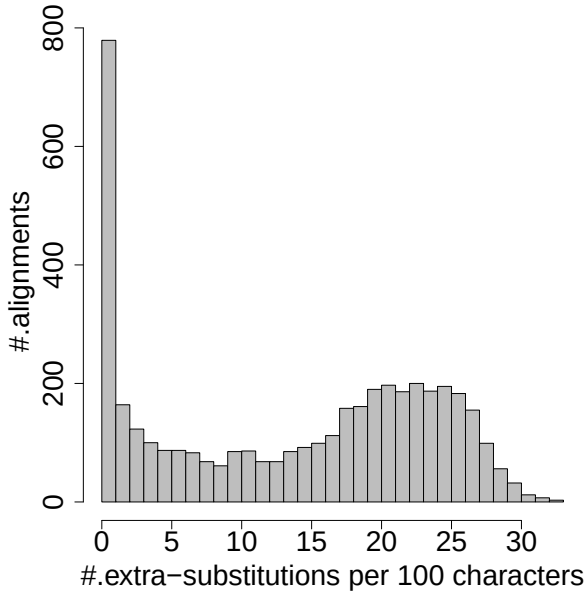
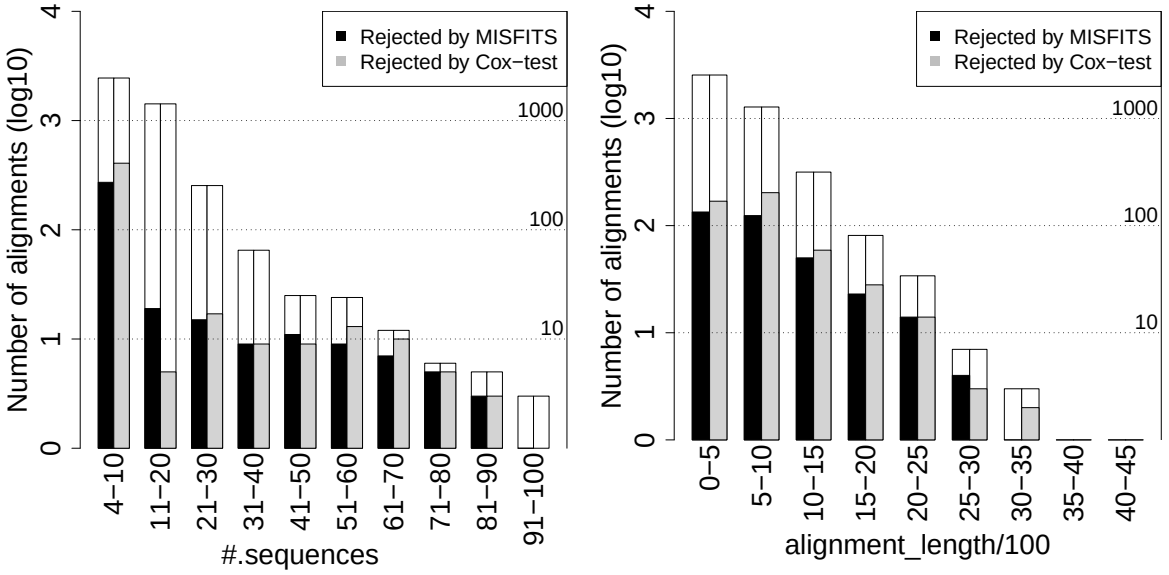


Figure 3.9: Results on PANDIT database under the selected models for the 4,268 alignments where over-represented patterns were observed and there were enough under-represented patterns to exchange with them. The histogram displays the number of alignments (the y-axis) versus the number of extra substitutions per 100 characters (the x-axis).

the 4,268 alignments, MISFITS showed that the number of assigned extra substitutions was not significant for 3,918 alignments (91.80%) and significantly high for 350 alignments (8.20%). This means our approach would reject 350 models. The Goldman-Cox test rejected 478 models (11.20%), which is in the same order of magnitude. Two-hundred and seventeen models were rejected by both approaches. Figures 3.10a-b display the number of alignments (the height indicated by the non-filled bars) and the number of alignments (models) being rejected by MISFITS (black bars) and by the Goldman-Cox test (gray bars) with respect to the number of sequences in the alignment (a) and to the alignment length (b). These figures (see also Figures A.1 and A.2) show that the proportion of models being rejected by both methods tends to increase when the number of sequences grows as well as when the alignment length becomes longer. This implies that it becomes more and more difficult to have a single model that can adequately explain the data.

We learned from this survey that in a number of instances (8.20%), the selected models and the resulting trees do not really fit the data. Moreover, typically singleton



(a) (b)

Figure 3.10: Results on PANDIT database under the selected models for the 4,268 alignments where over-represented patterns were observed and there were enough under-represented patterns to exchange with them. The height indicated by the non-filled bars display the number of alignments in logarithm to base 10 (the corresponding decimal values are depicted by the dashed horizontal lines together with the numbers on the right). The filled bars show the number of alignments (models) being rejected by MISFITS (black bars) and by the Goldman-Cox test (gray bars) with respect to the number of sequences in the alignment (a) and to the alignment length (b).

patterns are over-represented. One reason for this is the discrete nature of the patterns. Occasionally, some patterns have a very small likelihood to occur on the inferred tree. Therefore, it is more plausible to explain the occurrence of such a pattern by extra-subsitutions, which are not covered by the model but are more likely to happen on the tree.

3.4 Discussion

MISFITS provides a guided efficient way to pinpoint to site patterns in the alignment which are not captured well by the substitution model and the inferred tree. The differ-

ences (residuals) between their observed frequencies and the corresponding expectation manifest themselves in a clear deviation from the identity line (c.f. Figure 3.1). We then introduced a computationally feasible method which puts extra substitutions on the tree to explain the residuals. The extra substitutions reduce over-represented site patterns in the alignment and at the same time increase under-represented patterns. This has the ultimate effect that these extra substitutions pull over-represented patterns and simultaneously push under-represented patterns into the confidence region.

A big advantage of the approach is the possibility to map the extra substitutions on the tree. Moreover, the extra substitutions required give a biological interpretation why the data may not be adequately described by the tree-model. The reasons for significant deviations, however, may be different for every single instance. They depend on the selected sequences, the selected organisms and the unknown evolutionary history of the sequences. This needs to be elucidated on a case-by-case basis, and our tool points to potential regions in the alignment that may deserve a closer analysis. On the other hand, the assignment of extra substitutions to branches of the tree provides additional information concerning the interpretation of the inferred phylogeny: the branches of the tree where such extra substitutions would help to justify the residuals.

The approach we suggest also sheds additional light on the goodness of fit in model testing approaches that are discussed, for example, by Goldman (1993b); Posada (2008). It even may point to the risk of overfitting the data that may lead to biologically implausible results such as in the fungi, metazoa *CDC45*-like region example.

From the computational point of view, it is practical in terms of running time to apply MISFITS routinely to alignments. Firstly, the computational complexity required to find the number of substitutions that change a pattern in $\mathcal{D}^+$ into a pattern in $\mathcal{D}^-$ is indeed the complexity of the preliminary phase in the Fitch algorithm, i.e. $O(n)$, where n is the number of sequences (Fitch, 1971). Thus, computing the number of substitutions for every pair of patterns between $\mathcal{D}^+$ and $\mathcal{D}^-$ has complexity $O(n \cdot |\mathcal{D}^+| \cdot |\mathcal{D}^-|)$, where $|.|$ denotes the cardinality of a set. Secondly, finding an optimal matching between patterns in $\mathcal{D}^+$ and $\mathcal{D}^-$ such that the total number of extra substitutions is minimal according to the Munkres algorithm runs in the worst-case time $O(V^3)$, where $V = max\{|\mathcal{D}^+|, |\mathcal{D}^-|\}$. Moreover, the number of patterns in $\mathcal{D}^+$ is not larger than the number of distinct patterns observed in the alignment. The number of distinct patterns in the alignment cannot exceed both the alignment length and the total number of possible patterns (4^n).

Hence, $|\mathcal{D}^+| \leq M = min\{\ell, 4^n\}$. The cardinality of $\mathcal{D}^-$ is in the same order of magnitude with the cardinality of $\mathcal{D}^+$, as $\mathcal{D}^-$ contains patterns whose likelihood under the tree-model is larger than $1/\ell$. Therefore, in the worst case where all alignment sites are distinct and over-represented, computing the number of substitutions for every pair of patterns and then finding the optimal matching between $\mathcal{D}^+$ and $\mathcal{D}^-$ requires $O(nM^2)$ and $O(M^3)$ complexity, respectively. Nevertheless, while studying alignments with a large number of sequences from the PANDIT database we never observed 4^n patterns in the alignment. On average the computation of m_0 for one of the 4,268 alignments going through all analysis steps took 8.4 seconds on a single core of a 3-year old dual core AMD Opteron CPU 2220 SE. A more detailed depiction of the computing time with respect to the sequence length and number of sequences is given in Figure A.3.

We have discussed so far the biological implications and the computational complexity MISFITS may cause. It should be noted that there is also room for methodological extensions. For example, different locations of the root on the tree may result in different numbers of extra substitutions as there is a constraint about the character state at the root while employing the Fitch algorithm in our approach. It is feasible to implement an exhaustive or heuristic search for the location of the root which gives the minimal number of extra substitutions. However, it is more useful to provide a biologically meaningful rooting based on preliminary knowledge about the data.

One limitation of our approach is the restriction to the K3ST model for nucleotide characters. For more complex models of nucleotide evolution, the algebra does not work (see Section 2.2). Nevertheless, this is not a true drawback of the method, since the method is applied after tree reconstruction and model selection. If we have by statistical standards the best model selected, then it is pointless to have a second model that is again complex. We simply want to know where we still observe deviations; hence, MISFITS is a final step to find significant deviations.

It will be interesting to see how the phylogeny changes if we systematically introduce additional signals into the alignment. We may put a number of extra substitutions on several branches of the tree to change a number of patterns in the alignment accordingly. Each extra substitution will be placed on one branch and will change one site in the alignment. Thus, the sample (pattern) space varies in a controlled manner and the impact of such a variation on phylogeny reconstruction can be observed easily.

Chapter 4

ImOSM: Imbedding of Intermittent Evolution and Robustness of Phylogenetic Methods

Without deviation, progress is not possible.
Frank Zappa

4.1 Introduction

4.1.1 Performance evaluation in phylogeny inference

Phylogeny reconstruction implies inference of the tree topology, branch lengths, and model parameters for model-based methods. Thus, statistical criteria to evaluate the performance of inference methods in general can be used to assess phylogeny reconstruction methods. Nevertheless, performance of phylogenetic methods is typically judged by the inferred tree topology. In this perspective, phylogeny inference methods are usually evaluated by (see also Yang, 2006, pp. 185-188):

Consistency: the ability to estimate the correct tree with sufficient data. Thereby, a phylogenetic method is considered consistent if the probability that the estimated tree is the true tree approaches 1 when the alignment length approaches infinity. For model-based methods such as ML, consistency assessment implies the correct-

ness of the model, i.e the model used for the inference is the same as the model used to generate the data.

Efficiency: the ability to quickly converge on the correct phylogeny. Normally, efficiency of a phylogenetic method is gauged in relation with another method. The relative efficiency of two methods can be measured by the ratio of the amount of data or the alignment length they require to recover the true tree with the same probability P. It is, however, difficult to estimate the amount of data that a method requires to achieve a given P although there have been several attempts to tackle this problem (e.g. Churchill *et al.*, 1992; Yang, 1998; Fischer and Steel, 2009). On the other hand, the probability of recovering the true tree given the alignment length ℓ, $P(\ell)$, can be easily obtained via simulations. $P(\ell)$ is considered the reconstruction accuracy and is calculated as the proportion of the simulated alignments yielding the true tree. Therefore, one can measure the relative efficiency of method 1 over method 2 as:

$$E_{12} = \frac{1 - P_2(\ell)}{1 - P_1(\ell)} \tag{4.1}$$

where $P_1(\ell)$ and $P_2(\ell)$ are the reconstruction accuracies of methods 1 and 2, respectively (Yang, 2006, pp 188).

Robustness: the ability to infer the correct tree in the presence of model violation. A model-based method is considered robust to model violation if it still performs well even when its assumptions are wrong. Maximum parsimony tree reconstruction does not explicitly assume any model or any assumption except its own criterion that evolution requires as least changes as possible. Nonetheless, studies of the robustness in phylogeny inference usually include MP to investigate the behavior of MP in such circumstances.

To assess a tree reconstruction method, the true tree should be known. The most accessible way is to conduct simulations where the evolutionary process is simulated along a "true" tree under a "known" model of evolution. Thanks to the continuous effort to develop more realistic models of sequence evolution (Section 1.1.3) and to implement efficient phylogeny inference programs (Section 1.1.4), abound studies of the performance of phylogenetic methods under a variety of evolutionary scenarios have been carried out. These studies, though aiming at different purposes and sometimes reporting contradicting observations, have arrived at several common records of the performance of

different phylogenetic methods, such as (see e.g. Felsenstein, 1988; Huelsenbeck, 1995a; Nei, 1996; Yang, 2006, pp. 188-190, for reviews):

(i) MP is inconsistent when the true tree contains long nonadjacent branches (Felsenstein, 1978). In such a case, MP biases towards the long branch attraction tree. Similarly, distance- and likelihood-based methods under oversimplified models are also prone to long branch attraction. Later studies with trees of more than four taxa showed that MP is also inconsistent in the existence of a molecular clock (Hendy and Penny, 1989; Zharkikh and Li, 1993; Takezaki and Nei, 1994).

(ii) Likelihood methods are most of the time consistent given that the model used for the inference is the true model (Felsenstein, 1978; Yang, 1995; Swofford *et al.*, 2001). ML under the no common mechanism model can be inconsistent as ML is equivalent to MP in such a case (see Section 1.1.3.4).

(iii) Under the true model likelihood methods are often more efficient than MP and distance-based methods (Hasegawa *et al.*, 1991; Kuhner and Felsenstein, 1994; Tateno *et al.*, 1994; Huelsenbeck, 1995a) except cases in which the true tree itself is always favored by the latter methods. For example, if the true tree contains a cluster of long branches, MP can be the most efficient method.

(iv) The relative branch lengths in the true tree greatly affect the success of tree reconstruction methods. Trees with long internal branches compared to the external branches are much "easier" to recover than trees with short internal branches versus long external branches.

Undoubtedly, such insights may help to avoid wrong interpretation regarding the inferred phylogeny from real data due to reconstruction artefacts such as long branch attraction (see e.g. Anderson and Swofford, 2004; Brinkmann *et al.*, 2005).

4.1.2 Overview of studies of the robustness in phylogeny inference

Among the criteria to evaluate the performance of phylogenetic methods, robustness to incorrect assumptions about the underlying evolutionary model is of particular practical importance as complete and accurate *a priori* knowledge of evolutionary processes is typically not available. An evolutionary model makes a variety of assumptions about the substitution rates, character (e.g. nucleotide) frequencies, rates across sites, and even

heterotachy and heterogeneous substitution rates across branches. Violation of some assumptions affect tree inference methods more than violation of other assumptions. Studies of robustness are therefore abundantly diverse but can be summarized according to which assumptions are violated.

The first studies used an evolutionary model and a tree to generate alignments and then assessed the accuracy of phylogenetic methods using different models of sequence evolution. Several points can be highlighted from previous studies:

(i) Overall, these studies suggested that ML is more robust to model violations than other methods such as NJ and MP (Fukami-Kobayashi and Tateno, 1991; Hasegawa *et al.*, 1991; Tateno *et al.*, 1994; Kuhner and Felsenstein, 1994; Gaut and Lewis, 1995; Huelsenbeck, 1995b; Yang, 1995).

(ii) Violation of transition transversion difference and violation of unequal base compositions has minor (almost no) impact on tree reconstruction accuracy (Fukami-Kobayashi and Tateno, 1991; Huelsenbeck, 1995b).

(iii) Violation of rates across sites heterogeneity greatly influences tree reconstruction methods in recovering the true tree. ML assuming a homogeneous rate across sites is inconsistent when data are generated with more than one rates across sites e.g. two rates (Chang, 1996) or Γ-distributed rates (Huelsenbeck, 1995b; Yang, 1995; Bruno and Halpern, 1999; Sullivan and Swofford, 2001). On the other hand, when the inferring model includes rates across sites heterogeneity, ML is consistent.

