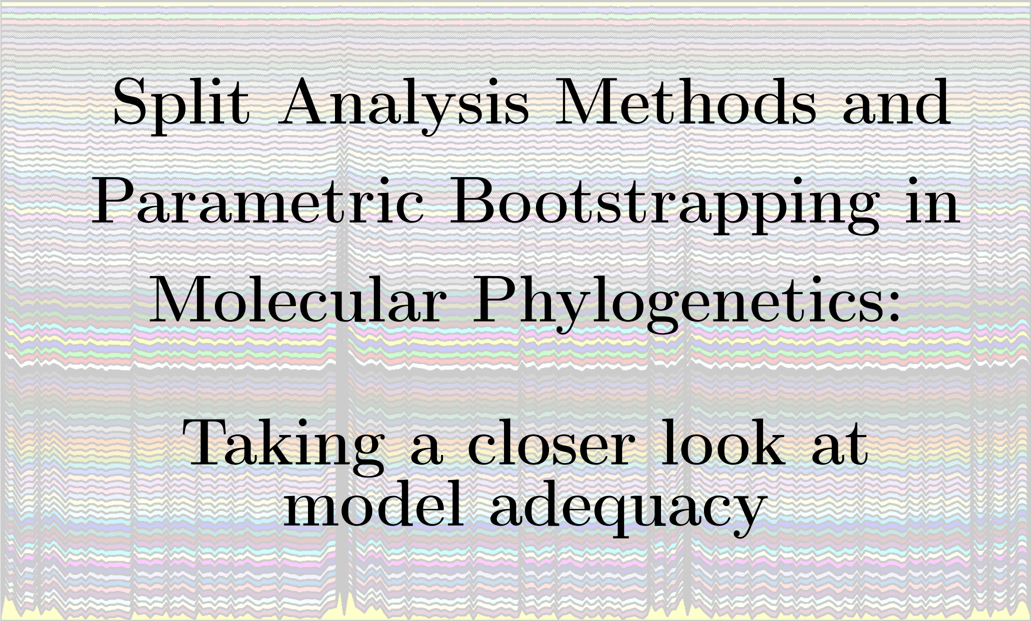
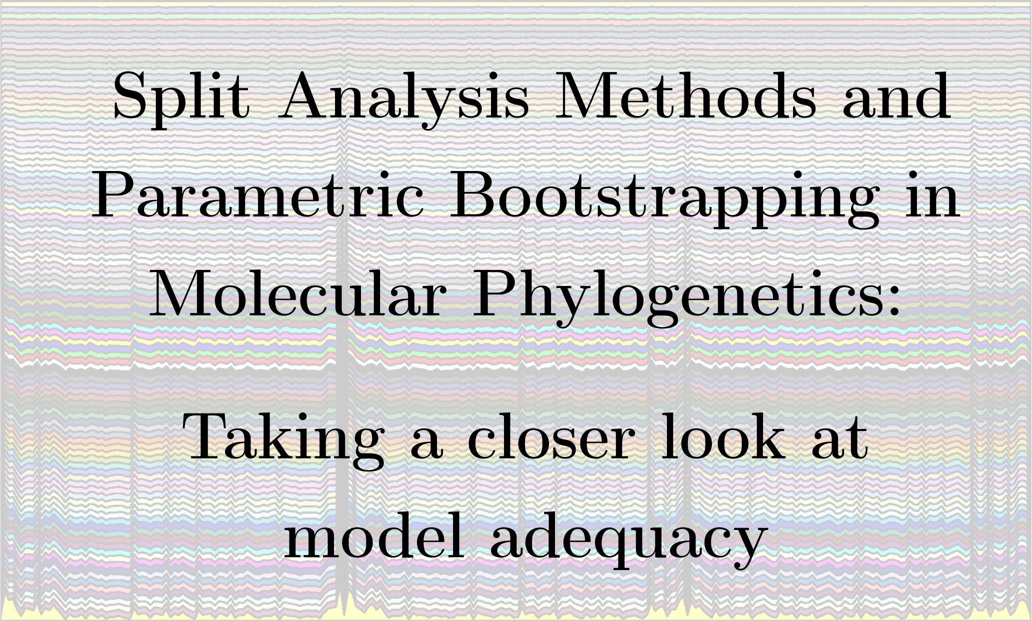


Split Analysis Methods and Parametric Bootstrapping in Molecular Phylogenetics:

Taking a closer look at model adequacy

Dissertation

Sandra A. Meid



Split Analysis Methods and Parametric Bootstrapping in Molecular Phylogenetics:

Taking a closer look at model adequacy

Dissertation

zur

Erlangung des Doktorgrades (Dr. rer. nat.)

der

Mathematisch–Naturwissenschaftlichen Fakultät

der

Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von

Sandra A. Meid

aus

Andernach

Bonn, 2014

Angefertigt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Rheinischen Friedrich-Wilhelms-Universität Bonn

FORSCHUNGS museum KOENIG

Die Dissertation wurde am Zoologischen Forschungsmuseum
Alexander Koenig (ZFMK), Bonn durchgeführt.

1. Gutachter: Prof. Dr. Bernhard Misof
2. Gutachter: PD Dr. Lars Podsiadlowski

Tag der mündlichen Prüfung: 28.04.2015
Erscheinungsjahr: 2015

"The most exciting phrase to hear in science,
the one that heralds new discoveries,
is not 'Eureka!' ('I found it!')
but rather 'hmm....that's funny..."

Isaac Asimov

FÜR ALL JENE, DIE MICH AUF DIESEM WEG BEGLEITET UND UNTERSTÜTZT HABEN.

Contents

List of Figures

List of Tables

Abbreviations

AIC	Akaike information criterion
ASRV	among-site rate variation
BIC	Bayesian information criterion
DNA	deoxyribonucleic acid
DT	decision theory
GoF	goodness-of-fit
GTR	general time-reversible
hLRT	hierarchical likelihood-ratio test
Indel	insertion or deletion of bases in nucleotide sequences
ML	maximum likelihood
MP	maximum parsimony
mRNA	messenger RNA
MSA	multiple sequence alignment
RNA	ribonucleic acid
RY	purine–pyrimidine
SRH	stationary, reversible and homogeneous

Summary

Even though the size of datasets in molecular analyses increased rapidly during the last years, undetected systematic errors as well as unsolved problems concerning the evaluation of data quality and adequate substitution model selection still persist. This not only hampers the correct analysis of these datasets but leads to undetectable effects in phylogenetic tree reconstruction.

Model-based tree reconstruction methods like maximum likelihood estimation and Bayesian inference have become the methods of choice for reconstruction of phylogenetic trees. Although maximum likelihood methods are known to be consistent if all necessary conditions are met, it depends strongly on the quality of the multiple sequence alignment and the ability of the chosen evolutionary model to reflect the underlying historical processes. This thesis addresses the assessment of model adequacy of estimated evolutionary models to multiple sequence alignments in the light of parametric bootstrapping and aims to find new methods for detection of model misspecifications with the help of split analyses.

The second chapter focuses on the influence of the number of gamma rate categories used in modelling among-site rate variation when trying to assess model adequacy using an absolute goodness-of-fit test. The analyses of simulated alignments show that the Goldmann-Cox test rejects models which were only approximated by four discrete gamma rate categories for various tree shapes and branch length setups, if they were simulated with a continuous gamma distribution. Increasing the number of discrete rate categories leads to an acceptance of model adequacy for stationary datasets and a correct detection of non-stationarity and inhomogenetity in simulated data. The results illustrate that the application of the proposed Goldmann-Cox test to evaluate model adequacy might be too strict and rigorous with empirical data, in particular for large phylogenomic datasets.

Approaches such as the Goldman-Cox test evaluate the absolute fit of data and model but, do not deliver a deeper insight into the structure of the misfit. The third chapter presents the visualisation of overrepresented splits within splits graphs, which provides a good tool for gaining an overview of possible patterns and contradictory signal or noise within datasets. The analysis of these split

residuals, observed by comparison to parametric bootstrap datasets based on the estimated models can help to gain a deeper insight into model adequacy. Highly overrepresented splits can give hints whether heterotachy applies or non symmetric substitution processes.

The fourth chapter aims to define a new split weighting scheme by formalising aspects like 'contrast of character states' or 'character state homogeneity' within split subsets. Splits which are detected by the proposed SAMS (Splits Analysis MethodS) algorithm are re-evaluated for a more objective and formal split weighting. A comparison of the published and the new approach showed that the developed weighting scheme delivers reasonable results but needs further improvement. The development of a new GUI offers a much more capable tool to perform a split analysis and visualise the results. The shape of a visualised split spectra can indicate, whether a dataset delivers a clear split signal or if there is a lot of noise present.

Essentially, all models are wrong, but some are useful.

George E. P. Box

1. Introduction

The theory of evolution is still subject of intense discussions, debated by numerous parties: scientists, religious persons and as well as by everyday people from all over the world. The fact that evolutionary processes like natural selection can explain the diversity of life on earth is often overshadowed by questioning how likely this development was and is, and how subtle complex structures, e.g. eyes or chloroplasts (which conduct the process of photosynthesis in plants) evolved.

Someone once told me that he struggles with accepting that evolution worked out like it can be observed today, because he thinks this is as unlikely as being inside of a room, opening the window and a wind gust comes in and builds a house of cards out of a pile, right off the bat. But this allegory is flawed. It misses the important part of the concept of evolution: the evolution of creatures, diversifications one by one, the ones which lead to higher fitness of individuals within populations and are therefore established and the majority, which vanishes. It is not the case that organisms adapt to a changing environment, which implies that evolution is progressive, directional or even deliberate. Evolutionary theory operates exactly the other way round. Mutations happen by chance, which can coincidentally lead to a better adaptation or to an increase of attractiveness for other members of their species, the organisms which inherit this change will reproduce, the others might cease. Evolution has no goal, it is an ongoing process driven by chance and necessity and we are only able to catch a glimpse of a short fragment of it.

'Nothing in biology makes sense except in the light of evolution.'
(Theodosius Dobzhansky, 1973)

If we want to use the pile of cards allegory, we are standing in this room right now and the house of cards is already existing. The question is not, whether it can happen, since it already has happened. The crucial question is therefore how all of this has happened. Merely all the intermediate stages are indiscernible. Because of this, because evolution 'cleans up after itself', it is tough to grasp and to figure out the true historical singular processes. In fact, it is a major challenge to unravel the history of this process, considering that the majority of steps is inaccessible. There is less information left compared to the information which has vanished. This makes it clearly impossible to prove what has happened. Scientists simply search for a reliable and comprehensive theory to explain the recent situation in the best possible way. Theories are of course themselves always subject to update and improvement. Evolutionary theory has been adapted over the years, but at its core it remains unchanged. Within the field of phylogenetic research, scientists try to reconstruct the evolutionary process and reveal phylogenetic relationships of e.g. species based on the information which is accessible nowadays. Therefore, morphological or molecular models of evolution are used to master this balancing act and to compensate the missing information in various ways. Principles for scientific methods were presented (Tillyard, 1921; Hennig, 1950; Cain & Harrison, 1960; Sokal & Sneath, 1963; Edwards & Cavalli-Sforza, 1963; Edwards & Cavalli-Sforza, 1964; Hennig, 1966; Sokal, 1966) in order to determine ancestry out of phylogenetic relationships based on characters. Steps involved in phylogenetic analysis include data acquisition of morphological characters or molecular sequences and their evaluation, whether characters of different organisms are comparable (homology, orthology). These sets of characters can then be arranged to morphological character matrices or in case of molecular data, multiple sequence alignments (MSA).

Table 1.1: Example of a multiple sequence alignment based on nucleotides. Four sequences (rows) are arranged within a matrix. Every site (column) acts as character that can be checked for matches (same state) or mismatches (different states), in order to infer evolutionary relationships between the sequences.

taxon 1	A	C	G	T	...
taxon 2	A	T	G	T	...
taxon 3	A	T	C	T	...
taxon 4	A	T	C	A	...

Morphologists study, for example, the anatomy of organisms, detect certain structures with associated functions and compare these with other species or groups. This is a complex task due to missing ancestral or intermediate stages. Moreover, decisions shall be made in an objective way leading to reproducible results (Aichele & Schwegleb, 2008). A corresponding position, structure or functionality of characters or of their development can indicate homology, a common origin and possibly a common ancestry. Then, if homologous, these characters can be used for inferences on the relationship of e.g. species.

Molecular phylogenetic analyses, on the other hand, use the simplicity of the molecular state space (A, C, G, T for nucleotide sequences, the classical 20 amino acids for amino acid sequences) that allows a clear formal description of evolutionary substitution processes used in molecular phylogenetics. This formal description has been extensively studied and forms the basis of model-based tree reconstruction methods. A formalisation of evolutionary processes is as well important in understanding the power of tree reconstruction methods. Moreover, it can help to compensate missing information caused by extinction while reconstructing ancestral states. All together, this is certainly a strength of the molecular approach. However, the finite character state space is also a practical weakness as the probability of overlooking homoplasy, sequences sharing the same character states by chance which were not present in the last common ancestor, is high. This is due to the stochasticity of the substitution process. Multiple substitutions and finally the saturation of the evolutionary substitution processes lead to a high similarity by chance of non-homologous characters. Lots of efforts have been put into the investigation of these problems resulting in heaps of algorithms and software packages to properly account for unobserved patterns of homoplasy.

Many phylogenetic tree reconstruction methods analyse datasets site by site ('columns' of the MSA as shown in table 1.1) as independent character. Alternatively, distance based methods reduce site differences to a scalar value of difference between sequences. A distance is therefore a measure of change from an ancestral status to the present one. Maximum parsimony (MP) counts the number of character changes and tries to find the tree with the least number of implied changes. It is often claimed, that MP is a 'model free' method, but in fact, this method makes implicit assumptions concerning the character state space and character transformations (Steel & Penny, 2000; Tuffley & Steel, 1997; Steel, 2002). Moreover, MP is known for its consistency problems (Felsenstein, 1978; Hendy & Penny, 1989).

Maximum likelihood (ML) (Fisher, 1958; Edwards & Cavalli-Sforza, 1964; Neyman, 1971; Felsenstein, 1981) and Bayesian tree reconstruction methods (Rannala & Yang, 1996) are known to be the most efficient and accurate methods to analyse phylogenetic datasets (Ogden & Rosenberg, 2006a). Both are based on statistical models of character substitutions (nucleotides or amino acids) trying to capture the underlying sequence evolution by assigning rates at which bases of one type change into bases of another type. While the simplest model of DNA sequence evolution assumes equal substitution rates and base frequencies (JC69 model) (Jukes & Cantor, 1969), other models distinguish between transitions and transversions (K80 model) (Kimura, 1980) or allow unequal base frequencies at equilibrium (F81 model) (Felsenstein, 1981). The HKY85 model (Hasegawa *et al.*, 1985) allows both, variation of base frequencies and differentiation between transitions and transversions. Extended models like the T92 model (Tamura, 1992) try to handle GC-content biases or distinguish between two different types of transition, such as the TN93 model, (Tamura & Nei, 1993). The most general but also most complex model allows six different substitution rate parameters and variation of base frequencies (General Time-Reversible, GTR) (Tavaré, 1986).

In all these approaches, various sets of parameters are tested to find the one model, which explains the outcome best. Therefore, different parameters such as the substitution rates, branch length and of course the topology are estimated, calculated and optimised. It is like having two dice, one with values from 1 to 6 (D6), and another one with values from 1 to 12 (D12). If you know, that one die was rolled two times and the summed result is for example '12', then it is much more likely, that the D12 was used, because for the D6 only one combination, two times rolling a '6' ($\frac{1}{36}$, ≈ 0.028), can lead to this result. The likelihood of the D12 ($\frac{11}{144}$, ≈ 0.076) is thus much higher for the present result and therefore the maximum likelihood.

Maximum likelihood methods are known to be consistent if certain conditions are met (Fisher, 1922; Chang, 1996). That means, that the reconstruction of the true tree is guaranteed if the sequences are infinitely long and evolved under the assumed evolutionary model that is used for the tree reconstruction. Empirical data are of finite length. But also for limited data, ML can reconstruct the correct tree (Felsenstein, 1978) and it has been shown that the ML method on average requires less data than other consistent methods such as minimum evolution (Steel & Penny, 2000; Tuffley & Steel, 1997; Steel, 2002). This is referred to as the efficiency of ML.

Model choice itself remains a critical step (Kelchner & Thomas, 2007). A model is by its nature a limited representation for a certain purpose and can never

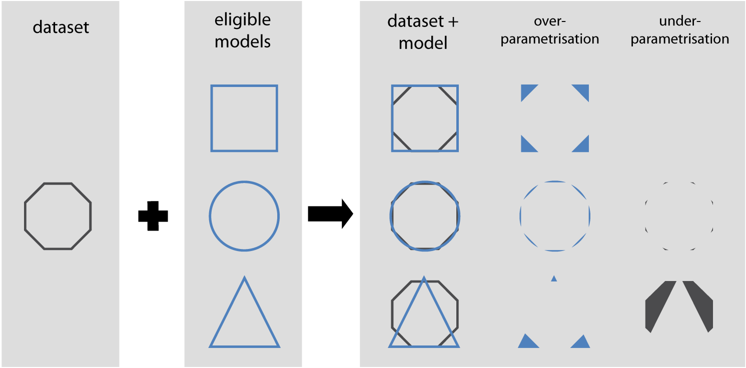


Figure 1.1: Analogy of model fitness: A dataset (black octagon) is compared to eligible models (blue square, circle and triangle). If the dataset can be explained exactly by an adequate model, it would have the same shape and there would not occur any over- or under-parametrisation. If the best fitting model out of the eligible models, in this case the circle, can explain the underlying dataset well, then it covers it without causing to much over- or under-parametrisation, it is adequate for a certain propose.

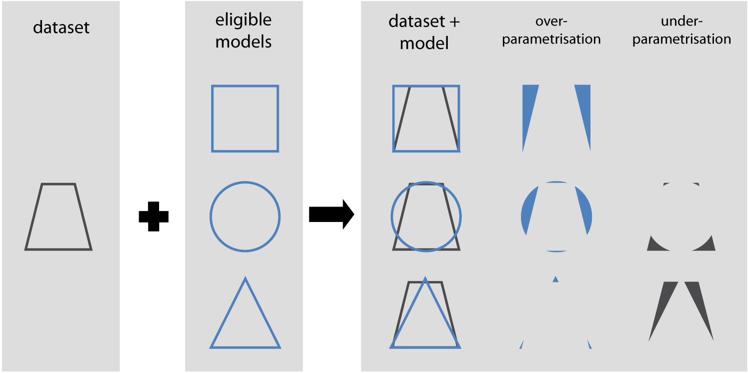


Figure 1.2: Analogy of model fitness: If unlike for the dataset in Fig. 1.1 there is no model which can explain the dataset (black trapeze) without over- or under-parametrisation, it might be the case, that there is no adequate model available.

cover every detail of the data. As for example, a topological map is a good model of a landscape for orientation but, it can never give for instance the impression of how it really looks like and how many people are living there. As visualised in Fig. 1.1, if a model out of all possible models captures the historical background with neither being too simple nor complex relative to the underlying truth, it is adequate for a dataset. In comparison, if there is no model which adequately fits (Fig. 1.2), a chosen model can become over- or under-parametrised. Over-parametrisation, on the one hand, can lead to an unnecessary sampling variance (stochastic error) which may affect phylogenetic accuracy (Cunningham *et al.*, 1998). Under-parametrisation, on the other hand, can cause more severe bias (Huelsenbeck & Rannala, 2004; Lemmon &

Moriarty, 2004; Brown & Lemmon, 2007) and leads to systematic error, which cannot be revised by just adding more data. Even worse, the more data is included, the higher the confidence for incorrect results (e.g., long-branch attraction (Felsenstein, 1978; Hendy & Penny, 1989)) can be achieved (Swofford *et al.*, 2001).

To take among-site rate variation, ASRV, (Sullivan & Swofford, 2001) into account, evolutionary models use gamma-distributed site rates (Γ) (Yang, 1994). Due to computational limitations, the continuous gamma distribution is approximated by a discrete distribution using a fixed number of rate categories (*ncat*). This partitioning results in *ncat* categories of uniform weight ($1/ncat$) to which the sites are equally allocated. The rate for each site of a category is then represented by the mean or median rate of all sites of this category. The approximation improves the more categories are used. However, usually four categories are applied since it has been proposed that this is sufficient (Yang, 1994). Furthermore, reducing the number of gamma rate categories drastically accelerates the computation of an ML analysis, which is proportional to the number of used categories (Jia *et al.*, 2014). Additionally, a proportion of invariable sites (I) can be estimated which leaves the remaining variable sites with gamma-distributed rates (Γ+I) (Steel *et al.*, 1993; Waddell & Penny, 1996).

Although the best fitting of all possible models is estimated (relative goodness-of-fit) (Kelchner & Thomas, 2007), this does not imply that the model fits well (absolute goodness-of-fit) (Gatesy, 2007). All substitution models mentioned above assume that the aligned nucleotides evolved under stationary, time-reversible and homogeneous (SRH) conditions (Jayaswal *et al.*, 2005; Ho *et al.*, 2006). This includes that the sites are independent of one another and that the codon structure of protein-coding sequences does not have an impact on substitution processes. Moreover, it is assumed that the processes are homogeneous along and as well across sequences, having a constant evolutionary "speed". For example, datasets that include different genes, these are assumed to share the same history. Violations of these conditions can lead to biased results (Felsenstein, 1978; Huelsenbeck & Hillis, 1993; Yang *et al.*, 1994; Swofford *et al.*, 2001; Ho & Jermiin, 2004; Jermiin *et al.*, 2004).

Nevertheless, it had been claimed that ML based methods are robust with respect to most model violations (Sullivan & Swofford, 2001). In fact, a model does not have to fit perfectly, but it has to be sufficient for the purpose of an unbiased phylogenetic reconstruction. Anyhow, it became clear that in empirical data substitution processes by themselves evolve which makes the proper formal description additionally harder and error prone (Wu & Susko, 2009; Kolaczkowski & Thornton, 2008; Whelan, 2008; Zhou *et al.*, 2007).

> *'We do not like to ask, 'Is our model true or false?', since probability models in most data analyses will not be perfectly true. [...] The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'.'*
>
> *(Gelman et al., 2013)*

In phylogenetic analyses, bootstrapping (Felsenstein, 1985; Hillis & Bull, 1993; Efron *et al.*, 1996), i.e., resampling the data, is often interpreted as measure of accuracy. Though, if the underlying model does not fit well, the results will become unreliable. 'Bootstrap support of 100% is not enough; the tree must also be correct' (Phillips *et al.*, 2004). In fact, bootstrapping is a measure of repeatability (Felsenstein, 2004), nothing more. This easily explains the fact that different phylogenomic studies infer maximal bootstrap support, but for incongruent trees (compare, e.g., Pick *et al.* (2010), Schierwater *et al.* (2009), and Dunn *et al.* (2008)).

Bootstrapping is not just a method of testing tree inference, it can also be used as a method of testing model parameters which were estimated to describe how observed MSAs may have evolved in a statistical way. For ML analyses, the Goldman-Cox test was proposed (Goldman, 1993b; Whelan *et al.*, 2001) to evaluate the absolute goodness-of-fit of data and model using so-called 'parametric bootstrapping'. Parametric bootstrap datasets are artificial datasets, generated using sequence simulation based on the estimated relative goodness-of-fit model parameters of the analysed (empirical) dataset. Thus, they represent a perfect fit of model and data. These bootstrap replicates can then be compared to the original dataset, for which the model was estimated. The Goldman-Cox test calculates and compares the entropy of site patterns (i.e., the set of character states of a site) of all datasets, the original one and the parametric bootstrap datasets. For example, the alignment in table 1.1 contains four patterns, 'AAAA', 'CTTT', 'GGCC' and 'TTTA'. If the degree of entropy of the dataset fits within 95% of the distribution of entropy-degrees of the bootstrap datasets, the model is statistically accepted as adequate for the dataset.

Table 1.2: Example of two different splits within an MSA based on nucleotide sequences. Five taxa are splitted twice into subsets. The split within the table on the left (indicated by a blue line) separates the first two sequences from the rest (1 2 | 3 4 5), the second split (1 2 3 | 4 5) in the table on the right (green line) groups the sequences of taxon 1,2 and 3 into a distinct subset, separated from taxon 4 and 5, the other subset.

split 1 2 | 3 4 5

taxon 1	A	C	G	T	...
taxon 2	A	C	G	T	...
taxon 3	A	T	G	T	...
taxon 4	A	T	C	T	...
taxon 5	A	T	C	A	...

split 1 2 3 | 4 5

taxon 1	A	C	G	T	...
taxon 2	A	C	G	T	...
taxon 3	A	T	G	T	...
taxon 4	A	T	C	T	...
taxon 5	A	T	C	A	...

Another possibility to evaluate MSAs is to analyse their possible splits. A split is a partitioning (bipartition) of sequences of an MSA into a pair of distinct subsets (A|B) of the complete sequence set (A∩B). In a phylogenetic context, splits are distinguished by the occurrence of e.g. nucleotides (character states) within sites (characters). In table 1.2, two splits are visualised with lines separating the subsets. The first site of the MSA shows the pattern 'AAAAA', which is an invariable site and 'counts' (delivers information of phylogenetic interest) for neither of the two splits. The second site 'CCTTT' supports a first split (12 | 345). This holds also for the third site 'GGGCC' supporting a second split (123 | 45). The last site indicates a third split (1234 | 5), which defines a terminal branch of a phylogenetic tree leading to taxon 5. Such a split is called a 'trivial' split. It does not provide any information relating to a possible topology, because it is always present and has only impact on the branch length within a possible phylogenetic tree including this taxon. For a set of n sequences, there are 2^{n-1} possible bipartitions. Per definition, for one split of all possible splits, one of the subgroups is empty and the other one contains all sequences (invariable sites). Furthermore, there are n trivial splits. The remaining $2^{n-1} - n - 1$ splits correspond to bifurcations of possible topologies. These are the most interesting ones, since they may represent phylogenetic information. If two splits A|B and C|D are compatible, i.e., if A∩C, A∩D, B∩C and B∩D are empty, this set of splits (A|B and C|D) fits on a topology.

First approaches based on splits were developed based on the distance method of split decomposition (Bandelt & Dress, 1992), and spectral analysis (Hendy & Penny, 1993; Hendy *et al.*, 1994), closely related to the Hadamard conjugation. The Hadamard conjugation is a transformation of the split spectrum which allows to correct state changes by including evolutionary models. Lento *et al.* (1995) used this approach to filter out conflicting signal. However, the processing effort grows exponentially with the number of taxa and is too computationally intensive for more than 30 sequences.

1.1. Aim of the thesis

Since substitution models can only be rough estimates of the underlying evolutionary processes, selecting the models with relative best fit does not guarantee that the selected model fits the empirical data sufficiently well (Gatesy, 2007). Covarion evolution (Fitch & Markowitz, 1970; Penny *et al.*, 2001) or heterotachy (differential evolutionary rates among organisms) (Lopez *et al.*, 2002; Sims *et al.*, 2009; Zhong *et al.*, 2011), which result in non-symmetric substitution processes and processes that have a high variation throughout the tree, are among possible sources leading to model misspecification. As consequence, if the model does not sufficiently describe the underlying history, tree reconstruction can become unreliable. Therefore, it is vitally important to develop reliable methods of testing model fit and model adequacy.

Chapter 2 addresses the assessment of adequacy of estimated evolutionary models using parametric bootstrapping. In particular, it focuses on the influence of the number of gamma rate categories used in modelling among-site rate variation (ASRV). Conventionally, a discrete gamma model with mostly four rate categories to approximate a gamma distribution is considered to be sufficiently complex (Yang, 1994). The proposed Goldman-Cox test of model adequacy is applied to various simulated datasets. The datasets are based on different tree shapes and branch length setups, either based on a stationary or non-stationary composition. Further, all of these datasets are simulated using a continuous gamma distribution or a discrete gamma distribution with four rate categories. All datasets are analysed by ML using different numbers of rate categories to approximate the gamma distribution. The aim is to analyse, whether the Goldman-Cox test provides reliable results for model adequacy of stationary datasets and whether the reliability increases with the number of used rate categories.

Visualising results of model adequacy tests comparing patterns for phylogenetic data is a complex task because of the high dimensionality of the character state space. For instance, for nucleotide MSAs and a dataset with n taxa there are 4^n possible site patterns. The Goldman-Cox test, which is applied in **chapter 2**, boils the variance of patterns down to a single number, which is then compared to those within parametric bootstrap datasets. Besides discarding a lot of information, the method delivers only an answer whether or not the estimated model is rated adequate. If the model is rejected (i.e., not adequate), there is no possibility to analyse putative causes and what would improve the MSA in order to find an adequate model.

Therefore, **chapter 3** aims the identification of possible causes why a model does not fit adequately. It is common practice in regression analysis to propose a model with some kind of residual diagnostics. This could be applied as well in phylogenetics. It corresponds to the comparison of observed pattern frequencies to those expected under the model (the model in this sense includes the tree, branch lengths and the model of nucleotide substitution). However, a comparison of site patterns is difficult to visualise with a growing number of sequences. Methods using split spectra and the Hadamard conjugation (Hendy & Penny, 1993; Hendy *et al.*, 1994) are only applicable for a small number of taxa as well, because split analysis also requires an exponentially growing processing effort. This can be solved by using purine–pyrimidine (RY) splits, for which the nucleotide character states are recoded to a two-letter character space, purines (A or G) and pyrimidines (C or T). This decreases heavily the computational effort of split analysis. Furthermore, splits can also be used to visualise incongruent signal present within a dataset via split networks, which provide a good compromise between looking at only one topology or all possible patterns. If a chosen model for an empirical dataset is correct, the split spectra should be similar to the ones obtained from data simulated with the same model specifications including tree topology, substitution rates and modelling of rate heterogeneity (parametric bootstrapping). Therefore, simulated and empirical datasets are analysed with an ML implementation and thereby estimated models are used for generating parametric bootstrap replicates. The split spectra of the original datasets and their corresponding bootstraps are analysed and compared. Splits, which occur more often in the original dataset than in every bootstrap dataset are declared as overrepresented. If splits occur less often in the original dataset than in every bootstrap dataset, this split is marked as underrepresented. Over- and underrepresented splits are then each transferred to a new MSA according to the number of their deviating occurrences. The new alignments are visualised within Neighbor-Net networks for further analyses in order to find out how chosen models may misfit the underlying evolutionary processes.

Reducing the character state space from patterns within a dataset by recoding to a RY code space for split analysis, as applied in **chapter 3** allows a much easier handling and a clear arrangement within split spectra or split networks. Although this is a useful approach, the recoding balances base composition bias (systematic error) and bias caused by heterotachy (Sims *et al.*, 2009) and is therefore more likely to be consistent with evolution under globally SRH conditions (Jayaswal *et al.*, 2005; Ho *et al.*, 2006). Even though this method decreases the computational effort, it ignores important 'challenges' which should be included while studying how to assess model adequacy.

In **chapter 4**, a new approach is described to re-evaluate the splits found by the SAMS algorithm for a more objective split weighting. SAMS (Splits Analysis MethodS) (Mayer & Wägele, 2005) considers all nucleotide character states, but unlike the Hadamard transformation, the software analyses an MSA only for 'observed' splits. This decreases the computationally effort significantly. Unfortunately, the weighting of splits in SAMS lacks a formal foundation. Limiting parameters have default values, but they can be freely adjusted. This study aims to define a new split weighting scheme by formalising aspects like 'contrast of character states' or 'character state homogeneity' within split subsets. With three different simulation setups the published and the new approaches are tested and compared.

Unter allen menschlichen Entdeckungen
sollte die Entdeckung der Fehler die wichtigste sein.
Stanisław Jerzy Lec

2. Are Four Categories Enough? Rethinking the Discrete Gamma Distribution Model in Assessing Model Adequacy

2.1. Introduction

The use of statistical models of nucleotide substitution which take biochemical processes into account strongly increased the efficiency and adaptability of methods for phylogenetic tree estimation. While working with data in limited supply, sampling error was a major concern. Nowadays, large scale datasets are often composed of numerous genes which then helps to gain control of statistical errors. Consequently, systematic biases come to the fore. Since different subsets of datasets, i.e. genes or domains, may have evolved under different conditions, the challenge is to find evolutionary models, which are able to handle highly heterogeneous data across alignments or taxa.

When analysing a dataset, different issues have to be addressed. Maximum likelihood (ML) methods are known to be principally consistent. Nevertheless, these methods strongly depend on the quality of the given multiple sequence alignment and the properties of the chosen evolutionary model to reflect the underlying historical evolutionary processes which led to the observed data.

The first proposed models trying to interpret the phylogenetic information existing within a sequential alignment addressed the rates of substitution and distribution of base frequencies. Jukes and Cantor introduced the JC69 model (Jukes & Cantor, 1969) which assumes equal substitution rates and base frequencies. The most general but also most complex model allows six different substitution rates and equal base frequencies (General Time-Reversible, GTR; Tavaré (1986)). The more degrees of freedom a model permits, the better it is able to describe the data, but the more data is necessary to achieve accurate results. Moreover, if a model is chosen which ignores important biological processes the results can become biased (systematic errors) (Swofford *et al.*, 2001) and lead to effects like long-branch attraction (Felsenstein, 1978; Hendy & Penny, 1989). Otherwise, over-parametrized and therefore too complex models can cause stochastic error (Cunningham *et al.*, 1998).

Besides substitution rates and base frequencies, among-site rate variation (ASRV) can be taken into account. To allow a correction for unequal rates across sites for ML analyses, the discrete gamma distribution (Γ) has been introduced by Yang (1994). Additionally, a proportion of sites can be estimated as invariable (I, p_{inv}= proportion of invariable sites) which leaves the variable sites with Γ-distributed rates (Γ+I; (Steel *et al.*, 1993; Waddell & Penny, 1996)). Using a discrete gamma model with a fixed number of rate categories to approximate the gamma distribution has been proposed and is usually applied with four categories (Yang, 1994).

Tools like ModelTest (Posada & Crandall, 1998), MrModelTest (Nylander, 2004) or ModelGenerator (Keane *et al.*, 2006) select the best fitting model from a predefined set with the help of relative model selection methods (Kelchner & Thomas, 2007), using different criteria (Akaike information criterion; AIC; (Akaike, 1974)), Bayesian information criterion (BIC; (Schwarz, 1978)), decision theory (DT) and hierarchical likelihood-ratio test (hLRT)).

Evolutionary models are only averaged approximations trying to capture evolutionary processes. It is frequently the case, that there is no fitting (absolute goodness-of-fit) model for a wide range of datasets. 'Given the simplicity of most models, it is possible that model selection in modern systematics is analogous to an overweight man shopping in the petites department of a women's clothing store. A particular garment might fit the portly man best, but this does not imply a good overall fit.' (Gatesy, 2007). Nevertheless, the methods which are currently in use will propose the relatively best fitting model (relative goodness-of-fit). Thus, even the best chosen model may not adequatly represent the dataset and in this case, the results of the applied methods can become untrustworthy.

Most of the common substitution models are based on similar assumptions, such as site independence, process-homogeneity across sequences and subtrees, partition homogeneity and an insignificance of certain functional structures such as codon-based substitution rates and codon usage bias. In case one or more of these assumptions are violated for an analysed dataset, the relative model selection criteria are not able to point this out (Felsenstein, 1978; Huelsenbeck & Hillis, 1993; Yang *et al.*, 1994; Swofford *et al.*, 2001).

Literature dealing with model fit is sparse. Only a small number of publications (Goldman, 1993b; Whelan *et al.*, 2001; Bollback, 2002; Huelsenbeck & Ronquist, 2001; Shavit Grievink *et al.*, 2010; Nguyen *et al.*, 2010) adresses model choice and model fit for either empirical (Ripplinger & Sullivan, 2008) or simulated data (Abdo *et al.*, 2005). Currently, two approaches have been designed to evaluate the absolute fit of data and model: the Goldman-Cox test (Goldman, 1993b; Whelan *et al.*, 2001) in an ML setting, and posterior predictive simulations (Huelsenbeck & Ronquist, 2001; Bollback, 2002) in a Bayesian setting. The Goldman-Cox test uses parametric bootstrapping, a method of testing model parameters which were estimated to describe how sequence alignments may have evolved in a statistical way. The bootstraps are artificial datasets which are generated using the estimated model parameters of the analysed dataset. Therefore, they represent a perfect fit of model and data. The entropy of patterns of all datasets, the original one and the parametric bootstraps, is calculated and compared. If the degree of entropy of the dataset fits within 95% of the distribution of entropy-degrees of the bootstraps, the model is accepted as adequate for the dataset.

Within the present study a range of MSA datasets on nucleotide level was simulated and analysed with respect to increasing non-stationarity caused by mixtures of evolutionary models (GTR model (substitution rates, base frequencies), shape of the gamma distribution (α) and proportion of invariable sites (p_{inv})) for different branches or subtrees nested in a topology. GTR was chosen because it is the most general and also the most commonly used time-reversible model for phylogenetic inference (Sumner *et al.*, 2012). The simulations were generated (i) with continuous gamma distribution and (2) using discrete gamma distribution with four rate categories. The trees were chosen to have unequal branch lengths, because it is known, that combination of long and short branches can cause biased attraction (Felsenstein, 1978; Hendy & Penny, 1989). For datasets simulated with a continuously gamma distribution analysed by ML methods given the correct substitution rates, base frequencies, Γ-shape parameter (α) and proportion of invariable sites, but only using four categories of rates to approximate the gamma distribution, the reconstruction may not always lead to the true tree (Kück *et al.*, 2012). Therefore the use of a higher number of rate categories should be considered: four categories might not be accurate enough to sufficiently cover different evolutionary rates as they might be present in large scale datasets. Therefore, tree inference was performed via ML estimation with different numbers of categories for estimating the discrete gamma distribution: 4, 12 or 25 categories. The model parameters, which were estimated within the scope of the ML analysis, were used for parametric bootstrapping to perform the Goldman-Cox test. First it was checked, whether the estimated models for the stationary datasets are accepted more often than the models of the non-stationary datasets and if the Goldman-Cox test is able to capture the model misspecification regarding stationarity. Furthermore, it was analysed, whether an increased number of categories for ML analysis with discrete gamma distribution has an impact on model-acceptance by the Goldman-Cox test for datasets simulated using either (i) discrete or (ii) continuous gamma distribution.

2.2. Materials and Methods

2.2.1. Sequence Data

Multiple Sequence alignments (MSA) on nucleotide level were simulated using INDELible V1.03 (Fletcher & Yang, 2009), Seq-gen v1.3.2 (Rambaut & Grassly, 1997) and MultimoSeqSim (Mayer, 2010). The GTR model of sequence evolution and either a continuous gamma distribution or discrete gamma distribution with four rate classes for ASRV was chosen. Insertion or deletion (indel) events were not simulated. The modelled datasets using continuous gamma distribution were either simulated with a specified proportion of invariant sites (see table 2.1) or p_{inv} was set to 0. Different combinations of topologies, branch lengths and model parameters were used to simulate these data (table 2.1). The combinations of model misspecifications differed in their extend of non-stationary base composition and substitution processes, shape of gamma distribution and p_{inv} (see table 2.1 and figures 2.1, 2.2, 2.3 and 2.4).

The first simulation setup (setup 1, Fig. 2.1) includes 14 taxa, with one sequence representing the outgroup. The first model arrangement (Fig. 2.1a) is stationary, with all branches evolving under the same evolutionary model (see table 2.1), whereas the second one (Fig. 2.1b) differs in model parameters from the outgroup taxon to the rest. The third setting (Fig. 2.1c) is highly non-stationary and evolves according to four different model parameter settings.

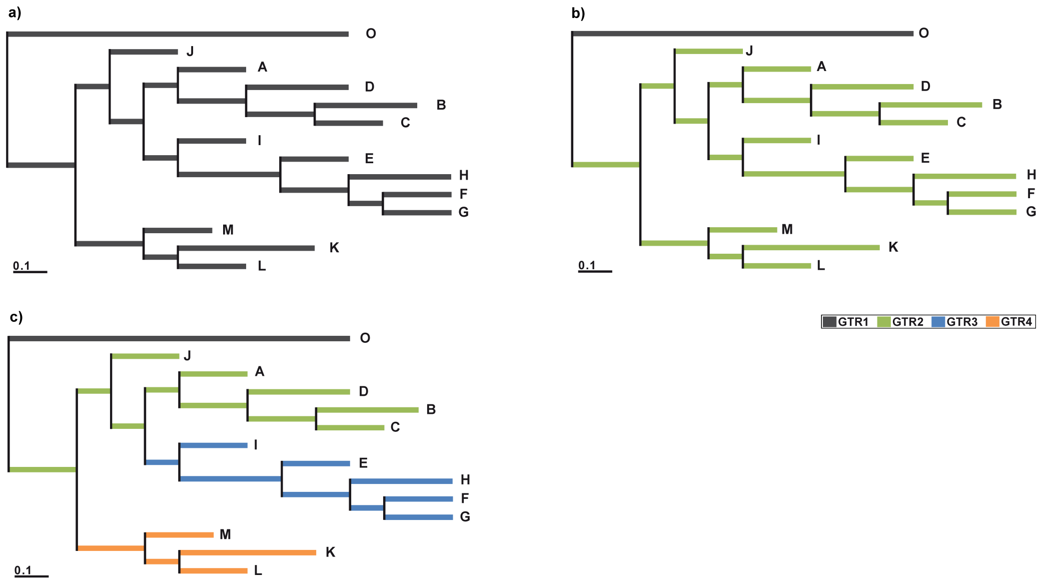


Figure 2.1: Setup 1: Topology, branch lengths and models used in analyses, covering a spectrum of stationary (a) and non-stationary (b and c) datasets. The topology includes 14 taxa with one taxon acting as outgroup. Three sets of parameters differ in the local application of evolutionary models for different clades (GTR1-4, for details see table 2.1).

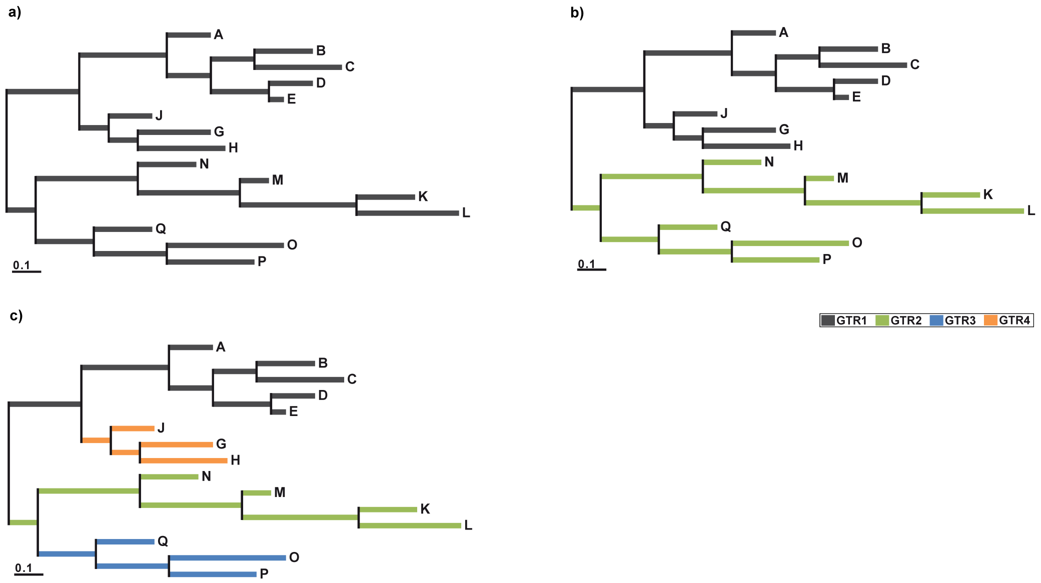


Figure 2.2: Setup 2: Topology, branch lengths and models used in analyses, covering a spectrum of stationary (a) and non-stationary (b-c) datasets. The unrooted topology includes 15 taxa. These three parameter sets differ in the local application of evolutionary models for different branches (GTR1-4, for details see table 2.1).

Table 2.1: Model specifications and parameters used for the simulation of different datasets. All parameter sets consist of an α value for gamma distributed rate heterogeneity (Γ), the proportion of invariant sites (I), GTR (general time-reversible) substitution rates and base frequencies.

			Substitution rates						Base frequencies			
Model	Γ	I	AC	AG	AT	CG	CT	GT	f_A	f_C	f_G	f_T
GTR1	0.75	0.3	0.2	1.0	0.7	0.3	0.5	1.0	0.25	0.20	0.25	0.30
GTR2	1.0	0.2	0.6	1.0	0.4	0.1	0.5	1.0	0.40	0.10	0.25	0.25
GTR3	0.8	0.1	0.8	1.0	0.6	0.1	0.3	1.0	0.25	0.30	0.25	0.20
GTR4	0.5	0.5	0.4	1.0	0.3	0.5	0.2	1.0	0.40	0.25	0.10	0.25
GTR5	1.0	0.4	0.4	1.0	0.9	0.2	1.0	1.0	0.30	0.10	0.20	0.40

The second simulation setup (setup 2, Fig. 2.2) is an unrooted tree with 15 taxa and an increasing number of evolutionary model specifications for different branches from a stationary composition (Fig. 2.2a) to non-stationary compositions with two (Fig. 2.2b) and four (Fig. 2.2c) different parameter arrangements.

The third simulation setup (setup 3, Fig. 2.3) is based upon a tree including 15 taxa that varies in combination of evolutionary models on different clades (b-d). The first arrangement (Fig. 2.3a) is a stationary dataset for which the same model was applied to all branches, whereas the second dataset (Fig. 2.3b) includes a paraphylum (highlighted in red color) which evolves according to a second model. It is nested within clades of the same model parameters. Four model-heterogeneous datasets (Fig. 2.3c-f) include branches evolving under different model parameters at several positions within the tree. For two datasets (Fig. 2.3g and h) these clades are combined ($g \approx c \cap e$ and $h \approx d \cap f$) resulting in two different clades within one tree evolving according to the same model parameters (see table 2.1).

All simulations were carried out three times in which the start seed for the generation of a random number was set to 1568746, 444444 and 555555, respectively. This was done to avoid a model rejection by chance conducting 100 bootstrap samples. The start seed was set constant for every simulation of a dataset and its corresponding parametric bootstrap datasets to avoid structural differences which may potentially be caused by different seeds.

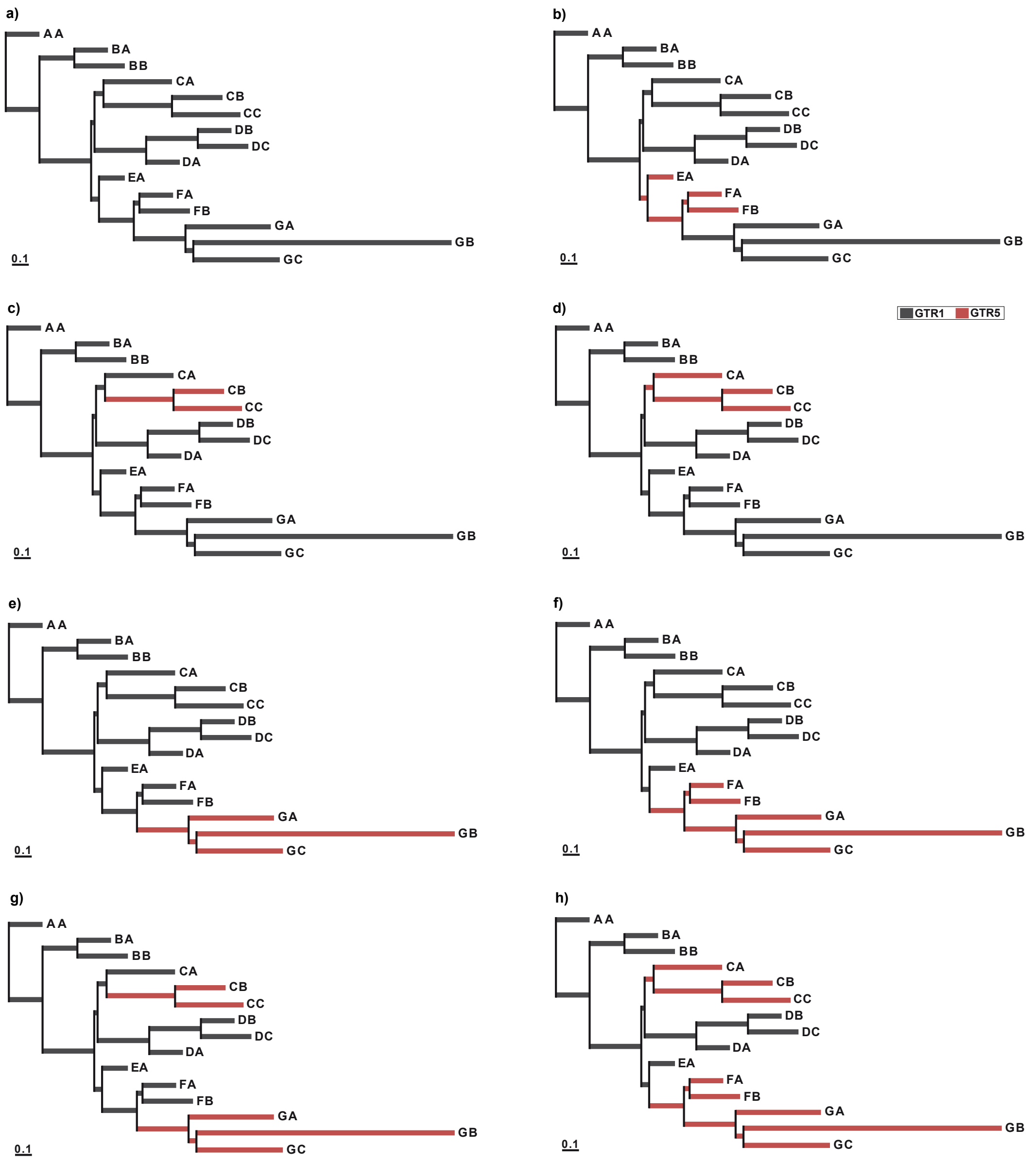


Figure 2.3: Setup 3: Topology, branch lengths and models used in analyses, covering a spectrum of stationary (a) and non-stationary (b-h) datasets. These eight sets of trees differ in the local application of evolutionary models (GTR1 and GTR5, for details see table 2.1).

Additionally, a number of datasets was designed (setup 4), all being stationary but having different numbers of taxa, sites or shorter branch lengths (see Fig. 2.4a-d) to check the impact on the results of the Goldman-Cox test with increasing amount of data up to provision of a pattern equilibrium. These simulations were carried out only with seed 1568746 for random number generation but with different numbers of bootstrap datasets.

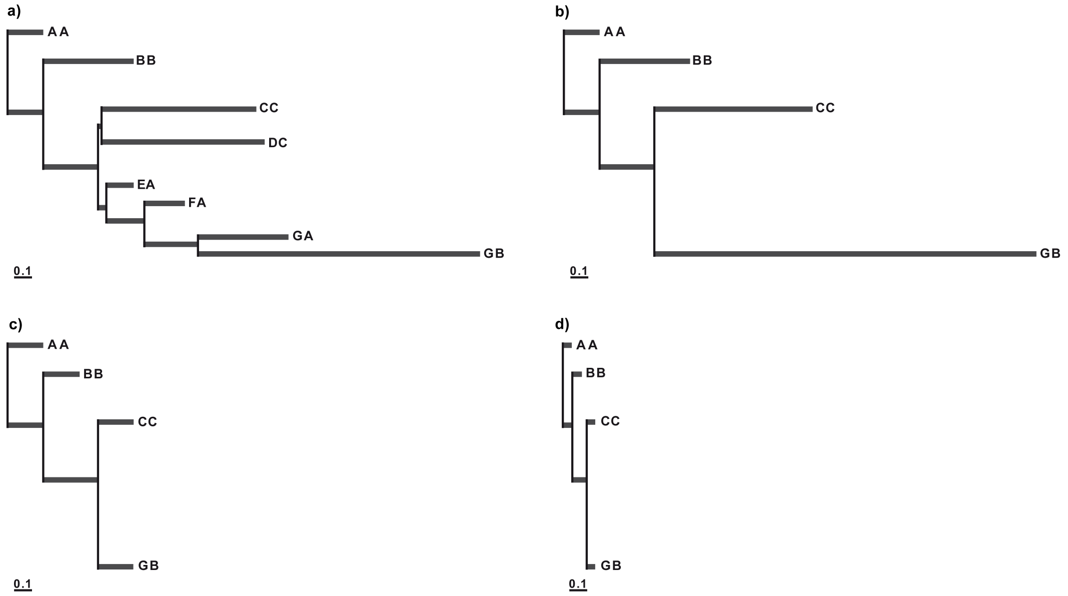


Figure 2.4: Setup 4: Topologies, branch lengths and models used in analyses in order to cover a spectrum of stationary datasets with 8 (a) or 4 (b-d) taxa and different branch lengths.

2.2.2. Phylogenetic Analyses and Parametric Bootstrapping

ML analyses were conducted with PhyML 3.0 (Guindon *et al.*, 2010; Guindon & Gascuel, 2003), RAxML standard version 7.3.5 (Stamatakis, 2006) and PAUP 4.0b10-console (Swofford, 2002), estimating the GTR+Γ+I or only GTR+Γ parameters, base frequencies and the best ML tree (for further details see Appendix, table A.1 and table A.2). For simulation setups 1, 2 and 3 all ML analyses were performed six times, using i) 4, ii) 12 or iii) 25 rate categories using either the mean or median for approximating gamma distribution. For simulation setup 4 all analyses were performed using four rate categories using either the mean or median for modelling gamma distribution. For each dataset 100, 1,000, 10,000 or 100,000 parametric bootstrap replicates were generated with INDELible V1.03, Seq-gen v1.3.2 and MultimoSeqSim. Different tools and options were used to compare the results and to check, if these depend on the used implementations or chosen parameters.

2.2.3. Goldman-Cox test and Pattern Analysis

The Goldman-Cox test (Goldman, 1993b) was applied using a Perl script (Mayer, available from the author upon request) to check if the estimated models are adequate for the datasets. This test evaluates the entropy of pattern distribution within a dataset and ranks the analysed dataset within the values of its corresponding bootstrap replicates. If the original dataset ranks outside the 95% confidence interval of the bootstrap distribution, the model is rejected as ’not adequate’ for the dataset.

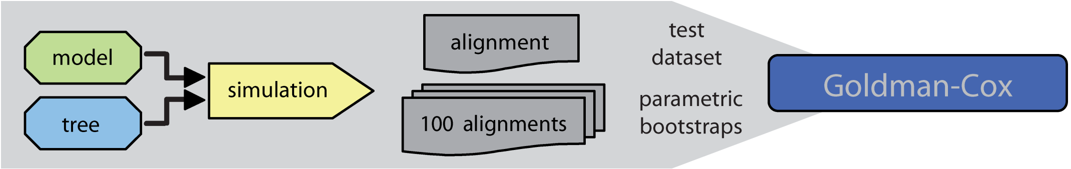


Figure 2.5: Flowchart of testing the Goldman-Cox test. On base of every tree and evolutionary model parameters, a nucleotide alignment is simulated. Then these parameters are used again to build 100 parametric bootstrap alignments. These datasets are then used as a reference rest set for the Goldman-Cox test.

Every tree and model composition was used to simulate (i) one and (ii) 100 datasets (see Fig. 2.5), all based on the designed model parameters rather than estimated values. This was done to check, whether the Goldman-Cox test accepts the true model in case the parametric bootstrap datasets were generated with identical parameters and therefore an adequate model. This arrangement constitutes a reference test for the Goldman-Cox test.

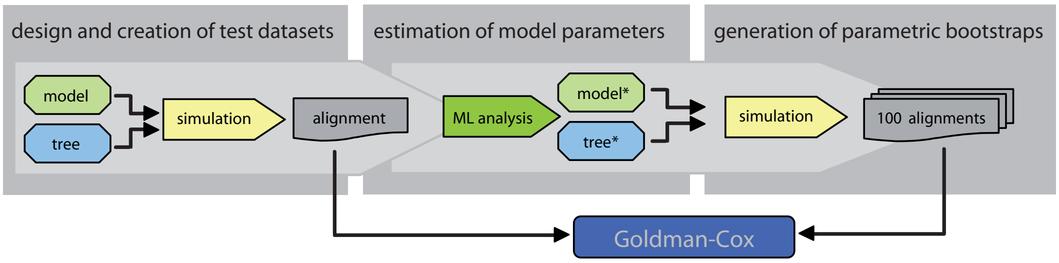

design and creation of test datasets
estimation of model parameters
generation of parametric bootstraps
model
tree
simulation
alignment
ML analysis
model*
tree*
simulation
100 alignments
Goldman-Cox

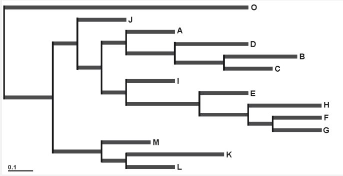
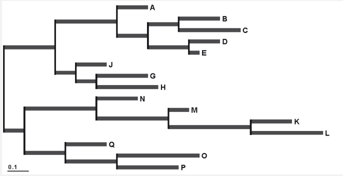
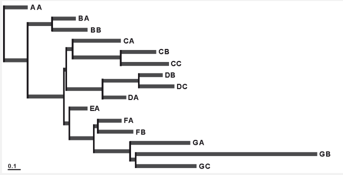
Figure 2.6: Flowchart of the analysis. On the basis of a tree and evolutionary model parameters alignments are simulated. Then parametric bootstraps are generated by analysing these test datasets, estimating the best fitting tree (tree*) and the evolutionary model (model*) with a ML analysis. These estimated parameters (tree* and model*) form the basis for a second simulation step, in which parametric bootstraps are generated for every dataset.

All simulated stationary and non-stationary datasets and their corresponding bootstraps were analysed with the Goldman-Cox test. Additionally, the distribution of site patterns and the frequencies of ACGT and RY splits within datasets and corresponding parametric bootstrap replicates were checked. Therefore, the number of occurrences of a single pattern or split within a simulated dataset was compared to all the numbers of occurrences of the same pattern or split in each of its corresponding parametric bootstraps.

2.3. Results

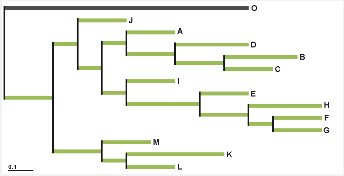
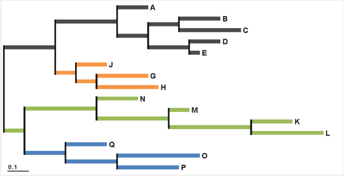
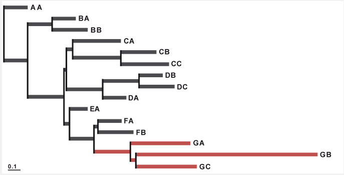
While all dataset-model combinations of the reference sets passed the Goldman-Cox test, the detailed results of the Goldman-Cox test for the actual analyses varied across tree shapes and degree of stationarity (for all results see Appendix, table A.3 and table A.4).

Table 2.2: Results of the Goldman-Cox test for stationary datasets. All datasets were simulated with continuous GTR+Γ and analysed with GTR+Γ+I. cat = rate categories for the Γ-distribution used in the ML analysis. s15, s44, s55 = runs with different seeds for random number generation (1568746, 444444, or 555555). The values within the table show, which rank the test result of the original dataset achieved within the distribution of the results of the corresponding bootstrap replicates. If the result of the original dataset ranks within a confidence interval (95%) the model is accepted as adequate (ranks 3 - 99, highlighted green), otherwise it is rejected (ranks 1,2,100 and 101, highlighted yellow).