Using one evolutionary model for the whole tree and for all sites to generate data is evidently a simplification (see e.g., Lopez *et al.*, 2002). More sophisticated studies of robustness have employed several techniques to violate the underlying model such as adding different GC content to different parts of the simulated data, changing the proportions of variable sites along branches of the tree, and using different sets of branch lengths to simulate partitioned data (a type of heterotachy). Kolaczkowski and Thornton (2004) demonstrated, on a four-taxon tree with different branch length sets, that MP is immune to heterotachy; however, contradictions to their "general conclusion" have been established (Spencer *et al.*, 2005; Gadagkar and Kumar, 2005; Gaucher and Miyamoto, 2005; Philippe *et al.*, 2005; Lockhart *et al.*, 2006; Shavit Grievink *et al.*, 2010). Kolaczkowski and Thornton (2009) showed that Bayesian phylogenetics exhibited long branch attraction bias when sequence sites evolve heterogeneously (e.g. with

different GC content), even when this complexity is incorporated in the inferring model; whereas, ML inference is more robust. Shavit Grievink *et al.* (2010) reported that for a four-taxon tree in which two nonadjacent branches undergo a change in the proportion of variable sites, tree reconstruction under the best-fit model often infers a wrong tree.

In the next chapters we show that the reconstruction accuracy of ML, MP, and BIONJ is hampered by a violation of rates across sites (RaS) heterogeneity and a simultaneous violation of the transition transversion ratio and RaS heterogeneity along two nonadjacent external branches of a four-taxon, clock-like tree. For an eight-taxon balanced tree, these violations cause each of the three methods to infer a different topology: ML and MP reconstruct wrong trees while BIONJ recovers the true tree. In addition, we report that tests of model homogeneity and model fit have enough power to detect such model violations. Based on the results, we draw recommendations for phylogenetic analyses of real data.

4.1.3 A call for a flexible tool to introduce model violation

Clearly, studies of performance in phylogenetics require tools to simulate sequence evolution. Currently available sequence simulation programs incorporate increasingly complex evolutionary scenarios to account for insertion and deletion events (e.g., Fletcher and Yang, 2009), lineage-specific models (Shavit Grievink *et al.*, 2008) or site-specific interactions (Gesell and von Haeseler, 2006). Nonetheless, studies of the robustness need an additional utility: a systematic means to introduce model violation to the simulated alignments. Unfortunately, it is not so straightforward how to use programs like the ones above to incorporate complex scenarios of model violation such as violating the relative substitution rates between the character states or violating rates across sites along several branches of the tree. We therefore introduce ImOSM, a flexible tool to "pepper" a model tree with well-defined deviations from the original model on arbitrary branches.

4.2 The ImOSM method

Assume that we have a phylogenetic tree $\mathcal{T}$ and an alignment $\mathcal{A}$ that evolved along $\mathcal{T}$ under a model of sequence evolution $\mathcal{M}$. ImOSM simulates *intermittent evolution*,

i.e. extra substitution(s) which are thrown on arbitrary branch(es) of the tree, thus changing the otherwise ideal alignment with respect to the substitution process defined by $\mathcal{M}$. Extra substitutions are modeled by the one step mutation (OSM) matrix (Klaere *et al.*, 2008) applied to the K3ST model (Kimura, 1981) (see Section 2.2). Thus, ImOSM actually *Im*beds *O*ne-*S*tep *M*utations into the alignment.

We now explain the different options ImOSM offers. Given a rooted tree and an alignment, one can, on the one hand, explicitly introduce an extra substitution to change a given alignment site by specifying a substitution class and a branch. For example, an extra substitution s_2 occurring on the external branch leading to taxon 1 of the rooted four-taxon tree (Figure 4.1a) changes the site pattern $AACA$ at the first position (column) of the alignment (Figure 4.1b) into the pattern $CACA$. Another extra substitution s_3 on the internal branch leading to taxa 3 and 4 changes the site pattern $GGAC$ at the second position into the pattern $GGTG$. Figure 4.1c depicts the resulting (disturbed) alignment. This explicit specification is worthwhile if one wants to study the effect of a (small) number of extra substitutions.

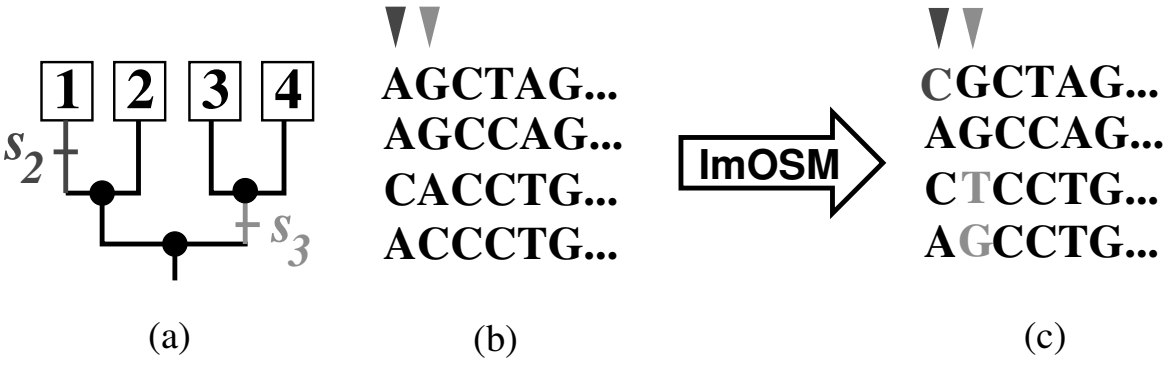


Figure 4.1: An example of an explicit setting in ImOSM. An extra substitution s_2 occurring on the external branch leading to taxon 1 of the rooted four-taxon tree (a) changes the site pattern $AACA$ at the first position of the alignment (b) into the pattern $CACA$. An extra substitution s_3 on the internal branch leading to taxa 3 and 4 changes the site pattern $GGAC$ at the second position into the pattern $GGTG$. The disturbed alignment is depicted in (c).

On the other hand, one may want to introduce the extra substitutions systematically and in a more convenient way. ImOSM provides a variety of settings to accomplish this. Firstly, for each branch different substitution classes may have different probabilities as described in Section 2.2.2, Equation 2.1. By providing equal probabilities for all

the three substitution classes or for the two transversion classes, the more specialized models JC69 (Jukes and Cantor, 1969) or K2P (Kimura, 1980) are derived, respectively. Thereby, ImOSM can simulate relative substitution rate heterogeneity (Section 1.1.3.4) across branches of the tree. Secondly, one can assign the number of extra substitutions per site to each branch by providing the corresponding branch length in the input tree. A branch is free from intermittent evolution by setting its length to zero. Lastly, the extra substitutions can be distributed to alignment sites according to a user defined rates across sites distribution.

Accordingly, ImOSM introduces various model violation scenarios to the data: (i) putting extra substitutions on a specific subset of branches violates the assumption of model homogeneity along the tree, (ii) the probabilities of the three substitution classes of the K3ST model violate the underlying relative substitution rates between the nucleotides along these branches, and (iii) distributing extra substitutions to alignment sites under a different rate distribution violates the underlying RaS distribution. For example, if intermittent evolution is distributed uniformly to alignment sites and along several branches, the assumption of RaS heterogeneity is violated. This also implies heterotachy as the rate at a site shifts along these branches (Philippe and Lopez, 2001, see also Section 1.1.3.4).

Implementation of ImOSM

The k-power of the $M_{\mathcal{T}}$ matrix (Section 2.2.2) provides the weighted exchangeability for all pairs of patterns given that k extra substitutions have occurred on the tree $\mathcal{T}$. The generation of $M_{\mathcal{T}}$ as in Equation 2.1 is analytically feasible. However, for a large number of sequences, computing the power of $M_{\mathcal{T}}$ (a $4^n \times 4^n$ matrix for n nucleotide sequences) is unnecessarily computationally expensive as we hardly observe all 4^n patterns in the alignment.

In practice, as the tree is rooted, ImOSM encodes every branch $e \in E$ as a vector of n binary numbers, v_e, where entries of 1 indicate the taxa (sequences) descended from this branch. If an extra substitution $s_i \in \{s_1, s_2, s_3\}$ occurs on branch e and is assigned to a given alignment site, it will change the nucleotide states of this site pattern at the sequences, whose entries in v_e are 1, according to the substitution class s_i.

If we want to introduce br_{ie} extra substitutions per site to branch e then in the tree

input to ImOSM, branch e must have branch length br_{ie}. We assume the number of extra substitutions that occur on this branch across the whole alignment is Poisson distributed with associated parameter $br_{ie}\ell$. br_{ie} corresponds to the rate parameter, that is the expected number of extra substitutions (events) per site (time unit). The alignment length represents the "time" interval (ℓ sites mean ℓ time units). ImOSM simulates waiting times for the extra substitutions under an exponential distribution with mean $1/br_{ie}$. While the remaining time $t_r > 0$ (at the beginning t_r is set to ℓ), ImOSM repeatedly draws a waiting time t_w from this distribution and compares t_w with t_r. If $t_w > t_r$, no extra substitution happens within time t_r; ImOSM finishes with this branch. If $t_w \leq t_r$, an extra substitution occurs. The extra substitution will be assigned to alignment sites with probabilities proportional to their relative rates (if given) or under a uniform distribution, otherwise. The substitution class (s_1, s_2 or s_3) of the extra substitution is selected with probabilities $\alpha_{s_1,e}, \alpha_{s_2,e}, \alpha_{s_3,e}$, respectively. The alignment is updated right after the extra substitution takes place and the remaining time is reset to $t_r - t_w$.

4.3 Simulations

We study the robustness of three phylogeny reconstruction methods ML, MP, and BIONJ against model violation yielded by ImOSM. Intermittent evolution is introduced to two nonadjacent external branches of a four-taxon tree and an eight-taxon balanced tree. The four-taxon tree allows for a unique choice of two nonadjacent external branches (ignoring the leaf labels); the eight-taxon tree allows for two possibilities (Figure 4.2). We call the trees C4, C8, and C8F, respectively. The internal branch lengths are set to 0.05 substitutions per site; while the external branch lengths (br) vary in $\{0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.75, 1.00\}$.

SEQ-GEN (Rambaut and Grassly, 1997) generates 100 alignments of length $\ell \in \{10^4, 10^5\}$ under the K2P + Γ model, assuming a transition transversion ratio (Ts/Tv) of 2.5 and a Γ-shape parameter α of 0.5 to model RaS heterogeneity. ImOSM then disturbs each of these alignments by putting br_{ie} extra substitutions on the indicated external branches such that $br_{ie} + 0.05 = br$. Thus, the trees are "clock-like" but two nonadjacent external branches evolve only partially according to the original K2P + Γ model. We note that 100 replicates are sufficient to produce stable results in our study

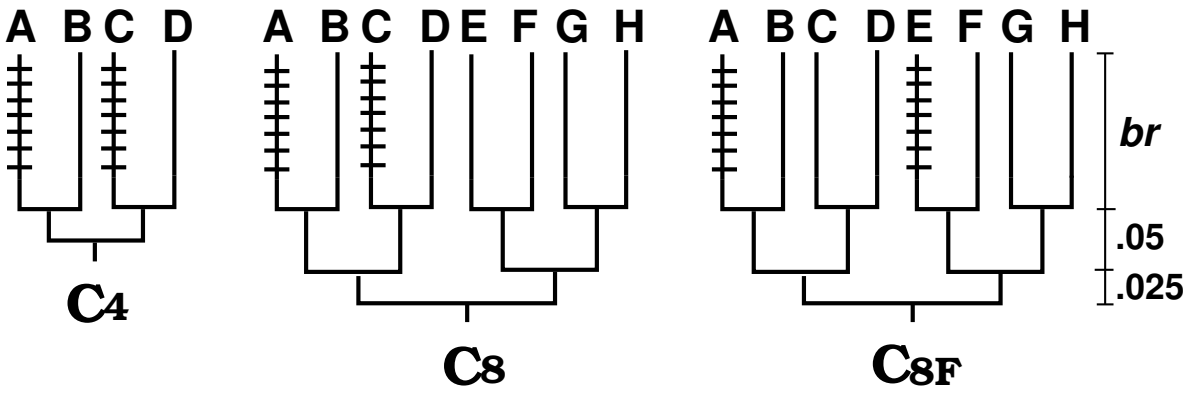


Figure 4.2: Trees used in our simulation and the corresponding abbreviations. Extra substitutions are introduced to the indicated external branches.

as the sequence length is substantially large.

Abbr.	Model	ImOSM setting	Extents of violation
`vNONE`	K2P + Γ*	$Ts/Tv = 2.5$ and RaS	No violation
`vTsTv`	K2P + Γ	$Ts/Tv = 1.0$ and RaS	Ts/Tv violation
`vRaSV`	K2P + Γ	$Ts/Tv = 2.5$ and no RaS	RaS violation
`vBOTH`	K2P + Γ	$Ts/Tv = 1.0$ and no RaS	Violating both Ts/Tv and RaS

Table 4.1: Different settings illustrate different extent of model violation introduced by ImOSM. *The underlying model is K2P + Γ with transition transversion ratio of $Ts/Tv = 2.5$ and Γ-shape parameter α of 0.5 to model rates across sites (RaS) heterogeneity.

Table 4.1 summarizes the different simulation settings. First, intermittent evolution retains $Ts/Tv = 2.5$ and the extra substitutions follow the site-specific rates as determined by SEQ-GEN. Hence, the simulation does not introduce any model violation. We refer to this simulation setting as `vNONE`. Second, extra substitutions are selected uniformly from the substitution classes (JC69 model) but site-specific rates are not changed. Thus, ImOSM "violates" Ts/Tv ratio on the indicated branches. We abbreviate this setting as `vTsTv`. Third, intermittent evolution retains $Ts/Tv = 2.5$ but now the extra substitutions are uniformly distributed. Therefore, ImOSM violates RaS heterogeneity assumption on the indicated branches. This setting is referred to as `vRaSV`. Lastly,

extra mutations are selected uniformly from the substitution classes and distributed uniformly to alignment sites. Thus, both Ts/Tv and RaS heterogeneity are violated on the indicated branches. This setting is abbreviated as `vBOTH`.

The disturbed alignments are subject to tree reconstruction. We use IQPNNI (Vinh and von Haeseler, 2004; Minh *et al.*, 2005) and PAUP* (Swofford, 2002) to estimate the ML and MP trees, respectively. For the ML inference we use the K2P + Γ model and estimate the model parameters. Neighbor joining trees are computed with BIONJ (Gascuel, 1997) with the ML distances based on the inferred model parameters from the ML tree estimation. This means the ML and BIONJ inferences are conducted under a misspecified model for the `vTsTv`, `vRaSV`, and `vBOTH` settings. In addition, we perform Model-Test (Posada and Crandall, 1998), test of model homogeneity across branches (Weiss and von Haeseler, 2003) and goodness of fit tests (Goldman, 1993b; Nguyen *et al.*, 2011).

4.4 Results

4.4.1 Tree reconstruction accuracy

Figure 4.3 presents tree reconstruction accuracy for all simulation settings: `vNONE` (first row), `vTsTv` (second row), `vRaSV` (third row), and `vBOTH` (last row). The accuracy, i.e. the proportion of alignments that yield the true tree, is shown on the y-axis. The x-axis displays the external branch length br or (br_{ie}+0.05). The first two columns show the results for the four-taxon tree C4 with the sequence length of 10^4 and 10^5, respectively. The last two columns show the results for the eight-taxon tree C8. Results for C8F are similar to those for C8 and can be found in the supplementary material, Figure B.1.

No model violation and Ts/Tv violation

The first two rows of Figure 4.3 show the accuracy for simulations with no model violation (`vNONE`) and with the violation of the transition/transversion ratio (`vTsTv`), respectively. For sequence length $\ell = 10^4$, the accuracy of all the three tree-building methods decreases as br increases for both scenarios (`vNONE`, `vTsTv`). ML performs the best while MP performs the worst on the eight-taxon tree (C8). Nonetheless, as the sequence

length increases to 10^5 all the methods successfully recover the true topology. Thus, the violation of the Ts/Tv ratio has almost no impact on reconstruction accuracy; the accuracy is governed by the sequence length. This observation corroborates previous results (Fukami-Kobayashi and Tateno, 1991; Huelsenbeck, 1995a).