Dataset	ML analysis options		Goldman-Cox		
			s15	s44	s55
setup 1a 	reference set		51	76	78
	GTR+Γ+I, 4 cat	mean	2	1	1
		median	1	1	1
	GTR+Γ+I, 12 cat	mean	48	28	42
		median	3	4	6
	GTR+Γ+I, 25 cat	mean	56	38	59
		median	37	11	26
setup 2a 	reference set		62	97	43
	GTR+Γ+I, 4 cat	mean	1	1	1
		median	1	1	1
	GTR+Γ+I, 12 cat	mean	36	48	33
		median	3	6	5
	GTR+Γ+I, 25 cat	mean	35	45	55
		median	30	27	20
setup 3a 	reference set		26	83	41
	GTR+Γ+I, 4 cat	mean	1	1	1
		median	1	1	1
	GTR+Γ+I, 12 cat	mean	34	32	29
		median	4	8	5
	GTR+Γ+I, 25 cat	mean	36	60	36
		median	23	29	18

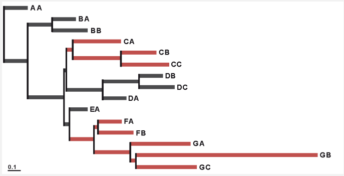
Stationary datasets simulated with a continuous gamma distribution were mostly rejected, if analysed with four rate categories for a discrete gamma distribution (see table 2.2). Likewise, the estimated model parameters for the datasets of setup 4 were mainly rejected by the Goldman-Cox test for various combinations (see Appendix, table A.2) of simulation software and ML implementations using four rate categories for gamma distribution (results not shown).

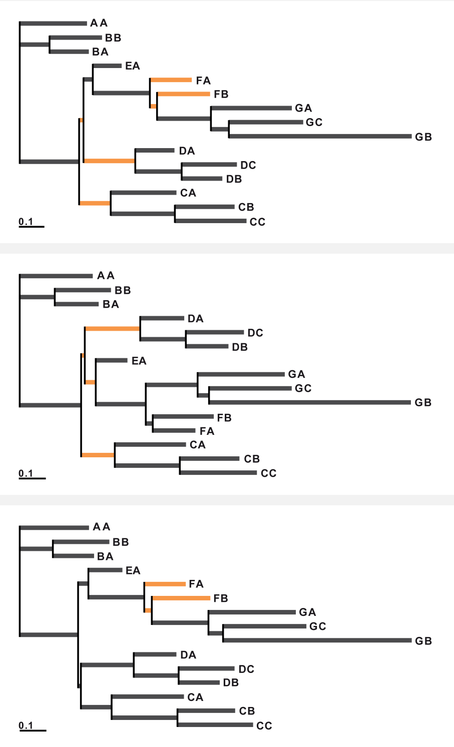
Table 2.3: Results of the Goldman-Cox test for model-heterogeneous datasets. All datasets were simulated with continuous GTR+Γ+I and analysed with GTR+Γ+I. cat = rate categories for the Γ-distribution used in the ML analysis. s15, s44, s55 = runs with different seeds for random number generation (1568746, 444444, or 555555). For explanation of values and color-code within the table see the heading of table 2.2.

Dataset	ML analysis options		Goldman-Cox		
			s15	s44	s55
setup 1b  O J A D B C I E H F G M K L 0.1	reference set		54	76	54
	GTR+Γ+I, 4 cat	mean	101	98	99
		median	36	40	49
	GTR+Γ+I, 12 cat	mean	101	101	101
		median	101	101	101
	GTR+Γ+I, 25 cat	mean	101	101	101
		median	101	101	101
setup 2c  A B C D E J G H N M K L Q O P 0.1	reference set		68	25	19
	GTR+Γ+I, 4 cat	mean	99	98	95
		median	23	21	10
	GTR+Γ+I, 12 cat	mean	101	101	101
		median	101	101	100
	GTR+Γ+I, 25 cat	mean	101	101	101
		median	101	101	101
setup 3e  AA BA BB CA CB CC DB DC DA EA FA FB GA GB GC 0.1	reference set		32	92	38
	GTR+Γ+I, 4 cat	mean	98	99	91
		median	6	15	8
	GTR+Γ+I, 12 cat	mean	101	101	101
		median	100	101	101
	GTR+Γ+I, 25 cat	mean	101	101	101
		median	101	101	101

The estimated models of those datasets simulated with branches that evolve according to two or more evolutionary models were accepted if analysed with four rate categories for the gamma distribution, but rejected if analysed with 12 or 25 rate categories (see table 2.3).

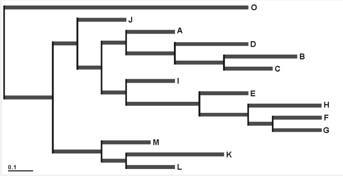
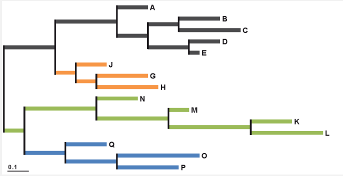
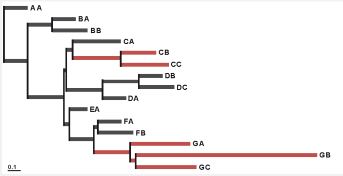
Table 2.4: Results of the Goldman-Cox test for model-heterogeneous datasets. All datasets were simulated with continuous GTR+Γ and analysed with GTR+Γ. cat = rate categories for Γ-distribution used in the ML analysis. s15, s44, s55 = runs with different seeds for random number generation (1568746, 444444, or 555555). For explanation of values and color-code within the table see the heading of table 2.2. Below the table, three topologies are shown, one for every seed. Although the models were accepted by the Goldman-Cox test, the topologies calculated by a ML analysis were all incorrect (incompatible branches highlighted orange).

Dataset	ML analysis options		Goldman-Cox		
			s15	s44	s55
setup 3h  AA BA BB CA CB CC DB DC DA EA FA FB GA GB GC 0.1	reference set		55	72	17
	GTR+Γ, 4 cat	mean	29	57	46
		median	1	3	2
	GTR+Γ, 12 cat	mean	92	96	95
		median	56	81	74
	GTR+Γ, 25 cat	mean	93	95	98
		median	86	93	86



One dataset, which is based on a non-stationary process (heterogeneous model composition; see table 2.4), generated with GTR+Γ and p_{inv}=0, showed peculiar results. For all analyses, all models were accepted except for two. A detailed analysis of the trees on which the analysed dataset and the corresponding bootstraps were based on, showed, that all ML tree reconstruction yielded a wrong topology.

Table 2.5: Results of the Goldman-Cox test for model-homo- and heterogeneous datasets. All datasets were simulated with 4 categories for GTR+Γand analysed with GTR+Γ. cat = rate categories for Γ-distribution used in the ML analysis. s15, s44, s55 = runs with different seeds for random number generation (1568746, 444444, or 555555). For explanation of values and color-code within the table see the heading of table 2.2.

Dataset	ML analysis options		Goldman-Cox		
			s15	s44	s55
setup 1a  O J A D B C I E H F G M K L 0.1	reference set		73	32	13
	GTR+Γ, 4 cat	mean	51	38	30
		median	3	1	1
	GTR+Γ, 12 cat	mean	101	101	101
		median	95	96	95
	GTR+Γ, 25 cat	mean	101	101	101
		median	100	101	101
setup 2c  A B C D E J G H N M K L Q O P 0.1	reference set		77	27	20
	GTR+Γ, 4 cat	mean	88	95	85
		median	10	10	3
	GTR+Γ, 12 cat	mean	101	101	101
		median	101	101	101
	GTR+Γ, 25 cat	mean	101	101	101
		median	101	101	101
setup 3g  AA BA BB CA CB CC DB DC DA EA FA FB GA GB GC 0.1	reference set		52	22	12
	GTR+Γ, 4 cat	mean	72	78	66
		median	4	5	3
	GTR+Γ, 12 cat	mean	101	101	101
		median	101	101	100
	GTR+Γ, 25 cat	mean	101	101	101
		median	101	101	101

The results for the datasets simulated with four rate categories for a discrete gamma distribution were similar for stationary and non-stationary setups. Estimated models of the ML analysis performed with four rate categories were widely accepted, whereas models estimated with more rate categories are rejected (see table 2.5).

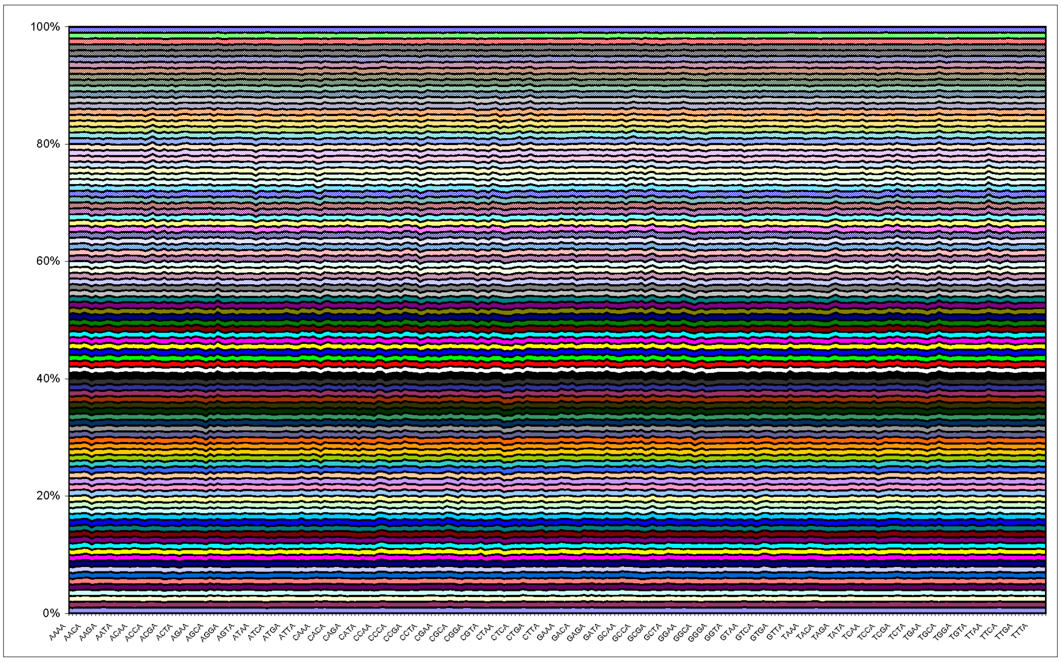


Figure 2.7: Comparison of pattern frequencies of the parametric bootstraps. X-axis: patterns from AAAA to GGGG. Y-axis: bootstrap datasets. The pattern frequencies within the parametric bootstraps are homogeneously distributed.

The distribution of pattern frequencies for datasets of setup 4a and 4b (Fig. 2.4) including corresponding bootstrap replicates which were based on estimated model parameters using four rate categories were analysed and for setup 4b visualised in Fig. 2.7: The 100% stacked area chart shows the pattern spectrum of 100 bootstrap replicates of the analysed dataset of setup 4b, simulated with one million sites. On the x-axis the patterns are listed, from 'AAAA' to 'TTTT'. The proportionate amount of occurrence of every pattern is shown on the y-axis. The diagram sums up to 100% and therefore displays the percentage of the total number of occurrences that each bootstrap contributes per pattern. The homogeneous distribution shows that the pattern frequencies are evenly distributed across all bootstrap datasets.

By adding the original dataset to the area chart (yellow coloured data at the bottom of chart of Fig. 2.8), the balanced distribution gets completely inhomogeneous. This shows, that the proportion of patterns differ extremely between analysed dataset and the parametric bootstrap replicates. If the

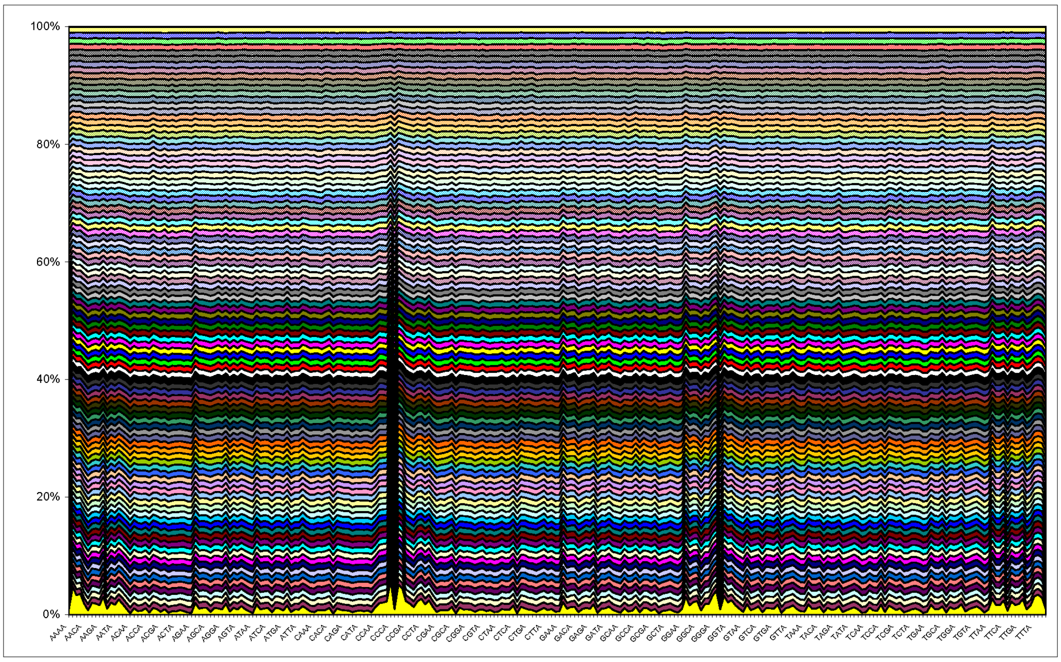


Figure 2.8: Pattern comparison of original dataset of setup 4b and its parametric bootstraps. X-axis: patterns from AAAA to GGGG. Y-axis: original (yellow coloured dataset at the bottom) and bootstrapped datasets. The patterns are distributed homogeneous within the bootstraps, but they deviate strongly from the original dataset.

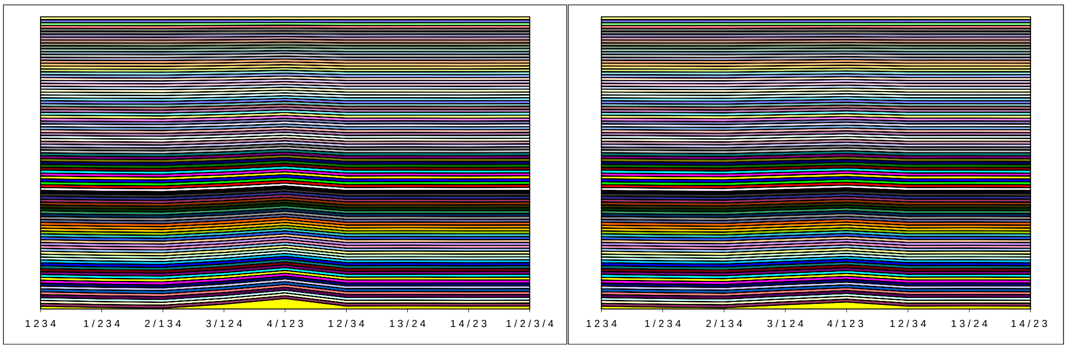


Figure 2.9: Split comparison of original dataset of setup 4b and its parametric bootstraps. On the left hand side the patterns are combined to ACGT splits, whereas RY-coding can be found on the right hand side. X-axis: splits. Y-axis: analysed (yellow coloured row at the bottom) and bootstrapped datasets. The more the patterns are generalized to splits, the lower the difference between data and bootstraps.

patterns would be distributed equally, all rows would be homogeneous as shown in Fig. 2.7, including bootstraps only.

Grouping the patterns to ACGT and RY splits balances the proportional distribution, but does not lead to an equilibrium (see Fig. 2.9). A comparison of the total number of occurrences of splits (by grouping pattern to ACTG or RY splits) to the mean number of occurrences within the bootstraps shows, that there is a distinct incongruence for trivial splits (splits which separate one taxon from all other taxa). The step from ACGT splits to RY splits balances the differences only slightly.

Table 2.6: Patterns found in the original dataset of setup 4b and its parametric bootstrap replicates condensed to ACGT splits.
Dataset = number of split occurrences, Mean btsps = mean value of number of split occurrences of the parametric bootstraps,
Difference = absolute value of (dataset - mean), SD = standard deviation.

Split	Dataset	Mean btsps	Difference	SD
1 2 3 4	280976.00	294645.29	13669.29	475.20
1 / 2 3 4	39071.00	72341.58	33270.58	236.13
2 / 1 3 4	43687.00	139495.94	95808.94	297.32
3 / 1 2 4	73164.00	42699.19	30464.81	215.16
4 / 1 2 3	140549.00	37880.00	102669.00	176.32
1 2 / 3 4	171355.00	168616.27	2738.73	405.08
1 3 / 2 4	112022.00	105782.57	6239.43	331.94
1 4 / 2 3	109038.00	109261.57	223.57	324.40
1/2/3/4	30138.00	29277.59	860.41	177.73

Table 2.7: Patterns found in the original dataset of setup 4b and its parametric bootstrap replicates condensed to RY splits.
Dataset = number of split occurrences, Mean btsps = mean value of number of split occurrences of the parametric bootstraps,
Difference = absolute value of (dataset - mean), SD = standard deviation.

Split	Dataset	Mean btsps	Difference	SD
1 2 3 4	427379.00	438943.71	11564.71	523.38
1 / 2 3 4	62724.00	96714.87	33990.87	286.26
2 / 1 3 4	67600.00	149252.44	81652.44	326.43
3 / 1 2 4	98730.00	65722.18	33007.82	258.94
4 / 1 2 3	150732.00	61125.96	89606.04	243.77
1 2 / 3 4	83824.00	82585.80	1238.20	304.31
1 3 / 2 4	55258.00	52037.47	3220.53	223.86
1 4 / 2 3	53753.00	53617.57	135.43	223.18

2.4. Discussion

We simulated different nucleotide sequence alignments with heterogeneous composition and non-stationary processes to examine compositions based on a mixture of models and its impact on the results of estimated model adequacy.

Out of all models which were estimated for the datasets which were simulated using continuous gamma and analysed using ML with a four rate discrete gamma distribution 81.25% (see Appendix, table A.4) were rejected by the Goldman-Cox test. Neither the use of different simulation or ML software nor a higher amount of bootstrap replicates did have a remarkable effect. A closer look at the distribution of patterns within the datasets and their corresponding bootstrap replicates showed, that the Goldman-Cox test is performing properly in this case. While one of the test datasets (see Appendix table A.2, simulated with INDELible, p_{inv}=0, analysed with PhyML estimating GTR+Γ) contained 7196 different patterns, the pattern variety within the corresponding bootstrap replicates fluctuated between 6915 and 7118, and the medium quantity amounted to 7007.91 patterns (results not shown). With a standard deviation of 48.06 there is an enormous deviation between the number of patterns of the original dataset and its bootstrap replicates. Thus, the Goldman-Cox test results are correct. Further, the deviation for the number of occurrences of the patterns one by one shows, that the original dataset and its parametric bootstraps show largely different pattern frequencies. Of course, even though the alignment was simulated with GTR+Γ, the datasets contain invariant positions. Remarkable is the fact, that all bootstrap replicates contain more invariant positions than the original dataset. For each dataset the number of each invariable pattern deviates significantly (always less than the median minus the standard deviation) from the numbers found in the corresponding bootstrapped data. Summarised, there are fewer different patterns and more invariant patterns within the bootstrap replicates.

Since the original test dataset did not allow a pattern equilibrium, the number of taxa was reduced and the number of sites was increased. Although a dataset with 4 taxa and 1 million sites provided a pattern frequency equilibrium, the pattern distribution of the original dataset and the bootstrap replicates varied highly. Likewise the estimated model was rejected by the Goldman-Cox test.

The results (rejected model adequacy although providing pattern equilibrium) were tested by performing analyses with 1,000, 10,000 and 100,000 bootstraps and it was checked, whether the random-seed function of INDELible has any influence on the comparability of our datasets. Both tests, varying the number of bootstrap replicates and the random-seed number, produced the identical results. The estimated parameters and trees were very similar to the models and trees of the simulations. Nevertheless, the difference was sufficient to be detected by the Goldman-Cox test.

A look at the distribution of the patterns of the dataset with four taxa, 1 million sites without invariant sites showed, even given the possibility of a pattern equilibrium, that the pattern distribution differs significantly between the original alignment and its parametric bootstraps. The visualisation in a 100% stacked area chart showed that the pattern distribution of the bootstrap replicates are very balanced and distributed homogeneously (Fig. 2.7). Adding the analysed original dataset to the area chart (yellow coloured data at the bottom in Fig. 2.8) showed that the pattern frequencies dramatically differs between the original dataset and its parametric bootstrap replicates. Grouping the patterns in ACTG splits, the differences for different patterns are compensated (distribution bias). Recoding them to RY splits balances the differences even more. The highest deviation takes place within splits where one subgroup contains only one taxon and the other subgroup the remaining ones (see tables 2.6 and 2.7). Since these so called 'trivial' splits and thus the corresponding patterns are not of phylogenetic interest and affect only the branch lengths, it might be a good idea to perform the Goldman-Cox test only for the sites which represent internal nodes (branching points). The fact that recoding the nucleotide alignments to an RY code balances the differences within the distribution indicates, that the 'challenges' which were counterbalanced might be linked to time-reversible and non-stationary processes, since RY-coded alignments are more likely to be consistent with evolution under globally stationary, reversible and homogeneous (SRH) conditions (Ho *et al.*, 2006).

However, apparently multiple substitutions within terminal taxa with longer branches are challenging for testing model adequacy. Datasets generated with much shorter branches (Fig. 2.4c and d) to be expected having less multiple substitutions also balanced the pattern distribution; differences decreased, but were still present.

Analyses of datasets simulated with a continuous gamma distribution and for which all branches are based on the same substitution rates, base frequencies, invariant sites and rate heterogeneity (SRH conditions) showed that model adequacy can be assessed much better by using more than four rate categories when performing ML analysis with gamma distribution. While only 18.75% of the models estimated with four gamma rate categories passed the Goldman-Cox test (see Appendix, table A.4), increasing the number of categories representing the gamma distribution cured this behaviour (see table 2.2).

On the other hand, the use of four gamma rate categories for ML analysis of datasets with heterogeneous composition and non-stationary processes led to an increased number of accepted estimated models. This result is alarming, since the results for the estimation with 12 or 25 rate categories were mostly rejected. This implies, that this might be due to a type II error which leads to false positive model acceptance.

The results for the datasets simulated with four gamma rate categories rather than using continuous gamma distribution, indicate that it is not necessarily better to use more categories. The challenge is rather to figure out how many categories truly represent the dataset. Here, the models estimated with the same number of categories as used for the simulation passed the Goldman-Cox test (see table 2.5). The models for the stationary datasets sometimes passed the test, even if analysed quite 'over-rated' with 12 gamma categories, whereas all models estimated with 25 gamma categories were rejected. Rejections of some estimated models for datasets simulated with four gamma rate categories which were analysed by taking the median of the categories instead of the commonly used mean can simply be explained by the fact that for the simulation the mean option was used for modelling the discrete gamma rates. In such cases the use of the 'median option' has to be considered as a misspecification.

In contrast to the other datasets with heterogeneous composition and non-stationary processes generated with GTR+Γ and p_{inv}=0, all estimated models except for two of all analyses of datasets based on setup 3h were accepted by the Goldman-Cox test (see table 2.4). A further evaluation of the topologies (Puigbò *et al.*, 2007) on which the original dataset and the corresponding bootstraps are based on showed, that all topologies calculated by ML were wrong. Matsen and Steel (2007) described for 4-taxon trees, and later Matsen *et al.* (2008) for trees with more taxa, that mixture models applied on one tree can 'mimic' standard models for certain parameter choices on a different tree. This is not only true for highly heterogeneous datasets, but also for data with low heterogeneity, for which processes can potentially get indistinguishable from stationary and homogeneous ones on a different tree. Both trees, the standard model and the mixed model tree can differ by at most one nearest neighbour interchange (Matsen *et al.*, 2008). This is apparently the case for datasets of this study (setup 3h) as well, and is highly alarming, since the use of the Goldman-Cox test assures to have an adequate model. It might be an adequate model, but for another dataset and a different tree.

Consequently, present analyses lead to the conclusion that testing for model adequacy of nucleotide datasets on base of parametric bootstraps based on evolutionary models which require stationary, reversible or homogeneous conditions (Jayaswal *et al.*, 2005) is not a good choice, if it can be assumed, that these conditions are violated by the data. In general, it is well known, that those violations increase the risk of phylogenetic errors (Ho & Jermiin, 2004; Jermiin *et al.*, 2004). Since empirical datasets are highly likely to be heterogeneous and non-stationary, there is always the possibility, that the estimated model, even if accepted by the test, may actually fit another tree. Furthermore, the proposed Goldman-Cox test it is very strict, rigorous and much more sensitive to violations of model assumptions than model-based tree reconstructions. In our simulations, the models had been often rejected, even though the calculated topology was correct.

In summary, the results indicate that the number of rate categories for modelling gamma distribution in ML analysis has to be considered very carefully and should be adequate for the dataset. The simulations of this study have shown that modelling gamma distribution with four rate categories when testing for model adequacy can be sufficient, but only for datasets, which were simulated with the same number of rate categories. The datasets, which were simulated with four gamma categories could also be approximated by 4 gamma rate categories in phylogenetic reconstruction. For the simulations which are based on a continuous gamma distribution, increasing the number of rate categories and therefore better approximating a continuous distribution for analysing stationary datasets, leads to better model estimates regarding the pattern distribution. For non-stationary datasets, which is to be expected from empirical datasets, the analysis with four gamma rate categories led to an acceptance of the estimated model, while the models which were estimated with 12 or more categories were rejected. Since some of the reconstructed trees were wrong, this implies, that an analysis based on a gamma distribution approximated by four rate categories under these conditions can lead to type 2 errors and gives therefore false positive results. Thus, it is advisable to inspect the dataset carefully before analysing and adjust the number of rate categories to the dataset, rather than to rely on a fixed number.

Die Menschen, die den richtigen Weg gehen wollen, müssen auch von Irrwegen wissen.
Aristoteles

3. Split Residual Diagnostics for Phylogenetics - Taking a More Detailed Look at Model Adequacy

3.1. Introduction

Results of phylogenomic analyses illustrate that systematic errors remain a major challenge for phylogenetic inference, with model choice as a critical step. For example, Philip *et al.* (2005) and Philippe *et al.* (2005) published their studies in the same issue of the journal 'Molecular Biology and Evolution', showing trees with 100% support for incongruent relationships. Phillips, Delsuc and Penny (2004) looked at yeast data taken from Rokas *et al.* (2003) and found that a slightly different model gave 100% support for different trees containing contradictory relationships. Goremykin *et al.* (2005) found that very minor changes of model parameters switched 100% bootstrap support for *Amborella* to 100% support for grasses basal in the phylogeny of angiosperms.

Poorly fitting substitution models cause numerous problems, from incorrect posterior probabilities (in Bayesian analyses), to biased estimates of branch lengths and in the worst case incorrect topologies. Simulation studies have shown that in case of well-fitted models popular measures of support, such as bootstrapping (Efron, 1979; Efron & Tibshirani, 1986; Felsenstein, 1985; Efron *et al.*, 1996) and the posterior probabilities, can be interpreted as measures of accuracy (Alfaro *et al.*, 2003; Erixon *et al.*, 2003). Unfortunately, in case of poorly-fitted models such an interpretation is potentially misleading.

It is therefore inevitable to analyse substitution model adequacy in order to address potential biases in tree reconstruction. In this chapter, an approach to assess model adequacy in empirical sequence data and an analysis of potential measures to reduce the source of biases in these data is presented. Due to its central part in the tree reconstruction procedure, model selection has received much attention. By far the most popular way of addressing goodness-of-fit of substitution models to observed empirical sequence data is a test for the relative best model fit, as exemplified by programs such as ModelTest (Posada & Crandall, 1998), MrModelTest (Nylander, 2004) and ModelGenerator (Keane *et al.*, 2006). The question is: which model, out of a predefined set of models, is most appropriate for my data?

Ripplinger and Sullivan (2008) compared the performance of relative model-selection methods and analysed how the choice of alternative well fitting models affects tree reconstruction (Posada & Buckley, 2004; Sullivan & Joyce, 2005). They found that different model selection criteria, the hierarchical likelihood-ratio test (hLRT), the Akaike information (AIC), the Bayesian information (BIC), and the decision theory (DT) criterion, preferred different models for the same datasets in almost every instance. The use of these 'alternative best-fit' models changed the optimum tree topology in about half of the cases. They advised to use the simplest supported substitution model selected by BIC or DT for ML tree reconstructions.

Besides evaluating the relative best-fitting model, approaches to assess absolute goodness-of-fit, i.e. the Goldman-Cox test (Goldman, 1993b; Whelan *et al.*, 2001) in a maximum likelihood (ML) setting and posterior predictive simulations in a Bayesian frame work (Bollback, 2002; Huelsenbeck & Ronquist, 2001) show how good a model describes the data. These absolute tests of model fit are used less commonly, perhaps because it is not clear what should be done if the answer to the question 'Is this model adequate?' is no. It would be useful to have a method that rather than just saying – the model fits poorly – delivers information on the quality of the misfits.

Using so called 'model free' methods, such as parsimony, is certainly not a good solution. These methods have implicit assumptions (Steel & Penny, 2000; Tuffley & Steel, 1997; Steel, 2002) and are known to have consistency problems (Felsenstein, 1978; Hendy & Penny, 1989). Several possible sources for model misspecification are known such as covarion evolution (Fitch & Markowitz, 1970; Penny *et al.*, 2001) or heterotachy (Lopez *et al.*, 2002; Zhong *et al.*, 2011), i.e. a change of patterns of rates across sites through time. Another cause is reticulate evolution, which includes hybridisation, horizontal gene transfer and recombination and leads to the fact, that relationships can not be seen as clades, but as a network of evolutionary lineages. Non-stationary substitution processes and processes that have a high variation throughout the tree can lead to model misspecification as well (Ho & Jermiin, 2004; Jermiin *et al.*, 2004; Squartini & Arndt, 2008). Exploring the ways in which data deviates from a particular model of substitution can tell an important story (e.g. homoplasy as pattern; Faith, 1989).

In the recent years it became obvious that substitution processes by themselves evolve (along trees), which makes a proper formal description difficult and error-prone (Wu & Susko, 2009; Kolaczkowski & Thornton, 2008; Whelan, 2008; Zhou *et al.*, 2007). Most likely the application of globally fitted substitution models not accounting for taxon or clade specific patterns may lead to biased reconstructions and artefacts. The identification of these artefacts in empirical data is almost impossible. The consequence must be to study the evolution of substitution parameters before the application of globally fitted models and to understand what the difference between local taxon or clade specific variation of the substitution parameters and a globally fitted model is. We have to conceptionally shift from just fitting a single substitution model to analysing the adequacy and the misfitting of the model or even submodels. The idea is to reduce the source of potential biases prior to tree reconstructions.

In theory, given a tree and substitution model, the expected frequency values of all site patterns can be calculated and compared to the observed pattern frequencies. In regression analysis, it is common practice to propose a model with some kind of residual diagnostics. In phylogenetics, this corresponds to comparing the observed pattern frequencies to those expected under the model (here, the model includes the tree, branch lengths and the model of nucleotide substitution). Absolute tests of model fit use variants of this approach. They simulate nucleotide MSAs under the best fitting tree and best fitting model and compare this simulated data with the observed empirical MSAs.

In a similar sense, MISFITS (Nguyen *et al.*, 2010) has been developed to evaluate the goodness-of-fit of substitution models. Based on a multiple nucleic acid alignment, the proposed evolutionary model and the inferred tree, the pattern frequencies of the empirical data and the expected ones by the model and tree are determined. Derived sets of over- and underrepresented patterns are examined and the alignment is checked if it can be adjusted by additional substitutions to fit the ML tree. Thus, the method is able to detect site patterns, which are not captured well, neither by the evolutionary model nor the inferred tree. By mapping the additional substitutions on the tree, this method gives a biological interpretation, why the model may not cover the alignment adequately. However, this is just a first step towards an appropriate treatment of model misspecification. At this time, there is no feasible solution how to handle datasets which are not represented by a single evolutionary model but only a combination of, either known or predicted, models.

Phylogenetic data is high dimensional, e.g. for nucleotide alignments with n sequences there are 4^n possible site patterns. Compared to this, the absolute goodness-of-fit tests use fairly brutal summary measures. They look at the pattern distribution of all datasets and boil the real data and each simulated dataset down to a single number which forms the basis for a decision. This way a lot of information is lost. Split graphs have been shown to be a particular useful tool in this context. They allow conflicting phylogenetic signals to be displayed in a network and therefore provide a good compromise between looking at all possible patterns and summarising datasets by one value. This idea has been explored to some extent in Wägele *et al.* (2003) and Goremykin *et al.* (2005).

However, it is not enough to simply consider conflict or compatibility among splits observed in empirical data, because for many combinations of tree shape (e.g. ones with combinations of short and long branches) and substitution models we can predict the presence of highly supported incompatible splits. Incompatibility does not necessarily have to be alarming. Nevertheless, incompatibility that is not predicted by the model is a cause for concern. Therefore, the methods of split analysis and parametric bootstrapping are combined for this study to perform split residual analyses of simulated and empirical nucleotide MSAs. After analysing all datsets with an ML implementation, the estimated evolutionary models including the tree, the substitution parameters and the rate heterogeneity parameters were used as base for simulating 100 parametric bootstrap datasets. The split spectra of the original datasets were compared the spectra of their corresponding bootstraps and all splits classified as over- or underrepresented were recorded and visualised in split networks. Simulated datasets were used to compare the estimated model parameters to the known statistical background, while the empirical datasets were analysed to get an idea, how the findings can be applied to assess model adequacy in phylogenetic analysis.

3.2. Materials and Methods

To test model adequacy based on split residuals for datasets with known statistical background, ten artificial datasets (see section 3.2.1) based upon tree shapes of different length and combinations of short and long branches were generated. This was done to involve complexity caused by conflict among the observed splits. Five published datasets (see section 3.2.1) were chosen for testing the method on empirical data. First, the datasets were analysed using an ML approach (section 3.2.2). The estimated model parameters were then used to generate parametric bootstrap datasets (described in section 3.2.2). Split spectra of all analysed datasets and their corresponding bootstraps were then compared and over- or underrepresented splits (described in section 3.2.3) were counted and visualized as split networks (see Fig. 3.1).

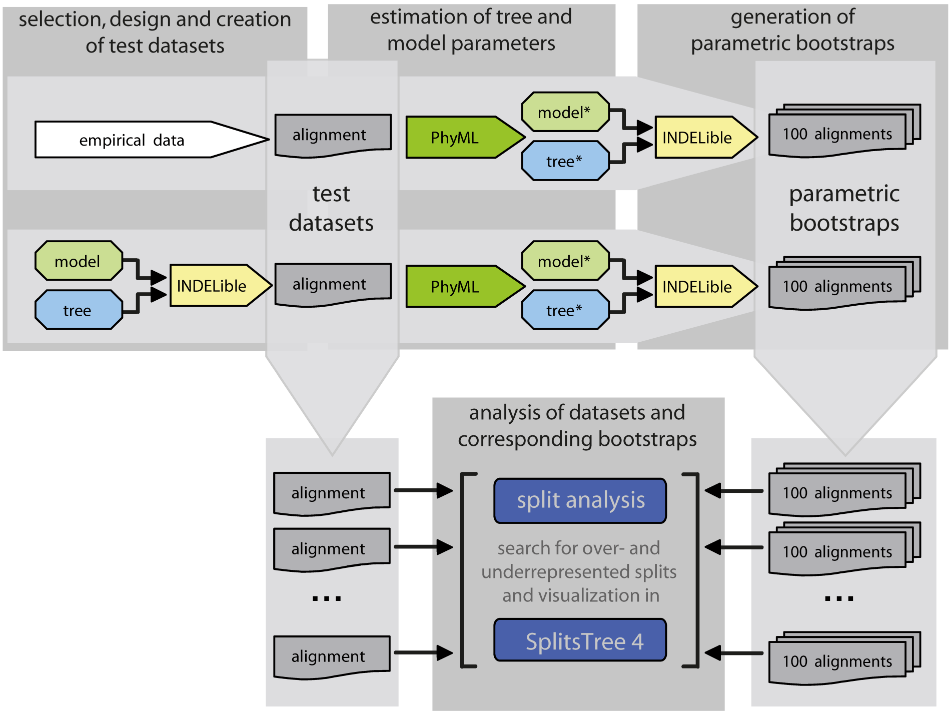


Figure 3.1: Overview of used methods and data. This flowchart shows how empirical and simulated nucleotide MSAs are processed throughout different analyses and processes. After a ML analysis the estimated model (model*) and tree (tree*) of every dataset is used to generate 100 parametric bootstrap datasets. The split spectra of the datasets and their corresponding bootstraps are then compared to check, if they can lead to statistical inference for model adequacy.

3.2.1. Sequence Data

Simulated Data Three different setups of topologies, branch lengths and number of taxa (Fig. 3.2, 3.3 and 3.4) were used to simulate ten nucleotide MSAs with 10.000 sites using INDELible V1.03 (Fletcher & Yang, 2009) with (i) four rate categories for approximating the gamma distribution or (ii) continuous gamma distribution modelling. The tree shapes combine short and long branches to involve complexity caused by conflict among the observed splits.

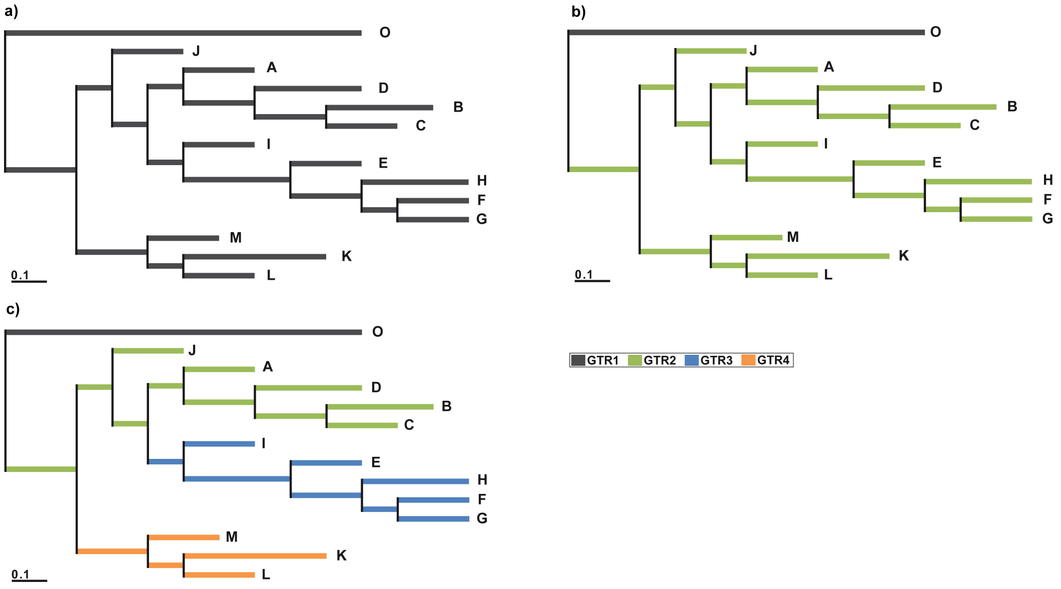


Figure 3.2: Topology 1 (datasets 1a, 1b and 1c, 14 taxa), branch lengths and models used in analyses, covering a spectrum of stationary (a) and non-stationary (b-c) datasets. Three sets of parameters differ in the local application of evolutionary models (see table 3.2 and 3.1).

All datasets of each setup were based on the same topology and differed only concerning evolutionary models for different branches or clades. One dataset of every setup (a) is simulated as evolving stationary and based upon only one model (GTR1, see table 3.1), the other datasets (b-c or d) include one or more clades which evolved under different model parameters (see table 3.1).

The nucleotide MSAs were simulated either with four rate categories approximating a gamma distribution and with a proportion of invariable sites (p_{inv})

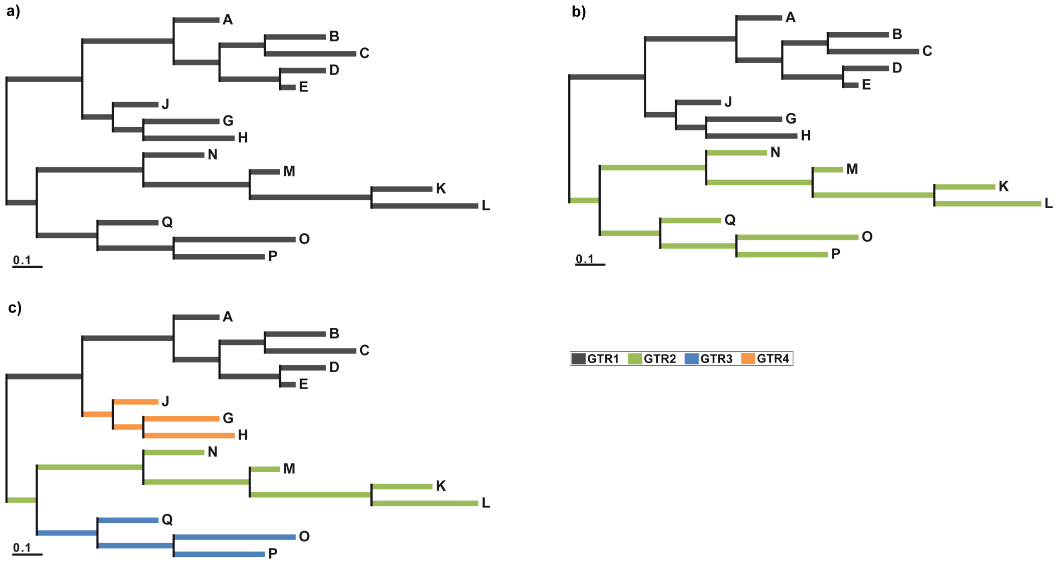


Figure 3.3: Topology 2 (datasets 2a, 2b and 2c, 15 taxa), branch lengths and models used in analyses, covering a spectrum of stationary (a) and non-stationary (b-c) datasets. These three sets of parameters differ in the local application of evolutionary models (see table 3.2 and 3.1).

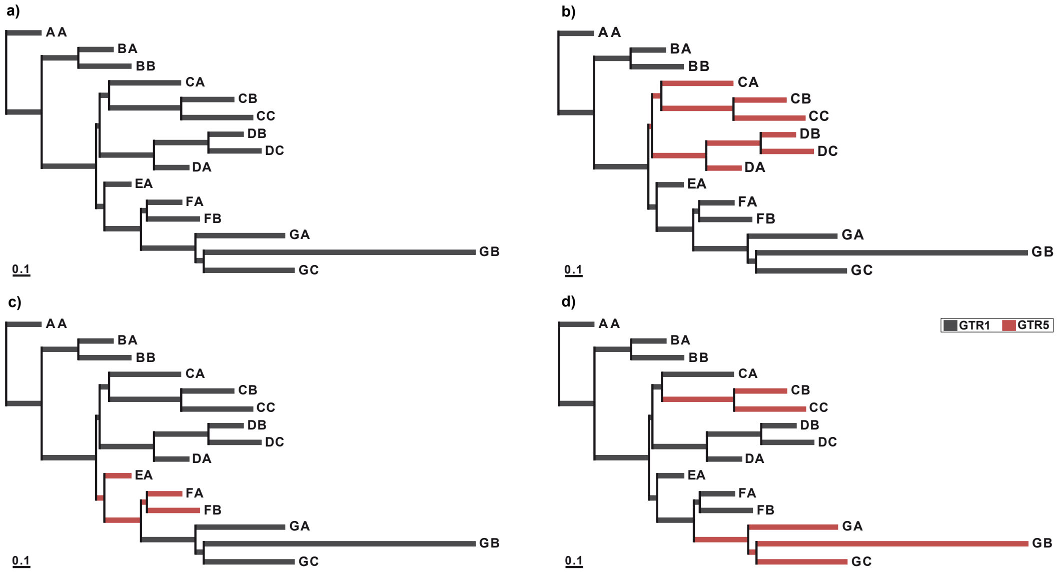


Figure 3.4: Topology 3 (datasets 3a, 3b, 3c and 3d, 15 taxa), branch lengths and models used in analyses, covering a spectrum of stationary (a) and non-stationary (b-d) datasets. These four sets of parameters differ in the local application of evolutionary models (see table 3.2 and 3.1).

set to 0 or with a continuous gamma distribution with a specified proportion of invariable sites (see table 3.1).

Table 3.1: Model specifications and parameters used for the simulations. All parameter sets consist of GTR (general time-reversible) substitution rates, base frequencies, an α value for Γ-distributed rate heterogeneity (α) and the proportion of invariant sites (p_{inv}).

			Substitution rates						Base frequencies			
Model	α	p_{inv}	AC	AG	AT	CG	CT	GT	f_A	f_C	f_G	f_T
GTR1	0.75	0.3	0.2	1.0	0.7	0.3	0.5	1.0	0.25	0.20	0.25	0.30
GTR2	1.0	0.2	0.6	1.0	0.4	0.1	0.5	1.0	0.40	0.10	0.25	0.25
GTR3	0.8	0.1	0.8	1.0	0.6	0.1	0.3	1.0	0.25	0.30	0.25	0.20
GTR4	0.5	0.5	0.4	1.0	0.3	0.5	0.2	1.0	0.40	0.25	0.10	0.25
GTR5	1.0	0.4	0.4	1.0	0.9	0.2	1.0	1.0	0.30	0.10	0.20	0.40

Table 3.2: Combinations of setups and simulation options. Several combinations of topologies, branch lengths and evolutionary models were simulated. All simulations using continuous gamma-distribution were generated with a proportion of invariant sites (p_{inv}). The simulation runs using discrete gamma-distribution with four rate categories were all performed using p_{inv}=0 for all different models (see table 3.1).

Topology	Dataset	Models (see table 3.1)	Parameters
1	1a	GTR1	GTR+Γ+I (continuous Γ) or GTR+Γ (discrete Γ)
	1b	GTR1/2	
	1c	GTR1/2/3/4	
2	2a	GTR1	
	2b	GTR1/2	
	2c	GTR1/2/3/4	
3	3a	GTR1	
	3b	GTR1/5	
	3c	GTR1/5	
	3d	GTR1/5	

Empirical Data Several published datasets were chosen for parametric bootstrapping and split analyses. Nucleotide MSAs presenting a diverse combination and mixture of organisms were selected to cover a spectrum of different species and branches within the tree of life (table 3.3).

Table 3.3: Empirical datasets used in the analyses. They were chosen accordingly to cover a spectrum of different species and clades within the tree of life.

Dataset	original		masked		Reference
	#Taxa	#Sites	#Taxa	#Sites	
Monocots/Dicots	14	26,976	14	26,976	Goremykin *et al.* (2005)
Cormorants/Shags	33	1,141	28	1,000	Holland *et al.* (2010)
Vertebrata	25	13,856	25	13,855	Phillips *et al.* (2004)
Fungi/Metazoa	32	36,180	19	19,707	Rokas *et al.* (2005)
Malacostraca	28	2,365	28	1,413	Wägele *et al.* (2003)

The published MSAs had to be preprocessed for the analysis. The impact of gapped or missing positions and ambiguous data can neither be proven nor accurately be reflected by the parametric bootstrap datasets to be comparable. Therefore, the empirical MSAs were screened for missing and ambiguous data, and gaps. To keep as much data as possible, all sequences of the MSAs were checked if they contain over-proportional amount of missing or ambiguous data or gaps and either taxa or sites were masked. Table 3.3 shows the size of the datasets before and after masking.

3.2.2. Phylogenetic Analyses and Parametric Bootstrapping

ML Simulated datasets were analysed with PhyML 3.0 (Guindon *et al.*, 2010) using different options. The gamma-shape parameter (α) was estimated by approximating with (i) 4, (ii) 12 or (iii) 25 rate categories, either using mean or median to analyse whether this makes a difference for the adequacy of the estimated model. Since the number of rate categories for approximating the gamma distribution turned out to produce no significant difference for the occurrence of over- or underrepresented splits, all empirical datasets were analysed with PhyML with only four rate categories for approximating the gamma distribution. The α was estimated by using the mean option. All described analyses of the simulated datasets were carried out estimating either

p_{inv} or fixing p_{inv}=0 (all sites are assumed variable and allocated to gamma-rates).

Parametric Bootstrapping To constitute as a reference set to test if over- or underrepresented splits occur using the true parameters, for all simulated datasets parametric bootstrap datasets were generated based on the model conditions of the original dataset (see Fig. 3.5). For these datasets and corresponding bootstrap MSAs it is expected that no over- or underrepresented splits occur for the split residual diagnostics.

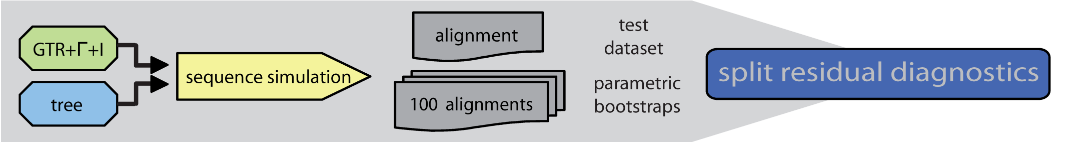


Figure 3.5: Flowchart of testing the analysis. For every simulated dataset the used model parameters are used again to build 100 parametric bootstrap alignments. These datasets are then used as a reference rest set for the split residual diagnostics.

Subsequently, for all ML analysed datasets, i.e. the simulated and the empirical datasets, the best fitting trees including branch lengths, base frequencies, substitution parameters, α and p_{inv} were used to generate parametric bootstrap datasets (see Fig. 3.1).

3.2.3. Split Analysis: Over- and Underrepresented Splits

Both, the empirical datasets and the corresponding parametric bootstrap datasets, were analysed for RY splits. Only splits which occurred more than five times were taken into account, since an occurrence of less than five times is expected as not significant for datasets with mostly more than 10,000 sites.

All occurring splits were recorded and the sites supporting these were summed up and listed as observed positions. Based on this, the bootstrap datasets were screened for the recorded splits and all supporting sites were summed up. Splits, which occurred in the original dataset more often than in every corresponding parametric bootstrap dataset were classified as overrepresented. Splits which occurred less often were classified as underrepresented.

The extent of deviation of split spectra between both, the analysed dataset and its bootstraps, can be estimated by calculating the difference between the observed and expected (mean) number of split occurrences. The mean value of occurrence for a split ($O_{S_{mean}}$) present in the parametric bootstrap datasets was calculated by dividing the total number of occurrence by the number of bootstrap datasets (see equation (1)).

$$O_{S_{mean}} = \sum_{i=1}^{n} \frac{O_{S_i}}{n} \tag{1}$$

where

O_{S_i} = amount of occurrence for a split S in bootstrap replicate i
S = each recorded split S in the original dataset
n = number of bootstrap replicates

If the estimated evolutionary model fits the dataset adequately, we would expect that the mean amount of occurrence for a split S over all n bootstrap replicates ($O_{S_{mean}}$) is similar to the observed amount of occurrence for a split S in the original dataset ($O_{S_{org}}$, see equation (2)).

$$O_{S_{org}} \approx O_{S_{mean}} \tag{2}$$

Two new MSAs are created, which represented the over- or underrepresented splits equal to the difference in amount of occurrence ($O_{S_{diff}}$) of a split S. This difference is determined between the observed amount of occurrence (original dataset) and the expected (mean) amount of occurrence over all bootstrap replicates (see equation (3)).

$$O_{S_{diff}} = O_{S_{org}} - O_{S_{mean}} \tag{3}$$

The new MSAs are visualized as Neighbor-Net networks (Bryant & Moulton, 2004) (OrdinaryLeastSquares variance) in SplitsTree 4 (Huson & Bryant, 2006), version 4.12.3.

3.3. Results

3.3.1. Simulated Datasets

All tests with the reference sets for the simulated MSAs showed no over- or underrepresented splits. The results for the discrete gamma-distributed datasets (table 3.4) showed a homogeneous distribution of over- and underrepresented splits for both, stationary and non-stationary datasets. In contrast to this, there is a certain increase of over- and underrepresented splits for the continuously simulated datasets (table 3.5) from stationary to non-stationary datasets. Though, the number of over- or underrepresented splits does not vary significantly between datasets which were analysed using different numbers of rate categories to approximate the gamma distribution.

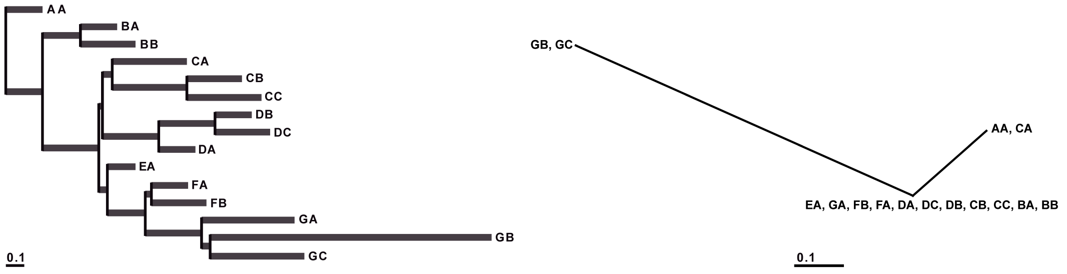


Figure 3.6: Stationary simulation 3a: all branches were simulated with the same evolutionary model. Left: underlying tree/model combination of analysed dataset, estimated: GTR+Γ+I, with four rate categories and using the mean to approximate the gamma distribution. Right: Neighbor-Net visualization of overrepresented splits resulting from parametric bootstraps and RY split analysis.

The overrepresented splits are always more frequent, while the underrepresented splits are in most cases rare (see tables 3.4 and 3.5).

The generated MSAs which reflect the over- and underrepresented splits proportionally to their deviation to the mean occurrence of the bootstrap datasets could be visualized and analysed within Neighbor-Net networks. Simulation 3a (see Fig. 3.6, left panel) is based on a tree whose branches are all evolving under the same evolutionary model. In this case only a few overrepresented splits are present (see Fig. 3.6, right panel).

The red coloured branches of the topology of the other datasets (Figs. 3.7, 3.8 and 3.9) were simulated with different evolutionary parameters. Taxa, which are not captured by the estimated model can be detected within the

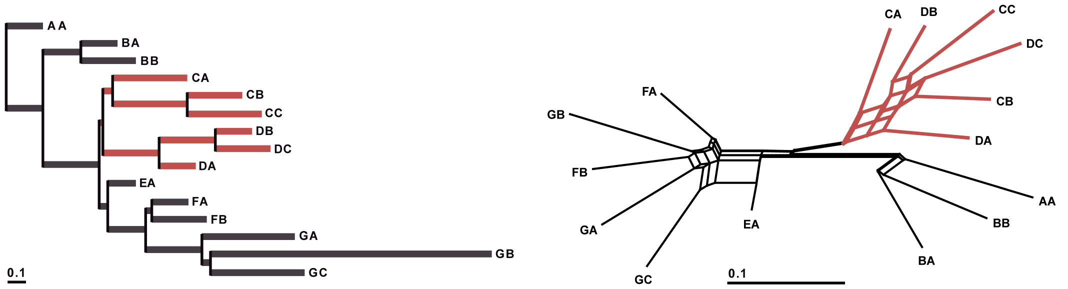


Figure 3.7: Non-stationary simulation 3b: the red coloured branches were simulated with different evolutionary model parameters than the black coloured branches. Left: underlying tree/model combination of analysed dataset, estimated: GTR+Γ+I, with four rate categories and using the mean to approximate the gamma distribution. Right: Neighbor-Net visualization of overrepresented splits resulting from parametric bootstraps and RY split analysis.

overrepresented splits showing long branches. For simulation setup 3d (Fig. 3.9) there are splits present, which group branches that evolve according to same parameters, although they had no monophyletic origin.

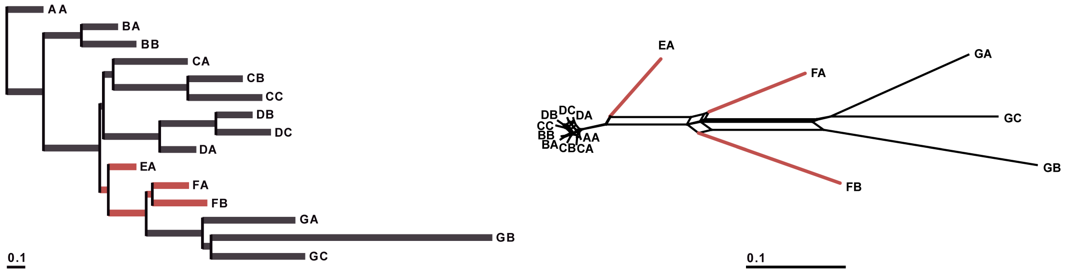


Figure 3.8: Non-stationary simulation 3c: the red coloured branches were simulated with another evolutionary model than the grey coloured branches. Left: underlying tree/model combination of analysed dataset, estimated: GTR+Γ+I, with four rate categories and using the mean to approximate the gamma distribution. Right: Neighbor-Net visualization of overrepresented splits resulting from parametric bootstraps and RY split analysis.

Table 3.4: Results of the analysis for over- and underrepresented splits of model-homo- and heterogeneous datasets. All datasets were simulated with discrete GTR+Γ (four rate categories) and analysed with GTR+Γ+I.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or underrepresented splits;
split = amount of splits detected as over- or underrepresented;
sites = number of sites which represent all over- or underrepresented splits;
green cells = no over- or underrepresentation.