Rates across sites violation

The third row of Figure 4.3 displays the accuracy for simulations with rates across sites heterogeneity violation (`vRaSV`). For the four-taxon tree C4 (the first two columns) the reconstruction accuracy, independent of the methods and independent of the alignment length, dramatically drops to 0 as *br* exceeds 0.4. Thus, the violation of RaS heterogeneity causes dramatic changes in tree reconstruction accuracy.

Surprisingly, for the eight-taxon tree C8 (Figure 4.3, third row, last two columns) BIONJ constantly performs the best and recovers the true tree once the sequence length is large. ML performs slightly better than MP. However, they both suffer from the RaS heterogeneity violation: their accuracy drops to 0 if *br* exceeds 0.4.

It should be noted that we have checked and recorded no possible bias of BIONJ due to the input order of the sequences in the distance matrix. All runs with the "randomized input order" option in the NEIGHBOR program (the PHYLIP package, Felsenstein, 1993) produced the same tree as the BIONJ tree.

Both RaS and Ts/Tv violation

The last row of Figure 4.3 shows the accuracy for simulations with violation of both RaS heterogeneity and the Ts/Tv ratio (`vBOTH`). Similar to the `vRaSV` setting, this simultaneous violation yields not only a dramatic change in the accuracy, but also distinct patterns for the C4 and C8 trees. For C4 the accuracy of all methods decreases independently of the sequence length as *br* increases. Interestingly, we observe a slow recovery of the accuracy for ML and BIONJ as *br* exceeds 0.75; nonetheless, their accuracy never exceeds $\frac{2}{3}$, even when we extend *br* to 2.0 (Figure B.2). The reason for the increase in the accuracy of ML and BIONJ remains unclear to us. Nevertheless, we note that Ho and Jermiin (2004) observed similar behavior of ML in their studies with compositional and/or rate heterogeneity among two external branches on a four-taxon tree.

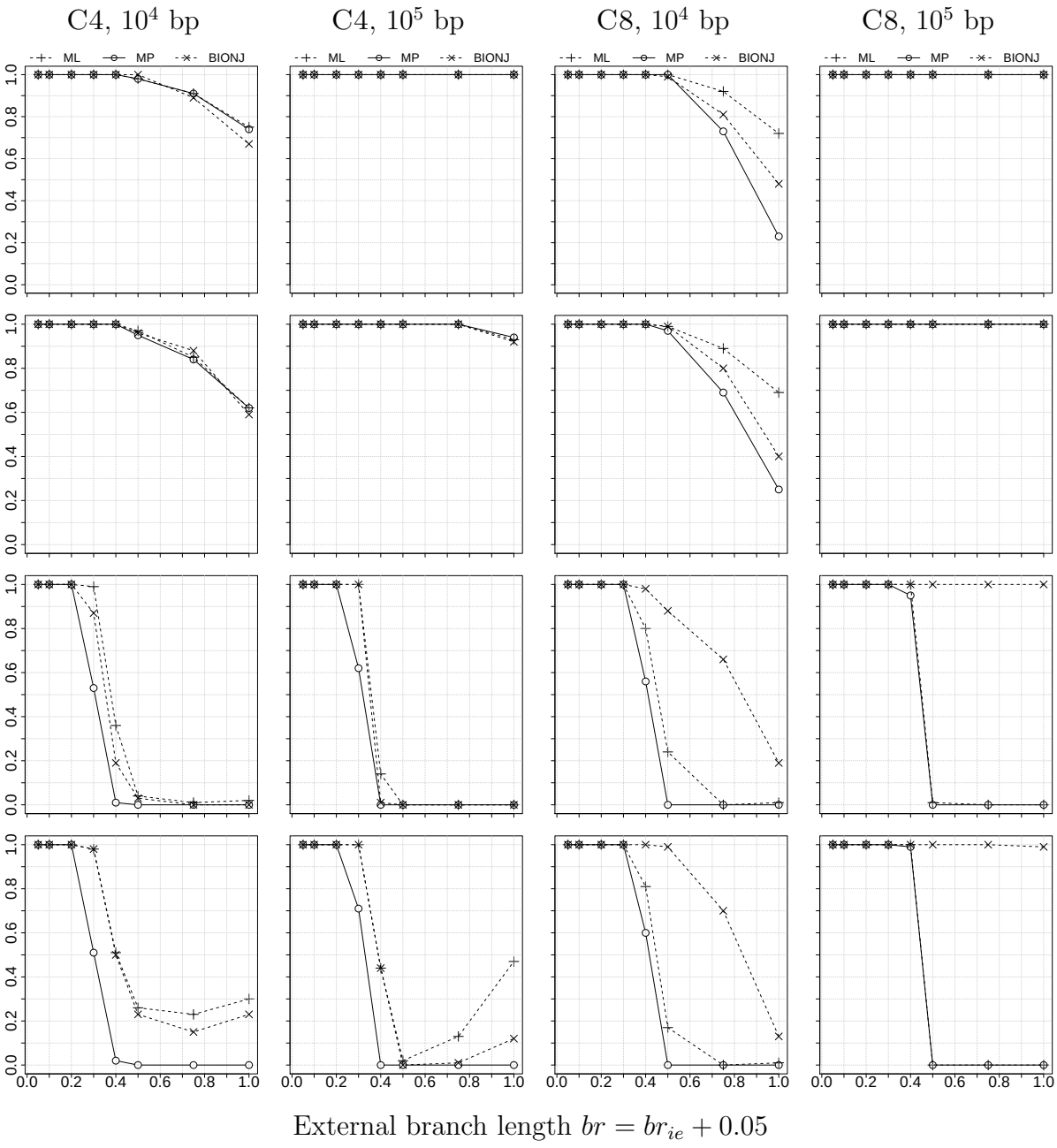


Figure 4.3: Tree reconstruction accuracy, i.e. the proportion of alignments that yield the true tree, is shown on the y-axis for simulations with no model violation (`vNONE`, first row), with Ts/Tv violation (`vTsTv`, second row), with RaS violation (`vRaSV`, third row), and with both Ts/Tv and RaS violation (`vBOTH`, last row). The first two columns show the results for the four-taxon tree C4 with alignment length 10^4 and 10^5, respectively. The last two columns show the results for the eight-taxon tree C8. The x-axis displays the external branch length br or (br_{ie}+0.05). Accuracy of ML is depicted by +, MP by ∘, and NJ by ×.

For C8, the accuracy of ML and MP suffers severely from the violation `vBOTH` while BIONJ's accuracy is not affected for large sequence lengths.

4.4.2 Parameter estimation

The observed behavior of ML and BIONJ provokes a further investigation of the ML estimated model parameters. Without any kind of model violation, `vNONE`, the ML estimations of both parameters the transition transversion ratio and the Γ-shape α are very close to the corresponding true values (Figure B.3). This confirms the statistical consistency of ML inference for the model parameters if the sequence length is large enough.

Transition transversion ratio violation, `vTsTv`, has no influence on the estimation of α: the inferred α is very close to the true value 0.5 (Figure 4.4, first row). However, the inferred Ts/Tv ratio substantially decreases from approximately 2.50 to 1.67 (C4) and to 2.07 (C8) as br_{ie} increases (Figure 4.4, second row). We note that the estimated Ts/Tv ratio roughly agrees with the branch length weighted average of the two Ts/Tv ratios that were used in the simulations.

Notably, rates across sites heterogeneity violation, `vRaSV`, influences the estimation of α but also the Ts/Tv inference (Figure 4.5, first and last row, respectively). The estimated α for the C4 and C8 trees are both larger than 0.5 reflecting lower RaS heterogeneity induced by ImOSM. A substantially larger α is inferred for C4 than for C8. For the C4 tree, the inferred α grows almost linearly with increasing external branch length. Whereas the estimated α for C8 increases to a maximum of 1.11 and then decreases. Similarly, the inferred Ts/Tv deviates from 2.5 more dramatically for C4 than for C8. Note that the proportion of intermittent evolution, i.e. the total branch length of the parts along which sequences evolve under intermittent evolution divided by the tree length, is larger on the four-taxon tree ($\frac{2(br-0.05)}{4br+0.05}$) than on the eight-taxon tree ($\frac{2(br-0.05)}{8br+0.25}$). This leads to the above differences and results in the distinct patterns of behavior (in terms of reconstruction accuracy) of BIONJ between the C4 and C8 trees.

Finally, the estimation of α and Ts/Tv under the violation of both RaS and Ts/Tv (`vBOTH`) shows similar patterns to those under `vRaSV` (Figure B.4). The parameters estimated for the C8F tree are similar to those for C8 as summarized in the supplementary material, Figure B.5.

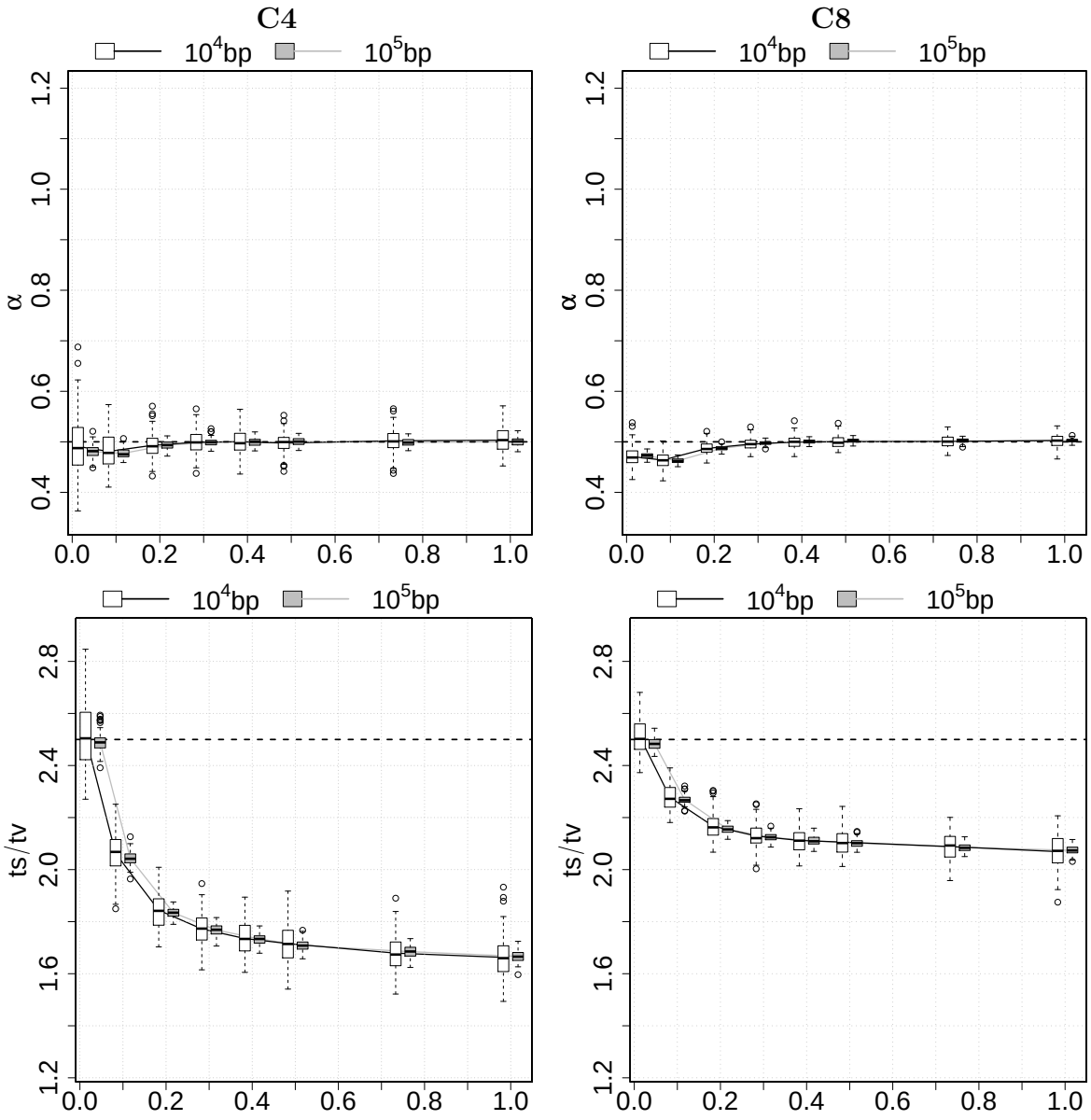


Figure 4.4: ML parameter estimation in the presence of transition transversion ratio violation (`vTsTv`). The first and the last rows show estimation of the Γ-shape parameter α and the Ts/Tv ratio, respectively. Results for the four-taxon tree C4 are presented on the left and for the C8 tree on the right. The x-axis displays the external branch length br or (br_{ie}+0.05).

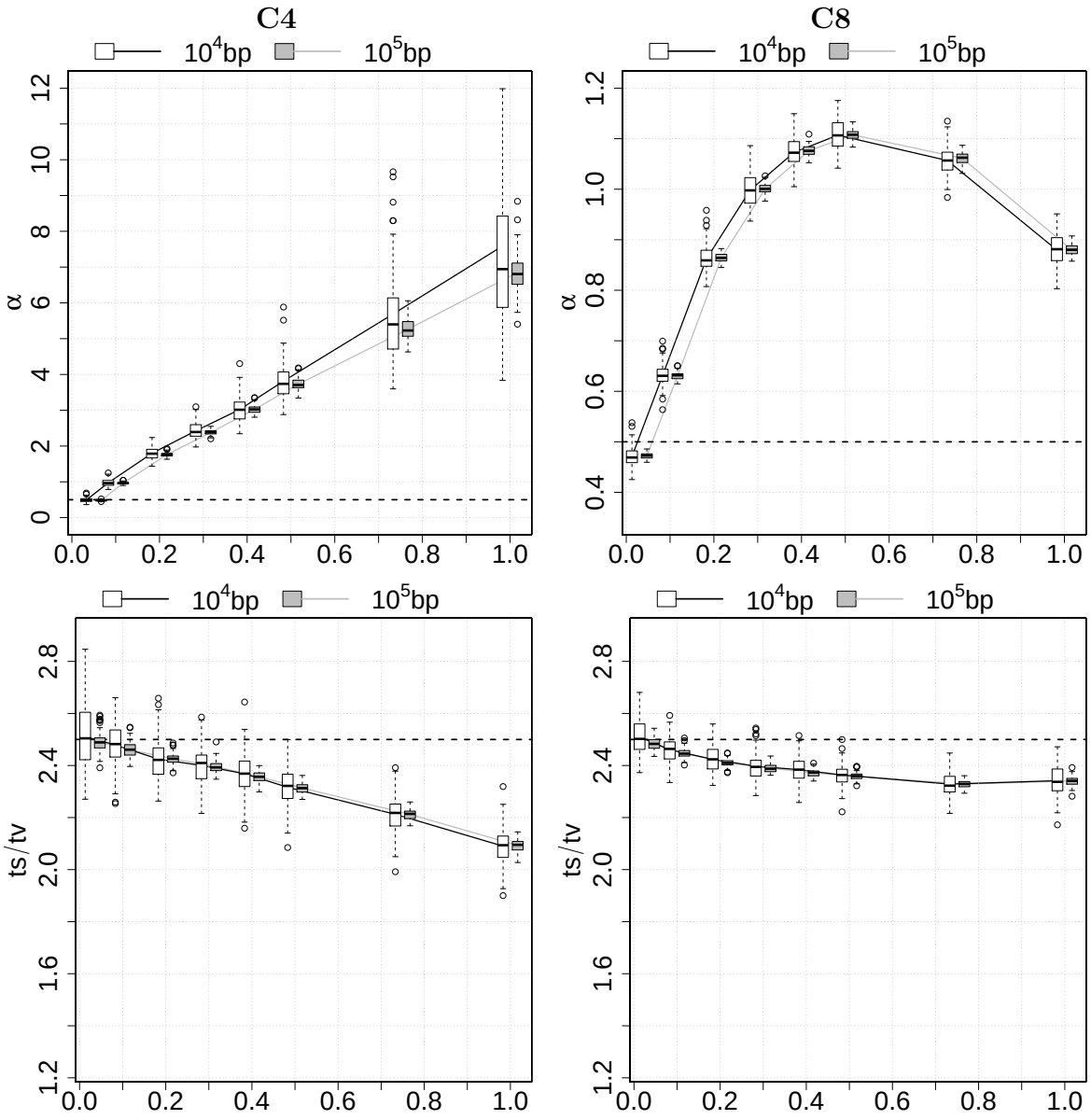


Figure 4.5: ML parameter estimation in the presence of rates across sites violation (`vRaSV`). The first and the last rows show estimation of the Γ-shape parameter α and the Ts/Tv ratio, respectively. Results for the four-taxon tree C4 are presented on the left and for the C8 tree on the right. The x-axis displays the external branch length br or (br_{ie}+0.05).