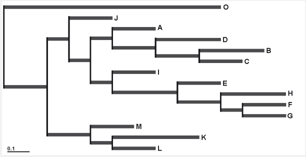
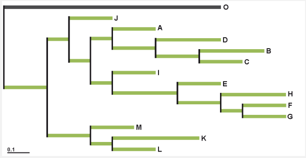
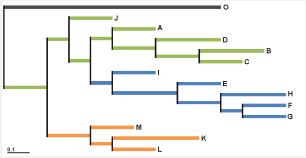
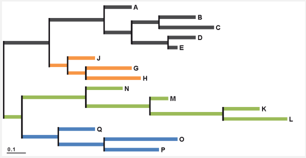
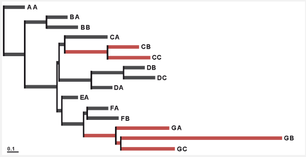
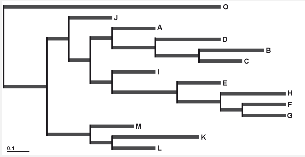
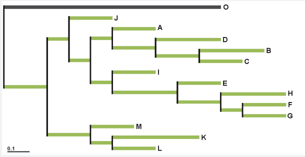
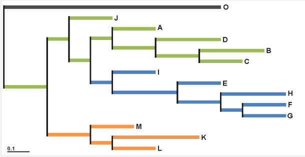
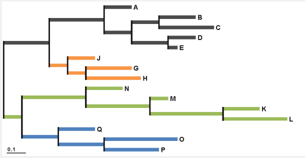
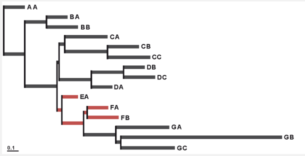
Dataset	ML analysis options		Over		Under	
			splits	sites	splits	sites
setup 1a  O J A D B C I E H F G M K L 0.1	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	3	22	0	0
		median	3	22	0	0
	GTR+Γ+I, 12 cat	mean	3	147	0	0
		median	3	24	0	0
	GTR+Γ+I, 25 cat	mean	4	160	0	0
		median	2	120	0	0
setup 1b  O J A D B C I E H F G M K L 0.1	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	3	16	0	0
		median	6	36	1	116
	GTR+Γ+I, 12 cat	mean	4	154	0	0
		median	4	107	0	0
	GTR+Γ+I, 25 cat	mean	3	154	0	0
		median	4	139	0	0
setup 1c  O J A D B C I E H F G M K L 0.1	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	6	65	1	21
		median	7	66	1	29
	GTR+Γ+I, 12 cat	mean	6	171	2	48
		median	5	59	2	48
	GTR+Γ+I, 25 cat	mean	6	175	2	47
		median	7	174	2	47
setup 2c  A B C D E J G H N M K L Q O P 0.1	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	8	89	3	62
		median	6	91	3	156
	GTR+Γ+I, 12 cat	mean	3	39	1	20
		median	3	66	1	20
	GTR+Γ+I, 25 cat	mean	3	40	1	20
		median	4	78	1	20
setup 3d  AA BA BB CA CB CC DB DC DA EA FA FB GA GB GC 0.1	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	7	195	0	0
		median	6	190	2	143
	GTR+Γ+I, 12 cat	mean	8	257	1	17
		median	5	210	1	18
	GTR+Γ+I, 25 cat	mean	8	262	1	17
		median	8	256	2	39

Table 3.5: Results of the analysis for over- and underrepresented splits of model-homo- and heterogeneous datasets. All datasets were simulated with continuous GTR+Γ+I and analysed with GTR+Γ+I.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or underrepresented splits;
split = amount of splits detected as over- or underrepresented;
sites = number of sites which represent all over- or underrepresented splits;
green cells = no over- or underrepresentation.
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Dataset	ML analysis options		Over		Under	
			splits	sites	splits	sites
setup 1a 	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	5	49	1	151
		median	7	61	1	181
	GTR+Γ+I, 12 cat	mean	6	53	0	0
		median	5	39	0	0
	GTR+Γ+I, 25 cat	mean	4	34	0	0
		median	8	63	0	0
setup 1b 	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	5	565	13	402
		median	6	545	12	450
	GTR+Γ+I, 12 cat	mean	3	543	10	187
		median	4	537	10	306
	GTR+Γ+I, 25 cat	mean	6	565	12	213
		median	7	567	7	129
setup 1c 	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	40	885	6	231
		median	42	896	5	288
	GTR+Γ+I, 12 cat	mean	38	844	5	162
		median	45	918	5	163
	GTR+Γ+I, 25 cat	mean	37	831	7	185
		median	38	851	4	128
setup 2c 	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	43	930	7	182
		median	45	963	7	311
	GTR+Γ+I, 12 cat	mean	43	917	8	186
		median	41	914	9	185
	GTR+Γ+I, 25 cat	mean	40	916	6	199
		median	39	872	10	229
setup 3c 	reference set		0	0	0	0
	GTR+Γ+I, 4 cat	mean	51	1706	21	931
		median	53	1771	24	1070
	GTR+Γ+I, 12 cat	mean	52	1725	25	867
		median	53	1713	20	824
	GTR+Γ+I, 25 cat	mean	53	1714	22	851
		median	53	1713	23	861

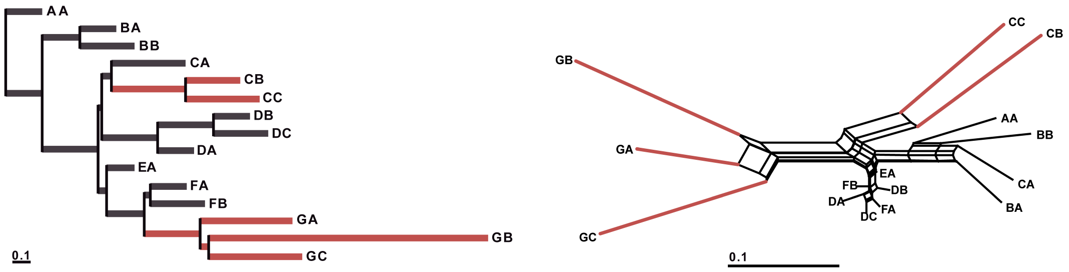


Figure 3.9: Non-stationary simulation 3d: the red coloured branches were simulated with another evolutionary model than the grey coloured branches. Left: underlying tree/model combination of analysed dataset, estimated: GTR+Γ+I, with four rate categories and using the mean to approximate the gamma distribution. Right: Neighbor-Net visualization of overrepresented splits resulting from parametric bootstraps and RY split analysis.

3.3.2. Empirical Datasets

For the datasets of Holland *et al.* (2010) and Wägele *et al.* (2003) neither over- nor underrepresented splits were found.

The Neighbor-Net network of the overrepresented splits of the dataset of Goremykin *et al.* (2005), containing monocots and dicots, shows very long branches for the outgroup species. Further, the grasses (*Zea*, *Triticum* and *Oryza*) are grouped together, but they share strong split support with *Oenothera*, which belongs to the eudicots. *Calycanthus* shows more splits in common with *Amborella*, than the estimated model shows.

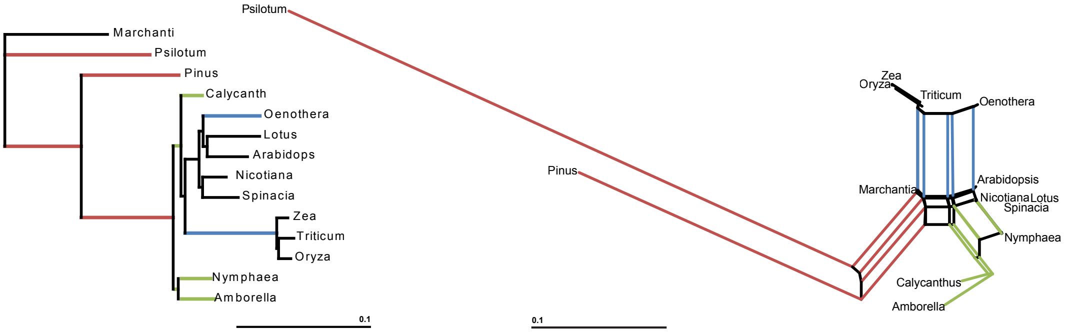


Figure 3.10: Monocots/Dicots, (Goremykin *et al.*, 2005) - ML tree and split network. On the left hand side, the split networks for the overrepresented splits are shown.

Rokas *et al.* (2005) published a dataset spanning a wide range of Fungi and Metazoa. A long branch within the ML tree clearly separates Metazoa from

Fungi. Fig. 3.11 shows that this separation is also present within the overrepresented splits.

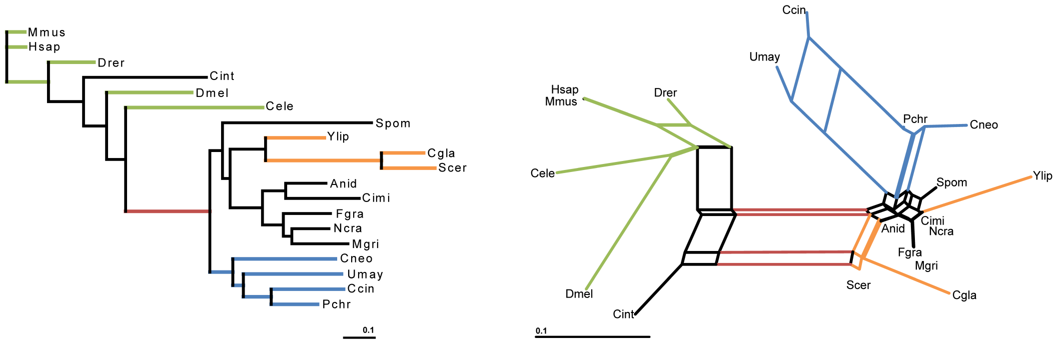


Figure 3.11: Fungi/Metazoa, (Rokas & Carroll, 2005) - ML tree and split network. On the left hand side, the split networks for the overrepresented splits are shown.

The dataset published by Phillips *et al.* (2004) includes a wide range of Eukaryota. The overrepresented splits (see Fig. 3.12, right hand side) show that the chosen evolutionary model does not seem to fit adequately for several taxa. There are long branches for Mammalia as well as for aves, which separate them from the amphibians.

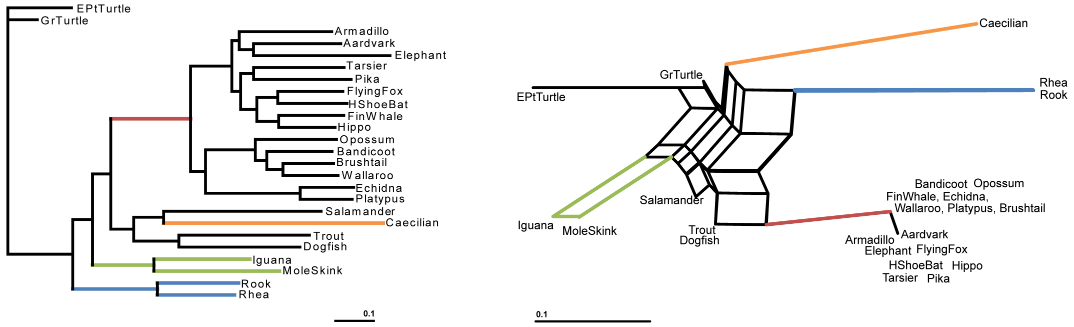


Figure 3.12: Vertebrata, (Phillips *et al.*, 2004) - ML tree and split network. On the left hand side, the split networks for the overrepresented splits are shown.

3.4. Discussion

For this study a number of different nucleotide MSAs, based on different topologies, branch lengths and evolutionary models were simulated to test the effect of model heterogeneity caused by different extends of non-stationarity on model adequacy. While the arrangements of evolutionary models varied for different branches or clades, the ML analysis estimates only one set of model parameters to cover the evolutionary processes of the whole tree. These estimated models were used as basis for parametric bootstrap datasets, which then represent a distribution of MSAs that reflect the estimated model parameters. Thus, those datasets represent what can be expected, if the model fits adequately to the analysed original dataset. If the model covers the data well, the original and the bootstrap datasets should show the same characteristics, such as the split spectrum. The split spectra of the original and the bootstrap datasets were screened for splits, which occur over- or underrepresented within the original dataset compared to their occurrence within the parametric bootstrap datasets.

The analyses of the simulated datasets showed in most cases more over- than underrepresented splits. This implies that the irregularities of model fit have a much higher impact on an increase of occurrence of certain splits. This could be a hint, that the chosen models are not to complex or over-parametrised. A significant variance in amount of overrepresentation depending on the number of used categories for modelling the gamma distribution in the ML analysis was not observable. This can be due to the balancing effect of using an RY recoding and splits instead of analysing the pattern distribution, for which analysing with only four rate categories for approximating the gamma distribution can be a cause for model-misspecification (see chapter 2).

For the datasets which were simulated using four rate categories for approximating the gamma distribution only a few overrepresented or underrepresented splits were observed (see table 3.4). Also there were only marginal differences between stationary and non-stationary datasets. Contrary to this, the analyses of the simulated datasets with a continuous gamma distribution (see table 3.5) showed a similar distribution of over- or underrepresented splits for stationary datasets, but a tremendous increase of overrepresented splits for non-stationary datasets. This indicates, that the estimated model can not entirely handle the differences of the used model parameters and consequently causes a deviation of occurrence for certain splits. Moreover, since the results for the datasets which were simulated with a discrete gamma distribution show split residuals to a lesser extent than the simulations based on a continuously gamma distribution, it seems as if the heterogeneity of model parameters combined with a continuous gamma distribution and therefore non-stationarity leads to an inadequacy of the estimated models.

For a simulation setup with short branch lengths, there was no noticeable divergence of over- or underrepresentation and only few splits identifiable (see Appendix, Fig. A.2 and tables A.7, A.8, A.9 and A.10). This indicates, that model adequacy is harder to assess the longer the branches are. However, there is no clear pattern which can be derived to decide, if a dataset evolved too heterogeneous (in the aspect of stationarity) to be assessed correctly by only one chosen statistical model. But a certain amount of overrepresented splits may indicate an inadequacy.

To get a closer look at the splits which are overrepresented, the MSAs representing the residual splits were visualized in SplitsTree using Neighbor-Net. For the stationary dataset the Neighbor-Net network showed only overrepresented splits for the outgroup and the clade containing the longest branch (Fig. 3.6). This shows, that the estimated model does not lead to a concerning amount of split residuals and therefore explains the data quite well. Only the complexity of taxa which cause long branch distances based on outgroup choice are clearly observable and cannot be be explained in every detail by the model.

With a clade evolving according to different model parameters within the sequence set (Fig. 3.7), the estimated model has to balance these differences with only a single parameter set for the whole tree. This leads to an increase of overrepresented splits and divides the Neighbor-Net network into three groups, separating the sequences which evolve according to a different set of evolutionary parameters from its sister group and the outgroup sequence. This implies, that the model is not able to adequately cover a wide range of splits, in most cases trivial splits, if a certain amount of sequences evolves differently. A dataset for which the sequences are evolving according to two different sets of evolutionary parameters in a ratio of one to two leads to an estimated model which is not able to fit adequately neither the first nor the second clade. Instead it leads to balanced occurrences of split residuals (over- or underrepresented splits) for the whole sequence set.

Clades with nested branches which evolve according to a different set of model parameters, lead to an increased number of trivial splits for the involved sequences and the 'following' sequences with the longest branches. For the clade with the highest amount of sequences which evolved according to the same evolutionary parameters (see Fig. 3.8) this leads in contrast only to small split deviations. This indicates, that the chosen model fits better for the largest stationary sequence set, whereas the model seems to be inadequate for the nested sequences which evolved under different parameters, as well as for the 'following' sequences, since a lot of split occurrence can not be explained by the model parameters.

For a dataset with two nested clades evolving under the same model conditions, some overrepresented splits are supporting a clustering of both clades, even though they are clearly separated in the underlying tree (Fig. 3.9). This indicates that the chosen model is not able to describe this condition of the original dataset, which can lead to bias during tree reconstruction.

All Neighbor-Net networks of non-stationary datasets show an increase of overrepresented trivial splits, which divide one sequence and the rest, mostly for the sequences evolving according to different parameters than the largest stationary clade. It seems as if this behaviour is caused by the amount of sequences evolving stationary or the ratio of amount of sequences, which evolve under different conditions. Datasets 3c and 3d include different clades, even paraphyletic ones, which do not induce an increase of overrepresented trivial splits. Here, it should be analysed further, if the occurrences of trivial splits are due to an increase of invariant sites for the sequences which evolved according to a second model (red coloured branches for setup 3).

The datasets by Holland *et al.* (2010) and Wägele *et al.* (2003) showed neither over- nor underrepresented splits. This could be caused by the high amount of sequences in relation to the number of sites. Further, the chosen sequences are more closely related than the sequences of the other datasets allowing for a single model to better fit the underlying evolutionary processes.

The dataset of Goremykin *et al.* (2005) contains only 14 sequences with 26,976 sites, which is more likely to offer a big variety of splits in comparison to the smaller datasets of Holland *et al.* (2010) and Wägele *et al.* (2003). The Neighbor-Net networks of overrepresented splits (Fig. 3.10) show long branches for the outgroup species. The grouping of grasses (*Zea*, *Triticum* and *Oryza*) and *Oenothera*, which belongs to the eudicots, and *Calycanthus* sharing splits with *Amborella* imply, that there is signal within the dataset, which cannot be described well by the estimated model.

The Neighbor-Net networks of overrepresented splits for the datasets of Rokas *et al.* (Fig. 3.11) and Phillips *et al.* (Fig. 3.12), both including a wide range of organisms in their datasets, reflect well the separation of Metazoa from Fungi (Rokas & Carroll, 2005), and Mammalia from aves and amphibians (Phillips *et al.*, 2004). It is striking that the Fungi/Metazoa split appears dominantly within the set of overrepresented splits. This indicates, that the use of a single model for such a wide range of taxa may balance differences, which covers phylogenetic information insufficiently and in consequence can lead to aftereffects. A further study should analyse Metazoa and Fungi seperately and how this has an effect on the estimated model parameters as well as on the overrepresented splits.

For the vertebrate dataset (Phillips *et al.*, 2004) the overrepresented splits show a comparable pattern to the ones of setup 3c and 3d. The largest subgroup (Mammalia) which can be supposed to be closely related and therefore is most likely to share similar evolutionary parameters, are clustered together within the overrepresented splits. Similar to this, a strong overrepresented split clusters the sequences of aves together and divides them from Mammalia and amphibians. In contrast to this, the sequences of the amphibians show a huge amount of trivial splits which are present within the split residuals. This can be observed as well within datasets of setup 3c and 3d for the smaller group of sequences which evolved stationary. In fact, in comparison to Mammalia the amphibians constitute a smaller subgroup of the complete sequence set. Similar to the dataset of Rokas *et al.* (2005), the strong splits which separate Mammalia from aves and amphibians and aves from Mammalia and amphibians should be covered by the model instead of appearing within the split residuals. Thus, the overrepresentation of these splits can indicate model-misspecification.

In summary, the application of residual diagnostics in combination with split analysis seems to be a reasonable method for analysing model adequacy. The comparison of observed split frequencies to those expected under the model (tree, branch lengths and model of nucleotide substitution) can indicate, if the underlying datasets are covered well by the chosen model. If this is not the case, it can give impressions whether the sequence set is too unbalanced (i.e. occurrence of overrepresented trivial splits) or if there is a shift of evolutionary processes, which might be a reason to apply different submodels (strong split separating clades, Rokas *et al.*, 2005, Phillips *et al.*, 2004). Conveniently, the results can be interpreted by visualizing them within Neighbor-Net networks, but they do not approve a formalisation. Nevertheless, since RY recoding does not take SRH conditions fully into account, the method can potentially be improved by analysing splits without recoding.

Nicht alles, was gezählt werden kann, zählt, und nicht alles, was zählt, kann gezählt werden.
Albert Einstein

4. Split Analysis Methods and Split Weighting - Split Analysis of Aligned Nucleotide Sequences

4.1. Introduction

Nucleotide MSAs can easily be suboptimal and lead to erroneous conclusions of phylogenetic relationships if based on wrong assumptions or if they contain uninformative regions due to strong substitutional saturation or strongly conserved sections without phylogenetic signal (Dress *et al.*, 2008). Additionally, empirical data rarely support just one unique tree because of the presence of diverse and incongruent signal (Goremykin *et al.*, 2005). All of these phenomena can influence the accuracy of the tree inference (Flook & Rowell, 1997; Felsenstein, 2004; Susko *et al.*, 2005; Löytynoja & Goldman, 2005; Ogden & Rosenberg, 2006b; Misof & Misof, 2009).

Since tree inference by itself cannot help to identify confounding factors, Bandelt and Dress proposed to combine phylogenetic analyses with a non-approximative distance method of split decomposition (Bandelt & Dress, 1992), and Hendy, Penny and Steel (1993; 1994) proposed the application of spectral analysis, related to the Hadamard conjugation, to accompany tree inferences.

Both approaches investigate an MSA based on splits without direct tree inference. A split is a bipartition with two subsets of sequences of the complete sequence set characterized by distinct features. Within a phylogenetic context based on nuclear sequence data these features are differences in nucleotide or amino acid character states of an MSA. Theoretically, for a set of n taxa, there are 2^{n-1} possible bipartitions (splits). In empirical datasets, however, the number of realized splits is generally smaller. Splits are directly related to tree topologies in a simple form. If there is exactly the same number of splits in a dataset as there are edges in a possible binary topology, and if all the splits are compatible to each other, only one tree topology is supported by the dataset. Two splits A|B and C|D are compatible, if one of A∩C, A∩D, B∩C and B∩D is empty.

The Hadamard conjugation transforms a split spectrum (an enumeration of observed splits, given n taxa) into a split vector describing the transformation between split spectrum and MSA using evolutionary models. Lento *et al.* (1995) used this approach to filter conflicting signals. But since this method estimates support for all possible splits of a dataset (2^{n-1}), the processing effort grows exponentially with the number of taxa, this is clearly too computationally intensive when handling more than 25-30 sequences.

Besides this, the complexity of nucleotide character states can be reduced to binary character states by considering just purines and pyrimidines (RY splits). This approach is much more computationally efficient, but lacks signal disclaiming the information within the subsets. Software packages offering split methods are for example Spectrum (Charleston, 1998), Spectronet (Huber *et al.*, 2002), SplitsTree 4 (Huson & Bryant, 2006), PHYSID (Wägele & Rödding, 1998), and SAMS (Mayer & Wägele, 2005). They calculate and display for example Lento-plots (Lento *et al.*, 1995), median networks (Bandelt *et al.*, 1995) and Neighbor-Net networks (Bryant & Moulton, 2004).

SAMS (*S*plits *A*nalysis *M*ethod*s*) is a tool using a version of the so-called 'PHYSID' method developed first by Wägele and Röding (1998), which was later refined and extended by Mayer and Wägele (2005). Both tools try to dissect phylogenetic signal in a dataset (Wägele & Mayer, 2007). SAMS considers all four nucleotide character states but unlike the Hadamard transformation, analyses an MSA only for observed splits.

In a first step, all splits present in the data are listed. Then, all sites of the MSA are evaluated if they support each or some of those splits. If a site supports a split, SAMS categorises this site support according to three different degrees of quality. It differentiates between 'binary' and 'noisy' splits. The latter can be refined to 'noisy only for one subset' or 'noisy for both subset'. Sites account for 'binary' support, if a certain character occurs only within one subset, e.g. AAA/CCC or AAA/GGT, and not within the other. They account for 'noisy outgroup', if a certain character occurs to 100% within one subset, and up to 25% (by default) within the other subset, e.g. AAAA/ACCC or CCCC/CGTT. Sites account for 'noisy in- and outgroup' support as well, if both subsets contain character states up to a certain amount, which are predominant for one of the subset, e.g. AAAC/ACC or GGGT/GTTC. By default, one subset must have at least 75% character identity and the nucleotide which is predominant within this subset is allowed to occur in the other subset up to 25%. After determination of supporting positions, the sequences are re-evaluated to exclude split support caused by chance similarities. Therefore, supporting positions of every sequence are checked pairwise against the corresponding positions of all other sequences or a consensus sequence of the other split subset. A consensus sequence is established by counting the occurring characters for each site. If the most frequent character of a site occurs with a proportion larger than or equal to a chosen consensus threshold (default=50%), and if no other character occurs with this same proportion, this character is chosen as the consensus character. Otherwise the consensus character is set as missing state. If the similarity of this comparison is higher than a certain threshold (default=25%), the detected positions will not longer be counted as supporting positions. This is done to avoid biased positions with accumulated 'outgroup states' which could be plesiomorphies. In a final step, all supporting positions for every observed split are counted.

The software delivers an output file, which lists all identified splits and their total support, as well as values for different support qualities of which the total support is composed. The proportions which limit how many dominant characters are allowed to occur within the other subset and of all other parameters, have default values, but can be adjusted without restrictions. SAMS is written in *C++* and is command line based, therefore it is flexible and platform independent.

To enhance user-friendliness of the software package, a platform independent general user interface (GUI) was developed, SAMS GUI (Meid *et al.*, 2012), which allows to adjust easily the parameters offered by SAMS (Mayer & Wägele, 2005) and visualises the results within a split support spectrum. This plot can be exported as image or scalable vector graphic file. Moreover, it offers the possibility to evaluate the identified splits for compatibility.

The weighting of splits in SAMS lacks a formal foundation. All limiting parameters have default values, but can be freely adjusted by the user. Therefore, a new approach was developed which re-evaluates the splits found by the SAMS algorithm (Mayer & Wägele, 2005) to assess a more objective split weighting. The support of a split is rated based on the contrast between the subsets of a split and on base of the nucleotide variance within a split subset.

To evaluate the new approach, four topology setups were designed and used to simulate nucleotide MSAs. All setups were analysed with SAMS and the new weighting scheme.

4.2. Materials and Methods

4.2.1. Software Development

For split search SAMS version 1.4.3 (Mayer & Wägele, 2005) was used and implemented within SAMS GUI (Meid *et al.*, 2012). As programming language *C++* was used including the library Qt UI framework version 4.7.3 (Nokia *et al.*, 2011) for the development of the new GUI.

4.2.2. Weighting of Splits

A new approach was developed to rate the support of a certain split S within a multiple nucleotide MSAs. In this approach, the support of the split patterns is estimated via two criteria, the contrast weighting and the quality weighting. The contrast weighting (C value) of a split S is calculated for every site s of a given nucleotide MSA. The quality weighting (Q value) refers to each subset (*g1* and *g2*) of a split S for every site s within the MSA. To assess the 'weight' (CQ value) for a certain split S, the C and Q values are multiplied for every subset of every site s of the MSA and then summed up for each subset of the split (*g1* and *g2*).

C value weighting of the characters, contrast between two subsets
Given a nucleotide MSA which is divided into two subsets of the complete sequence set, *g1* and *g2*. The contrast C_s of this split with subsets *g1* and *g2* for a site s, is defined by the difference of the 'contribution' of an event i of one nucleotide in proportion to the total number of nucleotides, i.e. the relative frequency (or empirical probability) of an event i for a site s. In this context, an event i is an occurrence of a certain nucleotide for a site.

$$C_s = \frac{|f_{s_{A_{g1}}} - f_{s_{A_{g2}}}| + |f_{s_{C_{g1}}} - f_{s_{C_{g2}}}| + |f_{s_{G_{g1}}} - f_{s_{G_{g2}}}| + |f_{s_{T_{g1}}} - f_{s_{T_{g2}}}|}{2} \qquad (4)$$

where

C_s = contrast of subsets *g1* and *g2* for a certain site s
f_{s_i} = relative frequency (or empirical probability) of an event i for a site s

The C value can then be calculated by summing over all sites:

$$C = \sum_{s=1}^{s_n} C_s \tag{5}$$

C = the contrast of the two subsets of a split S
s_n = total number of sites s

***Q* value** quality of a pattern within a subset, rating of homogeneity
Given a nucleotide MSA which is divided into two sequence subsets, *g1* and *g2*. The quality of pattern Q is calculated from the pattern homogeneity h of each subset of a split S for every site s within the nucleotide MSA:

$$h_{s_g} = \frac{\left(f_{s_{A_g}} \cdot n_A\right) + \left(f_{s_{C_g}} \cdot n_C\right) + \left(f_{s_{G_g}} \cdot n_G\right) + \left(f_{s_{T_g}} \cdot n_T\right)}{N_g} \tag{6}$$

where

h_{s_g} = rating of homogeneity of a pattern of a subset for a certain site s
n_i = number of occurrence of event i
i = an event, occurrence of a nucleotide
N_g = total number of events, e.g. size of the subset

The relative frequency ($f_{s_{i_g}}$) of an event i for a site s within a subset g is calculated by the total number of occurrences of an event (n_i) in proportion to the total number of events:

$$f_{s_{i_g}} = \frac{n_i}{N_g} \tag{7}$$

Therefore, equation (6) can be reduced to:

$$h_{s_g} = f^2_{s_{A_g}} + f^2_{s_{C_g}} + f^2_{s_{G_g}} + f^2_{s_{T_g}} \tag{8}$$

For example, a subset for which all sequences are having the same nucleotide for a certain site, i.e. AAAA, would be classified as homogeneous, and h_{s_g} would therefore be 1 (*max*) according to equation (8). A site with different events for all sequences within a subset, i.e. ACGT, has the highest level of disorder and is therefore the lower boundary (*min*). According to equation (8), $h_{s_g} = (\frac{1}{4})^2 + (\frac{1}{4})^2 + (\frac{1}{4})^2 + (\frac{1}{4})^2 = 0.25$.
To normalise these boundaries for a range between 0 and 1 (see equation (9)), the quality within a subset pattern Q_{s_g} for a certain site is calculated by subtracting the *min* from h_{s_g}. The result is divided by $max - min$ (see equation (10)).

$$[min..max] \longrightarrow [0..1] : f(x) = \frac{x - min}{max - min} \tag{9}$$

$$Q_{s_g} = \frac{\left(h_{s_g} - 0.25\right)}{0.75} \tag{10}$$

The quality of patterns for a certain site s is then calculated by summing up Q_{s_g} of both subsets:

$$Q_s = Q_{s_{g1}} + Q_{s_{g2}} \tag{11}$$

The quality of patterns for all sites, i.e. a split S is then assessed by summing up over all sites:

$$Q = \sum_{s=1}^{s_n} Q_s \tag{12}$$

where

Q_{s_g} = quality of a pattern of a subset g for a certain site s for a split S
Q_s = quality of patterns for a certain site s for a split S
Q = quality of patterns for a split S

***CQ* value** weighting for a certain split

Given the C values for every site s for every analysed split S and the Q values for every subset g for every site for every analysed split S, the CQ support values can be inferred. To calculate the CQ 'weight' of a certain split S the C and Q values are multiplied for every subset of every site s of the MSA (equation (13)) and then summed up for each subset (*g1* and *g2*, equation (14)).

$$CQ_{s_g} = C_s \cdot Q_{g_s} \tag{13}$$

$$CQ = \sum_{s=1}^{s_n} CQ_{s_{g1}} + CQ_{s_{g2}} \tag{14}$$

where

CQ_{s_g} = weighting for a certain split for site s for the subset g

CQ = weighting for a certain split S

By calculating the support of certain splits for a nucleotide MSA using these equations, the CQ values can be calculated and compared.

4.2.3. Sequence Data

To compare the old and new weighting scheme, four MSAs with a length of 10,000 nucleotide sites were simulated using INDELible V1.03 (Fletcher & Yang, 2009) with the JC model of sequence evolution (Jukes & Cantor, 1969), a specified proportion of invariant sites (p_{inv}= 0.3) and a continuous Γ-distribution for among-site rate variation (ASRV) among non-invariant sites with shape parameter α=1.0.

The tree for setup 1 (Fig. 4.1a) contains balanced branch lengths. The aim is to create a dataset with clear phylogenetic signal. Setup 2 is based on a tree (Fig. 4.1b) which includes long terminal branches. This was selected to simulate a dataset incorporating an increased phylogenetic signal-to-noise ratio.

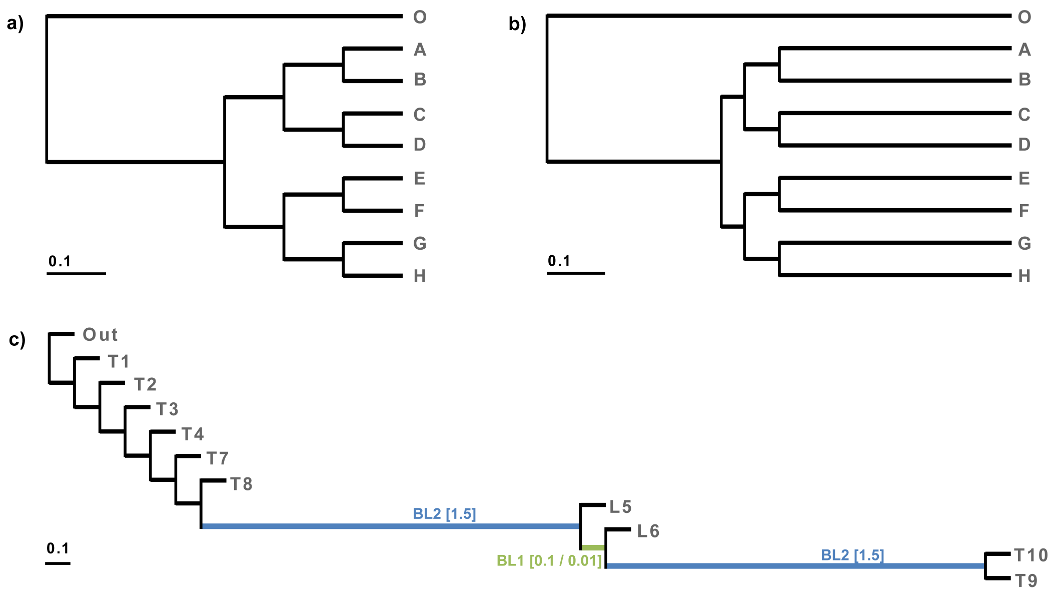
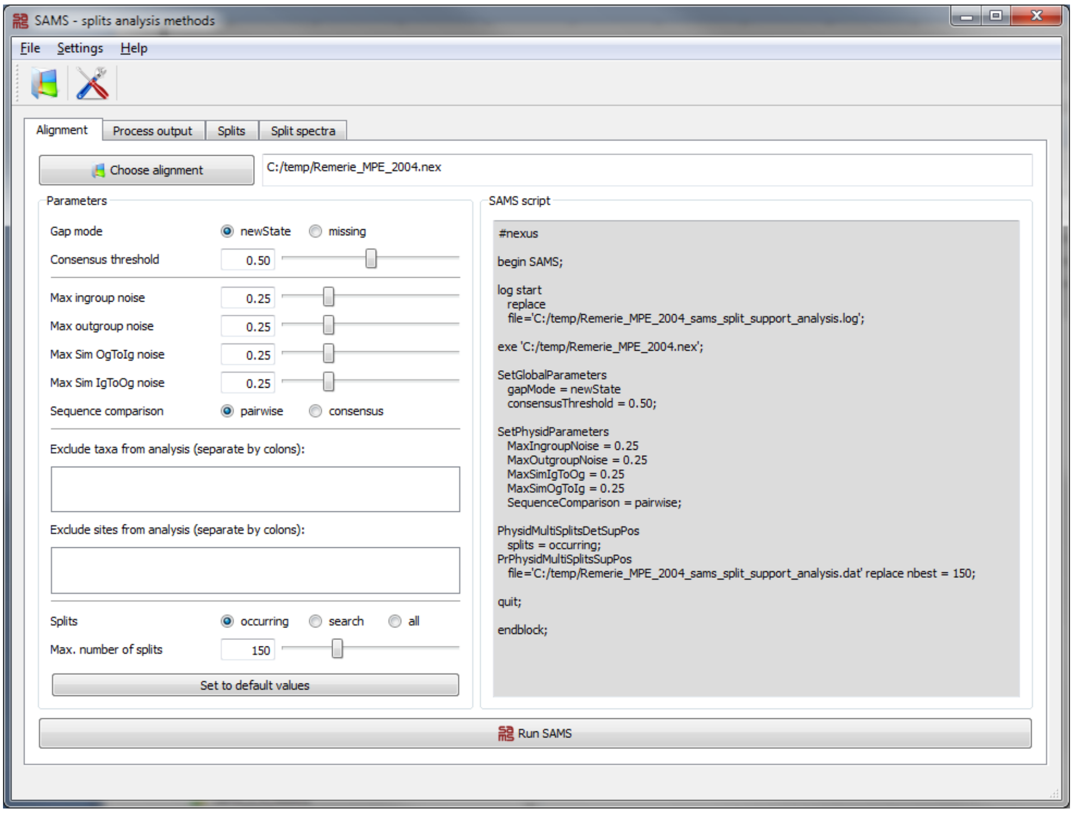


Figure 4.1: Simulation setups used for testing split search and weighting. The first tree (a) for setup 1 is based on balanced branches while setup 2 is based on the second tree (b) which includes long terminal branches. The third topology (c) is used for setups 3 and 4, once with a moderate (setup 3, $BL1 = 0.1$) and with a strong intermediate short branch (setup 4, $BL1 = 0.01$) between two long internal branches.

The third topology (Fig. 4.1c) was taken from a study by Kück *et al.* (2012). It was designed to test tree inference for datasets which are likely to cause long-branch artefacts of class I effects (symplesiomorphy effect) (Wägele & Mayer, 2007). The tree includes two long branches with a nested short branch connecting them. Kück *et al.* (2012) showed that in case of two internal long branches (Fig. 4.1c, *BL2*, in blue) with a length of $\sim$ 1.5 and an extremely short internal branch ($BL1$) of length $\sim$ 0.01 (Fig. 4.1c, in green) the majority of the maximum likelihood analyses of tested datasets failed to infer the correct topologies. The topology was used for (i) setup 3 with a moderate ($BL1 = 0.1$) and for (ii) setup 4 with an extremely intermediate short branch ($BL1 = 0.01$) between the two long internal branches.



C:/temp/Remerie_MPE_2004.nex

Gap mode ● newState ○ missing
Consensus threshold 0.50
Max ingroup noise 0.25
Max outgroup noise 0.25
Max Sim OgToIg noise 0.25
Max Sim IgToOg noise 0.25
Sequence comparison ● pairwise ○ consensus
Exclude taxa from analysis (separate by colons):
Exclude sites from analysis (separate by colons):
Splits ● occurring ○ search ○ all
Max. number of splits 150

```
#nexus

begin SAMS;

log start
  replace
  file='C:/temp/Remerie_MPE_2004_sams_split_support_analysis.log';

exe 'C:/temp/Remerie_MPE_2004.nex';

SetGlobalParameters
  gapMode = newState
  consensusThreshold = 0.50;

SetPhysidParameters
  MaxIngroupNoise = 0.25
  MaxOutgroupNoise = 0.25
  MaxSimIgToOg = 0.25
  MaxSimOgToIg = 0.25
  SequenceComparison = pairwise;

PhysidMultiSplitsDetSupPos
  splits = occurring;
PrPhysidMultiSplitsSupPos
  file='C:/temp/Remerie_MPE_2004_sams_split_support_analysis.dat' replace nbest = 150;

quit;

endblock;
```

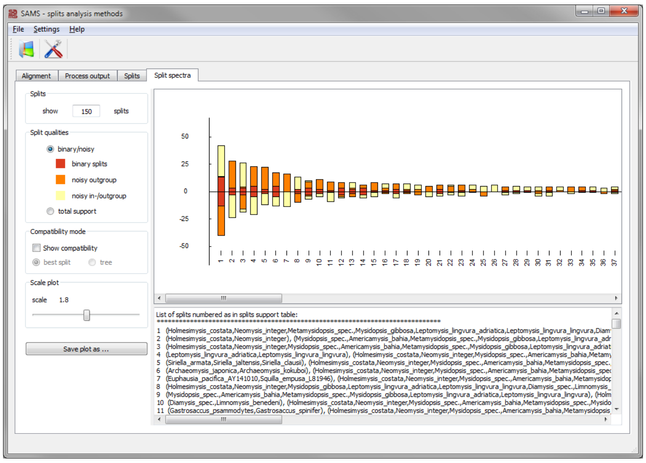
Figure 4.2: SAMS GUI - a nucleotide MSA file is chosen. In the *Parameter* interface different options can be chosen or adjusted. On the right hand side the current SAMS NEXUS block corresponding to the parameter values is shown.

4.3. Results

4.3.1. SAMS GUI

Implementing an intuitive handling for various options remarkably improves the user-friendliness of the software and can be summarized into three major points:

1. It offers an automatic generation of a SAMS NEXUS block (Maddison *et al.*, 1997) by adjusting the parameters by text boxes, value sliders or radio buttons within the parameter interface (see Fig. 4.2 on the left hand side). By starting the SAMS analysis, the generated NEXUS block is directly passed as parameter file to the SAMS analysis. The process output can be tracked within the *Process output* tab. When the analysis is finished, the GUI changes automatically to the tab *Splits*, which shows the output file of SAMS .



show 150 splits

● binary/noisy

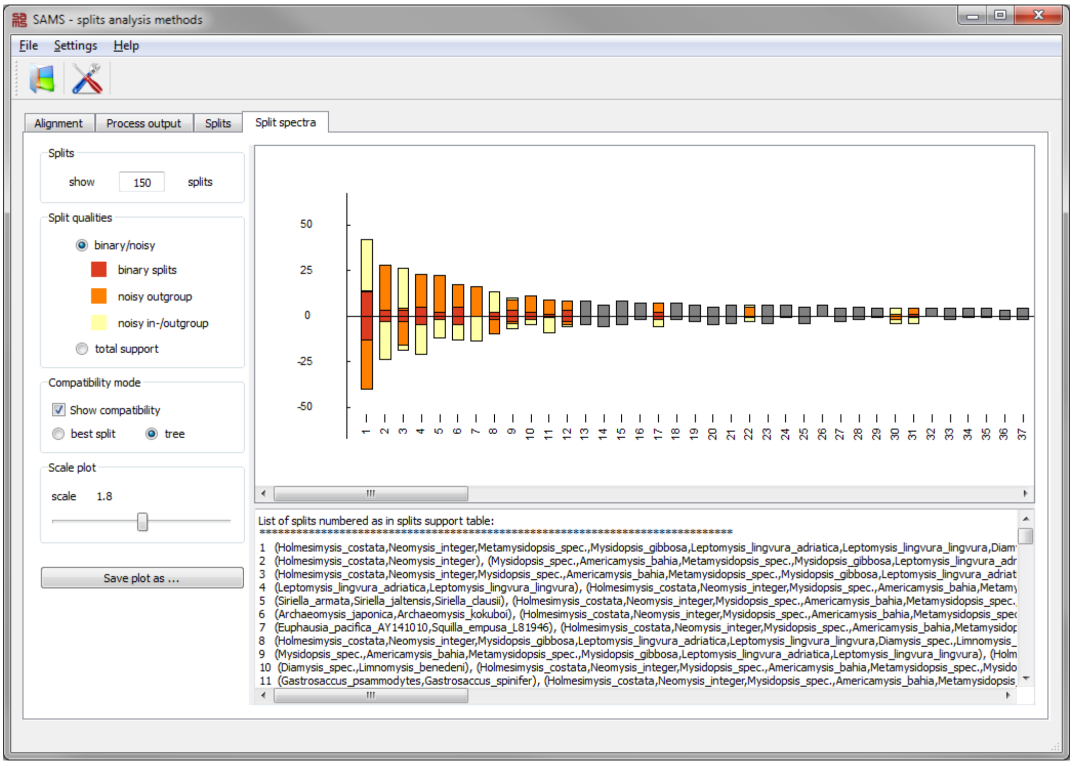
○ total support

□ Show compatibility

○ best split ○ tree

scale 1.8

```
List of splits numbered as in splits support table:
****************************************************************************
1  (Holmesimysis_costata,Neomysis_integer,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura,Diam
2  (Holmesimysis_costata,Neomysis_integer), (Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adr
3  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adriat
4  (Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamy
5  (Siriella_armata,Siriella_jaltensis,Siriella_clausii), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.
6  (Archaeomysis_japonica,Archaeomysis_kokuboi), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec
7  (Euphausia_pacifica_AY141010,Squilla_empusa_L81946), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidop
8  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_gibbosa,Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura,Diamysis_spec.,Limnomysis_
9  (Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura), (Holm
10  (Diamysis_spec.,Limnomysis_benedeni), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysido
11  (Gastrosaccus_psammodytes,Gastrosaccus_spinifer), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_
```



show 150 splits

● binary/noisy

○ total support

☑ Show compatibility

○ best split ● tree

scale 1.8

```
List of splits numbered as in splits support table:
****************************************************************************
1  (Holmesimysis_costata,Neomysis_integer,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura,Diam
2  (Holmesimysis_costata,Neomysis_integer), (Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adr
3  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adriat
4  (Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamy
5  (Siriella_armata,Siriella_jaltensis,Siriella_clausii), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.
6  (Archaeomysis_japonica,Archaeomysis_kokuboi), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec
7  (Euphausia_pacifica_AY141010,Squilla_empusa_L81946), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidop
8  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_gibbosa,Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura,Diamysis_spec.,Limnomysis_
9  (Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lingvura_adriatica,Leptomysis_lingvura_lingvura), (Holm
10  (Diamysis_spec.,Limnomysis_benedeni), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysido
11  (Gastrosaccus_psammodytes,Gastrosaccus_spinifer), (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_
```

Figure 4.3: Top: SAMS split spectrum of an analysis of an MSA file from Remerie *et al.* (2004) within the SAMS GUI. Bottom: The *Compatibility mode* is activated: non-compatible splits are coloured in grey.

2. The output file is directly visualised as split spectrum (see fourth tab *Split spectra*, Fig. 4.3, top). On the left hand side of the tab, different options can be chosen. The interface allows to adjust the number of printed splits and to switch between displaying the total split support or a representation which differentiates between the different support qualities. These

spectra can be scaled and stored in several graphic file formats, raster formats like JPEG, BMP and PNG as well as in scalable vector format (SVG).

3. By activating the *Compatibility mode*, splits that are not compatible turn into a grey colour (see Fig. 4.3, bottom). Either (i) all splits can be tested whether they match to the best split or (ii) every split is tested separately whether it is compatible to all earlier compatible splits, starting with the first one and resulting in a set of splits, from which a topology can be derived.

4.3.2. SAMS Analyses

Comparing both split spectra of datasets of setup 1 and setup 2, analysed with SAMS default values, shows that setup 1 delivers a clear signal whereas dataset of setup 2 contains much more noise. The dataset of setup 1 shows a few splits with high support (see Fig. 4.4a spectra scheme at the top of the split spectra, left hand side) and a high number of splits with lower supportive sites. The best splits (1-6) match the bifurcations within the topology of the underlying tree. When activating the *Compatibility mode* (Fig. 4.3, bottom), all following splits are marked as not compatible with the ones which reflect the topology. The shape of the second spectrum (see Fig. 4.4b) shows that the signal-to-noise ratio is increased, which is clearly visible since the majority of splits have similar support values. The splits which correspond to the underlying topology of dataset of setup 2 are widely spread among the split spectra. Two of the best supported splits are ranking first and second, the following ones received ranks 50, 70, 99 and 103. Nevertheless, when checking for tree compatibility, only the splits reflecting the true topology remain coloured.

4.3.3. Comparison of SAMS Support and *CQ* Weighting

The results for dataset of setup 1 show that the split which obtains the highest support is identical for the current SAMS and the new *CQ* method. The remaining splits which match the bifurcations of the true topology (= the topology on which the dataset is based), are ranked directly following the best split for the SAMS support scheme. If arranged according to the *CQ* values, these splits are placed on rank 11, 12, 14 and 17.

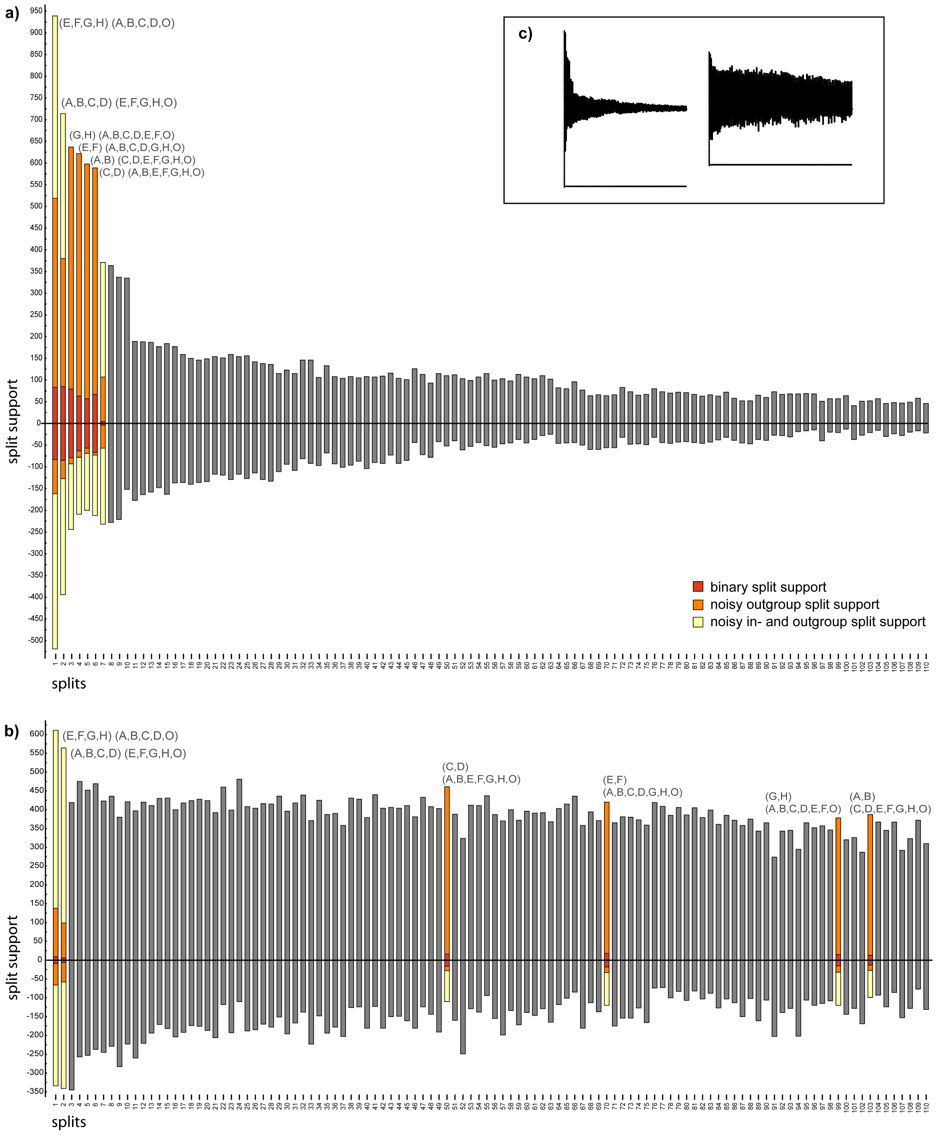


Figure 4.4: Split spectra for the datasets of setup 1 (a) and 2 (b). The coloured splits are compatible to a tree, while non-compatible splits are marked in grey colour. Within the box (c) on the top right hand side the shapes of both spectra are shown side by side.

The support values for the splits of setup 1 are listed in table 4.1 and 4.2. Both tables contain the same information but are arranged differently, they are sorted in descending order once according to SAMS (table 4.1) and according to *CQ* values (table 4.2). All splits which are compatible to the underlying topology are highlighted. The best supported splits (highlighted in green) are ranked first and second for both methods. The yellow highlighted rows mark remaining splits which fit the true topology.

Table 4.1: Split support of dataset based on setup 1. Splits are ranked according to SAMS split values. The green coloured rows mark the splits with highest support for both methods, SAMS and *CQ* weighting. The yellow coloured rows show the remaining splits which reflect the true topology.

SAMS		*CQ*			
Rank	Support	Rank	Support	g_1	g_2
1	1137	1	11281.56	(E,F,G,H)	(A,B,C,D,O)
2	804	2	11135.98	(A,B,C,D)	(E,F,G,H,O)
3	687	11	10236.18	(G,H)	(A,B,C,D,E,F,O)
4	671	12	10082.90	(E,F)	(A,B,C,D,G,H,O)
5	640	14	10028.63	(A,B)	(C,D,E,F,G,H,O)
6	623	17	10006.84	(C,D)	(A,B,E,F,G,H,O)
7	522	44	8063.75	(A,B,E,F)	(C,D,G,H,O)
			...		

Table 4.2: Split support of dataset based on setup 1. Splits are ranked according to *CQ* split values. For the colour code see table 4.1.

CQ		SAMS			
Rank	Support	Rank	Support	g_1	g_2
1	11281.56	1	1137	(E,F,G,H)	(A,B,C,D,O)
2	11135.98	2	804	(A,B,C,D)	(E,F,G,H,O)
3	10540.35	17	253	(B,O)	(A,C,D,E,F,G,H)
4	10513.03	16	272	(C,O)	(A,B,D,E,F,G,H)
5	10456.00	18	249	(D,O)	(A,B,C,E,F,G,H)
6	10395.48	20	237	(G,O)	(A,B,C,D,E,F,H)
7	10373.94	21	233	(A,O)	(B,C,D,E,F,G,H)
8	10372.44	24	224	(F,O)	(A,B,C,D,E,G,H)
9	10354.39	22	231	(H,O)	(A,B,C,D,E,F,G)
10	10325.69	19	244	(E,O)	(A,B,C,D,F,G,H)
11	10236.18	3	687	(G,H)	(A,B,C,D,E,F,O)
12	10082.90	4	671	(E,F)	(A,B,C,D,G,H,O)
13	10055.01	11	317	(C,D,E,F,G,H)	(A,B,O)
14	10028.63	5	640	(A,B)	(C,D,E,F,G,H,O)
15	10009.46	13	298	(A,B,E,F,G,H)	(C,D,O)
16	10008.25	14	289	(A,B,C,D,E,F)	(G,H,O)
17	10006.84	6	623	(C,D)	(A,B,E,F,G,H,O)
			...		

Table 4.3: Split support of dataset based on setup 2. Splits are ranked according to SAMS split values. For the colour code see table 4.1.

SAMS		*CQ*			
Rank	Support	Rank	Support	g_1	g_2
1	825	33	8567.51	(E,F,G,H)	(A,B,C,D,O)
2	783	35	8392.43	(A,B,C,D)	(E,F,G,H,O)
3	689	61	7897.53	(C,G,H,O)	(A,B,D,E,F)
			...		
49	525	2	10238.38	(D,O)	(A,B,C,E,F,G,H)
50	524	10	9358.53	(C,D)	(A,B,E,F,G,H,O)
51	524	115	7680.40	(B,F,G,O)	(A,C,D,E,H)
			...		
69	480	125	7633.70	(C,E,H,O)	(A,B,D,F,G)
70	478	9	9413.77	(E,F)	(A,B,C,D,G,H,O)
71	478	4	10055.79	(H,O)	(A,B,C,D,E,F,G)
			...		
98	443	119	7668.54	(A,C,F,O)	(B,D,E,G,H)
99	439	12	9150.65	(G,H)	(A,B,C,D,E,F,O)
100	439	118	7669.01	(B,C,G,H)	(A,D,E,F,O)
			...		
102	436	69	7862.65	(C,D,F,O)	(A,B,E,G,H)
103	435	11	9220.41	(A,B)	(C,D,E,F,G,H,O)
104	435	80	7790.70	(A,D,E,O)	(B,C,F,G,H)
			...		

Table 4.4: Split support of dataset based on setup 2. Splits are ranked according to *CQ* split values. For the colour code see table 4.1.

CQ		SAMS			
Rank	Support	Rank	Support	g_1	g_2
1	10248.22	42	529	(B,O)	(A,C,D,E,F,G,H)
2	10238.38	49	525	(D,O)	(A,B,C,E,F,G,H)
3	10222.26	57	509	(A,O)	(B,C,D,E,F,G,H)
4	10055.79	71	478	(H,O)	(A,B,C,D,E,F,G)
5	10040.54	75	474	(E,O)	(A,B,C,D,F,G,H)
6	9988.16	73	475	(C,O)	(A,B,D,E,F,G,H)
7	9947.36	89	457	(F,O)	(A,B,C,D,E,G,H)
8	9902.80	72	477	(G,O)	(A,B,C,D,E,F,H)
9	9413.77	70	478	(E,F)	(A,B,C,D,G,H,O)
10	9358.53	50	524	(C,D)	(A,B,E,F,G,H,O)
11	9220.41	103	435	(A,B)	(C,D,E,F,G,H,O)
12	9150.65	99	439	(G,H)	(A,B,C,D,E,F,O)
13	8977.20	113	412	(A,D)	(B,C,E,F,G,H,O)
14	8958.03	133	369	(B,D)	(A,C,E,F,G,H,O)
15	8949.7	119	396	(F,G)	(A,B,C,D,E,H,O)
			...		
30	8678.56	136	368	(C,H)	(A,B,D,E,F,G,O)
31	8637.37	146	351	(C,F)	(A,B,D,E,G,H,O)
32	8592.65	127	375	(C,G)	(A,B,D,E,F,H,O)
33	8567.51	1	825	(E,F,G,H)	(A,B,C,D,O)
34	8537.60	149	348	(C,E)	(A,B,D,F,G,H,O)
35	8392.43	2	783	(A,B,C,D)	(E,F,G,H,O)
36	8360.79	140	361	(A,C,D,O)	(B,E,F,G,H)
			...		

The split support values of dataset of setup 2 are listed in table 4.3 and table 4.4. The tables contain the same values, but differ in sorting by SAMS (table 4.3) or CQ split support values (table 4.4) in descending order. Splits which match the bifurcations of the true topology are highlighted. The splits which are highlighted in green are ranked first and second for the SAMS support values, while these splits are downgraded by the CQ weighting to ranks 33 and 35. The remaining splits which fit the true topology (yellow highlighted rows) are spread across the split spectra range (50,70,99 and 103) if sorted by SAMS support. While the best supported splits of the SAMS method are downgraded by the CQ method, the yellow highlighted splits rise in rank to positions 9 to 12. The upper ranks (1-8) are splits, each grouping the outgroup (O) with only one of the other sequences.

The split support for datasets of setup 3 and 4 are sorted by current SAMS (table 4.5 and 4.7) or the new CQ (table 4.6 and 4.8) weighting. Here, the crucial splits leading to long branch attraction are marked. The split highlighted in green (*L6,T10,T9*|*rest*) reflects the true topology, whereas orange highlighted rows show the support for the split which defines $L5$ and $L6$ as sister groups (Fig. 4.1c). The yellow highlighted split represents as well a wrong topology, $L5$ is grouped together with $T9$ and $T10$. For both setups and analyses, the splits which separate the taxa evolving along long branches (*T10,T9*|*rest* and *L5,L6,T10,T9*|*rest*) receive the highest ranks.

For the dataset of setup 3 with the medium short internal branch ($BL1 = 0.1$) the current SAMS scoring scheme ranks the split reflecting a bipartition of a wrong topology on the third position, whereas the split which is compatible with the true topology follows on rank 8 (table 4.5). Sorted by CQ value, the split matching the true topology is ranked much higher than the incompatible splits (table 4.6).

The SAMS support results for the dataset of setup 4 with the extremely short internal branch ($BL1 = 0.01$) are similar to those of setup 3, whereas the split which is compatible with the true topology is ranked worse among the other splits (table 4.7). Arranged in descending order by CQ value, the split which matches the true topology (*L6,T10,T9*|*rest*), highlighted in green) is ranked much higher than the incompatible split (*L5,L6*|*rest*), highlighted in orange), which is valued highly by the SAMS algorithm. However, another incompatible

Table 4.5: Split support of dataset based on setup 3. Splits are ranked according to SAMS split values. The green coloured row marks a split which reflects the true topology. The orange coloured split conflicts the green coloured one and leads to a false topology. The row which is coloured yellow shows the remaining alternative split of the two competing splits, also incompatible with the true topology. All rows highlighted in grey mark other splits which reflect the true topology as well.

SAMS		*CQ*			
Rank	Support	Rank	Support	g_1	g_2
1	2143	1	23082.29	(T10,T9)	(Out,T1,T2,T3,T4,T7,T8,L5,L6)
2	2117	2	22295.45	(Out,T1,T2,T3,T4,T7,T8)	(L5,L6,T10,T9)
3	1520	5	18300.75	(L5,L6)	(Out,T1,T2,T3,T4,T7,T8,T10,T9)
4	533	45	11925.15	(Out,T1)	(T2,T3,T4,T7,T8,L5,L6,T10,T9)
5	506	17	14644.15	(T1,T2,T3,T4,T7,T8,L5,L6)	(Out,T10,T9)
6	482	33	13532.83	(Out,T1,T2,T3)	(T4,T7,T8,L5,L6,T10,T9)
7	465	59	11236.57	(T7,T8)	(Out,T1,T2,T3,T4,L5,L6,T10,T9)
8	432	3	19527.83	(Out,T1,T2,T3,T4,T7,T8,L5)	(L6,T10,T9)
9	422	16	14777.71	(T8,T10,T9)	(Out,T1,T2,T3,T4,T7,L5,L6)
10	401	39	12580.40	(Out,T1,T2)	(T3,T4,T7,T8,L5,L6,T10,T9)
11	399	4	19325.89	(Out,T1,T2,T3,T4,T7,T8,L6)	(L5,T10,T9)
12	392	36	12842.50	(T8,L5,L6)	(Out,T1,T2,T3,T4,T7,T10,T9)
				...	

Table 4.6: Split support of dataset based on setup 3. Splits are ranked according to *CQ* split values. For the colour code see table 4.5.

CQ		SAMS			
Rank	Support	Rank	Support	g_1	g_2
1	23082.29	1	2143	(T10,T9)	(Out,T1,T2,T3,T4,T7,T8,L5,L6)
2	22295.45	2	2117	(Out,T1,T2,T3,T4,T7,T8)	(L5,L6,T10,T9)
3	19527,83	8	432	(Out,T1,T2,T3,T4,T7,T8,L5)	(L6,T10,T9)
4	19325.89	11	399	(Out,T1,T2,T3,T4,T7,T8,L6)	(L5,T10,T9)
5	18300.75	3	1520	(L5,L6)	(Out,T1,T2,T3,T4,T7,T8,T10,T9)
6	17488.81	15	376	(Out,T1,T2,T3,T4,T7)	(T8,L5,L6,T10,T9)
				...	

Table 4.7: Split support of dataset based on setup 4. Splits are ranked according to SAMS split values. For the colour code see table 4.5.

SAMS		*CQ*			
Rank	Support	Rank	Support	g_1	g_2
1	2172	1	23030.98	(T10,T9)	(Out,T1,T2,T3,T4,T7,T8,L5,L6)
2	2086	2	22457.18	(Out,T1,T2,T3,T4,T7,T8)	(L5,L6,T10,T9)
3	1662	5	18808.03	(L5,L6)	(Out,T1,T2,T3,T4,T7,T8,T10,T9)
4	539	48	12005.58	(Out,T1)	(T2,T3,T4,T7,T8,L5,L6,T10,T9)
				...	
17	328	9	16951.22	(T1,T2,T3,T4,T7,T8)	(Out,L5,L6,T10,T9)
18	320	3	19206.37	(Out,T1,T2,T3,T4,T7,T8,L6)	(L5,T10,T9)
19	314	4	19159.13	(Out,T1,T2,T3,T4,T7,T8,L5)	(L6,T10,T9)
20	307	61	11332.11	(T7,T8,L5,L6)	(Out,T1,T2,T3,T4,T10,T9)
				...	

Table 4.8: Split support of dataset based on setup 4. Splits are ranked according to *CQ* split values. For the colour code see table 4.5.

CQ		SAMS			
Rank	Support	Rank	Support	g_1	g_2
1	23030.98	1	2172	(T10,T9)	(Out,T1,T2,T3,T4,T7,T8,L5,L6)
2	22457.18	2	2086	(Out,T1,T2,T3,T4,T7,T8)	(L5,L6,T10,T9)
3	19206.37	18	320	(Out,T1,T2,T3,T4,T7,T8,L6)	(L5,T10,T9)
4	19159.13	19	314	(Out,T1,T2,T3,T4,T7,T8,L5)	(L6,T10,T9)
5	18808.03	3	1662	(L5,L6)	(Out,T1,T2,T3,T4,T7,T8,T10,T9)
6	17621.47	13	377	(Out,T1,T2,T3,T4,T7)	(T8,L5,L6,T10,T9)
7	17116.48	24	285	(Out,T1,T2,T3,T4,T7,T8,T10)	(L5,L6,T9)
				...	

split (yellow highlighted) is ranked higher than the split which is compatible with the true topology (table 4.8).

4.4. Discussion

With this work the split search tool SAMS is now available with a platform independent and user friendly interface. All improvements offer a much more convenient tool to perform and visualise a split analysis. By displaying the split spectra within the GUI, the user can easily discriminate between datasets containing clear or conflicting signal. The shapes of the split spectra visualise, if strong splits are dominating the dataset. The presence of noise becomes apparent by many splits with similar support values as shown in Fig. 4.4c). While the first setup (Fig. 4.4a) delivers a clear signal, the split support values of the second setup (Fig. 4.4b), however, shows an ambiguous split spectrum.

For the tested setups the algorithm works well. The compatible split set of setup 1 (clear signal, Fig. 4.4a) contains only splits, which match the bifurcations of the true tree. While these splits are all ranked best, the compatible split set for the second setup (Fig. 4.4b), which was designed to contain more noise than the first one, is widely spread within the split distribution. Nonetheless, the compatible split set matches as well the true topology.

The comparison of the current SAMS and the new *CQ* split weighting for the first setup (Fig. 4.1a) shows that the SAMS split weighting clearly outperforms the new *CQ* values. Splits matching the true topology are all placed on ranks 1 to 6, whereas sorted according to *CQ* values, splits matching the true topology are spread within ranks 1 up to 17. Though, since the two best supported splits (*T9,T10|rest* and *L5,L6,T9,T10|rest*), are identical for both methods, the splits which do not fit the true topology are rejected as incompatible.

The test for compatible splits delivers a set of splits which match the topology of the dataset of the second setup (Fig. 4.1b) ordered by SAMS support. However, this does not work for a *CQ* ranking order. Even though, the splits which are compatible with the true topology are spread within range 1 up to 103 according to the SAMS ranking and 1 up to 35 for the *CQ* weighting. The ranks 1 to 32 indicate that grouping of one sequence with the outgroup and splits with subsets containing only two sequences in general are ranked higher. This might be caused by similarities that occur for only two sequences, which match by chance and in case all other sequences have different nucleotides. This is very likely the case, since the outgroup as well as the terminal sequences have long branches. Subsequently, the C values are falsely scored very high.

For dataset of setup 3 (Fig. 4.1c) the new weighting scheme clearly outperforms the current SAMS support scheme. Both splits which separate the taxa evolving along long branches received the highest ranks for both weighting methods. The split containing the subset ($L5$,$L6$) falsely groups the two sequences together, which are separated only by a short intermediate branch between two long branches and is ranked third by the SAMS support. In contrast, the new CQ support calculation weights the true split ($L6, T9, T10|rest$) much higher and therefore leads to a split set containing the true rather than the false splits ($L5,L6|rest$ and $L5, T9, T10|rest$). In case the intermediate branch between the two long branches is extremely short ($BL1 = 0.01$, Fig. 4.1c), both methods fail to recover the true split. While the standard SAMS algorithm again overrates the split $L5,L6|rest$, by grouping the terminal taxa with long branches together, the CQ scoring can handle this artefact. However, the new weighting method is not able to distinguish between the grouping of $L6, T9, T10|rest$ and $L5, T9, T10|rest$. This is probably due to the lack of signal within the sequence showing the intermediate short branch.

The results illustrate, that the new weighting scheme performs very well in most cases. But the results are not always satisfying. The CQ values, especially the C value, is sensitive to the size of the bipartitions, as it is much more likely, that a smaller subset can have similarities by chance and therefore falsely induce a strong contrast. Therefore, the weighting criteria should be adjusted to take the size of the subsets into account.

Without using evolutionary models it is not possible to account for multiple substitutions. A comparison of pairs of sequences, not only for one but for a few sites, could partly counterbalance these effects. This could be realised with a sliding window approach, which would allow to compare not only one site, but several positions of two or more sequences. Similar to the row testing algorithm of SAMS, sequence similarities could be taken into account to detect outliers, supporting 'wrong' splits. Additionally, identifying and removing columns of random similarity within MSAs with tools like ALISCORE (Misof & Misof, 2009) could remove sites without usable phylogenetic information and therefore considerably improve the split analysis.

Moreover grouping the compatible splits to sets and comparing their total support could possibly allow for a statement useful for a phylogenetic reconstruction.

In conclusion, the development of a heuristic and easy-to-use split analysis that is not only based on RY splits, but takes all nucleotides into account, would offer a reasonable and hopefully enlightening tool to analyse datasets from a different perspective.

Überhaupt ist es für den Forscher ein guter Morgensport, täglich vor dem Frühstück eine Lieblingshypothese einzustampfen - das erhält jung.

Konrad Lorenz

5. General Discussion and Future Prospects

In most analyses, model adequacy is only addressed if phylogenetic trees do not meet the expectations. In the best case, a relative model goodness-of-fit test is performed to identify the best model of a set of available (preselected) models for the chosen reconstruction method. However, it has been shown, that models which have been identified as best fitting are not necessarily adequate for the dataset (Gatesy, 2007), especially, if the parameter assumptions of the model are violated (Felsenstein, 1978; Huelsenbeck & Hillis, 1993; Yang *et al.*, 1994; Swofford *et al.*, 2001; Ho & Jermiin, 2004; Jermiin *et al.*, 2004).

The Goldman-Cox test has been suggested to identify model model adequacy. In chapter 2 simulated data was used to investigate the performance of this test. Different nucleotide MSAs with heterogeneous base composition and non-stationary substitution processes were simulated to examine a mixture of models and its impact on the results of model adequacy estimation. For all sets of topologies and branch lengths the GTR model has been used, combined with a gamma distribution model to assess rate heterogeneity across sites and either a proportion of invariable sites ($p_{inv}> 0$) or $p_{inv}= 0$.

For the composition with a proportion of sites fixed to be invariable ($p_{inv}> 0$), the gamma distribution determines the rates for the remaining variable sites. Either the continuous or the four rate category discrete gamma distribution model was used to investigate whether the use of an adequate number of categories has a significant influence on model estimation and the results of model adequacy tests.

Each topology was once simulated to evolve (i) stationary (according to one model composition), (ii) with an increasing number of clades and (iii) with some clades evolving according to different conditions of substitution rates, base frequencies, invariant sites and rate heterogeneity. All datasets were analysed with ML using different implementations, different number of rate categories (4, 12 or 25) for the discrete gamma distribution and a number of different parameters (each run using mean or median as type of average for modelling gamma distribution; GTR+Γ or GTR+Γ+I).

Parametric bootstraps for each simulated dataset were generated, based on the respective ML estimates. The obtained pattern distributions were reviewed using the Goldman-Cox test. To ensure that the Goldman-Cox test accepts the true model parameters, every simulated dataset and 100 reference bootstraps based on identical model settings were evaluated. To exclude possible sources of error and to ensure that the obtained results are reliable, a number of different software implementations for sequence simulation and ML analysis were tested and various parameter settings were selected. All variations had no influence on the outcome.

The Goldman-Cox test can deliver false positive results. The predictions of the test do not stringently correlate with the correctness of tree inference in case of non-stationary datasets which were analysed with a model assuming stationarity. 'Mimicking' behaviour, a pattern distribution from a mixture model which is identical to one generated from a different tree or trees (e.g., in case of different gene trees) can be a possible cause. This implies that without knowing that a dataset is stationary beforehand, the Goldman-Cox test does not deliver fully reliable results.

The Goldman-Cox test itself is very conservative and sensitive to small model deviations such as approximating the gamma distribution by using only four categories, if it can be assumed, that rate heterogeneity is likely to follow a continuous distribution. For datasets which did not violate the assumptions of the model, the results show that the closer the number of chosen gamma rate categories for a discrete distribution is to the true number, the more accurate the results of the Goldman-Cox test. For datasets which violated the required conditions of the chosen models, the model which is estimated during ML analysis is often accepted by the Goldman-Cox test, although the resulting topology of the ML tree is inaccurate.

While this shows that already assessing reliably model adequacy is not trivial, it can just as little address the complexity of the causes for misspecification. Even worse, using the Goldman-Cox test can lead to type II errors when returning false positive results. Furthermore, present absolute model goodness-of-fit tests, such as the Goldman-Cox test, deliver only a binary answer - either the model is accepted as adequate or it is rejected as not adequate - but they do not provide implications on what to do, if the model is rejected. Therefore, it is of outstanding importance to develop new methods which are able to provide insight into the misspecification and hints how to proceed with the data.

Numerous molecular phylogenies published during the past decade are contradictory, even when large datasets where used and after applying elaborate analyses. Undetected systematic errors and unsolved problems of data quality evaluation and adequate substitution model selecting still persist. Therefore, the second study (chapter 3) focuses on a new method to gain insight into possible reasons or avoidance of model misspecification. Split decomposition, the study of split support and its visualization within splits graphs provide a valuable tool for gaining an overview of possible patterns and contradictory signal or noise within datasets. Instead of analysing all site patterns of the datasets, the amount of possible splits (2^{n-1} possible splits instead of 4^n possible site patterns for n taxa) is a reasonable recoding of phylogenetic information in order to increase clarity and reduce computational effort. The method of split analysis was therefore chosen and combined with the method of parametric bootstrapping to perform split residual analyses for further analyses on how models may be misspecified.

Simulated and empirical nucleotide MSAs were analysed with an ML implementation. The estimated evolutionary models including the tree, the substitution parameters and the rate heterogeneity parameters were used for simulating parametric bootstrap datasets. Both, the original datasets and the bootstraps were recoded to RY code and the split spectra of the original datasets and the spectra of their corresponding bootstraps were compared. Splits occurring more or less often in the analysed dataset as in every corresponding bootstrap dataset were classified as over- and underrepresented, respectively, and visualised in phylogenetic split networks. Simulated datasets were used to compare the estimated model parameters to the known statistical background. While the empirical datasets were analysed to understand, how the findings can be applied to assess model adequacy in phylogenetic analysis.

The analysis of model adequacy is influenced by a number of things. First, the chosen models seem to fit best to the largest subset of sequences which evolve according to the same evolutionary parameters. Datasets which evolved under globally SRH conditions (Jayaswal *et al.*, 2005) mostly produced only a few splits which deviated in their amount from the parametric bootstrap datasets, representing the statistical model. For datasets which are based on compositions including one or more clades of sequences evolving according to different model parameters as the rest of the tree, the amount of over- and underrepresented splits (split residuals) is increasing. It is noticeable, that mostly trivial splits are overrepresented for those sequences, which belong to the smaller subgroup sharing the same model conditions.

Using incorrect a priori assumptions about the dataset, i.e. discrete gamma distribution for analysis of datasets which can be expected to follow a continuous gamma distribution, can also have an influence on model adequacy. For the datasets which were simulated with a discrete gamma distribution with four rates only a small number of split residuals were occurring, for stationary as well as for non-stationary compositions.

Moreover, taxon sampling can have a strong influence on the assessment of model adequacy. The use of only one model for datasets containing sequences of a wide range of organisms may balance differences and therefore leaves out potential phylogenetic signal. The split networks of the empirical datasets show, that overrepresented trivial splits are present here as well. Moreover, it seems as if splits which divide Fungi from Metazoa in the dataset of Rokas *et al.* (2005) are not covered well by the estimated model, even though it can be expected that these evolved unequally.

Although split residual diagnostics as applied in this study seems to have the power to make headway in better understanding model adequacy, using RY splits does not completely exploit all possible sources of misspecification. This recoding may balance effects, such as base composition bias (systematic error) and bias caused by heterotachy (Sims *et al.*, 2009; Ho *et al.*, 2006), which should also be taken into account. The third study focuses on establishing a new formal split weighting scheme to re-evaluate splits, which were detected by the split search tool SAMS. In order to test this new weighting scheme, SAMS was extended and is now available with a platform independent and user friendly interface to provide an overview of all parameters which are easily adjustable by text boxes, value sliders or radio buttons. The results can be visualised within a split spectrum and stored in several graphic file formats. By using the *Compatibility mode* the splits are evaluated if they are compatible with the best split or all splits fitting together as tree. A new split weighting scheme, which formalises aspects like 'contrast of character states' and 'character state homogeneity' within split subsets was introduced and tested with simulated MSAs.

All of these improvements offer a much more capable tool to perform a split analysis and visualise the results. Moreover, the split analysis itself offers a valuable tool to analyse datasets from a different perspective. The shape of the visualised split spectra can demonstrate, if the dataset delivers a clear split signal or if there is a lot of noise present. In the latter case, the spectrum consists of splits with almost similar support. Whereas a clear signal shows a few splits with strong support, while the other split support values follow a strongly declining trend. Overall, the new split weighting scheme performs quite well, but should be adjusted to take the size of the subsets into account. For the datasets which combined a long branch with a medium short intermediate branch, the new scheme outperformed the old split weighting. For the dataset with an extreme short intermediate branch, both schemes failed to detect the split which corresponds to the true topology.

Both presented split weighting methods and as well a Hadamard approach should be the starting point for a new study in which datasets and corresponding bootstraps are screened for split residuals. This should yield deeper insights on model misspecification. Additionally, trivial splits could be excluded to gain a better view on splits of phylogenetic interest. Furthermore, one could apply alignment masking methods on the simulated datasets before the analysis and to design simulation setups with different partitions.

In summary, phylogenetic tree reconstruction should not be considered as a 'black box'. There are already methods available which can be applied at various steps during the analysis. Especially recent advances in the field of genomics produce empirical datasets which are by no means simplified homogeneous data sets of perfect quality without any misleading evidence of relationships or saturation. The analysis of MSAs with the Goldman-Cox test delivers plausible results for datasets, which match all conditions assumed by the chosen evolutionary model. The model, however, must not be limited to undervalued fixed values, because this can lead to misspecifications. Unfortunately, there is no method available to test how many categories should be modelled for approximating the gamma distribution. Ignoring this factor or relying on a traditionally used value can lead to wrong results. For simulated data with rates following a continuous gamma distribution, however, the results seem to be reliable if at least 12 discrete gamma categories are used. This behaviour should be studied in more detail in future analyses. Most likely, the categorization of the gamma distribution should by itself be subject to a maximum likelihood approach. This is currently not implemented in any of the available software packages.

Furthermore, if the SRH assumptions are known to be incompatible with the data, models should be considered which do not require these assumptions (Jayaswal *et al.*, 2014). If these can be used as well for sequence simulation, the Goldman-Cox test could perform much better for data which do not meet the SRH condition.

The analyses of non-stationary datasets show that the application of globally fitted substitution models which do not account for taxon or clade specific patterns increases the risk of biased tree reconstructions and misleading artefacts. The identification of these artefacts in empirical data is almost impossible.

As a consequence, the difference between local taxon or clade specific variation of the substitution parameters and a globally fitted model should be studied. A possible follow up action could be a conceptional shift from simply fitting a single substitution model to the analyses of model adequacy and the subsequent application of submodels to reduce potential biases in phylogenetic tree inference.

References

Abdo, Zaid; Minin, Vladimir N; Joyce, Paul & Sullivan, Jack. 2005. Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. *Molecular biology and evolution*, **22**(3), 691–703.

Aichele, D. & Schwegleb, H.-W. 2008. Die Taxonomie der Gattung Pulsatilla. *Repertorium novarum specierum regni vegetabilis*, **60**(1-3), 1–230.

Akaike, H. 1974. A new look at the statistical model identification. *Ieee transactions on automatic control*, **19**(6), 716–723.

Alfaro, Michael E; Zoller, Stefan & Lutzoni, François. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Molecular biology and evolution*, **20**(2), 255–66.

Bandelt, H J; Forster, P; Sykes, B C & Richards, M B. 1995. Mitochondrial portraits of human populations using median networks. *Genetics*, **141**(2), 743–53.

Bandelt, Hans-Jürgen & Dress, Andreas W M. 1992. Split decomposition: A new and useful approach to phylogenetic analysis of distance data. *Molecular phylogenetics and evolution*, **1**(3), 242–252.

Bollback, Jonathan P. 2002. Bayesian model adequacy and choice in phylogenetics. *Molecular biology and evolution*, **19**(7), 1171–80.

Brown, Jeremy M & Lemmon, Alan R. 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic biology*, **56**(4), 643–55.

Bryant, David & Moulton, Vincent. 2004. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Molecular biology and evolution*, **21**(2), 255–65.

Cain, A J & Harrison, G A. 1960. PHYLETIC WEIGHTING. *Proceedings of the zoological society of london*, **135**(1), 1–31.

Chang, J T. 1996. Full reconstruction of Markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, **137**(1), 51–73.

Charleston, Michael. 1998. Spectrum: spectral analysis of phylogenetic data. *Bioinformatics (oxford, england)*, **14**(1), 98–9.

Cunningham, Clifford W; Zhu, H & Hillis, D. 1998. Best-fit maximum-likelihood models for phylogenetic inference: empirical tests with known phylogenies. *Evolution*, **52**(4), 978–987.

Dress, Andreas W M; Flamm, Christoph; Fritzsch, Guido; Grünewald, Stefan; Kruspe, Matthias; Prohaska, Sonja J & Stadler, Peter F. 2008. Noisy: identification of problematic columns in multiple sequence alignments. *Algorithms for molecular biology : Amb*, **3**(Jan.), 7.

Dunn, Casey W; Hejnol, Andreas; Matus, David Q; Pang, Kevin; Browne, William E; Smith, Stephen A; Seaver, Elaine; Rouse, Greg W; Obst, Matthias; Edgecombe, Gregory D; Sø rensen, Martin V; Haddock, Steven H D; Schmidt-Rhaesa, Andreas; Okusu, Akiko; Kristensen, Reinhardt Mø bjerg; Wheeler, Ward C; Martindale, Mark Q & Giribet, Gonzalo. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**(7188), 745–9.

Edwards, A W F & Cavalli-Sforza, L L. 1963. The reconstruction of evolution. *Annals of human genetics*, **27**, 105–106.

Edwards, A W F & Cavalli-Sforza, L L. 1964. Reconstruction of evolutionary trees. *Pages 67–76 of:* Heywood, V H & McNeill, J (eds), *Phenetic and phylogenetic classification.* Systematics Association Publ. No. 6, London.

Efron, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *The annals of statistics*, **7**(1), 1–26.

Efron, B & Tibshirani, R. 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical science*, **1**(1), pp. 54–75.

Efron, B; Halloran, E & Holmes, S. 1996. Bootstrap confidence levels for phylogenetic trees. *Proceedings of the national academy of sciences of the united states of america*, **93**(23), 13429–34.

Erixon, Per; Svennblad, Bodil; Britton, Tom & Oxelman, Bengt. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Systematic biology*, **52**(5), 665–73.

Felsenstein, Joseph. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Systematic zoology*, **27**(4), 401–410.

Felsenstein, Joseph. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of molecular evolution*, **17**(6), 368–76.

Felsenstein, Joseph. 1985. Confidence Limits on Phylogenies: An Approach Using the Bootstrap. *Evolution*, **39**(4), 783.

Felsenstein, Joseph. 2004. *Inferring Phylogenies.* Sinauer Associates, Inc., Sunderland, Mass.

Fisher, R. A. 1922. On the Mathematical Foundations of Theoretical Statistics. *Philosophical transactions of the royal society a: Mathematical, physical and engineering sciences*, **222**(594-604), 309–368.

Fisher, Ronald Aylmer Sir. 1958. *Statistical Methods for Research Workers.* 13 edn. Hafner, New York.

Fitch, W M & 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. *Biochemical genetics*, **4**(5), 579–93.

Fletcher, William & Yang, Ziheng. 2009. INDELible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, **26**(8), 1879–88.

Flook, P K & Rowell, C H. 1997. The effectiveness of mitochondrial rRNA gene sequences for the reconstruction of the phylogeny of an insect order (Orthoptera). *Molecular phylogenetics and evolution*, **8**(2), 177–92.

Gatesy, John. 2007. A tenth crucial question regarding model use in phylogenetics. *Trends in ecology & evolution*, **22**(10), 509–10.

Goldman, Nick. 1993b. Statistical tests of models of DNA substitution. *Journal of molecular evolution*, **36**(2), 182–198.

Goremykin, Vadim; Holland, Barbara R; Hirsch-Ernst, Karen I & Hellwig, Frank H. 2005. Analysis of Acorus calamus chloroplast genome and its phylogenetic implications. *Molecular biology and evolution*, **22**(9), 1813–22.

Guindon, Stéphane & Gascuel, Olivier. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, **52**(5), 696–704.

Guindon, Stéphane; Dufayard, Jean-François; Lefort, Vincent; Anisimova, Maria; Hordijk, Wim & Gascuel, Olivier. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic biology*, **59**(3), 307–21.

Hasegawa, M; Kishino, H & Yano, T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of molecular evolution*, **22**(2), 160–74.

Hendy, Michael D & Penny, David. 1989. A framework for the quantitative study of evolutionary trees. *Systematic zoology*, **38**(4), 297–309.

Hendy, Michael D & Penny, David. 1993. Spectral analysis of phylogenetic data. *Journal of classification*, **10**(1), 5–24.

Hendy, Michael D; Penny, David & Steel, Mike A. 1994. A discrete Fourier analysis for evolutionary trees. *Proceedings of the national academy of sciences of the united states of america*, **91**(8), 3339–43.

Hennig, Willi. 1950. Grundzüge einer Theorie der Phylogenetischen Systematik.

Hennig, Willi. 1966. Phylogenetic Systematics. *University of illinois press, urbana.*

Hillis, David M. & Bull, James. J. 1993. An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis. *Systematic biology*, **42**(2), 182–192.

Ho, Joshua W K; Adams, Cameron E; Lew, Jie Bin; Matthews, Timothy J; Ng, Chiu Chin; Shahabi-Sirjani, Arash; Tan, Leng Hong; Zhao, Yu; Easteal, Simon; Wilson, Susan R & Jermiin, Lars S. 2006. SeqVis: visualization of compositional heterogeneity in large alignments of nucleotides. *Bioinformatics (oxford, england)*, **22**(17), 2162–3.

Ho, Simon Y & Jermiin, Lars. 2004. Tracing the decay of the historical signal in biological sequence data. *Systematic biology*, **53**(4), 623–37.

Holland, Barbara R; Spencer, Hamish G; Worthy, Trevor H & Kennedy, Martyn. 2010. Identifying cliques of convergent characters: concerted evolution in the cormorants and shags. *Systematic biology*, **59**(4), 433–45.

Huber, Katharina T; Langton, Michael; Penny, David; Moulton, Vincent & Hendy, Michael D. 2002. Spectronet: a package for computing spectra and median networks. *Applied bioinformatics*, **1**(3), 159–61.

Huelsenbeck, J P & Ronquist, F. 2001. MrBayes: Bayesian inference of phylogenetic trees. *Bioinformatics (oxford, england)*, **17**(8), 754–5.

Huelsenbeck, John P & Hillis, David M. 1993. Success of Phylogenetic Methods in the Four-Taxon Case. *Systematic biology*, **42**(3), 247–264.

Huelsenbeck, John P & Rannala, Bruce. 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Systematic biology*, **53**(6), 904–913.

Huson, Daniel H & Bryant, David. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, **23**(2), 254–67.

Jayaswal, Vivek; Jermiin, Lars S & Robinson, John. 2005. Estimation of phylogeny using a general Markov model. *Evolutionary bioinformatics online*, **1**(Jan.), 62–80.

Jayaswal, Vivek; Wong, Thomas K F; Robinson, John; Poladian, Leon & Jermiin, Lars S. 2014. Mixture models of nucleotide sequence evolution that account for heterogeneity in the substitution process across sites and across lineages. *Syst biol*, **63**(5), 726–742.

Jermiin, Lars; Ho, Simon Y; Ababneh, Faisal; Robinson, John & Larkum, Anthony W. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Systematic biology*, **53**(4), 638–43.

Jia, Fangzhi; Lo, Nathan & Ho, Simon Y W. 2014. The impact of modelling rate heterogeneity among sites on phylogenetic estimates of intraspecific evolutionary rates and timescales. *Plos one*, **9**(5), e95722.

Jukes, Thomas H & Cantor, C R. 1969. Evolution of protein molecules. *Pages 21–132 of:* Munro, HN (ed), *Mammalian protein metabolism.* New York: Academic Press.

Kück, Patrick; Mayer, Christoph; Wägele, Johann Wolfgang & Misof, Bernhard. 2012. Long Branch Effects Distort Maximum Likelihood Phylogenies in Simulations Despite Selection of the Correct Model. *Plos one*, **7**(5), e36593.

Keane, Thomas M; Creevey, Christopher J; Pentony, Melissa M; Naughton, Thomas J & Mclnerney, James O. 2006. Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *Bmc evolutionary biology*, **6**(Jan.), 29.

Kelchner, Scot A & Thomas, Michael A. 2007. Model use in phylogenetics: nine key questions. *Trends in ecology & evolution*, **22**(2), 87–94.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, **16**(2), 111–20.

Kolaczkowski, Bryan & Thornton, Joseph W. 2008. A mixed branch length model of heterotachy improves phylogenetic accuracy. *Molecular biology and evolution*, **25**(6), 1054–66.

Lemmon, Alan R & Moriarty, Emily C. 2004. The importance of proper model assumption in bayesian phylogenetics. *Systematic biology*, **53**(2), 265–77.

Lento, G M; Hickson, R E; Chambers, G K & Penny, D. 1995. Use of spectral analysis to test hypotheses on the origin of pinnipeds. *Molecular biology and evolution*, **12**(1), 28–52.

LOPEZ, P; CASANE, D & PHILIPPE, H. 2002. Heterotachy, an important process of protein evolution. *Molecular biology and evolution*, **19**(1), 1–7.

LÖYTYNOJA, ARI & GOLDMAN, NICK. 2005. An algorithm for progressive multiple alignment of sequences with insertions. *Proceedings of the national academy of sciences of the united states of america*, **102**(30), 10557–62.

MADDISON, D R; SWOFFORD, D L & MADDISON, W P. 1997. Nexus: An Extensible File Format for Systematic Information. *Systematic biology*, **46**(4), 590–621.

MATSEN, FREDERICK A & STEEL, MIKE A. 2007. Phylogenetic mixtures on a single tree can mimic a tree of another topology. *Systematic biology*, **56**(5), 767–75.

MATSEN, FREDERICK A; MOSSEL, ELCHANAN & STEEL, MIKE. 2008. Mixed-up trees: the structure of phylogenetic mixtures. *Bulletin of mathematical biology*, **70**(4), 1115–39.

MAYER, CHRISTOPH. 2010. *MultimoSeqSim*. `https://www.zfmk.de/en/research/research-centres-and-groups/multimoseqsim`.

MAYER, CHRISTOPH. available from the author upon request. *'multiphylip-Cox-Goldman.pl'*.

MAYER, CHRISTOPH & WÄGELE, WOLFGANG. 2005. *SAMS (Splits analysis methods) Version 1.4 beta*.

MEID, SANDRA A; MAYER, CHRISTOPH & WÄGELE, WOLFGANG. 2012. *SAMS GUI*. `https://www.zfmk.de/en/research/research-centres-and-groups/sams`.

MISOF, BERNHARD & MISOF, KATHARINA. 2009. A monte carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion. *Systematic biology*, **58**(1), 21–34.

NEYMAN, JERZY. 1971. Molecular studies of evolution: a source of novel statistical problems. *Pages 1–27 of:* GUPTA, S S & YACKEL, J (eds), *Statistical decision theory and related topics*. New York: Academic Press.

NGUYEN, MINH ANH THI; KLAERE, STEFFEN & VON HAESELER, ARNDT. 2010. MISFITS: Evaluating the goodness of fit between a phylogenetic model and an alignment. *Molecular biology and evolution*, July.

Nokia Corporation; Nord, Haavard; Chambe-Eng, Eirik; Trolltech; Digia & Qt project. 2011. *Qt*. `http://www.qt-project.org/`.

Nylander, Johan A. A. 2004. *MrModeltest v2*.

Ogden, T Heath & Rosenberg, Michael S. 2006a. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic biology*, **55**(2), 314–28.

Ogden, T Heath & Rosenberg, Michael S. 2006b. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic biology*, **55**(2), 314–28.

Penny, David; McComish, Benett J; Charleston, Michael A & Hendy, Michael D. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *Journal of molecular evolution*, **53**(6), 711–23.

Philip, Gayle K; Creevey, Christopher J & McInerney, James O. 2005. The Opisthokonta and the Ecdysozoa may not be clades: stronger support for the grouping of plant and animal than for animal and fungi and stronger support for the Coelomata than Ecdysozoa. *Molecular biology and evolution*, **22**(5), 1175–84.

Philippe, Hervé; Lartillot, Nicolas & Brinkmann, Henner. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Molecular biology and evolution*, **22**(5), 1246–53.

Phillips, Matthew J; Delsuc, Frédéric & Penny, David. 2004. Genome-scale phylogeny and the detection of systematic biases. *Molecular biology and evolution*, **21**(7), 1455–8.

Pick, K S; Philippe, H; Schreiber, F; Erpenbeck, D; Jackson, D J; Wrede, P; Wiens, M; Alié, A; Morgenstern, B; Manuel, M & Wörheide, G. 2010. Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships. *Molecular biology and evolution*, **27**(9), 1983–7.

Posada, David & Buckley, Thomas R. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic biology*, **53**(5), 793–808.

Posada, David & Crandall, Keith A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics*, **14**(9), 817–818.

Puigbò, Pere; Garcia-Vallvé, Santiago & McInerney, James O. 2007. TOPD/FMTS: a new software to compare phylogenetic trees. *Bioinformatics (oxford, england)*, **23**(12), 1556–8.

Rambaut, Andrew & Grassly, N C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : Cabios*, **13**(3), 235–8.

Rannala, B & Yang, Z. 1996. Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *Journal of molecular evolution*, **43**(3), 304–11.

Remerie, Thomas; Bulckaen, Bart; Calderon, J; Deprez, Tim; Mees, Jan; Vanfleteren, Jacques; Vanreusel, Ann; Vierstraete, Andy; Vincx, Magda; Wittmann, KJ & Wooldridge, T. 2004. Phylogenetic relationships within the Mysidae (Crustacea, Peracarida, Mysida) based on nuclear 18S ribosomal RNA sequences. *Molecular phylogenetics and evolution*, **32**(3), 770–7.

Ripplinger, Jennifer & Sullivan, Jack. 2008. Does choice in model selection affect maximum likelihood analysis? *Systematic biology*, **57**(1), 76–85.

Rokas, Antonis & Carroll, Sean B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. *Molecular biology and evolution*, **22**(5), 1337–44.

Rokas, Antonis; Williams, Barry L; King, Nicole & Carroll, Sean B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, **425**(6960), 798–804.

Rokas, Antonis; Krüger, Dirk & Carroll, Sean B. 2005. Animal evolution and the molecular signature of radiations compressed in time. *Science (new york, n.y.)*, **310**(5756), 1933–8.

Schierwater, Bernd; Eitel, Michael; Jakob, Wolfgang; Osigus, Hans-Jürgen; Hadrys, Heike; Dellaporta, Stephen L; Kolokotronis, Sergios-Orestis & Desalle, Rob. 2009. Concatenated analysis sheds light on early metazoan evolution and fuels a modern 'urmetazoon' hypothesis. *Plos biology*, **7**(1), e20.

Schwarz, Gideon. 1978. Estimating the Dimension of a Model. *The annals of statistics*, **6**(2), 461–464.

Shavit Grievink, Liat; Penny, David; Hendy, Michael D & Holland, Barbara R. 2010. Phylogenetic tree reconstruction accuracy and model fit when proportions of variable sites change across the tree. *Systematic biology*, **59**(3), 288–97.

Sims, Gregory E; Jun, Se-Ran; Wu, Guohong Albert & Kim, Sung-Hou. 2009. Whole-genome phylogeny of mammals: evolutionary information in genic and nongenic regions. *Proceedings of the national academy of sciences of the united states of america*, **106**(40), 17077–82.

Sokal, Robert R. 1966. *Numerical taxonomy.* W. H. Freeman, San Francisco.

Sokal, Robert R & Sneath, Peter H. 1963. *Principles of numerical taxonomy.* W. H. Freeman, San Francisco.

Squartini, Federico & Arndt, Peter F. 2008. Quantifying the stationarity and time reversibility of the nucleotide substitution process. *Molecular biology and evolution*, **25**(12), 2525–35.

Stamatakis, Alexandros. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (oxford, england)*, **22**(21), 2688–90.

Steel, Mike A. 2002. Some statistical aspects of the maximum parsimony method. *Exs*, 125–139.

Steel, Mike A & Penny, David. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. *Molecular biology and evolution*, **17**(6), 839–50.

Steel, Mike A; Székely, Laszlo; Erdös, Peter L & Waddell, Peter J. 1993. A complete family of phylogenetic invariants for any number of taxa under Kimura's 3ST model. *New zealand journal of botany*, **31**, 289–296.

Sullivan, Jack & Joyce, Paul. 2005. Model selection in Phylogenetics. *Annual review of ecology, evolution, and systematics*, **36**(1), 445–466.

Sullivan, Jack & Swofford, David 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? *Systematic biology*, **50**(5), 723–9.

Sumner, Jeremy G; Jarvis, Peter D; Fernández-Sánchez, Jesús; Kaine, Bodie T; Woodhams, Michael D & Holland, Barbara R. 2012. Is the general time-reversible model bad for molecular phylogenetics? *Systematic biology*, **61**(6), 1069–74.

Susko, Edward; Spencer, Mathew & Roger, Andrew J. 2005. Biases in phylogenetic estimation can be caused by random sequence segments. *Journal of molecular evolution*, **61**(3), 351–9.

Swofford, David L. 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. *Sinauer associates, sunderland, massachusetts.*

Swofford, David L; Waddell, Peter J; Huelsenbeck, J P; Foster, P G; Lewis, P O & Rogers, J S. 2001. Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. *Systematic biology*, **50**(4), 525–39.

Tamura, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular biology and evolution*, **9**(4), 678–87.

Tamura, K & Nei, M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Molecular biology and evolution*, **10**(3), 512–26.

Tavaré, Simon. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Pages 57–86 of:* Miura, RM (ed), *Lectures on mathematics in the life sciences, volume 17.* Providence (RI): American Mathematical Society.

Tillyard, R J. 1921. A new classification of the order perlaria. *The canadian entomologist*, **53**(2), 35–43.

Tuffley, C & Steel, M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. *Bulletin of mathematical biology*, **59**(3), 581–607.

Waddell, Peter J & Penny, David. 1996. Evolutionary trees of apes and humans from DNA sequences. *Pages 53–73 of:* Lock, Andrew & Peters, Charles R (eds), *Handbook of human symbolic evolution.* Clarendon Press, Oxford.

Wägele, Johann Wolfgang & Mayer, Christoph. 2007. Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects. *Bmc evolutionary biology*, **7**(Jan.), 147.

Wägele, Johann Wolfgang & Rödding, Friederike. 1998. A priori estimation of phylogenetic information conserved in aligned sequences. *Molecular phylogenetics and evolution*, **9**(3), 358–65.

Wägele, Johann Wolfgang; Holland, Barbara; Dreyer, Hermann & Hackethal, Beate. 2003. Searching factors causing implausible non-monophyly: ssu rDNA phylogeny of Isopoda Asellota (Crustacea: Peracarida) and faster evolution in marine than in freshwater habitats. *Molecular phylogenetics and evolution*, **28**(3), 536–51.

Whelan, Simon. 2008. Spatial and temporal heterogeneity in nucleotide sequence evolution. *Molecular biology and evolution*, **25**(8), 1683–94.

Whelan, Simon; Liò, Pietro & Goldman, Nick. 2001. Molecular phylogenetics: state-of-the-art methods for looking into the past. *Trends in genetics*, **17**(5), 262–272.

Wu, Jihua & Susko, Edward. 2009. General heterotachy and distance method adjustments. *Molecular biology and evolution*, **26**(12), 2689–97.

Yang, Ziheng. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of molecular evolution*, **39**(3), 306–314.

Yang, Ziheng; Goldman, Nick & Friday, A. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Molecular biology and evolution*, **11**(2), 316–324.

Zhong, Bojian; Deusch, Oliver; Goremykin, Vadim V; Penny, David; Biggs, Patrick J; Atherton, Robin A; Nikiforova, Svetlana V & Lockhart, Peter James. 2011. Systematic error in seed plant phylogenomics. *Genome biology and evolution*, **3**(Jan.), 1340–8.

Zhou, Yan; Rodrigue, Nicolas; Lartillot, Nicolas & Philippe, Hervé. 2007. Evaluation of the models handling heterotachy in phylogenetic inference. *Bmc evolutionary biology*, **7**(Jan.), 206.

A. Appendix

A.1. Simulated Datasets

A.1.1. Topologies and Simulations

For testing the analyses several simulation setups were generated and analysed. Within the chapters, the simulation setups were renamed. Different coloured branches are highlighting sequences which evolved according to varied evolutionary parameters.

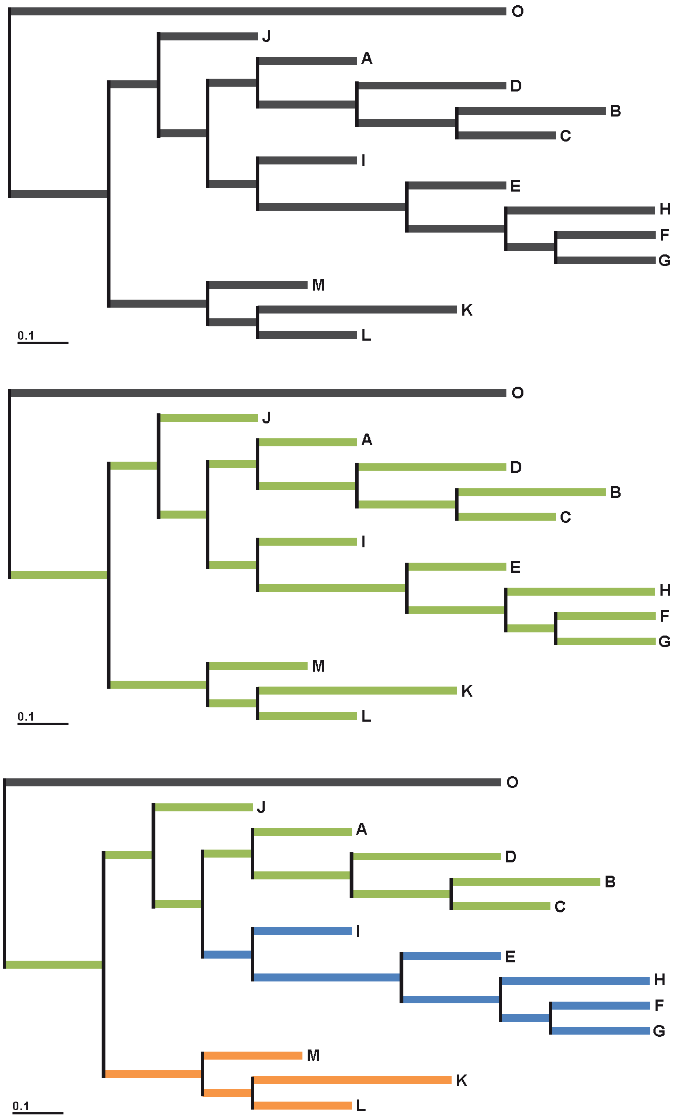


Figure A.1: Simulation setups S_50, 51 and 52

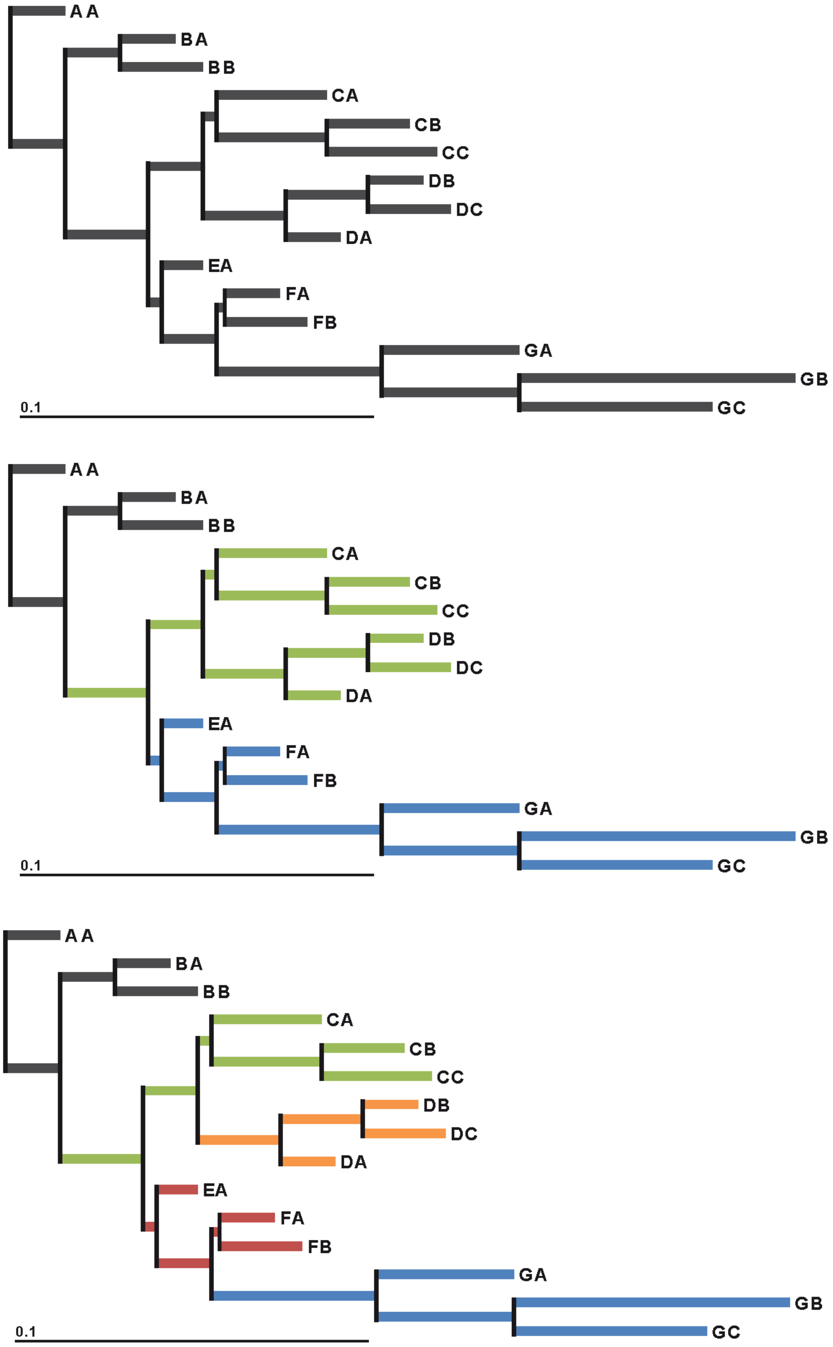


Figure A.2: Simulation setups S_60, 61 and 62

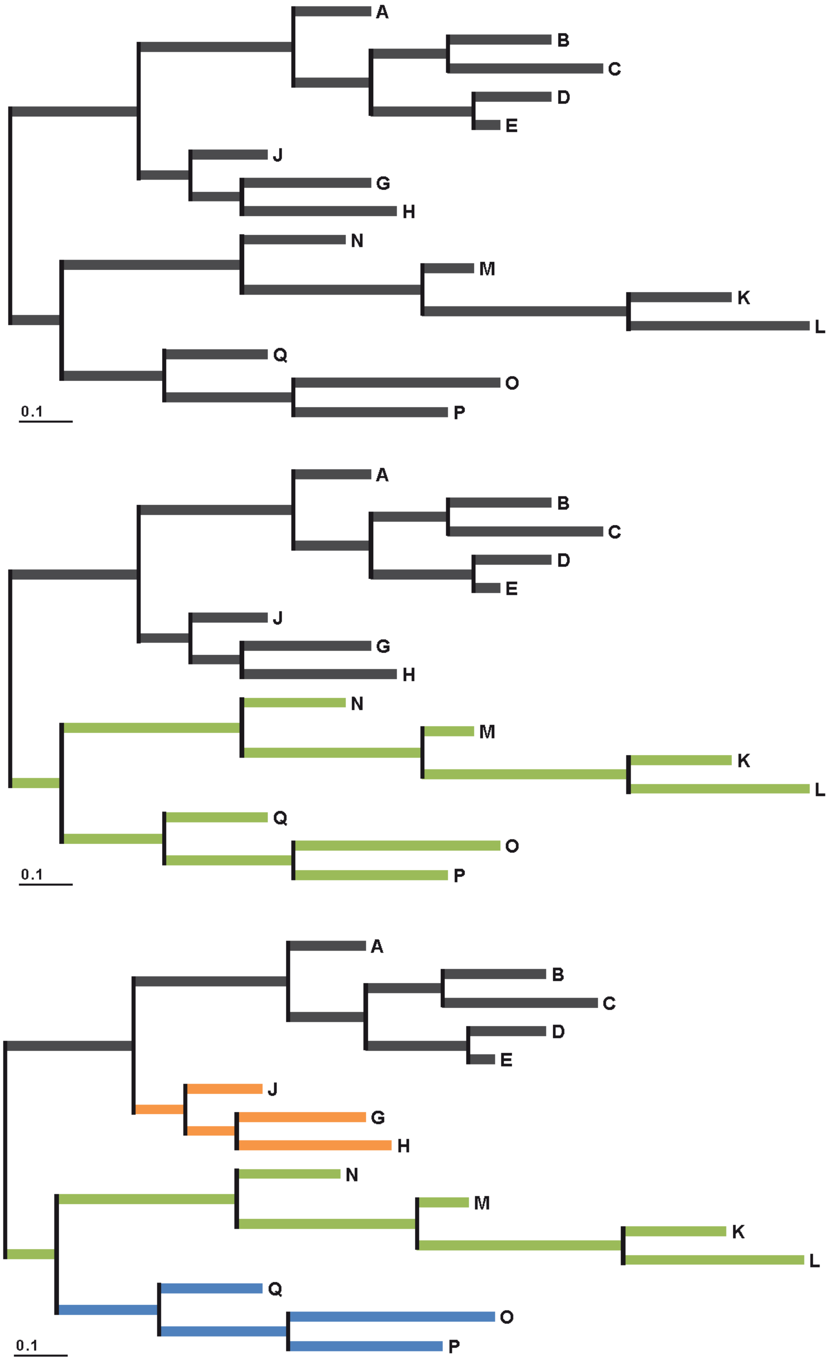


Figure A.3: Simulation setups S_70, 71 and 72

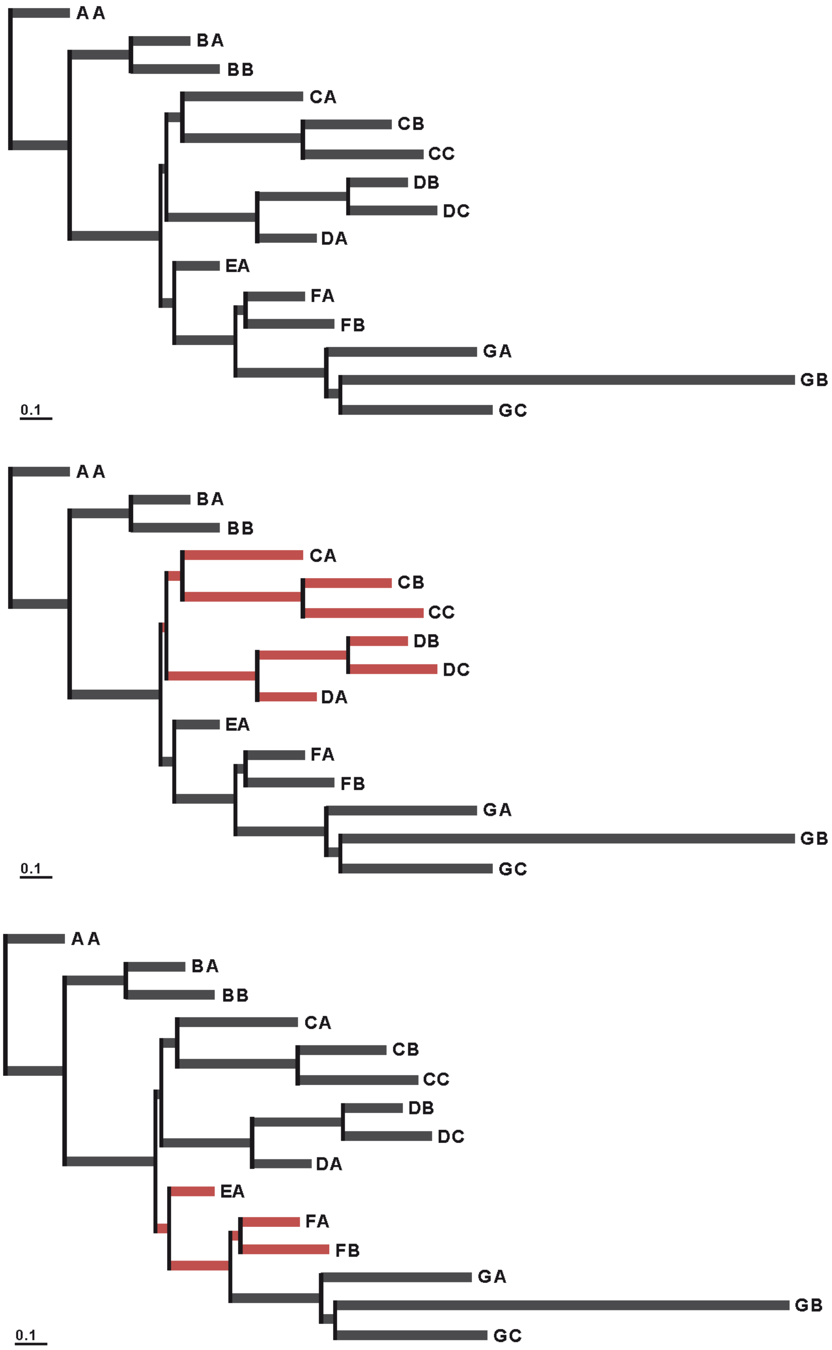


Figure A.4: Simulation setups S_80, 81 and 82

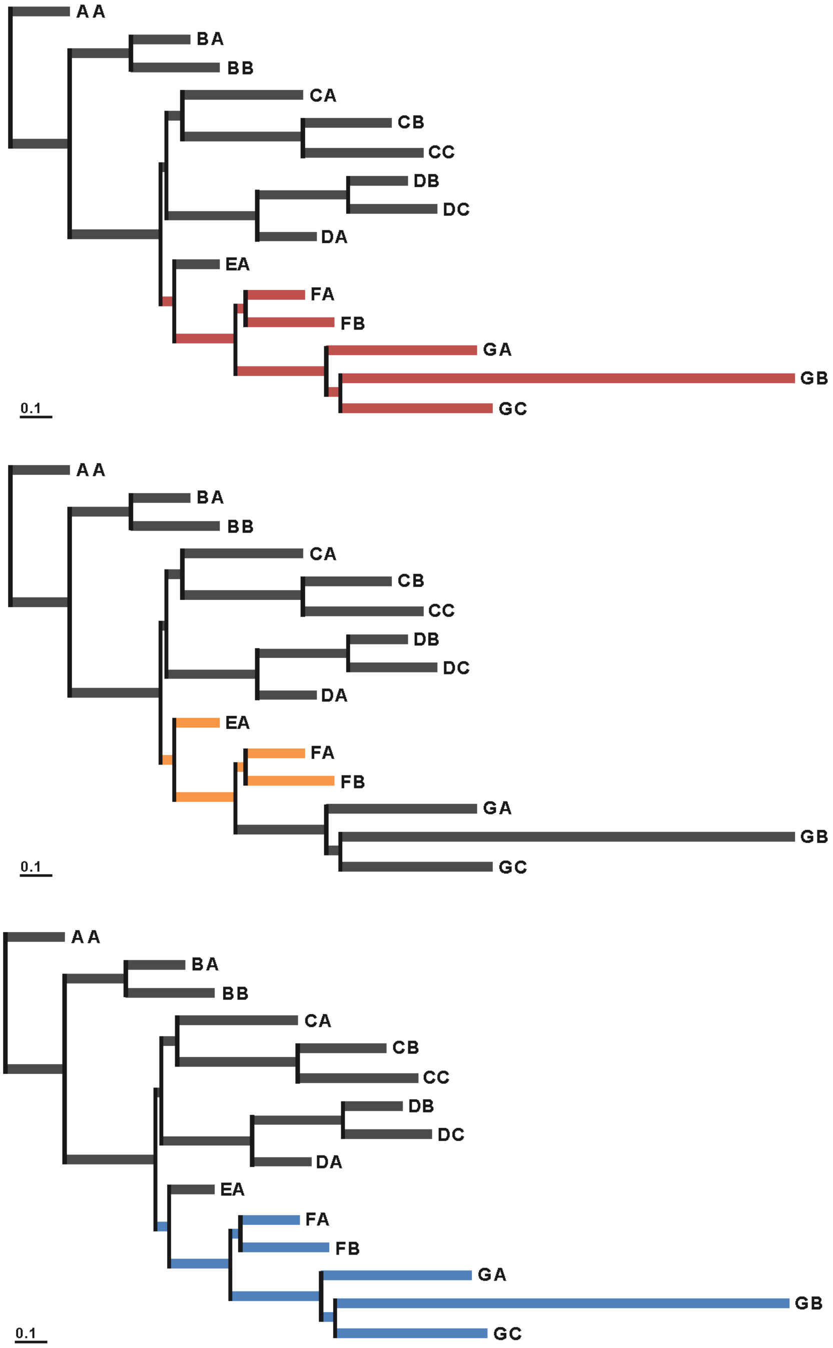


Figure A.5: Simulation setups S_83, 85 and 86

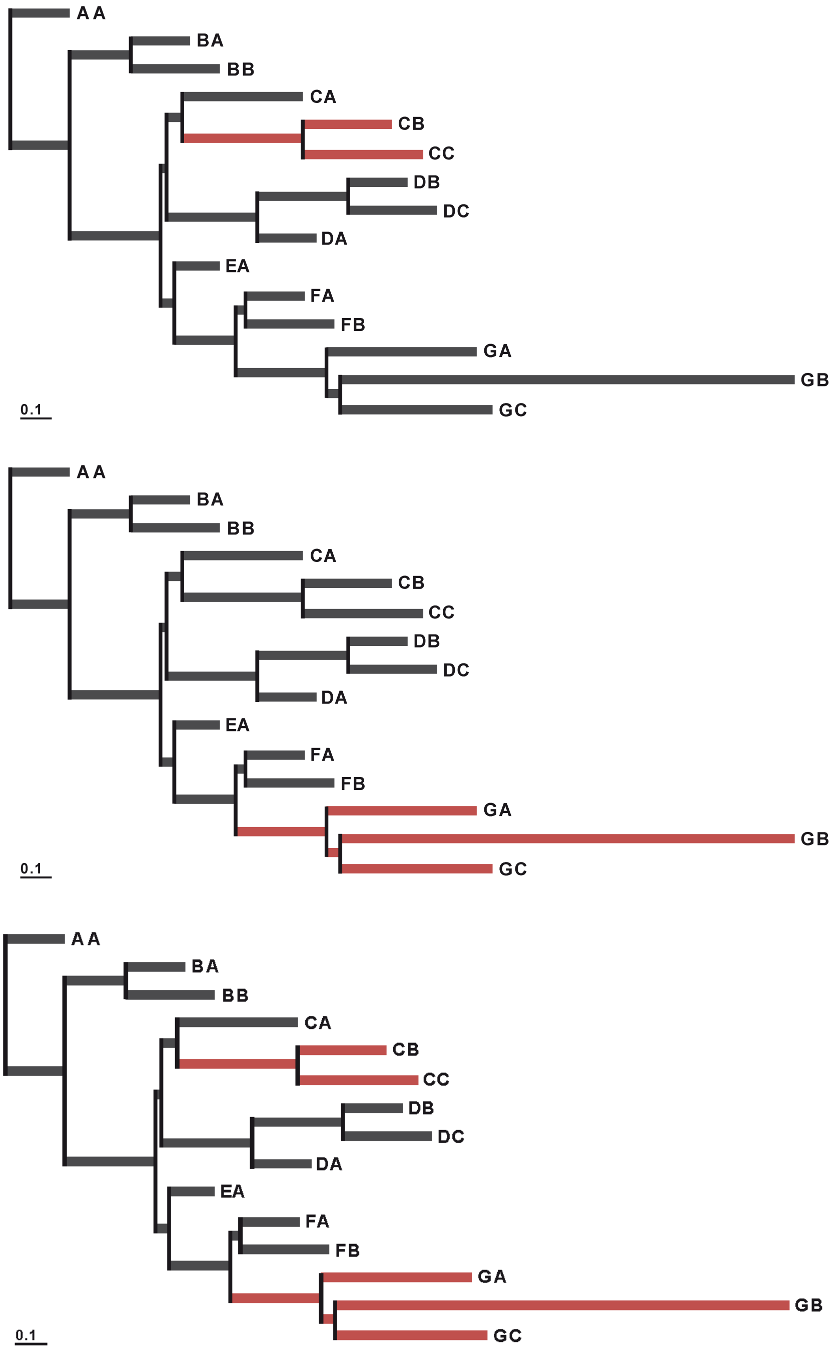


Figure A.6: Simulation setups S_94, 97 and 88

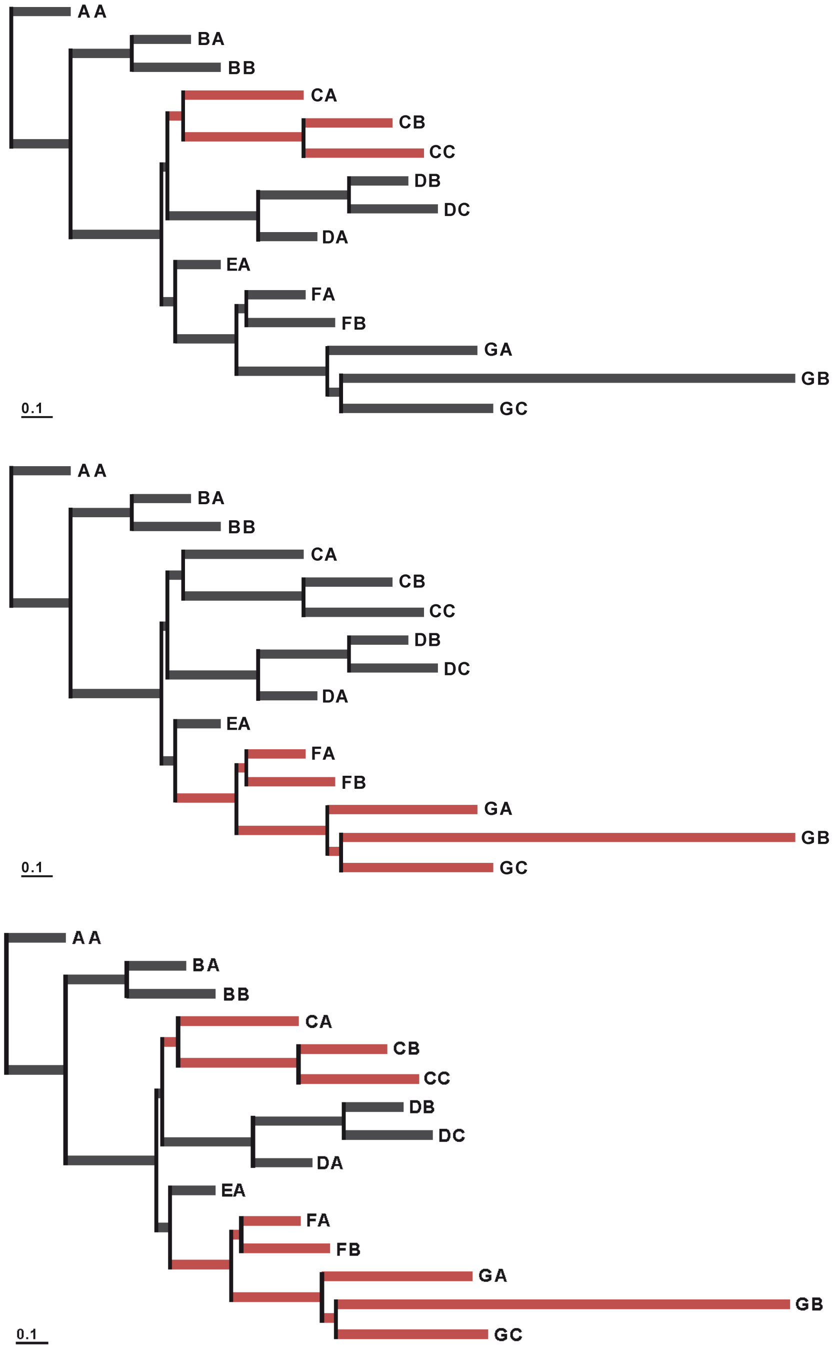


Figure A.7: Simulation setups S_84, 87 and 89

Table A.1: Combinations of datasets and analysis options used. Several combinations of topologies, branch length and evolutionary models were simulated and analysed. All simulations using continuous gamma-distribution were generated i) using a proportion of invariant sites (p_{inv}) or ii) with p_{inv}=0 for all different models (see table 2.1). The simulation runs using discrete gamma-distribution with 4 rate categories were all performed using p_{inv}=0 for all different models. The maximum likelihood analyses were performed six times, using i) 4, ii) 12 or iii) 25 rate categories and using either the mean or median.