4.4.3 Possible topological bias

We further check for possible topological bias, i.e. consistently inferring a "wrong" topology, under the `vRaSV` setting. For the four-taxon tree C4, as the sequence length increases to 10^5 and br exceeds 0.4, all three methods always infer the wrong topology `(A,C,(B,D));` which groups taxa that evolve similarly, i.e. `(A,C)`, and `(B,D)`. We noted that a unique MP tree is reconstructed for each of the alignments. Remarkably, although evolution was clock-like, all methods infer substantially larger branch lengths for the external branches leading to `A` and to `C` than the other external branch lengths. Moreover, the estimated internal branch length is significantly larger than zero (the average internal branch lengths inferred by all the three methods are larger than 0.03, Table 4.2). This means we did not observe any unresolved or multifurcating internal node in the inferred trees.

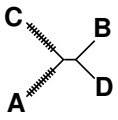
Inferred tree	Method	Mean external branch length				Internal branch length	
		to `A`	to `B`	to `C`	to `D`	mean	st. deviation
	ML	0.600	0.278	0.599	0.280	0.030	0.003
	MP*	0.289	0.180	0.289	0.180	0.127	0.001
	NJ	0.596	0.276	0.595	0.275	0.039	0.004

Table 4.2: Trees and branch lengths inferred by ML, MP and BIONJ for the four-taxon tree (C4) with external branch length $br = 0.5$ under the `vRaSV` setting for sequence length $\ell = 10^5$. All methods infer the same wrong tree as depicted. Recall that ImOSM introduced extra substitutions to the indicated external branches. * Branch lengths for MP are the numbers of substitutions assigned to the branches as reported by PAUP* divided by the sequence length.

For the eight-taxon trees BIONJ always infers, independently of the external branch lengths, the true tree as ℓ grows to 10^5. In contrast, as br exceeds 0.4 neither ML nor MP estimation converges to a single tree. Therefore, we increased ℓ up to 10^7. Table 4.3 shows the number of tree topologies reconstructed by ML and MP for the C8 and C8F trees with $br = 0.5$. As ℓ increases to 10^7 ML inference converges to a single tree, whereas, MP reconstructs more than one tree.

Method	Tree	Seq. length ℓ		
		10^5	10^6	10^7
ML	C8	2	1	1
	C8F	9	4	1
MP	C8	2	2	2
	C8F	4	3	2

Table 4.3: Number of tree topologies inferred by ML (first block) and MP (second block) for the C8 and C8F trees with external branch $br = 0.5$ under the `vRaSV` setting for sequence length $\ell \in \{10^5, 10^6, 10^7\}$.

Table 4.4 shows the tree topologies and their frequency inferred by ML (first block) and MP (second block) for the C8 tree (left) and C8F (right) with $\{br = 0.5, \ell = 10^6\}$. For both the C8 and C8F trees, ML constantly recovers the innermost branch. On each side of the innermost branch, ML then groups taxa that evolve under the pure K2P + Γ model. For C8, the sub-tree `((E,F),(G,H))` is accurately reconstructed, however, taxa `B` and `D` are always incorrectly clustered in the other sub-tree. In addition, ML cannot resolve the positions of taxa `A` and `C`, thus yielding a multifurcating node in the tree. For C8F the two cherries `(C,D)` and `(G,H)`, each in one sub-tree of the innermost branch, are correctly inferred. However, in 67% the cherry `(C,D)` is wrongly grouped with taxon `B` in one sub-tree and the cherry `(G,H)` is erroneously clustered with taxon `F` in the other sub-tree. The remaining 33 trees are multifurcating. Nonetheless, as ℓ grows to 10^7, the ML reconstruction converges to the first (the highlighted) tree. Hence, ML fails to recover the true tree for both the C8 and C8F trees.

MP also fails to reconstruct the true tree for both the C8 and C8F trees but shows a different behavior than ML. For C8, MP infers two tree topologies for $\ell = 10^6$ (Table 4.4, second block, left column). In both topologies, the two taxa `A` and `C`, which are affected by intermittent evolution, erroneously form a cherry. For C8F, three topologies are reconstructed and they all group taxa `A` and `E` (Table 4.4, second block, right column); therefore, MP cannot recover the internal branch separating {`A,B,C,D`} from {`E,F,G,H`}.

Thus, MP does not converge to a single tree (even if $\ell = 10^7$) and always clusters taxa evolving with lower RaS heterogeneity (induced by ImOSM) regardless of their positions in the tree (refer to the C8 and C8F trees) and regardless of the tree size

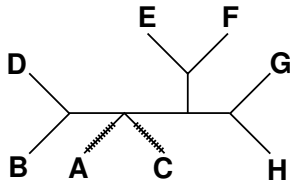
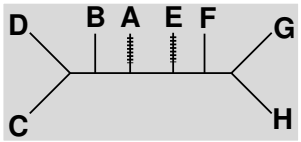
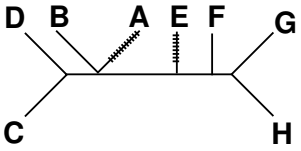
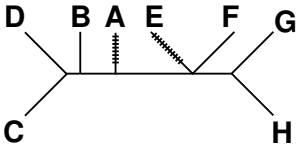
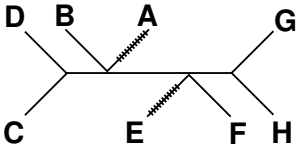
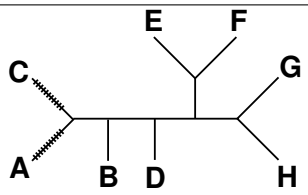
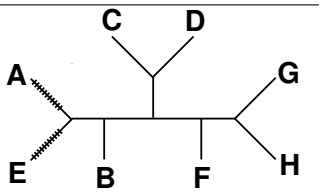
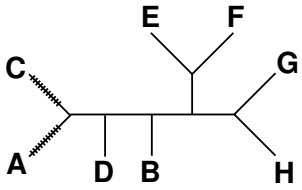
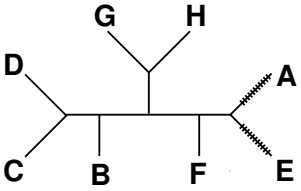
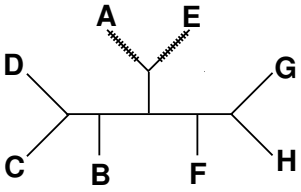
Method	Inferred trees for C8		Inferred trees for C8F	
	#trees	Topology	#trees	Topology
ML	100		67	
			19	
			12	
			2	
MP	55		50	
	45		45	
			5	

Table 4.4: Tree topologies inferred by ML (first block) and MP (second block) for the C8 (left) and C8F (right) trees with external branch $br = 0.5$ under the `vRaSV` setting for sequence length $\ell = 10^6$. Recall that ImOSM introduced extra substitutions to the indicated external branches.

(four- and eight-taxon trees). In contrast, ML infers a single wrong tree and tends to group "relatively close" taxa (on the same side of the innermost branch of the eight-taxon trees) evolving with larger RaS heterogeneity, i.e. taxa evolving under the pure K2P + Γ model. Finally, we note that the behavior of the methods is similar under the `vBOTH` setting.

4.4.4 Model test and goodness of fit evaluation

		$br = (br_{ie} + 0.05)$							
Test	Tree	*0.05*	*0.10*	*0.20*	*0.30*	*0.40*	*0.50*	*0.75*	*1.00*
	C4	0.99	1	1	1	1	1	1	1
(a)	C8	1	1	1	1	1	1	1	1
	C8F	1	1	1	1	1	1	1	1
	C4	0.05	0.92	1	1	1	1	1	1
(b)	C8	0.04	1	1	1	1	1	1	1
	C8F	0.04	1	1	1	1	1	1	1
	C4	0.03	1	1	1	1	1	1	1
(c)	C8	0	1	1	1	1	1	1	1
	C8F	0	1	1	1	1	1	1	1
	C4	0	0	0.21	0.79	0.7	0.65	0.28	0.13
(d)	C8	0	0.32	1	1	1	1	1	1
	C8F	0	0.35	1	1	1	1	1	1

Table 4.5: Proportion of alignments of length $\ell = 10^5$ under the `vRaSV` setting for which (a) the BIC criterion selects the K2P + Γ model, (b) the assumption of model homogeneity along the tree is rejected, (c) the K2P + Γ model is rejected by the Goldman-Cox test, and (d) MISFITS rejects the model and the inferred tree.

We perform several tests to complete the ML analysis for $\ell = 10^5$ under the `vRaSV` setting. The Bayesian information criterion, BIC, (Schwarz, 1978) selects K2P+Γ for more than 99% of the alignments (Table 4.5a). This means BIC does not identify local deviation from the original model. Markedly, the test proposed by Weiss and von Haeseler

(2003) rejects the assumption of model homogeneity across lineages (significance level $\alpha = 0.05$) for almost all alignments (more than 99% on average) if $br_{ie} > 0$ (Table 4.5b).

We further investigate the goodness of fit of the K2P + Γ model and the inferred ML tree to the data using the Goldman-Cox test (Goldman, 1993b) and MISFITS (Nguyen *et al.*, 2011). For each of the 100 disturbed alignments, we performed parametric bootstrap with 100 replicates. The Goldman-Cox test rejects independently of the tree size the K2P + Γ model for all alignments if $br_{ie} > 0$ (Table 4.5c). MISFITS rejects the K2P + Γ model and the inferred tree for a smaller proportion of alignments from the four-taxon tree (an average of 46% for $br_{ie} > 0$) than from the eight-taxon trees (90%, Table 4.5d).

4.5 Discussion

We introduced ImOSM, a tool to imbed intermittent evolution into phylogenetic data in a systematic manner. The intermittent evolution may possess an arbitrary number of distinct sets of relative substitution rates (within the K3ST model) and different distributions of rates across sites across branches on the tree. This is possible because the OSM matrix allows ImOSM to work directly on the alignment site patterns, i.e. the character states at the leaves, instead of evolving an ancestral sequence at the root along the tree like other sequence simulation programs. Thereby, ImOSM provides a convenient means to simulate relative substitution rate heterogeneity across branches (e.g. the `vTsTv` setting) and heterotachy (e.g. the `vRaSV` setting). For studies of robustness in phylogeny inference, ImOSM complements currently available sequence simulation programs by providing a flexible utility to incorporate various types of model violations into the simulated alignments.

We investigated the robustness of ML and BIONJ inference under a misspecified model as well as MP to model violations introduced to a four- and eight-taxon clock-like trees. We showed that the accuracy of all methods was unaffected by the violation of the Ts/Tv ratio on two nonadjacent external branches. RaS heterogeneity violation hampered all methods to recover the true topology for the four-taxon tree as the external branch length increased. For the eight-taxon balanced trees, the violation of RaS heterogeneity and the simultaneous violation of RaS and the Ts/Tv ratio on two nonadjacent external branches caused each of the three methods to infer a different topology. BIONJ using the

ML estimated distances always returned the correct tree; MP incorrectly grouped the two branches undergoing the intermittent evolution (i.e. with lower RaS heterogeneity) whereas ML tended to cluster close taxa evolving with higher RaS heterogeneity. In addition, if the affected branches are close, i.e. on the same side of the innermost branch in the C8 tree, ML inferred a multifurcating tree.

Previously, Kolaczkowski and Thornton (2004) reported that MP outperforms misspecified ML inference and is resistant to a specific setting of heterotachy in which concatenated data are generated from the same four-taxon tree but with different branch length sets. Their result stimulated numerous discussions about the performance of MP and ML tree estimation in the presence of heterotachy. Contradictions to this result were demonstrated for many other combinations of branch lengths (see e.g., Spencer *et al.*, 2005; Gadagkar and Kumar, 2005; Gaucher and Miyamoto, 2005; Philippe *et al.*, 2005; Lockhart *et al.*, 2006). More recently, Wu and Susko (2009) proposed a pairwise alpha heterotachy adjusted (PAHA) distance approach such that NJ with PAHA distances outperformed ML in several settings of heterotachy including the one from Kolaczkowski and Thornton (2004). Here we reported cases in which all methods (ML, MP and BIONJ) incorrectly grouped two nonadjacent branches affected by RaS violation for the four-taxon clock-like tree if the external branch length exceeds 0.4. Moreover, they all estimated larger branch lengths for these two branches. This implies that quartet based analyses, where different methods reconstruct the same tree with long branch attraction, should be interpreted with caution for real data.

The superiority of BIONJ over ML and MP for the eight-taxon trees is surprising. ML was reported in previous studies (e.g., Hasegawa *et al.*, 1991; Huelsenbeck, 1995b) to be more robust to model violation than distance methods such as NJ; nonetheless, the simulation settings (one evolutionary model) and the model trees (four-taxon trees) used in these studies were different from our simulations. Unfortunately, as the three methods infer three different topologies (Figure B.6), the joint analysis of such alignments by different tree reconstruction methods does not provide any indication of which tree may be the correct one. Thus, a more detailed analysis of the data is advised. Model-Test (Posada and Crandall, 1998), which selects a model from a collection of available models but makes no statement about the goodness of fit, did not help in these cases. BIC consistently selected K2P + Γ as the best model for the disturbed alignments. Fortunately, the test proposed by Weiss and von Haeseler (2003) does reject the assumption

of a homogeneous substitution process along the tree. This indicates that the data show model violation. Subsequently, the Goldman-Cox test (Goldman, 1993b) and MISFITS (Nguyen *et al.*, 2011, see also Chapter 3) demonstrated that the violation is so severe that the selected model and the inferred tree cannot explain the data adequately; hence, one should be careful in interpreting the tree. Therefore, we recommend tests of model homogeneity test when applicable and tests of model fit to complete practical phylogenetic analyses. These tests may help to avoid wrong conclusions about the evolutionary relationship as some inferred phylogenies might be unreliable due to model violation and model inadequacy to data.

Finally, we note that our simulations imply a kind of heterotachy. Thus, an interesting extension of this work would be to evaluate the accuracy of mixed branch length models that aim to account for heterotachy (Kolaczkowski and Thornton, 2008; Pagel and Meade, 2008). We also note that this is not an exhaustive simulation study for different model violations. We provide a tool to introduce model violations and show that already very simple violations of the model on two branches of the tree can lead to bewildering results, like the three different trees inferred by the three different phylogeny reconstruction methods.

Chapter 5

Conclusions and Outlook

I am still confused. But on a higher level.
Enrico Fermi

After providing an excursion into the basic tasks of phylogenetic analyses (Chapter 1) and an introduction to statistical tests of evolutionary models (Section 1.2 and Section 3.1), this thesis investigates two issues in phylogeny reconstruction. First, testing the absolute fit of an evolutionary model to an alignment enables us to reject inadequate models and to gain confidence in an accepted model. We proposed a new method, called MISFITS, to evaluate the goodness of fit between an evolutionary model and an alignment (Chapter 3). Different from previously developed methods, which assess fit based on the unconstrained likelihood and the likelihood under the model over the whole alignment, MISFITS compares the two likelihoods site-wise. This allows our method to pinpoint to site patterns that are inadequately captured by the model. Simultaneously, MISFITS assists the model in elucidating the presence of these site patterns in the alignment by introducing a minimum number of extra substitutions to the inferred tree. To the best of our knowledge, this is the first time the inadequacy of an evolutionary model to the data is quantified and explained in a maximum parsimony framework. The software program implementing MISFITS (freely available at http://www.cibiv.at/software/misfits) makes it possible to apply the method routinely in practical phylogenetic analyses. Our survey of the goodness of fit conducted on the PANDIT alignments (Whelan *et al.*, 2006) indicates that in a number of instances (350 of 4,268 alignments), the best-fit models do not fully explain the data. Therefore, more thorough investigations are advisable in such cases.

Second, awareness of the performance of different phylogenetic methods aids in avoiding incorrect conclusions from the inferred phylogenies due to reconstruction artefacts. We briefly summarized previous simulation-based studies of the performance in phylogeny inference putting emphasis on studies of the robustness to model violation, a practically important criterion to assess a phylogenetic method (Chapter 4). We then developed a flexible utility named ImOSM (http://www.cibiv.at/software/imosm), an add-on tool to sequence simulation programs, to introduce various kinds of model violations for studies of the robustness. Our simulation study with model violation caused by ImOSM provides additional insights into the robustness of ML, MP and BIONJ methods. On the one hand, it is an interesting observation that each of the three methods infers a different topology for the eight-taxon tree in the presence of rates across sites heterogeneity violation along two nonadjacent external branches. Nonetheless, such a result directly provokes further examination. On the other hand, the fact that the three methods agree on the same tree topology does not necessarily mean they recover the true tree (c.f. simulations with violation of rates across sites heterogeneity along two nonadjacent branches of the four-taxon tree). A hasty conclusion with regard to the reconstructed tree in such a case would be misleading. Fortunately, tests of absolute model fit including MISFITS do reject the best-fit model in both instances. This implies that the inferred trees are possibly unreliable due to the inadequacy of the model to the data. Thus, these tests should be applied to every practical phylogenetic analysis regardless of how pleased one might be about the inferred phylogeny.

Both MISFITS and ImOSM utilize the concept of *extra substitutions* occurring on arbitrary branches of a phylogeny which we call *intermittent evolution*. We demonstrated how to model intermittent evolution for nucleotide characters (Chapter 2) using the one step mutation (OSM) matrix proposed recently by Klaere *et al.* (2008). The OSM matrix integrates both components of an evolutionary process, the phylogeny and the substitution model, into a unified framework. The two applications, MISFITS and ImOSM, show how this framework sheds additional light on our understanding of the complexity of sequence evolution.

From the methodology point of view, modelling sequence evolution using the OSM matrix has one limitation, that is the permutation matrices from the substitution model and the identity matrix have to form a commutative group with respect to matrix multiplication. For nucleotide substitution, the K3ST model is the most general model

that allows for the formation of such a group (i.e. an Abelian group of order four with matrix multiplication as the group operation). Abelian groups of order 20 exist (Humphreys, 1996, Appendix B); therefore, extensions of MISFITS and ImOSM to amino acid data are possible. In general, the OSM matrix is applicable to any alphabet where an Abelian group can be constructed. The important point is to assign the permutation matrices of the group to the character states that make biological sense. This will be analysed in future work.

Bibliography

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

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control,*, **19**, 716–723.

Anderson, F. E. and Swofford, D. L. (2004) Should we be worried about long-branch attraction in real data sets? Investigations using metazoan 18S rDNA. *Mol. Phylogenet. Evol.*, **33**, 440–451.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2011) GenBank. *Nucl. Acids Res.*, **39**, D32–37.

Blanquart, S. and Lartillot, N. (2006) A Bayesian compound stochastic process for modeling nonstationary and nonhomogeneous sequence evolution. *Mol. Biol. Evol.*, **23**, 2058–2071.

Blanquart, S. and Lartillot, N. (2008) A site- and time-heterogeneous model of amino acid replacement. *Mol. Biol. Evol.*, **25**, 842–858.

Bofkin, L. and Goldman, N. (2007) Variation in evolutionary processes at different codon positions. *Mol. Biol. Evol.*, **24**, 513–521.

Bollback, J. P. (2002) Bayesian model adequacy and choice in phylogenetics. *Mol. Biol. Evol.*, **19**, 1171–1180.

Brinkmann, H., van der Giezen, M., Zhou, Y. and Poncelin de Raucourt, G. (2005) An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. *Syst. Biol.*, **54**, 743–757.

Bruno, W. J. and Halpern, A. L. (1999) Topological bias and inconsistency of maximum likelihood using wrong models. *Mol. Biol. Evol.*, **16**, 564–566.

Buckley, T. R., Simon, C. and Chambers, G. K. (2001) Exploring among-site rate variation models in a maximum likelihood framework using empirical data: Effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.*, **50**, 67–86.

Camin, J. H. and Sokal, R. R. (1965) A method for deducing branching sequences in phylogeny. *Evolution*, **19**, 311–326.

Cavalli-Sforza, L. L. and Edwards, A. W. F. (1967) Phylogenetic analysis: Models and estimation procedures. *Evolution*, **21**, 550–570.

Chang, J. T. (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. *Math. Biosci.*, **134**, 189–215.

Churchill, G. A. (1992) Hidden Markov chains and the analysis of genome structure. *Computers Chem.*, **16**, 107–115.

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

Cox, D. R. (1961) Tests of separate families of hypotheses. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, pages 105–123, Berkeley, CA, USA.

Cox, D. R. (1962) Further results on tests of separate families of hypotheses. *J. R. Soc. Stat. B*, **24**, 406–424.

Darwin, C. (1859) *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray, London.

Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978) A model of evolutionary change in proteins. In Dayhoff, M. O. (ed.), *Atlas of Protein Sequence Structure*, vol. 5, pages 345–352, National Biomedical Research Foundation, Washington DC.

Edgar, R. C. and Batzoglou, S. (2006) Multiple sequence alignment. *Curr. Opin. Struct. Biol.*, **16**, 368–373.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1963) The reconstruction of evolution. *Ann. Hum. Genet.*, **27**, 104–105.

Efron, B. (1979) Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, **7**, 1–26.

Farris, J. (1970) Methods for computing Wagner trees. *Syst. Zool.*, **19**, 83–92.

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

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

Felsenstein, J. (1981b) Evolutionary trees from gene frequencies and quantitative characters: Finding maximum likelihood estimates. *Evolution*, **35**, 1229–1242.

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

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

Felsenstein, J. (1993) *PHYLIP (Phylogeny Inference Package) version 3.5c*. Department of Genetics, University of Washington, Seattle, Distributed by the author.

Felsenstein, J. (2004) *Inferring Phylogenies*. Sinauer Associates, Sunderland, Massachusetts.

Fischer, M. and Steel, M. (2009) Sequence length bounds for resolving a deep phylogenetic divergence. *J. Theo. Biol.*, **256**, 247–252.

Fischer, M. and Thatte, B. (2010) Revisiting an equivalence between maximum parsimony and maximum likelihood methods in phylogenetics. *Bull. Math. Biol.*, **72**, 208–220.

Fitch, W. M. (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst. Zool.*, **20**, 406–416.

Fitch, W. M. (2000) Homology – a personal view on some of the problems. *Trends Genet.*, **16**, 227–231.

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

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

Fleissner, R., Metzler, D. and von Haeseler, A. (2005) Simultaneous statistical multiple alignment and phylogeny reconstruction. *Syst. Biol.*, **54**, 548–561.

Fletcher, W. and Yang, Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**, 1879–1888.

Foster, P. (2004) Modeling compositional heterogeneity. *Syst. Biol.*, **53**, 485–495.

Frati, F., Simon, C., Sullivan, J. and Swofford, D. L. (1997) Evolution of the mitochondrial cytochrome oxidase II gene in collembola. *J. Mol. Evol.*, **44**, 145–158.

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

Gadagkar, S. R. and Kumar, S. (2005) Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. *Mol. Biol. Evol.*, **22**, 2139–2141.

Galtier, N. (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. *Mol. Biol. Evol.*, **18**, 866–873.

Galtier, N. and Gouy, M. (1998) Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.*, **15**, 871–879.

Gascuel, O. (1997) BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.*, **14**, 685–695.

Gaucher, E. A. and Miyamoto, M. M. (2005) A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. *Mol. Phylogenet. Evol.*, **37**, 928–931.

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

Gesell, T. and von Haeseler, A. (2006) In silico sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics*, **22**, 716–722.

Gold, R. Z. (1963) Tests auxiliary to χ^2 tests in a markov chain. *Ann. Math. Statist.*, **34**, 56–74.

Goldman, N. (1993a) Simple diagnostic statistical tests of models for DNA substitution. *J. Mol. Evol.*, **37**, 650–661.

Goldman, N. (1993b) Statistical tests of models of DNA substitution. *J. Mol. Evol.*, **36**, 182–198.

Goldman, N. and Whelan, S. (2000) Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.*, **17**, 975–978.

Gowri-Shankar, V. and Rattray, M. (2007) A reversible jump method for Bayesian phylogenetic inference with a nonhomogeneous substitution model. *Mol. Biol. Evol.*, **24**, 1286–1299.

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.

Guindon, S. and Gascuel, O. (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, **52**, 696–704.

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

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

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

Higgins, D. and Lemey, P. (2009) Multiple sequence alignment. In Lemey, P., Salemi, M. and Anne-Mieke, V. (eds.), *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pages 68–108, Cambridge University Press, Cambridge, Second edn..

Ho, S. Y. W. and Jermiin, L. (2004) Tracing the decay of the historical signal in biological sequence data. *Syst. Biol.*, **53**, 623–637.

Horai, S., Hayasaka, K., Kondo, R., Tsugane, K. and Takahata, N. (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc. Natl. Acad. Sci. USA*, **92(2)**, 532–536.

Horn, R. A. and Johnson, C. R. (1991) *Topics in matrix analysis*. Oxford University Press, New York.

Huelsenbeck (1995a) Performance of phylogenetic methods in simulation. *Syst. Biol.*, **44**, 17–48.

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

Huelsenbeck, J. P. and Crandall, K. A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Annu. Rev. Ecol. Syst.*, **28**, 437–466.

Huelsenbeck, J. P. and Nielsen, R. (1999) Variation in the pattern of nucleotide substitution across sites. *J. Mol. Evol.*, **48**, 86–93.

Huelsenbeck, J. P. and Rannala, B. (1997) Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. *Science*, **276**, 227–232.

Huelsenbeck, J. P. and Ronquist, F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R. and Bollback, J. P. (2001) Bayesian inference of phylogeny and its impact on evolutionary biology. *Science*, **294**, 2310–2314.

Humphreys, J. F. (1996) *A course in group theory*. Oxford University Press, New York.

Jermiin, L. S., Ho, J. W. K., Lau, K. W. and Jayaswal, V. (2009) SeqVis: A tool for detecting compositional heterogeneity among aligned nucleotide sequences. In Posada, D. (ed.), *Bioinformatics for DNA Sequence Analysis*, pages 65–91, Humana Press, Totowa, NJ.

Jermiin, L. S., Jayaswal, V., Ababneh, F. and Robinson, J. (2008) Phylogenetic model evaluation. In Keith, J. M. (ed.), *Bioinformatics, Volume I: Data, sequence analysis and evolution*, pages 331–364, Humana Press, Totowa, NJ.

Jobb, G., von Haeseler, A. and Strimmer, K. (2004) TREEFINDER: a powerful graphical analysis environment for molecular phylogenetics. *BMC Evol. Biol.*, **4**, 18.

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

Jukes, T. H. and Cantor, C. R. (1969) Evolution of protein molecules. In Munro, H. N. (ed.), *Mammalian Protein Metabolism*, vol. 3, pages 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.

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

Klaere, S., Gesell, T. and von Haeseler, A. (2008) The impact of single substitutions on multiple sequence alignments. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **363**, 4041–4047.

Kolaczkowski, B. and Thornton, J. W. (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature*, **431**, 980–984.

Kolaczkowski, B. and Thornton, J. W. (2008) A mixed branch length model of heterotachy improves phylogenetic accuracy. *Mol. Biol. Evol.*, **25**, 1054–1066.

Kolaczkowski, B. and Thornton, J. W. (2009) Long-branch attraction bias and inconsistency in bayesian phylogenetics. *PLoS ONE*, **4**, e7891.

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

Kumar, S., Nei, M., Dudley, J. and Tamura, K. (2008) MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief. Bioinformatics*, **9**, 299–306.

Landan, G. and Graur, D. (2007) Heads or tails: A simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.*, **24**, 1380–1383.

Lartillot, N. and Philippe, H. (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Mol. Biol. Evol.*, **21**, 1095–1109.

Le, S. Q. and Gascuel, O. (2008) An improved general amino acid replacement matrix. *Mol. Biol. Evol.*, **25**, 1307–1320.

Lemey, P., Salemi, M. and Anne-Mieke, V. (2009) *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*. Cambridge University Press, Cambridge, Second edn..

Li, W.-H. (1997) *Molecular Evolution*. Sinauer Associates, Sunderland, Massachusetts.

Liu, K., Raghavan, S., Nelesen, S., Linder, C. R. and Warnow, T. (2009) Rapid and accurate large-scale coestimation of sequence alignments and phylogenetic trees. *Science*, **324**, 1561–1564.

Lockhart, P., Novis, P., Milligan, B. G., Riden, J., Rambaut, A. and Larkum, T. (2006) Heterotachy and tree building: a case study with plastids and eubacteria. *Mol. Biol. Evol.*, **23**, 40–45.

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

Lopez, P., Casane, D. and Philippe, H. (2002) Heterotachy, an important process of protein evolution. *Mol. Biol. Evol.*, **19**, 1–7.

Löytynoja, A. and Goldman, N. (2009) Evolution. Uniting alignments and trees. *Science*, **324**, 1528–1529.

Metzler, D. and Fleissner, R. (2009) Sequence evolution models for simultaneous alignment and phylogeny reconstruction. In Rosenberg, M. S. (ed.), *Sequence Alignment: Methods, Models, Concepts, and Strategies*, pages 179–208, University of California Press, California.

Meyer, S. and von Haeseler, A. (2003) Identifying site-specific substitution rates. *Mol. Biol. Evol.*, **20**, 182–189.

Minh, B. Q., Vinh, L. S., von Haeseler, A. and Schmidt, H. A. (2005) pIQPNNI – parallel reconstruction of large maximum likelihood phylogenies. *Bioinformatics*, **21**, 3794–3796.

Munkres, J. (1957) Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, **5(1)**, 32–38.

Navidi, W. C., Churchill, G. A. and von Haeseler, A. (1991) Methods for inferring phylogenies from nucleic acid sequence data by using maximum likelihood and linear invariants. *Mol. Biol. Evol.*, **8**, 128–143.

Nei, M. (1996) Phylogenetic analysis in molecular evolutionary genetics. *Annu. Rev. Genet.*, **30**, 371–403.

Nguyen, M. A. T., Klaere, S. and von Haeseler, A. (2011) MISFITS: evaluating the goodness of fit between a phylogenetic model and an alignment. *Mol. Biol. Evol.*, **28**, 143–152.

Pagel, M. and Meade, A. (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst. Biol.*, **53**, 571–581.

Pagel, M. and Meade, A. (2008) Modelling heterotachy in phylogenetic inference by reversible-jump Markov chain Monte Carlo. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **363**, 3955–3964.

Penn, O., Privman, E., Landan, G., Graur, D. and Pupko, T. (2010) An alignment confidence score capturing robustness to guide tree uncertainty. *Mol. Biol. Evol.*, **27**, 1759–1767.

Philippe, H. and Lopez, P. (2001) On the conservation of protein sequences in evolution. *Trends Biochem. Sci.*, **26**, 414–416.

Philippe, H., Zhou, Y., Brinkmann, H., Rodrigue, N. and Delsuc, F. (2005) Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.*, **5**, 50.

Posada, D. (2008) jModelTest: Phylogenetic model averaging. *Mol. Biol. Evol.*, **25**, 1253–1256.

Posada, D. (2009) Selecting models of evolution. In Lemey, P., Salemi, M. and Anne-Mieke, V. (eds.), *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pages 345–361, Cambridge University Press, Cambridge, Second edn..

Posada, D. and Crandall, K. A. (1998) MODELTEST: Testing the model of DNA substitution. *Bioinformatics*, **14**, 817–818.

Rambaut, A. and Grassly, N. C. (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, **13**, 235–238.

Rannala, B. (2002) Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.*, **51**, 754–760.