Dataset	Model	Simulation software	ML software	Estimated parameters	# of bootstrap replica
S_50	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_51	GTR1/2	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_52	GTR1/2/3/4	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_60	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_61	GTR1/2/3	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_62	GTR1/2/3/4/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_70	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_71	GTR1/2	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_72	GTR1/2/3/4	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_80	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_81	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_82	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_83	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_84	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_85	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_86	GTR1	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_87	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_88	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_89	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_94	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100
S_97	GTR1/5	INDELible	PhyML	GTR+Γ or GTR+Γ+I	100

Table A.2: Combinations of datasets and software used in the analysis. Several combinations of topologies, branch length and evolutionary models were processed using different sequence simulation and maximum likelihood implementations in order to test their impact on the results.

Dataset	Model	Simulation software	ML software	Estimated parameters	# of bootstrap replica
S_80	GTR1	INDELible	PhyML	GTR+Γ+I	100
S_80	GTR1	INDELible	PhyML	GTR+Γ+I	1,000
S_80	GTR1	Seq-gen	PhyML	GTR+Γ+I	100
S_80	GTR1	Seq-gen	PhyML	GTR+Γ+I	1,000
S_80	GTR1	MultimoSeqSim	PhyML	GTR+Γ+I	100
S_80	GTR1	MultimoSeqSim	PhyML	GTR+Γ+I	1,000
S_80	GTR1	INDELible	PhyML	GTR+Γ+0.3	100

continued on next page...

Table A.2 – continued from previous page

Dataset	Model	Simulation software	ML software	Estimated parameters	# of bootstrap replica
S_80	GTR1	INDELible	PhyML	GTR+0.75+I	100
S_80	GTR1	INDELible	PAUP	GTR+Γ+I	100
S_80	GTR1	INDELible	RAxML	GTR+Γ+I	100
S_80	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100
S_80	GTR1, p_{inv}=0	INDELible	RAxML	GTR+Γ	100
S_80	GTR1, p_{inv}=0	INDELible	PAUP	GTR+Γ	100
S_80	GTR1, p_{inv}=0	Seq-gen	PhyML	GTR+Γ	100
S_80	GTR1, p_{inv}=0	Seq-gen	RAxML	GTR+Γ	100
S_80	GTR1, p_{inv}=0	Seq-gen	PAUP	GTR+Γ	100
S_80	GTR1, p_{inv}=0	MultimoSeqSim	PhyML	GTR+Γ	100
S_80	GTR1, p_{inv}=0	MultimoSeqSim	RAxML	GTR+Γ	100
S_80	GTR1, p_{inv}=0	MultimoSeqSim	PAUP	GTR+Γ	100
S_80-4	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100
S_80-4	GTR1	INDELible	PhyML	GTR+Γ+I	100
S_80-4	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100
S_80-4	GTR1	INDELible	PhyML	GTR+Γ+I	100
S_80-8	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100
S_80-8	GTR1	INDELible	PhyML	GTR+Γ+I	100
S_80-8	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100
S_80-8	GTR1	INDELible	PhyML	GTR+Γ+I	100
S_80-4	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	1,000
S_80-4	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	10,000
S_80-4	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100,000
S_80-4	GTR1	INDELible	PhyML	GTR+Γ+I	1,000
S_80-4	GTR1	INDELible	PhyML	GTR+Γ+I	10,000
S_80-4	GTR1	INDELible	PhyML	GTR+Γ+I	100,000
S_80	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	1,000
S_80	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	10,000
S_80	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100,000
S_80	GTR1	INDELible	PhyML	GTR+Γ+I	1,000
S_80	GTR1	INDELible	PhyML	GTR+Γ+I	10,000
S_80	GTR1	INDELible	PhyML	GTR+Γ+I	100,000
S_80-4-02	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100
S_80-4-005	GTR1, p_{inv}=0	INDELible	PhyML	GTR+Γ	100

A.2. Test Results

The simulation setups were analysed with different Maximum likelihood (ML) analysis options and the estimated models were evaluated with the Goldman-Cox test. Additionally, the datasets were checked for over- or under-represented splits.

A.2.1. Results of the Goldman-Cox Tests

Table A.3: Results of the Goldman-Cox test for simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and 4 rate categories for discrete Γ modelling, analysed using PhyML with 4, 12 or 25 categories (cat) for Γ-distribution, estimating the shape parameter (α) with or without p_{inv} and using the median or mean. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and parametric bootstraps. The results are listed for every seed, listing the rank which the original dataset achieved within the bootstrapped datasets (1-101). For ranks 1, 2, 100 and 101 the model was rejected, otherwise the model passed the test.

Simulated with 4 categories for Γ-distribution								
	ML options		Estimated Γ			Estimated Γ+I		
Sim	cat	GTR+Γ/+I	15	44	55	15	44	55
50_i	reference set		73	32	13	73	32	13
	4	mean	51	38	30	51	51	29
		median	3	1	1	3	1	1
	12	mean	101	101	101	101	101	101
		median	95	96	95	95	96	95
	25	mean	101	101	101	101	101	101
		median	100	101	101	100	101	101
50_s	reference set		82	14	36	7	16	60
	4	mean	56	40	38	62	47	69
		median	3	1	2	3	1	5
	12	mean	101	101	101	101	101	101
		median	93	98	96	100	100	100
	25	mean	101	101	101	101	101	101
		median	101	101	100	101	101	101
51_i	reference set		81	50	27	81	50	27
	4	mean	36	45	26	49	50	26
		median	1	1	1	1	1	1
	12	mean	101	101	101	101	101	101
		median	99	101	93	99	101	92
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
52_i	reference set		95	44	12	95	44	12
	4	mean	84	86	63	83	86	63
		median	5	1	1	5	1	1
	12	mean	101	101	101	101	101	101
		median	100	101	101	100	101	101
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
60_i	reference set		61	1	41	61	1	41
	4	mean	40	37	41	50	45	48
		median	3	7	6	4	3	7
	12	mean	96	98	93	88	86	88
		median	80	85	73	80	78	66
	25	mean	99	99	97	93	86	86
		median	93	98	89	81	82	84
60_s	reference set		61	26	46	19	46	88
	4	mean	39	50	45	47	49	60
		median	4	12	7	7	12	7
	12	mean	97	99	91	89	95	92
		median	81	88	70	71	81	79
	25	mean	99	101	95	93	95	93
		median	95	97	91	85	89	86

continued on next page...

Table A.3 – continued from previous page

Simulated with 4 categories for Γ-distribution								
	ML options		Estimated Γ			Estimated Γ+I		
Sim	cat	GTR+Γ/+I	15	44	55	15	44	55
61_i	reference set		79	1	35	79	1	35
	4	mean	52	55	43	61	74	49
		median	7	10	6	9	22	9
	12	mean	94	100	92	94	93	90
		median	83	87	74	80	82	70
	25	mean	97	101	95	95	93	83
		median	92	99	89	92	87	89
62_i	reference set		66	2	38	66	2	38
	4	mean	68	77	58	76	81	70
		median	9	15	13	15	18	13
	12	mean	101	101	96	99	98	93
		median	93	99	87	93	91	91
	25	mean	101	101	100	98	99	98
		median	100	101	95	96	96	96
70_i	reference set		84	13	23	84	13	23
	4	mean	56	59	31	56	56	35
		median	2	2	1	2	2	1
	12	mean	101	101	101	101	101	101
		median	99	98	89	99	98	89
	25	mean	101	101	101	101	101	101
		median	101	101	98	101	101	98
70_s	reference set		76	95	94	79	77	67
	4	mean	54	51	44	37	47	66
		median	2	3	2	2	2	2
	12	mean	101	101	101	101	101	101
		median	100	95	92	93	94	96
	25	mean	101	101	101	101	101	101
		median	101	101	99	101	100	100
71_i	reference set		85	34	19	85	34	19
	4	mean	88	89	57	88	80	64
		median	6	3	1	6	4	1
	12	mean	101	101	101	101	101	101
		median	100	101	95	100	101	95
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
72_i	reference set		77	27	20	77	27	20
	4	mean	88	95	85	88	94	92
		median	10	10	3	10	10	3
	12	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
80_i	reference set		47	12	28	47	12	28
	4	mean	41	50	45	40	51	46
		median	2	3	2	2	3	2
	12	mean	101	101	101	101	101	101
		median	96	99	98	96	99	98
	25	mean	101	101	101	101	101	101
		median	101	101	100	101	101	100
80_s	reference set		100	28	90	49	42	100
	4	mean	51	34	53	74	51	41
		median	3	3	4	7	5	4
	12	mean	101	101	101	101	101	100
		median	99	97	98	100	100	97
	25	mean	101	101	101	101	101	100
		median	101	101	100	101	101	99

continued on next page...

Table A.3 – continued from previous page

Simulated with 4 categories for Γ-distribution

	ML options		Estimated Γ			Estimated Γ+I		
Sim	cat	GTR+Γ/+I	15	44	55	15	44	55
81_i	reference set		51	20	29	51	20	29
	4	mean	44	56	60	46	47	62
		median	1	1	2	1	1	2
	12	mean	101	101	101	101	101	101
		median	100	101	100	100	101	100
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
82_i	reference set		45	19	23	45	19	23
	4	mean	48	57	40	48	57	53
		median	2	1	1	2	1	1
	12	mean	101	101	101	101	101	101
		median	99	101	98	99	101	98
	25	mean	101	101	101	101	101	101
		median	101	101	100	101	101	100
83_i	reference set		52	22	16	52	22	16
	4	mean	73	78	73	74	81	72
		median	7	5	3	7	5	3
	12	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
85_i	reference set		41	24	33	41	24	33
	4	mean	64	94	76	64	93	70
		median	8	14	5	8	16	6
	12	mean	101	101	100	101	101	100
		median	98	101	99	99	101	99
	25	mean	101	101	101	101	101	101
		median	101	101	100	101	101	100
86_i	reference set		49	34	22	49	34	22
	4	mean	62	90	72	61	83	64
		median	4	6	2	4	6	2
	12	mean	101	101	101	101	101	101
		median	100	101	99	100	101	100
	25	mean	101	101	101	101	101	101
		median	101	101	100	101	101	100
94_i	reference set		45	18	22	45	18	22
	4	mean	45	56	47	46	67	49
		median	2	1	2	2	1	2
	12	mean	101	101	101	101	101	101
		median	98	101	98	100	101	98
	25	mean	101	101	101	101	101	101
		median	101	101	100	101	101	100
97_i	reference set		52	18	13	52	18	13
	4	mean	71	80	76	73	81	77
		median	10	5	4	10	6	4
	12	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
88_i	reference set		52	22	12	52	22	12
	4	mean	72	78	66	72	75	70
		median	4	5	3	5	5	3
	12	mean	101	101	101	101	101	101
		median	101	101	100	101	101	100
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101

continued on next page...

Table A.3 – continued from previous page

Simulated with 4 categories for Γ-distribution								
	ML options		Estimated Γ			Estimated Γ+I		
Sim	cat	GTR+Γ/+I	15	44	55	15	44	55
84_i	reference set		35	17	25	35	17	25
	4	mean	46	56	52	46	56	71
		median	1	1	2	1	1	2
	12	mean	101	101	101	101	101	101
		median	100	101	99	100	101	99
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
87_i	reference set		54	23	14	54	23	14
	4	mean	74	75	72	76	82	75
		median	9	5	3	7	5	3
	12	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
89_i	reference set		47	28	15	47	28	15
	4	mean	68	77	69	68	70	70
		median	2	2	3	2	3	3
	12	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101
	25	mean	101	101	101	101	101	101
		median	101	101	101	101	101	101

Table A.4: Results of the Goldman-Cox test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and continuous Γ modelling, analysed using PhyML with 4, 12 or 25 categories (cat) for Γ-distribution, estimating the shape parameter (α) with or without p_{inv} and using the median or mean. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed, listing the rank which the original dataset achieved within the bootstrapped datasets (1-101). For ranks 1, 2, 100 and 101 the model was rejected, otherwise the model passed the test.

			Simulated continuous Γ						**Simulated continuous $\Gamma + I$**					
	ML options		**Estimated Γ**			**Estimated $\Gamma + I$**			**Estimated Γ**			**Estimated $\Gamma + I$**		
Sim	cat		15	44	55	15	44	55	15	44	55	15	44	55
50_i	reference set		46	90	29	46	90	29	84	90	14	84	90	14
	4	mean	1	2	1	1	2	1	95	67	35	3	2	3
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	35	53	39	35	48	39	50	56	41	34	42	34
		median	2	5	1	4	9	4	2	2	3	25	23	17
	25	mean	48	67	48	46	59	51	52	53	49	48	49	42
		median	19	33	22	25	34	23	19	26	15	45	39	36
50_s	reference set		76	78	95	51	76	78	30	71	78	78	2	41
	4	mean	1	1	1	2	1	1	22	61	20	3	1	2
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	56	41	41	48	28	42	54	52	50	47	42	43
		median	9	3	10	3	4	6	2	3	4	23	21	19
	25	mean	69	45	49	56	38	59	51	59	55	55	47	53
		median	38	19	23	37	11	26	18	23	13	38	34	37
51_i	reference set		39	18	56	39	18	56	54	76	54	54	76	54
	4	mean	40	16	9	40	17	15	99	98	99	101	98	99
		median	1	1	1	1	1	1	35	39	49	36	40	49
	12	mean	100	99	92	99	100	92	101	101	101	101	101	101
		median	84	83	67	86	84	62	101	101	101	101	101	101
	25	mean	101	100	97	101	101	97	101	101	101	101	101	101
		median	99	99	87	99	97	89	101	101	101	101	101	101
52_i	reference set		7	33	99	7	33	99	1	80	47	1	80	47
	4	mean	83	96	99	84	97	97	94	99	96	90	94	99
		median	2	12	10	2	13	15	2	4	4	5	4	8
	12	mean	101	101	101	100	101	101	101	101	101	101	101	101
		median	98	100	101	98	100	101	101	100	100	100	99	101
	25	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	100	101	101	101	101	101	101	101	101	101	101	101
60_i	reference set		53	51	47	53	51	47	68	28	35	68	28	35
	4	mean	3	4	3	6	5	6	1	1	1	13	14	8
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	31	41	35	31	41	35	13	27	14	36	40	34
		median	9	6	12	9	9	12	1	2	1	10	7	8
	25	mean	35	55	41	38	47	43	26	34	21	46	52	43
		median	20	30	21	20	31	22	3	11	6	25	29	27
60_s	reference set		36	30	98	83	88	98	45	23	40	10	4	15
	4	mean	1	6	3	21	7	4	1	1	1	16	5	11
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	10	43	44	44	35	35	14	22	22	44	35	37
		median	3	13	7	13	2	4	1	2	1	8	4	11
	25	mean	18	45	39	52	51	42	19	31	38	55	42	44
		median	10	28	24	23	23	22	7	9	10	16	24	23

continued on next page...

Table A.4 – continued from previous page

			Simulated continuous Γ						Simulated continuous $\Gamma + I$					
	ML options		Estimated Γ			Estimated $\Gamma + I$			Estimated Γ			Estimated $\Gamma + I$		
Sim	cat		15	44	55	15	44	55	15	44	55	15	44	55
61_i	reference set		53	9	14	53	9	14	4	30	11	4	30	11
	4	mean	29	46	19	28	48	31	22	29	40	62	30	37
		median	6	15	5	6	12	2	2	3	2	27	1	6
	12	mean	53	79	47	53	74	51	49	65	65	79	59	69
		median	29	62	26	41	61	36	48	34	42	68	44	42
	25	mean	52	78	50	52	76	67	73	70	78	80	74	78
		median	49	74	42	52	66	50	52	55	64	79	64	63
62_i	reference set		88	65	41	88	65	41	84	2	16	84	2	16
	4	mean	73	80	82	86	84	79	78	74	71	78	71	61
		median	51	51	63	80	55	65	33	38	34	60	40	57
	12	mean	80	89	91	87	89	91	95	90	90	87	84	91
		median	78	87	90	83	82	88	79	83	77	89	81	81
	25	mean	88	91	92	89	92	96	91	95	91	92	91	94
		median	84	90	87	78	84	87	87	89	89	88	73	83
70_i	reference set		62	97	43	62	97	43	90	85	21	90	85	21
	4	mean	1	1	1	1	1	1	96	100	66	3	1	1
		median	1	1	1	1	1	1	1	2	1	1	1	1
	12	mean	32	43	31	36	48	33	68	69	66	51	41	42
		median	7	5	5	3	6	5	5	7	3	23	24	17
	25	mean	46	52	46	35	45	55	78	74	73	47	54	41
		median	14	27	20	30	27	20	33	33	33	32	35	25
70_s	reference set		91	31	31	78	11	52	13	92	99	79	98	39
	4	mean	1	1	2	1	1	2	95	97	65	1	1	1
		median	1	1	1	1	1	1	1	2	1	1	1	1
	12	mean	32	36	36	50	45	43	70	63	65	33	50	36
		median	2	6	3	8	5	7	5	5	2	22	34	18
	25	mean	35	40	44	50	56	48	77	70	67	56	51	43
		median	12	20	21	26	32	26	38	36	26	39	43	32
71_i	reference set		24	22	17	24	22	17	95	78	33	95	78	33
	4	mean	11	25	5	5	14	1	27	47	27	26	26	31
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	89	94	78	67	88	53	100	100	99	90	88	90
		median	41	47	27	26	37	14	80	93	68	64	48	46
	25	mean	93	94	89	80	87	70	101	101	101	91	91	93
		median	74	82	70	65	75	46	100	100	99	84	87	79
72_i	reference set		72	90	39	72	90	39	68	25	19	68	25	19
	4	mean	87	100	94	78	93	87	101	101	100	99	98	95
		median	3	12	11	2	6	5	16	16	6	23	21	10
	12	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	94	101	99	95	100	94	101	101	101	101	101	100
	25	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	99	101	100	100	101	99	101	101	101	101	101	101
80_i	reference set		26	83	41	26	83	41	94	96	36	94	96	36
	4	mean	1	1	1	1	1	1	95	99	76	1	2	1
		median	1	1	1	1	1	1	1	1	1	1	2	1
	12	mean	30	44	28	34	32	29	51	68	63	48	38	31
		median	4	8	5	4	8	5	2	3	4	36	29	19
	25	mean	36	44	34	36	60	36	59	77	71	54	46	40
		median	16	31	18	23	29	18	22	45	26	48	46	32
80_s	reference set		31	21	67	21	78	81	58	66	11	98	31	85
	4	mean	1	1	1	1	1	1	93	89	99	1	1	1
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	40	49	44	15	32	35	65	54	62	56	29	27
		median	5	14	11	3	9	8	5	5	8	40	13	11
	25	mean	50	56	51	25	39	50	77	68	75	63	34	35
		median	19	33	24	8	22	24	30	32	34	54	31	35

continued on next page...

Table A.4 – continued from previous page

			Simulated continuous Γ						Simulated continuous Γ + *I*					
	ML options		Estimated Γ			Estimated Γ + *I*			Estimated Γ			Estimated Γ + *I*		
Sim	cat		15	44	55	15	44	55	15	44	55	15	44	55
81_i	reference set		51	23	23	51	23	23	93	63	53	93	63	53
	4	mean	50	47	37	18	13	27	99	100	101	100	93	97
		median	1	1	1	1	1	1	4	4	2	23	10	8
	12	mean	100	94	99	91	84	97	101	101	101	101	101	101
		median	75	63	68	54	42	61	101	101	101	100	101	101
	25	mean	99	100	98	94	89	99	101	101	101	101	101	101
		median	101	94	93	88	66	96	101	101	101	101	101	100
82_i	reference set		13	34	11	13	34	11	100	81	44	100	81	44
	4	mean	43	63	30	28	59	42	101	101	101	101	101	101
		median	1	1	1	1	1	1	14	45	51	29	60	58
	12	mean	100	101	97	99	101	99	101	101	101	101	101	101
		median	72	90	71	82	90	69	101	101	101	101	101	101
	25	mean	100	100	97	101	101	99	101	101	101	101	101	101
		median	100	100	96	100	100	93	101	101	101	101	101	101
83_i	reference set		11	50	24	11	50	24	48	14	11	48	14	11
	4	mean	35	36	36	18	19	21	98	96	86	96	94	77
		median	1	1	1	1	1	1	2	1	1	11	8	7
	12	mean	100	98	95	95	98	98	101	101	101	101	101	101
		median	73	76	71	65	76	74	101	101	99	101	101	100
	25	mean	100	101	100	101	99	100	101	101	101	101	101	101
		median	94	99	96	97	96	93	101	101	101	101	101	100
85_i	reference set		31	47	61	31	47	61	84	91	31	84	91	31
	4	mean	92	99	85	82	95	84	101	101	101	101	101	101
		median	1	3	2	1	2	1	55	56	55	83	83	73
	12	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	99	101	97	100	100	100	101	101	101	101	101	101
	25	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	101	101	101	101	101	101	101	101	101	101	101	101
86_i	reference set		4	40	59	4	40	59	84	52	13	84	52	13
	4	mean	20	70	42	20	37	17	101	100	96	101	98	97
		median	1	1	1	1	1	1	3	8	2	14	14	11
	12	mean	96	101	100	95	101	100	101	101	101	101	101	101
		median	61	93	75	64	88	72	101	101	101	101	101	101
	25	mean	98	101	101	98	101	99	101	101	101	101	101	101
		median	93	101	98	92	100	92	101	101	101	101	101	101
94_i	reference set		39	24	31	39	24	31	39	24	31	39	24	31
	4	mean	38	46	53	30	42	45	38	46	53	30	42	45
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	100	100	100	99	101	100	100	100	100	99	101	100
		median	76	81	85	71	86	84	76	81	85	71	86	84
	25	mean	101	101	100	101	101	100	101	101	100	101	101	100
		median	97	97	97	96	98	97	97	97	97	96	98	97
97_i	reference set		32	92	38	32	92	38	32	92	38	32	92	38
	4	mean	97	98	91	98	99	91	97	98	91	98	99	91
		median	6	13	2	6	15	8	6	13	2	6	15	8
	12	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	100	101	101	100	101	101	100	101	101	100	101	101
	25	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	101	101	101	101	101	101	101	101	101	101	101	101
88_i	reference set		18	40	4	18	40	4	35	85	19	35	85	19
	4	mean	74	70	58	77	71	69	101	101	101	101	101	101
		median	1	1	1	2	1	1	33	56	27	38	60	28
	12	mean	100	100	100	100	101	101	101	101	101	101	101	101
		median	93	94	85	93	93	85	101	101	101	101	101	101
	25	mean	100	101	100	101	101	100	101	101	101	101	101	101
		median	101	99	98	100	99	96	101	101	101	101	101	101

continued on next page...

Table A.4 – continued from previous page

			Simulated continuous Γ						Simulated continuous Γ + I					
	ML options		Estimated Γ			Estimated Γ + I			Estimated Γ			Estimated Γ + I		
Sim	cat		15	44	55	15	44	55	15	44	55	15	44	55
84_i	reference set		42	12	17	42	12	17	42	12	17	42	12	17
	4	mean	40	10	36	42	9	43	40	10	36	42	9	43
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	100	94	101	100	93	98	100	94	101	100	93	98
		median	81	48	84	78	44	84	81	48	84	78	44	84
	25	mean	101	94	99	101	99	100	101	94	99	101	99	100
		median	95	83	95	95	87	95	95	83	95	95	87	95
87_i	reference set		16	81	51	16	81	51	16	81	51	16	81	51
	4	mean	39	66	65	43	46	48	39	66	65	43	46	48
		median	1	1	1	1	1	1	1	1	1	1	1	1
	12	mean	100	101	101	101	101	101	100	101	101	101	101	101
		median	84	95	92	86	94	92	84	95	92	86	94	92
	25	mean	101	101	101	101	101	101	101	101	101	101	101	101
		median	100	100	101	98	100	100	100	100	101	98	100	100
89_i	reference set		55	72	17	55	72	17	84	50	5	84	50	5
	4	mean	29	57	46	20	45	45	101	101	101	101	99	99
		median	1	3	2	1	2	1	26	24	15	39	18	18
	12	mean	92	96	95	79	95	96	101	101	101	101	101	101
		median	56	81	74	46	75	73	101	101	101	101	100	101
	25	mean	93	95	98	89	93	97	101	101	101	101	101	101
		median	86	93	86	73	95	91	101	101	101	101	101	101

A.2.2. Results of the Residual Diagnostics Tests

Table A.5: Results of the test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and four categories for modelling gamma distribution. The data was analysed by a maximum likelihood approach using PhyML with 4, 12 or 25 categories for gamma-distribution, estimating the shape parameter (α) and using the median or mean. 100 parametric bootstraps were generated with INDELible using the estimated GTR model. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or under-represented splits;
sp = amount of splits detected as over- or under-represented;
dif = number of sites which represent all over- or under-represented splits;
green cells = no over- or underrepresentation;
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Simulated: four rate categories for Γ – Estimated: Γ

	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
50_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	29	1	13	3	16	0	0	4	27	0	0
		median	4	27	1	12	6	42	0	0	3	22	0	0
	12	mean	4	119	1	14	6	168	0	0	3	146	0	0
		median	4	26	1	13	5	113	0	0	3	24	0	0
	25	mean	4	125	1	13	5	170	0	0	4	162	0	0
		median	4	25	1	14	7	161	0	0	2	120	0	0
50_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	24	0	0	2	15	0	0	2	11	1	16
		median	1	4	0	0	2	15	1	100	2	14	1	15
	12	mean	2	123	0	0	4	132	0	0	5	152	1	16
		median	3	23	0	0	3	21	0	0	3	19	1	16
	25	mean	3	136	0	0	3	135	0	0	4	150	1	15
		median	2	105	0	0	4	115	0	0	5	133	1	16
51_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	0	0	0	0	8	49	1	6	2	11	0	0
		median	1	7	1	115	9	51	1	6	4	25	1	97
	12	mean	2	116	0	0	7	183	0	0	4	153	0	0
		median	0	0	0	0	7	141	0	0	4	108	0	0
	25	mean	2	122	0	0	7	191	0	0	3	152	0	0
		median	1	7	0	0	7	169	0	0	4	140	0	0
52_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	14	2	41	4	33	4	126	4	53	3	70
		median	2	15	2	113	6	45	3	108	7	66	1	29
	12	mean	3	124	1	16	6	188	4	121	6	168	2	47
		median	2	14	2	40	5	132	4	123	5	59	2	48
	25	mean	3	129	1	15	6	191	4	120	6	174	2	47
		median	2	14	2	41	6	167	4	121	7	174	2	47
60_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	6	1	16	1	11	0	0	0	0	0	0
		median	1	6	1	111	1	11	0	0	0	0	0	0
	12	mean	2	144	1	19	2	140	0	0	0	0	0	0
		median	2	12	1	18	1	11	0	0	0	0	0	0
	25	mean	2	168	1	19	2	160	0	0	0	0	0	0
		median	2	130	1	19	2	124	0	0	0	0	0	0

continued on next page...

			Simulated: four rate categories for Γ – Estimated: Γ											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
60_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	0	0	0	0	1	7	0	0	1	9	0	0
		median	1	6	1	113	0	0	0	0	2	14	0	0
	12	mean	1	139	0	0	2	137	0	0	1	9	0	0
		median	0	0	1	10	1	7	0	0	1	9	0	0
	25	mean	1	164	0	0	2	157	0	0	1	9	0	0
		median	1	126	0	0	2	122	0	0	1	9	0	0
61_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	13	0	0	1	6	0	0	0	0	1	19
		median	1	8	1	16	1	6	0	0	0	0	1	16
	12	mean	3	21	0	0	1	6	1	23	0	0	1	22
		median	2	13	0	0	1	6	0	0	0	0	1	20
	25	mean	3	150	0	0	2	151	1	23	0	0	1	22
		median	2	13	0	0	1	6	1	22	0	0	1	21
62_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	16	0	0	1	5	1	27	1	6	0	0
		median	3	24	1	111	1	5	0	0	1	6	0	0
	12	mean	5	165	0	0	1	5	2	65	2	12	1	28
		median	2	17	0	0	1	5	1	30	2	12	1	25
	25	mean	5	189	0	0	2	161	2	67	3	154	1	29
		median	4	143	0	0	1	5	2	63	2	12	1	27
70_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	26	0	0	3	21	0	0	4	25	0	0
		median	7	64	0	0	5	33	0	0	5	33	1	113
	12	mean	9	169	0	0	3	158	0	0	7	136	0	0
		median	7	62	0	0	3	21	0	0	4	25	0	0
	25	mean	7	164	0	0	3	164	0	0	5	134	0	0
		median	8	69	0	0	3	140	0	0	4	25	0	0
70_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	17	0	0	4	26	0	0	9	86	0	0
		median	3	17	0	0	4	26	0	0	6	41	1	118
	12	mean	3	135	1	7	5	146	0	0	7	74	0	0
		median	1	6	1	6	3	21	0	0	7	74	0	0
	25	mean	3	140	1	6	5	152	0	0	7	74	0	0
		median	3	116	1	7	4	122	0	0	7	73	0	0
71_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	48	0	0	4	36	1	17	7	51	0	0
		median	7	58	1	117	5	50	2	32	8	60	2	141
	12	mean	7	54	0	0	4	151	0	0	9	153	0	0
		median	8	62	0	0	4	33	0	0	8	58	0	0
	25	mean	6	48	0	0	6	181	0	0	9	160	0	0
		median	6	48	0	0	5	147	1	16	8	58	0	0
72_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	36	2	51	5	67	4	89	6	109	2	44
		median	2	10	2	51	7	82	3	82	6	91	3	156
	12	mean	4	126	3	63	7	232	4	87	3	39	1	20
		median	3	35	2	50	6	100	4	88	3	66	1	20
	25	mean	3	128	3	63	7	239	4	86	3	40	1	19
		median	3	36	3	64	7	215	4	88	4	78	1	20
80_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	26	0	0	4	23	0	0	2	16	1	9
		median	5	28	1	93	4	27	0	0	2	17	1	9
	12	mean	8	133	0	0	5	158	0	0	2	17	1	9
		median	7	35	0	0	4	29	0	0	3	22	1	9
	25	mean	7	132	0	0	5	164	0	0	2	17	1	9
		median	6	31	0	0	5	144	0	0	2	17	1	9

continued on next page...

			Simulated: four rate categories for Γ – Estimated: Γ											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
80_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	12	0	0	0	0	0	0	3	21	0	0
		median	1	6	0	0	0	0	0	0	5	40	0	0
	12	mean	3	140	1	16	1	112	0	0	3	28	0	0
		median	3	99	0	0	0	0	0	0	2	14	0	0
	25	mean	3	145	1	19	1	116	0	0	3	28	0	0
		median	4	129	0	0	1	96	0	0	3	28	0	0
81_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	81	0	0	7	59	1	24	6	69	2	50
		median	8	104	1	95	5	26	1	26	6	70	3	149
	12	mean	8	201	0	0	8	203	0	0	7	94	1	27
		median	8	97	0	0	7	160	1	23	6	70	1	28
	25	mean	8	208	0	0	8	209	0	0	7	183	1	27
		median	8	188	0	0	8	190	1	23	6	71	1	28
82_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	47	0	0	4	71	1	31	3	27	1	17
		median	3	16	1	93	3	39	3	112	6	45	1	17
	12	mean	9	156	0	0	6	221	1	30	7	180	2	25
		median	10	102	0	0	4	166	2	50	5	39	1	17
	25	mean	7	152	0	0	6	226	1	29	7	185	2	25
		median	11	176	0	0	6	207	2	49	5	39	0	0
83_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	175	0	0	4	82	3	61	7	241	0	0
		median	7	186	1	125	3	57	5	170	8	249	1	127
	12	mean	5	192	0	0	3	159	1	22	7	242	0	0
		median	7	197	1	20	2	52	3	69	8	243	0	0
	25	mean	5	194	0	0	3	163	1	21	7	246	0	0
		median	6	197	0	0	3	144	1	19	8	262	0	0
85_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	28	0	0	2	16	0	0	3	22	1	30
		median	3	17	1	19	1	5	0	0	3	22	1	31
	12	mean	2	11	0	0	2	117	0	0	4	27	1	29
		median	2	12	0	0	1	5	0	0	4	27	1	30
	25	mean	2	11	0	0	3	147	0	0	7	100	1	28
		median	2	11	0	0	3	112	0	0	5	32	1	29
86_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	43	0	0	5	29	1	12	5	38	1	17
		median	1	35	1	96	6	36	1	12	3	20	0	0
	12	mean	3	140	0	0	7	171	1	12	4	28	0	0
		median	2	44	0	0	7	47	1	12	5	37	1	9
	25	mean	3	144	0	0	6	170	1	12	4	29	0	0
		median	2	45	0	0	8	166	1	13	5	35	1	9
88_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	8	156	1	64	6	78	1	21	7	195	0	0
		median	10	184	1	132	4	60	3	130	6	190	2	143
	12	mean	9	169	0	0	6	191	0	0	8	257	1	17
		median	9	165	0	0	5	89	1	21	5	210	1	18
	25	mean	9	171	0	0	6	196	1	21	8	262	1	17
		median	9	168	0	0	5	91	0	0	8	256	1	18
89_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	97	0	0	6	70	2	56	7	180	2	42
		median	9	149	1	128	5	61	4	159	8	170	4	192
	12	mean	8	172	0	0	7	180	2	55	10	208	1	22
		median	7	114	0	0	6	70	2	56	7	175	0	0
	25	mean	8	176	0	0	7	185	2	54	8	197	0	0
		median	7	115	0	0	7	168	2	55	10	196	1	23

Table A.6: Results of the test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and four categories for modelling gamma distribution. The data was analysed by a maximum likelihood approach using PhyML with 4, 12 or 25 categories for gamma-distribution, estimating the shape parameter (α) with a proportion of invariant sites (p_{inv}) and using the median or mean. 100 parametric bootstraps were generated with INDELible using the estimated GTR model. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or under-represented splits;
sp = amount of splits detected as over- or under-represented;
dif = number of sites which represent all over- or under-represented splits;
green cells = no over- or underrepresentation;
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Simulated: four rate categories for Γ – Estimated: Γ+I

	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
50_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	29	1	13	2	11	0	0	3	22	0	0
		median	5	31	1	12	6	42	0	0	3	22	0	0
	12	mean	4	120	1	14	6	171	0	0	3	147	0	0
		median	4	26	1	13	5	114	0	0	3	24	0	0
	25	mean	4	126	1	14	5	170	0	0	4	160	0	0
		median	4	25	1	14	7	161	0	0	2	120	0	0
50_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	23	0	0	6	71	0	0	4	22	0	0
		median	5	30	0	0	8	52	1	11	6	32	0	0
	12	mean	6	158	0	0	10	216	1	19	4	124	0	0
		median	5	31	0	0	8	152	0	0	4	22	0	0
	25	mean	6	166	0	0	10	244	1	7	8	167	0	0
		median	6	142	0	0	8	185	0	0	4	22	0	0
51_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	7	0	0	6	36	0	0	3	16	0	0
		median	1	7	1	115	7	45	1	85	6	36	1	116
	12	mean	2	116	0	0	7	184	0	0	4	154	0	0
		median	0	0	0	0	7	141	0	0	4	107	0	0
	25	mean	2	123	0	0	7	191	0	0	3	154	0	0
		median	1	7	0	0	7	169	0	0	4	139	0	0
52_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	15	2	38	5	29	3	105	6	65	1	21
		median	2	15	2	112	6	45	3	108	7	66	1	29
	12	mean	3	126	1	15	6	188	4	121	6	171	2	48
		median	2	14	2	40	5	133	4	123	5	59	2	48
	25	mean	3	130	1	15	7	197	4	120	6	175	2	47
		median	2	14	2	40	6	168	4	121	7	174	2	47
60_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	22	0	0	2	18	0	0	0	0	0	0
		median	1	6	0	0	3	23	0	0	0	0	1	103
	12	mean	1	6	1	18	3	25	0	0	0	0	0	0
		median	2	12	0	0	2	18	0	0	0	0	0	0
	25	mean	1	6	1	18	1	11	0	0	0	0	0	0
		median	2	14	0	0	3	25	0	0	0	0	0	0

continued on next page...

Simulated: four rate categories for Γ – Estimated: Γ+I

	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
60_s	reference set		0	0	0	0	0	0	0	0	1	6	1	29
	4	mean	1	4	0	0	1	5	0	0	1	6	0	0
		median	0	0	0	0	2	16	0	0	3	16	0	0
	12	mean	1	6	0	0	2	16	0	0	2	9	0	0
		median	0	0	0	0	1	6	0	0	1	6	0	0
	25	mean	0	0	0	0	1	9	0	0	0	0	0	0
		median	2	12	0	0	2	14	0	0	1	6	0	0
61_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	7	0	0	2	11	0	0	0	0	1	21
		median	4	43	0	0	0	0	0	0	0	0	1	19
	12	mean	1	5	0	0	1	6	0	0	0	0	0	0
		median	0	0	1	19	0	0	0	0	0	0	0	0
	25	mean	2	16	1	19	0	0	0	0	0	0	0	0
		median	2	16	1	19	0	0	0	0	0	0	1	22
62_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	20	0	0	0	0	0	0	1	6	0	0
		median	3	19	0	0	1	7	0	0	4	60	0	0
	12	mean	2	18	0	0	1	5	1	25	2	29	1	23
		median	3	23	0	0	1	5	1	28	0	0	0	0
	25	mean	3	22	0	0	1	5	2	55	0	0	0	0
		median	3	22	0	0	1	5	1	26	2	31	0	0
70_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	26	0	0	3	30	0	0	8	53	0	0
		median	7	64	0	0	4	27	0	0	5	33	1	113
	12	mean	9	170	0	0	3	158	0	0	7	137	0	0
		median	7	62	0	0	3	21	0	0	4	25	0	0
	25	mean	7	165	0	0	3	164	0	0	5	133	0	0
		median	8	69	0	0	3	141	0	0	5	32	0	0
70_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	16	0	0	3	23	1	23	6	44	0	0
		median	2	18	0	0	2	15	2	126	5	40	0	0
	12	mean	3	135	0	0	3	120	0	0	6	143	0	0
		median	2	18	0	0	2	15	0	0	5	37	0	0
	25	mean	3	143	0	0	3	125	0	0	7	158	0	0
		median	3	115	0	0	2	15	0	0	6	46	0	0
71_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	7	57	0	0	7	51	1	17	7	50	2	39
		median	7	58	1	115	5	49	2	32	8	60	1	118
	12	mean	7	54	0	0	4	152	0	0	9	153	0	0
		median	9	68	0	0	4	33	0	0	8	58	0	0
	25	mean	6	48	0	0	6	183	0	0	9	162	0	0
		median	6	48	0	0	5	148	1	16	8	58	0	0
72_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	36	2	50	5	68	4	87	8	89	3	62
		median	2	10	2	51	6	74	3	82	6	91	3	156
	12	mean	4	127	3	63	7	233	4	86	3	39	1	20
		median	3	35	2	49	6	100	4	88	3	66	1	20
	25	mean	3	128	3	63	7	239	4	86	3	40	1	20
		median	3	37	3	64	7	216	4	88	4	78	1	20
80_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	26	0	0	4	23	0	0	4	50	0	0
		median	5	28	1	93	4	27	0	0	2	17	1	9
	12	mean	7	129	0	0	5	160	0	0	2	17	1	9
		median	6	30	0	0	4	29	0	0	3	22	1	9
	25	mean	7	135	0	0	5	163	0	0	3	140	1	9
		median	7	36	0	0	5	142	0	0	2	17	1	9

continued on next page...

Simulated: four rate categories for Γ – Estimated: Γ+I

	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
80_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	51	1	11	3	39	0	0	5	30	0	0
		median	2	21	1	11	2	33	1	86	5	30	0	0
	12	mean	3	117	1	11	3	133	0	0	4	24	0	0
		median	3	28	1	11	2	33	0	0	4	25	0	0
	25	mean	3	121	1	11	3	136	0	0	4	25	0	0
		median	3	42	1	11	3	41	0	0	4	24	0	0
81_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	82	0	0	8	68	1	22	5	33	2	45
		median	8	104	1	95	5	26	1	26	6	70	3	146
	12	mean	8	202	0	0	8	203	0	0	7	94	1	27
		median	8	97	0	0	7	162	1	23	6	70	1	28
	25	mean	9	215	0	0	8	209	0	0	7	183	1	27
		median	8	186	0	0	8	189	1	22	6	71	1	28
82_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	47	0	0	5	84	3	69	7	65	0	0
		median	3	16	1	92	3	39	3	112	6	45	1	17
	12	mean	9	158	0	0	6	222	1	30	7	181	2	25
		median	10	102	0	0	4	169	2	50	5	39	1	16
	25	mean	7	152	0	0	6	228	1	29	7	186	2	25
		median	11	175	0	0	6	208	2	49	6	59	0	0
83_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	175	0	0	3	55	3	62	7	242	0	0
		median	7	186	1	126	3	57	5	168	7	215	1	126
	12	mean	5	192	0	0	3	159	1	22	7	242	0	0
		median	7	197	1	20	2	52	3	69	8	243	0	0
	25	mean	5	194	0	0	3	166	1	21	7	246	0	0
		median	6	196	0	0	3	146	0	0	8	262	0	0
85_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	28	0	0	1	5	0	0	3	21	0	0
		median	3	17	1	19	1	5	0	0	3	22	1	31
	12	mean	2	11	0	0	2	118	0	0	4	27	1	28
		median	2	12	0	0	1	5	0	0	4	27	1	30
	25	mean	2	11	0	0	3	147	0	0	7	100	1	28
		median	2	11	0	0	3	113	0	0	5	32	1	29
86_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	43	0	0	6	33	1	13	6	44	0	0
		median	1	35	1	96	6	36	1	12	3	20	1	17
	12	mean	3	141	0	0	7	172	1	12	4	29	0	0
		median	2	44	0	0	7	47	1	12	5	37	1	9
	25	mean	3	145	0	0	6	170	1	12	4	29	0	0
		median	2	44	0	0	8	167	1	13	5	35	1	9
88_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	8	156	1	63	6	78	1	21	7	195	0	0
		median	10	184	1	132	4	60	3	127	6	190	2	143
	12	mean	9	169	0	0	6	190	0	0	8	257	1	17
		median	9	165	0	0	5	89	1	21	5	210	1	18
	25	mean	9	171	0	0	6	196	1	21	8	262	1	17
		median	9	168	0	0	5	91	1	19	8	256	2	39
89_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	97	0	0	5	82	2	55	7	180	2	42
		median	9	149	1	128	5	61	4	158	8	170	4	191
	12	mean	7	115	0	0	7	181	2	55	10	208	1	22
		median	7	114	0	0	6	70	2	56	7	175	0	0
	25	mean	8	176	0	0	7	186	2	54	9	203	0	0
		median	7	115	0	0	7	169	2	55	11	216	1	23

Table A.7: Results of the test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and continuous modelling gamma distribution. The data was analysed by a maximum likelihood approach using PhyML with 4, 12 or 25 rate categories for gamma-distribution, estimating the shape parameter (α) and using the median or mean. 100 parametric bootstraps were generated with INDELible using the estimated GTR model. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or under-represented splits;
sp = amount of splits detected as over- or under-represented;
dif = number of sites which represent all over- or under-represented splits;
green cells = no over- or underrepresentation;
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Simulated: continuous Γ − Estimated: Γ														
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
50_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	17	1	33	2	13	0	0	1	5	1	123
		median	4	46	2	255	1	6	1	209	3	17	1	236
	12	mean	1	7	1	34	0	0	0	0	1	5	0	0
		median	2	14	0	0	0	0	0	0	2	11	0	0
	25	mean	5	51	0	0	0	0	0	0	2	12	0	0
		median	4	44	1	33	1	7	0	0	1	7	0	0
50_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	25	1	92	6	39	2	101	3	24	1	107
		median	7	53	1	209	6	39	1	208	6	41	1	227
	12	mean	4	34	1	11	4	26	1	14	4	31	0	0
		median	6	45	0	0	6	37	1	13	7	47	0	0
	25	mean	1	11	1	11	3	19	1	13	5	33	0	0
		median	8	59	1	10	5	35	1	14	4	29	0	0
51_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	397	2	245	7	360	3	262	7	395	2	237
		median	4	394	1	335	10	384	3	382	5	384	2	342
	12	mean	4	410	1	139	5	354	4	185	6	397	2	160
		median	5	411	4	220	6	357	2	187	6	376	2	188
	25	mean	5	417	1	138	7	368	1	123	5	389	1	137
		median	4	403	1	148	5	355	2	165	9	417	1	156
52_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	40	589	2	65	29	508	1	9	34	538	3	77
		median	39	577	4	175	34	555	1	82	37	563	4	154
	12	mean	39	590	2	63	32	532	1	10	30	508	2	43
		median	38	543	2	63	28	453	1	9	28	491	3	54
	25	mean	34	515	3	70	28	455	1	9	29	507	2	45
		median	41	589	3	87	25	420	1	9	33	525	3	72
60_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	10	1	136	1	6	0	0	3	62	1	128
		median	1	5	1	258	1	36	1	251	5	101	1	262
	12	mean	1	5	0	0	2	11	0	0	2	15	0	0
		median	0	0	0	0	2	11	0	0	3	21	0	0
	25	mean	1	6	0	0	0	0	0	0	2	15	0	0
		median	1	5	0	0	2	11	0	0	1	6	0	0

continued on next page...

			Simulated: continuous Γ – Estimated: Γ											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
60_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	5	1	135	1	31	1	129	1	8	1	124
		median	0	0	1	251	1	22	1	245	4	73	1	256
	12	mean	3	19	0	0	0	0	0	0	1	5	0	0
		median	1	5	0	0	2	49	0	0	1	7	0	0
	25	mean	0	0	0	0	0	0	0	0	2	12	0	0
		median	2	13	0	0	1	25	0	0	1	4	0	0
61_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	7	92	0	0	7	69	0	0	4	47	0	0
		median	6	87	1	154	6	62	1	147	3	43	0	0
	12	mean	8	107	0	0	8	82	0	0	3	44	0	0
		median	4	45	0	0	6	51	0	0	3	18	0	0
	25	mean	5	52	0	0	5	57	0	0	1	8	0	0
		median	6	58	0	0	7	71	0	0	3	27	0	0
62_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	64	0	0	2	33	0	0	2	48	0	0
		median	5	59	0	0	3	53	1	23	2	47	0	0
	12	mean	5	59	0	0	1	11	0	0	2	47	0	0
		median	5	57	0	0	3	41	0	0	2	47	0	0
	25	mean	6	67	1	30	3	40	0	0	2	46	0	0
		median	5	57	0	0	3	41	0	0	2	47	0	0
70_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	24	1	113	4	26	1	110	5	39	1	116
		median	3	23	1	222	7	41	1	216	6	42	1	229
	12	mean	4	28	0	0	6	31	0	0	5	38	0	0
		median	2	11	1	23	3	20	0	0	4	24	0	0
	25	mean	2	18	0	0	4	27	0	0	6	48	0	0
		median	3	23	0	0	3	16	0	0	8	60	0	0
70_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	36	1	113	4	23	1	103	3	17	0	0
		median	6	58	1	225	3	17	1	216	3	16	1	216
	12	mean	3	34	0	0	2	11	0	0	5	30	0	0
		median	8	69	0	0	4	25	0	0	3	17	0	0
	25	mean	6	46	0	0	2	11	0	0	3	15	0	0
		median	5	48	0	0	5	30	0	0	3	16	0	0
71_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	51	441	3	46	48	493	1	70	51	483	2	84
		median	53	502	4	194	55	556	4	259	57	545	3	227
	12	mean	60	551	4	53	46	483	2	97	54	516	3	90
		median	50	485	4	60	45	466	3	106	55	516	1	60
	25	mean	55	502	1	28	44	463	3	104	47	466	1	57
		median	62	567	1	28	49	505	3	120	55	528	2	84
72_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	28	391	5	97	31	422	7	184	28	511	3	97
		median	31	437	5	160	36	453	6	162	29	520	7	203
	12	mean	29	410	4	74	32	427	5	127	25	486	6	122
		median	21	323	3	65	33	404	6	153	26	489	3	76
	25	mean	24	373	5	77	33	474	7	143	26	471	4	101
		median	26	395	3	63	32	484	4	74	28	485	6	137
80_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	5	2	145	3	18	1	121	3	17	1	131
		median	2	12	2	249	3	30	1	212	3	16	2	235
	12	mean	0	0	1	17	4	22	0	0	2	10	0	0
		median	2	12	2	31	4	24	0	0	2	11	0	0
	25	mean	0	0	1	17	5	27	0	0	2	10	0	0
		median	1	5	1	16	1	7	0	0	2	11	0	0

continued on next page...

			Simulated: continuous Γ – Estimated: Γ											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
80_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	0	0	1	113	8	51	1	112	1	10	1	123
		median	1	6	2	237	7	43	1	209	3	21	1	213
	12	mean	3	25	0	0	3	19	0	0	5	52	0	0
		median	1	6	0	0	7	45	0	0	1	10	0	0
	25	mean	4	32	0	0	3	19	0	0	5	52	0	0
		median	3	17	0	0	5	31	0	0	1	9	0	0
81_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	45	594	5	100	50	609	8	151	39	537	7	124
		median	51	649	4	158	46	588	10	234	42	577	6	192
	12	mean	48	649	4	77	46	589	5	85	36	520	5	101
		median	45	610	4	80	47	602	7	129	39	549	5	95
	25	mean	44	614	4	77	46	567	5	94	39	541	4	91
		median	51	652	4	79	52	611	5	104	41	564	5	106
82_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	39	1023	13	412	45	1082	14	494	32	939	17	558
		median	38	998	12	516	49	1112	12	582	36	926	16	641
	12	mean	37	1017	17	416	43	1080	13	429	35	993	15	455
		median	33	978	15	415	43	1081	12	418	32	936	17	514
	25	mean	39	1027	13	298	43	1083	13	416	31	928	17	492
		median	34	993	13	360	44	1085	13	433	32	975	17	510
83_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	41	1005	10	334	45	1015	13	455	45	1019	9	302
		median	45	1052	13	482	45	1059	13	580	38	965	9	424
	12	mean	43	1019	10	288	42	1030	9	307	45	1024	10	279
		median	43	1042	9	289	47	1101	10	390	43	1014	9	291
	25	mean	40	1029	12	300	41	1054	11	326	43	1005	10	239
		median	42	1042	11	308	43	1073	10	316	41	991	9	252
85_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	31	794	15	352	39	1067	15	477	40	927	11	424
		median	35	881	16	561	40	1051	13	563	40	921	12	535
	12	mean	33	893	16	347	37	1045	9	314	40	946	10	276
		median	34	891	18	407	41	1071	9	339	35	893	9	342
	25	mean	35	906	16	358	38	1077	12	329	41	939	9	285
		median	36	901	12	323	41	1077	11	376	38	907	11	347
86_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	41	855	9	256	46	974	14	351	38	842	8	175
		median	39	849	13	437	44	958	14	463	40	824	9	334
	12	mean	37	835	9	162	47	993	12	280	36	811	6	132
		median	42	865	10	249	44	968	9	226	39	849	6	128
	25	mean	43	877	9	169	43	954	10	228	37	804	7	131
		median	42	870	8	153	47	1005	10	236	40	847	5	98
88_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	26	402	4	175	29	468	5	139	33	514	7	191
		median	27	437	3	192	29	426	5	188	31	488	6	217
	12	mean	26	505	3	119	28	447	4	85	34	529	6	149
		median	25	467	3	133	29	487	7	149	29	461	7	171
	25	mean	27	492	6	170	25	432	4	83	33	501	5	129
		median	28	474	3	126	25	429	4	82	30	460	5	128
89_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	25	376	2	54	34	412	3	72	25	305	2	50
		median	23	351	4	143	31	392	4	131	26	259	3	120
	12	mean	20	306	2	43	26	342	4	74	27	315	3	73
		median	19	315	2	56	29	388	4	78	26	312	3	77
	25	mean	25	379	2	52	29	402	3	69	22	267	3	61
		median	20	322	2	43	37	459	3	69	29	351	3	74

Table A.8: Results of the test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and continuous modelling gamma distribution. The data was analysed by a maximum likelihood approach using PhyML with 4, 12 or 25 rate categories for gamma-distribution, estimating the shape parameter (α) with a proportion of invariant sites (p_{inv}) and using the median or mean. 100 parametric bootstraps were generated with INDELible using the estimated GTR model. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or under-represented splits;
sp = amount of splits detected as over- or under-represented;
dif = number of sites which represent all over- or under-represented splits;
green cells = no over- or underrepresentation;
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Simulated: continuous Γ − Estimated: Γ+I														
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
50_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	65	1	169	3	19	1	163	3	16	1	163
		median	3	40	1	228	0	0	1	218	3	17	1	235
	12	mean	1	6	0	0	1	22	0	0	1	5	0	0
		median	2	33	0	0	2	24	0	0	1	7	0	0
	25	mean	2	35	1	29	0	0	0	0	2	12	0	0
		median	2	14	1	16	1	7	0	0	1	7	0	0
50_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	70	1	164	2	15	1	171	4	39	1	166
		median	5	34	1	236	4	28	1	231	6	37	1	224
	12	mean	6	51	0	0	2	16	0	0	6	37	0	0
		median	6	52	1	82	1	6	0	0	3	15	0	0
	25	mean	5	47	0	0	5	34	0	0	6	37	0	0
		median	5	45	0	0	5	32	0	0	5	32	0	0
51_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	397	2	245	7	360	3	261	5	371	2	237
		median	5	403	1	336	5	347	2	362	5	384	2	341
	12	mean	4	410	1	141	5	357	4	183	6	397	2	160
		median	2	396	2	190	6	357	2	187	6	375	1	180
	25	mean	4	411	3	173	4	351	3	161	5	389	1	136
		median	4	403	1	148	4	347	3	175	5	388	2	168
52_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	39	594	2	64	31	504	0	0	35	557	3	73
		median	38	582	3	171	28	491	1	98	34	542	3	142
	12	mean	35	538	2	64	30	498	1	9	32	537	2	46
		median	38	543	2	63	26	473	0	0	29	488	2	47
	25	mean	33	511	3	88	28	494	0	0	33	564	2	45
		median	38	570	2	63	34	531	1	9	30	513	1	28
60_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	33	1	122	1	5	0	0	2	48	1	7
		median	1	40	1	228	0	0	1	220	3	87	1	223
	12	mean	1	5	0	0	2	11	0	0	2	15	0	0
		median	0	0	0	0	2	11	0	0	3	21	0	0
	25	mean	2	29	0	0	0	0	0	0	1	6	0	0
		median	1	5	0	0	2	11	0	0	1	6	0	0

continued on next page...

			Simulated: continuous Γ – Estimated: Γ+I											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
60_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	0	0	0	0	2	18	0	0	1	8	0	0
		median	0	0	1	221	7	43	1	228	0	0	1	222
	12	mean	0	0	0	0	3	23	0	0	3	21	0	0
		median	0	0	0	0	2	14	0	0	2	13	0	0
	25	mean	0	0	0	0	4	27	0	0	1	8	0	0
		median	0	0	0	0	3	24	0	0	2	12	0	0
61_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	6	97	0	0	6	63	0	0	5	47	0	0
		median	6	87	1	154	6	69	1	145	5	68	1	155
	12	mean	4	45	0	0	6	68	0	0	4	33	0	0
		median	9	115	0	0	3	30	0	0	3	47	0	0
	25	mean	5	52	0	0	6	64	0	0	3	28	0	0
		median	6	60	0	0	4	49	0	0	4	59	0	0
62_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	52	0	0	2	33	0	0	2	47	0	0
		median	5	60	0	0	4	75	0	0	2	47	0	0
	12	mean	4	52	0	0	4	74	1	24	3	54	0	0
		median	4	50	0	0	3	42	0	0	2	46	0	0
	25	mean	5	60	0	0	3	40	1	26	3	53	0	0
		median	3	44	1	29	4	72	1	26	4	59	0	0
70_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	48	1	169	3	35	1	168	6	48	1	168
		median	5	40	1	228	5	30	1	218	7	51	1	231
	12	mean	2	11	0	0	2	11	0	0	5	36	0	0
		median	3	23	0	0	3	20	0	0	4	24	0	0
	25	mean	4	28	0	0	5	31	0	0	6	43	0	0
		median	3	29	0	0	3	16	0	0	8	60	0	0
70_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	43	1	162	8	54	1	165	2	15	1	173
		median	2	28	1	225	7	54	1	206	6	37	1	224
	12	mean	1	14	0	0	8	59	0	0	3	19	0	0
		median	4	37	0	0	5	44	1	23	1	6	1	8
	25	mean	5	48	0	0	10	88	0	0	2	15	1	8
		median	2	27	0	0	7	45	0	0	1	5	0	0
71_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	53	535	3	139	53	524	3	188	53	514	5	200
		median	52	500	1	195	54	563	2	250	54	525	3	271
	12	mean	47	471	1	8	46	485	1	61	51	498	1	50
		median	57	558	0	0	45	477	1	63	53	503	2	58
	25	mean	54	481	1	7	50	489	2	87	45	430	1	52
		median	53	515	3	25	46	475	1	64	48	462	2	57
72_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	31	436	2	32	33	432	8	140	29	508	3	29
		median	28	414	4	147	37	482	7	165	25	458	3	153
	12	mean	24	396	3	65	32	433	5	97	25	461	4	60
		median	25	379	3	66	33	417	8	152	24	460	5	98
	25	mean	28	413	4	59	32	480	5	96	26	518	2	63
		median	28	407	4	73	27	390	5	106	27	537	4	99
80_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	46	1	201	4	53	2	205	3	15	1	191
		median	1	7	1	261	3	17	1	253	3	16	2	234
	12	mean	0	0	0	0	2	13	0	0	2	10	0	0
		median	2	11	1	16	3	17	0	0	2	11	0	0
	25	mean	1	5	1	16	3	18	0	0	2	10	0	0
		median	0	0	1	17	4	21	0	0	2	11	0	0

continued on next page...