Revell, L. J., Harmon, L. J. and Glor, R. E. (2005) Underparameterized model of sequence evolution leads to bias in the estimation of diversification rates from molecular phylogenies. *Syst. Biol.*, **54**, 973–983.

Ripplinger, J. and Sullivan, J. (2008) Does choice in model selection affect maximum likelihood analysis? *Syst. Biol.*, **57**, 76–85.

Robinson, D. F. and Foulds, L. R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Rodríguez-Trelles, F., Tarrío, R. and Ayala, F. J. (2006) Rates of molecular evolution. In Fox, C. W. and Wolf, J. B. (eds.), *Evolutionary Genetics: Concepts and Case Studies*, pages 119–132, Oxford University Press, New York (NY), First edn..

Ronquist, F., van der Mark, P. and Huelsenbeck, J. P. (2009) Bayesian phylogenetic analysis using MrBayes. In Lemey, P., Salemi, M. and Anne-Mieke, V. (eds.), *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pages 210–266, Cambridge University Press, Cambridge, Second edn..

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

Saha, P., Thome, K. C., Yamaguchi, R., Hou, Z.-h., Weremowicz, S. and Dutta, A. (1998) The human homolog of saccharomyces cerevisiae CDC45. *J. Biol, Chem.*, **273**, 18205–18209.

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

Sankoff, D. (1975) Minimal mutation trees of sequences. *SIAM J. Appl. Math.*, **28**, 35–42.

Schmidt, H. A., Strimmer, K., Vingron, M. and von Haeseler, A. (2002) TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, **18**, 502–504.

Schöniger, M. and von Haeseler, A. (1999) Toward assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *J. Mol. Evol.*, **49**, 691–698.

Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Shavit Grievink, L., Penny, D., Hendy, M. D. and Holland, B. R. (2008) LineageSpecificSeqgen: generating sequence data with lineage-specific variation in the proportion of variable sites. *BMC Evol. Biol.*, **8**, 317.

Shavit Grievink, L., Penny, D., Hendy, M. D. and Holland, B. R. (2010) Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. *Syst. Biol.*, **59**, 288–297.

Sokal, R. R. and Sneath, P. H. A. (1963) *Numerical Taxonomy*. W.H. Freeman and Co., San Fransisco, CA.

Spencer, M., Susko, E. and Roger, A. J. (2005) Likelihood, parsimony, and heterogeneous evolution. *Mol. Biol. Evol.*, **22**, 1161–1164.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.

Strimmer, K. and von Haeseler, A. (2009) Genetic distances and nucleotide substitution models. In Lemey, P., Salemi, M. and Anne-Mieke, V. (eds.), *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pages 111–141, Cambridge University Press, Cambridge, Second edn..

Sullivan, J. and Joyce, P. (2005) Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.*, **36**, 445–466.

Sullivan, J., Markert, J. A. and Kilpatrick, C. W. (1997) Phylogeography and molecular systematics of the Peromyscus aztecus species group (Rodentia: Muridae) inferred using parsimony and likelihood. *Syst. Biol.*, **46**, 426–440.

Sullivan, J. and Swofford, D. L. (2001) Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.*, **50**, 723–729.

Swofford, D. L. (2002) PAUP*: Phylogenetic analysis using parsimony (*and other methods). version 4.

Swofford, D. L. and Maddison, W. P. (1987) Reconstructing ancestral character states under Wagner parsimony. *Mathematical Biosciences*, **87**, 199–229.

Swofford, D. L., Olsen, G. J., Waddell, P. J. and Hillis, D. M. (1996) Phylogenetic inference. In Hillis, D. M., Moritz, C. and Mable, B. K. (eds.), *Molecular Systematics*, pages 407–514, Sinauer Associates, Inc., Sunderland, Massachusetts USA, Second edn..

Swofford, D. L., Waddell, P. J., Huelsenbeck, J. P., Foster, P. G., Lewis, P. O. and Rogers, J. S. (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Syst. Biol.*, **50**, 525–539.

Takezaki, N. and Nei, M. (1994) Inconsistency of the maximum parsimony method when the rate of nucleotide substitution is constant. *J. Mol. Evol.*, **39**, 210–218.

Tamura, K. and Kumar, S. (2002) Evolutionary distance estimation under heterogeneous substitution pattern among lineages. *Mol. Biol. Evol.*, **19**, 1727–1736.

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.

Tateno, Y., Takezaki, N. and Nei, M. (1994) Relative efficiencies of the maximum–likelihood, neighbor–joining, and maximum–parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.*, **11**, 261–277.

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

Tuffley, C. and Steel, M. (1997) Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bull. Math. Biol.*, **59**, 581–607.

Tuffley, C. and Steel, M. (1998) Modeling the covarion hypothesis of nucleotide substitution. *Mathematical Biosciences*, **147**, 63–91.

UniProt Consortium (2010) The Universal Protein Resource (UniProt) in 2010. *Nucl. Acids Res.*, **38**, D142–148.

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

Vandamme, A.-M. (2009) Basic concepts of molecular evolution. In Lemey, P., Salemi, M. and Anne-Mieke, V. (eds.), *The Phylogenetic Handbook: a Practical Approach to Phylogenetic Analysis and Hypothesis Testing*, pages 1–29, Cambridge University Press, Cambridge, Second edn..

Vinh, L. S. and von Haeseler, A. (2004) IQPNNI: Moving fast through tree space and stopping in time. *Mol. Biol. Evol.*, **21**, 1565–1571.

Waddell, P. J. (2005) Measuring the fit of sequence data to phylogenetic model: allowing for missing data. *Mol. Biol. Evol.*, **22**, 395–401.

Waddell, P. J., Ota, R. and Penny, D. (2009) Measuring fit of sequence data to phylogenetic model: gain of power using marginal tests. *J. Mol. Evol.*, **69**, 289–299.

Wang, H. C., Spencer, M., Susko, E. and Roger, A. J. (2007) Testing for covarion-like evolution in protein sequences. *Mol. Biol. Evol.*, **24**, 294–305.

Weiss, G. and von Haeseler, A. (2003) Testing substitution models within a phylogenetic tree. *Mol. Biol. Evol.*, **20**, 572–578.

Whelan, S. (2008) Spatial and temporal heterogeneity in nucleotide sequence evolution. *Mol. Biol. Evol.*, **25**, 1683–1694.

Whelan, S., de Bakker, P. I. W., Quevillon, E. and Rodriguez, N. (2006) Pandit: an evolution-centric database of protein and associated nucleotide domains with inferred trees. *Nucleic Acids Research*, **34**, 327–331.

Whelan, S., Blackburne, B. P. and Spencer, M. (2011) Phylogenetic substitution models for detecting heterotachy during plastid evolution. *Mol. Biol. Evol.*, **28**, 449–458.

Whelan, S. and Goldman, N. (1999) Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. *Mol. Biol. Evol.*, **16**, 1292–1299.

Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum likelihood approach. *Mol. Biol. Evol.*, **18**, 691–699.

Whelan, S., Liò, P. and Goldman, N. (2001) Molecular phylogenetics: State-of-the-art methods for looking into the past. *Trends Genet.*, **17**, 262–272.

Wu, J. and Susko, E. (2009) General heterotachy and distance method adjustments. *Mol. Biol. Evol.*, **26**, 2689–2697.

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: Approximative methods. *J. Mol. Evol.*, **39**, 306–314.

Yang, Z. (1995) Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. *J. Mol. Evol.*, **40**, 689–697.

Yang, Z. (1998) On the best evolutionary rate for phylogenetic analysis. *Syst. Biol.*, **47**, 125–133.

Yang, Z. (2006) *Computational Molecular Biology*. Oxford University Press, New York, USA.

Yang, Z. and Roberts, D. (1995) On the use of nucleic acid sequences to infer early branchings in the tree of life. *Mol. Biol. Evol.*, **12**, 451–458.

Zharkikh, A. (1994) Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.*, **39**, 315–329.

Zharkikh, A. and Li, W.-H. (1993) Inconsistency of the maximum parsimony method: the case of five taxa with a molecular clock. *Syst. Biol.*, **42**, 113–125.

Appendix A

Supplemental tables and figures to Chapter 3

Null hypothesis $\mathbf{H}_0$: A given evolutionary model $\mathcal{M}$ is an adequate model to describe a given alignment $\mathcal{A}$.

Step 1: Estimate the ML tree for $\mathcal{A}$ under $\mathcal{M}$ and all the parameters (model parameters and branch lengths).

Step 2: Calculate the unconstrained log likelihood for the original alignment $\mathcal{A}$ as in Equation (3.1) and compute its difference to the maximum log likelihood (L) from the ML inference: $\sigma_{obs} = \mathrm{L}_{\mathcal{A}}^{unc} - \mathrm{L}_{\mathcal{A}}$.

Step 3: Use the parameters in **Step 1** to generate say 100 alignments $\mathcal{A}_1 \cdots \mathcal{A}_{100}$, i.e. parametric bootstrap replicates under $\mathcal{M}$.

Step 4: For each generated alignment $\mathcal{A}_i$: (i) compute the unconstrained likelihood, (ii) estimate the ML tree and all parameters under the model $\mathcal{M}$, then (iii) calculate the difference $\sigma_j = \mathrm{L}_{A_j}^{unc} - \mathrm{L}_{A_j}$.

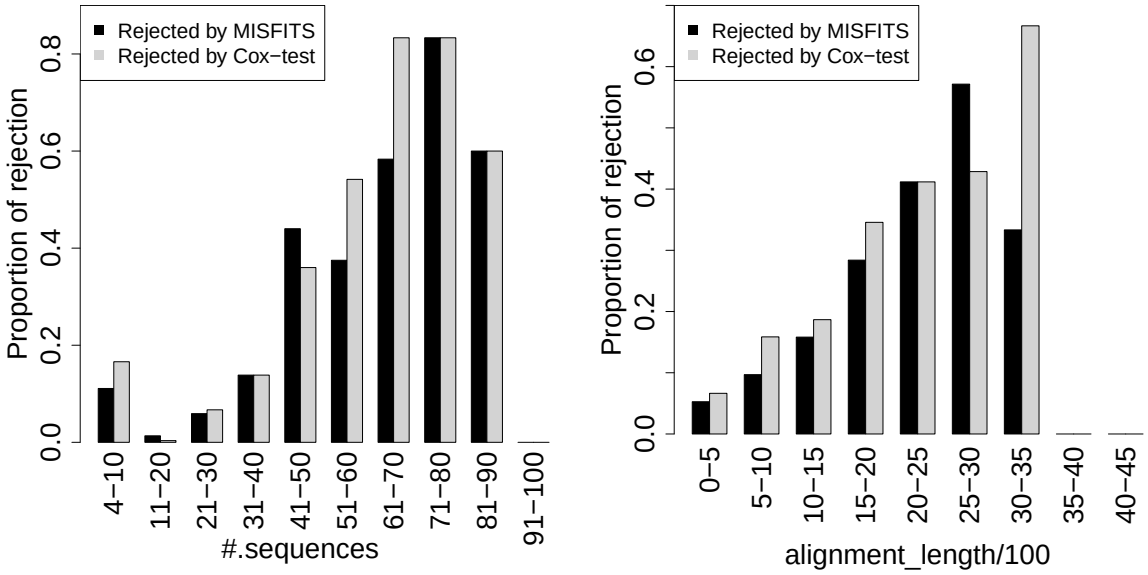
Step 5: Determine the P-value as the proportion of the generated alignments where $\sigma_j \geq \sigma_{obs}$.

Interpretation: The differences $\sigma_1 \cdots \sigma_{100}$ represent the expected difference under the null hypothesis as the model $\mathcal{M}$ was used to generate the data $\mathcal{A}_1 \cdots \mathcal{A}_{100}$. Thus, $\sigma_1 \cdots \sigma_{100}$ form the null distribution to which the observed difference σ_{obs} is compared. If the P-value attained at **Step 5** is smaller than 5% then $\mathbf{H}_0$ is rejected. This indicates that the model $\mathcal{M}$ (and the inferred ML tree) are insufficient to explain the data $\mathcal{A}$. On the other hand, a large P-value does not reject $\mathbf{H}_0$.

Table A.1: Recapitulation of the procedure to perform the Goldman-Cox test (Goldman, 1993b) of absolute model fit in phylogeny inference (see also e.g. Whelan *et al.*, 2001; Jermiin *et al.*, 2008).

	Rate				
Model[a]	One rate	I	Γ	I + Γ	$\sum_{Model}$
JC	0.07	0	0.02	0	0.09
F81	0.16	0.16	0.16	0.09	0.57
K80	0.02	0.35	0.68	0.31	1.36
HKY	0.23	1.20	3.87	3.47	8.77
K3ST	0	0.12	0.33	0.09	0.54
K3STuf	0.09	0.77	2.09	1.78	4.73
TN93ef	0.05	0.12	0.16	0.19	0.52
TN93	0.16	0.70	1.92	2.23	5.01
SYM	0.12	0.33	3.40	4.52	8.37
GTR	0.31	2.51	26.57	40.65	70.04
$\sum_{Rate}$	1.21	6.26	39.20	53.33	100.00

Table A.2: Percentages (%) of the selected models for the 4,268 PANDIT alignments going through all analyses. [a] refer to Table 1.3 for a detailed description of the models.



(a) (b)

Figure A.1: Results on PANDIT database under the selected models for the 4,268 alignments where over-represented patterns were observed and there were enough under-represented patterns to exchange with them. The barplots display the proportion of models being rejected by MISFITS (black bars) and by the Goldman-Cox test (light gray bars) with respect to the number of sequences in the alignment (a) and to the alignment length (b).

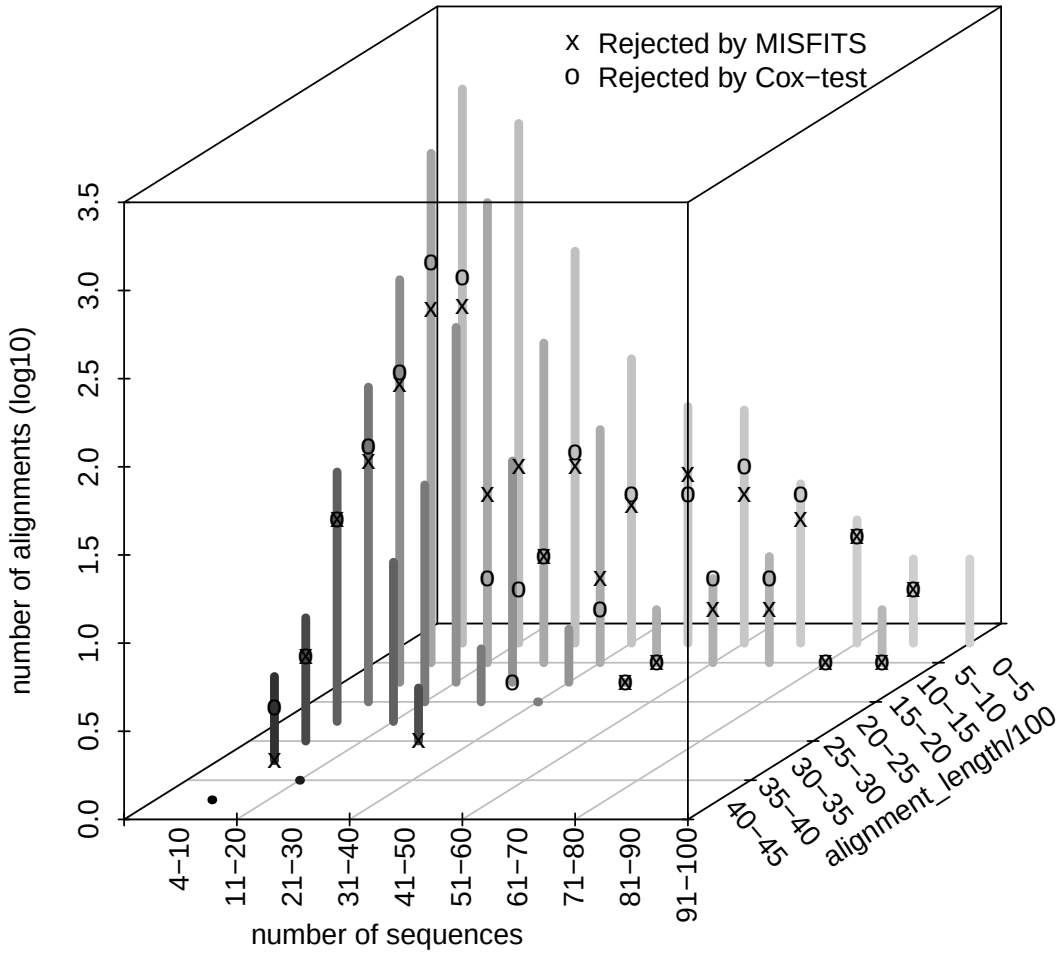


Figure A.2: Number of alignments and rejected alignments for the 4,268 PANDIT alignments where over-represented patterns were observed and there were enough under-represented patterns to exchange with them. The x-axis represents the number of sequences in the alignment, the y-axis shows the alignment length divided by 100. The bars along the z-axis displays the number of alignments (in logarithm to base 10) with certain ranges of the alignment length and the number of sequences. Points on each bar shows the number of alignments (models) being rejected by MISFITS (x) and by the Goldman-Cox test (o).