			Simulated: continuous Γ – Estimated: Γ+I											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
80_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	31	1	199	6	40	1	202	6	39	2	213
		median	2	13	1	243	3	18	2	248	4	25	1	244
	12	mean	2	15	0	0	6	39	0	0	5	31	0	0
		median	1	7	0	0	3	25	1	20	3	18	1	8
	25	mean	2	15	0	0	6	39	0	0	2	13	0	0
		median	2	17	0	0	3	18	1	21	5	34	0	0
81_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	48	636	5	94	52	640	8	169	38	570	8	170
		median	51	676	6	217	52	648	5	204	43	590	5	204
	12	mean	42	582	5	90	49	626	4	74	42	586	7	112
		median	43	626	4	80	51	614	4	94	39	556	5	108
	25	mean	54	689	7	114	49	626	4	59	39	561	4	79
		median	49	643	5	93	46	611	8	136	44	598	6	109
82_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	36	1012	13	452	48	1111	13	490	34	961	17	553
		median	36	1003	15	585	42	1056	12	576	36	958	16	650
	12	mean	33	993	15	377	43	1078	14	435	36	963	14	436
		median	37	1008	14	407	43	1081	12	418	35	959	17	519
	25	mean	33	983	14	384	42	1072	13	408	39	1018	15	413
		median	34	992	13	360	45	1091	12	356	34	978	14	449
83_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	42	1055	11	366	41	1087	9	405	44	1029	8	286
		median	44	1056	11	451	44	1054	14	582	42	1006	9	414
	12	mean	40	1039	10	231	44	1084	10	297	42	1005	9	269
		median	40	1002	12	338	47	1103	10	390	42	1004	11	322
	25	mean	44	1057	8	210	40	1029	11	326	44	1024	9	213
		median	43	1052	10	275	43	1068	15	430	41	994	10	288
85_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	33	893	15	419	36	1036	11	425	37	903	10	399
		median	33	878	15	508	37	1025	10	512	39	914	10	487
	12	mean	34	902	11	260	38	1069	14	408	38	926	11	325
		median	36	898	15	363	38	1058	11	406	38	911	12	414
	25	mean	33	890	15	352	36	1050	9	265	33	861	11	325
		median	33	887	16	371	36	1058	11	367	41	940	10	315
86_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	39	821	8	245	45	1008	9	294	36	818	6	210
		median	41	869	10	364	47	993	9	371	37	845	6	290
	12	mean	39	838	7	151	44	971	11	256	38	823	6	120
		median	41	827	11	251	48	1012	12	250	36	803	5	103
	25	mean	41	847	9	176	48	1012	8	184	38	841	7	140
		median	40	810	6	114	50	1026	10	237	38	833	4	87
88_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	28	456	5	189	31	477	4	142	35	530	7	180
		median	25	432	4	224	28	461	3	173	33	508	8	260
	12	mean	25	425	4	143	26	462	5	84	32	525	6	149
		median	28	454	4	158	31	486	4	121	33	498	5	146
	25	mean	27	465	4	143	31	510	3	63	32	527	4	105
		median	31	524	4	150	28	503	3	63	34	507	8	177
89_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	25	339	3	100	30	407	2	32	25	265	4	87
		median	23	314	4	146	38	460	4	142	30	334	5	164
	12	mean	27	398	3	64	32	391	4	74	27	314	1	25
		median	22	324	1	30	33	431	3	71	23	290	3	76
	25	mean	23	307	2	53	31	399	3	69	26	321	3	61
		median	19	290	3	64	33	424	3	69	26	315	2	36

Table A.9: Results of the test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and continuous Γ+I modelling. The data was analysed by a maximum likelihood approach using PhyML with 4, 12 or 25 rate categories for gamma-distribution, estimating the shape parameter (α) with a proportion of invariant sites (p_{inv}) and using the median or mean. 100 parametric bootstraps were generated with INDELible using the estimated GTR model. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or under-represented splits;
sp = amount of splits detected as over- or under-represented;
dif = number of sites which represent all over- or under-represented splits;
green cells = no over- or underrepresentation;
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Simulated: continuous Γ+I – Estimated: Γ

	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
50_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	241	9	318	8	231	10	366	4	167	6	279
		median	2	10	7	245	8	58	12	370	4	25	5	228
	12	mean	0	0	3	118	10	186	5	196	5	35	4	177
		median	1	7	6	188	8	58	7	246	7	48	2	134
	25	mean	6	156	5	159	9	202	5	196	3	134	3	147
		median	2	10	5	165	8	57	5	201	3	22	1	102
50_s	reference set		5	34	1	13	0	0	0	0	0	0	0	0
	4	mean	7	162	11	342	6	198	11	334	4	155	12	405
		median	6	37	7	234	7	44	9	281	7	45	7	266
	12	mean	7	167	6	194	3	22	7	196	7	45	6	199
		median	6	36	4	154	5	33	8	224	6	38	4	165
	25	mean	7	45	3	140	5	155	9	229	7	160	3	137
		median	6	36	3	139	4	27	10	254	4	27	4	162
51_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	8	511	8	303	5	537	10	334	2	474	12	340
		median	9	531	7	376	5	542	11	416	4	490	12	425
	12	mean	10	545	6	134	6	563	7	130	2	491	12	199
		median	9	530	8	272	4	537	9	294	4	503	11	283
	25	mean	10	551	9	185	6	565	12	212	4	513	7	133
		median	11	553	7	156	7	566	8	141	4	509	8	144
52_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	44	788	8	234	44	880	7	237	46	867	14	307
		median	44	802	7	290	48	916	6	280	54	957	9	359
	12	mean	43	779	9	235	38	818	6	203	44	839	13	288
		median	46	805	6	180	38	840	6	199	49	893	9	247
	25	mean	44	780	7	156	43	875	6	183	45	831	8	241
		median	44	786	8	233	41	791	6	189	48	892	10	261
60_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	4	1	147	5	40	1	141	2	11	1	159
		median	4	23	1	300	7	51	1	294	6	71	1	313
	12	mean	1	6	0	0	3	18	0	0	2	12	0	0
		median	0	0	1	107	5	41	0	0	2	12	0	0
	25	mean	0	0	0	0	3	29	0	0	2	10	0	0
		median	1	6	0	0	4	27	0	0	1	5	0	0

continued on next page...

			Simulated: continuous Γ+I – Estimated: Γ											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
60_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	26	1	142	2	9	1	147	2	15	1	136
		median	3	26	1	303	4	65	1	304	6	59	1	294
	12	mean	2	21	1	22	2	14	0	0	1	5	0	0
		median	3	26	1	101	4	40	0	0	1	6	1	115
	25	mean	2	21	0	0	1	6	0	0	2	13	0	0
		median	5	37	0	0	1	6	0	0	2	14	0	0
61_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	12	150	0	0	14	158	0	0	14	127	0	0
		median	13	181	1	161	17	213	1	170	12	124	1	174
	12	mean	11	144	0	0	15	154	0	0	13	120	0	0
		median	10	112	0	0	13	180	0	0	12	110	0	0
	25	mean	11	124	0	0	15	167	0	0	12	113	0	0
		median	10	122	0	0	11	135	0	0	15	147	0	0
62_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	7	75	0	0	8	115	0	0	9	148	0	0
		median	6	62	0	0	8	114	0	0	10	129	0	0
	12	mean	6	69	0	0	10	129	0	0	8	130	0	0
		median	6	67	0	0	9	121	0	0	9	120	0	0
	25	mean	6	67	0	0	10	127	0	0	7	108	0	0
		median	7	74	0	0	8	114	0	0	7	109	0	0
70_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	5	246	13	371	4	296	9	271	3	179	10	267
		median	4	33	8	217	3	17	10	262	3	17	8	212
	12	mean	7	187	8	200	5	160	8	202	6	160	6	133
		median	4	31	7	184	3	16	9	220	3	18	8	175
	25	mean	6	178	5	134	5	171	7	180	4	155	6	139
		median	3	23	5	137	2	16	9	214	3	17	7	161
70_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	208	12	331	5	267	12	350	5	185	8	237
		median	1	5	9	254	3	20	8	228	6	42	5	147
	12	mean	3	152	5	147	8	180	8	185	7	160	6	157
		median	1	5	9	227	7	43	7	180	7	47	4	117
	25	mean	3	15	6	159	6	38	6	147	7	48	4	115
		median	2	9	6	168	7	41	6	152	4	28	4	106
71_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	108	1148	6	274	117	1327	9	368	97	1096	7	282
		median	113	1223	5	236	114	1195	8	335	112	1244	9	397
	12	mean	109	1335	6	255	110	1369	8	326	115	1374	8	275
		median	108	1290	5	231	108	1292	8	328	106	1269	9	300
	25	mean	101	1296	7	264	109	1364	7	305	106	1312	7	269
		median	101	1257	6	262	107	1325	7	304	109	1317	6	245
72_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	44	845	10	314	44	884	9	277	45	901	8	245
		median	50	909	12	429	45	908	9	329	50	956	8	259
	12	mean	48	895	11	321	42	942	6	219	50	974	8	246
		median	50	918	10	303	39	849	7	237	48	928	8	240
	25	mean	49	988	9	297	44	937	8	239	44	948	7	233
		median	48	902	11	342	44	932	11	298	50	932	6	204
80_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	2	224	9	266	3	262	15	427	5	222	13	422
		median	1	7	9	273	2	13	12	353	4	27	10	332
	12	mean	4	28	4	126	3	152	6	166	3	161	8	229
		median	4	25	5	152	2	11	10	250	3	22	11	302
	25	mean	2	13	4	124	4	176	11	255	3	164	8	228
		median	3	20	5	150	4	25	10	251	2	12	9	266

continued on next page...

			Simulated: continuous Γ+I – Estimated: Γ											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
80_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	225	9	277	3	225	10	363	1	250	12	395
		median	1	9	8	256	2	11	10	352	0	0	11	361
	12	mean	4	145	2	97	4	24	7	245	1	143	7	233
		median	2	15	3	122	3	16	5	213	0	0	9	271
	25	mean	4	154	1	73	3	18	6	224	2	169	6	186
		median	2	15	3	117	2	12	5	205	2	10	6	192
81_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	94	1788	19	644	90	1810	18	604	93	1718	18	606
		median	86	1612	19	627	95	1751	17	578	91	1651	16	525
	12	mean	86	1798	18	611	90	1864	20	599	92	1776	17	565
		median	91	1778	20	641	93	1842	16	545	93	1744	17	553
	25	mean	82	1754	21	632	93	1892	17	522	92	1778	17	553
		median	92	1842	19	617	91	1873	15	534	93	1751	21	623
82_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	61	1666	20	930	51	1642	21	921	62	1743	20	933
		median	56	1651	18	1054	54	1681	23	1105	66	1791	21	1092
	12	mean	53	1604	20	903	52	1632	24	925	58	1715	22	933
		median	61	1662	20	915	51	1677	23	912	61	1748	20	919
	25	mean	53	1604	19	891	54	1690	21	888	56	1698	22	929
		median	56	1631	18	890	53	1677	23	917	55	1695	22	927
83_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	64	1568	18	789	62	1552	17	742	66	1452	16	714
		median	71	1622	18	755	68	1583	17	852	75	1546	16	815
	12	mean	69	1747	16	724	69	1592	16	700	71	1635	16	706
		median	64	1656	18	772	63	1533	15	703	71	1495	17	745
	25	mean	66	1747	18	752	65	1699	15	646	75	1669	16	706
		median	70	1757	17	744	67	1664	16	698	69	1625	16	685
85_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	52	1540	22	894	67	1695	20	929	58	1591	19	906
		median	57	1643	24	1050	67	1726	17	1006	60	1642	19	1023
	12	mean	56	1568	21	866	66	1776	19	894	51	1513	21	898
		median	55	1601	23	889	62	1660	18	893	57	1582	18	878
	25	mean	54	1558	21	860	64	1675	20	880	58	1572	23	916
		median	57	1607	23	890	65	1691	19	891	57	1602	18	873
86_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	49	1562	21	672	53	1601	25	743	52	1573	26	804
		median	51	1594	20	825	54	1633	24	900	51	1585	26	936
	12	mean	48	1546	20	641	52	1582	24	703	51	1556	23	669
		median	52	1580	19	636	56	1636	25	744	52	1566	25	705
	25	mean	50	1573	21	651	50	1582	22	699	52	1559	23	671
		median	50	1561	21	647	51	1566	24	735	49	1509	24	675
88_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	64	1236	12	360	58	1153	13	440	62	1170	12	406
		median	65	1228	12	482	61	1161	11	449	67	1195	13	525
	12	mean	57	1175	12	360	64	1272	14	415	57	1122	9	339
		median	59	1207	13	383	64	1212	14	448	69	1230	11	358
	25	mean	60	1217	12	386	65	1270	11	393	64	1216	11	393
		median	54	1164	11	353	65	1280	12	405	63	1199	11	368
89_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	55	1015	8	364	61	1295	10	462	54	1142	9	353
		median	59	1101	7	329	73	1299	9	455	54	1099	9	426
	12	mean	56	1157	8	362	62	1327	9	449	55	1150	8	339
		median	55	1109	10	371	61	1262	9	453	52	1070	8	337
	25	mean	58	1175	6	296	68	1380	9	454	59	1181	7	315
		median	56	1108	7	321	58	1303	8	442	48	1072	9	335

Table A.10: Results of the test for several simulated datasets (Sim, S50−S89). The data was generated with INDELible (_i) or Seq-gen (_s) using GTR and continuous Γ+I modelling. The data was analysed by a maximum likelihood approach using PhyML with 4, 12 or 25 rate categories for gamma-distribution, estimating the shape parameter (α) with a proportion of invariant sites (p_{inv}) and using the median or mean. 100 parametric bootstraps were generated with INDELible using the estimated GTR model. The whole process was performed using three different seeds 1568746 (15), 444444 (44) and 555555 (55) for monte-carlo simulation of data and bootstraps. The results are listed for every seed.
cat = rate categories for gamma-distribution used in the ML-analysis;
over, under = results for over- or under-represented splits;
sp = amount of splits detected as over- or under-represented;
dif = number of sites which represent all over- or under-represented splits;
green cells = no over- or underrepresentation;
The darker the orange cells, the more over- or underrepresented splits were observed. The darker the blue cells, the higher the deviation observed and expected amount of split occurrence.

Simulated: continuous Γ+I – Estimated: Γ+I

	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
50_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	5	0	0	5	49	1	151	2	30	1	144
		median	0	0	1	173	7	61	1	181	5	49	1	190
	12	mean	1	5	0	0	6	53	0	0	3	34	0	0
		median	2	11	0	0	5	39	0	0	2	14	0	0
	25	mean	1	5	0	0	4	34	0	0	3	36	0	0
		median	0	0	0	0	8	63	0	0	4	26	0	0
50_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	35	1	131	4	38	1	157	3	26	0	0
		median	3	23	1	182	1	9	1	185	4	23	1	155
	12	mean	2	16	0	0	2	17	0	0	5	27	0	0
		median	2	16	0	0	4	27	0	0	3	19	0	0
	25	mean	3	21	0	0	3	19	0	0	5	29	0	0
		median	4	27	0	0	4	29	0	0	5	33	0	0
51_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	10	563	9	340	5	565	13	402	6	522	12	366
		median	9	531	7	376	6	545	12	450	4	490	11	405
	12	mean	10	545	6	136	3	543	10	187	6	519	9	168
		median	9	530	8	272	4	537	10	306	4	503	11	282
	25	mean	10	551	9	185	6	565	12	213	4	513	7	134
		median	11	553	7	156	7	567	7	129	4	509	8	145
52_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	44	837	9	190	40	885	6	231	43	899	12	252
		median	48	882	5	280	42	896	5	288	50	944	11	375
	12	mean	42	797	7	189	38	844	5	162	45	896	11	254
		median	49	877	7	194	45	918	5	163	46	880	9	203
	25	mean	43	789	5	145	37	831	7	185	44	878	8	223
		median	43	815	6	172	38	851	4	128	43	835	11	223
60_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	0	0	0	0	5	38	0	0	3	44	0	0
		median	3	19	1	167	2	20	1	176	1	5	1	171
	12	mean	0	0	0	0	4	33	0	0	0	0	0	0
		median	0	0	0	0	5	40	0	0	3	17	1	8
	25	mean	0	0	0	0	3	27	0	0	3	16	0	0
		median	1	7	0	0	2	16	0	0	4	20	0	0

continued on next page...

			Simulated: continuous Γ+I – Estimated: Γ+I											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
60_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	45	0	0	1	5	0	0	1	8	0	0
		median	1	34	1	160	2	40	1	190	3	44	1	184
	12	mean	2	31	0	0	1	5	0	0	1	7	0	0
		median	1	8	0	0	3	22	0	0	1	7	0	0
	25	mean	3	24	0	0	0	0	0	0	0	0	0	0
		median	2	16	0	0	1	6	0	0	0	0	0	0
61_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	9	95	0	0	15	165	0	0	11	104	0	0
		median	10	152	0	0	13	177	1	175	14	128	1	172
	12	mean	8	103	0	0	12	145	0	0	12	127	0	0
		median	7	82	0	0	13	167	0	0	12	110	0	0
	25	mean	8	116	0	0	15	163	0	0	12	113	0	0
		median	8	87	0	0	16	199	0	0	13	119	0	0
62_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	7	65	0	0	7	107	0	0	10	144	0	0
		median	7	76	0	0	8	117	0	0	9	138	0	0
	12	mean	5	60	0	0	9	121	0	0	9	119	0	0
		median	6	60	0	0	9	120	0	0	8	115	0	0
	25	mean	6	62	0	0	9	119	0	0	9	138	0	0
		median	7	75	0	0	9	120	0	0	11	148	0	0
70_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	1	11	0	0	3	21	1	152	3	28	2	174
		median	2	22	2	180	3	22	1	186	3	27	2	213
	12	mean	2	19	0	0	2	14	0	0	1	15	1	21
		median	4	34	1	7	1	5	0	0	1	5	1	20
	25	mean	3	24	0	0	1	5	0	0	3	25	1	20
		median	4	34	1	7	3	21	0	0	2	20	0	0
70_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	22	1	151	3	75	1	149	2	11	1	153
		median	6	32	2	195	1	6	1	181	2	13	1	184
	12	mean	4	22	0	0	2	32	0	0	3	17	0	0
		median	5	28	0	0	1	6	0	0	3	18	0	0
	25	mean	4	22	0	0	1	6	0	0	1	7	0	0
		median	4	22	0	0	1	4	0	0	2	12	0	0
71_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	105	1186	3	96	108	1214	5	203	101	1138	4	209
		median	103	1142	3	266	108	1211	4	276	104	1211	5	331
	12	mean	97	1067	1	65	108	1196	3	113	99	1168	3	108
		median	101	1078	3	103	101	1117	4	122	106	1209	5	136
	25	mean	99	1114	2	86	105	1139	4	120	109	1223	5	149
		median	99	1036	1	66	104	1162	3	111	97	1120	3	116
72_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	51	1004	10	282	43	930	7	182	44	946	8	199
		median	50	1002	10	418	45	963	7	311	45	953	8	324
	12	mean	47	941	11	294	43	917	8	186	45	919	7	192
		median	50	975	10	281	41	914	9	185	39	904	5	152
	25	mean	47	905	9	265	40	916	6	199	43	912	5	155
		median	48	993	10	292	39	872	10	229	45	951	5	163
80_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	4	43	1	168	5	86	0	0	4	26	1	171
		median	1	20	1	196	2	51	1	199	3	18	1	208
	12	mean	3	29	0	0	3	50	0	0	2	14	0	0
		median	2	16	0	0	3	19	1	17	3	18	1	10
	25	mean	2	12	0	0	2	11	0	0	2	12	0	0
		median	2	13	0	0	2	13	1	17	3	20	0	0

continued on next page...

			Simulated: continuous Γ+I – Estimated: Γ+I											
	Analysis options		1568746				444444				555555			
			over		under		over		under		over		under	
Sim	cat		sp	dif	sp	dif	sp	dif	sp	dif	sp	dif	sp	dif
80_s	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	3	19	1	169	5	52	1	176	3	37	1	177
		median	2	11	1	187	3	18	1	215	3	37	1	217
	12	mean	3	18	1	9	3	29	0	0	3	29	0	0
		median	2	10	0	0	1	6	0	0	4	39	0	0
	25	mean	5	28	0	0	1	15	0	0	3	33	0	0
		median	0	0	1	10	1	6	0	0	1	9	0	0
81_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	87	1720	15	367	89	1731	14	388	92	1731	15	364
		median	85	1661	16	516	91	1743	13	460	90	1714	15	544
	12	mean	85	1658	16	364	88	1718	14	323	88	1651	17	394
		median	83	1629	15	359	91	1735	13	327	90	1674	15	377
	25	mean	84	1621	14	347	90	1744	14	332	91	1681	17	389
		median	86	1699	16	346	88	1719	12	321	90	1658	17	395
82_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	55	1720	21	910	51	1706	21	931	57	1782	24	967
		median	54	1717	21	1039	53	1771	24	1070	61	1836	21	1055
	12	mean	52	1663	20	824	52	1725	25	867	57	1739	20	843
		median	55	1670	21	854	53	1713	20	824	60	1769	22	870
	25	mean	53	1670	19	794	53	1714	22	851	61	1781	20	844
		median	55	1674	21	843	53	1713	23	861	56	1730	21	868
83_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	66	1680	14	535	62	1691	13	554	68	1604	13	520
		median	71	1750	13	630	67	1720	11	631	76	1680	12	601
	12	mean	64	1658	14	473	63	1680	12	457	63	1557	14	440
		median	66	1674	13	470	65	1681	13	492	67	1571	14	473
	25	mean	66	1673	14	478	69	1694	11	412	64	1559	15	479
		median	67	1682	13	442	64	1661	14	501	70	1587	13	471
85_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	50	1669	25	884	64	1783	20	846	51	1650	17	811
		median	57	1705	20	914	64	1796	19	963	54	1680	20	948
	12	mean	56	1672	23	782	60	1716	18	757	55	1652	21	795
		median	57	1660	23	789	61	1735	20	783	53	1640	20	801
	25	mean	51	1629	26	814	64	1759	23	817	57	1667	20	786
		median	54	1639	23	805	68	1775	19	789	54	1643	23	807
86_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	47	1630	21	709	53	1710	25	795	55	1658	24	772
		median	52	1685	21	816	59	1764	23	886	49	1610	26	909
	12	mean	46	1561	19	589	53	1621	22	662	49	1558	25	693
		median	47	1564	20	617	53	1645	23	687	50	1560	23	683
	25	mean	49	1595	22	635	54	1628	24	690	51	1566	24	680
		median	50	1591	21	621	53	1627	24	708	54	1567	25	711
88_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	63	1304	9	349	60	1227	10	306	66	1298	9	352
		median	58	1253	13	499	59	1227	10	411	70	1362	11	498
	12	mean	58	1220	9	266	61	1214	9	285	60	1272	10	308
		median	62	1254	12	315	62	1225	8	262	62	1289	10	313
	25	mean	57	1244	12	312	59	1194	9	282	68	1316	11	319
		median	61	1255	11	312	59	1202	11	320	62	1226	12	347
89_i	reference set		0	0	0	0	0	0	0	0	0	0	0	0
	4	mean	55	1181	7	329	58	1338	10	371	62	1288	7	297
		median	56	1166	7	402	60	1323	9	507	60	1231	7	425
	12	mean	57	1161	8	275	65	1348	8	331	54	1207	10	309
		median	55	1158	6	250	65	1365	9	377	58	1183	9	317
	25	mean	58	1186	7	276	61	1329	8	356	50	1165	8	291
		median	52	1158	6	251	68	1370	7	316	58	1194	11	338

A.3. SAMS GUI Manual

Zoologisches Forschungsmuseum
Alexander Koenig

SAMS tutorial:

SAMS

Splits Analysis Methods

version 1.4.3 beta

Sandra A. Meid,
Christoph Mayer and
Wolfgang Wägele

2012

Contents

1 Introduction

SAMS is an interactive analysis program for molecular sequence data. It implements several routines which, for a given set of aligned DNA sequences, estimate the phylogenetic signal present in the data that supports or contradicts putative splits, i.e. internal branches in putative phylogenetic trees. With this information it is possible to visualize the information content of the data set and the signal to noise relationship.

SAMS is based on ideas of Johann Wolfgang Wägele which have been published earlier (see references [1-6] below). These ideas have been refined and extended by Christoph Mayer in the development of the SAMS analysis program. Sandra Meid developed the SAMS GUI and the compatibility test methods implemented within the GUI.

1. Wägele JW, Rödding F: A priori estimation of phylogenetic information conserved in aligned sequences. Mol Phyl Evol 1998, 9: 358-365.

2. Wägele JW, Erikson T, Lockhart P, Misof B: The Ecdysozoa: artifact or monophylum? J Zool Syst Evol Res 1999, 37: 211-223.

3. Wägele JW: Foundations of phylogenetic systematics. Munich: Verlag Dr. F. Pfeil; 2005.

4. Wägele JW, Rödding F: Origin and phylogeny of metazoans as reconstructed with rDNA sequences. Progr Mol Subcell Biol 1998, 21:45-70.

5. Johann Wolfgang Wägele and Christoph Mayer, BMC Evolutionary Biology 2007, 7:147 doi:10.1186/1471-2148

The SAMS-program (version 1.4.3) is a beta-test version which is still under development. The authors are not aware of bugs that would cause the program to obtain incorrect results, but they could exist. This program is distributed in the hope that it will be useful, but of taxa that can be analysed is limited to 256.

Future versions of this program are intended to be released under the Gnu Public License (GPL). Since the current version is a pre-release version, it is not shipped with source code or the GPL.

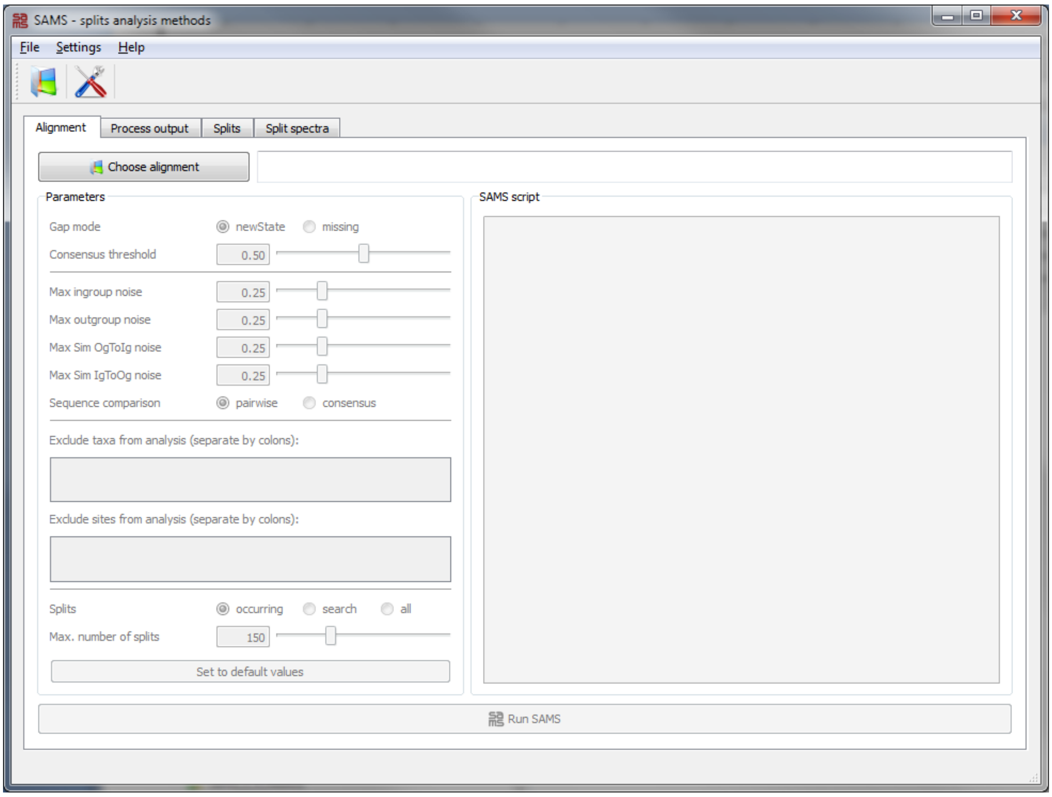
Please report any crashes, bugs, or problems you have with this program to Christoph Mayer or Sandra Meid.

When publishing results obtained via SAMS please cite: Christoph Mayer, Sandra Meid and Johann Wolfgang Wägele, 2012, Zoologisches Forschungsmuseum Alexander Koenig, Bonn, Germany.
https://www.zfmk.de/en/research/research-centres-and-groups/sams

2 Tutorial

2.1 Start SAMS GUI

Start the application. For Windows and Mac OS click on the SAMS executable, for Linux start the sh-script.



Gap mode ● newState ○ missing

Consensus threshold 0.50

Max ingroup noise 0.25

Max outgroup noise 0.25

Max Sim OgToIg noise 0.25

Max Sim IgToOg noise 0.25

Sequence comparison ● pairwise ○ consensus

Exclude taxa from analysis (separate by colons):

Exclude sites from analysis (separate by colons):

Splits ● occurring ○ search ○ all

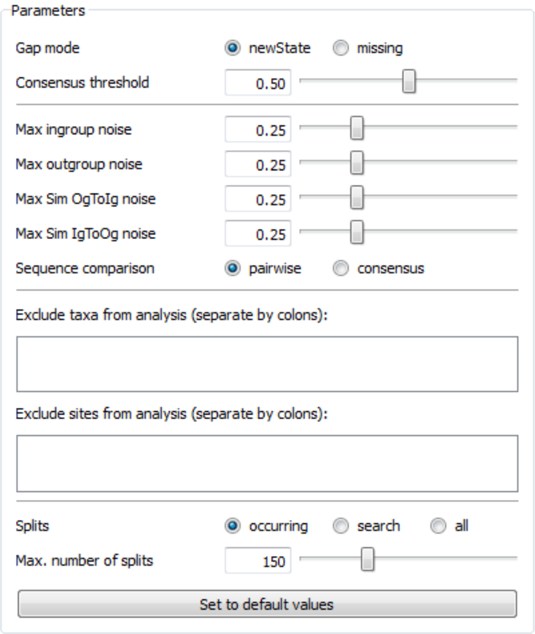
Max. number of splits 150

2.2 Choose alignment

By clicking on the <Choose alignment> button a window appears in which you can choose the file you want to process. SAMS is capable of reading data files in the NEXUS format. It can read all relevant NEXUS-blocks including Taxa-blocks, Data-blocks, Character-blocks, Assumptions-blocks and Tree-blocks. Blocks that are unknown to SAMS are ignored. SAMS defines its own block, the so called SAMS-block, in which the user can write any set of SAMS-commands in order to run the program.

2.3 Choose parameters

Adjust the parameters to your needs. The usage and effect of these parameters are described in detail in Chapter 2.



Parameters

Gap mode ● newState ○ missing

Consensus threshold 0.50

Max ingroup noise 0.25

Max outgroup noise 0.25

Max Sim OgToIg noise 0.25

Max Sim IgToOg noise 0.25

Sequence comparison ● pairwise ○ consensus

Exclude taxa from analysis (separate by colons):

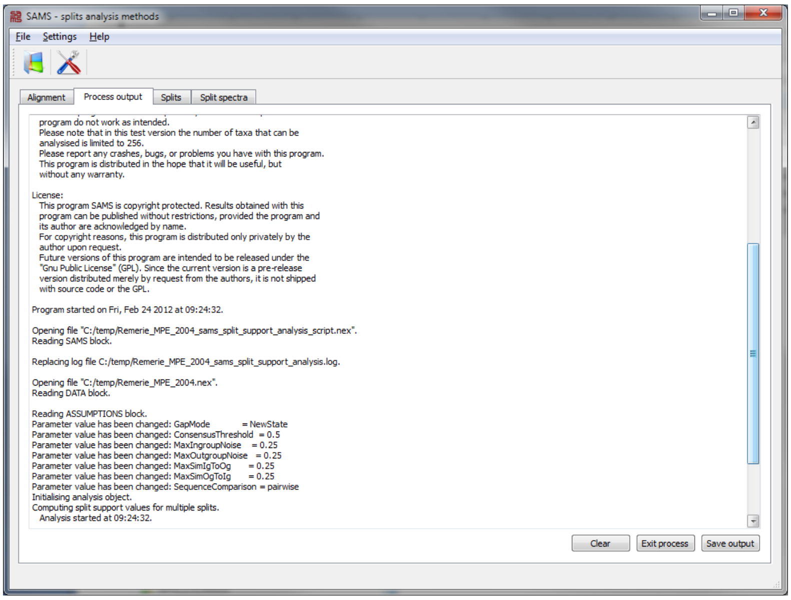
Exclude sites from analysis (separate by colons):

Splits ● occurring ○ search ○ all

Max. number of splits 150

2.4 Run SAMS

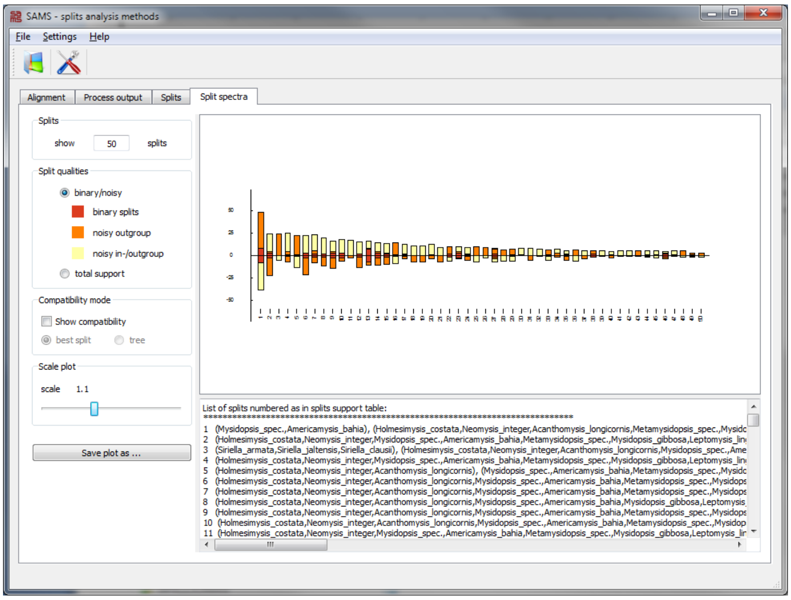
Wait until SAMS has processed the data. It will open the tab <Splits> when finished.



```
  program do not work as intended.
  Please note that in this test version the number of taxa that can be
  analysised is limited to 256.
  Please report any crashes, bugs, or problems you have with this program.
  This program is distributed in the hope that it will be useful, but
  without any warranty.

License:
  This program SAMS is copyright protected. Results obtained with this
  program can be published without restrictions, provided the program and
  its author are acknowledged by name.
  For copyright reasons, this program is distributed only privately by the
  author upon request.
  Future versions of this program are intended to be released under the
  "Gnu Public License" (GPL). Since the current version is a pre-release
  version distributed merely by request from the authors, it is not shipped
  with source code or the GPL.

Program started on Fri, Feb 24 2012 at 09:24:32.

Opening file "C:/temp/Remerie_MPE_2004_sams_split_support_analysis_script.nex".
Reading SAMS block.

Replacing log file C:/temp/Remerie_MPE_2004_sams_split_support_analysis.log.

Opening file "C:/temp/Remerie_MPE_2004.nex".
Reading DATA block.

Reading ASSUMPTIONS block.
Parameter value has been changed: GapMode           = NewState
Parameter value has been changed: ConsensusThreshold  = 0.5
Parameter value has been changed: MaxIngroupNoise    = 0.25
Parameter value has been changed: MaxOutgroupNoise   = 0.25
Parameter value has been changed: MaxSimIgToOg       = 0.25
Parameter value has been changed: MaxSimOgToIg       = 0.25
Parameter value has been changed: SequenceComparison = pairwise
Initialising analysis object.
Computing split support values for multiple splits.
  Analysis started at 09:24:32.
```

2.5 Split spectrum

After the analysis is finished the results are plotted as split spectrum within the tab <Split spectra>.



show 50 splits

● binary/noisy
🟥 binary splits
🟧 noisy outgroup
🟨 noisy in-/outgroup
○ total support

□ Show compatibility
● best split ○ tree

scale 1.1

```
List of splits numbered as in splits support table:
*******************************************************************************
1  (Mysidopsis_spec.,Americamysis_bahia), (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Metamysidopsis_spec.,Mysidc
2  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lin
3  (Siriella_armata,Siriella_jaltensis,Siriella_clausii), (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Mysidopsis_spec.,Ame
4  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lin
5  (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis), (Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidc
6  (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidops
7  (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidops
8  (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Mysidopsis_spec.,Americamysis_bahia,Mysidopsis_gibbosa,Leptomysis_
9  (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidops
10  (Holmesimysis_costata,Neomysis_integer,Acanthomysis_longicornis,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidop
11  (Holmesimysis_costata,Neomysis_integer,Mysidopsis_spec.,Americamysis_bahia,Metamysidopsis_spec.,Mysidopsis_gibbosa,Leptomysis_lir
```

3 SAMS Parameters

3.1 gapMode

default new state

The **gapMode** parameter controls how gaps are treated in molecular sequence data.

newState Gaps are treated as an additional state.

missing Gaps are treated as missing character.

3.2 consensusThreshold

default 0.50, real value in range [0..1]

The value of the **consensusThreshold** parameter is relevant whenever a consensus sequence has to be computed for a set of taxa during an analysis. This is done as follows: For each site, the occurring characters are counted. The most frequent character of a site is chosen as the consensus character, if the character occurs with a proportion larger than or equal to **consensusThreshold** and if no other character occurs with this same proportion. Otherwise the consensus character is set as missing character.

3.3 MaxIngroupNoise

default 0.25, real value in range [0..1]

This parameter influences whether alignment positions are counted as noisy in- and outgroup supporting positions for a certain split. The value indicates the percentage of the ingroup bases which do not have to be identical with the predominant base. In other words, the ingroup is allowed to contain 25% noise by default.

3.4 MaxOutgroupNoise

default 0.25, real value in range [0..1]

This parameter influences whether alignment positions are counted as noisy outgroup and noisy in- and outgroup supporting positions for a certain split. The value indicates the percentage of the predominant ingroup base, which is allowed to occur in the other group, 25% noise by default.

3.5 MaxSimIgToOg

default 0.25, real value in range [0..1]

The supporting positions of every taxon are compared to corresponding positions or the consensus sequence of the other group. The value indicates the percentage of sequence positions within the ingroup, which are allowed to be similar to sequence positions of the outgroup, 25% noise by default.

3.6 MaxSimOgToIg

default 0.25, real value in range [0..1]

The supporting positions of every taxon are compared to corresponding positions or the consensus sequence of the other group. The value indicates the percentage of sequence positions within the outgroup, which are allowed to be similar to sequence positions of the ingroup, 25% noise by default.

3.7 SequenceComparison

default pairwise

The sequences are compared to exclude split support caused by similarity by chance. If the similarity of the sequences is higher than the chosen threshold (MaxSimIgToOg and MaxSimOgToIg), the support of this positions is not counted as support.

pairwise Supporting positions of every taxon are compared to corresponding positions of the other group.

consensus Supporting positions of every taxon are compared to the consensus sequence of the other group.

3.8 Exclude taxa from analysis

Excludes a single taxon or a <set of taxa> from the analysis. The <set of taxa> can be any set of valid taxon names or numbers separated by colons.

3.9 Exclude sites from analysis

Exclude single or a <set of character positions> from the analysis. The <set of character positions> can be any set of valid position names or numbers separated by colons.

3.10 Splits

default occurring

If the number of taxa, here denoted by n, is too large, it becomes impossible to determine supporting positions for all 2^n possible splits. With the splits options it is possible to specify the set of splits to be analysed.

occurring Only the splits that occur in the data are analysed.

all All possible 2^n splits are analysed. Since the number of splits increases exponentially with the number of taxa, this option is only allowed for less than 15 taxa.

search In a first step, all splits are analysed which occur in the data. In a second step, a specified proportion of the highest supported occurring splits is used as starting points for a search, in which taxa are added and removed from the ingroups of splits, with the aim of finding splits not explicitly occurring in the data, but having a high number of supporting positions.

3.11 Max. number of splits

default 150

Number of splits which should be analysed.

3.12 Set to default values

By clicking this button all parameters are restored to default values.

4 Splits spectrum preferences

4.1 Splits

This parameter changes the number of splits, which are plotted. If the number is higher than the number of analysed splits, all available splits are shown.

4.2 Split qualities

Switch between total support and support for three different split qualities.

binary Each position in in- and outgroup have identical bases, bases between both groups are different.

noisy outgroup Taxa of the ingroup have homogeneous bases. The base which is predominant within the ingroup is allowed to occur in other group up to 25% by default.

noisy in- & outgroup Ingroup taxa must have at least 75% character identity and the base which is predominant within the ingroup is allowed to occur in the outgroup up to 25% by default.

4.3 Compatibility mode

Shows which splits are compatible with the best split or all splits fitting together as tree, starting with the best split.

A.4. Additional manuscript prepared during the projects (I)

5

Aspects of Quality and Project Management in Analyses of Large Scale Sequencing Data

Björn M. von Reumont, Sandra Meid and Bernhard Misof
*Zoologisches Forschungsmuseum Alexander Koenig,
Adenauerallee 160, 53113 Bonn,
Germany*

1. Introduction

We describe step-by-step the outline of a project, in which the evolutionary history of pancrustaceans (crustaceans and hexpods) was revisited using molecular methods. It was part of a larger program, the 'Deep Metazoan Phylogeny' priority program of the Deutsche Forschungsgemeinschaft (DFG), wich aimed to reconstruct the metazoan tree of life involving more than 30 subprojects. This chapter should be understood as a backbone, that clarifies important points to plan and to conduct projects in molecular biology, also using next generation sequencing data. The text is divided in four parts: 1) theoretical aspects to projects in molecular biology, 2) the process from the collection of material in the field to the final sequencing, 3) the process from the sequence to the reconstructed topology with a special emphasis on data quality, and 4) the conclusions to prevent pitfalls.

1.1 Fascination and complexity of molecular evolutionary biology

Working in molecular evolution to reconstruct the evolutionary history of organisms is a very fascinating, but also very complex issue. Per definition evolutionary biology, and respectively molecular evolutionary biology, is the division in science, which overlaps and intersects mostly with other areas of natural sciences, like chemistry, physics, informatics, mathematics, bioinformatics, geography but also philosophy and history. Exactly that complexity and intersection creates the fascination and addiction of many scientists to work in that area.

Being on field excursions and collecting specimens in their natural habitats is like travelling back in time into the century and time of classic field biology, geography and history. If once the laboratory part has started, technicial and laboratory skills are demanded, while in parallel the amount of characterized sequences starts to force one to become a sophisticated software user, partly applying bioinformatics knowledge or (the often much faster alternative) cooperating with bioinformaticians. The analyses, interpretation and discussion of the results represent the climax of the project by some (at least) publications in highly respected journals.

1.2 General management strategies applicable for scientific projects in molecular evolution

In general, scientists are highly educated in their specific disciplines, but are often 'freshmen' in managing projects with all involved aspects.

These eventually less developed soft skills can cause an underestimation of possible volume of work and subsequently lead to a massive lack of time, which finally degrades the results and the quality of the scientific project. A rigorous project management as conducted in economics featuring a global, yet detailed intersected time schedule with 'milestones' as anchor points and deadlines (including buffer-time in reserve) as general frame in a project roadmap is mandatory for a solid project. The 'golden triangle' of project management (e.g. Kerzner, 2009; Litke et al., 2010) illustrates interrelations that affect projects and their quality management: A) goals and qualitative results, B) planned time schedule and C) calculated costs. If one edge of that triangle becomes delicate, all could be at risk, and the quality of the project is affected (see figure 1).

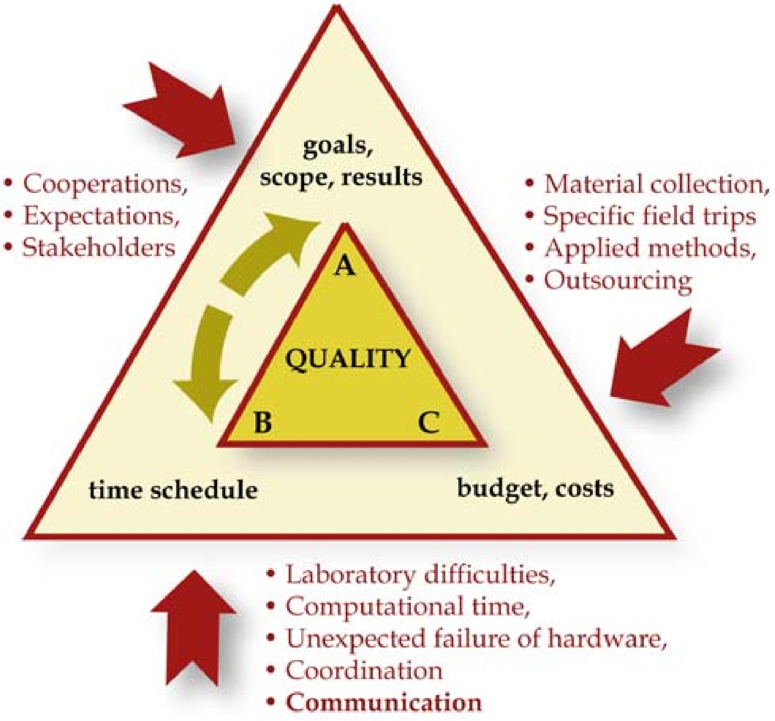


Fig. 1. The golden triangle of project management adapted to molecular projects. The red arrows indicate where the points written outside the second (red) triangle have most impact. However, some points have an impact on more than just one edge. Laboratory difficulties for example cost primarily time, but also stress the budget. If things go wrong (and mostly they unfortunately follow the law of Murphy in the scientific business) goals might also be affected by laboratory difficulties. The core triangle pictures the three main components, which are interwoven. If one edge is affected, the other ones are affected either. A major specification is probably, that A and B generally are more connected with each other in most aspects, while the budget is constant or not directly affected (golden arrows). If e.g. computational analyses of phylogenetic trees do not work or cause difficulties, a delay in the time schedule is created, that primarily affects the results, but not directly the budget

If a larger project is conducted, in which more persons are directly involved or third parties included (e.g. by outsourcing of sequencing to companies, etc.), additional aspects play a veritable role. Who is directly or indirectly involved in and linked to the project? Which interests and influence (negative and positive) have the different persons or parties in the project? All of these involved persons (with different expectations and interests) are stakeholders of the project. In general, a stakeholder analysis in the planning phase is extremely crucial and a standard approach in economics (Weaver, 2007; Freeman, 2010; Litke et al., 2010). Which risks might rise by involved persons? In science, competion between work groups must be considered. Is cooperation possible, which is always to prefer. If no cooperation is feasible, which risks exist subsequently for the project? If third parties are involved by outsourcing of e.g. sequencing, an exact analysis of possible candidate companies and their interests and capability are important (see also additionally paragraph 2.3). Last but not least, if you are a PhD student or postdoc do not forget one very important or even the key stakeholder (Bourne, 2010), the PI or supervisor. What are his interests, which are yours? Is there a risk or conflict you might have to deal with or to solve? What are his expectations? Perhaps an agreement on objectives is necessary. One major factor is an open discussion, regular (scheduled) communication and time for additional, intermediate meetings; also a clearly communicated agreement on objectives avoids difficulties or even disappointment of one or both parties in the project.

The communication strategy is a further key factor (Bourne, 2010), it is important to prevent typical pitfalls like 'just reporting', 'flood of detailed information' and that 'no feedback' is given. See also general principles of communication to transport information (Chapter 1.3.5/1.3.6 in: Wägele, 2005; Bourne, 2010). Communication is quite clearly time consuming, but it pays off. All points of the golden triangle are linked to communication, including budget and quality of results. Communication skills improve the general quality of the project, can save costs and time, and eventually most importantly: control and enhance the motivation of the involved persons.

Several software packages to coordinate communication, interaction and project work exist to provide an effective platform and frame to conduct and coordinate projects. Examples are Teamwork, OpenLab, Italy; Teamlab, Ascensio System (open source); Clarizen (web based); Endeavour software project management, Ezequiel Cuellar (open source). If you are a bioinformatician, the last package might be respectively interesting.

A characteristic of scientific projects is that new open questions and potentially new fields of methodologies are explored. Respectively, if additionally laboratory work is included, the risk to end without any or absolutely unexpected results (latter one might result in the desired nature paper) is part of the scientific business and in general hard to evaluate. That has to be calculated in advance and should be reflected in the time and risk management.

However, there is also a clear difference between projects in economics and science: scientific projects aim in most cases for fundamental and theoretical insights instead for a direct financial benefit of involved parties. Changing and evaluating laboratory methods for example, might be unexpected time consuming, but necessary and can at the same time establish a new state of the art method. Time and space to walk open minded on paths that seem to be ineffective, not suitable or even out of topic at first glance might bring the breakthrough and must be possible. Louis Pasteur (1822-1895) quoted on his accidentally discovery of penicillin, "chance favours the prepared mind", but one condition for this famous quote is, that the scientist needs the (mentally) freedom to meet chance. A too rigid framework and control might hinder that. Contrariwise many scientists focus often too much on details (as being trained for) and loose their track on the overall relations of the project, which provokes a rather high inefficiency. Consequently a compromise between efficiency and creativity/innovation has to be made. This is easy to write, but hard to transfer and to realize, as personally experienced.

2. Project phases from species collection in the field to sequencing

2.1 Collection and fixation of samples in the field – RNAlater or sooner?

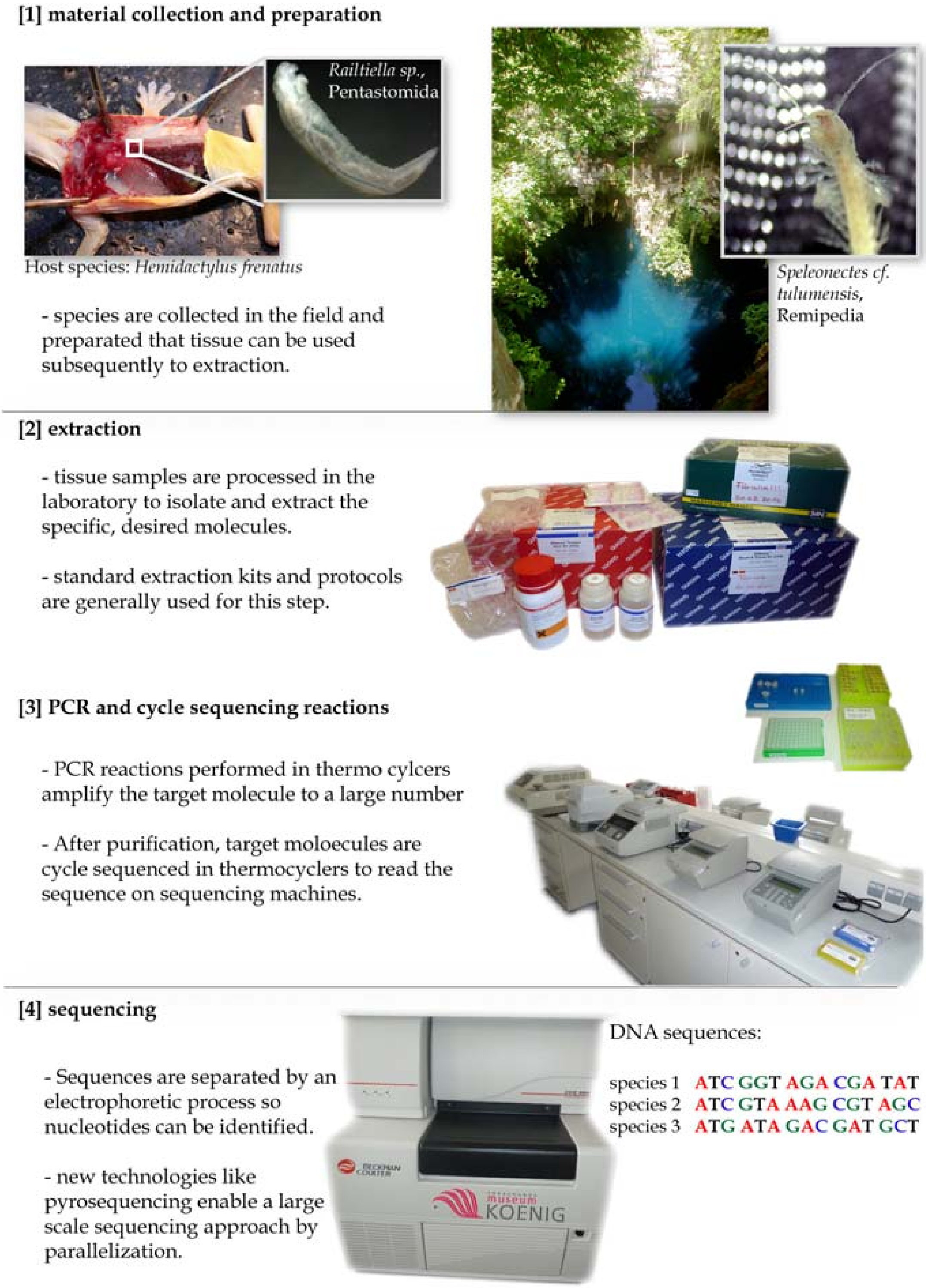
Normally, the planned molecular project starts with the extraction of molecules (DNA or RNA) from specimens (see figure 2) and every true biologist will do his very best to collect and preserve these specimens by himself in the field.

If the specimens or the tissue is preserved in Ethanol for DNA based work, 94% (or higher), ethanol p.a. should be used. This is true for every tissue collected in the field. Despite the rumour, that crustaceans are tricky to sequence in the laboratory, because the aggressive enzymes of the exocrine glands rapidly degrade the DNA, this specific experience was never made working with 94% ethanol p.a.. Working with material collected and sent by colleagues, difficulties appeared and could be linked to the quality (not p.a.) or low concentration of ethanol. Especially material of larger, vessel based expeditions, is obviously often stored in ethanol, which has been diluted due to ethanol shortage during the cruise. If you expect to join an expedition, plan enough quantities of 94% ethanol (and you better hide some of the ethanol in case colleagues did not properly calculate their ethanol contingents, they seem to tend to desperate actions in these situations). Storing the samples in -20 °C probably keeps degradation processes at a low level, but fieldtrip cooling is not obligatory to preserve high quality DNA.

However, cooling plays a veritable role, if you have to collect samples in the field for RNA based analyses. RNA as a single stranded molecule can be degraded very fast (and unfortunately very efficiently) by a group of enzymes, called RNAses. These enzymes are nearly omnipresent in our body including e.g. perspiration liquid. They have to be inhibited by cold temperatures or chemicals (or both) to stop RNA degradation. The best procedure to ensure good quality of RNA samples is consequently to collect the specimen and to extract the RNA immediately. Unfortunately this is in most cases not possible in the field. For example, many groups of crustaceans live in remote habitats.

For example, remipedes live in anchialine cave systems (see figure 2, top right picture) and require cave diving expeditions. They were collected by BMvR on the Yucatan peninsula in Mexico. Even the organization of the cooling chain to freeze the samples directly in the field and to ship them to the laboratory for RNA extraction was not possible: logistic companies that could have shipped the samples in time did not ship dry ice due to regulations of the International Air Transport Association (IATA), In general, the dry ice transportation by airplane is not officially authorized and problematic in some countries. Awareness and integration of such eventual logistic problems are eminent for a realistic project plan and time schedule.

Using RNAlater for RNA isolation is one solution to collect specimens. It is a non toxic, non flammable liquid that can be transported everywhere without any problems (even in airplanes) and it preserves RNA at room temperature at least for 5-7 days (Grotzer et al., 2000; product descriptions of e.g. Qiagen, Applied Biosystems) without loss of quality compared to frozen samples (Grotzer et al., 2000; Mutter et al., 2004; Gorokova, 2005). A closed cooling chain is not mandatory. For preservation of microcrustaceans of zooplankton like copepods, up to a month of storage time is possible without any losses of RNA quality if RNAlater is used (Gorokhova, 2005). Own experiences corroborate this study with samples


Railtiella sp., Pentastomida
Host species: *Hemidactylus frenatus*
Speleonectes cf. tulumensis, Remipedia
DNA sequences:
species 1 ATC GGT AGA CGA TAT
species 2 ATC GTA AAG CGT AGC
species 3 ATG ATA GAC GAT GCT

[1] material collection and preparation

- species are collected in the field and preparated that tissue can be used subsequently to extraction.

[2] extraction

- tissue samples are processed in the laboratory to isolate and extract the specific, desired molecules.

- standard extraction kits and protocols are generally used for this step.

[3] PCR and cycle sequencing reactions

- PCR reactions performed in thermo cylcers amplify the target molecule to a large number

- After purification, target moloecules are cycle sequenced in thermocyclers to read the sequence on sequencing machines.

[4] sequencing

- Sequences are separated by an electrophoretic process so nucleotides can be identified.

- new technologies like pyrosequencing enable a large scale sequencing approach by parallelization.

Fig. 2. Overview of the typical phases within a molecular project that start with material collection in the field and end with the final sequences. The two pictures in the left on top [1] show a dissected house gecko (*Hemidactylus frenatus*), which was parasitized by tongue worms (Pentastomida, small picture) in his lunge tract. On the right, a remote anchialine cave system in Mexico is shown. Within these caves live the enigmatic Remipedia (*Speleonectes cf. tulumensis*) that were collected by cave diving

of different sizes like copepods, ostracods, remipedes, and leptostracans, which were stored at room temperature for up to 14 days after collection (including transportation and shipping time). High temperatures may harm the sample quality despite RNAlater preservation, depending on the general temperature conditions of the expedition area. Good experiences were made with standard fridges (about 4°C), they are easy to organize and the sample is cooled, but not frozen.

RNAlater should have room temperature for preservation of tissue samples to enable a thorough penetration, and the liquid should not be cooled before and directly after preservation of material. Before preservation, tissue has to be cut into little fragments, additionally use a pestle (even some smaller crustaceans have a carapace that has to be cracked) to ensure a fast diffusion of the liquid into the tissue. After a few hours or a day, RNAlater can be moderately cooled. If frozen away after one day, a cooling chain must be guaranteed.

For marine organisms a careful sorting or sample preparation is crucial before the preservation of tissue to prevent larger amounts of salt water to dilute and affect the preservation liquid. In general, RNAlater should be sufficiently added to the sample, about 1:5-10 between sample and RNAlater (according to manufacturer protocols) turned out to be insufficient. Even for smaller specimens 15-25 ml tubes were at least used, depending on the collected numbers.

However, contrary to own good experience with RNAlater, other projects using RNAlater to preserve representatives of evolutionary early hexapod lineages report frustrating results, gaining degraded RNA or only very few EST sequences. As stated, the best method has to be tested for each species group. In that special case the best choice was liquid nitrogen, with all subsequent difficulties in the field. An interesting effect is, that RNAlater perfectly preserves DNA (Gorokova, 2005; Vink et al., 2005), which makes it an ideal alternative to ethanol preservation.

The main goal of many projects in molecular biology is the reconstruction of the evolutionary history of species. In this context so called large-scale next generation sequencing approaches have recently been used applying RNA based sequencing (see paragraph 2.3). The approach aims to randomly sequence expressed genes of a specimen when the tissue or specimen was collected and preserved (‘transcriptome shot’). One quality criterion to achieve a good coverage of different genes is, how fast the specimen was preserved. If the stress level of the specimen was high, a relatively high level of stress response proteins are the consequence, biasing the quantity but also quality of finally sequenced genes. Always ensure that stress is kept to a minimum level for organisms before preservation to guarantee a maximum number of represented genes. Another important method to achieve a maximum intersection of expressed genes is the collection of different larval and/or development stages of an organism to cover possibly different gene expression patterns. If parasitic forms are sampled, like tongue worms, that parasitize the respiratory tract of vertebrates (Pentastomida, see figure 2 top left picture), a careful preparation of the tissue is necessary to prevent contamination by the host tissue.

Collected specimens should carefully be determined before preservation. Additionally, collected and stored voucher species might enable a second identification after sequencing, if unexpected results or difficulties occur. This specific point is often forgotten. An approach to centralize the storage of voucher specimens and DNA including the linked collection and laboratory data is the DNA bank network (Gemeinholzer et al., 2011). This platform provides an efficient and practical solution to access and exchange data and tissue in an extended form, compared to classical accession sheets like in GenBank. This storage allows a general traceability of DNA sequences, and their quality concerning specimen identification and the DNA itself, like concentration, signal strength, electropherogram etc. In most cases this information is missing in published NCBI data (see figure 4).

2.2 Extracting DNA, RNA and subsequent amplification of the molecules

The extraction of DNA or RNA from tissue follows standard protocols and available kits (e.g. Mülhardt, 2008; Sambrook & Russel, 2000). Eventually it is reasonable, to test different kits and protocols to be time efficient.

A fast and specifically tested method is needed to isolate RNA from tissue. Only few studies mainly from the medical/clinical field are published, which show that quality and quantity of RNA yields are dependent on used preservation/isolation method and extraction kits; additionally both parameters can improve using RNAlater (Forster et al., 2008; Hemmrich et al., 2010, see also Gorokhova, 2005). One serious consideration should be outsourcing of RNA extraction and subsequent sequencing. Time is saved if one party or company provides service from extraction to the final sequences, also in cases of difficulties with the samples.

The PCR method is an established method and several specific adaptations exist to ensure the maximum sensitivity to amplify the desired fragments (e.g. Mülhardt, 2008; Palumbi in: Hillis et al., 1996).

Everyone who works in a molecular lab performing PCR knows that this step is the most sensitive and delicate one for possible contamination. Consequently, a rigorous management should be conducted to maintain high standards in working procedures (Mülhardt, 2008; Sambrook & Russel, 2000). The awareness that contamination can happen despite all efforts is important. If that is considered and influences a general risk management, in consequence all sequences, which are finally included in analyses are blasted in a standard procedure. Exactly this step is the last bastion to guarantee as first step the quality of phylogenetic analyses. If a contamination occurred, the contaminated sequences must be identified and excluded (see figure 4 and paragraph 3.1).

2.3 The sequencing process – A typical case for outsourcing

The term phylogenomics was coined by Eisen (1998) and is recently used for analyses including large scale sequencing data and large numbers of genes derived from cDNA libraries (see also Philippe et al., 2005). A new strategy is the sequencing of the 'transcriptome', which represents the set of expressed genes in an organism, that are encoded by mRNA molecules. Most mRNA molecules are tagged by Poly-A tails and thus easily to fish by specific adaptors if total RNA was isolated. These fished mRNA molecules are reverse transcribed in cDNA and finally libraries are reconstructed that represent ideally all expressed genes in an organism. These mRNA fragments are called expressed sequence tags (ESTs) because of their poly-A tail 'tag' (excellent reviews on that topic: Jongeneel, 2000; Bouck & Vision, 2007). With the new technology of pyrosequencing the possibility arose to directly sequence cDNA molecules in a large scale sequencing approach. Pyrosequencing is not based on the principles of the Sanger sequencing with chain termination reactions, but instead on an enzyme cascade, which generates light if deoxynucleotides are added and pyrophosphate is separated. This difference enables a highly miniaturized and parallelized procedure and technique (see figure 3). For more details see Ronaghi, 2001; Shendure et al., 2004; Ellegren, 2008; Hudson, 2008; Petterson et al., 2009; Voelkerding et al., 2009.

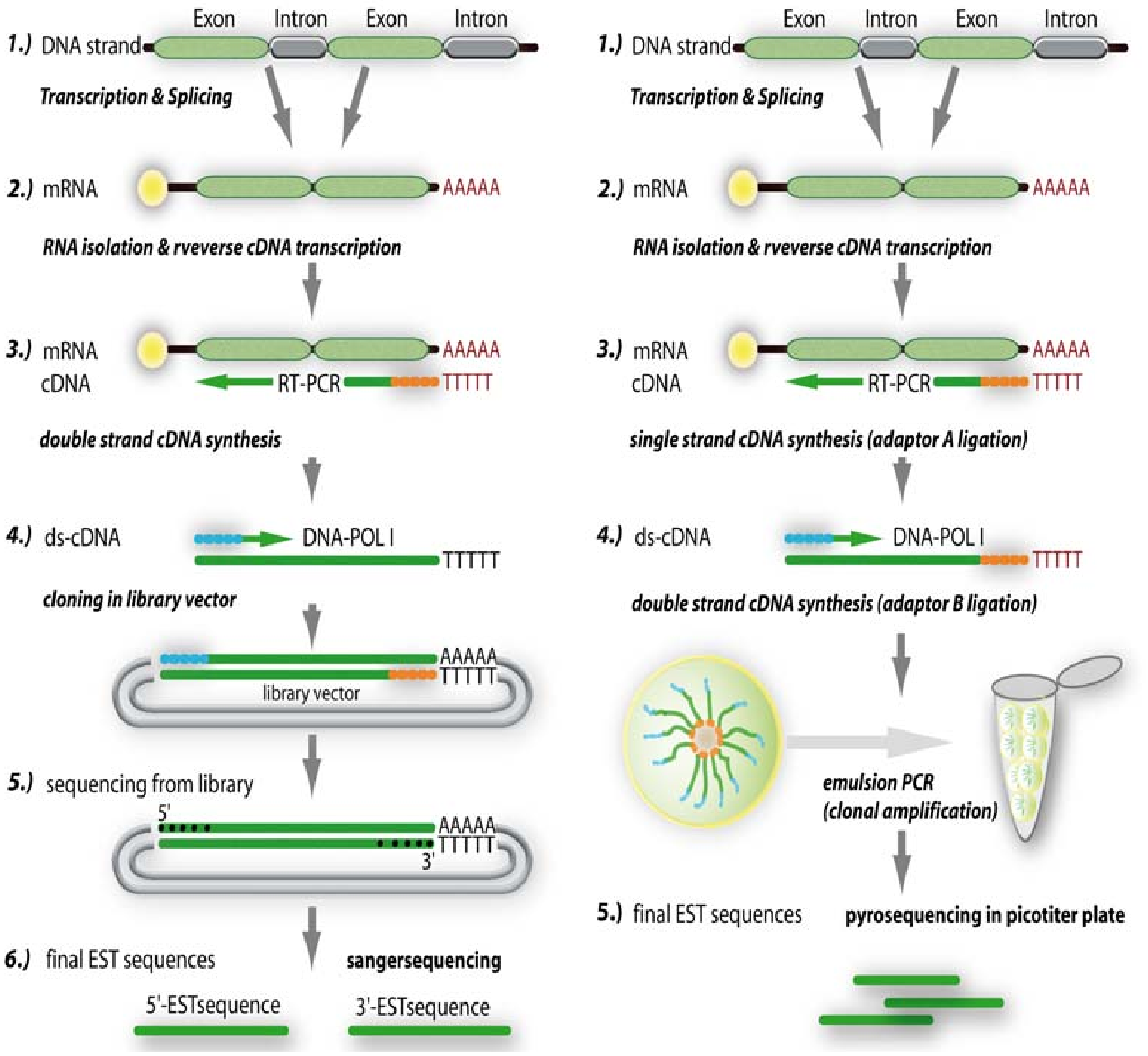


Fig. 3. Differences between standard sanger-sequencing (on the left) and the new pyrosequencing technology (on the right) of next generation sequencing (NGS). Both technologies use mRNA specific target sequences to extract mRNA form the total RNA, which is isolated from tissue. The main difference is that the time and cost intensive step of fragment cloning and sequencing from a subsequently picked library is skipped for pyrosequencing. Depending on the precise technology, double stranded cDNA is generated by an emulsion PCR, in which fragments are amplified in micro compartiments. The sequence fragments are finally transferred on picotiter plates for a massive parallel sequencing

Sequencing is frequently outsourced, which offers a price level that is hard to beat by do-it-yourself sequencing at universities or other research institutions. Focused on large scale or next generation sequencing, some points should be considered. In most companies laboratory procedures and steps are ISO certified ensuring a guaranteed high level of quality and reproducability.

It is a specific quality of molecular biological studies that often unique samples of species with rather unknown evolutionary history are analysed. The collection of these specimens is

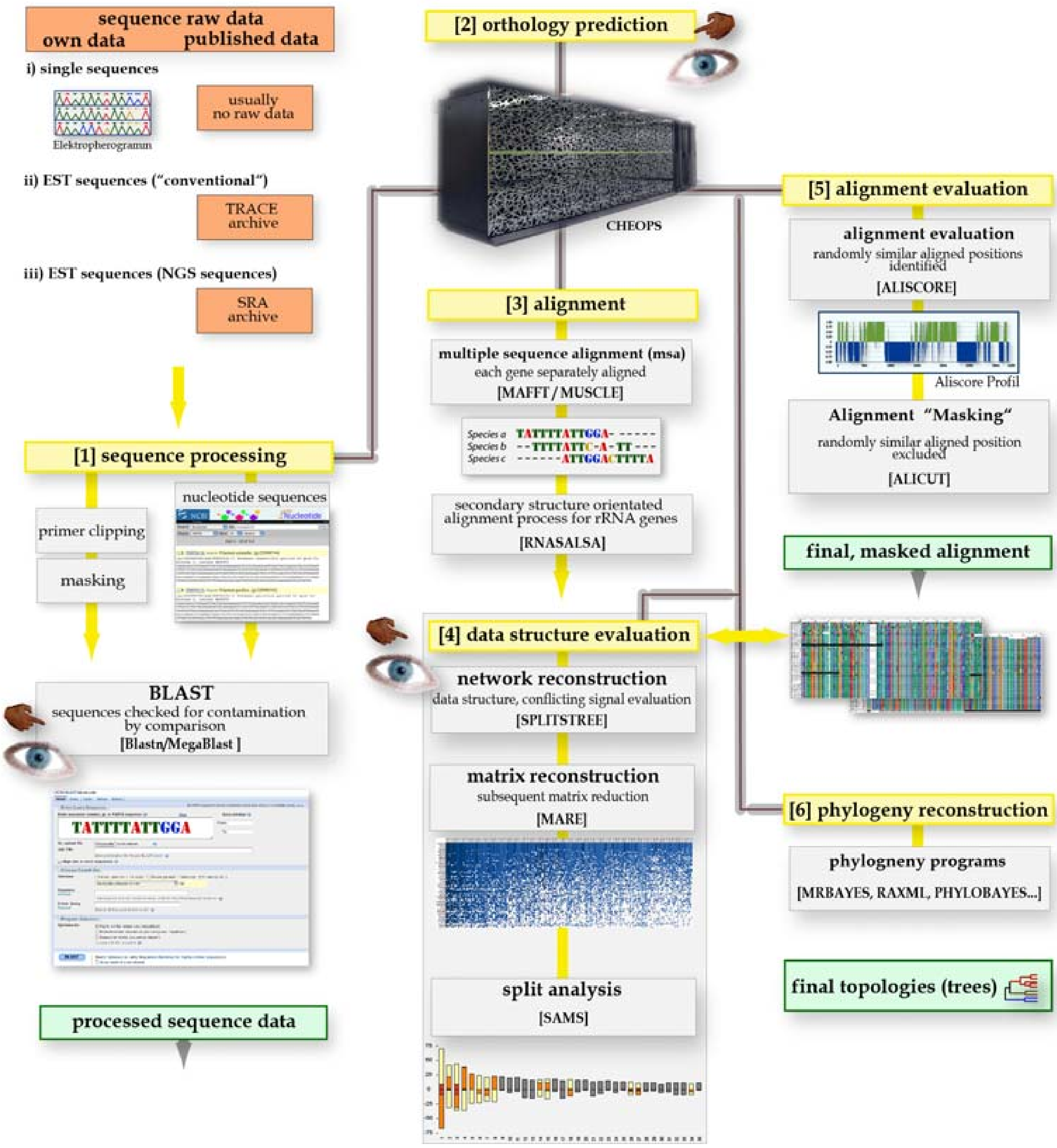

sequence raw data
own data
published data
i) single sequences
Elektropherogramm
usually no raw data
ii) EST sequences ("conventional")
TRACE archive
iii) EST sequences (NGS sequences)
SRA archive
[1] sequence processing
primer clipping
masking
nucleotide sequences
BLAST
sequences checked for contamination by comparison
[Blastn/MegaBlast]
TATTTTATTGGA
processed sequence data
[2] orthology prediction
CHEOPS
[3] alignment
multiple sequence alignment (msa)
each gene separately aligned
[MAFFT / MUSCLE]
Species a TATTTTATTGGA------
Species b --TTTTATTC-A-TT---
Species c ------ATTGGACTTTTA
secondary structure orientated alignment process for rRNA genes
[RNASALSA]
[4] data structure evaluation
network reconstruction
data structure, conflicting signal evaluation
[SPLITSTREE]
matrix reconstruction
subsequent matrix reduction
[MARE]
split analysis
[SAMS]
[5] alignment evaluation
alignment evaluation
randomly similar aligned positions identified
[ALISCORE]
Aliscore Profil
Alignment "Masking"
randomly similar aligned position excluded
[ALICUT]
final, masked alignment
[6] phylogeny reconstruction
phylogneny programs
[MRBAYES, RAXML, PHYLOBAYES...]
final topologies (trees)

Fig. 4. Working flow of a typical phylogenetic analysis, which starts from scratch with the raw data (gained sequences) and ends with the final topology. Finger and eye symbols pinpoint crucial points to control not only the quality of the process, but also the data quality in the meaning of potential information or conflicts within gene sequences (data structure). A major aspect is, that large scale sequencing and phylogenomic data requires enormous computational power. Supercomputers (in this case CHEOPS: Cologne High Efficiency Operating Platform for Science, RRZK University of Cologne) or large cluster systems (ZFMK Bonn) are an essential requisite in the conducted analyses. Bold bars shaded in grey with internal brown lines symbolize circuit paths and represent steps that are constraint by computational limitations. Own sequence raw data and published data (orange) are processed and quality controlled

often difficult and dependent on single favourable unpredictable conditions. Thus, if anything goes wrong during sequencing, the loss may be irreversible. The second aspect is that samples must not be contaminated by other samples before and after sequencing. If contamination happens, it might not be detectable at all with desastrous consequences. This aspect must be integrated in process flows of sequencing facilities, for example by using tagging techniques applied on each library prior to sequencing to identify immediately eventual contamination. BLAST procedures against other processed project samples or libraries must be a second manadatory strategy.

3. Quality management during molecular analyses

For phylogenomic data the presented figure 4 illustrates only a rough scheme or framework of analysis. Depending on applied techniques and the choice of different software packages an adaptation is needed. Detailed descriptions of the working process to analyse rRNA and phylogenomic data with an emphasis on data quality are given in: von Reumont et al., (2009), von Reumont, (2010) and Meusemann et al., (2010).

[1] Sequences from different sources are processed in software pipelines, quality checked and controlled. It is problematic, that normally electropherograms are not available for published single sequences selected from public databases i). Therefore sequence errors cannot be discovered in these data. ii) EST sequences are normally stored in the TRACE archive in NCBI including the trace files. These represent the raw data and are in general not quality checked. iii) NGS raw data is stored in the Short Read Archive (SRA), which accounts for the difference of sequences from next generation sequencing to the ‘conventional’ EST sequences. [2] Respectively for the phylogenomic data the prediction of putative ortholog genes is eminent important. This step is computationally intensive and different approaches can be used, see paragraph 3.2. [3] Processed sequence data is aligned applying multiple sequence alignment programs. In case of rRNA genes a secondary structure-based alignment optimization is suggested. [4] A first impression of the data structure is gained by phylogenetic network reconstructions. That point becomes problematic with phylogenomic datasets comprising hundreds of genes and alignment sizes larger than 100 MB! Consequently, a method to evaluate the structure for these datasets could be the software MARE that reconstructs graphics of the data matrix based on the tree-likeness of single genes for each taxon (Misof & Meyer, 2011). Subsequently, a matrix reduction is possible after the alignment evaluation. [5] The final alignment evaluation and processing is applied for each gene with ALISCORE (Misof & Misof, 2009) to identify randomly similar aligned positions and those positions are subsequently excluded (=masking) by ALICUT (www.utilities.zfmk.de). Single, masked alignments are concatenated to the final alignment or supermatrix. A matrix reduction for phylogenomic datasets is performed applying MARE to enlarge the relative informativeness and to exclude genes that are uninformative (Misof & Meyer, 2001; www.mare.zfmk.de). For most analyses it could be useful to compare data structure before and after the alignment process in a network reconstruction or unreduced matrix [4]. Information content in respect of signal that supports different splits in the alignment can be visualized by SAMS (Wägele & Mayer, 2007). [6] After this the phylogenetic tree reconstruction is performed with several software packages.

3.1 The processed sequences and their quality

Most phylogenetic studies use own and published sequences in their analyses. However, in both cases a rigorous control of the quality of the sequence is crucial. This is conducted in the steps of sequence processing (see figure 4, [1]). Different software tools guarantee quality by threshold value settings. A completely different aspect of quality is that the finally included sequence is indeed linked to the supposed species. Either misidentification of the specimen or the sequence can evoke serious bias in a subsequent analysis. If reaction in the laboratory were contaminated, the sequence is linked to the wrong species depending on the source of contamination. Both kinds of misidentification can be identified in general by careful BLAST procedures (Altschul et al., 1997, Kuiken & Corber, 1998). Yet, they are time intensive and in some cases difficult to interpret. For example, if you work with closely related species. In this case, the misidentification or contamination is rather impossible to detect, in particular if one species is unknown or only few or no sequences have been published. Other sources of data (like morphology) can also help to identify contamination (Wiens, 2004).

Several studies report that possible contaminations of taxa played a veritable role in studies, which proposed new evolutionary scenarios, but were actually based on contaminated sequences (von Reumont, 2010; Waegele et al., 2009; Koenemann et al., 2010). A careful control of sequence quality or a more critical interpretation of the reconstructed topologies could have prevented the (eventually repeated) inclusion of the contaminated sequences and subsequent publication of such suspicious phylogenetic trees. If contaminated sequences of older studies from rarely sequenced species are tacitly included into new analyses, this indeed can obscure phylogenetic implications. That is probably the case with the Mystacocarida, a crustacaean group with an still unclear phylogenetic position. They are rarely sequenced and the first and only published 18S rRNA sequence by Spears and Abele (1998) is very likely a contamination (von Reumont, 2010; Koenemann et al., 2010), which was impossible to identify for the authors in that study of 1998, which constituted the first larger analysis of crustaceans at all. A new study with completely sequenced 18S rRNA genes (von Reumont et al., 2009) including a new 18S rRNA gene sequence of the Mystacocarida revealed the contamination of the published sequence (von Reumont, 2010).

The search for contamination reaches a new dimension in phylogenomic data. A recent study (Longo et al., 2011) describes, that some non-primate genome databases, like the NCBI trace archive, provide sequences with human DNA contaminations, which can be traced back to pre-sequencing errors and/or low quality standards. Consequently, cross checking with published data might not help to be 100 percent sure about your own sequences. If you read the last sentence think about your own laboratory routines. Are they sufficient? If you outsource EST sequencing to an external company, which quality standard do they have and which risk management to handle possible contaminations?

This is respectively worrisome in cases of cross species analyses and genome analyses and indicates, that a better screening is generally needed (Phillips, 2011). The response of NCBI was, that trace archive data represents the raw data, which is not quality checked (http://www.ncbi.nlm.nih.gov/About/news/18feb2011.html). A careful processing of these sequences is obligate before analyses, including the control for possible contamination. An important conclusion is that every sequence from public databases should be treated suspiciously and a careful processing procedure is necessary to prevent errors by contamination. Do not trust your own data, but also do not trust public data.

3.2 Orthology prediction

Only homologous genes can be used in molecular phylogenetic studies. Homologous genes are further distinguished in two different classes: i) ortholog genes which originate in a single speciation event, and ii) paralog genes that originated from gene duplications independently of speciation events (Fitch, 1970; Sonnhammer & Koonin, 2002; see review: Koonin, 2005). The prediction of ortholog genes in the era of large scale and next generation sequencing is a very delicate and computationally intensive process. An overview of commonly used methods for prediction of putative ortholog genes and their efficiency assessment is given in Roth et al. (2008) and Altenhoff and Dessimoz (2009).

A difficulty for phylogenetic reconstructions within arthropods is that only few data bases include sufficient numbers of complete arthropod genomes (Altenhoof & Dessimoz, 2009). INPARANOID and OMA are the two leading projects concerning the number of included arthropods. For that reason the orthology prediction for an arthropod dataset (Meusemann et al., 2010; von Reumont, 2010) and a further pancrustacean dataset (von Reumont et al., 2011) were based on INPARANOID 6 and 7 (Ostlund et al., 2010). Identified ortholog gene sets were extended using the HaMStR approach (Ebersberger et al., 2009) relying on the INPARANOID project. A set of orthologous genes was constructed using the InParanoid transitive closure (TC) approach in HaMStR described by Ebersberger et al. (2009). This set based on proteome data of so called 'primer taxa', which are completely sequenced genome species. Sequences of primer taxa were aligned within the set of orthologs and used to infer profile hidden Markov models (pHMMs). Subsequently, the pHMMs were used to search for putative orthologs among the translated ESTs of all taxa in the data set.

For the pancrustacean dataset pre-analyses were performed to compare the influence of using the OMA or INPARANOID projects with the same settings in HaMStR and the previous processing pipeline. For both analyses the same five primer taxa (*Aedes aegypti, Apis mellifera, Daphnia pulex, Ixodes scapulatis, Capitella* sp.) were used in HaMStR to train hidden markov models to extent the putative orthologs for all included taxa. Relying on OMA, 344 putative ortholog genes were identified in contrast to 1886 genes using INPARANOID. The resulting, reduced topologies (RAXML, -f, a, PROTCATWAG, 1000 BS) differ clearly in their resolution: the OMA based topology shows less resolution.

However, these results demonstrate the importance of further, more detailed studies on the impact of ortholog gene prediction. The quality of the trees might be severely influenced in this step of the analysis. A problem is the enormous computational power needed for comparative analysis of phylogenomic datasets.

3.3 Evaluation of data structure and data quality

All steps described so far are important to obtain in a standardized, rigorous processing high quality of the data and finally gene sequences, which are subsequently aligned and used for phylogenetic analyses.

The term *data quality*, however, addresses a different level of quality. A given multiple sequence alignment (MSA, synonymously often named data matrix) can include processed genes that are finally (after the processing procedure) of high quality, but for the phylogenetic goal to reconstruct a specific evolutionary history maybe not usable, if not informative. *Data quality* indeed refers to the scale of information or signal within the alignment. The term *data structure* is sometimes used synonymously to the term *data quality*. Multiple substitution processes generally change sequences with time caused by random substitution processes, however, the extent of substitutions differs for parts of the DNA. In some parts of the DNA this substitution process erodes the former phylogenetic signal by multiple exchanges of nucleotides. After a long time nucleotides that represented synapomorphic characters to a sister taxon are by chance multiple substituted in the process of signal erosion (Wägele & Mayer, 2007). By this process a different, random signal (noise) can arise, that in most cases is in conflict (and obscures) the historical, phylogenetic signal. In contrast, other genes are extremely conservative and nucleotides barely change with time. In this case a phylogenetic signal is hardly to detect either, caused by too few substitutions or synapomorphic characters. The mathematical substitution models, which are applied to reconstruct phylogenetic trees from multiple sequence alignments, try to implement several aspects of the briefly described processes. However, they are always an approximation and respectively are unable to differ between phylogenetic signal and noise. For further details see (Felsenstein, 1988; Wägele, 2005; Wägele & Mayer, 2007).

A first and fast evaluation of the structure in a dataset is feasible with network reconstructions, in which conflicts are visualized that are not illustrated by the (forced) bifurcations in phylogenetic trees (Holland et al., 2004; Huson & Bryant, 2006). It was the first time proposed by Bandelt and Dress (1992) to combine every phylogenetic analysis with a non-approximative method, which allows not compatible, alternative groupings contrary to bifurcting phylogenetic trees. One approach, the method of split decompositon, was developed by Bandelt and Dress (Bandelt & Dress, 1992). Hendy, Penny and Steel published a second method, the split analysis (Hendy & Penny, 1993; Hendy et al., 1994). Both methods work with so called bifurcations or splits.

A split is a couple of two groups of taxa, which are distinct subsets of the whole taxaset. Within the molecular phylogenetic context splits are distinguished by the occurence of nucleotide bases within sites. For a set of n taxa, exist 2^{n-1} possible bipartitions, in real datasets occur normally fewer splits. If there is only split signal for one unique dichotomous tree within a dataset, the number of splits is of the same value as the edges of a possible phylogeny. Given a taxon quartet (A, B), (C, D) few synapomophies between B and C can cause a split for second, alternatively supported topology (A, D) (B, C). This split migth not be visualized in a reconstructed tree-topology. Software packages offering non-approximate methods are SplitsTree (Huson & Bryant, 2006), Spectrum (Charleston, 1998), Spectronet (Huber et al., 2002) and SAMS (Wägele & Mayer, 2007).

SAMS is a software approach that was developed by Wägele and Mayer (2007) to perform a split analysis on the alignment. It accounts for all states of bases but analyses the columns of an alignment for occurring splits in a efficient way. Hence you can generate a split spectrum showing conflicting signal simultaneously obtaining a good overview on the data quality. Real splits are additionally differentiated from the conflicting ones. The method is currently under development, at the moment large datasets are difficult to analyze. Additionally, only nucleotide data is possible as input format. Further development is necessary and in progress to establish a new system, which evaluates all sites of an alignment and weights them according to contrast and homogeneity aspects to address these aspects.

Yet, network reconstruction and split analysis is limited by the size of a dataset and with larger or phylogenomic datasets still beyond abilities of available programs. Additionally, networks give only a rough overview and illustrate the present data structure, answering the question if a conflict or noise exists. More details are often not to analyze, for example which single genes or partitions create a conflict within an alignment. This part becomes additionally delicate handling ‘supermatrices’ that are composed of phylogenomic data.

Several strategies exist to handle ‘supermatrices’, which mostly are data sets with a large number of taxa and genes, but also missing information or gaps. Often, concatenated ‘supermatrices’ are filtered and reduced using predefined thresholds of data availability

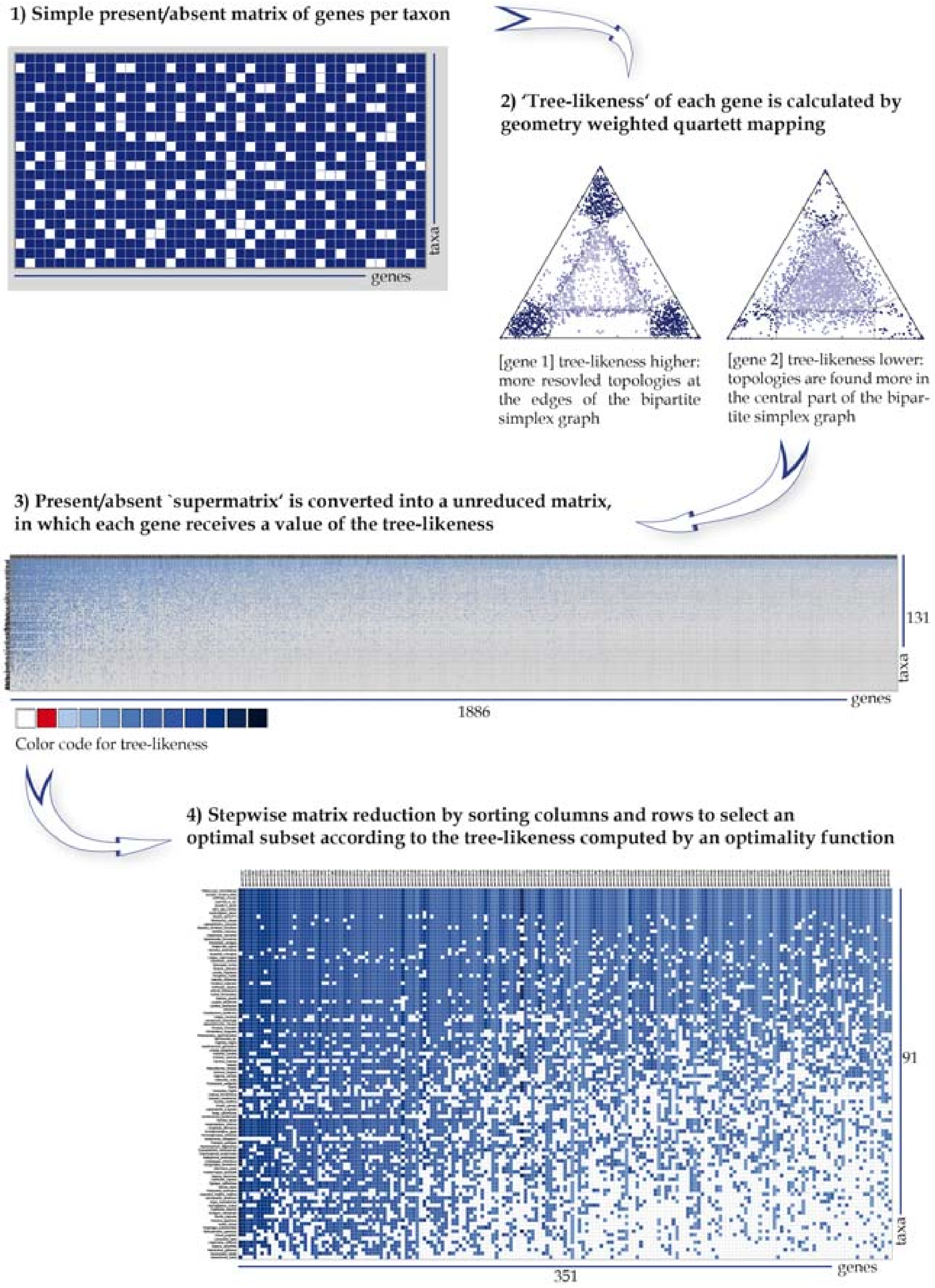

1) Simple present/absent matrix of genes per taxon
taxa
genes
2) 'Tree-likeness' of each gene is calculated by geometry weighted quartett mapping
[gene 1] tree-likeness higher: more resovled topologies at the edges of the bipartite simplex graph
[gene 2] tree-likeness lower: topologies are found more in the central part of the bipartite simplex graph
3) Present/absent `supermatrix' is converted into a unreduced matrix, in which each gene receives a value of the tree-likeness
131
taxa
genes
1886
Color code for tree-likeness
4) Stepwise matrix reduction by sorting columns and rows to select an optimal subset according to the tree-likeness computed by an optimality function
91
taxa
genes
351

Fig. 5. Work flow of the MARE software. All genes are concatenated to a supermatrix, which is transformed into a `supermatrix' composed of all genes that are represented by tree-likness value. A tree-likeness is calculated in the step before via geometry weighteed quartet mapping. This supermatrix` is reduced by selecting an optimal subset of genes and taxa relying on the calculated value of the tree-likeness. The reduction is stepwise performed using an optimality function. The matrices composed of the tree-likeness values for each gene are colour coded. White symbolizes an absent gene, red a value of 0. From light to dark blue the value increases, dark blue represents a value of 0.9 -1.0

(Dunn et al., 2008; Philippe et al., 2009) depending on the relational number of present genes for a taxon. Taxa are excluded, if they are represented by less genes than accepted with the defined threshold value. Software tools like MARE are a first step to evaluate the data more detailed and enable an objective reduction of 'supermatrices' (large MSA´s of phylogenomic data), by selecting subsets of genes. MARE utilizes an alternative approach to data reduction selecting a subset of genes and taxa from a supermatrix based on information content and data availability (Meyer & Misof, 2010; http://mare.zfmk.de; Meusemann et al., 2010; von Reumont et al., 2011). The approach yields a condensed data set of larger information content by maximizing the ratio of signal to noise, and reducing uninformative genes or poorly sampled taxa.

MARE evaluates in a first step the 'tree-likeness' of each single gene. Tree-likeness reflects the relative number of resolved quartets for all possible (but not more than 20,000) quartets of a given sequence alignment or alignment partitions. The process is based on geometry-weighted quartet mapping (Nieselt-Struwe & von Haeseler, 2001), extended to amino acid data. For each gene a value for the tree-likeness is calculated by summarizing the support values for each of the three possible topologies during the quartet mapping procedure. After this step the previous present/absent matrix is changed to a matrix that contains values of tree-likeness for each gene per taxon. In the second step the matrix reduction is performed. The connectivity of the matrix (the gene and taxa overlap) is monitored during this step: two genes must have connection with at least three taxa. The matrix is reduced stepwise, with each reduction a new matrix is generated. Within each reduction step the column or row with the lowest information content (sum of values for tree-likeness) is excluded. The procedure is guided by an optimality function, which represents a trade off between matrix density and retained taxa and genes. For further details on the procedure and the algorithm, see: (Meyer & Misof, 2011; http://mare.zfmk.de).

4. Conclusions

When conducting or managing a project in molecular evolution use the available elements of project managing to prevent mistakes at this basic level. Important are the time schedule and milestones with sufficient backup time. A careful stakeholder analysis provides a detailed risk analysis, which is important in general, respectively if many persons or working groups are involved. Fieldtrips and appropriate preservation methods of the collected species must be carefully planned either, to start the molecular analysis with qualitative successful isolated material.

A process flow with a rigorous concept of quality control contributes to the quality of the gained sequences or data. The final sequences should have been checked for contamination. If techniques of next generation sequencing or expressed sequence tags are used, pay sufficient attention to select the best strategy for the prediction of ortholog genes. The aligned sequences should always be processed in the multiple sequence alignment for each gene or partition. Software like ALISCORE identifies randomly aligned alignment positions. Before the reconstruction of phylogenetic trees the *data quality* should be evaluated applying software to visualize the data structure and potential conflicts. Software for a more specific split analysis capable of larger data is e.g. SAMS, which is still under development. Assessing the data structure and quality is an essential strategy to identify conflict in phylogenetic trees or their eventual inability to reflect the 'real' evolutionary history of a species group.

Large data matrices or MSAs should be reduced to subsets, which were selected by the tree-likeness of each gene applying the software MARE. The software MARE is a first step to utilize objective criteria to select informative subsets of genes from a partially 'supermatrix'. However, several aspects are still to address further in future. Procedures of orthology prediction and matrix reduction need for example further investigation.

5. Acknowledgement

BMvR and SAM thank J-W Wägele for the chance and support to conduct the projects within the DMP framework. We would like to thank all colleagues who have been involved in the priority program SPP 1174 'Deep Metazoan Phylogeny' of the Deutsche Forschungsgemeinschaft (DFG) and the members of the molecular lab and Zentrum für molekulare Biodiversität (zmb) at the Zooloogischen Forschungsmuseum Alexander Koenig (ZFMK), Bonn. Respectively cooperation with Karen Meusemann was prosperous. Open discussions and exchange of experiences was extremely fruitful in all fields, not only the molecular area. Michael Kube from the Max Planck Institute of Molecular Biology and Genetics, Berlin, Germany gave eminent help and tips for the work with RNA. For detailed explanations and answers regarding the NGS projects we would like to thank colleagues from following companies: GATC, Konstanz, Germany and LGC, Berlin, Germany. The work for this manuscript is granted by the DFG proposals WA530/34, WA530/33.

6. References

Altschul, S. F.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids and Research,* 25, 3389-3402

Altenhoff, A. M. & Dessimoz, C. (2009). Phylogenetic and functional assessment of orthologs inference projects and methods, *PLoS Computational Biology*, 5, 1

Bandelt, H. J. & Dress, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution* 1:242-252.

Bouck, A. & Vision, T. (2007). The molecular ecologist's guide to expressed sequence tags. *Molecular Ecology,* 16, 907-924

Bourne, L. (2010). Beyond reporting. The communication strategy, *PMI Global Congress Proceedings*, Melbourne, Australia

Budd, G.E & Telford, M.J. (2009). The origin and evolution of arthropods, *Nature*, 457, pp. 812-817

Charleston M. (1998). Spectrum: spectral analysis of phylogenetic data, *Bioinformatics (Oxford, England)* 14, 1, 98-9

Forster, J.L.; Harkin, V.B.; Graham, D.A. & McCullough, S.J. (2008). The effect of sample type, temperature and RNAlater (TM) on the stability of avian influenza virus RNA, *Journal of Virological Methods*, 149, pp. 190-194

Ebersberger, I.; Strauss, S. & Von Haeseler, A. (2009). HaMStR: profile hidden markov model based search for orthologs in ESTs, *BMC Evolutionary Biology*, 9, 157

Edgecombe, G.D. (2010). Arthropod phylogeny: An overview from the perspectives of morphology, molecular data and the fossil record, *Arthropod Structure and Development,* 39, pp. 74-87

Eisen, J. A. (1998). Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis, *Genome Research,* 8, 163-7

Ellegren, H. (2008). Sequencing goes 454 and takes large-scale genomics into the wild, *Molecular Ecology,* 17, 1629-1631.

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

Fitch, W. M. (1970). Further improvements in the method of testing for evolutionary homology among proteins, *Journal of Molecular Biology,* 49, 1-14.

Freeman, E.R. (2010). *Strategic management: a stakeholder approach.* ISBN 978-0521151740, Cambridge University Press (first published by Pitman Publishing, 1984)

Gemeinholzer, B.; Droege, G.; Zetzsche, H.; Knebelsberger, T.; Raupach, M.; Borsch, T.; Klenk, H.-P.; Haszprunar, G. & Waegele; J.-W. (2011). The DNA Bank Network: the start from a German initiative. Biopreservation and Biobanking. April 2011, 9 (1):51-55, available at http://www.dnabank-network.org

Gorokhova, E. (2005). Effects on preservation and storage of microcrustacenas in RNAlater™ on RNA and DNA degradation, *Limnology and Oceanography: Methods,* 3, 143-148

Grotzer, M.A.; Pati, R.; Georger, B.; Eggert, A.; Chou, T.T. & Philips, P.C. (2000), Biological stability of RNA isolated from RNAlater™-treated brain tumor and neuroblastoma xenografts, *Medical Pediatric Oncology,* 34:438-442

Hemmrich, K.; Denecke, B.; Paul, N.E.; Hoffmeister, D. & Pallua, N., (2010). RNA Isolation from Adipose Tissue: An Optimized Procedure for High RNA Yield and Integrity, *Labmedicine,* 41 (2), pp 104-106

Hendy, M. & Penny, D., (1993). Spectral analysis of phylogenetic data. Journal of Classification, 10, 1, 5-24

Hendy, M., Penny, D. & Steel, M., (1994). A discrete Fourier analysis for evolutionary trees. Proceedings of the National Academy of Sciences of the United States of America, 91, 8, 3339-43

Holland, B. R.; Huber, K. T.; Moulton, V. & Lockhart, P. J. (2004). Using Consensus Networks to Visualize Contradictory Evidence for Species Phylogeny, *Molecular Biology and Evolution,* 21, 1459-1461

Huber, K, Langton M, Penny D, Moulton V, & Hendy M., (2002). Spectronet: a package for computing spectra and median networks., Applied bioinformatics 1, 3, 159-61

Hudson, M. E., (2008). Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, 8, 3-17

Huson, D. H. & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution,* 23, 254-267

Jongeneel, C. V. (2000). Searching the expressed sequence tag (EST) databases: panning for genes. *Briefings in Bioinformatics* 1, 76-92.

Kerzner, H. (2009). *Project management: a systems approach to planning, scheduling and controlling,* ISBN 978-0470278703, John Wiley & Sons, 10th edition

Koenemann, S.; Jenner, R. A.; Hoenemann, M.; Stemme, T. & Von Reumont, B. M. (2010). Arthropod phylogeny revisited, with a focus on crustacean relationships, *Arthropod Structure and Development,* 39, 88-110

Koonin, E. (2005). Orthologs, paralogs and evolutionary genomics, *Annual Reviews of Genetics*, 39, 1, 209-338

Kuiken, C. & Korber, B. (1998). Sequence quality control, Los Alamos National Laboratory *HIV Compendium*, III, pp. 80-90

Litke, H.-D.; Kunow, I. & Schulz-Wimmer, H. (2010). *Projektmanagment,* ISBN 978-3-448-09949-2, Haufe-Lexware GmbH & Co. KG, Freiburg

Longo, M. S.; Longo, M. J.; O'Neill, R. J. & O'Neill (2011). Abundant Human DNA Contamination Identified in Non-Primate Genome Databases, *PLoS ONE,* 6, 2, e16410. doi:10.1371/journal.pone.0016410

Meusemann, K.; Von Reumont, B. M.; Simon, S.; Roeding, F.; Strauss, S.; Kuck, P.; Ebersberger, I.; Walzl, M.; Pass, G.; Breuers, S.; Achter, V.; Von Haeseler, A.; Burmester, T.; Hadrys, H.; Wagele, J. W. & Misof, B. (2010). A phylogenomic approach to resolve the arthropod tree of life. *Molecular Biology and Evolution* 27, 2451-64.

Meyer B. & Misof, B. (2011). MARE: Matrix Reduction – A tool to select optimized data subsets from supermatrices for phylogenetic inference. Zentrum für molekulare Biodiversitätsforschung (zmb) am ZFMK, Adenauerallee 160, 53113 Bonn, Germany, http://mare.zfmk.de

Misof, B. & Misof, K. (2009). A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion, *Systematic Biology*, 58, 1

Mülhardt, C. (2008). *Der Experimentator: Molekularbiologie/Genomics*, Spektrum Akademischer Verlag, 6. Auflage. ISBN-10: 9783827420367

Mutter, G.L.; Zahrieh; D., Liu; C.M.; Neuberg, D.; Finkelstein, D.; Baker, H.E. & Warrington, J.A. (2004). Comparison of frozen and RNAlater™ solid tissue storage methods for use in RNA expression microarrays, *BMC Genomics*, 5:88

Nieselt-Struwe K. & Von Haeseler A. (2001). Quartet-mapping, a generalization of the likelihood-mapping procedure. *Molecular Biology and Evolution* 18:1204-1219

Ostlund, G.; Schmitt, T.; Forslund, K.; Köstler, T.; Messina, D. N.; Roopra, S.; Frings, O. & Sonnhammer, E. L. L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis, Nucleid Acid Research, 38

Palumbi, S. R. (1996). Nucleic acids II: The Polymerase Chain Reaction, in: Molecular Systematics, Hillis, D. M., Moritz, C., Mable, B. K. 2nd edition, Sinauer Associates, ISBN 978-0878932825

Petterson, E.; Ludneber, J. & Ahmadian, A. (2009). Generations of sequencing technologies, *Genomics*, 93, pp. 105-111

Philippe, H.; Delsuc, F.; Brinkmann, H. & Lartillot, N. (2005). Phylogenomics, *Annual Review of Ecology and Evolutionary Systematics,* 36, 541-562

Philippe H; Derelle R; Lopez P; Pick, K.; Borchiellini, C.; Boury-Esnault, N.; Vacelet, J.; Renard, E.; Houliston, E.; Quéinnec, E.; Da Silva, C.; Wincker, P.; Le Guyader, H.; Leys, S.; Jackson, D. J.; Schreiber, F.; Erpenbeck, D.; Morgenstern, B.; Wörheide, G.

& Manuel, M. (2009). Phylogenomics revives traditional views on deep animal relationships. *Curr Biol.* 19:706-712.

Phillips, M.L. (2011). Contamination of non-primate DNA archives with human sequences indicates that better screening is needed, *nature news*, doi:10.1038/news.2011.99

Ronaghi, M. (2001). Pyrosequencing Sheds Light on DNA Sequencing, *Genome Research,* 11, pp. 3-11

Sambrook, J. & Russel, D. W. (2000). Molecular Cloning: A laboratory manual, 3rd reprint, ISBN 978-0879695774

Shendure, J.; Mitra, R.; Varma, C. & Church, G. (2004). Advanced sequencing technologies: methods and goals, *Nature Reviews in Genetics,* 5, pp. 335-344.

Sonnhammer, E. L. L. & Koonin, E. V. (2002). Orthology, paralogy and proposed classification for paralog subtypes, Trends in Genetics, 18, 12, 619-620

Spears, T. & Abele, L. G. (1998). Crustacean phylogeny inferred from 18S rDNA, In *Arthropod Relationships*, editors: R. A. Fortey and R. H. Thomas, ISBN 978-0412754203, Chapman and Hall, pp. 169-187, London

Thornton, J. W. & Desalle, R. (2000). Gene family evolution and homology: genomics meets phylogenetics, *Annual Reviews of Genomics and Human Genetics,* 1, 41-73

Vink, C.J.; Thomas, S.M.; Paquin, P.; Hayashi, C.Y. & Hedin, M. (2005). The effects of preservatives and temperatures on arachnid DNA, *Invertebrate Systematics*, 19, pp. 99-104

Voelkerding, K. V.; Dames, S. A. & Durtschi, J. D. (2009). Next-Generation Sequencing: From Basic Research to Diagnostics, *Clinical Chemestry*, 55, pp. 641-658

Von Reumont, B. M.; Meusemann, K.; Szucsich, N.; Dell'ampio, E.; Gowri-Shankar, V.; Bartel, D.; Simon, S.; Letsch, H. O.; Stocsits, R. R.; Luan, Y. X.; Wägele, J. W.; Pass, G.; Hadrys, H. & Misof, B. (2009). Can comprehensive background knowledge be incorporated into substitution models to improve phylogenetic analyses? A case study on major arthropod relationships, *BMC Evolutionary Biology* 9, 119.

Von Reumont, B. M. (2010). *Molecular insights to crustaecan phylogeny. A status quo of past, present and perspective prospects also covering phylogenomics,* ISBN 978-3-8381-1770-6, Südwestdeutscher Verlag für Hochschulschriften, Saarbrücken, Germany.

Von Reumont, B. M.; Jenner, R. A.; Wills, M. A.; Dell´Ampio, E.; Pass, G.; Ebersberger, I.; Meusemann, K.; Meyer, B.; Koenemann, S.; Iliffe, T. I.; Stamatakis, A.; Niehuis, O. & Misof, B. (2011). Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as a sister group to Hexapoda, accepted with minor revisions, in re-prep for MBE

Weaver, P. (2007). A Simple View of Complexity in Project Management, *Proceedings of the 4th World Project Management Week, Singapore*

Wiens, J. (2004). The Role of Morphological Data in Phylogeny Reconstruction, *Systematic Biology,* 53, 653-661

Wägele, J.-W. (2005). *Foundations of phylogenetic systematics*, ISBN-13: 9783899370560, Friedrich Pfeil Verlag, München

Wägele. J.-W. & Mayer, C. (2007). Visualizing differences in phylogenetic information content of alignments and distinction of three classes of long-branch effects, *BMC Evolutionary Biology*, 7, 147

Wägele, J. W.; Letsch, H.; Klussmann-Kolb, A.; Mayer, C.; Misof, B. & Wagele, H. (2009). Phylogenetic support values are not necessarily informative: the case of the Serialia hypothesis (a mollusk phylogeny), *Frontiers in Zoology*, 6, 12

A.5. Additional manuscript prepared during the projects (II)

Kück *et al. BMC Bioinformatics* 2014, **15**:294
http://www.biomedcentral.com/1471-2105/15/294

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

AliGROOVE – visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support

Patrick Kück[1*], Sandra A Meid[1], Christian Groß[2], Johann W Wägele[1] and Bernhard Misof[1]

Abstract

Background: Masking of multiple sequence alignment blocks has become a powerful method to enhance the tree-likeness of the underlying data. However, existing masking approaches are insensitive to heterogeneous sequence divergence which can mislead tree reconstructions. We present AliGROOVE, a new method based on a sliding window and a Monte Carlo resampling approach, that visualizes heterogeneous sequence divergence or alignment ambiguity related to single taxa or subsets of taxa within a multiple sequence alignment and tags suspicious branches on a given tree.

Results: We used simulated multiple sequence alignments to show that the extent of alignment ambiguity in pairwise sequence comparison is correlated with the frequency of misplaced taxa in tree reconstructions. The approach implemented in AliGROOVE allows to detect nodes within a tree that are supported despite the absence of phylogenetic signal in the underlying multiple sequence alignment. We show that AliGROOVE equally well detects heterogeneous sequence divergence in a case study based on an empirical data set of mitochondrial DNA sequences of chelicerates.

Conclusions: The AliGROOVE approach has the potential to identify single taxa or subsets of taxa which show predominantly randomized sequence similarity in comparison with other taxa in a multiple sequence alignment. It further allows to evaluate the reliability of node support in a novel way.

Keywords: Software, Alignment quality, Sequence heterogeneity, Topological node support

Background

Alignment masking as a measure of reducing noise in sequence alignments is regularly applied in phylogenetics. The idea behind the concept of masking blocks of sequence alignments is the reduction of the unpredictable influence of substitution saturation and/or ambiguously aligned blocks of sequence alignments on subsequent tree reconstructions [1-8] by increasing the tree-likeness of the data. Simulations and analyses of alignment masking of empirical data corroborate the correctness of this idea. Currently, software packages mask complete blocks of multiple sequence alignments applying either arbitrarily chosen thresholds of sequence variability within alignment columns (e.g. software Gblocks [1,2] and REAP [9]), or automatically adjusted thresholds depending on the input alignment (e.g. trimAl [4] and BMGE [6]), or applying a sliding window approach to identify blocks of predominantly high alignment ambiguity (ALISCORE [5,7]). All methods exclude complete alignment blocks instead of sequence subsets thus masking also potentially valuable data for subsets of taxa.

Due to their design all masking methods are relatively insensitive to heterogeneous sequence divergence of single taxa. This is an important deficiency of masking methods, because heterogeneous sequence divergence can cause strong biases in tree reconstructions, for example

*Correspondence: patrick_kueck@web.de
[1] Zoologisches Forschungsmuseum A. Koenig, Adenauerallee 160-163, 53113 Bonn, Germany
Full list of author information is available at the end of the article

long branch effects or the misplacement of rogue taxa. Therefore, a method which can visualize heterogeneous sequence divergence or alignment ambiguity related to single taxa or subsets of taxa would be a useful complement to currently used masking approaches. It offers the chance to identify taxa which are potentially misplaced in trees and reduce the tree-likeness of the data.

For this purpose, we developed AliGROOVE, a new tool to visualize the extent of sequence similarity and alignment ambiguity in pairwise sequence comparisons derived from a multiple sequence alignment. AliGROOVE can help to detect strongly derived sequences that have the potential to bias tree reconstructions and node support. We implemented an adaptation of the recently published ALISCORE masking algorithm [5,7] which has been successfully tested in simulations and on empirical data [5,7,8]. Using a simple match/mismatch scoring for nucleotide data and a BLOSUM62 scoring matrix for amino acid data ALISCORE uses a Monte Carlo resampling within a sliding window to generate profiles of pairwise sequence similarity for all pairwise sequence comparisons. AliGROOVE summarizes site scores of these profiles normalized over the whole alignment length for each pairwise comparison. The obtained scoring values between sequences are translated into a similarity matrix and thus deliver information on the extent of taxonomically heterogeneous alignment ambiguity or sequence similarity within a multiple sequence alignment.

We used simulated data to investigate if our application of the algorithm is able to detect ambiguously aligned taxa or groups of taxa and if the obtained sequence similarity scores can be used to tag unreliable nodes. For that purpose we tested AliGROOVE on data sets with and without indel events whereby tests on data sets with indel events are performed on correct and on realigned data sets that deviate from the true alignment. Additionally, we applied AliGROOVE on an empirical data set comprising five mitochondrial genes of 53 chelicerate ingroup taxa and eight myriapod outgroup taxa. With both the simulated and empirical data sets we also tested the potential of the approach to illustrate heterogeneous tree-likeness among data blocks within an alignment.

AliGROOVE Algorithm

Identification of sequence similarity/scoring

The algorithm of AliGROOVE is based on the scoring scheme of ALISCORE [5,7] which compares pairs of amino acid/DNA sequences for random similarity within a sliding window. In short, first, the observed mismatch within the sliding window is scored. This mismatch score is then compared with mismatch scores of the same window size generated by permutations of character states within the sliding window and a predefined sequence neighborhood. If the observed score is better than 95% of the score of all generated permutations, it is considered non-random, otherwise indistinguishable from random similarity. Each position within the sliding window receives a positive sign if the observed score was significantly better than scores of random sequence similarity, or if not, a negative sign. The number of single signs for each alignment position corresponds to the size of the sliding window. For each position signs are summed up and normalized by the sliding window size. A profile of sequence similarity between two sequences will thus show sections in which these two sequences might show non-random similarity indicated by a positive sum of signs and sections of random similarity expressed by a negative sum of signs for each position. Now, for each profile the AliGROOVE algorithm calculates an arithmetic mean of profile signs over all sites excluding globally invariant sites within the alignment and records these values in a matrix for a given set of sequences. The entries in this similarity matrix express the average amount of non-random versus random similarity in pairwise comparisons and can thus illustrate heterogeneous signal in the data.

The algorithm is based on either match/mismatch scores for nucleotide sequences or on amino acid substitution matrices (BLOSUM62, PAM250, PAM500) to score amino acid matches/mismatches. This scoring regime turned out to be efficient in alignment masking [5,7,8,10-18].

Identification of suspicious branches

AliGROOVE pairwise similarity scores can be used to tag potentially unreliable relationships in a pre-defined tree. Potentially unreliable relationships can be caused by extensive substitution saturation or extensive alignment ambiguity both causing long branches in a tree which can occurr in inner and terminal branches.

AliGROOVE tags terminal branches with the mean pairwise similarity score (S_{XY}) between the terminal taxon and all other taxa. For example, the terminal branch of taxon A in a six taxon topology (taxa A to F), is tagged with R_A defined as:

$$R_A = \frac{S_{AB} + S_{AC} + S_{AD} + S_{AE} + S_{AF}}{5} \tag{1}$$

To tag internal nodes, AliGROOVE calculates the mean similarity score from all pairwise comparisons across this node. The tagging of the internal nodes follows the hierarchy given by a topology and ends at the most central internal branch. Following a guiding topology effectively reduces the number of splits to be analyzed to the ones which are of special interest. This reduction of the complexity of analyses makes the approach computationally efficient. For example, to tag the internal branch separating taxa A and B from the remaining taxa (taxa C to F), AliGROOVE calculates $R_{AB|CDEF}$ defined as:

$$R_{AB|CDEF} = \frac{S_{AC} + S_{AD} + S_{AE} + S_{AF} + S_{BC} + S_{BD} + S_{BE} + S_{BF}}{8} \quad (2)$$

The calculation of the mean pairwise similarity score treats all pairwise comparisons as independent replicates. This assumption is not justified in every case. For example, taxa C and E might be closely related and S_{AC} and S_{AE} do not represent fully independent replicates.

Results

Testing the performance with simulated data (Setup A & B)

We simulated nucleotide and amino acid sequence alignments under two different topological conditions (Figure 1). Our first setup represents 4-taxon trees (setup A) containing long terminal branches BL2 (Figure 1a). This setup has been selected to reduce the complexity of phenomena and to demonstrate the ability of AliGROOVE to identify heterogeneous sequences which can cause long branch attraction of terminal branches. Our second simulation setup consists of 6-taxon trees (setup B) containing long internal branches BL2 (Figure 1b). The frequencies of correct and incorrect Maximum Likelihood tree reconstructions using nearly correct model assumptions (using four rate categories instead of a continuous Γ distribution) were recorded (Figures 2, 3). To simulate large-scale phylogenetic analyses based on concatenated supermatrices, setup A comprises alignment lengths of 250,000 sites, while setup B has alignment lengths of 50,000 sites. The shorter sequence lengths of setup B have been chosen to reduce computational time of our 6-taxon analyses.

In setup A (Figure 1a), we simulated data with increasing terminal branch lengths of two unrelated taxa. For increasing branch length conditions the similarity scores between sister taxa correlate with tree reconstruction success ((L1,S1) & (L2,S2) in Figure 2). The mean similarity scores for internal branches are as well correlated with the tree reconstruction success. Negative mean similarity scores are directly correlated with tree reconstruction errors. Using AliGROOVE with the tree tagging option to project the observed pairwise sequence similarity scores on a provided guiding tree, the internal branch connecting two groups of taxa is tagged as suspicious (red colored) when the observed similarity score of this branch receives a negative value. A complete overview of all results is given in the Additional files 1 and 2.

In setup B, we simulated multiple sequence alignments with two internal nodes using 6-taxon trees (Figure 1b). The results lead again to the conclusion, that there is a correlation between the similarity score of the two long internal branches and tree reconstructions, which were predominantly incorrect in case of negative scores ($BL2 \geq 1.1$) (Figure 3). For example, in setup B taxa L1 and L2 are connected to the remaining taxa via two long internal branches. With increasing internal branch lengths taxa L1 and L2 occur more often as sister group instead of being paraphyletic in relation to remaining taxa. In this case, taxa L1 and L2 will share character states which have been lost in other taxa inducing a wrong sistergroup relationship based on plesiomorphies. By using the ALIGROOVE approach with the tree tagging option, correctly reconstructed short internal branches assigning taxa L1

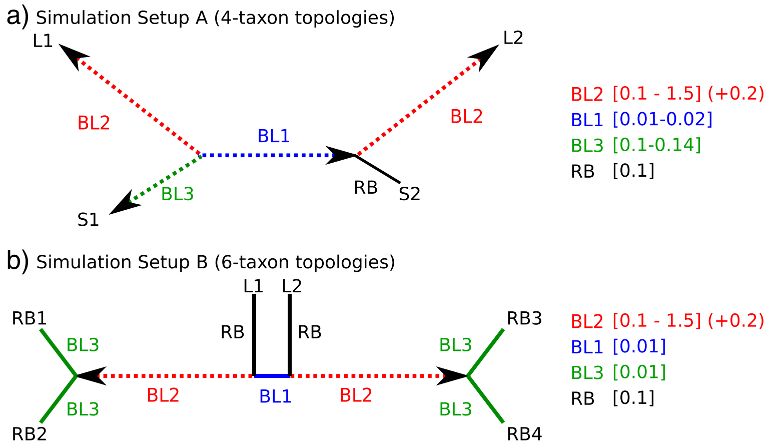


Figure 1 Simulation setup A and B. Two sets of nucleotide and amino acid data using **a)** 4-taxon (setup A) and **b)** 6-taxon topologies (setup B). Setup A contained two elongated, non directly related terminal branches ($BL2 = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$) under three different branch length conditions of remaining short branches ($BL3 = 0.1, 0.12, 0.14$ and $RB = 0.1$) and two different lengths of the very short internal branch ($BL1 = 0.01, 0.02$). Setup B contained two elongated internal branches ($BL2 = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$), separated by a short internal branch ($BL1 = 0.01$) while terminal branches are kept constant ($BL3 = 0.01$ and $RB = 0.1$).