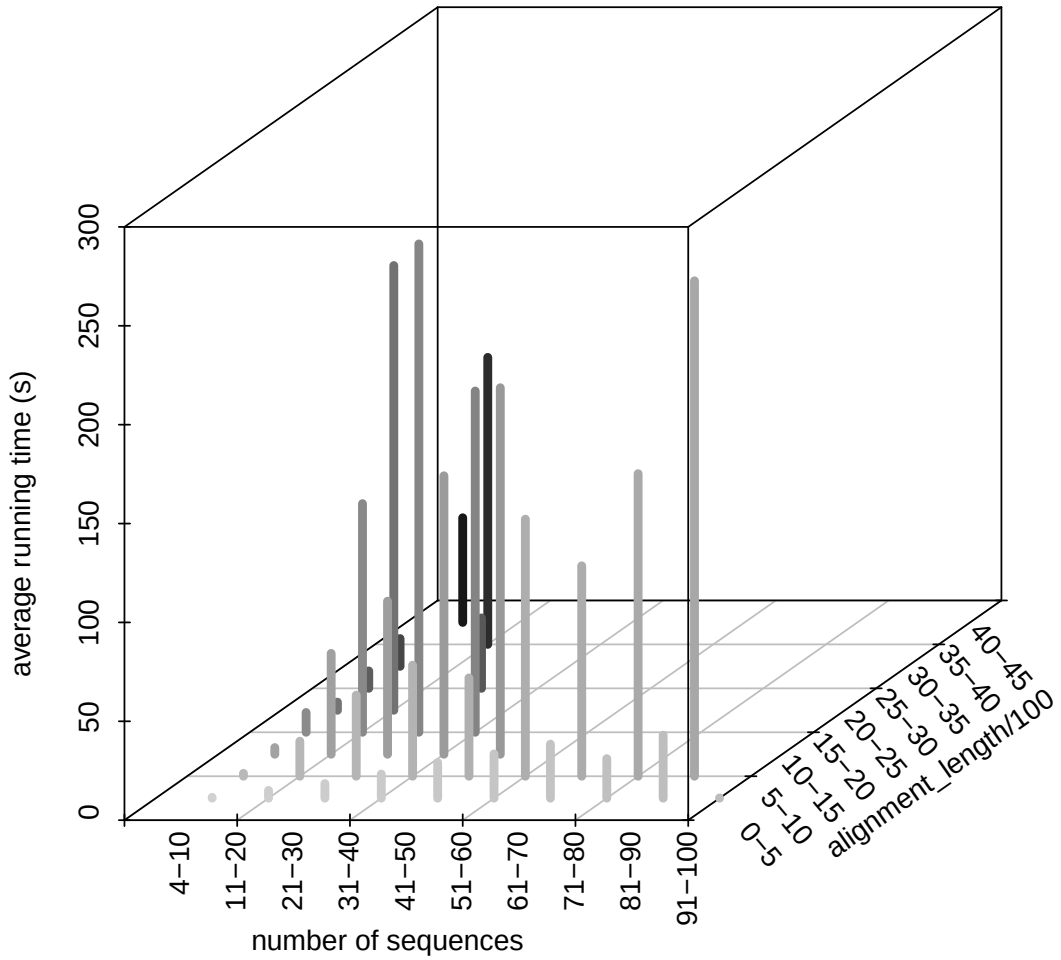


Figure A.3: Empirical computation time to compute m_0 given $\mathcal{D}^+$ and $\mathcal{D}^-$ for the 4,268 PANDIT alignments. The x-axis displays the number of sequences in the alignment, the y-axis shows the alignment length divided by 100. The bars along the z-axis displays the average running time to compute m_0 given $\mathcal{D}^+$ and $\mathcal{D}^-$ for these alignments with the corresponding ranges of the alignment length and the number of sequences.

Appendix B

Supplemental figures to Chapter 4

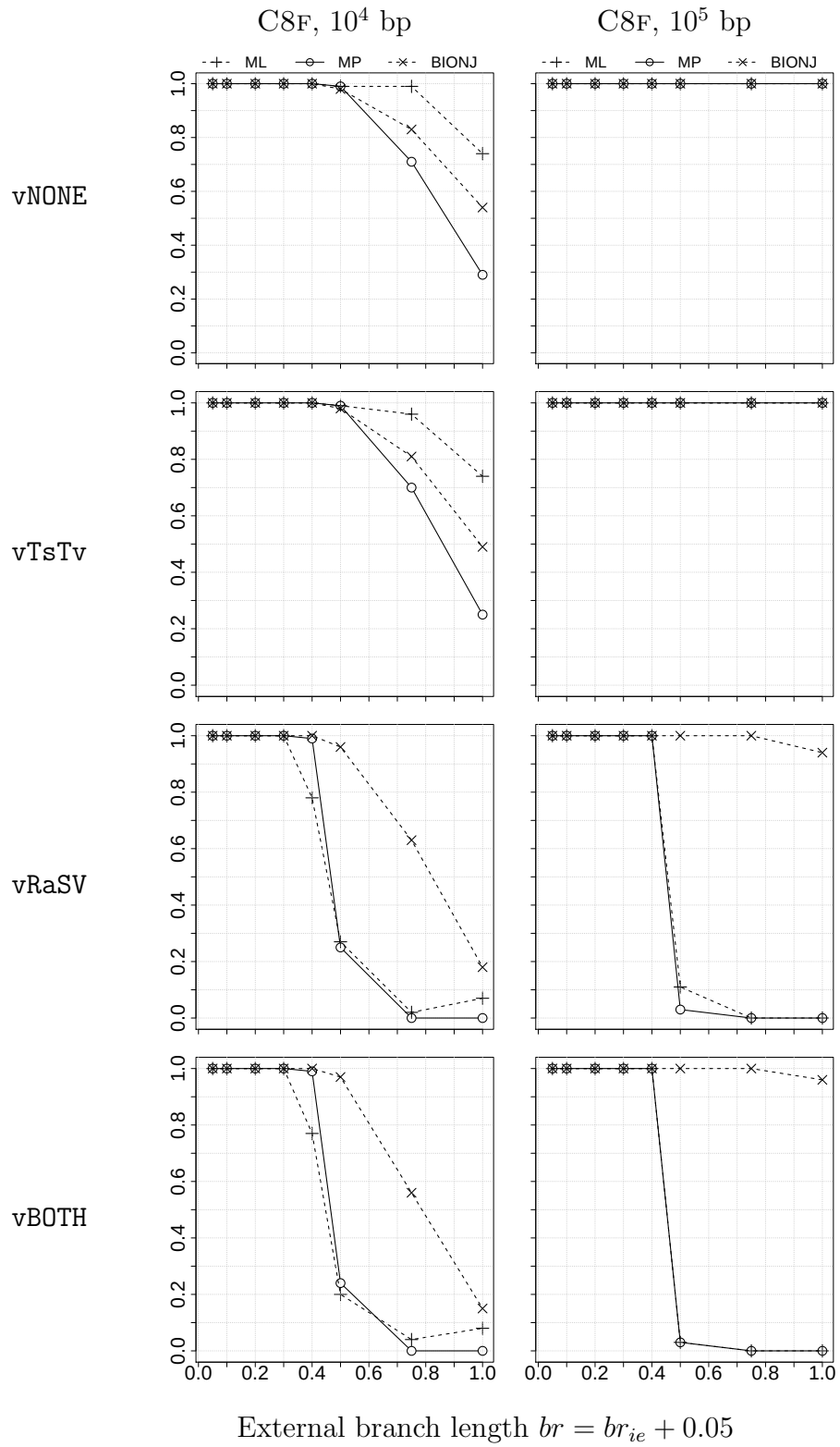


Figure B.1: Tree reconstruction accuracy for C8F. The accuracy, i.e. the proportion of alignments that yield the true tree, is shown on the y-axis for simulations with no model violation (`vNONE`, first row), with Ts/Tv violation (`vTsTv`, second row), with RaS violation (`vRaSV`, third row), and with both Ts/Tv and RaS violation (`vBOTH`, last row). The x-axis displays the external branch length br or (br_{ie}+0.05). Accuracy of ML is depicted by +, MP by ∘, and BIONJ by ×.

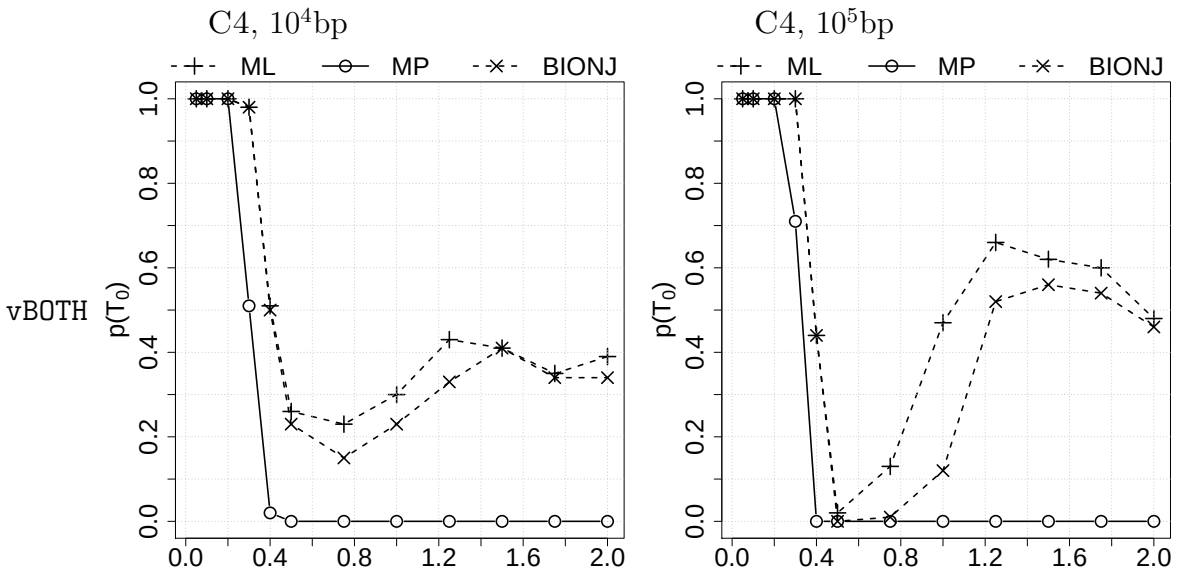


Figure B.2: Tree reconstruction accuracy $p(\mathcal{T}_0)$, i.e. the proportion of alignments that yield the true tree, is shown on the y-axis for simulations with the C4 tree with simultaneous violation of RaS and TsTv (`vBOTH`) and external branches grows to 2.0 branch length. The two columns show the results for sequence length 10^4 and 10^5, respectively. The x-axis displays the external branch length br or (br_{ie}+0.05). Accuracy of ML is depicted by +, MP by $\circ$, and BIONJ by $\times$.

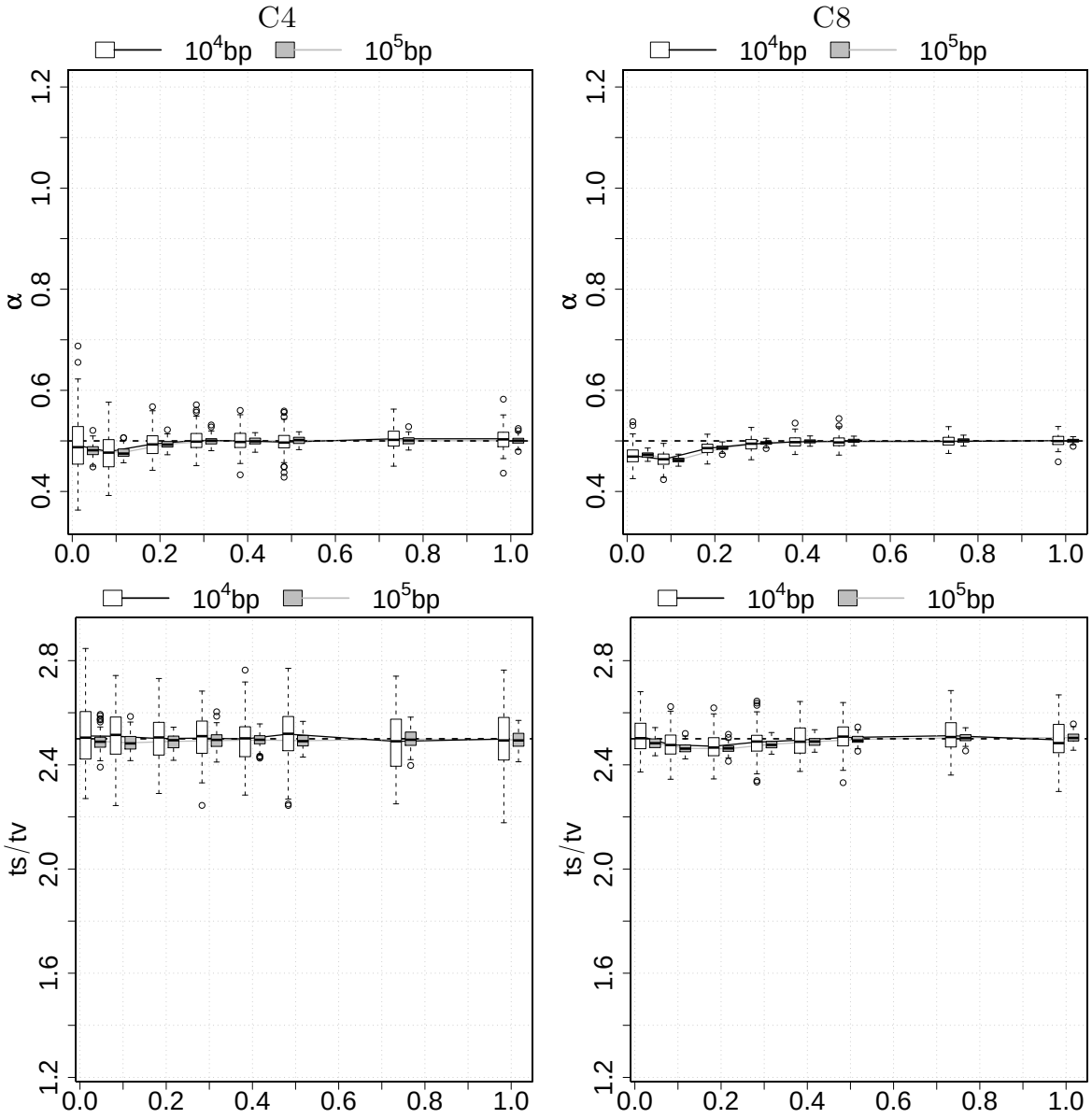
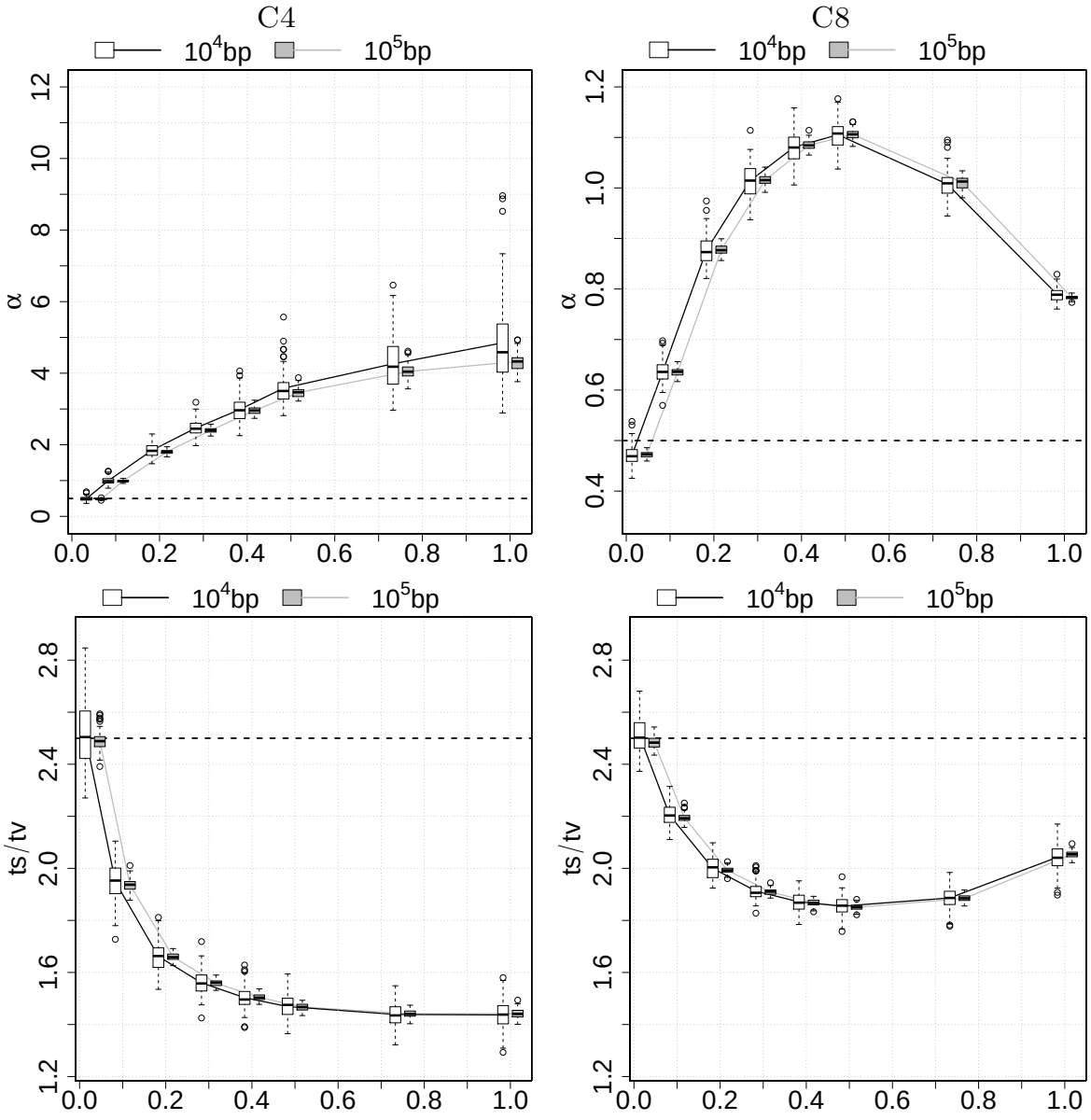