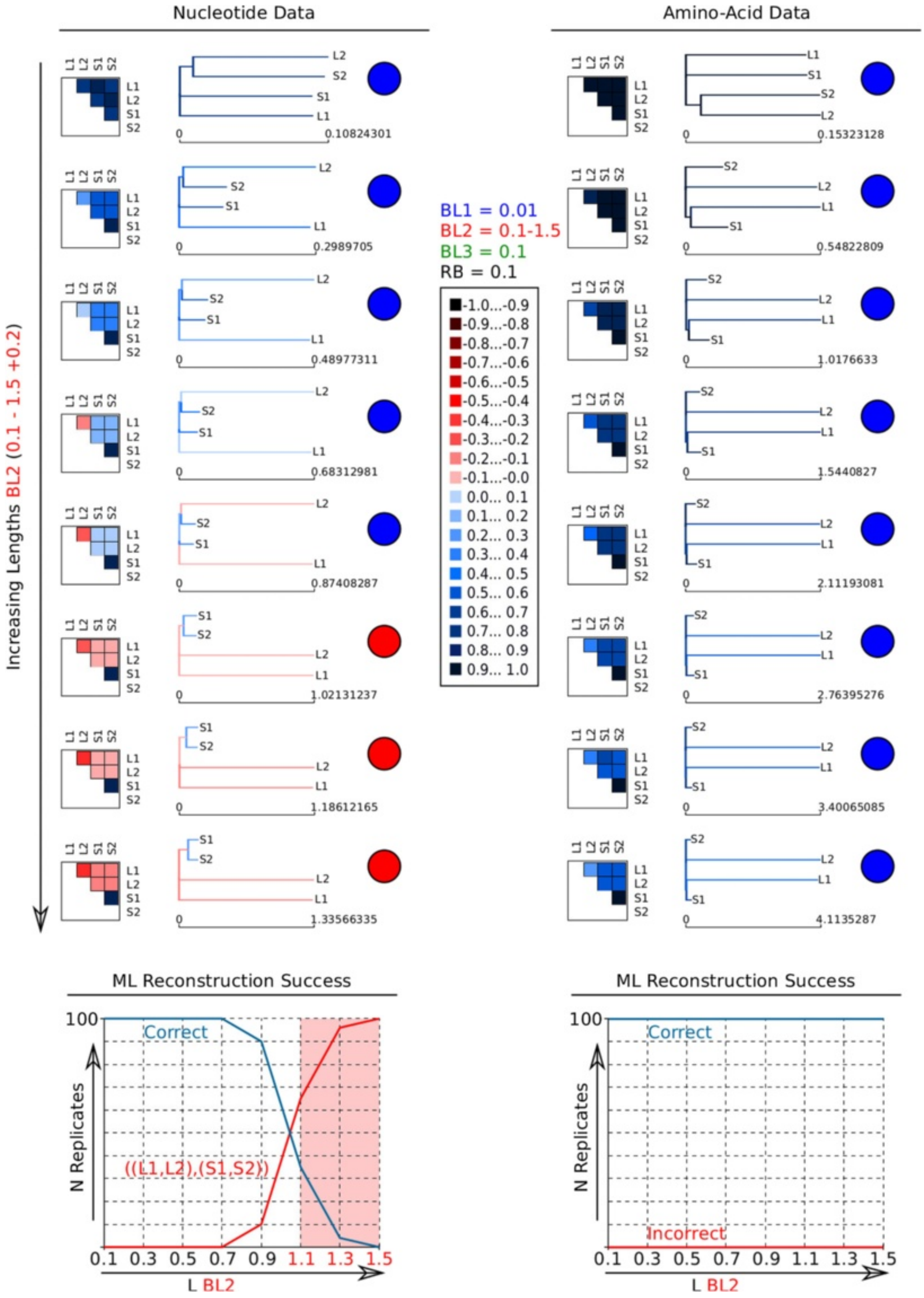


Figure 2 AliGROOVE example results of 4-taxon simulation tests (simulation setup A). AliGROOVE similarity scores and identified branch reliability of best Maximum Likelihood (ML) topologies obtained for different branch elongations of two non-directly related terminal branches (L1, L2) (Figure 1a) considering nucleotide and amino acid data (sequence length: 250,000 bp). The two graphs below show the reconstruction success in relation to the length of long branches (BL2). Note that amino acid sequences are more reliable. Colour coded similarity score ranges are shown in the center. Lacking reliability of internal branches (red internal branch) is observed for incorrect ML topologies predominating in 100 data replicates conducted for each length of BL2. Boxes with coloured squares show scores for pairwise sequence comparisons. In the corresponding topologies unreliable branches are shown in red. Circles indicate whether the topologies are correct (blue) or wrong (red). All results of 4-taxon simulations are given as Additional files 1 and 2.

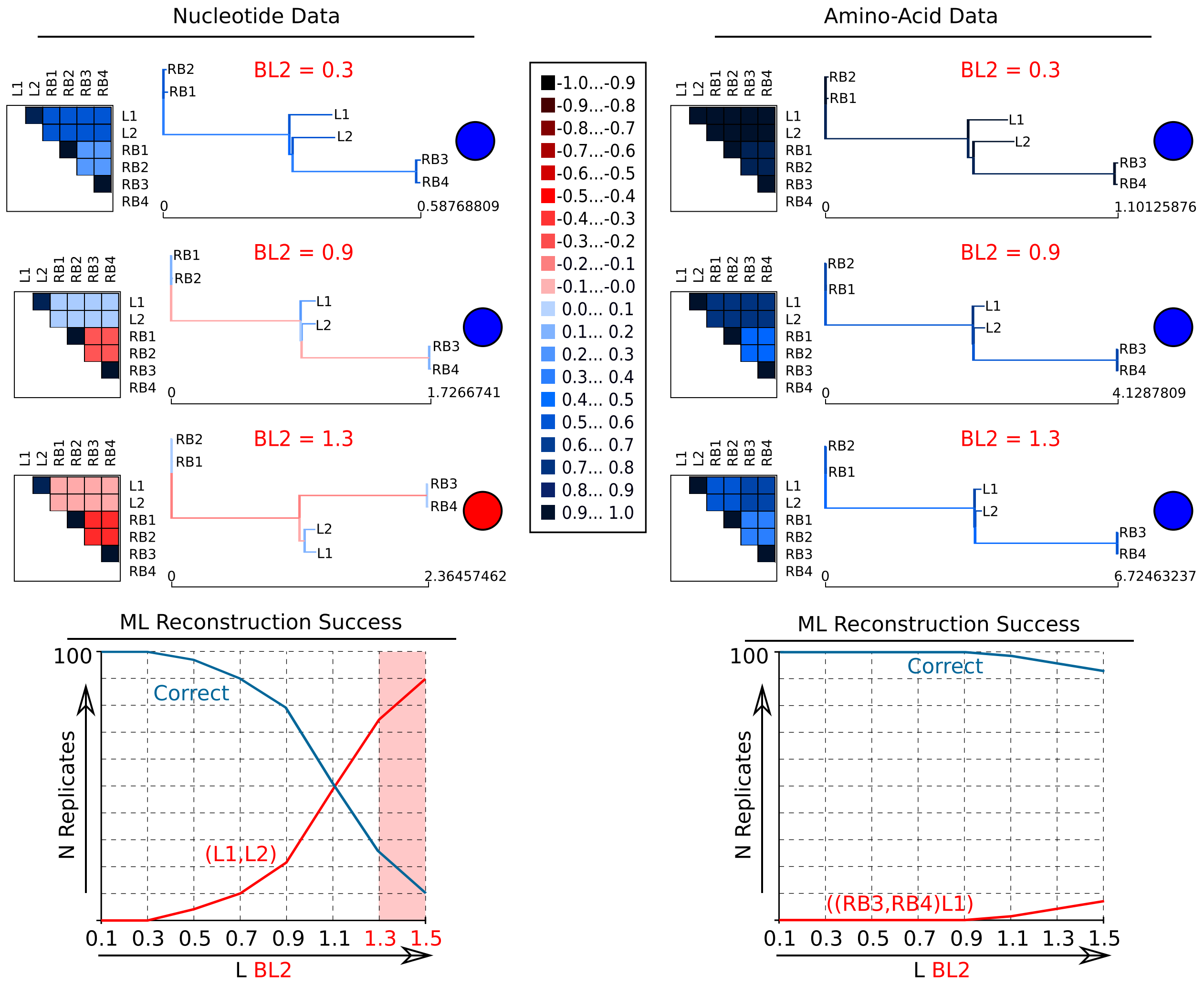


Figure 3 AliGROOVE example results of 6-taxon simulation tests (simulation setup B). Simulation setup B (Figure 1b) using six taxa (sequence length: 50,000 bp). The overall ML reconstruction success corresponds to pairwise similarity scores obtained between L1,L2 and remaining taxa (RB1-RB4) and decreases with a more frequent incorrect grouping of (L1,L2) with increasing lengths of BL2 (from dark blue to red). Colour coded similarity score ranges are shown in the center. Lacking reliability of internal branches connecting L1 and L2 scores negatively (red internal branch) in cases where the corresponding ML topologies are incorrect (red circles). All results of 6-taxon simulations are given as Additional file 3.

and L2 as paraphyletic groups have been tagged as non-suspicious, whereas incorrectly resolved short internal branches have been identified as suspicious. Whenever branch lengths are balanced, tree reconstructions have been continuously successful, which is also reflected by the similarity scores obtained for the alignments of these topologies (see Figure 3). All AliGROOVE results of the 6-taxon setup are shown in the Additional file 3.

Testing the performance on simulated data setup C

In setup C, we simulated data sets with and without indel events under four different branch length conditions of a 15-taxon topology (Figure 4) and two different models of sequence evolution (Jukes-Cantor and General Time Reversible model). Both models of sequence evolution used for data simulations led to similar AliGROOVE results (Additional file 4). Pairwise sequence comparisons of data sets simulated without indel events receive positive similarity scores in all four 15-taxon topologies and reconstructed trees are always correct (Additional file 4). Correctly aligned data sets simulated with indel events receive positive similarity scores when indel events are treated as fifth character (Figure 5). Strongly divergent sequences receive negative similarity scores if indel events are treated as ambiguous characters. The high overall reconstruction success obtained from ML analyses correlates with the AliGROOVE results obtained with indels as fifth character (Figure 5, Additional file 4). These data sets realigned receive negative similarity scores independently of the chosen indel scoring (Figure 6), whereas similarity

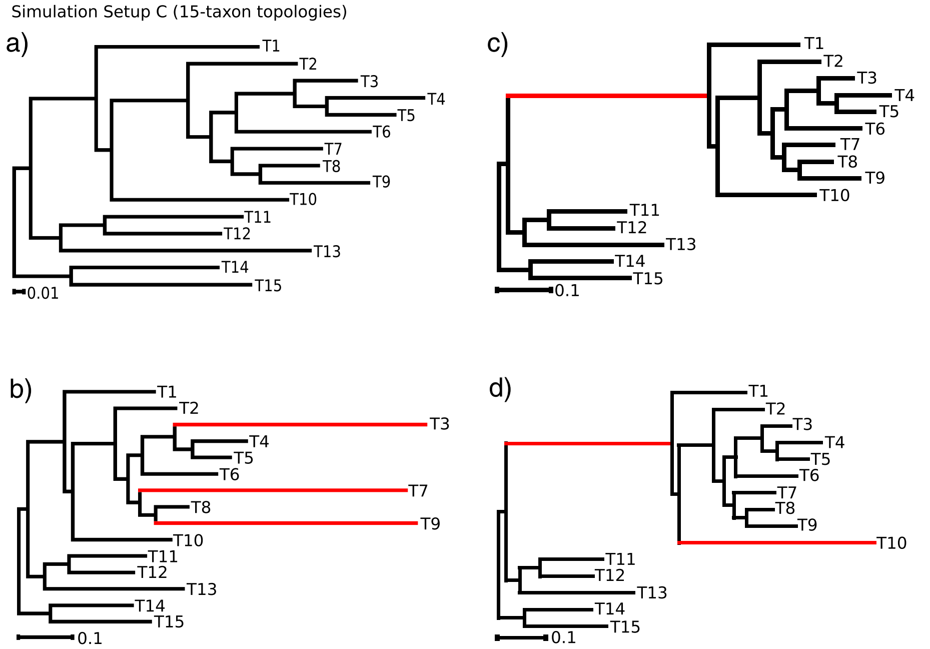


Figure 4 Simulation setup C. Nucleotide data simulation with and without indel events based on four different branch length conditions of a 15-taxon topology. Elongated branches of topology C2 **(b)**, C3 **(c)**, and C4 **(d)** in comparison to topology C1 **(a)** are highlighted red.

scores decrease under both settings compared to scorings inferred from correct multiple sequence alignments (Figures 5, 6). For all simulated branch length conditions, the incorrect placement of long internal and terminal branches could be identified successfully in realigned data sets with both scoring options (Figure 6, Additional file 4).

Testing the performance on simulated data setup D

In setup D, we simulated nucleotide and amino acid sequence data sets for a 61-taxon tree with data block sizes of 500, 1000, 1500, 2000, and 2500 sites under four different branch length conditions (BL2 = 0.1, 0.5, 0.9, 1.3) (Figure 7). Tree reconstruction analyses were performed with correct site rate heterogeneity and proportion of invariant sites model parameters. Maximum Likelihood tree reconstructions were correct for unreduced (all sequences) nucleotide and amino acid sequence data sets with branch lengths of BL2 = 0.1 and alignment lengths above 500 sites. For all other setups, Maximum Likelihood failed to find the correct tree for the nucleotide sequence data sets and delivered correct trees for amino acid sequence data sets only in case of data blocks larger than 2000 sites and branch length BL2 less or equal 0.9 (Additional file 5, Additional file 6). At least one of the five long branches was always misplaced in incorrect trees (Figure 8, Additional file 5, Additional file 6). Thus, the seven highly divergent sequences (T16, T25, T27, T39, T40, T41, and T42) were problematic in nucleotide and most amino acid sequence data sets.

With the AliGROOVE algorithm, the highly divergent seven nucleotide sequences did not consistently cause negative scores in all pairwise sequence comparisons if branch lengths of BL2 were set to 0.5, but got almost always negative scores if BL2 were set to ≥ 0.9 and data blocks to >1000 sites (Additional file 5). With amino acid datsets, the seven highly divergent sequences got only positive scores in all pairwise sequence comparisons, independently of the tree reconstruction success (Additional file 6).

The tree tagging algorithm tagged all highly divergent nucleotide sequences and associated long branches as unreliable for branch lengths BL2 $\geq$ 0.9, and tagged all incorrectly placed nucleotide sequences and associated long branches as unreliable if sequence length of nucleotide data blocks was set to 2500 sites and branch lengths BL2 = 0.5. In case of shorter data blocks and branch lengths set to BL2 = 0.5, tagging was less consistently correct (Additional file 5). For amino acid datsets, non of the seven highly divergent sequences and associated long branches were tagged as unreliable.

These results also apply to the concatenated nucleotide and amino acid supermatrix data sets which consist of all data blocks. The AliGROOVE pairwise distance similarity matrix of the concatenated nucleotide supermatrix shows the seven highly divergent sequences mostly red colored, however despite being misplaced on the tree,

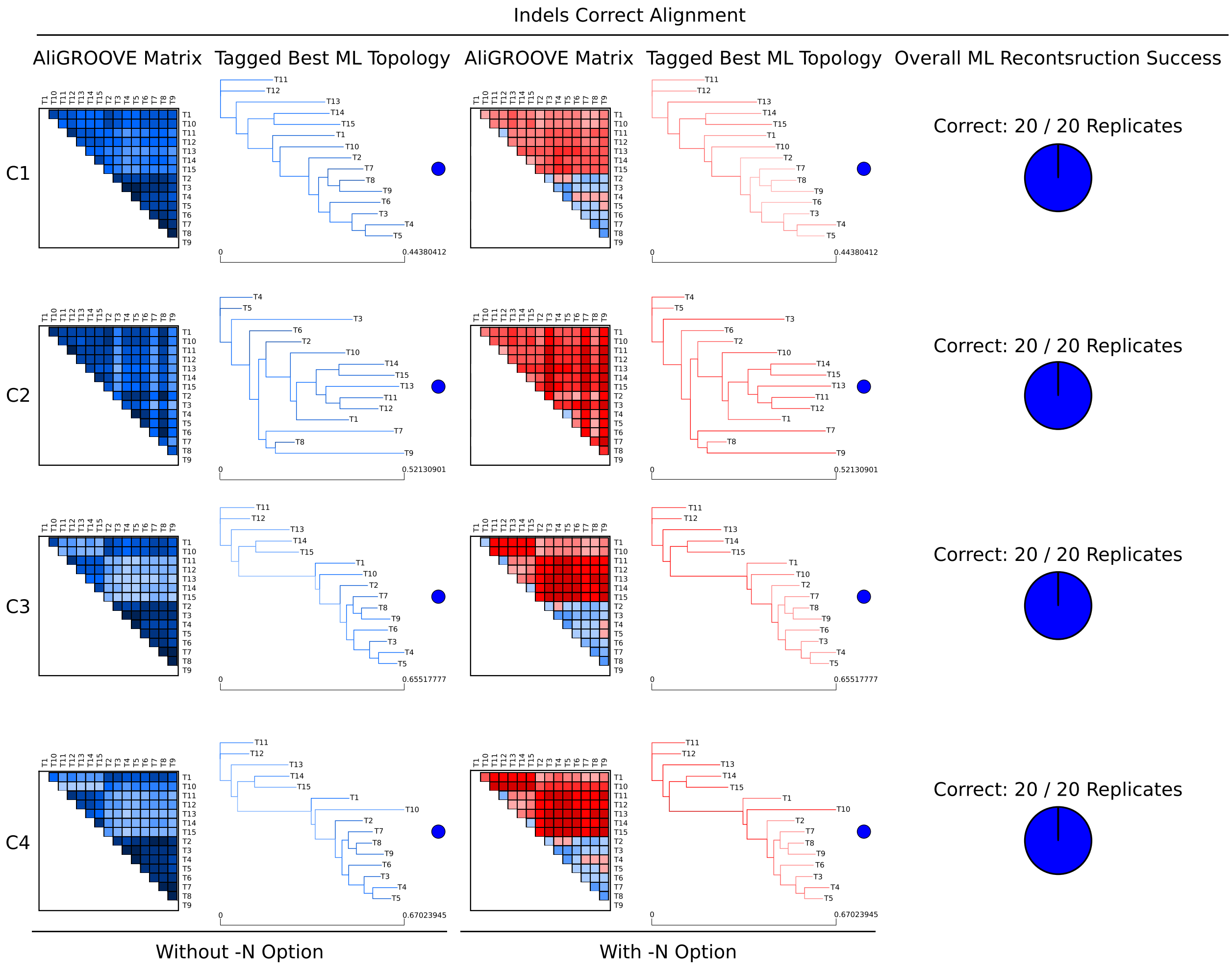


Figure 5 AliGROOVE example results of correctly aligned 15-taxon simulation tests (simulation setup C). AliGROOVE similarity scores and identified branch reliability of best Maximum Likelihood (ML) topologies as well as the obtained overall ML reconstruction success for four different branch length conditions of a 15-taxon topology (Figure 4) under the GTR model of sequence evolution used for data simulation and tree reconstruction. Correctly aligned data sets have been simulated with indel events and analysed by AliGROOVE without -N option (indel events are treated as fifth character state) or with -N option (indel events treated as ambiguous characters).

the branches associated with this seven highly divergent sequences are not consistently tagged as suspicious (Figure 8). With the amino acid supermatrix, the highly divergent sequences are not highlighted in the distance matrix and branches associated with these sequences are not tagged as suspicious, despite being wrong. For both nucleotide and amino acid supermatrices the exclusion of the seven divergent sequences led to correct topologies (Additional file 7).

In general, the AliGROOVE tagging algorithm is optimistic concerning the reliability of branching patterns and never tags a branch as unreliable if in fact correct.

Testing the performance with empirical mitochondrial data

We used mitochondrial DNA sequence data downloaded from the NCBI genome data base for 53 chelicerate ingroup taxa and eight myriapod outgroup taxa. It is known that among chelicerates the Acari (mites, ticks) are problematic [19,20]. AliGROOVE analyses of the concatenated supermatrix file and of gene partitions showed that pairwise sequence comparisons involving mite sequences received negative scores while pairwise comparisons between other sequences achieved mainly positive scores (Figure 9). Among all gene partitions, only the Cytochrome Oxidase I (COI) DNA sequence alignment shows positive similarity scores for nearly all taxon comparisons. While nearly all pairwise sequence comparisons of the ATP Synthase Subunit 6 (ATP6) yielded negative similarity scores, impacts of random sequence similarity and alignment ambiguity vary for mite subgroups in Cytochrome b (Cytb), Cytochrome Oxidase II (COII), and Cytochrome Oxidase III (COIII). For Cytb,

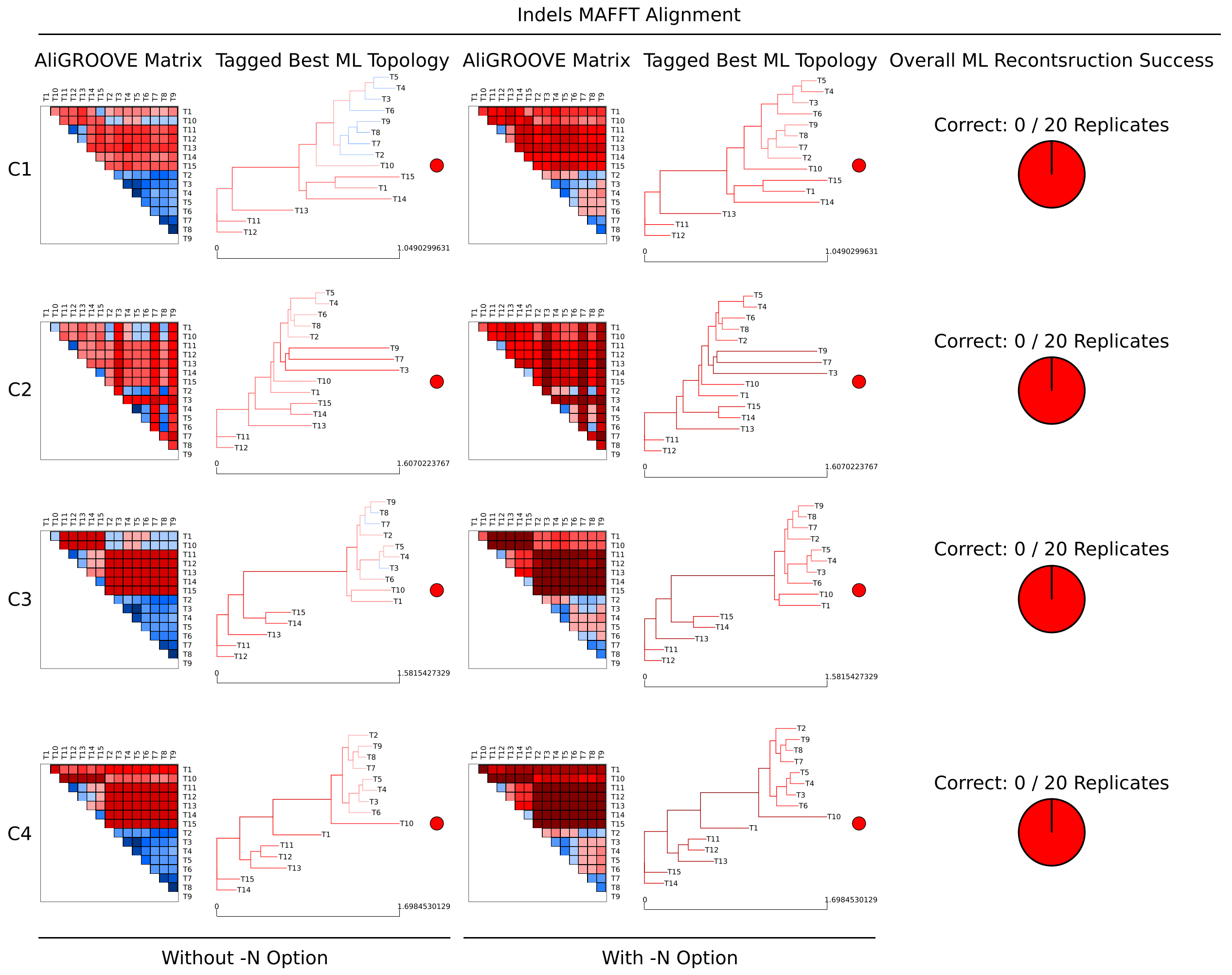


Figure 6 AliGROOVE example results of MAFFT realigned 15-taxon simulation tests (simulation setup C). AliGROOVE similarity scores and identified branch reliability of best Maximum Likelihood (ML) topologies as well as the obtained overall ML reconstruction success for four different branch length conditions of a 15-taxon topology (Figure 4) under the GTR model of sequence evolution used for data simulation and tree reconstruction. Realigned (MAFFT) data sets have been simulated with indel events and analysed by AliGROOVE without -N option (indel events are treated as fifth character state) or with -N option (indel events treated as ambiguous characters).

mite sequences are not highly divergent whereas specific mite subgroups appear strongly misaligned in COII (*Dermatophagoidae*) and COIII (*Panonychus* & *Tetranychus*). These three mite subgroups are also scored constantly negative in pairwise comparisons of the concatenated supermatrix. Nevertheless, the phylogenetic position of *Dermatophagoidae*, *Panonychus* and *Tetranychus* receives high bootstrap support in the tree reconstruction based on the concatenated supermatrix. The supermatrix sister group relationship of Acariformes and Ricinulei with a bootstrap support of 36 was as expected tagged as unreliable (red colored) with AliGROOVE. However, the supermatrix clade ((Ricinulei, Acariformes), Parasitiformes) that received a bootstrap support of 99 was tagged as unreliable as well (Figure 9).

Discussion

It has been shown that traditional masking of entire sequence alignment blocks can improve the signal-to-noise ratio or tree-likeness in sequence alignments. Here, we show that the sliding window approach as it is used in ALISCORE [5,7] can be modified to identify single taxa or subsets of taxa which show predominantly randomized sequence similarity in comparison with other taxa (Figure 9). Masking of these taxa can also improve the signal-to-noise ratio in sequence alignments. The approach implemented in AliGROOVE can be used to test the reliabilities of reconstructed topologies and to identify unreliable node support in a user specified tree (Figures 2, 3 5, 6, 8, 9, Additional files 1, 2, 3, 4, 5, 6). This possibility offers a convenient way of studying node

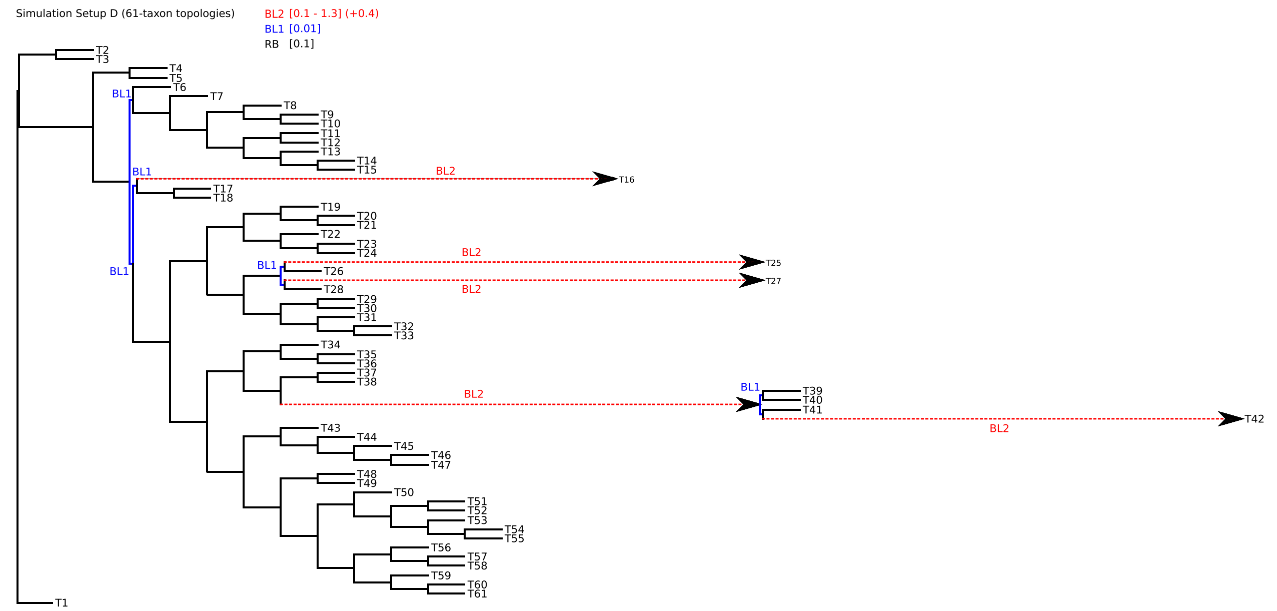


Figure 7 Simulation setup D. 61-taxon setup used for simulation of nucleotide and amino acid gene partitions. Alignment lengths of single gene partitions were set to 500, 1000, 1500, 2000, and 2500 character state positions. To simulate different substitution rates, internal and terminal branches were stepwise increased for each gene partition length (highlighted red), ranging from 0.1 to 1.3 ($BL2 = 0.1, 0.5, 0.9, 1.3$). Internal branches in close vicinity to elongated branches are kept very short ($BL1 = 0.01$). All remaining branches are kept equally long ($RB = 0.1$).

support in a given tree and multiple sequence alignment complementary to conventional bootstrap analyses. The identification of taxonomic subsets offers the possibility to mask only taxonomic sub-blocks of multiple sequence alignments that clearly contain the least signal due to alignment ambiguity, sequence saturation or excessive divergence.

Results of the analyses of simulated nucleotide data sets with indel events and/or missing data (coded as gaps) and correct sequence alignment showed that the AliGROOVE

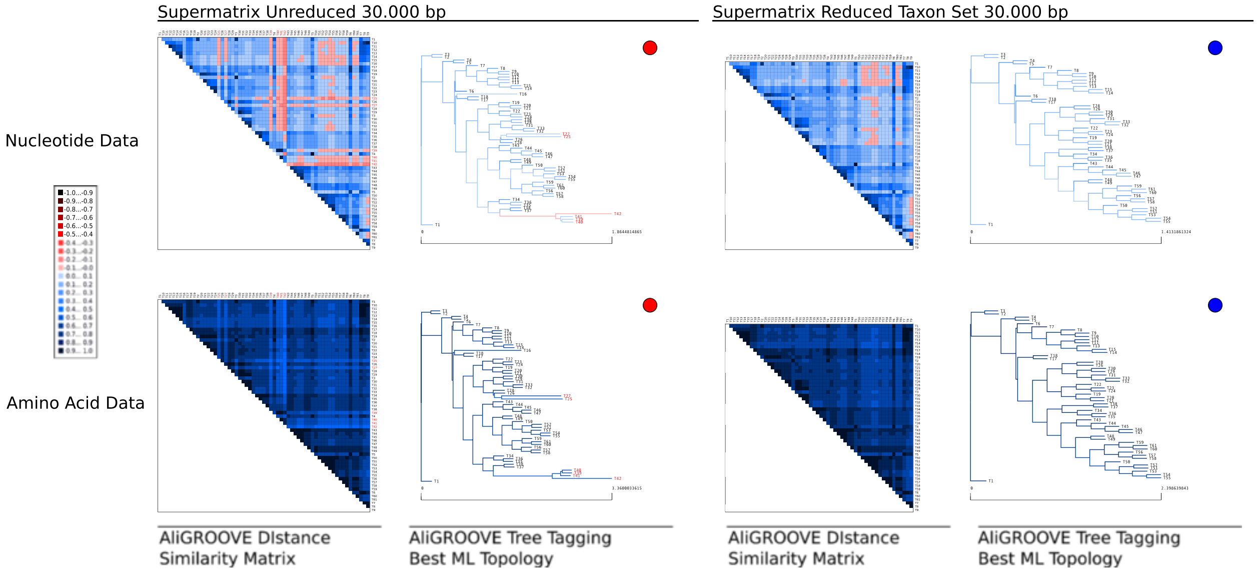


Figure 8 AliGROOVE results of concatenated gene analyses on unreduced and reduced 61-taxon simulation setup D. AliGROOVE similarity score distance matrices and associated ML topologies of the original supermatrix and the taxon reduced supermatrix, including all 20 gene partitions. The darker blue the colour coded similarity scores in AliGROOVE matrices, the higher the non-randomized accordancy between pairwise sequence comparisons. Red indicates the opposite. Tagged branch reliability of associated best ML topologies is given next to each matrix. Correct reconstructed topologies are pointed blue, incorrect trees red. Names of incorrectly resolved sequences are highlighted red.

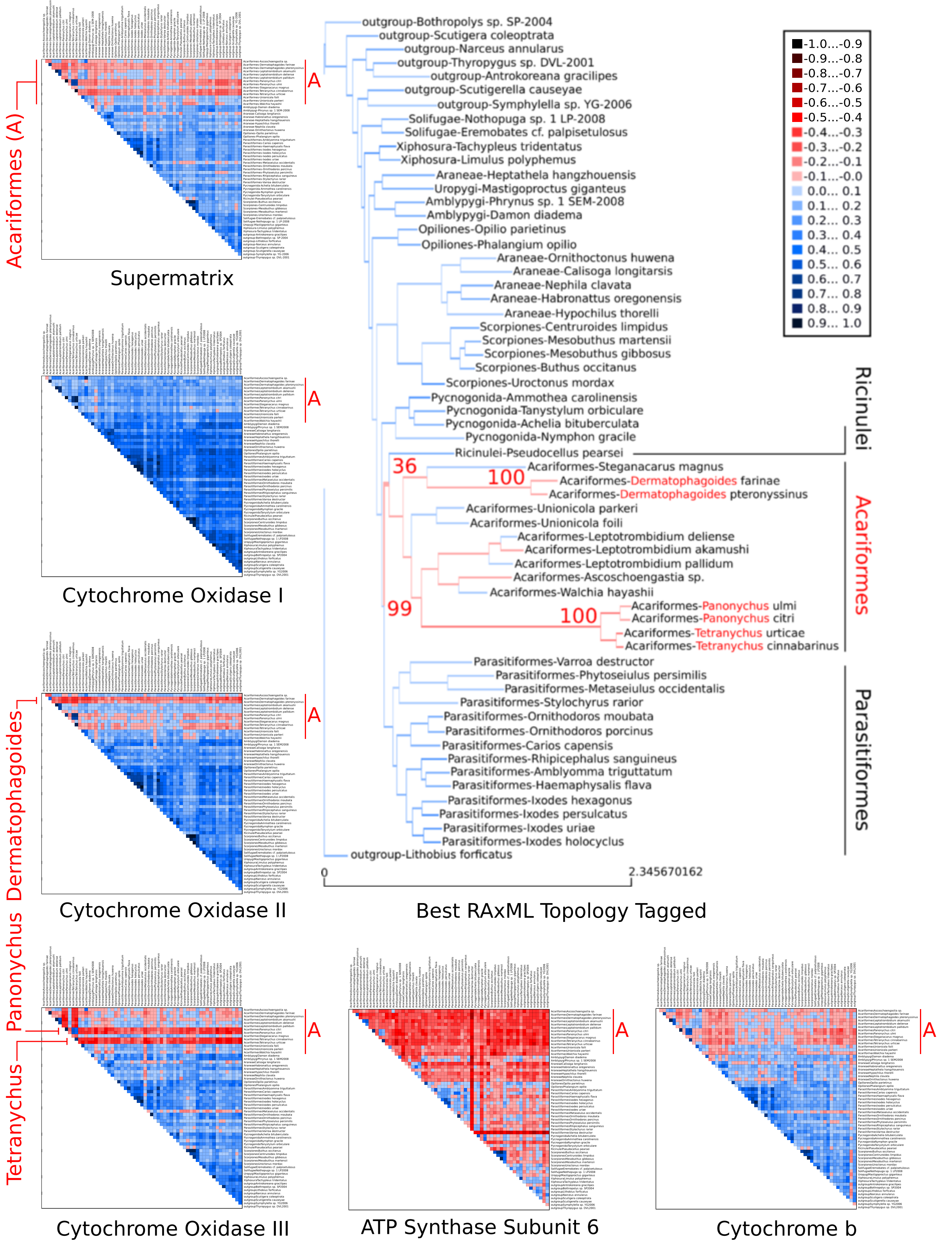


Figure 9 AliGROOVE results of concatenated and single gene analyses on mitochondrial data of Chelicerata taxon groups. Except of Cytochrome Oxidase I (mainly positive pairwise similarity scores) and ATP6 (mainly negative pairwise similarity scores), AliGROOVE identified mainly negative similarity scores in pairwise comparisons whenever sequences of Acariformes (A) are involved. Although subgroups within Acariformes get a higher bootstrap support in the best ML tree using the supermatrix data (shown here), they are tagged with AliGROOVE as unreliable. The misleading information accumulates for these taxa especially in the COII and COIII sequences, while ATP6 is generally to noisy. In the concatenated data set ("supermatrix") the misleading patterns are still dominant.

approach correctly identified excessively divergent sequences with treating indels as fifth character state (Figure 5, Additional file 4). After realigning these data, the difference between treating indels as fifth or ambiguous character state vanished. This may be explained by misplaced indels during the process of realignment which should be better treated as ambiguous character states. For empirical data, in particular indel-rich data in which we cannot discriminate between misplaced and correctly placed indels, this result implies that indels should be treated as ambiguous character state or completely removed from phylogenetic analyses [2,4,21].

The results concerning simulation setup D merit additional discussions. In these analyses, branch length differences between clades have been pushed to the extreme. With nucleotide sequences, the AliGROOVE algorithm correctly tagged misplaced branches if BL2 ≥ 0.9. With amino acid data even these long branches were never tagged as unreliable despite being incorrectly placed. Apparently, detectable substitutional saturation accumulated only if branch lengths BL2 were ≥ 0.9, and extremely short internal BL1=0.01 were insufficient to accumulate any signal. This phenomenon was pronounced for amino acid data. The extremely short internal branch lengths of BL1=0.01 can be interpreted as hard polytomies, for which tree reconstructions cannot deliver correct results. However, the frequency of hard polytomies limiting the application of the AliGROOVE algorithm in empirical data is currently unknown.

The mitochondrial DNA sequence data set of chelicerates shows strong heterogeneity of sequence divergence as indicated in the similarity matrix (Figure 9). Specimens of Acariformes display mostly random similarity to all other sequences. This observation implies that Acariformes cannot be robustly placed in the tree or are potentially misplaced despite robust bootstrap support. This is exactly what we see in the tree reconstruction using the concatenated supermatrix data set, as Acariformes are sister group to Ricinulei and form together with Parasitiformes the sister group to Pycnogonidae. This grouping which is considered implausible by many specialists [19,20,22,23] gets a high bootstrap support. The questionable sister group relationship between Ricinulei and Acariformes has been identified with AliGROOVE and is tagged as suspicious in the topology inferred from the supermatrix. The AliGROOVE algorithm clearly identified the most problematic sequences and gene partitions in the data set and demonstrates its usability with this data.

Conclusions

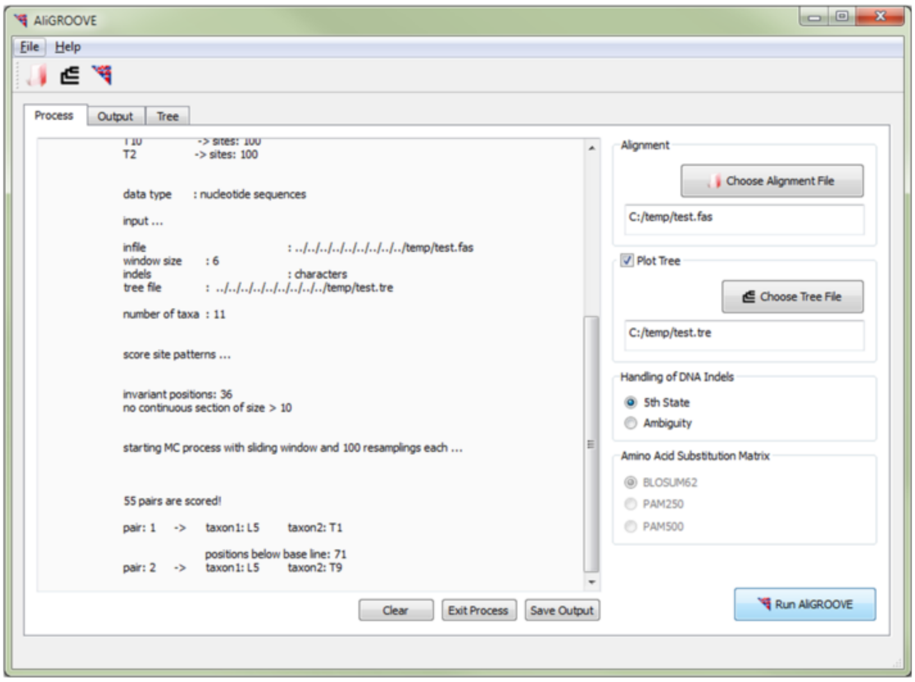
The analyses of the simulated and the empirical data show that the sliding window approach identifies relevant sources of reconstruction error. Therefore, we suggest our method as an important complement to all character based masking approaches in phylogenetics. It offers the possibility to exclude taxa or gene partitions based on a formal argument instead of excluding taxa based exclusively on the evaluation of branch lengths. The exclusion or exchange of conflicting sequences and/or gene partitions improves the signal-to-noise ratio of the alignment and, as a consequence of this, can lead to less biased, more realistic trees. The simple usage of the AliGROOVE program via graphical user interface (Figure 10) facilitates the identification of potentially problematic taxa or gene partitions for users which feel uncomfortable with command line based software while the alternatively available command line version of AliGROOVE can be easily integrated into automated analysis pipelines. AliGROOVE has no maximum limit in taxon number or sequence length.

Material and methods

Simulated data setup A & B

To test the efficiency of AliGROOVE we designed two sets of nucleotide and amino acid sequence data using 4-taxon and 6-taxon trees (Figure 1). The topology of the 4-taxon setup (setup A, Figure 1a) contained two long branches of unrelated taxa (with branch lengths $BL2 = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$) under three different branch length conditions for the other two short terminal branches ($BL3 = 0.1, 0.12, 0.14$ and $RB = 0.1$) and two different lengths of the short internal branch ($BL1 = 0.01, 0.02$). The 6-taxon setup (setup B, Figure 1b) contained two long internal branches ($BL2 = 0.1, 0.3, 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$), separated by a short internal branch ($BL1 = 0.01$) while the lengths of terminal branches are kept constant ($BL3 = 0.01$ and $RB = 0.1$). For both test setups, 100 alignments were generated for each step of $BL2$ branch elongation. Sequence length of each alignment of setup A was set to 250,000 character state positions and for setup B to 50,000 character state positions to reduce the calculation time. All alignments were generated with INDELible v.1.03 [24]. In order to simulate nucleotide sequence data we used the Jukes-Cantor model (JC) of sequence evolution and for amino acid sequence data the BLOSUM62 substitution model. All data were simulated with among site rate variation (ASRV), using a mixed-distribution model with a shape parameter $\alpha = 1.0$, and a proportion of invariant sites $\rho_{inv} = 0.3$. ASRV was modelled using a continuous Γ-rate distribution while indel events were not simulated.

Trees of simulated data were inferred with PhyML_ 3.0_linux64 [25,26]. We analyzed the data with a mixed-distribution model (JC+Γ+I) and correct parameter values ($\alpha = 1.0$, $\rho_{inv} = 0.3$), except for the categorization of the gamma distribution. The number of relative substitution rate categories was set to four (c = 4) and tree topologies and branch lengths were optimized. Maximum



```
T10           -> sites: 100
T2            -> sites: 100


data type     : nucleotide sequences

input ...

infile                          : ../../../../../../../../../temp/test.fas
window size     : 6
indels                          : characters
tree file       :  ../../../../../../../../../temp/test.tre

number of taxa  : 11


score site patterns ...


invariant positions: 36
no continuous section of size > 10


starting MC process with sliding window and 100 resamplings each ...



55 pairs are scored!

pair: 1    ->    taxon1: L5      taxon2: T1

                 positions below base line: 71
pair: 2    ->    taxon1: L5      taxon2: T9
```

Alignment: C:/temp/test.fas

☑ Plot Tree: C:/temp/test.tre

Handling of DNA Indels
● 5th State
○ Ambiguity

Amino Acid Substitution Matrix
● BLOSUM62
○ PAM250
○ PAM500

Figure 10 Graphical user interface (GUI) version of AliGROOVE. Overview of the AliGROOVE process window.

Likelihood analyses were performed and evaluated with a Perl pipeline. For each branch length-combination, we generated 100 data replicates and recorded the frequencies of correct and incorrect tree reconstructions using correct alignments and nearly correct substitution models (Figures 2, 3, Additional files 1, 2, 3).

Simulated data setup C

To test the efficiency of AliGROOVE when sequences contain gaps and missing data we simulated nucleotide sequence data sets for four different 15-taxon topologies (Figure 4). The -N option of AliGROOVE allows to toggle between scoring gaps as fifth character state or as ambiguity. The efficiency of AliGROOVE with and without the usage of the -N option was tested on correct alignments (Figure 5) and on realigned data sets using MAFFT [27,28] under default values (Figure 6). Additionally, alignments were also simulated without indel events under otherwise identical parameter settings. Topologies differed only in branch lengths. While topology C1 (Figure 4a) consisted of more or less well balanced branch lengths, three terminal branches (Taxon T3, T7, T9) have been strongly increased in topology C2 (Figure 4b). One internal branch separating taxa T1 to T10 from remaining taxa has been strongly increased in topology C3 (Figure 4c), and one internal branch separating taxa T1 to T10 from remaining taxa as well as an addtional terminal branch (taxon T10) has been strongly increased in topology C4 (Figure 4d). Alignment lengths of simulation setup C were set to 50,000 sites. All data were simulated with ASRV, using a mixed-distribution model with a shape parameter $\alpha = 0.5$, and a proportion of invariant sites $\rho_{inv} = 0.1$. ASRV was modeled using a continuous Γ-rate distribution while indel events were simulated using a Lavalette Distribution where the maximum indel length was set to 20. Insertion and deletion rate were both set to 0.2. Single state frequencies of GTR simulations were set to $T = 0.35$, $C = 0.15$, $A = 0.35$, $G = 0.15$.

Trees of simulated data were inferred with PhyML_3.0_linux64 [25,26] using either the JC or GTR model of sequence evolution (depending on the substitution model used for data simulations) with a mixed-distribution model by estimating the α shape parameter and the proportion of invariant sites. The number of gamma shape rate categories was set to four (c = 4) and tree topologies and branch lengths were optimized. Maximum Likelihood analyses were performed and evaluated with a Perl pipeline. For each topology and AliGROOVE setting, we generated 20 data replicates and recorded the frequencies of correct and incorrect tree reconstructions (Figures 5, 6, Additional file 4).

Simulated data setup D

To test the efficiency of AliGROOVE on large data sets and more realistic data block lengths, we simulated five different data block lengths of nucleotide and amino acid sequence data for a 61-taxon topology under four different internal and terminal branch length conditions

(Figure 7). Alignment lengths of single data blocks were set to 500, 1000, 1500, 2000, and 2500 sites. To simulate different substitution rates for specific branches we stepwise increased single internal and terminal branches for data block length from 0.1 to 1.3 ($BL2 = 0.1, 0.5, 0.9, 1.3$). To increase rate heterogeneity between long branches and nearest-neighbour branches we kept internal branches very short ($BL1 = 0.01$). All remaining branches are kept at $RB = 0.1$. Our simulation setup lead to a total number of 20 gene partitions with each alignment length of data blocks being represented four times, each time with another substitution rate for specific taxa due to increased branch lengths of the data underlying topology.

Like in simulation setup A and B we simulated all data with ASRV, using a mixed-distribution model with a shape parameter $\alpha = 1.0$, and a proportion of invariant sites $\rho_{inv} = 0.3$. ASRV was modeled using a continuous Γ-rate distribution. Indel events were not simulated. In order to simulate nucleotide sequence data we used the Jukes-Cantor model (JC) of sequence evolution and the BLOSUM62 substitution model for amino acid sequence evolution. For sequence concatenation we used FASconCAT v1.0 [29].

Trees of simulated data were again reconstructed with PhyML_3.0_linux64 [25,26] using the JC of sequence evolution (JC+Γ+I) with correct rate heterogeneity and invariant site proportion parameters ($\alpha = 1.0$, $\rho_{inv} = 0.3$). The number of gamma shape rate categories was set to four (c = 4). All Maximum Likelihood analyses were performed and evaluated with a Perl pipeline.

AliGROOVE was tested on complete as well as reduced data blocks and supermatrices. Reduced sequence blocks and supermatrices were used to test the overall quality improvement of given data and associated trees after removing sequences which have been identifed as potentially unreliable in the majority of the AliGROOVE analyses (Additional files 5, 6, 7, 8).

Empirical data

We used AliGROOVE without the -N option (indels coded as fifth character state) on a concatenated super alignment (5082 character state positions) as well as on corresponding single gene data sets of five mitochondrial genes (Atp6 ↪ 696 character state positions, COI ↪ 1575 character state positions, COII ↪ 783 character state positions, COIII ↪ 861 character state positions, and Cytb ↪ 1167 character state positions) downloaded from the NCBI genome data base for 53 chelicerate ingroup taxa and eight myriapod outgroup taxa. Single mitochondrial genes were aligned with ClustalW [30] and concatenated with FASconCAT [29]. The best ML topology of the mitochondrial data set was estimated using RAxML_7.2.2 [31] and the GTR+Γ model. Single node support has been evaluated by performing 1000 bootstrap replicates (Figure 9).

Computation time

Time complexity of AliGROOVE is given by:

$$O\left(M * N^2\right) \tag{3}$$

M means the sequence length of a given alignment, N the total number of aligned taxon sequences. For example, the AliGROOVE computation time of a single 4-taxon alignment with sequence lengths of 250.000 character states took 809 seconds using a GenuineIntel(R) Core(TM) i7, 2.60GHz processor. The computation time of a 64-taxon data set with an alignment length of 2500 characters, conducting 1830 pairwise sequence analyses, took 2578 seconds.

Implementation of AliGROOVE

AliGROOVE is implemented in Perl and runs on Linux, Mac OS, and Windows operating systems. It can be used via command line or graphical user interface (GUI). The GUI of AliGROOVE (Figure 10) is based on Qt, a cross-platform application and GUI framework in C^{++}.

Availability of supporting data and requirements

- **Project name:** AliGROOVE – visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support
- **Project home page:** http://zfmk.de/web/Forschung/Abteilungen/AG_Wgele/Software/AliGROOVE/index.en.html
- **Operating system(s):** Platform independent
- **Programming language:** Perl
- **Other requirements:** Perl 5.0 or higher
- **License:** GNU GPL version 2
- **Any restrictions to use by non-academics:** No restrictions

Additional files

Additional file 1: Complete results of 4-taxon simulations based on nucleotide data. Graphical result plots of all AliGROOVE analyses performed for nucleotide data based on 4-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 2: Complete results of 4-taxon simulations based on amino acid data. Graphical result plots of all AliGROOVE analyses performed for amino acid data based on 4-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 3: Complete results of 6-taxon simulations. Graphical result plots of all AliGROOVE analyses performed for nucleotide and amino acid data based on 6-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 4: Complete results of 15-taxon simulations. Graphical result plots of all AliGROOVE analyses performed for complete and indel events included nucleotide data based on 15-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 5: Complete results of unreduced 61-taxon gene partitions simulations based on nucleotide data. Graphical result plots of all AliGROOVE analyses performed for unreduced nucleotide gene partitions based on 61-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 6: Complete results of unreduced 61-taxon gene partitions simulations based on amino acid data. Graphical result plots of all AliGROOVE analyses performed for unreduced amino acid gene partitions based on 61-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 7: Complete results of taxon reduced gene partitions based on 61-taxon nucleotide data simulations. Graphical result plots of all AliGROOVE analyses performed for taxon reduced nucleotide gene partitions based on 61-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Additional file 8: Complete results of taxon reduced gene partitions based on 61-taxon amino acid simulations. Graphical result plots of all AliGROOVE analyses performed for taxon reduced amino acid partitions based on 61-taxon topologies. The pdf document can be opened with pdf readers like AdobeAcrobatReader, Xpdf, or DocumentViewer.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
PK and BM developed the AliGROOVE algorithm and conceived the study. BM and PK and SAM programmed AliGROOVE. SAM developed the graphical user interface of AliGROOVE. PK designed the setup. PK and CG performed all analyses. PK, BM, SAM, and JWW discussed and wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
We would like to thank Birthe Thormann for proof reading, Christoph Mayer for helping to determine the computational time complexity of AliGROOVE, all members of the ZFMK for inspiring discussions, and two anonymous reviewers for helpful comments.

Author details
[1]Zoologisches Forschungsmuseum A. Koenig, Adenauerallee 160-163, 53113 Bonn, Germany. [2]University of Amsterdam, Amsterdam, Netherlands.

Received: 8 November 2013 Accepted: 14 August 2014
Published: 30 August 2014

References

1. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17**(4):540–552.
2. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56**(4):564–577.
3. Dress AWM, Flamm C, Fritzsch G, Grünewald S, Kruspe M, Prohaska SJ, Stadler PF: **Noisy: identification of problematic columns in multiple sequence alignments.** *Algorithms Mol Biol* 2008, **3:**7.
4. Capella-Gutiérez S, Silla-Martinez JM, Gabaldón T: **trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.** *Bioinformatics* 2009, **25**(15):1972–1973.
5. Misof B, Misof K: **A Monte Carlo approach successfully identifies randomness in multiple sequence alignments: a more objective means of data exclusion.** *Syst Biol* 2009, **58:**21–34.
6. Criscuolo A, Gribaldo S: **BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments.** *BMC Evol Biol* 2010, **10:**210.
7. Kück P, Meusemann K, Dambach J, Thormann B, von Reumont B, Wägele JW, Misof B: **Parametric and non-parametric masking of randomness in sequence alignments can be improved and leads to better resolved trees.** *Front Zool* 2010, **7:**10.
8. Wu M, Chatterji S, Eisen JA: **Accounting for alignment uncertainty in phylogenomics.** *PLoS ONE* 2012, **7:**e30288.
9. Hartmann S, Vision TJ: **Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment?** *BMC Evol Biol* 2008, **8:95:**S13.
10. Schwarzer J, Misof B, Tautz D, Schliewen UK: **The root of the East African cichlid radiations.** *BMC Evol Biol* 2009, **9:**186.
11. Simon S, Strauss S, von Haeseler A, Hadrys H: **A phylogenomic approach to resolve the basal pterygote divergence.** *Mol Biol Evol* 2009, **26**(12):2719–2730.
12. Meusemann K, von Reumont BM, Simon S, Roeding F, Kück P, Strauss S, Ebersberger I, Walzl M, Pass G, Breuers S, Achter V, von Haeseler A, Burmester T, Hadrys H, Wägele JW, Misof B: **A phylogenomic approach to resolve the arthropod tree of life.** *Mol Biol Evol* 2010, **27**(11):2451–2464.
13. Murienne J, Edgecombe G, Giribet G: **Including secondary structure, fossils and molecular dating in the centipede tree of life.** *Mol Phylogenet Evol* 2010, **57:**301–313.
14. Dinapoli A, Zinssmeister C, Klussmann-Kolb A: **New insights into the phylogeny of the Pyramidellidae (Gastropoda).** *J Mollus Stud* 2011, **77:**1–7.
15. Kück P, Hita-Garcia F, Misof B, Meusemann K: **Improved phylogenetic analyses corroborate a plausible position of Martialis Heureka in the ant tree of life.** *PLoS ONE* 2011, **6**(6):e21031.
16. Nesnidal MP, Heimkampf M, Bruchhaus I, Hausdorf B: **The complete mitochondrial genome of Flustra foliacea (Ectoprocta, Cheilostomata) - compositional bias affects phylogenetic analyses of lophotrochozoan relationships.** *BMC Genomics* 2011, **12:**572.
17. Privman E, Penn O, Pupko T: **Improving the performance of positive selection inference by filtering unreliable alignment regions.** *Mol Biol Evol* 2012, **29:**1–5.
18. von Reumont BM, Jenner RA, Wills MA, Dell'Ampio E, Pass G, Ebersberger I, Meyer B, Koenemann S, Iliffe TM, Stamatakis A, Niehus O, Meusemann K, Misof B: **Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda.** *Mol Biol Evol* 2012, **29**(3):1031–1045.
19. Dabert M, Witalinski W, Kazmierski A, Olszanowski Z, Dabert J: **Molecular phylogeny of acariform mites (Acari, Arachnida): Strong conflict between phylogenetic signal and long-branch attraction artifacts.** *Mol Phylogenet Evol* 2010, **56:**222–241.
20. Pepato AR, daRocha CEF, Dunlop JA: **Phylogenetic position of the acariform mites: sensitivity to homology assessment under total evidence.** *BMC Evol Biol* 2010, **10:**235.
21. Capella-Gutiérez S, Gabaldón T: **Measuring guide-tree dependency of inferred gaps in progressive aligners.** *Bioinformatics* 2013, **29**(8):1011–1017.
22. Dunlop J, Alberti G: **The affinities of mites and ticks: a review.** *J Zool Syst Evol Res* 2008, **46:**1–18.
23. Talarico G, Michalik P: **Spermatozoa of an Old World Ricinulei (Ricinoides karschii, Ricinoidae) with notes about the relationships of Ricinulei within the Arachnida.** *Tissue Cell* 2010, **42**(6):383–390.
24. Fletcher W, Yang Z: **INDELible: A flexible simulator of biological sequence evolution.** *Mol Biol Evol* 2009, **26**(8):1879–1888.
25. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52**(5):696–704.
26. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **PhyML 3.0: New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59**(3):307–321.
27. Katoh K, Kuma Ki, Hiroyuki T, Miyata T: **MAFFT version 5: Improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Res* 2005, **33**(2):511–518.
28. Katoh K, Toh H: **Recent developments in the MAFFT multiple sequence alignment program.** *Brief Bioinform* 2008, **9**(4):286–298.
29. Kück P, Meusemann K: **FASconCAT: Convenient handling of data matrices.** *Mol Phylogenet Evol* 2010, **56:**1115–1118.

30. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**(22):4673–4680.
31. Stamatakis A: **RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models.** *Bioinformatics* 2006, **22**(21):2688–2690.

doi:10.1186/1471-2105-15-294
Cite this article as: Kück *et al.*: AliGROOVE – visualization of heterogeneous sequence divergence within multiple sequence alignments and detection of inflated branch support. *BMC Bioinformatics* 2014 **15**:294.

«Et tes amis seront bien étonnés de te voir rire en regardant le ciel. Alors tu leur diras: ›Qui, les étoiles, ça me fait toujours rire!‹
Et ils te croiront fou.»

"And your friends will be properly astonished to see you laughing as you look up at the sky! Then you will say to them, 'Yes, the stars always make me laugh.'
And they will think you are crazy."

„Und Deine Freunde werden sehr erstaunt sein, wenn sie sehen, dass Du den Himmel anblickst und lachst. Dann wirst Du ihnen sagen: 'Ja, die Sterne, die bringen mich immer zum lachen.'
Und sie werden Dich für verrückt halten.“

Antoine de Saint-Exupéry

Erklärung

Hiermit versichere ich an Eides statt, dass die vorgelegte Arbeit, abgesehen von den ausdrücklich bezeichneten Hilfsmitteln, persönlich, selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt wurde. Daten und Konzepte die aus anderen Quellen direkt oder indirekt übernommen wurden sind unter Angaben von Quellen kenntlich gemacht. Diese Arbeit hat in dieser oder ähnlichen Form keiner anderen Prüfungsbehörde vorgelegen und ich habe keine früheren Promotionsversuche unternommen. Für die Erstellung der vorliegenden Arbeit wurde keine fremde Hilfe, insbesondere keine entgeltliche Hilfe von Vermittlungs- bzw. Beratungsdiensten in Anspruch genommen.

Sandra Meid
Bonn, den 24.11.2014