Figure B.3: Parameter estimation for C4 and C8 settings in the absence of model violation. The first and the last rows show estimation of the shape parameter α and the Ts/Tv ratio, respective. Results for the four-taxon tree C4 are presented on the left and for the C8 tree on the right. The x-axis displays the external branch length br or (br_{ie}+0.05).



External branch length $br = br_{ie} + 0.05$

Figure B.4: Parameter estimation in the presence of the simultaneous TsTv and RaS violation (`vBOTH`). The first and the last rows show estimation of the shape parameter α and the Ts/Tv ratio, respective. Results for the four-taxon tree C4 are presented on the left and for the C8 tree on the right. The x-axis displays the external branch length br or (br_{ie}+0.05).

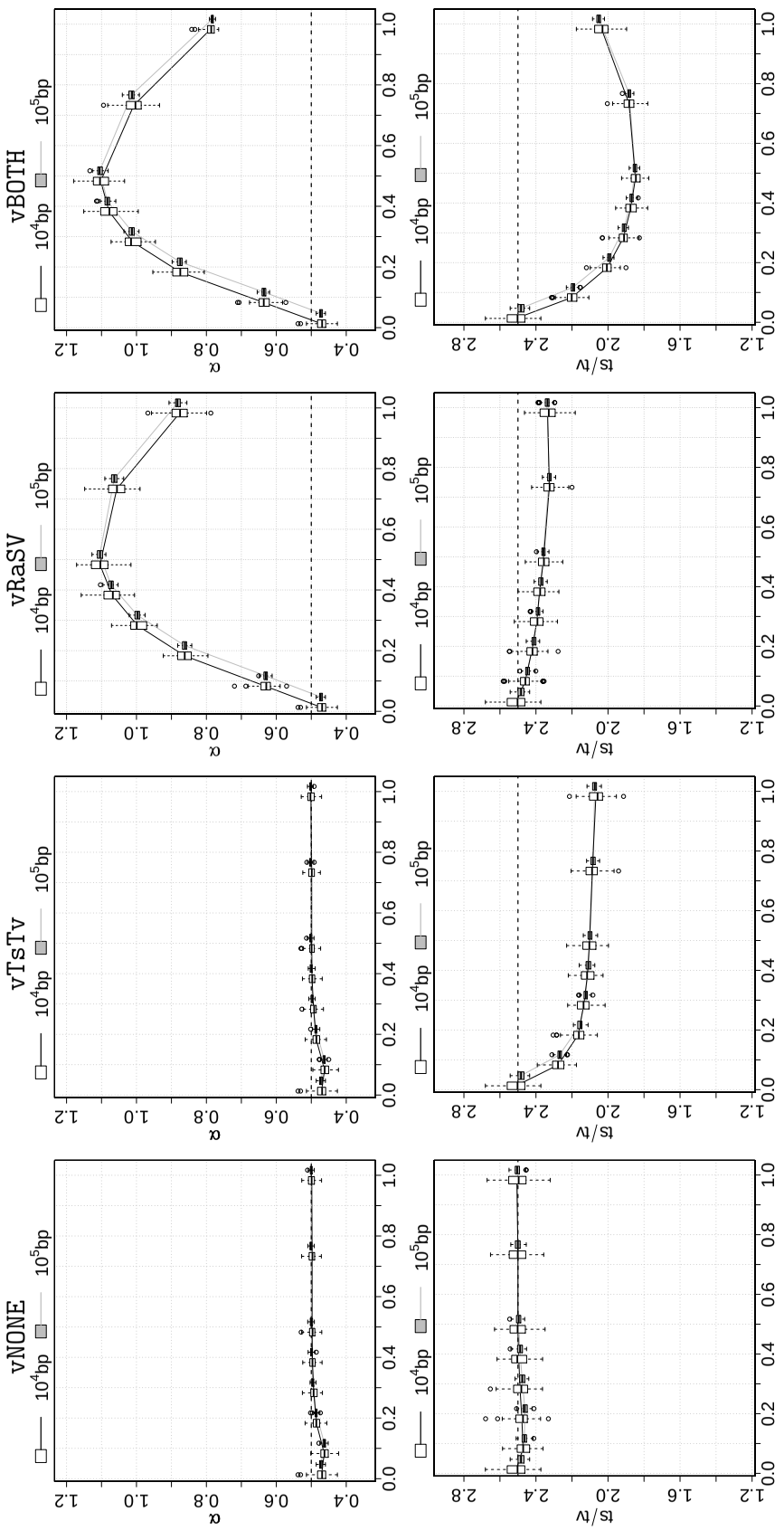


Figure B.5: Parameter estimation for the C8F tree without and with different kinds of model violation. The first and the last rows show estimation of the shape parameter α and the Ts/Tv ratio, respectively. Columns from left to right display the results for simulations with no model violation (the first column), with TsTv violation (second column), with RaS violation (third column) and with simultaneous TsTv and RaS violation (the last column).

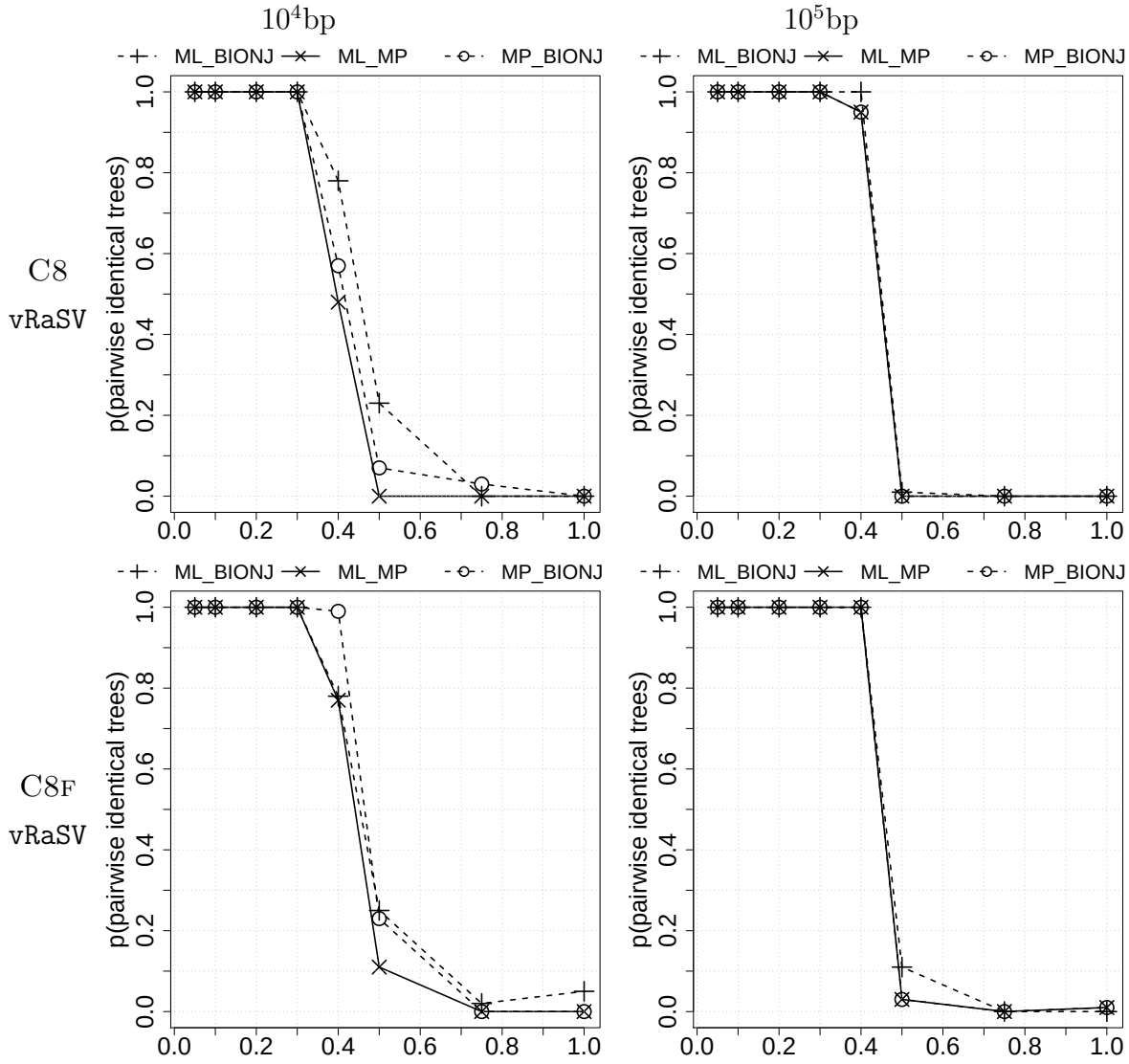


Figure B.6: Pairwise comparison between trees inferred by ML, MP and BIONJ for simulations under the `vRaSV` setting for the C8 tree (first row) and the C8F tree (last row). The first and second columns show the results for alignment length 10^4 and 10^5, respectively. The x-axis displays the external branch length br or (br_{ie}+0.05). The y-axis shows the proportion of alignments that yield the same tree topologies between: ML and BIONJ (+), ML and MP (×), and MP and BIONJ (○). As shown, all the lines drop to 0 as br exceeds 0.5. This means each of the three methods produces a different tree.

Curriculum Vitae

Contact information

Minh Anh Thi Nguyen
Center for Integrative Bioinformatics Vienna (CIBIV)
Max F. Perutz Laboratories
Dr. Bohr Gasse 9
A-1030 Vienna, Austria
Phone: ++43 +1 / 4277-24028
Fax: ++43 +1 / 4277-24098
Email: minh.anh.nguyen(AT)univie.ac.at

Date of birth March 10, 1983

Place of birth Hanoi, Vietnam

Nationality Vietnamese

Education

2001 - 2005: Bachelor in Computer Science, College of Technology, Vietnam National University, Hanoi.

1998 - 2001: High school for gifted students in Mathematics and Informatics, University of Natural Science, Vietnam National University, Hanoi.

Research experience

June 2006 - present: Ph.D. student at the Center for Integrative Bioinformatics Vienna (CIBIV), Max F. Perutz Laboratories (MFPL).
Supervisor: Univ-Prof. Dr. Arndt von Haeseler.
Topic: Goodness of fit and robustness of phylogenetic methods in light of intermittent evolution.

September 2005 - May 2006: Research and teaching assistant at the Faculty of Information Technology, College of Technology, Vietnam National University, Hanoi.

January 2004 - June 2005: Member of the Optical Character Recognition (OCR) project.
Supervisor: Dr. Nguyen Viet Ha.
Topic: Studying and applying neural network to Vietnamese hand-written character recognition.

Publications

1. M. A. T. Nguyen, T. Gesell and A. von Haeseler. ImOSM: Intermittent evolution and robustness of phylogenetic methods. Submitted to *Mol. Biol. Evol.*
2. M. A. T. Nguyen, S. Klaere and A. von Haeseler (2011) MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. *Mol. Biol. Evol.*, 28(1): 143-152.
3. Nguyen Thi Minh Anh, Dinh Viet Cuong, Ngo Tri Hoai and Nguyen Viet Ha (2005). A multiple neural network model for Vietnamese hand-written character recognition. *Proceedings of the eighth national conference on information technology and communication*, pp: 37-46. August 2005, Hai Phong, Vietnam. (in Vietnamese).

Talks

1. Cases in which parsimonious reconciliation underestimates and misinterprets gene family evolution. *Bertinoro Computational Biology 2010*, May 24-28, Bertinoro, Italy.

2. Evaluating the goodness of fit between a phylogenetic model and an alignment. *PhyloGroup 9 meeting*, December 11, 2009, Hinxton, UK.

Posters

1. M. A. T. Nguyen, T. Gesell and A. von Haeseler (2011) Imbedding additional substitutions into data to study robustness of phylogenetic methods. *Genome research in Austria (GEN-AU) evaluation conference*, May 02-04, 2011, Vienna, Austria.
2. M. A. T. Nguyen, S. Klaere and A. von Haeseler (2010) MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. *Annual Meeting of the Society for Molecular Biology and Evolution (SMBE)*, July 4-8, 2010, Lyon, France.
3. M. A. T. Nguyen, A. Fuehrer and A. von Haeseler (2007) Cases in which parsimonious reconciliation underestimates and misinterprets gene family evolution. *PhyloInformatics workshop*, October 22-24, Edinburgh, UK.
4. M. A. T. Nguyen, B. Q. Minh, A. von Haeseler and S. Klaere (2007) Il Buono, Il Brutto, Il Cattivo (The good, the bad, and the ugly in phylogenetic diversity conservation). *PhyloInformatics workshop*, October 22-24, Edinburgh, UK.

Tools developed

1. ImOSM: A C++ program to *Im*bed intermittent evolution as *O*ne *S*tep *M*utations into phylogenetic data. ImOSM is a useful tool to introduce model violation to study the robustness in phylogeny inference.

   Home page: http://www.cibiv.at/software/imosm.

2. MISFITS: A C++ program to evaluate the goodness of fit between a phylogenetic model and an alignment.

   Home page: http://www.cibiv.at/software/misfits.